PREDICTING TOTAL DELIVERY PERIOD

Toyosi Bamidele

Date: 01/25/2021

Table of Contents

I. Problem Statement

This machine learning exercise aims to predict the estimated time taken for delivery(total_delivery_period). Understanding that extraordinarily early or too late orders has a much worse impact than slightly early/late, the prediction approach involves getting a clear understanding of each feature and its contribution to predicting the total delivery period.

Features that contribute to the model's strong predictive power enable much more exact delivery times; resulting in improved customer and driver experience, customer retention, and increased driver referrals to join dasher program, resulting in an overall revenue increase.

Exploratory Data Analysis

The introduction of the following features below showed some improvements towards model performance.

1. Total Processing Time

By examining the dataset, it is clear that the total processing time would be a critical factor in determining the entire delivery duration; the only times accounted for were the estimated time for the restaurant to receive the order from DoorDash and the estimated travel time between the store and consumer.

The absence of these datapoints leaves out a significant portion of time, such as how long the restaurant takes for food preparation, the time duration from order pick-up to the dasher commencing their journey for delivery, and how long it takes dasher to locate the customer and vice versa, and any other time conditions that might affect the trip.

With this understanding, in a bid to improve model performance, additional features such as the mean and median processing time are included in the dataset.

2. New Date Attributes

Exploratory data analysis of the date attribute revealed a pattern of the count of orders depending on the days' time. Features such as the hour of delivery(Fig 1.1) and part of the day show the orders' spread throughout the day(Fig 1.2). The overall trend shows a higher frequency of orders between late night to dawn hours and lower frequency of orders in the early morning and early afternoon which could impact delivery times depending on the number of dashers available

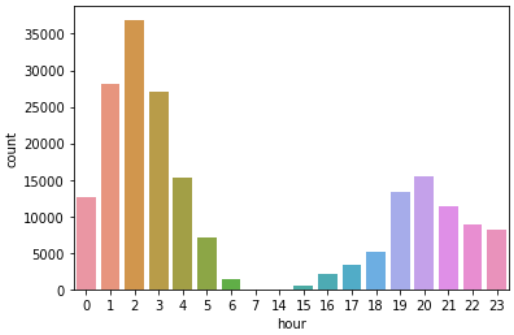

Fig 1.1-hour frequency

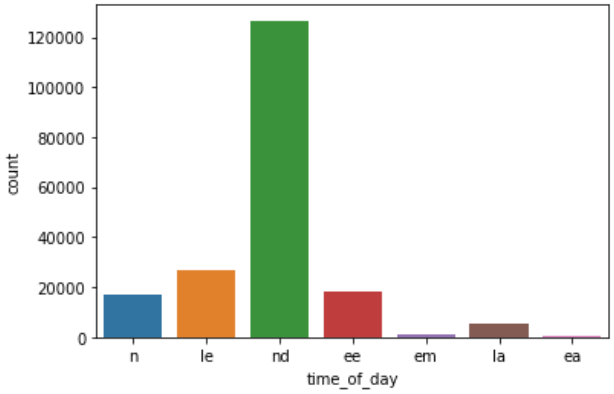

Fig 1.2 part of the day frequency

Key for Fig 1.2

1. nd night_dawn - 12am - 5am
2. em early morning - 5am - 9am
3. lm late morning - 9am -12pm
4. ea early afternoon - 12pm - 3pm
5. la late_afternoon - 4pm -5pm
6. ev early evening - 5pm -7pm
7. le late evening -7pm -9pm
8. night- 9pm -11pm

3. Scheduled Orders

With the assumption that the total processing time should take less than or equal to 3600 secs (1hr), if on the same-day delivery, it showed that a portion of the training data (7.4%)(Fig 1.3) had a total processing times greater than 1hr. This process enabled the data to split into scheduled or unscheduled orders. In this process, the trade-off is that any order scheduled between 30 mins to 1hr is categorized as not scheduled. given the available information and an understanding of general average delivery times, the indicated threshold is 1 hr. Overall, the purpose of introducing these new features enable a better prediction towards model performance improvements.

