VIT®

Vellore Institute of Technology

(Deemed to be University under section 3 of UGC Act, 1956)

Neural Networks and Fuzzy Control

TAXI DEMAND PREDICTION

By Team, **NoTaxiFound**

Submitted by - HARSH BHARDWAJ

Motivation

New York City has one of the busiest roads and highest demands for taxis. Despite the availability of a large amount of taxi trip data in New York City, there is still a lack of accurate and efficient methods for predicting the demand for yellow taxis. This presents a challenge for taxi companies and drivers who want to optimize their resources and provide better service to customers. Indian cities also face the same problem, thus by developing a machine learning model that can accurately predict the demand for yellow taxis in New York City based on historical trip data, we can make a universal algorithm that can be applied to Indian cities also.

Research gaps

1. There are few models for taxi demand prediction that focus on building a predictive algorithm for a city as a whole and those which do, have limited spatial resolution (grouping large areas together), reducing efficiency to trade off complexity.
2. Many existing models use complex machine learning algorithms that are difficult to interpret. This makes it challenging for stakeholders to understand the factors that drive demand and make decisions based on the model output.
3. Traditional taxi demand models often have limited generalizability to new contexts or environments.
4. Traditional taxi demand models often suffer from data quality issues such as missing values, outliers, and measurement errors. But a good machine learning model needs to be built after efficient pre-processing mechanism.

Aim

To develop an accurate and reliable model for predicting taxi demand in New York City, which could potentially be used by taxi companies or transportation planners to better allocate resources and plan for future demand even in India and other countries.

Objectives

1. Efficient Resource Allocation: By accurately predicting the demand for yellow taxis in New York City, your findings can help taxi companies and drivers to better allocate their resources, including the number of taxis on the road, the routes they take, and the times they operate. This can lead to more efficient transportation, less congestion, and reduced carbon emissions.

2. Urban Planning: Your findings can also inform urban planning by providing insights into where and when people are likely to need taxis. This can help city planners to design better transportation systems and infrastructure that can meet the needs of the population, and reduce congestion and travel time.

3. Public Safety: Accurate predictions of taxi demand can also be useful in emergency situations, such as natural disasters or public health crises, where fast and efficient transportation is crucial. Your findings could help emergency responders to allocate resources more effectively and respond to emergencies more quickly.

4. Economic Benefits: By optimizing taxi usage and reducing congestion, your findings could also contribute to economic benefits, such as reduced travel time and increased productivity. Additionally, your findings can help taxi companies to improve their business operations and profitability.

