Food Delivery Analysis

Christopher Way, Saurav Thapa

IS428: Data Mining Techniques and Applications

Professor Karen Chen

December 20, 2023

Abstract

This study presents an exploration into how one can properly take advantage of data from food delivery services in order to develop a model to predict food delivery times based upon a variety of factors. The study is motivated by the rapid growth of food delivery apps like UberEats, DoorDash, and Grubhub, especially in the context of the post-covid delivery sphere, which has significantly higher usage than pre-covid levels. Various factors are analyzed in order to understand their impact on delivery services. The findings are intended to assist both users and drivers by providing a clearer understanding of delivery times to enable better planning and service efficiency. Using a dataset containing 45,593 rows and 19 unique columns, the study conducts an exploratory data analysis, revealing key insights such as the impact of weather conditions on delivery times and the predominance of certain age groups among delivery personnel. Several regression models were used to analyze the data, with a specific focus on Gradient Boosting, Random Forest, and Decision Tree regression. The results were deduced by a focus on metrics like Mean Squared Error (MSE), Mean Absolute Error (MAE), and R-squared. The results indicate that driver ratings, multiple deliveries, and road traffic density are significant factors in determining delivery times. The study concludes that the Gradient Boosting and Random Forest models were most effective in predicting delivery times, with r-squared metrics of about .6 for each. The study gives valuable insights into the food delivery industry, with implications for both drivers and users to optimize their delivery experiences.

Background/Motivation

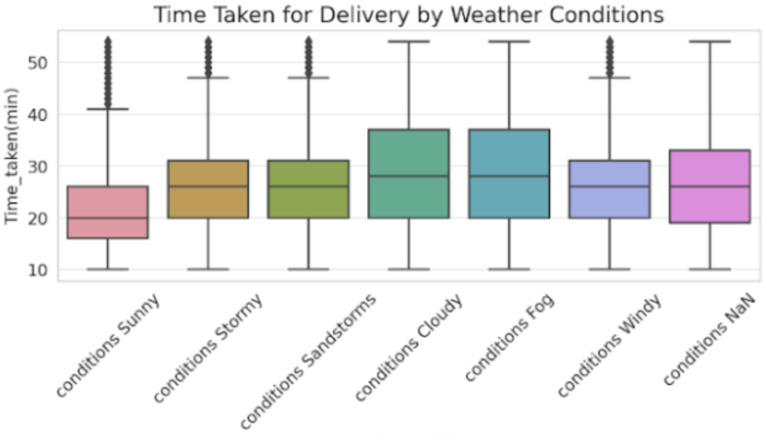
Food delivery apps such as UberEats, DoorDash, Grubhub, etc. have exploded in popularity in recent years. Their user base was slowly growing in the late 2010's; however, due to COVID-19, their usage increased exponentially. Most people were hesitant to leave their houses due to the fear of catching the virus, so they turned to food delivery apps to still get the food they craved for. Furthermore, for the people who wanted to earn some money, it was a way for them to do so with limited contact with the customers. It kept the economy going and kept both the drivers and users content with the service.

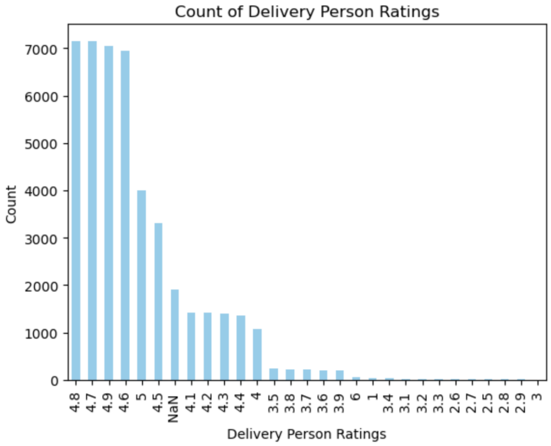
As we were looking for datasets to use, we came across this one and it piqued our interest. I, Saurav Thapa, have experience doing DoorDash in a small town and seeing this data made me curious as to how delivery time was impacted by the rating of the person, type of city, type of vehicle, order type, etc. We wanted to take a deeper look into the delivery drivers and how delivery time can be affected by numerous factors.

