ALY6110：Data Management and Big Data

Module 6 Assignment：Final Project Report

Group members：Chen-Yu Chu，Sean Mclean，Yin Tang，Jingyi Wang，Runcheng Yang

Instructor：Ajit Appari

June 25，2024

Introduction

This report explores the use of predictive analytics in the taxi service sector with a particular emphasis on fare prediction, which is a vital component of the business. Predicting taxi fares accurately is not only a technical issue but also a necessity that affects customer satisfaction, operational efficiency, and service pricing.

To address this, we have employed a variety of regression models to analyze a comprehensive dataset detailing numerous taxi rides. The dataset includes variables such as trip distances, toll amounts, and other related charges—all vital for developing robust models that can predict fare amounts with high precision. Our methodology is thorough, integrating data preparation, exploratory data analysis (EDA), model selection, training, and evaluation. Each model’s performance is meticulously analyzed using several metrics to ensure reliability and applicability.

Through this analysis, the report aims to not only present a comparative evaluation of the predictive power of each model but also to offer insights into the nuances of taxi fare dynamics. By doing so, we seek to contribute to the broader conversation on how big data can be harnessed to refine and revolutionize transportation logistics and customer service strategies.

Real-World Problem

The real-world problem addressed in the document is predicting taxi fares using various regression models. This involves analyzing data related to taxi rides, such as trip distances, toll amounts, and other related charges, to develop predictive models that can accurately forecast the total fare amount for a ride. We summarized three key points of the real-world problem:

1. Predictive Accuracy: The need to develop models that can predict taxi fares accurately to help taxi companies and ride-sharing services optimize pricing strategies and improve customer satisfaction.

2. Model Selection and Complexity: Evaluating different regression models (Linear Regression, Random Forest, Gradient Boosting) to determine which model best fits the data's characteristics, balancing between simplicity and the ability to capture complex patterns.

3. Data Quality and Availability: Ensuring that the data used in the models is complete and accurate, which impacts the predictive power of the models.

This problem is significant because it directly affects operational efficiency and profitability in the transportation industry, influencing fare pricing, customer service, and competitive strategy.

Methodology

The methodology for this analysis was structured to provide a comprehensive understanding of the factors influencing taxi fares. The process encompassed data preparation, exploratory data analysis (EDA), model selection, training, and evaluation. Each step was carefully designed to ensure the integrity and accuracy of the results, utilizing a combination of statistical techniques and machine learning models.

- Data Preparation

The initial stage involved cleaning and preprocessing the taxi dataset. This included converting pickup and dropoff timestamps into pandas datetime objects, removing rows with invalid datetime entries or duplicate records, and filtering out trips with non-positive values for distances or total amounts. Additionally, outlier treatment was performed by calculating the interquartile range (IQR) and removing values significantly outside this range.

- Exploratory Data Analysis (EDA)

The EDA phase aimed to uncover underlying patterns and insights within the data through visual and statistical methods. Histograms with kernel density estimates were used to examine the distribution of key variables like total amount and trip distance. Scatter plots explored the relationship between trip distance and total amount, and a correlation matrix provided insights into the linear relationships between numerical features, guiding the feature selection process for modeling.

- Model Selection

Three regression models were chosen based on their distinct characteristics and suitability for the dataset:

Linear Regression: Selected for its simplicity and interpretability, assuming a linear relationship between features and the fare amount.

Random Forest: Chosen for its ability to handle complex, non-linear relationships without explicit specification, utilizing an ensemble of decision trees.

Gradient Boosting: Employed for its optimization capabilities in reducing both bias and variance, building trees sequentially to correct previous prediction errors.

- Model Training and Evaluation

Each model was trained using the prepared dataset and evaluated using metrics such as Mean Squared Error (MSE), Mean Absolute Error (MAE), and the coefficient of determination (R^2). These metrics helped assess the accuracy, consistency, and explanatory power of the models. The models' performance was further analyzed through feature importance plots, highlighting which variables most significantly influenced the fare predictions.

Through this rigorous methodology, the study aimed to not only predict taxi fares accurately but also to provide insights into the efficiency of different modeling approaches in real-world applications.

Exploratory Data Analysis (EDA)

