IRBAS www.irbas.academyirmbr.com April 2022
International Review of Basic and Applied Sciences Vol. 10 Issue.2

Machine Learning Algorithms for Predicting the Impact of Business Attributes Over Business Stars of Yelp Data-Set

BUSHRA HAQ
Department of Computer Science
Balochistan University of Information Technology, Engineering & Management Science
Quetta 87100, Pakistan
Email: bushrahaq@ieee.org

SOHRAB KHAN
Department of Computer Systems Engineering
Balochistan University of Engineering and Technology, Khuzdar 89100, Pakistan
Corresponding Author Email: sohrab@buetk.edu.pk

RAHILA UMER
Department of Computer Science
Balochistan University of Information Technology, Engineering & Management Science,
Quetta 87100, Pakistan
Email: rahilaumer@buitms.edu.pk

NOOR UDDIN
Department of Software Engineering
Balochistan University of Engineering and Technology, Khuzdar 89100, Pakistan
Email: engr.nooruddin@yahoo.com

Abstract

YELP is a crowed-sourced reviewing platform which connects people to local businesses and let people to review the businesses and see the reviews of any business. In this paper, we are focused on the reviews of the restaurants. We aim to predict the business- stars rating for a restaurant from the business attributes, such as the ambience and intimate attributes. For this paper, we investigated 12th Round Yelp Business Data-set which is publicly available on the official YELP website. Through this data we will predict the stars (rating) of a review. The two feature selections, a) Correlation Coefficient and b) Wrapper Method, with the combination of two machine learning algorithms, a) Random Forest Tree and Decision Tree Model are used. After the analysis of the performance of each model, the Random Forest Tree algorithm is found to be the best model for predicting the stars/ratings of the reviews. Moreover, we tried to find the subset or combination of business attributes which impacts positively on business stars/ratings.

Keywords: *Correlation Coefficient, Decision Tree, Feature Selection, Machine learning, Predicting, Random Forest, Wrapper Method, Yelp Data-set.*

Introduction

While making choices to select any business, it is commonly observed that people make their choices based on the stars/ratings of the business. People will obviously choose the business with higher stars/ratings as it will ensure the high quality of service rendered by the business. In the YELP search, the judgement of the users is the first influence which can be seen through the business's stars/ratings. Moreover, the star meter under the business name can determine the behavior of the people that whether people will find out more

about the business or look for some other businesses. According to the economic research, an additional half star in the stars/ratings can allow the restaurants to seek more visitors' frequency, which shows that the stars/ratings is the key element of the YELP experience (Andreson & Magruder, 2012). YELP stars/ratings for any specific business is the mean of all the stars/ratings given to that specific business. However, it is very important to keep informed with the consequences of the representation of any business entirely through the single star/rating.

The difference can be caused as one particular user cares more about food, but the restaurant's review has 1 star/rating due to the people complaining about the poor or low service which can ruin their delicious meal (Fan & Khadmi, 2014). This 1 star/rating restaurant could be ideal but the user will keep looking for other restaurants. The different feature selections in the data analysis process will rectify the accuracy significantly in the learning models. Therefore, it's mandatory to identify the important features that influences the review stars/ratings. In this paper, we reviewed various similar studies which elaborated different ways to predict the business stars/ratings through different machine learning models. However, there was no suggested combination of multiple attributes in the YELP business data- set which directs the businesses to improve their business stars/ratings. Therefore, we applied different feature selection and machine learning models to analyze the strong attributes from the YELP business data-set which provides the impact of business stars/ratings.

Background

The aim of this study is to identify the important attributes of YELP business data-set which impacts positively on the business star/rating attribute. Singh et al. use feature selection but the criteria are different than the criteria in our research which is discussed in methodology section (Singh et al., 2019; Carbon, Fujiii & Veerina, 2014). The training and testing accuracy for their paper is found almost similar as ours. However, as in this paper, we are not discussing the training and testing accuracy and we are more focused on the feature selection in different ways. To find the best prediction results, we have used two different methods feature generation methods & machine learning models. The prediction of the business-star/review-rating through the text base data mining and/or sentiment analysis is also being reviewed by different researchers (Fan & Khadmi, 2014; Yu, Xue & Ouyang, 2015; Asgher, 2016; Li & Zhang, 2018; Singh et al., 2019).

Another research was carried out by Lu et al. which predicted that the restaurants will still open till 2017 using the YELP data of 2016. To predict the success and continuity of restaurants, the authors used the feature selection technique. For the predictions, the feature selection was divided into two categories, Text & Non-Text features, and to articulate the results from the research, the feature engineering and logistic regressions were performed on the data. The feature engineering was performed through the feature ablation study. However, both text and non-text features were used for the prediction of the ratings of restaurants during the study (Lu, Qu, Jiang & Zhao, 2018). Additionally, the Chen et al. evaluated the top five, lowest five, good and bad features, then applied five supervised learning algorithms (Linear Regression, Logistic Regression, Na¨ıve Bayes, Decision Tree, and Neural Network) in Rapid Miner to predict. The researcher concluded that the Neural Network & Decision Tree give better results as compared to other learning algorithms (Chen & Xia, 2020).

