What Drives a High Yelp Rating?

Predicting the rating and determining what factors give a high rating on Yelp?

Quraiz Najmi, Nirva Patel, Nghi Van Pan(Tyler)

Introduction:

The Internet has brought all the things on our fingertips, from booking reservations at restaurants to buying and researching the next new product purchase. The Internet has promoted a lot of new features and evolved our lives, changed our ways of living and doing businesses. Out of all these, reviews/customer feedback have become a crucial part of everyday's life as well as doing business. It helps businesses to grow and develop their strategies as per customer's feedback. Just like that, Yelp is one of the most popular and reliable apps for choosing restaurants/shops, reserving tables, seeing pictures, or getting reviews.

Yelp is a local business directory service and review site with social networking features. It allows users (customers) to give ratings and review businesses, whereas, for users (business owners), it will enable them to add their business to Yelp and help it grow. By using Yelp, one can find nearby restaurants, bars, shops, etc.[1], to see an overview of the restaurant/shops and compare their ratings at the same time to see which one is better so users can decide on where to go. The company was founded in 2004 to help the business owners grow their business, improve their services, and have better marketing and the consumers to provide their reviews with ease and choose the best business amongst available.

Another factor to consider is social networks. Social networks play an important role in influencing customers on where to spend their money. Customers check on multiple sites to check the review before they make a purchase. Apart from social media, scholarly papers sometimes tend to influence the decision[2]. Scholarly papers have become very common nowadays. When you search into how "Yelp reviews can affect a business, you'll repeatedly come across research studies that become a sort of epochal text highlighting the start of an era when people really began to notice the impact of consumer review sites on business". The research paper written by Michael Luca, Harvard Business School Associate's Professor, found out that "a one-star increase in Yelp ratings results in a 5 to 9 percent increase in an independent restaurant's revenue"[3]. Scholarly articles and their ratings for the business affects the consumer decision. Scholarly articles can be a very good source for business to grow but at the same time it can hurt the business as well.

---

[1] Humbarger, T. (2010, April 22). *Why is yelp important to your business?* Social Media Today. Retrieved October 31, 2021, from https://www.socialmediatoday.com/content/why-yelp-important-your-business.

[2] *(PDF) Analysis of Yelp Reviews - researchgate.net*. (n.d.). Retrieved December 19, 2021, from https://www.researchgate.net/publication/263736290_Analysis_of_Yelp_Reviews

[3] Thacker, K., & Kyle ThackerKyle handles marketing and PR for Uncorkd. Aside from bartending and restaurant management. (2015, October 10). *How yelp affects restaurant reputations*. Uncorkd iPad Wine and Beverage Menus for Restaurants. Retrieved December 19, 2021, from https://www.uncorkd.biz/blog/how-yelp-affects-restaurant-reputations/

The goal of this paper is to understand how different aspects of the business affect the ratings of a business. For this purpose, the dataset was acquired from Yelp (Link to the Dataset), which has several features of a business, and the ratings that different businesses have is indicated by the "stars" column, which will be the target column for this project. More detailed information about the dataset will be provided in the approach section of this paper. There were two external dataset used for accomplishing the goal and they were acquired from Missouri Census Data Center , Income dataset (Link to the Dataset), and Wikipedia, Public Income Data (Link to the Dataset).

Motivation:

"Yelp connects people with great local businesses. With unmatched local business information, photos and review content, Yelp provides a one-stop local platform for consumers to discover, connect and transact with local businesses of all sizes by making it easy to request a quote, join the waitlist, and make a reservation, appointment or purchase"[4]. As of 2021, there are 535k paying advertising locations. The Q3 2021 Net Revenue was $269M and Net income was $18M. As of 2020 there were 31M App unique devices and 224M cumulative reviews found on yelp. Restaurant is the top reviewed business category. Next to restaurants comes Travel, Home/Local Services, Health, and Shopping. Yelp is considered as the number 1 site for public directories online. The average yelp visit is 3:10 minutes long and an average visitor clicks through 7:10 pages. Businesses with 45 star ratings experienced largest revenue growth from 2016-2019[5].

Small businesses and restaurants generally have a short life span, unless the restaurant is a chain[6]. They may not last for more than 4 years[7]. It is very common for small businesses. According to BBC news, only 74 companies of the S&P 500 companies survived for more than 40 years. Blanding applauds features in Yelp in which reviewers have public profiles. As per the google trend yelp peaked in 2011. Yelp provides a level playing field for small restaurants, which may not be able to afford paying mass advertisements[8].

---

[4] *Yelp economic average shows early signs of recovery for local economies with more than 230,000 reopened businesses in 2020*. Yelp Inc. - Yelp Economic Average Shows Early Signs of Recovery for Local Economies with More than 230,000 Reopened Businesses in 2020. (n.d.). Retrieved December 19, 2021, from https://www.yelp-ir.com/news-releases/news-release-details/2021/Yelp-Economic-Average-Shows-Early-Signs-of-Recovery-for-Local-Economies-with-More-than-230000-Reopened-Businesses-in-2020/default.aspx

[5] Julia McCoyApril 26, 20206 min read V. I. P. C. O. N. T. R. I. B. U. T. O. R. J. M. C. W. (2020, July 2). *15 things you may not know about yelp*. Search Engine Journal. Retrieved December 19, 2021, from https://www.searchenginejournal.com/yelp-facts/355044/

[6] Business Appraisal Blog http://www.businessappraisalblog.com/14/

[7] Risky business:44% of small firm reach Year 4 http://www.nbcnews.com/id/16872553/#.U6gaHPPD-70

[8] BBC Report http://www.bbc.com/news/business-16611040

In the dataset, there is a skewed distribution of good and bad ratings across different business types and businesses' locations. There are differences in facilities that different businesses provide and this can either satisfy the customers demand of convenience or not, some common features such as Wifi, Parking, etc. This happens not only in restaurants or in metropolitan areas but also in any type of business or location these days. The motivation for this project is to predict what drives high yelp ratings. In order to have a better understanding of how these features may affect ratings on Yelp, rather than focusing on only restaurants, the whole dataset that includes different businesses and locations is required to achieve the goal of what drives a high Yelp rating for all kinds of businesses on Yelp review. For this purpose, the ratings of different businesses need to be predicted based on various different features. Ratings are the "stars" column present in the dataset, which ranges from 1, 1.5, 2, 2.5, 3, 3.5, 4, 4.5, and 5. The ratings in this paper will be predicted as binary classification, classifying ratings less than 3.5 to 0 and more than 3.5 to 1. Ratings will also be predicted as continuous variables between 1 to 5 using regression techniques because continuous variables give a better idea of the ratings to a user. For example, two businesses can have a rating of 3.5, and it would be hard for a user to determine which business is better. However, if one business has a rating of 3.567 and another has 3.711, the user can clearly understand that the latter is a better business.

Literature review

To complete this study and gain more knowledge about the subject, literature research needs to be done. Google Scholar and Rutgers Library System were used for regular research for deeper understanding of the project. Some of the articles that is found were relevant to our research are as follows: "Restaurants Rating Prediction using Machine Learning Algorithms", "Predicting Restaurants' Rating and Popularity Based on Yelp Dataset", "Zomato Restaurants Data Analysis Using Machine Learning Algorithms", "Data Driven Approach to Predicting Rating Scores of New Restaurants", and "A Machine Learning Model for Recommending Restaurants based on User Ratings". Out of all these research studies, there is a discussion of the first three studies below and how they are helpful for our research.

