

# Resurgence of NYC Taxis

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

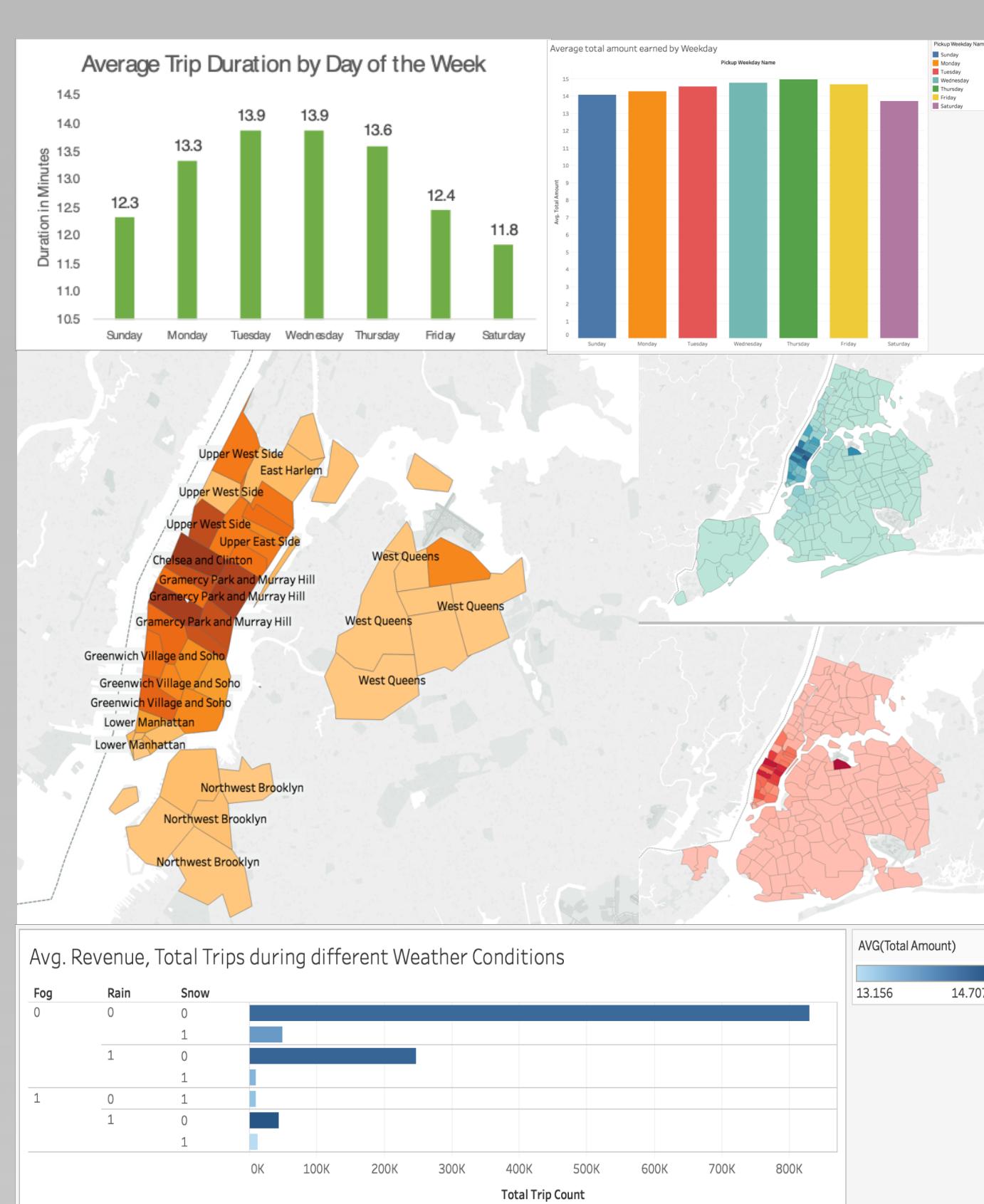
## Problems and Objectives

- Due to advent of smart cab companies such as Uber and Lyft, people are preferring 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

