COPENHAGEN BUSINESS SCHOOL

Master Thesis
MSc.IT

Machine Learning Applications for Assessing P2P Portfolio Default Risk

Author:
George Troicky

Supervisor:
Robert Kauffman

A thesis submitted in fulfilment of the requirements for the degree of MSc IT in

Business Administration and Information Systems (e-Business)
Copenhagen Business School

15th of May 2020
169,205 characters / 78 pages
Student ID: 124129

Table of Contents

Abstract

Crowd funding has contributed to a rapid transformation of the financial sector over the recent years. Although it is not a new business model, it has gained drastic and steady traction over the last few years. It provides alternative access to funds for individuals, companies, and entrepreneurial ventures (Crowdfunding Explained, 2017). A peer-to-peer(P2P) lending, as one source type of crowdfunding, has quickly taken market share in the consumer loan market, attracting attention from corporate investors, institutional customers, regulatory parties, and rating agencies.

In 2004, a UK based lending platform called Zopa, has generated approximately 5 billion GBP in cumulative loans issued in the last 15 years. With Morningstar and Moody's increasing their rating threshold, Zopa is the first P2P platform to achieve a AAA rating for their senior loans through a 245M GBP securitization program arranged by Deutsche Bank and partnered lenders (Krapf, 2018). With their claim of being able to originate high quality loans based on a proprietary model, it is not their first securitization of loans. Initially building a business models that diverted business away from conventional consumer lenders, P2P has grown to a size that is now converging with conventional institutional financing. With the growth a new sector raises new questions for methods on evaluating the sector, and more specifically the assets generated for financing on the secondary market. What value can machine learning (ML) provide to understanding securitized assets generated by P2P platforms? What variables are persistent drivers in predicting default rates? Will ML have predictive power in better understanding the factors of loan defaults from historical data? The research takes a positivistic approach in understanding the correlation of independent variables to the dependent variable. The results show that ML techniques can be applied with a disciplined approach to classify defaulted and completed loans. With the use of ML models, the drivers for predicting defaults included the sum of repaid interest and principle. This is consistent with industry standards, which project the highest rate of default for consumer loans falling within the first few months of a loan's term.

Key words: Marketplace lending, Peer-to-peer loans, securitization, digital economy, consumer lending

1 INTRODUCTION

Peer-to-peer (P2P) is a portion of the growing alternative finance market that has swiftly captured the attention of consumers, businesses, and banks in recent years. The new approach to funding has gained traction in markets around the world. Starting out as a platform to enable transactions for borrowing and lending between individual consumers has quickly turned into becoming a large player in the consumer finance sector. UK based lending platform, Zopa, began as a startup built on the idea that unprofessional investors with excess capital can generate income through lending to other individuals in need of cash. US companies sharing the same profile have also experienced rapid growth and transitions in the last decade, underwriting an abundance of loans and capturing a market that has caught the attention of many large investors. In addition to providing an intermediary platform that connects borrowers and lenders, their services include rating borrowers, and diversifying individual lenders' investments where risk can be averted. This intermediation provides a fluid, transparent, and standardized platform while products are customized to each investor and borrower. Not only have platforms *democratized* access to new investment opportunities, but access to an abundance of information and tools provides unprofessional investors the tools to assess investments on a professional level. A testament to its growth and market share in the UK consumer loan economy, makes it an interesting company to conduct a case study. Furthermore, recent news regarding the securitization of Zopa's loans showed their ability to originate high quality assets for the secondary market. These high-quality loans stem from the optimized approval process Zopa has in place, rejecting a high number of loans deemed to be too risky for its customers (the lenders).

1.1 P2P LENDING: GROWING MARKET

Between 2013 and 2016, the consumer-finance market rose from 10% to 13% in the UK[1](Murati, 2018). This growth rate in the consumer finance market in this same timeframe corresponds to Zopa's growth, where they have estimated to have generated a total of over 1 billion GBP by year 2015 solely in the UK market. Their steady growth, reaching over 5 billion GBP total loans since their conception, has gained attention from a variety of potential customers and investors alike. Specialists in the consumer finance market, such as Zopa, have been able to introduce cross-

[1] Outstanding consumer finance balance is measured as a share of the total GDP for corresponding years.

functionality of business functions, contrasting traditional business structures where IT, business, marketing, and risk are usually run in silos (Murati, 2018). The cross-functionality of a P2P platform ties in with the theory provided in the research, in which the phenomenon regarding platforms and their ability to quickly adapt to consumer needs, utilize powerful management systems, and generate insights to increase performance; therefore, making them a strong competitor to traditional firms(Murati, 2018) (Dhar, 2017) .

As there is an increasing demand for crowd funding platforms, innovations frequently require to be revised and improved by either their original developers/owners, or by new market entrants. The continuous improvements have led to significantly better performing platforms, products, and increasing credibility in the industry and the services provided by the platforms.

There is a wide selection of crowdfunding models such as equity, donations, rewards, debt securities, etc. Peer-to-peer (P2P) is a model that has been gaining traction where loans are verified quickly and administered in a simple fashion. The purpose of the P2P is to provide a platform by which buyers and sellers of credit can transfer capital between both parties via improved IT infrastructures of P2P platforms. The success of such platforms has attracted investors on the secondary market where large investors securitize loan portfolios entirely generated by P2P lenders. This has taken place in both the EU and US market (Alloway, 2013). Lending Club, the largest P2P lending platform in the US, had a cumulative 15.98 billion USD reported in 2015. Sharing congruent loan profiles and characteristics with Zopa, their returns ranged from 6.7% to 22.8%, whereas Zopa provides a narrower 10% - 18% expected return (Cohen et al., 2018).

Recent news regarding new arrangements to securitized 245m GBP of Zopa loans as of year-end 2019. These arrangements date back to 2018, in which, the framing is based on the initial securitization of a portfolio in 2016 (with maturity in 2024). Zopa is one of Europe’s largest consumer P2P lending platform, that is by measurement of cumulative amount lent. In order to broaden its product range and increase its customer proposition, Zopa has secured a banking license for the UK market.

Initially established as an unconventional model to circumvent institutional lenders, P2P lenders, such as Zopa, have turned back to these institutional lenders for funding. This new strategy

entails a 'controversial'[2] product of repackaging and bundling their assets for the secondary market – securitization (Alloway, 2013). Zopa, as well as other P2P lenders, has been able to increase returns to investors by providing faster process for securing loans and cheaper rates. Their optimization for issuing credit has increased their ability to reach a larger market in a shorter time-frame, and for smaller amounts, which is not something larger banks and lenders have the ability to do at an equally competitive price(Black, 2016). Their business model has also allowed for smaller scale investors to build their own portfolios and ultimately recreate the same strategies and construct portfolios which can be resold (Alloway, 2013) (Clemons et al., 2017). Overall, as P2P platforms mature, the sector can attract larger investors and via securitization they could decrease their funding costs while increasing funding and scale their operations (Weil et. al., 2015).

On a larger scale, given the vast amount of loans issued and overall transactions executed by Zopa, they have been able to accumulate asset pools that would interest many corporate and institutional investors looking for a return on investment (ROI) in a time when the market has been underperforming.

Just as Zopa has leveraged their IT infrastructure in mediating the convenient transfer of capital to consumers in the UK market, machine learning (ML) will be applied in analyzing a snapshot of the underlying assets generated by Zopa and their independent originators on their platform. This is the first step of the large undertaking of securitizing a company's assets and better understanding the performance of the income generated by the pool of assets. Via the application of supervised ML algorithms, Zopa's loan book, extracted in late 2019, will be separated using several classifiers to understand if there is a possibility to predict the outcome of the individual loans, and discuss the approach of generalizing these methods to the entire loan book. The predictive power of the algorithms will be tested via a two-tailed hypothesis test to understand if two sample populations are statistically significant.

The research herein, will take an inductive approach, meaning data will be collected and a potential conclusion will be drawn based on the aggregated data provided by Zopa. The case

---

[2] Securitization's process of repackaging illiquid assets for the secondary market means that static assets can be transformed into products for trading with external investors. This played a role during the 2007 global financial crisis (GFC), which resulted in (derivatives included) the underwriting of risky mortgages to be sold off as 'safe' products (Metz, 2016).

study is supported by Zopa's loan book with loans dating back to 2005, where approximately 56% have matured (known as 'completed' based on data set), approximately 37% are still active (issued between 2014 and 2019), and the remaining portion is between late and defaulted loans (a majority being defaulted). Over 600,000 loans are expected to be completed before the end of the 3rd quarter of 2019 that still have a status as 'active'. This snapshot of their assets provides a variety of loans at a low abstraction containing data that can be used to train the algorithms on loans that have been completed or defaulted.

1.2 Challenges for Secondary Market

Aggregate in an article published by Daniel Lanyon in 2019, P2P platforms have scaled their operations drastically in recent years, providing a gross new lending of over 6 billion GBP in 2018. This showed an approximate 20% increase from the previous year. With UK leading the industry in Europe with the most the largest market share and number of platforms, the article further states:

> *The peer-to-peer lending market is now funding more than 9 billion GBP of loans across Europe each year with [Approximately] 67% of this funding coming through UK platforms [as of 2019] (Lanyon, 2019).*

Previously in 2015, as the alternative UK finance market grew to 3.2 billion GBP, P2P platforms steadily became a conventional means for originating consumer loans. As these platforms grow and large investors accumulate loans originated by the platform, the same investors turn to securitizing the loans. Although securitization is not a new financial practice, it is usually applied to large asset pools as the fixed costs are high and require access to large sums of capital to initialize (Craughan et al., 2017). In the limited time that P2P platforms have been in the market, they have quickly accumulated market share at a lower cost base and with little to no regulatory capital requirements that traditional lenders are bound to. In the short run, securitized P2P loans have proven to be successful in providing liquidity to the large investors. The challenge lies in appropriately assessing the servicer risk associated P2P originated loans as traditional lenders have a longer historical track record than P2P lenders. Ensuring strict measures for assessing their performance and developing processes towards fully understanding credit risk, requires the involvement of several parties. Rating agencies are used in this aspect for applying their standardized methods in assessing these risks. The research herein will introduce an alternative method to analyzing a portfolio via projecting loan outcomes through ML classification to answer the following questions:

Do these ML methods increase the predictability of defaults of loans in an asset pool on a granular level?

Which models provide the most valid results?

Does this method provide an understanding for of what drives default in a consumer loan portfolio that has been originated by a P2P portfolio?

These questions are fundamental in analyzing the assets that may qualify for the secondary market and will carry through to the discussion where the context of the results will be analyzed. Given the infancy of the sector, an alternative approach towards extrapolating information can help future investors gauge opportunities in the P2P marketplace as there is a limited track record and available historical data. In applying ML algorithms, this research will explore the possibility for an alternative benchmarking method when rating an up-and-coming sector's performance.

1.3 Included Theory

With the emergence of the internet, businesses have become more interconnected than ever, growing on a global scale faster than ever. With this new interconnectedness, information is more accessible and overly abundant, even posing a challenge for many to understand how to pre-process the data and create meaning form it. Based on previous research in *Understanding the Information-Based transformation of Strategy and Society*, the literature provides a framework to better understand the success of a FinTech and the creation of a new currency, information (Clemons et al., 2017).

Along with specialized products, customized pricing, and a growing sharing economy, *FinTech Platforms and Strategy* examines the fundamentals of FinTech strategies and the components that make-up the sum of highly efficient and competitive start-ups (Dhar et al., 2017).

2 Topic Research: Literature review

2.1 Literature Overview

Section 4 will elaborate on relevant articles, journals, and research pertaining to the key components of the research. The purpose of this section is to highlight the research revolving around securitization, lending platforms, and the synergy that exist between the two in the

structure of three subsections: investment strategies, P2P FinTech platforms, and analytics for financial investments.

Articles published by major media outlets such as Financial Times and tech editorials, have been bringing attention to the fast growth of P2P lending companies and their ability to scale their business model to a degree that has allowed them to venture into new financial products with large financiers. Unconventional intermediaries that have optimized their ability to capitalize on micro transactions and reduce process cost, ultimately creating a valuable revenue stream. Furthermore, their fast scalability has attracted large corporate and institutional investors, ultimately realigning themselves with a business they had initially disrupted.

The following table entails the preliminary research that uncovered literature pertaining to the different perspectives of P2P lending platforms, ML methods, and the topics that were explored to connect the literature together (presented at the top of the table).

References	P2P securitization	Data mining / analysis methods & default risk	Driving forces behind default	Consumer loan - ABS	Leveraging ML for asset securitization
Data-Driven Investment Strategies for Peer-to-Peer Lending: A Case Study for Teaching Data Science		✓	✓	✓	
Notebook on Nbviewer		✓	✓		✓
P2P Lenders Turn to Securitisation Deals	✓				✓
Moody's: Demystifying Securitization for Unsecured Investors	✓		✓		
Cooperation and Competition in the US P2P Market	✓				
Prospectus: Deutsche Bank AG as Arranger	✓			✓	
Structuring a Marketplace Lending Platform Securitisation in Europe	✓				

The preliminary research narrowed the topic to three subcategories by which the following literature review is structured: investment strategies, P2P platforms, and analytics for investments. As the approach was narrowed some sources were excluded as they were not as

relevant for the research that follows. This includes a study published by an international UK based law firm, Hogan Lovells. It provided an additional vantage point to the complexities involved in securitizing assets generated on P2P lending platforms, as well as the opportunities.

A report published by Hogan Lovell, *Structuring a marketplace lending platform securitization in Europe*, provided a legal perspective to the securitization of P2P loans as well as the limits for the platforms given specific jurisdictions and the mechanisms of securitization (Craughan et al., 2017). With the UK alternative finance market growing to 3.2 billion GBP by year end 2015, P2P lending has not only captured the attention of large institutional investors but has become an integral part of the UK financial landscape. Jurisdictional measures must be understood for P2P loans as they can subject to different rules in the European market. An example includes P2P platforms in Germany, where the platform strictly acts as an intermediary between consumers and banks and the SPV, in a securitization structure, would be granted the loan receivables from a fully licensed bank rather than through a P2P platform. Additionally, there is the Simple, Transparent, and Standardized Securitization Proposal (STS), which sets the guidelines by which aim to ensure the obligation for disclosure of required investor information. Hogan Lovells' report comments on these standards "may encourage [P2P] platforms to promote and maintain high underwriting standards".

2.2 INVESTMENT STRATEGIES: REPACKAGING CONSUMER LOANS

As businesses grow, they turn to various methods of financing for a multitude of reasons, such as purchase of new income producing assets, repaying existing loans, increasing staff, or improving working capital. Securitization is known as a means of raising capital when trying to improve financial metrics and provide additional funding based on illiquid assets. The securitized assets of interest will be consumer loans, consisting of secured and unsecured loans. In context, this repackaging of an asset allows for investors to base risk and returns on the asset rather than the originating company. For example, a BBB rated business sells invoices to a AAA rated business. When the BBB rated business requires funding, they will be evaluated based on their current credit rating. By separating their rating from the risk tied to the individual asset, they are able to receive funding that is assessed solely on the risk of the asset rather than the originator (BBB business).

Consumer Laon market

Provided in a report by McKinsey & Co., challenger banks in Sweden alone, which includes FinTech's P2P lenders, accounted for 60% of the consumer finance market in 2016, a 200% increase from 2001. Although Major banks in the UK have only given up 30% of the market share in the recent years according to McKinsey' s estimates. Additionally, consumer finance revenues amongst European banks generated 56 billion EUR, placing second to payments in terms of returns for shareholders. Aside from stringent regulatory requirements imposed by the EU in the recent years, there have been significant loan loss impairment reductions, dropping to 130 BSP in 2017(compared to 210-220 BSP) (Murati et al., 2018).

2.2.1 Securitization

Based on a report by Global Credit Ratings (GCR), securitization will be introduced along with its structural features, advantages, associated risks, and its impact on the consumer loan originator. GCR's '101 to securitization' familiarizes the various terminology and continuously relate the results and data to practical mechanism of analyzing a portfolio that is to be securitized. Drawing on the graphical representations provided by GCR, similar illustration will be provided given the data from Zopa's public loan book and 2016 prospectus in order to give context to the terminology used in the analysis.

Securitization is the process by which assets are packaged into securities that are then resold to investors. The cash flows generated by the asset are then used to repay the investors for their provided funding. As a peer-to-peer lender acts as the intermediary of the assets, they do not retain the right to secure assets that are simply generated on their platforms. Alternatively, large investors, in this case being P2P Global Investment PLC(P2PGI), who own a large portion of the platform's loans can securitize their assets with the help of the platform on due diligence processes and as the underwriter to the generated loans. In the process, the investor(s) receives funding backed by the asset, giving rise to an asset-backed security (ABS) derived from consumer loans. Companies such as, P2PGI can then be publicly traded on the market as a security with a portion of their portfolio consisting of P2P loans (Deutsche, 2016).

History of Securitization

In a paper published by Stephen Quinn, from Texas Christian Universities Department of Economics, goes back to the late 17th century. The paper showed how the Bank of England, South Sea Company, and East India Company came to own 80% of Britain's national debt in a span of

30 years. The debt restructuring turned Great Britain's national debt from 'a poorly coordinated, heterogeneous, illiquid and expensive pool of funds into a modern-style national debt' (Quinn, 2016). Almost 200 years later, in the 1970's, the process of pooling assets for the secondary market began with mortgage loans that were guaranteed by government agencies. Approximately a decade later, 1.2 billion USD in ABS were issued on the long-term securitization market in the US. Shortly thereafter:

> *Since that time, the ABS market in the US has grown dramatically to $280 billion in new issuance in 2001 with about $350 billion anticipated for 2001. While the US market still accounts for the largest share of the global securitization market, it is a maturing sector and its growth rate has slowed considerably compared to the markets in Europe and Asia. This trend is expected to continue for the next several years (Parker et al., 2003). [corresponding table has been appended as figure 8]*

Mechanisms of Securitization

A special purpose vehicle or entity (SPV) is established in order to separate the risk associated with the originator of the asset. The SPV 's only purpose is to purchase the assets from the originator and then issues ABS to its investors. Via this 'absolute transfer' or 'true sale' of the assets to the SPV, the assets are 'de-linked' from the originator and credit risk is no longer associated with the originator, but rather the assets themselves. Additionally, this is a legal separation, which means if the originator files for bankruptcy, creditors to the originator would not have a claim to the assets or the cash flows generated by the assets. This mechanism applies both ways, as the investors in the ABS do not have a claim against the estate of the originator.

Credit Enhancement

Reiterating to the associated credit risk, credit enhancement is an important step in protecting investors from incurring loss on their ABS when losses occur with the underlying securitized asset. For the securitization of P2PGI's assets, credit enhancement took place in the form of subordination of junior ranking notes, cash reserves, liquidity reserve, and excess available interest proceeds (Deutsche Bank AG, 2016).

The subordination of the loans, also known as the capital structure, carves out the assets based on their risk. These levels in the capital structure are known as tranches, where the most senior tranche would have the first claim to the cash flows and preceding junior notes would follow

based on the arranged capital structure. Based on the associated risk, there is a congruent expected return on the note. The higher the risk, the lower the note sits in the capital structure, and the higher the return for that note. Based on the 2016 prospectus, a capital structure is provided with the following note classes structured in a hierarchy from top to bottom:

Notes Class	Initial Principle Amount (GBP)	Relevant Margin	Ratings(fitch/Moody's)
A	114,000,000	1.45%	AA-(sf)/ Aa3(sf)
B	7,500,000	2.90%	A(sf)/ A2(sf)
C	7,500,000	4.00%	BBB+(sf)/ Baa2(sf)
D	9,000,000	7.00%	BB (sf) / Ba3(sf)
Z	12,144,000	Variable interest amount	Unrated

This type of subordination also provides credit enhancement in the sense that losses from the underlying asset are absorbed firstly by the Z, D, C, and B notes before A.

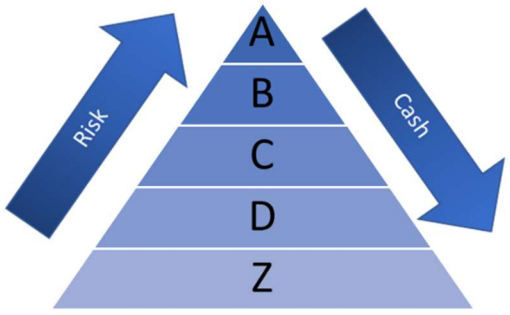


Figure 1 Absorption of risk versus dispersal of cash (Markovitz, 2019)

Reserves may be set aside to absorb any losses that transpire during the course of a portfolio's performance. The reserves are made up of cash set aside by the originator as a portion of the funding. The reserves can also be configured by excess cash flows generated by the asset to replenish a reserve to its specific requirement. If reserves are used to 'absorb' any losses from the underperformance of an asset, it may result in a limit to the funding provided to the originator until the reserves return to their specified level. Seen below is an illustration representing the allocation of risk and cash to the notes in a capital structure:

Additional to reserves being taken, there is also excess available interest proceeds stated in the prospectus. This pertains to any cash that is left over as a credit in the accounts structured in the

transaction that, on each Note Payment Date[3], will be paid out to stated parties as agreed and places the senior note fourth in line to the initial payment of taxes, fees, expenses, and other administrative costs in the structure(Parker et al., 2003).

Provided below is a simplified transaction structure for the securitization that took place starting in 2016 and maturing in 2024.

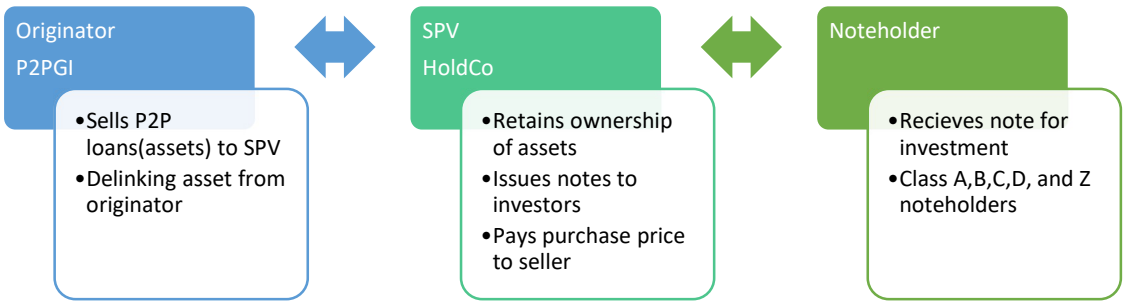


Figure 2 Simplified securitization transaction structure

This structured form of disintermediation is what allows businesses to raise capital from the conventional, on-balance sheet debt financing. In the true-sale of the loans to a bankruptcy remote entity, the parties have achieved a delinking where the funding is provided solely based on the credit risk of the assets in the portfolio (Parker et al., 2003). This generally provides funding at a lower cost in comparison to traditional debt financing when business take out a secured loan from a financial institution, collateralized by the borrowers' assets. Under conditions where the transfer receives off-balance-sheet(OBS) treatment, it may be subject to lower regulatory capital requirements for financing institutions, and from a company's perspective, OBS accounting practices allow for the increase funding backed by its on-sold assets while keeping leverage ratios low.

As this securitization deal is based on a set portfolio with maturing loans from Zopa's P2P platform, it would be characterized as an amortization structure. The portfolio will not be replenishing and principle along with interest will be repaid with a defined end date, contrary to a revolving period structure (Markovitz, 2019).

[3] "Note Payment Date" means the First Note Payment Date and, thereafter, the 20th day of each calendar month provided that if any Note Payment Date would otherwise fall on a day which is not a Business Day, it shall be postponed to the next day that is a Business Day (Deutsche Bank AG, 2016)

Secondary Market

Via the secondary market, investors can sell their loans to other investors and depending on the demand, the liquidity of the asset can fluctuate. This ultimately allows investors in a P2P platform, if all criteria are met, to liquidate their assets in return for cash that can be used for alternative investments. For example, a company like P2PGI can raise capital backed by their P2P asset portfolio at a relatively low cost for further investments in consumer loans. A conventional example is the Federal National Mortgage Association (FNMA) whose goal is to securitize mortgages that provides lenders with capital to reinvest and expand the secondary mortgage market through issuance of mortgage-backed securities (MBS) (DeGrave, 2016).

