Resurgence of NYC Taxis

Apurva Sharma, Sanman Yadav, Shama Kamat, Shivanshi Bajpai
IST 718- Big Data Analytics, Syracuse University

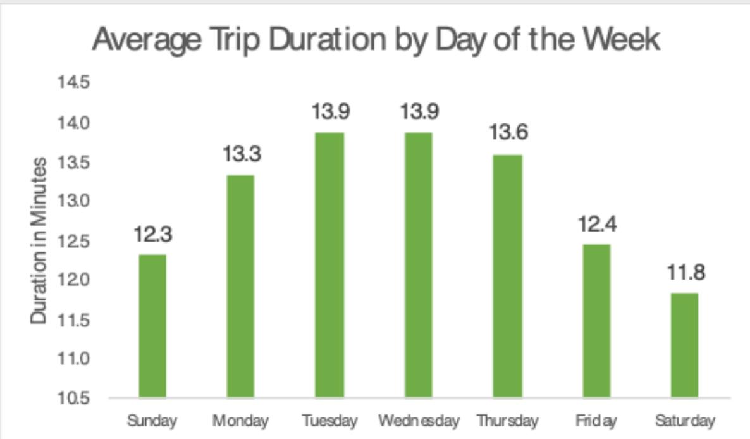
Problems and Objectives

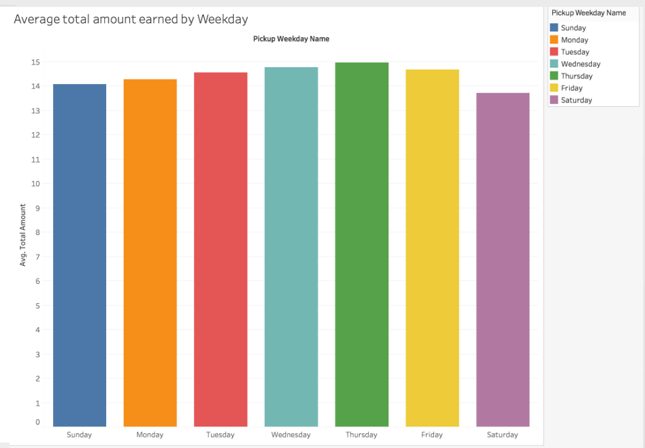
- Due to advent of smart cab companies such as Uber and Lyft, people are prefferring smart services over the traditional taxis in NYC
- The research throws light on the following important issues to be the reason in dip in the usage of Yellow taxis:
  - **Information Unavailability:**
    - The yellow taxi drivers do not get real-time updates on availability of passengers. The drivers spend majority of their time searching for customers
  - **Fare negotiation:**
    - Availability of smartphone apps for Uber and Lyft helps passengers to be informed about the ride prices
    - The yellow cab drivers spend time in making a proposal and waiting to make a deal post the effort of finding a customer

