CBS COPENHAGEN BUSINESS SCHOOL
HANDELSHØJSKOLEN

Hi stranger, can you recommend a restaurant?

- An investigation of social recommendations using a latent space model

15th September, 2020
Master´s Thesis
Elena Bundgaard Noltensmeier (102534)
MSc in Business Administration and Information Systems
Characters: 82.041
Pages: 82.041 / 2.275 =36

Supervisor:
Assistant professor
Thomas Frick
Department of Digitalization

Co-supervisor:
External Lecturer
Nicolai Frost Jacobsen
Department of Digitalization

Abstract

This paper investigates how a latent space model with link prediction as the network evaluator performs if implemented as a social recommendation system for restaurants on the Yelp platform. The model is investigated with both users and restaurants as the nodes of the networks. The model used is an optimized latent space model built for large scale networks by Nicolai Frost Jacobsen (2018) to be utilized in his master's thesis, *Large scale latent variable modelling for link prediction in complex networks*. Latent space models have had much success in other areas such as friend recommendations, movie recommendations and even proteins network. The latent space model and link prediction model performed well on all the networks investigated, though the conditions they performed well under indicated that the latent space model might not be the right fit to recommend restaurants within a large-scale network across states. The latent space became an expression of physical distance as the nodes within each state clustered together. The computational costs are too high compared to the value it creates in the form of social recommendations.

Executive summary

Latent information is hiding in the social networks people form and these latent variables might provide new insights about users or restaurants, information that cannot be obtained through analyzing the datasets on regular terms. Investigating large scale networks has a high computational cost and this might be one of the reasons that latent space networks are not the most applied. This paper investigates how a latent space model with link prediction as the network evaluator performs if implemented as a social recommendation system for restaurants on the Yelp platform.

The model investigated is a Latent Space Log-Likelihood model created by Nicolai Frost Jacobsen to be utilized in his master's thesis, *Large scale latent variable modelling for link prediction in complex networks*. The model is applied on several datasets. A large dataset containing data across the available American states; a dataset where data reduction is applied to the dataset with all the available states, the city Charlotte; and a reduced dataset of the city Charlotte dataset. When using a latent space model to model the networks of restaurants across states, the latent space becomes an expression of distance which makes it easier for the model to make good predictions. All train and test scores across the datasets were between 0,83 and 0,97, where train and test score within a pair only varied with 1 percent point. Despite good prediction rates, this provides poor recommendations. This problem was identified when the two-dimensional latent space was visualized, and the nodes were colored with a color representing their state. All states except Nevada in the reduced dataset of all American states highly clustered together. Changing the scope of the investigated data to only focus on the city Charlotte created a network of nodes, where it was not as prominent that location had a huge impact on the latent space as in the previous network. Upon applying multiple dimensions to the network, the train and test scores improved which indicates more latent information was exposed with more dimensions. A deeper investigation of the nodes in the Charlotte network would have been a help to investigate the network at local level and could have provided greater insight into how more dense and overlapping networks would perform, especially if weighing the links and doing data reduction, creating less overlapping and maybe intensify the patterns in the networks. This would be the next step in future work.

The computational costs are too high for the latent space model to be applied successfully as a social recommendation system. When built as one network across states, location weighs too heavy on the latent space, making it not scalable. Therefore, datasets based on smaller areas should be

implemented, which are even more computationally expensive than to implement one large model. Moreover, due to the computational costs, it should not be deployed as a live model as it is too heavy to run real time. The model that only contained 100 random users and the restaurant they had visited took around half an hour to compute and a webpage is expected to load within at least a couple of seconds.

Disclaimer

This paper was written during the period of start 2020 with the deadline of 15th September 2020. During this period the virus Covid-19 was a big concern. At one point all public institution closed down libraries and universities included. A computer and GPU that was borrowed from CBS just before the lockdown unfortunately was locked up with the rest of CBS's facilities. Therefore, another solution was sought, and a computer and a GPU were borrowed from a person in the author´s private network. This solution came with some restraints as the owner of the computer used it for work when working from home. Therefore, it was a limited amount of time the models could run over. Due to the circumstances the author decided to focus and investigate the publicly available dataset from Yelp.

Table of content

Introduction

Ranking information is important today more than ever. Let me ask you: “When did you last go to the second page on a google search?” We are creating data at a speed where no one can keep up and we are doing this together, by collectively creating large scale networks. Yelp is a crowdsourced platform where the crowd is delivering information in the form of reviews and tips in exchange for a platform that provides the wisdom of the crowds and some extra services. Yelp is a great example of a platform with a large-scale network where the information is in the center, and how they visualize and provide the data to the user is of great importance.

Recommendation systems have shown to be a crucial part of the user experience for online companies in order to sort through the data that the platforms make available to the users. Leading companies, most notably Amazon, YouTube, and Netflix, have demonstrated the value of recommendation systems and have radically transformed what customers expect from a digital experience (Blueshift, 2017).

Recommending restaurants is an interesting machine learning case because of its high practicality and rich context. Every dining experience is unique, subjective and composed of numerous factors. In this paper, we will take a closer look at recommending restaurants with the help of link prediction applied to a latent space model that has modelled a network from the available data. Building the model on connections rather than variables allows this paper to focus on the unspoken similarities between restaurants rather than price, type of kitchen, etc. The data investigated in this paper is from Yelp and the goal is to predict relations between the nodes in the network created by the latent space model using link prediction as the evaluator. Therefore, the research question investigated is:

How does a latent space model perform as a social recommender for the restaurants on the platform Yelp?

The work process behind this research paper was built around the CRISP process and therefore the structure of the paper resembles this process. First, all the literature used for this paper is introduced. The literature is split into two areas: the more technical literature that needs to be

explained in order to understand the machine learning models, and the literature behind recommendation systems in order to understand how this field of work is approached in general. Yelp will be analyzed to help understand the data. Next, the data preparation is explained in order to understand the process of selecting the data. After this, the models and the flow of developing the models will be explained before the whole process will be evaluated prior to the final conclusion and future work of the research being presented.

Literature review

Software and hardware

To investigate the data and build the model, the software Jupyter Notebook was used. Jupyter Notebook is an open-source web-based interactive computing notebook environment with the default programming language python. This is a tool where you can execute human readable documents as you can add text documents without interrupting the executed code in between the code, thus making this software a great tool for data analysis as you can execute and write about the processes in the same document. A lot of libraries built for machine learning are also available when programming in python.

The minimum specs recommended for a computer if you wish to do machine learning is 16 GB RAM. Also, the CPU processor is advised to be or be above Intel Corei7 7th Generation as it is more powerful and delivers high performance. When working on deep learning models, a GPU is indispensable as the matrices created to mimic neural networks are computationally expensive. The utilization of a GPU enables parallel processing of these matrices and significantly shortens the time it takes to run the models. According to Kislay Keshari, a Big Data and Data Science expert, it can go from days and months to hours (Keshari, 2020).

Two different computers were used for this project as it was only possible to borrow a computer with the specs for machine learning for a limited amount of time. The specs for both computers have been set up in Table 1 below.

Table 1 Computer specs

Type	**Macbook pro**	**Windows**
RAM	8 GB	16 GB
Processing (CPU)	2,9 GHz Dual-Core Intel Core i5	3,6 GHz Intel(R) Core(TM) i9-9900
Graphics (GPU)	Intel Iris Graphics 6100 1536 MB	NVIDIA Geforce RTX 2070
Used	For all phases except when modelling	Modelling phase

Adjacency matrix

A regular matrix describes the relationship between two variables while an adjacency matrix identifies whether or not a node-link relation exists (Weisstein, 2020). An example of an adjacency matrix can be seen in Table 2. The rows and columns of the adjacency matrix are the nodes of the network and the value 1 indicates that a link between the two nodes exists and a 0 indicates no connection. This data is visualized in Figure 1. The example of the adjacency matrix is an undirected graph as this is what will be investigated in this paper, meaning that the links are bidirectional and therefore half of the matrix is not filled out. An example of a directed graph could be a platform where you follow a user rather than befriending them, in this example a direct link is important to be able to identify.

Table 2 Adjacency matrix example undirected network

	1	2	3	4	5
1	-	-	-	-	-
2	1	-	-	-	-
3	0	1	.-	-	-
4	1	0	0	-	-
5	1	1	1	1	-

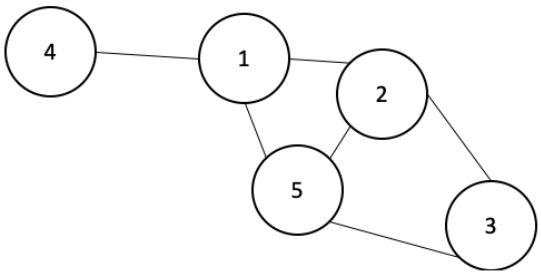


Figure 1 Network visualizing the adjancency matrix from Table 2

Latent space model and Link Prediction

The latent space model is an unsupervised machine learning model. This model has proven to make good predictions in multiple fields such as friend recommendations (Provost & Fawcett, 2013, p. 303), movie recommendation (Provost & Fawcett, 2013, p. 307), and criminal networks (Jacobsen, 2018, p. 7) and. Within the data science field, the term 'latent' means "*Relevant but not observed explicitly in the data"* (Provost & Fawcett, 2013, p. 27). The latent variable is inferred from other variables which are observable. The Latent space model is built on the primary assumption that the more alike two nodes in a network are the closer together they will be placed to each other. In other words, the distance between nodes in the network i.e. the latent space, is a metric for similarity and thus parity between the nodes (Jacobsen, 2018, p. 11). In this research paper, the observable data is connections between restaurants and users. These connections are noted in an adjacency matrix representing the social network. Both restaurants and users will act as both nodes and links in different models to see how the performance of the model changes due to this. The Latent space model builds the latent space between the nodes of the social network. A direct link between two nodes is an indication of a connection. It is up to the model to place all the nodes according to each other, visualizing the network. The strength of this model is that it can be used to look into connections rather than a target value. It strips the data of information and only takes into account if you set foot in the restaurant, indicating some sort of likeness. As the latent space model is an unsupervised model i.e. there exists no target value to evaluate on, we need to find another way to evaluate the network. To do this we can make use of the supervised machine learning method Link prediction.

Evaluating the network

Link prediction is the inference of the existence of a link between two nodes based on the already existing links in the network and on the node's properties (Jacobsen, 2018, p. 15). The latent space model is built on links and we can therefore predict whether or not a link between two nodes in the latent space exists in order to evaluate how well the network is built and give an indication on how well this model can be used to predict possible new links in the future. The model evaluates links in the between the nodes, and uses the latent space as the nodes property and predicts whether or not a link or a non-link should be formed. When applying link prediction, 10 percent of the links equally

split into links and non-links are removed from the training of the model and kept as test data to evaluate the performance.

ROC scoring

The performance of the Link prediction model is based on how well the Link prediction model can recreate links in the network. To evaluate the performance of the link prediction model ROC scoring, Receiver Operating Characteristics curve, is introduced. This is a graphical plot that illustrates the model's ability to predict the links from non-links. The outcome of how a link or non-link is predicted can be separated into four sections better known as a confusion matrix. True negatives, true positives, false negatives and false positives (Provost & Fawcett, 2013, p. 190). This is also illustrated in the confusion matrix in Table 3 inspired from to book *Data Science for Business (Provost & Fawcett, 2013, p. 189)*

Table 3 Confusion matrix

		Actual	
		Positive	Negative
Predic-	Yes	True Positive	True Negative
ted	No	False positive	False Negative

TRUE NEGATIVES (TN) 0 was predicted, 0 was the right answer. Correctly predicted the class is negative.

TRUE POSITIVE (TP) 1 was predicted, 1 was the right answer. Correctly predicted the class was positive.

FALSE NEGATIVES (FN) 0 was predicted, 1 was the right answer. Wrongly predicted the class was negative.

FALSE POSITIVE (FP) 1 was predicted, 0 was the right answer. Wrongly predicted the class was positive.

To calculate the accuracy of the model the true positive rate is calculated: $TPR = \frac{TP}{TP+FN}$

Then the false positive rate is calculated: $FPR \frac{FP}{FP+TN}$

AUC stands for Area Under the ROC Curve and is used to summarize the performance of the model (Provost & Fawcett, 2013, p. 219). This is the number that will be used to evaluate the models investigated in this paper.

Recommendation systems

The purpose of a recommendation system has been described as *"... reduce consumers' search costs in light of the increasing product variety on the Internet (Resnick and Varian 1997)*" (Zheng, Provost, & Ghosee, 2007, p. 1). The recommendation system is an information filtering system that should predict a rating or preference of the user in order to rank the information in such a way that the user finds it useful. Search engines are a great example. They use algorithms to predict the best possible fit between the search information and the ranked results. We expect the best matching results to be ranked the highest and if we cannot find the result on page one, we change our search query rather than go to page two to see if result 23 was better than the 20 first results on the first page.

