Advanced Machine Learning Coursework: Kaggle Competition NYC Taxis

The Unicorn Hunters
(Christian Frantzen, Guilherme Barreiro Vieira,
Mortimer Sotom and Théo Verhelst)

May 8, 2018

Abstract

This is the abstract

1 Introduction

2 Data Analysis

The first step of a machine learning project is to explore and analyse the data, in order to better understand the problem. Our dataset is composed of 11 features: a unique identifier for each row, the identifier of the taxi company, the pickup time, the pickup and drop-off locations, the number of passengers, and a boolean flag indicated if the trip data has been stored on-board or directly sent to the data server.

Figures 1 to 5 show the distribution of some of these features. We did not show the pickup month, minute and seconds, as their graphs are uniformly distributed and therefore not visually informative. Outliers have been removed in order to properly display the pickup and drop-off location distribution, as well as the trip duration. These outliers will be discussed in the next section. We can see in figure 1 and 2 that the location seems to be roughly normally distributed, and the logarithm of the trip duration also appears to be normally distributed in figure 7. The smaller bumps outside of the main bell in the location curves correspond to trips to the city airport.

It is also important to make sure that the training and test sets are independently and identically distributed. For this, we compared the above distributions with those from the test set. If the distribution from the training and testing set significantly overlap, then we can consider that the I.I.D. assumption is verified. As shown for example in figure 8 and 9, it is indeed the case.

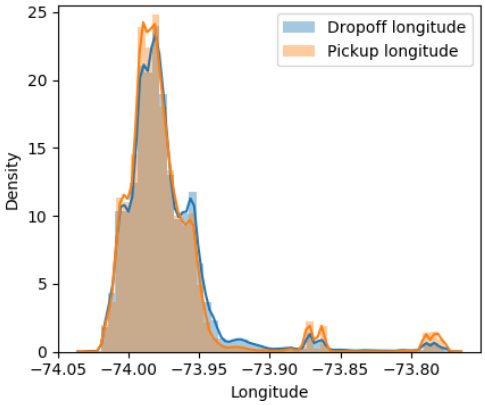

Figure 1: Distribtion of the longitude.

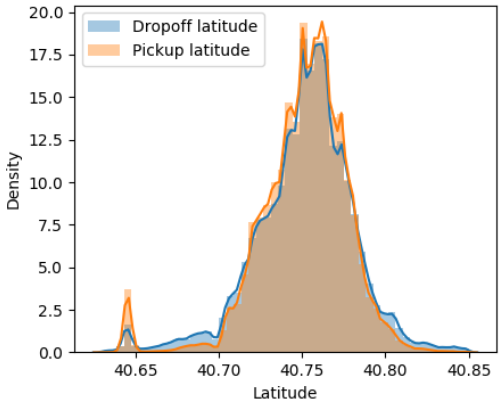

Figure 2: Distribution of the latitude.

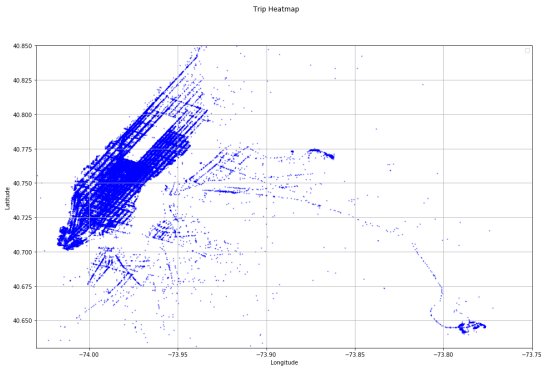

Figure 3: Trip locations as distributed on a map.

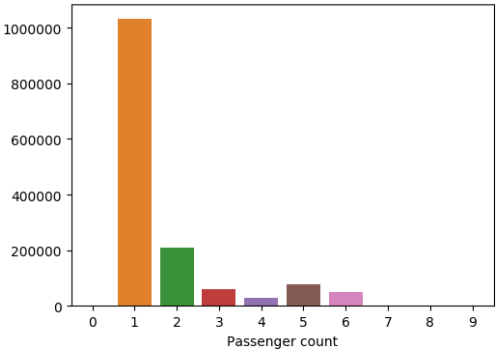

Figure 4: Distribution of the number of passengers.

