Turkish Journal of Forecasting vol. 05 no. 1 (2021) pp. 01-11

Content list available at JournalPark

Turkish Journal of Forecasting

Journal Homepage: tjforecast.xyz

Application of Random Forest Algorithm for the Prediction of Online Food Delivery Service Delay

H. Şahin[1,*], D. İçen[1]

[1]*Hacettepe University, Faculty of Science, Department of Statistics, Beytepe Campus, 06800 Ankara, Turkey*

A R T I C L E I N F O

Article history:

Received	17	December	2020
Revision	10	January	2021
Accepted	19	January	2021
Available online	29	August	2021

Keywords:

Ensemble learning
Random forest
Online food ordering

R E S E A R C H A R T I C L E

A B S T R A C T

Online shopping industry nowadays has been growing rapidly with the evolution of technology. Consumers have started to shop according to certain criteria with the spread of the online shopping sector. One of the sectors that enlightens the future customer in terms of service quality by getting feedback from purchases (comments or ratings) is the online food sector. In this study, a classification study is conducted to investigate the observance with the fast delivery criteria, which is one of the cornerstone criteria in the online food industry. Random Forest (RF) algorithm is applied for the classification. The most important advantage of the RF is it handles a large number of input variables also it is speediness. More than that RF algorithm reduces the overfitting problem and as a result variance is small and therefore it improves the accuracy. The application is implemented through R programming language. In this study, online food delivery variable is created as two categories (On time or Early and Late) and is estimated by RF algorithm that is applied this data for the first time. According to the results, the correct classification rate of the testing data for the estimation of the online food delivery status variable is found as 95.85%. In addition, the performances of the restaurants are compared for the customers. It turns out that the traffic situation does not greatly affect the result of the delivery status. As a result, RF algorithm is applied to the data obtained by web scraping techniques and the delivery status performance of restaurants is revealed with this study.

1. Introduction

The increase in internet usage has leaded up the way for online shopping with the development of technology. Online shopping is used frequently today. Furthermore, this type of shopping opportunities can reach to the consumers in the easiest way with different categories such as food, clothing and cosmetics wherever they want and in every time of the day. As time passes, the online shopping sector is growing rapidly and competition between companies is increasing speedily. Therefore, consumers attach importance to shopping according to certain criteria such as: product and service quality, fast delivery, reliability and user satisfaction.

People's eating needs with online food ordering has also affected the online food industry as it is practical and affordable. For this reason, fast delivery and product quality in online food ordering are effective in consumers' food ordering.

* Corresponding author.
E-mail addresses: hanifesahinnnnn@gmail.com (Hanife Şahin), duyguicn@hacettepe.edu.tr (Duygu İçen)

https://doi.org/10.34110/forecasting.842180

Ensemble learning is a method that combines the results of multiple learners for more accurate results. The Random Forest (RF) algorithm, which is one of the ensemble learning methods, produces trees with the feature of randomness by selecting samples from the data with the bootstrap technique. With the trees created, the forest is obtained and the decisions taken by each tree are combined and concluded. Random forest is a preferred algorithm because it is very strong against over learning and missing values in data. The importance of variables can also be calculated with this algorithm.

This study aims to see the competence of popular food brands for fast delivery in online food ordering. For this purpose, it identifies brands that are good at fast delivery. Moreover, important variables that effects online food ordering quality is identified by the help of RF.

2. Literature Review

Currently there are many scientific papers assessing in many scientific areas with RF which is one of the ensemble learning methods. In this section, RF applications for various fields are summarized and the advantages of RF in these studies are briefly explained.

Akman et al. applied random forest, bagging and Classification and Regression Tree (CART) methods to a data obtained from the health field and obtained the highest accuracy rate as 95.4% with random forest [1].

Akar and Güngör applied RF, Gentle AdaBoost (GAB), Maximum Likelihood Classification (MLC) and support vector machine (SVM) methods to the data related to satellite images. Random forest showed a higher classification accuracy than other methods [2].

Özdemir used the RF algorithm for the process of obtaining potential distribution maps. The AUC value of the potential distribution model obtained by the random forest method was determined as 97.8% [3].

Kalaycı compared the multilayer perceptron, support vector machine, decision tree, k-nearest neighbour, Naive Bayes and RF using a data set consisting of 1353 samples containing 9 different features to predict whether a website has an identity thief. The RF algorithm was found to be more successful than other methods [4].

Irmak and Aydilek used RF, decision tree, support vector machine, k-nearest neighbour, artificial neural network, linear, heap, harmonious enhancer, inclined enhancer and sampled total regression methods in order to estimate Adana's air quality index correctly in their study. The method that best estimates the air quality index was found to be random forest [5].

Canaz and Sevgen reached 70.2% correct classification rate by applying the RF to the light detection and change data of İzmir Bergama district [6].

Çömert et al. used RF algorithm for the data of Adrasan and Kumluca fires between 24-27 June 2016 for mapping the burnt forest area. The overall accuracy rate was found to be 99% [7].