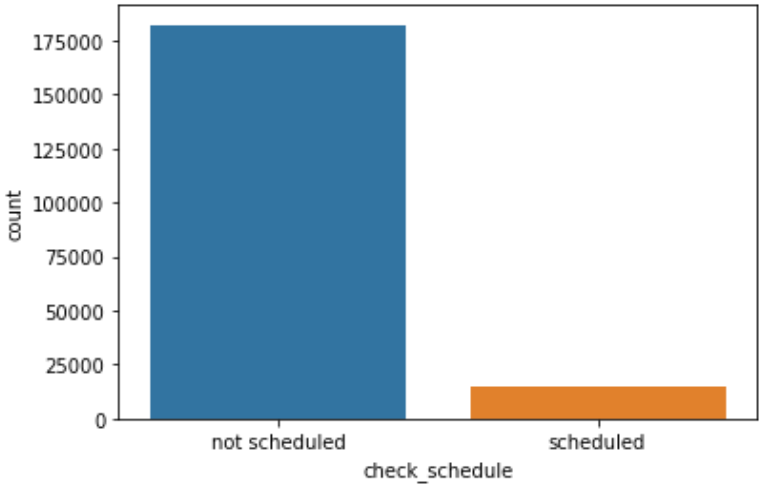

Fig 1.3 scheduled vs not scheduled frequency

Feature Selection

Selecting Top features

The best features selected are on the basis of an f-test using an f-regression function that ranks features by how significantly it improves the model with respect to the p-value.

The top features are:

- Schedule attributes
- Processing time attributes
- estimated_store_to_consumer_driving_duration'
- subtotal
- total_outstanding_orders
- hour
- 'time_of_day_nd'

Feasibility of model performance

Using the train data and extracting a test set, the model for feasibility reveals the following results:

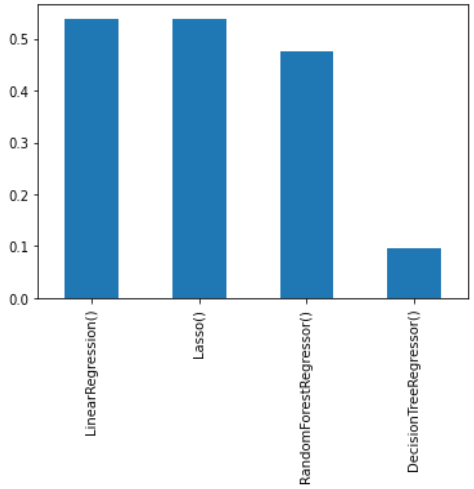

Fig 1.4 R squared

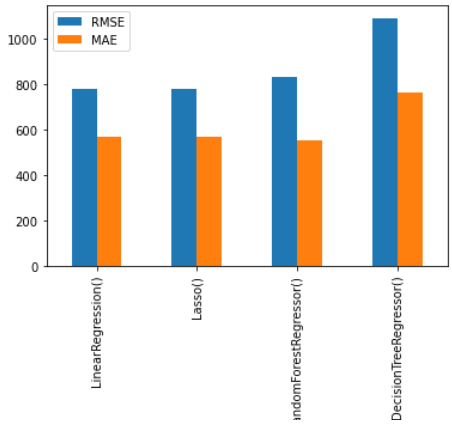

Fig 1.5 RMSE and MAE model

Following the data preprocessing step and feature selection, the model revealed linear regression models followed by the random forest model as a better fit for the in terms of R squared(Fig 1.4) and RMSE (error term)(Fig(1.5)).

II. Feature Recommendations

The following are the five features recommended for inclusion in the training dataset.

1. Scheduled vs. not scheduled
    a. These feature's give a clear understanding of which data points will have a longer processing time and create better separability within the dataset.

2. Processing time
    a. This accounts for every block of time from order creation to delivery that will enable a more definite value for the total delivery period.
    b. The times to account for are as follows :
        a. Time from order received to order ready
        b. Time from order pick up to delivery commencement
        c. Time from delivery location located to order received by the customer

3. Time of the day
        a. This shows how many orders are placed during the hour. The volume of dashers available in comparison to the number of orders could also impact how fast orders will get delivered.

4. Ratings and Reviews (Sentiment data)
    a. Pattern/Behavioral data by customer id, store id and driver id can suggest the rating of the restaurant, customer, and the driver. Having these data points can indicate if there is a behavioral pattern for these three categories, and whether orders are:
        1. early/ late/ on time to be prepared
        2. early/ late/ on time to the reached customer location
        3. early/ late/on time to be received by the customer

5. Location and Weather Conditions (Sentiment data)
    a. This can enable the driver to get to the delivery point on time; all factors below will enable an incremental time or shortage based on these details below.
        1. Information such as road diversion, accidents or general road conditions
        2. Optimal delivery route
        3. Temperature data

III. Assessing Model Performance

Performance measurements can be obtained by comparing the predecessor's predicted total delivery period and the actual total delivery period and calculating an error term. My model's fit can be checked by obtaining its predicted delivery times given the predecessor model's same data points.

Its error term can be obtained in comparison to the actual delivery periods. Suppose my model produces a lesser error term (i.e., RMSE and MAE, that shows how close the model fits to the data is to the actual data points). In that case, it can be inferred that it will be better than the predecessor model given the available information and data points.