**Predicting Time of Arrival for Food Delivery Service**

Daniel Carpenter, Sonaxy Mohanty, Zachary Knepp

University Of Oklahoma

Data Science and Analytics Institute

DSA 5103-995 | Intelligent Data Analytics

Fall 2022 | Charles Nicholson

Table of Contents

[**Executive Summary** 3](#_Toc118966463)

[**Problem Background** 4](#_Toc118966464)

[**Results** 7](#_Toc118966465)

[**Conclusion** 8](#_Toc118966466)

[**References** 9](#_Toc118966467)

## **Executive Summary**

(1 page)

**Problem Statement**

This analysis aims to predict estimated delivery times for a food delivery service.

**Model Assumptions**

**Summary of Findings**

**Recommendations**

## **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 are 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**

*INCLUDE EXPLORATORY DATA ANALYSIS HERE*

**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 boxcox 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 boxcox 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.

**Modeling Choices**

**Model Validation Methods (e.g., 5-fold CV)**

## **Results**

**Model Performance Summary**

**Key Findings of Analysis!**

## **Conclusion**

Summary of problem, approach, findings

Key issues, limitations, etc.

Final recommendation

## **References**

**Appendix**

Data visualizations, tables, transformations, etc. which support the work, but are not of primary importance

Important code excerpts or algorithms used / developed (if any).