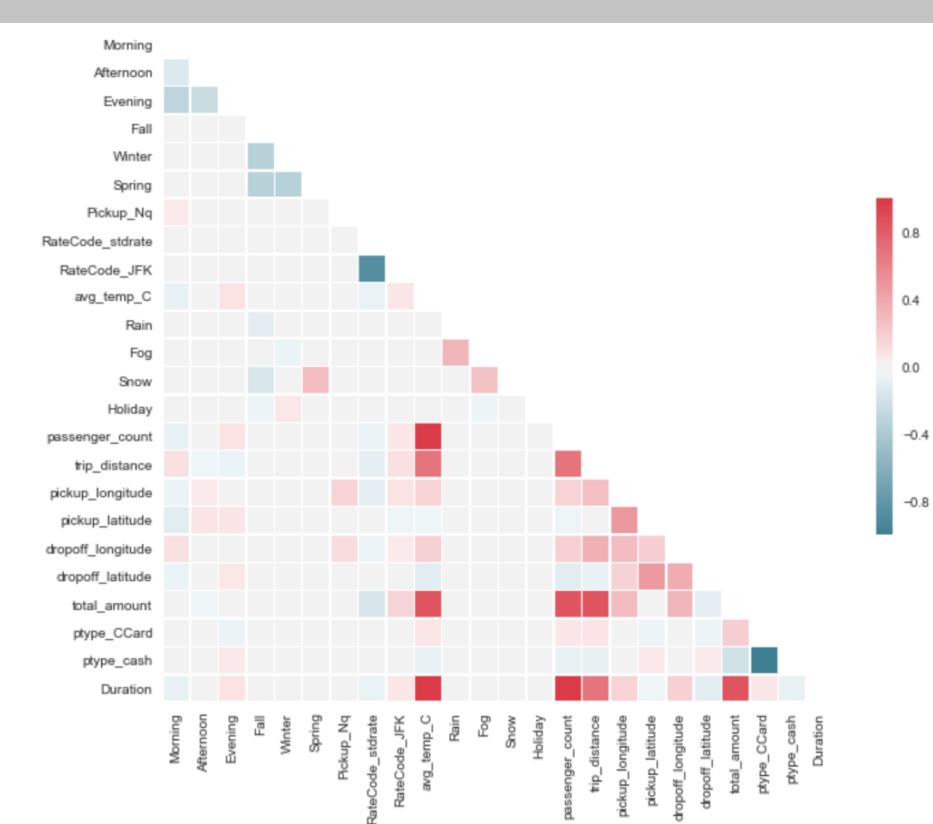
- Full Merged Dataset : NYC Yellow Taxi Trip Records + USA Federal Holidays Dataset + NYC Weather Data
- Number of data points : 12 M
- Training data points out of Sample Dataset: 600 K
- Total Features: After cleaning, the total features are 61
- Cleanliness: Merging, Date time conversion, Imputation, 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