Hegde et al. also worked on the feature selection and attributed from the Yelp Dataset on Restaurant Businesses and highlighted the most crowded day by reviews through the review date attribute, nearest locations of the restaurants by kd- tree algorithm applying on location attributes (longitude and latitude) and high frequency attributes (credit card, parking, wheelchair, and Wi-Fi) by weightage attribute derived from Yelp Data-set. However, the researchers did not explain that how they get the results from the

research (Hedge, Satyappanavar, & Setty, 2017). While looking further for more deep learning techniques on Yelp Data-set, we found that Rafay et al. applied various extraction features & pre-processing techniques prior to using the LSTM and MNB algorithms for text classification on Yelp Data-set. There are many other researches that contribute in machine learning techniques which we also used in our research but we included only those papers in which these techniques are being applied on Yelp Data-set (Rafey, Suleman & Alim, 2020). Moreover, Reddy et al. worked on Yelp Review Data in which the researchers pre-processed it first and then classified the customer reviews category based on machine learning approaches like Multinominal Na¨ıve Bayes, and Random Forest. The findings evaluated that Random Forest generates more accurate results as compare to other classifiers. Therefore, we decided to apply Random Forest on our data-set as well (Elkouri, 2015).

Another study Elkouri et al. applied Na¨ıve Bayes, Support Vector Machines, and Logistic Regression as models to per- form positive/negative classifications and 5-Star Classification accuracy on reviews data-set of Yelp. The researchers also suggested Random Forest for 5-Star Classification accuracy which we are performing in our business data-set of Yelp (Elkouri, 2015). Furthermore, Liu (2020) predicts rating from the review text with four machine learning methods which includes Na¨ıve Bayes, Logistic Regression, Random Forest, and Linear Support Vector Machine (SVM) and four transformer-based models, including BERT, DistilBERT, RoBERTa, and XLNet (Liu, 2020). Kong et al. also did feature selection to identify the business attributes that link to high star ratings for each country. Na¨ıve Bayes, Support Vector Machines (SVM), Decision Tree, Logistic Regression, and Gaussian Discriminant Analysis (GDA) models were used to classify the data and calculate the strength of the feature sets, some researchers used the chi-square scoring function to select the most important feature (Kong, Nguyen & Xu, 2015).

However, we observed that a very low importance is given to the business attributes of YELP business data-set which strengthen our research. Shellenberger (2017) worked on the business attributes through the selected attributes that have a significant correlation to business star/rating and then applied linear regression in IBM-SPSS (Shellenberger, 2017). Moreover, during the research, we found that there are rich attributes available in the YELP business data-set which can help to predict the business star/rating better.

Methodology

A. Data

As large service provider companies like Amazon are growing their online content very fast, are also very interested in the system, i.e., recommendation system services that enhance the chances of the user to buy a product or avail any service. Not like the online stores or other websites which offer millions of options, users are interested to avail any system which helps them to reduce the options virtually. As this is the era of recommendation system, so these types of content are used intelligently. Our research purpose is also to utilize this type of information significantly to predict the best features which are influenced positively in restaurant growth. In this research, we used the Yelp Data-set of the 12th Round. Yelp is a reviews and ratings platform developed in San Francisco in July 2014 and it is focused on the reviews and rating of local businesses from 32 different countries. Yelp have approximately more than 90 million reviews on different business categories like, restaurants and hotels, and have approximately Mobile visitors are approximately 86 million and around 75 million unique visitors from desktop. The automated software to recommend personalized suggestions which is based on reliable reviews for each visitor is used in the Yelp platform. The data set content like, images, review texts, ratings, locations, collectively make a recommendation and Out of all reviews only 3 quarter of reviews aid in model training program (Maia, Ferreira & Martins, 2019; Rendle, 2010). However, the Yelp Data-set is way vast having six different data files, namely, business, photos, review, tip, checkin and user, but as out research is based on the impact of business attributes, so we worked on the Yelp Business data-set.

The Yelp Business data file is in JSON format and the data looks likes:

{
// string,22characters unique businessId "business_id":"tnhfDv5Il8EaGSXZGiuQGg",

// string,the business's name "name": "Garaje",

// string,the full address of the business "address": "475 3rd St",

// string, the city
"city": "San Francisco",

// string, 2-character state code.
//if applicable "state": "CA",

// string, the postal code "postal code": "94107",

// float, latitude "latitude": 37.7817529521,

// float, longitude "longitude": -122.39612197,

// float,star rating, rounded to half-stars "stars": 4.5,

// integer, number of reviews "review_count": 1198,

// integer, 0 or 1 for closed or open. "is_open": 1,

// object, business attributes to values.
// some attribute values might be objects "attributes": {
"RestaurantsTakeOut": true, "BusinessParking": {
"garage": false,
"street": true,
"validated": false,
"lot": false,
"valet": false

},
},

// array of strings of business categories
"categories": [
"Mexican",
"Burgers",
"Gastropubs"
],

// an object of key day to value hours,
//hours are using a 24hr clock "hours": {
"Monday": "10:00-21:00",
"Tuesday": "10:00-21:00",
"Friday": "10:00-21:00",
"Wednesday": "10:00-21:00",

"Thursday": "10:00-21:00",
"Sunday": "11:00-18:00",
"Saturday": "10:00-21:00"
}
}

B. Data Processing

The data file was in JSON format, so it was required to convert the file into the CSV format. For converting JSON into CSV, we used the Python script. After the conversion, the CSV file had 74 variables with 156,639 observations. Whereas, for this specific research. Whereas, for this specific research, we are only considering the food and restaurant observations, after filtering according to food and restaurant we remained with 23,569 observations.

Moreover, we are focused on the business attributes that have business-stars so that we can analyze the impact of business attributes over the business star. Therefore, we decided to research on the 32 business attributes and business- stars variables. After the selection of business attributes and business stars attribute, we remained with 13,981 observations and then after elimination of null observations gave us 3,431 observations with 32 variables. Additionally, business-stars have 9 levels between the value of 1 to 5. Further, in the data cleaning process, we remained with very low observations in business-stars level of 1, 1.5, 2, and 2.5, therefore we place all these observations in 2. We repeated the same for 4.5 and 5 which placed together in 5. After this process, we arranged our target variable in 5 level factors, i.e. 2, 3, 3.5, 4 and 5.

