

Customer Service Recommendation to the Restaurants based on Predictive Analysis

Tejaswini Dhupad
Information Technology
Savitribai Phule Pune University
Pune, India

Ayushi Deshmukh
Business Analytics
Deakin University
Melbourne, Australia

Girimal Dhupad
Computer Science
Savitribai Phule Pune University
Pune, India

Abstract—Over the recent years, using the electronic networks has been increased and the internet has changed into one of the main channels of obtaining the information, hence today the decision making process of the consumer is highly driven by the information that has been catered to them over the internet. To effectively monetize the customer base and maximize revenues, carriers have become increasingly interested in how they can leverage customer analytics to ensure that the best offer is presented to the right customer at the most appropriate time across any channel. Therefore, predictive analysis can be used to identify the service strategy for the new and established restaurants to implement profitable business model.

I. INTRODUCTION

To determine the factors that explain customer satisfaction in the full service restaurant industry, the data was gathered from the Bangalore food assist (BFA) in association with zomato, who was interested in precisely identifying if restaurants should provide online order or book table service to their patrons. BFA has provided with a sample of 40,000 established restaurants data which are represented using address, rate, votes, neighbourhood, cuisines, online_order, book_table etc. The restaurants are rated according to the quality of service they provide and customer satisfaction level. These services are valued by customers and can be used as an added advantage over its competitors. BFA is trying to analyze if online order and book table service should be appropriate for Bangalore restaurants to be inherited in their business model. The findings suggested that the customer satisfaction was greatly influenced by the customer review and other factors (such as cuisines, meal-type, restaurant-type, restaurant name, address etc). These services impact the attractiveness of the restaurant and its customer rating. The study tests the existing model and enhances the literature on to implement profitable business model for restaurants.

A. Duncan's multiple range test

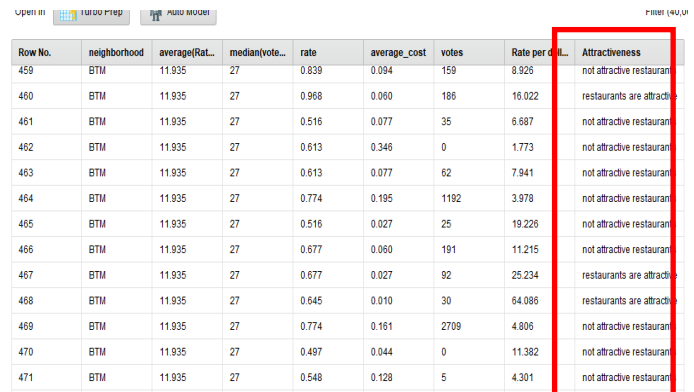
When the analysis of variance test on an accuracy measure produces evidence to reject the null hypotheses, H_0 and H_a , we can accept the alternative hypothesis-that all of the mean accuracies are not equal. However, we still do not know which of the means are significantly different from which other means, so we will use Duncan's multiple range test to separate significantly different means into subsets of

homogeneous means. For the difference between two subsets of means to be significant it must exceed a certain value. This value is called the least significant range for the p means, R_p , and is given by

$$R_p = r_p \sqrt{V/s^2}, \tag{1}$$

where the sample variance, s^2 , is estimated from the error mean square from the analysis of variance, s^2 , r the number of observations (rows), and r_p the least significant studentized-range for a given level of significance, and the degrees of freedom [1].