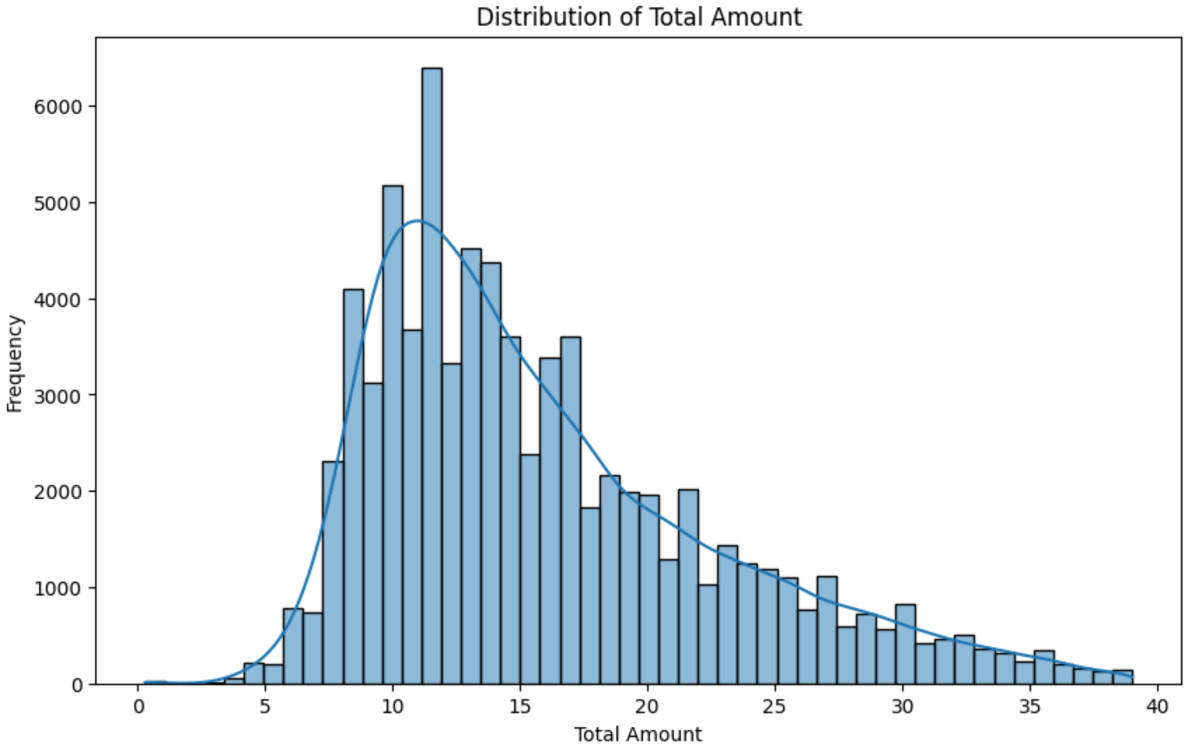
In the EDA stage of the taxi dataset, a thorough investigation was conducted to prepare the dataset for deeper analysis and modeling. The process began by converting the taxi ride pickup and dropoff timestamps to pandas datetime objects, ensuring all subsequent time-based analyses would be accurate. This step was crucial for performing time-based calculations and filtering effectively.

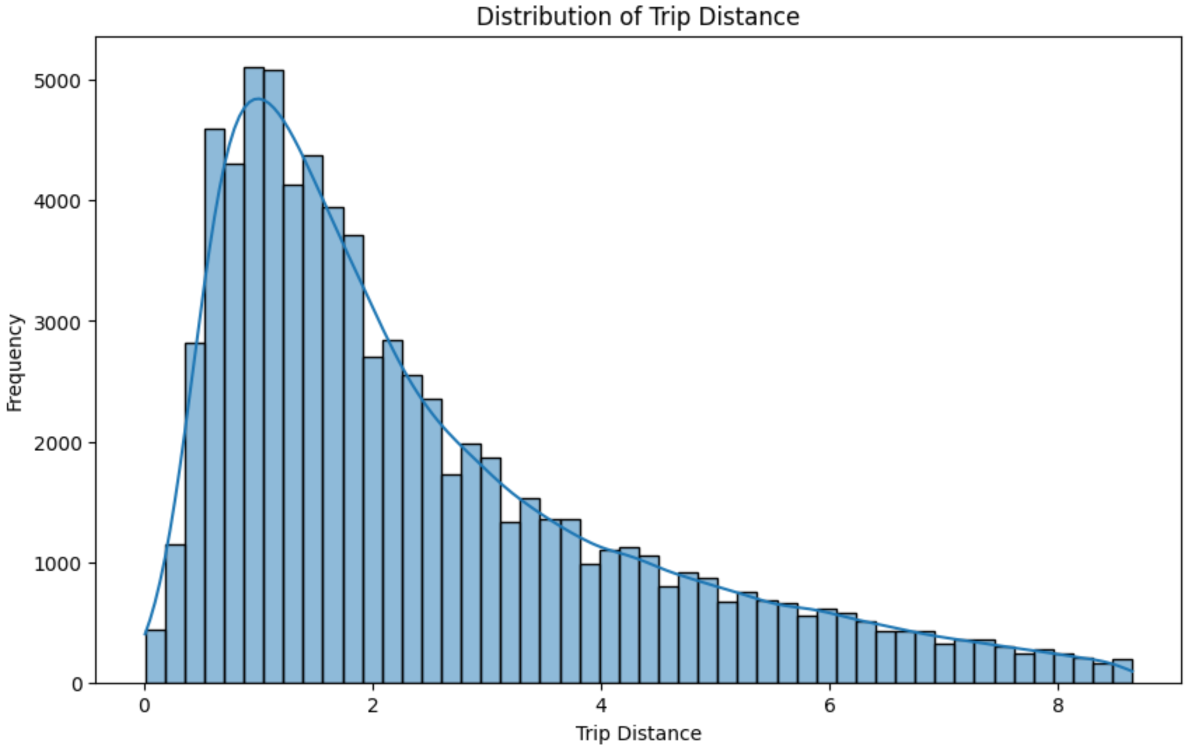
Following the datetime conversion, the analysis progressed with meticulous data cleaning. The dataset was cleansed of rows containing invalid datetime entries (NaT), which might indicate incomplete records. Duplicate entries were removed to prevent skewed results that could distort statistical measures. Additionally, trips with non-positive values for distances or total amounts were filtered out, eliminating likely data entry errors or instances of cancelled trips.