A general problem that recommendation systems try to solve is the data overload problem. Data overload is a known problem within every industry that deals with data in some sort of way, let it be internal or external data. To get the most value out of the data, sorting through it in a smart way is necessary. Otherwise, important data is lost in the jungle of metric combinations and pages of results. Yelp is no stranger to this problem. Their product is information and they depend on new updated information to stay relevant. How they make the information available on the platform has a direct impact on the users' user experience. The algorithms used in recommendation systems are often content-based filtering, collaborative filtering or a hybrid (He & Chu, 2011, p. 1). An illustration of the content-based filtering approach and collaborative filtering approach can be seen in Figure 3. Other ways to build a recommendation system could be a popularity based system where you simply recommend the item that has the highest sold count or most views depending on how you classify popular. Another way could be a classification based recommendation system where you classify people into target users groups, based on age, gender, etc., and match the target

item to the target user group (M, 2019). Collaborative filtering is built on the assumption that people have the same likes and dislikes as people whom they historically have tended to agree with which is also known as homophily (Schafer, Frankowski, Herlocker, & Sen, 2007, p. 300). This

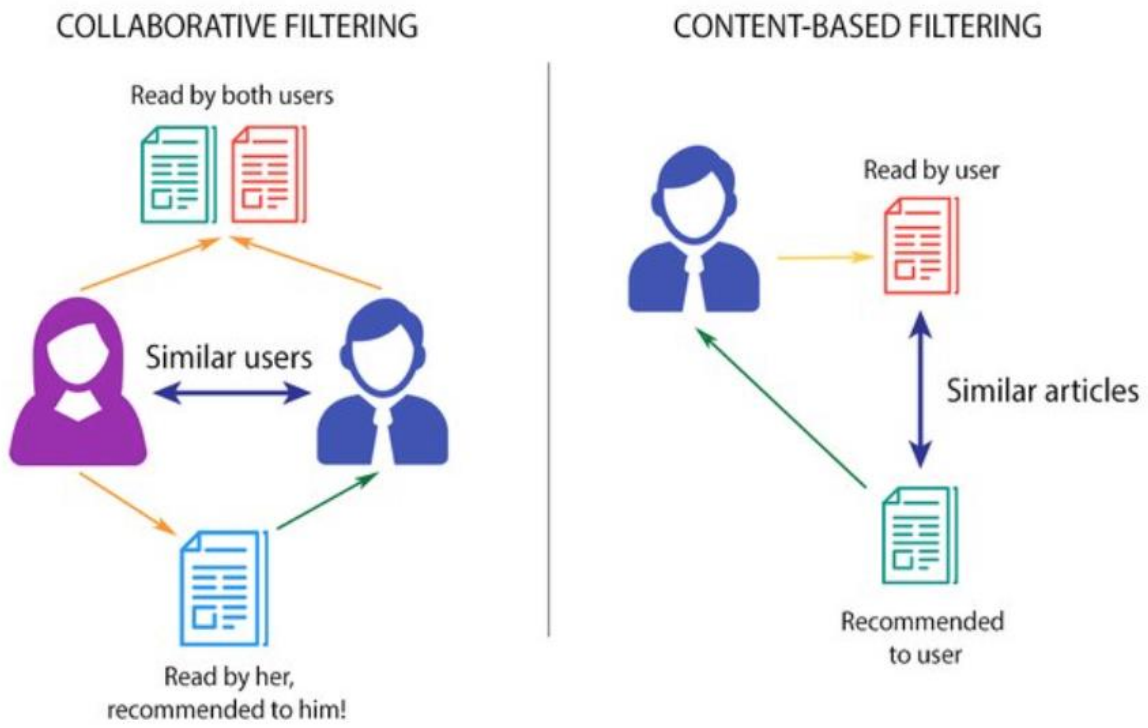


Figure 3 Examples of collaborative and content-based filtering

especially comes in handy when a platform has a lot of data from many users recommending what their friends like (Lee & Brusilovsky, 2018, p. 393). This approach is able to recommend more complex items such as movies without attributes to understand the item it is recommending these attributes could be genre, cast members, imdb score etc. (Schafer, Frankowski, Herlocker, & Sen, 2007, p. 300). A real recommender example of this could be when you are looking at an item on a web shop and they show you "similar items".

Social links are important, literature states that users of a platform tend to pay attention to the links/friends they have formed on the social platform, making them more receptive to the input from their links. When a consumer wants to buy a new product, the consumer tends to consult with their friends whom they know has experience with the product. The advice we take from our friends are considered truthful and can influence our decision (He & Chu, 2011, p. 2-3). Put in another way homophily explains that people tend to make connections with other people whom have similar

characteristics. May this be "*age, sex, religion, ethnicity, educational and occupational class, social positions, etc., in a process of 'social selection'*" (Lee & Brusilovsky, 2018, p. 398).

Content-based filtering, on the other hand, is based on the assumption that items with similar objective features will be rated similar (Chen, 2015, p. 100). A challenge with a content-based filtering approach is extracting the objective features that provides the best prediction.

Recommendation system challenges

Collaborative filtering techniques perform well when there is sufficient rating information. Though the lack hereof result in the data sparsity problem. The data sparsity problem where the data in the network is sparse. Most real life networks are sparse. (Chen, 2015, p. 100) (Zheng, Provost, & Ghosee, 2007, p. 2). With a rapidly increasing number of users coming online and joining the platform the cold start problem and the data sparsity problem have been increasingly intractable. The Cold start problem is when there is no data regarding users or items (Zhao, Qian, & Feng, 2014). Collaborative filtering approaches are also known to be computationally expensive because to built the model all target user's taste is compared with all other users (Lee & Brusilovsky, 2018, p. 397-398). The sparse data in a collaborative filtering model creates the challenge of accurately measuring user similarities based on a limited number of reviews. The problem occurs when there is a relatively higher number of items than users. Literature highlights that this is especially a hard model to implement if the items have a short life cycle such as job openings, events and news articles. These items might simply have too little time to accumulate enough ratings before their value expires (Lee & Brusilovsky, 2018, p. 398). The problem also relates to the cold start problem. Even for a system that is not particularly sparse, when a user initially joins, the system has no reviews from this user. Therefore, the system cannot accurately interpret this user's preference (He & Chu, 2011, p. 2).

Another problem with collaborative filtering is that a model like this is more vulnerable to shilling attacks and copy-profile attacks. The shilling attack occurs when a business user wants to reinforce their establishment's rating or dethrone a competitor and intentionally distorts recommendation predictions to their own advantage. A malicious user can create multiple profiles and create fake user-item ratings to achieve their desired goal (Lee & Brusilovsky, 2018, p. 397). This is considered illegal on all the review sites and they will try to flag them and if it can be proven they will be removed from the platform. Though a good example of how YELP is not always able to detect fake

reviews is with the story 'The Shed at Dulwich' (Rosenberg, 2007). This was London's top-rated restaurant. Just one problem: It didn't exist. All reviews and pictures were all fake and the food pictures were not even edible as the food was made with bleach tablets and shaving cream. Malicious users can also target the recommendations of a specific user by copying their reviews, creating a similar rating profile and thereby having the collaborative filtering method pick them out as a perfect peer and therefore what new items they may have rated will be suggested to them.

Methodology

This paper and model has been developed with a pragmatic view using one of the most common methodologies within data mining. The cross-industry standard process for data mining (CRISP) provides a structured iterative approach to planning data mining projects. The diagram portrayed in the book *Data Science for Business: What you need to know about data mining and data-analytic thinking* by Provost and Fawcett (2013) can be seen below in Figure 4. The model consists of six phases: Business Understanding, Data Understanding, Data Preparation, Modelling, Evaluation and Deployment. Creating a model ready for deployment is one large iterative process until the model is declared ready for deployment. After deployment, the model should still be revised and when flaws or new insights are discovered, a new iteration of the model should be initiated in order to optimize the performance.

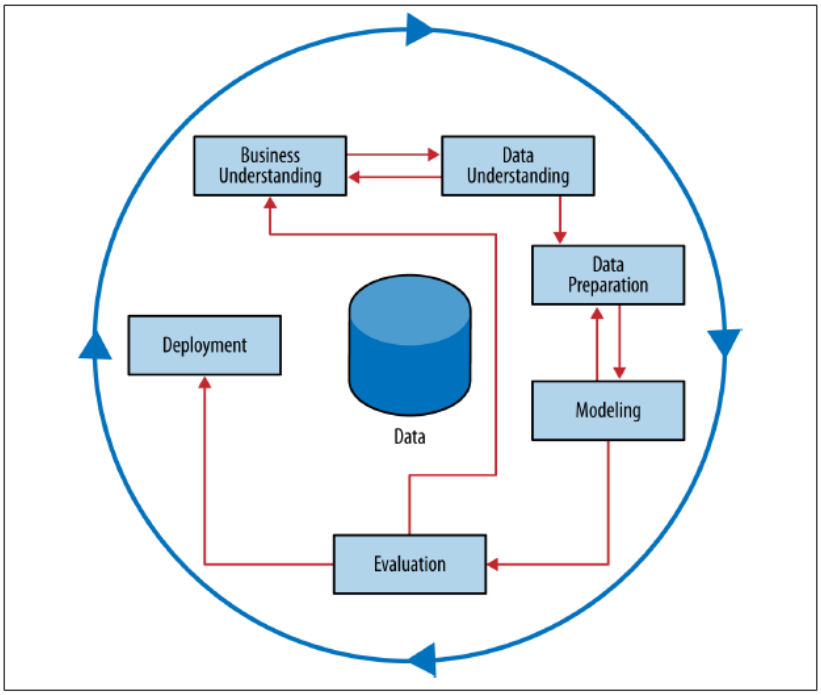


Figure 4 CRISP

Iterations between phases are also important as they strengthen the understanding of each other, specifically Business understanding and Data understanding, and Data preparation and Modelling. Provost & Fawcett (2013) describe: "Often the entire process is an exploration of the data, and after the first iteration the data science team knows much more" (p. 27). Not all iterations lead to a

product ready to be deployed. It takes multiple iterations and gaining of new insights to end up with a model ready for deployment. The many iterations built into the process are to make sure no stone is unturned as the possibilities within this field are endless. It all comes down to how well-informed the decisions made were (Provost & Fawcett, 2013, p. 27-33).

In the first phase of the CRISP model, the Business Understanding phase, the business objective is determined, data mining goals are determined, and an alignment between these two are sought. The next phase is Data Understanding. In this phase, the focus is on collecting the initial data, describing it, exploring it and verifying the quality of the data. The understanding of business and data is achieved through working with both phases in an iterative process. Understanding the business helps investigate and understand the data. New insights from the data help form the business understanding. This is highlighted as an especially important part of the process by Provost and Fawcett, as it is in these steps that it should be defined what exactly it is that we want to do and how (Provost & Fawcett, 2013, p. 28). As moving forward from this step you need to have a direction and the means to move in that direction. When the business understanding and data understanding is reached the model proceed to the Data preparation phase. In this phase the data is cleaned and formatted into a format that is easier for the model to comprehend. It is normal that datasets are with string values as people have an easier time understanding this. But for a machine learning model to be able to use the words in the database it must be translated into dummy variables creating integer values in the dataset. Two examples of a human readable variable transformed to a dummy variable can be seen below. In figure 5, there are only two values occurring within a column and are therefore changed into Boolean values within the same attribute column. If more values occur as in Figure 6, the values inside the attribute column is divided into their own attribute column and a boolean value is indicating whether or not this attribute is true for the instance. Other ways to create dummy variables is to hash a string, which is often done with id's when doing machine learning.

Gender
Male
Female
Female
Male

→

Gender
0
1
1
0

Figure 6 Example of binary dummy variables

Gender
None
Male
Female
Female

→

Gender1	Gender2	Gender3
1	0	0
0	1	0
0	0	1
0	0	1

Figure 5 Example of dummy values if multiple values

In the Modeling phase, the focus is on choosing, building and assessing the model, and generating a test design by which to assess the model. The model should reflect the business goal that was determined in the first two phases, Business understanding and Data understanding. The iterative process between Data preparation and Modelling ensures that the model is optimized and tweaked on all parameters, and that the best possible variable composition is chosen when evaluating the dataset or the model. The next phase is the Evaluation phase where the results from the models are evaluated, but also the overall process and how to proceed next. Is the model ready for deployment or should the process start over, going back to the Business Understanding. The last phase, Deployment, is where the deployment is planned, including planning about monitoring and maintenance, the project is reviewed and a final report is written. For this paper, the Deployment phase is redundant as it is not a real business case that shall be implemented, though implications with deploying this model will be discussed as a part of the Evaluation phase.

As mentioned, this research paper is structured so it is as true to the process as possible without going through the large iterations for readability. The CRISP process takes both the business aspect and model performance into account when developing the model, allowing for both the business needs as well as optimizing the performance to influence the final output. The iterative process creates a situation where there is a constant trade off between the business side and what the data and model can deliver to meet the expectations from the business.

Equipment

As mentioned in the disclaimer the computer with enough processer power to built the adjacency matrixes and latent space models were only accessible in shorter intervals of time lasting a couple of days. The computer has the specs as already introduced in the hardware software section and this was maxed out on multiple occasions which will be elaborated in the data preparation section.

Data

The dataset used for this paper is a publicly available dataset supplied by Yelp. There has been no direct contact with Yelp, which limits the amount of knowledge about the dataset to what Yelp has publicly released about the data. The dataset is a fraction of their actual data, both quantity wise but

also information wise. It might be because of legislation, privacy for their users, or business advantages that they have decided to exclude from the dataset. An example could be that the reviews in the dataset do not have an attribute that indicates whether or not the review is recommended. How a review becomes a recommended review will be explained in the Business Understanding phase under the Process section. Had I had a contact inside Yelp, I might have been able to collect further data or had certain inside knowledge that would lead me to take the model into a different direction. Because of this, some assumptions will be made on how the data is collected on the platform based on how the data is structured in the dataset.

The model

By Co-supervisor Nicolai Frost Jacobsen, it was advised to work with the model without adding weighted links between the nodes first as the focus of the process should be to investigate how the model did with only tweaking the dimensions and the data as the time for an assignment like this is rather limited when working with large scale network. Implementing weighted links is some iterations down the road after the general understanding of the data network has been gained. Therefore, the weight of the links in the network has not been implemented though it is believed that this would strengthen the model making better predictions because that is not where you start with a model.

Process

For this research paper, the general process was built around the CRISP process. This will reflect on the next section of the paper as the headings used will be named after the CRISP phases to understand the process from start to finish. First, the Business understanding is investigated and explained. The business understanding is not achieved through a collaboration with Yelp, therefore the business understanding has been achieved through an investigation of Yelps official information, such as Yelp FAQ sites, secondary sources and an investigation of the platform. Next, a section on the Data understanding will be elaborated. Then, the process of how the data was prepared will be explained. After that, the modelling of the mode. Then, the evaluation of the results and process and, lastly, the deployment. The Deployment phase is more theoretical for this paper as it will not be deployed but aspects of deployment will be discussed.

Business understanding

The design choices of the recommendation system are not only about optimizing the performance of the model itself. Every digital platform serves some kind of purpose and implementing a new feature should ideally support this purpose. In the first phase, Business Understanding, the platform Yelp was investigated to understand how to optimize the performance in a way that supports the platform and to find out what bias the platform gave the data that is used to build the model. The investigation of Yelp as a business was deemed necessary as only one dataset was investigated which heavily biased the model and, therefore, the results. To avoid the heavy bias, more datasets should have been investigated to see if the same patterns were across platforms or if the results only applied to Yelp. As there is no direct informant connected to this research paper, it was crucial to make a structured investigation of the platform to understand the features and how users are incentivized by Yelp to use their platform. This knowledge helps us understand the data structure and format of the dataset which will be explained in the next section Data Understanding. First, a general investigation was initiated to understand the structure and functions of the platform, the observations made were written down as seen in appendix 2. The observations from the platform were combined with the Yelp FAQ knowledge, which Yelp themselves provide and the articles and papers written about Yelp should combine to create an understanding of Yelp. The general

observations of the platform is elaborated in the next section to understand the general functions of the platform

Three types of users:

As a multisided platform, Yelp has multiple segments of customers, which their platform caters for. A multisided platform is defined as a platform with *"… two or more clearly distinct groups of platform users (E.g., content providers and consumers accessing the content)*" (Constantiou & Kallinikos, 2015, p. 233-234). It is important to identify the groups of the platform in order to be able to identify the potential business value a recommendation system can add to the platform for the individual user groups. It was decided to distinguish users into three groups based on the actions of the different users.

