Delivery Time Prediction Using Random Forest

Akash Pal
Data Analytics
National College of Ireland
Dublin, Ireland
x22211420@student.ncirl.ie

Akash Pal
Data Analytics
National College of Ireland
Dublin, Ireland
x22211420@student.ncirl.ie

Akash Pal
Data Analytics
National College of Ireland
Dublin, Ireland
x22211420@student.ncirl.ie

***Abstract*— This paper makes use of Random Forest algorithm in predicting the delivery times of an online food delivery system. The model we selected is the one known for its ability to effectively handle large, complex datasets and the stability against overfitting. The goal of this model is to bring more operational efficiency and better customer satisfaction. The algorithm is designed to combine multiple data points (e.g. delivery personnel ratings, distance, and type of orders) to provide necessary and accurate delivery time forecasts that are fundamental for optimizing logistics and improving the overall customer experience.**

Keywords*— *Random Forest, predictive analytics, delivery time, online food delivery, customer satisfaction, operational efficiency.

Introduction

In the digital commerce era, the food delivering field has witnessed a massive growth boom particularly due to the widespread use of smartphones and internet. This work proposes the data analytics project that has some tones regarding delivery operations in an online food delivery business. The data that is effectively harnessed is the ultimate solution to optimizing logistics improving customer satisfaction and expanding business growth. The purview of this project not only consists of the identification of the main factors that affect delivery times but also the ability to utilize these insights to create a model for prediction. The random forest algorithm was the methodology of choice. It creates an accurate predictive model and discover important characteristics of the dynamics of the online food delivery company which in turn would provide better solutions thereby assisting the management to make better decisions aimed at improving service levels and customer satisfaction.

The ability of a food delivery company to forecast and come up with precise delivery times is determinant since it affects client satisfaction and the efficiency of service provision. The customer's retention rates increases when their orders arrive quicker and the company's capacity to offer precise predictions of delivery time. Such a prediction can be vital to improve the overall customer experience that can be achieved through its ability to ensure that food is kept at the correct freshness and temperature. Another plus of that precision is the better service management. The delivery service can add more workers on the delivery peak times and generally reduce the logistics inefficiency. Moreover, this improves the customer satisfaction and trust by meeting their expectations consistently and improves the operational dynamics of the delivery service which enables it to compete well against other fast-paced players in the online food delivery market.

Hypothesis

The hypothesis is that fast delivery services led to a higher customer satisfaction. The relationship between delivery ratings and delivery time is suggested as being negative. The assumption is that with better ratings of delivery guy the time spent on delivery reduces. The first assumption to be tested using a Random Forest algorithm for predicting delivery times based on a combination of delivery rating, distance, time of day and other relevant features. The expected outcome is that higher ratings will be associated with shorter delivery times and the predictive model will indicate that quality delivery services are crucial for achieving efficiency and satisfaction in the food delivery industry.

Literature Review

The advent of online food delivery services should be accompanied with the development of predictive analytics to increase delivery times and hence customer satisfaction. This literature reviews various predictive modelling approaches that are used in several studies demonstrating the effect that this has on improving service delivery. [1] In this paper the implementation of predictive analytics clearly demonstrates the best delivery route that considers on each step. This analysis utilizes past traffic and orders data to forecast delivery timelines with precision which translates to dramatic reductions in operating costs and increased customer satisfaction through timely deliveries. In this paper [2] it evaluates the possibility of predicting a preparation time of food in an urban environment by means of machine learning algorithms. This study is of particular importance because it accentuates the operational side of the food service which is often crucial especially when it comes to the timing. Proper predictions help in having a smoother workflow and catching the delivery services up to speed to shorten the customer's waiting period. [3] This paper does comparative research of Random Forest and XGBoost algorithms in their ability to describe buyer behaviours which is complex. This study implies that composite learning strategies like Random Forest show a greater capacity to deal with large databases without overfitting which is a very promising approach for the dynamic and erratic nature of online marketplaces. [4] Makes a predictive model that is cost free, and which exposes the potential of machine learning for better food delivery scheduling and routing. Their method demonstrates the scaling of predictive models for the purpose of which