Outlier treatment was performed by calculating the IQR for both trip distances and total amounts. Outliers—defined as values below or above 1.5 times the IQR from the first and third quartiles

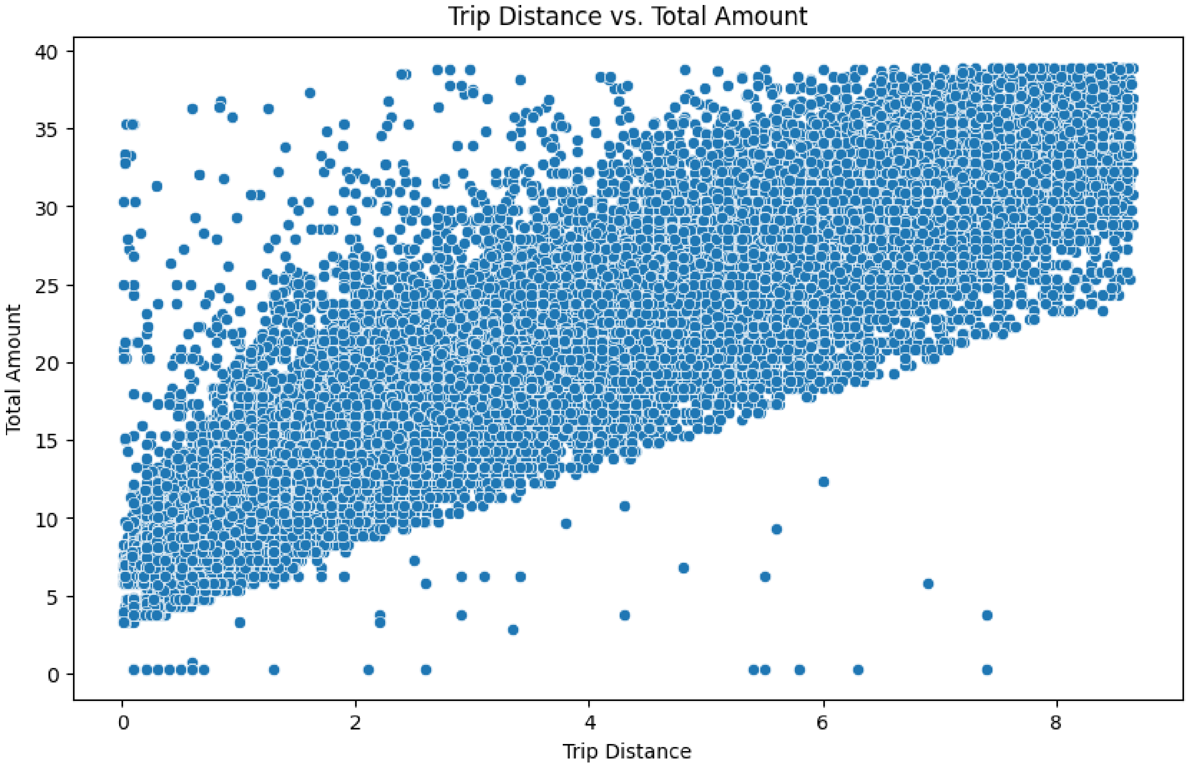
, respectively—were removed to ensure a robust analysis, as they can significantly affect the mean, variance, and other statistical tests.

Statistical summaries provided by the describe() function offered insights into the central tendency, dispersion, and shape of the dataset’s distribution, excluding NaN values. This was particularly useful for quickly gauging the scale and spread of data, such as understanding the typical trip distance and fare amount.