2.3 P2P FinTech Platforms

P2P lending platforms are categorized as a debt crowdfunding platform. P2P lending also goes by market-based lending and lending-based crowdfunding. The loans issued on these platforms are generally unsecured and the platforms mitigate default risk via disbursing a loan across multiple lenders and repackaging the loans in portfolio pools (Jenik, 2013). These pools will generally consist of various loan profiles that are catered to match the risk profile and expected returns of the investor. These platforms have not been limited to personal lenders and borrowers but have extended to businesses and small and medium enterprises (SME) raising capital from personal lenders, and business providing capital to other business.

As cited in the working paper, debt crowdfunding," is estimated to have raised 3.6 billion EUR [in Europe], [and] 909 million pounds in the UK [in 2015]". The value in the UK has gone up from 547 million pounds from the previous year and continuously growing.

Representative of the risk associated with lending on a platform predominantly made up of unsecured loans, P2P platforms offer quicker and easier access to both lenders and investors looking for a higher return. Further citing Kriby and Warner, the diversification of the portfolios and ability to partition assets helps mitigate systematic risk (Jenik, 2013).

These deals generally take a long time to construct with the addition of incorporating more than one financier for loan syndication, risk hedging, and formal rating agency assessments of the asset portfolios to-be securitized for secondary market investors.

The transition of a FinTech's for a new market is summarized by a Financial Times article as the following:

> *The move towards securitization highlights a shift in the growing P2P industry – even as they eschew traditional banking the biggest P2P lenders have been increasingly supported by Wall Street and large institutional investors (Alloway, 2013).*

The article also mentions that the recently expanding P2P lending platforms have been recruiting experienced individuals from the banking sector for developing more efficient systems of analysis and creating competitive processes. The improvements to the processes are reflected in congruently competitive prices and products.

With the increasing growth of in the FinTech sector, a less recent report publishes by McKinsey & Company in 2016 showed an increase of 205% increase in venture-capital investments from 2013 to 2014(Dietz, 2016). It outlines the six markers of success for the 'FinTech Attackers' as stated in the report: advantaged modes of customer acquisition, step-function reduction in cost to serve, innovative uses of data, segment-specific propositions, leveraging existing infrastructure, and managing risk for regulatory stakeholders.

The first marker is identified as a FinTech's ability to acquire new customers while keep costs low. In the past this has worked through partnerships with established firms, which gave exposure to the growing start-up as it was able to integrate with reputable systems and platforms.

Tying in with lack of existing overhead costs incurred by FinTech startups, they retain a cost margin advantage to conventional business and institutions, allowing them to provide a product at a more competitive price.

The next marker emphasizes the importance of data as it allows for tracking and projection of outcomes, dependent on various attributes provided to the analysis. With over 90% of the data in the world having been created in the past few years, the ability to sift through and create value from the data is vital for growing organizations.

In targeting specific groups and products, a successful FinTech focuses on a single aspect of the banking sector where the price-sensitive customer makes up a large portion of the market and is open to remote solutions that a large institution may lack the capabilities to provide.

The fifth marker relates to the idea that there is no need to 'recreate the wheel'. This recounts the point of 'advantaged modes of customer acquisition' but with a focus on infrastructural developments. FinTech's can leverage existing platforms and establishments, where both firms can create a symbiotic relationship to improve their customer proposition.

Finally, a FinTech's ability to conform to regulatory requirements as they scale-up in a heavily regulated market. With different market areas imposing different restrictions, this is a deciding factor for a FinTech that plans on expanding in different market areas.

In *Understanding the Information-Based Transformation of Strategy and Society*, information has become the currency of today's economy adding monetary value to a firm that has the capabilities to leverage their systems to better understand the immense amount of data (Clemons et al., 2017). As information is the "glue that binds economic activities together", FinTech platforms play a vital economic role as they aggregate data and decrease their costs in developing infinitely customizable products for their users. As it was also previously mentioned in the McKinsey & Company report, a FinTech that partners with established institutions gains more leverage in market competition due the additional access gained to data as well as the exposure they gain through the partnership and potential access to more resources.

2.4 Previous Methods for Analysis

Previous research, pertaining to two of the largest US based P2P lending platforms, introduces portfolio optimization using machine learning techniques in a case study. The research takes the perspective of the individual investor who is looking to understand the metrics used in a comparative analysis. The analysis parallels that of this research as its objective is to project defaults among existing loans created in a P2P platform. Cohen, et al., extends their research to projecting the risk and return on investment.

The case study uses data mining and machine learning algorithms to understand patterns from which certain conclusions were drawn. It outlines the generic classification methods used by the P2P platforms and aims to understand the risks associated with the loans.

The variables Cohen used, related to the attributes of the borrows and the loan profiles. The research is strongly rooted in tying in data science methods to a business case; it further abridges technology, IT, e-business, and investment strategies. The results from the case study utilized machine learning to obtain a return on investment of 3%. It further discussed a potential increase in improving the return to 3.21% via the incorporation of optimization.

The strongest methods in practice trained 2 data sets, defaults and completed loans. Based on the following criteria, the split the models, predicting default and non-defaulted loans, where joined

in classified via random forest. This method had a relative 7% increase in improvements and returns on investment of 3.21%.

Ultimately, the research provided by Cohen used machine learning to optimize return on investment, its focus was aimed at identifying the optimal return given a set of P2P loans form a data set. Although this model requires extensive fine tuning for the benefit of an individual's portfolio performance, it provides extensive background on establishing a base case in fine tuning a model's predictive strength and performance. This correlates to supplementary research pertaining specifically to machine learning algorithms devised for machine learning strategies.

With the inherently vast amount of information provided by P2P lending platforms in a debt sharing economy, proposing a "new and more accurate credit risk models to protect consumers and preserve financial stability… [where they see to]… enhance credit risk accuracy of peer-to-peer platforms by leveraging topological information" (Giudici et al., 2019). With the development of modeling software and abundance of data, what is interesting to note here is the depth in which one can go to in aggregating relationships, patterns, and conclusions from a vast amount of data. Such models taking a topological approach could include k-nearest neighbor. In this approach, the data is formed in clusters and segregated into groups based on the provided independent variables. This method of clustering results is established by a parameter such as Euclidian distance, also known as the Pythagorean theorem. Using the coordinates of a two-dimensional plane, the distance between two points on a set plane would constitute the cluster to which they belong. Along with using such a method for identifying riskier borrowers, this is also a method used by e-commerce sites and streaming services to pair user results with one and other.

An alternative study conducted on comparing defaulted loans and their comparative FICO score (Emekter, 2014). The main drivers for borrowers being attracted to the P2P lending platform was to consolidate debt, consisting of 54% of Lending Clubs asset pool. The research continues to mention that one of the additional reasons was the rates were substantially lower and borrowers were able to borrow amounts that credit cards would otherwise not have provided. The second purpose for borrowing was due to paying home mortgages or home remodeling/repairs. This class of borrowers only amounted to 7% of the outstanding amount. In comparison the 47 million USD borrowed for home repairs is a fraction of the 387 million USD borrowed by individuals looking to consolidate debt with additional debt (Emekter, 2014).

The data used Emekter et al. study applied micro-economic variables that were then analyzed via a logistic regression model and the results showing key drivers for default being credit score, income-to-debt ratio, and credit grade. The model was split between two dependent variables,

defaulted or not, and setup as binary. Loans that were either defaulted or in arrears were categorized in one group while the other included borrowers who were up to date or completed their payments in full. With the initial conjecture that riskier borrowers would most likely default on their loans was proven via the provided model (Emekter, 2014).

Decisions trees, J48, and random forest are additional methods that have been used in recent studies as they are recursive and rely upon splitting variables with the smallest entropy, all of which will be explained in section titled Machine Learning. The case study conducted by Cohen et.al, on Lending Clubs loans, random forest, naïve Bayes, and logistic regression were used as these are common methods among researchers in machine learning.

2.4.1 Data Mining

Data mining is understood as the general concept of aggregating data effectively and using metrics to better summarize results that are found. Provided below is a simplified description what data mining entails:

> *Data mining is the process of sorting through large data sets to identify and establish relationships to solve problems through analysis using various computational tools (Interface Technologies, 2018)*

When sifting through and mining the data, it is important to look back and see quantity of data provided that can support the results, as well as the number of times that the output is accurate. This term of accuracy should not be confused with the accuracy in machine learning models as it is a generalized term in this case. As will be explained in Machine Learning, accuracy, precision, and recall are separate metrics used to determine the predictive strength of a machine learning model. Data mining can consist of sequencing, clustering, or classifying data all of which can be done without machine learning algorithms but are the underlying principles that which ML algorithms are based on.

Missing values can also lead to inconsistencies with a model and it is important handle the data in the preprocessing stage. There are several techniques that can be applied to smoothing out the inconsistencies in the data during the preprocessing stage (Kampf,2016). Depending on how large the data set is and the small portion of values that might have a missing attribute, implementing a filter to ignore the instance all together is one method.

Alternatively, a constant can be incorporated for the missing values if removing the attribute or instance is not feasible. One type of constant can include ‘N/A’ or 0, as the value does not apply.

Using a mean or median can also be incorporated as constant value. For example, a data set pertaining to a residential data might have instances were some income values are missing and an average can be introduced. Furthermore, the constant mean can be more than one value. Referring to a residential data set, an average income can be applied to a class of instances which could include four different averages among the entire data set where the income is based on which region the individual is from(Kampf,2016).

2.4.2 Machine Learning

Prior to explaining the data processing that took place, this section will outline the methods and applicable terminology that was previously described under section 4.4 in more detail. This will draw on tangential examples and aim to provide context to the applied methods and their results. Furthermore, it will provide a better understanding when analyzing the various metrics applied to models (and their outputs) and attributes in relation to the data (the preprocessed inputs). This section will also explain key identifiers of a model's strength and predictive abilities when tested with external data sets. Additional to the metrics, the following will clarify the difference between supervised and unsupervised ML and the various case in which models under both categories can be applied to.

Supervised Machine Learning

In supervised machine learning there are various ways of going about classifying the data. One method can include a regression model, and another would be classification. In a regression model, the objective is to plot the inputs provided to a continuous output and a common algorithm would include logistic regression (can be both discrete and continuous). For classification, common methods used include naïve Bayes, decision trees, and neural networks. The objective with classifiers relies on plotting the introduced date to a discrete output (Soni, 2019).

An additional point to supervised machine learning is understanding the trade-off between bias and variance. These two parameters generally have a negative correlation. A model that is consistent wrong would prove to have a high bias but low variability in errors based on different training sets. Alternatively, If the errors in a model provided different data sets variability in the output's errors, it would be known to have a low bias and high variance (Soni, 2019).

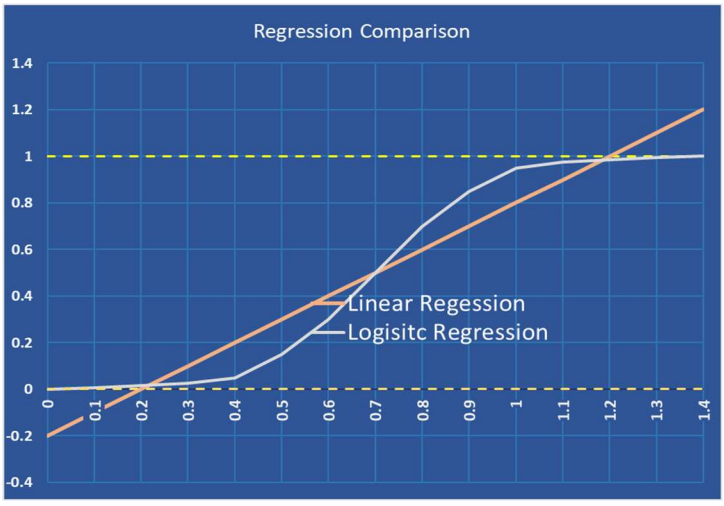
As the models will be trained towards predicting the final loan status it is a discrete value output that would be plotted by the model. The 'loan status' class of discrete values consists of 'late',

'default', 'active', and 'completed'. Previous research has transposed a larger set of values in a binomial model, categorizing 'late' and 'default' into one group, and 'completed' into another. In the diagnostics section of this research, a base case will be trained by a multinomial and binomial classification. An additional steady sample of 100 loans with known status will be tested in the model.

Setting Baselines

ZeroR, also known as Zero Rule, is a base line classifier used in benchmarking classification algorithms. It is the simplest classifier compared to naïve Bayes, decision trees, etc. It is not a linear nor logistic regression model but calculates the most frequent output in the data set. For example, if defaults appear 80% of the time, the model will presume all instances are default and would be correct 80% of the time. This same benchmark can be used to against other classifiers. For instance, if an algorithm such as naive Bayes correctly predicts the instances less than 80% of the time, it could be discerned as not a viable model (Egnelschall, Applied Technology Research).

Logistic Regression as a Classifier



Logistic regression is a classification when incorporated into ML and release on the discrete input values to provide an output. A regression model constructs a line to which a probability is assigned for a variable belonging to that category, being the output. The difference between linear and logistic regression is that the output has constraints in a logistic regression while a linear regression can surpass the bounds. This is represented as a sigmoid as it is compared to a regression line in the image to the right (pant, 2019) (Gupta, 2019).

As the hypothesis test states that the dependent variable must be with a bound, a linear model would not suffice appropriate classification of a data set that contained non-discrete values. A hypothesis expectation for a regression model can be generalized by the following, $0 \leq h_\theta(x) \leq 0$.

$$y = \beta_0 + \sum_{i=1}^{n} \beta_i X_i$$

$$Sigmoid\ Function = \frac{1}{1 + e^{-y}}$$

The function above depicts a linear regression model, where the beta at 0 is the y-intercept and the variables after display the sum of the population slope coefficient multiplied by the independent variable(input/attribute) (Pant, 2019). The sigmoid function incorporates limits and based on the drawing above can be understood as such: $\lim_{x \to \infty} (sigmoidFunction) = 1.$ Substituting the linear regression for the *y* variable in the sigmoid function, a linear model assumes the limits of the logit function a logit function which limits the dependent variable to only two possible outcome and their probability.

"Logistic regression becomes a classification technique only when a decision threshold is brought into the picture" (Gupta, 2019). *Network Based Scoring Models to Improve Credit Risk Management in P2P Lending Platforms*, uses various model in increasing the accuracy for P2P customer risk scoring and states that, "logistic regression model is one of the most widely used method for credit scoring"(Guidici et al., 2019).

Naïve Bayes Classifier

Naïve Bayes is a simple Bayesian network that has the initial assumption that attributes are independent, meaning that any one attribute in the data is unrelated to another. This is a model that, although simple, tends to have comparable performance to other classifiers as it is a robust model and is not easily screwed by outliers and increase in data sample sizes (Soni, 2019). The robustness translates to overfitting, which is where the classification of instances matches to closely to the overall data and would not by useful making generalization about the relationships in the data. An example of underfitting can also be present, which would relate to a linear model applied to a non-linear data set and ultimately have little predicative ability.

The model is used to calculate the maximum posterior probability($P_{(x|c)}$ as shown below):

$$P_{(c|x)} = \frac{P_{(x|c)}\, P_c}{P_x}$$

In using this theorem, the model predicts the probabilities of the attributes in relation to the outcome. In context, the outcome would either be defaulted or not defaulted loans. The denominator would be a constant value and can be introduce as a proportionality as follows:

$$P_{(c|x)} \, \alpha \, P_c \prod_{i=1}^{n} P_{(x|c)}$$

The value for the P_x in this case equates to the attributes that will be used to create a probability for the outcome. Provided by WEKA, this formula can be adjusted to cope with multivariable outputs which is known as multinomial naïve Bayes classifier type.

Decision Trees

Decision tress are based on a method of splitting instances via decision nodes where the algorithm decides on the driver with the highest probability to split an instance based on each attribute for each node that is created. At the top of this structure is the root node containing the entire population and the strongest split in the sample which continues branching out into a tree-like structure. When the algorithm decides that it can no longer split on the sample, it creates a terminal node, also known as the leaf node (Soni, 2019).

These models are built on the concept of information gain which is based on calculating the entropy of the split in the data set. As an example, if the model splits on a data set of eight instances in the following manner, 3:5 and 2:6, the first split would have an entropy of approximately 0.9544 and the second would be 0.8113. The first split has a higher value meaning it was more capable discerning that the sample belonged to one group over the other. When the value for entropy is close to or equal to .5, this means that the decision is as good as a coin toss and the data has yet to be classified further.

The following formula can be used in Excel; =-((A/(A+B))*LOG((A/(A+B)),2))-((B/(B+A))*LOG((B/(B+A)),2)), where A denotes the first sample and B denotes the second. Entropy formula provided below:

$$Entropy \; = -\left(\frac{A}{A+B} * \log_2\left(\frac{A}{A+B}\right)\right) - \left(\frac{B}{A+B} * \log_2\left(\frac{B}{A+B}\right)\right)$$

J48 is an algorithm that is based on an earlier developed classification of C4.5. The C4.5 is also an improved version of the ID3, where it is capable of handling both continuous and discrete attributes. A threshold is determined automatically where the list is split for attributes lying

outside the constraints. The model is recursive, meaning the reiterates in finding the best split on a given attribute until it has either reached a leaf node, or more likely in large samples with multiple attributes, it has reached its next two decision nodes for another split(Jain, 2017). Towards Data Science simplifies the C4.5 algorithm via the following pseudo code:

1. Check for the above base cases.
2. For each attribute a, find the normalized information gain ratio from splitting on a.
3. Let a_best be the attribute with the highest normalized information gain.
4. Create a decision node that splits on a_best.
5. Recur on the sublists obtained by splitting on a_best and add those nodes as children of node.

Random forest is another method which is derived from the basic decision tree but creates multiple decision tress that it then creates multiple decision trees, each one based on a random subset of features. The average is then pooled, and an estimate is taken from all trees (Koehrsen, 2017).

Unsupervised

Unsupervised machine learning is more commonly used in exploratory analysis and in dimensionality reduction. Dimensionality reduction is a method that pertains to the precisely what it is called, reducing variables (also known as dimensions) in each input to construct a simpler model via this reduction. The alternative method would be for exploratory data analysis (EDA), as it is intended to automatically structure the data, either in clusters or segments (Soni, 2019). An exploratory approach is effective for instances when the provided raw data does not have a concrete output and the user simply wants to gain insight at a higher abstraction. Receiving a large data sample of consumers for a certain retail company might initially undergo a clustering method as this might draw relationships between individual consumers and create segments to which individuals would be classified to (Soni, 2019).

Some common models mentioned by Towards Data Science that are used for unsupervised ML include neural networks, principle component analysis, and k-means clustering. In the following research, unsupervised machine learning will not be applied as categories that are of interest are known. In a case where demographic data was available, clustering borrowers into groups to better understand the relationship between the groups would make good use of these unsupervised methods.

2.4.3 Industry standards: Moody's Methods

In a closed request for comment(RFC) proposal on updating Moody's approach to rating consumer loan-backed ABS, they propose minor changes to their sovereign risk analysis to secured consumer loans for two benchmarking methods: one method references their Idealized Expected Loss((IEL) Table for Standard Asymmetric Range, and the other uses the Symmetric Range for the IEL. Based on a standardized table set out for rating the assets on the loss performance, the two approaches consider the historical performance and categorize them using logarithmic scales in setting the bounds as per the IEL table (Krapf,2018).

For an asset class that are strongly linked to the credit rating of the underlying asset and/or entity, the method of Symmetric Range is applied to rating the asset. In this approach, the asset assumes the properties of the rating initially applied to the asset. The lower bound rating is the lowest IEL for a rating, while the upper bound is the highest IEL for the same rating.

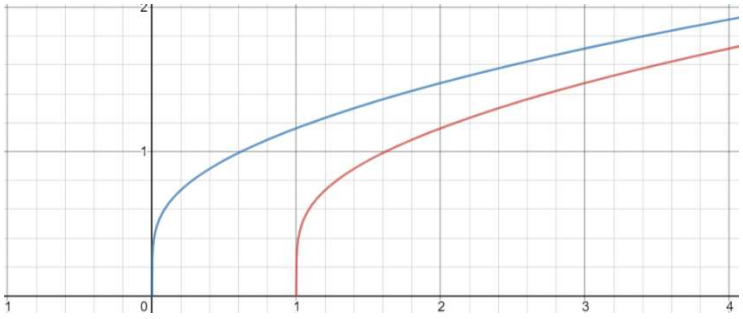


Figure 3 Moody's Approach 1: Symmetric Range Bounds

The red line in figure 3 depicts a general representation of the function of the lower bound of a rating and the upper bound of a rating is represented by the blue line. As seen in the x- and z-axis, the values are not to scale in accordance with the IEL and is strictly used as a visual representation for the designated values to each rating. This method for scaling the graphs is also used in the explanation for approaches two and three.

As stated in the report, this method of Symmetric Range is applied to rating an asset class that is strongly linked to the credit rating of the underlying asset and/or entity. In this approach, the asset assumes the properties of the rating initially applied to the asset.

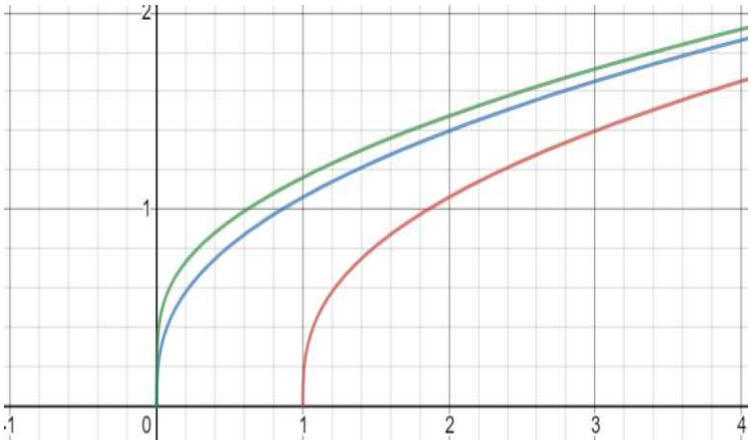


Figure 4 Moody's Approach 2: Standard Asymmetric Range

Depicted in figure 4 is the change in the function with a weight distribution of 80/20 which is an adjustment from the previous 50/50 split in the second approach. This new method reduces the range for an asset's classification. For example, a given year's loss performance being close to the upper rating bound from the Symmetric Range Approach, this new method would classify that asset pool in a lower rating.

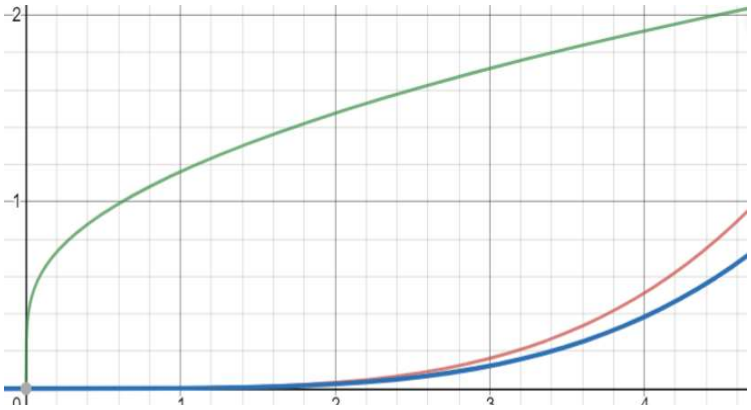


Figure 5 Moody's Approach 3: Wide Asymmetric Approach

As the equations in Moody's report does not list the proposed bounds in a logarithmic function, the blue and red lines in figure 5 strictly show an approximated visual representation of the contents in the IEL table.

The following bounds were presented in the report for the Wide Asymmetric Approach:

1) Rating Lower Bound$_r$ = IEL$_{r-1}$
2) Initial Rating Upper Bound$_r$ = IEL$_r$

These bounds were recalculated to show the relationship to the new method via logarithmic function:

1) Rating Lower Bound$_r$ = *exp* (log (IEL$_{r-1}$))
2) Initial Rating Upper Bound$_r$ = *exp* (log (IELr))

This showed a graph that was comparable to that of the second approach but had a reduced threshold for losses incurred over the duration of issued loans from a particular year.