Our findings from this project will help users and delivery drivers alike in seeing how delivery is affected by various factors. For example, users can look at the results to determine how long it would take for their food to be delivered, so they can time it more accurately. If they are out of their home and plan to get back at a certain time, they can more precisely order the food so it gets to their doorstep not too long after they get home. For delivery drivers, they can better estimate how long it will take to deliver, so if a customer is waiting on an order, they can give them a better guess. Additionally, they can decide to take on multiple orders if they realize that they can deliver multiple foods in time.

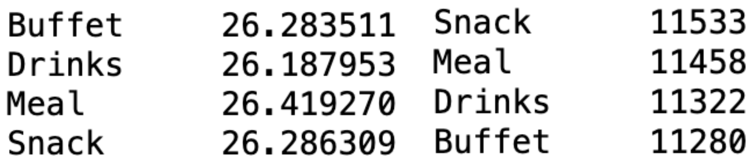
Exploratory Data Analysis

Looking through the dataset, it contains 45,593 rows and 19 unique columns, them being ID, Delivery_person_ID, Delivery_person_Age, Delivery_person_Ratings, Restaurant_latitude, Restaurant_longitude, Delivery_locaiton_latitude, Delivery_location _longitude, Order_Date, Time_Ordered, Time_Order_picked, Weatherconditions, Road_traffic density, Vehicle_condition, Type_of_order, Type_of_vehicle, Multiple_deliveries, Festival, City, and Time_taken(min). The data was collected from February to April, but March had the higher order amount, which was around 31,989 orders.

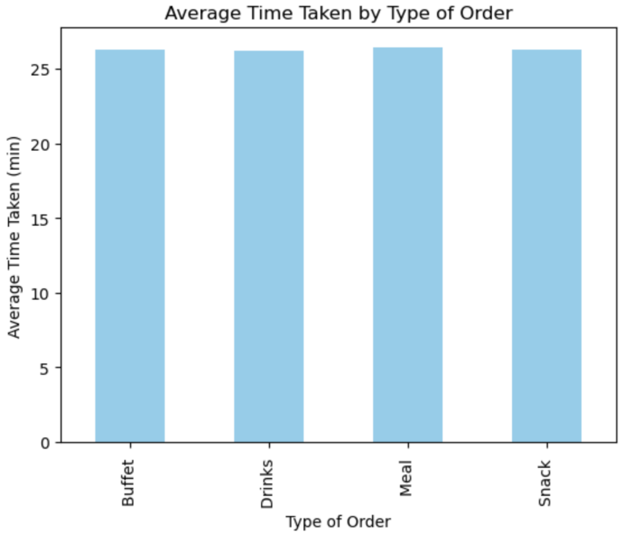




Our initial findings showed that cloudy and foggy conditions resulted in longer delivery time, so sunny days allowed for faster time. The data contained roughly the same amount of food types: 11,533 snack orders, 11,458 meal orders, 11322 drink orders, and 11,280 buffet orders. Additionally, it took roughly around 26.28 minutes for each type of order. Ages 20-39 were the primary delivery drivers, and ages below or above were not recorded. Next, food was picked up within 15 minutes from order placement no matter the conditions. Motorcycle was also the most popular delivery method, while bicycle was the least preferred by the drivers. Lately, the dataset skewed left for ratings, meaning most were 4.5 or higher.



```
Buffet      26.283511  Snack      11533
Drinks      26.187953  Meal       11458
Meal        26.419270  Drinks     11322
Snack       26.286309  Buffet     11280
```



Model Development

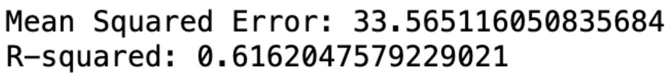
Before analyzing our dataset more in depth, we had to prepare it for modeling. First, we removed “ID”, “Delivery_person_ID”, and “Weatherconditions” from the dataset. The two different ID columns were removed because they were irrelevant to our experiment. “Vehicle_condition” was removed because there was no context surrounding it. For example, we searched up vehicle condition 0 on google and it showed that it meant the vehicle was inoperative, which made no sense so our only option was to drop it. Next, we removed “(min)” from the rows and converted it to integers because it made analyzing harder as the values would be string and not numbers. Additionally, any columns that contained numbers were converted to integers for the purpose of running the model. We also removed any null, “NaN”, and 0 values from the columns except for a couple columns where it made sense to have 0 as a value.