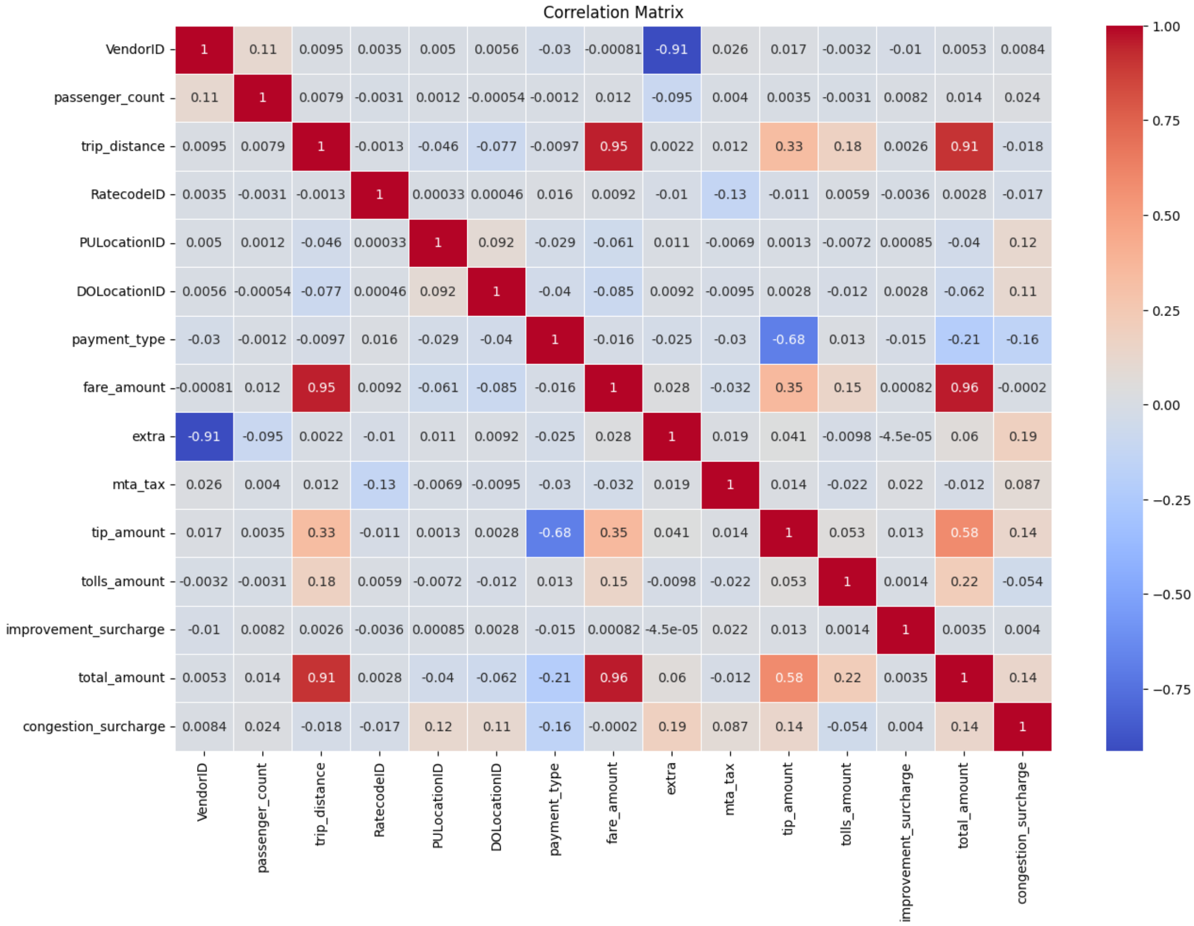




Visual explorations were integral to the EDA process. Histograms with a kernel density estimate (KDE) were generated for total_amount and trip_distance to examine their distributions, revealing any skewness or outliers and the general distribution shape of these key variables.



A scatter plot between trip distance and total amount helped explore their relationship, highlighting the expected positive correlation where longer trips generally entail higher costs. This plot also helped in spotting any unusual patterns, such as unexpectedly high fares for short trips, which could indicate traffic delays or special pricing.