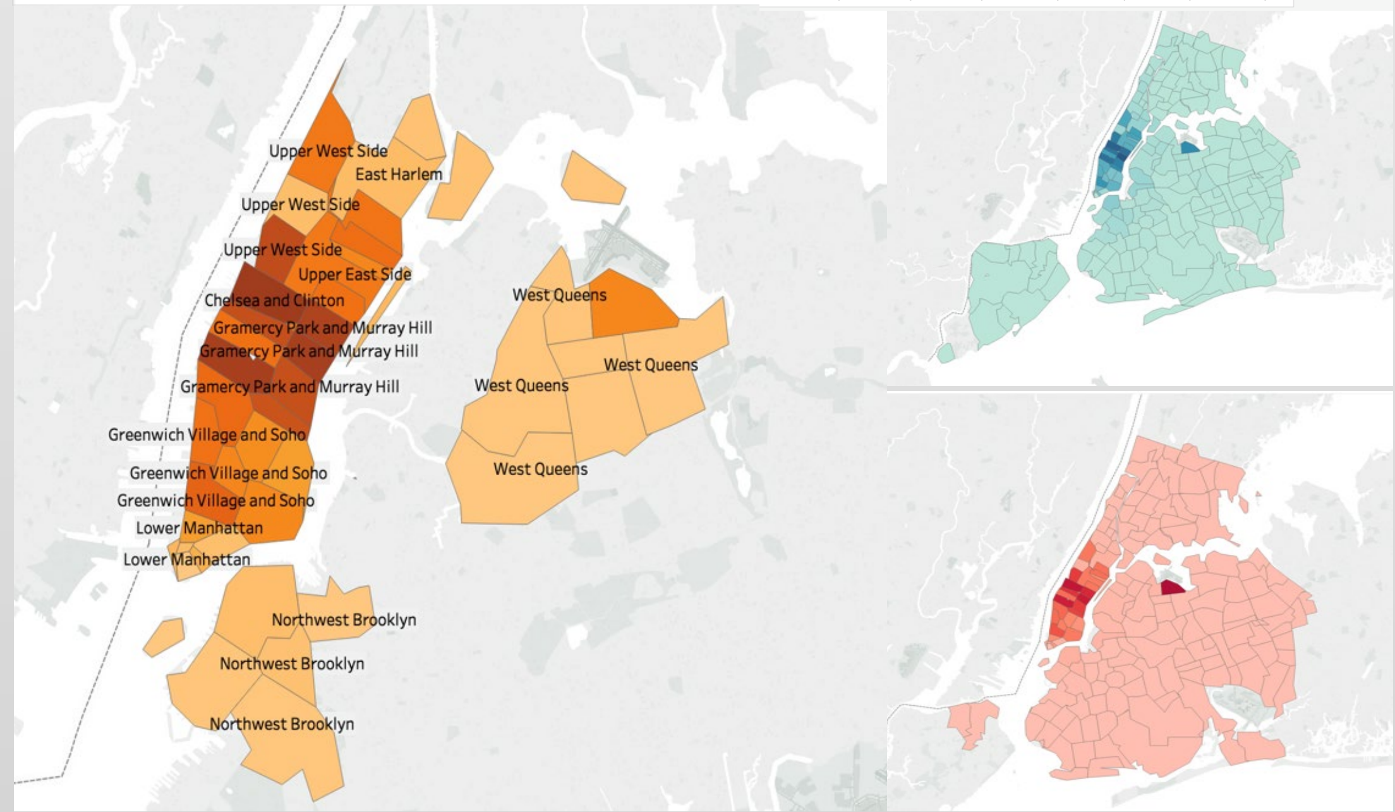
Data Description

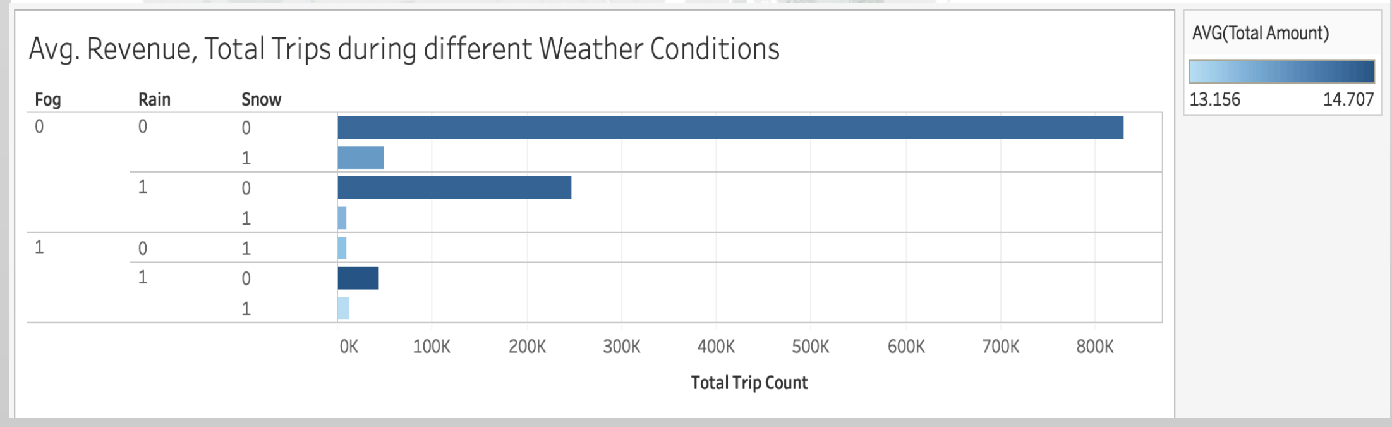
- Full Merged Dataset : NYC Yellow Taxi Trip Records + USA Federal Holidays Dataset + NYC Weather Data
- Number of data points : 12 M
- Training data points out of Sample Dataset: 600 K
- Total Features: After cleaning, the total features are 74
- Cleanliness: Merging, Date time conversion, Imputation, Dummy variables, Standardization
- Label/output to predict: Duration and Total Amount