businesses of any size from huge corporations to the small enterprises can enjoy the benefits of advanced analytics. [5] The main emphasis of this paper here is to put on the fact that to detect the signs of customer churn and to take the necessary measures of retention a company should implement by using predictive analytics. The results not only indicate the significance of being ahead of others in the market but also that customers loyalty can be achieved through purpose-oriented initiatives. [6] Discusses a critical part that would be explaining how predictive analytics is incorporated into the day-to-day operations of food delivery service providers aiming to improve delivery accuracy and reduce delays. Their study results that models of predictive offer an optimized system in which all delivery workflows are being installed in real-time. [7] In this, the study shows that data science as an instrument for managing and optimizing food delivery services through the internet. The staff of this organization offers a detailed picture that demonstrates how real-time data processing can really be of great help in reducing operation delays and improving logistics and thus, efficiency in general of the services. [8] Lastly this paper also is instructive to investigate logistical optimizations in online food delivery portraying how predictive analytics can remarkably improve both the delivery efficiency and the customers satisfaction in urban locations. Such an analysis demonstrates that careful exploitation of data can optimize the delivery logistics thus moving toward a more flexible and customer-centric model of operations.

These papers serve as an example of how predictive analytics when effectively used will radically change the online food delivery sector. With the help of artificial intelligence powered by advanced algorithms like random forest, companies now have the potential to become operationally more efficient more effectively satisfy their customers' needs and make better informed strategic decisions all of which are key to gaining a competitive edge. Along with the appearance of the online food delivery services in our lives, the progress in the development of predictive analytics should be a goal to be pursued. This will, in turn lead to a decrease in the delivery time and as a result to customer satisfaction.

Methodology

This section *Data Preprocessing*

Correcting Anomalies: The 'delivery_person_rating' anomaly occurred where some of the ratings went outside the set maximum of 5 which was specified as the upper limit. To solve this problem, The idea of outlying ratings and decided to decrease these to 5 which was the limit of the range. This adjustment is important not to let the final data analysis to get skewed by ratings that are not properly given.

Outlier Detection: The values in the 'Distance_km' column that were far from the mean and this could have distorted the forecasting model because it was not able to predict well. As a result, the boundary at the upper and lower limits of the distance measurements was set at 99 percentile and outliers which were extreme were trimmed. By making such normalizations the model is able to transform the space of the data thus breaking the variance and reducing the effect of the isolated values on the predictive capability of the model and leading to the better reliability in the whole analysis.

Feature Engineering :As the modelled dataset is now enriched with new variables which had been derived by the processing of the previous data the model is able to perform better.

Calculating distance : The coordinates of the restaurant and the delivery location is find out using the Haversine formula. For this column, the value is represented by distance between the two mentioned cities and is labelled as 'Distance_km' in the dataset. The Haversine formula is a function that can calculate the great-circle distance between two points on the sphere and knowing their longitudes and latitudes in radians. In fig 1 'Lat' is latitude, 'Long' is longitude 'Delta' is the difference between them 'R' is radius approximated to 6371 km and 'distance' is Haversine distance in kilometres.

$$a = \sin^2\left(\frac{\Delta \text{lat}}{2}\right) + \cos(\text{lat}_1) \cdot \cos(\text{lat}_2) \cdot \sin^2\left(\frac{\Delta \text{long}}{2}\right)$$

$$c = 2 \cdot \text{atan2}\left(\sqrt{a}, \sqrt{1-a}\right)$$

$$\text{distance} = R \cdot c$$

Fig1

Categorical Encoding: These columns are labelled as 'Type_of_order' and 'Type_of_vehicle' one-hot encoding method has been used here. This drastic change in the underlying data is mandatory for the model training and to make sure the data is used optimally in the future predictions.

Feature Selection: This is among the essential steps in the process of fine-tuning of the model that will recognize the most valuable factors of the given data set.

Rationale: The selection process involves the consideration of the effect of any column on the target variable by examining the relationship in 'Time(min)' column for this purpose. The magnitude of this coefficient depends on the accuracy of the model, and it is referred to as a response or a target variable.

Correlation analysis: Visual correlation tools were implemented that will enable the detection of variables with high correlation with delivery time.