The two models were built using the IEL Table from Moody's, one using the existing formulas provided by Moody's in the Wide Symmetric Approach, and the other converts the upper and lower bounds to a logarithmic function. The purpose for using the logarithmic functions is that it shows the change in the percentage and normalizes any outliers/skewness in the data. Since the graphs are show 2 groups of ratings, one from the higher end of the IEL and one from the lower end, one can see the correlation of the bounds in a stacked relationship. As ratings between Aa2 and Baa1 were excluded from the model, this is the explanation to the gap seen in the "Wide Approach Applied" model in the appendix.

2.4.4 Defaults

Moody's report states that the highest concentration of loans that default can be seen in the short term of a vintage's portfolio. Given a certain year, this number fluctuates but the highest rate of increase is predominantly seen in the first 12-24 months of a loan and a gradual decrease later in the portfolio's life. This rate of default over the course of a loan portfolio is demonstrated via a default timing curve, showing whether it is a front end of back end profile. A back-end profile is where the spike in defaults is seen earlier in the life of the loans to a base case. The contrary statement pertains to a font end default timing curve. Additionally, recoveries tend to be higher for loans deemed default earlier in their life rather than loans with an extended delinquency period.

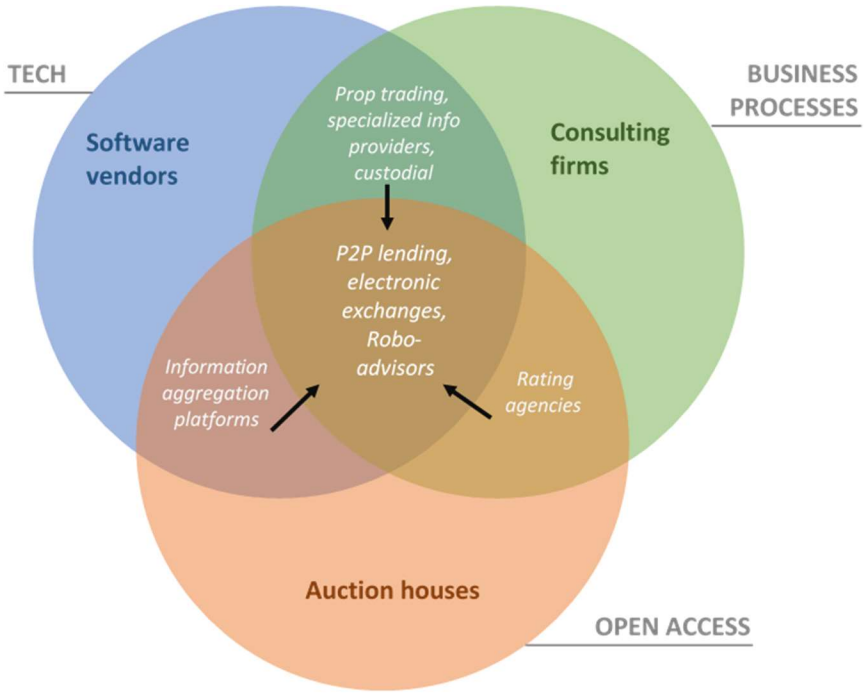
When assessing an asset pool's default variability, Moody's typically examines the servicers ability to apply consistent practices to servicing loans in arrear and in default. Historical loss performance can be used to illustrate a servicers ability to maintain performance on its loans over their lifetime. Historical loss performance graphs have been provided in section 6.1.2 to show the accumulated losses that have been incurred in each vintage in Zopa and compared to the data provided in the prospectus. As the non-extrapolated data is based on current historical data, it is limited to projecting months in advance depending on the year that is being presented. For example, loans from last year may only show accumulated defaults until the current data and would be cut-off at month 18.

3 THEORIES IN FINTECH

3.1 PLATFORM PHENOMENON: FINTECH STRATEGIES

The emergences of new business's and business models in the last few decades has been enabled by the emergence of the internet and connectivity. This new phenomenon has been denoted as the "*platform*" business (Dhar, 2017). Dhar and Stein outline the following attributes as essential components to a FinTech platforms:

1. *They are open; allowing easy participants;*
2. *They implement key business and operational processes, some of which typically exhibit network effects that increase in value as participation increases; and*
3. *They implement these business processes automatically using enabling technology*



The components have been organized in the following Venn-diagram with corresponding business models that fall within and their cross-over to other with other segments. The firms

along the outside denote incomplete models such as consulting firms lacking access to aggregate data, auction house lacking specialized information, and so forth.

Existing firms can adapt to competitive platforms that pose a strategic advantage in their positioning between technology, accessibility, and process implantation. This may include an investment search platform introducing the option for trading and ultimately leveraging its current capabilities to provide a fuller process for its users. This approach to increase one's current model, is considered "*platform completion*" and the firm would strive to position itself at the center of the diagram.

An aspect of this theory that is interesting, is "*component replacement*". This is where platform completion cannot compete with incoming platforms and cannot meet its customers' new expectations regarding price and value for service. This is ultimately because incoming disrupters have specific competencies in a niche segment of the business where established firms lack resources in developing. Relating to the previous example of a platform that transitions from strictly providing investment information to providing its users with the ability to trade. With 'component replacement', this may entail that the search platform has not only facilitated a fuller process but can complete trades and provide information faster due to a change in the IT infrastructure. With the advancement to the process, the newly established trading platform has elevated the benchmark for performance expectations.

The introduction of more efficient processes that leverage technology are a component that is separating platforms from conventional business models. These components can include software that able to capture, aggregate, and compute information so that the firm can increase its own value generating strategies.

P2P lending platforms have taken advantage of the ability to streamline lending process and created value in the consumer loan market at a competitively low cost that a conventional financial institution could not. As the financial sector advances, more partnerships and mergers will transpire, combining each firm's competencies and resources to create competitive synergies.

3.2 Evolving Economies

With the connectivity of the world economy providing abundant information not only available to business, but individuals alike. This new age of information has shifted businesses to rethink their strategies in creating a competitive advantage. The framework in *Understanding the Information-Based Transformation of Strategy and Society* outlines an optimistic approach to future

opportunities as information-based strategies transform (Clemons et al., 2017). Improvements in process efficiencies and automated processes in pricing has made it possible for numerous individuals to participate either individually or collaboratively in business models that were once only available to professionals. Reiterating a previous point about information being the glue that binds economies together, the digital age has shown the creation of this newly monetized, and hyperabundant commodity.

3.2.1 Customized Pricing

Historically, the optimization of processes, reduction in cost, and expedience in product development has led to businesses increasing their value proposition to customers as they become more capable of catering to customer needs. Leveraging IT systems, firms can provide hyper-differentiated products at a fractional cost. Lending platforms consider personal, as well as generalized information to aggregate the final interest/price of the loan dependent on the risk of the potential borrower. Using additional data sources on a continuous basis allows for increased automation of underwriting and servicing loans.

3.2.2 Rise of sharing economy

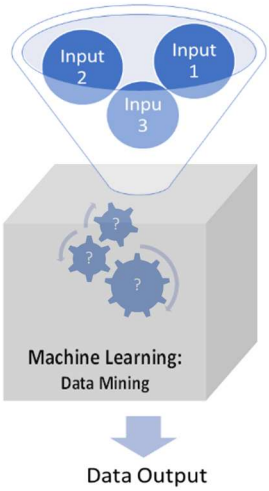
In relation to lending platforms, the ability to hyper-differentiate between various borrows has allowed for the platforms to provide customized credit analysis and develop proprietary systems with the ability to handle assessments on a granular level. In connecting lenders and borrowers, the platform can partition loans, and corresponding risk, across multiple lenders and investors. Lending platforms can disperse the risk to an investor by reducing the amount lent. Ultimately, this allows "consumers to adjust their consumption to their actual needs and increase the economic usefulness of their assets". In turn, this incentivizes the lenders to use Zopa while the platforms diversify their risk and reward in providing them a risk averse portfolio, never incorporating more than 1% of an individual's portfolio to a high-risk investment.

3.2.3 Democratization by access

"This implies a thicker tail for the scope of consumers' demand interests and product offerings in the market". This concept intertwines with the previous point of a growing sharing economy, where unprofessional individuals now have access to platforms that allow them to enter a market only professional business participated in.

3.3 BLACK BOX PHENOMENON

The black box phenomenon is understood as the lack of understanding around the cognitive mechanism in understanding the outputs generated by an algorithm given a set of inputs (Villani, 2018). With the advancement of ML, there comes the challenge of understanding how it is that decision is made by that particular machine. There is contradicting interest in this field as our interests in improving efficiency meets the challenge of being to explain how it is that a conclusion was drawn. There is a variety of models used in data mining where some are more



explainable than other, but as ML develops into more sophisticated models such as neural networks, the understanding becomes limited *(Villani, 2018, p.114-115).* This concept can be illustrated by the diagram below:

Algorithmic Discrimination: are machines prejudice?

The concept of discrimination among machine learning algorithms has been in question in numerous use-cases such as the legal and financial sector. Although personal information will not be incorporated into the methodology, it is important to understand the prejudices that may be ingrained in the algorithm's classification and selection process. Based on a study in 2019, price discrimination seemed to be more prevalent among face-to-face interactions. While prejudice existed among FinTech algorithms, it showed a 40% reduction in rejections, and when considering applicants, could also increase a loss in revenue from face-to-face rejections (Bartlett, 2019).

4 CASE STUDY & DATA

Section 4 will introduce Zopa's business process, a historical outlook of their loans, and the additional data incorporated from the UK postal codes. The business process includes an introduction of Zopa's growth and an overview of the lending process as per the prospectus. The historical data is going to provide further context to support the claims of Zopa on the basis of the loan book. The data includes some minor preprocessing for a better understanding of the Zopa's portfolio in comparison to industry standards and methods used in the prospectus. Section 4 will conclude with explaining the UK data that was incorporated into the sample set.

4.1 ZOPA'S STORY

Zopa was founded in 2004 and launched its peer-to-peer(P2P) lending platform. The idea began with founders, Tim Parlett, Giles Andrews, James Alexander, Richard Duvall, and David Nicholson, seeing an opportunity to simplify lending practices in the consumer loan market. Since starting in 2005, Zopa has generated over 5 billion GBP, issuing loans to approximately half a million borrowers. Over the course of their underwriting, they have generated an additional 280 million GBP in interest for its investors. Their value lies in the product they can offer to their customers at low cost and provide a transparent user experience for borrowers and lenders (Zopa, 2020) (Northzone, 2020). The system is based on reducing costs for customers while increasing returns for professional and non-professional investors alike. Via its 'soft' credit check, Zopa can pre-approve over 60% of its customers in less than 12 seconds, making it substantially more efficient than a face-to-face approval process conducted by conventional financial institutions (Zopa, 2020).

Another testament to Zopa's accomplishments includes its market share in the UK. Zopa currently has lent approximately 5 billion GBP in the UK market, and with the UK consumer loan market to be estimated at approximately 27 billion GBP, they account for approximately 18% of the market (United Kingdom Consumer Credit, 2020). In comparison to a more prominent P2P platform in the US, Lending Club, having funded ~58 billion USD(Lending Club, 2020), their

market share of the US consumer loan market equates to approximately 1.4%(Board of Governors of the Federal Reserve System,2020)[4].

Selling to Customers

By creating appeal for consumers and potential borrows, Zopa published a report with the intention to stimulate consumer lending as a means of investment. The research found that individuals who borrowed for the purpose of renovating their homes saw a return on their investment in the form of the resale value of their homes (O'Neill, 2019). Claiming the kitchen to be the 'heart of the home', as their suggestion showed various returns on investment based on a sample study of 1,550 homeowners who have taken a loan for specifically for home improvements. Various improvements and corresponding average returns were as follows: kitchen (51%), loft conversion (70%), décor (62%), and garden (14%) (O'Neill, 2019). Additionally, almost two thirds of homeowners decided to continue living in their homes post-renovation showing that P2P lending makes personal financing affordable to elevate people's lives.

Selling to Investors

Their claim-to-fame also lies in being the first peer-to-peer platform in the world, bring forth approximately 15 years of experience (Judith, 2015). Its initial introduction into the market showed slow growth and resulted in all founders but one leaving Zopa's in its start-up phase. Knowing the potential of the company, Andrews persevered in communicating the value of the company and convincing investors of Zopa's worth. Facing challenges during their blow-back in entering the US market, and the 2007-2008 financial crisis, Zopa's default rates topped-off at 5.54% in 2008 (Judith, 2015). Post-crisis, Zopa began turning a profit.

Growth and Expansion

Zopa's products include Zopa Core, Zopa Plus, and an interest savings account (ISA). As Zopa originated more loans on their platform surpassing the 1 billion GBP around 2015, they began attracting the attention from institutional investors. In 2016, they finalized their first securitization

---

[4] US consumer loan market estimate is based on a snapshot from December 2019 as this is the most relevant data provided for the UK consumer loan market. Lending Clubs total amount of loans lent has been divided by the 4.191 trillion USD consumer loan market size, equate to 1.4%.

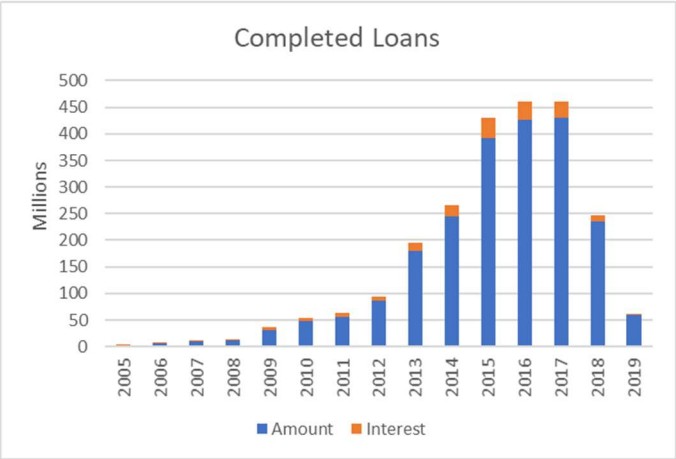
deal. This was also the first European securitization deal pertaining to unsecured consumer loans originated on a P2P platform backed by 150 million GBP (Hale, 2016). This access to capital has provided additional capital to Zopa as well as the ability for lenders on the platform to sell their existing loans before maturity. From this new business model, Zopa can provide more flexibility to its users while retaining a 1% fee for any transition of credit ownership.

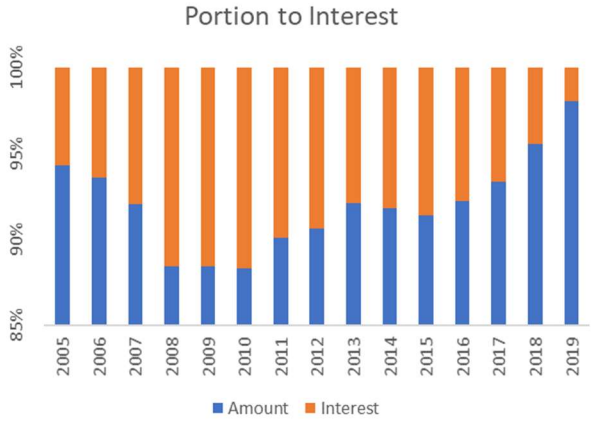
Although other P2P lending platforms have securitized loans in the past, a recent headline that sets Zopa apart is that they are the first to achieve a AAA rating on assets strictly generated on their platform. It is additionally the only program globally to receive this grade for funding equivalent equal to 245 million GBP (O'Neill, 2019). As this is the third securitization of asset backed securities (ABS) for a P2P lender, it was preceded by an increase in the rating of the assets in the portfolio based on Moody's rating methods during the second securitization deal(O'Neill, 2019).

4.1.1 Historical comparison of Zopa performance to their claims

Given the granularity of the issued loans dating back to 2005, Zopa's claims to numbers can be validated to a degree to replicate a similar generalization about the overall performance of the business over the last 15 years.

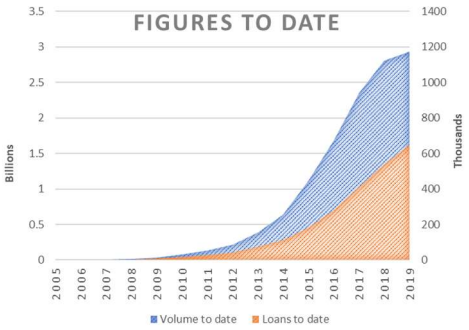
Depending on how the total sum of loans had been calculated, the volume of facilitated loans to date can either be estimated to 4.75 billion GBP or 5.18 billion GBP. The difference between the aggregated values is a result of accounting for the interest charged on all loans. Provided below is a chart(left) demonstrating the increase in issued loans that have been completed. 2018 and 2019 show a low amount which is understood as only showing loans that have matured from those years.

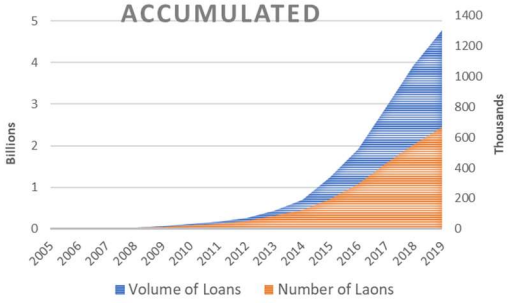




The chart on the right shows the portion of interest charged and collected on the cumulative loans that matured from each year. The average interest can be seen here as being approximately 8% average across all issued and completed loans. The current completed loans in 2018 and 2019 show a small portion of interest that was collected which is understood to be due to early loan payments and reducing the amortization schedule of the loan.

In reference to claims made by Zopa, the current return advertised by the company ranges from 3.4 – 6% (range includes both Zopa Core and Zopa Plus products). Although the charts above show a greater return, it does not include fees taken by Zopa, separation of risk profiles, nor defaulted loans that may pertain to capital loss. Included below is the total number of loans created on the platform each year. Zopa's growth over the years has accumulated to approximately 5 billion GBP as seen in the graph to the right. Showing the same loans for each year from the 'Figures to Date' graph except as a sum of the previous year(s) for completed loans.

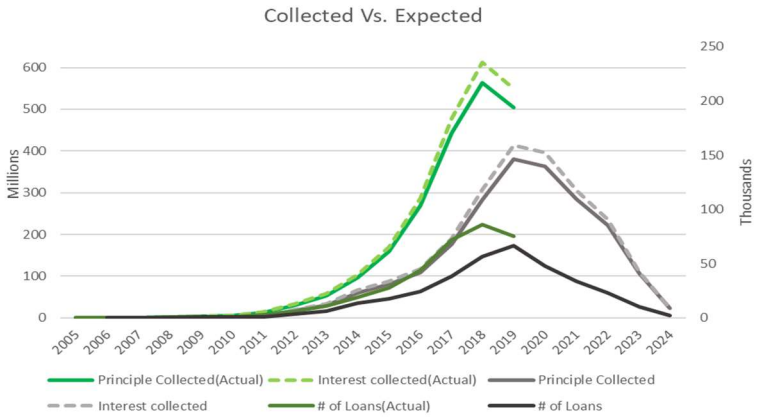




Although the graph to the left has been constructed on modified data, it is an approximated representation that would have shown a larger value before processing the original data set. Additionally, it correlates to the claims Zopa has made in quantity of loans that have been signed (seen on orange) and total volume of accumulated since they began. The values here have been aggregated by adding from previous year to the current for all issued loans (include defaulted, late, active, and completed).

Additional business performance is displayed below for loans that have been labeled as 'completed' in the original data set. Both graphs show the same volume and value of loans via sorting by date issued and due date. Firstly, the graph 'Collected Vs. Expected' shows all completed loans charted by the year they were meant to mature in grey, while the green colored graph shows the same loans in the year they were completed. It illustrates the quantity of loans

that were paid early in comparison to their terms and the loss in revenue from interest not collected on outstanding principle (interest show as dotted line).



4.1.2 Initial overview of Zopa's Loan Book

The entire loan book consists of 667,000 loans. Includes miscellaneous data points for defaulted loans where the loan has been closed with an incorrect date. Additional research has been done via contacting Zopa and no further information was given on the matter. The difference is minor and includes approximately 8,000 of the default loans which only reduces the defaults by approximately 1.2% as seen below:

Type	All Loans		Filtered	
Active	269,185	40.35%	269,185	40.88%
Completed	359,727	53.93%	359,727	54.63%
Default	31,474	4.72%	22,933	3.48%
Late	6,694	1.00%	6,694	1.02%
Total	667,080		658,539	

The previous table shows the total sum of loans separated by the types that were provided by Zopa. The filtered column includes loans where the default date is either a system error or an incorrect value. An incorrect value was identified by the fact that the date included the years between 1900 and 1912 which would not be applicable. The filtering of the data only reduces the sample size by a small amount and therefore does not impact the overall quality of data that is to be used for the algorithms.

Investigating Interest collected

From the data additional investigation led to a better understanding of the interest paid on loans and the difference between what should have been collected versus the amounts that were actually paid by the borrowers for completed loans. The values were calculated via formulas based on APR and APY.

Completed Loans	Number of invoices	Actual Interest collected	Interest calculated	Average difference
Paid Early	258,931.00	135,207,136.65	289,872,521.51	-597.32
Paid on time	257.00	101,579.65	102,746.42	-4.54
Paid Late	100,539.00	45,368,081.05	45,353,890.96	0.14
Total	667,080.00	383,506,138.37	796,340,840.85	-618.87

The table above summarizes the values for strictly completed loans and the interest amounts that were collected. In the 'Actual Interest collected', this is the sum of interest payments that were collected from the three types of loans. The three subcategories for completed loans were based on when the loan was paid in accordance with the initial terms agreed. The formula used to categories the loans is an 'IF' statement:

$$= \text{IF}\ (\#\text{ofPayments} < \text{Term}\ ,\ 1\ ,\ \text{IF}\ (\ \#\text{ofPayments} > \text{Term}\ ,\ 3\ ,\ 2\ )\ )$$

The variable, '#ofPaymetns', is how many payments the borrower made before the loan was entirely paid back and categorized as completed. The 'Term' is duration of the loan and since the payments are made on a monthly basis, both values translate to months as the unit of measure. For example, a loan that has a '#ofPayment' less than the term such as 6 payments on a 12 month term would have paid back the loan in half the time and would be categorized with a 1(equal to Paid early). If the quantity of payment exceeds the term, it is labeled as being paid late (thus the value 3 is assigned). Finally, this leaves the loan with being categorized with a 2 if neither are true, and the loan would be understood as having paid back the principle and interest as it was agreed.

The column 'Interest Calculated' shows the expected interest that should have been collected given the term, interest rate, and principle of the loan. At this level of investigation, it is evident that loans on average that pay back earlier, cost the borrowers less in expense and more for investors expecting returns on the predetermined arrangements. Overall, the difference in interest amounts for loans paid on time can be denoted to various outliers as well as rounding errors.

As it should be, loans that are considered late are generally paying more in interest. This calculation is limited to applying 30 days to each month to identify the date when the loan matures and excludes months with 31, 28, and 29 days. Because of this, loans could be misidentified as late by as little as one day when in fact it is paid on time.

The following subsections will provide an overview of the process under which a Zopa borrow undergoes when being issued a loan and provides context to the loans seen in the data that is to be pre-process, processed, analyzed, and discussed.

4.1.3 Zopa's lending process

The following section will summarize the process Zopa carries out when providing a loan to its applicants. The is an important step in the entire process as this is the fundamental process that has placed Zopa as a competitive P2P loan platform. For each of the five stages in the loan application process, an excerpt from the Prospectus as well as further explanation of what the process entails and its relationship to the research at hand. This process is extracted from the issued prospectus on Zopa's securitization deal in 2016 with Deutche Bank AG, who was leading the program to finance the P2P platform (Deutche Bank AG, 2016).

Stage 1: Zopa Borrower Loan application

The first stage of the application process for a potential borrower to receive a loan begins as any process with the fulfillment of an application. The information required by Zopa is an industry standard[5] to better gauge the creditworthiness of the borrower.