The first study, "Restaurants Rating Prediction using Machine Learning Algorithms," talks about rating prediction using machine learning. As the world has become digitized, food popularity apps are also increasing. Everyone orders their food in a few clicks, so it is essential to know what restaurants to order it from. The critical factors that matter are their rating, cleanilessness, staff, environment, and food. This is where the predictive rating analysis comes in. This case study focuses on mining customer ratings, authenticating them, classifying the reviews in positive and negative thoughts, and finding the worthiness of the restaurant. Various algorithms have been used in the study. They have used the dataset from Kaggle. The information talks about the information of restaurants in the city of Banglore. They performed exploratory data analysis to maximize insights into a dataset, uncover the underlying structures, extract important variables, detect outliers and anomalies, develop parsimonious models, and determine optimal factor settings. They have used different kinds of graphs for better visualization. The bar graph helps them to understand the trend and box plot to look for the outliers. The project performs both multinational classifications regarding rating prediction and binary classification in terms of rating popularity change predictions. This paper studies a

number of features about existing restaurants in different areas in a city and analyses them to predict rating of the restaurants.[9]

The second study, "Predicting Restaurants' Rating and Popularity Based on Yelp Dataset," talks about predicting restaurants' ratings and popularity based on the Yelp dataset. This study is quite relevant to the project that is being performed. They discussed how the popularity changes based on restaurant features, such as available services, price level, locations, opening hours, and closing hours. They specifically did their studies in Toronto city, representing various cuisines as their main target. The study uses linear regression, logistic regression, Naive Bayes, Neural Network, and SVM to output a predicted yelp rating of the restaurant. The main focus of their project was on restaurants in the Toronto area. For the rating prediction, they used linear regression and logistic regression. The reason for applying these two methods is because it is more robust to problematic model specifications. The project performs both multinational classifications regarding rating prediction and binary classification in terms of popularity change. Their overall prediction accuracy is around 26 to 32 percent for the multinational prediction and around 70 percent for the binary prediction. Their logistic regression test result can be used to provide restaurant improvements suggestions for business owners in Toronto. They decided not to use the text of restaurant reviews in their projections since the text of reviews directly contains relevant information about the rating reviews.[10]

The third study, "Zomato Restaurants Data Analysis Using Machine Learning Algorithms," predicts the restaurant rating based on the Zomato Bangalore dataset. This research gives us a very good input when using the Yelp dataset, but there is one difference: they focus on restaurants only, and the Yelp dataset is used and that includes more than just restaurants. However, it gives us an insight on how to predict similar data and compare the accuracy between models. This research uses a 70/30 split and focuses mainly on supervised learning methods, such as Linear Regression, Random Forest, and Decision Tree. They split into two types of Categorical Decision Tree and Continuous Decision Tree. They applied the algorithm to 17 attributes: URL, address, name, online order accept(binary), book table option(binary), rate, votes, phone, location, restaurant type, favorite dish, cuisines, cost approximation for 2, review list, menu item, type of meal and location. For the linear regression, they only perform simple linear regression by comparing two variables and choosing the one with the highest coefficient; as a result, the highest accuracy for these algorithms is only 24%. Next is the Decision Tree, which they use to predict by applying categorical and continuous values in the dataset, and the accuracy score is 85%. However, to improve the accuracy score of the Tree, they apply Random Forest, where they use the dataset and create multiple trees and give out the best Tree for the prediction of the dataset. By doing this, the accuracy score is improved by 2%, and it gives the best result out of three methods with 87% accuracy. They finally state that the prediction could

---

[9] Journal, I. J. I. T. E. E. (2021, August 4). *Volume-9 issue-6*. International Journal of Innovative Technology and Exploring Engineering (IJITEE). Retrieved October 31, 2021, from https://www.ijitee.org/download/volume-9-issue-6/.

[10] *CS 229 Machine Learning Final Project ...* - *cs229.stanford.edu*. (n.d.). Retrieved October 31, 2021, from https://cs229.stanford.edu/proj2017/final-reports/5244334.pdf.

be more accurate if there were more critics in the dataset, which could improve the research's prediction.[11]

Approach

1. *Data Collection, Cleaning and Preparation*

To achieve the purpose of this project, a dataset from Yelp was used through Yelp open dataset. Four datasets, business.json, checkin.json, user.json, tips.json were taken. After dropping the null values, for checkin.json, the day count was obtained by splitting the column value by comma, which gave the number of check-ins at every business. Business and checkin data were joined on business_id resulting in a big messy dataset with a lot of texts. This dataset was used to get the number of users visiting a particular business. It was also used to get the number of days a business stayed open and also the number of hours a business stayed open in a week. Tips and user data were joined on user_id as well. From business data, a lot of new features were created from the “attributes" column, which was a column full of dictionaries in the form of strings that required substantial amount of time and thinking to extract the meaningful attributes and use them as features such as- does a business have parking or not, does it have free Wi-Fi, paid Wi-Fi, or no Wi-Fi, does it accept credit card or not, and finally, how noisy is a business on a scale of 0 to 4 indicating no noise, quiet, average, loud and very loud. Features indicating whether a business is a restaurant, a health-related business, an educational institute, a car service business, a shop, a spa, or a fitness related business were also created by iterating through the “categories' ' column. This took a long time as NLP was not used since a lot of businesses served multiple purposes such as health and education, and NLP algorithms were not suitable for this purpose.

From the tips and user data, the number of tips a particular business received from different users was obtained by using groupby. Features showing the number of years the user has been elite, and the number of days the user has been active on Yelp were also created. Finally, the dataset containing business and checkin data, and user and tips data were merged.

Income of people really determines how a particular business would do in that area. For example, a very expensive restaurant would not do well in an area where the average income of its residents is low. Similarly, a very cheap restaurant might not do well in a region where people have a very high income as people would not like cheap food and the restaurant’s ambience. That is why income data was collected externally for various different areas present in the dataset. For the US, the average income based on the postal code was collected, and for Canada, income based on the city was collected as income for Canada based on postal code was not available. This information was then merged to our existing dataset. The dataset was then supplied for modelling. Models used were Linear Regression, Decision Tree, Random Forests, XGBoost, and LightGBM. All the models gave a very poor performance, and hence, something needed to be changed in the dataset. So, more analysis of the data was needed.

One thing that really affects ratings of a business are what customers say about the business. Tips dataset have an attribute called “text” that has a short review of what people have

[11] *© 2021 JETIR February 2021, volume 8, issue 2 zomato ...* (n.d.). Retrieved October 31, 2021, from https://www.jetir.org/papers/JETIR2102170.pdf.

to say about the business. The only way to make good use of this information was through Natural Language Processing. Most of the algorithms found were not serving the purpose of just classifying the reviews into good or bad because they all required a target variable, which the data did not have in the tips dataset. After a lot of research on different algorithms, an algorithm called zero shot text classifier developed by Hugging Face was found to be suitable for achieving what was needed. It can classify texts into different categories or classify them into positive or negative without a target column. This was used to properly classify the reviews in the tip's dataset to positive and negative. Later, everything was grouped by the business_id, and the number of positive and negative reviews were added for a particular business. This was a very good piece of information as it helped in improving the accuracy of the models by a lot more as shown in the modelling section of the paper.

2. Modeling

For modeling, the dataset was splitted into two categories by training and testing the dataset in 75/25 ratio for more than 86,000 rows of instances and the target column to predict was "stars," which is the rating for businesses that ranges from 1 to 5 and with incremental step of 0.5 given by Yelp (1, 1.5, 2, 2.5, 3, 3.5, 4, 4.5, and 5). For modelling, supervised learning was used and both Regression and Classification models were used. Finally, Ensembling was used to know if the result can be better on the test set.

1. Classification

   For classification, the ratings were converted in the "stars" column, which were the target, to 0 if they were below 3.5 and 1 if they were above 3.5. 0 indicating a low rating and 1 indicating a high rating. According to a chart from yelp showing average star rating by region, we can tell that the difference between regions is not significant. However, it can be seen that the region with lowest average star rating is Ontario with 3.56 and therefore it was decided to choose this as a measure of good and bad rating, with the prepared dataset with a target column of stars is 0.5 difference, it was decided to round it down to 3.5 and it still gave the same result. The dataset was then split into a train and test set[12]. The models used were:

   Logistic Regression:

   Since it was decided to use classification by converting stars (ratings) to binary as mentioned above, logistic regression was the obvious first choice. Logistic regression was selected as it "is often used for predictive analytics and modeling, and extends to applications in machine learning"[13]. The various features created above were used to

[12] Blog, Y. – O. (2021, August 31). *Restaurant ratings on yelp are remarkably consistent, no matter who's writing them, when, and where*. Yelp. Retrieved December 20, 2021, from https://blog.yelp.com/news/restaurant-ratings-on-yelp-are-remarkably-consistent-no-matter-whos-writing-them-when-and-where/

[13] *Logistic regression*. IBM. (n.d.). Retrieved December 19, 2021, from https://www.ibm.com/topics/logistic-regression#:~:text=It%20is%20used%20in%20statistical,or%20a%20choice%20being%20made

make predictions on the "stars," which are the ratings of a particular business. Evaluation metric used here was the accuracy score from python's sci-kit learn and confusion matrix and classification report were displayed as well, which will be discussed further in the results section.

Decision Tree Classifier:

Apart from logistic regression, the next model used was the decision tree on the final dataset to predict "stars" again as a binary variable as mentioned above. The goal was to capture trends by dividing the space into smaller regions. Trees can lead to a superior classification by capturing the division, however, "when classes are not well-separated, trees are susceptible to overfitting the training data, so that Logistic Regression's simple linear boundary generalizes better"[14]. Similar kind of problem seemed to have occurred in the case of this dataset as the accuracy was worse than the logistic regression model. However, it did have an edge in certain aspects, which will be discussed in the results section. Evaluation metric used here was the accuracy score from python's sci-kit learn and confusion matrix and classification report were displayed as well, which will be discussed further in the results section.

