Mid-Sem Report ML project

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

1. Motivation

Urban mobility is experiencing a dramatic shift in the direction of taxi services and carpooling platforms. However, the recurring occurrence of increased fares and surges for so many factors, such as traffic conditions, fuel prices, and demand in rides, becomes noticeable concerns to the commuter. Increment changes like those make customers unhappy and lose their trust in the pricing mechanisms. This project focuses on the design of a machine-learning model that can predict taxi fares fairly by considering are dynamic variables. Basically, the focus is on strengthening consistency and transparency in pricing, which will lead to more satisfactions of customers and allowing service providers to maintain trust and fairness.

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.
  - Log Transformation: The 'Trip_distance' feature was subjected to a log transformation. This is a common technique to handle skewed distributions, where a few very large values can disproportionately influence the model. The log transformation compresses the range of the data, reducing the impact of outliers.

4.1. Visualization of correlation of features

:

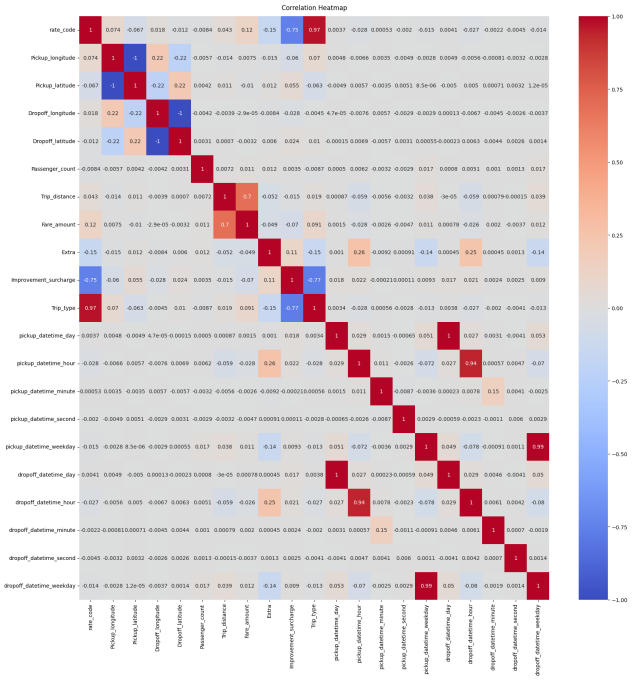

Figure 1. Correlation Heatmap of dataset

5. Methodology and 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:

1. Linear Regression: This is a basic statistical model that assumes a linear relationship between the input features and the target variable (fare amount). It serves as a baseline model to compare with more complex algorithms.

2. Polynomial Regression: This extends linear regression by adding polynomial terms (e.g., squared or cubed features) to the model. This allows the model to capture non-linear relationships between the features and the target variable. Models with polynomial degrees ranging from 1 to 3 were tested.

3. Decision Tree: This model uses a tree-like structure to make decisions based on the input features. Each node in the tree represents a decision based on a feature, and the leaves of the tree represent the predicted fare amount.

4. Random Forest: This is an ensemble learning method that combines multiple decision trees to improve prediction accuracy and generalization ability. Each tree is trained on a random subset of the data, and the final prediction is made by averaging the predictions of all the trees.

5. XGBoost: This is a gradient boosting algorithm known for its efficiency and performance in various machine learning tasks. It works by sequentially adding decision trees, with each tree attempting to correct the errors made by the previous trees.

6. Multi-Layer Perceptron (MLP): This is a type of artificial neural network that can learn complex non-linear relationships in the data. It consists of multiple layers of interconnected nodes that process the input features and produce a prediction for the target variable.

7. L2 Regularization (Ridge Regression): This is a technique used to prevent overfitting in linear regression models by adding a penalty term to the model's loss function. This penalty encourages the model to have smaller weights, which can make it more robust to noise in the data and improve its generalization performance.