In Ünlü's study, support vector machine (SVM), RF, bagging and decision trees methods were applied to classify historical coins. RF showed a higher performance than other methods with 71% correct classification rate [8].

Ekelik and Altaş created a classification model by applying RF to user data obtained as a result of digital advertising broadcasts of a construction company. The correct classification rate for the created classification model was found to be 82%, and the AUC value was found to be 66% [9].

Akın and Terzi applied cox regression and RF to the data obtained from leukemia patients hospitalized in Ondokuz Mayıs University Chest Diseases Department to determine the mortality risk of leukemia patients. It was concluded that the RF can be used as an alternative to Cox regression [10].

Baba and Sevil compared RF and robust regression methods to estimate the initial returns of public offerings published on Borsa Istanbul. Prediction results showed that the RF achieved better prediction accuracy than other methods [11].

3. Data Set

In this section the detailed information about the data set used in this study is given. Firstly, this data is built for the work titled: “A Web Mining Approach to Collaborative Consumption of Food Delivery Services” which is the official institutional research project of professor Juan C. Correa at Fundación Universitaria Konrad Lorenz [12], [13]. The data is also associated with the work titled as “Evaluation of Collaborative consumption of food delivery

services through web mining techniques" [14]. The raw data is downloaded from Mendeley repository [13]. The importance of data consists of several important techniques to consider. Firstly, this data is a novel set of records that consists the online food delivery service records with the local traffic conditions that are captured from Google Maps API. Secondly, data is collected with the help of web scraper technique.

The data set is the sample of 787 restaurants and 4296 customers in Bogota city. Besides, it includes the key performance indicators of restaurants and their traffic descriptions, as captured by Google Maps with the traffic conditions taken three times on Saturdays at rush hours [12], [13]. The variable names and their explanations used in the analysis are given in Table 1.

Table 1. Variables and their explanations

Variable	Explanation
Moment	It has three possible values "Morning", "Noon" and "Afternoon" that describes the captured moment by Google Maps API on three different times during the day.
Name of Provider	It is the commercial name of each restaurant.
Number of Comments	It describes the total number of customers who ordered service from that restaurant.
Minimum Charge Ordering	This value represents the minimum Colombian pesos that required for providers to deliver their orders to customer.
Cost Delivery	It is the amount of Money in Colombian pesos that is need for the delivery of the food from restaurant to customer.
Distance	This is the distance that is between restaurant and the address of the customer that is measured as meters.
Typical Traffic Afternoon	It represents the traffic conditions around each restaurant on rush hours on afternoon. (G: Free traffic, O: Average traffic, R: Heavy traffic)
Typical Traffic Noon	It represents the traffic conditions around each restaurant on rush hours on noon. (G: Free traffic, O: Average traffic, R: Heavy traffic)
Typical Traffic Morning	It represents the traffic conditions around each restaurant on rush hours on morning. (G: Free traffic, O: Average traffic, R: Heavy traffic)
DTF_catg	It has two possible values "Late" and "On time or Early" describing the delivery performance of restaurant.

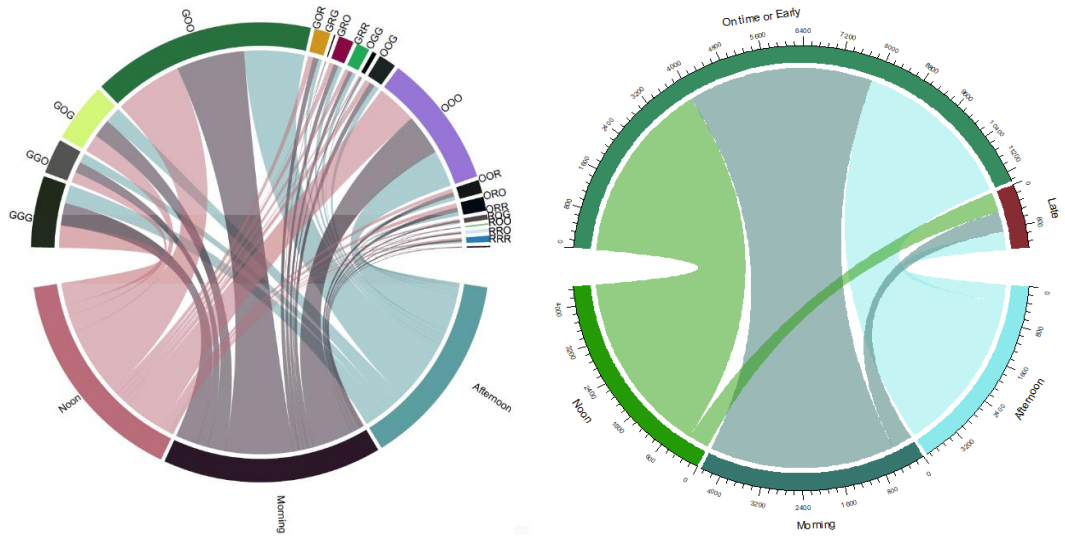
The variable "Delivery Time Fullfilment (DTF)" is created in order to compare the efficiency of online food service of popular restaurants in Bagota city [12], [13]. This variable represents the difference in seconds between restaurants own declared delivery times and expected travel time given by Google Maps API as recorded.