Random Forest Classifier:

The poor performance of logistic regression and decision tree led to the use of random forest for classification of "stars" using the dataset prepared. Again, binary classification was performed as mentioned above. Results were quite better than logistic regression and decision tree with a good increase in the accuracy. This could be because "the decision tree model gives high importance to a particular set of features. But the random forest chooses features randomly during the training process. Therefore, it does not depend highly on any specific set of features."[15] Random forest was used twice once with the dataset prepared and once again after doing resampling of the data using SMOTE.

2. Regression

For regression, different approaches were followed. The target "stars" was predicted as a continuous variable as this gives a more accurate prediction, less error, and a better R-squared score. Also as discussed in the motivation, predicting the ratings as continuous variables would only result in a more accurate result for a customer to compare and decide between different businesses. However, a different approach of predicting the "stars" as discrete was also used for the decision tree model to check if it results in a lesser error.

---

[14] Cheesinglee. (2016, October 3). *Logistic regression versus decision trees*. The Official Blog of BigML.com. Retrieved December 19, 2021, from https://blog.bigml.com/2016/09/28/logistic-regression-versus-decision-trees/

[15] *Decision Tree vs. Random Forest - which algorithm should you use?* Analytics Vidhya. (2020, May 12). Retrieved December 19, 2021, from https://www.analyticsvidhya.com/blog/2020/05/decision-tree-vs-random-forest-algorithm/

Linear Regression:

Linear regression was the first regression model that tried to predict the "stars" as continuous variables. This is because, it easily "allows you to understand the strength of relationships between variables"[16]. It also allows one to easily visualize the important features in the dataset, which can be very helpful as seen in the discussion section of this paper. The evaluation metrics used were mean squared error, root mean squared error, and R-squared.

Decision Tree Regression:

To compare the results of the linear regression, a decision tree model was implemented. Decision tree regression was used to predict the "stars" as a discrete variable and check if it is better or worse than the linear regression model. The dataset prepared was supplied to the decision tree, and the evaluation metrics used were mean squared error, root mean squared error, and R-squared.

Random Forest Regression:

Just like classification, for the regression approach to this dataset, a random forest model was decided to be used for implementation to predict "stars" as a continuous variable. Results were quite better than linear regression and decision tree with a good increase in the R-squared score and a decrease in mean squared error and root mean squared error, which were the metrics used for its evaluation.

XGBoost:

To further try to increase the accuracy of the predictions, gradient boosting was necessary. XGBoost is a gradient boosting algorithm that was used here for predicting the "stars," which are the target ratings of the dataset. XGBoost has a good execution speed and is a high performance model. It is a type of ensemble learning algorithm called boosting, "ensembles are constructed from decision tree models. Trees are added one at a time to the ensemble and fit to correct the prediction errors made by prior models"[17]. Given the fact that the dataset prepared from Yelp data is tabular, XGBoost works well with the same. The evaluation metrics used were mean squared error, root mean squared error, and R-squared.

LightGBM:

---

[16] Shin, T. (2021, August 10). *3 reasons why you should use linear regression models instead of neural networks*. Medium. Retrieved December 19, 2021, from https://towardsdatascience.com/3-reasons-why-you-should-use-linear-regression-models-instead-of-neural-networks-16820319d644

[17] Brownlee, J. (2021, March 6). *XGBoost for regression*. Machine Learning Mastery. Retrieved December 19, 2021, from https://machinelearningmastery.com/xgboost-for-regression/

After XGBoost, LightGBM was implemented to see if the accuracy of the predictions can be improved even further. It gave promising results as the mean squared error was lower than XGBoost and the R-squared score also increased by a little bit. This model was chosen because it offers a very high accuracy and training speed, handles overfitting very well, and it also utilizes low memory[18]. The evaluation metrics used were mean squared error, root mean squared error, and R-squared.

Class: AveragingModels:

Finally, a custom class was created that was used for ensembling the above regression models except decision tree because of its poor performance, which will be discussed later. This class took the average of the predictions that the models made, and returned its own predictions. It gave substantially better results than any of the models did alone, which will be shown in the results section. The evaluation metrics used were mean squared error, root mean squared error, and R-squared.

Result

1. Classification

For classification, as discussed earlier, the ratings in the “stars" column were divided into 1 if they were above 3.5 and 0 if they were below 3.5, making the problem a binary classification problem. The resultant column of 1 and 0, was called “star_bin” and the “stars” column was dropped. Features used for training and prediction of “star_bin” were- review_count, is_open, checkincnt, ofdaysopen, Totaltime, AcceptCreditCard, Wifi_free, Wifi_no, Wifi_pay, Noise, parking, Restaurant, Health, Education, Spa, Fitness, Car_services, Shop, nu, negative, positive, elite_years, days_yelped, income.

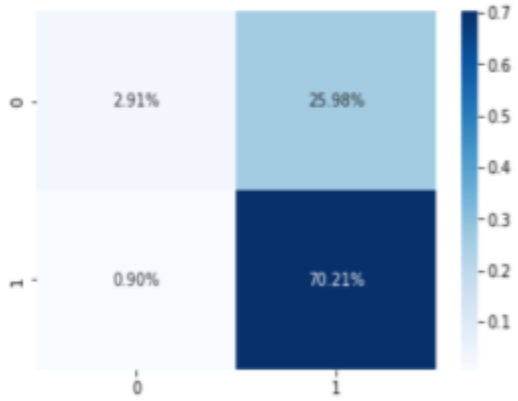
Baseline:

For a baseline classification model, a dummy classifier from sklearn.dummy was used and it gave an accuracy of 71%.

Logistic regression:

Logistic regression is one of best and goto models to start with for any classification problem in machine learning. Therefore, it was decided to be used as the first model to see if better results than the baseline model can be achieved. Accuracy score was used to evaluate the model. It resulted in an accuracy of 0.73 or 73%, which is not good given that the baseline model gave an accuracy of 71%. The following are the confusion matrix and the classification report of logistic regression model-

[18] *LightGBM in python: Complete guide on how to use lightgbm in python*. Analytics Vidhya. (2021, August 18). Retrieved December 19, 2021, from https://www.analyticsvidhya.com/blog/2021/08/complete-guide-on-how-to-use-lightgbm-in-python



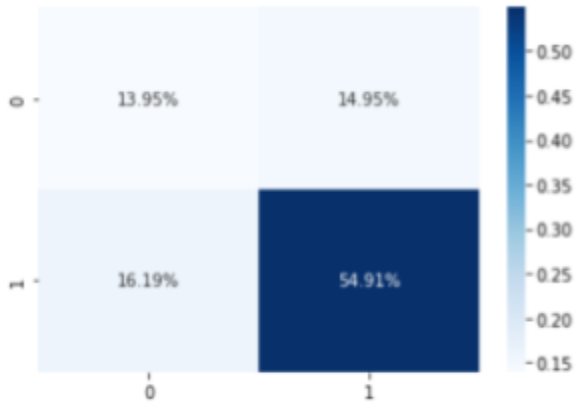
	precision	recall	f1-score	support
0	0.76	0.10	0.18	6259
1	0.73	0.99	0.84	15400
accuracy			0.73	21659
macro avg	0.75	0.54	0.51	21659
weighted avg	0.74	0.73	0.65	21659

Figure 1

The model has very high false positives (25.98%) as seen in the confusion matrix that most of the times the model predicts 1 even if the actual value is 0. So basically the model tends to predict high ratings a lot more than low even though the actual value in the test set is low. This model is not a good model given the low accuracy and high false positive rate. This is also because only about 6259 0s are present in the class versus 15400 1s as seen in the support of the classification report. The recall for the minority class is very low

Decision Tree:

After trying logistic regression, the decision tree was used to check if it gives better predictions or not. It was important to get a better performance than the baseline and logistic regression both, and also reduce the high number of false positives. Accuracy score was used to evaluate the model. Decision tree resulted in even worse accuracy of only 0.69 or about 69% compared to logistic regression and the baseline model. Below are the confusion matrix and classification report of decision tree model-



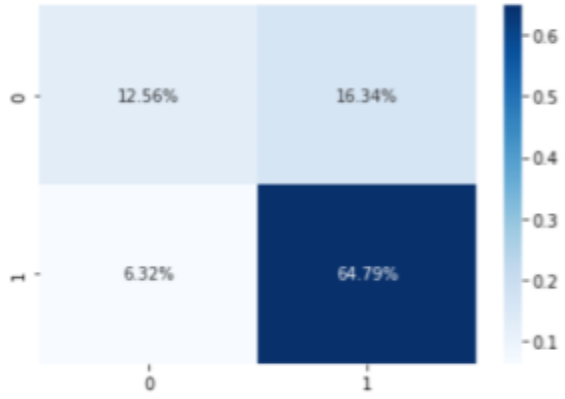
	precision	recall	f1-score	support
0	0.46	0.48	0.47	6259
1	0.79	0.77	0.78	15400
accuracy			0.69	21659
macro avg	0.62	0.63	0.63	21659
weighted avg	0.69	0.69	0.69	21659

Figure 2

It is clear that it performed badly, however, the false positives (14.95%) and true negatives (13.95%) have a little balance as compared to logistic regression where false positives were way higher than true negatives. The recall for the minority class is very low again.