The below table predicts the classification of attractive restaurants in the given neighbourhood[2]



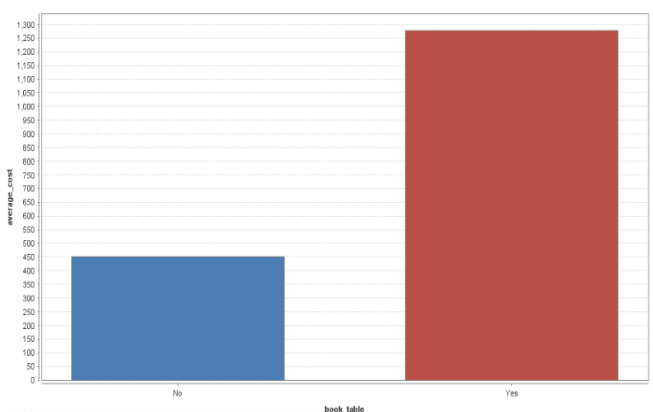
Row No.	neighbourhood	average(Rat...	median(vote...	rate	average_cost	votes	Rate per d...	Attractiveness
459	BTM	11.935	27	0.839	0.094	159	8.926	not attractive restaurant
460	BTM	11.935	27	0.968	0.060	186	16.022	restaurants are attract
461	BTM	11.935	27	0.516	0.077	35	6.887	not attractive restaurant
462	BTM	11.935	27	0.613	0.348	0	1.773	not attractive restaurant
463	BTM	11.935	27	0.613	0.077	62	7.941	not attractive restaurant
464	BTM	11.935	27	0.774	0.195	1192	3.978	not attractive restaurant
465	BTM	11.935	27	0.516	0.027	25	19.226	not attractive restaurant
466	BTM	11.935	27	0.677	0.060	191	11.215	not attractive restaurant
467	BTM	11.935	27	0.677	0.027	92	25.234	restaurants are attract
468	BTM	11.935	27	0.645	0.010	30	64.086	restaurants are attract
469	BTM	11.935	27	0.774	0.161	2709	4.806	not attractive restaurant
470	BTM	11.935	27	0.497	0.044	0	11.382	not attractive restaurant
471	BTM	11.935	27	0.548	0.128	5	4.301	not attractive restaurant

Figure 1: Classification of attractive restaurants in the given neighbourhood

II. DATA EXPLORATION AND PREPARTION IN RAPIDMINER

Based on the business problem identified and stated in executive summary, identified attributes are considered as significant: rate(numerical), votes(numerical), average_cost(numerical), neighbourhood(categorical), online_order, book-table for answering attractive restaurants in a neighbourhood.

The dataset provided by BFA is showing few errors while



predicting the book table attribute(label). As seen from the plot the restaurants providing book table services are too high on average as compared to restaurants not providing the service when plotted against average cost. This implies for unbalanced data in the data-set. To deal with this problem, we have made use of SMOTE operator in the model while predicting the book_table label.

The vote attribute which is used as one of the predictors to identify attractive restaurants in the neighbourhood has few extreme values (i.e. outliers) as clearly seen below.

