Problem Statement

The Taxi Limousine Commission(TLC) provided yellow taxi service for the city of New York. But

these taxi services were concentrated to the Manhattan region of the city and did not ride to the

outskirts. The green taxi service was hence initiated to cater to the demands outside of Manhattan

which started in 2013. The data is vast and there are various conclusions we can draw from it. We

aim to maximize the revenue of the company by implementing various machine learning techniques

of supervised and unsupervised algorithms to identify the ideal model to the data.

Data

The dataset is about TLC green taxi service for years 2016 to 2018. The data is been consolidated

from NYC Taxi Limousine Commission for each of the above years. The data set consists of 19

columns, which consist of numerical and categorical values. There are total of 22.5 million instances

overall. The domain of the dataset is Public Transportation.

|  |  |
| --- | --- |
| **Header** | **Data Description** |
| VendorID | Indicates the LPEP vendor that provided the record  **1: Creative Mobile Technologies, LLC**  **2: VeriFone Inc.** |
| lpep\_pickup\_datetime | The date and time when the meter was first engaged |
| lpep\_dropoff\_datetime | The date and time when the meter was disengaged. |
| passenger\_count | The number of passengers riding. This is a driver-entered value |
| trip\_distance | The distance of the trip stored by the taxi meter |
| PULocationID | The taxi zone where the taxi meter was engaged |
| DOLocationID | The taxi zone where the taxi meter was disengaged |
| RatecodeID | The rate code that is in effect according to the type of ride:  **1: Standard rate**  **2: John F. Kennedy Airport**  **3: Newark Airport**  **4: Nassau or Westchester**  **5: Negotiated Rate**  **6: Group Ride** |
| store\_and\_fwd\_flag | This flag indicated whether the trip record was held in vehicle memory before sending to the vendor aka “store and forward” because the vehicle did not have a connection to the server.  **Y: store and forward trip**  **N: not a store and forward trip** |
| payment\_type | An integer indicating the type of payment method used by the customer.  **1: Credit Card**  **2: Cash**  **3: No charge**  **4: Dispute**  **5: unknown**  **6: Voided trip** |
| fare\_amount | The fare of the ride calculated by the taxi meter on the basis of distance travelled and time spent. |
| extra | Miscellaneous extras and surcharges like rush hour surcharge and overnight surcharge |
| mta\_tax | A $0.50 tax automatically added based on metered use |
| improvement\_surcharge | A $0.30 improvement surcharge assessed on hailed trips at the flag drop |
| tip\_amount | The tip amount which is automatically added in a credit card transaction. |
| tolls\_amount | Total amount paid to tolls on the trip. |
| total\_amount | Total amount charged to customers, not inclusive of cash tips. |
| trip\_type | An integer indicating the kind of trip automatically assigned to the ride based on the metered rate but that can altered by the driver.  **1: Street-hail**  **2: Dispatch** |
| congestion\_surcharge | A $2.75 charge on Green Taxis that pass through Manhattan south of 96th street |

**Data Preprocessing:**

1. Raw data is a monthly data which we consolidated by year.
2. Next, we combine all 3 years (June 2016 – June 2019) dataset into a single dataset which results into 22.5 million instances.
3. Valid Data:
4. ‘tolls\_amount’ – Since toll amount cannot be negative, we are filtering positive values for toll amounts.
5. ‘fare\_amount’ – Since the initial base fare charge is $2.5, we have taken fare amount values greater than or equal to $2.5.
6. ‘passenger\_count’ – Since an XL ride can take up to 6 passengers, we have limited the maximum passenger count to 6.
7. ‘RatecodeID’ –  Valid Rate code ID range from 1-6. Hence, a constraint has been put to remove any invalid category values
8. ‘trip\_distance’ – Rides whose trip distance lesser than or equal to zero are considered as cancelled rides. Hence, trip distances which are greater than zero are considered.

1. Date and Time format: ‘lpep\_pickup\_datetime’ and ‘lpep\_dropoff\_datetime’ store pickup and drop off date and time of every unique ride. Each of these columns were separated according to year, month, day, hour and minutes.
2. Dropped columns:
3. ‘ehail\_fee’ – This column consists entirely of zeros values which can be ignored.
4. ‘lpep\_pickup\_datetime’ – Since date and time of pickup of customer has been incorporated by the separate columns created for year, month, date and time, this column for removed.
5. ‘lpep\_dropoff\_datetime’ -- Since date and time of drop off of the customer has been incorporated by the separate columns created for year, month, date and time, this column for removed.
6. ‘store\_and\_fwd\_flag’ -- The store and forward flag indicates if the record was initially held in vehicle memory due to some connection issue. This column does not add any value to our research and hence it is dropped.