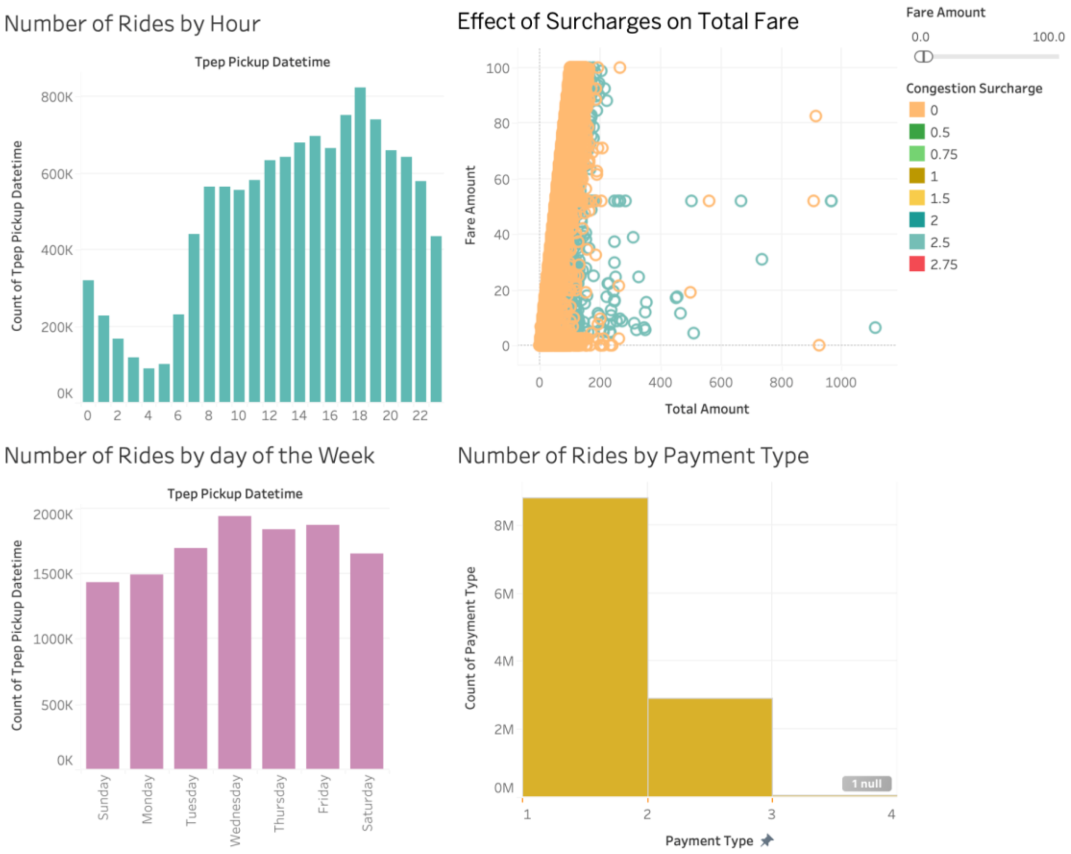


	VendorID	passenger_count	trip_distance	RatecodeID	PULocationID	DOLocationID	payment_type	fare_amount	extra	mta_tax	tip_amount	tolls_amount	improvement_surcharge	total_amount	congestion_surcharge
VendorID	1	0.11	0.0095	0.0035	0.005	0.0056	-0.03	-0.00081	-0.91	0.026	0.017	-0.0032	-0.01	0.0053	0.0084
passenger_count	0.11	1	0.0079	-0.0031	0.0012	-0.00054	-0.0012	0.012	-0.095	0.004	0.0035	-0.0031	0.0082	0.014	0.024
trip_distance	0.0095	0.0079	1	-0.0013	-0.046	-0.077	-0.0097	0.95	0.0022	0.012	0.33	0.18	0.0026	0.91	-0.018
RatecodeID	0.0035	-0.0031	-0.0013	1	0.00033	0.00046	0.016	0.0092	-0.01	-0.13	-0.011	0.0059	-0.0036	0.0028	-0.017
PULocationID	0.005	0.0012	-0.046	0.00033	1	0.092	-0.029	-0.061	0.011	-0.0069	0.0013	-0.0072	0.00085	-0.04	0.12
DOLocationID	0.0056	-0.00054	-0.077	0.00046	0.092	1	-0.04	-0.085	0.0092	-0.0095	0.0028	-0.012	0.0028	-0.062	0.11
payment_type	-0.03	-0.0012	-0.0097	0.016	-0.029	-0.04	1	-0.016	-0.025	-0.03	-0.68	0.013	-0.015	-0.21	-0.16
fare_amount	-0.00081	0.012	0.95	0.0092	-0.061	-0.085	-0.016	1	0.028	-0.032	0.35	0.15	0.00082	0.96	-0.0002
extra	-0.91	-0.095	0.0022	-0.01	0.011	0.0092	-0.025	0.028	1	0.019	0.041	-0.0098	-4.5e-05	0.06	0.19
mta_tax	0.026	0.004	0.012	-0.13	-0.0069	-0.0095	-0.03	-0.032	0.019	1	0.014	-0.022	0.022	-0.012	0.087
tip_amount	0.017	0.0035	0.33	-0.011	0.0013	0.0028	-0.68	0.35	0.041	0.014	1	0.053	0.013	0.58	0.14
tolls_amount	-0.0032	-0.0031	0.18	0.0059	-0.0072	-0.012	0.013	0.15	-0.0098	-0.022	0.053	1	0.0014	0.22	-0.054
improvement_surcharge	-0.01	0.0082	0.0026	-0.0036	0.00085	0.0028	-0.015	0.00082	-4.5e-05	0.022	0.013	0.0014	1	0.0035	0.004
total_amount	0.0053	0.014	0.91	0.0028	-0.04	-0.062	-0.21	0.96	0.06	-0.012	0.58	0.22	0.0035	1	0.14
congestion_surcharge	0.0084	0.024	-0.018	-0.017	0.12	0.11	-0.16	-0.0002	0.19	0.087	0.14	-0.054	0.004	0.14	1

Finally, a correlation matrix was visualized through a heatmap to provide a clear view of the linear relationships between numerical features. This step was critical in identifying which variables most strongly influenced the total amount, guiding the feature selection process for subsequent predictive modeling.

Through this comprehensive EDA, we gained valuable insights into the dataset's structure and quality, identifying potential issues such as outliers or anomalies that could influence model training. This foundational understanding of the data's basic trends, patterns, and relationships informed hypothesis testing and mode ling assumptions in the subsequent stages of analysis, ensuring that predictive models would be based on well-understood and cleanly prepared data.

Visulization


Number of Rides by Hour

Effect of Surcharges on Total Fare

Number of Rides by day of the Week

Number of Rides by Payment Type

The dashboard presents four distinct visualizations related to taxi ride data, providing insights into the dynamics of taxi usage based on various factors such as time, surcharge effects, and payment types:

1. Number of Rides by Hour: This bar chart shows a clear diurnal pattern in taxi usage, with the lowest demand in the early morning hours and a peak in the evening around 18:00. This likely corresponds to the end of typical working hours, suggesting tha

t a significant portion of taxi usage may be for commuting home.

2. Effect of Surcharges on Total Fare: The scatter plot illustrates the relationship between total fare amounts and the congestion surcharge applied. The color gradient and the size of the circles, which represent fare amounts, indicate that higher surcharges generally correlate with higher total fares, particularly noticeable in the denser cluster of points around lower surcharge values. This plot suggests that while surcharges add to the fare, the overall fare might also be influenced by other factors like distance or time of day.

3. Number of Rides by Day of the Week: The histogram indicates relatively consistent taxi usage throughout the weekdays with a slight increase on Fridays. The weekend shows a notable decrease in usage, which might reflect a reduced demand for commuting services outside of the traditional workweek.

4. Number of Rides by Payment Type: This bar chart shows a stark preference in payment methods, with an overwhelming majority of rides paid using method ’1’ (presumably cash or a common credit card). There is a significant drop in usage for payment types labeled ’2’, ’3’, and a negligible count labeled as ’null’, indicating either an error in data entry or a rarely used method.