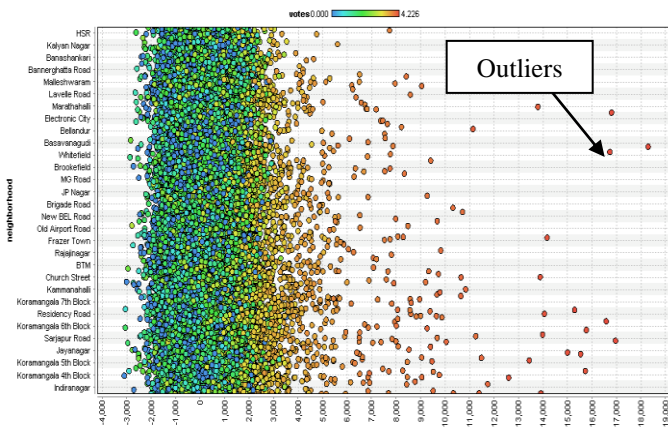


Figure 3: Scatter plot for votes vs neighbourhood

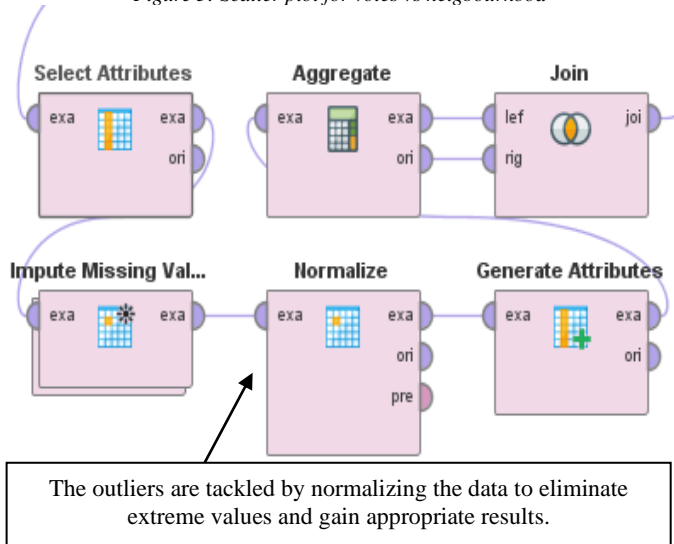


Figure4: Impute missing value and outlier

As seen from the bar chart of the attractive restaurants are found in Koramangala 7th block, BTM followed by Jayanagar. This result was formulated using two key strategies: rate per dollar charge (normalized) attribute must be higher than the average rate per dollar charge and votes gained by the restaurants must be greater than the threshold votes to classify restaurants into “Attractive restaurant” and “Non-Attractive restaurant” within Attractiveness attribute. Missing values are dealt by using impute missing value operator (sub-process KNN schema). Reason behind using KNN schema was to predict appropriate closest k neighbour’s for all missing values.

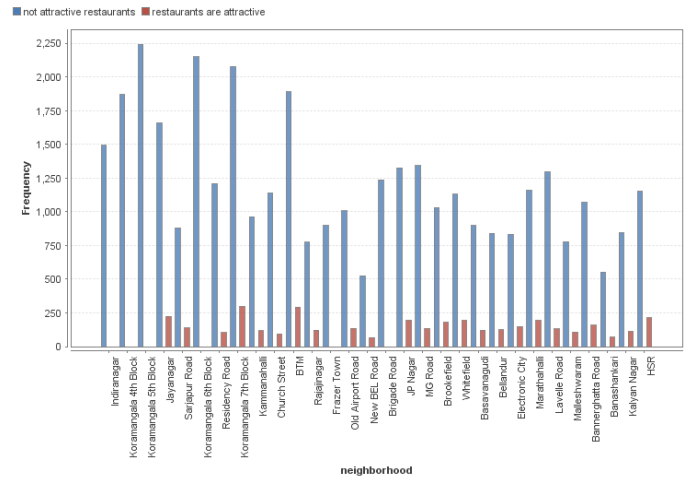


Figure 5: Attractive restaurants in the neighbourhood

III. DISCOVERING RELATIONSHIPS AND DATA TRANSFORMATION IN RAPIDMINER

As per the BFA interest in determining if the restaurants must or must not provide online booking services in Bangalore. Considering - online order and book table as the labels. Out of all the 17 attributes most prominent once are address, name, phone, cuisines, menu_item, (weighted against the labels). Follow below graphs for more insights about attributes relationships:

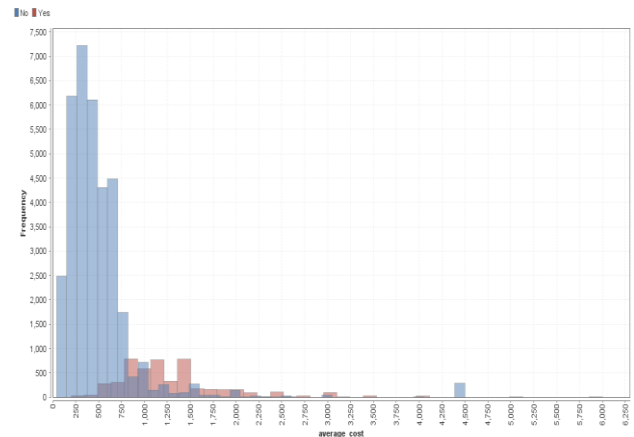


Figure 6: Histogram for featuring relationship between book table and average_cost