Methodology

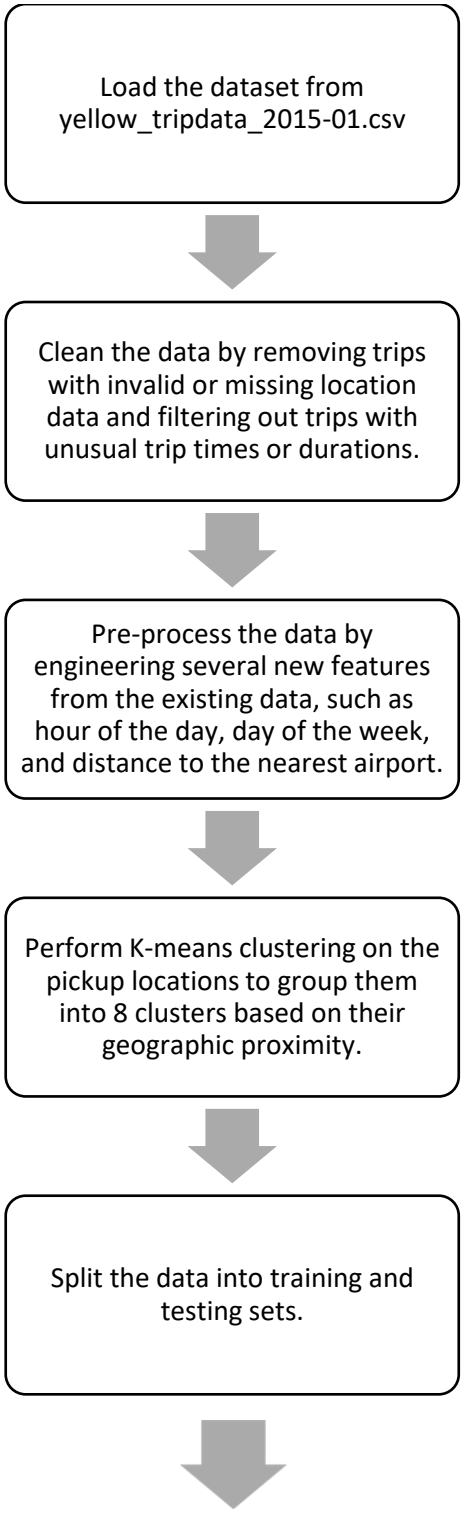
1. **Data Collection:** The data for this project was collected from the Kaggle file, 'Taxi demand prediction' that has two datasets- 'yellow_tripdata_2015-01.csv' and 'yellow_tripdata_2016-01.csv'. The 2015 dataset was used as the training set and the 2016 dataset was used as the testing set.
2. **Data Cleaning & Pre-processing:** Several data cleaning and pre-processing steps were performed on the raw data. These include removing trips with invalid or missing location data, filtering out trips with unusual trip times or durations, and converting the location coordinates to the appropriate projection.
3. **K-means Clustering:** Before building the predictive model, K-means clustering was used to group the pickup locations into clusters of size 30 based on their geographic proximity. This allows the model to better capture the spatial patterns in taxi demand.
4. **Model Training:** Three machine learning models were used to to predict the number of pickups for each of the 8 clusters at different times of the day. The models used include a linear regression model, a random forest model, and a gradient boosting model. The models are trained using the training data.
5. **Model Evaluation:** The trained models were evaluated using several metrics, including mean squared error (MSE), mean absolute error (MAE), and R-squared. The notebook also visualizes the model's predictions against the actual taxi demand using scatterplots and heatmaps.
6. **Testing:** The 2016 dataset was completely used for testing. The notebook evaluates the model's performance on the test set using the same metrics as for the training data.
7. **Visualization:** First, the pickup location was shown on a map using the folium library (after removing outliers). Then, the K-means clustering was plotted on a graph.