Random Forest:

Given the poor performance of the decision tree, random forest was decided to be implemented as it is an ensemble of decision trees, and it could result in a better performance. Accuracy score was used to evaluate the model. It gave an accuracy of 0.77 or 77%, which is better compared to the previous models used and also the baseline model. Below are the confusion matrix and classification report of random forest model-



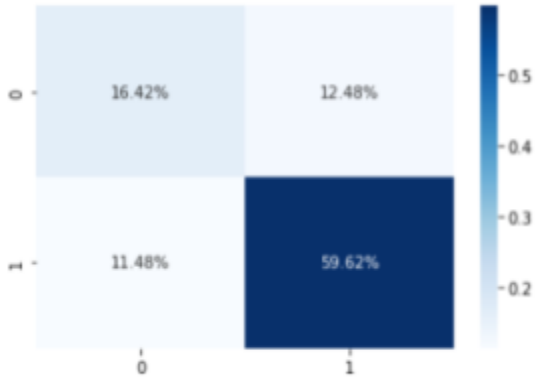
	precision	recall	f1-score	support
0	0.67	0.43	0.53	6259
1	0.80	0.91	0.85	15400
accuracy			0.77	21659
macro avg	0.73	0.67	0.69	21659
weighted avg	0.76	0.77	0.76	21659

Figure 3

It is evident that the accuracy is the best for random forest models compared to other models. True positives (64.79%) have decreased compared to logistic regression model, however, the false positives (16.34%) have also decreased, and true negatives (12.56%) have increased but not by much. The reason for these results will be discussed in the discussion section of the paper.

Random Forest Classification with SMOTE:

In the efforts of reducing the number of false positives and increasing the recall for the minority class (0.43), random forest classifier was trained again after resampling the data to evenly distribute the 0s and 1s. For this purpose, SMOTE was used. It resulted in accuracy dropping from 0.77 to 0.76. Below are the results of the same-



	precision	recall	f1-score	support
0	0.59	0.57	0.58	6259
1	0.83	0.84	0.83	15400
accuracy			0.76	21659
macro avg	0.71	0.70	0.71	21659
weighted avg	0.76	0.76	0.76	21659

Figure 4

After resampling, recall for the minority class increased from 0.43 to 0.57, so this makes it a better model. False positives are lower and true negatives are higher for this model, hence, the little loss of accuracy can be a useful compromise.

2. Regression

As discussed in the approach, models used for regression predicted the "stars" (ratings) as a continuous variable, and that is the approach followed in order to try to increase the accuracy of the models, and displaying the ratings in a better manner to the user. If a user can see the ratings of a business as a continuous number rather than discrete, it will help them compare different businesses better. Features used for training and prediction of "stars" were- review_count, is_open, checkincnt, ofdaysopen, Totaltime, AcceptCreditCard, Wifi_free, Wifi_no, Wifi_pay, Noise, parking, Restaurant, Health, Education, Spa, Fitness, Car_services, Shop, nu, negative, positive, elite_years, days_yelped, income. For this purpose, the first model that was used is linear regression.

Linear regression:

Linear regression is one of best and goto models to start with for any regression problem in machine learning. Therefore, this was the first model to be used to predict "stars" for the dataset that was prepared. The results of linear regression were evaluated by using mean squared error, root mean squared error, and R-squared. The results are displayed below-

MSE	RMSE	R-squared
0.567	0.753	0.156

Table 1

Based on the MSE, RMSE, and R-squared, it can be inferred that regression model is not performing very well as the MSE and RMSE may be low given the nature of the problem, however, R-squared value is too low to call this model a good one. The MSE of 0.567 and RMSE of 0.753 show that the error is not that high since the continuous variable being predicted is not very far off on an average. Recall that the actual "stars" values are 1, 1.5, 2, 2.5, 3, 3.5, 4, 4.5, and 5. Hence, on an average the predictions are far off from the actual values, they are not good enough, and the R-squared definitely needed to be increased.

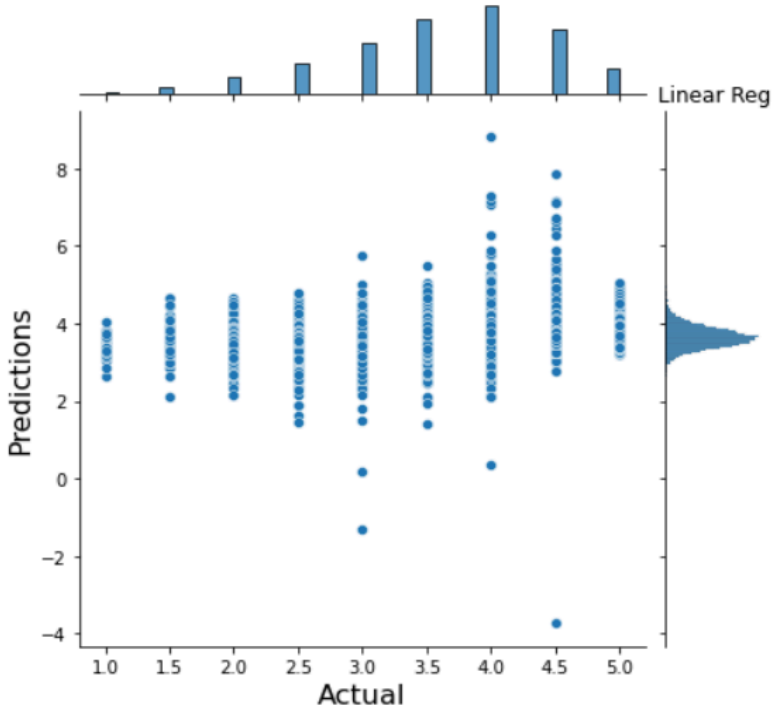


Figure 5

It is evident from figure 1 that linear regression is not able to predict very well. Lots of predictions happen to be between 3.5 and 4.5, and there are also certain negative values for ratings, which does not make sense. There are also several values more than 5, which suggested that a better model was needed to solve this issue.

Decision Tree:

Next model that was tried was the decision tree as it was a good option to consider after using a linear model like linear regression. The results of the decision tree were evaluated by using mean squared error, root mean squared error, and R-squared. The results are displayed below-

MSE	RMSE	R-squared
0.908	0.953	-0.353

Table 2

It is clearly evident that the decision tree is performing even worse than linear regression with a terrible mean squared error and root mean squared error, and a negative R-squared, which was very disappointing to see. The reason can be explained by the graph below-

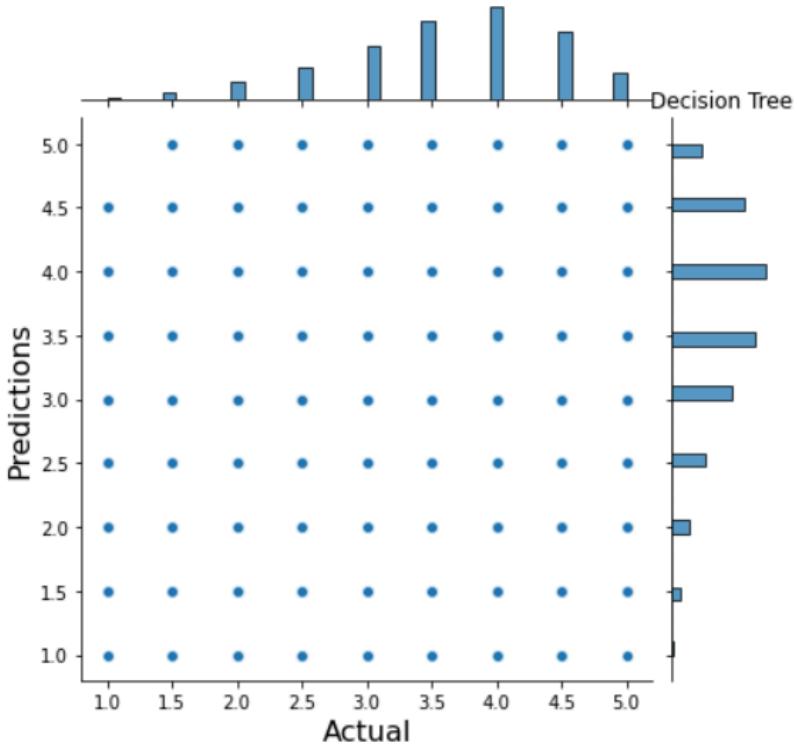


Figure 6

Decision tree regressor predicted the “stars” as a discrete variable rather than continuous. In this case, like linear regression most values are between 3.5 and 4.5, which is depicted by the histogram in the right of the scatter plot.

