Data Analysis and Prediction of New York Taxi Trip Duration Using Machine Learning Models

Saumya Gautam
MCM1, School of Computing
Dublin City University
Dublin, Ireland
saumya.gautam3@mail.dcu.ie

***Abstract*—There have been a lot of studies on how to predict the trip duration of the cabs running in a city. But in this study, we compared the efficiency of different predictive models and the effects of different features on the prediction. This motivation was to study and come up with best model to the predict trip duration. In this paper, we analyze multiple models to compare and predict the duration of taxi rides from one area of a city to another based on the recent demand and other relevant information. Remembering information that has been collected from the past is critical here, since taxi requests in the future are correlated with information about actions that happened in the past. To analyze this information, the data is used from the New York city taxi service to predict the duration of each taxi ride. This data can be used by taxi vendors for better services to the users. The research work not only uses a prediction model but also gives an in-depth analysis of the factors associated with the New York City taxi trips. A city like New York is expected to have various factors and variations with respect to the trip durations. The dataset used for training and testing purposes in multi-dimensional and requires a lot of pre-processing. This research work involves application of relevant machine learning algorithms such as linear regression, random forests, etc. and the link for the same can be found on Github.**

***Index Terms*— trip_duration, Linear Regression, Decision Tree, Random Forest, RMSE, R^2**

I. INTRODUCTION

In major cities, the major concern always begins with the issues related to travel and travel time. There are moments where passengers are waiting for cabs and none-available and sometimes the same cabs are usually empty. This is due to mismanagement and less knowledge of the location which demand the cabs at a particular time.

Predicting taxi duration and demand throughout a city can help to organize the taxi fleet and minimize the wait-time for passengers and drivers.[1] Method to employ a learning model is based on historical GPS data in a real-time environment. This GPS data provides the start and end coordinates to provide the distance travelled by the cab along with the areas from which the passengers usually demand the cabs. This data can further help in matching the demand and supply of available cabs in the area.

Our approach consists of comparing multiple predictive models and finding out the most efficient predictive model to help solve the cab demand imbalance problem. Given the pick-up and drop-off location, our goal is to predict the travel duration based on the dataset X that contains the taxi travel records of the New York City. Other information may also be introduced to dataset X, e.g. weather, holiday and distance. Our main focus is to find the best model.

This paper has further been divided as follows:

Exploratory analysis of the dataset. Here dataset explored using test and train data.

Methods used - Multiple regression models were used here for prediction of Trip duration.

II. RELATED WORK

Luis Moreira-Matias, João Gama, Michel Ferreira, João Mendes-Moreira, and Luis Damas suggested informed driving is the key feature for the increase in the sustainability of taxi companies. the data obtained from the sensors were used for 2 time series forecasting techniques to originate the prediction.[1] the results provided the framework for a important insight into spatiotemporal distribution of demand and duration. To do so, well-known time-series forecasting techniques were used and adapted to this problem, such as the time-varying Poisson model and the autoregressive integrated moving average (ARIMA).

Further to this, an experiment was carried out by Yongxin Tong, Yuqiang Chen[2], where their theory is based on to precisely balance the supply and the demand of taxis, online taxicab platforms need to predict the Unit Original Taxi Demand (UOTD), which refers to the number of taxi-calling requirements submitted per unit time.[2] They proposed LinUOTD, a unified linear regression model with more than 200 million dimensions of features. The simple model structure eliminates the need of repeated model redesign, while the high-dimensional features contribute to accurate UOTD prediction.

Mohammad Nozari Zarmehri [3]and Carlos Soares recognized A data mining approach can be used to generate models for trip time prediction. but, generating a separate model for individual taxi could be very expensive, hence the authors propose an algorithm that predicts a model for each taxi type.

Model output: { T_i , *j, G*}

Where T_i is the taxi, *j* being the recommended level and *G*

being the recommended algorithm.