First, distinguishing between private people joining the platform to interact in the social network of the platform and the users that are created belonging to a business. Next, the private people were divided into two groups: the Active users and the Lurkers, the distinction of all three user types can be seen below:

- Business users: To create a business user the company must claim a business on the Yelp site before they can have a company user and employee users. The users created belong to the business, have a company tag and can reply to people who have interacted with their site. They can reply publicly as a comment directly on a review or in a private message sent directly to the user. They can also react to reviews written with a "thanks", which is sent privately to the user.
- Active users: people who have created a profile and have written at least one review. The distinction of "active" is made because of the next user type where you do not interact (lurkers). Having a profile is not interacting with the platform in this scenario as it is the activity of reviewing that can be tracked.
- Lurkers: It is common that a large part of the members of an online community do not participate but still find great value in the community. This is known as the 90-9-1 rule about partition inequality stating that 90 percent of people lurk, 9 percent are editing or modifying content and the 1 percent are content creators (Haklay, 2016).. Yelp has allowed people to view their content without creating an actual profile, creating the opportunity to simply lurk and not take part in the content creating or network building. The definition "lurker" also covers over the created users that do not create content (Nielsen, 2012).

It is important to identify the sides of the platform to identify the potential business value a recommendation system can add to the platform. What incentives it can create and create more value to the site.

In this paper, the knowledge about the businesses are collected through Yelp's FAQs and articles as it was not possible to not create a business profile and investigate the platform from that angle. Businesses can post general information about their business, for example opening hours and special offers. For a fee, the businesses can advertise with banners and search ads (Miller, 2009).The active users with a profile can review businesses, interact with other users and their reviews, create personalized profiles and achieve platform specific goals, such as yelp elite and write area tips. The Lurkers can access all the same webpages as the active user without creating a profile but without a profile you cannot write information on the yelp profile. Further details of the observations from the analysis of the platform can be found in appendix 2.

Yelp has implemented a lot of functionality in order to create a trusted environment. Their users are incentivized to create a profile with pictures and text about themselves. They enforce this behavior by having an algorithm that marks reviews as recommended or not-recommended and a deficient profile might be a reason for your review to be marked as not recommended. Yelp are incentivizing creating a profile with a lot of information that seems more trustworthy than a profile with a first name and no profile picture. Yelp Elite is also enforcing trustworthy profiles as you need an adequate profile in order to be considered for Yelp Elite (FAQ, What is Yelp's Elite Squad?, 2020).

Some studies mention Yelp as a tool that can make or break a retailer. Research has found a positive correlation between a higher rating and more customers during peak hours. The difference of 0,5 stars, averaging from 3 stars to 3,5, was shown to increase a restaurant's chance of selling out during prime dining times from 13% to 34%. Looking into the change from 3,5 to 4 stars on average increased the same chance by 19 percentage points. Prime time was set to 7 PM and the research paper looked into how many reservations were made compared to their capacity of tables (Anderson & Magruder, 2012). Therefore, it is in the businesses' interest to gain a high average rating score. This might lead some to write good reviews about themselves and bad about competitors. This is a known problem and there algorithms that are trying to filter these types of recommendations away. Incentivizing people to create user profiles and write detailed reviews is a way to deal with this type of problem as you cannot just copy-paste long prewritten reviews as they would get caught in the algorithm that recommends reviews.

For an easy overview of users' review behavior, a statistical fact box about the users' review habits is displayed on their profile as well as their written reviews. Yelp is incentivizing people to create a user that shows you as a person, your hobbies and likes, and displays each user's reviews on their user page. This should incentivize people to think twice about their language use and to fairly justify their reviews as they are not hiding behind an anonymized user. Yelp rewards their users when they engage in the platform with long reviews, pictures, connecting with people and reacting to other reviews, though they have to nominate themselves or be nominated by another user if they want to be invited to the inner circle. The users get rewarded with a badge for their Yelp-elite acquirement. This means that their reviews will be highlighted compared to reviews by others. At the same time, this reward system diminishes reviews made by people who are not very active or provide insufficient reviews. In general, this incentivizes users to be active and create adequate reviews, though it might also result in users trying to cheat the system by reviewing restaurants they did not visit to keep their elite badge. The regulator for behavior like this is the social network where people can react to each other's reviews, but at the same time users might form a network where they just like everything they put up because, as the saying goes, "If you scratch my back I'll scratch yours". To be rewarded as a Yelp elite, you must be evaluated by a physical panel that looks through your writing style, picture, quality, etc. All of this is subjectively evaluated by the panel to choose yelp elite members (FAQ, What is Yelp's Elite Squad?, 2020).It is not all the groups of users on the platform that Yelp are charging for their services. Yelp is a content crowdsourcing platform therefore they are dependent on their users to continuously add more data because the data looses the relevance with time as it should reflect the current service at the restaurant. This is also reflected in how the algorithm for top restaurants are formed, here the more up to date the review is the higher the star rating is weighed (FAQ, What is Yelp's recommendation software?, 2020).

The interrelationship between user groups on the platform determine the revenue model of the platform. Yelp has no value if they have no content and they have competitors that provide platforms where you can share reviews. Therefore, Yelp must subsidize the content creators in order for the business users to see value in being active on the platform and are willing to pay to participate. "Optimality will call for subsidies, [...] and one should subsidize more the less profitable side of the market" (Sanchez-Cartas & Leon, 2019, p. 3) It is free for businesses to join, creating a business profile and "claiming their page", but if they in any way want to interact with the people who have written the reviews on their page, they have to pay. Also, they can pay to get

an advertisement in the top list, or pay for services such as reservations and virtual queuing systems for their customers.

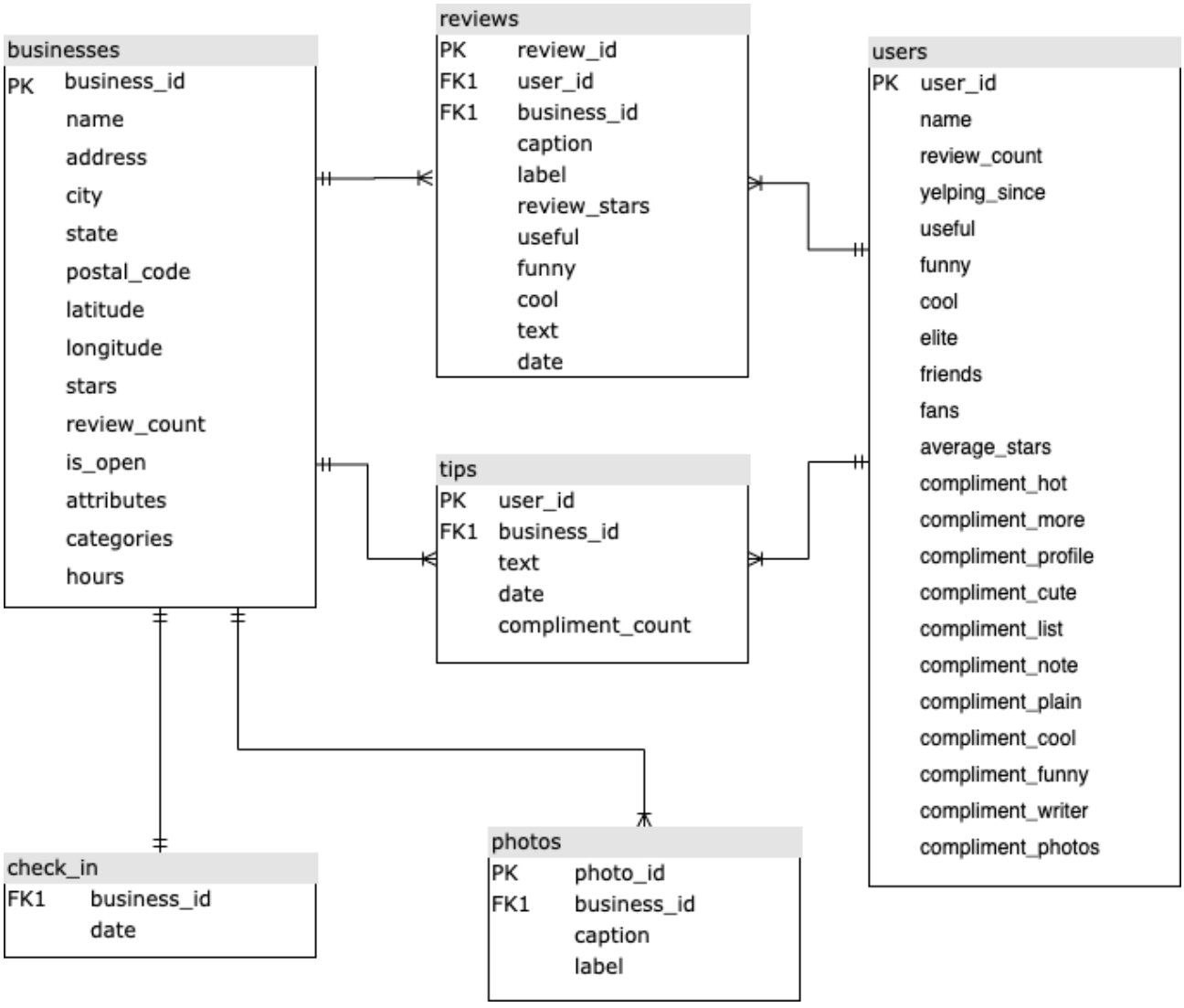
To summarize, Yelp can be described as a crowdsourced local business review and social networking site. Yelp has social networking functions where you can befriend people and incentivizes the active users to engage with each other, creating social networks amongst their users. As the businesses on the platform are delivering a physical service it is characterized as a local business review site because all searches are made within an area. You need to physically be able to transport yourself to the location, or a delivery guy from the location to you, for this site to provide value. An assumption is that you are looking for recommendations within a physical area because you are going to be physically present at some point and you should be able to consider all possible options near you. Yelp is interested in providing incentives for their users to be active on Yelp creating more content as this enable Yelp as a platform to charge for services that the business wants to have available on their profile on the platform as studies have showed that this increase your customer flow.

Data understanding

Now that a better understanding of Yelp as a business has been achieved, it is time to apply this to the data that is available for this paper. The data used for this paper was a publicly available Yelp dataset released for their yearly competition where students all over the world can conduct research and analysis on the dataset and submit their findings to compete for a prize from Yelp. The dataset used for this paper was from their competition in 2019. Yelp provided data distributed on 6 .json files containing data on users, tips, check-ins, photos, businesses and reviews. The .json files have been visualized in an entity relationship diagram as seen below in Figure 7 to help create an easy overview of the structure and available data. PK in the diagram stands for primary key and is a unique value for that row in the database table used to identify relationships across the database. The FK1 is short for foreign key and is a value that is unique for a single instance in another table. The entity diagram shows the attributes in each table and the relationship between the tables. The lines between the tables identify the relationship between the tables. For example, one business has many reviews but a single review is only written about one business and a user can write many reviews but the specific review is only written by one user (Lucidchart, 2020). The relationship between the businesses and check_in is 1:1. This is because the business only occurs once in the check_in table with an accumulated list of dates of the different check-ins. This accumulation was

probably done by Yelp when they released the data to make the file smaller. It is common to accumulate data in order to anonymize the users. Though as we have names and details on the users in another .json file, the accumulation might just be to compress the data. Comparing all the identified users and functionality with the data files, it becomes clear that only a fraction of the data Yelp is collecting is part of the dataset made available.

Figure 7 ER diagram



The review is an indication of a person dining at the restaurant creating a connection between the restaurant and the user. Looking at Figure 2, we can see a .json file named check_in. The reason that this paper focuses on the review as an indication of presence at the restaurant rather than the actual check in data is because the data in the check_in file is bound to a business in an accumulated

manner, where the only indication of the user is a timestamp making it impossible to identify which user has visited what place when and connecting this data.

The data was investigated by looking into the values of the attributes of each .json file. The files were loaded into python, the method head() was used to explore the first rows of data and the columns names. The review .json file was too large to load into the active memory and therefore the first lines of the document were investigated to identify attributes and their format without loading it into python. The python file where this was investigated can be found in the external Appendix named all json files.

The focus of this paper is the business file and the review file. These two files have information about the restaurants and the reviews, and the business ID links both files together. The reason for leaving out the user table is because the identification of the user is already identified in the review table as a foreign key.

Business json file and Review json file.

The Business dataset included 192.609 different businesses and spanned over multiple business categories such as restaurants, shopping, home services, etc. with a total of 1.300 different categories, as seen on the length of the list Figure 8.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 192609 entries, 0 to 192608
Data columns (total 14 columns):
business_id     192609 non-null object
name            192609 non-null object
address         192609 non-null object
city            192609 non-null object
state           192609 non-null object
postal_code     192609 non-null object
latitude        192609 non-null float64
longitude       192609 non-null float64
stars           192609 non-null float64
review_count    192609 non-null int64
is_open         192609 non-null int64
attributes      163773 non-null object
categories      192127 non-null object
hours           147779 non-null object
dtypes: float64(3), int64(2), object(9)
memory usage: 20.6+ MB
```

Figure 8 .info() information about the business dataset

