Where will You Setup Your Business Next?

A Machine Learning Approach to Suggest Ideal Geographical Location for New Restaurant Establishment

SUBMISSION DATE: 22.07.18

SUBMITTED BY

Ibne Farabi Shihab (14201002)

Maliha Moonwara Oishi (14301011)

Department of Computer Science and Engineering

SUPERVISOR

Hossain Arif
Assistant Professor
Department of Computer Science and Engineering

DECLARATION

We, hereby declare that this thesis is based on results we have found ourselves. Materials of work from researchers conducted by others are mentioned in references.

Signature of Supervisor **Signature of Authors**

Hossain Arif
Assistant Professor
Department of Computer Science and Engineering
BRAC University

Ibne Farabi Shihab – 14201002
Maliha Moonwara Oishi – 14301011

ABSTRACT

A restaurant business is a very prospective and profitable business nowadays. However, ensuring quality food, good stuff, inner-environment etc. is a big concern and most importantly before facing all these, the trickiest part is to choose a perfect place where a restaurant business will flourish. Without doing a perfect research on this area, setting up a restaurant may lead to an immediate downfall. Not only for choosing a preferred restaurant location, people are now hiring professionals to do ground check and here the data scientists are coming into play as a bigshot. This research is focused on suggesting a suitable place for setting up a restaurant business based on the existing data from Yelp where 75 features have been extracted for analysis. Several machine learning algorithms (Support Vector Machine, Decision Tree, Logistic Regression and Decision Tree with presort) have been used and juxtaposed to nurture out the suitable one. As yelp's review is authentic and it is maintained regularly, we have considered the rating of a business as the point of suggestion. We have also looked at the comparative analysis of this algorithm and searched for an algorithm which gives us the best result.

ACKNOWLEDGMENT

First of all, we would like to express our deepest sense of gratitude to Almighty Allah. Secondly, we would like to express our sincerest gratitude to our advisor Hossain Arif for his continuous support, patience, motivation, and immense knowledge of our research. His guidance helped us in all parts of the progress.

Finally, we would like to express our sincere gratefulness to our beloved parents, brothers, and sisters for their love and care. We are grateful to all of our friends who helped us directly or indirectly to complete our thesis.

TABLE OF CONTENT:

LIST OF EQUATIONS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1
OVERVIEW

1.1 **INTRODUCTION**

The ever-growing amount of available digital information and the number of visitors to the internet have created a potential challenge of information overload which effects our timely access to the items we want [1]. Choosing the best approach that is suited for a specific application or product is still in its early stage [2]. Only a few studies have been done comparing evidence for deciding among different systems [3]. Companies like Google, Bing and Duck Duck Go have managed to optimize the search and make things much more concise. Often, we want an optimum search result based on specific keywords and which is exactly what traditional search engines do. However, sometimes we do not want a generalized solution, rather a solution that will be prioritized based on the user's preference and their choice [4]. Keeping an eye on these choices, recommender systems have emerged [5]. The recommender systems which generalize the search are called non-personalized recommender system and the ones that deal with personal choices are called personalized recommender system [6]. After the emergence of recommender systems, the term Recommender System has gained popularity among people. As human psychology varies from person to person, so do their choices and this is exactly why they tend to prefer a system like which can suggest better recommendations based on their personal data [7]. Many product-based websites (online grocery stores, tech shops, app stores etc.) are nowadays using recommender systems to deal with individual preferences and suggest the products they are most likely to be interested in. In fact, we even see the use of

networking sites such as Facebook. They use this type of recommendation for showing relevant pages, advertisements or even for suggesting groups [8]. A location-based recommender system is also becoming a popular branch of its application after the emergence of GPS [5,7]. Depending on the need, different kinds of recommender systems have been deployed. The top N item recommendation and rating prediction are also turning out to be very popular. POI (point of interest) is one such N item recommendation system which mainly focuses on the geographical similarity of one user with others and suggests things to a user which they might what to have a look at [9]. On the other side, there are rating prediction systems too [10]. How a user will rate a certain item, which they have not rated before, is the main concern of this system. In our work, we are trying to do the rating prediction in the context of the restaurant business. If anyone wants to set up a business, our system would tell them the suitable places to set up that business depending on the average rating of the users using machine learning.

1.2 RELATED WORK

In a research article [11], Lunkad predicted the rating of a restaurant using data from Yelp. His concentration was on the two sub-datasets (business and review) of yelp dataset. In his work, he tried to predict the rating of the restaurant using the review subset. The business sub-dataset contains state, city, name, average stars, business id etc. and review file contains user id, business id and review. Six major attributes were considered from these two sub-datasets. Later, he used a support vector machine, linear regression, and naïve bias model to have a look at how it works. Among these three algorithms, linear regression and support vector

machine works well though linear regression (53.13%) was slightly better than support vector machine (52.35%). He divided his dataset among 3 sections called development set, cross-validation set, and testing set. He used the development set to develop the model, the testing set was to test the model and the cross-validation set to eliminate the possibility of overfitting.

Another work by Wang [12] et al. have aimed to predict new restaurant success as well as rating. Moreover, their goal was to recognize the restaurant features that control the success of the restaurant business. They wanted to predict the range in which a restaurant business can be successful. They have used yelp dataset for the prediction of restaurant success and rating as this dataset is very standard and easy to use. For classifying restaurant features, that contains most weight, they have run the Chi-squared test as well as stochastic gradient descent. For this purpose, they have used a variety of binary and multi-class classification algorithms such as Support Vector Machine (SVM), Random Forest, Logistic Regression, and Multilayer Neural Networks. Their aim was to guess a restaurant's success and they have rounded ratings to the nearest star. After performing these algorithms, they have found that two algorithms among the four, performed really well. These two algorithms are Random Forest and Multilayer Neural Networks. The accuracy of these two algorithms is 56% for multi-class classification and 60% for binary classification. Later, they have performed sentiment analysis on restaurant reviews and after undergoing several processes the accuracy increased to 85% using clustering algorithms combined with the mentioned work.

In another paper, Yu [13] et al. have used yelp dataset for predicting business success and rating, as yelp dataset provides the connection of people in the local business. In their paper, the authors have mainly concentrated on the reviews for the restaurants. Moreover, they have also noticed that the sentiment features are very useful for rating prediction. Star rating is the most useful option where users

can make their choices among all the businesses that are available. The higher the rating is, the higher the chance to be liked by the users because the users know that higher star rating ensures the high quality of service. So, in yelp dataset, the star rating encourages the users to judge specific business as they know that people will only give a higher rating if they are actually satisfied with the business. Therefore, the star rating is the way to evaluate a business success. For the entire investigation, they have used the yelp dataset. In their project, the authors aimed to predict the star rating for the review of the restaurant. For this purpose, they have used three machine learning algorithms: Linear Regression, Random Forest Tree, and Latent Factor Model, which were then combined with sentiment analysis. They have evaluated each individual model to check which algorithm gives the best accuracy. After the evaluation, they have found that the random forest tree algorithm gives the maximum accuracy of around 85%.