Random Forest:

Despite a very poor performance of the decision tree model for regression analysis, random forest was decided as it is an ensemble of several decision trees, and the idea was to be

able to achieve a better prediction through random forest as it works on a random subset of features, and this is very suitable for large datasets such as this one. The results were better than expected as below-

MSE	RMSE	R-squared
0.451	0.672	0.328

Table 3

This was a good model to use as it is evident that there was quite a big difference between random forest and decision tree for prediction. Also the joint plot in figure 3 clearly shows that the random forest solved the problem that arises in linear regression of getting negative values as predictions and getting predictions of more than 5. The problem over here is that very few values get predicted correctly for 1, 1.5, and 2 stars.

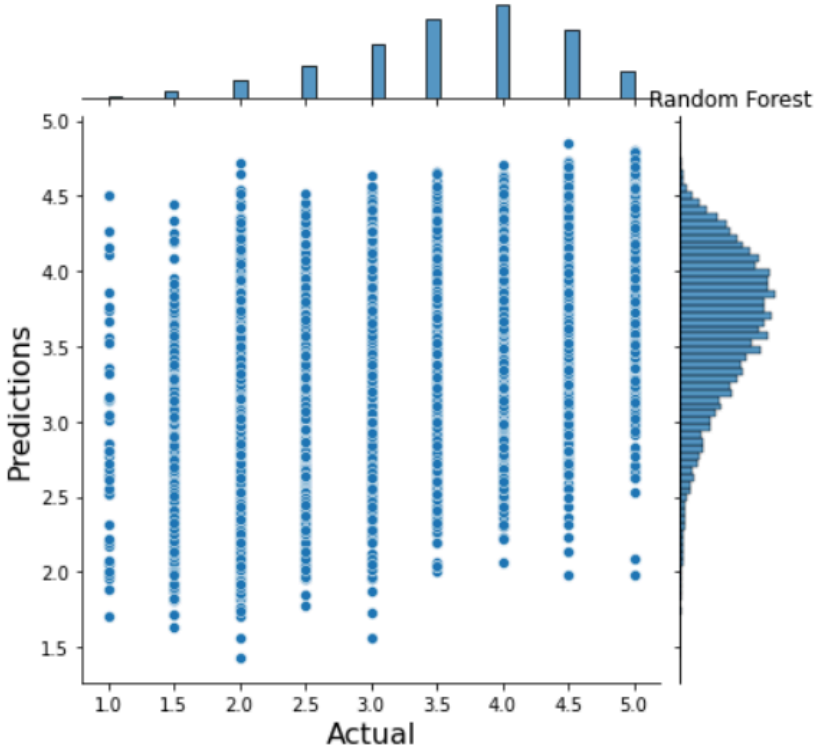


Figure 7

XGBoost:

XGBoost as discussed in the approach section of the paper is an ensemble learning algorithm and uses gradient descent to boost the accuracy. Hence, this model was decided to be used for checking if it gives a better prediction than random forest. The results of the XGBoost were evaluated by using mean squared error, root mean squared error, and R-squared. The results are displayed below-

MSE	RMSE	R-squared
0.423	0.650	0.370

Table 4

XGBoost was able to give good results as the mean squared error and root mean squared error decreased even more than random forest, but also R-squared value increased. The distribution of the prediction is displayed in the figure below-

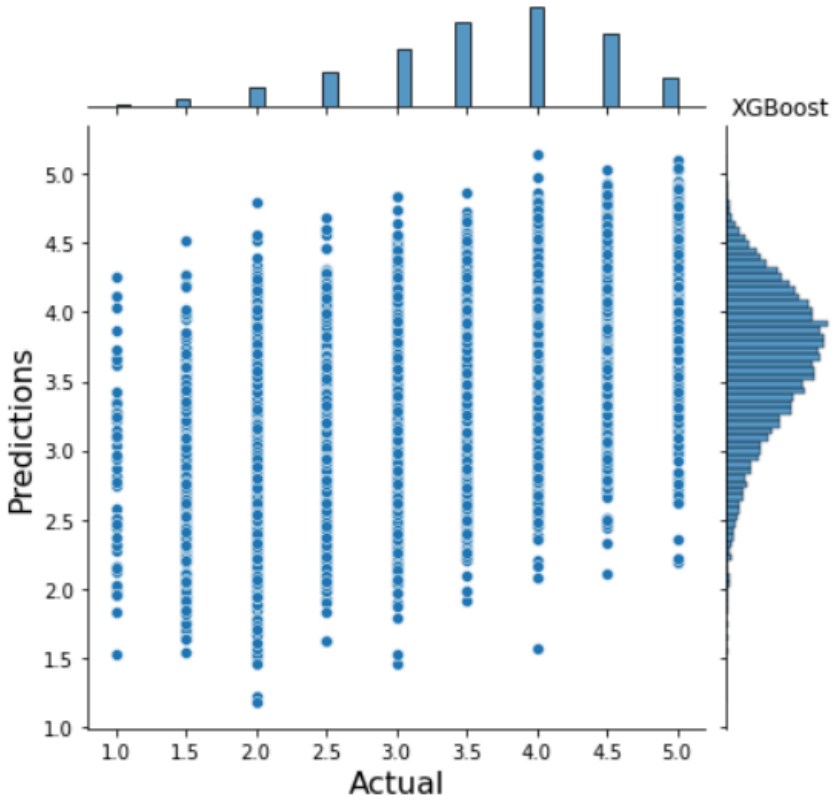


Figure 8

Model seemed to be doing a better job than random forest at predicting values for “stars” between 3.5 and 4.5, however, this model also suffered while predicting from 1 to 2.5.

LightGBM:

One of the highest performing models and one of the fastest ensembling algorithms- LightGBM- was used to see if the accuracy can be increased and a model better than XGBoost can be used. LightGBM supports the use of categorical variables, encoded as numbers, and the features that were supplied as categorical features from the prepared dataset were- 'AcceptCreditCard', 'Wifi_free', 'Wifi_no', 'Wifi_pay', 'Noise', 'parking', 'Restaurant', 'Health', 'Education', 'Spa', 'Fitness', 'Car_services.' The results were promising and the best ones compared to any other models.

MSE	RMSE	R-squared
0.422	0.650	0.383

Table 5

LightGBM gave the best results with the lowest mean squared error and root mean squared error, and a higher R-squared value. The jointplot below of actual versus predicted values also looks better than other models as it is a bit more even. However, the lower “stars” values were still not being predicted as good as they should be.

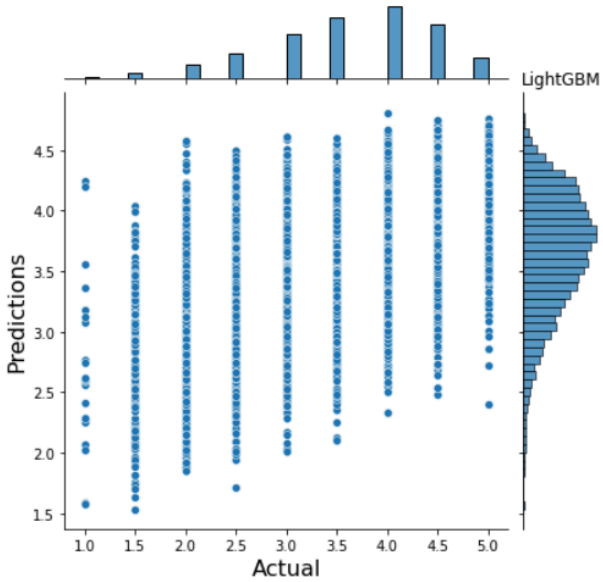


Figure 9

Custom class: AveragingModels (ensemble):

The LightGBM and XGBoost models were performing fairly well, however, there was still scope of improvement in the values of mean squared error and root mean squared error, and increase in R-squared value. For this purpose, a class called AveragingModels was written that combined the predictions of different models and average of the predicted values from different models. It basically created an ensemble model of different numbers of models that were supplied to it as arguments while instantiating the AveragingModels object. Linear regression, random forest, XGBoost, and LightGBM were used for this class. Decision tree was not used due to its poor accuracy. The results are displayed below-