Data Exploration



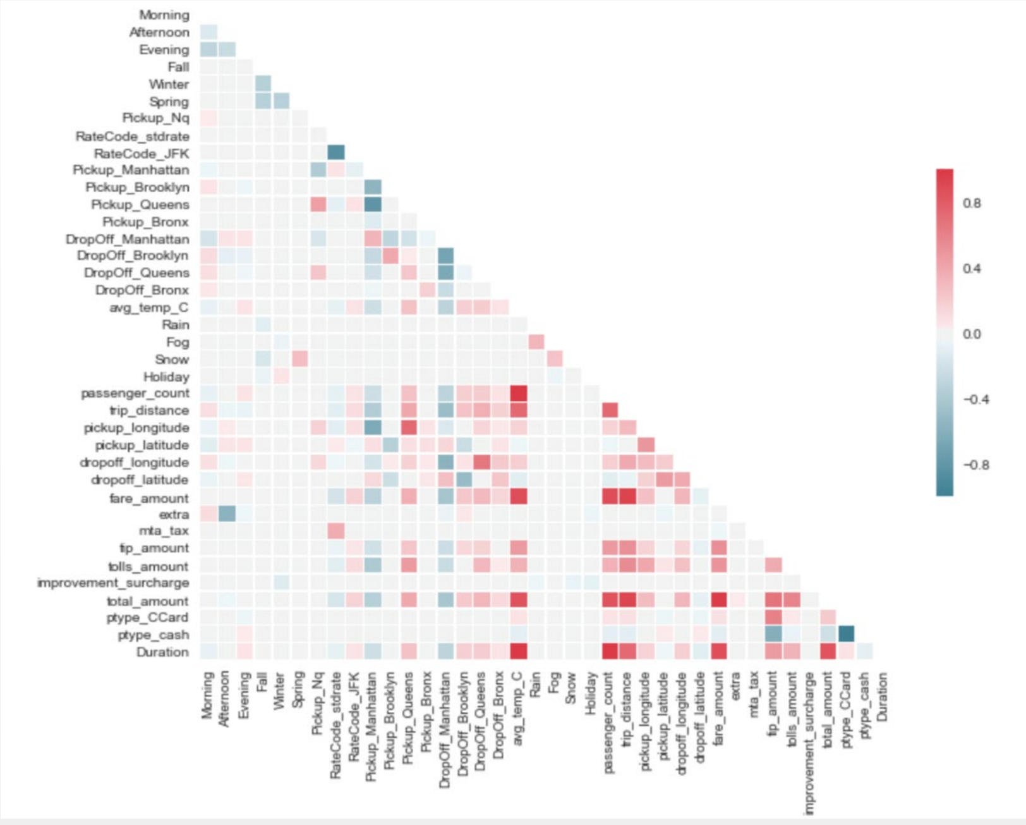




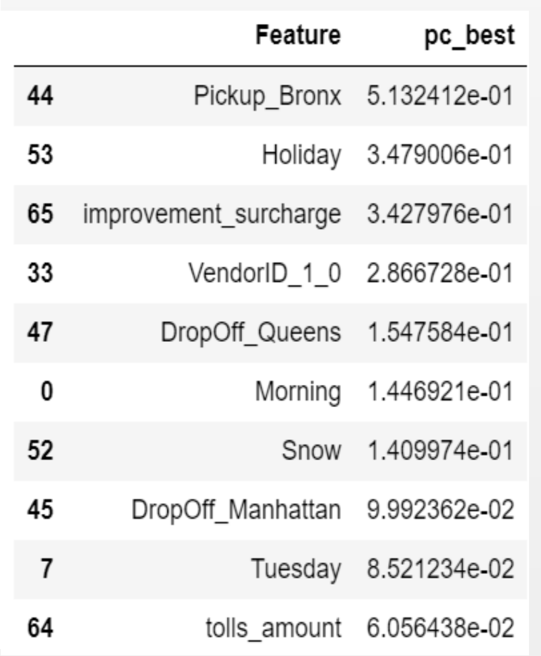


Model Description

- To find the relevant set of linearly dependent features, we implemented Principal Component analysis technique on the entire dataset
- Implementation of linear regression, random forest, gradient boosting and XGBoost models required selection of features from the best PCA component
- A threshold was then set to choose the features to be passed through the pipeline for predicting duration as well as the total ride amount
- Following is the correlation matrix used to cross validate the significance of relevant features obtained through PCA