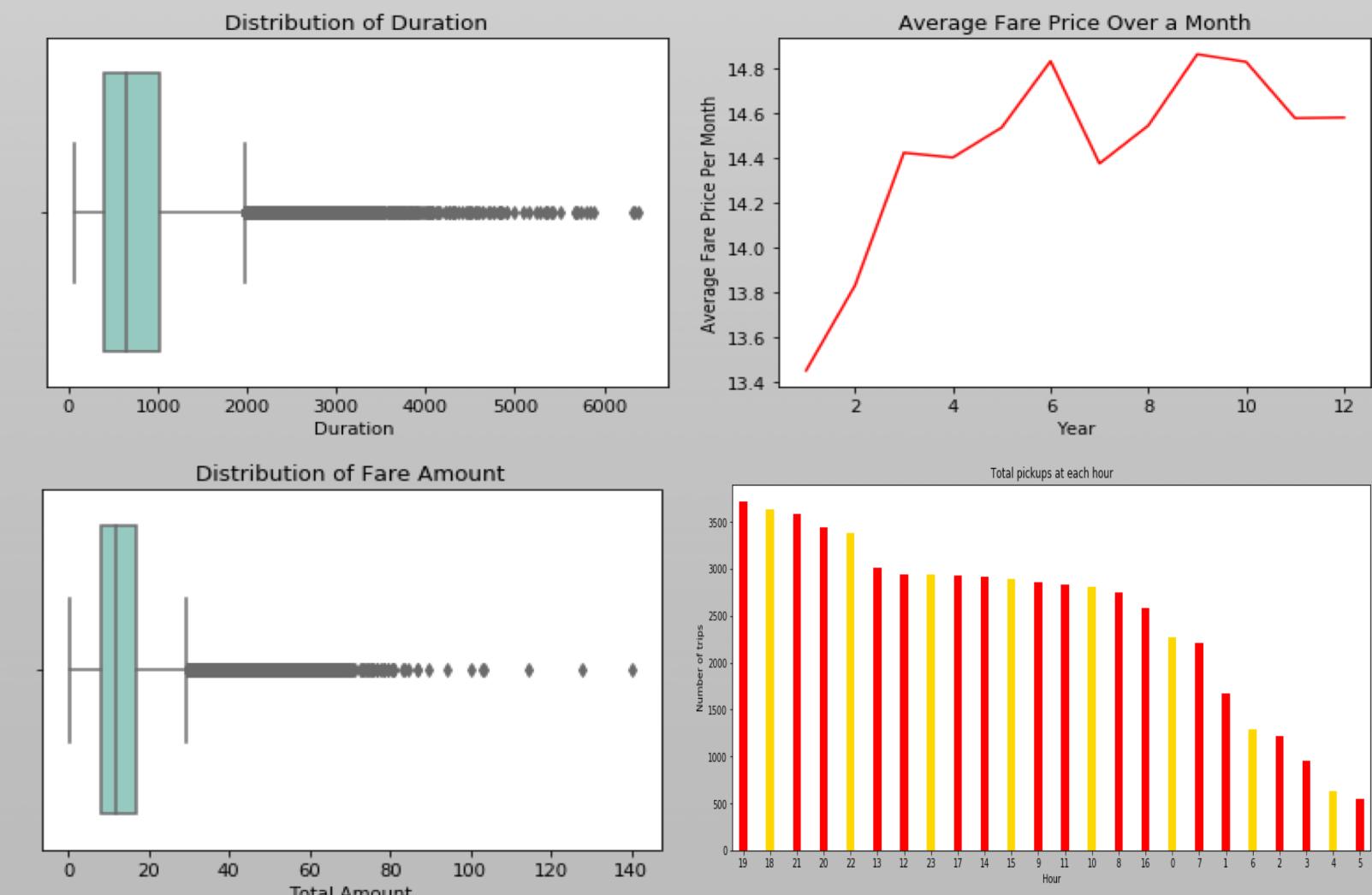


These are the top features used by four models:

feature_engineering_duration	Feature	pc_best
5	total_amount	0.690945
3	dropoff_longitude	0.067075
1	pickup_longitude	0.001308
2	pickup_latitude	0.004390
4	dropoff_latitude	0.021739
0	trip_distance	0.719446