Integrating these insights, the visualizations collectively depict a taxi service heavily utilized during typical commute hours, especially in the late afternoon and early evening. The impact of surcharges on fares points to potential policy or operational adjustments to manage traffic or environmental impacts. The consistency in weekday usage versus the drop on weekends could help in planning fleet availability and operational hours. The dominance of a single payment method suggests customer prefer ence or technological limitations in payment infrastructure. Together, these insights provide a comprehensive understanding of how taxis are used across different times, days, and payment preferences, which could guide strategic decisions in service provision and fare structuring.

Predictive Models

- ### Models we used and why

In the analysis of the taxi dataset, three distinct regression models were employed for predictive modeling: Linear Regression, Random Forest, and Gradient Boosting, each chosen for its unique capabilities and suitability for handling different aspects of the dataset.

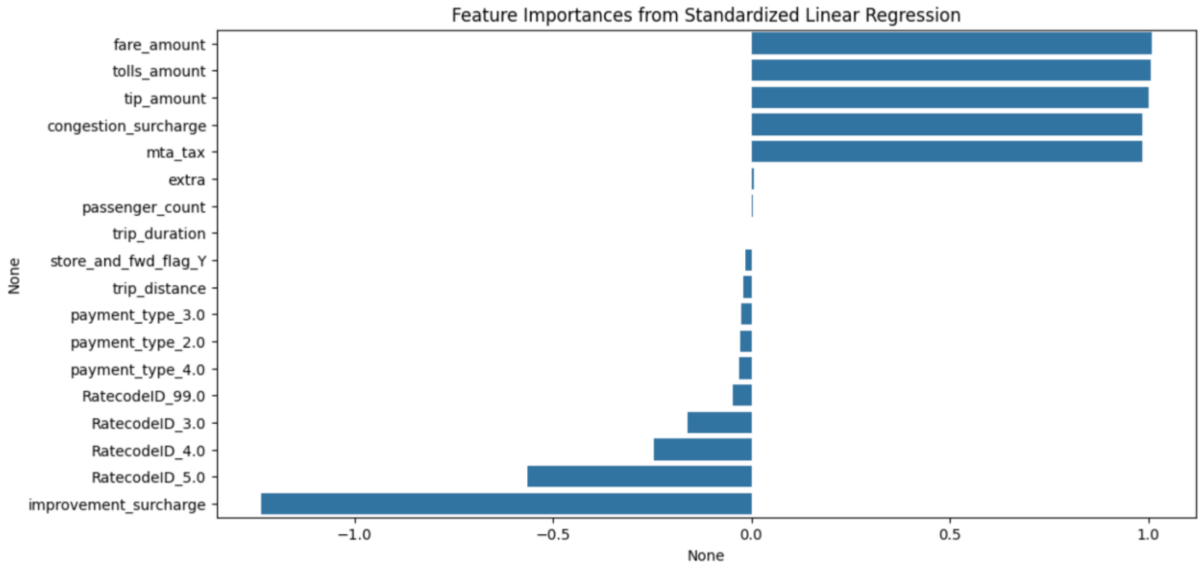
Linear regression was utilized primarily for its simplicity and interpretability, assuming a straightforward linear relationship between features such as trip distance, duration, and other related charges, and the dependent variable, the total fare amount. The model was trained and subsequently evaluated on a test set using metrics such as MSE, MAE, and R^2. These metrics provided insights into the model's accuracy and the variance in fares it could explain. The visualization of the model's coefficients offered a clear depiction of which features most significantly impacted the fare, thus aiding in understanding the direct relationships within the data.
The random forest was chosen due to its resistance to overfitting and its capacity to represent intricate, nonlinear relationships without requiring the explicit articulation of these interactions. The performance of this model was evaluated using the same set of metrics during training and testing, taking advantage of the ensemble technique of averaging numerous decision trees.

Gradient boosting was incorporated into the study for its advanced optimization capabilities on both bias and variance, building trees sequentially to correct previous errors and progressively improve predictions. Like the other models, it was evaluated on how well it could predict unseen data and how much of the fare variance it could explain. The feature importance from this model were particularly valuable, revealing nuanced insights into which factors most strongly predict fares, often highlighting different aspects than the other models due to its method of emphasizing correction of prior errors.

