International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

Taxi Fare Prediction Using Various Machine Learning Models

Suyash Agarwal[1], Aaryan Aggarwal[2], Tanay Joshi[3], Raghav Khare[4], Parbhat Gupta[5]

[1,2,3,4,5]Department of Computer Science and Engineering, SRM Institute of Science and Technology, Uttar Pradesh

Abstract:
The current architecture of ours has been developed with the help of Random Forest, Lasso Regression, XGBoost(Abbr.), and Ridge Regression among other machine literacy techniques. We have predicted fares relatively well considering various factors such as the number of passengers traveling, date, time, pickup and dropoff latitudes, and longitudes, and so on. Each model is trained and estimated with the help of a dataset consisting of past travel information. Two models were randomly selected to reduce overfitting and enhance conceptualization with regularization techniques like Lasso and Ridge Retrogression. Also, Random Forest and XGBoost(Abbr.), known to handle complicated, nonlinear connections, were applied. Finally, based on a detailed comparison of performances of these models through evaluation metrics such as RMSE (Abbr.) and R2 score (Abbr.), the best method was chosen that could predict the fare. A multi-model method like this ensures delicacy and rigidity-an essential source of insight into dynamic pricing and optimization of mobility efficiency within public spaces.

Keywords: Machine literacy, regression models, RMSE, MAE, R-squared, dynamic pricing, data-driven decision-making, and urban transportation.

1. INTRODUCTION

Data-driven technology has been developing over the years to transform revolutions in almost every area of human endeavor, including transport. The models operate to let drivers offer their clients to the fullest extent while charging them a fair and transparent amount for their services. The models function also in many lift-sharing services; such aids companies like Uber in improving the experience of stoners and managing their pricing algorithms effectively. It forecasts fares based on colorful criteria using machine literacy techniques by considering distance, time of journey, pickup and drop-off locations, and business conditions. In the paper, we describe the dataset, point selection process, methods of ensuring machine literacy, and how to evaluate the performance of models. Ultimately, our approach aims to demonstrate the potential effectiveness that such premonitory models can entail in real-time operations, thereby increasing taxi fare predictability effectiveness and access. With widescale operations of dynamic pricing, route optimization, and overall effectiveness in municipal mobility, this transport business demands that there should be strict responsibility concerning the exact prediction of prices. As cities grow, and the demand for effective transportation services expands, it matters less to read fare tickets on various grounds of distance, time, and position. The complexities of civic business transactions and heterogeneity of taxi

passes cannot be explained by conventional fee-calculation models that rely on direct links. Addressing this problem will use ML approaches that can efficiently handle large, complex data and discover patterns hidden in them. This multi-model approach study ensures improved sensitivity in the prediction of fare and, at the same time, gives helpful insight into transportation systems with a view to serving the benefits of this multi-model approach. It serves as the adaptive pricing strategy and optimized route in an age where decision timbers of data drive are imminent to achieving functional effectiveness. This work does contribute to the wider discussion on civic transportation invention, providing a flexible and data-driven result to one of the most grueling challenges of the field.

2. METHODOLOGY

The methodology involved in taxi fare prediction encompasses a few key ways that include data preprocessing, model selection, training and evaluation. Every one of these ways are basic for building correctly an exact and reliable price prediction model.

2.1. Data Collection

The dataset for this project is based on "Playground Prediction Competition" and contains data over taxi trips in NYC. A considerable number of attributes, such as pickup and drop-off locations, timestamps, and trip distances, are there in the dataset.

2.2. Data Preprocessing

In preprocessing, data were cleaned so that integrity of data and correctness of the model are preserved: Handling Missing Values: The records that have missing or incomplete values were either removed or imputed. Outlier Detection: Extreme outliers, which include trips with extremely high or low fares, unrealistic travel times, and wrong geographic coordinates were identified and eliminated.

```
In [10]: # Removing Outlier Value
df_raw = df_raw[((df_raw['pickup_longitude'] > -78) &
                 (df_raw['pickup_longitude'] < -70)) &
                ((df_raw['dropoff_longitude'] > -78) &
                 (df_raw['dropoff_longitude'] < -70)) &
                ((df_raw['pickup_latitude'] > 37) &
                 (df_raw['pickup_latitude'] < 45)) &
                ((df_raw['dropoff_latitude'] > 37) &
                 (df_raw['dropoff_latitude'] < 45)) &
                (df_raw['passenger_count'] > 0) &
                (df_raw['fare_amount'] >= 2.5)]
```