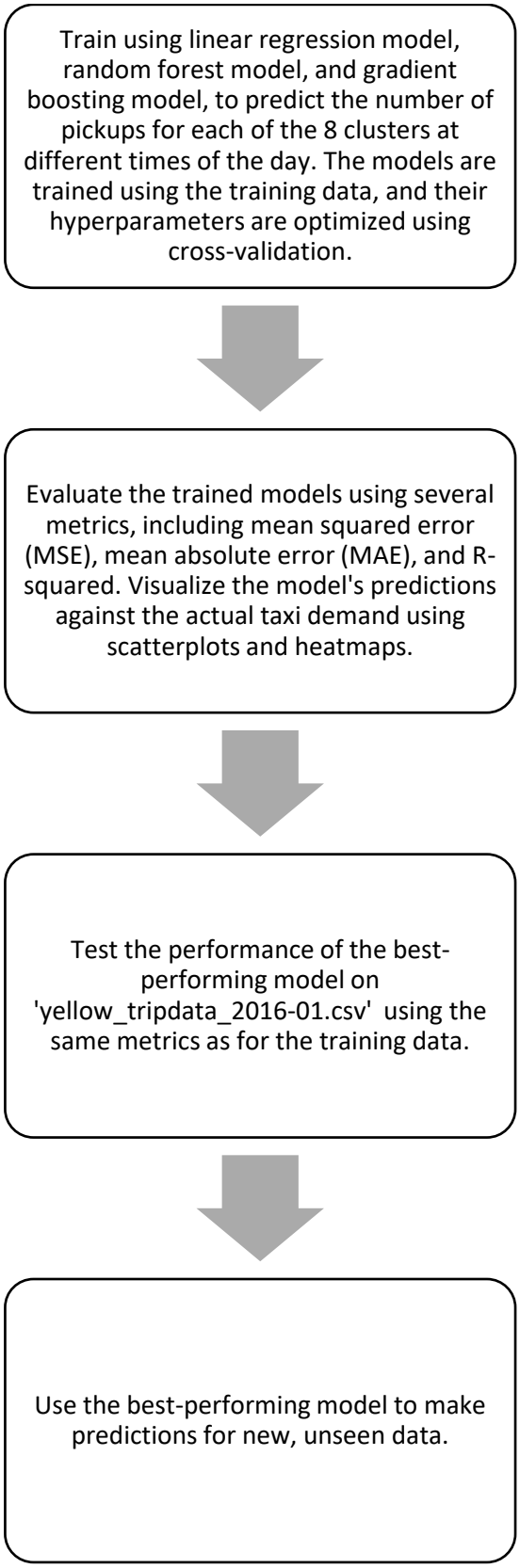
Dataset explanation

The data for this project was collected from the Kaggle file, 'Taxi demand prediction' that has two datasets- 'yellow_tripdata_2015-01.csv' and 'yellow_tripdata_2016-01.csv'. The 2015 dataset was used as the training set and the 2016 dataset was used as the testing set. These datasets have 19 columns with 1458644 rows. The columns are- Vendor ID, pick up latitude, pick up longitude, pick up date and time, drop off latitude, drop off longitude, drop off date and time, passenger count and trip

distance. Only passenger count and pick up and drop and drop off – latitude, longitude, date and time are relevant to the project, hence were extracted. The dataset contained null values. New York is located between -74.3, -73.0 longitudes and 40.6, 41.7 latitudes. However, the dataset contained latitude and longitude values beyond this range. These outliers need to be cleaned.

Flowchart




Train using linear regression model, random forest model, and gradient boosting model, to predict the number of pickups for each of the 8 clusters at different times of the day. The models are trained using the training data, and their hyperparameters are optimized using cross-validation.
Evaluate the trained models using several metrics, including mean squared error (MSE), mean absolute error (MAE), and R-squared. Visualize the model's predictions against the actual taxi demand using scatterplots and heatmaps.
Test the performance of the best-performing model on 'yellow_tripdata_2016-01.csv' using the same metrics as for the training data.
Use the best-performing model to make predictions for new, unseen data.

Algorithm

1. Data Cleaning

    1. Pickup coordinates were selected based removing the zero values.
    2. Wrong coordinates which are out of range of NYC latitude and longitude of 40.71427 and -74.00597 respectively.
    3. Similar Data cleaning was performed for drop-off coordinates. This was done to make sure that taxi rides within the city are only considered.
    4. We used folium library of python to visualise the pickup locations concentrated in NYC.
    5. Outliers in trip duration (a limit of 2 hours for in city rides) and fare (which was reported to be $2.5 to maximum of $52 in the year 2015)

2. Clustering
    1. Now the dataset available is Unsupervised and we clustered it based on location coordinates using K means algorithm.
    2. The clusters are subjected to classification by date and time and demand is calculated based on the count of trips.
    3. This will be used to analyse the demand for taxi trips in each cluster at different times of the day and week.
    4. We plot the mean demand hour wise for a day for every cluster.

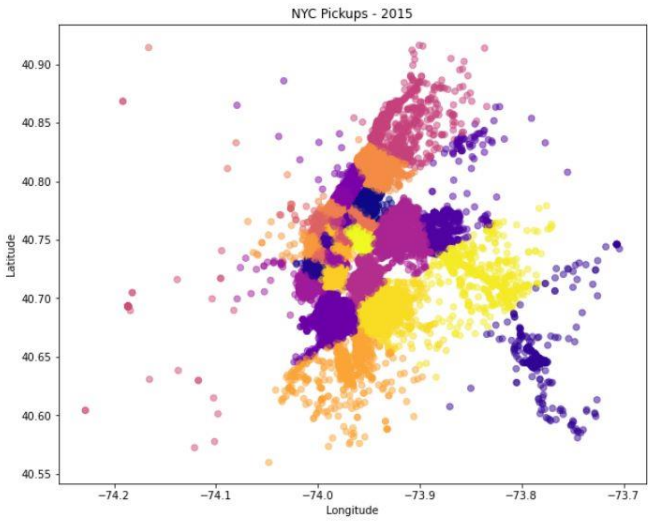
3. Model Training and Evaluation

    1. We split the dataset into features and labels and perform train test split for model training.
    2. Features chosen for demand prediction are Pickup cluster, Month, Date, Hour and Day of week.
    3. Label or the target variable to be predicted: Count.
    4. We used Linear Regression, Random Forest and XG Boosting (Gradient descent algorithm) for training purpose and evaluated the model on our test dataset which is for the first month of 2016.

