MEF UNIVERSITY

PREDICTING YELP STARS BASED ON BUSINESS ATTRIBUTES

Capstone Project

Ahmet Tek

İSTANBUL, 2018

MEF UNIVERSITY

PREDICTING YELP STARS BASED ON BUSINESS ATTRIBUTES

Capstone Project

Ahmet Tek

Advisor: Dr. Ebru Arısoy Saraçlar

İSTANBUL, 2018

MEF UNIVERSITY

Name of the project: Predicting Yelp Stars Based On Business Attributes
Name/Last Name of the Student: Ahmet Tek
Date of Thesis Defense: 03/09/2018

I hereby state that the graduation project prepared by Ahmet Tek has been completed under my supervision. I accept this work as a "Graduation Project".

03/09/2018
Dr. Ebru Arısoy Saraçlar

I hereby state that I have examined this graduation project by Ahmet Tek which is accepted by his supervisor. This work is acceptable as a graduation project and the student is eligible to take the graduation project examination.

03/09/2018
Prof. Dr. Özgür Özlük

Director
of
Big Data Analytics Program

We hereby state that we have held the graduation examination of Ahmet Tek and agree that the student has satisfied all requirements.

THE EXAMINATION COMMITTEE

Committee Member	Signature
1. Dr. Ebru Arısoy Saraçlar	……..
2. Prof. Dr. Özgür Özlük	……..

Academic Honesty Pledge

I promise not to collaborate with anyone, not to seek or accept any outside help, and not to give any help to others.

I understand that all resources in print or on the web must be explicitly cited.

In keeping with MEF University's ideals, I pledge that this work is my own and that I have neither given nor received inappropriate assistance in preparing it.

Ahmet Tek 03.09.2018 Signature

EXECUTIVE SUMMARY

PREDICTING YELP STARS BASED ON BUSINESS ATTRIBUTES

Ahmet Tek

Dr. Ebru Arısoy Saraçlar

SEPTEMBER, 2018, 24 Pages

Yelp is a business review website where consumers can comment on a business from their point of view. This allows other consumers to have prior knowledge of the business. Whenever we search something we try and hope to get the most relevant results, and recommender systems can achieve this. Review websites, such as Yelp and TripAdvisor allow users to post online reviews for various businesses, products and services and have been recently shown to have a significant influence on consumer shopping behavior [1]. This paper aims to predict restaurant ratings using their attributes such as alcohol, noise level, Wifi, music, a smoking area and to find the most important attributes for higher ratings.

Yelp dataset has lots of information about businesses and consumer behaviors and it is free for academic usage. For these reasons, Yelp dataset has been selected in this project.

Machine Learning models have been executed for two-star label classification. Since we aim to find the most important features for a higher rating we only choose 4 and 5-star labels from the dataset. In our research, restaurant rating prediction is implemented as binary-class classification where the class labels are the star ratings. Restaurant attributes are the input features of the classifier. We will investigate Decision Trees, Naive Bayes Classifier, Two-Class Decision Forest, Two-Class Boosted Decision Trees, Two-Class Neural Network, Two-Class Support Vector Machine, Two-Class Logistic Regression and choose the most important 10 attributes resulting in high ratings.

Key Words: Content-Based Filtering, Yelp, Restaurant Reviews

ÖZET

PREDICTING YELP STARS BASED ON BUSINESS ATTRIBUTES

Ahmet Tek

Dr. Ebru Arısoy Saraçlar

EYLÜL, 2018, 24 sayfa

Yelp bir işletme inceleme ve yorumlama sitesidir. Tüketici, bir işletmeyi kendi bakış açısıyla yorumlayabilir. Bu durum diğer tüketicilerin işletmeler hakkında önceden bilgi sahibi olmasını sağlar. İnternetten bir şey aradığımız zaman, aradığımız nesneyle ilgili en yakın sonuçları elde etmeyi umarız ve günümüzde tavsiye sistemleri bunu başarabilir. Yelp, TripAdvisor gibi web siteleri, kullanıcılara işletmeler, ürünler ve kullandıkları servisler hakkında çevrimiçi yorum yapmaya izin vermektedir. Araştırmalar, bu durumun tüm tüketicilerin alışveriş yapma davranışı üzerinde önemli bir etkiye sahip olduğunu ortaya koymaktadır [1]. Bu makale, alkol, gürültü seviyesi, Wifi, müzik, sigara içme alanı gibi işletmenin özelliklerini kullanarak restorana verilen puanları öngörmeyi ve daha yüksek puan için en önemli özellikleri bulmayı amaçlamaktadır.

Yelp, bizlere büyük bir veri seti sağlamaktadır. Ayrıca işletmeler ve müşteriler hakkında bir çok bilgiyi içermekte ve akademik alan için ücretsiz kullanım sağlamaktadır. Bu nedenlerden dolayı, bu projede Yelp veri kümesi seçilmiştir.

Projedeki makine öğrenmesi modelleri iki-yıldız etiketli sınıflandırma için çalıştırılmıştır. Yüksek yıldız etiketleri için en önemli restoran özelliklerini bulmaya amaçladığımızdan, datasetinde yalnızca 4 ve 5 yıldızlı etiketleri seçmeyi tercih ediyoruz. Araştırmamızda, restoran derecelendirme tahminlemesi, yıldızların etiket olarak kullanıldığı iki sınıflı bir sınıflandırma problem olarak ele alınmıştır. Restoran özellikleri, sınıflandırma modelleri için girdi olarak kullanılmıştır. Makalede Karar Ağaçları, Naïve Bayes Sınıflandırma, İki Sınıflı Decision Forest, İki Sınıflı Boosted Decision Trees, İki Sınıflı Neural Network, İki Sınıflı Support Vector Machine ve İki Sınıflı Logistic Regression algoritmaları üzerinde çalışılacaktır ve yüksek yıldız derecesi sağlayan en önemli 10 özellik seçilecektir.

Anahtar Kelimeler: İçerik Tabanlı Filtreleme, Yelp, Restoran Değerlendirme

TABLE OF CONTENTS

1. INTRODUCTION

Yelp, founded in 2004, is a multinational corporation that publishes crowd-sourced online reviews on local businesses. A Yelp user can share his or her experience with a business by posting a review about the business and, also by rating the business with stars ranging from 1 to 5. As of 2018, Yelp.com had 155 million reviews and 145 million monthly visitors. Yelp dataset has information about users, business, and reviews. Studies show that review websites have a significant impact on consumer purchase decisions as well as on product sales and business revenues [1].