**Normalization:**

The continuous variables were normalized using the formula:

*Xnorm =*

**RESEARCH QUESTION 1**

**Which rides are the most profitable for the drivers on the basis of location, type of ride and time?**

The aim of this research question is to explain the prediction of the total amount (profit) by analysing how each feature affects the output instance. We judge the feature importance using the SHapley Additive exPlanations (SHAP) that uses game theory to interpret the model chosen. SHAP has two estimation approaches KernalSHAP and TreeSHAP. TreeSHAP is the estimation approach used to predict the Shapley values here as Tree based models as that would help us correctly estimate SHAP values when features are dependent. Also it os computationally less expensive when compared against KernalSHAP . The features can be interpreted on a global as well as a local level.

The model chosen to fit the data is Random Forest Tree Regressor, where our input features are 'VendorID', 'RatecodeID', 'PULocationID', 'DOLocationID', 'passenger\_count', 'trip\_distance',’duration’ , 'payment\_type', 'trip\_type', 'pickup\_year', 'pickup\_month', 'pickup\_day', 'pickup\_hour','pickup\_minutes', 'dropOff\_year', 'dropOff\_month', 'dropOff\_day', 'dropOff\_hour', 'dropOff\_minutes',’speed’.‘total\_amount’ is the output variable as is constituted of approximately the sum of the input features given below. Hence these features, 'fare\_amount', 'extra', 'mta\_tax', 'tip\_amount', 'tolls\_amount', and 'improvement\_surcharge', have been removed to build the model.

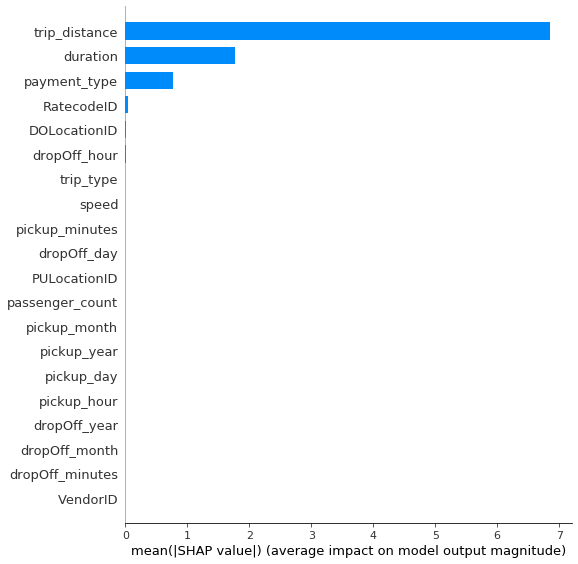
The model gives 89.55% accuracy when trained on data from 2018 and tested on 2019 data hence this model was chosen.

At global level, we get Shapley value matrix of equivalent size as that of your dataset. One value corresponding each data point, these values can be used to analyse the model.

* SHAP feature importance

The features are given importance based on their individual effect on the output variable, ‘total\_amount’. This summary plot depicts the global importance of every feature on the output. The y-axis has the features and the x-axis has the mean absolute Shapley values per feature.

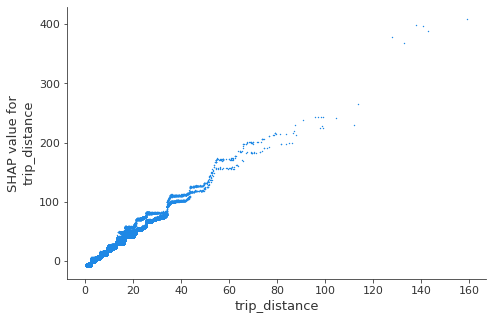
Ij= |ϕ(i)j|



* SHAP dependence plot

# A dependency plot is also displayed which shows the effect of a single feature on the predictions of the model. Since our summary plot indicates that ‘trip\_distance’ effects the total amount the greatest, a dependency plot of ‘trip\_distance’ is created.