```
Restaurants                     59371
Shopping                        31878
Food                            29989
Home Services                   19729
Beauty & Spas                   19370
                                ...
Geneticists                         1
Calligraphy                         1
Churros                             1
Sauna Installation & Repair         1
Court Reporters                     1
Name: categories, Length: 1300, dtype: int64
```

Figure 9 screenshot from jupyter notebook of the attribute "Categories" unfodled and intances counted

Through the investigation of Yelp's platform, it was noticed that a business could have a max of three categories associated and these could be suggested by Active users and the businesses themselves could change this. To inspect the attributes of the businesses' dataset, a general method that describes the dataset was applied.

As seen above in Figure 9, the category which occurred most was Restaurant with 59.371 instances, the second was Shopping with 31.878 instances and the third third was the category Food with 29.989. Later, when merging the business file with the review file, it showed that more than half of the reviews were associated with the category Restaurants. Even though only one third of the businesses are categorized as a restaurant, the restaurant category is the most reviewed category. This shows that it might be the most interesting group of business owners for Yelp to investigate on. As the recommendation system should recommend restaurants, all other businesses were removed to investigate the restaurants. A new dataset was created where only businesses with the category "Restaurant" were included leaving 59.371 restaurants to be investigated.

There were a total of 14 attributes connected to the business. The method also showed that not all values of the attribute were filled out. The attributes attributes, categories and hours had non-null values. These attributes are not required when creating a profile and can be added later. This might explain that not all businesses had these values. A description of each attribute can be seen in Table 4 below.

Table 4 business.json column descriptions

Column name	Description
business_id	The unique id that identifies the business
name	The name of the business
address	The address of the business written full
city, state, postal_code	The city, state and postal code that the restaurant is located at
latitude, longtitude	The latitude and longtitude of the restaurants location
stars	An algorithmic average of stars assigned to the restaurant rounded to the closest 0,5
review_count	The total number of reviews that the restaurant has received. This is not the same numbers as the total count in this dataset.
is_open	A Boolean value indicating if the restaurant is closed for good
attributes	Attributes of the restaurant. Are they serving alcohol, is dogs allowed etc.
categories	A list of categories to the restaurant. Maximum 3. Free writing
hours	A list over the opening hours per day of the week

How the attribute attributes is getting its values has been rather hard to identify from the investigation of the platform. While the attribute categories gets its values from a pre-created category list which both the particular business users and individual users can manipulate, the attributes values are not added the same way. Below in Figure 11 is a screenshot of how a category is added to the restaurant.

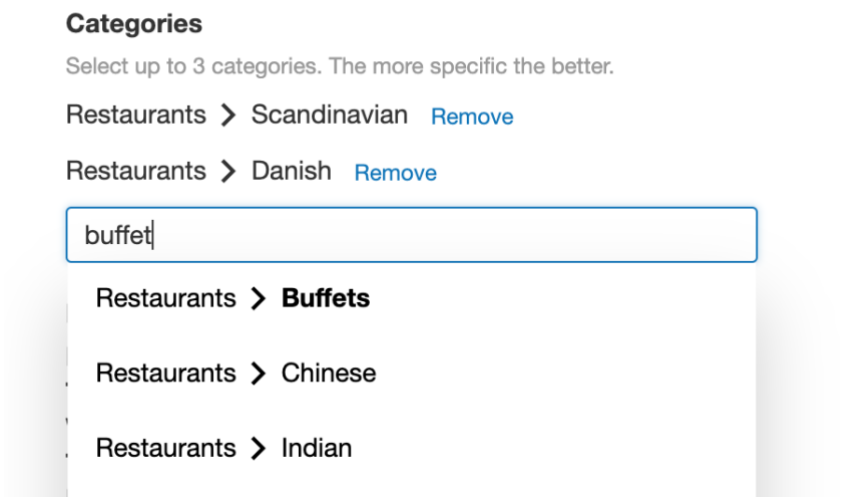


Figure 10 Screenshot of the category dropdown list from Yelps website.

Therefore, the assumption can be made that these amenities are either added by Yelp, the restaurant or an algorithm that searches through the reviews. Restaurants that have not been claimed have

amenities and that would point in the direction of an algorithm highlighting certain words or Yelp adding the amenities as there could not be found a way for Active users to add amenities. Though some of the categories must be the restaurant themselves who have added them as they seem too specific to be something that a Yelp employee would know or a reviewer would mention in their review such as “Bike parking” to which one might wonder at the relevancy of this as a part of a review.

The attribute attributes was in the format of a dictionary and to investigate the different attributes a search for the longest was made in order to find the instance that was most likely to have the largest amount of values. The longest list of attributes can be seen below in Figure 12.

```
Out[50]: '{\'WheelchairAccessible\': \'True\', \'RestaurantsTableService\': \'True\', \'BestNights\': "{\'monday\': False, \'t
         uesday\': False, \'friday\': False, \'wednesday\': False, \'thursday\': False, \'sunday\': False, \'saturday\': Fals
         e}", \'DietaryRestrictions\': "{\'dairy-free\': False, \'gluten-free\': True, \'vegan\': False, \'kosher\': False,
         \'halal\': False, \'soy-free\': False, \'vegetarian\': True}", \'BusinessAcceptsBitcoin\': \'False\', \'Smoking\': "u
         \'outdoor\'", \'DogsAllowed\': \'False\', \'CoatCheck\': \'False\', \'HasTV\': \'True\', \'Alcohol\': "u\'full_bar
         \'", \'NoiseLevel\': "u\'average\'", \'RestaurantsGoodForGroups\': \'True\', \'RestaurantsDelivery\': \'True\', \'Hap
         pyHour\': \'True\', \'BusinessParking\': "{\'garage\': True, \'street\': True, \'validated\': True, \'lot\': False,
         \'valet\': True}", \'RestaurantsTakeOut\': \'True\', \'RestaurantsAttire\': "u\'casual\'", \'Music\': "{\'dj\': Fals
         e, \'background_music\': False, \'no_music\': False, \'jukebox\': False, \'live\': False, \'video\': False, \'karaoke
         \': False}", \'WiFi\': "u\'free\'", \'Ambience\': "{\'touristy\': False, \'hipster\': True, \'romantic\': False, \'di
         vey\': False, \'intimate\': False, \'trendy\': True, \'upscale\': False, \'classy\': True, \'casual\': True}", \'Good
         ForKids\': \'False\', \'OutdoorSeating\': \'False\', \'GoodForMeal\': "{\'dessert\': False, \'latenight\': False, \'l
         unch\': False, \'dinner\': True, \'brunch\': False, \'breakfast\': False}", \'BikeParking\': \'True\', \'RestaurantsR
         eservations\': \'True\', \'Caters\': \'True\', \'RestaurantsPriceRange2\': \'2\', \'GoodForDancing\': \'False\', \'Bu
         sinessAcceptsCreditCards\': \'True\'}'
```

Figure 11 Longest instance of the attribute "attributes" with the restaurant category

The distribution of the star rating was investigated at this stage to make sure that the smaller data test sets and random calculated test sets had the same distribution. The distribution is seen in Figure 13 below. This is the average rating for the restaurant which Yelp provided and is therefore not an average rating of the reviews in the dataset. The score a restaurant gets is also based on an algorithm that takes time and relevance into account.

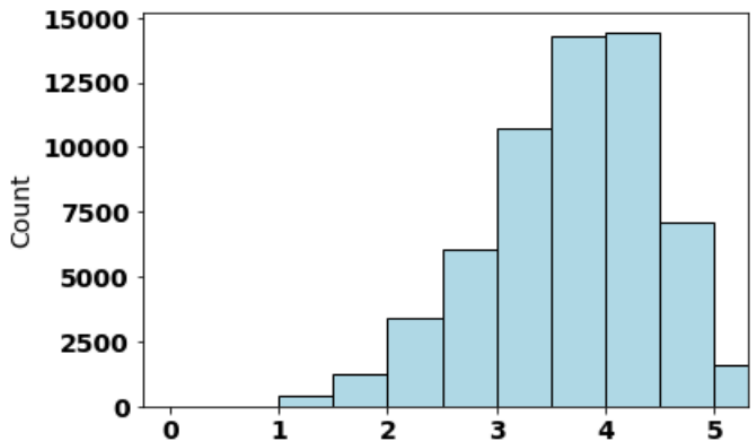


Figure 12 distribution of average stars

To investigate the reviews data, the files Reviews and Businesses were merged. The new merged file in columns has been marked in Figure 13 below and in Table 5 a description of the columns.

```
hours           object
review_id       object
user_id         object
review_stars    int64
useful          int64
funny           int64
cool            int64
text            object
date            object
dtypes: float64(3), int64(6), object(13)
memory usage: 705.2+ MB
```

Figure 13 .info() of the review columns

Table 5 column names and descriptions of the values from review.json merged with business.json

Column name	Description
review_id	Unique id to the review. This is the only unique if in the dataframe as a restaurants and a users id can occur multiple times.
user_id	A unique id that identify a user
review_stars	The stars for the particular review. The scale is 1-5 in whole numbers
usefull, funny, cool	Reactions from other users to the review
text	The text body of the review
date	The date of when the review was uploaded.

Taking a closer look at the review attributes file, we can see the reactions that the different reviews have gotten. All these are fairly positive, enforcing a more positive environment on the platform.

The max and minimum dates were investigated, and it showed that the oldest review dated back to the 19th of October, 2014, and the latest data to be the 14th of November, 2018. The attribute date was divided into a year attribute and a month attribute. This was done to filter data older than three

years from the latest date away for one of the models. The distribution of reviews per year can be seen in the Figure 14 below.

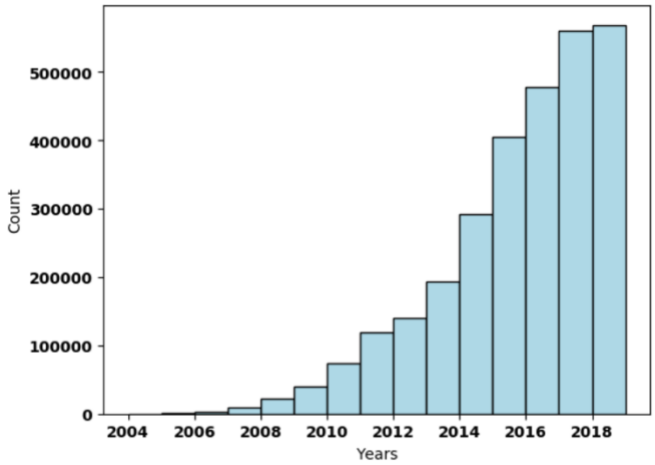


Figure 14 Distributed dates

Then the states was investigated and states from both Canada and USA was present in the dataset as seen in Figure 15.

```
In [6]: states = raw.state.unique()

Out[6]: array(['ON', 'NC', 'AZ', 'OH', 'PA', 'NV', 'QC', 'AB', 'WI', 'IL', 'SC',
               'NY', 'NM', 'CA', 'WA', 'XWY', 'BC', 'VT', nan, 'AR'], dtype=object)
```

Figure 15 All states in the dataset.

Next the amount of businesses with each state was investigated. This insight gave the impression that some data had been left out by Yelp as some of the states had less than 100 restaurants registered at Yelp as seen in Figure 16.