MSE	RMSE	R-squared
0.313	0.560	0.532

Table 6

These results are better than any of the results given by the models individually. The lowest mean squared error and root mean squared error values were obtained, and the highest R-squared value was observed. The jointplot below shows the distribution of the actual versus predicted values-

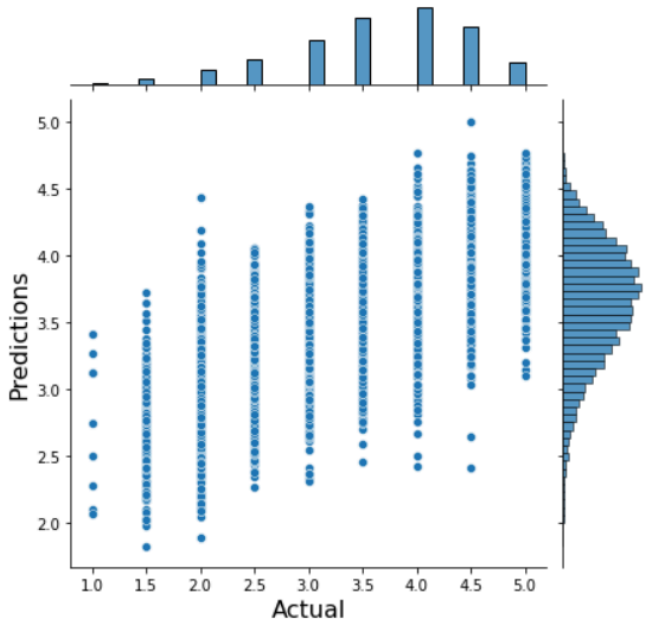


Figure 10

It is evident that the model was able to predict quite well between 3 to 4.5. However, even this combination of different models is not able to do well for test values of 1 to 2.5. It is doing better than most models, but still not good enough. The reason behind this will be discussed in the discussion section of the paper.

Discussion

The reason why the models for both classification and regression are not able to perform that well is because the data is biased as it has more number of "stars" ratings values ranging from 3.5 to 4.5. A point that needs to be mentioned is about the distribution of ratings in the dataset. Yelp dataset includes most businesses with ratings running from 3.5 to 4.5 stars ratings and less businesses with ratings below 2.5 stars. This results in a low accuracy score in classification problems because the algorithms are not getting enough data for "stars" ranging from 1 to 2.5. When the "stars" below 3.5 are converted to 0, and those above 3.5 are converted to 1, as discussed in the previous sections, it introduces a lot more 1s than 0s, resulting in a high false positives. It results in a very low recall for the minority class. However, by using SMOTE to resample the classes, the recall for the minority class for random forest classifier was increased to 0.57 as opposed to the one without resampling (0.43).

The same goes for regression models as well. Since most of the instances in the data have "stars" ratings ranging from 3.5 to 4.5, the models tend to predict less between 1 to 2.5. This could be because customers tend to give most businesses a mediocre to high rating on Yelp. This is clearly visible in the distribution of the "stars" rating in the graph below, which shows a skewed distribution between 3.5 to 4.5 stars rating.

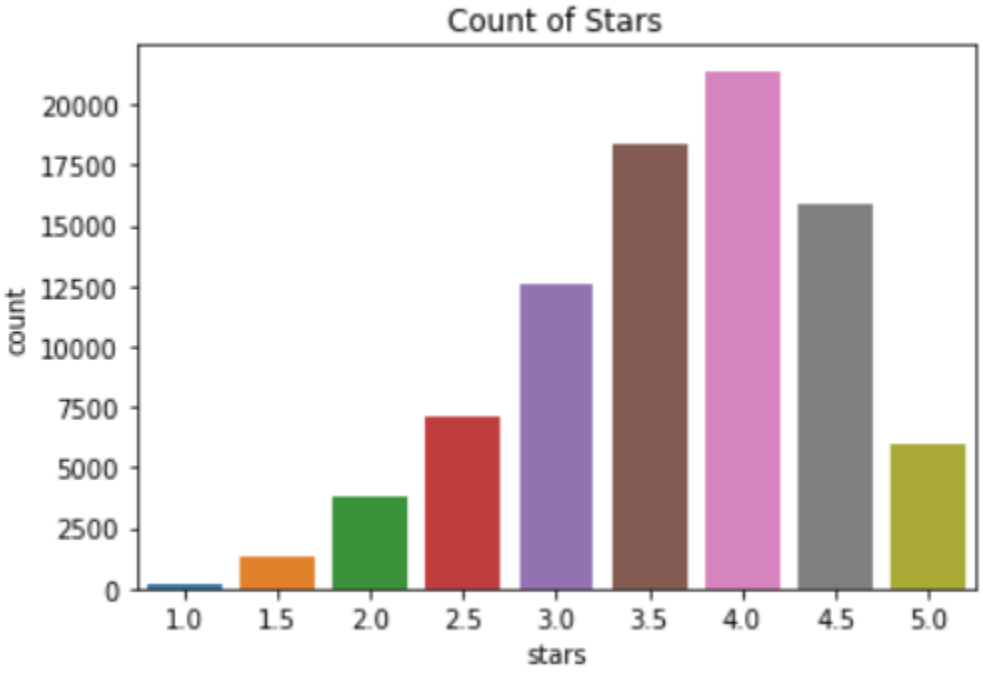


Figure 11

There are several factors that affect the performance of the models and this can be inferred by finding the importance of different features in the dataset. This would also help answer the question- what drives a high Yelp rating? Below is the graph showing feature importance for the random forest regression model-

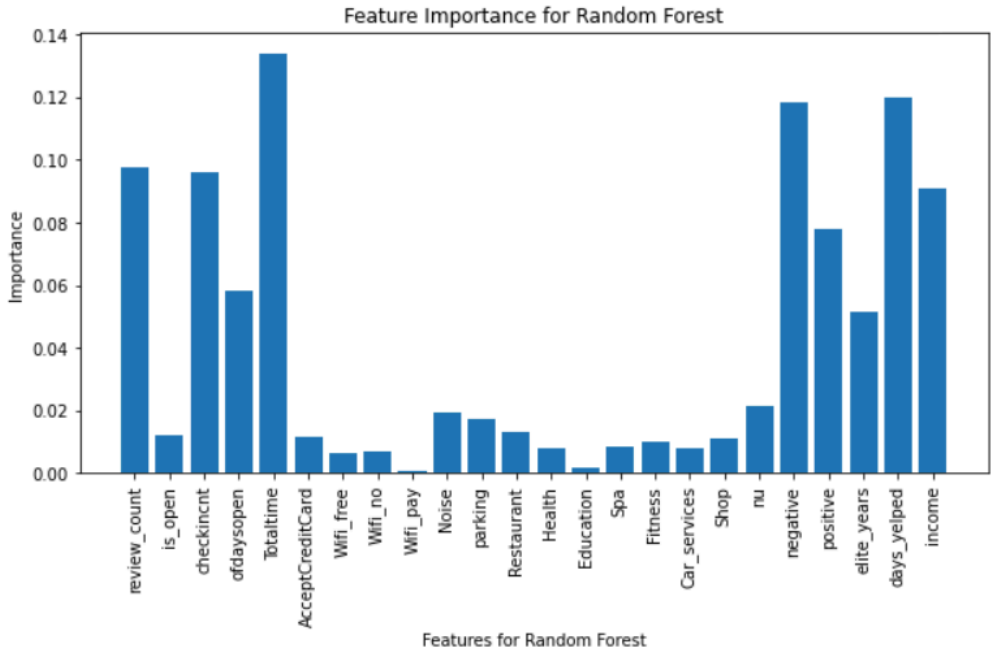


Figure 12

It can be inferred that the most important features for the accuracy of random forest model are review_count, checkincnt, Totaltime, negative, positive, days_yelped, and income. This means that most of the features in the dataset are not contributing much to the accuracy of the model. It makes sense that these features are important and one can say that they drive a high Yelp rating for a business as well. Especially, the number of negative reviews tend to really affect the ratings of a business. Number of reviews, number of customers (check-ins), total number of hours a business stays open in a week, number of positive reviews, the number of days a user has been active on Yelp, and income of the area where a business is located do matter, and that is what the features shown important by random forest depict. This can be compared to the feature importance of the LightGBM model as well.

