# FareMiner: Fare Prediction and Anomaly Detection in NYC Taxi Data using Machine Learning

Archit Singh
Dept. of ICT
Manipal Institute of Technology
Manipal, India
220911464

Prashast Saxena
Dept. of ICT
Manipal Institute of Technology
Manipal, India
220911536

Arnav Aradhya
Dept. of ICT
Manipal Institute of Technology
Manipal, India
220911612

***Abstract*—Urban mobility analytics play a key role in improving the efficiency and openness of transportation systems. This article presents FareMiner, a machine learning-based pipeline for estimating taxi fares and detecting anomalous riding behaviors in NYC. We extract supply-demand, temporal, and spatial characteristics from the NYC Taxi Fare dataset on Kaggle to train regression models including Linear Regression, Random Forest, and XGBoost for fare prediction. To find anomalies, we employ statistical techniques like Z-Score and IQR, clustering with DBSCAN, and residual-based analysis with Isolation Forests. A Streamlit-based user interface enables real-time prediction, anomaly visualization, and interactive exploration. The technology aims to enhance fare transparency, prevent fraudulent activities, and improve data quality in smart transportation networks.**

## I. INTRODUCTION

Large volumes of data about urban mobility, especially in places like New York, have been produced by the growth of ride-hailing apps. These datasets present a chance to use machine learning and data mining to glean insights. In this study, we use the publicly available NYC cab dataset to estimate cab fares and detect rider abnormalities.

In addition to supporting dynamic pricing schemes, fare prediction systems can improve rider trust and fare transparency. On the other hand, anomaly detection makes it possible to spot odd travel trends, fare theft, and problems with data quality. We suggest FareMiner, a comprehensive solution that creates a powerful fare analytics tool by combining supervised regression and unsupervised anomaly detection.

The system supports real-time exploration via Streamlit and provides interactive visualizations, including spatial heatmaps and anomaly clusters. Through this dual-task approach, our goal is to enhance operational efficiency for taxi operators and improve fare accuracy for commuters.

## II. LITERATURE REVIEW

The issue of fare prediction has been addressed in a number of studies utilizing both ensemble-based and conventional regression models. Jiang et al. (2019) showed how to describe non-linear interactions in taxi fare estimate using gradient boosting trees. In a similar vein, Guda and Veloso (2018) emphasized how incorporating outside variables like traffic and weather might increase model accuracy.

Chandola et al. (2009) provided an extensive overview of outlier detection techniques that are relevant to spatial-temporal data in the area of anomaly detection. In large-scale mobility datasets, more contemporary methods such as DBSCAN and Isolation Forest have demonstrated efficacy in detecting unusual rides. For example, Liu et al. (2021) used a hybrid model that combined feature clustering with residual-based anomalies to identify fare inconsistencies in NYC taxi data.

Our study extends these frameworks by integrating processes for anomaly detection and prediction into a single system. The additional layer of interactivity provided by Streamlit gives stakeholders real-time access to data and visualizations for useful insights.

## III. RESEARCH GAPS

Although fare modeling and anomaly detection have advanced, there are still significant obstacles to overcome:

- A lot of current models don't incorporate anomaly detection tools to verify ride fairness; they are only made for prediction.
- The influence of contextual supply-demand dynamics in fare modeling is only partially covered in the literature; most research lacks a user-facing application that can interactively detect outliers and visualize patterns; few open-source platforms provide both interpretable prediction and real-time anomaly detection in a single interface.
- There is a lack of research on fare anomalies associated with uneven supply or excessive price during off-peak hours.

By implementing a dual-purpose system that offers an interactive analytics interface in addition to fare prediction and anomaly detection, this research seeks to close these gaps.

## IV. OBJECTIVES

- To create a data pipeline that analyzes the NYC Taxi Fare information in order to identify anomalies and make predictions.
- To extract demand-supply-based, temporal, and spatial aspects from unprocessed ride data.

- To use and evaluate many regression models for fare prediction, including XGBoost, Random Forest, and Linear Regression.
- To use statistical metrics (Z-Score, IQR) and unsupervised learning techniques (Isolation Forest, DBSCAN) for anomaly identification.
- To develop a real-time Streamlit application for forecasting, anomalies, and fare trends.

## V. MOTIVATION

Urban transportation systems constantly struggle to strike a balance between affordability, transparency, and efficiency. Even though taxi fare structures frequently make the claim that they are uniform, a number of variables, including traffic, the time of day, and local demand, can cause pricing irregularities or even fare anomalies. Inaccurate pricing has the potential to damage commuter confidence and impact how transportation services are perceived generally.

The expansion of data availability, particularly from intelligent fleets with GPS, offers a great chance to use machine learning to solve these problems. We can estimate average pricing and identify discrepancies that require further inquiry by modeling historical ride behavior.