> *The first stage is the online application by the borrower on the Zopa Platform, when he or she provides required information such as identity, address history, gross employment annual income, desired loan term and amount. Upon submitting the application, the Zopa Borrower provides consent to Zopa to carry out a "soft search" with the credit bureau.*

As per the European Banking Authority's recommendations, the creditor must gauge the requestor's ability to fulfil their obligations to the loan. This includes the following ratios and metrics that must be evaluated: loan to income, loan service to income, debt to income, debt

[5] As per the European Banking Authorities recommendations, stated in document EBA/CP/2019/04 in section 5.2.1, *General Requirements for Lending to Consumers*

service to income ratio. This initial step in the process is an informal 'soft search' as Zopa states and leaves no mark nor trace on the applicant's record and potentially having a negative effect on their rating as a borrower in the UK.

Stage 2: Eligibility screening

As stated below, this is a generic screening process ensuring that Zopa, as a lender, meets further requirements for assessing the applicant meets minimum requirements to receive a loan.

> *The second stage is the eligibility screening, according to certain eligibility criteria. These criteria define generic eligibility requirements, including but not limited to: minimum age, minimum UK residency history, and minimum income. If a loan application successfully passes this screening, the loan application then goes to the classification stage.*

Annex 2 in EBA/CP/2019/04 outlines these same criteria for a lender to be able to gather in order to verify the identity of the applicant.

Stage 3: Classification: Application of the proprietary scoring model

Once data is submitted by the applicant and all relevant information is collected, it passes through a model that aggregates data from additional parties where applicable. This process is customized to the specific applicant. The loan will then be placed in either of the seven categories listed in the except below.

> *A loan application that passes eligibility criteria goes through the "scorecard": a proprietary scoring model using both data submitted by the applicant and provided by third-party credit-reporting agencies, to generate a credit score related to that specific loan application by the Zopa Borrower. Scores translate to "Zopa Markets". As of the Provisional Loan Portfolio Cut-Off Date[6] there are 7 Zopa Markets (A*/A1/A2/B/C1/D/E), plus the N markets for scores that fall below the lowest Zopa credit market cut-off and which are score declines.*

[6] "Provisional Loan Portfolio Cut-Off Date" means 31 August 2016

The eighth category is "N" which places loans in the lowest category of Zopa's credit market. This consists of loans that would not be eligible for the portfolio as well as declined from receiving funding.

Stage 4: Ratecard, Quote and Reservation

The Ratecard is also a proprietary model where some factors are disclosed but are not exhaustive. This includes all loans accepted into Zopa Markets as mentioned previously. The price is then determined via the following process:

> *Applications which are mapped to a non-N Zopa Market are then matched to one or many lender(s) (either as a whole loan or set of microloans) and subsequently go through the Ratecard, which determines the price to be shown to the borrower (consisting of rate of interest, a borrowing fee, and a resulting APR). The Ratecard is driven by multiple factors including, but not limited to: Zopa Market, loan term and loan amount. For whole loans, the borrower is presented with a quote, alongside pre-contract and regulated contract information, including the name of lender. If the borrower chooses to accept the quote (a "Reservation" of the quote), he/she electronically signs the contract, provides bank account details and direct debit mandate, and consents to the terms and conditions.*

As per Zopa's investment page, loans categorized from A* to E include obligors that have an income equal to or greater than the UK's average income. At this stage, the requesting party, borrower, can decline the offer provided by Zopa. If the Loan is categorized in the riskier E or D class, and the potential borrower will be matched to many lenders. From an investor's perspective, this splits a riskier loan into several microloans and diversifying the risk for the investor. Furthermore, Zopa diversifies your investment as a lender on the platform so that the a 'single borrower holds no more than 1% of [the] initial investment'.

Stage 5: Final stage underwriting

The loan goes through a partially automated process that will retract any outstanding information that is still required by the 'Underwriter Guidelines'.

> *Once a Reservation occurs, the loan application goes through final stage underwriting for final approval. This stage in the Zopa eligibility process includes data verification and final review to ensure adherence to Zopa's Underwriting Guidelines. A portion of Reservations are auto-validated where no*

'flag' is triggered requiring a referral for manual review of the application data. The applications which are the subject of a referral flag go to manual underwriting for a review of the specific feature(s) related to the referral flag (for example identity verification, proof of income, clarification on credit items, etc.). At this stage, further data will be gathered, and additional checks made on eligibility.

Although this is the final stage and Reservations have been accepted, this final review of credentials and validation of information might present further insight into the borrower's creditworthiness based on additional data points revealed herein.

4.1.4 Added Context to Zopa's Loans

Credit Evaluation

At Zopa's discretion, all aspects of the loan are conditional during the process and new quotes can be presented. The variety and type of data that can be used during the evaluation is dependent on the applicant and the weights assigned will also vary. The same information such as credit score, Zopa Market rating, or income are not available on a granular level. Alternatively, Zopa published various metrics that resulted in defaults, such as loan term length. These metrics are attributes and applied to the ML models to align the output of the algorithms.

Defaults Vs. Arrears: as per Zopa

Both terms relate to the servicing of the loan that has an outstanding balance. This loan is monitored to ensure all payments are made in a timely manner as agreed upon by all parties. The Prospectus defines the procedures that take place for servicing loans in arrears as 'Collections' and those in default are labeled as 'Recoveries'. Zopa's collections teams handle outstanding loans that are currently in arrears while the servicing of loans that have, or are about to go into default, can be outsourced to a third-party.

Regarding the downloaded data, there are four categories: *Completed*, *Default*, *Active*, and *Late*. The loans that have been labeled as *default* would consist of 2 additional subcategories. This includes secured and unsecured loans.

Secured loans become defaulted after 4 months beginning from the last period that a payment was either not received or only partially made. These loans are general paired with a vriaty of collateral but more commonly include real estate, car or other major assets of value owned by the borrower. Unsecured loans have a default period of 3 months after last full payment and are

generally not backed by any sort of collateral. This type of loan poses a higher risk for the lender but is also more difficult to obtain as there are less point of creditworthiness to establish for the potential borrower.

Although the guarantee of payback is higher with secured loans, Zopa offers loans at a higher interest rate to make up for the risk. This higher return for higher risk appetite can translate over to investors on their platform. The return, dependent on risk appetite of the investor, can range from a 2% -5% return with the least risky loans, and the riskiest loan profile has an expected net return of 10% - 18% return. These returns are of course paired with the fact that the A*, lowest risk loan profile, has an expected default rate of less than 5%, while the riskiest, E, profile has an expected annual default rate of 10% - 15%(Zopa, 2019).

Prior to becoming a default, a loan is considered in arrears. This classification can correspond to loans in the data set that are labeled as *late*. A loan will generally stay in this status if the borrower has made substantial payments to cover the missed amount. Dialogue and compelling evidence of payment may prevent the loan from going into default for collections to takeover. A loan will generally receive this status after 15 days (one-month average) from the time that a payment has been missed. In this stage, a debt collection agency may be contacted after 45 days have passed.

Completed status is tied to loans where principle and interest have been paid back. Instances where the loan is paid back in advance, the terms for expected interest can be agreed on between all parties who are stakeholders in receiving the full interest payment. This would of course benefit the borrower in reducing the interest paid and for this reason, would require that terms be reevaluated for the lender.

Active loans are critical to this as those are the loans that the research herein will aim to predict the outcome for. Active loans are currently maintaining payments and are following the schedule agreed on when signing the terms and conditions of the lent amount.

Understanding Zopa's Interest calculation

In order to understand the business model of Zopa and the profits acquired from the loans issued via their platform, the interest paid was assessed for each of the individual assets. Zopa loans operate on an amortized loan profile. This type of loan includes a fixed payment of the initial principle issued at a certain date and an interest rate tied to the loan at the given date. The interest rate, as mentioned in subsection *Origination and Underwriting of Loans on the Zopa Platform*, is

a proprietary calculation part of Zopa's **Ratecard**[7], which would have been beneficial to understanding the credentials under which Zopa issues their loans.

The initial numbers provided by Zopa include the following:

- Principle borrowed – principle amount of outstanding loan
- APR – required calculation as the APY was provided instead
- Term of loan – number of monthly payments that needed to be made until loans maturity

Zopa stated their interest rates as being calculated based on annual percentage rate (APR) but is in fact calculated via annual percentage yield (APY). APR needed to be converted backwards via the following formulas provided herein. APY, which is more commonly advertised to lenders takes into consideration annual compounding, unlike APR.

As stated by Frankel, APR is generally a metric one may see when borrowing money, while APY pertains to individuals how are loaning money or depositing funds into a savings account. Therefore, it is important to understand the difference between the two as the calculations advertise APR, but the data collected must be calculated on an APY.

For context, interest rates in general can be understood via the following example. A loan amount of $100 has an annual interest rate of 12% that is not compounded monthly. The interest will then be divided by the 12 months in a year, providing you with a 1% interest rate monthly on the outstanding $100 initially borrowed. Thus, your payments on interest would translate to $1 per month as your cost for borrowing or lending the initial $100.

Monthly interest payments = Principle x (interest / 12 months)

$100 x (12% / 12 months) = $100 x 1% = $1 in monthly payments on interest

Cost after 1 year of borrowing equates to $112 dollars

(assuming principle is paid back at the end of the term)

The annual percentage rate (APR) for a loan is the amount that would either be charged to the borrow or earned by the lender on an annual basis. Loans with fixed APRs contain rates that are

[7] Ratecard pertains to a proprietary process which Zopa uses in identifying borrower risk. Their ability to underwrite loans at customized interest rates is a part of their success as P2P loan platform.

guaranteed not to change during the life of the loan. Fixed rates are generally higher than variable rates at the time of loan origination (Frankel, 2019).

Variable APR based loans can change at any time and when analyzing the total interest paid on accounts within the loan book, this may denote the difference between calculated and actual amount paid for completed loans. As these loans are usually tied to an index, such as LBOR, they usually tend to fluctuate accordingly.

APR in this case is the future return or value on your investment given the compounding interval (APR Calculator, 2020). Referring to the example above, the compounding of interest on an investment worth $100 would utilize the following formula at the given rate of 12% annually:

$$APR = P\left(\left(1+\frac{r}{n}\right)^{\mathrm{nt}} - 1\right)$$

$$APR = Principle\ Amnt\left(1+\frac{Interest\ Rate}{Quant.Copmounded\ in\ period}\right)^{Quant.Compounded\ *\ Interest\ Rate}$$

$$APR = 100\left(1+\frac{.12}{12}\right)^{12*1}$$

$$APR = 100(1+.12)^{12}$$

100*1.127 = $112.7 will be earned at year-end

After investigating the interest values provided in the loan book on loans that were complete, it was discovered that APY was provided instead of APR and thus, the APR needed to be calculated in order to see the amount that should be charged on principle borrowed. More specifically, the value provided the effective APR and required the following formula, from which provided the APR. The *i* variable in this case was the Provided APY.

$$APR = ((i+1)^{1/n} - 1) * n$$

After the APR has been calculated, the monthly payments could be calculated via the following formula:

$$MonthlyPayments = \frac{\frac{APR}{n} * \mathrm{P}}{1-\left(1+\frac{APR}{n}\right)^{-\mathrm{P}}}$$

The monthly payment value would then be multiplied by the number of months for the loan. This value gives the total amount paid as well as the interest when subtracting the total payment from the principle amount.

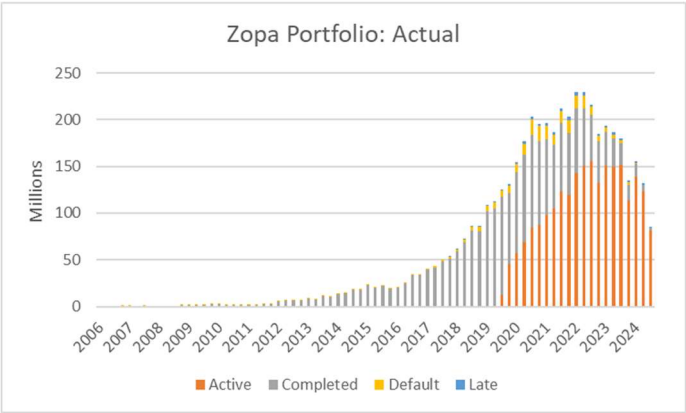
The *Amortized Interest Calculation* can be seen in its entirety via the provided formula below:

$$Amortized\ Interest\ Caclculation\ = \left(n * \left(\frac{((i+1)^{\frac{1}{n}} - 1) * P}{1 - \left((i+1)^{\frac{1}{n}}\right)^{-P}} \right) \right) - \mathrm{P}$$

The formula provided above was conducted in several steps via excel, where each calculated amount, such as monthly payments and APR conversion, could be seen. The formula above is a combination of all steps that was later utilized for each loan to calculate the expected interest amount to be paid and the difference presented between what was collected.

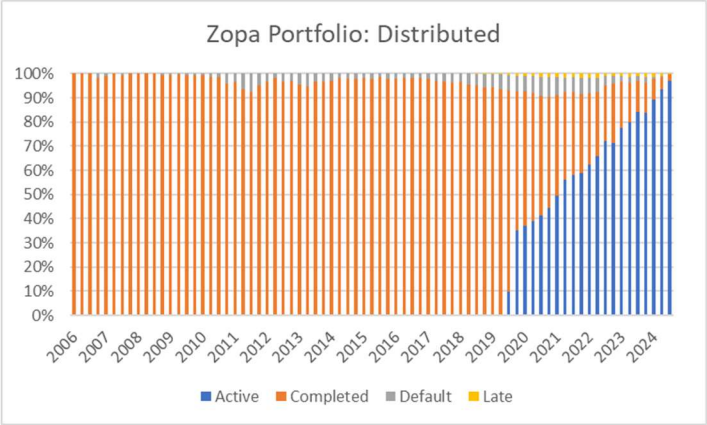
4.2 Historical Data Analysis

Historical performance of the portfolio is used as standard practice for analyzing a portfolio of assets and looks to identify general characteristics which are arrears, defaults, and recoveries (Markovitz, 2019). The analysis here will touch upon the defaults of Zopa's loans as per the approach found in Moody's report due to the limited data provided by the public loan book.



The first illustration in this section provides an overview of the loans issued over the years. This not only shows the overall growth of the company, but the growth of defaults as well. Since Zopa began, they have been able to generate a steady flow of new loans maintaining a steady completion of loans, defaults to a minimum. In the first graph, it is evident that Zopa has incurred more defaults at a faster rate than they have grown, but alternatively, the next graph represents the loans as a small percentage of the total loan make for a given year.

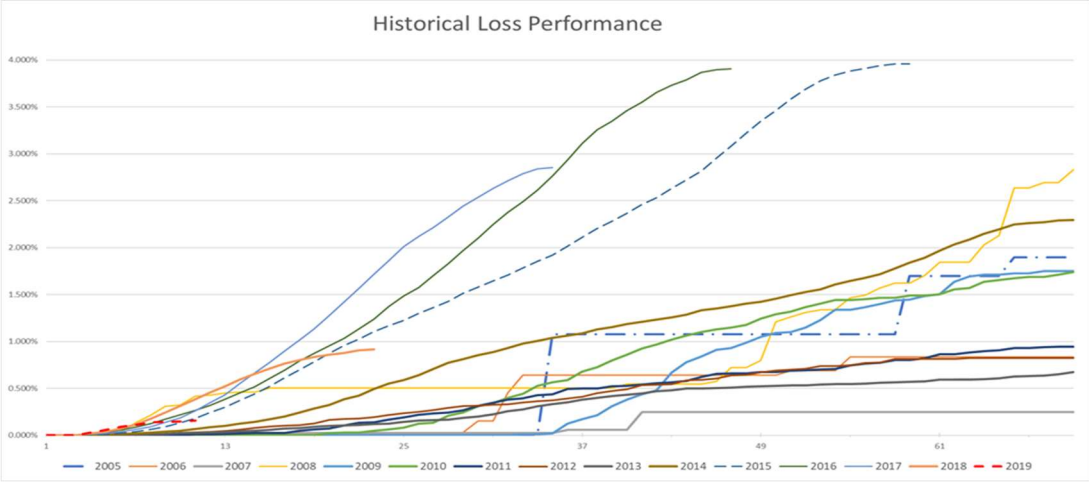
For loans that are defaulted, there might be spikes due to extended time that the loans have been on the books as late, and then deemed default, or even loans that have been on the books as default which later are finally recorded as a loss and deemed default. This relates the processes that Zopa has in place and maintaining a consistent portfolio, which is crucial when projecting performance on historical data.



The second aspect is loans that are deemed late. These are loans that, after a 3-month period, end up in arrears for after 15 days. Only making-up a smaller portion of the sum, this could either be due to the borrowers incurring a slight delay in payments and/or Zopa being able to service the loans efficiently. In servicing the loans, Zopa would either identify a loan as in arrears or default based on evidence provided by the borrower.

The historical loss performance takes into consideration only loans that have been originated in a given year. This illustration shows a smoothing of the curve which can be denoted to several factors. One of which can be deduced to Zopa's servicing of loans in that are behind on payments has become more efficient and thus having the ability to maintain their process in updating loans to default status. Another reason for this can be due to the fact when Zopa first began, a loan of 1,000 GBP would have a greater weight on the vintage once deemed default, which is represented in the spike for 2005, at month ~36. Historically, the loss performance has transitioned into a smoother graph as a single loan defaulting would have less weight in the total portfolio and normalizing the spikes when deemed default.

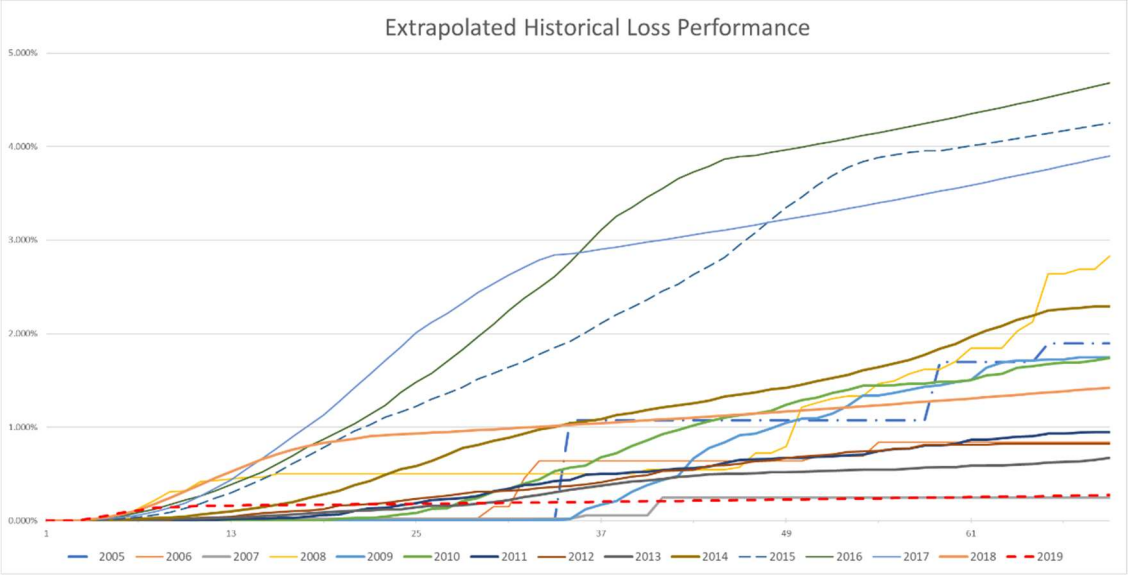
Current historical performance shown in graph on next page



In addition to the historical loss performance becoming more stable, it has also increased where we see the cumulative default being much higher at a sooner stage over the last few year. In comparison to the models provided in the prospectus, data only provided up to 2016, this data is not segregated divided by the rating provided by Zopa but encapsulates a global perspective on the total portfolio. Regardless of this limitation on the validating the results, the gap between the more recent years in comparison to the trend seen in Zopa's earlier year can be validated from a shallow perspective. From the prospectus, there is a paralleling gap over the years. In segregating the loans based on their assigned ratings, this model can be applied to assigning the appropriate rating as designated by Moody's recently proposed methods.

With the gap between more recent years and earlier years of Zopa, the increase in its customer base relates to the increase in loans from their public loan book. To mitigate this, Zopa has undergone a change to its approach in becoming more selective in its customers, as their current claim on acceptance to rejection is approximately 20:80. More importantly, the representation of the data above is preliminary charting, intended to construct a full default timing curve. The mean default would strive to show the expected performance based on the current economic outlook (Udot et al., 2017). Moody's report specifies that the defaults can be extended for each vintage that has not reached its maturity by taking the last data point and multiplying it by one plus the average growth rate of cumulative defaults. This would be a recursive process in extending the graphs to project expectations, via data extrapolation as seen below.

Extrapolation method provided in graph on next page



4.3 UK

Appended under 'Postal Code Data: Full overview', is a table explaining all data points and categories derived from the source, DeGrave. This table gives extensive insight into generalized factors of each individual postal code. Consisting of postal codes and district codes, each post code can have a variety of inputs such as population, deprivation index, distance to nearest public transport, longitude, latitude, etc. Most of the categories have not been used as they do not add value comparative to previous research that was conducted when analyzing default risk of assets. Due to the granularity of the data for each post code, the districts needed to be summated in order to help establish a base line from the information that was provided by Zopa. An example of this includes a sum of the values such as houses and population for all districts within each post code. The amount of district codes can vary from 1 to over 1,500 for each postal code. Additional to consolidating the population and households per postal code, the Index for Multiple Deprivation has been analyzed to show the minimum, maximum, average, and standard deviation values for each postal code. The Standard deviation was relatively high for most districts, therefore showing that the deprivation is not strictly dependent on the district.

Combining UK public data for added attributes to Zopa Loan Book

The Index of Multiple Deprivation (IMD) was a figure presented in the postal code data and uses a set list of metrics to define the economic and sociological deprivation within each district. IMD is the official measurement for poverty in the UK (Penny, 2019). This metric for poverty was aggregated into four separate fields as mentioned. These methods were best suited for compiling

the data from all districts into single postal codes. Providing both ends of the spectrum with minimum and maximum, the data quickly shows an average, as well as the distribution of the IMD for a given postal code (Penny, 2019).

As stated by the Ministry of Housing, Communities, and Local Governments in England, '[the IMD] follows an established methodological framework… to encompass a wide range of an individual's living conditions. Ratings are not proportional and do not measure on an absolute scale, but relative. An example of this is that a district rated 200th versus 100th does not mean that it is twice as deprived as the other. IMD considers seven categories in its assessment: income, employment, education, health, barriers to housing & services, crime, and living environment.

5 Applied Methods

5.1 Process Overview

Theoretical behind process

From David Spiegelhalter's work in The Art of Statistics, Learning from Data, he introduces the process of 'Learning from Data' as inductive inference. Firstly, his distinction between deductive and inductive is made on the basis of conclusions being drawn, or deduced, on concrete evidence, whereas the later draws on the fact that there is general uncertainty. Spiegelhalter provides a simplified diagram that inductive inference would follow to arrive at a conclusion that can help better understand the outcome of data processing.

The diagram outlines a high abstraction for approaching the conclusion from the given dataset of Zopa's loan in four stages. Additionally, as stated by Spieglhalter, the arrows in between each stage, can be interpreted as "tell us something about [the observations]". More specifically, the process will serve to understand the driving factors behind a loan's status, and to predict the outcome of a loan prior to maturity. This process is focused towards understanding something for a data set that is not entirely clear from immediate observations.

Figure 6 Process of inductive inference

The processes that take place in this model include a variety of inferences and obstacle to continuously understand if whether, or not the data and inference made about the data can be misconstrued as bias or if the results can be replicated. This is where it is important to understand variables external to the data set. Such variables can include consumer spending patterns, borrowing trends, and the general loan market. These attributes can only corelate to the economic snapshot at the time that any data is collected and analyzed.

Zopa's public loan book provided a large sample size of over 660,000 loans that have been issued on Zopa's platform with a large portion of the portfolio consisting of loans from 2019. The initial data sample shows a good variation of loans that have been paid back, paid late, outstanding, and defaulted. As the initial data sample from Zopa only had 11 attributes, the data set was combined with an additional nine attributes that added more depth to the data. These additional eight attributes are country, population, number of households, average Index of Multiple Deprivation (IMD), calculated interest, prepayment identifier, due date, months on books, and rating class. The additional data was provided by an independent developer who has developed a user-friendly platform that aggregates data in CSV form from the Office for National Statistics (Bell, 2020).

