Cab Fare Prediction

CIS 563 Introduction to Data Science

Bindushree Huruhihakkalu Ambikanath
Dr. Reza Zafarani, Assistant Professor
December 17, 2021

INTRODUCTION

This project helps in predicting the cab fare in New York City and detecting trends based on Latitude and Longitude limitations. Insight into cab utilization and fare distribution in New York's five boroughs: Manhattan, Queens, Brooklyn, Bronx, and Staten Island. Initial steps in this project will be cleaning the huge data set and preparing them for training and testing. With help of Keras, neural network will be created to predict cab fare depending on the user's position, destination address, and distance. This prediction model will be evaluated by using Decision tree, K Means, Random Forest, Regression techniques like AdaBoost and bagging.

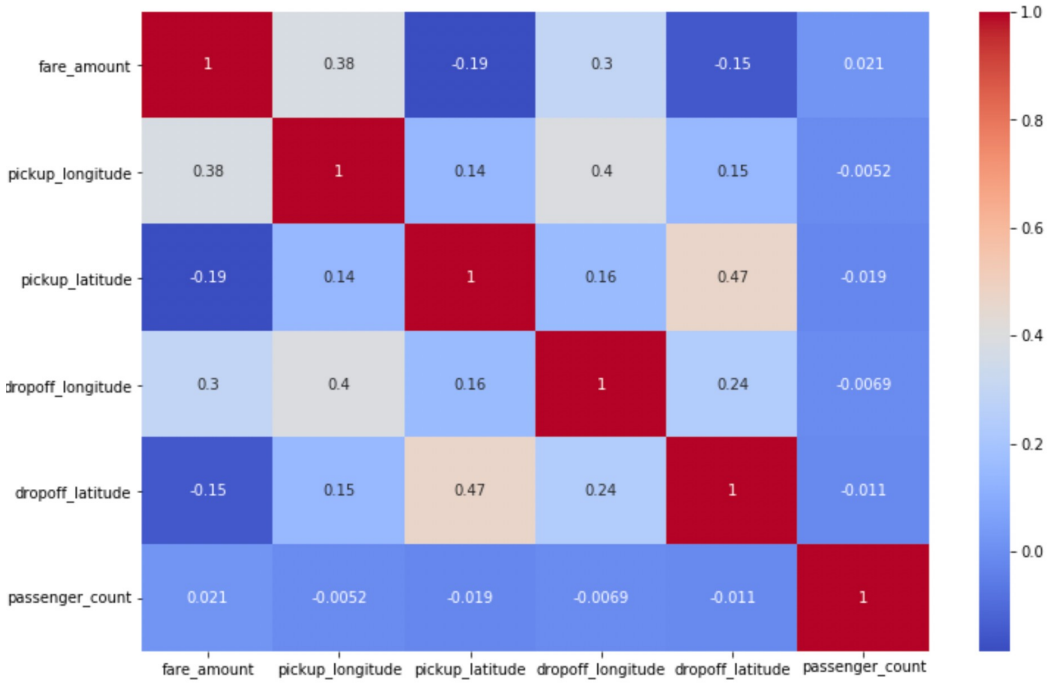
Mean squared error will be calculated. MSE is used to check how close estimates or forecasts are to actual values. Lower the MSE, the closer is forecast to actual. Certain questions mentioned below and many more will be answered in this project,

- Is pickup fare higher or drop off?
- Is onward and backward trip cost more or the overall trip cost?
- Which city has the highest cab fare?
- Difference between airport pickup cab and normal cab.

To better understand and visualize the prediction, graphs plotted are portrayed in the following sections.

DATASET FOR THE MODEL

Below are the fields present in the data.