We brought a different perspective to this recorded value and divided it into two categories as "Delivery is Late (Late)" if the DTF variable is less than zero, and "Delivery arrived early or just on time" if it is equal to or greater than zero. By doing so, we intend to identify the performances of popular restaurants that deliver the fast food on time or early. Moreover, the effects of traffic conditions about the delivery result is examined. In addition to this, we want to suggest consumers on which restaurant to choose by taking into account the time ordered time in the day for the most possible early delivery.

In order to summarize the data and see its internal dynamics, the graphs below are drawn and their details are explained. There are two chord diagrams in Figure 1 as labelled (a) and (b). First of all, the following chord diagram in Figure 1 (a) shows to see the traffic density in the determined time periods of the day (Morning-Noon-Afternoon). To draw this graph, a new variable (TT_MNA) is created by combining three typical traffic variables.

For example, OGR represents the daily traffic changes from Orange in the morning to Green at noon and Red in the afternoon. Thus, it is seen that the daily traffic situation is mostly observed as GOO and OOO. It indicates that the intensity of these traffic situations is observed at about the same frequency in the morning noon and afternoon. The second chord diagram in Figure 1 (b) indicates that the majority of customer orders arrive on time or early. It is seen that there is no noticeable difference regardless of whether the delivery is taken place in the morning, noon, or evening for the cases where the customer order arrives late.

The Sankey diagram is another popular type of chart used to visualize the flows and frequencies of variables proportionally. The width of the lines is used to show the magnitude of the densities; therefore, the larger the line, the greater the observed frequency. In addition, it is used to see the flow of the value to be investigated within the data. In order to examine the delivery status of food (DTF_catg), time of day (Moment) and traffic situation (TT_MNA) together the the sankey diagram is drawn in Figure 2. This figure actually gives a nested and compact view of the cases (a) and (b) given in Figure 1.



(a) (b)

Figure 1. Chord diagram of the variables (a) Traffic Density in the rush hour and Time periods of day (b) Delivery performance and Time periods of day

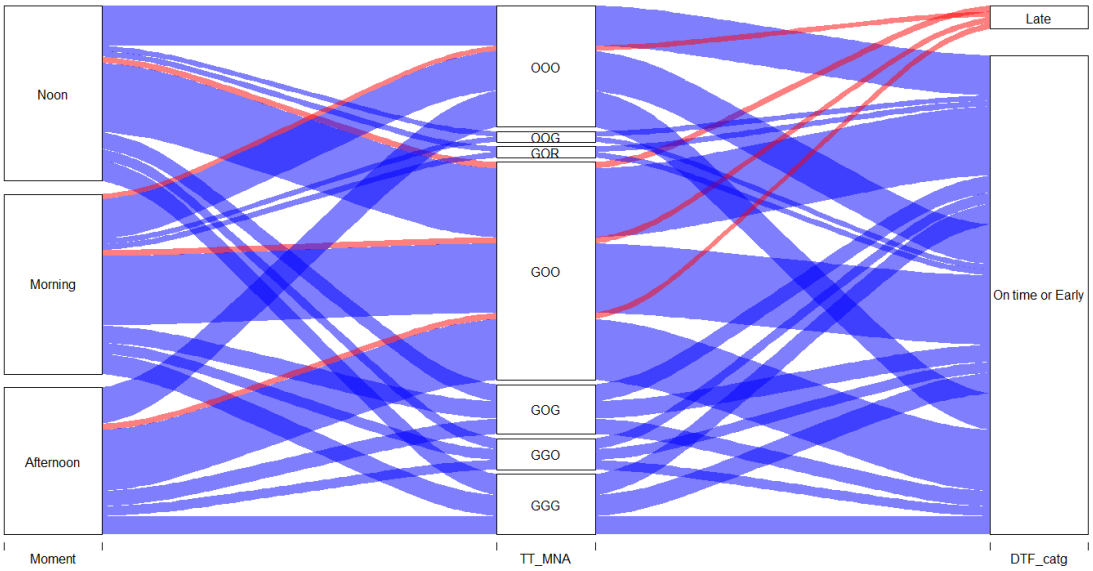


Figure 2. Sankey diagram of the variables moment, traffic density and food delivery performance

Figure 2 indicates the time of day (Moment) and the traffic situation (TT_MNA) together in this specialized situation when the delivery is late (DTF_catg = Late). This figure focuses on the late food delivery situation and this is indicated by the red lines in the sankey diagram. This indicates that the delay in food delivery is mostly in the case of OOO and GOO situations of the traffic density, but this delay has equal intensity at the specified times of the day.

4. Ensemble Learning

In this section, the RF algorithm which is one of the ensemble learning methods for classification is explained in detail. Ensemble learning is a learning method that enables creating a model with more than one learner, rather than training a model with one learner. Hence, it aims that the models will make the right decision together [15]. There are a variety of ensemble learning methods used in the literature. Since we focused on RF in our study, RF algorithm is detailly explained in this section.