Sample Overview: Count of Loans per Loan Status

Sample (1000)			Full Data List		
Status	Count	%	Status	Count	%
Completed	615	61.5%	Completed	359,727	53.90%
Default	80	8.0%	Default	31,474	4.70%
Late	7	0.7%	Late	6,694	1.00%
Active	298	29.8%	Active	269,185	40.40%
Total	1.000[8]	100.0%	Total	667,080	100.00%

In order to build the models and test results, a small sample of 1,000 loans have been randomly removed and added to a list that would have the algorithms applied. The random value was

[8] The final sample was 1,500 samples as there were an additional 500 default samples used in the set to balance the model between defaulted and completed loans. This changed the profile to the following: Completed 41%, Default 38.67%, Late 0.47%, and Active 19.87%

assigned via Microsoft Excel's RAND () function. Further examination of the list was analyzed to ensure there were no duplicates.

The values that were assigned at random to consolidate a sample data set of 1,000 loans are summarized in the table above(left) and the proportional makeup of the smaller sample size to the full sample. Although there was a shift in the sample sets loan makeup, the distribution of loan types within the randomized sample shows a sufficiently distributed spread, as well as a large enough data set for ML training and testing. In order to optimize the process in testing the models, the smaller sample set will be used. It will also determine the processing capabilities and limits to the number of instances that can be used in a model at one time. The 'Diagnostics' section will elaborate if changes will be required in reducing the quantity of instances or attributes.

5.1.1 Controlling Sample Variability

Variability in the random sample data set was compared to the proportion of varying loan statuses in the initial data sample of 667,000 samples[9], to ensure that the variability between loan statuses would not exceed that of the original sample set. This step in processing the data at random strives to uphold the quality of the data to which the models would be tested against and support the internal validity of the research. The sample set also holds external validity as the data has a direct relationship to the loans which the research will strive to understand.

5.1.2 Process Flow

The framing to the research process is derived from is categorized as inductive inference as to which inductive reasoning will be applied to the entire population based on smaller extracted samples. The generalized conclusion will stem from the fact that the sample, the study population, has a strong correlation to the target population, that which assumptions are being made about.

In the case of machine learning, the predictive strength of models increases with the size of the tested data set. This provides more scenarios to which the model can be more attuned to the variability in a single input that may be considered an outlier to the entire sample.

---

[9] Original data set extracted consisted of 667,080 loans, all having a unique 'loan ID'. Data was extracted at year-end 2019

In figure 2, the process by which data will be extracted parallels that of inductive inference diagram provided in figure 3. ML research project conducted on language processing in 2019, The Moderation of Social Media Platforms, followed a similar process in staging the data in separate parts after having aggregated from multiple sources, versus there only being one data source for this framework.

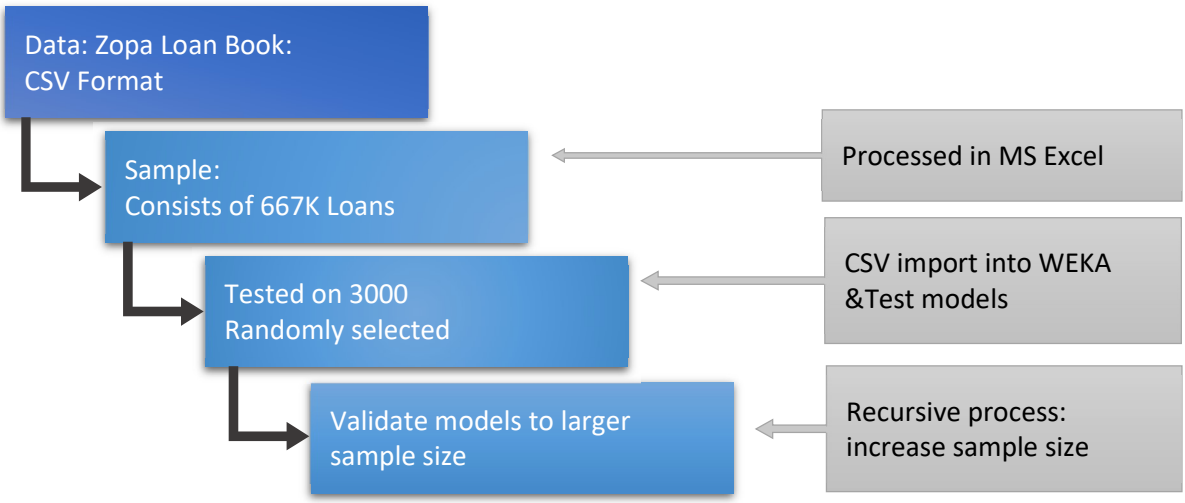


Figure 7 Process framework

5.2 PREPROCESSING

About the Tools

WEKA is an academic tool developed for the purpose of understanding and teaching the application of common machine learning algorithms. The java-based application has pre-loaded classifiers where raw data can be easily imported for training and testing a user's models. WEKA's interface also allows for changes to be made to the parameters of models, custom exclusion of values, assignments of weights, etc. The details of WEKA's application are mentioned in the data preprocessing and processing. The preprocessing incorporates the initial sample test on the formatted datasets and fine tuning of inputs prior to training the final model.

The processing of the data involved investigating the context of the data prior to and providing the reasoning behind the models used. Explanations to newly introduced terminology and elements of the research will be paired with examples, graphs, tables, and explanations.

WEKA's incorporates a built-in classification to the supplied data set. These algorithms ultimately allow for the automated calculation of a problem with minimal human intervention,

providing an output given the parameters and algorithm used in the mechanism. The parameters, or data discrimination, are a part of supervised machine learning. Only after the output is generated, the analysis can begin with better understanding the input provided and adjustments to be made for model accuracy. Presented in the appendix under Zopa Data Categories, is a table listing the data provided by Zopa and its context.

The dependent variable will be determining the status of the loan given the following attributes. The goal of the algorithms will aim to draw relationships between dependent and independent variables on a granular, invoice by invoice, basis. This mapping of the data in order to come to a relevant conclusion is part of a process called *predictive analysis*, where machine learning is leveraged to minimize human input and data processing, ultimately leading into the precursor of *Artificial Intelligence*.

When trying to reach a conclusion about data, this goes beyond the shallow descriptions of explaining what the data is, but rather strives towards understanding how this data serves to project future outcomes. In the instance of Zopa loans, understanding pattens behind existing data in order to better understand the outcome of unknown results will be an approach that requires modeling with a degree of predictive strength that could potentially serve to add value in assessing the underlying risks associated with the borrowers and lenders on their platform.

Incorporating UK Data

To provide additional input for the ML algorithm in constructing the models, UK post codes were retrieved from the Zopa Loan Book and duplicates were removed. Excel's built-in functions were used to extract and remove duplicate values to match with the Loan Book. As the UK postal code CSV file was too large for Excel, the UK postal code workbook was extracted in parts separated in alphabetical order and loaded into separate workbooks: A-G.xlsx, H-L.xlsx, M-P.xlsx, Q-T.xlsx, and U-Z.xlsx.

From this, the UK lists were minimized with post codes that were no longer in service and memory was reduced by eliminating categories such as longitudinal, and latitudinal coordinates for all post codes. Then, the posts codes that were relevant to the Zopa Loan Book were filtered out using a VLOOKUP function to match post codes provided in the Zopa book with the post codes in the UK CSV file. After consolidating the lists and 'cleaning up' the workbooks, it was possible to consolidate all post codes in one workbook into two tabs, A-L and M-Z. Pivot tables were then created to quickly aggregate the average IMD, population, country, and households.

The reason to the aggregating of data per post codes was that each regional identifier in the post code, usually represented as B3, EH14, U5, etc., may have several unique post codes within. Provided below is a table of the categories and equated values shown for each post code which reference table in section 12.1 and will be referred to as *UKTable*.

Code	Country	Pop	House	IMD
AB11	Scotland	21,209(sum)	10,915(sum)	3,417.004(Average)

After consolidating data for post codes into one list with aggregated values per post code, the VLOOKUP function was applied a second time to the list of loans in the additional categories that would be populated by the post code list. The following formula(pseudo) is provided below for each category.

Category	Pseudo Formula
Country	`=VLOOKUP (PostCode; UKTable; Country; Exact Match)` *Return exact country match to the provided post code*
Population	`=VLOOKUP (PostCode; UKTable; Population; Exact Match)` *Return exact population match to the provided post code*
Households	`=VLOOKUP(PostCode; UKTable; Households; Exact Match)` *Return exact households match to the provided post code*
Index of Multiple Depravation (IMD)	`=VLOOKUP(PostCode; UKTable; IMD; Exact Match)` *Return exact IMD match to the provided post code*

Handling missing values

After incorporating postcode data, it was evident that various post either, did not match to what was provided in the UK data set, or the Zopa Loan book was missing post code values all together. Investigating the missing postcodes in the loan book showed that there was an even distribution of missing post codes over the last decade and can be denoted as a lack of information provided by the customer. The following shows the newly processed loan book where missing values have been taken out.

Table provided on next page

Adjustment Totals Due to Missing Post Code

Status	Loan Count	%Δ	Loan Amount	%Δ	Interest Amount	
Active	269,185.00	-1.26%	2,289,157,348.16	-1.27%	167,081,203.81	-1.42%
Value Decrease	3,391.00		29,182,270.44		2,373,239.71	
Completed	359,727.00	-3.83%	2,218,218,178.11	-3.38%	180,676,797.35	-3.75%
Value Decrease	13,776.00		75,035,724.33		6,768,267.02	
Default	31,474.00	-1.72%	221,961,504.34	-1.79%	27,886,674.95	-1.75%
Value Decrease	540.00		3,982,290.00		488,999.00	
Late	6,694.00	-1.72%	50,097,251.08	-1.65%	7,861,462.27	-1.71%
Value Decrease	115.00		825,870.00		134,093.87	
Grand Total	667,080.00	-2.67%	4,779,434,281.69	-2.28%	383,506,138.37	-2.55%
Value Decrease	17,822.00		109,026,154.77		9,764,599.60	
New Grand Total	649,258.00		4,670,408,126.92		373,741,538.77	

As it was explained in section Data Mining above, removing instances entirely that did not have a post code was the approach in this case as it only had a small impact in reducing the sample size. It was also important to see the impact across all 'loan status types' to ensure that the reduction was consistent throughout all loans and did not fall heavily on loan status. Additional to posing a drastic change to the loan book, a proportional decrease across all groups shows that there are not significant changes to the data models provided previously.

Incorporating an Additional Class

The incorporation of an additional class included the rating score that is distributed to Zopa borrowers. Unfortunately, this value is only provided to investors, so it is a rounded estimated classifier based on claims made by Zopa. The data set was categorized by the interest rate that was charged to the borrowed amount. This of course is only a piece of the puzzle as the interest calculation considers current credit score, outstanding debt, granular demographic data, amount being borrowed, term of loan, etc. It is understood that this is a very simplified and limited approach to analyzing the data. Its main purpose is to provide an additional value to the recursive models. It will be investigated further in the preliminary 'Diagnostics' section later in the methodology.

The applied classification method split the data into groups based on interest rate. As shown below in the table, the section under classification details the parameters for splitting the loans into their designated class. Loans with interest rates below 5% were A*, those below 9% were A,

and so on. In comparison to the rates Zopa specified for each class, there are minor differences. Due to overlaps and gaps in the rates, such as C and D rated loans, the classification method used a wider range for C rated loans and a narrows range for D rated loans.

	Zopa's Classification			Research Classification		
Class	Typical Interest (%)	Loan Term	Default (%)	Typical Interest (%)	Loan Term	Default (%)
A*	3-5	12-60	<0.5%	<5	47	0.41%
A	5-9	12-60	0.7 - 2.6%	5-9	41	1.93%
B	9-14	12-60	2 - 5%	9-13	39	4.40%
C	13-17	12-60	4 - 8%	13-19	43	6.57%
D	19-26	12-60	7 - 15%	19-23	36	6.89%
E	21-29	14-48	10 - 15%	>23	38	10.61%

The results were compared with the published loan term and default rate for each loan class. The average loan term for all classes fell within the range that Zopa listed. With the exception of the C and D rated loans, the historical default rates for the classes that were calculated, fell with the ranges that Zopa had claimed on their website: (SUM Class Principle + SUM Class Interest) / SUM Class Default. Although a majority of the data could be reconciliated to hold more validity, it will be important to note these limitations if, in any case, the algorithms derive this attribute as a driver to a conclusion.

5.2.1 Data Formatting

Once the comma separated value (CSV) file was downloaded from Zopa's website, it was examined in Microsoft Excel from which unneeded values were removed. There are several ways by which one can process the data for use in WEKA.

ARFF

One of the approaches included an ARFF (Attribute-Relation File Format) file which is created via preprocessing in Excel and then imported into Visual Studios in order to eliminate unreadable characters, as well as implement proper formatting for Java based WEKA to process. This is a file format that was developed for the Machine Learning Project in the department of Computer Science at the University of Waikato. The functionality begins with an @relation line. The attributes, starting with an @ symbol, are specified in the header. In creating the ARFF file for Zopa's loan data, 17 attributes were created. As shown blow, each attribute is defined with @attribute, given a name, and defined by type. The first attribute is named 'Disbursal' and

defined as a date. The type can be further specified via formatting of the value. With the case of 'Disbursal', the date was formatted in Excel as year-month-day.

```
@relation default-rates-p2ploans
@attribute Disbursal date "yyyy-MM-dd"
@attribute Loanamount numeric
@attribute Principle numeric
@attribute Interest numeric
@attribute Payments numeric
@attribute Lastpmnt date "yyyy-MM-dd"
@attribute term {'12','24','36','48','60'}
...
```

Each instance in this case is a loan, that contains 17 values separated by a comma and delimited to a column when imported into WEKA. The is the same principle by which CSV files are converted to .xlsx format in MS Excel. The instances are defined by an @data line, declaring where the data begins. An example of one instance has been provided below. It only contains the first 7 attributes corresponding to those above, with alternating highlighting:

```
@data
2016-10-24,7500,5371.937493,467.0225073,35,2019-09-26,48,...
```

Since the objective is to teach the models in predicting each loan's loan status, the attribute 'Loanstatus' is 'set as class' in WEKA, where then WEKA's algorithms will build its model dependent on that value among the instances.

The model can either be built and tested using existing instances, or a new data set containing '?' characters for the 'Loanstatus' attribute can be added.

The workflow of this process includes processing data in Excel, loading it into Visual Studios, from which it is saved in ARFF format. When working with large data sets, errors that occur upon loading them into WEKA requires that this process be reiterated. This is due to Excel having predefined operators that will not read as text values but are required for the ARFF file to be imported properly.

CSV

Although WEKA is developed to work with ARFF files, CSV files can be loaded into the application where it is then converted. The conversion separates the CSV into tabular format

where it then automatically assigns attributes to each value. When errors occur, the process that is reiterated requires less steps, and reduces processing overhead.

Specific attributes can be defined by type within WEKA, simplifying the preprocessing and formatting of multiple files. As this research does not contain a variety of string values, the chance for potential errors is drastically reduced. An example of challenges revolving around importing as a CSV file includes language processing as there may be characters found in the data which are considered operators for ARFF file format. The loan book data mainly consists of dates and numeric values. The additional string (text) values, are limited in their variability and include a limited number of loan statuses and UK post codes (short string values).

5.2.2 Overview of Preprocessed for Diagnostics

Once the data has been preprocessed to remove missing values and apply a constant value for missing values, sample set for processing was sorted based on the random character assigned. The total count of the sample size, as explained in section 7.1.1, was 1,500 instances. The additional 500 was added to the sample as it was strictly loans that had defaulted. This ties into the practice of balancing a model when constructing a binomial classification model.

An example that supports the method for balancing can be an image recognition model that must identify whether and image is a dog or a cat. In order, to know what a dog is, the alternative must also be provided. In providing the data for the two classes, the idea is to create a model that can perform in classifying both outcomes to a high degree. This type of structuring is dependent on both sets being equally as diverse and well suited in providing the necessary input for the model, as well as having the capability of discerning between a cat and a dog. The binomial split will be made up of approximately 600 completed loans and 600 defaulted loans, providing a slightly larger total set compared to the multinomial sample.

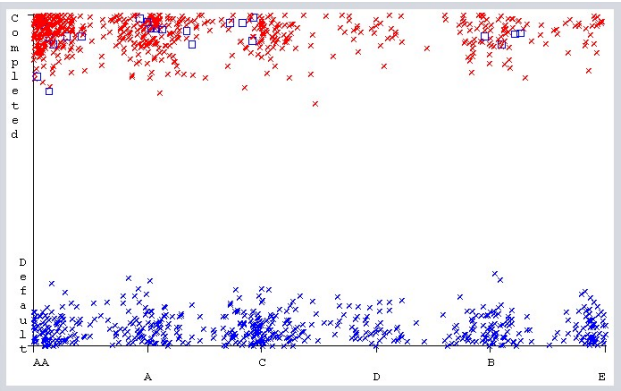
When balancing sets, it must also be important to consider that real world instances do not occur in a 50% chance. As it was seen in the initial 1,000 instance sample set of loans, defaults amounted to only a portion of the portfolio, while completed or active loans made the largest portion of Zopa's loan book. To study the models' performance, the classification was conducted on both data sample, realistically distributed multinomial data, and binomial balanced data. The final Data set to be used for initial testing was made of 21 attributes, including the final attribute set to class, 'Loan Status'. Table to which attributes are referring to can be seen in the Appendix, under section 12.2.

Labels such as PPF, DUE DATE, and MonthsOnBook, were initially used as helper columns while investigating the data and configuring models. The data has been kept for the first sample test in order to see the performance and if they bear any indicators. The results captured in the first trial of processing will explain any outliers, so that the model can be retrained from a more practical approach.

5.3 PROCESSING

In the processing stage of the research, four models were used: logistic regression, naïve Bayes, J48, and random forest. This process was done in wo stages, one with a relatively small sample size of approximately 1000 instances for a binomial data set and a multinomial data set. From previous research that was conducted, other process included converting statuses such as 'late' into default due to a lack of data to manipulate a binary classification of loans. Alternatively, the binary data set was made of actual loans that were deemed default or completed from historical data.

The initial test was done on a set of 1000 instances that included a multinomial class: default, complete, late, and active. The model was accurate in defining the class with a 10-fold cross validation. All results were above 90% accuracy. The binomial sample also included an extremely high accuracy rating. The binomial sample set was made up of approximately 1000 instances but included samples that were either default or completed. The model was balanced, meaning that the instances provided were close to 50% split between both class types. Weka was then used to balance the data set to exactly 50% defaults and completed loans.



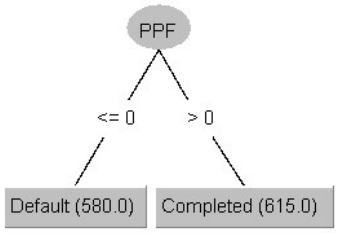
A logistic regression on a binomial data set of 1195 instances was trained and had an output result of 97.99%. Looking into the results further, a visualization of the model's classification of loans was investigated. Between the two classes, the profiles for classification were remarkably similar, meaning that an attribute does not skew the data in any one direction. For example, an x-axis for status, and y-axis for loan rating show an

even comparison between the two as seen in the chart provide here. The jittered[10] red and blue values depict default and completed loans. The density for completed loans is greater at the AA end of the scale, defaulted loans show a denser population in the E segment. Unfortunately, the data is not in order and places B as second last, and C in third.

Naïve Bayes was the second model to be trained on the binomial data set. Just as logistic regression was run via a 10-fold cross validation. The results showed a 100% success rate, with all metrics in a detail results matrix showing a value of 1.00.

1195 instance data set was a J48 classifier. On the same 10-fold cross validation the model was just as successful as naïve Bayes. Before looking into the decision tree constructed for J48, random forest classifier was run and amazingly showed equally as high results. As astonishing as it may be, this is not practical, and is an instant red flag.

Firstly, to examine this problem, the J48 tree was investigated. WEKA can generate a decision tree from the provided data and show how the root was first split, and were the terminal nodes stopped. The splits also provide a threshold value determined as being the amount that would set the highest or lowest entropy for a given value, and increasing the information gained from the decision. With the decision tree provided here from the latest J48 model, it is able to distinguish between defaults and completed loans on one split, the prepayment factor. If it were not the PPF, it would have been MonthsOnBook, or DueDate as the next attributes to provide a 100% accuracy.



The problem with the first sample test showed that the data could have been overfitting, but more importantly, it contained values that were based on a loan having been completed. The result of this was removing all attributes that were derived from the default date. Furthermore, in this case the algorithm identified any relationship between the final loan status and integer ID, the unique identifier was also removed. The encrypted borrower ID was left as it was a unique string value.

[10] Jitter is a random dispersion of data points in a given area that would otherwise be overlapping and unseen. Increasing the jitter, as it was down for the graph above, helps show the population size for the various rated loans in each class.

5.3.1 Diagnostics

The initial test provided a clear understanding for what the data is supposed to represent and where one must tread carefully. In cases where data can be represented several different ways, more is usually not better, as there might be cases of overfitting, or data derived from a value that can only be assigned to a specific class would mostly skew results unfavorably.

As it was seen here, the attributes derived from the default date, used in building non-ML based models went unnoticed until used as inputs for ML. Going forward, the four attributes were removed, leaving a total of 17 attributes that will be utilized in training the models via the same methods to determine the strongest outputs. In comparing the models, a training set of 50 loans that are known to be defaulted and a separate set of 50 loans that are known to be completed will be tested to see which of the two models provide the most accurate prediction for both binomial and multinomial. This will be followed by a comparative analysis.

5.4 DATA ANALYSIS

This section will analyze the models' accuracy and precision based on the output of the model in a 10-fold cross-validation. In a conducting a cross-validation via 10 folds, the partitions the data into 10 proportionally split groups(folds). The model provides 10 separate evaluations from which an average is calculated. The final models are evaluated one more time which is seen under 'stratified cross-validation' within WEKA. The printouts of a model's evaluation and performance is shown for each model constructed under binomial and multinomial data sets. The following section will first present the results of binomial and multinomial models on the smaller data set. Once the best performing model is captured, based on the summary, two models from each section will be reconstructed via the same method, but with a larger data set of 50,000 instances. This will show a change in precision and recall as more values are introduced. All screen prints of the model's performance in WEKA have been appended. The following section will also explain the measures used for evaluating the models in relation to the best performing model for the binomial data set as these will be referred to throughout the analysis.

5.4.1 Binomial

The results for the four models are as follows in order of least to most accurate: random forest, logistic regression, J48, and naive Bayes. The result for naïve Bayes can also be assessed via a 2-by-2 confusion matrix from which the precision, accuracy, and F-score can be derived. The matrix consists of true positive, false positive, false negative, and true negative. True positives

pertain to the instances that were considered default and identified by the model as default when they were tested during the training which amounted to 491 defaults. False negative has been specified as loans that were completed and those consisted of 564 correctly identified as such. The outstand false negative and false positives amount to 139 instances that were incorrectly classified. The naïve Bayes model also had the highest f-measures, also known as the f-score. The f-measure is the optimal measure between accuracy and precision where the model would perform best in the trade between the two.

It is critical to understand the difference between the two as there is a tradeoff between precision and accuracy. An example of such includes a machine that can detect a faulty product leaving the assembly line and withhold it from being sent out. Out of 1,000 products, 100 are known to be faulty. For a method that uses high precision, this presents the risk of a portion of products that are faulty, to be sent out. Since all 60 products the machine detected were faulty, its precision is at 100% but customer satisfaction might not be as 4% of products sent our would be defective. If a machine is more heavily weighed on detecting based on recall, it may overcompensate and remove more products than necessary. Since it is not as precise, this would leave the manufacturer with a loss in products shipped out. The f-measure in this case finds the ideal point were the highest precision can be achieved while taking into consideration the recall of the machine.

Along with testing at an 88.34% accuracy, the model was 10% more accurate in precision for identifying defaulted loans and 13% less accurate in its recall of defaults. The overall precision and recall for identifying a completed loan were 84.2% and 94.5%.

The next best performing model in the group was a decision tree, J48. With an overall accuracy of 86.93%. the model was 25% more accurate in precision for identifying defaulted loans but 24% less accurate in its recall of defaults. The overall precision and recall for identifying a completed loan were 79.8% and 98.9%.

