NYC Taxi Fare Prediction

Project Report

Akarsha Sehwag, Harsha.M.S,Kriti Bhardwaj, Kajal Sharma

October 25, 2021

Contents

List of Figures

List of Tables

1 Abstract

NYC Taxi is one of the cab services in New York, which is growing rapidly online car rental and taxi hire companies in New York. To keep up the pace at which they are growing they have decided to optimize their pricing to gain a good customer base and stay competitive. This project is aimed at predicting the optimized and thus competitive fare price depending on several factors like time, day of the week, peak hours, etc. Other factors to consider are distance, weather, availability of cabs, etc. If the car is rented for ¿1 day, then the fare should be calculated accordingly. We also find insights about the taxi bookings, their pickup/drop locations, times to understand our user behaviour.

2 Introduction

In New York, the taxi services are segmented into 5 major categories. These categories include **medallion taxis** ("yellow cabs"); **street hail liveries** ("boro or green taxis"); **black cars, liveries and luxury limousines** ("FHVs"); **commuter vans; paratransit vehicles**; and **wheelchair accessible vehicles** ("WAVs").

Medallion Taxicabs are often referred to as yellow cabs. Fares are set by the TLC and based on an initial charge, distance, and time, plus surcharges. Many yellow cabs are owned and operated by a garage as part of a fleet. In this arrangement, drivers pay to lease the taxi and medallion on a daily or weekly basis. There are also Individual Owner-Operators who own both the medallion and the vehicle, and can lease them to other drivers when not in use.

Street Hail Liveries are also known as green or boro taxis. They began providing service to New Yorkers in August 2013. Boro taxis can accept street hails and electronic trips, as well as pre-arranged trips, in Manhattan above E. 96th St. and W. 110th St., and anywhere in the other boroughs. They cannot pick up passengers at airports unless the trips are pre-arranged through a base. Fares are set by the TLC on street hails and e-hails; the dispatching base sets the fare when service is pre-arranged.

For-Hire Vehicles, known as FHVs, are divided into three categories: black cars, liveries, and luxury limousines. Black cars provide pre-arranged service, typically through smartphone apps or agreements with corporate clients; liveries and luxury limousines also provide pre-arranged service. FHVs cannot accept street hails in New York City. All FHVs must be affiliated with a base and can be dispatched by any FHV base of the same class. All FHV fares are set by the dispatching base; in addition, fares for liveries must be given upfront and are binding. FHV services which provide more than 10,000 trips per day are classified as High Volume For-Hire Services.

Commuter vans operate throughout New York City but within specific geographic boundaries approved by the Department of Transportation. Fares are set by the licensed authorities. Commuter vans typically provide rides in areas lacking in other public transportation options for a flat rate. Commuter vans are operated by a Commuter Van Authority.

Paratransit vehicles provide pre-arranged service for medical-related purposes. Trips are usually to or from healthcare facilities. Vehicles

must be dispatched by a paratransit base.

From January 2016 to June 2018, TLC-licensed vehicles and drivers completed nearly 780 million trips. The combined daily average trips in the taxi and FHV sectors increased 30 percent over the same period, from 766,000 in 2016 to over 1,000,000 in 2018. The fastest growing industry segment was High Volume For-Hire Services, which increased daily average trips 137 percent from 2016 to 2018.

The fewest number of trips were given on January 23rd, 2016, when 27 inches of snow fell in New York's worst snowstorm since 1869. Only 171,000 trips were given that day, compared to the highest recorded day, April 14th, 2018, which had over 1.23 million trips. Daily trip patterns vary across industries. In the medallion and High Volume sector, weekday trips are most frequently taken during morning and evening peak periods with mid-day lows. Saturdays see the most trips overall in these sectors. Traditional FHV trips are more likely to occur early in the day and decrease into the evening.

More than half of all medallion and SHL trips are less than two miles. Yellow cab trips tend to be about one mile longer: the average yellow cab trip was 3.7 miles while the average SHL trip was 2.8 miles.

Each industry segment serves New York City's boroughs differently. Medallion taxis dominate service in Manhattan, while other segments distribute trips more evenly. The average yellow cab fare was $13.61. The average SHL fare was lower at $12.78. Roughly half of all SHL fares were paid by credit card in 2016 and 2017, whereas two-thirds of medallion riders paid with credit card. Fares for High Volume For-Hire Services are paid entirely by credit card.

Passengers can request medallion and SHL vehicles anywhere in New York City through the use of mobile applications by TLC-approved companies. In a typical month, 70 to 75 percent of all e-hail requests are completed.

The most popular locations for e-hail requests tend to be in Manhattan. The five neighborhoods that rank highest for total daily requests are Upper West Side North and South, the East Village, Lincoln Square West, and TriBeCa/Civic Center. The most popular locations for requests outside of Manhattan were Park Slope and Crown Heights North in Brooklyn.

3 Problem Statement

NYC Taxi is one of the cab services in New York, which is growing rapidly online car rental and taxi hire companies in New York. To keep up the pace at which they are growing they have decided to optimize their pricing to gain a good customer base and stay competitive.

3.1 Objective

Predict the optimized fare price: The system should be able to predict the competitive fare price depending on several factors. For example, during peak hours, the prices of the fare should be comparatively high or if the travel is at night, then the night fares might apply. Other factors to consider are distance, weather, availability of cabs, etc. If the car is rented for ¿1 day, then the fare should be calculated accordingly.