In another paper [14], the authors wanted to predict the rating of a neighbor of an already rated business and the reason behind this is that there is a high possibility that they are more likely to go nearby that business. They came to a decision that there is a weak correlation between the rating of a business and its neighbor. They used two kinds of factors (intrinsic and extrinsic) of latent factor model for deciding this. Using geographical location model, they have achieved much less error compared to the state of art models like social MF, biased MF and SVD++. They have used matrix factorization for this task. They talked about three observations. Firstly, most businesses have neighbors within a short geographical distance from their five locations. Secondly, observations are weakly positively correlated, and lastly, the observation is for all type of business. They also used factors like business category, popularity, and review content. As future work, they have shown interest in investigating the influence of geographical neighborhood in POI recommendation and sentiment analysis of business reviews.

In this research paper Jain [18] et al., they wanted to predict whether a movie will be successful or not depending on the data from Twitter (360 million ‘Tweets’ and 160 million search per day) before the release of the movie which they called is hype. We as people love to share our opinions in social media and they wanted to analyze this as the parameter on giving the decision of success. As filmmaking business is considered to be one of the riskiest business they wanted to give those people a certain kind result depending on the pre-release hype on Twitter. They also addressed a task where the prediction was based on the motion movies initial release. The focused on this social media as the activity of people is highest on the social media. They have extract data regarding movie before pre-release and show their analysis in a graphical model which considered to be easy for the business people. For their measurement, they took the ratio of the user and the number of “Tweets’ and considered a movie as a success if the ratio is more close to 1. They claimed that their way has outperformed the collaboration technique which attempts to learn from the past user relationships (how the user will rate a new movie given their previous rating.). Using their equation they came to the conclusion that the Skyfall movie’s hype is 0.497 and they also predicted the opening week collection will be 3976000$ per day. Lastly, they did not talk about any future work or any kind of improvisation which can be done on their paper.

CHAPTER 2

MACHINE LEARNING APPROACHES & ALGORITHMS

2.1 MACHINE LEARNING AND ITS APPROACHES

In the introduction part, we have said that we will follow a machine learning approach. Now, the first question is what is machine learning? According to Tom M. Mitchell [19], "a computer program is said to learn from experience *E* with respect to some class of tasks *T* and performance measure *P* if its performance at tasks in *T*, as measured by *P*, improves with experience *E*." In-plane words, machine learning is an approach where we try to give intelligence to the machine so that it can do a specific task more like a human. Machine learning itself is a huge field and thus it is divided into 4 branches. Below this branches will be explained briefly.

2.1.1 SUPERVISED MACHINE LEARNING

In the field of machine learning, most of them are supervised learning. Supervised learning is where we have input variables (X) and output variable (Y) and we use an algorithm to learn the mapping from the input to the output.

$$Y=f(X) \quad (1)$$

The goal of this approach is to learn the mapping well so that for the new input (X), we can accurately predict the output (Y) of that corresponding input. It is called supervised as the algorithm learn the data and correct itself from the output during the learning stage and keep learning until it reaches an acceptable stage. The main

difference between supervised learning with unsupervised is that the supervised one is labeled data which means we know the outcome. Add to that, we also have the idea of the distribution. This supervised learning further gets divided into two parts. They will be discussed below:

2.1.1.1 CLASSIFICATION

In this type of supervised learning, the output is categorical which means they are divided into different classes. Suppose we are identifying the color of items assuming all items have only one color. Red, blue, green etc. are the output classes for our classifier. Another example is when we take binary type decision. For example, whether I will go to school or not depending on the weather condition. If the weather is good, I will go and if the weather is bad, I will not go. This is a binary type classifier.

2.1.1.2 REGRESSION

In case of classification, the values were categorical whereas, in regression, the values are continuous or in other words real value like taka or dollar. For example, let’s say we are trying to predict the price of the stock market. We want to see what will be the price of shares in the stock market in coming 7 days. We cannot divide this into fixed classes unless we grouped the price range. So, in this context, our algorithm will output continuous or real values like taka.

Some common types of problems built on top of classification and regression include recommendation and time series prediction respectively. In supervised learning, we can also measure the accuracy of our model and thus, we have the

control over the environment. Usually, we do this accuracy measurement splitting the data into train and test set.

2.1.2 UNSUPERVISED LEARNING

In supervised learning, we have the input variables (X) and the output variable(Y). In unsupervised learning, though we have the input variables (X) we do not know the output (Y). At first look, it might look a bit awkward as there is not an output label. Then the question arises, how this learning work? In unsupervised learning, what we do is, we divided the inputs in some sort of groups using different approaches. Algorithms are left to their own devices to discover and present the interesting structure in the data. It might be on similarities or how close they are to each other. The goal for unsupervised learning is to model the underlying structure or distribution in the data in order to learn more about the data. This unsupervised learning further can be divided into two classes. They are explained below:

2.1.2.1 CLUSTERING

In this type of unsupervised learning, what we do is try to discover the inherent groupings in the data. For example, many customers buy different kinds of foods from a super shop. We want to have a look at the different customer's type. We want to hold our regular customers. To do so, we will give them a small amount of discount automatically from our system. For this, the purpose we need to find out those customers from our selling record and have to do it manually. But, if we use

a clustering we can easily get to know about our regular customers and can give them a small amount of discount easily without any hassle. So, thus clustering can be used to get inside about the customers.

2.1.2.2 ASSOCIATION

Another type of unsupervised learning is the association. We use Facebook and YouTube on daily basis. If we click on an advertisement, we often see that similar kind of advertisements is showing in our side panel. Another example is when we listen to music or watch any kind of videos on YouTube. We see a similar kind of music or video is showing in the side panel. These things are done by the association. When a user listens to a group of music's or clicks on several advertisements, this advertisement or music's or videos get grouped. Then, when a new user clicks on an item of this group others are shown in the side panel. Thus, most of the times, we do see the things we like and credit goes to the association.

In unsupervised learning, we have less control over the environment as there is not a way to measure the accuracy of the model.

2.1.3 SEMI-SUPERVISED LEARNING

In the field of machine learning, semi-supervised learning occupies the middle ground, between supervised learning (in which all training examples are labeled) and unsupervised learning (in which no label data are given). Interest in Semi-supervised learning has increased in recent years, particularly because of application domains in which unlabeled data are plentiful, such as images, text, and

bioinformatics. Basically, it is the mixture of labeled and unlabeled data. It is useful in many ways. Labeling a huge amount of data is time-consuming and expensive. Add to that, it also can impose human biases on the model. Thus, it means that supervised learning improves the accuracy of the model. The little bit of labeled data is the starting point for the algorithm and kind of a clue to start its labeling and to label a large amount of data. This approach is widely used in genetic sequencing, web page classification and so on.