Logistic regression was used as a base case and was positioned as third best model among the four. In applying its statistical method to the binomial set, 1,011 instances were correctly classified, with 184 as incorrectly classified. From the confusion matrix, there were 486 correctly identified defaulted loans, 111 were incorrectly classified. The result is approximately 18% of default loans having been misclassified and about 12% of misclassified complete loans. A more detailed analysis will show that the area under curve (AUC) was a weighted average of 89.1%. This value is derived from the f-measure and show the entire area under the receiver operating

characteristic curve (ROC). The greater the area under the curve, the better the algorithm was able to identify the dependent variables with the trade-off between precision and recall.

Random forest performed the worst[11], with an AUC of 88.3% which corresponds to the confusion matrix showing a misclassification of 28% for defaulted loans, and 14% were misclassified completed loans. If, for example, the AUC was at 50% would mean that there is a 1:1 trade-off between recall and precision and is as good as flipping a coin to decide on the dependent variable. Thus, a model with AUC ≤50% would not be acceptable.

Model (binomial classification)	Correctly classified	AUC (weighted average between 2 classes)
Naïve Bayes	1,082	.951
J48	1,032	.923
Logistic Regression	1,011	.891
Random Forrest	947	.883

5.4.2 Multinomial

As stated in section 7.1.1, the multinomial models were trained using a four classes that were, for the most part partitioned equal to that of the overall data sample of ~660,000 loans on a sample set that was an even 1,000 instances. The reason for creating a sample that was equally balanced was to see the performance of the algorithms when assigning the probability for a particular outcome to the entire data set. In other words, loans statuses that are more prevalent would assume new data for testing as being partitioned in a similar manner. Although, random forest was not the best performing model, it was able to correctly classify all defaults without misclassifying other loans in the default category. Unfortunately, both decision trees, J48 and random forest, were not able to classify any loans labeled as late. Instead, J48 classified the seven late loans as completed, and random forest classified six as completed and one as active.

Although J48 had the best accuracy at 94.1%, its AUC provided the lowest value which can be reduced to the fact that its f-measure for late loans was unidentifiable and reducing its weighted average for the entire sample size. The performance of the model was based on its accuracy, false positive rate, and true positive rate. When compared to logistic regression, which place third in AUC metric at 95.3%, J48 performed about 37% in reducing the false positive rate and about 10% better in its true positive rate.

[11] Random Forrest– 79.25% accuracy: 947 correctly classified and 248 incorrectly classified

Naïve Bayes performed second best in terms of identifying the least number of incorrect instances. Its ability to classify the instance correctly was 25% better than the succeeding model, random forest. Although random forest had a 1% increase in AUC measure, Naïve Bayes was selected for the testing a large data sample as the classification outweighed the 1% improvement on harmonizing between precision and recall.

Model (multinomial classification)	Correctly classified	Incorrectly Classified	AUC (weighted average between 4 classes)
J48	941	59	.935
Naïve Bayes	933	67	.979
Random Forest	918	82	.980
Logistic Regression	859	141	.953

6 RESULTS

Continuing from the previous two section, the selection for the best two models to be retrained on a larger data set included J48 and naïve Bayes for binomial and multinomial data sets. The same methodology was applied to creating a larger set on instances of 50,000 for both cases. In the binomial set, 25,000 randomly assigned defaulted loans were selected and 25,000 randomly assigned completed loans were selected. As for the multinomial data set, 50,000 random loans were extracted from the loan book that have not been incorporated into the model previously. The results showed that J48 outperformed Naïve Bayes on both fronts, bi- and multi-variable classification.

For the binomial classification, naïve Bayes' performance dropped by approximately 4% in correctly identifying the classification. In a multinomial model, naïve Bayes decreased to 90%, from its original 93% accuracy in predicting classification. These percentages are based on the difference between the weighted average in accuracy from the entire sample size and distribution of instance classification[12]. The result based on accuracy, false positives, and ROC provide the following comparative analysis to be between J48 binomial and J48 multinomial models as the best performers.

[12] The difference between the percentages is a higher abstraction on evaluating the weighted averages between the larger and smaller sample size would differ slightly. A more accurate approach in identifying the change in performance, the percentage increase or decrease would need to be assessed per class.

The screen print of the results for J48 in a binomial classification have been presented in the following table[13]. The overall accuracy as well as the f-measure had an improvement of a 5.5% change in the accuracy to correctly identify a loan classification, and a 5.7% increase to the f-measure. The AUC also increased by an additional 2.8%. The model's performance places it in second to the multinomial J48 model.

Summary of J48 Results in Binomial data set of 50,000 instances

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
1	0.837	0.002	0.997	0.837	0.91	0.846	0.949	0.962	Default
2	0.998	0.163	0.86	0.998	0.924	0.846	0.949	0.921	Completed
Wtd. Avg.	0.917	0.083	0.928	0.917	0.917	0.846	0.949	0.942	

In this instance, the tradeoff between precision and recall can be seen here, where the precision and recall between the two classes are almost inversely related to each other. For defaults, the model classified more less defaults but was more precise in its classification. Relative to the analogy of a machine detecting defects in the assembly line, the model can identify the defaults to a high degree of certainty (97%), whilst lacking the correct classification for of total defaults in the data set. Completed loans on the other hand might not be as precise, and might classify more loans as complete than default, its ability to classify most loans as complete is high (99.8%).

When investigating the output decision tree constructed by the model, J48 Binomial has its first split on the attribute 'principle collected'. Although it is difficult to explain, this can be understood as the model identifying a loan with larger payments to the loan principle as the first step in greatest information gain. The split, provided in the appendix, with the greatest entropy is at approximately 1,000 GBP. Throughout the model, a majority of the decision nodes are seen to be based on 'principle collect' as the model narrows its approach in identifying the loans until a number of misclassified loans reaching a leaf where the number misclassified is accepted by the model as the optimal point before it begins to over fit its classification on an instance by instance basis. A large portion of misclassified default loans can be seen in the third and fourth layer of

[13] J48 Binomial data set of 50,000 loans correctly classified a total of 45,869 instances and misclassified 1,738 instances

the tree for loans having a 'principle collected' that is greater than 1,000 GBP (first decision node).

J48 multinomial data set increased its performance on all metrics provided below, in addition to being able to identify the late payments but at a low accuracy given that it is a substantially smaller sample size to the other three classes. ROC increased in its weighted average by approximately 4.4%, and the number of correctly classified instances (true positives), increase by 2.5%. Although the algorithm was only able to achieve 61.3% in precision and 22.6% in accuracy, the weighted average in the last row of the table below takes the instances' weight into consideration. Thus, not bearing heavily on the overall performance of the model.

Summary of J48 Results in Multinomial data set of 50,000 instances

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
1	0.961	0.046	0.957	0.961	0.959	0.915	0.97	0.961	Completed
2	0.998	0.055	0.931	0.998	0.963	0.936	0.985	0.971	Active
3	0.275	0.006	0.674	0.275	0.39	0.415	0.775	0.349	Default
4	0.119	0.001	0.479	0.119	0.191	0.236	0.744	0.115	Late
Wtd. Avg.	0.939	0.048	0.929	0.939	0.929	0.896	0.966	0.931	

Both naïve Bayes and J48 underperformed in classifying the default class from the same data set. Like the binomial model, the table above shows the recall-precision trade-off for default loans. With precision at 67.4% and recall at 27.5%, this model lacks in accuracy for predicting defaults as 72% of default loans are classified under another class. Based on the provided data and 10-fold cross-validation, the model can only predict the defaulted loans at a 67.4% accuracy. Compared to the multinomial naïve Bayes classifier, 7% less of the default loans where predicted by J48.

A large difference between the decision trees of multi- and binomial J48 pertained to the first decision node splitting on date of last payment and was able to correctly identify the 'active' class of loans at a 94.8% precision, and recalling for 99.8% of loans that were actually late. It would be important to keep in mind the decision tree constructed via multinomial classification had a substantially different structure from that of the binomial classification. For multinomial J48 began with 'principle collected', just as binomial, but went further into splitting the next large portion of the data by last payment date and overfits based on a single instance basis in many cases. This overfitting is represented in the appendix where a structural overview of the number

of leaf nodes and decision nodes used. At the second level, the tree does not effectively split on the principle of greatest gain, but rather to increase the output accuracy. This would ultimately only increase the final, weighted result but may present a problem when new data is introduced. A method to overcome the overfitting of a decision tree would include pruning, which is the opposite of splitting. It would entail removing nodes from the model but would need to be tested as each node is removed because of this may drastically decrease the predictive accuracy of the tree with the existing data or when new instances are introduced(Jain, 2017).

6.1 INTRODUCING NEW DATA

The following section will take the models a step further in testing their capabilities with classifying a small data set that has not yet been introduced. The data set will consist of one sample that contains. An equal split between the four classes and another set where the samples are randomly split. The randomly split set resulted in the following:

	Randomized sample	Even Sample
Status	**Count**	**Count**
Active	25	25
Complete	67	25
Default	8	25
Late	0	25
Total	**100**	**100**

This final comparative analysis will show the performance of the models and the value of the results. It will also demonstrate if there were instances of overfitting. The sample will also be applied to the binomial models. As there is a mismatch between the classified data being used, the binomial models will not be able to identify whether a loan is active or late since the original model was only trained on defaulted and completed loans. Alternatively, the results for binomial models will display how successful the model is at classifying the default and completed loans at the very least.

6.1.1 Sample Testing with External Data Set: 100 Instances

As it was previously presumed about the J48 model, the supplied test set showed drastically inaccurate results. The recall of the results is fairly high, but the precision showed otherwise:

Matrix table provided on next page

	Multinomial J48		Multinomial NB				Original sample
	Active	Completed	Active	Completed	Default	Late	
Active	1	24	4	19	1	1	*25*
Completed		67	8	55	1	3	*67*
Default	2	6	3	2	3		*8*
Late							0
	3	97	15	76	5	4	100

When incorporating the randomized data sample into both models, J48 heavily weighed the original data sample towards completed loans and was not able to identify any defaults. When precision and recall was calculated, J48 had 100% recall on completed loans but only a 67% precision. The overall accuracy of the model for active (74%), completed (70%), default (92%), and late (100%) loans is seen here to not be an appropriate measure for defining this model as an accurate classifier since it could not identify any defaults provide an f-score of 0.22[14].

Naïve Bayes showed similar result in accuracy being high with a large tradeoff for precision. Given that active loans have yet to mature, this model provides interesting results on the active loans, showing that 19 out of 25 loans as completed. Additionally, the 12 misclassified loans from the data set only misclassified 33% of loans in the default or late class. When referring to a previous study, a multinomial data set was modified where active and complete loans were categorized in one group while late loans were added to defaulted loans. The models showed a high ability to classify completed loans, but both underperformed in classifying defaulted loans.

The evenly distributed sample set was used to see how the model split the classification of loans given a set that did not share the same distribution as the training set. This was intended to see if the models would disregard or incorporate the probability of an instance appearing in a data set.

Matrix of multinomial model performance provided on next page

[14]External confusion matrix calculator used from, https://confusionmatrixonline.com/

	Multinomial J48				Multinomial NB				Original
	Active	Completed	Default	Late	Active	Completed	Default	Late	sample
Active	25				23	1		1	*25*
Completed	1	24				21	3	1	*25*
Default	4	12	8	1	2	8	15		*25*
Late	12	10	3		10	5	4	6	25
	42	46	11	1	35	35	22	8	100

In applying the results into an error matrix, the J48 model had 100% recall for active loans and 96% for defaulted loans, but a substantially lower precision as the additional misclassifications in the group fell in the active loan category. This is also not entirely wrong considering the loan is active, and the classification in the completed and default category would result in a prediction based on collected principle and date paid back. The average f-measure derived from the following matrix resulted in a weak score of .47[15] and accuracy of only 57%. More importantly, the J48 model was not able to identify defaulted loans. Naïve Bayes performed a little better with a resulting accuracy of 65% and an f-score of .63[16]. The multinomial J48 model's inability to classify can relate back to the complex decision tree, showing an obvious case of overfitting and lack of pruning that took place in the training process.

	Binomial J48		Original
	Completed	Default	sample
Active	14	11	*25*
Complet	67		*67*
Default	2	6	*8*
	83	17	100

J48 binomial model extrapolates results based on several factors, but a few attributes that were seen to repeat in the decision nodes included, 'principle collected' and 'interest collected'. 'Principle collected' appeared at the root, second, forth, and sixth level in the tree. This carries limitations, as a loan that has just started may have only paid a small portion of its principle and interest, therefore labeling a portion of new loans as default. In the matrix above, we see that the algorithm classified 2 defaults as completed which could be to the amount of principle that has been paid back. When investigating these instances, one was certainly default, but the other (D rated) had actually paid back its principle in full and the interest paid back was more than what was calculated on the term and interest. Thus, it is possible the had

[15] F-score for Multinomial J48: Active 0.75, Completed 0.68, Default 0.44, Late 0.0

[16] F-score for Multinomial Naïve Bayes: Active 0.77, Completed 0.70, Default 0.64, Late 0.36

either been mislabeled or there is additional background information not presented in Zopa's loan book.

The model classified all completed loans correctly, most of default loans correctly (including Zopa's potential misclassification). Although prediction may be skewed for 'active' loans to appear as default, due to less principle and interest paid, 56% were identified as complete. Additionally, the evenly distributed sample set, the model was able to identify 84% of defaulted loans and 100% of completed loans. It shared a similar identification of 'late' and 'active' loans, labeling more 'late' loans as default and 'active' loans as complete. The performance values (precision, recall, and f-measure) are not ideal which can be due to the small sample size provided and that a multinomial sample is provided to a binomial classifier. In a 3-by-3 matrix, the model had a weighted average f-measure of 0.60, and when excluding the 'active' category, this measure jumps to 0.80, and in a 2-by-2 matrix equates to 0.90.

	Binomial NB		Binomial J48		Original sample
	Completed	Default	Completed	Default	
Active	15	10	14	11	*25*
Completed	**19**	6	**25**		*25*
Default	3	**22**	3	**22**	*25*
Late	3	22	7	18	25
	40	**60**	**49**	**51**	100

The final two binomial models in an equally distributed sample test showed performance on a large sample in each category. In a 2-by-2 matrix that excludes the results for active and late loans, naïve Bayes had an average f-measure of 0.82 and J48 averaged an f-measure at 0.94. With the exclusion of 2 variables, the predictions by the binomial models can be validated. Further validation would include seeing the 'active' and 'late' loans' statuses after maturity. When incorporating the results into a 4-by-4 matrix, the results decreased, but individual identifiers for recall still maintained a high value. Precision decreased because of the increase of instances that the model identified as incorrect but results in classifying instances that the model(binomial) was not trained for remained consistent for all trained classifiers.

7 DISCUSSION: WHAT IS GAINED & HOW ITS APPLIED

The training and testing that took place can, to a limited degree, be applied in a generalized manner as the method for data preprocessing, model construction, and model testing followed a similar footprint to previous research in classifying loans in binomial and multinomial cases. The main difference seen between the binomial and multinomial classifiers was their ability to classify instances in a new data set.

The models assessed in the research included naïve Bayes, logistic regression, J48 decision tree, and random forest. All models included a binomial and multinomial classification, totaling 8 models. Prior to constructing the models, preprocessing and initial testing on the preprocessed data was conducted to understand the scope of the attributes. The initial training, mentioned in the diagnostics section, allowed for a preliminary investigation of how the models interact with the data and first set of 22 attributes.

A review of previous literature on machine learning provided a basic understanding for how the outputs of various models can be understood as well as the metrics that can play a vital role in discerning the strengths of one model over another. Previous methods for ML model construction provided an abundance of previously sources in a variety of programming languages. Unfortunately, previous methods were limited in their application to consumer loans, and the application of the model needed to be adjusted for this case study. More specifically, this ties into increasing a classifiers accuracy for extrapolating the status of a loan prior to maturity.

After conducting the diagnostics, model performance showed inconsistencies with expectations. These inconsistencies where not minor metrics but could be quickly identified as an overfitting or misapplied attributes. When all models in the initial training began providing outputs of 100%, there is little confidence in the fact that performance would be consistent across all methods at such a high degree. This adds to the strength of using several different models such as naïve Bayes and decision trees in one research method. Where one performs better as a robust model in most cases, be it large or small samples, the latter's performance increases with data size. Although performance increases when more instances are introduced for training, comparing results and how they came to be, allowed for a deeper knowledge of the model's feasibility when new data is introduced. This also entailed an investigation into the decision trees and understanding any potential for overfitting as was seen with the multinomial decision tree.

Prior to testing the models with smaller uncategorized data, the improvements to the model were evident when a larger sample was added. The increase of the data size and diversity of instances have shown to both improve a model's overall accuracy, and in other instance it showed a decrease in performance. When applying the test set to the models, it brings to light the output of

the mechanisms, from which it becomes possible to work from a bottom-up approach in adjusting model parameters. As WEKA is considered more of an exploratory learning tool to help introduce amateurs to machine learning algorithms and classifiers, customizing parameters becomes limited. Although data processing tools are provided in WEKA, using alternative methods to sift and data mine large files allows for a more efficient data processing. The methods for converting files to AARF might have allowed for improved results as the integer type could have been specified rather than WEKA discerning the value type on its own. This may include specifying strings in a limited array over an unspecified set or labeling the attributes as dates rather than numbers. With this multi-model approach, the methods technique can be reaffirmed to hold validity in the sense that more than one strategy is pointing to a similar conclusion, payment on principle and interest is a strong identifier in loans achieving a completed status. In the case seen here, once the models were trained on a larger data set, the performance for decision trees increased while the naïve Bayes classifiers decreased in performance. To validate performance of the models', new data was incorporated for testing on the final models. This approach extended beyond just testing the two best models from the initial approach in order to see a more holistic comparison between performance on an unclassified data set.

The final testing was based on generalizing results towards a larger sample and should be understood as such. What the result on the unclassified data sets derived was that performance of a binominal dataset was substantially more consistent among the different classifiers in comparison to the multinomial classifiers. At this stage, the results were investigated and back tracked to the original instances that had been misclassified. Firstly, this back tracking was done to investigate the misclassification of the multinomial models and some of the results made it difficult to explain the rationale behind the model's classification. Some of these loans showed no signs of being default based on the current amount paid compared to what was borrowed and applied interest rate. This brings more questions that, in later research, should be addressed with Zopa clarifying data outliers.

The binomial models proved to be robust as they classified 'default' and 'complete' instances at a greater accuracy than multinomial models. Although this validates the models' ability to predict statuses, what is more important, is predicting statuses of loans that are currently open. This includes 'late' and 'active' loans. The multinomial models lose reliability when tested against the smaller data set, while binomials prove otherwise. The binomial models had a high reliability in deciphering between defaults and completed loans, and they were consistent in providing similar results in classifying instances that were not used in the training process. In order to further

validate the classification of the instances, the loans that are active and late at the time of this research would need to be evaluated in the future to see the loans' final outcome.

The main goal was for the models to classify loans and predict the outcome of a loan prior to becoming completed or defaulting. As this is a snapshot of the loans from 2019 year-end, validating the results would require assessing the loan book in future research.

Future research would require that the classification method be based on a binomial model. In doing so, optimizing a model would entail manipulating loans from a snapshot of a previous date where the outcome of the loans is known. For example, a loan book extract from 2017 would consist of active, completed, default, and late loans, as does the loan book used in this research. The difference would be that the 2017 snapshot would consist of loans that have only paid a small portion of their principle and interest, but their status would be adjusted from 'active' and 'late', to the actual outcome, 'completed' and 'default', as this would be presently known. Once the models are constructed, this would help set a baseline for the performance of the models applied to currently outstanding loans where the outcome is yet to be known. Additionally, the attributes pertaining to principle and interest paid should be a percentage of the total amount borrowed rather than the nominal value. The purpose behind using the proportional amount rather than the value paid is to separate smaller loans that may have defaulted from newer loans that have yet to accumulate these amounts.

The approach taken here is intended to generalize a small sample study to Zopa' entire loan book. The research is an example use-case of machine learning to help project the outcome of loans and the dependent variables that hold the most weight in determining that outcome. Given the limited number of attributes provided by the loan book posed a challenge to improving a model's performance when incorporating new instances. The added data for UK post codes presented no clear drivers in determining the classification of loans. In particular, the added attribute used for indexing the deprivation in the UK showed that loans were dispersed across the entire spectrum. This is inclusive of all models in the study.

As in most cases, machine learning models are particularly challenging to observe on a low abstraction as the mechanisms regarding how a model is constructed to classify data can be obscure. Investigating these intricacies can even lead to more unknowns, but what is important to machine learning is leveraging previous research methods to replicate, validate, and re-test methods. This allows for new research to test alternative methods and take different approaches that have not been conducted previously. Considering the data that one is researching, and the outcome that is to be achieved, parameter adjustments can be made to expand on previous

methods. In all, this helps assure a consistent environment, it establishes a control, and advances the communities understanding of machine learning in practice. Since it is not always possible to trace the results back to a one-to-one relationship, predictive techniques rely on historical snapshot data. Consumer loans depend on broader variables that can be used to determine outcomes, but incorporating surrounding variables also requires that the model be flexible and robust enough to accommodate the changes.

P2P loans have identified a niche in consumer banking that has allowed them to underwrite loans at a fraction of the cost compared to conventional banks and underwriters. Their credit rating system and automation of various processes in distributing loans has made them successful in assuming a strong market position. As it was outlined by Dhar's theory of the phenomenon revolving emerging platform, P2P platforms have been successful in competing in this market because of their ability to integrate systems, onboard customers quickly, efficiently aggregate information, and provide open access to users. Moreover, as it has been seen in the UK consumer loan market, network effect plays a role in user switching from conventional loan providers to P2P platforms. This transition in customer base can also be understood as *positive feedback*. Where the loss of one firm's customer base adds to another firm's customer acquisition; a two-fold impact felt by both business parties. This transition of customer from one provider to another has its barriers given the confidence that must first be gained by the transitioning market. This barrier was evident when Zopa first started but turned into a snowball effect as the platform began gaining traction. This traction can be denoted to the overall market confidence in the increase in the sharing economy. Specifically, with P2P lending, risk is averted by splitting risk among many parties and is also limited to the number of participants involved. Through incentives, transparency, and low barriers to entry for the individual borrower and lender, P2P platforms like Zopa streamline a process that, at one point only involved professionals, has been *democratized* for small amateur investors.

Zopa's gain in the market share has been drastic and their growth in the market has increased substantially when evaluating their proportional growth to the market size in the UK compared to the largest P2P lenders in the US. As P2P platforms gain more traction, their involvement with large investors, and institutional clients will increase with time. With increased focus from institutional investors, lending platforms have gained the attention of regulators and it is now under question as to how much additional attention these platforms will be receiving from regulatory bodies. A benefit to conducting the case study on Zopa is that they restrict lending by and to UK residence, closing of additional variability when assessing the company's growth and

the risk of the assets. The attention from large investors and the risk associated with the underlying asset brings in an additional party, rating agencies. Since P2P platforms have grown to their current size, large investors have acquired a stake in growth of the platforms. In doing so, this introduces sharing economy into the secondary market, where underlying assets can be traded between external parties. This entire process requires for the assets to be assessed and rated by established agencies. The research here draws on the securitization of a portfolio consisting of Zopa's loans that is owned by a single investor. In the deal that transpired in 2016, a prospectus showed methods of analysis for rating the asset pool. One of the methods involved a loss performance measure of extrapolated loans and their comparative vintages. These methods were provided by Moody's who published their revised methods for projecting loan outcomes base average defaults prior to current date that needs to be extrapolate. As seen above, the method is limited in explaining the outlook of a portfolio, particularly for a growth company that has seen a drastic increase in customers, revenue, and defaults. To counter this method, an alternative approach to using machine learning to defaulted loans whose current status is either 'active' or 'late' might present more relevant results. Alternative to using average growth rate of defaults as a projection tool, using ML algorithms that predicts on additional variables beyond that of what is used in loss performance extrapolation can increase the relevance of a forecast as well as assist in categorizing a loan portfolio rating. For this type of forecasting to be relevant, additional attributes pertain to each loan be required. Such elements would include borrower income, current debt outstanding, past defaults, past loans, loan types, agency credit rating, education, marital status, home ownership, etc. As seen with the combination of UK data, an algorithm may be able to predict a loans status on even fewer variables, but this would need to be tested as it is a speculative assumption.