4 Exploratory Data Analysis

4.1 Data Acquisition

The first step is to acquire the data from the data sources and then start working on it. There can be many techniques to gather data such as open-source platforms, APIs, RSS feeds or Web Scraping. We have gathered data from TLC Trip Data. The yellow and green taxi trip records include fields capturing pick-up and drop-off dates/times, pick-up and drop-off locations, trip distances, itemized fares, rate types, payment types, and driver-reported passenger counts.
Data Dictionary is available here:
Yellow Trips Data Dictionary

4.2 Data Cleaning and Data Preparation

Data cleaning is the process of fixing or removing incorrect, corrupted, incorrectly formatted, duplicate, or incomplete data within a dataset. Data preparation is the process of cleaning and transforming raw data before processing and analysis. It is an important step before processing and often involves reformatting data, making corrections to data and combining data sets to enrich data. We have removed negative values and null values from the obtained data source. For data cleaning, we have installed Pandas which is a software library written for the Python programming language for data manipulation and analysis.

4.3 Data Preprocessing

Data Preprocessing is the method which is used to transform the raw data into useful data. We extract features from the raw data which are used for creating the model. A feature is a property shared by independent units on which analysis or prediction is to be done. Features are used by predictive models and influence results. The features selected for this model are :

```
trip_duration,trip_distance,hour_of_day,day_of_week,fare_amount
```

4.4 Data Mapping

Data mapping is the process of matching fields from one database to another. It's the first step to facilitate data migration, data integration, and other data management tasks. Before data can be analyzed for business insights, it must be homogenized in a way that makes it accessible to decision-makers. Data

comes from many sources, and each source can define similar data points in different ways.

5 Insights(EDA)

Data Attributes –

- **Drop-off Datetime vs Tip Amount** – The pattern we have come across regarding the drop-off time is that on weekdays the drop-off is at a significant spike at the office times, i.e, 9 am and 6 pm, there is a dip during the lunch/afternoon hours. While on weekends we can see that the number of drop-offs is greater during the late-night hours. The tips people give also follow the same pattern.

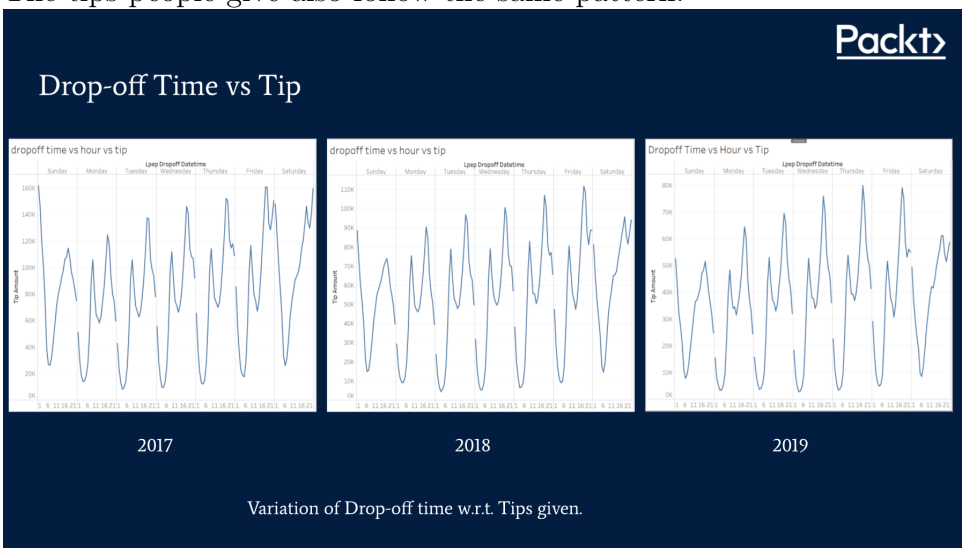


Figure 1: Drop-off Datetime vs Tip Amount

- **Hourly Division** – The number of Passengers and Total Amount is always at a spike around the morning and evening peak period.

Figure 2: Hourly Distribution

- **Weekly Division** - The number of Passengers and Total Amount is highest on Saturday and lowest on Monday.

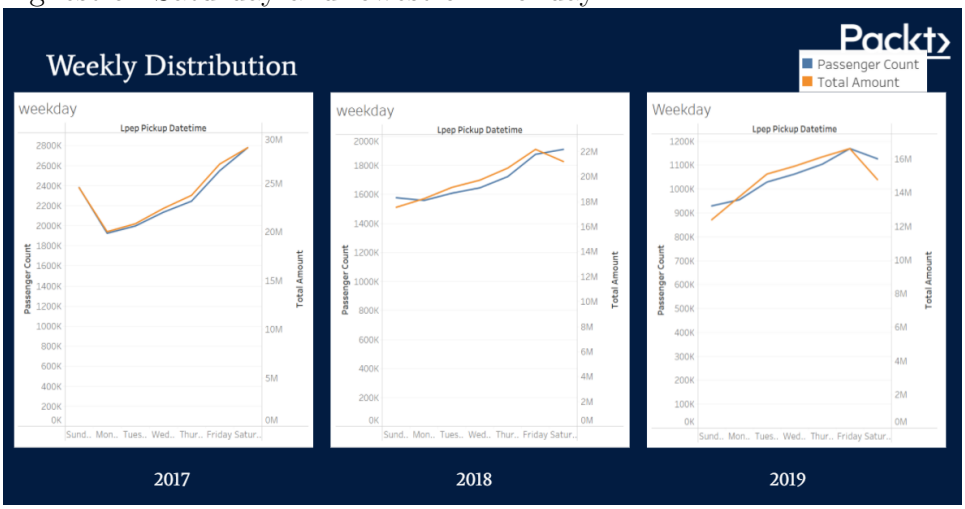


Figure 3: Weekly Distribution

- **Monthly Division** – The number of Passengers and Total Amount is highest during Quarter 1 in March and lowest during Quarter 3 in August.