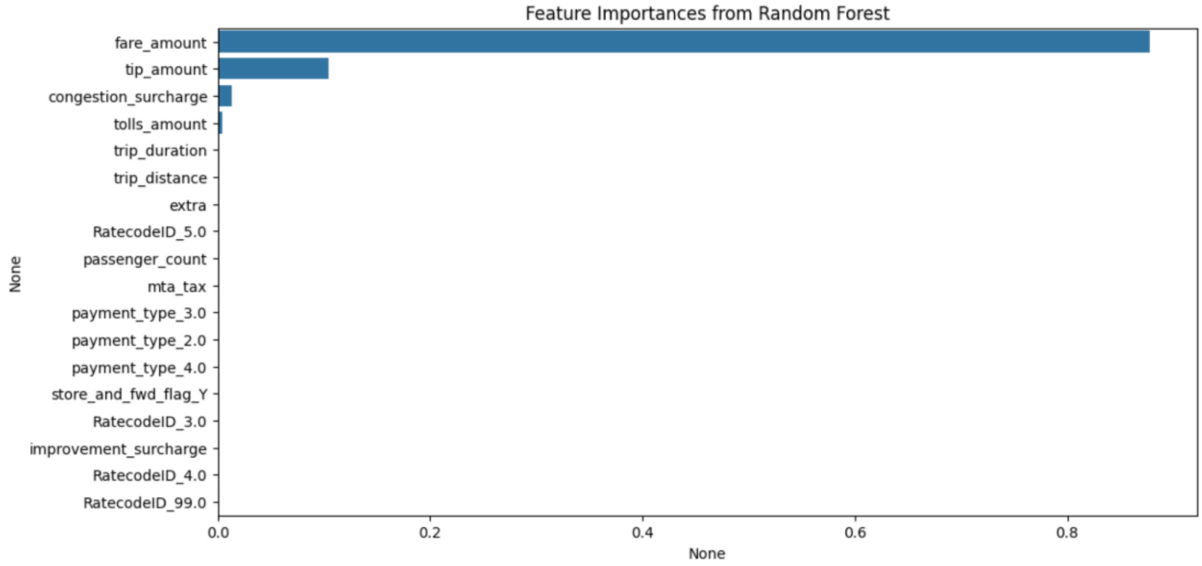
The integrated use of these models not only facilitated a comprehensive understanding of the key factors influencing taxi fares but also highlighted the strengths and limitations of each modeling approach. Linear Regression provided a baseline understanding, Random Forest added depth with its handling of complex interactions, and Gradient Boosting offered precision through progressive improvements. This multi-model approach enriched the analysis, allowing for a well-rounded understanding and robust predictive performance, crucial for operational optimizations and strategic decision-making in taxi service operations.

- ## Feature Importance

The feature importance plots for the Linear Regression, Random Forest, and Gradient Boosting models provide a detailed visualization of the variables most impactful in predicting taxi fare amounts in the dataset. Each model presents a slightly different perspective on what drives fare predictions, reflecting their individual analytical strengths and modeling approaches.

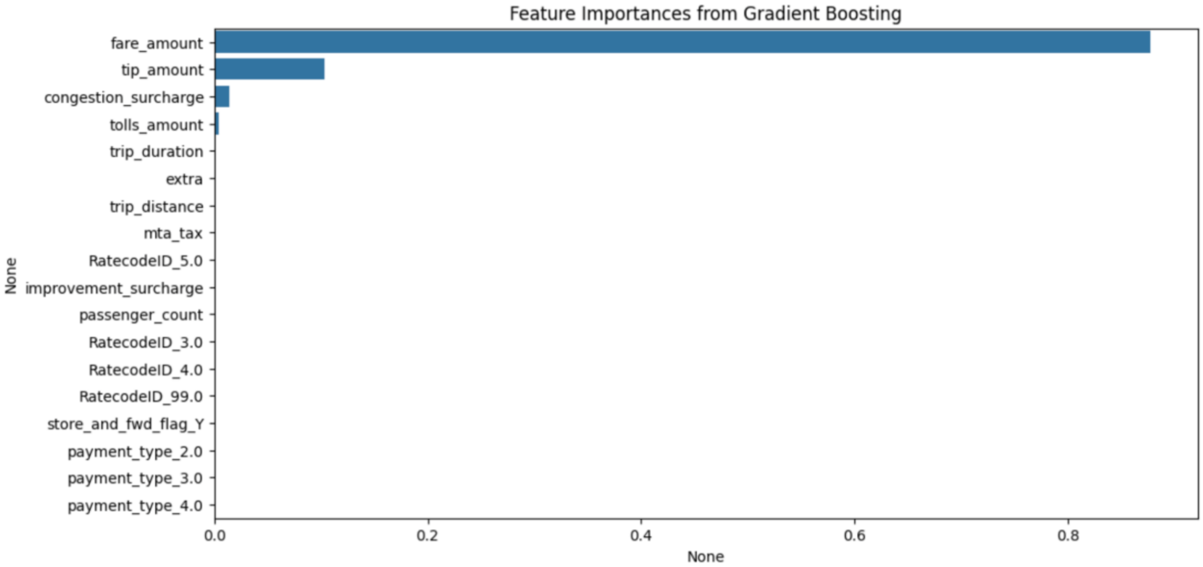


The plot shows that in the linear regression, fare_amount, tolls_amount, and tip_amount are the most influential predictors. This indicates a strong linear relationship between these variables and the total amount charged, which is expected as these components are direct determinants of the final fare. The plot also reveals some negative influences from other features, although these effects are relatively minor compared to the positive impacts.



In the random forest model, fare_amount still appears as the most significant predictor, followed closely by tip_amount. However, congestion_surcharge and tolls_amount also show notable importance, suggesting that the random forest model is capturing more complex interactions between the features than the linear model.

The ability of random forest to handle non-linear dependencies allows it to leverage a broader spectrum of features effectively.

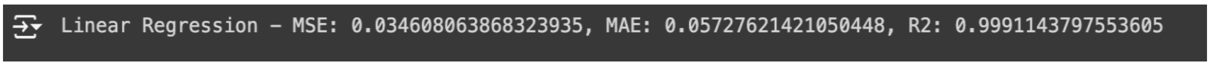


Similar to Random Forest, Gradient Boosting emphasizes the importance of fare_amount and tip_amount, with congestion_surcharge also appearing as a significant predictor. This model fine-tunes the influence of each feature through successive corrections of errors from previous trees, which likely enhances its focus on features that frequently correlate with higher errors in fare predictions.