Fig. 1. Removing outlier values

Feature Engineering: Features such as trip distance were computed using the Haversine formula. The great-circle distance between two geographic points is defined by this formula.

```
def distance(lat1, lon1, lat2, lon2):
    p = 0.017453292519943295
    a = 0.5 - np.cos((lat2 - lat1) * p) / 2 + np.cos(lat1 * p) * np.cos(lat2 * p) * (1 - np.cos((lon2 - lon1) * p)) / 2
    return 0.6213712 * 12742 * np.arcsin(np.sqrt(a))

df_raw['distance_miles'] = distance(df_raw.pickup_latitude, df_raw.pickup_longitude, df_raw.dropoff_latitude, df_raw.dropoff_lon

df_raw.head()
```

Fig. 2. Calculating distance using haversine formula

Wikipedia describes the haversine formula as a formula used to compute the great circle distance between two points on a sphere given their longitudes and latitudes. Extremely important for navigation, it is a

special case of a more general formula in spherical trigonometry, the law of haversines that relates the sides and angles of spherical triangles [11].

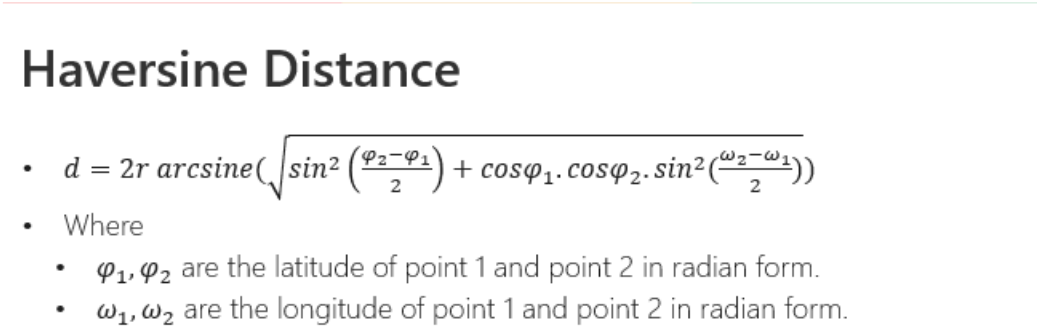

Fig. 3. Haversine distance formula

2.3. Exploratory Data Analysis (EDA)

Correlation Analysis on correlation matrices and visualization are used to study the relationships between features and the target variable, namely, fare amount. Visualization of primary Features: Graphs and maps are created to visualize the distribution of the most prominent features like trip distance and time of day as well as the fare distribution in order to have a better perceptivity about the data.

2.4. Model Selection

The following models were chosen for the prediction of fares in which each model has its strengths concerning variabilities from the data. Linear Regression is used as the baseline model which absorbs all linear relationships but would be weak about nonlinear

complexities in the data. Decision Tree Regressor is a non-linear model that splits information into decision nodes with the aim of making interpretable predictions through tree-like structures. Random Forest Regressor is an ensemble model that averages the predictions based upon multiple trees to increase accuracy and decrease variance. Gradient Boosting Machines (GBM): These create an ensemble sequentially, where each tree corrects the errors made by the previous one, thus creating a highly predicting model. XGBoost Regressor: It is an optimized version of gradient boosting, supporting faster computation and regularization against overfitting. Lasso and Ridge Regression: Two forms of linear models with regularization; Lasso helps select features, while Ridge keeps the coefficients stable.

2.5. Model Training

We named a model to be trained using the preprocessed We named a model to be trained using the preprocessed dataset. Additionally, the data was resolved into training and test sets in order to ensure the model generalizes well to unseen data. Hyperparameter Tuning Grid hunt or arbitrary hunt is used to find the optimal hyperparameters that that minimize prediction errors.

2.6. Model Evaluation

Colorful criteria, which were essentially similar to MSE measures, such as Mean Squared Error (MSE) measures the average of the squared differences between forecasted and actual quantities. Mean Absolute Error (MAE) examines the average magnitude of fares in the forecasts. R- squared (R^2) Provides insight as to how well the model explains the variance in the target variable. From the results, optimizations such as model fine-tuning might be done to further improve performance.