In today's fast-changing world, understanding user expectations and preferences is a very important issue for businesses' sustainability and profitability. People are looking at ratings of businesses via Yelp.com before going somewhere and making a choice between high-rating businesses. They prefer high-rating businesses to ensure they can get the best service. Restaurant star rating is the most important attribute to decide to go to a business or not. For this reason, Yelp.com website puts the star rating button beneath the business's name. If it has a high rating, users can click on business to get more information about. In fact, M. Anderson and J. Magruder (2011) have shown that star ratings are so central to the Yelp experience that an extra half star allows restaurants to sell out 19% more frequently [2].

New opening restaurants fail at a surprisingly high rate: %59 within the first year of opening and about %80 within the first five years [7]. Thus, investigating the relationship between a restaurant's success and the attributes provided by the restaurant is important.

This study aims to create models that predict business ratings based on various business attributes. Since the provided dataset is quite large for the available computational infrastructure, we will only focus on restaurant business which keeps the biggest amount in the dataset.

2. LITERATURE REVIEW

Whenever we search something on the internet, we try to get the most relevant results, and this can be achieved using recommender systems. Recommender systems help users find their interests among many preferences. Recommendation systems are valuable to understand people's preferences. There are two kinds of recommendation systems: Content-Based Filtering (CBF) and Collaborative Filtering (CF).

2.1. Content-Based Filtering

Content-Based Filtering (CBF) is based on the profile of the user's preference and the item's description. CBF algorithms recommend items that are similar to those items that were liked in the past. Figure 1 shows that if 'Product A' is similar to 'Product B' and 'User' buy 'Product A' then CBF algorithms recommend 'Product B' to that user.

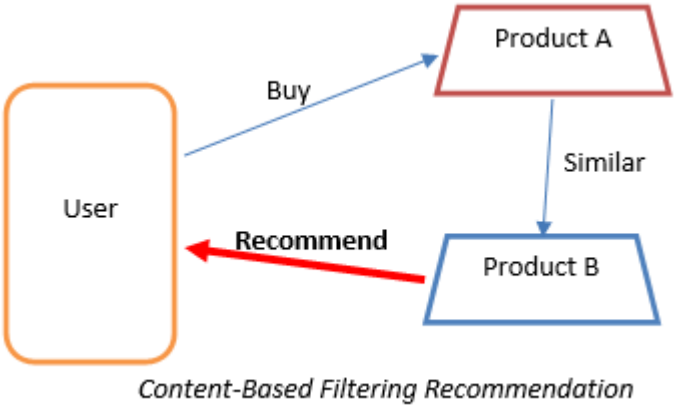

Figure 1: Content-Based Filtering Recommendation

Suppose that, there are four categories of movies such as Action, Adventure, Comedy and Horror and there is a user UA who has watched movies related to Action and Comedy. The content-based recommendation engine will only recommend the movies related to these categories. However, it never recommends other categories to the user which is a deficiency of the CBF approach. This problem can be solved with Collaborative Filtering recommendation system.

2.2. Collaborative Filtering

Collaborative Filtering (CF) is based on finding similar users with similar preferences in a community. If two users have the same or almost the same rated items in

common, then they may have a similar taste. The CF algorithm groups the users together and recommends an item that a user hasn't rated before but was rated by other users in his groups. Figure 2 illustrates the CF algorithm. 'User A' and 'User B' can be clustered in the same group when the users both buy 'Product' and Product B'. If 'User B' buy 'Product C' then the CF algorithm recommends the same product to the 'User A'.

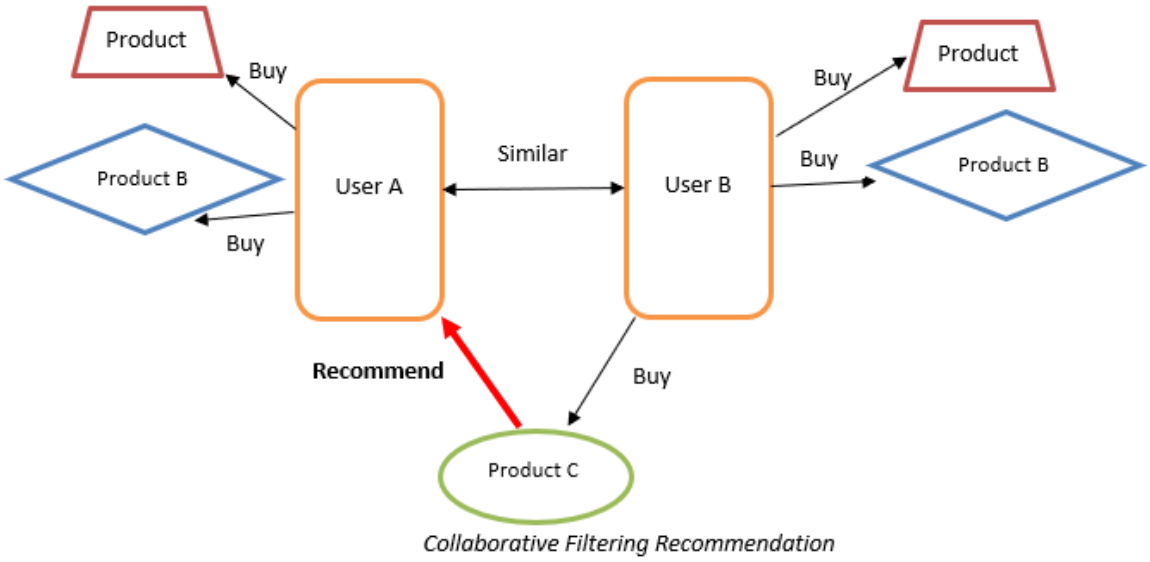

Figure 2: Collaborative Filtering Recommendation

2.3. Literature Review

This study will use the Content-Based Filtering recommendation system. There are some studies on Yelp business attributes. Mathieu, Grillet, Passerini, Tiwari (2016) [5] searched yelp dataset to define business types and business location for business success. Using yelp dataset, business's success score was calculated and utilized to find a relation between location and success score using OpenStreetMap tool. The paper's goal was to predict the success of a location for business features. Multiple Linear Regression, Support Vector Regression and Random Forest Regression algorithms were applied for modelling. Random Forest Regression gave the highest accuracy with the lowest variation at %95 confidence level among all the three machine learning algorithms. The main difference between Mathieu, Grillet, Passerini, Tiwari (2016) and this paper is the use of the map for the business location. We predict the business success based on business features out of the location.

Badrul Sarwar, George Karypis, Joseph Konstan, and John Riedl [9] searched item-based recommender systems in their paper. Recommender systems can be used for personalized recommendations for any products. New recommender system technologies can respond quickly to the user with a high-quality recommendation. This paper investigated different recommendation techniques for computing item-based similarities. The first step was to find the similarity between items and to select the most similar items. The paper presented three similarity methods: Cosine-based Similarity, Correlation-based Similarity, Adjusted Cosine Similarity. The paper observed that the item-based similarity methods provide better prediction results than the user-based similarities methods.