Figure 4: Monthly Distribution

- **Payment Type** – People use Credit Card and Cash as their preferred modes of payment.

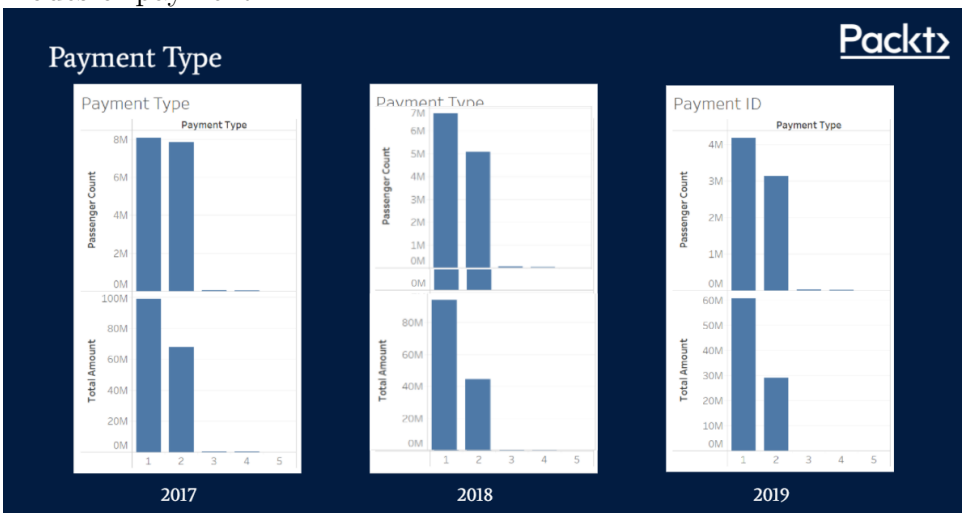


Figure 5: Payment Type

- **Payment Type vs Tip** – Most of the people who use Credit Cards tend to give tips as compared to other payment methods.

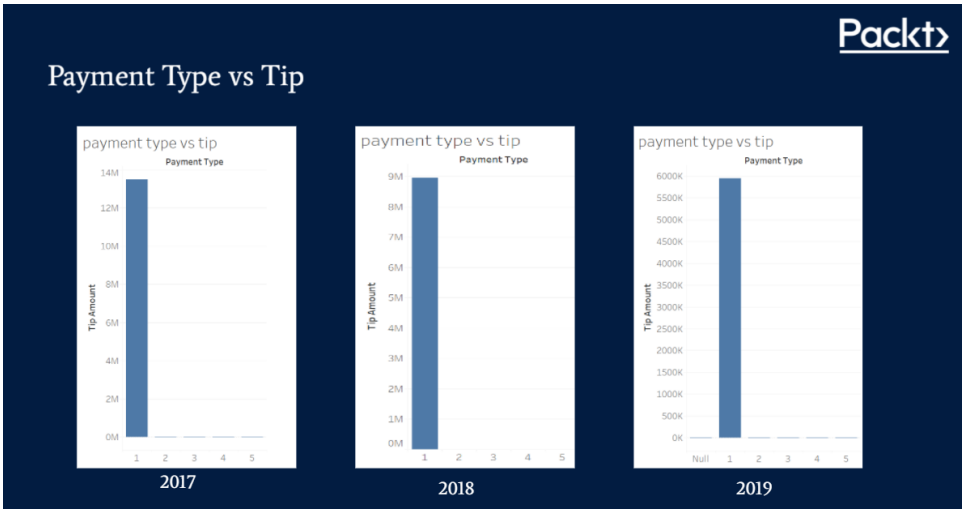


Figure 6: Payment Type vs Tip

- **Ratecode ID** – People prefer taking the standard rate taxis most of the time.

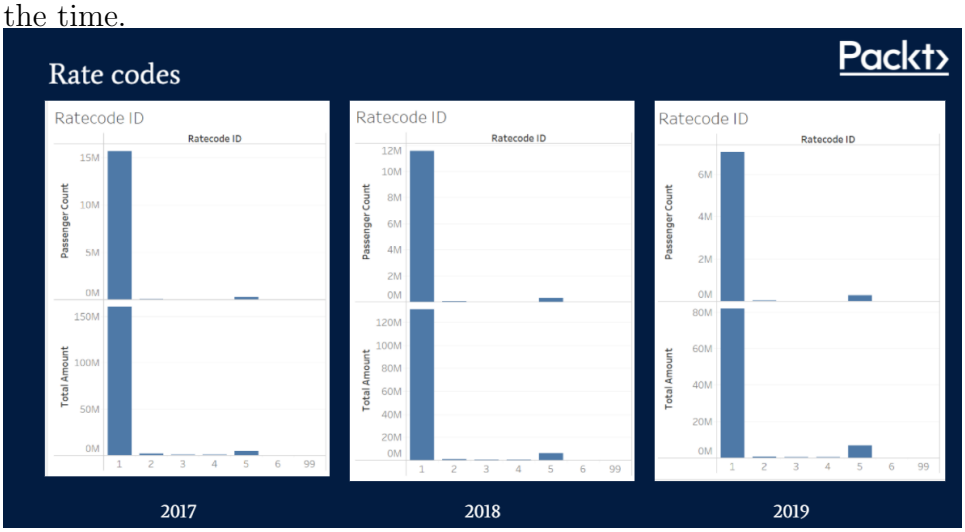


Figure 7: RatecodeID

- **Taxi Vendors** – People take taxis from VeriFone Inc. more than from Creative Mobile Technologies, LLC. The number of passengers and fare amount for VeriFone Inc. is significantly higher than for Creative Mobile Technologies, LLC.