With the increase of quality of results from neural networks, authors Xian Zhou, Yanyan Shen, Yanmin Zhu, Linpeng Huang (2018) propose an end-to-end deep neural network solution to the prediction task.[4] They employ the encoder-decoder framework based on convolutional and ConvLSTM units to identify complex features that capture spatiotemporal influences and pickup-drop-off interactions on citywide passenger demands. A novel attention model is incorporated to emphasize the effects of latent citywide mobility regularities. Their predictions provide a much efficient result then the traditional prediction framework.

With Traces study in picture, Marco V and Carlos B.[5] used naïve Bayesian classifier carry out the analysis of predictability of taxi trips for the next pick-up area type given history of taxi flow in time and space. They explore the relationships between pick-up and drop-off locations; and analyze the behavior in downtime (between the previous drop-off and the following pick-up).

III. DATASET AND EXPLORATORY ANALYSIS

3.1 Dataset Information

To get a better understanding of the topic, we visited multiple taxi and transport datasets of different counties and cities such as LA, Melbourne, etc. This is when I came across the current dataset. The dataset is based on the New York Trip records of 2016. This data was published by the NYC Taxi and Limousine Commission (TLC) and was made publicly available for analysis on Kaggle.[6] The dataset consists of train and test.csv big data files with 2 million records combined divided in 70-30 ratio. Each csv contains a unique identifier for each trip, the vendor code associated to that trip, the pick-up and drop off date and time. The data also includes the number of passengers in that trip present and the location details of their pickup point and drop off point. Each trip unique trip has its trip duration mentioned. The data set is relatively clean with minimal missing values.

3.2 Exploratory Analysis

3.2.1 Basic Statistical Properties

From early exploration we observed that there are no missing observations to be imputed and the dataset is quite a clean one. Basic statistical summary, such as minimum, maximum, mean, count, standard deviation of the dataset for cab trips for the year 2016 is given in Figure 1.

Variable	Count	Mean	SE Mean	StDev	Minimum	Maximum	Median
vendor_id	1458644	1.535	0.000413	0.4988	1	2	2
passenger_count	1458644	1.6645	0.00109	1.3142	0	9	1
pickup_longitude	1458644	-73.973	0.000059	0.0709	-121.933	-61.336	-73.982
pickup_latitude	1458644	40.751	0.000027	0.0329	34.36	51.881	40.754
dropoff_longitude	1458644	-73.973	0.000058	0.0706	-121.933	-61.336	-73.98
dropoff_latitude	1458644	40.752	0.00003	0.0359	32.181	43.921	40.755
trip_duration	1458644	959	4.34	5237	1	3526282	662

Figure 1. Statistical Summary

3.2.2 Outliers

After analyzing data based on Mean, std deviation and Min and Max observations, we identified the outliers for the trip Longitude and Latitude Coordinates, trip duration, and passenger count.

Passengers: For the passenger count we observed that values were from 0 to 9. The value '0' does not make any sense from business case point of view. Also, the maximum value '9' is also highly unlikely since the taxi is usually are for 5 adults and exceptions can be for children aged under 7 who can sit in lap. These observations are most likely to be errors which we need to remove.

Latitude and longitude Coordinates: Based on the various New York City (NYC) coordinate projections, the latitude and longitude ranges are as follows:
Latitude is between 40.7128 and 40.748817
Longitude is between - 74.0059 and - 73.968285
But the Max and Min for pickup- and drop-off coordinates fall outside the range stated above. We can exclude these as well since the study is based on NYC only.

Trip Duration: The Max trip duration is around 3526282.00 seconds i.e. 980 hours approx, which is clearly unrealistic and shows that outliers are present. Such outliers were removed by excluding all data points which were more than two standard deviations away from the mean trip duration time.

3.2.3 Data Pre-processing and Feature Engineering

Pre-processing:

Variables were reviewed for their datatypes to see if they were correctly. The pickup and drop-off timestamp variables are being treated as non-null objects. These features needed to be specified as date objects for further feature engineering and analysis.

Feature-Engineering:

The pickup_datetime and dropoff_datetime variables both combine date and time information into the same column. For our study, we delimited this information into additional features such as **month**, **week**, **day** and **hour** because the month, or hour or day of pick up may influence the trip duration in response to underlying trends.