C. Objective of Research

The main objective of this research was to analyze the business attributes and observe the impact of the attributes on the performance of businesses on Yelp through various models of machine learning. Additionally, through this research we also tried to predict the best possible combinations of business attributes which could help the food/restaurant businesses to analyze their business and grow efficiently. Lastly, we also wanted to investigate the impact of different type of attributes on business-stars ratings.

Implementation

In this section, we have explained the rationale of each model we have used during our research work.

A. Random Forest

We used random forest because it is a combo of multiple decision trees, it is very prominent in both of the classification and regression problems because of its strong and correct predictions (Paing, et al., 2018). Breiman (2001) in his research also explained the features of Random Forest which are very well explained and therefore many researchers used that approaches further research works (Genuer, Poggi & Tuleau-Malot, 2010; Strobl et al., 2008; Breiman (2001).

Brieman (2001), in his research, quoted that “It gives useful internal estimates of error, strength, correlation and variable importance”. Keeping this in mind we used random forest to identify the importance of our attributes to choose the most appropriate features which hits our target attribute. Random Forests are an ensemble of distinct Decision Trees, Gini Importance can be held to analyze Mean Decrease Gini and Mean Decrease Accuracy which is a measure of attributes importance for predicting a target attribute (Naghibi, Pourghasemi, & Dixon, 2016; Kuhn, Egert, Neumann & Steinbeck, 2008; Lee, Kim, Jung, Kim & Kim, 2019; Shuai, Yang, Chen, Yuan, & Song, 2019; Han, Guo & Yu, 2016).

Which is done by Gini Impurity which is a metric used in measure of random forest which is beneficial to minimize when splitting each node. Represents the probability that a randomly selected sample from a node will be incorrectly classified according to the distribution of samples in the node. If all the results in a group from a single category, the Gini Impurity reaches zero (i.e., if there is only one possible attribute in a group, result will be given that label 100% of the time). This measure is the probability of a new result being incorrectly classified at a given node in a Random Forest, based on the training data

The formula for calculating the Gini impurity of a data set or feature is as follows:

$$G(k) = \Sigma P(i) * (1 - P(i))i = 1 \qquad (1)$$

Where P(i) is the probability of a certain classification i, per the training data set.

We implement this in R with Gini-based importance methods. Result are explained in Mean Decrease Gini and Mean Decrease Accuracy

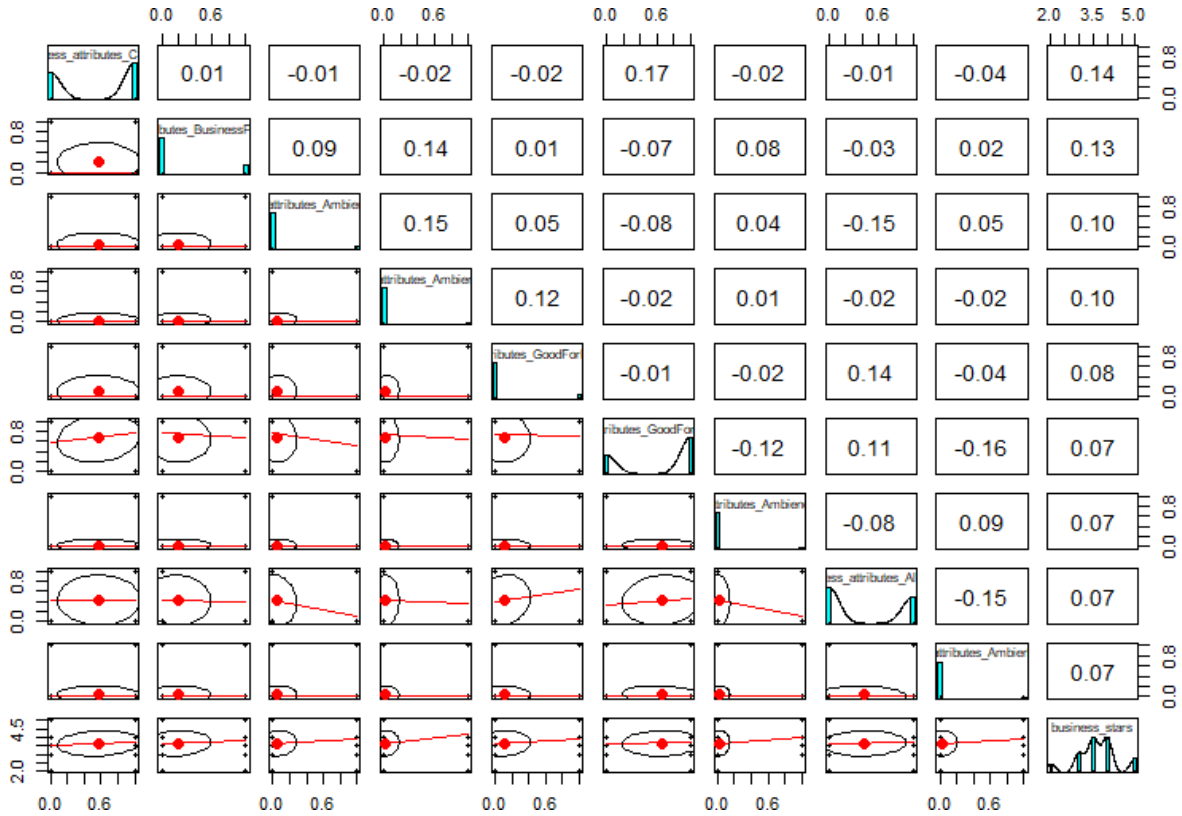


Fig. 1: Result of correlation between business attributes and the target attribute

- Mean Decrease Gini is the average of an attribute's total decrease in node impurity, based on the proportion of samples reaching that node in each individual decision tree in the random forest. A higher Mean Decrease Gini indicates higher importance of the attribute which effectively measuring of the importance of an at-tribute for estimating the value of the targeted attribute across all the trees that make up the forest.