4.1. Bagging

Bagging method is developed in 1996 by Leo Breiman. This method retrains the basic learner by deriving new training data from an existing training data. The purpose of the bagging method is to create differences by deriving

new data sets and thus increase the success of total classification. Each training dataset is applied to the same basic learner and the decisions taken are combined by weighted voting method [16].

The main steps of bagging method are given below:

- Multiple subsets are created from the original data set.
- A basic model (weak model) is created in each of these subgroups.
- Models run in parallel and independently from each other.
- Final estimates are determined by combining estimates from all model [17].

Bagging algorithms include Bagging meta-estimator and Random Forest that are the most frequently used popular algorithms.

4.2. Random Forest

The RF was developed as a new method in 2001 by Leo Breiman [18]. In this method, the Bagging method developed by Leo Breiman in 1996 and the Random Subspace technique used to select random subgroups proposed by Tin Kam Ho in 1998 were combined. A random forest can be defined as a collection of tree type classifiers. It is an improved version of the Bagging method by adding the randomness feature [19]. Rather than branching each node to branches using the best branch out of all the variables, the random forest branches each node using the best of the randomly selected variables in each node. Each new training data is generated by bootstrap method from the original training data. Trees are then grown using random feature selection. Developed trees are not pruned [20], [19]. The random forest uses the CART algorithm to generate trees. The random forest is also very fast, resistant to over-adaptation, and works with as many trees as desired [21].

The RF is based on two parameters. These parameters are the number of trees to be created (N) and the number of variables that will be randomly selected in each node (m). The p value represents the total number of predictive variables, and the assumed m value was proposed as p/3 when constructing regression trees and $\sqrt{p}$ when constructing classification trees [1].

The idea in the random forest is to improve the variance reduction of the Bagging method by reducing the correlation between trees without increasing the variance too much. This is achieved by random selection of input variables in tree growth [22].

5. Results

The 4.0.2 version of the R program is used in the study. A random forest algorithm is applied to the data mentioned in Section 3. Missing observations are deleted during the data preprocessing phase. The random forest algorithm is not sensitive to outliers. Therefore, it is seen that the outliers do not affect the correct classification rate and the outliers were not deleted from the data. Numeric variables in the data are normalized. Then the data is divided into two as 80% training and 20% test data.

The number of trees (N), which is one of the parameters required for the application of the method, is selected as 500. The number of randomly selected variables (m) in each division is determined as 3 by taking the square root of the number of variables in the data.

Figure 3 shows the error rate depending on the number of trees. The optimal number of trees can be chosen as 500. Because the high number of trees made the forest more stable, although it did not cause a much decrease in the error rate.

As seen in Figure 4, it is determined that the number of randomly selected variables (m) in each division in the case of the least test error is 3. Here the parameter expressed as m is expressed as mtry in the R program.

After determining the optimal values for the parameters, the algorithm is applied again to estimate the target variable (DTF_catg) in the data using the specified parameters. Table 2 gives the confusion matrix of training data.

Confusion matrix of the training data is given in Table 2. The correct classification rate for training data is 96.24%, however true negative rate is 7.35%. This is because of the imbalanced data. Random forest algorithm tends to classify majority class with a more accurate percentage than the minority class. Some other algorithms are proposed

for eliminating this problem, however it is demonstrated that there is not any distinct winner[23]. In addition to this, Kappa statistic is calculated and confirming that the fit is good with 77.58%.