These are the top 10 of the total features used by four models:

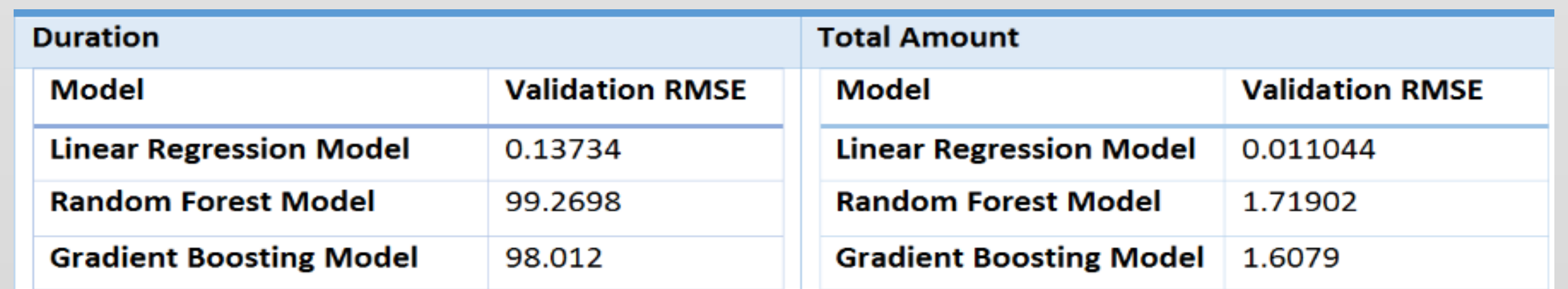


	Feature	pc_best
44	Pickup_Bronx	5.132412e-01
53	Holiday	3.479006e-01
65	improvement_surcharge	3.427976e-01
33	VendorID_1_0	2.866728e-01
47	DropOff_Queens	1.547584e-01
0	Morning	1.446921e-01
52	Snow	1.409974e-01
45	DropOff_Manhattan	9.992362e-02
7	Tuesday	8.521234e-02
64	tolls_amount	6.056438e-02

Model Comparison Metrics

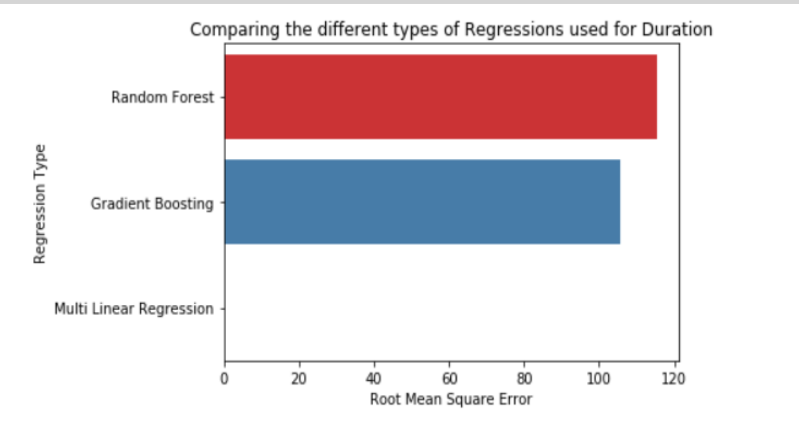
- For predicting both Duration and total amount we split the dataset into training, validation and testing in the 60%, 30% and 10% of the data respectively
- For predicting Duration and Total Amount, following models were implemented and the generalization performances were evaluated:
  - Linear regression : root mean squared error
  - Random Forest : root mean squared error
  - Gradient boosting regression : root mean squared error
  - XGBoost : accuracy