Attributes that are sorted and displayed in the Attributes Importance Graph generate for the Random Forest by this measure. The most important attribute to the model will be highest in the graph and have the largest Mean Decrease Gini values, vis-a-vis, the least important attributes will be lowest in the graph, and have the smallest Mean Decrease Gini values

- Mean Decrease Accuracy attribute basis can be determined out of the bag error calculation phase. Higher accuracy in the random forest can be reduced by excluding a single attribute means the attribute which is more important is required most, therefore attribute with a large Mean Decrease Accuracy are more important for classification of the data.

B. Decision Tree

Decision tree is the best model so far to learn and split the data-sets into smaller subsets, and those smaller sub-sets are used to predict the target attribute. The test in the decision tree is represented as the node and the possible results as the edges of the tree. This process of splitting the subsets remain continue until there is no space for further gain or a present result is met, e.g. the maximum subsets are formed.

In our research, we are undergoing the selection process of attributes where various feature selection methods have appeared highlighting that the use of sets of the decision trees which learnt from a data set could be an important and useful tool for selection of the relevant and informative attributes in line with target attributes.

C. Correlation Coefficient

Correlation coefficient is a technique which shows whether and how strongly pairs of attributes are related to each other. The perfectly correlated features have the correlation equal to 1, and the features will be totally independent, if the correlation is zero. The dependent attributes have almost the same effect and are more linearly dependent when the features have the high correlation.

In our research, we selected the correlation coefficient method for feature selection, because we need to gauge the impact of all the attributes over the target attribute with the help of correlation coefficient. Moreover, correlation coefficient predicts the dependency and influence of one attribute over the other (Hall, 1999).

Table 1: This table is self-explanatory showing the attributes which are selected after applying the feature selection methods

Wrapper	Correlation Coefficient
business attributes Alcohol	business attributes Caters
business attributes Ambience casual	business attributes BusinessParking street
business attributes Ambience classy	business attributes Ambience trendy
business attributes Ambience hipster	business attributes Ambience hipster
business attributes Ambience intimate	business attributes GoodForMeal brunch
business attributes Ambience trendy	business attributes GoodForMeal lunch
business attributes Ambience upscale	business attributes Ambience intimate
business attributes BusinessAcceptsCreditCards	business attributes Alcohol
business attributes BusinessParking garage	business attributes Ambience classy
business attributes BusinessParking lot	
business attributes BusinessParking street	
business attributes Caters	
business attributes GoodForKids	

business attributes GoodForMeal breakfast
business attributes GoodForMeal brunch
business attributes GoodForMeal dinner
business attributes GoodForMeal latenight
business attributes GoodForMeal lunch
business attributes HasTV
business attributes RestaurantsGoodForGroups
business attributes RestaurantsReservations
business attributes RestaurantsTableService
business attributes WiFi

There are a few other researches which are using the Decision Tree model for the feature selection because decision tree is the most important CART algorithm which helps to find the best attributes from our data-sets and to show us the impact of the attributes over the target attributes (Praditya & Saputro, 2019; Wickramarachchi, 2016).

D. Wrapper Method

Wrapper strategies are those strategies of feature selection which are carried out by applying different algorithms. After executing the algorithms, it is one of the best feature for analysis of data-seta. Wrapper method approach is more appropriate method to find the selected features which predicts more accurate results (Kursa, et al., 2010).

Furthermore, Wrapper Method is built-in algorithms, in which we to provide the data-set and these algorithms filters the features from the given data-set. The wrapper method while executing repeatedly and consistently reruns the algorithm and the process runs automatically which further provides the important features of the data-set. The features abstracted from the results of wrapper method will then be used in other models which further provide more accurate and relevant results. Therefore, we selected the wrapper method algorithms for our research work.

Result and Discussion

A. Feature Selection

It is obvious that the feature selection is a major step in any application of machine learning method. In the modern data-sets, there are often too many variables are being described for practical model building, and most of these variables comes out to be irrelevant for the classification. This is also obvious that their relevancy is not known in the start. There are always several disadvantages while working with large feature sets. The two important disadvantage of large features data-set include, the algorithm can be slow down because of the large feature data-sets and if the number of variables is ominously higher than optimal, accuracy is decreased by many machine learning algorithms (Kursa, et al., 2010).

Feature selection is performed through various method, whereas, for this specific research, we used two strong feature selection methods,

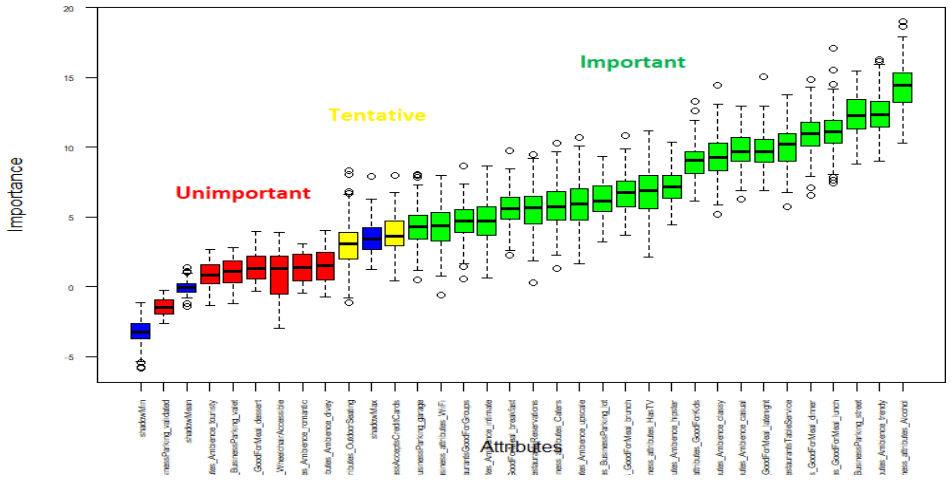
- Filter Method.
- Wrapper Method.