[https://slundberg.github.io/shap/notebooks/plots/dependence\_plot.html]



RESEARCH QUESTION 3:

Model Used: Quantile Regression

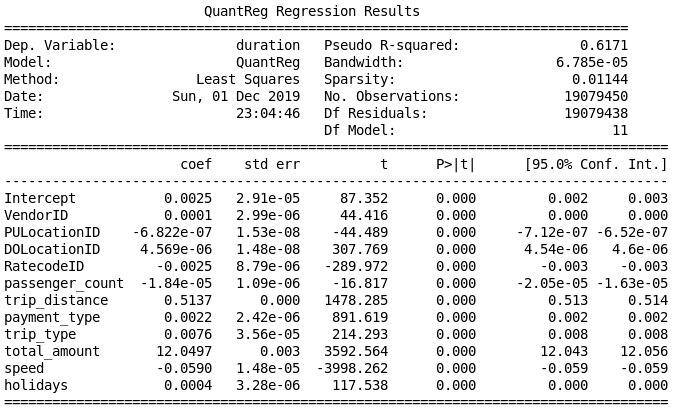
Linear Regression depicts the relationship between the dependent and independent variables, providing a mean estimate for the independent variable as a depiction of the strength of the model. While this has been a classic approach to understanding the relationship between the predictor and response, a mean estimate does not depict what is truly depict the what occurs in different ranges of data. Given our vast number of data points, dividing and analysing the data in different ranges was imperative. This is what quantile regression allows—we estimate coefficients of our model to estimate and conditional median.

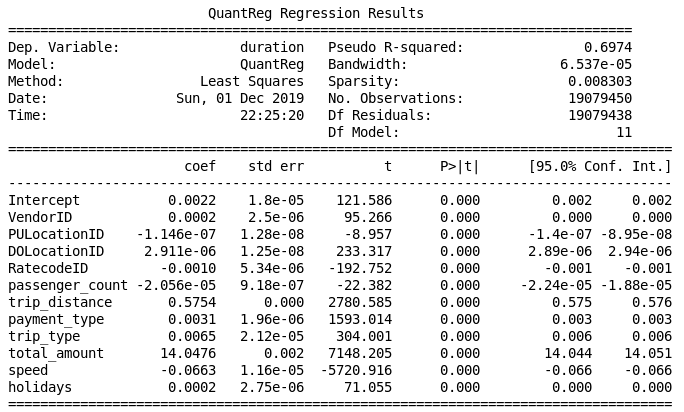
Through this research question we explored the relation between the trip distance and the trip duration, where the basic assumption is that a greater trip duration and lesser distances indicates greater traffic in the area. We see how shorter and longer trip durations are affected by trip distance and total amount (cab fare).

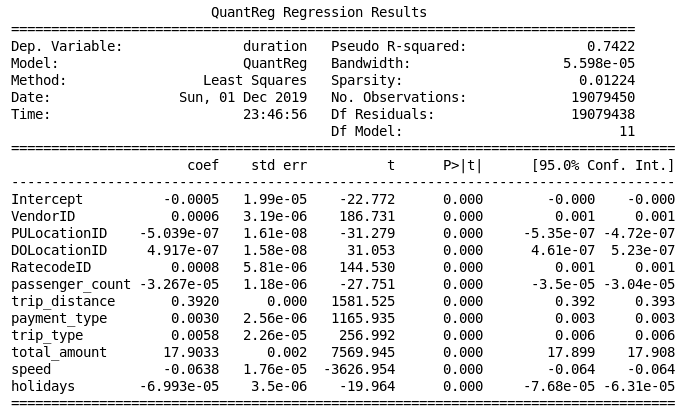
Analysis:

For initial analysis, the 25th, 50th and 75th quantiles regression models were created.

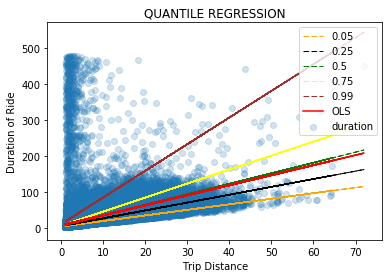
(The categorical variables were not one hot-encoded, instead considered label encoded as that they did not prove to contribute heavily to the model, and would add too many features considering the large number of groups with each categorical feature.)

0.25 quantile

0.5 quantile

0.75 quantile

Considering the fluctuating values of the coefficient of trip\_distance and most instinctive relationship between trip distance and duration, we further analysed the relationship between trip distance and different ranges in duration of ride. We compare the quantile regressions with the linear regression to draw conclusions.



We also see that the slopes low incline because of the majority of the trip durations are between 0-150.

The analysis indicates that the estimated mean and median (0.5 quantile) coincide. The different slopes with respect their quantiles indicate that different ranges of trip distances affect the trip distances differently. However we may conclude that for this data, linear regression maybe a suitable model in this case considering the coinciding of mean and median.