2.1.4 REINFORCEMENT LEARNING

Reinforcement learning is the last branch among the 4 branches of machine learning. It allows machines and software agents to automatically determine the ideal behavior within a specific context, in order to maximize its performance. Simple reward feedback is required for the agent to learn its behavior; this is known as the reinforcement signal. This behavior can be learned once and for all, or keep on adapting as time goes by. If the problem is modeled with care, some Reinforcement Learning algorithms can converge to the global optimum; this is the ideal behavior that maximizes the reward. **T**his automated learning scheme implies that there is little need for a human expert who knows about the domain of application. Much less time will be spent designing a solution, since there is no need for hand-crafting complex sets of rules as with *Expert Systems*, and all that is required is someone familiar with Reinforcement Learning. For example, a baby usually does not know what heat is. Now, if he touches a glass of water with high temperature, he will feel the heat and will identify that I have a dangerous thing. Next time, whenever he comes across this glass, will not touch it. This is called reinforcement learning.

2.2 ALGORITHMS

In the last chapter, we discussed the different types of machine learning approach. To use this approaches, we need to use machine learning algorithms. In this chapter, we will discuss regarding those algorithms and their pros and corns.

2.2.1 LINEAR REGRESSION

Linear Regression attempts to model the relationship between two variables by fitting a linear equation to the observed data. One variable is considered to be an explanatory variable or independent variable and the other is considered to be an independent variable. The dependent variable is the variable which value we want to focus on or which value we have to calculate. On the other hand, the independent variable is the variable which does not depend on other variables. The values are independent. This algorithm performs very for the continuous data in the output column or dependent variable. The dependent variable is dependent on the independent variable. Suppose we have five people's height and weight. If we draw a graph:-

TABLE 1 : LINEAR REGRESSION EXAMPLE

Height(cm)	65	65	62	67	69
Weight(pound)	107	124	109	120	155

Least Square Criteria:

In the following graph, there is height in the X-axis and weight in the Y-axis and the points are put depending on the weight. If we want to put a line that matches the least points it is called least squares criteria.

Linear Model will have the smallest value for D.

D=(X1)^2+(X2)^2+…….(X8)^2 (2)

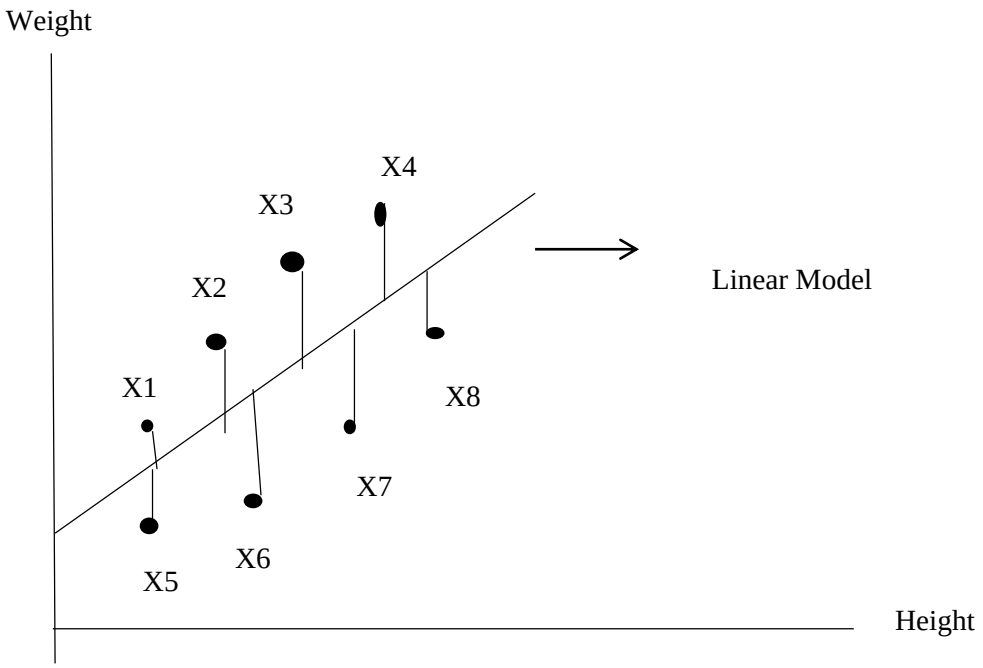

Figure 1: Linear regression explanation

2.2.2 LOGISTIC REGRESSION

For exploring the dataset, a statistical method is used which is called logistic regression. So in another word, logistic regression is the statistical method of

evaluating the particular dataset. Logistic Regression has similarity with the concept of probability. If an event occurs, the probability will be:-

P(event)=occurred event/total number of event

But in Logistic Regression, the probability will be the event is occurring versus the event is not occurring.

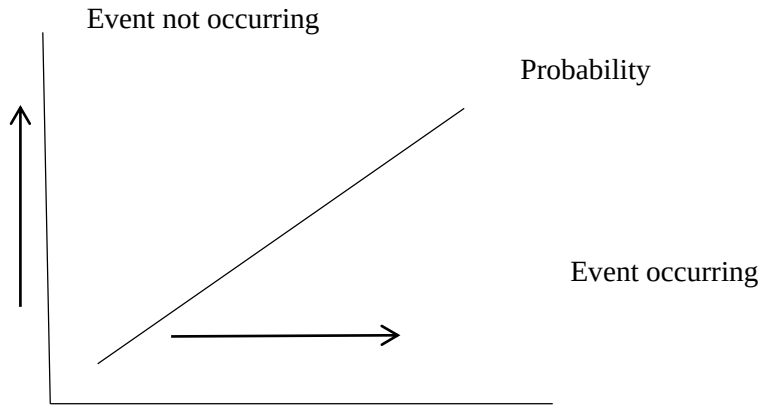

Figure 2: Logistic regression explanation

Logistic Regression usually gives two types of data. Here the data can be categorical or the data can be numerical. Therefore, Logistic Regression is used to determine the categorical outcome variable. The categorical variable only takes the limited values. So it is called the categorical variable. The example of a categorical variable is gender. Gender can be male and female. Another example is temperature. Temperature can be hot, cold, humid etc. Numerical data can be in a rage for example 0 to 10 or 1 to 5 etc.