Initially, when approaching the problem of how to calculate food delivery time, we were unsure of the best type of model to use. We tested a large variety of different regression models such as linear, lasso, ridge, elasticnet, support vector, k-nearest neighbor, decision tree, gradient boosting, quantile, huber, and gaussian process. After running the models, we settled on three models to use as they provided the most accuracy when performing the regression, which were gradient boosting, random forest, and decision tree.

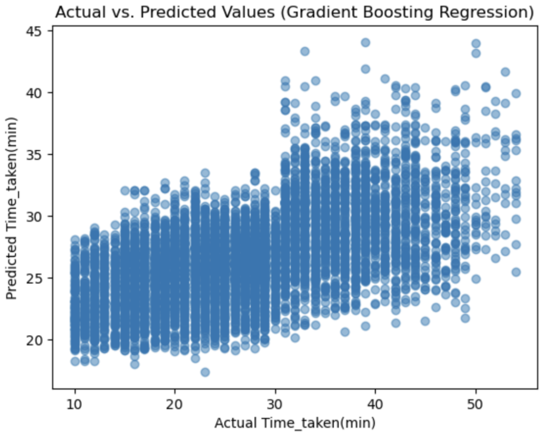
Result and Insights

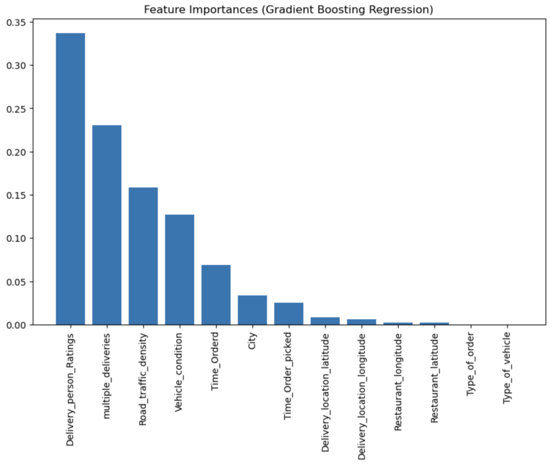
When we ran the various different models, we were looking for three key aspects: mean squared error, mean absolute error, and r-squared. Mean Squared Error (MSE) is the average of the squares of errors. It gives more weight to large errors due to the squaring of each term and is more sensitive to outliers. A lower MSE indicates a better fit for the model to the dataset. Mean Absolute Error (MAE) is the average magnitude of errors in a set of predictions. This is the average over the test sample of the absolute differences between predictions and actual observation which is less sensitive to outliers compared to MSE. Lastly, R-squared is a statistical measure that represents the proportion of variance for the dependent variable explained by the independent variables in a regression model. It ranges from 0 to 1, where a higher value indicates a better fit of the model of the dataset. However, it does not indicate whether the model is appropriate. In terms of graphing, we were looking for a linear upward trend as the actual time needed to match up with the predicted time. Lastly, feature importance told us what the model considered which values were most important when running it. A higher number obviously meant more importance.

Firstly, running Gradient Boosting gave us a MSE of 33.5651 and R-squared of 0.6162, which was the most accurate of any model we ran. Running a feature importance showed us that the rating of the driver was the most important factor followed by multiple delivered and road traffic density. This makes sense because drivers tend to have a rating around 4.5 and users give 4 or 5 on average. If they have a rating close to a 5, it means they have exceptional service and deliver food in time without messing anything up. For graphing, it had a general upward trend, but there were too many errors when predicting the delivery time to the actual time.