Farhan (2014) searched a linear regression prediction model that gave important restaurant features and predicted business reviews to improve customer satisfaction. Farhan (2014) used Naive Bayes, neural networks, random forest and linear regression algorithms on prediction modelling. The linear regression model had the highest accuracy. The author used the time dimension for annual trends and people's repetitive behavior throughout the years. We differed from this article by not using the time dimension and 15 years of data. Farhan (2014) predicted the star ratings for a given review based on the time and restaurant attributes, but we predict the star ratings based on only restaurant attributes [6].

Aileen Wang, William Zeng, Jessica Zhang (2016) aimed to predict the important features from text reviews and sentiment analysis for restaurant success. They tried to find the optimal subset of features to get the highest accuracy score for restaurant success. They calculated a chi-squared test on the samples to get the best set of features. The paper built Logistic Regression, Decision Tree, Random Forest, MLP Classifier, SVM, K-Means Clustering, AdaBoost Classifier and Naive Bayes on prediction modelling. MLP and AdaBoost Classifier had the most accuracy on predicting star model. The highest chi-squared features are airport, amazing, bad, buffet, buffets, chinese, delicious, denny, falafel, gyro, horrible, hummus, ihop, manager, minutes, pita, terrible, waitress, wings, worst. The sentiment analysis algorithms improved accuracy by 42% compared to binary classification accuracy. Sentiment features from customer text reviews yielded more accurate information about the restaurant success [8].

3. EXPLORATORY DATA ANALYSIS AND METHODOLOGY

3.1. Methodology

The paper analyzes MySQL and JSON file types that Yelp provides. We install a database on MySQL and name it as 'Yelp_db'. We import 11 tables to MySQL Yelp_db and check the data. But it is very difficult to process the MySQL tables on python. We begin with coding in R to get the data in JSON files. After we get the data in R, we convert the JSON file to RDS file format which provides us with a memory efficient format. After cleaning the data in R, we save the final file format as CSV to work easily on this dataset with Python and R.

3.2. Dataset

We use the dataset provided by Yelp as part of their Dataset Challenge Round 11 (Dataset, 2018) for the start rating prediction models. This dataset includes information about local businesses in 11 metropolitan areas across 4 countries and contains information about 174K businesses, 1.3M business attributes, 5.3M user reviews, 1.1M tips, 1293 distinct business category (e.g. restaurant, food, dentist, hair salons etc.), 16.7M check-in and 200K pictures.

Concretely, the dataset consists of six JSON files: business.json, checkin.json, review.json, tip.json, user.json, photo.json. Business.json file has all the information about businesses, business attributes and business categories, therefore we will not use the rest of the JSON files.

We can see all the information that we use for ML modelling in Table 1, Table 2, Table 3. Business Table has business id, business name, address, city, state, review count, star rating and is_open flag information. We can filter only open businesses, find total review count for the star rating.

id	name	neighborhood	address	city	state	postal_code	latitude	longitude	stars	review_count	is_open
--6Mefn...	John's Chinese BBQ Restaurant		328 Highway 7 E. Chalmers Gate 11. Unit 10	Richmond Hill	ON	L4B 3P7	43.8409	-79.3996	3	37	1
--7zmmk...	Primal Brewery		16432 Old Statesville Rd	Huntersville	NC	28078	35.4371	-80.8437	4	47	1
--8LPVS...	Valley Bone and Joint Specialists		3941 E Baseline Rd. Ste 102	Gilbert	AZ	85234	33.3795	-111.728	4.5	3	1
--9e1ON...	Delmonico Steakhouse	The Strip	3355 Las Vegas Blvd S	Las Vegas	NV	89109	36.1232	-115.169	4	1451	1
--9OOL...	Great Clips		1835 E Guadalupe Rd. Ste 106	Tempe	AZ	85283	33.3617	-111.91	3.5	11	1
--ab39Ii...	Famous Footwear		1800 E Rio Salado Pkv 110. Tempe Marketplace	Tempe	AZ	85281	33.4301	-111.905	4	10	1
--coVkb...	Eazor's Auto Salon		616 Long Rd	Pittsburgh	PA	15235	40.4531	-79.8389	5	12	1
--ciBEbX...	Howl at the Moon	Downtown	125 7th St	Pittsburgh	PA	15222	40.4439	-80.0002	3	51	1
--cZ6Hh...	Pio Pio	Dilworth	1408 E Blvd	Charlotte	NC	28203	35.1999	-80.8448	4	317	1
--DaPTJ...	Sunnyside Grill	Corso Italia	1218 Saint Clair Avenue W	Toronto	ON	M6E	43.6778	-79.4447	3.5	39	1
--Ddme...	World Food Championships	The Strip	3645 Las Vegas Blvd S	Las Vegas	NV	89109	36.1143	-115.171	3	5	1

Table 1: Business Dataset

Business Category Table has business id and category information. The Category column has information as Dentists, Health & Medical, Hair Salons, Shopping, Restaurants, Burgers, Italian, Automotive, Auto Repair, Sports Clubs. We can filter the business categories that we want through this column.

id	business_id	category
57	l09JfMeO6vnYs5MCJtrcmO	French
58	l09JfMeO6vnYs5MCJtrcmO	Restaurants
59	IOSlT5iGE6CCDhSG0zG3xg	Beauty & Spas
60	IOSlT5iGE6CCDhSG0zG3xg	Nail Salons
61	b2I2DXtZVnpUMCXp1JON7A	Tires
62	b2I2DXtZVnpUMCXp1JON7A	Oil Change Stations
63	b2I2DXtZVnpUMCXp1JON7A	Auto Repair
64	b2I2DXtZVnpUMCXp1JON7A	Automotive
65	0FMKDOU8TJT1x87OKYGDTg	Barbers
66	0FMKDOU8TJT1x87OKYGDTg	Beauty & Spas
67	Gu-xs3NIOTi3Mi2xYoN2aw	French
68	Gu-xs3NIOTi3Mi2xYoN2aw	Food
69	Gu-xs3NIOTi3Mi2xYoN2aw	Bakeries
70	Gu-xs3NIOTi3Mi2xYoN2aw	Restaurants
71	lHYiCS-v8AFiUitv6MGpxg	Food
72	lHYiCS-v8AFiUitv6MGpxg	Coffee & Tea
73	94KziT6DO9XlBET3WzIv_w	Shopping

Table 2: Business Category

Business Attributes Table has business id, attribute name and attribute values information. Attribute name column has information as Bike_Parking, Business_Accepts_Credit_Cards, Alcohol, Has_TV, Noise_Level, Music, WiFi, Good_For_Kids. Attributes can have TRUE, FALSE binary values.