So here is an example of Logistic Regression. Suppose there is a data of students where there are two attributes. One is hours of study and another is pass or fail. 0 means fails and 1 means pass. The data is given below:

TABLE 2: LOGISTIC REGRESSION EXAMPLE

Hours	result
1	0
2	0
3	0
4	0
5	0
6	1
7	1
8	1
9	1
10	1

So from the table [2], we can see a student can pass if he/she study 6 hours to 10 hours. So the scatter plot of the data is given below:

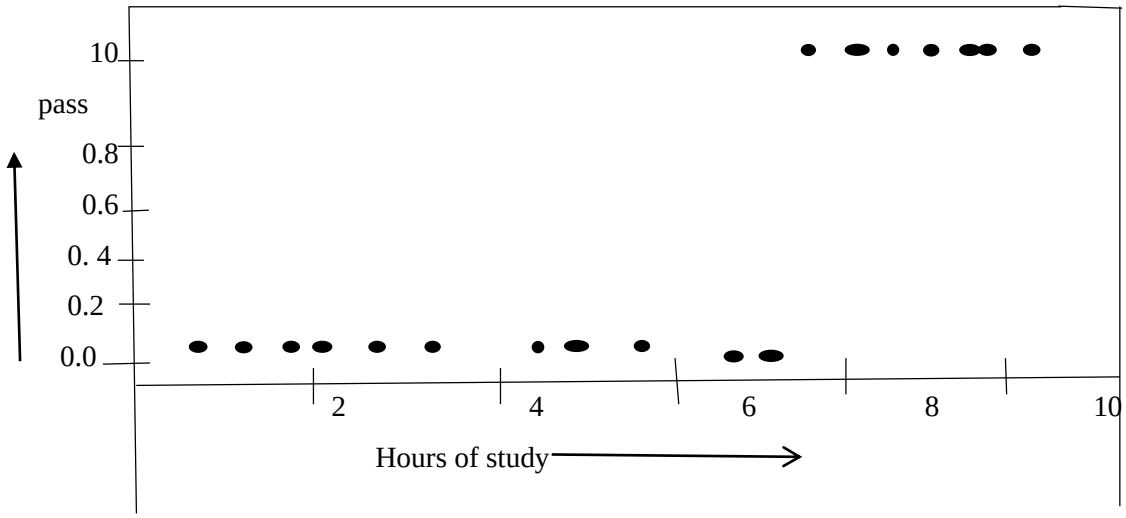

Figure 3: Scatter plot of the data (Logistic Regression)

If we do normal Linear Regression in our data, then the graph will be like this:-

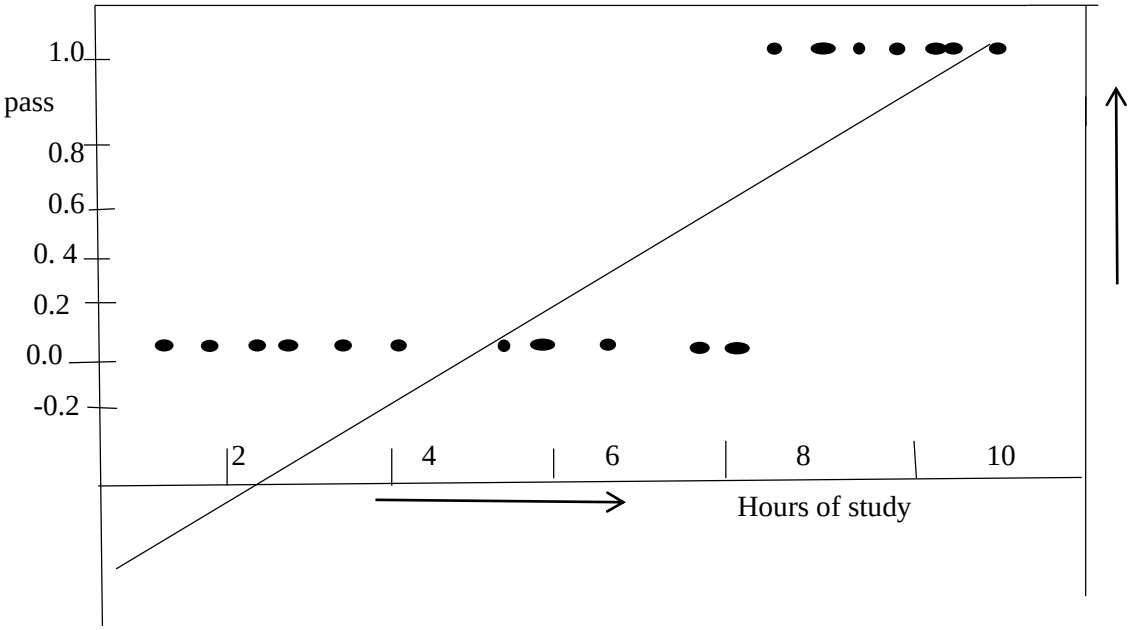

Figure 4: Scatter plot of the data (Linear Regression)

Linear Regression produces all types of values between 0 and 1. Moreover, it also takes the negative values. Therefore, Linear Regression produces negative values as well as values greater than 1, but negative values have no meaning. Therefore, Logistic Regression plots the data in the following way:-

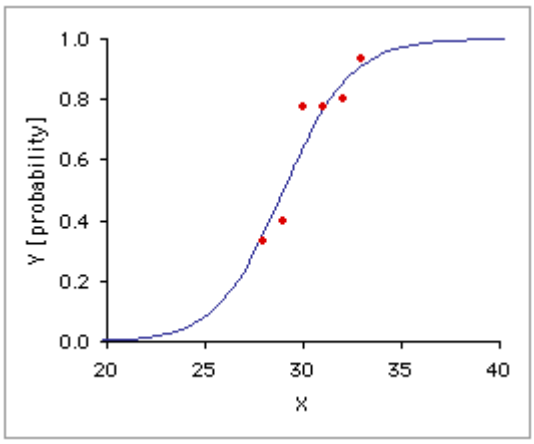

Figure 5: Logistic regression value ploting

2.2.3.1 Logistic Function

We usually predict the outcomes as yes or no or 1 or 0. The formula of logistic function is:

F(x)=L/(1+e^-k(x-x0)) (3)

Where,

L = curves maximum value

K = steepness of the curve

X0 = x value of sigmoids midpoint.

A standard logistic function is called sigmoid function. If k=1, x0=0 and L=1 then the equation then becomes:

S(x)=1/1+e^-x (4)

2.2.3 DECISION TREE

A decision tree is an algorithm which is used to reach in a decision or to get the target value. Ross Quinlan invented this algorithm. The target values can be set in a tree model called classification tree. In this tree, there are leaves and parents like other trees. Here the leaf nodes represent the class labels and branch represents features of that label. Decision trees where target values can take continuous values is called a regression tree. We use a decision tree to describe the whole decision-making process in a particular way which makes the entire thing easily

Understandable. Decision tree normally used operation research especially decision analysis. For example, suppose a bank will give loans depending on age. So the age will be first distribute in 3 sections. The division will be age 20, age 20-50 and lastly, age 50 up. The people who are 20 years old will get a 2% loan. People who are above 50 will get 5% loan. Those, whose age are between 20 - 50 will be classified into two categories. One is married and the other is unmarried. Those who are unmarried will get 10% loan and those who are married will again classify into two leaves, one those who have kids and those who don't have kids. The people who have kids will get 80% loan and who do not have kids will get a 20% loan.

2.2.3.1 ADVANTAGES OF DECISION TREE

It performs the future selection. When a new node is written, it can be classified into future classes or new leaves or labels.

It’s simple but it can actually describe lots of things. So describing a large thing in a particular way, a decision tree is very useful. It needs little effort to write the whole thing in the decision tree.

