Predicting Time of Arrival for Food Delivery Service

Daniel Carpenter: danielcarpenter@ou.edu

Sonaxy Mohanty: sonaxy.mohanty@ou.edu

Zachary Knepp: Zachary.M.Knepp-1@ou.edu

University Of Oklahoma

Data Science and Analytics Institute

DSA 5103-995 | Intelligent Data Analytics

Fall 2022 | Charles Nicholson

Table of Contents

Problem Background

Problem Description

This analysis aims to predict estimated delivery times for a food delivery service. A firm can give the option for delivering the food to a customer's house; using current technology, the company may give the consumer an estimated time of arrival to help manage their expectations, which could lead to enhanced retention of customers for future orders. Companies like DoorDash or GrubHub give customers an estimated time of delivery for food and beverage orders. Most consumers may expect a few conditions to affect the time to deliver, but there may be many circumstances that impact the delivery time. For example, if the algorithm knows that there is a crash impeding traffic between the major routes of the customer and the delivery service, then that may impact the transport time. Additionally, severe weather may delay the ability of a driver to deliver the food to the destination. Overall, providing accurate estimates to the customer will help manage expectations, which may lead to retained customers.

Data Description

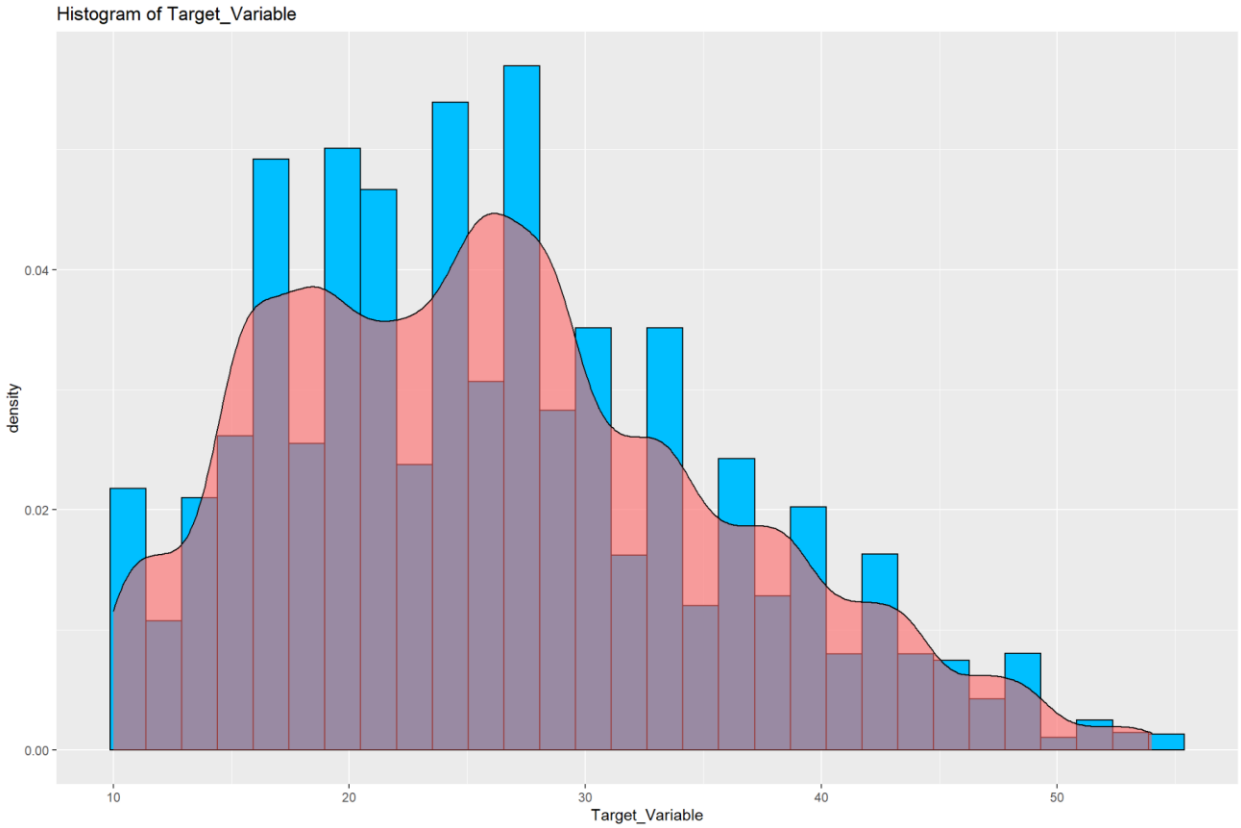
The data set for this problem consists of 19 columns describing the characteristics of the delivery driver and the conditions that they face while driving to the destination. The target or predicted variable is the minutes taken to deliver the food. Location data such as the latitude and longitude of both the source restaurant and delivery location are included. The data offers details such as the time the that the customer placed the order and the time that the delivery service picked it up. Additionally, the data describes the type of order placed. Other characteristics about the city or known festivities occurring during the time of delivery are