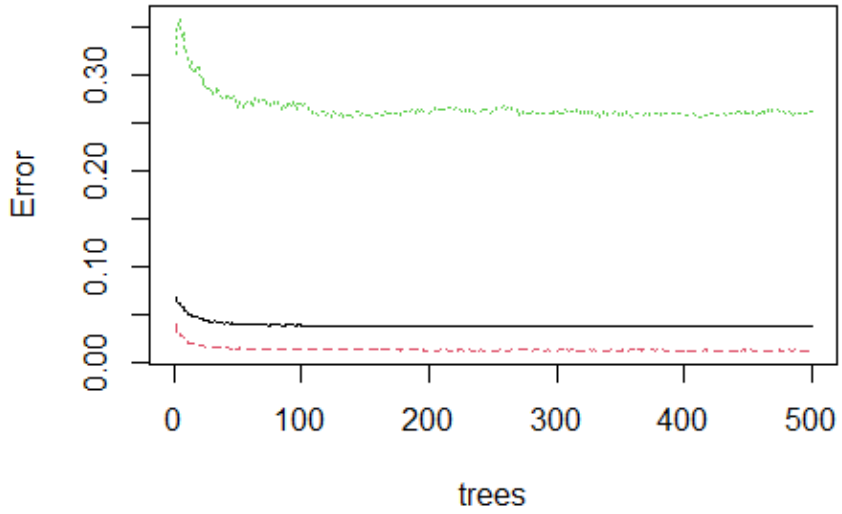


Figure 3. Error rate of random forest algorithm depending on the number of tree

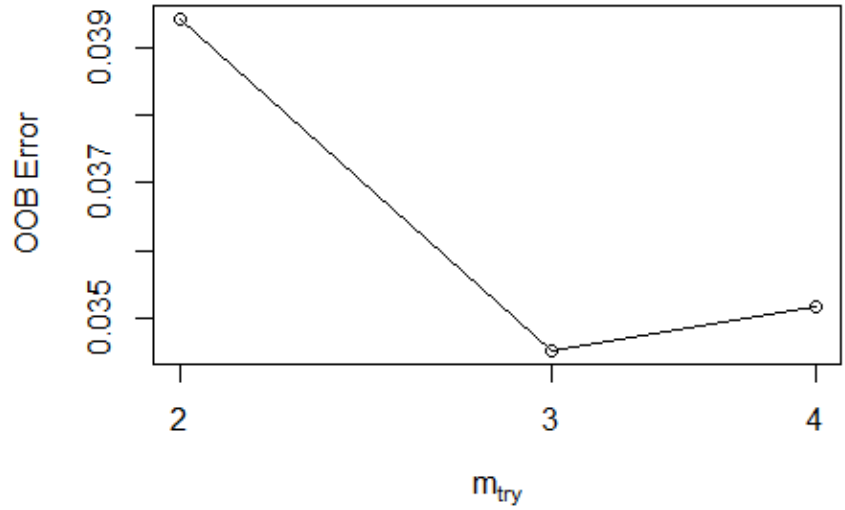


Figure 4. Number of randomly selected variables in each division during the composition of tree

Table 2. Confusion matrix of train data

Predicted	Actual	
	On Time or Early	Late
On Time or Early	9001 (88.88%)	265 (2.62%)
Late	116 (1.14%)	745 (7.35%)

Figure 5 (a) shows that the variable Distance is the most important variable for the dependent variable. The other variables that follow this variable in terms of importance are Number of comments, Name of the provider, Minimum charge ordering and Cost delivery, respectively. Furthermore, Traffic conditions seemed not to have big effect on the prediction of food delivery performances. On the other hand, Figure 5 (b) measures the purity of the classification results obtained at the end of the tree in the absence of any variables. In other words, when one of the variables leaves the model, it shows the decrease that may occur in the Gini coefficient. According to this graph, the distance variable is the variable that makes the biggest contribution to the classification result. In addition to this it is clearly seen that the contribution of other variables to the classification result is close to each other and significantly higher.

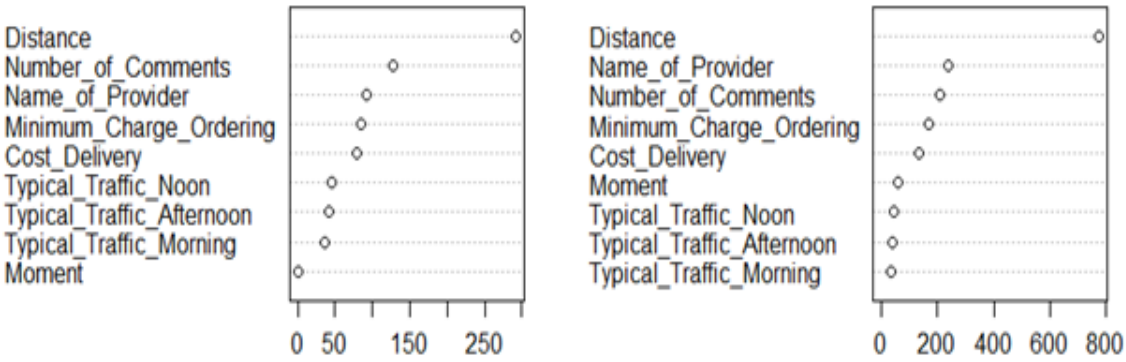


Figure 5. Variable performances in RF (a) Contribution of variables to accuracy (b) Mean Decrease in Gini Value

The confusion matrix of test data is given in Table 3. The correct classification rate for the test data is found to be 95.85%. In addition to this, Kappa statistic is calculated and confirming that the fit is good with 74.58%.

Table 3. Confusion matrix of test data

Predicted	**Actual**	
	On Time or Early	**Late**
On Time or Early	2253 (88.98%)	79 (3.12%)
Late	26 (1.02%)	174 (6.87%)

Cross Validation and Bootstrap methods are applied to the data to evaluate the model performance in more detail. Results are given and explained in Figure 6 and Figure 7, respectively.

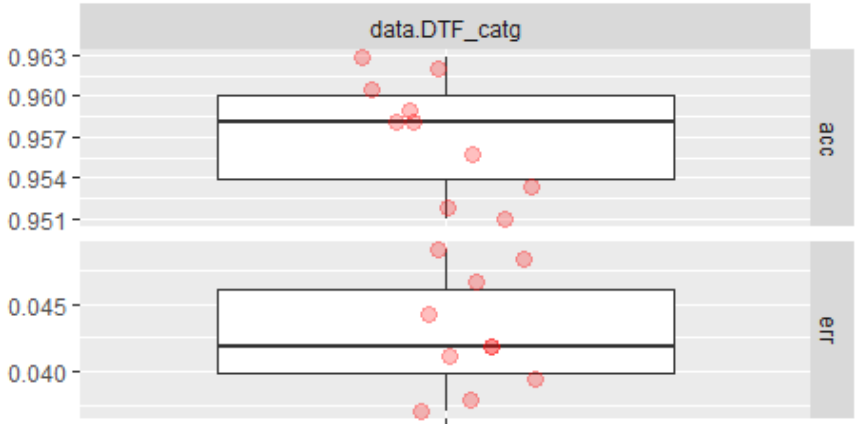


Figure 6. Cross validation

Figure 6 shows the accuracy and error values of the Cross Validation method. It is seen that the correct classification rate for the Cross Validation method is 95.72%, and the error rate is 4.28%.