In the support vector machine, the relationship between parameters does not affect trees performance.

2.2.3.2: DISADVANTAGES OF DECISION TREE

The plan and the tree may not the same. Sometimes it happens the outcome does not fulfill the expectation level.

A large tree can be complex to execute.

2.2.4 SUPPORT VECTOR MACHINE

Support vector machine or SVM is one of the machine learning algorithms. SVM introduced in COLT-92 by Boser, Guyon and Vapnik. It is a well-motivated algorithm which developed from Statistical Learning Theory. SVM is a supervised learning algorithm that means the machine will be supervised by us to produce future predictions like other supervised learning algorithms do. SVM is mainly used for solving classification and regression challenges. Though it is used for both classification and regression, it is largely used in classification problems. SVM is used to split the data. Spitting process can be of many types but in SVM, the splitting is not random. The best possible way to split the data is to choose a line between the data that is in maximum distance from both the data. If we put a random line that is close to one side and far from another side, it's not the proper way of splitting. Moreover, if we put the straight line in a place where the distance between the line and the data is narrow is not also the exact way of splitting the data. So the best way is to place the straight line in a place where the line is far from both data. Hence, the line should be as far as possible from both datasets. The distance between the straight line and the dataset can be named as margin. The margin should be as wider as possible to get the best result. The straight line that separates the two datasets is known as a hyperplane. The datasets separating by the hyperplane or the points near to the hyperplane are the support vectors. For example, let's look at the below figure.

2.2.4.1 WHEN AND WHERE TO USE:

We can use support vector algorithm to identify the right hyper-plane. For example:-

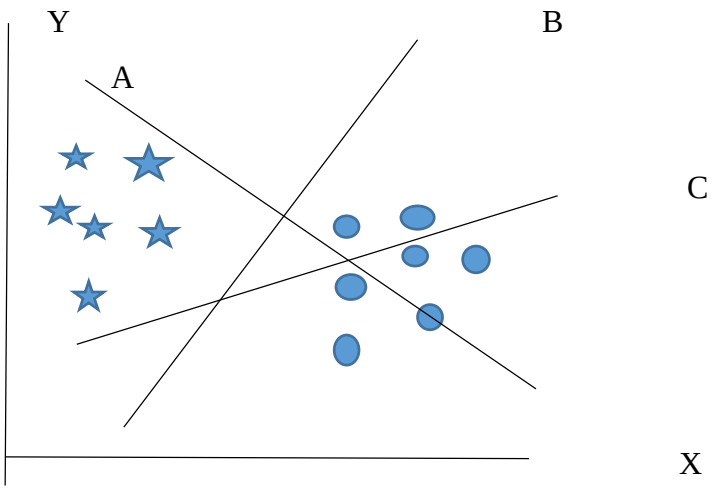

Figure 6: Support vector machine hyper-plane

In figure 6, there are three lines between star and circles. Among this three hyperplane, SVM is used to choose the right hyperplane that separates starts and circle equally. In this graph, B separates the start and circle equally. We need the thumb rule to identify the right hyper-line.

SVM has a technique which is called the kernel trick. These function which takes low dimensional input space and transforms it into a higher dimensional space. It is mostly used in non-linear separation problem.

2.2.4.2 ADVANTAGES OF SUPPORT VECTOR MACHINE

i) Prediction accuracy is mostly high.

ii) Robust works when training examples contain errors.

iii) Fast evaluation of the learned target function.

2.2.4.3 DISADVANTAGES OF SUPPORT VECTOR MACHINE

i) Long training time.

ii) Difficult to understand the learned function.

iii) Not easy to incorporate domain knowledge.

CHAPTER 3
DATASET & FEATURE SELECTION TECHNIQUE

3.1 DATASET

In the field of machine learning, collecting the data is a crucial factor. To obtain a good result, data is a key element as authentic and bias less dataset plays a vital role in the performance of a model. For this reason, we used a renowned dataset [15] from the website called YELP. This site is a USA-based site, storing different kinds of business information from 2005.

3.2 FEATURE SELECTION TECHNIQUE

Feature selection or in other words subset selection is an important fact in the field of machine learning, wherein subsets of the features available from the data are selected for application of a learning algorithm. This is an important stage of preprocessing and is the way of avoiding the curse of dimensionality and feature extraction. There are two approaches in feature selection. One is the forward selection and the other one is the backward selection. The main idea of feature selection is to eliminate the variables with little or no predictive information.

Feature selection method can be decomposed into 3 broad cases [7]. These filter methods, embedded methods and wrapper methods. These 3 methods are explained below:

3.2.1 FILTER METHODS

These methods select features based on discriminating criteria that are relatively independent of classification. Several methods use simple correlation coefficients similar to Fisher's discriminant criterion. Others adopt mutual information or statistical tests (t-test, F-test).

3.2.2 WRAPPER METHODS

Wrapper methods utilize the classifier as a black box to score the subsets of features based on their predictive power. Wrapper methods based on SVM have been widely studied in the machine-learning community. SVM-RFE (Support Vector Machine Recursive Feature Elimination), a wrapper method applied to cancer research is called, uses a backward feature elimination scheme to recursively remove insignificant features from subsets of features. In each recursive step, it ranks the features based on the amount of reduction in the objective function. It then eliminates the bottom-ranked feature from the results. A number of variants also use the same backward feature elimination scheme and linear kernel.

3.2.3 EMBEDDED METHODS

The inducer has its own FSA (either explicit or implicit). The methods to induce logical conjunctions provide an example of this embedding. Other traditional machine learning tools like decision trees or artificial neural networks are included in this methods.

3.3 BASIC FEATURE SELECTION ALGORITHMS:

3.3.1 CHI-SQUARE

Chi-Squared is the common statistical test that measures divergence from the distribution expected. It is a statistical test of independence to determine the dependency of two variables. Chi-square test is only applicable to categorical or nominal data. Let's say, we have a target variable and some features or in other words columns. Now, we calculate the chi-square between each and feature and the target variable. If the target variable is dependent on the feature, then this feature is important and if it is not, then we can discard this feature.

3.3.2 EUCLIDIAN DISTANCE

Euclidean Distance is the most common use of distance. In most cases when people said about distance, they will refer to Euclidean distance. Euclidean distance or simply 'distance' examines the root of square differences between coordinates of a pair of objects. For each feature Xi calculate the Euclidean distance from it to all other features in a sample. Euclidean distance d (Xi; Yi) between features Xi and Yi is calculated using the formula:

$$distance(x,y)=\{\Sigma i(xi-yi)2\}½ \quad (5)$$

One thing to keep in mind that Euclidian distance is used for raw data, not to standardize data.