In the Filter Method, we filtered and took only the subset of the relevant features. The model is built after selecting the features. The filtering at this stage is executed through correlation coefficient.

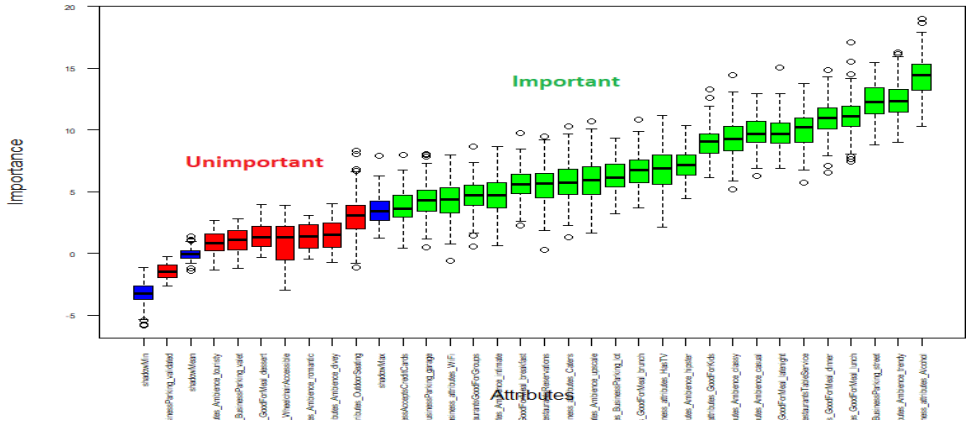
1. Correlation Coefficient

The first method which we used in the Correlation Coefficient with the help of Stephen Turner's two functions "flattenSquareMatrix" and "cor. prob". The flattenSquareMatrix will be used to flatten the combinations from square matrix into a data frame of four columns made from column names, row names, correlation value and p-value, and the cor. prob will be used to create a correlation matrix with the p-values. Result show not very strong correlation between business_stars and business attributes. We cutoff correlation on 0.05 then got 9 attributes greater than 0.05 correlation out of 31 business attributes, Figure 1 show the correlation between the attributes.

The maximum value of correlation is 0.14 which is based on business_attributes_caters; this attribute has the strong influence on business_stars. Afterwards we analyzed that business_attributes_BusinessParking_street is correlated to business_stars. Additionally, business_- attributes_Ambience_trendy, business_attributes Ambience_hipster, business_attributes_GoodForMeal_brunch, business_attributes_GoodForMeal_lunch, business_at- tributes_Ambience_intimate, business_attributes_Alcohol and business_attributes_Ambience classy has the strong influence on business_stars respectively.



(a) Showing the picture of first stage of Boruta's result important attributes, unimportant attributes and tentative attributes



(b) It shows the result after adjusting tentative attributes

Fig. 2: comparative view of Wrapper Method (Boruta function)

B. Wrapper Method

Second one is wrapper method we perform it with Boruta function in R it provides us with a clear view of which business attributes are really important. After running 200 iterations Boruta confirm 22 attributes important 7 unimportant and 2 placed in tentative shown in figure 2. Then we adjust tentative attributes by TentativeRoughFix function so the final confirm important attributes are 23(i.e. business_attributes Alcohol, business_attributes_Ambience_casual, business_attributes_- Ambience_classy, business_attributes Ambience_hipster, business_attributes_Ambience_intimate, business_attributes_- Ambience_trendy, business attributes__Ambience__upscale, business__attributes__BusinessAcceptsCreditCards, business__attributes BusinessParking_garage, business_attributes_BusinessParking_lot, business_attributes_BusinessParking_street, business_attributes_Caters, business_attributes_GoodForKids, business_attributes_GoodForMeal_breakfast, business_attributes_GoodForMeal_brunch, business_attributes_GoodForMeal_dinner, business_attributes GoodForMeallatenight, business_attributes_GoodForMeal_lunch, business_attributes_HasTV, business attributes_RestaurantsGood- ForGroups, business_attributes_RestaurantsReservations, business_attributes RestaurantsTableService, business_at- tributes_WiFi) and confirm unimportant attributes are 7 and this detail is very well visualized in Figure 2. Figure 2a show all attributes with tentative where as in Figure 2b we can see all the important and unimportant attributes. Important point to be noted here is we have all 9 attributes which selected in Correlation Coefficient (like business_attributes_Alcohol, business_attributes Ambience_classy, business_attributes_- Ambience_hipster, etc.) also included in confirm important list of wrapper method, so these 9 attributes are strongly recommended.

Table 2: show the attributes which are selected from feature selection. Correlation Coefficient column attributes are descending order according to the r value.