Figure 7 gives the accuracy and error values of the Bootstrap method. It is seen that the correct classification rate of the Bootstrap method is 95.78%, and the error rate is 4.22%.

The above analysis results show that the correct classification rates found by performance measurement methods coincide with the correct classification rate found with the RF algorithm. After these values are obtained ROC curves are drawn for the performance research.

In Figure 8 given above, the red curve indicates On Time or Early delivery and the green curve indicates Late delivery. The AUC value is found to be 0.9733 for two of the categories of response variable. This result indicates that the model performance has a high successful performance in predicting both categories of the two-category target variable.

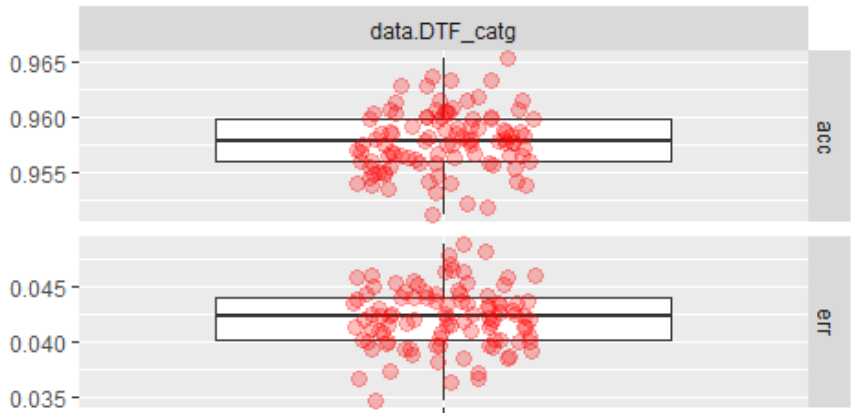


Figure 7. Bootstrap

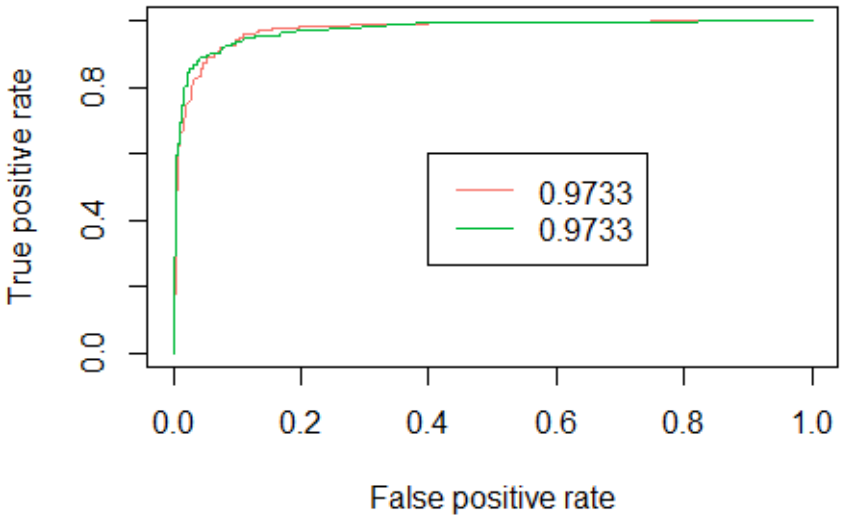


Figure 8. ROC curves

After the application of RF algorithm to online food delivery services data, the providers' performances are given according to the delivery result with Table 4 below.

The most ten frequent provider names preferred by customers are selected and the delivery performances of these restaurants are shown in Table 4 according to RF algorithm. Generally, all restaurants here have a high on time or early delivery percentage. Also, Cali Mio, Donde Beto El Original, Jenos Pizza, La Monapizza and Tacos & Bar BQ providers can be chosen for early delivery by customers who needs online food delivery services. In addition, Subway restaurant has a higher percentage in terms of late delivery than other providers.

The percentage of on time or Early deliveries and Late deliveries occurring in the time frame of the day are given in Table 5 according to RF algorithm.

Table 5 shows that online food deliveries made during the day at noon time are late. In addition, it is seen that On Time or Early online food delivery has with the highest percentage in the Afternoon time frame. Table 6 shows the time frame for which restaurant will deliver the online food On Time or Early or Late.