The pick-up and drop-off coordinates were used to calculate the **distance** of the trips using **Haversine formula** and **Euclidean distance formula** to get the distance and bearing between 2 GPS points. Distance can also be key factor to affect the trip_duration as more the distance, more the duration.

Target Variable:

The target variable we will be predicting is 'trip_duration'. We checked its distribution to see if there are transformations that need to be applied. Figure 2 shows that trip_duration is highly skewed to the right as evident by the skewness value > 1.0 and long right tail.

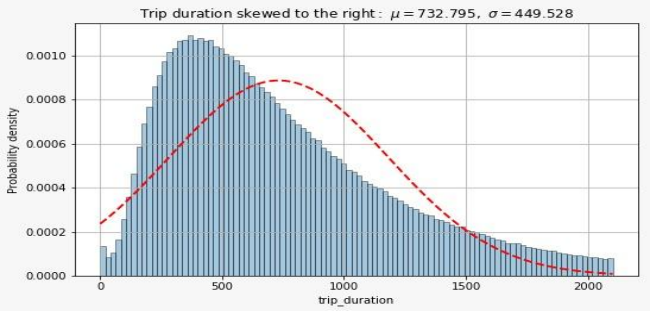

Figure 2. Target Variable Skewness

This skewness is due to the trip_duration observation as high as 3 hours which were not removed previously as they were still within two standard deviations the mean.
To normalize the distribution, log transformation was applied to trip_duration.

Feature Variables:
The variables that will be fed into our machine learning model to predict the dependent variable 'trip_duration' are the feature variables. All these variables were explored to understand them and to see if any transformations are required before proceeding with machine learning.
Id: The id variable is a unique identifier of each trip.
vendor_id: There are two vendors 1 and 2 where vendor 2 is slightly popular than vendor 1, although the difference is not significant and trips durations are also similar for both the vendors.
Store and fwd flag: This was checked to see if store and forward trips were contributing to trip_duration outliers for the vendors. However, server connections did not have much contribution to the outliers.
Passenger Count: Passenger count varies from 0 to 9, where 0 and 9 are considered as outliers and highest frequency is for 1 passenger. Although trip_duration doesn't vary much with passenger count increase.
Trip_duration by hours and day: Trips tend to be longer during office hours (8 am to 6 am), in evening (6pm to midnight) from Thursday to Saturday as people go out during weekends and early Saturdays and Sundays.
Trip_duration by Months: We checked this for seasonality but Trip_duration doesn't vary much with months.
Distance: Distance variable derived from feature engineering may directly impact the trip_duration.

3.2.4 Patterns or Trends Observed
After doing the exploratory analysis and feature engineering, below patterns were observed:

1. Although minimum and maximum number of passengers is from 0 to 9 which we ruled out as outliers, the max count is for 1 passenger per trip.
2. Day of the week had slight effect on the number of trips with Monday reporting least number of pickups and Friday with highest number of pickups.
3. There was noticeable trend in number of pickups with respect to pick up hour. For weekdays, the count was least in early morning from hours 3am to 5am, which then starts to grow at 6am and increases with office hours and stays almost same during the day and then increases to highest at 6pm in the evening and keeps busy till 11pm.
4. Over the weekends, there are peaks at mid-night on Saturday and Sunday as people usually go out to parties.

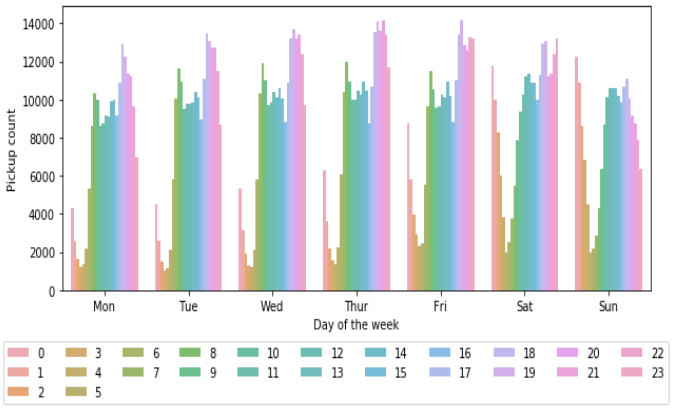

Figure 3. Pickup Count based on Week days and Pick Hours

5. When we look at months, there is no significant change in pick up count, only for one dramatic drop in pickup counts in January end or February start which could be assumed as seasonality, but the decline is specific to a date. This could be an outlier as a result of erroneous entry or some event that happened hat day. Since the data is of 6 months, there is not much to say about the trend as it does not vary much.