id	business_id	name	value
1	FYWN1wneV18bWNgQjJ2GNg	AcceptsInsurance	1
2	FYWN1wneV18bWNgQjJ2GNg	ByAppointmentOnly	1
3	FYWN1wneV18bWNgQjJ2GNg	BusinessAcceptsCreditCards	1
4	He-G7vWjzVUysIKrfNbPUQ	BusinessParking	{"garage": false, "street": false, "validated": false, "lot": true, "valet": false}
5	He-G7vWjzVUysIKrfNbPUQ	HairSpecializesIn	{"coloring": true, "africanamerican": false, "curly": true, "perms": true, "kids": true, "extensions": true...
6	He-G7vWjzVUysIKrfNbPUQ	BusinessAcceptsCreditCards	1
7	He-G7vWjzVUysIKrfNbPUQ	RestaurantsPriceRange2	3
8	He-G7vWjzVUysIKrfNbPUQ	GoodForKids	1
9	He-G7vWjzVUysIKrfNbPUQ	ByAppointmentOnly	0
10	He-G7vWjzVUysIKrfNbPUQ	WheelchairAccessible	1
11	8DShNS-LuFqpEWIp0HxijA	BusinessAcceptsCreditCards	1
12	8DShNS-LuFqpEWIp0HxijA	RestaurantsPriceRange2	2
13	8DShNS-LuFqpEWIp0HxijA	BusinessParking	{"garage": false, "street": false, "validated": false, "lot": true, "valet": false}
14	8DShNS-LuFqpEWIp0HxijA	BikeParking	1
15	PfOCPiBrlOAnz_NXi9h_w	Alcohol	full_bar
16	PfOCPiBrlOAnz_NXi9h_w	HasTV	1
17	PfOCPiBrlOAnz_NXi9h_w	NoiseLevel	average

Table 3: Business Attributes Dataset

The businesses describe in the Yelp dataset belong to different categories, such as restaurants, food, dentist and hair salons, etc. The text reviews or business attributes for different business categories may be very different. For example, a typical coiffure review may contain the words ‘haircut’, ‘balayage’ and ‘hair color’, but these words would not

occur in a restaurant review. That is why the model training and testing for each category should be separate. Figure 3 shows the distribution of business categories in the dataset. Restaurants make up almost 31% of the 174K businesses. Therefore, in this paper, we restrict to a business recommendation system for restaurants only.

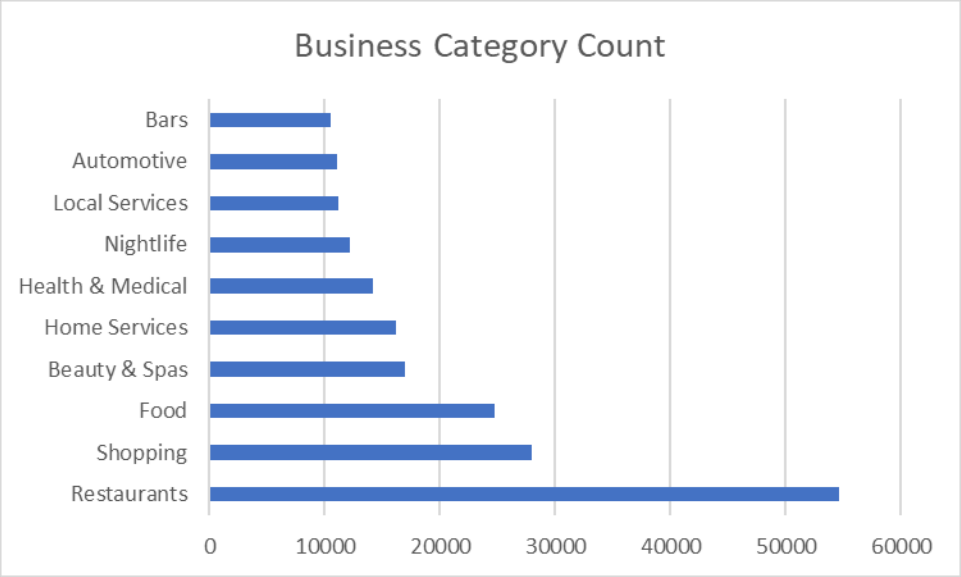

Figure 3: Top 10 Business Count

3.3. Data Cleaning

We start data cleaning by choosing only the restaurant part from the Yelp data. Business.json file has 174567 rows and 15 attributes. Since it is in JSON format, "categories" and "attributes" properties are written in a single column side by side with comma delimited. JSON file format is as shown below Figure 4:


```
{
"business_id": "-ooEO2YqDQVYNHnSF2BPfw",
"name": "Restaurant Lucca",
"neighborhood": "Rosemont-La Petite-Patrie",
"address": "12 Rue Dante",
"city": "Montr\u00e9al",
"state": "QC",
"postal_code":
"H2S 1J5",
"latitude": 45.5328965,
"longitude": -73.6136037,
"stars": 4.5,
"review_count": 11,
"is_open": 1,
"attributes": {"GoodForMeal": {"dessert": false, "latenight": false, "lunch": false, "dinner": false, "breakfast": false, "brunch": false},
"RestaurantsTableService": true, "Caters": false, "HasTV": false, "RestaurantsGoodForGroups": true, "WiFi": "no",
"RestaurantsAttire": "casual", "RestaurantsReservations": true, "OutdoorSeating": false, "BusinessAcceptsCreditCards": true,
"RestaurantsPriceRange2": 3, "BikeParking": true, "RestaurantsDelivery": false,
"Ambience": {"romantic": false, "intimate": false, "classy": false, "hipster": false, "touristy": false, "trendy": false,
"upscale": false, "casual": false},
"RestaurantsTakeOut": true, "GoodForKids": false},
"categories": ["Italian", "Restaurants"],
"hours": {}
}
```


Figure 4: JSON File Format

The recommendation system develops in this project is restricted to restaurants recommendation. So, 54618 out of 174567 dataset rows are set apart as the restaurant data. We also get the only "is_open=1" rows, since we don't care about the close restaurants. We get the 40394 rows for modelling the data.