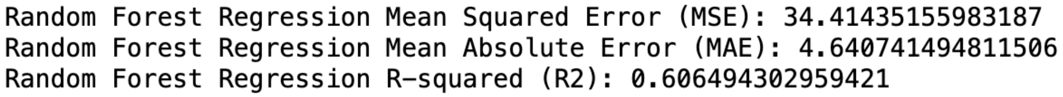


```
Mean Squared Error: 33.565116050835684
R-squared: 0.6162047579229021
```

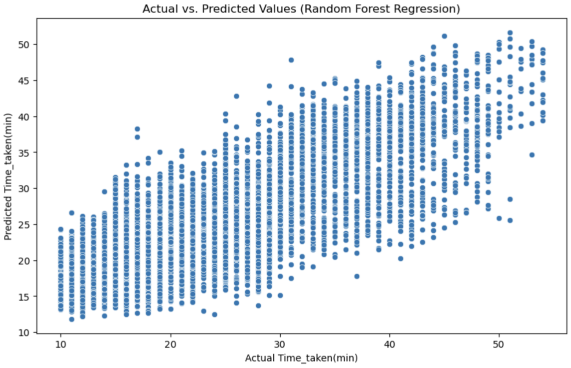


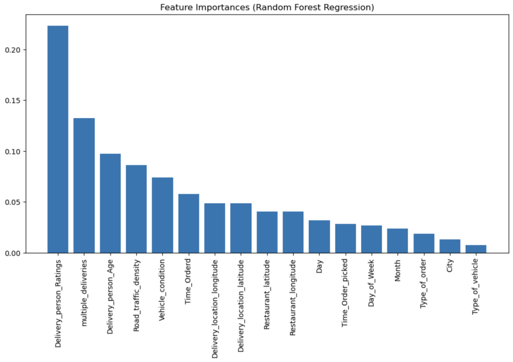


Next, running Decision Tree gave us a MSE of 36.1197, MAE of 4.6407, and R-squared of 0.6065, which was the second most accurate of the models we ran. The graph had a general positive trend, but like Gradient Boosting, it made many errors when predicting the time. For example, when predicting a 10 minute delivery time, the model ranged from approximately 13 minutes to 25 minutes. It also ranked the features as delivery driver rating, multiple deliveries, and delivery person age. This is the same as Gradient Boosting and it makes sense because if a driver were to have multiple deliveries, it would take them additional time to deliver the other food. If a driver predicted wrong and the two (or more) orders were in opposite directions, they would have to spend more time driving from one to the next.

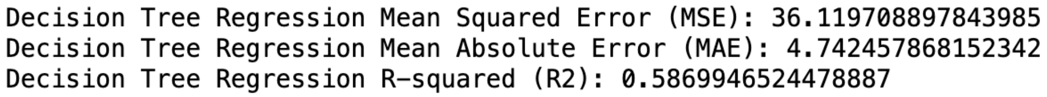


```
Random Forest Regression Mean Squared Error (MSE): 34.41435155983187
Random Forest Regression Mean Absolute Error (MAE): 4.640741494811506
Random Forest Regression R-squared (R2): 0.606494302959421
```