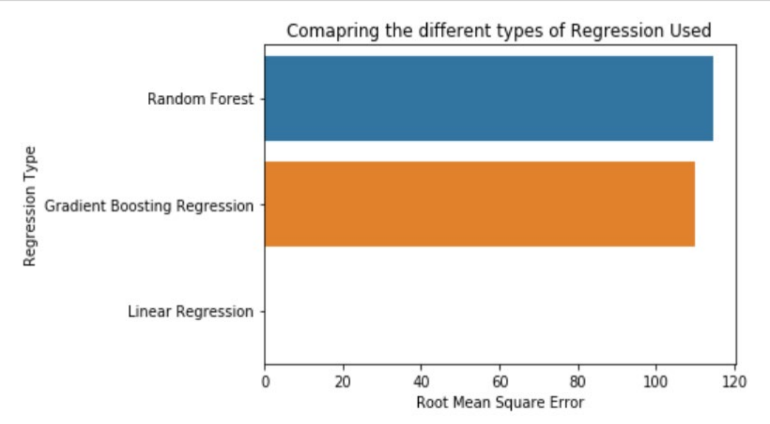
Results- Prediction Performance



Duration		Total Amount	
Model	**Validation RMSE**	**Model**	**Validation RMSE**
Linear Regression Model	0.13734	**Linear Regression Model**	0.011044
Random Forest Model	99.2698	**Random Forest Model**	1.71902
Gradient Boosting Model	98.012	**Gradient Boosting Model**	1.6079

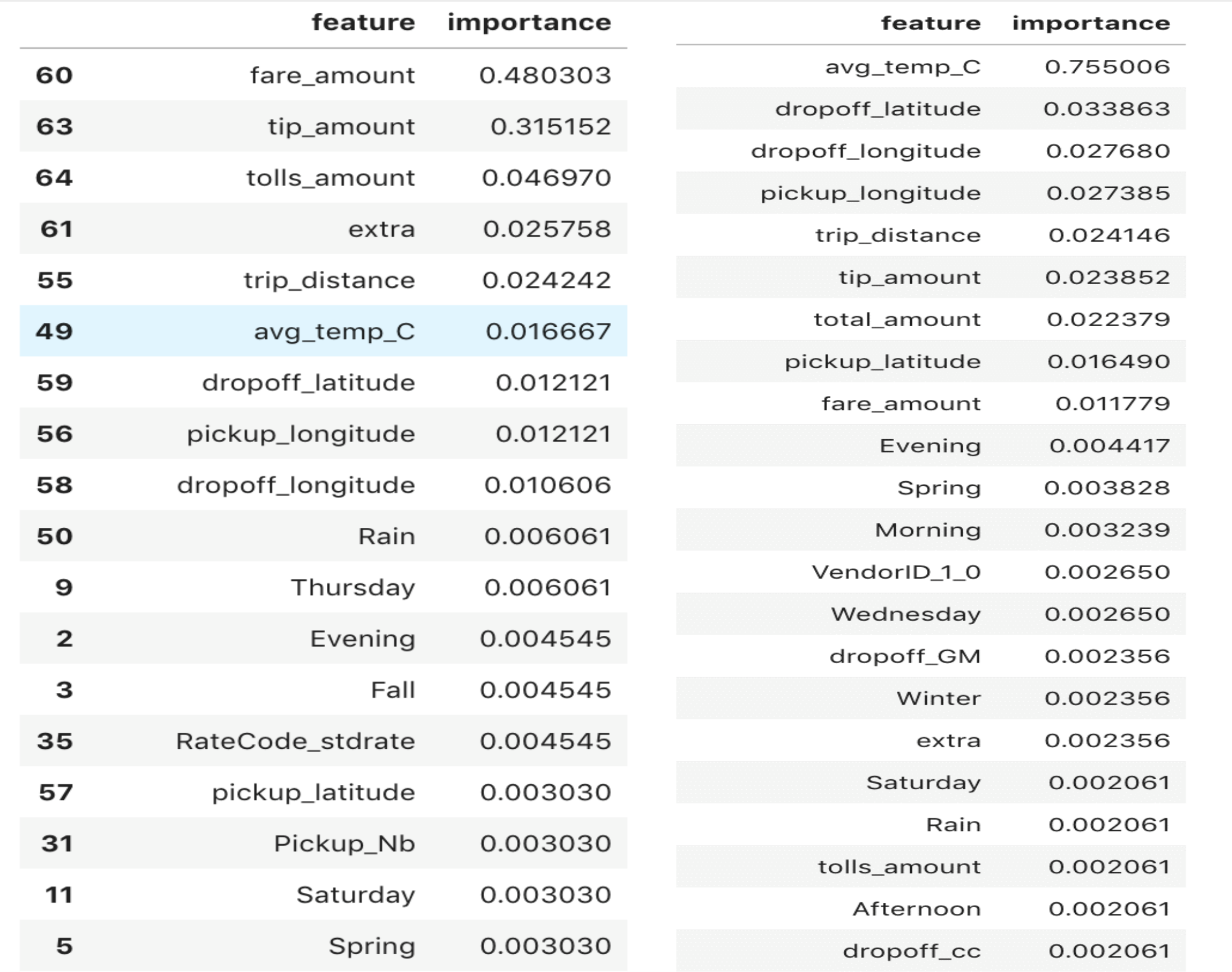
The rmse for gradient boosting regression model was achieved to be 1.71178 for amount and 87.3163 for total duration. The accuracy for the XGBoosting model is 67.9% for duration and 63% for total amount