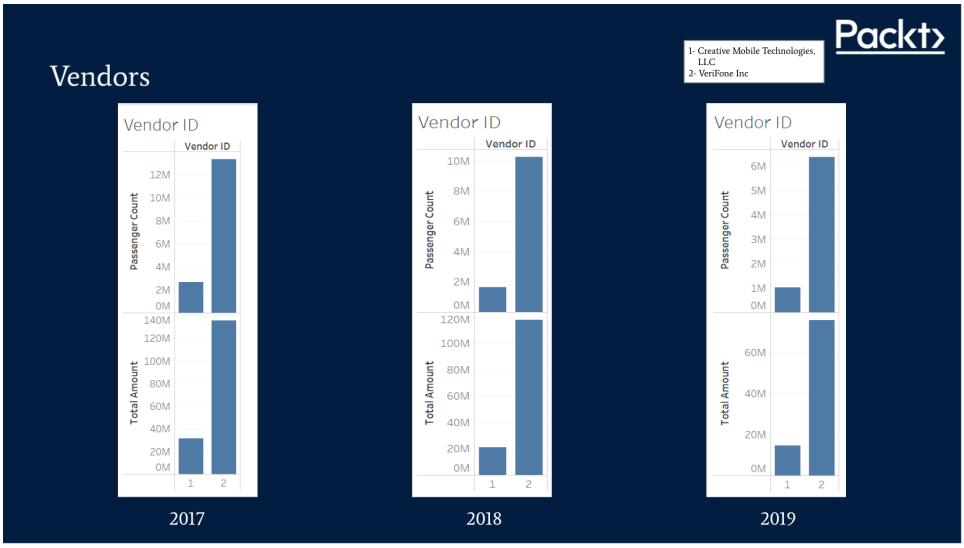


Figure 8: Vendor Distribution

- **Price/km for Vendors for Passengers** - Price/km each vendor charges for a different number of passengers.

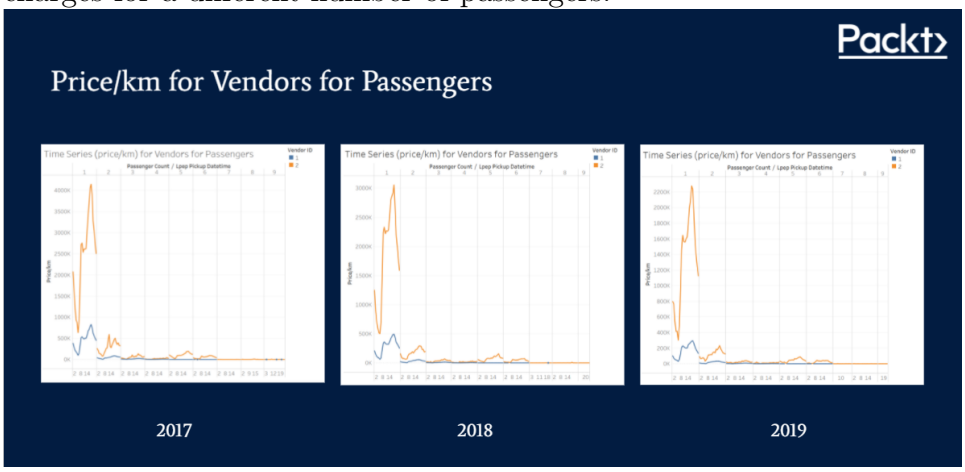


Figure 9: Price Per km for Passengers

- **Green Taxis** started serving in Manhattan from 2015, so the difference can be observed from the two years:

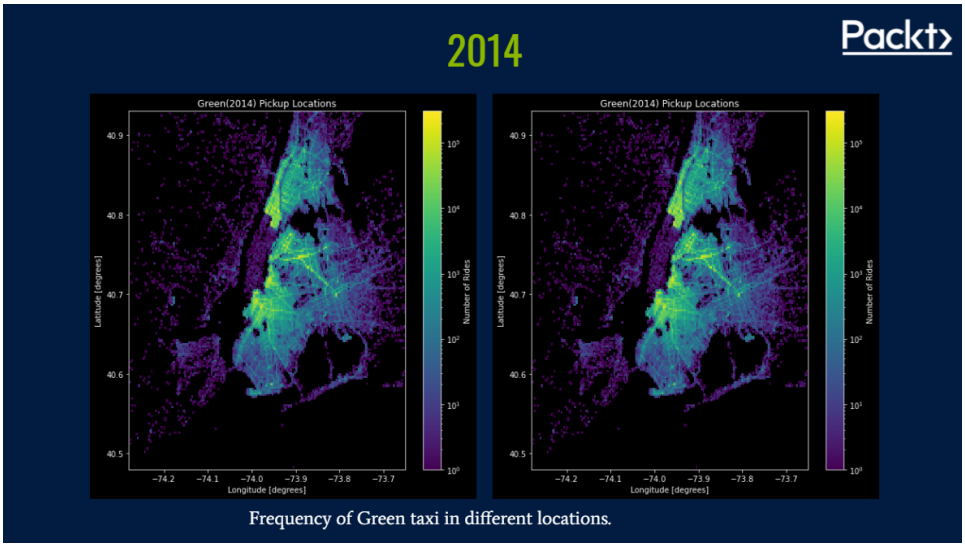


Figure 10: Taxi Pickup Density in 2014

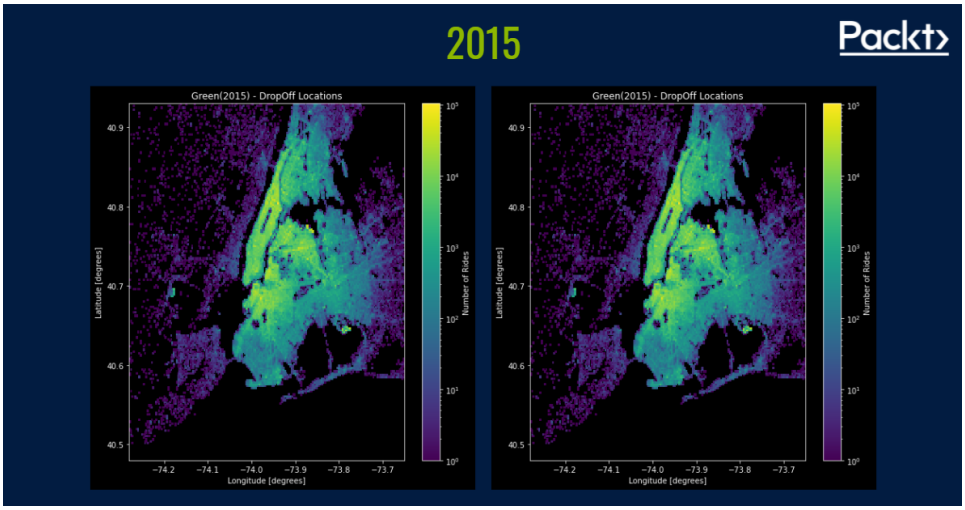


Figure 11: Taxi Pickup Density in 2015

6 Modelling

6.1 Probablistic Models

6.1.1 Linear Regression

There are two types of Linear Regression, i.e, Simple and Multiple. In Simple Linear Regression only one independent variable is present and the model has to find the linear relationship of it with the dependent variable. While, in Multiple Linear Regression there are more than one independent variables for the model to find the relationship.

6.1.2 Polynomial Regression

Polynomial regression is a special case of linear regression where we fit a polynomial equation on the data with a curvilinear relationship between the target variable and the independent variables.

6.1.3 Ridge Regression

Ridge regression is a model tuning method that is used to analyse any data that suffers from multicollinearity. This method performs L2 regularization. When the issue of multicollinearity occurs, least-squares are unbiased, and variances are large, this results in predicted values to be far away from the actual values.

6.2 Machine Learning Models

6.2.1 Decision Tree Regressor

Decision Tree is a decision-making tool that uses a flowchart-like tree structure or is a model of decisions and all of their possible results, including outcomes, input costs, and utility. Decision-tree algorithms fall under the category of supervised learning algorithms. It works for both continuous as well as categorical output variables.

6.2.2 KNeighbors Regressor

KNN regression is a non-parametric method that, in an intuitive manner, approximates the association between independent variables and the continuous outcome by averaging the observations in the same neighbourhood.

6.2.3 Support Vector Regressor

Support Vector Regression is a supervised learning algorithm that is used to predict discrete values. Support Vector Regression uses the same principle as the SVMs. The basic idea behind SVR is to find the best fit line. In SVR, the best fit line is the hyperplane that has the maximum number of points.

6.2.4 Artificial Neural Networks

6.2.5 Ensemble Techniques

A technique that takes multiple base models to combine their output using different statistical methods to produce one optimal predictive model. These Ensemble methods usually perform better than individual learning methods. There are usually two kinds: Bagging and boosting.

6.3 Time Series models

The Augmented Dickey-Fuller Test with a p-value 0.0 indicated that time series models such as **Auto-regressive(AR)**, **Moving Average(MA)**, **ARIMA**, **Space-Time ARIMA(STARIMA)** and other simliar models would not be appropriate.

7 Modelling Insights

Augmented Dickey-Fuller Test indicated that the data is stationary, which meant that the time-series models would not be applicable in this case.

Simple linear models - Multiple and ridge regression models are moderately effective in capturing the variance, while the maximum error is moderately high.

Support Vector Regressor takes a significant amount of time to train, while performing no better than the simple regression models and also has the highest maximum error.

Polynomial regression performs better than the simple linear models as it captures the predictor variables in higher order and also the interaction between the variables.

The following is the regression equation obtained:

```
18.332258071053843(trip_duration)
 + 11.486268095963036(trip_distance)
 + 1.455473311763102(hour_of_day)
 - 0.15030612907598329(day_of_week)
 - 19.046618352531986(trip_duration^2)
 + 0.7440054612299148(trip_duration * trip_distance)
 + 0.3656198237964201(trip_duration * hour_of_day)
 + 0.2587405701363147(trip_duration * day_of_week)
 + 3.338853314504214(trip_distance^2)
 - 0.06794972775721675(trip_distance * hour_of_day)
 + 0.9861245929130833(trip_distance * day_of_week)
 - 1.2728536859797197(hour_of_day^2)
 - 0.009020206242299267(hour_of_day * day_of_week)
 - 0.3798260919243237(day_of_week^2)
```

Multilayer perceptron is able to capture a good amount of variance giving a high R2 value, but the maximum error is slightly higher than the previous regression models. Though the model gives a good R2 value but the explainability of such a model is generally low.