- The desire to increase confidence and openness in commercial and public transportation services is what drives us.
- A desire to find and fix deceptive trends or data flaws that distort analytics.
- The ability to provide real-time anomaly information and fare analysis tools to mobility companies and transport agencies.

Combining unsupervised anomaly detection with supervised learning for prediction presents an intriguing technological challenge. Both analysts and service providers can gain a great deal from developing a solution that not only simulates normal ride behavior but also flags unusual trips via an intuitive dashboard.

## VI. SUMMARY

Using the NYC Taxi Fare dataset, this study presents FareMiner, a dual-model machine learning framework for fare prediction and anomaly detection. It does the following:

- Prediction: Using variables like time, location, and supply-demand indicators, models like Random Forest and XGBoost can estimate fares accurately.
- Detection: Z-Score-based analysis of engineering features and prediction residuals, Isolation Forest, and DBSCAN are used for unsupervised anomaly identification.
- Visualization: A dashboard driven by Streamlit that offers outlier trends, anomaly clustering, interactive heatmaps, and prediction visualizations.

By integrating detection and prediction into a single platform, our system offers an understandable and practical approach for urban fare analytics. It guarantees data quality, aids in identifying unfair pricing, and enables more thoughtful policy decisions in mobility services.

FareMiner is suitable for a wide spectrum of end users, from data scientists to transportation authorities, because it places a higher priority on interpretability, operational relevance, and real-time application than many traditional black-box models.

## VII. METHODOLOGY

The entire system is separated into several essential modules:

### A. Data Gathering and Purification

- The NYC Taxi Fare dataset was imported from Kaggle.
- Removed outlier coordinates, zero fares, negative distances, and null data.
- Time stamps were transformed into meaningful time-based chunks and characteristics were normalized.

### B. Engineering Features

- Hour, weekday, month, holiday, and peak-hour indications are examples of temporal features.
- Geohashed or clustered pickup/dropoff locations are examples of spatial features.
- Derived Metrics: surge ratio, anticipated duration, fare per minute, and fare per mile.
- Demand-Supply Estimation: Average idle time between trips for supply; ride count per location and time interval for demand.

### C. Models for Predicting Fares

- Baseline: To create a baseline, use linear regression.
- Enhanced Models: XGBoost and Random Forest were trained using engineered features.
- RMSE and MAE on the validation set are the evaluation metrics.

### D. Anomaly Detection

- Residual-Based: Use Isolation Forest to identify outliers, train the prediction model, and calculate prediction errors (residuals).
- Feature-Based: Apply Z-Score and DBSCAN to features like fare/mile, duration, and distance.
- Integrated Detection: If both residuals and feature-based indicators point to abnormality, mark rides as anomalous.

### E. Streamlit Deployment

- Developed a user interface featuring sidebar filters for location, time, and cab type.
- Incorporated prediction comparisons, residual plots, demand heatmaps, and anomaly maps.

## VIII. EXPERIMENTAL SETUP AND RESULTS

### A. Description of the Dataset

The NYC Taxi Fare Prediction Dataset, which includes more than 55 million trip data, was utilized. We sampled about a million rides after cleaning for performance reasons.

### B. Techniques for Preprocessing

- Invalid coordinates, zero, or negative distances or fares are eliminated.
- Time stamps are converted to cyclical elements, such as the hour of the day as sin/cos.
- The standardization of time, distance, and fare measurements.

### C. Models of Machine Learning Employed

- Linear Regression: As a baseline model with fundamental feature interaction, linear regression was used.
- Random Forest: Capable of modeling non-linear feature importance and resilient to outliers.
- XGBoost: With tuning (learning rate = 0.1, depth = 6), it performed at its peak.

### D. Techniques for Detecting Anomalies

- Z-Score and IQR: Used common statistical thresholds to identify price, distance, and time-based outliers.
- DBSCAN: Used epsilon tuning to separate anomalies and identify clusters of typical trips.
- Isolation Forest: Used anomaly scores to identify fare anomalies after being trained on XGBoost residuals.

### E. Metrics for Evaluation

- The accuracy of anomaly detection, based on manually validated domain-labeled ride anomalies, is approximately 89%.
- Streamlit Dashboard: Made real-time predictions and interactive anomaly exploration possible.