Dataset instances

Dataset is found in Kaggle: https://www.kaggle.com/c/new-york-city-taxi-fare-prediction/data

SPECULATIONS

Here a list of hypotheses is created which will affect the cab fare:

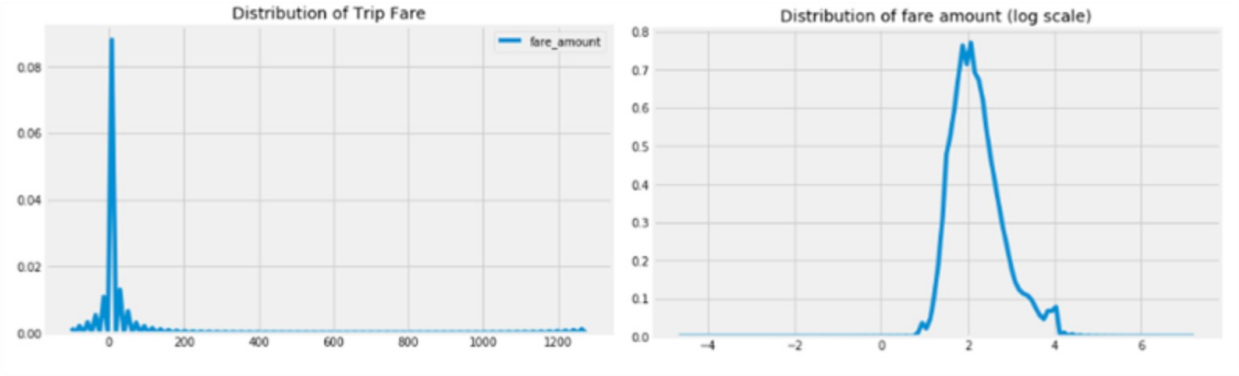
- **Distance travelled:** Greater the distance, higher the cost.
- **Time of Travel:** Cab fares might be higher during peak hours.
- **Airport travel or Regular:** Airport pickup and drops will have higher cost
- **Neighborhood:** Neighborhood affects the fare.

CLEANING AND EXPLORATION OF DATA

All outliers, negative and Not a Number values will be excluded.

Cab Fare Amount Distribution

In the dataset which is taken, few instances are negative. It is removed because the cost of a journey cannot be negative.

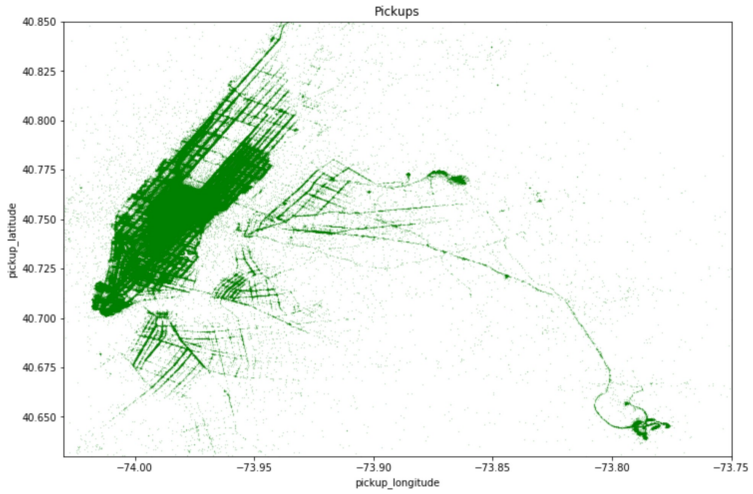


Distribution of fare amount

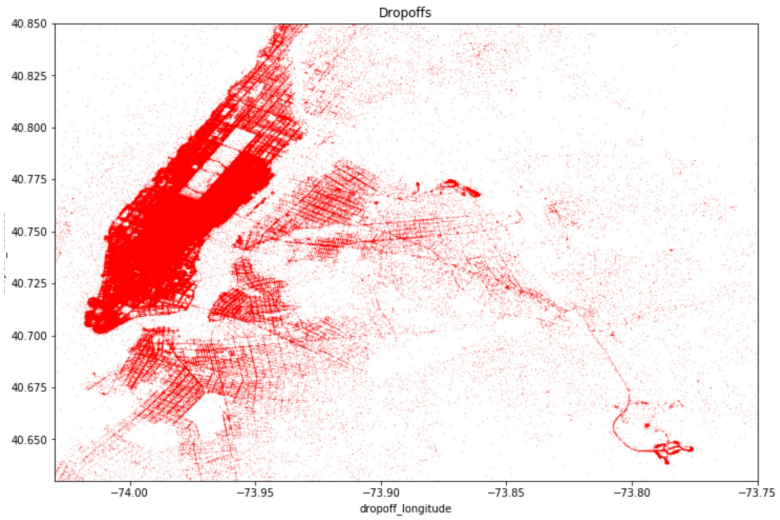
Geographical features Distribution

The latitude and longitude ranges are -90 to 90 degrees and -180 to 180 degrees, respectively. In the dataset, latitudes and longitudes are in the range of

(-3488.079513, 3344.459268) which is not possible. These incorrect values are excluded.