3.3.3 RECURSIVE FEATURE ELIMINATION

As the name suggests, this method recursively removes features which do not contribute to the accuracy. In this method, the algorithm repeatedly trains a model (usually support vector machine or linear regression) and eliminate the features which have less contribution towards the accuracy until my given condition has met. Let's take an example. I am trying to select the best 4-6 features, resulting in the best accuracy. I start with 10 features and run the recursive feature elimination algorithm. In every iteration, it will eliminate the lowest ranked features and will go on until it meets the given condition. Depending on the result, I need to decide whether I will take 4 features or 5 features and so on. It is kind of a trial and error method.

3.3.4 CORRELATION BASED FEATURE SELECTION

This method searches feature subsets according to the degree of redundancy among the features. The evaluator aims to find the subsets of features that are individually highly correlated with the class but have low inter-correlation. The subset evaluators use a numeric measure, such as conditional entropy, to guide the search iteratively and add features that have the highest correlation with the class. Usually, Pearson's correlation is being used in this method.

CHAPTER 4
PRE-PROCESSING AND WORKING PROCEDURE

4.1 PRE-PROCESSING

To build up a successful machine learning model, we need to do pre-processing. The accuracy of a model largely depends on these pre-processing steps. It is the 50% of a machine learning model. There are usually 3 steps in pre-processing. Data cleaning, feature selection, feature scaling are the 3 steps. We followed this 3 steps which will be discussed below.

4.1.1 DATA CLEANING

Yelp dataset contains six sub-datasets: business, check-in, tips, review, photos, and user. For our work business and check-in, subsets were taken into consideration. In the business subset, we have 15 columns with 156635 instances of businesses. The columns that we have in our business subset are address, attributes, business_id, categories, city, hours, is_open, latitude, longitude, name, neighborhood, postal_code, review_count, stars and state. In the check-in subset, we have 135148 rows and 2 columns: business_id and time. In total, we have 16 columns as business_id, which is common in both of these subsets. There are some businesses without its data in the check-in subsets. We merged these two subsets by matching their business id. Among these columns, there are some irrelevant columns which are not necessary for our inferential work. business_id was kept for the identification of the business. Address, postal code, latitude, longitude, neighborhood, hours, is_open, and name are not necessary since these are for the identification of the business. Thus, we are left with 8 columns. Category column

denotes the type of the business (i.e. restaurants, shop etc.). As our concern is regarding the restaurant business, we picked the businesses where there is the word ‘restaurant’ in the category. Thus, we ended up with 49536 restaurants and dropped the category column as it was not necessary anymore. The interesting and tricky part is, there are 2 nested columns (attributes and time columns) among these 7. In attributes columns, there are 82 nested columns which define the specific attributes (i.e. car parking, alcohol etc.) of a restaurant. We flattened these columns to get rid of these 82 nested columns. The time column contains the time of check-in at different times of a day on an hourly basis. For example, 7 pm to 7.59 pm is counted as an hour and check-in count for this particular range is stored with respect to it. Our concern was to find the total check-in count of certain business and to do so, we aggregated this hourly check-in data into a single value of total check-in count. Lastly, we dropped the time column as the aggregated check-in count was calculated. Thus, we ended up these 88 columns.

4.1.2 HANDLING MISSING VALUES AND SELECTION OF FEATURES

There were some columns with missing values which could cause the anomaly in our dataset. For example, missing data can introduce a substantial amount of bias, making the handling and analysis of the data more arduous and can create reductions in efficiency and most importantly, can cause errors which need to be handled. For this purpose, we followed different approaches for different types of

values. To deal with the missing values of star column, we have used imputation. Imputation is a process to substitute the missing values with available information. We took the average of the whole column and substituted the missing values. For the ease of classification, we rounded those average values to their nearest value to maintain similarity with dataset columns value (1, 1.5, 2 and so on). Basically, for all the values which are numerical like review count, we have used imputation. The next type of values are the categorical values. In case of this type of columns, we substituted the missing values with the most frequent ones. Thus, we cleaned our data for our work. As we mentioned earlier, the motive of our work is to assist a person in choosing the suitable place for a restaurant opening. To obtain this, we have used the existing restaurant's data that we have cleaned. Through inference, we can let the user know the suitable places. In the previous paragraph, we discussed the procedure for cleaning the data. After cleaning the data, we were left with 88 features or in other words columns.

Later, the features we took into consideration were Is_open, review_count, checkin_count, AcceptsInsurance, AgesAllowed, Alcohol, 9 features related to ambience (casual, classy and so on), BYOB, BYOBCorkage, 7 features related to BestNights (Friday, Monday and so on), BikeParking, 5 features related to businessParking (garage, lot and so on), BusinessAcceptsBitcoin, BusinessAcceptsCreditCards, ByAppointmentOnly, Caters, CoatCheck, Corkage, 6 features related to DietaryRestrictions (gluten-free, vegetarian and so on), DogsAllowed, DriveThru, GoodForDancing, GoodForKids, 6 features related to GoodForMeal (breakfast, brunchand and so on), 8 features related to HairSpecializesIn (africanamerican, asian and so on), HappyHour, HasTV, 9 featuresMusic (dj, background_music and so on), NoiseLevel, RestaurantsDelivery, RestaurantsTableService, OutdoorSeating, Open24Hours, RestaurantsCounterService, RestaurantsAttire, state, city, RestaurantsTakeOut,

WheelchairAccessible, WiFi, RestaurantsGoodForGrous, RestaurantsPriceRange2, Smoking, RestaurantsReservations, RestaurantsTableService. After that, we did a feature selection which also played a vital role. There are different methods for feature selection and among those, we have used the chi-square test [16]. From the 88 features, the star is our output label and we excluded this from our feature selection procedure. Chi-square test is a statistical tool for measuring the importance of features. For this purpose, we set the value of alpha of chi-square to 0.001 to strongly discard the possibility of a null hypothesis which is suggested by many experts. After performing the chi-square test, some features got eliminated such as checkin_count, AcceptsInsurance, AgesAllowed etc. Eventually, after doing all of the processes we left with 75 features(there is an extra column which is for the business identification); and data is similar to the snapshot of figure 7 (a snapshot is given because of the space limitation; actual data consists of 75 columns/features and 49536 instances). For location identification, we used the business id to find the location of that respective place from the actual business sub-dataset.

is_open	review_cc	stars	checkin_c	AcceptsIn	AgesAllov	Alcohol	Am
0	0.000573	0.9	6	0	0	0	
0	0.001433	0.9	52	0	0	3	
1	0.003009	0.4	63	0	0	2	
1	0.00043	0.6	4	0	0	3	
1	0.002149	0.6	23	0	0	3	
1	0.00086	0.5	10	0	0	0	
1	0.006448	0.6	148	0	0	2	
1	0.005875	0.6	43	0	0	2	
1	0.007308	0.5	138	0	0	2	
1	0.000573	0.8	8	0	0	3	
1	0.016765	0.6	332	0	0	2	

Figure 7: Snapshot of data

4.1.3 FEATURE SCALING

At the last part, we did the scaling of the features and converted the categorical values to an integer number as algorithms only understand numerical values. In case of scaling, we divided all the values with maximum number in the respected column to keep the values between 0 and 1 which is known as zero mean normalization.