All Dataset consists of categorical values and we flatten all rows into columns to use them in ML models. As the categories are separated by a comma, firstly, "categories" column is divided into 10 different new columns. For 65 major restaurant categories (such as Fast Food, Pizza, Mexican, Nightlife, Bars etc.), we create new dummy columns to flatten the data rows to columns. See Appendix B for the whole list. If a category includes Italian and Restaurants values, we flag these 2 column values as 1 and other 63 dummy column values as 0. We trim the blank values in the dataset. There are so many business attributes as categorical variables in the dataset. We also flatten them into new columns and if an attribute exists in the row, we flag its value as 1 (TRUE), otherwise as 0 (FALSE). When we analyze the dataset column by column to get N/A (Not Applicable) count, we get 59 columns which have more than %50 N/A values and we decide to drop these 59 columns from the dataset. We flag rest of N/A as FALSE otherwise the model that we run give us very low accuracy. We take some columns out of the final dataset that we don't use them in the model: name, latitude, longitude, neighborhood, address, postalcode and isopen. In the JSON file, all attribute values are normally stored as TRUE-FALSE, but some attributes - restaurantspricerange2, alcohol, noiselevel, restaurantsattire, wifi, smoking, byobcorkage, agesallowed - have also other values rather than TRUE-FALSE. We flatten these values to new columns as new attributes. For example, WIFI attribute has FREE, PAID, NO, N/A values in the dataset. We create new attribute columns named WIFI_FREE, WIFI_PAID, WIFI_NO. In this way, all dataset has 1-0 values. The business star label has values of 2.5, 3.5, 4.5 and we round all them upper bound. In the final dataset, we have 40394 rows and 169 attributes and all of them ready to use. After all, the final dataset is ready to further ML models with 40394 rows and 116 attributes (See Appendix A).

3.4. Exploratory Data Analysis

Final dataset has 40394 rows and 116 attributes. We analyze the dataset based on star attribute as target label.

We see that the consumers give higher star ratings to the restaurants that serve alcohol than the restaurants that don't. Also, the consumers prefer the restaurants that have parking valet. We can see that reviewers give higher star ratings for restaurants that serve dinner and have a quiet noise level. We can see all the attributes and average star ratings in Appendix E.

Figure 5 shows the restaurant star counts. 4-star ratings are dominating all dataset. People tend to give 3 or 4 stars rating rather than 1 or 5.

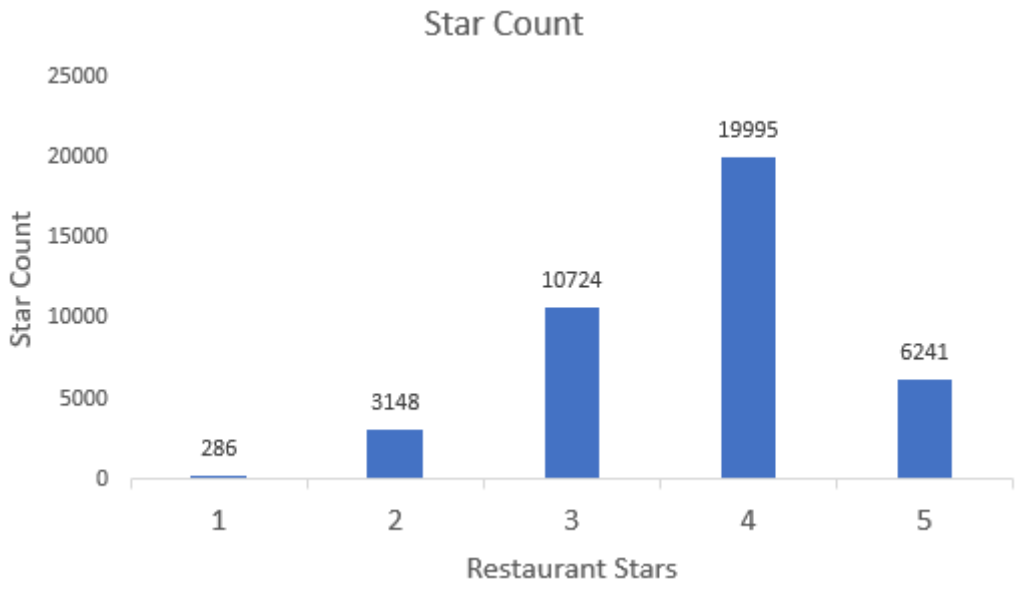

Figure 5: Restaurant Star Count

Figure 6 shows us review count per star. Most of the reviews give 4 stars to the businesses. As expected, this figure has the same bar chart as ‘Figure 5: Restaurant star count”.

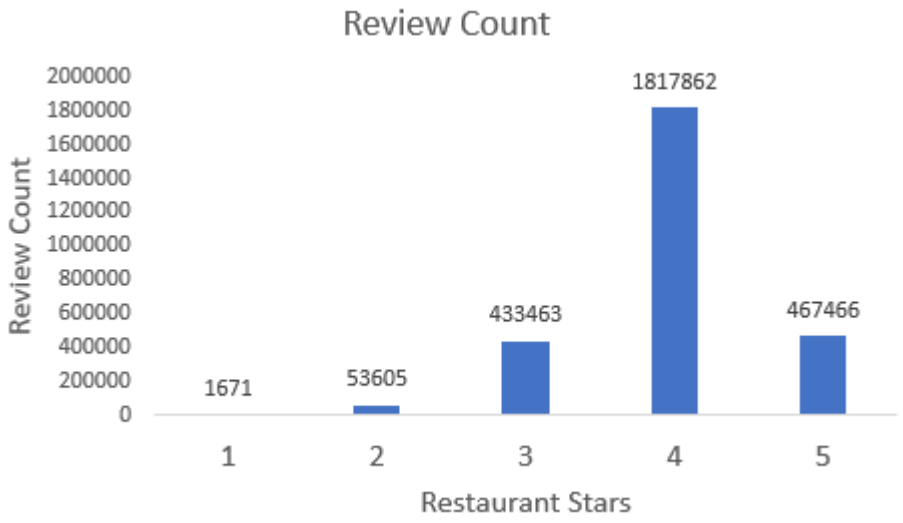

Figure 6: Business Reviews Count

4. MODEL SELECTION

The final dataset has 116 attributes which include 65 major restaurant categories that shown in Appendix B. We have decided to remove these major categories from the dataset and the final dataset has 51 attributes. All attributes have True-False binary values but we apply Min-Max Scaler function to the city, state and review count that have their own values and it's not applicable to flatten all values into new dummy columns. We use supervised algorithms. These classification algorithms is applied as machine learning model: Decision Trees, Naive Bayes Classifier, Two-Class Decision Forest, Two-Class Boosted Decision Trees, Two-Class Neural Network, Two-Class Support Vector Machine, Two-Class Logistic Regression.

4.1. Evaluation Metrics

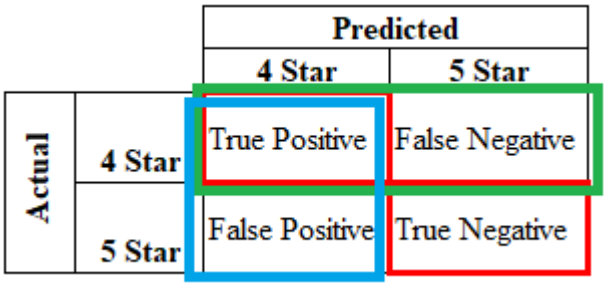

Figure 7: Evaluation Metrics

Accuracy: The accuracy metric can be defined as the percentage of correctly classified instances. The accuracy metric is shown in red color in Figure 7.