A combination of an exploratory case study of P2P loans and common practices for machine learning can be used to classify loans in a vintage for projecting data. This takes into consideration that the output would provide a more holistic understanding of a portfolio where forecasting can be used for businesses to optimize strategies, introduce new portfolio rating methods, provide additional projections for credit enhancement, and allow investors to strategies for worst case scenarios. Within the machine learning standards that were applied, additional preprocessing and processing of data establishes an overview for the type of data that was processed and additional data that would improve the predictive strength of the models. The comparison between binomial and multinomial classifiers elaborated on the lack of relevancy a multinomial classifier would have when applied in the aforementioned practice of predicting defaulted loans. Furthermore, to abridge the context of loss performance as per industry

standards and the application of machine learning, models had been constructed based on parameters of rating agency methodology, providing a deeper understanding for what machine learning algorithms must work towards in order to be a practical and competitive tool in assessing loans for securitization.

7.1 LIMITATIONS

The limitations of this research did not include in-depth interviews with specialists on the company have been conducted. These interviews would entail a walkthrough of the data that is to be processed and the context behind the values provided therein. Discussing the business model with Zopa stakeholders would establish a more comprehensive understanding of the loan pairing process, diversification of loans among borrowers, management of large investors in a P2P environment, and the underwriting process.

The research would need to be validated on an invoice-by-invoice basis, where predicted outcomes have been matched with loans post-maturity status. Potentially, in some cases, the status of loans me be finalized prior to as a portion of the outstanding portfolio is projected to default. In applying similar methods of data mining, results would differ slightly. Additional data is critical for improving results, but is after accessing additional computational power, alternative programming languages, and using alternative applications for constructing classification models. With alternative applications, training can be dynamic with inputs that are continuously added. Adjusting the models and testing samples at higher limits would be possible on alternative applications.

A limitation that has been stated throughout the research, is the models' ability to differentiate between new loans that have made small payments to principle and interest, compared to loans that have defaulted. This approach would include shifting parameters for classifying loans within the ML algorithm as well as assigning weights to attributes that may help differentiate between the two. Such attributes could include the date of last payment and an additional variable for current date to give contextual variables to the classifier. Another aspect to this would include taking the match between current term of loan and months passed since loans issuance, as this would be an additional method for classifying new loans having accomplished a smaller portion of their term.

8 CONCLUSION

The contribution to new knowledge includes a systematic approach to preprocessing data relevant for understanding consumer loans. The process entails the reconstruction of rating agency models to provide context to the mined data, as well as the ML models designated for this research. Findings include a prevalence in loans being classified as completed due to a higher value of principle and interest being paid back by the narrower. Having followed a standardized approach to outlining the research method, the preprocessing specific to the data from Zopa's public loan book has provided several classifiers for labeling existing loans that are either default or complete with high validity. Predicting loans that are late and active is possible by the model but must be validated via filtering for a vintage that has been fully completed. Although P2P lenders have a strategic advantage to conventional consumer lenders, their loan portfolios share a commonality when measured with loss performance. The theories for FinTech platforms presented here correspond to their ability to quickly gain a large customer base at a lower cost; however, they currently face the same challenges as traditional banks when measuring their performance for the purpose of securitization.

In practice, the methods here are intended to be applied towards assessing consumer loan portfolios and projecting potential losses over the life of a portfolio that has been structured for the secondary market. These methods are intended to provide guidance on the success rate of various models and reasoning behind that success rate. Via the incorporation of additional variables and classifiers, such as those built here, may have a positive impact on the current methods that which rating agencies undergo when establishing a baseline projection for consumer loan portfolios sold to the secondary market.

REFERENCES

Alloway, T. “P2P Lenders Turn to Securitization Deals.” *Financial Times*, Financial Times, 1 Oct. 2013, www.ft.com/content/9a8e427e-2a07-11e3-9bc6-00144feab7de.

“APR Calculator.” Calculator.net: Free Online Calculators - Math, Health, Financial, Science, Maple Tech. International LLC, 2020, www.calculator.net/apr-calculator.html?cloanamount=2020&cloanterm=2&cloantermmonth=0&cinterestrate=5.86&ccompound=annually&cpayback=month&cloanedfees=0&cfrontfees=0&type=1&x=93&y=18#generalapr.

Bartlett, Robert P., Adair Morse, Richard Stanton, and Nancy Wallace. "Consumer-Lending Discrimination in the FinTech Era." Introduction. *Consumer-lending Discrimination in the Fintech Era*. Cambridge, MA.: National Bureau of Economic Research, 2019. N. page. Print.

Bell, C. “UK Postal Codes.” Https://Www.doogal.co.uk/UKPostcodes.php, IP Address: 109.228.61.150 Feb. 2020.

Black, T., Brodsky, L., Mole, K. “Cooperation and Competition in the US P2P Market.” McKinsey & Company, McKinsey & Company, Sept. 2016, www.mckinsey.com/industries/financial-services/our-insights/cooperation-and-competition-in-the-us-p2p-market.

Clemons, E.K., Dewan, R.M., Kauffman, R.J., Weber, T.A. “Understanding the Information-Based Transformation of Strategy and Society.” *Journal of Management Information Systems*, vol. 34, no. 2, 2017, pp. 425–456., doi:10.1080/07421222.2017.1334474.

Cohen, M.C., Guetta, D.C., Jiao, K., Provost, F. “Data-Driven Investment Strategies for Peer-to-Peer Lending: A Case Study for Teaching Data Science.” Big Data, vol. 6, no. 3, 2018, pp. 191–213., doi:10.1089/big.2018.0092.

“Crowdfunding Explained.” Internal Market, Industry, Entrepreneurship and SMEs - European Commission, European Commission, 30 Aug. 2017, ec.europa.eu/growth/tools-databases/crowdfunding-guide/what-is/explained_en.

Craughan, J., Reid, E., Palmer, D. “*Structuring a Marketplace Lending Platform Securitization in Europe*.” *Lexology*, International Law Office, 17 May 2016, www.lexology.com/library/detail.aspx?g=a0becd4e-15a9-407f-8d75-6927861846a2.

DeGrave, K. “Notebook on Nbviewer.” *Jupyter Notebook Viewer*, Kyle Dergrave PhD., 12 Nov. 2016, nbviewer.jupyter.org/github/degravek/notebooks/blob/master/project_loans.ipynb?flush_cache=true.

Deutsche Bank AG, London Branch, ARRANGER, and REGISTERED OFFICE OF THE ISSUER Marketplace Originated Consumer Assets 2016-1 PLC. “Prospectus: marketplace originated consumer assets 2016-1 plc.” Irish Stock Exchange, ISE, 3 Oct. 2016.

Dhar, V., Stein, R. “FinTech Platforms and Strategy.” *Communications of the ACM*, vol. 60, no. 10, 2017, pp. 32–35., doi:10.1145/3132726.

Dietz, M. “FinTechnicolor: The New Picture in FinTech.” *McKinsey & Company*, McKinsey & Company, Feb. 2016.

Engelschall, R. ".MSG: Machine Learning Catalogue." *Machine Learning Catalogue*. Applied Technology Research, n.d. Web. 25 Apr. 2020.

Frankel, M. “Interest Rate vs. APY vs. APR: What's the Difference?” *The Ascent*, The Ascent, 2 Aug. 2019, www.fool.com/the-ascent/credit-cards/articles/interest-rate-vs-apy-vs-apr-whats-the-difference/.

Giudici, P., Hadji-Misheva, B., & Spelta, A. Network Based Scoring Models to Improve Credit Risk Management in Peer to Peer Lending Platforms. 2019 Retrieved from https://www.frontiersin.org/articles/10.3389/frai.2019.00003/full

Gupta, A. "Understanding Logistic Regression." *GeeksforGeeks*. GeeksforGeeks, 30 May 2019. Web. 25 Apr. 2020.

Hale, Thomas. "Subscribe to the FT to Read: Financial Times Debut Securitization for Zopa Loans." *Financial Times*. Financial Times, 26 Sept. 2016. Web. 25 Apr. 2020.

Interface Technologies. "Data Mining and Why It's Important." *Interface Technologies*. Interface Technologies, 05 Apr. 2018. Web. 25 Apr. 2020.

Jain, R. " Decision Tree. It begins here." *Medium*. Towards Data Science, 20 Mar. 2017. Web. 2020.

Jenik, I., Lyman, T., Nava, A. “Crowdfunding and Financial Inclusion.” *CGAP Working Paper*, Apr. 2017

Judith, E. "My Money - Giles Andrews, Zopa Chief." *Financial Times*. Financial Times, 14 Aug. 2015. Web. 25 Apr. 2020.

Kampf, Eran. "Data Mining - Handling Missing Values the Database." *Medium*. DeveloperZen, 23 Dec. 2016. Web. 25 Apr. 2020.

Koehrsen, W. " Random Forest Simple Explanation." *Medium*. Towards Data Science, 27 Dec. 2017. Web. 2020.

Krapf, Armen, and Daniel Kolter. “Proposed Update to Moody’s Approach to Rating Consumer Loan-Backed ABS.” *Moodys.com*, Moody's Investors Service, Inc., 14 Nov. 2018, www.moodys.com/researchdocumentcontentpage.aspx?docid=PBS_1130074.

Lanyon, Daniel. “Investors Are Putting £9bn to Work in P2P Lending across Europe, UK Still Dominating - AltFi News.” AltFi, 2019, www.altfi.com/article/5510_investors-are-putting-9bn-to-work-in-p2p-lending-across-europe-uk-still-dominating.

Lending Club. "Personal Loans Borrow up to $40,000 and Get a Low, Fixed Rate." *Lending Club*. N.p., 2020. Web. 25 Apr. 2020. <https://www.lendingclub.com/info/statistics.action>.

Markovitz, Y., and Yohan, A. “GCR Research Securitization 101.” Global Credit Rating Company Limited, Sept. 2019.

Metz, Caroline. “Turning Debts into a Market: The Wonderful Promises of Securitization.” The Broker Online, 2016, www.thebrokeronline.eu/turning-debts-into-a-market-the-wonderful-promises-of-securitization/.

Murati, A., Skau, O., Taraporevala, Z. “Disruption in European Consumer Finance: Lessons from Sweden.” McKinsey & Company, McKinsey & Company Financial Services, Apr. 2018, www.mckinsey.com/industries/financial-services/our-insights/disruption-in-european-consumer-finance-lessons-from-sweden.

Northzone, Investment Firm. "Zopa." *Northzone*. N.p., 15 Apr. 2020. Web. 25 Apr. 2020.

O’Neill, T. “New Securitization of Zopa Loans Receives First Ever AAA Rating for P2P Loans Globally.” *Zopa Blog*, Zopa Bank Limited, 9 Dec. 2019, blog.zopa.com/2019/12/09/new-securitisation-of-zopa-loans-receives-first-ever-aaa-rating-for-p2p-loans-globally.

O'Neill, T. “Zopa Research Confirms That the Kitchen Really Is the Heart of the Home...and a New One Can Deliver Homeowners More than 50% Return on Investment.” *Zopa Blog*, Zopa Bank Limited, 20 July 2017, blog.zopa.com/2017/07/20/zopa-research-confirms-kitchen-really-heart-homeand-new-one-can-deliver-homeowners-50-return-investment/.

"Our Story." *Zopa.com*. Ed. Press Team Zopa. Zopa, 2020. Web. 25 Apr. 2020.

Pant, A. "Introduction to Logistic Regression." *Medium*. Towards Data Science, 22 Jan. 2019. Web. 25 Apr. 2020.

Parker, A., Sandback, A., Havlicek, B., Clarkson, B., Fanger, D., Laszlo, E., Zarin, F., Siegel, J., Gluck, J. , Becker, K., Merl, M., Mack, P., Jones, S., Rouyer, S. “Demystifying Securitization for Unsecured

Investors.” Moodys.com, Moody's Investor Service: Special Comment, 31 Jan. 2003, www.moodys.com/sites/products/AboutMoodysRatingsAttachments/2001700000415918.pdf.

Penney, B. Statistical Release. In *www.gov.uk/MHCLG*. Retrieved from Ministry of Housing website, 2019, https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/835115/IoD2019_Statistical_Release.pdf.

Quinn, Stephen. (2008). Securitization of Sovereign Debt: Corporations as a Sovereign Debt Restructuring Mechanism in Britain, 1694-1750. SSRN Electronic Journal. 10.2139/ssrn.991941.

Riza, E., Tu, T., Jirasakuldech, B., Lu, M. "Evaluating Credit Risk and Loan Performance in Online Peer-to-Peer (P2P) Lending." *Applied Economics* 47.1 (2014): 54-70. *Static.tongtianta.site*. Taylor & Francis, 2014. Web.

Soni, Devin. "Supervised vs. Unsupervised Learning." *Medium*. Towards Data Science, 16 July 2019. Web. 25 Mar. 2020.

Spiegelhalter, D. J. The Art of Statistics: Learning from Data. Pelican, an Imprint of Penguin Books, 2020.

Udot, L., Parry, A. “Moody's Has Assigned Definitive Ratings to UK Consumer Loan Securitization LaSer ABS 2017 PLC.” Moody's Investors Service, Inc., 16 Mar. 2017.

"United Kingdom Consumer Credit1993-2020 Data: 2021-2022 Forecast: Calendar." *United Kingdom Consumer Credit | 1993-2020 Data | 2021-2022 Forecast | Calendar*. N.p., 2020. Web. 25 Apr. 2020. <https://tradingeconomics.com/united-kingdom/consumer-credit>.

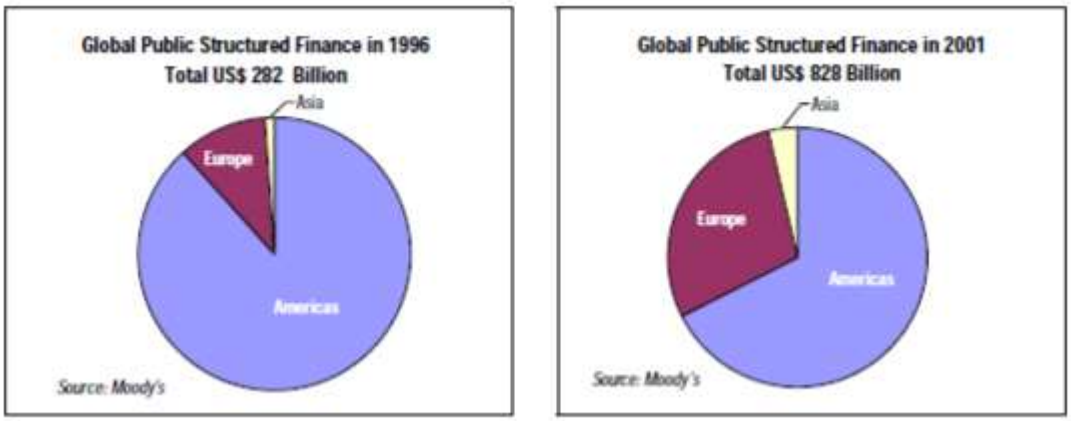
Villani, C., Schoenauer, M., Bonnet, Y., Berthet, C., Cornut, A., Levin, F., Rondepierre. B. *For A Meaningful Artificial Intelligence*. Paris: French Parliament, Mar. 2018. PDF.

Weil, Ariel, and Thorsten Klotz. "Moody's: Growth in Europe's Peer-to-peer Lending Could Benefit ABS Market." *Moodys.com*. Moody's Investor Service, 07 June 2015. Web. 25 Apr. 2020.

https://www.zopa.com/invest/risk/markets

APPENDIX

Figure 8 Moody's report of historical growth of securitization market from 1996-2001



ZOPA DATA CATEGORIES

Data Categories Form Zopa

Attribute	Explanation
Loan ID (encrypted)	Unique identifier
Borrower(encrypted)	Duplicates detected
Disbursal date	Date that loan was provided
Original loan Amount	Amount borrowed
Principle collected	Total principle collected, regardless of status
Interest collected	Total interest collected on balance
Total number of payments	Payments made (can include several payments in one month

Last payment date	
Term	Original length of loan
Lending rate	Interest rate assigned to loan
Latest status	Complete, default, active, late
Date of default	If defaulted
Post code	Regional postcode ID

FULL TRAINING SET WEKA

Data Set 1	Explanation
Encrypted Loan ID	Value changed from encrypted string to integer for memory purposes
Encrypted Borrower ID	Original encrypted string kept incase relationship was found due to repeating values identified
Disbursal date	Date loan was provided
Original Loan Amount	Principle amount applied for and given to borrower
Principal Collected	Amount paid toward principle balance outstanding
Interest Collected	Amount paid toward interest balance outstanding
Total number of payments	Cumulative number of payments made
Last payment date	Date on which most recent payment was recorded
Term	Agreed number of installments to be paid (per month)
Lending rate	Interest rate on borrowed principle
PostCode	Regional postcode of borrower

country	Country related to postcode's units
Pop	Sum population of postcode's units
House	Sum of households in postcode's units
AVGIndx	Average of IMD, calculated on sum of IMD score, divided by number of units in postcode
CALC	Calculated interest based on provided rate
PPF	Prepayment factor identifies is for completed loans, all others are labeled with a constant '0'. If the 1, then loan was completed early, if 2, loan was on time, and 3 is for loans completed paste their due date.
DUE DATE	Date loan should be completed: Disbursal date + Term
MonthsOnBook	Only for defaulted loans: Last payment date – Disbursal date
RateClass	Split loans into groups based on interest rate measures provided by Zopa
Latest Status	Most recent status of loan since February 2020

SECURITIZATION PROCESS MAPS

Figure 9 GCR, Securitization 101, Securitization Structure

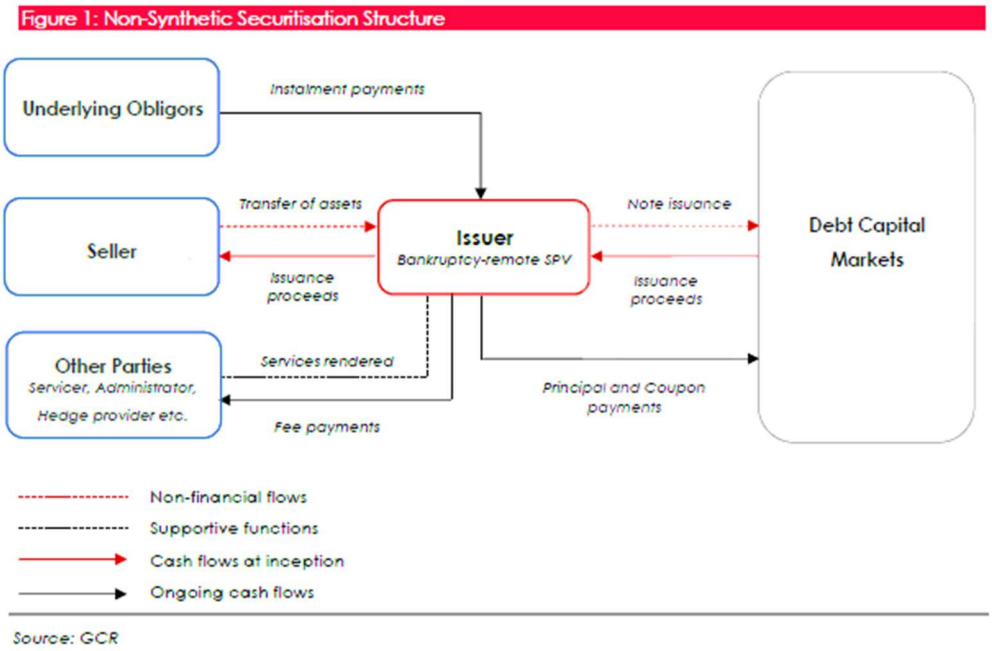


Figure 10 Zopa Prospectus, Diagrammatic Overview of the Transaction at Issue, 2016