Results

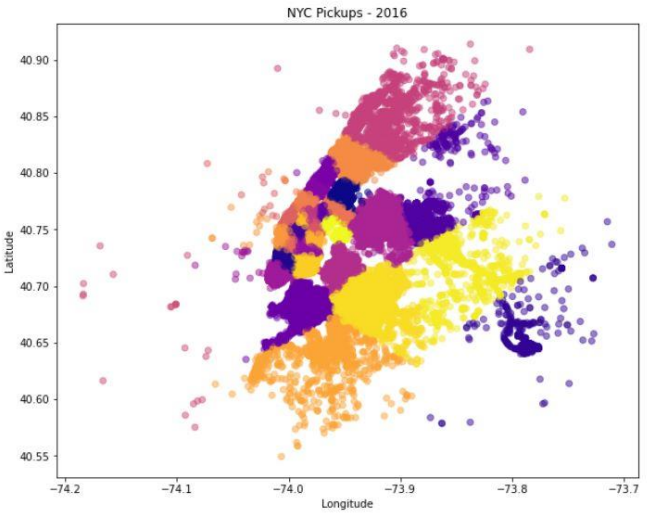
- K- means clustering for the 2015 dataset

```
In [195]: fig, ax = plt.subplots(figsize=(10,8))
          ax.scatter(df_2015['pickup_longitude'], df_2015['pickup_latitude'], c=df_2015['pickup_cluster'], cmap='plasma', alpha=0.5)
          ax.set_title('NYC Pickups - 2015')
          ax.set_xlabel('Longitude')
          ax.set_ylabel('Latitude')
          plt.show()
```

