IST707 Applied Machine Learning

Green Taxi Trip Analysis

Harmish Doshi hadoshi@syr.edu

Kunwar Uday Singh Sikarwar ksinghsi@syr.edu

Anvitha Shashidhar aks100@syr.edu

Abstract

This project examines the Green Taxi Trip Data for New York City. The goal of this research is to use a regression algorithm to predict the fare amount for a taxi trip based on factors related to the trip and predict if the trip will have a surcharge using classification. The data was preprocessed to eliminate missing or incorrect values, outliers, and extraneous columns before feature selection using correlation matrices. Cross-validation was used to train and evaluate three regression models: Linear Regression, Lasso, and Random Forests. Identifying defective meter trips, negative distances, overfitting models, and processing big datasets were among the difficulties encountered during the analysis. According to the study, travel time, distance, peak hours, and holiday flags are all significant predictors of fare amount. Pickup location(latitude and longitude), peak hours, payment type, pickup days in a week and holiday flags are all the significant predictors for surcharge. The NYC Green taxi trip data provides significant insights for anybody interested in the patterns and trends of taxi rides in New York City, and this information can be utilized to inform policy decisions, reduce fees and idle time in order to help taxi operators and drivers run their businesses more profitably. Additionally, by offering more trustworthy and transparent fare estimates, this will enhance the overall customer experience.

Green Taxi Trip Analysis

The taxi industry is essential to each city's transportation sector, and New York City (NYC) is no different. Green Taxi Trip Data contains abundant information regarding taxi rides taken throughout the city. By developing a regression model based on pickup sites, trip distances, and fare amounts, this study attempts to estimate the fare amount for a taxi trip based on multiple attributes and a classification model to predict if a taxi ride would have a surcharge. The collection, which spans November 2022 to January 2023, has over 200,000 trip logs. This report describes the data preprocessing, feature selection, and model training processes, as well as the obstacles encountered during the analysis of the NYC Green Taxi Trip Data. The insights can be used to increase the efficiency and efficacy of the city's taxi sector.

Literature Review

The New York City Taxi and Limousine Commission (TLC) makes trip record data for taxi and for-hire car rides in the city available to the public. The data provides precise information about each trip's date, time, location, and fare. This data has been utilized in several studies to investigate patterns and trends in taxi and for-hire car use, as well as the influence of ride-sharing services on the taxi business (Tirachini & Cats, 2018). The trip record data dictionary contains an exhaustive list of the variables included in the dataset, such as pickup and drop-off locations, trip distance, and payment method. Researchers interested in examining traffic patterns and creating models for anticipating demand for taxi and for-hire vehicle services will find this information useful.

The Kaggle notebooks investigated prediction models for various areas of the New York City taxi system. The first notebook is concerned with predicting taxi trip durations, while the second is concerned with predicting taxi rates. These notebooks are significant to our research since they show how to utilize machine learning techniques to predict taxi-related factors, which might potentially be applied to our own research issue of predicting taxi demand in different parts of the city. These notebooks also provide insights into the types of data that can be used to predict taxi-related factors, such as weather data and taxi location data.

Hypothesis

For regression:

There is a significant correlation between various features such as pickup and drop-off locations, trip distance, and pickup and drop-off datetime, and the fare amount for a taxi trip.

For classification:

There is a statistically significant relationship between the variables of pickup location, peak hours, payment type, pickup days in a week, holiday flags and the presence of surcharge for green taxis. A classification model trained on these variables can accurately predict the surcharge for green taxi rides with a high degree of accuracy."

Methods

Individual taxi trips recorded in the dataset served as the study's unit of analysis. We used the whole dataset of taxi trip records to sample for the study. We added coordinate attributes (Longitude/ Latitude) using the geopy library to our dataset using pick up locations. Further, we created a surcharge flag based on congestion_surcharge attribute (0 for no surcharge and 1 for surcharge).

To balance our classes for the surcharge flag, we used RandomOverSampler to balance the minority class within our data to avoid bias. Further, with the help of LabelEncoder library, we converted attributes 'peak_hours', 'pickup_dayofweek', 'holiday_flag', 'payment_type' into categorical variables.

For regression models, we have split our data into 60% training data and 40% testing data. Based on the factors such as pickup and drop-off locations, trip distance, and pickup and drop-off datetime, we predicted the fare amount for each trip in our test data. We deployed machine learning algorithms such as Lasso, Linear Regression, Decision Tree, K-NN, Gradient Booster, XGBoost Regressor, Ridge and Random Forest to forecast the cost of a cab ride.

We trained the models on a subset of the data before testing them on a holdout set. To preprocess the data and extract additional features, we also used feature engineering approaches.