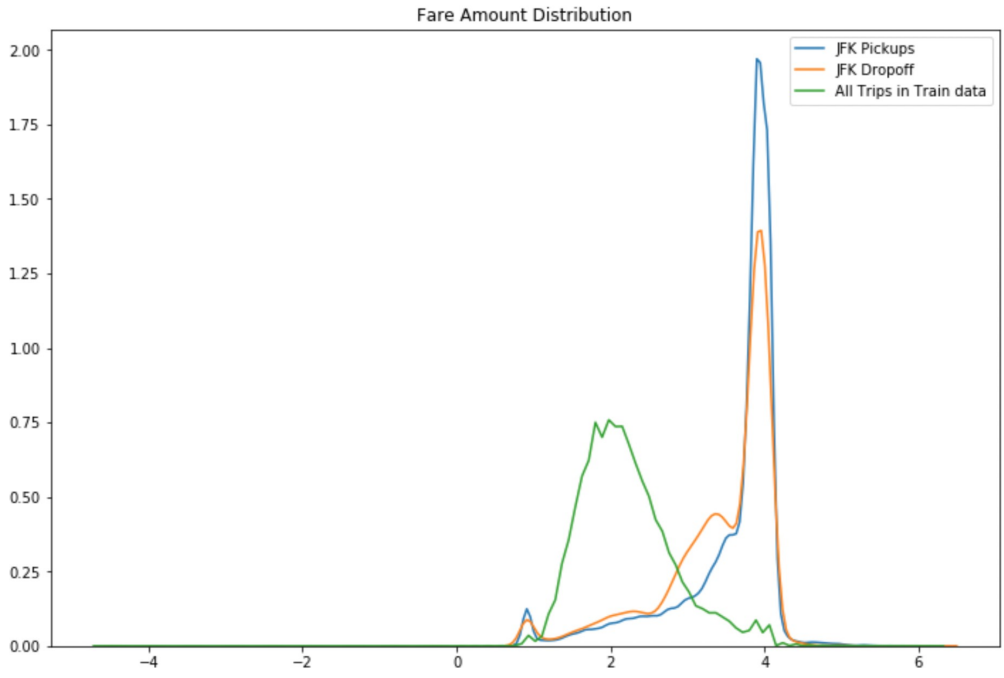
Distribution of pick-ups across NY



Distribution of drop-off across NY

From the above graph it is clear that JFK Airport and LaGuardia have a high density of pickups. From the graph we can understand that the fare for airport pickup/drop is higher. A model is built to assess whether a pickup or drop-off is from one of