Decision Tree regressor is able to capture the variance to a very good extend while maintaining a moderate maximum error.

Ensemble technique was used to combine the following weak learners:

- Multiple Linear Regressor
- Random Forest Regressor
- Gradient Boosting Regressor

The ensemble model gives the highest explained variance while the maximum error is in order of the decision tree regressor model.

8 Evaluation Hypothesis

The following table summarizes the modelling evaluation details:

Model name	**Explained Variance**	R^2 **Value**	**Max Error**
Ensemble Regressor	93.37%	0.9371	16.67
Decision Tree Regressor	89.45%	0.8945	90.02
Multilayer Perceptron Regressor	93.47%	0.9346	17.28
KNeighbors Regressor	93%	0.9300	17.57

Table 1: Model Evaluation

The following graphs depicts the performance of the models about various metrics:

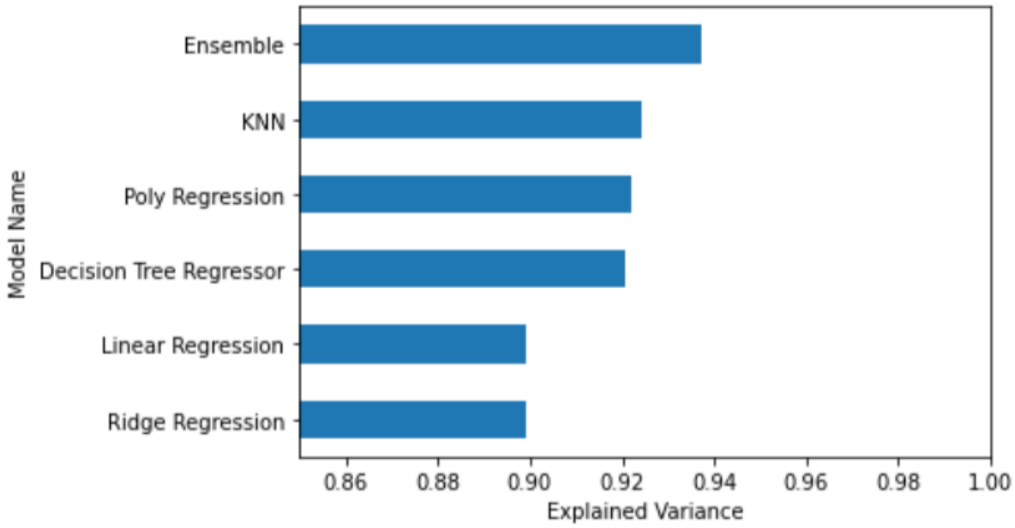


Figure 12: Explained Variance

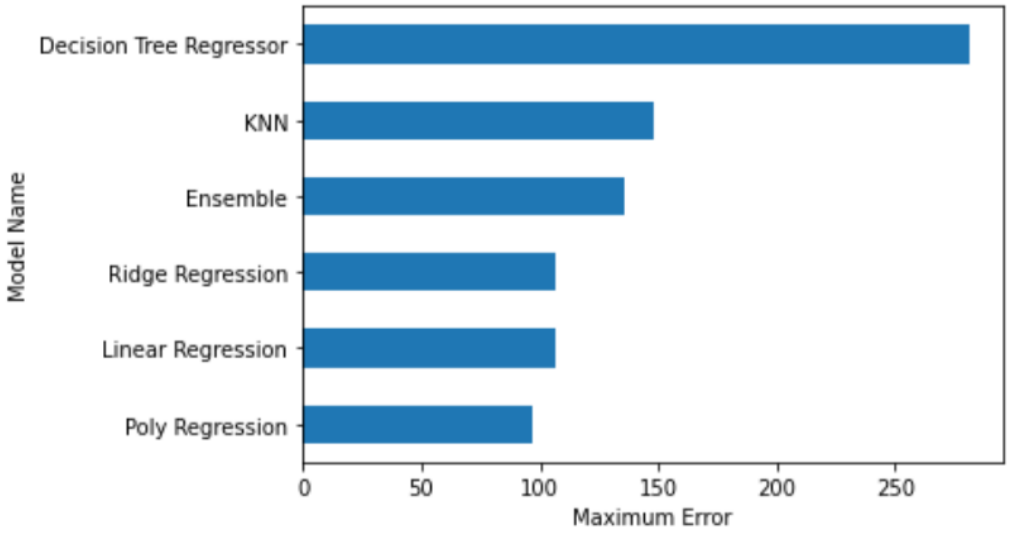


Figure 13: Maximum Error

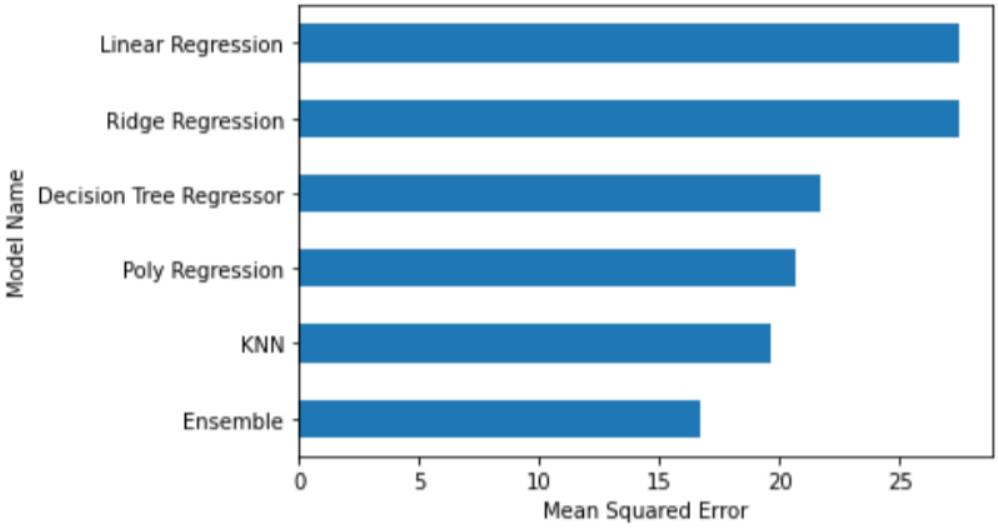


Figure 14: Maximum Squared Error

9 User Interface

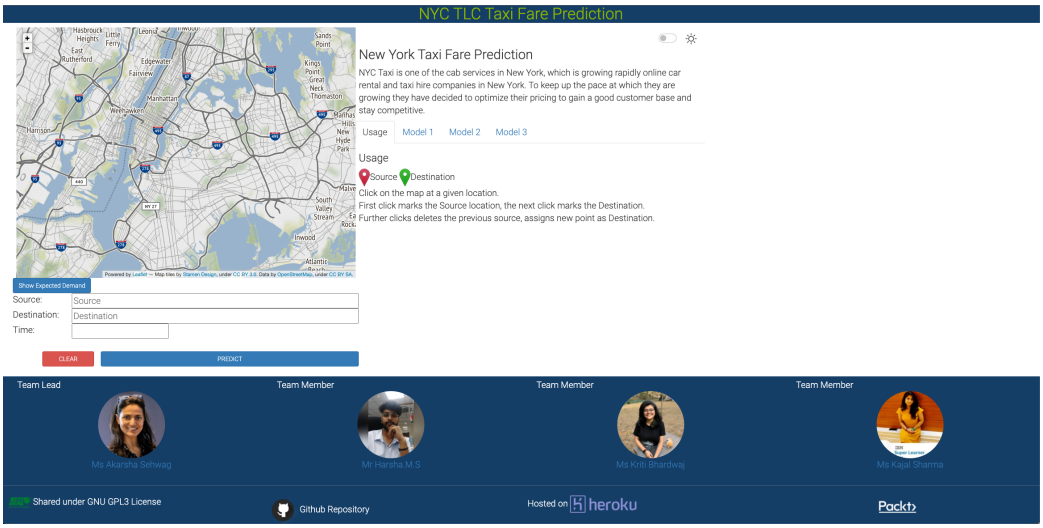


Figure 15: User Interface

Figure 16: Fare Prediction between 2 points at the lowest demand hour

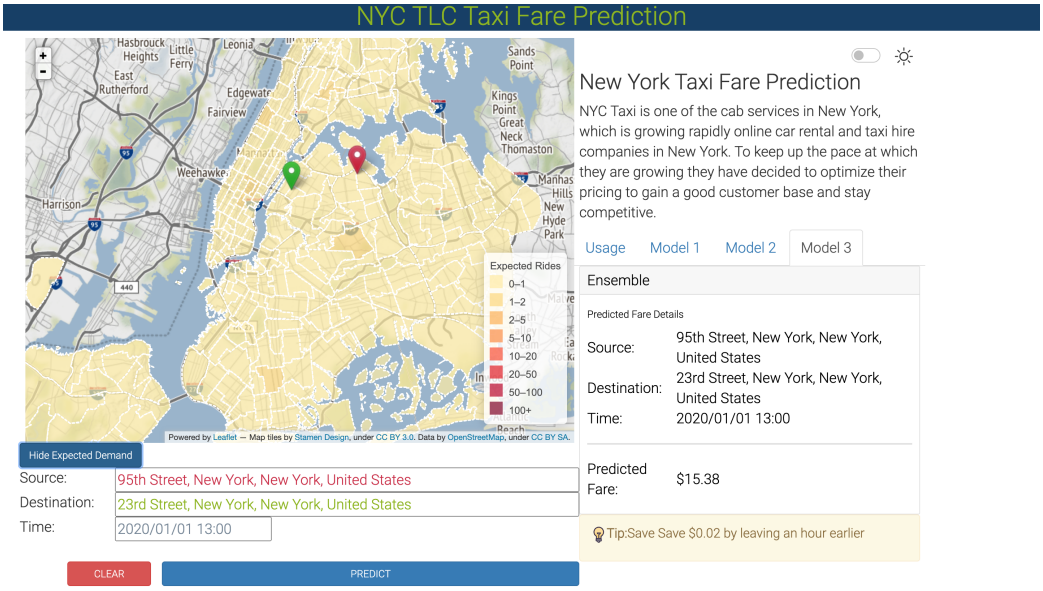


Figure 17: Demand Prediction at the lowest demand hour

Figure 18: Fare Prediction between 2 points at the highest demand hour