Further, to analyze the surcharge for the given trip ride in Green Taxi, we created classification model on 80% of training data, for Naïve Bayes, Logistic Regression, K-NN, Decision Trees, Random Forest and XGBoost Classifier and further evaluated their performances based on accuracy, precision, recall and f1- score metrics.

Based on the existing dataset of taxi trip records, our approaches intended to produce reliable predictions of taxi fares and find profitable pickup places for green taxis.

Results

Sample description: The sample consisted of over 908,613 record trips from the period November 2022 to January 2023. The Green Taxi Trip Data includes information about the date, time, pickup and drop-off locations, distance traveled, fare, payment method, and other attributes of each trip taken by a green taxi.

Regression models results:

Model	MSE	R-Squared	Adj R-squared	MAE
Linear Regression	18.981418	0.8791	0.8791	1.9822
KNN Regression	11.82	0.92	0.92	1.30
Lasso Regression	18.98	0.88	0.88	1.98
Ridge Regression	18.98	0.88	0.88	1.98
XGB Regressor	12.997	0.9172	0.9172	1.5299
Random Forest Regression	1.662	0.938	0.938	0.646

To answer the first research question, a linear regression and lasso regression were used to estimate the price amount for a taxi journey based on factors such as pickup and drop-off locations, trip distance, and pickup and drop-off datetime.
From the results, we see that Random Forest has the lowest MSE, R-squared, and MAE values, which indicates that it has performed better than other models and has the best fit for the given dataset and problem being addressed. The KNN and XGB models also show good performance, but not as good as Random Forest. Linear Regression, Lasso, and Ridge models have higher MSE, R-squared, and MAE values compared to Random Forest, KNN, and XGB, indicating that they have performed relatively worse in predicting the numerical value.

Classification models results:

Model	Precision	Recall	F-1 score	Support
Naïve Bayes	0.69	0.57	0.59	149558
Decision Trees	0.76	0.68	0.70	149558
Logistic Regression	0.70	0.55	0.57	149558
KNN	0.76	0.69	0.70	149558

Random Forests	0.76	0.69	0.70	149558

Classification Analysis:

To answer the second research question, Naïve Bayes, Decision Trees, Logistic regression, Random Forest, and K-NN algorithms were used to predict the surcharge fee.

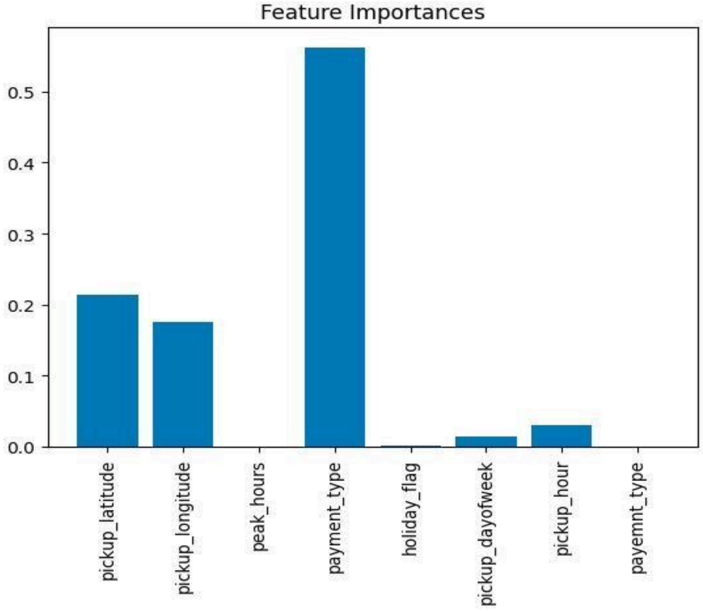
Based on the metrics mentioned above we can say that Decision Trees and Random Forest have the highest precision and recall, followed by KNN, indicating that they have made fewer false positive and false negative predictions. Naive Bayes and Logistic Regression, on the other hand, have lower precision and recall, indicating that they have made more false positive and false negative predictions.