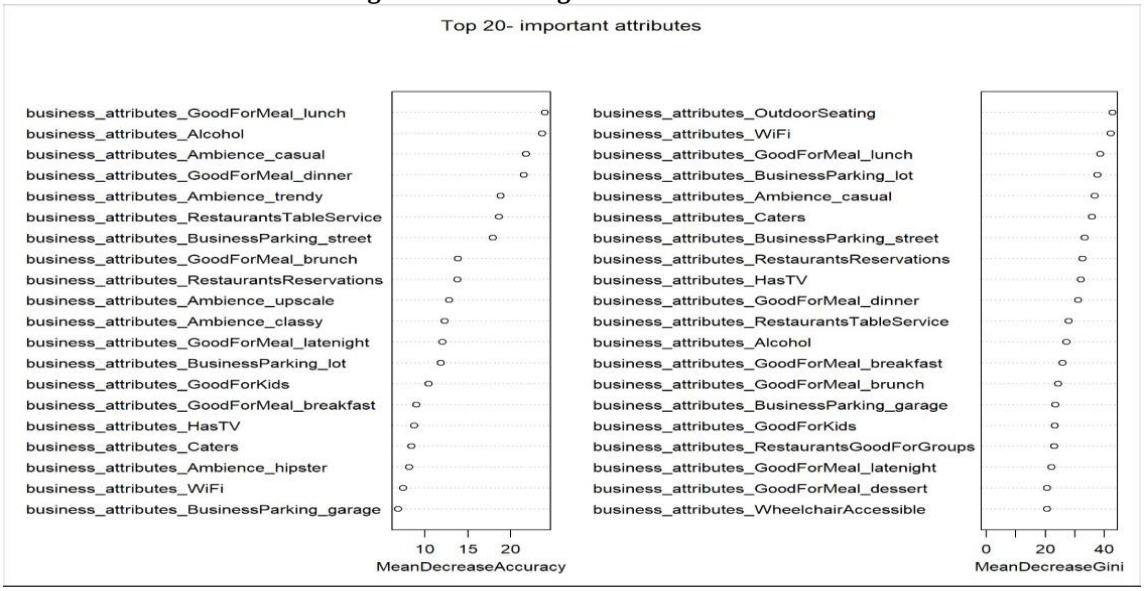


Fig. 3: MeanDecreeaseAccuracy (This graph tests how worse the model performs without each variable). MeanDecreaseGini (This graph measures how pure the nodes are at the end of the tree without each variable)

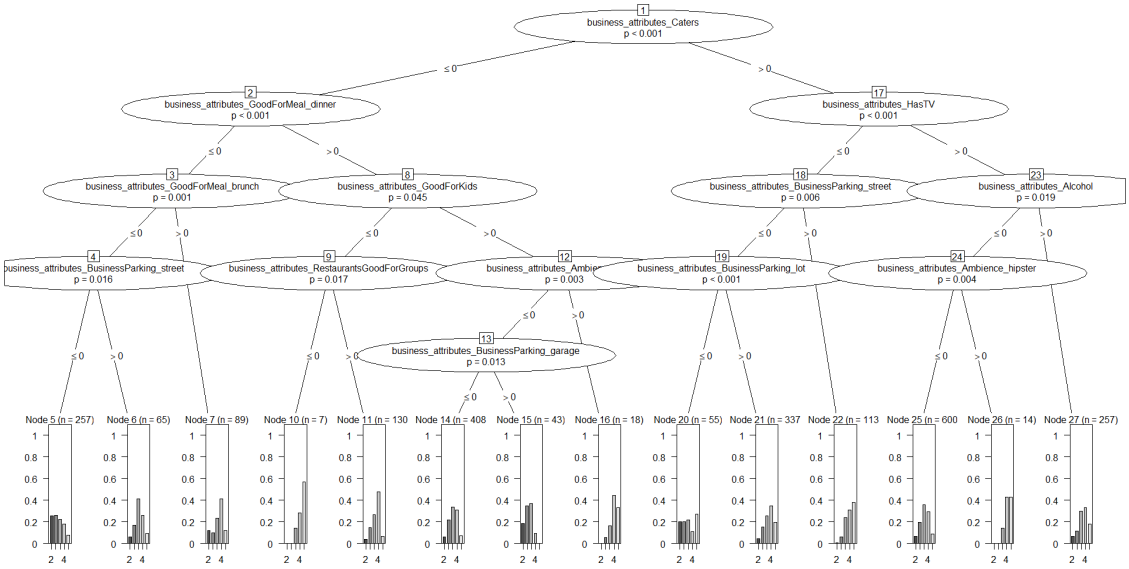


Fig. 4: Decision Tree to determine the best possible combination of business attribute to achieve the good business stars rating.

```
1) business_attributes_Caters <= 0; criterion = 1, statistic = 61.809
  2) business_attributes_GoodForMeal_dinner <= 0; criterion = 1, statistic = 47.862
    3) business_attributes_GoodForMeal_brunch <= 0; criterion = 0.999, statistic = 25.125
      4) business_attributes_BusinessParking_street <= 0; criterion = 0.984, statistic = 19.844
        5)*  weights = 257
      4) business_attributes_BusinessParking_street > 0
        6)*  weights = 65
    3) business_attributes_GoodForMeal_brunch > 0
      7)*  weights = 89
  2) business_attributes_GoodForMeal_dinner > 0
    8) business_attributes_GoodForKids <= 0; criterion = 0.955, statistic = 17.615
      9) business_attributes_RestaurantsGoodForGroups <= 0; criterion = 0.983, statistic = 19.725
        10)*  weights = 7
      9) business_attributes_RestaurantsGoodForGroups > 0
        11)*  weights = 130
    8) business_attributes_GoodForKids > 0
      12) business_attributes_Ambience_trendy <= 0; criterion = 0.997, statistic = 23.476
        13) business_attributes_BusinessParking_garage <= 0; criterion = 0.987, statistic = 20.451
          14)*  weights = 408
        13) business_attributes_BusinessParking_garage > 0
          15)*  weights = 43
      12) business_attributes_Ambience_trendy > 0
        16)*  weights = 18
1) business_attributes_Caters > 0
  17) business_attributes_HasTV <= 0; criterion = 1, statistic = 40.56
    18) business_attributes_BusinessParking_street <= 0; criterion = 0.994, statistic = 22.199
      19) business_attributes_BusinessParking_lot <= 0; criterion = 1, statistic = 28.209
        20)*  weights = 55
      19) business_attributes_BusinessParking_lot > 0
        21)*  weights = 337
    18) business_attributes_BusinessParking_street > 0
      22)*  weights = 113
  17) business_attributes_HasTV > 0
    23) business_attributes_Alcohol <= 0; criterion = 0.981, statistic = 19.569
      24) business_attributes_Ambience_hipster <= 0; criterion = 0.996, statistic = 22.771
        25)*  weights = 600
      24) business_attributes_Ambience_hipster > 0
        26)*  weights = 14
    23) business_attributes_Alcohol > 0
      27)*  weights = 257
```