4.2 WORKING SETUP

After doing all the pre-processing, we started our work with 55 thousand restaurants with their 4 features (city, state, review count, check in the count). In our output column named ‘rating’ have values from 0 to 5(0, 0.5, 1, 1.5 and so on).

We have considered 0-2 as a single class as this signifies bad rating. For the next part, we went 0.5 for each step (2.5, 3, 3.5, 4 to 5). After doing so, we ended up with 7 types of output in our rating column. In our work, 4 features were the independent variables and the rating column was the output variable. After that, we applied supervised machine learning using linear regression, logistic regression, decision tree, support vector machine, multi-layer perceptron algorithms and looked at their accuracy to choose which algorithm works better on our dataset. Add to that, the rating was the output label of our work. The goal of our work is, depending on the 4 features how accurately our algorithm can predict the rating. From this model, we will say the place is better for setting up a restaurant.

CHAPTER 5
RESULT AND ANALYSIS

We began our work with a simple linear regression model. We used the Linear Regression model from Scikit learn [17], where we got 15.14% accuracy on the training set which is better than the random guess of 10%. We also got 15.8% accuracy using 10-fold cross validation which is similar to the training set, which proves that there was no under fitting. This algorithm will not work for sure since we are doing a classification problem (discrete values) whereas linear regression is for continuous values.

TABLE 3. SCORE TABLE

Name	**Accuracy score**	**Precision score**	**Recall score**	**Cross validation score**
Decision tree	60.48%	49.10%	60.48%	61.00%
Decision tree(with pre sort)	60.50%	49.08%	60.50%	60.76%
Logistic regression	60.30%	37.58%	61.30%	61.87%
Non-linear svm	97.02%	95.29%	97.25%	97.01%

Next, we used the Decision Tree. For this, we used the decision tree from Scikit learn. Here we got better results and the accuracy percentage was 60.48%. Overfitting was also controlled as the cross-validation is almost similar to the

accuracy score which handles one of the problems of the decision tree [Table 3]. The problem with the decision tree is that it does not work with out-of-sample values on most of the occasion. Thus, the decision tree does not work very well. Increasing the data will not increase the accuracy as the training and validation error goes parallel [Figure 8].

Figure 8: Learning curve for decision tree model

Figure 9: Learning curve for decision tree with presort model

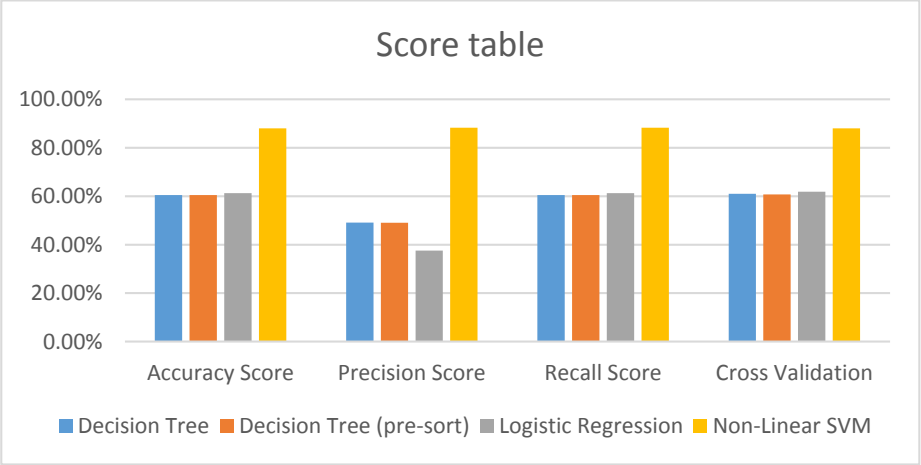

Figure 10: Score table

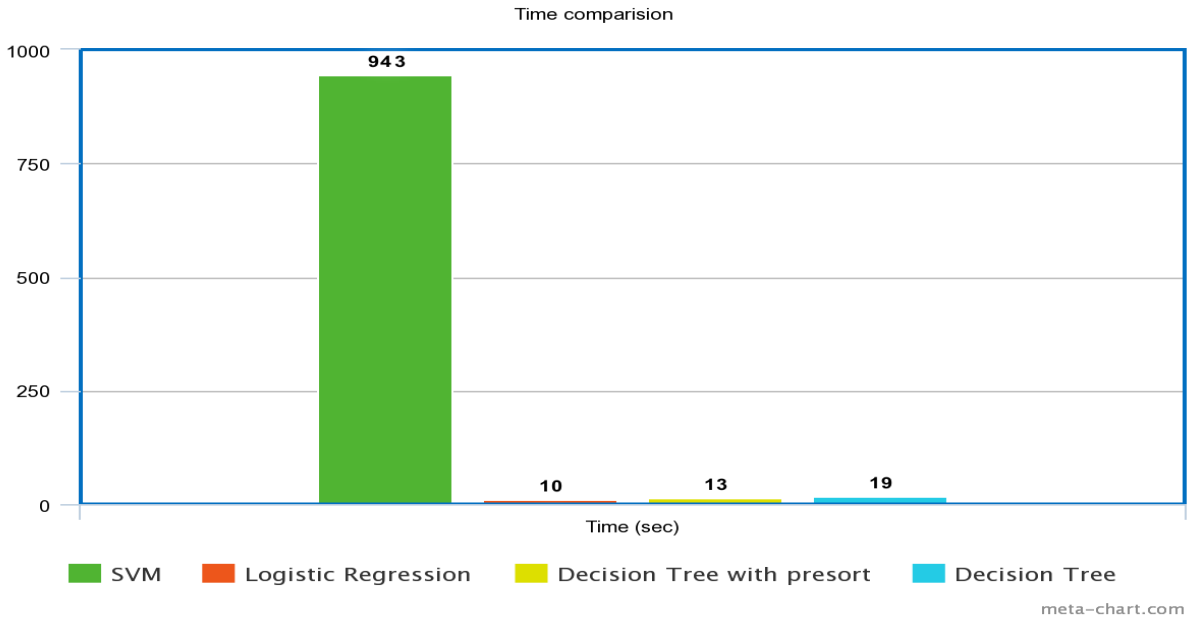

Figure 11: Time complexity graph

After that, we tried to use a variation of the decision tree (Presort) with the hope that we might find the best splitting condition. But, we were still unable to find the best splitting although we found a slightly better splitting condition and thus increased the accuracy by around 0.7% [Table 3]. In this algorithm, the increase in data size will not improve the accuracy for the same reason as figure 8 [Figure 9].

SVM is the algorithm which gives us accuracy to an expected level. It outperforms others algorithms in the context of precision, recall and accuracy score. [Figure 10] In case of time complexity, support vector machine is on the slower side [figure 11]. Since time is not a huge concern for our purpose, we preferred the support vector machine. Cross validation score is 97.01% which proves that it is not due to over fitting.