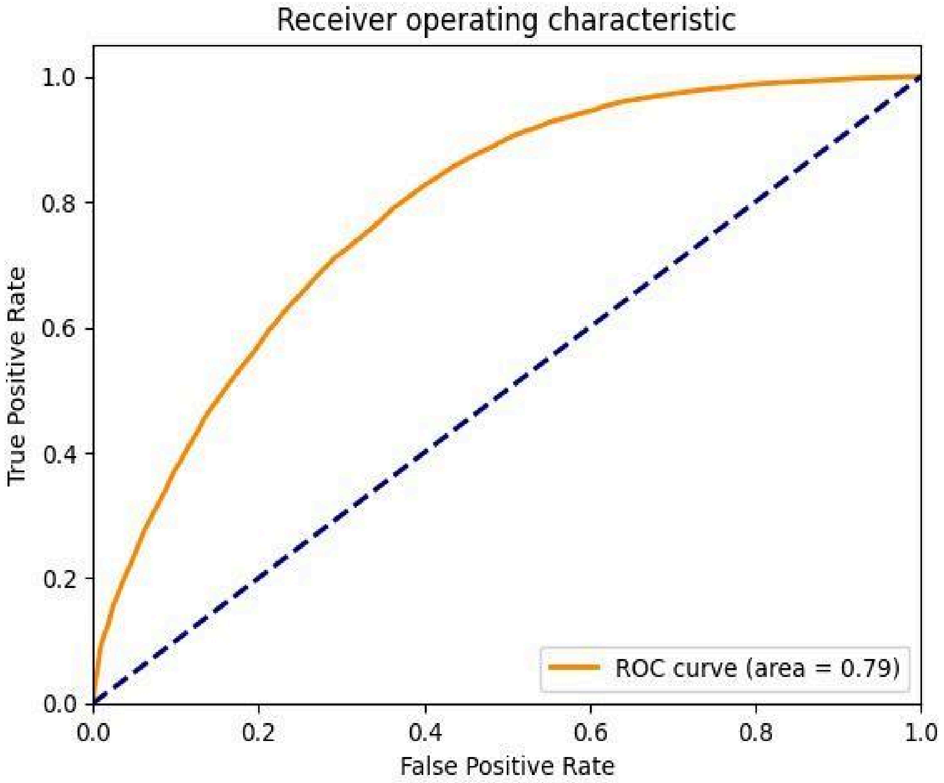
Discussions

The regression results show how well different models predicted a fare amount. Random Forest Regressor had the best performance, meaning it predicted the fare amount most accurately. Linear Regression, Lasso, and Ridge models didn't do as well as the other models in predicting the fare amount.

The classification results compared how different models performed in making predictions. Decision Tree, KNN, and Random Forest did the best job, with less inaccuracies. Naive Bayes and Logistic Regression gave weaker results in terms of precision, recall, accuracy and f1- score.

We use feature importances in XGBoost classifier to understand which attributes have more importance for the surcharge flag. We see that payment type is highly correlated to pickup_latitude and pickup_longitude.





The ROC curve is used to evaluate based on their true positive rate and false positive rate. The area under the curve is used for evaluating classifier performance and our result is 0.79 which indicates good classification.

When we look back at the literature, we see that our findings are consistent with previous studies that used machine learning algorithms to predict taxi fares. Our research adds to the body of knowledge by comparing the performance of various regression models in predicting taxi fares. We have also predicted the surcharge fee based on pickup locations using classification that could not be compared with the previous research.

Despite the value of our research, there are some limitations to consider. For starters, because our dataset only included taxi journeys in New York City, our findings may not be applicable to other cities or areas. Second, our model does not take into consideration things like traffic congestion or weather conditions, which could affect taxi fares. Future study could investigate including these parameters in the analysis to increase prediction accuracy.

Overall, our research provides useful insights into estimating taxi fares using machine learning algorithms, which can help taxi businesses and politicians improve the efficiency and effectiveness of the taxi industry in the city.

References

- NYC TLC. (n.d.). About TLC Trip Record Data. Retrieved from https://www.nyc.gov/site/tlc/about/tlc-trip-record-data.page
- NYC TLC. (n.d.). Trip Record Data (Green Taxis). Retrieved from https://www.nyc.gov/assets/tlc/downloads/pdf/data_dictionary_trip_records_green.pdf
- Nitin194. (n.d.). NYC Taxi Trip Duration Prediction. Kaggle. Retrieved from https://www.kaggle.com/code/nitin194/nyc-taxi-trip-duration-prediction#Model
- Abdurraffay00. (n.d.). NYC Taxi Fare Predictor. Kaggle. Retrieved from https://www.kaggle.com/code/abdurraffay00/nyc-taxi-fare-predictor
- Tariq, U., Naeem, M. A., Gulzar, M. A., & Khan, A. (2020). Machine Learning with Oversampling and Undersampling Techniques: Overview, Study, and Experimental Results. International Journal of Advanced Computer Science and Applications, 11(3), 442-451. https://www.researchgate.net/publication/340978368_Machine_Learning_with_Oversampling_and_Undersampling_Techniques_Overview_Study_and_Experimental_Results