```
In [18]: raw.groupby('state')['business_id'].count().sort_values()
Out[18]: state
         WA           3
         BC           3
         CA           4
         VT           5
         AR           7
         XWY          8
         NM          14
         NY          91
         SC       12021
         IL       23511
         AB       53565
         WI       78768
         QC      119520
         PA      175524
         OH      197414
         NC      235244
         ON      470435
         AZ     1009611
         NV     1171204
         Name: business_id, dtype: int64
```

Figure 16 Number of restaurants in each state

Now that the data has been investigated, we are ready to move on to the data preparation phase as it has become more clear how the data could clean for the model that supports it. From the start it has been known that it is the latent space model that is going to be used in the modelling phase. Therefore, the investigated data is the business_id and user_id in connection to each other. However, it is still just as important to investigate the available data to support the business understanding but also to be able to discuss how other models can support the same purpose with other means.

Data preparation

Data preparation and Modelling are phases that are overlapping each other. Therefore, these two phases are going to be explained in the iterative process they were investigated in. First, in the data preparation phase the data was cleaned. The goal of this phase is to iteratively prepare data and train the model to gain new insights and start over.

The first data file that was prepared was the business file. This was done to select only the relevant instances when merging the two data files. In the earlier step, the only business type that was investigated was of the category restaurants. The reason being that the recommendation system should recommend restaurants. Therefore, the first thing in the Data preparation step was to remove all businesses that did not have the category restaurant. This was done by creating a new list

copying all instances where the attribute category had the word restaurant in it. After this, the closed restaurants were removed, meaning all instances that had the attribute is_open that equaled 0 were removed. This was done simply because the model should not recommend restaurants that do not exist anymore. The distribution between closed and open restaurants were 42.237 instances with the value 1 = open and 17.134 instances with the value 0 = closed.

Next, it was decided to only focus on states inside of the USA and the states of Canada were removed. "ON", “QC", "AB", "XWY", "BC", "nan" were all removed. Studies find that people have different eating habits when on holiday compared to dining out in your hometown (Kocevski & Risteski, 2012). As such, it was decided to only focus on the states within one country, as there might be people who have written reviews in both Canada and the USA. By removing one country, the amount of cross-country reviews are limited. This was done as there was no other way to differentiate nationality or country. Next, reviews three years older than the most current date in the dataset was removed. This decision was made because as time passes the environment changes: new restaurants are opening, old ones closes, they change their concept, the users preferences change, etc.[en28] After the business dataset was cleaned, the review file was loaded in and merged with the business dataset based on the business left in the dataset. The size of the review.json file was too large to handle on the regular laptop. Therefore, it was decided to only load the reviews that were linked to a restaurant and merge the restaurant attributes with the review attributes. Approximately 50 percent of the reviews were connected to restaurants. Leaving 3.546.952 reviews for further investigation. After this, the columns review_id, business_id and user_id were hashed because the latent space model only handled integers and as explained in the data understanding section, it is normal to create dummy values. The amount of reviews left per user was counted and the users that had only left one review was removed. To create a matrix of connections between users and restaurants the only attributes needed are business_id and user_id, therefore the rest of the attributes were removed. First, a matrix was calculated and converted to an adjacency matrix.

As already mentioned, most real-life networks are sparse. As such, in the first iteration of preparing the data it was decided to use the business_id’s as nodes and the user_id’s as links in order to diminish the sparsity of the matrix because there was a larger amount of connections to form as there were more links than nodes. The network statistics from the first clean dataset, which included data from all the available American states, can be seen in the Table 7 below.

Table 6 Network statistic all available American states

All American states	
Nodes	*23629*
Links	*7979790*
Average clustering coefficient	*0,5355869647407252*

True enough, the statistics of this network showed a sparse matrix as we know real life networks often are, even though the number of edges to the number of vertices were 337 times larger. The Average clustering is an expression of how many neighbors Vi is connected to within its neighborhood. A score of 1 indicates that every neighbor connected to Vi is also connected to every other vertex within the neighborhood. A score of 0 indicates that no vertices are connected to any other vertices that are connected to V1 (GeeksforGeeks, 2018). The average clustering coefficient was 0,53 in this network, indicating dense local networks. To be part of the neighborhood network you only need one link to be included. An illustration of three different clustering coefficients from three different network configurations from Fundamentals of spreading processes in single and multilayer complex networks by de Arruda, G. F., Rodrigues, F. A., & Moreno, Y. can be seen in Figure 17 below.

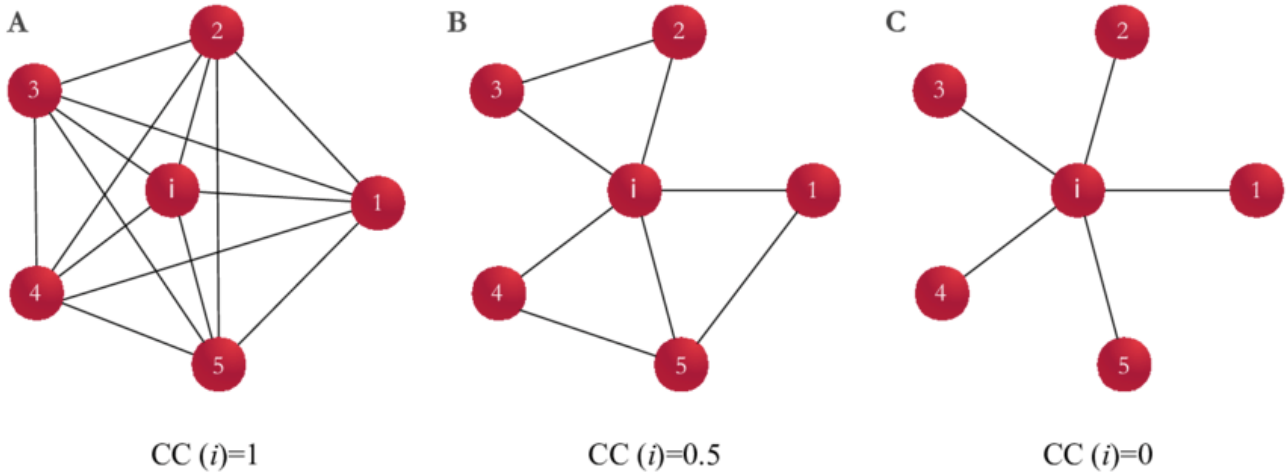


Figure 17 (de Arruda, Rodrigues, & Moreno, 2018)

Modelling

The first model was trained on the adjacency matrix containing data from all available American states. The model was trained three times with a different number of dimensions(k). The train and test scores for the three models can be seen in table 8. The results are disturbing as the model performs above all expectations on training and test data. It is neither over- nor under-fitted but there is something off and it occurs on all three trained models. To see the visualization of AUC graph have a look at appendix 2. Overfitting happens when a model learns the train data to a very detailed degree and can therefore not generalize on to the test data while underfitting is when it cannot model the training data nor generalize on the test data which is the result of a poor choice of model (Provost & Fawcett, 2013).

Table 7 Test and train scores for all available American states

K	Train score	Test score
2	0,9549	0,9539
3	0,9636	0,9623
4	0,9678	0,9668

Visualizing the two dimensional model in the latent space gave an indication on why the model performed so well on both test and train. As seen in Figure 16 below, the nodes cluster together in what might seem to be the same amount of states investigated. The clustering is not as clear when the model becomes multi dimensional as visualized in Figure 15 and 16 below. With the average clustering coefficient on 0,53, it became clear that networks were more dense area wise. To investigate this further, a new dataset was prepared.

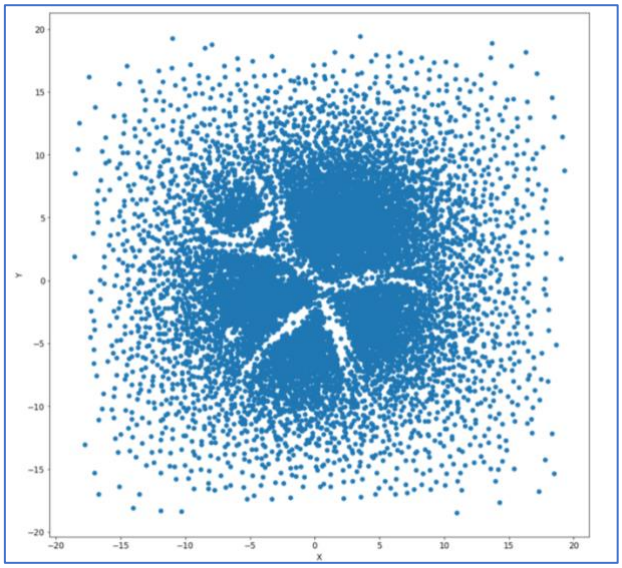


Figure 18 two- dimensional space all available American states

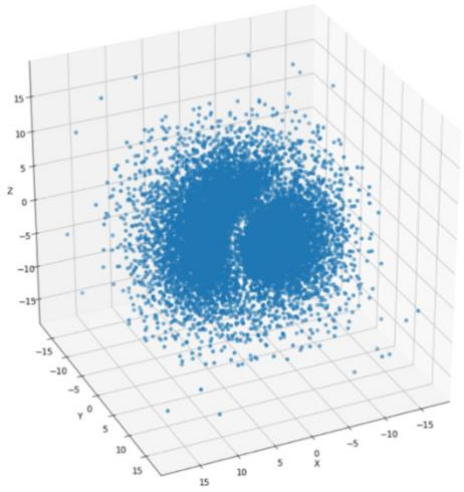


Figure 19 k =3, all available American states

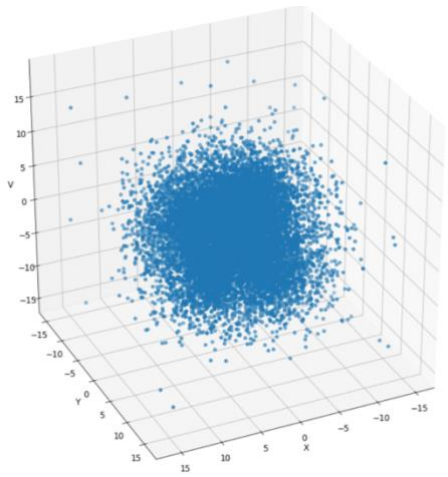


Figure 20 k = 4, all available American states

Data preparation

To investigate the assumption about the data clustering into clusters in the latent model data, reduction is applied. Data reduction is a general task that is used when dealing with large datasets, in which one slices the dataset to a size where important information is not lost and is easier to process. (Provost & Fawcett, 2013) Small datasets might also better reveal the information within. This is not done without sacrificing information, as this works as a trade off between the manageability gained against the information lost. Provost and Fawcett (2013) write that this is often a trade worth making (p. 304). The data was cleaned as in the first data preparation and then 100 random users who had at least written 2 reviews were chosen and their reviews were collected into a new dataset.

An adjacency matrix was created. The network statistics can be seen below in Table 8 where the average clustering coefficient is 0,95, indicating dense networks within the model.

Table 8 Network statistic Reduced all available American states

Reduced all American states (100 random users=	
Nodes	*2284*
Links	*193404*
Average clustering coefficient	*0,9555170234484153*

Modelling

Then, the latent space model was built. The train result for the model was 0,9761 and the test result was 0,9721, which was higher than the earlier models built, though this makes a lot of sense if the networks within the model are more connected. To see the model visualized, go to Appendix 3. Next, the state abbreviation was added to the two dimensional model visualized in Figure 21. This was done in order to visualize in color what state the plot in the latent space belonged to.

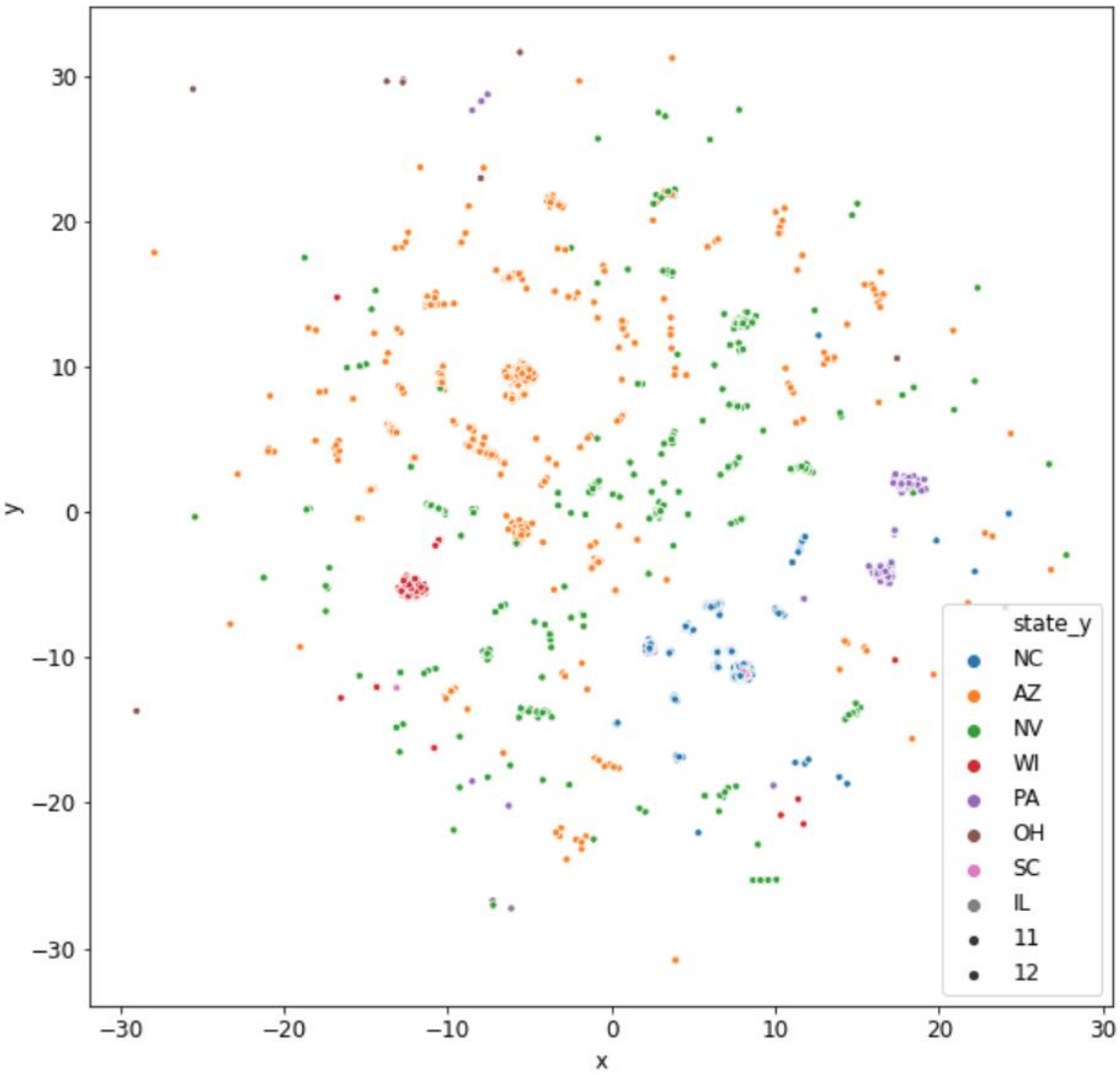


Figure 21 Reduced all available American states two-dimensional

The notion that the latent space was an indication of distance, and thereby clustering restaurants that are physically closer, seems to be viable when looking at the figure above. We see clusters of each color representing the state clustering together and then with a radius of empty space before hitting