6. Results and Analysis

6.1. Linear Regression

Metric	Test Set	Train Set
MAE	1.888191285176301	1.9012855842785725
MSE	1.888191285176301	101.96563891345184
R^2 Score	0.5700722997131895	

Table 1. Performance Metrics

- Clear underfitting pattern
- Higher MSE values compared to more sophisticated models
- Lower R² score indicating poor fit to the data pattern

6.2. Polynomial Regression : Degree 2

Metric	Test Set	Train Set
MAE	1.1827	1.121
MSE	1.182	85.79
R^2 Score	0.390	

Table 2. Performance Metrics

- Slight improvement over degree 1
- Still exhibits underfitting
- Moderate improvement in R² score

6.3. Random Forest

:

Metric	Test Set	Train Set
MAE	1.459	0.558
MSE	1.459	17.632
R^2 Score	0.5995	

Table 3. Performance Metrics(of random forest)

- Better generalization than single decision tree
- Lower MSE values than simpler models
- More stable R^2 scores

6.4. Multi-Layer Perceptron (MLP)

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

Table 4. Performance Metrics

- Shows underfitting in current configuration
- Consistent but suboptimal performance

6.5. XGBoost

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

Table 5. Performance Metrics

- Most balanced metrics overall
- Best performance across all evaluation criteria

6.6. Ridge Regression

Metric	Test Set	Train Set
MAE		
MSE	48.0717	87.598
R^2 Score	0.57006	

Table 6. Performance Metrics

- Similar performance to linear regression
- Higher MSE values compared to advanced models
- Lower R^2 score indicating poor fit

7. Graph of XGBoost and MLP

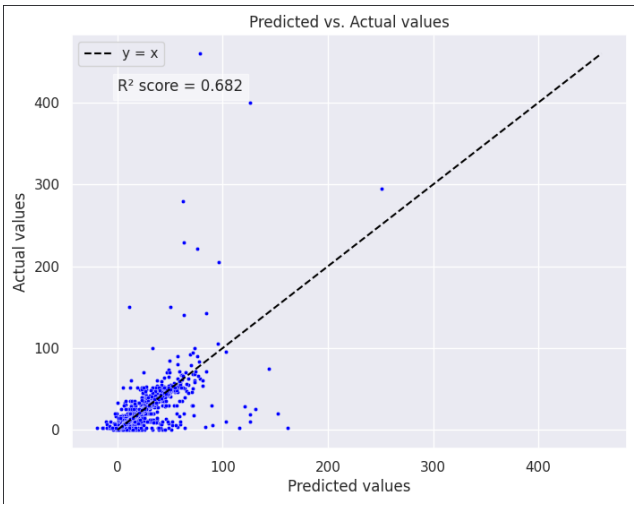

Figure 2. XGBoost prediction vs actual values

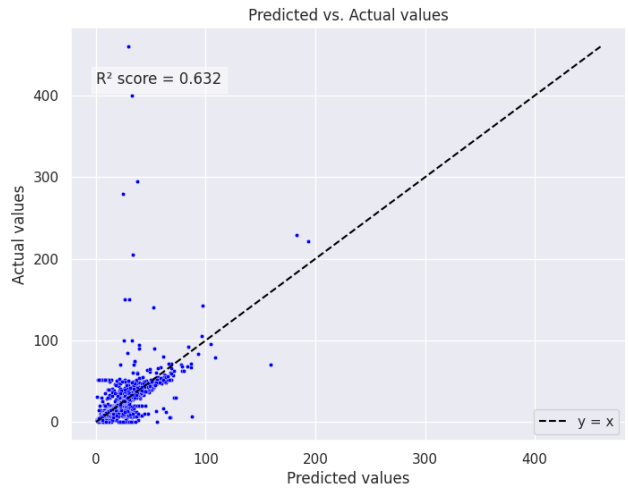

Figure 3. Prediction of Multilayer perceptron

8. Conclusion

In this interim report, we explored multiple machine learning models for predicting taxi fares, focusing on their ability to account for complex factors such as traffic, distance, and time of travel. Among the models tested, XGBoost stood out as the best performer, delivering the most accurate fare predictions with the lowest Mean Squared Error (MSE) of 0.933, highest R^2 score of 0.682, and lowest Mean Absolute Error (MAE) of 0.933. Its strong generalization capabilities allowed it to effectively capture nonlinear relationships and deliver consistently accurate predictions across diverse conditions. Multi-Layer Perceptron

(MLP) also demonstrated potential, with an MSE of 38.85 and an R^2 score of 0.652, but lagged behind XGBoost in accuracy and consistency. The performance gap between the two models indicates that XGBoost's superior handling of feature interactions and better bias-variance trade-off make it the most reliable model for this fare prediction task so far.

9. 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)