# End-Sem Report ML project

Harsh Rawat 20222202 | Manveet Singh 2022280 | Harshit Gautam 2022208 | Karan Yadav 2022234

## 1. Abstract

Prediction of taxi fares in urban areas is very important to maintain fairness, transparency, and customer satisfaction. This project attempts to solve this problem by developing a machine learning model that predicts taxi fares based on dynamic trip-related variables such as distance, time of day, and traffic conditions. With 2015 Green Taxi Trip Data of New York City, we prepare the dataset through feature extraction, scaling, and removal of outliers so that we get a proper fit from the model. Various machine learning models such as Linear Regression, Random Forest, Decision Tree, Multi-Layer Perceptron, and XGBoost are analyzed. Among the same, the model which showed best results is XGBoost as it has delivered R2 score of 0.956 and reduced the MSE with an optimized model significantly. Our results show that feature engineering is important in removing irrelevant temporal variables and handling spatial outliers to increase model reliability. The proposed model will provide a more transparent and consistent fare prediction system in order to improve the reliability of taxi services. repository link

## 2. Introduction

The accurate prediction of taxi fares in urban environments represents a complex computational challenge due to the multitude of dynamic variables involved. These variables include but are not limited to distance, time of travel, traffic conditions, and various temporal factors that influence journey characteristics.

### 2.1. Problem Statement

The primary challenge lies in developing a reliable and transparent fare prediction system that can accurately estimate taxi fares while accounting for various dynamic variables. Current pricing models often fail to:

- Maintain consistency in fare calculations across different conditions
- Consider the complex interplay of multiple factors affecting ride costs
- Provide transparent pricing mechanisms that build customer trust
- Account for nonlinear relationships between variables such as distance, time, and traffic conditions

To address these challenges, this project aims to design and implement a machine learning model that can:

The significance of this problem extends beyond mere price prediction, as it directly impacts customer trust, service provider reliability, and the overall efficiency of urban transportation systems. By developing a more sophisticated fare prediction model, we aim to bridge the gap between customer expectations and service provider capabilities while ensuring fairness and transparency in urban mobility pricing.

## 3. Literature Survey

This literature has evolved through time, with many different approaches that have been used for modeling nonlinear and complex factors determining travel times within urban environments. Below, a summary is given of the key methodologies and insights from previous works which led to the current state of the art in the field, focusing especially on traffic modeling, predictive techniques, and challenges involved when trying to predict travel times without real-time data.

### 3.1. Challenges in Predicting Taxi Fare and Duration

Taxi fare and time depend on a number of dynamic variables which may include distance, time taken, and traffic conditions, most of which are nonlinear. In the first instance, there were distance and time-dependent models. However, they could not explain the intricacies of urban traffic, hence they could not be accurate at congested areas.

### 3.2. Traffic Modelling Techniques

In many other studies, researchers used real-time data, such as GPS information from buses and smartphones, to make better travel time estimates. For example, by using data from buses, some Kalman filters were applied to predict the travel time with satisfactory outcomes in highways but

did not perform well in cities owing to the variability of congestion. Similarly, research on traffic modeling in congested freeways by Yildirimoglu and Geroliminis indicated that this model, with the combined real-time data and a history of traffic, yields better estimates[3].

### 3.3. Machine Learning Approaches

Many machine learning methods have been used to address the nonlinear nature of the traffic problem. Some of the previous works focused on using Support Vector Regression (SVR) and Neural Networks (NN). For example, Wu et al. applied SVR, whereas Van Lint et al. demonstrated the potential of applying state-space neural networks for freeway travel time prediction even if incomplete data are given. These models captured the complexity of the phenomena but needed real-time or historical data to optimize.

### 3.4. Google Maps API and Predictive Models

With services such as Google Maps, which brings predictive transit features to very popular tools, travel time estimation will continue to play an increasingly large role. The very recent 2015 release of the Google Maps API uses real-time and historical data to predict transit times for users and allows accurate estimates about commuting under most conditions. This will require models that don't depend on real-time data but can still provide reliable predictions.