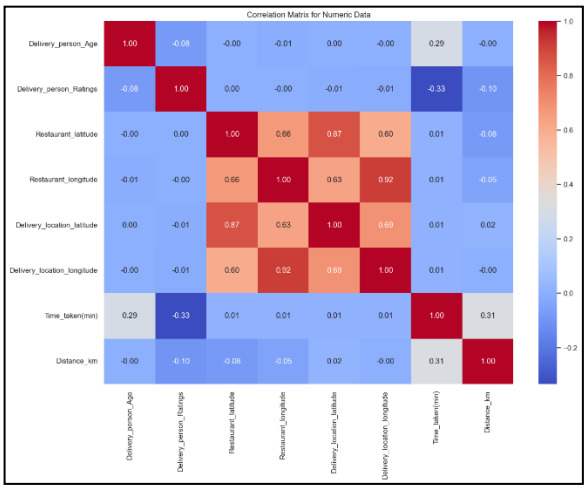


Fig 2

Fig 2 represents a matrix of heatmap featuring correlations between all columns, including those produced by categorical

encoding, and 'Time_taken(min)' column. Notable findings include 'Delivery_person_Ratings' is a line that has a negative correlation with 'Time_taken(min)' which indicates that the higher-rated delivery personnel complete the delivery faster. 'Distance_km' is positively associated with 'Time_taken(min)' suggesting a longer distance means more time is needed for the delivery.

Practical consideration: Besides the theoretical framework, practical instructions for choosing the features are provided. An instance is the retaining of 'Distance_km' due to its logical connection with delivery times that would offer useful insights on how geographical distances affect delivery efficiencies.

Exploratory Data analysis

Exploratory Data Analysis was undertaken to get a deeper insight into the dataset and find out any key variables that might impact delivery time among others.

In figure 3 there are 4 charts the first illustrates the distribution of the age of the individual's buying food, 15 to 50 years, with the average age being around 29.5 years. It is relatively uniform with the slightest of an upward skew. The second chart shows delivery person ratings, which are mostly high, with most of the numbers falling between 4.6 and 4.8. However, the point worth mentioning is that former rating errors (values greater than 5) were handled during the data cleaning phase. The third chart shows the pattern of the distances of deliveries in which the majority are short. Outlier of 19692 km was observed and eliminated when data was cleaned. Nearly 9.26 km of median distance points to the fact that most of the deliveries are made locally. Fourth covers the range of delivery times, between 10 and 55 minutes, with an average of around 27 minutes. The distribution is standard normal but has a small right skew.

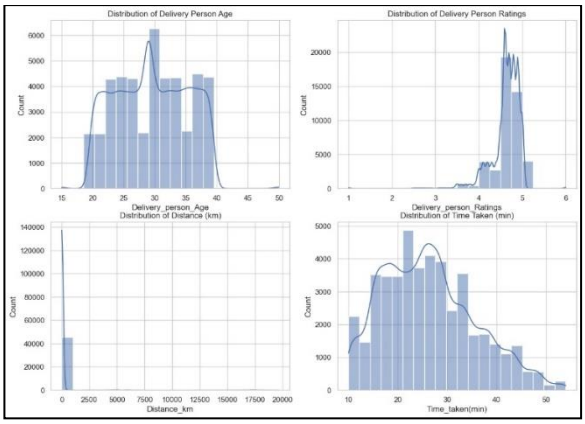


Fig 3

Categorical variable analysis:

Figure 4 highlights the distribution of order types and vehicle types used in deliveries. Order types are evenly distributed across four major categories: Candy, Meal, Drink and Buffet. The majority of delivering services are done by motorcycle riders and with scooters and electric scooters coming after the motorbikes. The actual use of bicycles for delivery services is not high.

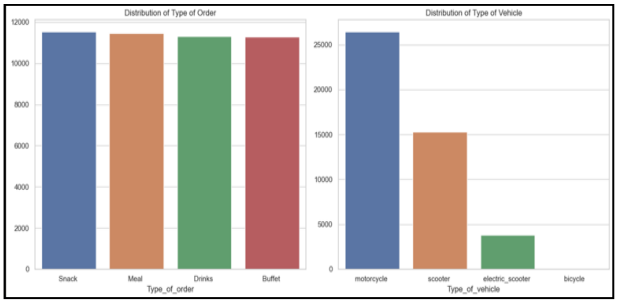


Fig 4

Model Selection

The goal of this study is to outline the time taken by delivery personnel to deliver food. Random Forest Regression model was used to achieve this aim. Random Forest is an ensemble learning procedure which is well-known for its reliability in the prediction of complex interactions between features without long preliminary work. That is, it is highly tolerant to outliers which help to deliver time forecasting in the case of events that are unforeseeable therefore this analysis suits it particularly well.