included. Finally, the remaining data reveals observed weather, traffic, and vehicle conditions. Altogether, this information helps create a model to predict the time to deliver the food or drinks.

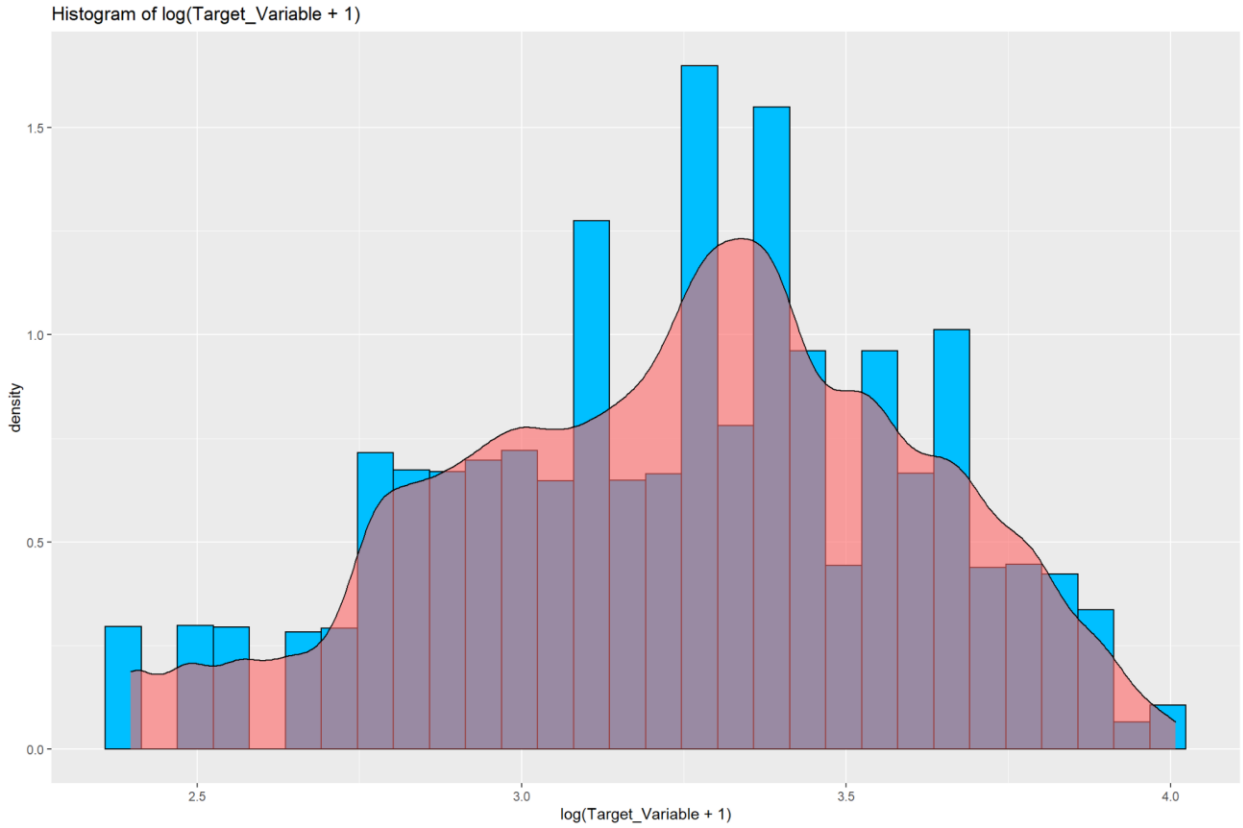
Exploratory Data Analysis

Analysis 1: Exploring the Target Variable