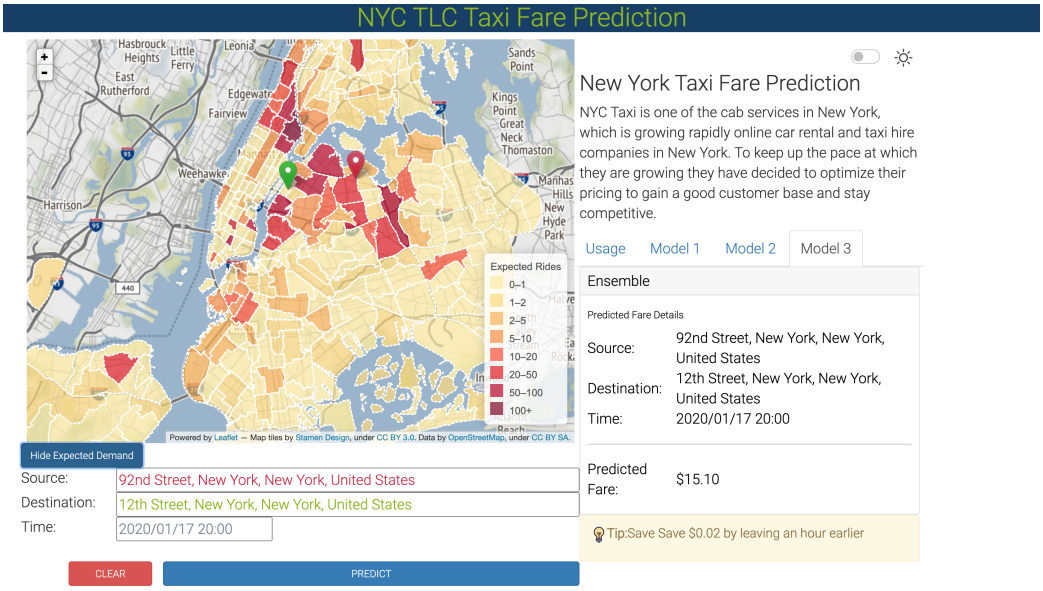


Figure 19: Demand Prediction at the highest demand hour

10 Technology Stack

Purpose	Tools/Technologies Used
Data Storage and Code Execution	
Data Manipulation and Visualization	
Model Training	
Maps API	
Map Rendering	
Backend Server	
Cascading Style Sheets	
HTML DOM manipulation / API calls	

Table 2: Technology Stack

11 Conclusion

Augmented Dickey-Fuller Test indicated that the data is stationary, time-series models would not be applicable in this case.

Simple linear models - Multiple and ridge regression models
moderately effective in capturing the variance,
maximum error - moderately high.

Support Vector Regressor
significant amount of time to train
performs no better than the simple regression models
has the highest maximum error.

Polynomial regression
performs better than the simple linear models

Multilayer perceptron
able to capture a good amount of variance giving a high R2 value,
maximum error is slightly higher than the previous regression models.

Decision Tree regressor
able to capture the variance to a very good extent while maintaining a moderate maximum error.

12 Future Scope

The project is just a starting point for a lot of business use cases:

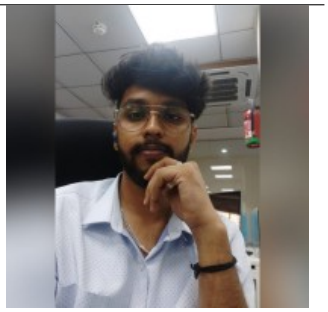
- Predicting the expected demand of taxi at a particular time and particular place
- Recommending time to the user when the given location will be less in demand and thus no surcharge.