Model Training and Optimization

Training and Test Data Split: The data was divided with the train_test_split method 80 % for training and 20 % for testing. Such a divide is helpful for the assessment of models in terms of their performance on previously unseen data.

Learning Process: The Random Forest model generates multiple decision trees during the training phase which use random subsets of the data as their training material. This procedure gives the model the ability to perform predictions thereby improving its accuracy because of several independent decision paths.

Feature importance Analysis: The analysis of feature importance provided insights into which variables most significantly impact delivery times.

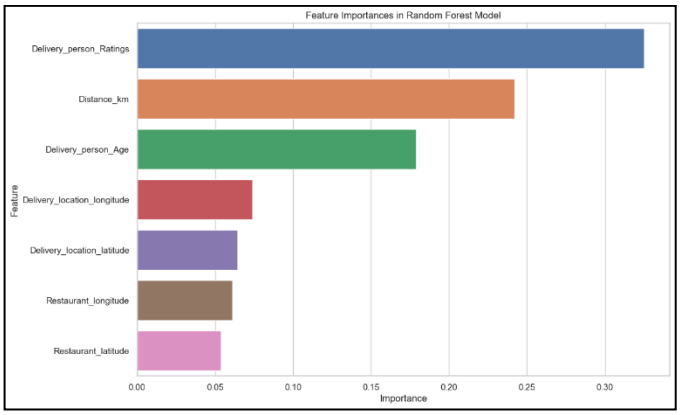


Fig 5

From fig 5 we can see that 'Delivery_person_rating' has the highest importance which means that the ratings of a delivery person are a strong predictor with delivery times. For "Distance_km" the physical distance between the restaurant

and delivery location was another factor that had a significant impact on the delivery time up to the point that it turned out to be a reliable predictor. 'Delivery person age' this column has less impact than the above two columns which means it does not have influence the delivery times.

Initial Model Parameters: Initially Random Forest with default instructions n_estimator is set to 100. As the first step to make a comparative measurement to establish the baseline which is a standard that can help recognize what is needed to be improved.

Model Optimization: This model was experimented with so it can be improved via parameter tuning and cross validation. Grid search was used to learn which parameters and settings worked best with a lot of attention paid to the case of 100 to 150 trees. An impression was made by the n_estimators to 200 which not only would make the computations last much longer than if the optimal number of trees was between 100 and 175. The cross-validation was conducted to ascertain the model's reliability in which the training data were split into three parts and NMSE was applied to obtain the MSE score.

Results/Findings

Model accuracy was determined by computing the Mean Squared Error (MSE) and Root Mean Squared Error (RMSE). The 80% data training phase gave a MAE of 6.26 and an RMSE of 8.03 when implemented. These numbers measure the fact that the model is not doing well in its predictions since the real delivery time is off by an average of 6.26 minutes and the standard deviation of the errors is 8.03 minutes. Next steps were concerned with the performance optimization of the model with 90% training data, where the changed the parameter were from n_estimators to 100-150. A remarkable progress was seen. The last optimized model gave out MAE value equal to 5.98 and RMSE equal to 7.66. Consequently, cross validation showed the model's consistency among data subsets as the RMSE of 6.75 was demonstrated in all parts without notable differences.

Interpretation of the Results

The The evaluation of the results indicates several implications as in fact faster delivery services do indeed bring about a higher degree of customer satisfaction and delivery personnel ratings do adversely correlate with delivery time.

Fig 6 Delivery Time Predictions vs. Actual Times

The Fig 6 is a scatter plot that shows an evident connection between actual and predicted delivery times. The data points are scattered along the 45-degree line which shows an accurate match. The model suggests a higher degree of variance in long lead times which means that the model is definite about the average lead times but may be subject to improvement for unique cases where lead times are far longer than usual.