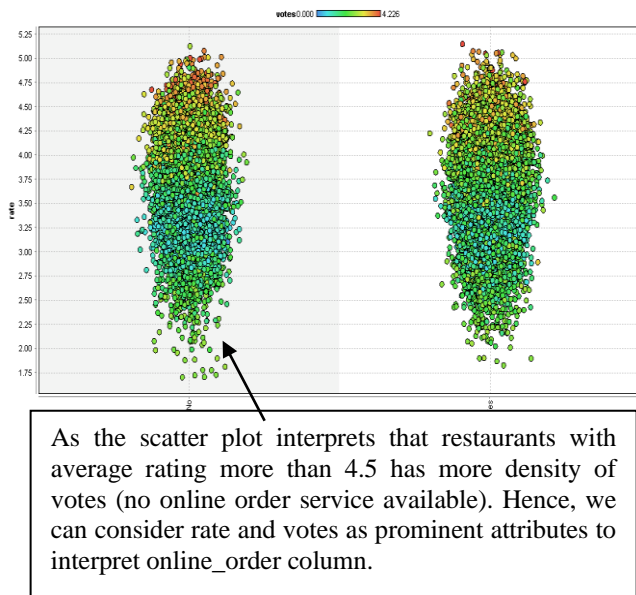


Figure 7: Relationship between online order, rates and votes

The high weights signify that the provided predictor attributes (numerical and non-text) are more important for predicting the label attributes: online order, book table while building the model. In the models, the attributes which weight more than 0.3 are considered as prominent, as they have some relation with the label attribute. Furthermore, the correlation between numerical attribute is also weighted. For example, book_table and average_cost have high correlation between them i.e. 0.617.

Book table		Online order	
attribute	weight	attribute	weight
average...	0.617	average...	0.081
rate	0.426	rate	0.070
votes	0.408	votes	0.035
attribute	weight	attribute	weight
reviews_...	0.590	reviews_...	0.926
phone	0.527	address	0.910
address	0.513	phone	0.900
menu_it...	0.510	name	0.840
name	0.509	dish_liked	0.405
dish_liked	0.483	cuisines	0.334
cuisines	0.371	meal_type	0.062
rest_type	0.234	rest_type	0.049
meal_type	0.063	location	0.037
location	0.061	neighbor...	0.015
neighbor...	0.010	book_ta...	0.000
online_o...	0.000	menu_it...	0