A histogram was drawn for the target variable to show the distribution. According to the histogram, the data appears to be right skewed.

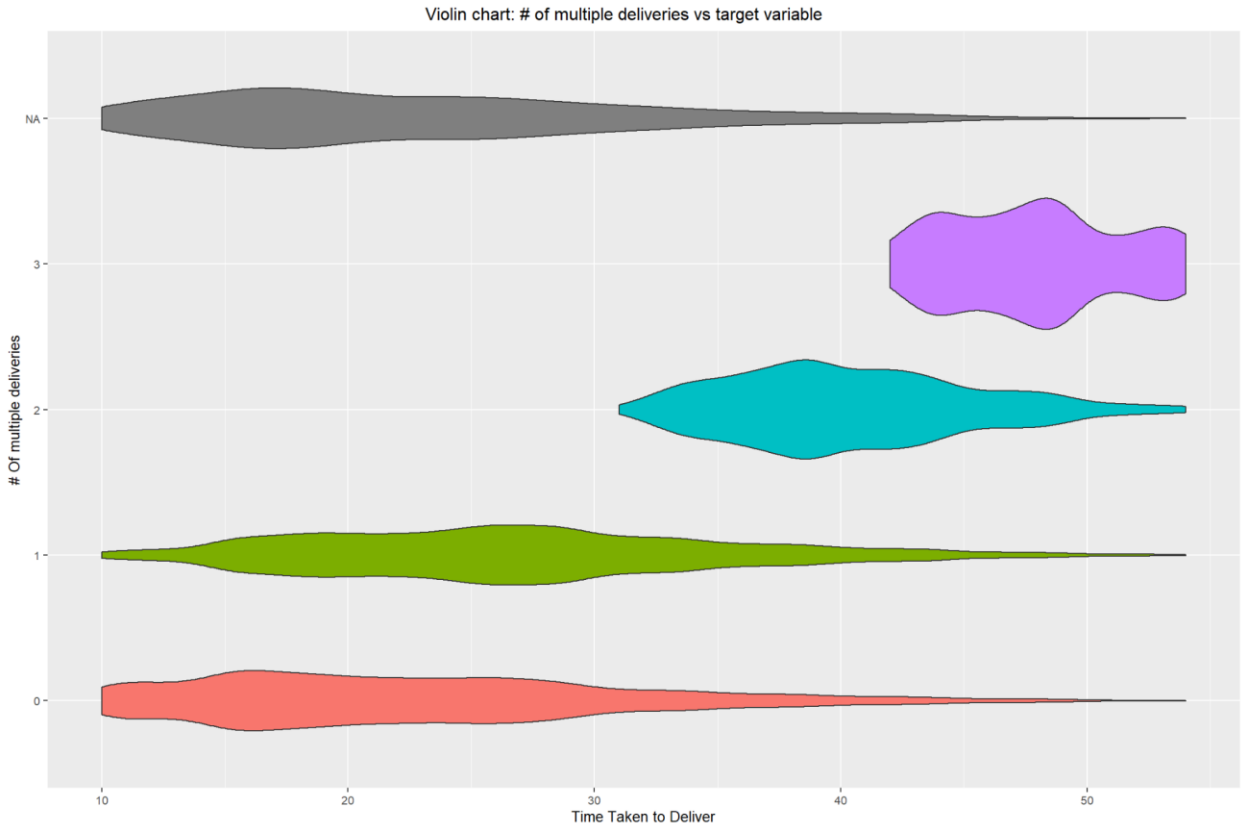


In an attempt to fix the right skewness, and to make the data look more normal, a logarithmic transformation was performed. The results are as follows:

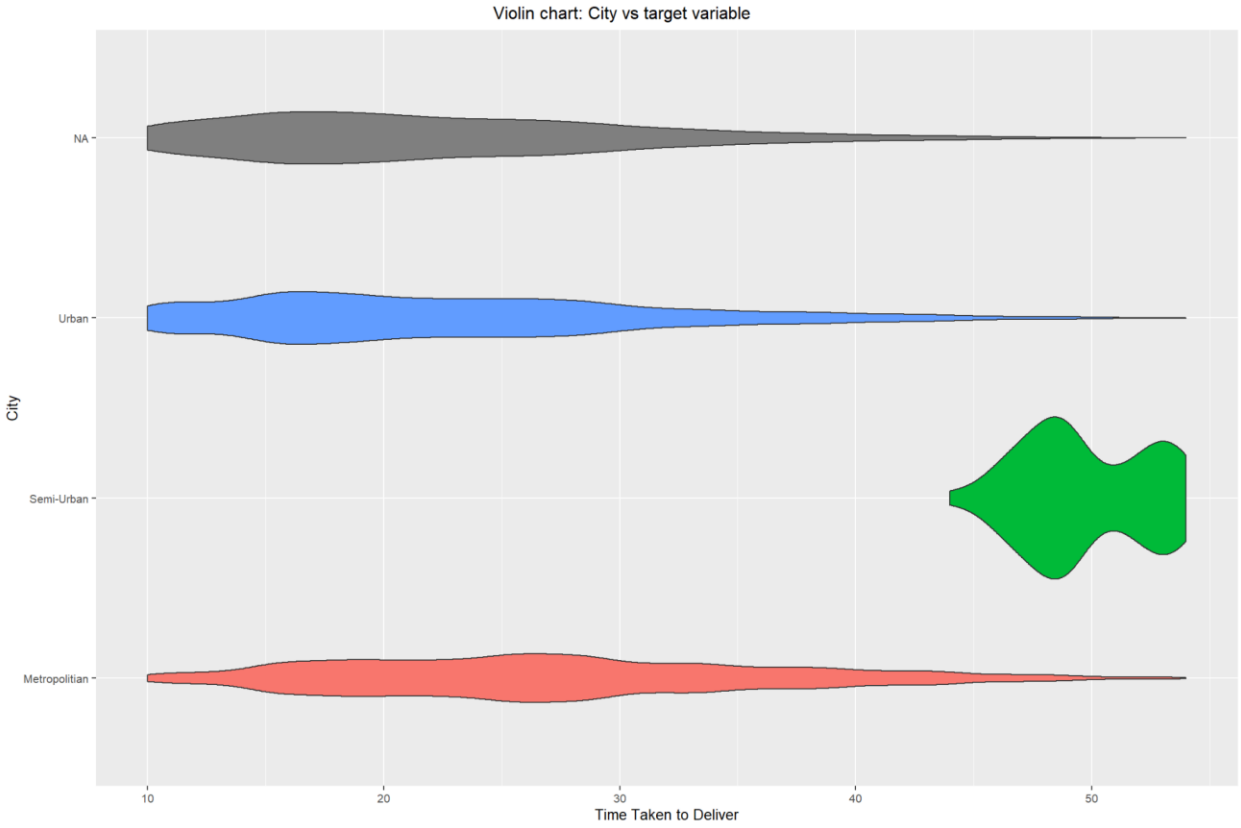

Histogram of log(Target_Variable + 1)
density
1.5
1.0
0.5
0.0
2.5
3.0
3.5
4.0
log(Target_Variable + 1)