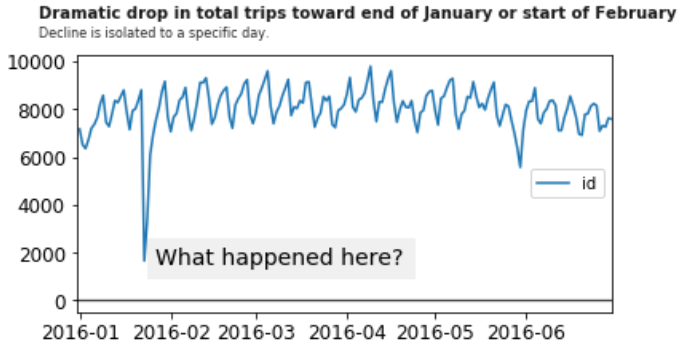

Figure 4. Pickup Count over Time

After analyzing dataset features and target variable, we need to ensure the data is 'model ready' by checking data for missing values and removing unnecessary features. We split the data into training and test sets and feed the sets into few regressions algorithms to determine which model is best to predict the target variable 'trip_duration'.

IV. METHODOLOGY

4.1 Linear Regression

Linear regression is the most well-known regression learner. Following assumptions were made while using this model:

- There is linear relationship between the independent and dependent variables.
- The variables (features) are normally distributed. If not, a non-linear transformation (e.g., log-transformation) may be needed to fix the issue.
- No or little multicollinearity: The independent variables (features) are not highly correlated with each other.
- Residuals are independent of one another.
- Residuals are equal across the regression line.

4.2 Decision Tree

Decision Trees (DTs)[7] are a non-parametric supervised learning method used for classification and regression. It works by creating a model that predicts the value of a target variable by learning simple decision rules inferred from the data features.

Some advantages of using decision tree are:

- it is the easiest to interpret
- does not require feature scaling
- computationally less expensive than other methods

4.3 Random Forrest

A Random Forest[8] is an ensemble technique capable of performing both regression and classification tasks with the use of multiple decision trees and a technique called Bootstrap Aggregation, commonly known as **bagging**. The basic idea behind this is to combine multiple decision trees in determining the final output rather than relying on individual decision trees.

Advantages of Random Forrest are:

- Random Forest increases predictive power of the algorithm and helps prevent overfitting.
- Ability to handle multiple input features without need for feature deletion.
- Prediction is based on input features considered important for classification.

4.4 Performance Metrics and Model Comparison

Coefficient of Determination

The coefficient of determination, denoted R^2 or r^2 and pronounced "R squared", is the proportion of the variance in the dependent variable that is predictable from the independent variable. This correlation, known as the "goodness of fit. R-squared has the useful property that its scale is intuitive: it ranges from zero to one, with zero indicating that the proposed model does not improve prediction over the mean model, and one indicating perfect prediction.

Root Mean Squared Error (RMSE)

The RMSE is the square root of the variance of the residuals. It indicates the absolute fit of the model to the data–how close the observed data points are to the model's predicted values. Whereas R-squared is a relative measure of fit, RMSE is an absolute measure of fit.

RMSe is defined as:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N} (Predicted_i - Actual_i)^2}{N}}$$

4.4 Model Comparison and Deployment

Comparison: First, the dataset with no new features was fed into the models and performance was compared with when training dataset with added new features was fed into the models. After comparing the results of models with and without feature engineering, all the models were compared based on their respective performance metrices and the model with best performance was used for prediction.