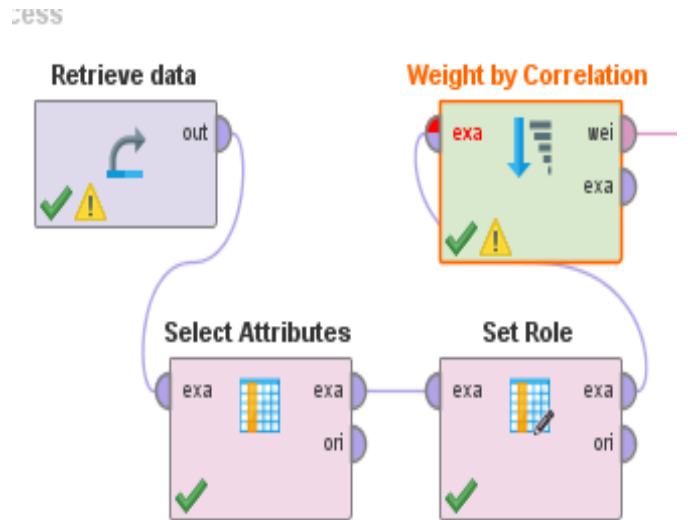


Figure 8: Model to find correlation between numeric variables

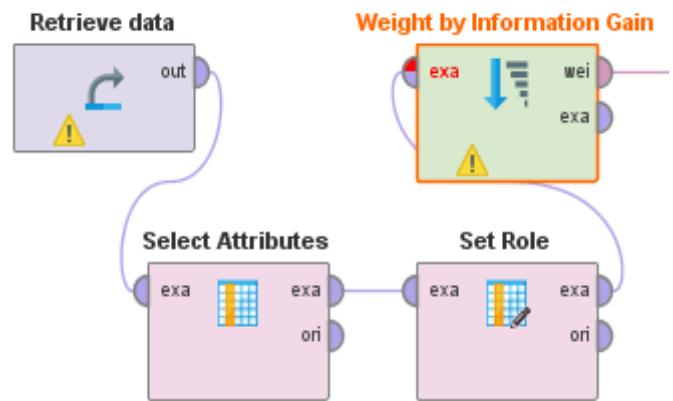


Figure 9: Model for weighing categorical variables

IV. CREATE A MODEL(S) IN RAPIDMINER

A. K-NN Classification:

Based on the Data retrieved from BFA, required predictors are identified by weight (address, average_cost, cuisines, dish_liked, name, online_order, phone, rate, votes). In the designed model missing values of rate and average_cost are taken care by replace missing value operator by averaging the values and role is set as "Online_order" as a label attribute. Sampling type is chosen as stratified sample as the label is nominal attribute, as this will ensure both the classification of online_order are selected proportionally in training and testing data set. The sampling data is split into two segments of 70%(training) and 30%(testing). The model is designed using K-NN classification technique. The distance type and mixed measure are taken as Mixed Measure and Euclidean distance respectively, as it is most optimal and less biased. In K-NN model k value is decided based on accuracy and kappa of the model. The testing data is applied to the model to get a good accuracy and kappa of 93.73 and 0.875 respectively.

SMOTE upsampling is done to handle unbalanced data. It creates a balance between restaurants who provide book-table service and those who doesn't provide. Moreover, it is applied on training data and not on validation data-set, to keep the accuracy end results as practical to true outcome as possible. For SMOTE upsampling the number of neighbourhood parameter is kept as 5.

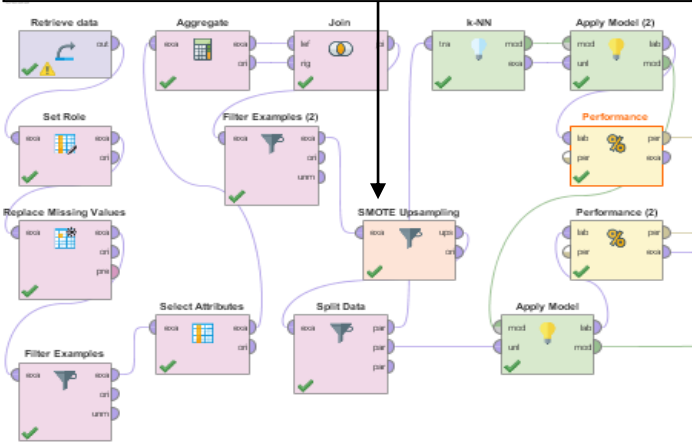


Figure 10: Model for established restaurants for label book tabel.

The value of “k” is calculated using a loop parameter with 10 iterations, best k and kappa value is found, accordingly that k value was picked. As per the observation as the k value increases the accuracy is decreasing but kappa is increasing. As more kappa accounts for occurring of correct predictions by chance. Hence it is one of the robust measure of calculating the performance of the model.