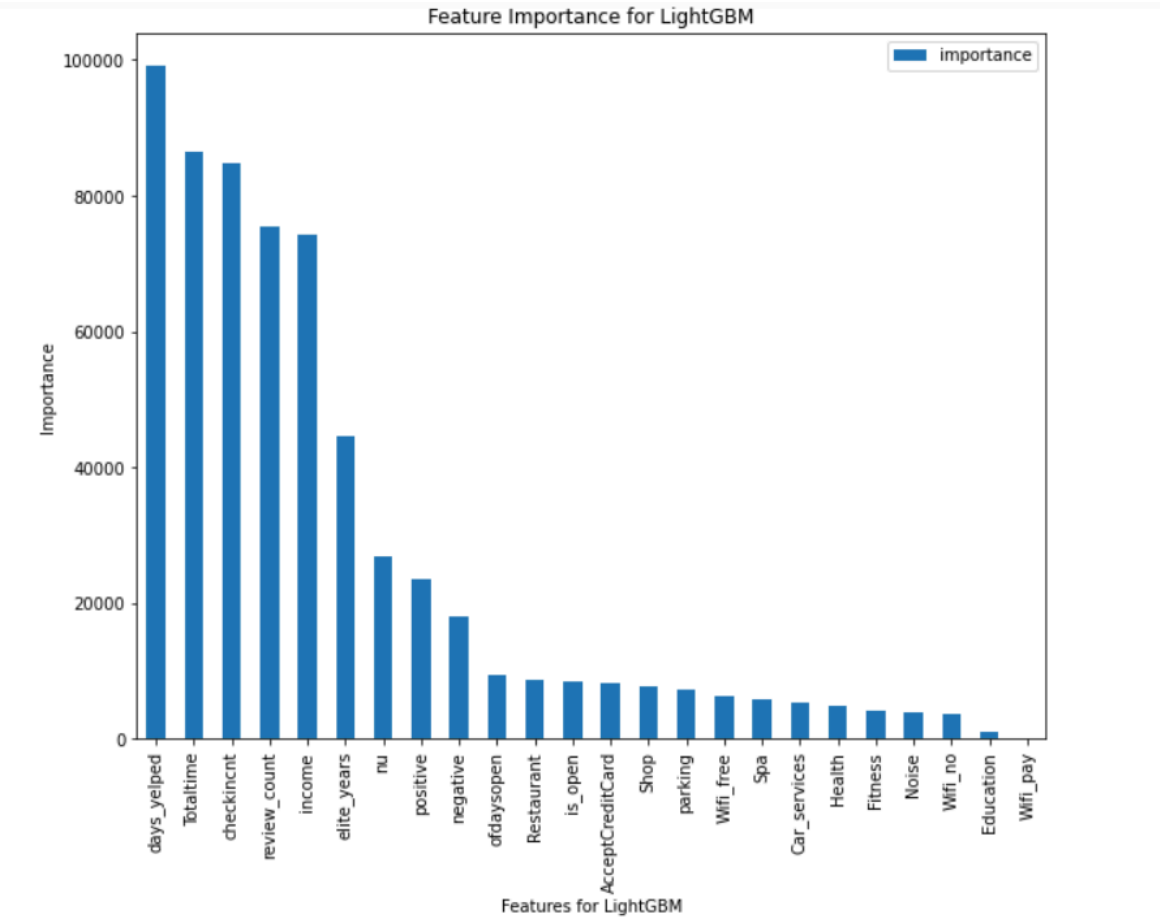


Figure 13

The LightGBM model also gives importance to those features that random forest considers important. Only exception here is the “nu” column having more importance than “positive” and “negative” columns, which was not shown by random forest model. Since, LightGBM performs better than random forest, it can be said that “nu,” which are the number of tips received by a particular business, also contributes to model accuracy. Number of tips can also significantly change the ratings of a business on Yelp, as good businesses tend to receive more tips than bad businesses. It is also evident that whether a business has parking, accepts credit cards, and has free Wi-Fi, tends to have some impact on the ratings but not much.

Number of visits and check-ins, represented by checkincnt, also create a huge impact on the ratings of a business, this indicates a strong relationship between the attributes check in count with the target of star rating. The graph below shows that the “stars” for a business increases with the average number of visits or it could be vice versa. However, there is an exception of 5 stars rating because the data collection of 5 stars rating is lesser compared to the rest of the dataset.

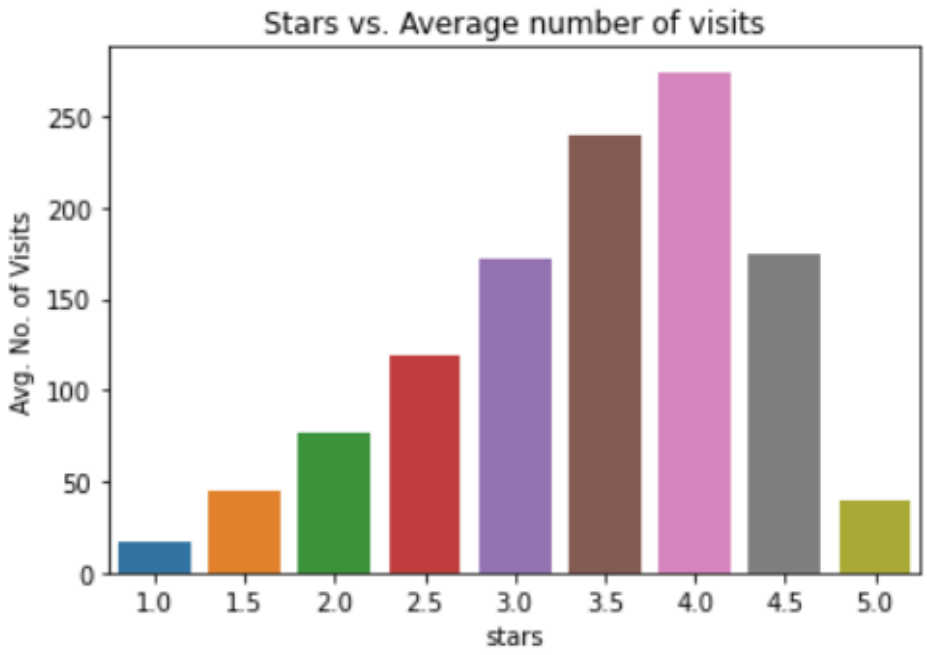


Figure 14

The number of reviews and income are also very important features for determining the ratings on Yelp, which is supported by both random forest and LightGBM feature importance graphs above. The scatter plot below shows the distribution of review count based on ratings, and this is strong evidence why the results of the models are able to predict more between 3.5 and 4.5. The average income is high mostly from metropolitan areas that are densely populated and tend to have more local businesses than other areas. This leads to customers having more alternate options available to choose from when their choice is not available enabling them to try a lot of new businesses, products, and services. The bar graph below shows that as the average income increases, the “stars” rating increases, which again shows the reason for high importance of the income in both random forest and LightGBM. It shows that it is better to open a business where the average income of the people living in the area is high.

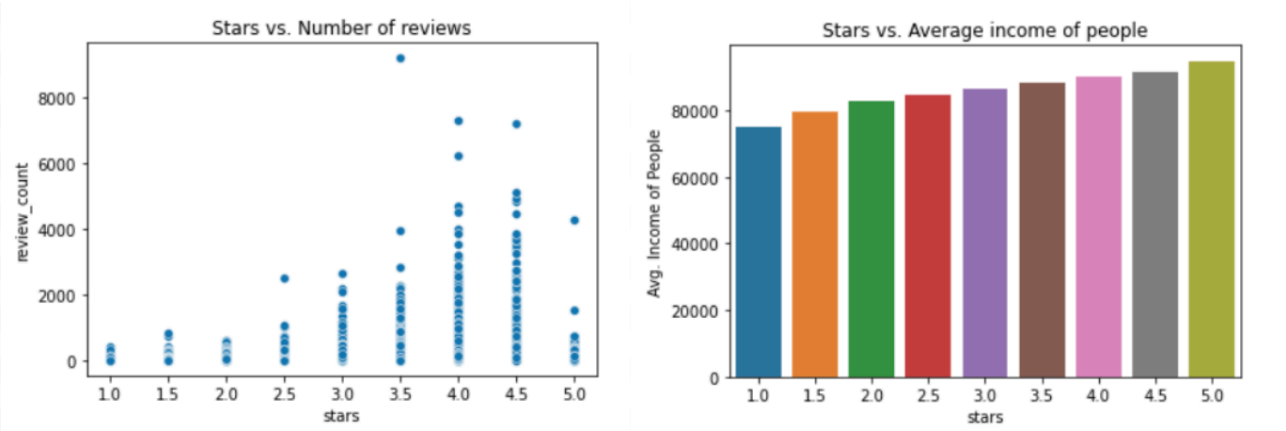


Figure 15 & 16

The high income could mostly be in metropolitan cities, and businesses opened in cities tend to get more reviews. The bar graph shows the top twenty cities with the highest number of businesses collected in the Yelp dataset. Number of businesses in the dataset basically represents the number of reviews as every business in the dataset has at least one review.