- Optimizing the number of taxis according to the demand.
- Optimizing the location of taxis according to the demand.

13 References

- https://www.analyticsvidhya.com/blog/2021/05/all-you-need-to-know-about-your-first-machine-learning-model-linear-regression/
- https://www.mygreatlearning.com/blog/what-is-ridge-regression/#:~:text=Ridge%20regression%20is%20a%20model,away%20from%20the%20actual%20values
- https://www.analyticsvidhya.com/blog/2020/03/polynomial-regression-python/
- https://www.geeksforgeeks.org/python-decision-tree-regression-using-sklearn/
- https://bookdown.org/tpinto_home/Regression-and-Classification/k-nearest-neighbours-regression.html
- https://towardsdatascience.com/unlocking-the-true-power-of-support-vector-regression-847fd123a4a0#:~:text=Support%20Vector%20Regression%20is%20a,the%20maximum%20number%20of%20points
- https://www.tableau.com/learn/articles/time-series-analysis
- https://www.geeksforgeeks.org/data-preprocessing-in-data-mining/
- https://www.talend.com/resources/data-mapping/#:~:text=Data%20mapping%20is%20the%20process,and%20other%20data%20management%20tasks.&text=Data%20now%20comes%20from%20many,data%20points%20in%20different%20ways

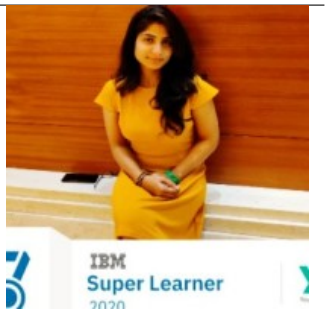
14 Team

Team Lead
Ms. Akarsha Sehwag



Team Member
Mr. Harsha.M.S

Team Member
Ms. Kriti Bhardwaj



Team Member
Ms. Kajal Sharma