## Results- Inference

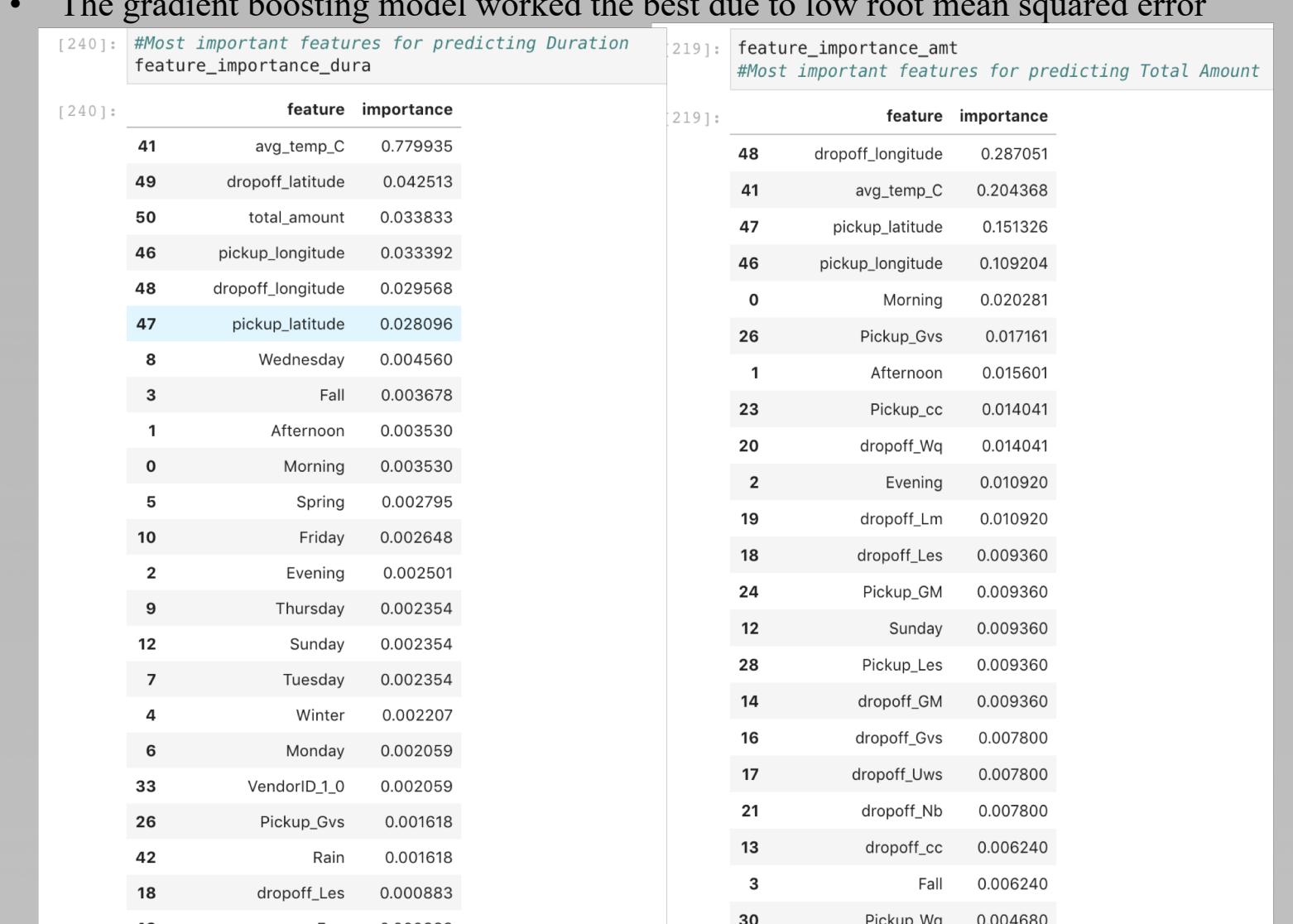
- Exploration Findings:**
  - A bright sunny day is the best day for yellow cab drivers. In an event when there is no Snow, & no Fog but rain then, the total amount incurred is more which makes rain a significant factor
  - On Thursdays, the average total amount incurred is the most
  - Maximum revenue earned is in Queens possibly because of drop-offs at LaGuardia Airport
  - Similarly, (Wq) West 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