Analysis 2: Visualizations of interactions between target variable and factor variables

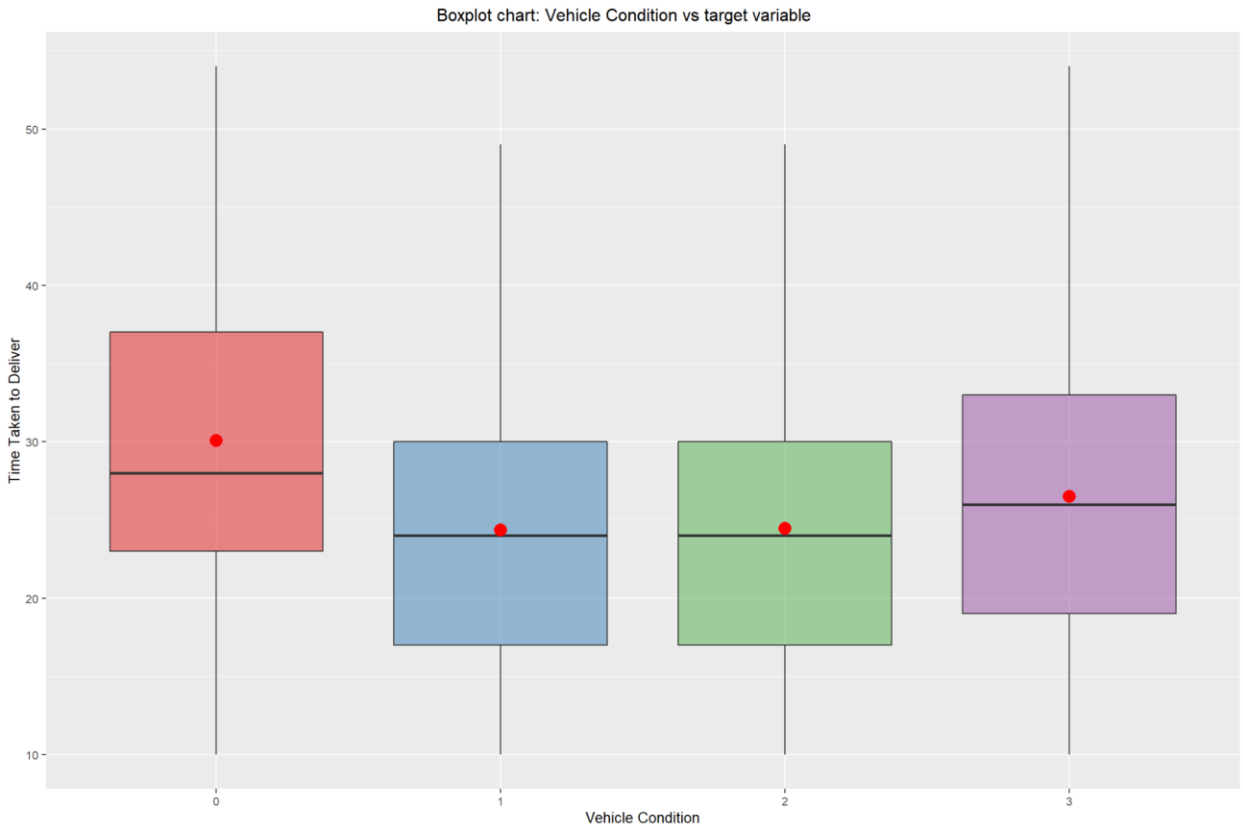
The violin chart of "*# of multiple deliveries vs target variable*" shows that the more deliveries the delivery drivers make, the higher the target values they achieve, i.e., the minutes taken for the delivery of food will increase if there are a greater number of deliveries assigned to the delivery drivers.



The Violin chart of "*City vs target variable*" indicates that Semi-Urban areas have the highest amount of the target value, i.e., the time taken to deliver food or drinks in semi-urban areas is the longest. It is a slim distribution, meaning it does not vary as much as the other factors.