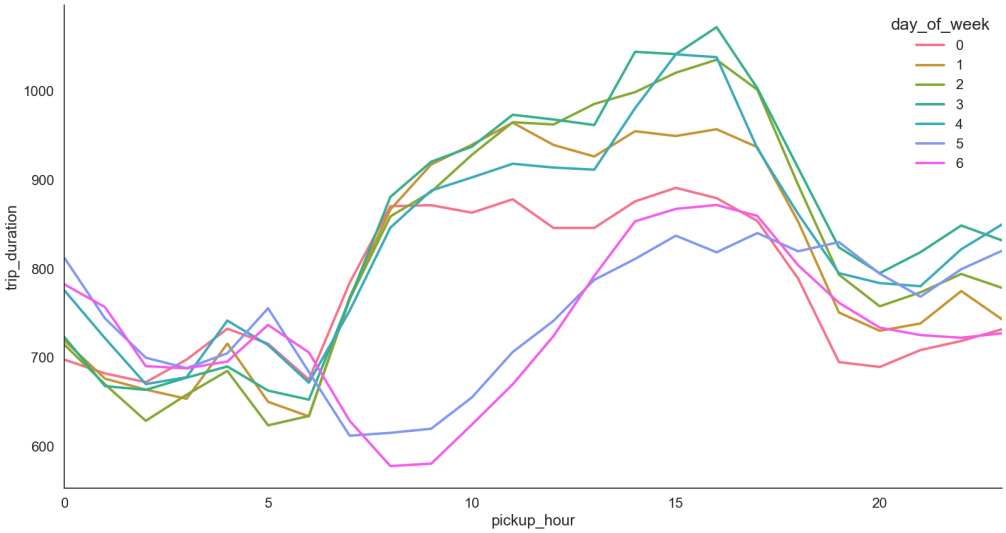

Figure 5: Distribution of the pickup time, for different days of the week.

3 Preprocessing

4 Feature Selection

In order to ensure that the model was not being fed useless features, a feature selection elimination procedure was undertaken. The results are shown in the above figure - the lower the ranking the better the feature is - by using the scikit-learn package recursive feature elimination. This function attempted to find a ranking of the features by doing a prediction while removing one feature at the time. As expected the most important features such as drop-off and pickup locations as well as different measures of distance were ranked as very important. In contrast, features that didnt contain much information such as `pickup_year`, `store_and_fwd_flag`, `vendor_id` and `pickup_month`, were ranked as very low importance and hence were removed from the model. Interestingly enough, the precipitation feature providing an idea of the weather conditions, didnt rank very low leading to believe that it didnt have a significant effect in the duration of the trips within New York and specifically Manhattan which is where most of the journeys were recorded.

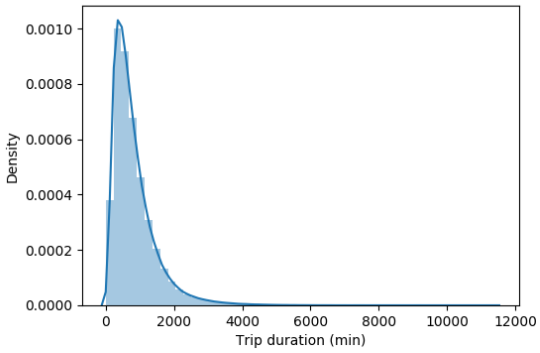

Figure 6: Distribution of the trip duration, in seconds.

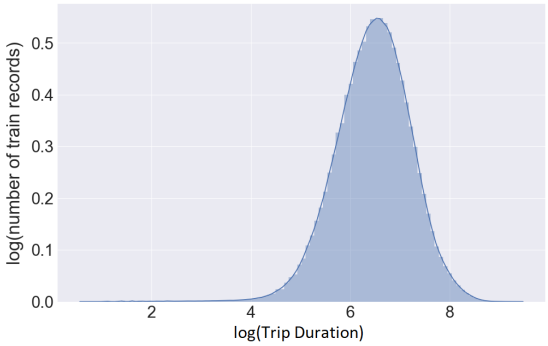

Figure 7: Distribution of the logarithm of the trip duration, in seconds.

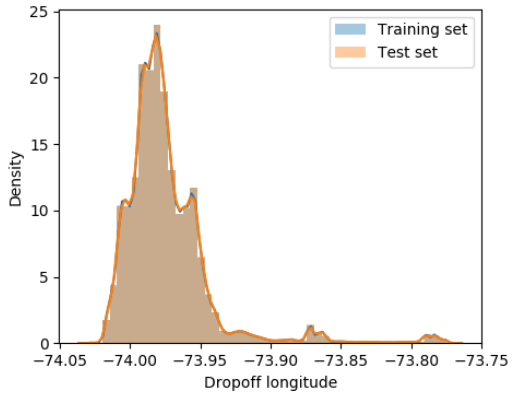

Figure 8: Drop-off longitude in the training and test set.

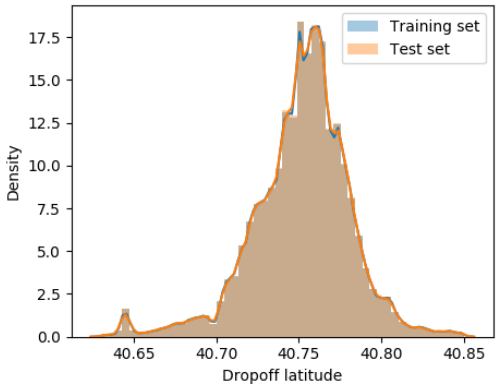

Figure 9: Drop-off latitude in the training and test set.

5 Methods for Model Selection

6 Results

7 Conclusion

References

A Additional Figures