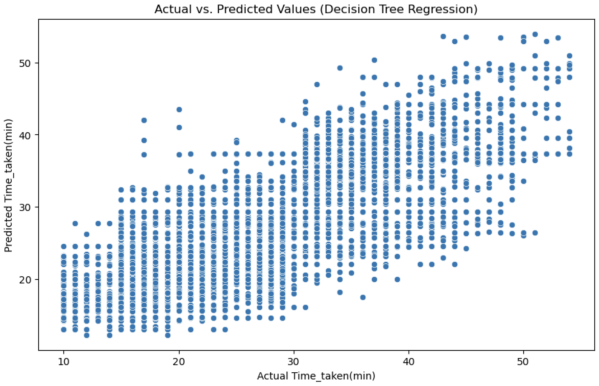


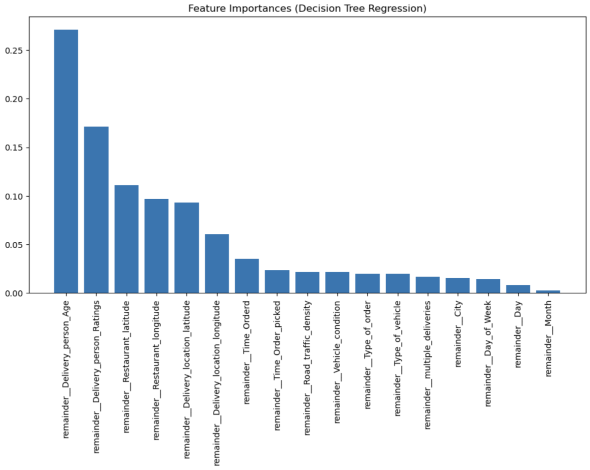


Lastly, our third most accurate model was Decision tree where the MSE was 36.1197, MAE was 4.7425, and R-squared was 0.587. The graph it produced also had a general positive trend; however, it was a lot less visually accurate than the other ones. For example, its predictions for 10-25 minute deliveries were about the same and only started guessing higher after that. For feature ranking, it listed them in the order of delivery person age, delivery person ratings, and restaurant latitude and longitude, which is very different compared to the other models. It is interesting that it ranked the person's age as the most important because we never guessed the driver's age correlated to how fast they deliver. However, based on the results, we can guess that younger people tend to drive more recklessly than older people, so they deliver food faster.



```
Decision Tree Regression Mean Squared Error (MSE): 36.119708897843985
Decision Tree Regression Mean Absolute Error (MAE): 4.742457868152342
Decision Tree Regression R-squared (R2): 0.5869946524478887
```





After running the different models, we decided to use Gradient Boosting and Random Forest to predict the time taken for deliveries. The models were able to predict within 1-2 minutes of one another. Based on the R-squared, we can conclude that they are about 60% correct in terms of predicting, so they should be up or down about 10-15 minutes from the actual time.

Conclusion

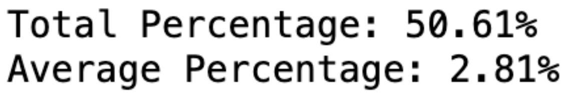
In conclusion, this study presents a comprehensive examination of factors influencing food delivery times. After examining the various models available, Gradient Boosting and Random Forest were determined to be the most effective at predicting food delivery times. It has been found that the rating of a driver is a good indicator of how fast they deliver food. Vehicle type has been found to be a poor indicator of how long it takes a driver to deliver food. Developing a quality model requires significant data cleanup and feature selection, a significant amount of the original dataset had to be purged in order for the model to have any value in predicting delivery times. It appears that for a dataset of this type involving many columns, gradient boosting and random forest regression are useful algorithms to apply when attempting to develop a predictive model.

Future Work

In terms of future work, we would like to begin exploring larger and more complex datasets. Our dataset was only 45,000 rows maximum, with thousands removed during the data cleanup process. We could likely generate more interesting results through larger datasets. We also would like to use more resource intensive algorithms in order to develop models. The ones we used only required a small amount of computational power, we could possibly get more useful models and information by working with more powerful computers. We also strongly desire to achieve more accuracy with our results. Our highest accuracy achieved in our model development was only a .616 r-squared, we would like to get much closer to 1 on that statistical measure, as well as a lower MSE and MAE.

Team Reflection

Based on the feedback we received from the milestone report from the professor and TU Dublin students, we made various changes to our project. Firstly, we better outlined the motivation and background for why we chose the data and use cases for why it was important for us to analyze it. We admit that our initial report was not the best in terms of conveying that. Next, there were questions about how the data was collected. We downloaded the data from Kaggle, which tends to give insight into how the data was collected and what the different variables mean; however, the user who uploaded the dataset failed to report how they collected the data and what some columns mean, like vehicle condition. We looked through the discussion and it seems other users are also confused on what specific columns mean and question how the data was collected. We mentioned needing to exclude some columns or rows previously but failed to mention why. Some rows contained erroneous values because they contained "NaN" or 0 for columns that required specific information. After running some code, we found that there was a total of 50.61% missing values from the data and approximately 2.81% missing values from each column



```
Total Percentage: 50.61%
Average Percentage: 2.81%
```

Reference

https://www.kaggle.com/datasets/gauravmalik26/food-delivery-dataset