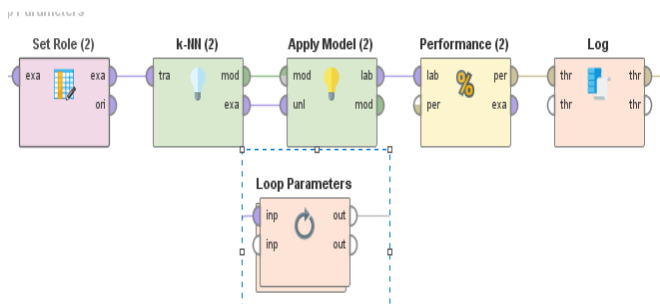


Figure 11: Loop parameter to calculate k value

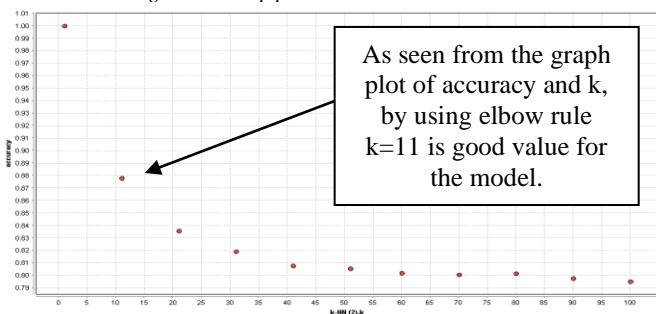


Figure 12: Elbow plot

B. Decision Tree:

To compare the results, alternative model has been built using decision tree, in which restaurants data-set is passed and label is set to “online_order”. We have not dealt with missing values, as the decision tree algorithm is well capable of handling missing values. The data is split into two sub-segments of 70% and 30%. With 70% dataset model is built and the 30% data is passed as a testing data to the trained model to observe and track its performance. The performance parameters were observed to get accuracy and kappa.

Within the decision tree depth parameter is set to 10, as we are limiting the number of nodes in a branch to 10. This is giving us more appropriate results by saving cost as compared to other integer values. However, the pruning parameter is set to 0.1 to get a good fit of training data and testing data as well. If the prune value increases, the tree size decreases and may give us overfitting model which will not be good for testing data.

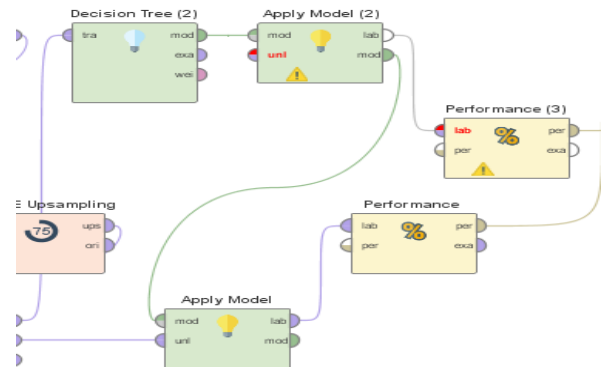


Figure 13: Decision tree model

V. EVALUATE AND IMPROVE THE MODEL(S) IN RAPIDMINER

As it can be analyzed from the table below that decision tree is the best fit as it has comparatively low accuracy and kappa, but it gives more accuracy on validation data. Moreover, K-NN is a lazy classification technique which learnings all its training data. The accuracy is achieved more using cross validation (with number of folds=10).

PARAMETER	MODEL		
	K-NN (k=10)	DECISION TREE	CROSS VALIDATIO (Decision tree)
ACCURACY	90.73%	87.48 %	89.07%
KAPPA	0.775	0.75	0.781
SENSITIVITY	91.58%	91.02%	89.11%
SPECIFICITY	83.88%	83.93%	89.03%

Table 1: Comparison table

Even the accuracy and kappa are increased with the use of cross validation along with it specificity has increased exponentially which means that the data is less biased. Cross validation model sub-process are applied as:

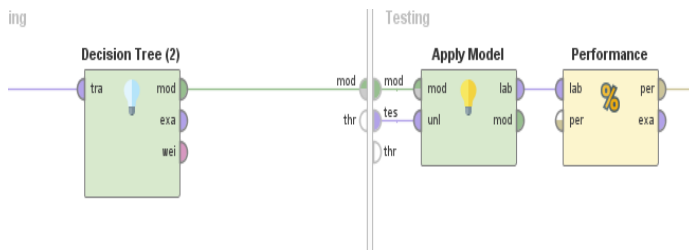


Figure 14: Cross validation model

The precision of the model (one of the other most important validation factor) increases with Cross Validation, this can be analyzed from the confusion matrix as follows:

kappa: 0.750

	true No	true Yes	class precision
pred. No	2983	319	90.34%
pred. Yes	571	3235	85.00%
class recall	83.93%	91.02%	

Figure 15: Decision tree with 70:30 split -Confusion matrix

As it can be analyzed using a confusion matrix of decision tree the predicted no (i.e. book table service is not provided) was 2983 cases and prediction yes (service provided) was 3235 cases when true positive. But, after cross validation values are 15169 and 15182 which are significantly higher. Thus, it is an upgrade performance over already existing models.

	true No	true Yes	class precision
pred. No	15169	1856	89.10%
pred. Yes	1869	15182	89.04%
class recall	89.03%	89.11%	

Figure 16: Cross Validation with decision tree -Confusion Matrix

ROC is a plot which provides insights about stability and feasibility of the model (BRADLEY, 1997)[1]. Higher lift and more area under the curve, better the model. AUC value close to 1 is considered as a good model. Accuracy of the model is low as compared to other still it is a better one because we should not solely focused on gaining high accuracy while checking the performance evaluation table. Performance evaluation other major factors like class

precision, class recall and kappa should also be taken into consideration.

Hence, the cross validation decision tree is best model on considering all these statistics and also after analysing the Below ROC plot:

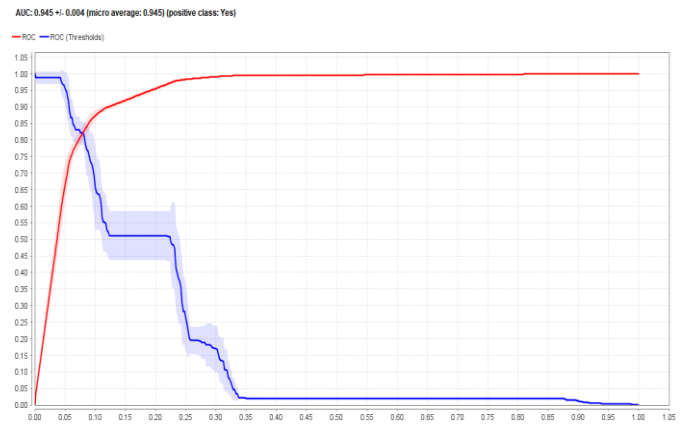


Figure 17: ROC curve and AUC

PerformanceVector

```

PerformanceVector:
accuracy: 89.07% +/- 0.52% (micro average: 89.07%)
ConfusionMatrix:
True: No Yes
No: 15169 1856
Yes: 1869 15182
kappa: 0.781 +/- 0.010 (micro average: 0.781)
ConfusionMatrix:
True: No Yes
No: 15169 1856
Yes: 1869 15182
AUC: 0.945 +/- 0.004 (micro average: 0.945) (positive class: Yes)
sensitivity: 89.11% +/- 0.82% (micro average: 89.11%) (positive class: Yes)
ConfusionMatrix:
True: No Yes
No: 15169 1856
Yes: 1869 15182
specificity: 89.03% +/- 0.87% (micro average: 89.03%) (positive class: Yes)
ConfusionMatrix:
True: No Yes
No: 15169 1856
Yes: 1869 15182
    
```

Figure 18: Performance of cross validation decision tree.