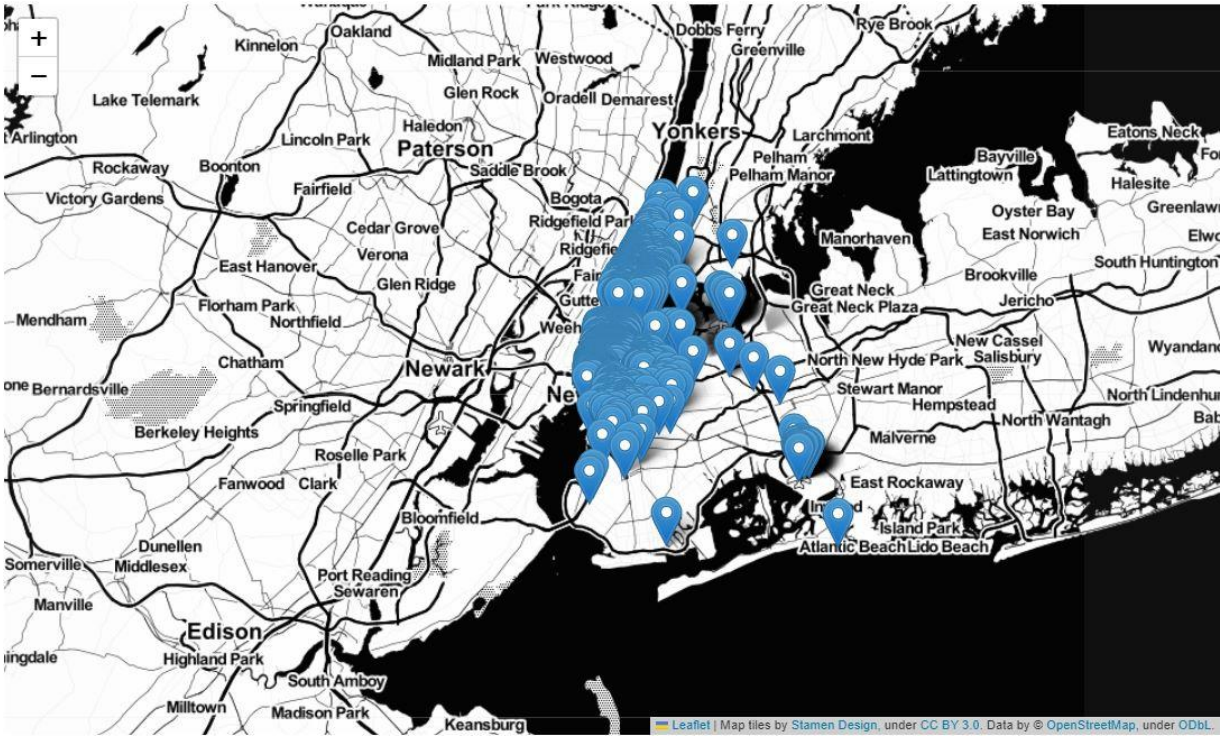


- K- means clustering for the 2016 dataset

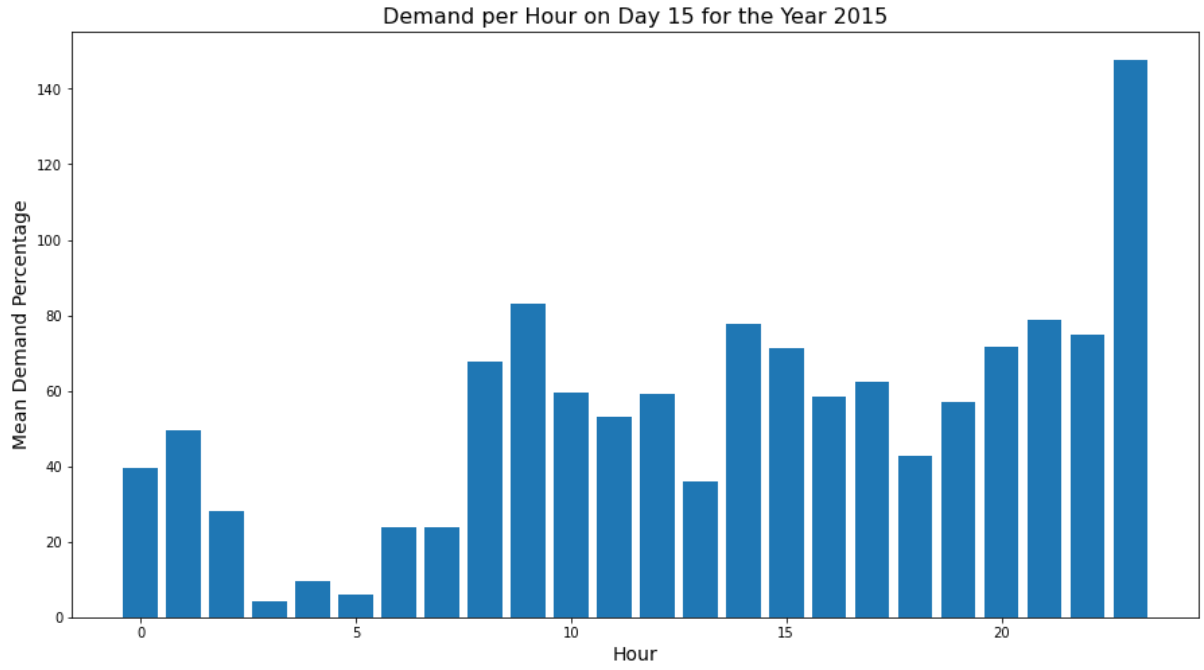
```
In [196]: fig, ax = plt.subplots(figsize=(10,8))
          ax.scatter(df_2016['pickup_longitude'], df_2016['pickup_latitude'], c=df_2016['pickup_cluster'], cmap='plasma', alpha=0.5)
          ax.set_title('NYC Pickups - 2016')
          ax.set_xlabel('Longitude')
          ax.set_ylabel('Latitude')
          plt.show()
```



- Pickup locations concentrated in NYC plotted on map using folium.