Table 4. Restaurants and food delivery result

Name of Provider	Food Delivery Result	
	On Time or Early	Late
Cali Mio	48 (100%)	0 (0%)
Cali Vea	47 (97.92%)	1 (2.08%)
Donde Beto El Original	48 (100%)	0 (0%)

Jenos Pizza	48 (100%)	0 (0%)
KFC	61 (92.42%)	5 (7.58%)
Kokoriko	47 (97.92%)	1 (2.08%)
La Monapizza	48 (100%)	0 (0%)
Oma	43 (89.58%)	5 (10.42%)
Subway	39 (81.25%)	9 (18.75%)
Tacos & Bar BQ	48 (100%)	0 (0%)

Table 5. Moment and food delivery result

Moment	**Food Delivery Result**	
	On Time or Early	**Late**
Afternoon	158 (96.93%)	5 (3.07%)
Morning	160 (95.24%)	8 (4.76%)
Noon	159 (95.21%)	8 (4.79%)

Table 6. Restaurants, moment, and food delivery result

Name of Provider	**Moment**	**Food Delivery Result**	
		On Time or Early	**Late**
Cali Mio	**Afternoon**	16 (100%)	0 (0%)
	Morning	16 (100%)	0 (0%)
	Noon	16 (100%)	0 (0%)
Cali Vea	**Afternoon**	16 (100%)	0 (0%)
	Morning	16 (100%)	0 (0%)
	Noon	15 (93.75%)	1 (6.25%)
Donde Beto El Original	**Afternoon**	16 (100%)	0 (0%)
	Morning	16 (100%)	0 (0%)
	Noon	16 (100%)	0 (0%)
Jenos Pizza	**Afternoon**	16 (100%)	0 (0%)
	Morning	16 (100%)	0 (0%)
	Noon	16 (100%)	0 (0%)
KFC	**Afternoon**	17 (89.47%)	2 (10.53%)
	Morning	24 (100%)	0 (0%)
	Noon	20 (86.96%)	3 (13.04%)
Kokoriko	**Afternoon**	16 (100%)	0 (0%)
	Morning	15 (93.75)	1 (6.25%)
	Noon	16 (100%)	0 (0%)
La Monapizza	**Afternoon**	16 (100%)	0 (0%)
	Morning	16 (100%)	0 (0%)
	Noon	16 (100%)	0 (0%)
Oma	**Afternoon**	13 (81.25%)	3 (19.75%)
	Morning	14 (87.50)	2 (12.50%)
	Noon	16 (100%)	0 (0%)
Subway	**Afternoon**	16 (100%)	0 (0%)
	Morning	11 (68.75)	5 (31.25)
	Noon	12 (75%)	4 (25%)
Tacos & Bar BQ	**Afternoon**	16 (100%)	0 (0%)
	Morning	16 (100%)	0 (0%)
	Noon	16 (100%)	0 (0%)

This table summarizes the ten restaurants most frequently preferred by customers for online food ordering and the performance achieved by the RF algorithm of their deliveries according to the time of day. According to this table, customers who want to order food online are offered different alternatives according to the order moment. In fact, this table is used to inform customers which restaurants to choose for fast delivery by using all the variables given in Table 1 using the RF algorithm. In other words, the delivery performance of popular restaurants at the specified times of the day according to the distance of the customers has been revealed by including the traffic situation at the time of the order.

6. Conclusion

In this study, online food ordering data collected from the city of Bagota is analyzed. Raw data are downloaded from Mendeley repository [13]. Considering this data, it is aimed to predict the "Late" or "On Time or Early" arrival of the online food order with the RF algorithm. The variables used in the analysis are explained in detail in Table 1 in Chapter 3. Considering the "Late" or "On Time or Early" of the food delivery, the new variable (DTF_catg) is created and this variable is determined as the dependent variable. Other variables described in Table 1 are used as independent variables.

With the results of this application, customers who makes an online food order can be informed about the performance of the restaurants instantly. Because in order to predict the "late" or "early or on time" arrival of online food delivery with the RF algorithm several important variables are taken into account in this study. These variables are the distance between the customer and the restaurant, also the instant traffic information of this distance, the cost of delivery, the moment of day and the number of comments made to the restaurant by customers also the minimum order amount required for food ordering are also taken into account. Along with these variables, the traffic conditions at the specified moments of the day are also taken into account.

The RF algorithm results declares that the traffic conditions do not affect the online food delivery results too much, although it is thought that it plays an important role for the result of the delivery. This result complies with the results of the study by Correa et al. [14]. Considering that the online food orders delivered to customers are usually provided by single-person vehicles, it is a natural result that the delivery is not affected by traffic.

As a result, this study provides a suggestion about the performance estimation of delivery by RF algorithm at the point of giving customers different alternatives for online food delivery of big data collected by web scraping techniques. In addition, it also helps rival restaurants determine on which variables they should improve their performance to gain advantage over their competitors.

References

[1] M. Akman, Y. Genç, H. Ankaralı, Random Forests Yöntemi ve Sağlık Alanında Bir Uygulama, Turkiye Klin. J. Biostat. 3 (2011) 36–48. https://www.turkiyeklinikleri.com/article/en-random-forests-yontemi-ve-saglik-alaninda-bir-uygulama-59725.html (accessed July 19, 2020).