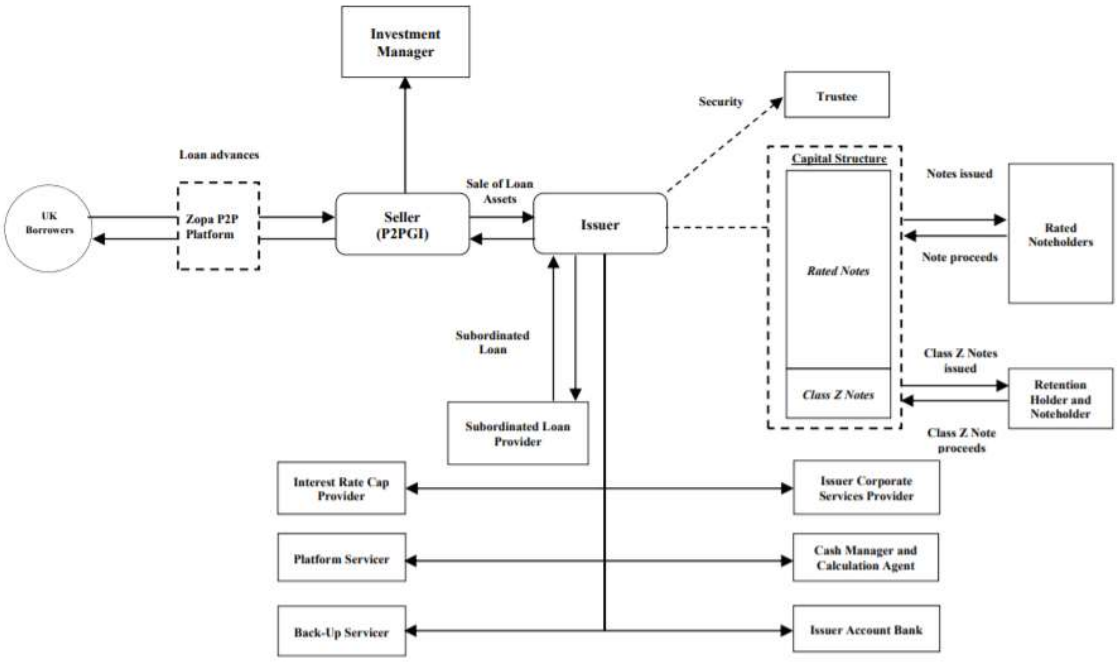


Figure 11 Zopa Prospectus, Diagrammatic Overview of Ongoing Cash Flow

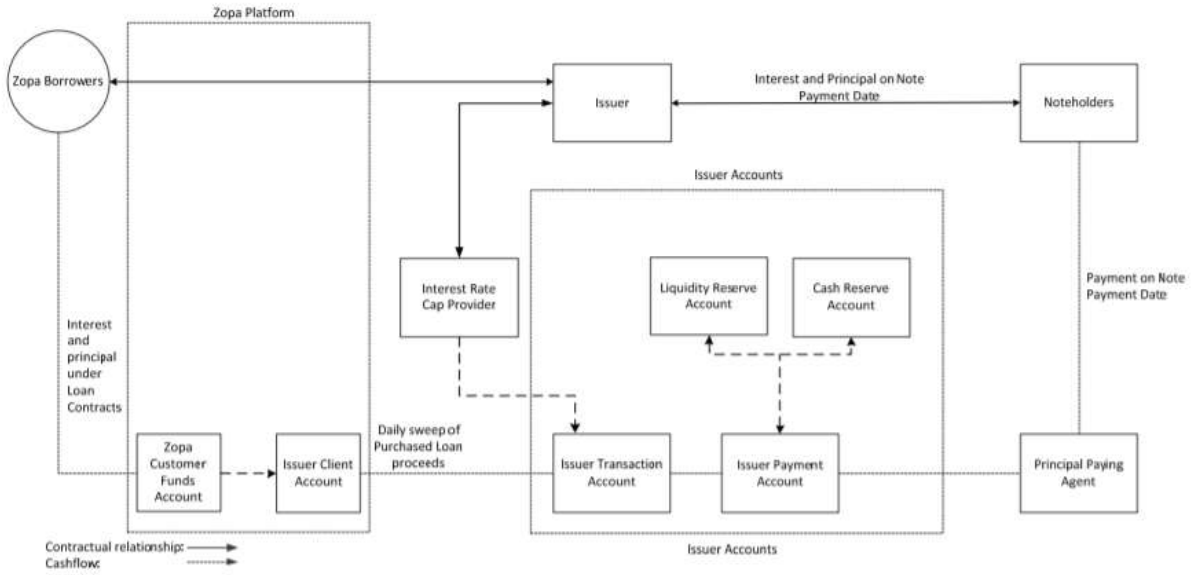


Figure 12 Moody's 3rd Approach to Ratings in Practice

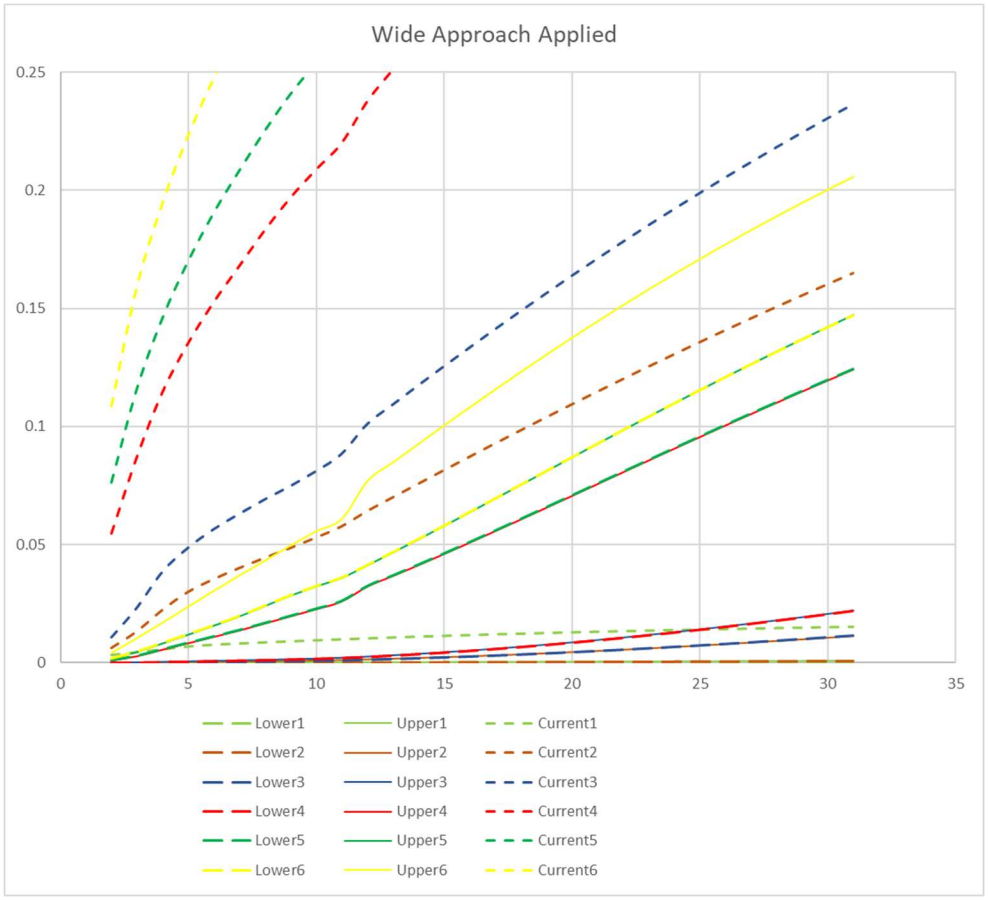
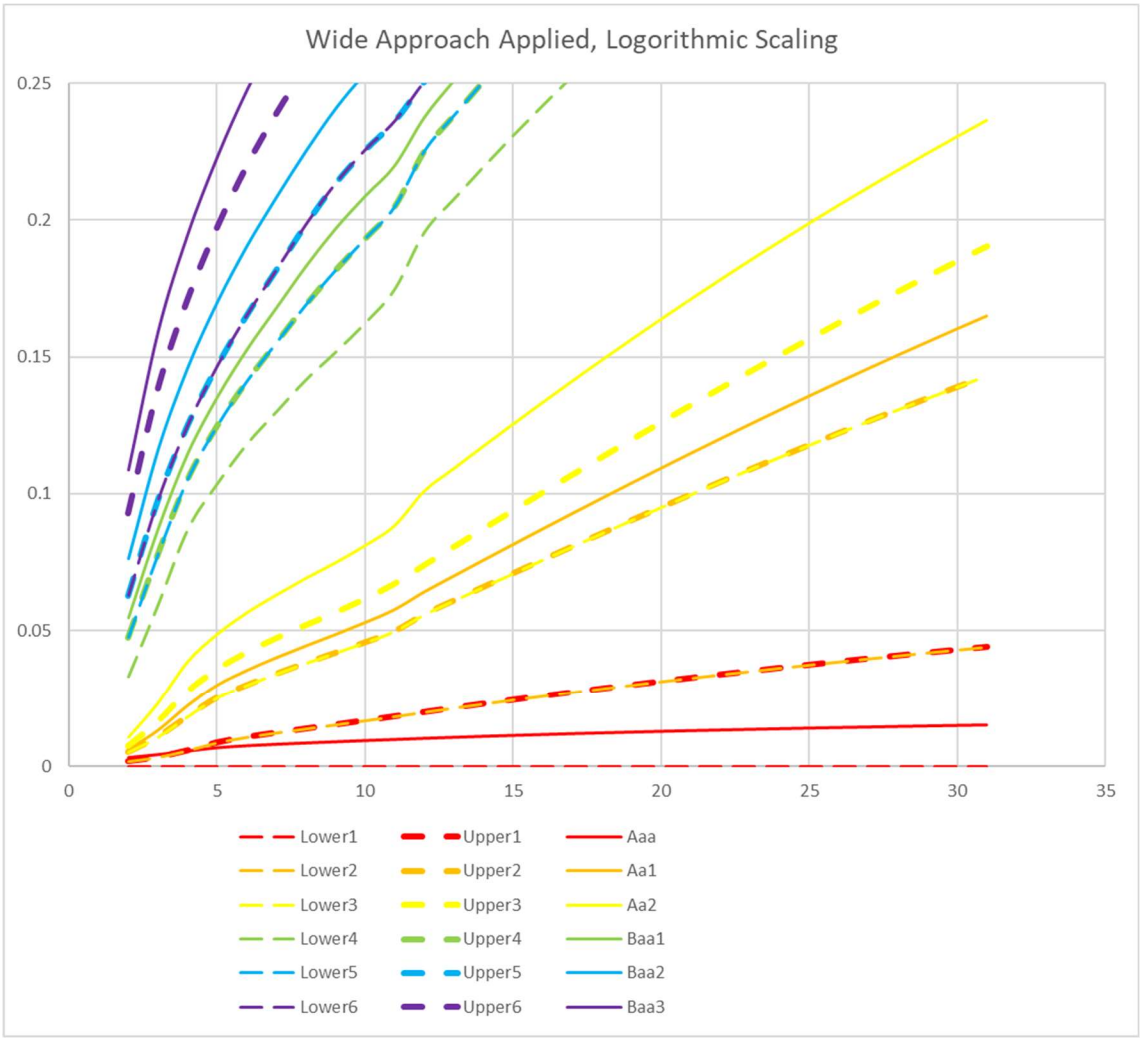


Figure 13 Applying Logarithmic Scaling to Approach 3



APPENDED RESULTS: WEKA SCREEN PRINTS

Figure 14 Logistic Regression: Binomial sample of 1195 instances

```
=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.814    0.122    0.870      0.814    0.841      0.693    0.893     0.860     Default
                 0.878    0.186    0.825      0.878    0.851      0.693    0.889     0.831     Completed
Weighted Avg.    0.846    0.154    0.847      0.846    0.846      0.693    0.891     0.846

=== Confusion Matrix ===

   a      b      <-- classified as
 486.24 111.26 |      a = Default
  72.87 524.63 |      b = Completed
```

Figure 15 Naive Bayes: Binomial sample of 1195 instances

```
=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.822    0.055    0.937      0.822    0.876      0.773    0.951     0.941     Default
                 0.945    0.178    0.842      0.945    0.890      0.773    0.951     0.937     Completed
Weighted Avg.    0.884    0.116    0.889      0.884    0.883      0.773    0.951     0.939

=== Confusion Matrix ===

   a      b      <-- classified as
 491.39 106.11 |      a = Default
  33.03 564.47 |      b = Completed
```

Figure 16 J48: Binomial sample of 1195 instances

```
=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.750    0.011    0.985      0.750    0.852      0.761    0.923     0.938     Default
                 0.989    0.250    0.798      0.989    0.883      0.761    0.923     0.888     Completed
Weighted Avg.    0.869    0.131    0.892      0.869    0.867      0.761    0.923     0.913

=== Confusion Matrix ===

   a      b      <-- classified as
 448.12 149.38 |      a = Default
   6.8  590.7  |      b = Completed
```

Figure 17 Random Forest: Binomial sample of 1195 instances

```
=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.722    0.138    0.839      0.722    0.777      0.590    0.883     0.891     Default
                 0.862    0.278    0.756      0.862    0.806      0.590    0.883     0.867     Completed
Weighted Avg.    0.792    0.208    0.798      0.792    0.791      0.590    0.883     0.879

=== Confusion Matrix ===

   a   b   <-- classified as
 432 166 |   a = Default
  83 515 |   b = Completed
```

Figure 18 Logistic Regression: Multinomial sample of 1000 instances

```
=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.987    0.123    0.774      0.987    0.867      0.814    0.979     0.946     Active
                 0.829    0.068    0.951      0.829    0.886      0.743    0.949     0.937     Completed
                 0.688    0.027    0.688      0.688    0.688      0.660    0.928     0.794     Default
                 0.000    0.004    0.000      0.000    0.000      -0.005   0.519     0.049     Late
Weighted Avg.    0.859    0.080    0.871      0.859    0.858      0.753    0.953     0.922

=== Confusion Matrix ===

   a   b   c   d   <-- classified as
 294   3   1   0 |   a = Active
  78 510  23   4 |   b = Completed
   5  20  55   0 |   c = Default
   3   3   1   0 |   d = Late
```

Figure 19 Naive Bayes: Multinomial sample of 1000 instances

```
=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.987    0.068    0.860      0.987    0.919      0.885    0.982     0.930     Active
                 0.914    0.018    0.988      0.914    0.949      0.880    0.978     0.989     Completed
                 0.950    0.009    0.905      0.950    0.927      0.921    0.987     0.970     Default
                 0.143    0.004    0.200      0.143    0.167      0.164    0.834     0.141     Late
Weighted Avg.    0.933    0.032    0.937      0.933    0.933      0.880    0.979     0.964

=== Confusion Matrix ===

   a   b   c   d   <-- classified as
 294   2   0   2 |   a = Active
  43 562   8   2 |   b = Completed
   1   3  76   0 |   c = Default
   4   2   0   1 |   d = Late
```

Figure 20 J48: Multinomial sample of 1000 instances

```
=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.970    0.037    0.917      0.970    0.943      0.918    0.958     0.891     Active
                 0.961    0.075    0.953      0.961    0.957      0.888    0.936     0.935     Completed
                 0.763    0.004    0.938      0.763    0.841      0.834    0.865     0.744     Default
                 0.000    0.000    ?          0.000    ?          ?        0.660     0.011     Late
Weighted Avg.    0.941    0.058    ?          0.941    ?          ?        0.935     0.900

=== Confusion Matrix ===

   a   b   c   d   <-- classified as
 289   6   3   0 |   a = Active
  23 591   1   0 |   b = Completed
   3  16  61   0 |   c = Default
   0   7   0   0 |   d = Late
```

Figure 21 Random Forest: Multinomial sample of 1000 instances

```
=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.977    0.013    0.970      0.977    0.973      0.962    0.996     0.986     Active
                 0.987    0.190    0.893      0.987    0.937      0.832    0.978     0.984     Completed
                 0.250    0.000    1.000      0.250    0.400      0.484    0.980     0.873     Default
                 0.000    0.000    ?          0.000    ?          ?        0.574     0.009     Late
Weighted Avg.    0.918    0.120    ?          0.918    ?          ?        0.980     0.969

=== Confusion Matrix ===

   a   b   c   d   <-- classified as
 291   7   0   0 |   a = Active
   8 607   0   0 |   b = Completed
   0  60  20   0 |   c = Default
   1   6   0   0 |   d = Late
```

Figure 22 Naive Bayes: 50,0000 instances, Binomial

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances       42052               84.104  %
Incorrectly Classified Instances      7948               15.896  %
Kappa statistic                          0.6821
Mean absolute error                      0.1906
Root mean squared error                  0.3454
Relative absolute error                 38.1257 %
Root relative squared error             69.0845 %
Total Number of Instances            50000

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.791    0.109    0.879      0.791    0.833      0.685    0.920     0.912     Default
                 0.891    0.209    0.810      0.891    0.849      0.685    0.920     0.918     Completed
Weighted Avg.    0.841    0.159    0.844      0.841    0.841      0.685    0.920     0.915

=== Confusion Matrix ===

     a     b   <-- classified as
 19780  5220 |     a = Default
  2728 22272 |     b = Completed
```

Figure 23 J48: 50,0000 instances, Binomial

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances       45869               91.738  %
Incorrectly Classified Instances      4131                8.262  %
Kappa statistic                          0.8348
Mean absolute error                      0.1331
Root mean squared error                  0.2584
Relative absolute error                 26.629  %
Root relative squared error             51.6873 %
Total Number of Instances            50000

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.837    0.002    0.997      0.837    0.910      0.846    0.949     0.962     Default
                 0.998    0.163    0.860      0.998    0.924      0.846    0.949     0.921     Completed
Weighted Avg.    0.917    0.083    0.928      0.917    0.917      0.846    0.949     0.942

=== Confusion Matrix ===

     a     b   <-- classified as
 20931  4069 |     a = Default
    62 24938 |     b = Completed
```

Figure 24 Naive Bayes: 50,0000 instances, Multinomial

```
=== Summary ===

Correctly Classified Instances       45149               90.298  %
Incorrectly Classified Instances      4851                9.702  %
Kappa statistic                          0.8212
Mean absolute error                      0.064
Root mean squared error                  0.1975
Relative absolute error                 23.4715 %
Root relative squared error             53.4782 %
Total Number of Instances            50000

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.934    0.085    0.922      0.934    0.928      0.849    0.967     0.966     Completed
                 0.940    0.045    0.940      0.940    0.940      0.895    0.976     0.972     Active
                 0.349    0.017    0.488      0.349    0.407      0.390    0.908     0.417     Default
                 0.085    0.015    0.052      0.085    0.065      0.056    0.812     0.041     Late
Weighted Avg.    0.903    0.064    0.903      0.903    0.902      0.841    0.967     0.936

=== Confusion Matrix ===

     a     b     c     d   <-- classified as
 24201   784   621   307 |     a = Completed
   811 20146   124   354 |     b = Active
  1116   244   762    61 |     c = Default
   121   253    55    40 |     d = Late
```

Figure 25 J48: 50,0000 instances, Multinomial

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances       46955               93.91   %
Incorrectly Classified Instances      3045                6.09   %
Kappa statistic                          0.8856
Mean absolute error                      0.0491
Root mean squared error                  0.1686
Relative absolute error                 17.9878 %
Root relative squared error             45.657  %
Total Number of Instances            50000

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.961    0.046    0.957      0.961    0.959      0.915    0.970     0.961     Completed
                 0.998    0.055    0.931      0.998    0.963      0.936    0.985     0.971     Active
                 0.275    0.006    0.674      0.275    0.390      0.415    0.775     0.349     Default
                 0.119    0.001    0.479      0.119    0.191      0.236    0.744     0.115     Late
Weighted Avg.    0.939    0.048    0.929      0.939    0.929      0.896    0.966     0.931

=== Confusion Matrix ===

     a     b     c     d   <-- classified as
 24910   746   238    19 |     a = Completed
    24 21389     9    13 |     b = Active
   970   584   600    29 |     c = Default
   118   252    43    56 |     d = Late
```

Figure 26 J48, 50,000 Data set, Multinomial Root

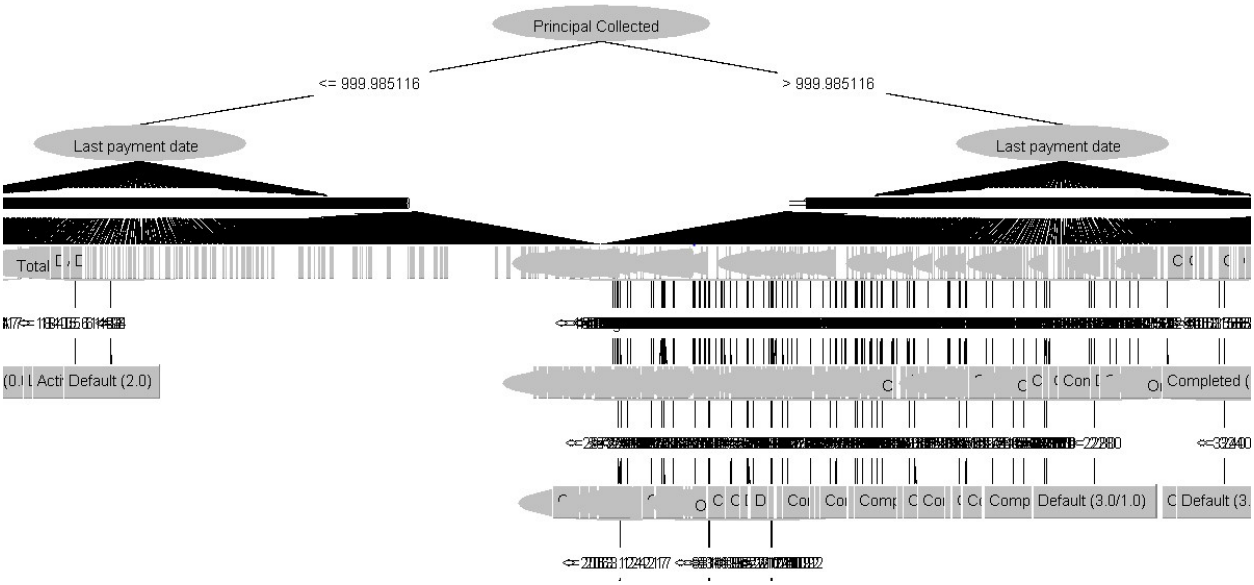


Figure 27 Structural overview of J48 Multinomial decision tree

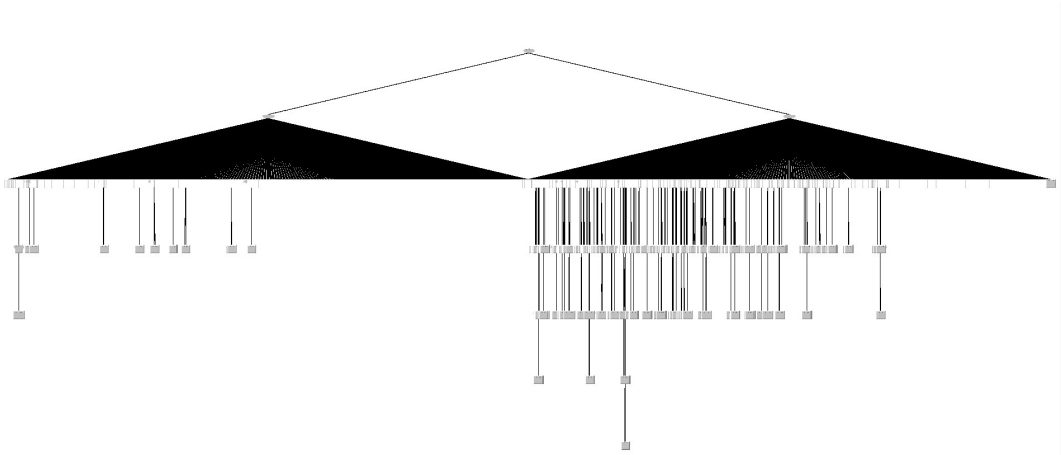


Figure 28 J48, 50,000 Data set, Multinomial Split on Disbursal date decision node

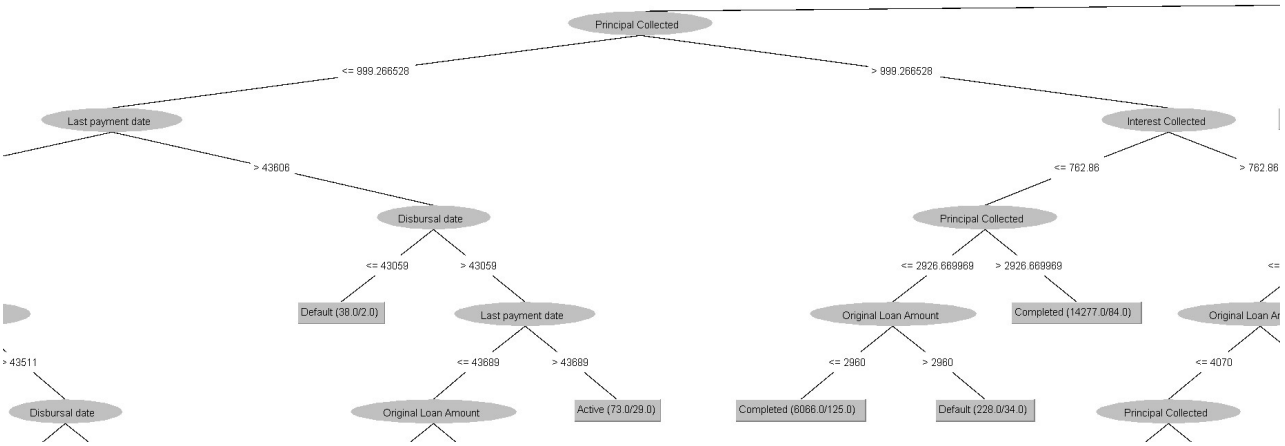


Figure 29 J48, 50,000 Data set, Binomial, Image 1 of 2

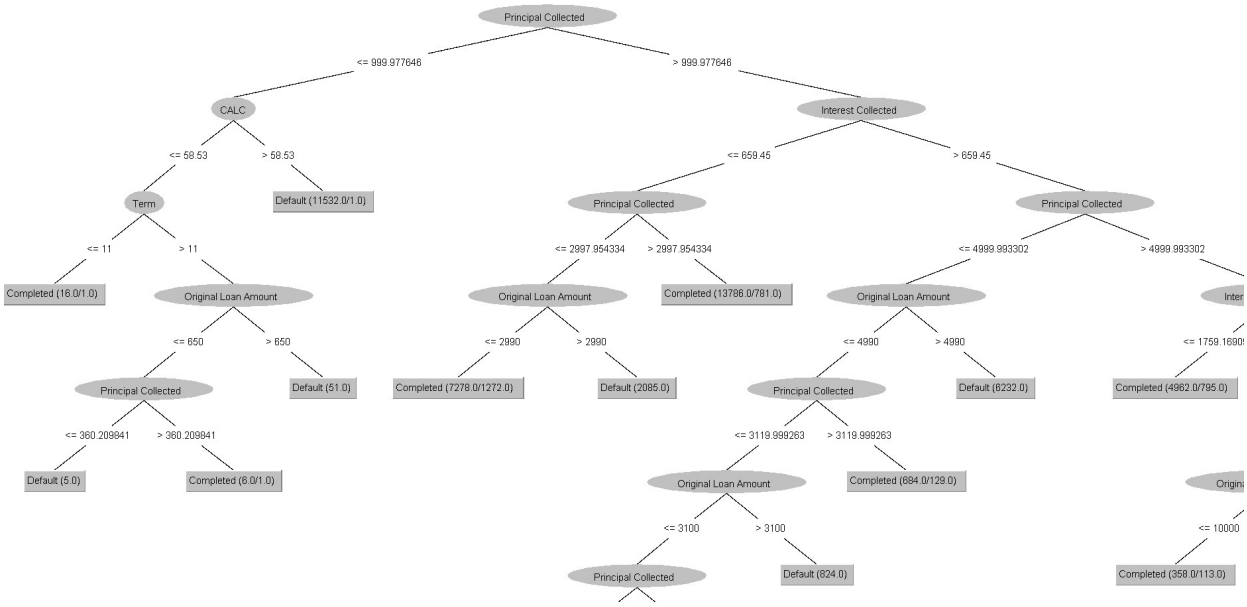


Figure 30 J48, 50,000 Data set, Binomial, Image 2 of 2