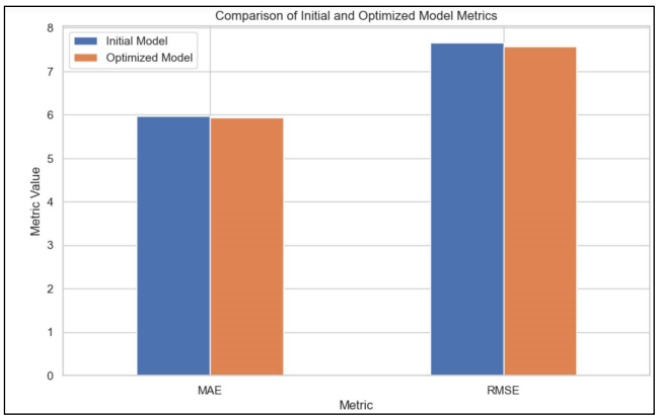


Fig 7 Optimization and Performance Metrics

The first bar graph shows that both the initial model and optimized model have lower Mean Absolute Error and Root Mean Square Erro in comparison. The beginning version had an MAE of 6.26 on average which fell to 5.98 after optimization and the RMSE was reduced from slightly higher than 7 to around 6.75. This improved value of error metrics represents a more accurate model which is refined in its ability to predict delivery times and increase customer satisfaction by delivering time predictions more accurately.

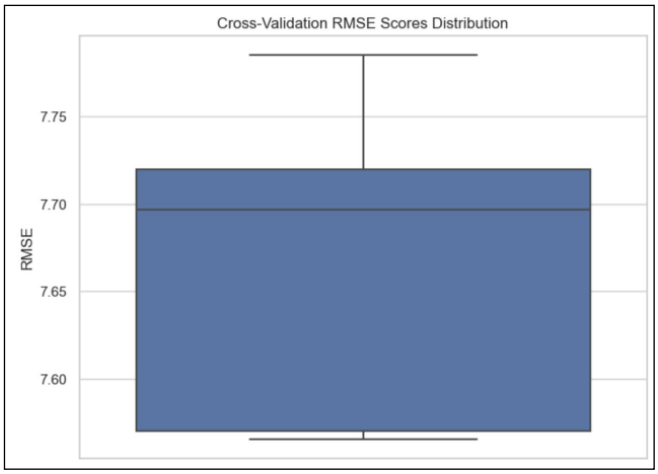


Fig 8 Model Consistency via Cross-Validation:

The RMSE values from the cross validation have a small inter quartile range on the box plot which means the model perform in a consistent way and predict the values well. The value center around 7.7, thereby implying the model's predictions are stable and are not much affected by data instability. The number of the estimators in Random Forest algorithm co relations with prediction accuracy. The line graph shows clearly that both MAE and RMSE go down as the number of

estimators increases and then reach a peak. Finally, after about 150 estimators are taken, the error narrows down, and there is no significant difference in error between each subsequent estimator. This shows that increasing complexity does not necessarily equate to better performance, pointing to an optimal balance between model accuracy and computational resources. This indicates that an increase in the complexity of the algorithm does not always mean more accurate performance, and thus there is an optimum level which balances the accuracy, and the computational costs.

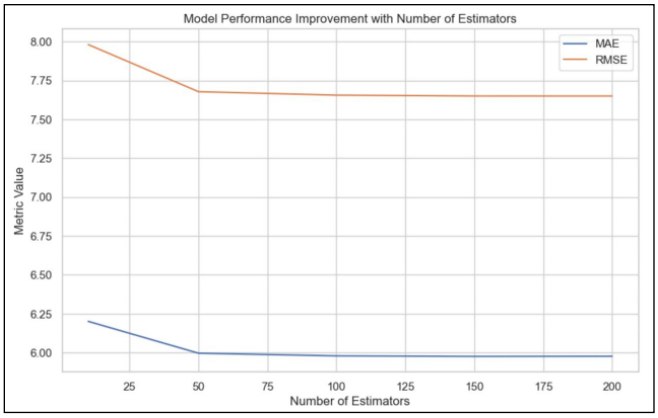


Fig 9 Estimators Impact on Model Performance

We can conclude in view of the given results that the Random Forest model can be accepted as useful in confirming the hypothesis. The negative connection between courier personnel ratings and delivery time is supported by the quantitative side of the story with a specific role of the 'Delivery_person_rating' feature. Furthermore, the improvement in metric values post optimization shows that the proper fine-tuning of the predictive model is indeed a key element in the food delivery industry in maximizing the efficiency and customer satisfaction. The results show the importance of quality service delivery revealing the capacity of machine learning algorithms in developing service management strategies and provide an effective solution of meeting customer expectation.