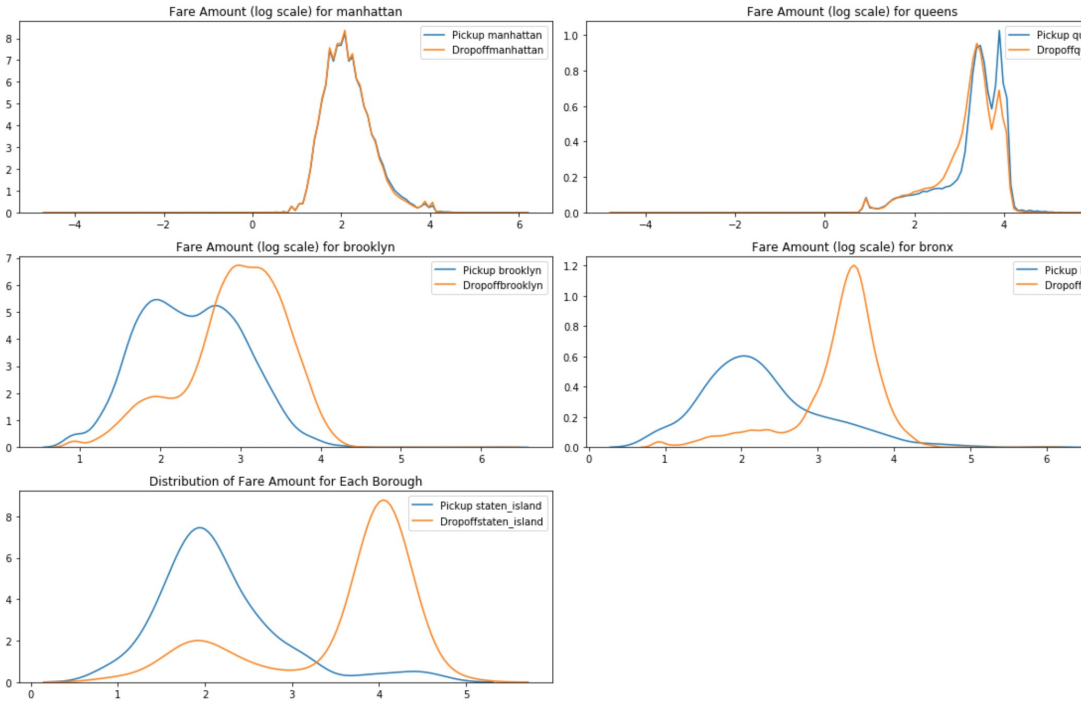
New York's three airports.



Fare amount for JFK pick-ups and drop-off

Next step is to see if we get higher cab fare predictions from certain neighborhoods and check the correctness of it. Every pickup and drop info given is divided into 5 boroughs : Manhattan, Queens, Brooklyn, Staten Island, and the Bronx.

From the graph below we can see clearly which city has the highest cab fare compared to others.



Distribution of Fare Amount across different neighborhoods of New York

Distribution of Trip Distance

Haversine Distance is used to determine the journey distance in miles using the pickup and drop-off coordinates.

BUILDING MACHINE LEARNING MODELS

How to build machine learning models to predict cab fare:

1. **Data Preparation:** Cleaning the data, removing unnecessary columns, and translating categorical variables to a machine-readable format.
2. **Model Evaluation:** This model is evaluated using below algorithms.

Model Used	Mean Squared Error	R2 Score
Linear Regression	70.0099383485925	0.256136295709813
Ridge Regression	70.0096676409456	0.256139172009918
Lasso Regression	87.1745701167874	0.0737600949725478
K-Nearest Neighbor	20.4105255142992	0.783135802234552
Decision Tree	33.4881455513385	0.644184574544359
Random Forest	18.0869850695378	0.80782368859866
AdaBoost	71.7504310388732	0.237643359272194
Gradient Boosting	71.5031229667471	0.240271036184528
Bagging Regression	18.1585861056763	0.807062919296054

CONCLUSION AND FUTURE WORKS

Model built to predict cab fare is fairly accurate. Random Forest represents the nonlinearities of traffic and location effect, this model outperforms all other models employed.

To further improve this model below points can be considered.

- Absolute location can be used as compared to relative location.
- Analyze the data using a non-linear model capture the data's intricacies.
- Utilize a larger training dataset.

REFERENCES

(1) NYC Taxi Trip and Fare Data Analytics using BigData - Umang Patel, Anil Chandan

(2) Estimating Taxi Demand-Supply Level using Taxi Trajectory Data Stream - Dongxu Shao, Wei Wu, Shili Xiang, Yu Lu

(3) Trip Fare Estimation Study from Taxi Routing Behaviours and Localizing Traces - Ce Liu (Innopolis University, University of Pittsburgh) Qiang Qu (Innopolis University)

(4) Predicting Taxi Demand at High Spatial Resolution: Approaching the Limit of Predictability - Kai Zhao, Denis Khryashchev, Juliana Freire, Cl´audio Silva, and Huy Vo