Accuracy *= (True Positive + True Negative) / (True Positive + False Positive + False Negative + True Negative)*

We will use the accuracy metric to find out how many of 4 and 5 star labels are correctly classified.

Recall: The recall metric calculates how many of actual positive stars are calculated as the positive star. The recall metric is shown in green color in Figure 7.

***Recall** = True Positive / (True Positive + False Negative)*

We will use the recall metric to find out how many of 4 star label are correctly classified.

Precision: The precision metric calculates how many of predicted positive stars are the actual positive star. The precision metric is shown in blue color in Figure 7.

***Precision** = True Positive / (True Positive + False Positive)*

We will use the precision metric to find out how many of predicted 4-star label are actual 4-star.

F1 Score: If we need to find a balance Recall and Precision scores, we can use F1 score.

***F1** = 2* (Precision * Recall) / (Precision + Recall)*

We need a high F1 score to get a better model result. Both Recall and Precision scores need to be high value or need to be a balance between these two scores.

4.2. Decision Tree Classifier

Decision Trees are basic classification algorithms that create a model to predict the label class by learning simple decision rules from the data features. We create some models with different parameters. Maximum depth of the tree parameter sets as default, 1, 8, 10, 12, 15 and 20. Decision Tree model with default parameters has 0.67 accuracy, 0.56 precision, 0.56 recall and 0.56 F1 scores for the test dataset. Decision Tree model with a maximum depth of parameter for 8 has 0.75 accuracy, 0.61 precision, 0.53 recall and 0.51 F1 scores for the test dataset. Other parameters have lower evaluation scores, so we will calculate the feature importance for maximum depth of parameters for 8.

10 most important attributes for restaurant businesses are as follows: state, review count, business accepts credit cards, good for kids, restaurants takeout, restaurants table service, business parking street, ambience trendy, alcohol full bar, noise level quiet.

4.3. Naïve Bayes (GaussianNB and BernoulliNB)

The Naive Bayesian classifier is based on Bayes' theorem with the independence assumptions between predictors. A Naive Bayesian model is easy to build and useful for very large datasets. There are 3 types of Naive Bayes which are Gaussian, Multinominal and Bernoulli. Gaussian Naive Bayes assumes that features have a normal distribution. Bernoulli is useful for binary (zeros and ones) features. We create Gaussian and Bernoulli Naïve Bayes models. Gaussian Naïve Bayes model has 0.59 accuracy, 0.58 precision, 0.61 recall and 0.56 F1 scores for the test dataset. Bernoulli Naïve Bayes model has 0.68 accuracy, 0.58 precision, 0.59 recall and 0.58 F1 scores for the test dataset. Bernoulli model has better evaluation result than Gaussian.

10 most important attributes for restaurant businesses are as follows: state, review count, business accepts credit cards, good for kids, bike parking, restaurants good for groups, restaurants takeout, restaurants price range 22, noise level average, restaurants attire casual.

4.4. Two-Class Decision Forest

Decision forests are fast, supervised ensemble models. Decision forest is a good choice if the dataset has only two class labels. The parameters that we used for Decision Forest model are resampling method (bagging), maximum depth of the decision trees (100), number of decision trees (100). Two-Class Decision Forest model has 0.75 accuracy, 0.48 precision, 0.15 recall and 0.22 F1 scores for the test dataset.

10 most important attributes for restaurant businesses are as follows: alcohol full bar, outdoor seating, review count, noise level average, noise level loud, ambience trendy, business accepts credit cards, ambience intimate, good for kids, state.

4.5. Two-Class Boosted Decision Trees

A boosted decision tree is an ensemble learning method. A boosted decision tree work iteratively and the second tree corrects for the errors of the first tree, the third tree corrects for the errors of the first and second trees [10]. The parameters that we used for Two-Class Boosted Decision Trees model are the maximum number of leaves per tree

(20), the minimum number of samples per leaf node (10), learning rate (0.2), number of trees constructed (100). Two-Class Boosted Decision Trees model has 0.75 accuracy, 0.45 precision, 0.27 recall and 0.33 F1 scores for the test dataset.

10 most important attributes for restaurant businesses are as follows: review count, state, restaurants takeout, alcohol none, caters, ambience trendy, business accepts credit cards, noise level average, good for meal lunch, noise level quiet.

4.6. Two-Class Neural Network

A neural network model that can be used to predict a target that has only two class labels. The Neural network consists of multiple layers. A single layer is called as Perceptron and multi-layer perceptron is called Neural Networks [11]. The parameters that we used for Two-Class Neural Network model are number of hidden nodes (100), learning rate (0.1), number of learning iterations (100), the initial learning weights diameter (0.1). Two-Class Neural Network model has 0.78 accuracy, 0.58 precision, 0.37 recall and 0.45 F1 scores for the test dataset.

10 most important attributes for restaurant businesses are as follows: outdoor seating, restaurants reservations, restaurants attire casual, alcohol none, restaurants price range 22, good for kids, state, restaurants takeout, good for meal lunch, business parking street.

4.7. Two-Class Support Vector Machine

Support vector machines (SVMs) are a set of supervised learning methods used for classification, regression and outlier detection. SVM try to maximize distance to nearest point and it's called margin [12]. The parameters that we used for Two-Class Support Vector Machine model are the number of iterations (50), lambda (0.001). Two-Class Support Vector Machine model has 0.76 accuracy, 0.50 precision, 0.002 recall and 0.003 F1 scores for test dataset.

10 most important attributes for restaurant businesses are as follows: alcohol none, review count, wifi no, restaurants delivery, good for meal dessert, ambience trendy, alcohol full bar, good for meal dinner, restaurants good for groups, restaurants attire casual.

4.8. Two-Class Logistic Regression

Logistic regression is used to predict the probability of an outcome and is especially popular for classification tasks. The Logistic regression predicts the probability of occurrence of an event using a logistic function [13]. The parameters that we used for Two-Class Logistic Regression model are L1 regularization weight (1), L2 regularization weight (1). Two-Class Logistic Regression model has 0.76 accuracy, 0.48 precision, 0.03 recall and 0.056 F1 scores for the test dataset.

10 most important attributes for restaurant businesses are as follows: alcohol none, restaurants table service, state, outdoor seating, review count, good for meal dinner, good for meal dessert, alcohol full bar, restaurants price range 21, ambience trendy.