Results- Inference

- Exploration Findings:
  - A bright sunny day is the best day for yellow cab drivers. In an event when there is no Snow, & no Fog but rain then, the total amount incurred is more which makes rain a significant factor
  - On Thursdays, the average total amount incurred is the most
  - Maximum revenue earned is in Queens possibly because of drop-offs at LaGuardia Airport
  - Similarly, Queens borough having LaGuardia Airport had highest revenue because of the number of pickups and drop offs rides at the airport
  - The average trip duration was the highest during Tuesdays and Wednesday
  - Most number of trips occurred during the month of October
- Significance of Correlation:
  - The correlation matrix between each variable to every other variable identifies the highly correlated features
  - This helps in identifying the significant features while efficiently reducing the number of features required to run the model
- Significance of PCA:
  - Principal Component Analysis is a significant method to identify features through dimensionality reduction
  - The features chosen by PCA can be cross-validated through logical understanding of correlation matrix
- Best Model:
  - The best model is chosen by comparing the generalization performance values of all the implemented models
  - The gradient boosting model worked the best due to low root mean squared error

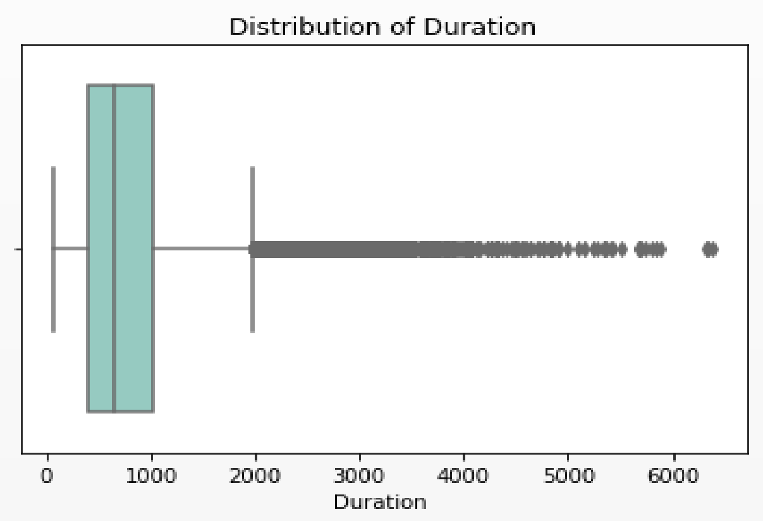


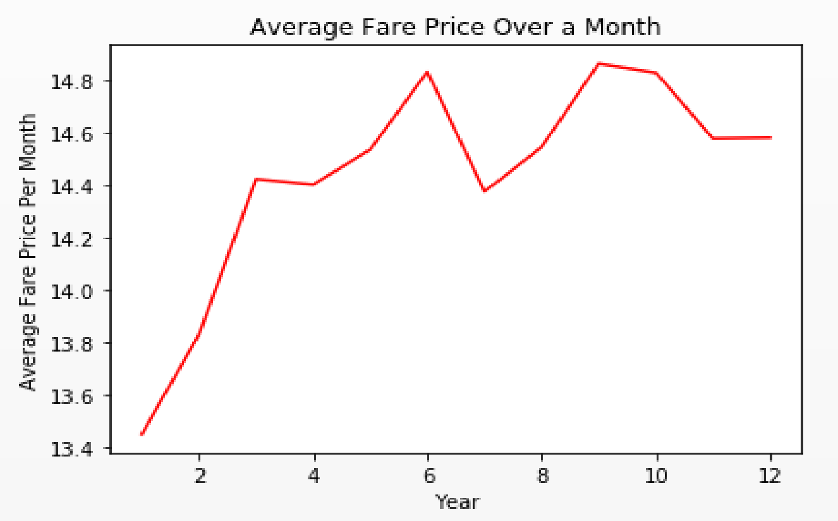
	feature	importance
60	fare_amount	0.480303
63	tip_amount	0.315152
64	tolls_amount	0.046970
61	extra	0.025758
55	trip_distance	0.024242
49	avg_temp_C	0.016667
59	dropoff_latitude	0.012121
56	pickup_longitude	0.012121
58	dropoff_longitude	0.010606
50	Rain	0.006061
9	Thursday	0.006061
2	Evening	0.004545
3	Fall	0.004545
35	RateCode_stdrate	0.004545
57	pickup_latitude	0.003030
31	Pickup_Nb	0.003030
11	Saturday	0.003030
5	Spring	0.003030