In our task, we have used 75% of data as for training purpose and 25% for test purpose. Support vector machine was able to correctly classify the rating of 12014 from 12384 restaurants from the test dataset. We took this number of restaurants and tried to find the top 5 cities for establishing the restaurant. Restaurants rating between 3.5 and 5 are considered to be the good one. Thus, we tried to find out the cities with the most restaurants in it with stars between 3.5 and 5. Correctly classified restaurants numbers of respecting ratings (3.5, 4, 4.5,5) are given in Table 4.

TABLE 4: CORRECTLY CLASSIFIED STARS

Stars	Correctly Classified Stars
5	302
4.5	1432
4	3123
3.5	1131

Now, according to our goal, we tried to find top 5 cities to establish a restaurant business. For this, we calculated the cities which have the most number of restaurants with our above-mentioned range. The top 5 cities for establishing a restaurant is shown in Table 5.

TABLE 5: TOP 5 CITIES BASED ON MOST RESTAURANTS

City Name	Restaurants Numbers
Richmond Heights	15
Charlotte	30
Toronto	20
Phoenix	16
Las Vegas	29

Later, scoring has been done. For this, we multiplied the number of restaurants with the respective rating (restaurants with rating 3.5 got multiplied by the number restaurants having the rating of 3.5 and so on) from this scoring, top 5 cities have been found and these are listed below in Table 6.

TABLE 6: TOP 5 CITIES BASED ON SCORE

Name Of city	Rating3.5	Rating 4	Rating 4.5	Rating 5	Score
Las Vegas	5	3	13	9	133
Toronto	3	5	7	5	87
Phoenix	3	2	8	3	69.5
Montreal	2	3	3	7	64.5
Mississauga	6	3	3	3	61.5

According to our analysis, we found that Las Vegas, Toronto, Phoenix, Montreal and Mississauga are the top 5 cities to set up a restaurant business.

Chapter 6
FUTURE WORK AND CONCLUSION

Thus far, we only considered the business and check-in sub-datasets. Later, we plan on adding the review sub-dataset to get more insights on the restaurant business. We are also thinking of adding the user object to analyze the behavior of users and have a look at the user's preference level to get more insights in the context of the restaurant business. Lastly, we are also considering to predict where a user will go if he/she go to another state based on their previous preference. To conclude it all, in our work we have shown that by using 75 features, we can predict the rating of a business using SVM algorithm on the data collected from YELP. We did not use a group of complex variable combinations which keeps our model simple and with good accuracy. To accomplish this, we took 75 features to predict the average rating of a business, in which we achieved 97.02% accuracy. We have also shown comparative analysis among different classification algorithms. From the comparative analysis, we found that for our dataset, SVM is the best in terms of accuracy. Lastly, we suggested suitable places for setting up a restaurant business. For this purpose, we suggested top 5 cities in 2 ways. First one is based on the amount of restaurants in a city and last one is based on score. We hope that our work will facilitate the business people in taking decisions regarding setting up a restaurant business.

REFERENCES

[1] [1] K. M. G. Hoq, “Information overload: causes and consequences: a study”, 2014 Philosophy and Progress: Vols. LV-LVI, ISSN 1607-2278 (Print)

[2] [2] G. Adomavicius , A. Tuzhilin, 2005, “Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions”, IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, volume: 17, pages: 734-749

[3] M. D. Gemmis, L. Laquinta, P. Lops, C. Musto,F.Narducci, G. Semeraro, “Preference Learning in Recommender Systems”, 2009

[4] J. Zeng, F. Li, H. Liu, J. Wen and S. Hirokawa, "A Restaurant Recommender System Based on User Preference and Location in Mobile Environment," 2016, 5th IIAI International Congress on Advanced Applied Informatics (IIAI-AAI), Kumamoto, page: 55-60.

[5] A. Nürnberger,K. Turowski , A. Zuccala, “Towards Effective Research-Paper Recommender Systems and User Modeling based on Mind Maps”,2015

[6] S. Khatwani and M. B. Chandak, "Building Personalized and Non Personalized recommendation systems," 2016 International Conference on Automatic Control and Dynamic Optimization Techniques (ICACDOT), Pune, 2016, pp. 623-628.

[7] L. Anitha, M. K. Kavitha Devi, P.Anjali Dev, “A Review on Recommender System”, 2013, International Journal of Computer Applications (0975 – 8887) , volume 82 – No3

[8] J. T. Jacquesa, M. Perry, P. O. Kristensson, “Differentiation of online text-based advertising and the effect on users’ click behavior”, 2015, Computers in Human Behavior, volume: 50, pages 535-543

[9] R. D. Gottapua , L. Venkata, S. Monangi, “Point-Of-Interest Recommender System for Social Groups”, 2017, Complex Adaptive Systems Conference with Theme: Engineering Cyber Physical Systems, CAS Chicago, Illinois, USA,volume:114, page: 159–164

[10] Al Mamunur Rashid, Istvan Albert, Dan Cosley, Shyong K. Lam, Sean M. McNee, Joseph A. Konstan, and John Riedl, “Getting to know you: learning new user preferences in recommender systems” 2002 In Proceedings of the 7th international conference on intelligent user interfaces,ACM, New York, NY, USA, page: 127-134

[11] K. Lunkad, “Prediction of Yelp Rating using Yelp Reviews”,2015

[12] A. Wang, W. Zeng, J. Zhang, “Predicting New Restaurant Success and Rating with Yelp”,2016

[13] Yu, M., Xue, M., & Ouyang, W. (2015). Restaurants Review Star Prediction for Yelp Dataset.

[14] L. Hu, Y. Liu, A. Sun, “Your Neighbors Affect Your Ratings: On Geographical Neighborhood Influence to Rating Prediction, 2014 Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval . pages 345-354

[15] “Yelp Dataset”.Yelp.com,2017. [Online]. Available: https://www.yelp.com/dataset

[16] M. L. McHugh, “The Chi-square test of independence”, 2013, Biochemia Medica, 23(2),page:143–149.

[17] Scikit-learn: Machine Learning in Python, Pedregosa et al., JMLR 12, pp. 2825-2830, 2011.

[18] A. Jain, P. Kasat, A. S. S. Reddy, “Box-Office Opening Prediction of Movies Based on Hype Analysis through Data Mining,” International Journal of Computer Applications, vol. 56, October 2012.

[19] R. Burke, M. Ramezani, Recommender Systems Handbook, Springer, 2010, in print, Ch. Matching Recommendation Technologies and Domains

[20] B. Xiao, I. Benbasat, E-commerce product recommendation agents: Use, characteristics, and impact, MIS Quarterly 31 (1) (2007) 137–209.

[21] F.O. Isinkaye, Y.O. Folajimi, B.A. Ojokoh, Recommendation systems: Principles, methods and evaluation, Egyptian Informatics Journal, Volume 16, Issue 3,2015, Pages 261-273, ISSN 1110-8665, https://doi.org/10.1016/j.eij.2015.06.005