what seems to be a wall of nodes. An interesting thing, is that the restaurants in the state of Nevada are holiday locations due to their proximity to Las Vegas. This state is the most scattered out in between the rest of the clusters. The scattering of the restaurants in this state indicates that people who have visited a restaurant in the state of Nevada have multiple connections to restaurants outside of the state and therefore are not clustered closely together. This visualization gives the understanding that this model is hard to scale across states as the networks seem to be dense local networks.

The high performance from the models built so far makes more sense when the latent space is visualized both the full dataset and the reduction. When investigating 10 percent of a large dataset that clusters together like this there is a higher probability of two instances to be investigated as non links to be from two different states and links to be within the same cluster. The further increase in model where data reduction was applied was also expected as this would expose the data and the patterns in it.

The high performance from the models built so far makes more sense when the latent space is visualized. It becomes clear that the data is clustering in a pattern that seems to match the size and amount of states. Doing data reduction made the clustering even more prominent. When investigating 10 percent of a large dataset that clusters together like this, there is a higher probability of two instances to be investigated as non links to be from two different states and links to be within the same cluster. The further increase in models where data reduction was applied was also expected as this should expose the data structure and the patterns when visualized.

Data preparation

This suspicion led to the further investigation of a city to see how sparse or dense a matrix was if the links of the networks were restaurants. And after that, changing the edges and nodes to investigate how sparse the matrix would become if it was people that were the links and restaurants as nodes.

Madison	30522
Cleveland	36553
Gilbert	38283
Chandler	43638
Mesa	46848
Tempe	52662
Henderson	58374
Pittsburgh	75312
Scottsdale	101531
Charlotte	102566
Phoenix	199019
Las Vegas	531505

Figure 22 Largest cities

The next step would have been to investigate the largest city and using the user_id as nodes and business_id as links. It was not possible to build the matrix from this data as it maxed out the active memory of the computer. Therefore, a smaller city was chosen to investigate in order to be able to compare a model where the nodes and links were turned around. The city Charlotte was chosen instead. The data for the city of Charlotte was cleaned as the previous models. Charlotte had 102.566 instances. The statistics of the network can be seen below in Table 9. The average clustering coefficient was 0,59. The train score of the link prediction model was 0,8414 and the test score 0,8397. A drop in performance was expected as narrowing down the data set to a city means more networks closer together and a higher amount of links per node is present. Location might not have as big an impact on the latent space when local networks physically far away from each other are removed.

Table 9 Network statistic Charlotte. Users are nodes.

Charlotte (user_ids are nodes)	
Nodes	18118
Links	6417952
Average clustering coefficient	*0,5876377484619045*

The results of the train and test scores can be seen in table 10 below.

Table 10 Charlotte train and test score with dimension change

K value	Test score	Train score
K=2	0,8397	0,8414
K = 3	0,8598	0,8619
K = 4	0,8738	0,8756

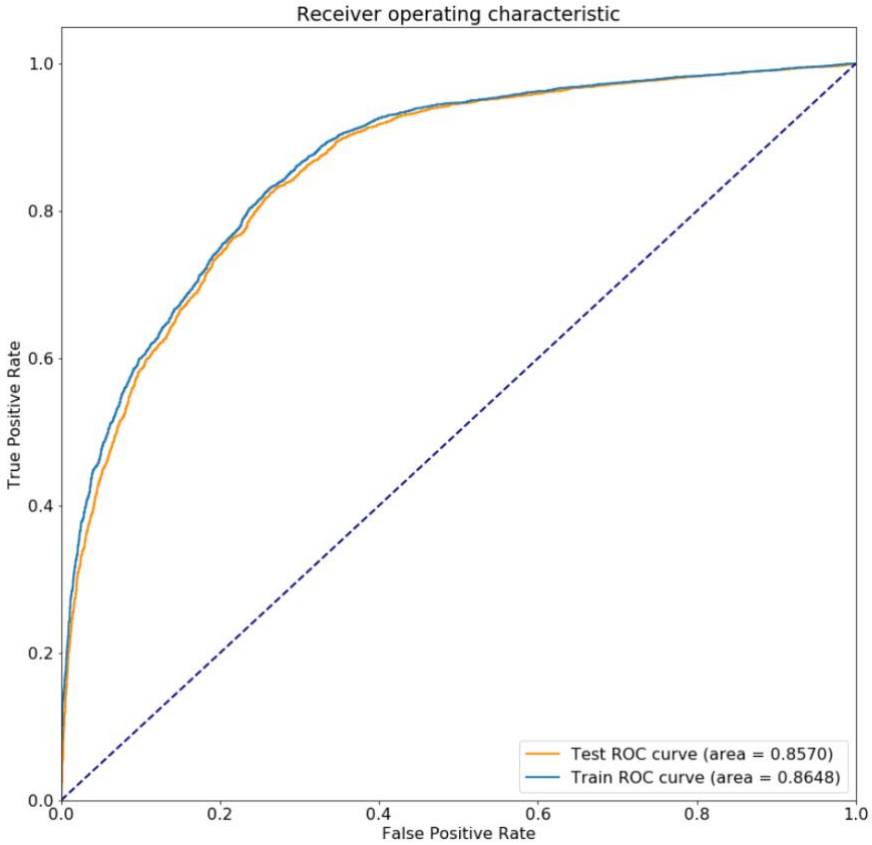
The more dimensions added the better the performance. This gives an indication that there is some depth to the model than can differentiate the nodes with more variables. To see the visualized roc score graph go to Appendix 4.

The last model to be investigated was a random selection of users within Charlotte. This was chosen to be investigated because of the knowledge of a more dense network within local areas. As such, 100 random people were selected within the Charlotte dataset and all the reviews which they had written were used to form a new dataset. Sizing down the data set might reveal more general patterns. The statistical data can be seen in Table 11.

Table 11 Network statistics reduced Charlotte

Small Charlotte small users_id are nodes	
Nodes	890
Link	51342
Average clustering coefficient	0,7884047983618162

The average clustering coefficient was higher in the network where the data had been reduced. The model was trained in a two dimensional network. The train score was 0,8570 and the test score was 0,8648 as seen in Figure 23 below. The higher average clustering coefficient reveals that the



networks existing within the latent space are more dense which might explain the little increase in performance on the two dimensional model compared to the model trained on the Charlotte data without a data reduction.

Figure 23 ROC Graph Reduced charlotte

With the network visualized, it becomes more clear that smaller networks do appear, though there still is a large cluster in the center as seen in Figure 24 below. The author realized that by changing the links and nodes but not removing instances where a business only occurred once, it might create nodes of users in the network that did not have any links to other users.

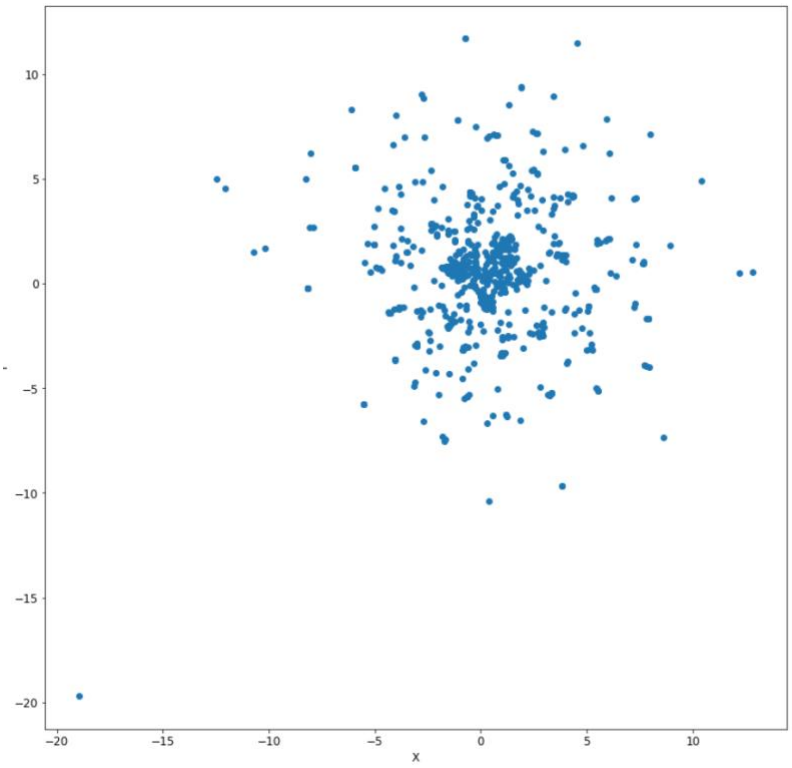


Figure 24 Latent space two-dimensional reduced Charlotte

Evaluation

In this section the findings and performance will be discussed in order to evaluate how the latent space model would perform as a social recommendation system for restaurant and users in the Yelp data. Using the CRISP model allowed the author to have both the technical aspect of how it is to investigate a social recommendation system but also to see what the business needs were and think about how it would fit into the Yelp platform and where it could create values. Some of these ideas and thoughts will be evaluated in this next section

Model performance

How well the link prediction performs on the network is an indication of how well the networks have placed the connected nodes according to each other. The baseline of all the models are 0,5, as already explained. All the models generally perform well on the train and test data, ranging between 0,97 to 0,83 in train and test scores where the train and test score as a pair has the largest span of 1 percent point. When training on all the available data filtered on the American states, it becomes clear that restaurant networks in the latent space are clustering together based on location. The clustering is investigated in a reduced dataset to see if this might make the patterns clearer, as the first visualization of the two-dimensional latent space of all American states was a very node dense model. When visualizing the reduced dataset and Nevada was added to the dataset, it was interesting to see how many of the Nevada nodes were scattered across the middle of the latent space in a diagonal line. The reason for a large part of the Nevada nodes to be placed in the center of the latent space might be due to their cross-state links to other restaurants in other states as people from other states travel to Las Vegas for a short holiday trip. This would arguably cause many of the Nevada nodes to be scattered like this.

Too good to be true

The high performance of the model built with the large dataset that included all American states was, as already explained, not a satisfying result. The latent space became an indication of distance, making it easier to predict if two nodes were connected or not. If the two nodes were to be part of the same network clustering close together, chances of a link were high, and if the nodes were not within the same cluster of nodes, chances were that there was no link. Therefore, if the model simply predicts a link every time there is a node within the cluster, chances are higher than the baseline of 0,5 to predict a link. Likewise, if it predicts a none-link every time the node is away

from the cluster, the baseline of this prediction is also higher 0,5 making the high performance of the model less impressive.

Blame the physical distance

Users that are creating the links between the restaurants are to some degree more likely to visit restaurants within a certain radius of where they live or work. The radius might be time-related. In cities with better infrastructure, we might see networks of restaurants in which the physical distance between the restaurants are larger, because people are able to travel longer and have more options scattered out within the city. Taking into account that most of the restaurant experiences happen within a radius of their home, they participate in creating more dense local networks that, when compared to other states, will cluster together based on distance because the local networks are more interconnected.

Scaling problems

The clustering of networks in states indicates that even if a latent space model performed well on a local network, the model does not scale well when large physical distances are a reality in the network. This might explain why the model has not performed at a satisfying level so far compared to the other earlier mentioned success stories of implementing latent space models. The nodes and edges are both restrained to a physical presence. If the target user, the one who asks for a recommendation, travels outside of his or her local network, the connections become fewer. These poor connections will then result in less reliable recommendations. To eliminate the local networks, users should be rethought to fit a larger group of people that exist across cities.

Some of the blame is in the design of the model

As already identified, the physical location and the likelihood of the user dining out within this network create dense local networks. The design choices for the model have some blame when investigating the local dense network, in this case the Charlotte dataset as that was the most local area investigated in this paper. Data reduction was applied to the Charlotte dataset to see if we saw a distinctive pattern in the latent space. The visualization of the two-dimensional model could still point in the direction of a latent space that was somewhat influenced by location. Though it cannot be said with certainty, this will be explained further in the next section, Charlotte. Weighing the links in the model would have created a new different latent space, if first the adjacency matrix had been created with the focus of weighing the links. Creating an adjacency matrix where all links are weighed equal no matter how many connections two users or two restaurants have had creates a