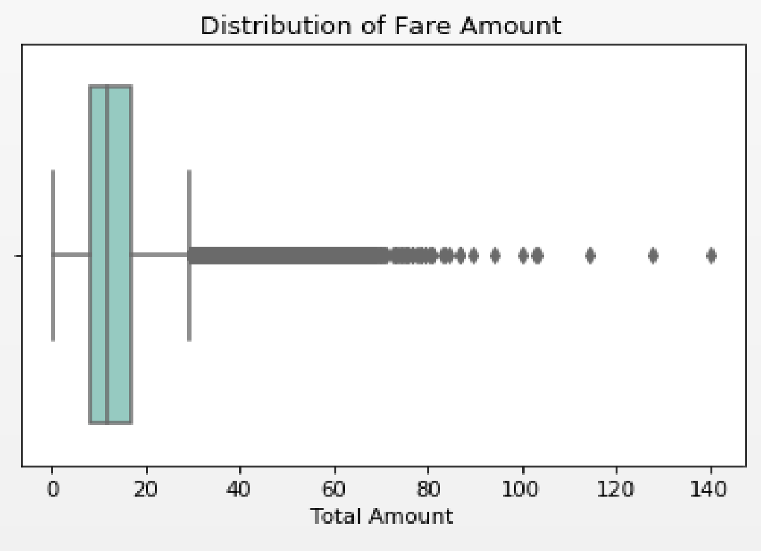
feature	importance
avg_temp_C	0.755006
dropoff_latitude	0.033863
dropoff_longitude	0.027680
pickup_longitude	0.027385
trip_distance	0.024146
tip_amount	0.023852
total_amount	0.022379
pickup_latitude	0.016490
fare_amount	0.011779
Evening	0.004417
Spring	0.003828
Morning	0.003239
VendorID_1_0	0.002650
Wednesday	0.002650
dropoff_GM	0.002356
Winter	0.002356
extra	0.002356
Saturday	0.002061
Rain	0.002061
tolls_amount	0.002061
Afternoon	0.002061
dropoff_cc	0.002061

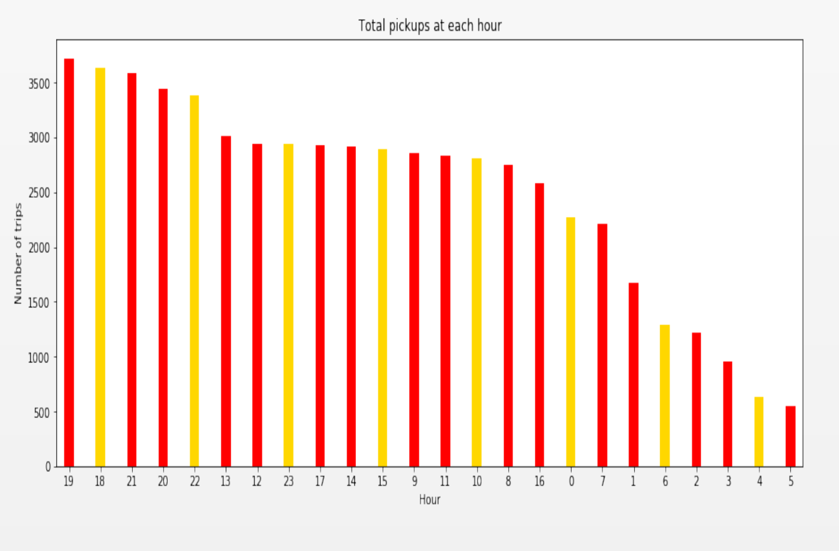
Novel Question

- Focus: Resurgence of Yellow Cabs
- Whether the ride will be profitable for the driver?
- Does weather and holidays really affect the profitability of driver?
- Why these questions:
  - Uber/Lyft Drivers and smart apps have destroyed the remarkable icon cabs
  - Stringent Interviews and Background Checks to achieve driving permits for above mentioned smart cabs. Not everyone gets to become their drivers.
  - Hence, adverse effect on employment rate and crime rate