Fig. 5: Decision Tree view which shows statistics of attributes

C. Random Forest

We use Random forest for train the model. Over all training result is not very significant but it gives a very good image about the business attributes on target attribute. For that we use varImpPlot () and importance () function in Rstudio. These function shows which variable play an important role in the model and this is done by Mean-DecreaseAccuracy (this test is used to identify how the model performs without any single attribute), MeanDecreaseGini (this gives us information about how nodes perform at the end of the tree without any attribute).

Figure 3 elaborates the importance graph of the attributes, which is a critical output of the random forest algorithm. For each attribute in the matrix, it tells how important that attribute is in classifying the data. The graph shows each attribute on the y-axis, and their importance on the x-axis. The attributes are ordered top-to-bottom as most-to-least important. Therefore, the most important attributes are at the top and an estimate of their importance is given by the position of the dot on the x-axis. graph's MeanDecreaseAccuracy section shows the importance of business_- attributes_GoodforMeal_lunch and business_attributes_Alcohol over business_attributes_BusinessParking_garage and its MeanDecreaseGini section show the importance of business_- attributes_OutdoorSeating and business_attributes_WiFi over business_attributes_WheelchairAccessible.

Moreover, the Figure 3 shows the results that is higher values which means that the attributes are more important. In model we take all 31 attributes but random forest model plotted just 20 important attributes, these attributes are important for getting better business_stars ratings. As we can see clearly over here that all these attributes are present in feature selection result. Random forest supports our previous results.

D. Decision Tree

We used Decision Tree Model for analyzing the impact of all attributes on our target attribute. The decision tree model provided us a clear view. With the help of decision tree, we found out the top attributes which have high impact on our target attribute. Moreover, the results of the decision tree are very close to the results of Correlation Coefficient results, which shows that decision tree also support feature selection model.

Figure 4 elaborate that there is one root node (i.e. business_attributes_Carters), and many branches to reach the nodes. However, in this case, we have 14 terminal nodes and each terminal node has a probability distribution of 2, 3, 3.5, 4, and 5 business_stars rating categorical information. Depending on the probability assigned to the leaf, it shows which business attribute or combination of business attributes plays an effective role in increasing the business stars rating. We can clearly see that the tenth terminal node has highest probability of a five-star rating. Function "Minsplit" is used to trim the decision tree, it uses the condition of a minimum criterion of 0.99 Figure 5. Moreover, we can also set" Minsplit" manually. For splitting tree, we must have at least 500 observations and for this research, we had 3,431 observations as previously mentioned. The weight and their values show the performance of attributes within the Model.

The overall results discussion is to identify which attribute or combination of attributes can be a business-focused and help in increasing business_star rating.

- With the above results, we concluded that the business_at- tributes_Caters have the highest impact on the business_- Stars. Additionally, we have observed that business_- attributes_BusinessParking street having relation with business_attributes_Caters also have high impact on the business_Stars. Following, business_attributes_Ambi- ence_hipster with the business_attributes_HasTV having the relation with business_attributes_Caters also have high impact on the business_Stars. Finally, the business_attributes_Alcohol with business_attributes_HasTV having relation with business_attributes Caters is also making an impact on the business_Stars.

- Moreover, business_attributes_Ambience_trendy with the business_attributes_GoodForKids having relation with business_attributes_GoodForMeal_dinner also making a high impact on business Stars.

- We have observed that those businesses, which do not offer dinner and brunch and do not have caters going a way down in their ratings. Additionally, it was also seen that these businesses which are rated low also do not have parking streets as well.

Conclusion

This research aimed to investigate various models of machine learning to determine that which business attributes have the highest impact on business-stars rating and the results can be used to help businesses to improve their ratings. We limited the research to the data from business data-set to raise the control on the problem. Further, we used two most important methods of the feature selection and two machine learning models to determine that which feature/model is best to predict business performance. In the last, we implemented and evaluated various prediction models which clearly high-lighted the effective business attributes. The results also showed us the most effective business attributes (like, BusinessParking Street, Caters, GoodForMeal brunch, etc.) that have strong impact on increasing the business-stars ratings.

References

Andreson, M. & Magruder, J. (2012). Learning from the crowd: Regression discontinuity estimates of the effects of an online review database. *The Economic Journal, vol. 122, no. 563, pp. 957–989*.

Asghar, N. (2016). Yelp dataset challenge: Review rating prediction. arXiv preprint arXiv:1605.05362.

Breiman, L. (2001). Random forests machine learning, vol. 45.

Carbon, K. , Fujii, K. & Veerina, P. (2014). Applications of machine learning to predict yelp ratings.

Chen, Y. & Xia, F. (2020). Restaurants' rating prediction using yelp dataset, in 2020 IEEE International Conference on Advances in Electrical Engineering and Computer Applications (AEECA), IEEE, pp. 113–117.