more simplified version of the reality and a more flat model. If the links were weighed, we might see more obvious clusters of take-away places, kid-friendly restaurants or maybe plant-based places.

Charlotte

The city Charlotte was reduced and visualized in a two-dimensional latent space to investigate if any clear patterns emerged. Though in this model, it might in higher degree be based more on the popular city areas where people like to "go out" and how much people go out. The nodes in the center of the model might represent the people that has above a certain number of reviewed restaurants within the center of the city. Unfortunately, I do not have any good area knowledge of the city of Charlotte nor of the restaurants. Therefore, it is hard to offer any truly good observations on what the latent space might be an expression of. Except, of course, for the obvious. Which in this case would be that users with many reviews are overlapping with a lot of different users, as well as each other, and are, therefore, placed in the center of the model. A more in depth investigation of restaurants and areas in the specific investigated local network would help decide what the latent spaces might be an expression and if we see that location also has a large impact on the latent space in a more local network.

The classic recommendation problems

This model struggles with the classic recommendation system problems such as data, the sparsity problem, and the cold start problem. This is nothing out of the ordinary as this is a common problem for all recommendation systems.

The cold start problem occurs for both the new users and new restaurants joining the platform. If recommendations are based on a social network and the users have yet to review anything, they do not have any connection to other users or restaurants. Therefore, they are unable to get recommendations and be recommended. This also occurs for the new restaurants as they have not yet gotten any reviews. Therefore, they are not part of the network and cannot be recommended through the social recommendations. Neither the user nor the restaurant will appear in any networks if there is not data to form a connection with. From a business perspective, the cold start problem could be solved by asking the user to identify their likes and deliver this data to the platform as a start point so the user is able to use the implemented features. Or for restaurants, a special list for "New restaurants in town" could be implemented to ensure they were noticed. In fact, Yelp rewards people who are the first to review a restaurant so a list like this might create a situation where they

are quickly reviewed by someone and become part of the network. The cold start problem is a theoretical research problem that is not solved because a design decision in the recommendation system was made to avoid an empty recommendation. The user or restaurant still experience that they are not part of the network when first entering the site.

As mentioned earlier, real life networks are sparse. This is also the case for the Yelp dataset. Even if you have reviewed a thousand restaurants, which some people in the Yelp community have done in order to contribute to the platform, it is only a fraction of how many restaurants you can visit. This is what makes it sparse. Even when looking at the dense local networks in the bigger picture, this is again only a fraction of the restaurants out there. You can look at a small area of the network that appears to be less sparse, but that does not change the fact that it is a sparse network.

Connections rather than a target value?

As we know, dining experiences depend on a sea of attributes, many which the restaurant themselves may have very little influence on or ability to avoid. An unexpected busy night might create a less attentive personnel, resulting in a bad review. Had it been any other evening, this would not have happened. Users write reviews based on their singular experience and when they write their reviews, they rarely account for unexpected difficulties encountered on that specific night by the restaurant. Their rating is then an expression of that evening and their experience, which is not wrong. The user should not have to do an analysis, considering whether this happens every night or once every couple of months because someone calls in sick. The wisdom of the crowds generalize the collected opinion and it is okay that bad reviews exist. Therefore, if you rate a restaurant 4 or 5 stars based on this singular experience, it does not really make any difference in the big picture, as one's next experience might differ due to various elements, such as a new menu, a change of staff, etc. By investigating connections rather than ratings, and by removing instances that the user rated with 1 or 2 stars, we build a positive network. Ensuring that the recommendations do not recommend places that have previously been rated poorly by the user is no more than filtering these results away before it reaches the front end. Today, more often than not, we do our research before setting foot in a restaurant, which indicates that we must have had some positive expectations when we entered and thereby having displayed some general interest in the concept of the restaurant prior to entering. From a business perspective, it makes more sense to focus on telling the users what they might like rather than dislike.

Deployment

The Deployment phase was sized down to a section within the Evaluation phase, as the results of this paper do not end in a product that should be deployed. The importance of how a model should be deployed is still an interesting topic, especially considering a heavy model such as the latent space model.

The computational cost of latent space models is high. The model is heavy to run and, due to the expectations of how fast something should load online, this model cannot be implemented live on Yelp's platform. The computational cost Yelp themselves should pay to implement such a recommendation system is also far from feasible as the Active users and the Lurkers are the groups of users on Yelp that are unwilling to pay for the service as other review platform alternatives exist. If Yelp were to direct this model towards the paying group of the platform, it could be implemented as a part of their grouping of their users to optimize the reached customers to be interested in a restaurant when advertising on page. As explained, the model does not perform well across states, it is not the size but the physical limitations that that the network is marked by in reality that structures the network making the recommendation poor. If a more local network is investigated, this might provide other insights than distance but it is important to remember that local datasets do not mean small datasets and thereby less heavy to run. It is the amounts of nodes in the network that determine how heavy it is to run.

Conclusion

This paper has described how link prediction on a latent space model optimized to large scale networks could perform as a model behind a social recommendation system. Though this type of model has been praised when implemented in other networks, the restaurant networks have the disadvantage that users demand local recommendations according to their current whereabouts rather than which pizzaria they should try close to home.

If the results are solely looked at from a technical aspect, the networks are performing well, but with the important catch that the structures in the network make it easier for the model to predict well. The performance problem occurs when the model is looked at from a business perspective. It is a heavy infeasible model to implement on a website where users' expectations to live performance is high. Also, as a recommendation system, the latent space model is not scalable. Though as a social recommender on a local scale, there is still information to be gained and

investigated. The performance of the model built on the Charlotte data showed promising results with a lower prediction rate and an increase in performance when more dimensions were added.

This investigation was hardly the size of what an internal team in Yelp could achieve with their resources. This is only the beginning of what a latent space model evaluated by a link prediction model can unfold of knowledge. It would be highly recommended to continue working with this type of model. Perhaps not as a social recommendation system implemented on page, but as a means to uncover new latent variables in the networks.

Future work

Local areas

The last investigated model was the small Charlotte dataset. Despite the computational costs and the fact that it is a time consuming model to use, it might still have interesting insights to offer to the Yelp platform. Therefore, even if it does not make a feasible social recommendation system even on small networks, it is still recommended to investigate this use of the model to gain interesting latent insights. Maybe it can be used to locate interesting networks of restaurant grouping on other latent factors than the category and price.

Classifying users and restaurants

Moving forward, as the model is not feasible to implement on the platform as a social recommender, there is still potential. As mentioned earlier, recommendation systems are often created by multiple models evening out the performance. Drawing inspiration from how advertisements are matched with users on Facebook, it could be interesting to classify the users based on their profile, text reviews, or the tags from places they have visited and then using a latent space model to investigate how the network of classified users are visiting restaurants. This breaks down the physical radius that users are creating links within and is not user specific, which means that if a user creates a profile with the right information, they can be labelled their classification and participate in the network, even if the model is calculated offline and not live on the website. However, the cold start problem for restaurants is slightly increased if this form of social recommendation model is implemented offline and only calculated ones every month or so.