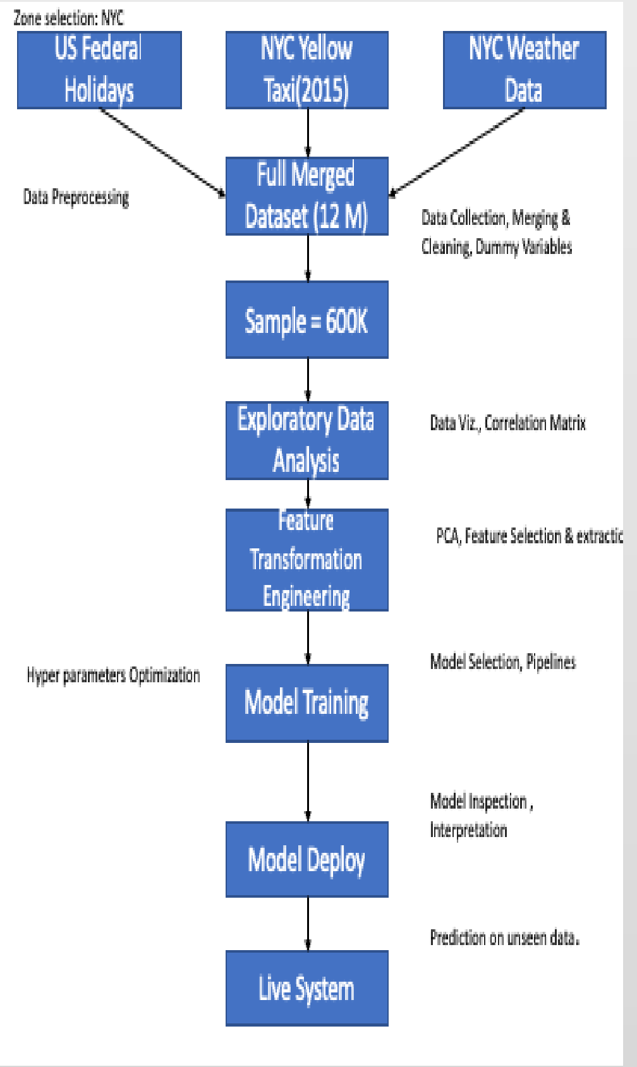


- The average duration of a trip is about 15-18 minutes
- The average total amount is between $10-$19
- The average fare price is highest for the month of October
- Longest rides occur at 3:00 PM in the evenings and at 6:00 AM in the morning, the distance of the rides is the shortest

Conclusion and Future Work

- Overall, our models for predicting taxi pickups' total amount and duration in New York city performed well
- The gradient boosting regression model closely followed by the random forest model performed the best
- This was likely due to their unique ability to capture complex feature dependencies
- The rmse for gradient boosting regression model was achieved to be ~$1.71178 for amount and ~87.3163 sec for total duration
- Our results and error analysis for the most part supported our intuitions for the usefulness of features with the exception of holiday feature which was not found important for model performance
- Our model can be used by city planners and yellow taxi drivers in determining where to position yellow taxi cabs and understanding patterns in ridership
- In our future work we plan to implement two models: neural network regression for its capability to automatically tune and model feature interactions
- Although we used k-means clustering, we aim to achieve better results in the future

Our Project Approach



Zone selection: NYC

US Federal Holidays

NYC Yellow Taxi(2015)

NYC Weather Data

Data Preprocessing

Full Merged Dataset (12 M)

Data Collection, Merging & Cleaning, Dummy Variables

Sample = 600K

Exploratory Data Analysis

Data Viz., Correlation Matrix

Feature Transformation Engineering

PCA, Feature Selection & extractic

Hyper parameters Optimization

Model Training

Model Selection, Pipelines

Model Inspection , Interpretation

Model Deploy

Prediction on unseen data.

Live System

Data Sources and References

City of New York Dataset :https://data.cityofnewyork.us/Transportation/2015-Yellow-Taxi-Trip-Data/ba8s-jw6u
Data Dictionary:
http://www.nyc.gov/html/tlc/downloads/pdf/data_dictionary_trip_records_yellow.pdf
Federal Holiday Dataset: https://catalog.data.gov/dataset/federal-holidays
Weather Dataset: https://www.kaggle.com/mathijs/weather-data-in-new-york-city-2015