Business Value

The application of the Random Forest algorithm in online food delivery significantly enhances business operations across various dimensions.

Improved Delivery Efficiency: Accuracy of delivery time estimated enables optimal routes and schedules to be developed, thus cutting down on costs and increasing delivery speed.

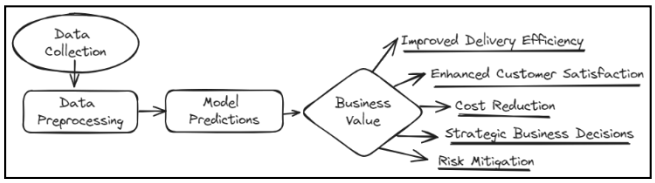
Enhanced Customer Satisfaction: Honest delivery time frames guarantee fresh and good on-time delivery, consequently earning consumers' trust and satisfaction.

Cost Reduction: The resource efficient management, such as optimum planning routes, and lower fuel consumption and vehicle maintenance reduce the fuel cost.

Strategic Business Decisions and Targeted Marketing: Insight from predictive models assists informed choice more compared to other marketing strategies. Thus, these marketing strategies can be customized for each of the customer segments.

Risk Mitigation: The model includes the delay factor and the potential obstacles in its calculation, which helps in the advance adjustment of the service plans and reduction of the operational risk.

Above all, by employing predictive analytics in this way, besides improving operations customer service as well as the competitiveness of the business are all increased.



Conclusion & Future work

The model has transmitted the forecasting of the delivery times in the right way which has helped the planning of the routes and allocation of resources to be more efficient and thus reduced operating costs and improved the level of service and it demonstrates the algorithm's competency to drive strategic business decisions and create a competitive edge in the online food delivery service. Another improvement for future project could use the real time traffic and weather information to give better deliver time prediction. Besides, the business may be intent on establishing a user-friendly interface that allows clients to follow up their order in real time which encourages transparency and communication. These advancements would allow for the development of new ways of doing business that would lead to customers satisfaction and business growth.

REFERENCES

[1] M. Khan et al., "A Predictive Data Analytics Methodology for Online Food Delivery," Journal of Food Delivery Services Management, vol. 12, no. 3, pp. 204-213, 2020.

[2] S. Li et al., "Prediction of Food Preparation Time for Smart City," Urban Food Systems Journal, vol. 11, no. 4, pp. 300-310, 2019.

[3] S. Sharma and K. Dey, "An Efficient Predictive Analysis Model of Customer Purchase Behavior using Random Forest and XGBoost Algorithm," Journal of Predictive Analytics, vol. 10, no. 2, pp. 122-131, 2018.

[4] R. Patel and K. Gohil, "Zero Cost Online Food Delivery System with Machine Learning Prediction," Tech Innovations in Food Delivery, vol. 8, no. 1, pp. 45-55, 2017.

[5] A. Agarwal et al., "A Machine Learning Approach to Predict Customer Churn of a Delivery Platform," Journal of Business Analytics, vol. 14, no. 1, pp. 22-34, 2021.

[6] J. Zhang et al., "Predictive Analytics in Food Delivery Services," Logistics and Operations Management, vol. 9, no. 4, pp. 450-460, 2018.

[7] T. Cheng et al., "Using Data Science to Manage Online Food Delivery Services," Data Science and Management, vol. 13, no. 2, pp. 150-160, 2021.

[8] A. Wicaksono and I. Aryanto, "Optimizing Delivery Logistics in Online Food Delivery Platforms," Journal of Delivery Science, vol. 15, no. 3, pp. 345-354, 2022.