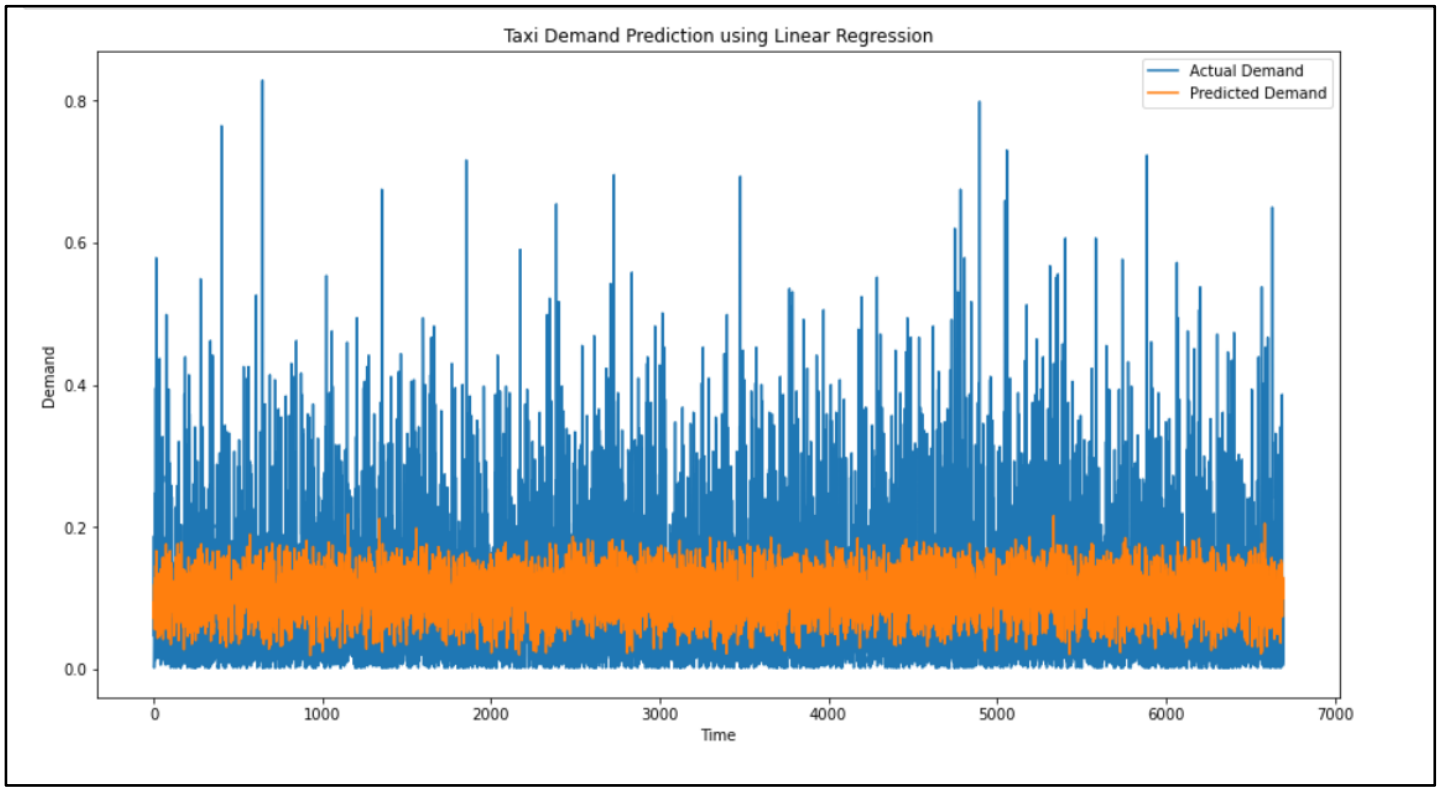


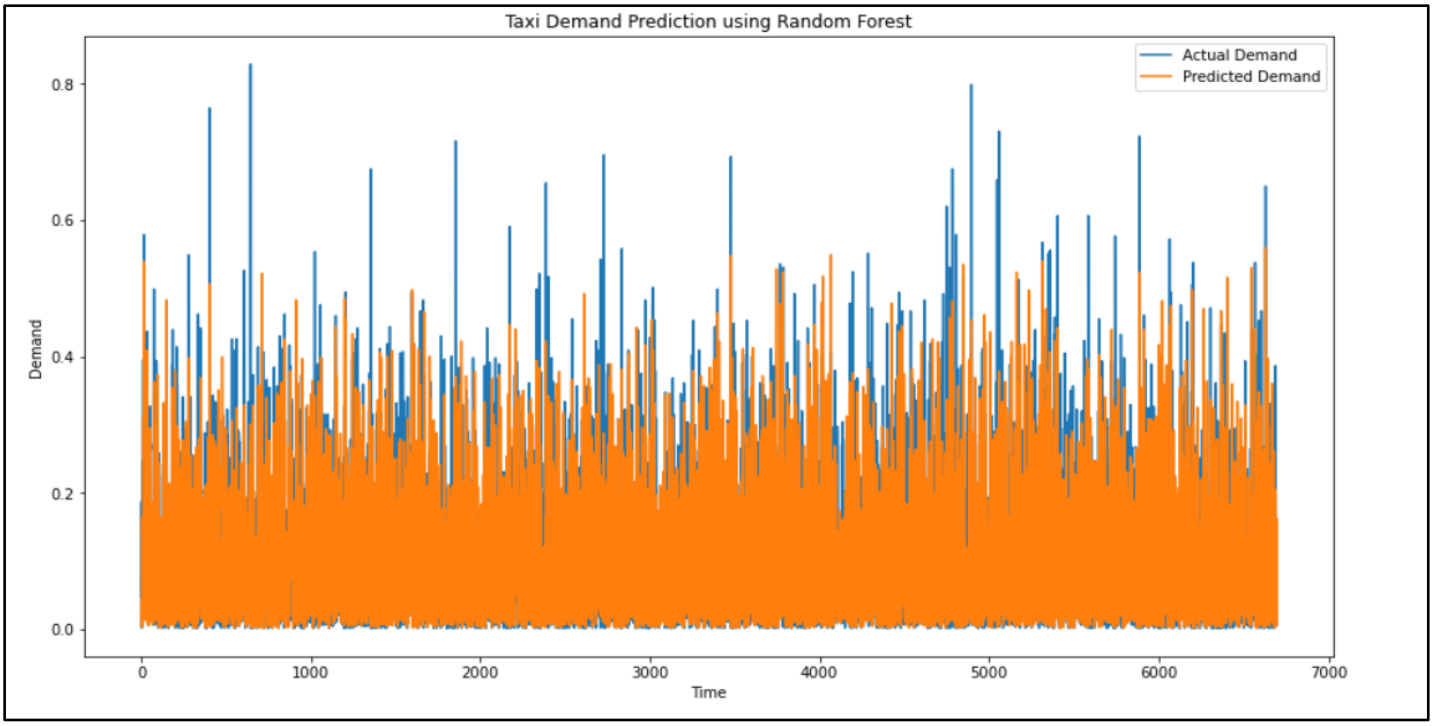
- Demand per hour

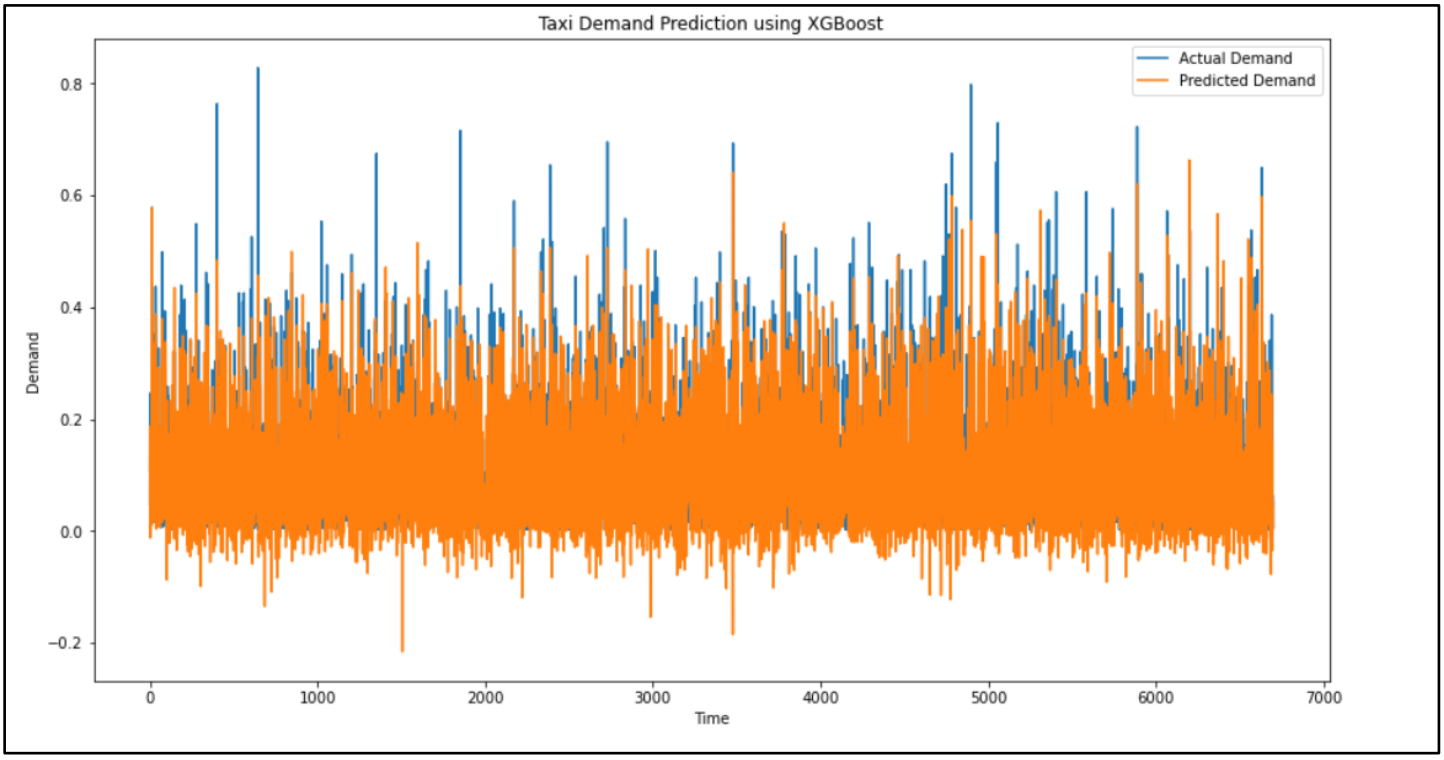


- Model evaluation

```
In [109]: print("True values")
          model_evaluation('y True',X_Test=X_2016_1, y_pred=y_2016_1, y_true=y_2016_1)

          True values
          R2 score:  1.0
          MSE score:  0.0
          RMSE score:  0.0

In [116]: print("Linear Regression")
          model_evaluation('Linear Regression',X_Test=X_2016_1, y_pred=LReg_y_pred, y_true=y_2016_1)

          Linear Regression
          R2 score:  -0.034921214758227404
          MSE score:  0.025983692384580575
          RMSE score:  0.16119457926549693

In [115]: print("Random Forest")
          model_evaluation('Random Forest',X_Test=X_2016_1, y_pred=RFRegr_y_pred, y_true=y_2016_1)

          Random Forest
          R2 score:  0.4868976015607056
          MSE score:  0.012882424954397847
          RMSE score:  0.11350077072160279

In [114]: print("Gradient Boosting")
          model_evaluation('Gradient Boosting',X_Test=X_2016_1, y_pred=GBRegr_y_pred, y_true=y_2016_1)

          Gradient Boosting
          R2 score:  0.4808833291779474