4.9. Machine Learning Algorithm Comparison

We evaluate all 7 ML algorithms with Accuracy, Precision, Recall and F1 scores metrics. Almost every model has the same accuracy score, but it's not enough to claim that every model has a good result. Logistic Regression has 0.76 accuracy score but has 0.03 recall score which is a very low result. This result makes Logistic Regression worse than other ML models. Neural Network and Decision Trees have higher F1 and accuracy scores. This two ML algorithms have 4 same important features: state, good for kids, restaurant takeout, business parking street.

ML Algorithms / Evaluation Metric	Boosted Decision Trees	Decision Forest	Decision Trees	Logistic Regression	Naive Bayesian	Neural Network	Support Vector Machine
Accuracy	0.750	0.759	0.75	0.762	0.68	0.788	0.762
Precision	0.457	0.478	0.61	0.487	0.58	0.586	0.500
Recall	0.270	0.151	0.53	0.030	0.59	0.375	0.002
F1 Score	0.339	0.229	0.51	0.056	0.58	0.458	0.003

Table 4: ML Algorithms Comparison

4.10. Common Restaurant Attributes

We create seven machine learning model as shown in Table 5 and find the 10 most important restaurant features. Some features are common in ML algorithms that shown in Table 6. Review count and state features are important for 6 ML algorithms. When a restaurant gets a high review count on yelp.com and opens its restaurant at a specific state, the restaurant can be successful. Also, ambience trendy, alcohol none, alcohol full bar, restaurant takeout, accept credit card and good for kids features are common features for almost every ML algorithms.

ML Algorithms Name	Boosted Decision Trees	Decision Forest	Decision Trees	Logistic Regression	Naive Bayesian	Neural Network	Support Vector Machine
ML Algorithms ID	ML1	ML2	ML3	ML4	ML5	ML6	ML7

Table 5: ML Algorithms Name

Attributes	ML1	ML2	ML3	ML4	ML5	ML6	ML7	Grand Total
reviewcount	1	1	1	1	1		1	6
state	1	1	1	1	1	1		6
ambiencetrendy	1	1	1	1			1	5
alcoholnone	1			1		1	1	4
alcoholfull_bar		1	1	1			1	4
restaurantstakeout	1		1		1	1		4
businessacceptscreditcards	1	1	1		1			4
goodforkids		1	1		1	1		4
restaurantsattirecasual					1	1	1	3
outdoorseating		1		1		1		3
noiselevelaverage	1	1			1			3
goodformealdinner				1			1	2
restaurantstableservice			1	1				2

restaurantspricerange22					1	1		2
noiselevelquiet	1		1					2
goodformealdessert				1			1	2
businessparkingstreet			1			1		2
goodformeallunch	1					1		2
restaurantsgoodforgroups					1		1	2
restaurantspricerange21				1				1
restaurantsdelivery							1	1
bikeparking					1			1
ambienceintimate		1						1
wifino							1	1
restaurantsreservations						1		1
caters	1							1
noiselevelloud		1						1
Grand Total	**10**	**10**	**10**	**10**	**10**	**10**	**10**	**70**

Table 6: Restaurant Attributes and ML Relationship

5. CONCLUSION

We investigated seven Machine Learning Classification Model for yelp dataset. Evaluation metrics show that Decision Trees and Neural Networks models have higher scores than others. For a restaurant to be successful, it should have attributes as follows: state, review count, business accepts credit cards, good for kids, restaurants takeout, restaurants table service, business parking street, ambience trendy, alcohol full bar, noise level quiet, outdoor seating, restaurants reservations, restaurants attire casual, alcohol none, restaurants average price range, good for meal lunch.

Aileen Wang et al. (2016) [8] predict restaurant stars based on business attributes. The paper performs binary-classification, multi-classification and sentiment analysis on restaurant reviews. The paper works on Yelp Challenge 2017 dataset, so it has slightly different restaurant attributes than our Yelp Challenge 2018 dataset. The paper performs best results with Random Forest and Multilayer Neural Network algorithms. Binary-classification accuracy score is %60 and multi-classification accuracy score is %56 and sentiment analysis accuracy score based on business attributes is %85. The result shows that sentiment analysis has more accurate information about the restaurant success. We get better binary-classification accuracy score result than Aileen Wang et al. (2016) [8] result. Also, Neural Network algorithm is the best algorithm for these two paper.

New opened restaurants fail through the first year with a %59 probability [7]. This paper helps restaurants to success with the most restaurant important features. In order to get 5-star and have successful, the restaurant needs to have the features mentioned in part “4.10. Common Restaurant Attributes”.

APPENDIX A: All Attributes

ALL_ATTRIBUTES_1	ALL_ATTRIBUTES_2	ALL_ATTRIBUTES_3
city	sushibars	vegetarian
state	delis	salad
stars	steakhouses	hotdogs
reviewcount	seafood	middleeastern
restaurants	chickenwings	eventplanningservices
fastfood	sportsbars	specialtyfood
pizza	coffeetea	lounges
mexican	mediterranean	korean
americantraditional	barbeque	canadiannew
nightlife	thai	artsentertainment
sandwiches	asianfusion	winebars
bars	french	glutenfree
food	buffets	latinamerican
italian	indian	british
chinese	pubs	gastropubs
americannew	greek	icecreamfrozenyogurt
burgers	diners	southern
breakfastbrunch	bakeries	vegan
cafes	vietnamese	desserts
japanese	texmex	hawaiian
german	businessparkinggarage	restaurantspricerange22
bagels	businessparkingstreet	restaurantspricerange23
caterers	businessparkingvalidated	restaurantspricerange24
juicebarssmoothies	businessparkinglot	alcoholbeer_and_wine
fishchips	businessparkingvalet	alcoholfull_bar
ethnicfood	ambienceromantic	alcoholnone
tapasbars	ambienceintimate	noiselevelaverage
soup	ambienceclassy	noiselevelloud
halal	ambiencehipster	noiselevelquiet
businessacceptscreditcards	ambiencedivey	noiselevelvery_loud
goodforkids	ambiencetouristy	restaurantsattirecasual
bikeparking	ambiencetrendy	restaurantsattiredressy
hastv	ambienceupscale	restaurantsattireformal
restaurantsgoodforgroups	ambiencecasual	wififree
caters	goodformealdessert	wifino
restaurantsreservations	goodformeallatenight	wifipaid
restaurantstakeout	goodformeallunch	
restaurantstableservice	goodformealdinner	
outdoorseating	goodformealbreakfast	

restaurantsdelivery	restaurantspricerange21	

APPENDIX B: Major 65 Restaurant Categories

Major Restaurant Categories_1	Major Restaurant Categories_2	Major Restaurant Categories_3
restaurants	juicebarssmoothies	desserts
pizza	tapasbars	bagels
nightlife	fastfood	fishchips
food	mexican	soup
americannew	sandwiches	americantraditional
cafes	italian	bars
delis	burgers	chinese
chickenwings	japanese	breakfastbrunch
mediterranean	steakhouses	sushibars
asianfusion	sportsbars	seafood
indian	barbeque	coffeetea
diners	french	thai
texmex	pubs	buffets
hotdogs	bakeries	greek
specialtyfood	vegetarian	vietnamese
canadiannew	middleeastern	salad
glutenfree	lounges	eventplanningservices
gastropubs	artsentertainment	korean
vegan	latinamerican	winebars
german	icecreamfrozenyogurt	british
hawaiian	caterers	southern
ethnicfood	halal	