Bibliography

Anderson, M., & Magruder, J. (2012). Learning from the crowd: Regression discontinuity estimates of the effects of an online review database. I *The Economic Journal, 122(563)* (s. 957-989).

Bagheri, E., & Du, W. (2020). Temporal Latent Space Modeling for Community Prediction. In: Jose J. et al. (eds). I *Advances in Information Retrieval.* Springer, Cham.

Bender, E. A., & Williamson, S. G. (2010). I *Lists, Decisions and Graphs.*

Chen, L. C. (2015). Recommender systems based on user reviews: the state of the art. In User Modeling and User-Adapted Interaction (25).

Clement, J. (24. July 2020). *Global digital population as of July 2020.* Hentet fra Statista: https://www.statista.com/statistics/617136/digital-population-worldwide/

Clement, J. (7. January 2020). *Hours of video uploaded to YouTube every minute as of May 2019.* Hentet fra Statista: https://www.statista.com/statistics/259477/hours-of-video-uploaded-to-youtube-every-minute/#:~:text=As%20of%20May%202019%2C%20more,for%20online%20video%20has%20grown.

Constantiou, I. D., & Kallinikos, J. (2015). New games, new rules: big data and the changing context of strategy. 30(1). In Journal of Information Technology,.

de Arruda, G. F., Rodrigues, F. A., & Moreno, Y. (2018). Fundamentals of spreading processes in single and multilayer complex networks. I *Physics Reports, 756* (s. 1-59).

FAQ, Y. (8. September 2020). *What is Yelp's Elite Squad?* Hentet fra Yelp: https://www.yelp-support.com/article/What-is-Yelps-Elite-Squad?l=en_US

FAQ, Y. (8. September 2020). *What is Yelp's recommendation software?* Hentet fra Yelp: https://www.yelp-support.com/article/What-is-Yelp-s-recommendation-software?l=en_US

FAQ, Y. (7. September 2020). *Why would a review not be recommended?* Hentet fra Yelp: https://www.yelp-support.com/article/Why-would-a-review-not-be-recommended?l=en_US

GeeksforGeeks. (8. February 2018). *Clustering Coefficient in Graph Theory*. Hentet fra GeeksforGeeks: https://www.geeksforgeeks.org/clustering-coefficient-graph-theory/

Haklay, M. (2016). Why is participation inequality important? I C. H. Capineri, *European Handbook of Crowdsourced Geographic Information* (s. 35–44). London.

He, J., & Chu, W. W. (2011). Design Considerations for a Social Network-Based Recommendation System (SNRS). Los Angeles CA 90095:: Computer Science Department University of California.

Jacobsen, N. F. (2018). *Large scale latent variable modeling for link prediction in complex networks.* Appendix A.

Keshari, K. (28. April 2020). *edureka!* Hentet fra Everything You Need to Know about the Best Laptop for Machine Learning: https://www.edureka.co/blog/best-laptop-for-machine-learning/

Kocevski, J., & Risteski, M. (2012). Eating out on vacation. I *Procedia-Social and Behavioral Sciences, 44* (s. 398-405).

Lee, D., & Brusilovsky, P. (2018). Recommendations Based on Social Links. In Social Information Access. I *Lecture Notes in Computer Science, vol 10100.* (s. 391-440). Springer, Cham.

Lucidchart. (8. September 2020). *What is an Entity Relationship Diagram (ERD)?* Hentet fra Lucidchart: https://www.lucidchart.com/pages/er-diagrams

M, M. (16. February 2019). *Introduction to recommendation systems and How to design Recommendation system,that resembling the Amazon.* Hentet fra Medium: https://medium.com/@madasamy/introduction-to-recommendation-systems-and-how-to-design-recommendation-system-that-resembling-the-9ac167e30e95

Miller, C. C. (8. January 2009). *Yelp Jumps the Pond*. Hentet fra Bits: https://bits.blogs.nytimes.com/2009/01/08/yelp-jumps-the-pond/

Nielsen, J. (23. October 2012). *The 90-9-1 Rule for Participation Inequality in Social Media and Online Communities*. Hentet fra NN/g Nielsen Norman Grouå.

Provost, F., & Fawcett, T. (2013). I *Data Science for Business: What you need to know about data mining and data-analytic thinking.* O'Reilly Media, Inc.

Rosenberg, E. (2007). 'The Shed at Dulwich' was London's top-rated restaurant. Just one problem: It didn't exist. *Washington Post*.

Sanchez-Cartas, J. M., & Leon, G. (2019). Multisided Platforms and Markets: A Literature Review.

Schafer, J. B., Frankowski, D., Herlocker, J., & Sen, S. (2007). Collaborative filtering recommender systems. In In The adaptive web. Berlin, Heidelberg: Springer.

TONDJI, L. N. (2018). Web recommender system for job seeking and recruiting. Researchgate.

Weisstein, E. W. (14. September 2020). *Adjacency Matrix*. Hentet fra MathWorld--A Wolfram Web Resource: https://mathworld.wolfram.com/AdjacencyMatrix.html

Zhao, G., Qian, X., & Feng, H. I. (2014). Personalized recommendation by exploring social users' behaviors. In International Conference on Multimedia Modeling.

Zheng, R., Provost, F., & Ghosee, A. (2007). Social network collaborative filtering: preliminary results. I *To appear in Proc. Sixth Workshop on eBusiness (WeB2007).*

Appendix 1

Notes from investigation of the functionality of the yelp platform

Users → users without a profile can look at all the data the only thing they cant do is

You can access the site and the reviews without logging in or leaving any information.

You can login as a company

You can create a user and login as that user ✍ either with an existing user from another platform such as facebook, apple or google, or you can create your own user with name, email, zip code and birthday.

From the observations on how you can access the Yelp platform three types of users has been defined. Write users, read users and company users.

Restaurants

Is it closed

Link to website

Link to directions

Link to menu

Phone number

Opening hours

Message the business

Edit business detail: name, address, phone number, links to website, categories (only 3), opening hours, duplicate of other restaurant, is it permanently closed, is location inside mall or airport. ✍ to edit these details you have submit your email and a message indicating why this should be updated or changed.

You can see if the business has claimed their Yelp site.

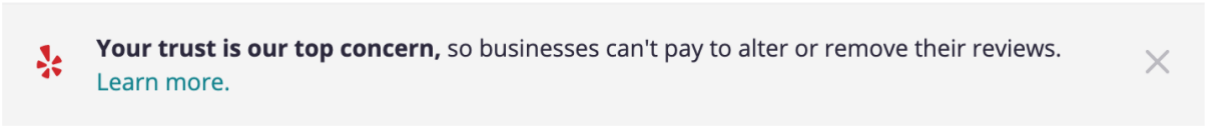
See the amenities of the establishment

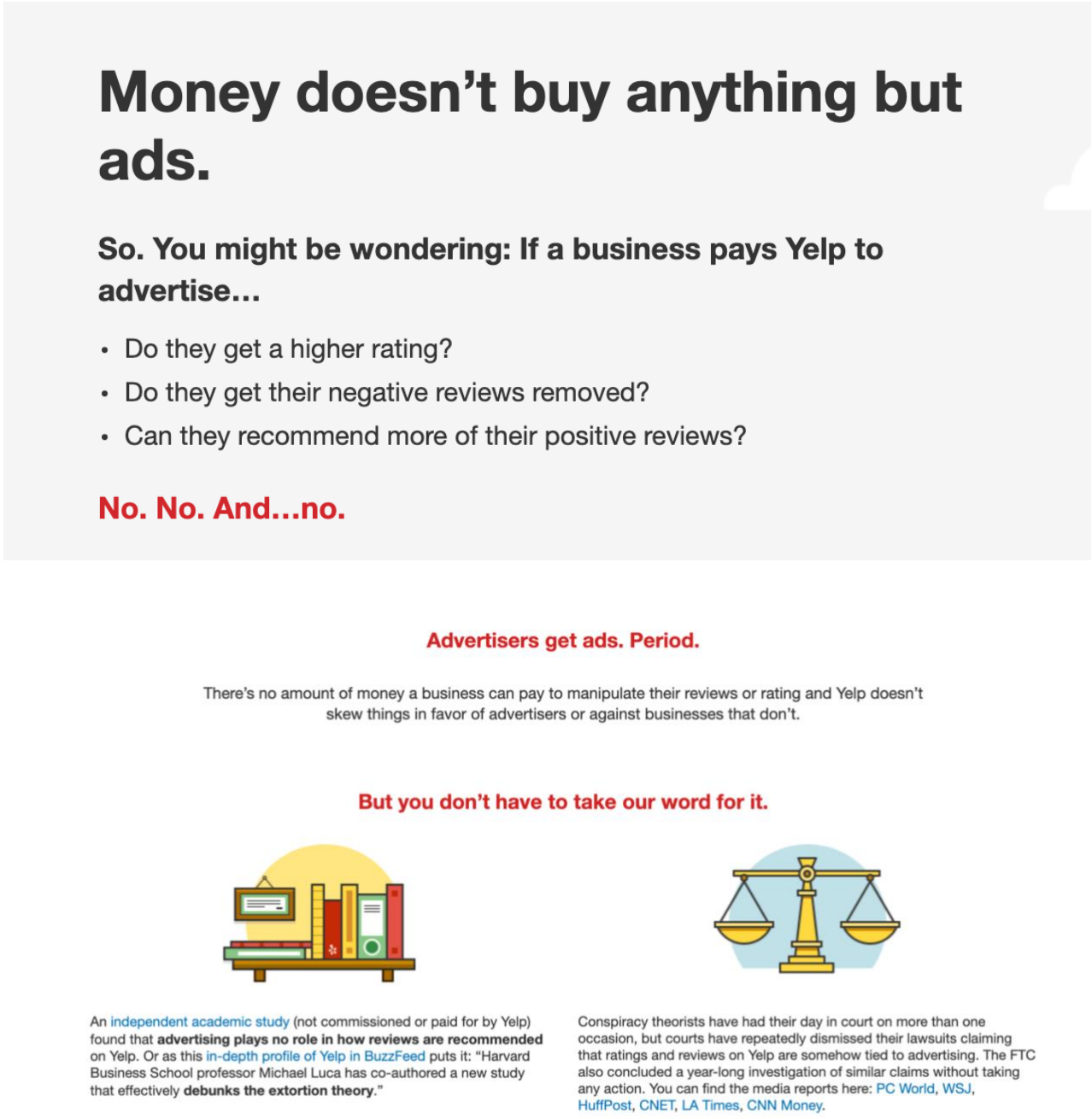
Amnesties

Reviews

The headline of the list of reviews is “Recommended reviews” ✍ this indicates some sort of ranking

Yelp has a block of text as a pop-up see picture below





You can search for reviews, sort and choose what language you want.

The default sort is Yelp sort you can change this to: newest first, oldest first, highest rated, lowest rated, elitists

With every review is the profile of the person ad general statistics such as number of friends, number of reviews and number of uploaded pictures and an elite tag if they have achieved this.

You can share other peoples reviews or embed them

You can see “currently not recommended reviews” ✍ These are reviews that have been flagged and are not factored into the business's overall star rating.

You can check in

Profile

You can add pictures to your profile to personalize it.

You can add a text about yourself

You can add a lot of different funny facts about yourself in the form of answers to questions that yelp has asked you such as: your first concert, favorite movie, favorite meal, the best book you have read, your last meal on earth would be, latest’s crush, what you do besides being on Yelp etc.

You can befriend other yelp users

You can receive notifications about new friend requests or compliments.

You can see your already published reviews and delete, forward or edit them

You can see events

You can see check ins

See your bookmarks

Between users

Compliment others reviews or pictures with tags and a message.

Rate a review: Usefull, cool, funny

Follow another user

See other users profile including: their connections, reviews and the included pictures, statistic on reviews, similar reviews, message

Block or report other users.

Services

You can order take away or take away as delivery through yelp though not available everywhere.

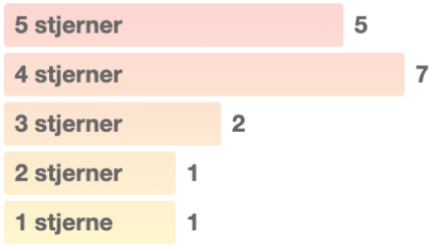
You can reserve a table through yelp

Create events and show interest in them

Tips

Om Mikkel J.

Vurderingsfordeling



Se flere diagrammer

Stemmer på anmeldelser

- Nyttig **45**
- Sjov **8**
- Cool **9**

else

Statistik

- Anmeldelsesopdateringer **1**
- Første **1**
- Følgere **8**

10 komplimenter

1 2 1 3 2 1

10 komplimenter

1 2 1 3 2 1

Lokalitet

København S

Yelper siden

juli 2012

Ting jeg er vild med

Gourmet, Design, Teknik, Helst en kombination af disse

Min hjemstavn

Islands Brygge

Når jeg ikke Yelper...

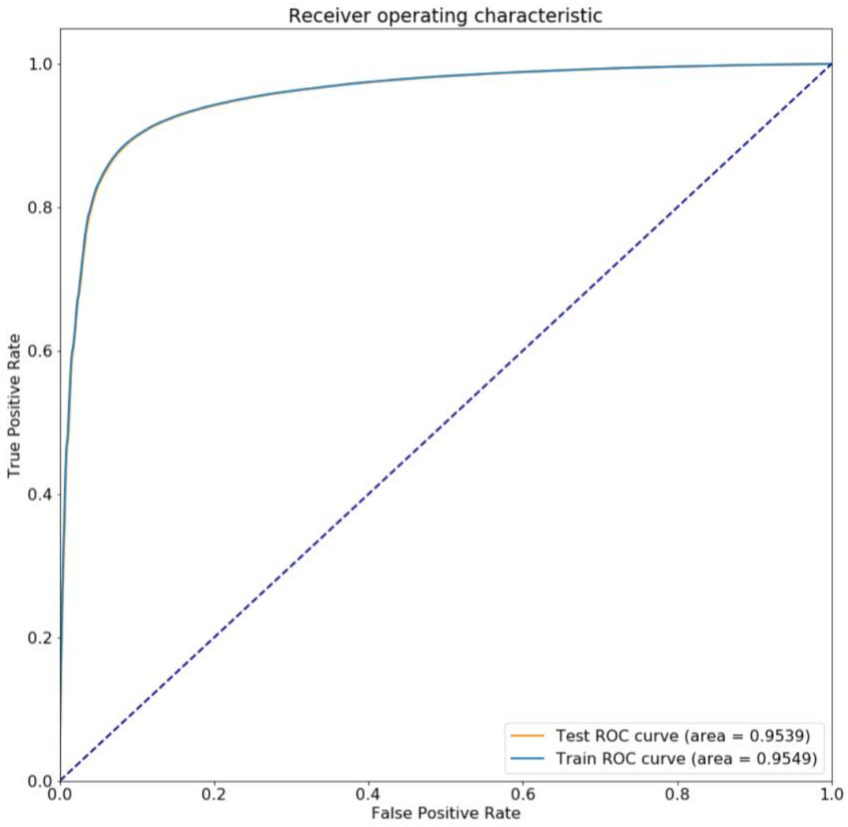
Kajaker jeg den

Den sidste gode bog jeg læste

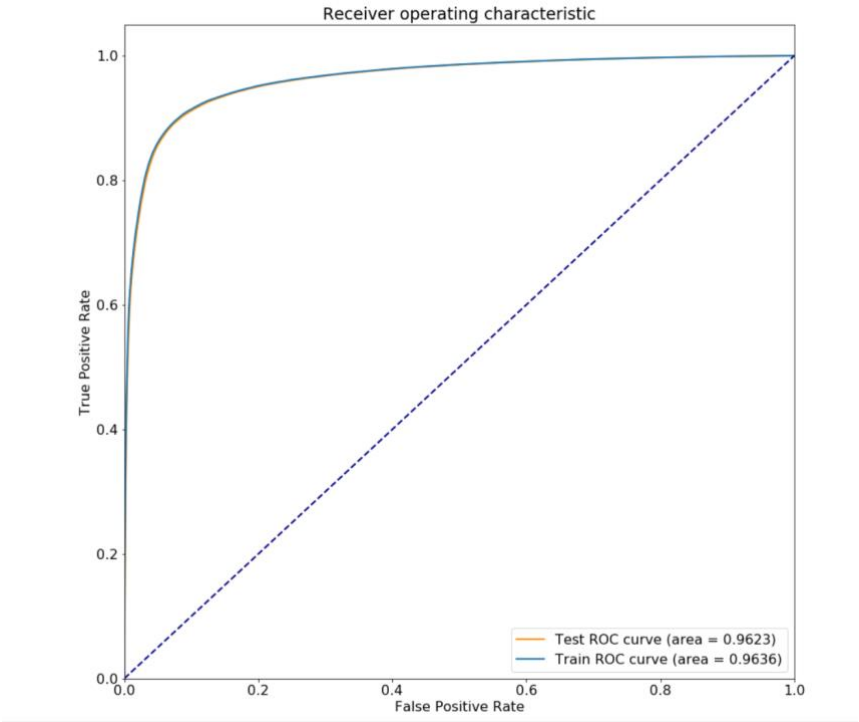
Modernist Cuisine

Appendix 2

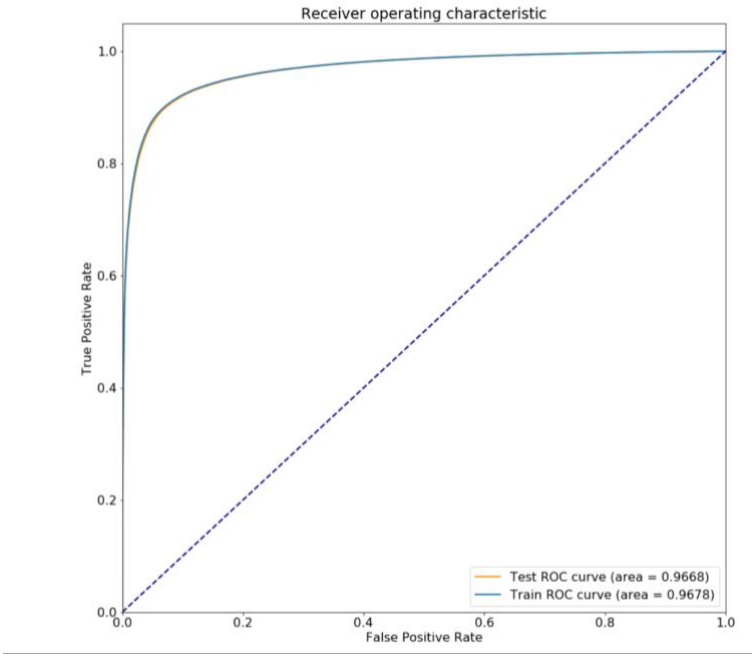
K = 2



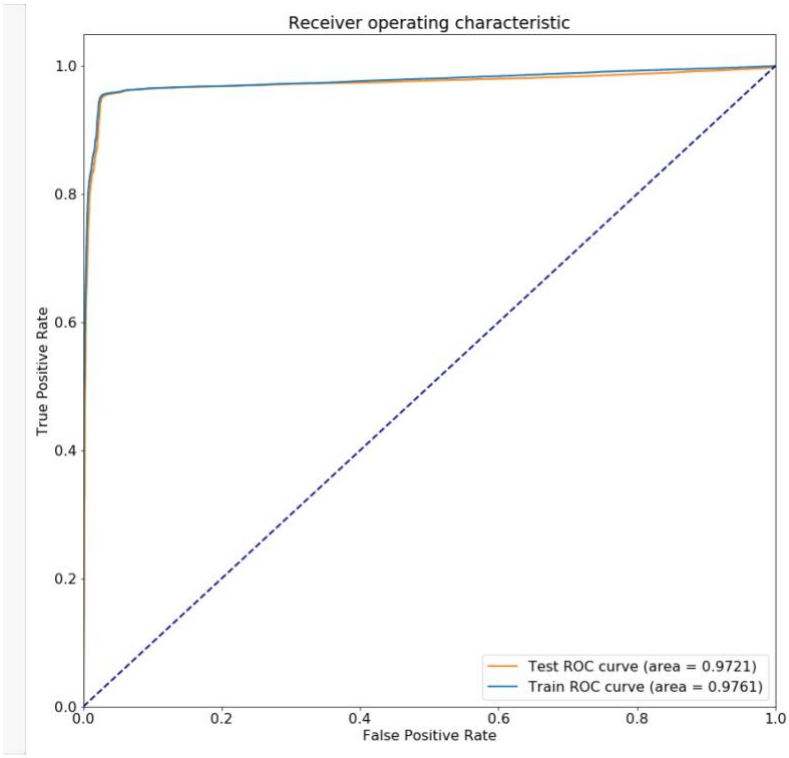
K = 3

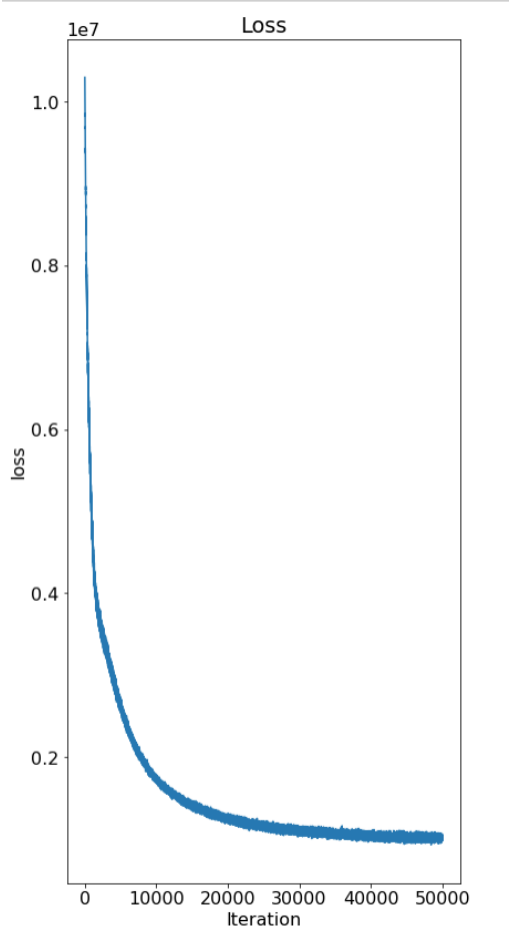

Receiver operating characteristic
True Positive Rate
False Positive Rate
Test ROC curve (area = 0.9623)
Train ROC curve (area = 0.9636)

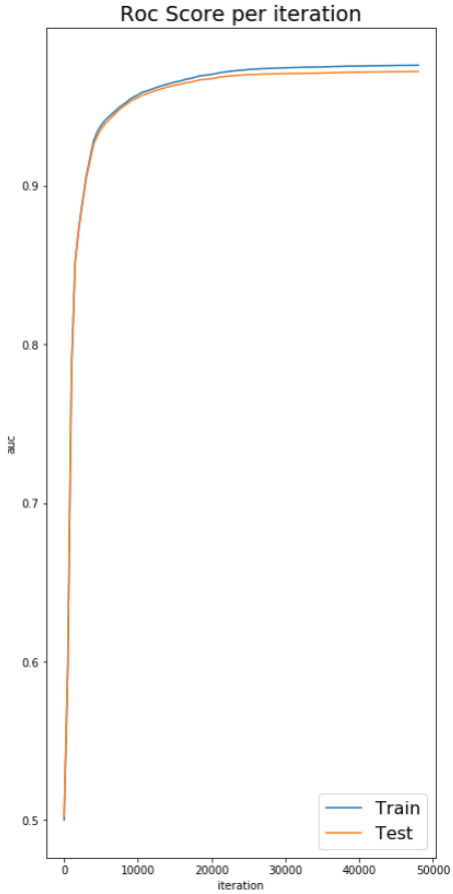
K = 4


Receiver operating characteristic
True Positive Rate
False Positive Rate
Test ROC curve (area = 0.9668)
Train ROC curve (area = 0.9678)

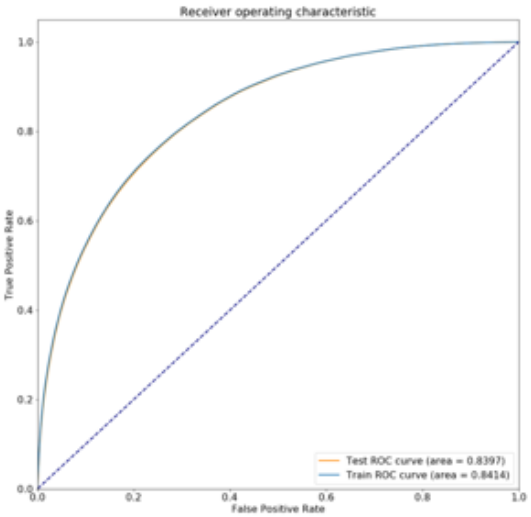
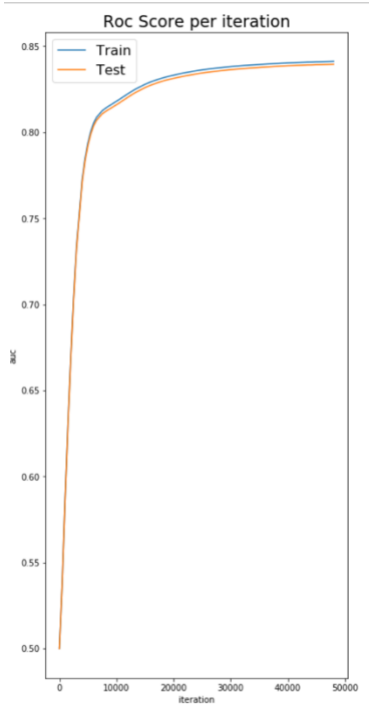
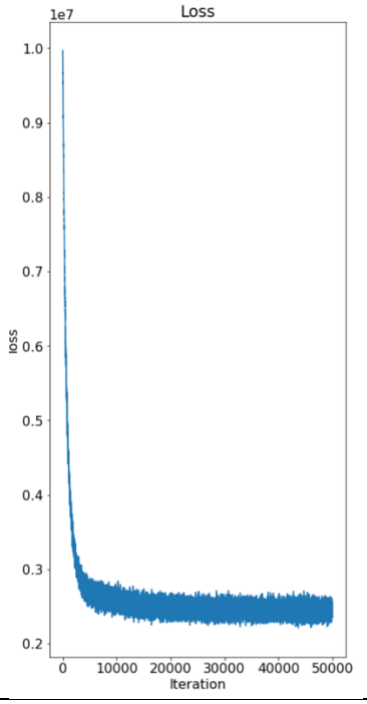
Appendix 3


Receiver operating characteristic
True Positive Rate
False Positive Rate
Test ROC curve (area = 0.9721)
Train ROC curve (area = 0.9761)

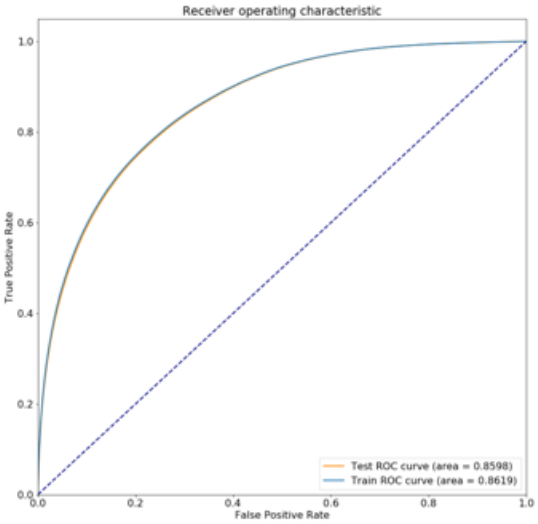

Loss
1e7
loss
Iteration

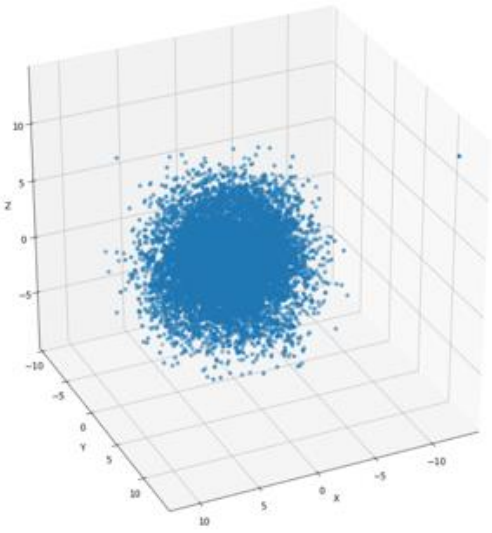

Roc Score per iteration
auc
iteration
Train
Test

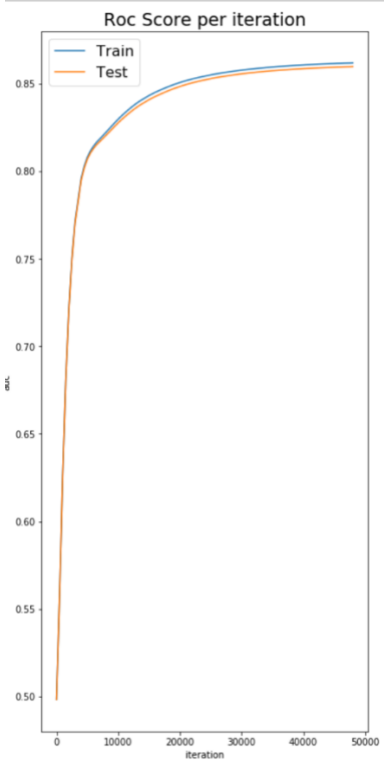
Appendix 4

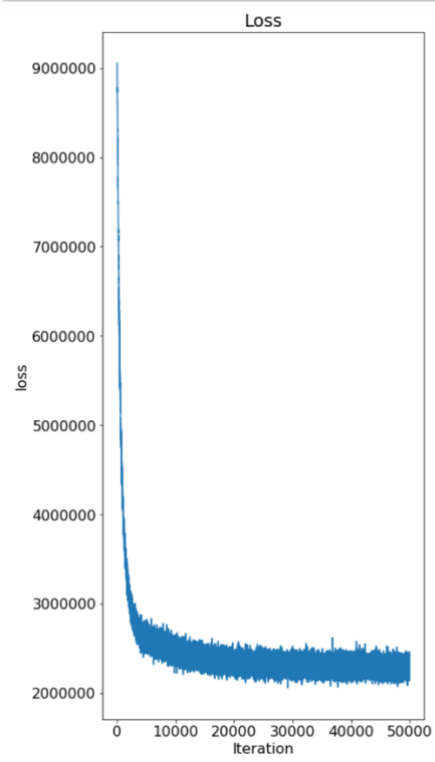
Charlotte models		
K2 		

K3


Receiver operating characteristic
True Positive Rate
False Positive Rate
Test ROC curve (area = 0.8598)
Train ROC curve (area = 0.8619)




Roc Score per iteration
Train
Test
iteration


Loss
loss
Iteration