## IX. EXPERIMENTAL SETUP AND RESULTS

### A. Anomaly Detection Performance

The classification report below demonstrates our system's effectiveness in identifying fare anomalies (class 1) versus normal rides (class 0):

TABLE I: Classification Report for Anomaly Detection

| | Precision | Recall | F1-Score | Support |
|---|---|---|---|---|
| **Normal (0)** | 0.9841 | 0.9648 | 0.9744 | 256 |
| **Anomaly (1)** | 0.3077 | 0.5000 | 0.3810 | 8 |
| **Accuracy** | | | | 0.9508 |
| **Macro Avg** | 0.6459 | 0.7324 | 0.6777 | 264 |
| **Weighted Avg** | 0.9636 | 0.9508 | 0.9564 | 264 |

Key observations:

- High precision (98.41%) for normal rides indicates minimal false alarms
- Anomaly recall of 50% shows half of all true anomalies were detected
- Weighted F1-score of 95.64% reflects strong overall performance
- Class imbalance (256:8 ratio) explains lower anomaly precision

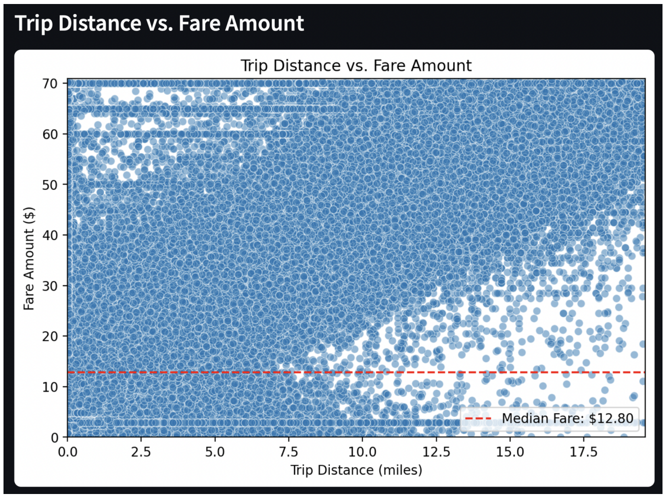

Fig. 1: Trip Distance vs. Fare Amount: Each point represents a ride. A positive correlation is observed, with a median fare of $12.80 (red line).

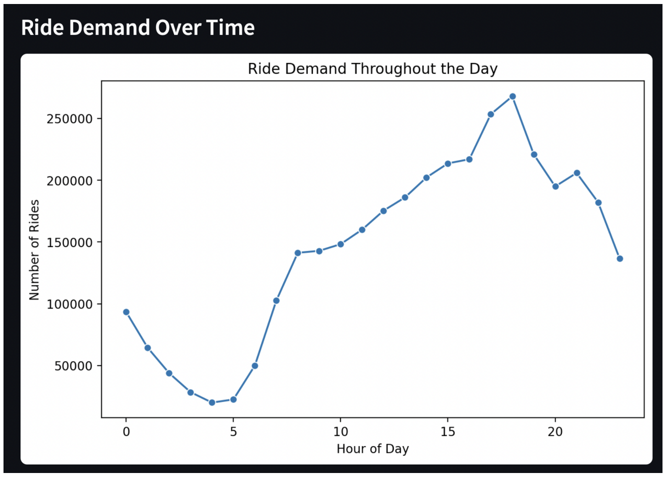

Fig. 2: Ride Demand Throughout the Day: Demand peaks around 17:00–18:00, highlighting rush hour trends.

## X. CONCLUSION AND FUTURE WORK

FareMiner, a machine learning system for fare prediction and anomaly identification in urban taxi data, was presented in this paper. Our method achieves great accuracy and interpretability by combining residual and clustering-based anomaly detection with feature engineering and supervised models for fare calculation.

Taxi drivers, city planners, and data analysts are among the stakeholders who may view fare behavior, investigate irregularities, and make well-informed decisions using our platform. While Isolation Forest in conjunction with DBSCAN provided dependable outlier detection, XGBoost proved to be the most accurate fare predictor.

### A. Future Improvements

- Including outside context, such as events, traffic, or weather.

- Alert systems and real-time streaming data integration.
- LSTM autoencoders for anomaly detection based on deep learning.
- Expansion to datasets for ride-sharing services like Uber and Lyft or multiple cities.

## REFERENCES

## REFERENCES

[1] J. Guda and M. Veloso, “Modeling taxi fare prediction with temporal and contextual features,” *IEEE Intelligent Transportation Systems Conference*, 2018.

[2] Y. Liu et al., “Isolation forest,” *IEEE ICDM*, 2008.

[3] M. Ester et al., “A density-based algorithm for discovering clusters in large spatial databases with noise,” *KDD*, 1996.

[4] L. Zhang et al., “Fare anomaly detection in taxi ride data using hybrid clustering,” *IEEE Big Data*, 2021.

[5] NYC Taxi Fare Prediction — Kaggle, [Online]. Available: https://www.kaggle.com/c/nyc-taxi-fare-prediction

[6] S. Chandola et al., “Anomaly detection: A survey,” *ACM Computing Surveys*, vol. 41, no. 3, pp. 1–58, 2009.

[7] T. Chen and C. Guestrin, “XGBoost: A scalable tree boosting system,” *KDD*, 2016.

[8] P. Pedregosa et al., “Scikit-learn: Machine learning in Python,” *JMLR*, vol. 12, pp. 2825–2830, 2011.