APPENDIX C: Attributes that have more than %50 N/A Values

N/A Attributes_1	N/A Attributes_2
acceptsinsurance	businessacceptsbitcoin
goodfordancing	corkage
dogsallowed	hairspecializesincurly

byob	hairspecializesinextensions
hairspecializesinafricanamerican	musicdj
hairspecializesinkids	musickaraoke
hairspecializesinstraightperms	musicjukebox
musicnomusic	bestnightsfriday
musicvideo	bestnightssunday
bestnightstuesday	dietaryrestrictionsglutenfree
bestnightsthursday	dietaryrestrictionshalal
dietaryrestrictionsdairyfree	smokingoutdoor
dietaryrestrictionskosher	byobcorkageyes_corkage
dietaryrestrictionsvegetarian	agesallowed19plus
smokingno	happyhour
byobcorkageno	drivethru
agesallowed18plus	open24hours
agesallowedallages	restaurantscounterservice
byappointmentonly	hairspecializesincoloring
wheelchairaccessible	hairspecializesinperms
coatcheck	hairspecializesinasian
goodformealbrunch	musicbackgroundmusic
dietaryrestrictionsvegan	musiclive
dietaryrestrictionssoyfree	bestnightsmonday
smokingyes	bestnightswednesday
byobcorkageyes_free	bestnightssaturday
agesallowed21plus	

APPENDIX D: Final 51 Attributes List

Final_Attributes_1	Final_Attributes_2
restaurantsreservations	businessparkingvalidated
outdoorseating	businessparkinglot
restaurantsdelivery	ambienceintimate
businessparkingvalet	ambiencehipster
ambienceromantic	ambiencedivey
ambienceclassy	ambiencetouristy
restaurantspricerange22	ambiencetrendy
noiselevelquiet	ambiencecasual
restaurantsattiredressy	goodformealdessert
wifino	goodformealdinner

city	goodformealbreakfast
state	restaurantspricerange21
restaurantsgoodforgroups	restaurantspricerange23
caters	restaurantspricerange24
restaurantstakeout	alcoholfull_bar
businessparkinggarage	alcoholnone
ambienceupscale	noiselevelaverage
goodformeallatenight	noiselevelloud
goodformeallunch	noiselevelvery_loud
alcoholbeer_and_wine	restaurantsattirecasual
stars	restaurantsattireformal
reviewcount	wififree
businessacceptscreditcards	wifipaid
goodforkids	hastv
bikeparking	restaurantstableservice
businessparkingstreet	

APPENDIX E: Restaurant Attributes Star Average Values

Restaurant Attributes	FALSE Star Avg	TRUE Star Avg
businessacceptscreditcards	3,80	3,69
goodforkids	3,73	3,70
bikeparking	3,58	3,80
hastv	3,68	3,75
restaurantsgoodforgroups	3,66	3,72
caters	3,62	3,88
restaurantsreservations	3,63	3,87
restaurantstakeout	3,77	3,70
restaurantstableservice	3,61	3,82
outdoorseating	3,65	3,82
restaurantsdelivery	3,71	3,70
businessparkinggarage	3,70	3,80
businessparkingstreet	3,65	4,00
businessparkingvalidated	3,71	3,99
businessparkinglot	3,65	3,83
businessparkingvalet	3,70	3,93

ambienceromantic	3,70	4,16
ambienceintimate	3,70	4,25
ambienceclassy	3,70	4,12
ambiencehipster	3,70	4,21
ambiencedivey	3,70	3,85
ambiencetouristy	3,71	3,33
ambiencetrendy	3,69	4,08
ambienceupscale	3,70	4,10
ambiencecasual	3,62	3,84
goodformealdessert	3,71	3,75
goodformeallatenight	3,71	3,61
goodformeallunch	3,63	3,84
goodformealdinner	3,63	3,88
goodformealbreakfast	3,71	3,72
restaurantspricerange21	3,76	3,63
restaurantspricerange22	3,65	3,77
restaurantspricerange23	3,69	3,94
restaurantspricerange24	3,70	4,00
alcoholbeer_and_wine	3,68	3,89
alcoholfull_bar	3,67	3,80
alcoholnone	3,74	3,66
noiselevelaverage	3,65	3,77
noiselevelloud	3,72	3,51
noiselevelquiet	3,68	3,85
noiselevelvery_loud	3,72	3,22
restaurantsattirecasual	3,72	3,70
restaurantsattiredressy	3,70	4,02
restaurantsattireformal	3,71	3,67
wififree	3,67	3,78
wifino	3,69	3,74
wifipaid	3,71	3,56

REFERENCES

[1] P. Y. Chen, S. Y. Wu, and J. Yoon. The impact of online recommendations and consumer feedback on sales

[2] M. Anderson and J. Magruder. Learning from the crowd. 2011.

[3] Parul Aggarwal, Vishal Tomar, Aditya Kathuria - Comparing Content Based and Collaborative Filtering in Recommender Systems

[5] Mathieu, Grillet, Passerini, Tiwari (2016) Uncovering Business Opportunities from Yelp and Open Street Map Data

[6] Farhan (2014) - Predicting Yelp Restaurant Reviews

[7] King, Tiffany, Njite, David, Parsa, H.G., and Self, John T. (2005). Why Restaurants Fail Management, 46:304–322.

[8] Aileen Wang, William Zeng, Jessica Zhang - Predicting New Restaurant Success and Rating with Yelp

[9] Badrul Sarwar, George Karypis, Joseph Konstan, and John Riedl - Item-Based Collaborative Filtering Recommendation Algorithms

[10] Two-Class Boosted Decision Tree - https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/two-class-boosted-decision-tree

[11] Two-Class Neural Network - https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/two-class-neural-network

[12] Two-Class Support Vector Machine - https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/two-class-support-vector-machine

[13] Two-Class Logistic Regression - https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/two-class-logistic-regression

[14] Ms Azure Capstone Project – Yelp Final Project - https://gallery.cortanaintelligence.com/Experiment/Yelp-Final-Project