Violin chart: City vs target variable
NA
Urban
Semi-Urban
Metropolitian
City
10
20
30
40
50
Time Taken to Deliver

The Boxplot of "*Vehicle Condition vs Target Variable*" shows that vehicle conditions 0 and 3 have a higher target value.



Analysis 3: Visualizations of interactions between Target variable and numeric variables

The first scatterplot drawn shows the relationship between multiple deliveries and the increasing amount of the target variable. Delivery drivers who make multiple deliveries tend to have more time taken to deliver food or drinks. The scatterplot also provides an overview of how the ratings of delivery persons are affected concerning the delivery time. The majority of the deliveries take between 15 and 30 minutes. Additionally, most of the delivery persons have a high rating of more than 4.

Another scatterplot was drawn with the city being used as the determination for color. This visual shows semi-urban having the highest amount of time taken for delivery, followed by urban and metropolitan areas.

Methodology

Data Cleansing

Overall, the data combines and cleans sixty-five thousand text files containing food delivery data to provide a framework for effective predictive modeling.

Missing Data

To handle missing data, the model imputed both factor and numeric data separately. To impute factor data, the algorithm called K-Nearest Neighbors uses a distance measure to infer missing values based on the five closest “neighbors” in the data. To impute the numeric data, the model leverages predictive mean matching, which creates a hybrid approach between regression-based imputation, while limiting the range of variation to the data in the training set. This

method maintains the variation in the training data while limiting the production of outlying data. Choosing these methods for factor and numeric data allows for completeness in the data so that the model can maximize all possible data for enhanced prediction.

Skews and Outliers

The model normalizes three skewed variables to better model the training and test data. Implementing the Box-Cox function for the Time_Ordered, Time_Order_Picked, and the Target_Variable helps normalize these distributions. Note that the Target_Variable experienced outliers, but the normalization transformation results in no outliers present. Please note that the location data relating to latitude and longitude have skewed distributions; however, since negative values persist, normalizing these distributions with the Box-Cox method is impossible, so the model does not transform these skewed columns. Overall, these transformations help enhance the model predictivity.

Factors

Since the factor data contains few unique values, the model does not "factor lump" the data. Factor lumping is the process of reducing the number of unique factors within a certain variable. If there were many unique values in related variables, then the model would factor lump to help fit the models more efficiently.

Description of Modelling Approach

The models that are used for this dataset are as follows:

- Ordinary Least Squares (OLS)
- Multivariate Adaptive Regression Splines (MARS)

- Elastic Net
- Principal Component Regression (PCR)
- Decision Trees using Classification and Regression Trees (CART) algorithm
- Random Forests
- Gradient Boosting

Using the OLS model, a stepwise variable selection process is carried out to simplify the model without impacting much of the performance. Except for OLS and PCR modeling approach, all other models are created for the data set using a 5-fold cross-validation technique for model validation and hyper tuning purposes. All the models use the same set of predictors on the training data set to determine the best model using Root Mean Squared Error (RMSE) and R^2 statistical measures.

Initial Results

Model	Notes	Hyperparameters	RMSE	Rsquared
OLS	lm	N/A	0.7861	0.551
MARS	caret and earth	Degree = 1 nprune = 23	0.761	0.5795
Elastic Net	caret and elasticnet	Alpha = 0.4 Lambda = 0.00111291986423209	0.7865	0.5508
PCR	pcr	ncomp = 15	0.7866	0.5508
CART	rpart	cp = 0.00655468073118936	0.8403	0.4873
Random Forest	rf	mtry = 6	0.6489	0.4642
Gradient boost	xgbTree	max_depth = 3 eta = 0.3 nrounds= 150	0.669	0.6751

The Random Forest model performs better as compared to other models based on the RMSE value of 0.65 which explains 46% of the variance in the target variable using the predictors of the data.

Future Scope of Work

To validate whether the RMSE value can be further reduced, Neural Network will be implemented and then the model with the least RMSE score will be fitted to the test set.