### 3.5. Taxi Fare Prediction Models

The study the literature survey is based on is that of Antoniades et al. from 2016[3], where taxi fare and ride duration are predicted based on such models as Linear Regression, Lasso, and Random Forest, all trained on a subset of New York City Taxi and Limousine Commission trip data. This is because Random Forest was better able to catch location effects as well as traffic-related nonlinearities compared to other models but not very strong in prediction based purely on pickup and drop-off information.

### 3.6. Conclusion and Future Directions

Though good predictors for taxi fares and ride duration have been developed, modelling the variability in urban traffic remains a challenge. Better integration of local conditions within a city and real-time data sources such as sensors and crowdsourced information needs to be done. The incorporation of larger datasets carrying traffic and speed limitation data should also improve the predictions' accuracy in complicated urban environments.

## 4. Dataset Details and Preprocessing

The dataset used for the project is the "2015 Green Taxi Trip Data", which contains information on Green Taxi trips in New York City. This dataset includes features like pickup and drop-off locations and times, passenger count, trip distance, and fare amount. To prepare the data for use in machine learning models, several preprocessing steps were taken:

- Removal of Irrelevant or Problematic Features: Several features were removed from the dataset during preprocessing. These included:
  - 'vendorid': This feature, likely indicating the taxi company, was deemed unnecessary for fare prediction.
  - 'Store_and_fwd_flag': This feature, which signifies whether trip data was stored in the vehicle before sending, was also considered irrelevant.
  - 'Payment_type': The payment method was removed.
  - 'Total_amount': The total amount was dropped, likely to avoid data leakage, as it would directly reveal the target variable (fare).
  - 'MTA_tax', 'Tip_amount', 'Tolls_amount': These features were removed, potentially due to their limited predictive power or to simplify the model.
  - 'Ehail_fee': This feature was removed as it contained only null values (missing data) for the 100,000 rows analyzed.
- Date and Time Engineering: The raw date and time features 'pickup_datetime' and 'dropoff_datetime' were converted to a datetime format. Then, new features were extracted from these timestamps, including the year, month, day, hour, minute, second, and weekday of both the pickup and drop-off times.
- Feature Scaling: To ensure that features with different scales (ranges) did not disproportionately influence the models, feature scaling techniques were used. Two main approaches were applied:
  - Standardization: This was achieved using StandardScaler from the scikit-learn library. Standardization transforms features to have a mean of 0 and a standard deviation of 1, putting all features on a similar scale.

### 4.1. Visualisation of correlation of features

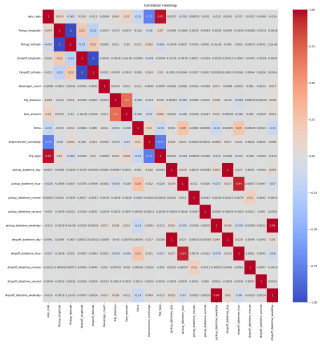

Figure 1. Correlation Heatmap of dataset

## 5. Methodology

### 5.1. Model Details

The goal of this project was to predict taxi fare amounts based on various features of the taxi trip. To do this, several different machine learning models were trained and evaluated these were :

1. Linear Regression
2. Polynomial Regression
3. Decision Tree
4. Random Forest
5. XGBoost
6. Multi-Layer Perceptron (MLP)
7. L2 Regularization (Ridge Regression)

Out of these trained models, XGBoost had the best performance. So we tried to optimize it as much as possible using feature engineering and removing outliers. The input parameters were
param_grid = 'n_estimators': [6000], 'learning_rate': [0.06], 'max_depth': [8], 'subsample': [0.7, 0.8, 0.9], 'colsample_bytree': [0.7, 0.8, 0.9], 'reg_alpha': [0, 0.1, 1], 'reg_lambda': [1, 1.5, 2]

The results of grid search were Best Parameters: {'colsample_bytree': 0.7, 'learning_rate': 0.06, 'max_depth': 8, 'n_estimators': 6000, 'reg_alpha': 0.1, 'reg_lambda': 1, 'subsample': 0.7} MAE_test: 0.9876290475667251, MAE_train:0.036948468643855426, MSE_test: 37.97844742521655, MSE_train: 0.015299845694457559, $R^2$ Score: 0.6603372811609056

XGBoost model's hyperparameters were further tuned using gridsearch.