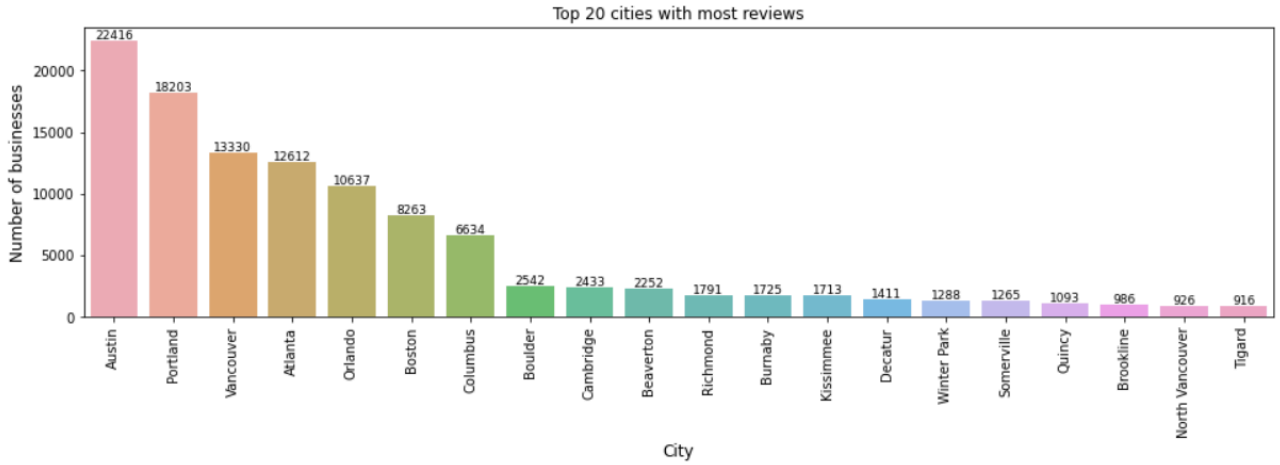


Figure 17

Lastly, it is worth exploring the features that indicate whether a business is a restaurant, shop, health related, etc. The number of different types of businesses is a factor that needs to be considered. This is because from the Yelp dataset there is more data of restaurants compared to other types of businesses, which could be one of the reasons that created inaccuracy for machine learning models. The bar graph below shows different types of businesses that are present in the data and the total number of counts by each type.

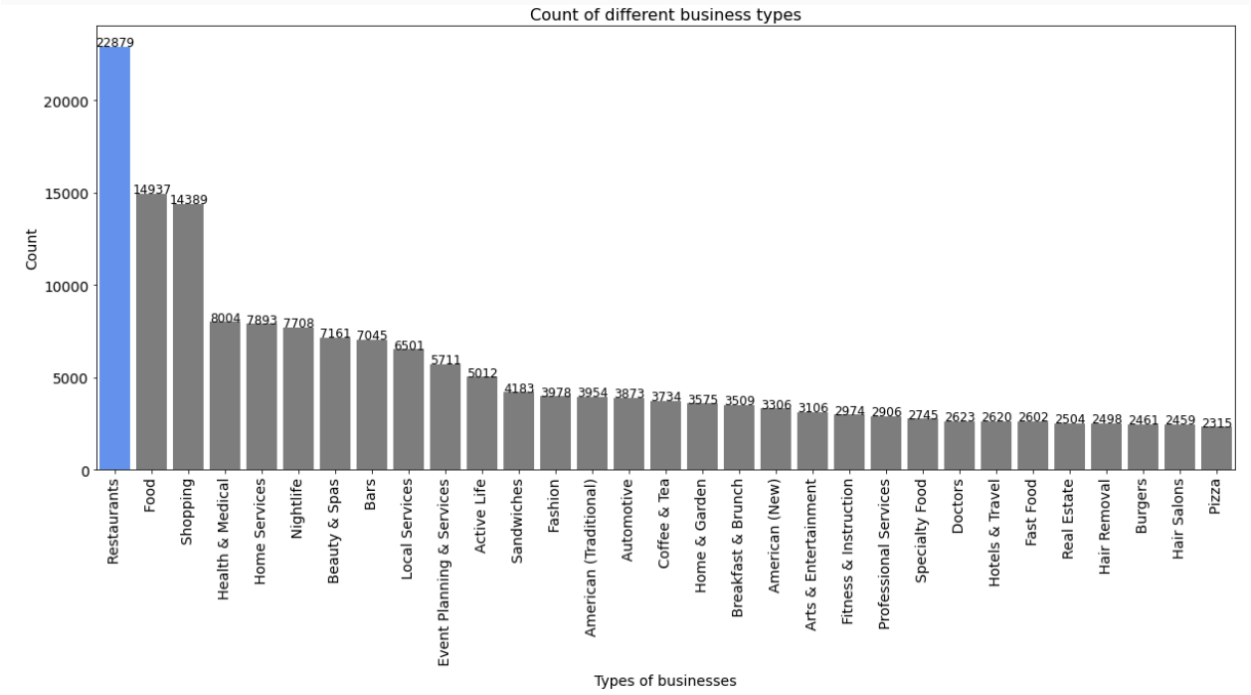


Figure 18

Conclusion

In conclusion, from the results and discussion, it can be inferred that there is not a very strong relationship and correlation between most of the attributes with the target “stars,” which is the ratings of a particular business. Classification algorithms did not give a satisfactory performance except for random forest, which was somewhat better than logistic regression and decision tree classifier. False Positives is the biggest problem in classification learning, even with the accuracy score of 77% for Random Forest Classifiers, there is still a high percentage of False Positives. This indicates that there are some outliers and also lack of data in target value ranging from 1 to 2.5, and 5 stars rating, which makes the result skew toward target value of 3.5 to 4.5. This concludes that more data should be collected from businesses with ratings ranging from 1 to 2.5 and 5 stars rating. However, in the current dataset the best model for the classification approach is Random Forest Classifiers because of the highest prediction accuracy. After resampling the data using SMOTE, random forest was able to achieve a better result. Even though the accuracy went down a little bit, it helped increase the recall for the minority class, hence, reducing the number of false positives in the predictions. There was a tradeoff between accuracy and recall, but the accuracy was not that low even after SMOTE and there was a substantial increase in recall for 0s, which makes random forest classifier, after applying SMOTE, the best model for binary classification approach of this problem.

In regression analysis, linear regression outperformed decision trees by a big margin, however, it still was not a good model, which drove the usage of random forest. Random forest did perform better than linear regression as it is an ensemble of several decision trees, but it still was not performing that well given its low R-squared value. XGBoost and LightGBM tend to

work better than the other algorithms for regression giving the lowest mean squared error and root mean squared error, however, R-squared value was still not very good with the highest R-squared value being 0.383 in case of LightGBM. Finally, the custom ensembling class AveragingModel that created a model by taking the average of the predictions of linear regression, random forest, XGBoost, and LightGBM resulted in the lowest mean squared error (0.313) and root mean squared error (0.56), and the highest R-squared value (0.532) making it a decent model for this problem.

Natural Language Processing is one of the most essential methods that were used in the approach of this project because it was very informative to include the "text" from the tips data when doing analysis. It helped the models to train and predict better by including the number of positive reviews and negative reviews that a business receives. Feature importance graphs also show that these attributes are really important for the models. Thinking about it, even in the real world, what people have to say about a particular business really impacts the ratings and success of a business.

Apart from what users have to say, different aspects such as number of tips received by a business, number of reviews, number of customers visiting (check-ins), total number of hours a business stays open in a week, the number of days a user has been active on Yelp, and income of the people of the area where a business is located are very important factors in determining the success of a business. These aspects along with the positive and negative feedback of people for a particular business should be considered very carefully as they are what drives a high Yelp rating for a business.

Features such as- number of days a business stay open, the type of business (restaurant, shop, health, fitness, etc.), having Wi-Fi or not, is a business noisy or not, whether a business accepts credit card or not- do not matter much in impacting the ratings of a business, according to the models used. For future work, more data can be collected of businesses having a rating of 0 to 2.5, and 5 stars rating to even the number of "stars" rating distribution in the dataset. This can help to achieve better prediction accuracy and performance in every model. Along with this, analyzing different business types can also be helpful in getting a better insight on what are the factors that give a certain business a decent rating, and comparison between businesses can become useful. Moreover, by having a better understanding of each business geographically, machine learning based on a certain city can be more informative to learn about what is the most important feature for a business at a certain location. Just like random forest classifier, for the binary classification discussed in this paper, SMOTE can be used to resample the classes, giving a more balanced data of high (1) and low (0) "stars" ratings. As with the binary classification approach, SMOTER can be used for the regression approach to create a more balanced dataset for regression, which can help the models to predict better and work well with lower "stars" ratings as well.