- 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

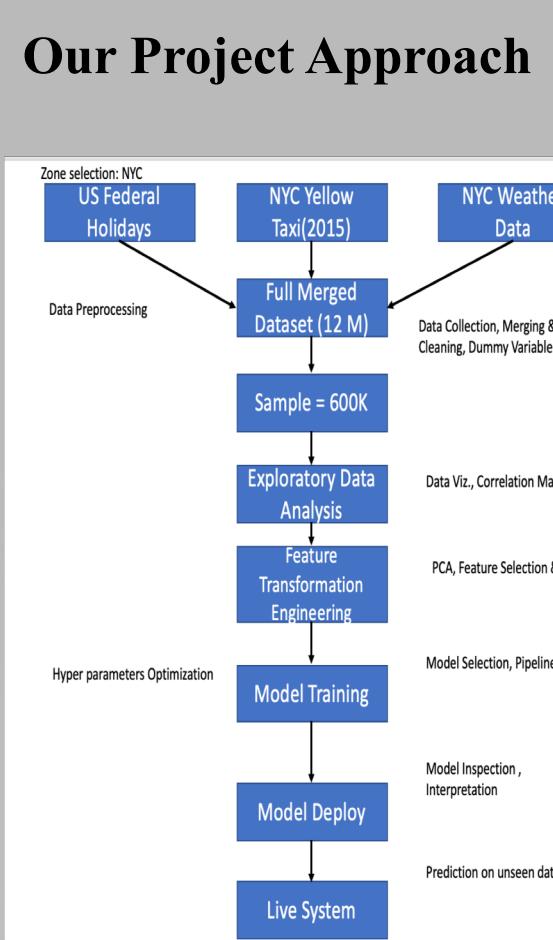
## Model Comparison Metrics

- 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

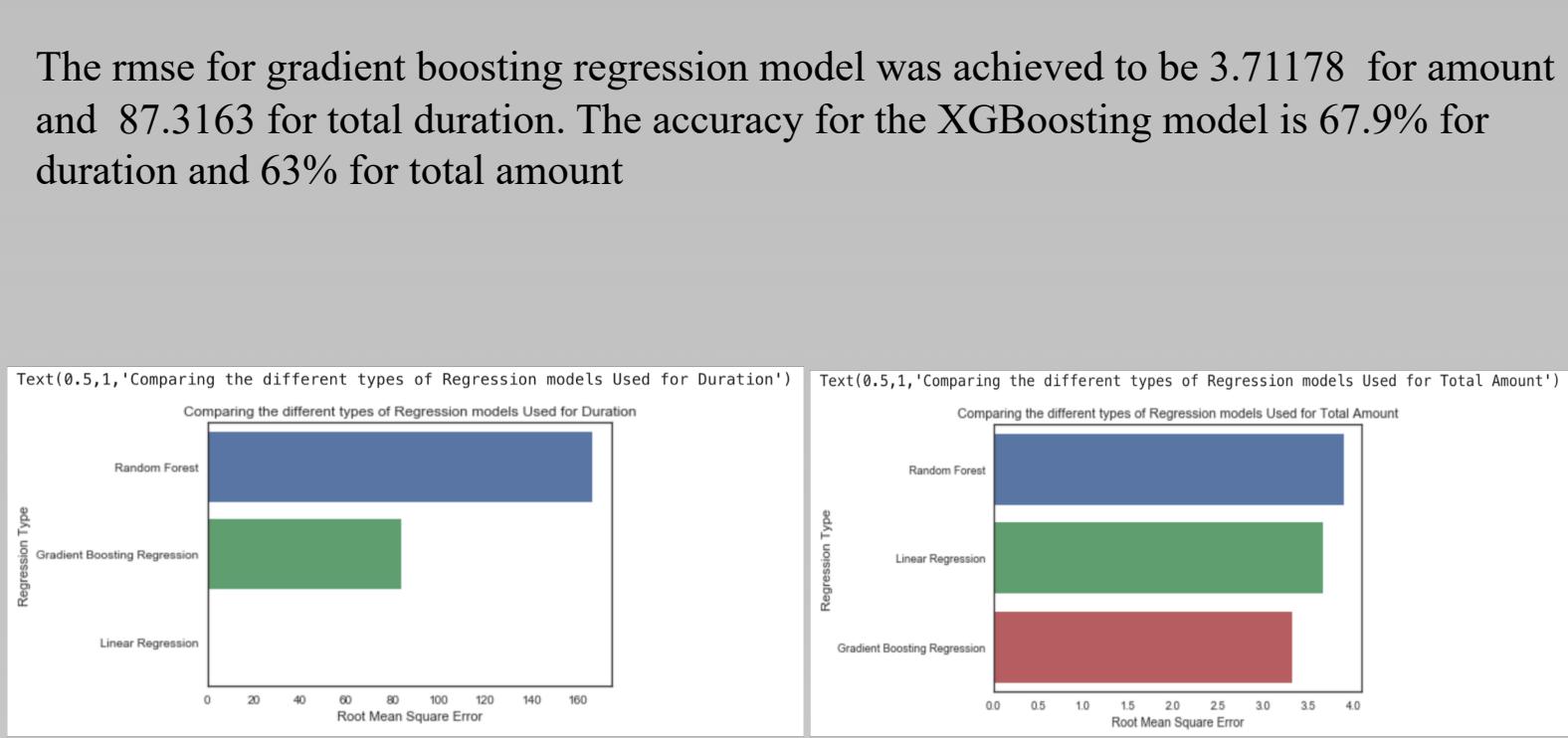
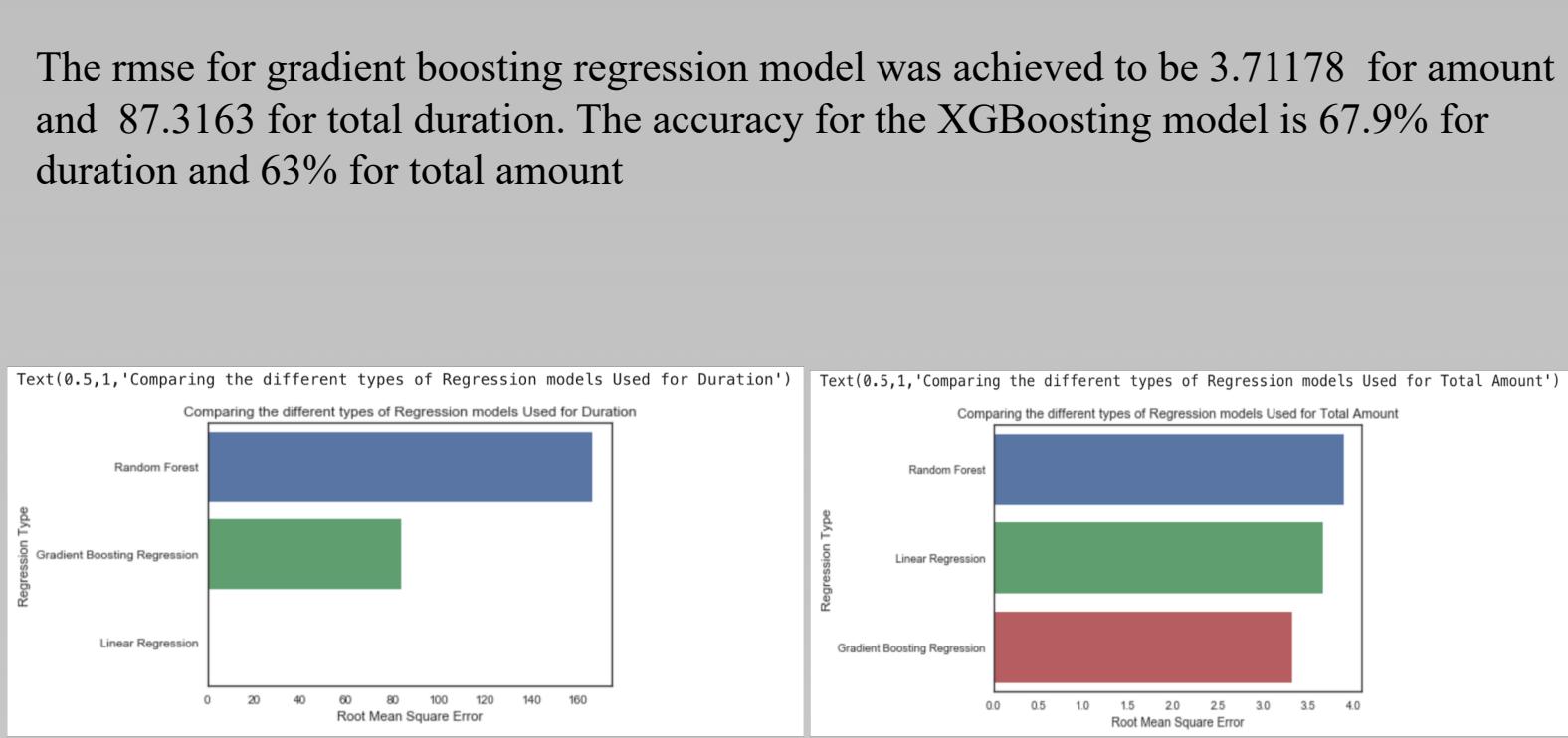


## Conclusion and Future Work

- 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 ~\$3.71178 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



## Results- Prediction Performance



## 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

## 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](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/mathijss/weather-data-in-new-york-city-2015>