Elkouri, A. (2015). Predicting the sentiment polarity and rating of yelp reviews, arXiv preprint arXiv:1512.06303.

Fan, M. & Khademi, M. (2014). Predicting a business star in yelp from its reviews text alone. arXiv preprint arXiv:1401.0864.

Genuer, R., Poggi, J.-M. & Tuleau-Malot, C. (2010). Variable selection using random forests, Pattern recognition letters, vol. 31, no. 14, pp. 2225– 2236..

Hall, M. A. (1999). Correlation-based feature selection for machine learning.

Han, H., Guo, X. & Yu, H. (2016). Variable selection using mean decrease accuracy and mean decrease gini based on random forest, in 2016 7th ieee international conference on software engineering and service science (icsess). IEEE, 2016, pp. 219–224.

Hegde, S., Satyappanavar, S. & Setty, S. (2017). Restaurant setup business analysis using yelp dataset, in 2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI). IEEE, 2017, pp. 2342–2348.

Kong, A., Nguyen, V. & Xu, C. (2015). Predicting international restaurant success with yelp.

Kuhn, S., Egert, B., Neumann, S. & Steinbeck, C. (2008). Building blocks for automated elucidation of metabolites: Machine learning methods for nmr prediction, BMC bioinformatics, vol. 9, no. 1, p. 400.

Kursa, M. B. & Rudnicki, W. R. et al., (2010). Feature selection with the boruta package, J Stat Softw, vol. 36, no. 11, pp. 1–13.

Langley, P. et al., (1994). Selection of relevant features in machine learning, in Proceedings of the AAAI Fall symposium on relevance, vol. 184, pp. 245–271.

Lee, H., Kim, J., Jung, S., Kim, M. & Kim, S. (2019). Case dependent feature selection using mean decrease accuracy for convective storm identification, in 2019 International Conference on Fuzzy Theory and Its Applications (iFUZZY). IEEE, 2019, pp. 1–4.

Li, C. & Zhang, J. (2014). Prediction of yelp review star rating using sentiment analysis.
Liu, Z. (2020). Yelp review rating prediction: Machine learning and deep learning models" arXiv preprint arXiv:2012.06690..
Lu, X., Qu, J., Jiang, Y. & Zhao, Y. (2018). Should i invest it? predicting future success of yelp restaurants. in Proceedings of the Practice and Experience on Advanced Research Computing, pp. 1–6.
Maia, R., Ferreira, J. & Martins, A. (2019). Rating prediction on yelp academic dataset using paragraph vectors, Rating prediction on yelp academic dataset using paragraph vectors, pp. 56–62.
Naghibi, S. A., Pourghasemi, H. R. & Dixon, B. (2016). Gis-based groundwater potential mapping using boosted regression tree, classification and regression tree, and random forest machine learning models in iran," Environmental monitoring and assessment, vol. 188, no. 1, p. 44.
Paing, M. P., Pintavirooj, C., Tungjitkusolmun, S., Choomchuay, S. & Hamamoto, K. (2018). Comparison of sampling methods for imbalanced data classification in random forest, in 2018 11th Biomedical Engineering International Conference (BMEiCON). IEEE, 2018, pp. 1–5.
Praditya, N. & Saputro, A. H. (2019). Hyperspectral band selection based on decision tree algorithm in beeswax identification on rome beauty apple, in 2019 International Seminar on Research of Information Technology and Intelligent Systems (ISRITI). IEEE, 2019, pp. 430–433.
Rafay, A., Suleman, M. & Alim, A. (2020). Robust review rating prediction model based on machine and deep learning: Yelp dataset in 2020 International Conference on Emerging Trends in Smart Technologies (ICETST). IEEE, pp. 8138–8143.
Reddy, C. S. C., Kumar, K. U., Keshav, J. D., Prasad, B. R. & Agarwal, S. (2017) Prediction of star ratings from online reviews, in TENCON 2017-2017 IEEE Region 10 Conference. IEEE, 2017, pp. 1857–1861.
Rendle, S. (2010). Factorization machines," in 2010 IEEE International Conference on Data Mining. IEEE, pp. 995–1000.
Shellenberger Jr, P. M. (2017). Predicting yelp food establishment ratings based on business attributes.
Shuai, Y.,, Yang, C., Chen, J., Yuan, C. & Song, T. (2019). A hybrid electronic components detection parameter sequence optimization method based on random forest and mean decrease accuracy, in 2019 IEEE 10th International Conference on Software Engineering and Service Science (ICSESS). IEEE, 2019, pp. 1–4.
Singh, R., Woo, V., Khan, N., Kim, J. Lee, H. J., Rahman, H. A. & Park, J., Suh, J. Eom, V., Gudigantala, N. et al. (2019). Applications of machine learning models on yelp data. Asia Pacific Journal of Information Systems, vol. 29, no. 1, pp. 117–143, 2019.
Stephenturner. explore-correlations.r. [Online]. Availablehttps://gist.github.com/stephenturner/3492773
Strobl, C., Boulesteix, A.-L., Kneib, T., Augustin, T. & Zeileis, A. (2008). Conditional variable importance for random forests," BMC bioinformatics, vol. 9, no. 1, p. 307.
Wickramarachchi, D., Robertson, B., Reale, M., Price, C. J. & Brown, J. (2016). Hhcart: An oblique decision tree," Computational Statistics & Data Analysis, vol. 96, pp. 12–23.
Yelp. Yelp dataset. [Online]. Available: http://www.yelp.com/dataset challenge
Yu, M., Xue, M. & Ouyang, W. (2015). Restaurants review star prediction for yelp dataset. Technical Report 17. UCSD, Tech. Rep.