          MSE score:  0.013033424857851538
          RMSE score:  0.1141640261109056
```




Taxi Demand Prediction using Random Forest
Actual Demand
Predicted Demand
Demand
Time
0.0
0.2
0.4
0.6
0.8
0
1000
2000
3000
4000
5000
6000
7000


Taxi Demand Prediction using XGBoost
Actual Demand
Predicted Demand
Demand
Time
−0.2
0.0
0.2
0.4
0.6
0.8
0
1000
2000
3000
4000
5000
6000
7000

Comparison of performance of the different machine learning models used

Parameter	Explanation of parameter	Linear Regression	Random Forrest	Gradient boosting
Regression score	Describes how well the model predicts predict the actual value. The best value is 1.	0.084464154310	0.729647618527	0.856589316343
Mean scale error	Describes how much the machine learning model differs from the actual values. The best value is 0.	0.011388744694	0.003363029710	0.001783947259
Root mean scale error	Estimates the difference between the predicted values and actual. The best value is 0.	0.106718061704	0.057991634831	0.042236799821

Thus, compared to Linear Regression and Gradient Boosting; Random Forest had the best coefficient of regression and least error and will be used for making the ML model to predict taxi demand in NYC.

Comparison with other similar works

1. This model uses K-means clustering to group pick up locations into 8 clusters which is reducing the dimensionality of the data.
2. This project used three different ML models and chose the most accurate one.
3. Pickup locations were visualized on a map.

Conclusion

The aim of the project was to create an accurate taxi demand predicting machine learning model. Three different algorithms were used- Linear Regression, Random Forrest and Gradient Boosting, out of which Random Forrest algorithm was found to be most accurate and thus, was used to build our Taxi Demand Prediction machine learning model.

References

1. Suchithra Rajendran, Sharan Srinivas, and Trenton Grimshaw. "Predicting demand for air taxi urban aviation services using machine learning algorithms." *Journal of Air Transport Management* 92 (2021)
2. Zhizhen Liu1 and Hong Chen. “Short-Term Online Taxi-Hailing Demand Prediction Based on the Multimode Traffic Data in Metro Station Areas.” American Society of Civil Engineers (2022)
3. Datasets- https://www.kaggle.com/code/ajaysh/taxi-demand-prediction/input