VI. DEPLOYMENT IN RAPIDMINER

For deploying the model and validating it in the Zomato score data, following steps needs to be implemented in RapidMiner:

- Use read CSV operator in a new process. By using import configuration wizard, read the csv file that needs to be scored.
- Check missing values in data and ensure that data doesn't contain any missing values.
- Apply all the pre-requisites applied on the training data.
- For applying the entire constructed model, use apply model operator.
- Use write model operator in the process to back the scored data.
- Connections: Connect the Read CSV to input of Apply Model operator.
- Connect apply model to write model.
- Run the model.

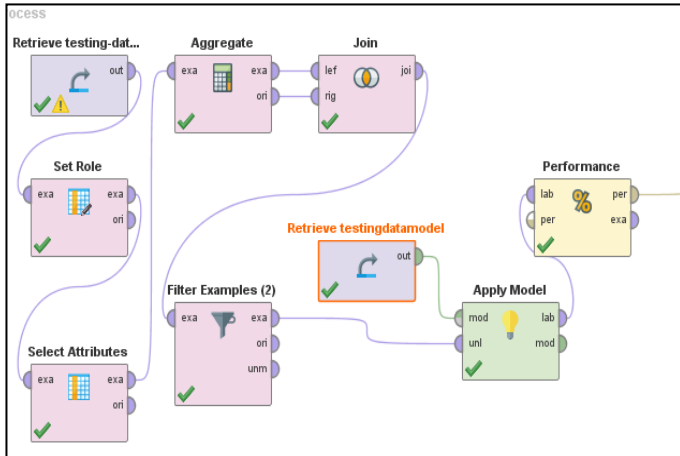


Figure 19: Honest testing using test data.

The honest testing is working good on build model as it is giving us 85.33% accuracy and 0.667 kappa.

```

PerformanceVector:
accuracy: 85.33%
ConfusionMatrix:
True:  No   Yes
No:    1979  110
Yes:   370   814
classification_error: 14.67%
ConfusionMatrix:
True:  No   Yes
No:    1979  110
Yes:   370   814
kappa: 0.667
ConfusionMatrix:
True:  No   Yes
No:    1979  110
Yes:   370   814
    
```

Figure 20: Performance Vector

VII. CONCLUSION

As asked from BFA end that what strategy must be implemented by new restaurants to provide online order and table booking services.

To do so, we have analyzed the established restaurants whose votes are more significant than the threshold (i.e. 24).Have

even considered attributes such as address, name, phone, cuisines and discarded customer review attributes because, it will help in identifying the restaurants which are well established and provide good customer satisfactory services such as online order and book table. Furthermore, it will also provide insights about strategic approach towards implementing these services — thus giving a competitive edge over the rivals. This strategy is beneficial and can be achieved only by new restaurants which have crossed the threshold vote parameter bar. As these services are more aligned towards customer satisfaction, the new restaurants must only take these services when they have established themselves in the industry and are ready to compete with the old restaurants.

Another BFA’s query was regarding the strategy for already established restaurants:

By analyzing the data, the classification of date is being done based on the customer ratings. Restaurants which have more average rating are doing good in their business as compared to restaurants with low rating. The Lower rated restaurants provide basic online order or book table services but the services are not up to the mark, as they are facing late delivery or wrong food parcel getting submitted. Hence, restaurants with less average rating should focus more on solving the above problem while providing online delivery and book table services as they will built a good customer review platform and can compete with their competitors in terms of more valued services. This strategy is beneficial only for established restaurants who haven’t marked the bar of good ratings and are willing to expand their business.

REFERENCE

- [1] BRADLEY, A. E., 1997. The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pergamon, Volume 30.
- [2] Hoffman, K., Kelley, S. and Rotalsky, H. (1995), "Tracking service failures and employee recovery efforts", Journal of Services Marketing, Vol. 9 No. 2, pp. 49-61
- [3] Syed Saad Andaleeb and Carolyn ConwaySam and Irene Black School of Business, Penn State Erie, The Behrend College, Erie, Pennsylvania, USA "Customer satisfaction in the restaurantindustry: an examination of thetransaction-specific model".