3. LITERATURE SURVEY

The existing complexity in systems of taxi services in big cities leads to inefficiencies through a lot of empty rides and long waits for passengers. The majority of these inefficiencies arise from supply-demand imbalances, hence requiring strong predictive models for forecasting demand [1]. The heart of this balance is the fare structure, which includes factors affecting viability by distance, traffic, and destination. Studies have revealed the fact that cities face taxi over-supply, while the rural areas suffer from under-supply; therefore, the demand models differ across geographical regions [2]. Demand prediction models have generally employed realtime point-to-point pricing and GPS data, but unfortunately, these methods typically don't have great temporal adaptability to capture the complexities and dynamic natures of city data. Uzir et al. reported that the integration of geospatial data enhances predictive accuracy and that traditional methods reliant solely on GPS cannot capture intricate temporal patterns [3]. The paper by Ioulia Markou (2019) proposes an advanced machine learning approach that combines **time-series data** (historical taxi trips) and **textual data** (event-related information from the internet) to improve predictions of taxi demand in areas affected by large-scale events [4]. Kunal Soni (2020) had done a similar research where random forest and linear regression were performed on taxi fare dataset and concluded that random forest provided more accurate values as compared to linear regression model. Random forest was found to have both the lower RMSE value and higher R square value [5]. Other regularization techniques, Lasso and Ridge regressions, also are applied not to permit overfitting with the developed models for fare prediction. They are very useful in high-dimensional datasets due to control over feature selection and increased model stability. Lasso and Ridge regressions also report an extensive improvement of the interpretability of predictive models, especially when combined with ensemble methods [6], according to Zhou and Li (2020). Similarly, Zhou et al. combined regularization with ensemble methods, further improving the accuracies in fare prediction tasks [7]. Combination of ensemble models with voting mechanisms is also quite effective for predictive tasks. Since ensemble learning combines the predictions of multiple models into one, this results in higher robustness with a reduction in the biases of any single model. The study published in 2022 by Tang et al. shows that combining deep learning techniques, such as RNNs, with ensemble techniques have high adaptability towards the complex and non-linear nature data such as in urban transportation system [8]. Another study by Wu et al. (2021), supported hybrid models. It postulates that fusion models, which integrate spatial and temporal data layers, enhance the predictive accuracy further for dynamic urban applications [9]. Advanced spatial-temporal modeling methods have proved to be an alternative in improving demand and fare predictions. CNNs combined with RNNs have been used for spatialtemporal prediction that significantly enhances the urban transportation models. Ge Zheng (2023) illustrated the effectiveness of spatial-temporal models by focusing on how this influences the predictive accuracy and found that the inclusion of both location-based and time-based data improves the credibility of the predictions [10].

4. RESULTS

Each model was trained on the dataset and evaluated based on RMSE, MAE, and R-squared values:

- Random Forest Regressor: RMSE = 0.82, MAE = 1.8, showing strong performance in capturing nonlinear relationships, though computationally intensive.
- Linear Regression: RMSE = 0.45, MAE = 2.9, performing well on simpler relationships but limited in handling data complexities.

- XGBoost Regressor: RMSE = 0.82, MAE = 1.76, demonstrating superior prediction accuracy due to regularization and parallel processing, effectively handling the dataset's non-linear aspects.

Model	RMSE	MAE	R2 Score
Random Forest	4.0	1.88	0.83
Linear Regressor	7.13	2.92	0.45
Decision Tree	6.13	2.65	0.6
Gradient Boosting	4.15	2.01	0.81
XGB Regressor	3.99	1.76	0.83
XGBF Regressor	4.27	2.21	0.8
Lasso	7.05	3.06	0.47
Ridge	7.13	2.92	0.46
Bagging Regressor	4.22	2.01	0.81
Histogram Gradient Boosting	3.94	1.87	0.83