[2] Ö. Akar, O. Güngör, Rastgele Orman Algoritması Kullanılarak Çok Bantlı Görüntülerin Sınıflandırılması, J. Geod. Geoinf. 1 (2012) 139–146. doi:10.9733/jgg.241212.1t.

[3] S. Özdemir, Random Forest Yöntemi Kullanılarak Potansiyel Dağılım Modellemesi ve Haritalaması: Yukarıgökdere Yöresi Örneği, Turkish J. For. | Türkiye Orman. Derg. 19 (2018) 51–56. doi:10.18182/tjf.342504.

[4] T.E. Kalaycı, Kimlik Hırsızı Web Sitelerinin Sınıflandırılması İçin Makine Öğrenmesi Yöntemlerinin Karşılaştırılması, Pamukkale Univ. J. Eng. Sci. 24 (2018) 870–878. doi:10.5505/pajes.2018.10846.

[5] M.E. Irmak, İ.B. Aydilek, Hava Kalite İndeksinin Tahmin Başarısının Artırılması için Topluluk Regresyon Algoritmalarının Kullanılması, Acad. Platf. J. Eng. Sci. 7 (2019) 507–514. doi:10.21541/apjes.478038.

[6] S. Canaz Sevgen, Airborne Lidar Data Classification in Complex Urban Area Using Random Forest: A Case Study of Bergama, Turkey, Int. J. Eng. Geosci. 4 (2019) 45–51. doi:10.26833/ijeg.440828.

[7] R. Çömert, D. Küçük Matcı, U. Avdan, Object Based Burned Area Mapping With Random Forest Algorithm, Int. J. Eng. Geosci. 4 (2019) 78–87. doi:10.26833/ijeg.455595.

[8] R. Ünlü, Classification of Historical Anatolian Coins with Machine Learning Algorithms, Alphanumeric J. 7 (2019) 275–288. doi:10.17093/alphanumeric.620095.

[9] H. Ekelik, D. Altaş, Dijital Reklam Verilerinden Yararlanarak Potansiyel Konut Alıcılarının Rastgele Orman Yöntemiyle Sınıflandırılması, İktisat Araştırmaları Derg. 3 (2019) 28–45. doi:10.24954/JORE.2019.27.

[10] P. Akın, Y. Terzi, Dengesiz Veri Setli Sağkalım Verilerinde Cox Regresyon ve Rastgele Orman Yöntemlerin Karşılaştırılması, Veri Bilim. 3 (2020) 21–25. https://dergipark.org.tr/tr/pub/veri/issue/55996/642147 (accessed July 19, 2020).

[11] B. Baba, G. Sevil, Predicting IPO Initial Returns Using Random Forest, Borsa Istanbul Rev. 20 (2020) 13–23. doi:10.1016/j.bir.2019.08.001.

[12] M.A. Segura, J.C. Correa, Data of collaborative consumption in online food delivery services, Data Br. 25 (2019) 104007. doi:10.1016/j.dib.2019.104007.

[13] J.C. Correa, Raw Data of A Web Mining Approach to Collaborative Consumption of Food Delivery Services, 1 (2018). doi:10.17632/M9Z9HW4NSC.1.

[14] J.C. Correa, W. Garzón, P. Brooker, G. Sakarkar, S.A. Carranza, L. Yunado, A. Rincón, Evaluation of collaborative consumption of food delivery services through web mining techniques, J. Retail. Consum. Serv. 46 (2019) 45–50. doi:10.1016/j.jretconser.2018.05.002.

[15] A. Güven, Topluluk Öğrenmesi (Ensemble Learning) Yöntemleri, (2019). https://medium.com/@anilguven1055/topluluk-öğrenmesi-ensemble-learning-3b71524297d5 (accessed July 27, 2020).

[16] L. Breiman, Bagging predictors, Mach. Learn. 24 (1996) 123–140. doi:10.1023/A:1018054314350.

[17] A. Singh, A Comprehensive Guide to Ensemble Learning, (2018). https://www.analyticsvidhya.com/blog/2018/06/comprehensive-guide-for-ensemble-models/ (accessed July 27, 2020).

[18] H. Yılmaz, Random Forests Yönteminde Kayıp Veri Probleminin İncelenmesi ve Sağlık Alanında Bir Uygulama, 2014.

[19] L. Breiman, Random forests, Mach. Learn. 45 (2001) 5–32. doi:10.1023/A:1010933404324.

[20] K.J. Archer, R. V. Kimes, Empirical characterization of random forest variable importance measures, Comput. Stat. Data Anal. 52 (2008) 2249–2260. doi:10.1016/j.csda.2007.08.015.

[21] L. Breiman, A. Cutler, Random forests - classification description, (2005). https://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm (accessed August 4, 2020).

[22] T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning, 2009. doi:10.1007/978-0-387-84858-7.

[23] J. Cohen, A Coefficient of Agreement for Nominal Scales, Educ. Psychol. Meas. 20 (1960) 37–46. doi:10.1177/001316446002000104.