Deployment: The test data was pre-processed in same manner as training data before feeding into model. Feature Engineering was applied on pickup_datetime variable of test data to derive 'date', 'time', 'month', 'hour' and 'day' features. Similarly, pickup_location and dropoff coordinates were used to calculate 'distance' feature. After processing test dataset, the data was fed into the Random Forest Model for predictions.

V. RESULTS AND FINDINGS

5.1 Checking performance with Feature Engineering

R^2 values of Linear Regression, Decision Tree and Random Forest came out be 0.048, 0.405 and 0.68 when the training data was fed into models without features from feature engineering. These values were less when compared to model performance with new features added to training data.

5.2 Performance Metrics of the Models used

Models	Performance Metrics			
	R^2	**MAE**	**MSE**	**RMSE**
Linear Regression	0.440	0.430	0.332	0.576
Decision Tree	0.563	0.354	0.259	0.509
Random Forest	**0.780**	0.250	0.130	**0.361**

Table 2. Performance Metrics

5.3 Model Comparison

Linear Regression: As per R-squared, only 4.4% of the variation in the dependent variable is explained by this model. Such a low score is an indication that the relationship between the features and independent variable may be better explained with a non-linear model. This model has the highest RMSE value amongst all three models.

Decision Tree: R^2 value of the decision tree is higher than linear regression and almost 56% of the variation is explained by decision tree. But this model is still not a good fit for this data and RMSE value is still higher.
Random Forest: This model provides highest R^2 value and can explain 78% of the variation in dependent variable. The RMSE of this model is the lowest of all three models.

After comparing the R^2 value of all the three models, we decided to go with **Random Forest Model** because of its higher R-squared value compared to other two models. Random Forrest also provides many advantages over other models because of high predictive accuracy and efficiency with large datasets.

VI. CONCLUSION

The research problem was finding the trip duration based on the given pick-up and drop-off locations. Different patterns were observed and features in the dataset were explored and new features were extracted using Feature Engineering. It can be concluded that feature engineering adds to better accuracy and makes the prediction algorithm perform better.
On comparing various models for predicting the target variable 'trip_duration' on the train dataset, it can be concluded that a nonlinear model is best suited for this dataset. Hence based on a high accuracy received from coefficient of determination and a low error value, Random forest was applied on the test dataset and we successfully predicted the trip duration of each trip in that dataset. Prediction done on this dataset and the use of machine learning algorithm could benefit the concerned authorities to improve the traffic issues and bring in new factors that could affect the increase or decrease in the duration time.

VII. REFERENCES

[1] Luis Moreira-Matias, João Gama, Michel Ferreira, João Mendes-Moreira, and Luis Damas. On Predicting the Taxi-Passenger Demand:A Real-Time Approach. DOI: 10.1007/978-3-642-40669-0_6 Issn: 0302-9743

[2] Yongxin Tong, Yuqiang Chen, Zimu Zhou, Lei Chen, Jie Wang, Qiang Yang, Jieping Ye, Weifeng Lv. The Simpler The Better: A Unified Approach to Predicting Original Taxi Demands based on Large-Scale Online Platforms. KDD'17, August 13–17, 2017, Halifax, NS, Canada

[3] Mohammad Nozari Zarmehri and Carlos Soares. Using Metalearning for Prediction of Taxi Trip Duration Using Different Granularity Levels

[4] Xian Zhou, Yanyan Shen, Yanmin Zhu, Linpeng Huang. Predicting Multi-step Citywide Passenger Demands Using Attention-based Neural Networks. WSDM '18: Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining.

[5] Marco Veloso, Santi Phithakkitnukoon, Carlos Bento. Urban mobility study using taxi traces. TDMA '11: Proceedings of the 2011 international workshop on Trajectory data mining and analysis September 2011

[6] New York city taxi trip duration can be found at https://www.kaggle.com/c/nyc-taxi-trip-duration/data

[7] Pedregosa et al. Decision Trees, JMLR 12, pp. 2825-2830, 2011. https://scikit-learn.org/stable/modules/tree.html

[8] Breiman Leo, Cutler Adele. Random Forests https://medium.com/datadriveninvestor/decision-tree-and-random-forest-e174686dd9eb