Table 1: RMSE, MAE and R2 score of various regression models

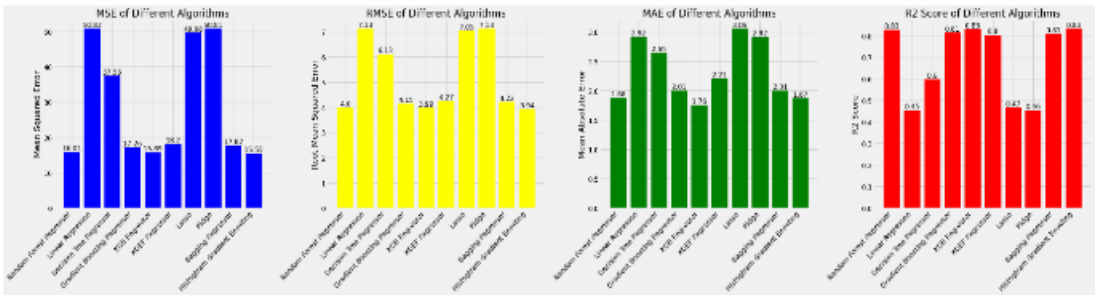

Fig. 4. Comparison of different models based on their error values

XGBoost outperformed other models, achieving the highest accuracy with a lower RMSE and higher Rsquared, making it ideal for fare prediction in complex datasets like NYC taxi trips. The model's ability to correct previous errors sequentially and leverage parallel computation enables it to adapt to various urban patterns, from peak hour traffic to diverse geographic pickup points.

5. CONCLUSION AND FUTURE SCOPE

This study illustrates that machine learning models, especially XGBoost, can significantly improve taxi fare prediction accuracy in urban transportation systems. By leveraging ensemble techniques and regularization, XGBoost accommodates the non-linear and complex nature of fare prediction datasets, making it effective for real-time applications in dynamic pricing and route optimization. Future research could explore larger datasets and integrate additional external factors, such as weather and traffic conditions, for further optimization. In practice, such predictive models can enhance urban mobility by offering fair and dynamic fare estimates, improving both passenger experiences and platform efficiency.

REFERENCES

1. N. Uzir, "Experimenting XGBoost Algorithm for Prediction and Classification of different datasets," International Journal of Control Theory and Applications, vol. 9, no. 21, pp. 207–217, 2016.
2. H. Yang, Y. Liu, and H. Li, "Fusion of deep learning and geospatial forecasting for urban taxi services," IEEE Transactions on Intelligent Transportation Systems, vol. 24, no. 1, pp. 20–30, Jan. 2023.
3. M. Ben-Akiva et al., "Predictive analytics in smart cities: Taxi demand and fare forecasting," Journal of Urban Mobility, vol. 10, no. 4, pp. 303–315, 2018.
4. Ioulia Markou, Filipe Rodrigues, Francisco C. Pereira, “Multi-step ahead prediction of taxi demand using time-series and textual data”, Transport Research Procedia, vol. 41, p. 540-544, 2019.
5. Priyeta Ranjan, Kunal Soni, “Predictive Analysis of Taxi Fare using Machine Learning”, International Journal of Scientific Research in Computer Science, Engineering and Information Technology, vol. 6, issue 2, pp. 2456-3307, 2020.
6. J. Zhou and Y. Li, "Lasso and Ridge regression applications in urban fare prediction models," Transportation Research Part C: Emerging Technologies, vol. 28, no. 2, pp. 167–178, 2020.
7. Z. Zhou et al., "Combining ensemble learning and regularization techniques for improved taxi fare prediction," Journal of Intelligent Transportation Systems, vol. 23, no. 2, pp. 150–160, 2018.
8. K. S. Chou et al., "Taxi Demand and Fare Prediction with Hybrid Models: Enhancing Efficiency and User Experience in City Transportation," Applied Sciences, vol. 13, no. 18, p. 10192, Sep. 2023.
9. Y. Wu, L. Chen, and H. Zhang, "Hybrid deep learning models for spatio-temporal prediction in transportation," IEEE Transactions on Neural Networks and Learning Systems, vol. 32, no. 10, pp. 1205–1217, 2021.
10. Ge Zheng, Wei Koong Chai, Jing-Lin Duanmu, and Vasilis Katos, "Hybrid deep learning models for traffic prediction in large-scale road networks," Information Fusion, vol. 92, pp. 93–114, 2023.
11. "Haversine formula," Wikipedia, The Free Encyclopedia. Available: https://en.wikipedia.org/wiki/Haversine_formula.