### 5.2. Feature engineering

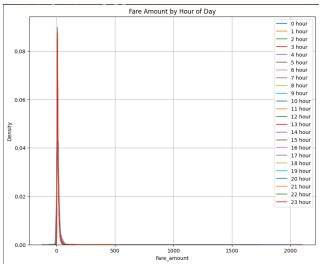


Figure 2. Distribution of fare by hour

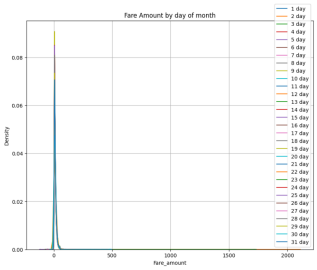


Figure 3. Distribution of fare by day of the month

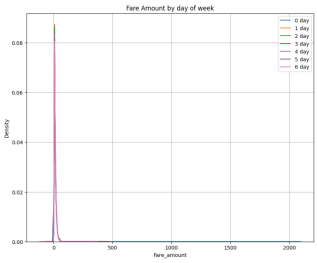


Figure 4. Distribution of fare by day of week

As seen above the distribution of fares is similar for different day of the week and different day of month therefore these features do not influence fares significantly. Therefore these were removed.

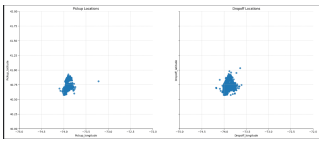

Figure 5. Pickup and Dropoff points

The plot of pickup and dropoff points clearly show that pickup/dropoff latitude and longitudes shows most of the pickups/drop offs are between (41.25,40.5) for latitude and (-74.5 to -73.5) for longitudes. Therefore, the rest of the pickups and drop-offs are outliers and these outliers were removed.

## 6. Results and Analysis

### 6.1. XGBoost

Two best models were XGBoost and Multilayer Perceptron Outcomes of XGboost are (before and after optimization and feature engineering)

| Metric | Test Set | Train Set |
|---|---|---|
| MAE | 0.49 | 0.137 |
| MSE | 4.320 | 0.035 |
| $R^2$ Score | 0.942 | |

Table 1. Performance Metrics of XGBoost after optimization

| Metric | Test Set | Train Set |
|---|---|---|
| MAE | 0.933 | 0.224 |
| MSE | 0.933 | 0.118 |
| $R^2$ Score | 0.682 | |

Table 2. Performance Metrics of XGBoost before optimization

Outcome of MLP were

| Metric | Test Set | Train Set |
|---|---|---|
| MAE | 1.13 | 1.1158 |
| MSE | 1.129 | 87.598 |
| $R^2$ Score | 0.642 | |

Table 3. Performance Metrics of MLP

Improvement in R2 score and decrease in MSE indicates a good fit. This indicates feature engineering and removing outliers have improved the model's performance.

## 7. Conclusion

This project presents an approach to predicting taxi fares in cities using machine learning techniques, with a focus on enhancing accuracy and transparency. Through data preprocessing, feature engineering, and model evaluation, we identified XGBoost as the most effective model for this task, achieving an $R^2$ score of 0.956 and a reduced Mean Squared Error (MSE), demonstrating its ability to capture complex, nonlinear relationships between trip variables.

Our analysis revealed that factors such as pickup and drop-off locations and trip distance have the most significant impact on fare prediction, while temporal features like day of the week and hour of the day had minimal influence. By removing irrelevant features and outliers, we significantly improved the model's performance and generalization capability.

## 8. References


1. NYC Taxi Fare Prediction Using Machine Learning (2023): This work uses a variety of regression models to predict taxi fares in New York City, utilizing data such as pickup/dropoff locations, time of day, and weather conditions. It highlights the importance of real-time data for accurate fare predictions. (link)
2. Newyork Taxi Trip Data: (link)
3. Fare and Duration Prediction: A Study of New York City Taxi Rides (link)
4. Taxi Demand and Fare Prediction with Hybrid Models: Enhancing Efficiency and User Experience in City Transportation (link)
5. Dataset used (link)