Across all three models, fare_amount, tip_amount, and tolls_amount consistently emerge as key predictors of total fare, underscoring their fundamental role in fare composition. The variations in feature importance rankings between models underscore different modeling capabilities: Linear regression provides a baseline understanding of direct linear relationships; random forest captures a wider array of interactions among more variables; and gradient boosting refines this understanding by iteratively improving on prediction errors.

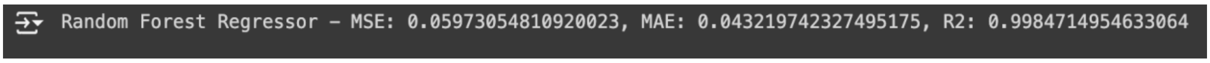
- **Performance of the models**

Our analysis involves three different regression models—Linear Regression, Random Forest, and Gradient Boosting—each applied to predict taxi fares using the same dataset. The performance of these models is evaluated based on three metrics: MSE, MAE, and R^2.



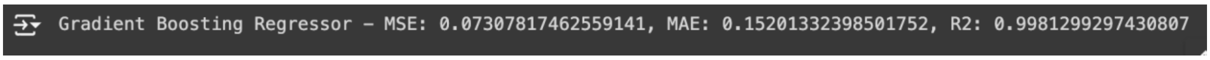
```
Linear Regression - MSE: 0.034608063868323935, MAE: 0.05727621421050448, R2: 0.9991143797553605
```

The simplest of the three models, linear regression demonstrates remarkably high performance with an R^2 value close to 1 (0.9991), indicating that the model explains nearly all the variance in the target variable. The MSE and MAE are quite low, at 0.034 and 0.057, respectively, which suggests that the model’s predictions are very close to the actual values.



```
Random Forest Regressor - MSE: 0.05973054810920023, MAE: 0.043219742327495175, R2: 0.9984714954633064
```

Random forest also shows excellent performance with an R^2 of 0.9987. However, its MSE (0.0597) and MAE (0.0432) are slightly higher than those of the linear regression model, indicating slightly less accuracy in prediction. This could be due to the model capturing more complex patterns, which might not necessarily contribute to improving the prediction in this particular context.



```
Gradient Boosting Regressor - MSE: 0.07307817462559141, MAE: 0.15201332398501752, R2: 0.9981299297430807
```

Gradient boosting also shows a very high R^2 of 0.9981, slightly lower than the other two models but still indicating very high explanatory power. The MSE and MAE are the highest among the three models, at 0.073 and 0.152, respectively, which could suggest that while gradient boosting is robust for varied data, it might be overfitting or too complex for this specific dataset where simpler models perform better.

In conclusion, all the three models perform exceptionally well in predicting taxi fares, as indicated by the R^2 values close to 1. However, the simpler Linear Regression model appears to provide the best balance between simplicity and accuracy for this dataset, achieving the lowest MSE and MAE. This suggests that the relationships in the dataset between the features and the fare ar e predominantly linear, or that the additional complexity introduced by the Random Forest and Gradient Boosting models does not significantly enhance the predictive accuracy for this particular data.

Results/Conclusions

Our analysis of the taxi dataset using three different regression models—Linear Regression, Random Forest, and Gradient Boosting—revealed high predictive accuracy in determining taxi fares. The linear regression model proved to be the most effective, demonstrating almost perfect explanatory power with an R^2 close to 1, and achieving the lowest MSE and MAE. This suggests that the relationships within the dataset are predominantly linear, making more complex models unnecessary for this context. The Random Forest and Gradient Boosting models, while robust, did not significantly improve the predictive accuracy beyond what was achieved by the Linear Regression model.

Recommendations

Based on the findings, we recommend:

1. Continued use of linear regression for similar datasets: For datasets with similar characteristics and linear relationships, linear regression should be the primary modeling approach due to its simplicity, interpretability, and high accuracy.
2. Cautious Application of Complex Models: While models like random forest and gradient boosting are valuable for their complexity and ability to capture nonlinear relationships, their application should be evaluated against the nature of the data to avoid unnecessary model complexity.

3. Enhanced Data Collection: Improving the accuracy and completeness of data collection regarding trip distances, toll amounts, and other fare-related factors could further refine the predictive accuracy of all models used.

Future Research

Future studies could focus on several areas to extend our understanding and application of predictive modeling in taxi fare calculations:

1. Integration of Additional Variables: Including variables such as weather conditions, day of the week, and real-time traffic updates could enhance the models' accuracy and applicability to dynamic pricing.
2. Cross-Model Validation with Larger Datasets: Applying these models to larger, more varied datasets could validate their effectiveness across different urban settings and fare structures.
3. Development of Real-Time Predictive Models: Research into real-time predictive modeling for dynamic fare pricing could provide actionable insights for taxi companies to optimize fares instantly based on current demand and supply conditions.

These sections aim to encapsulate the essence of your findings and suggest directions for continued research and application improvements.

References

Shaw, A. (2020). Taxi Dataset. Kaggle. Retrieved from https://www.kaggle.com/datasets/anandshaw2001/taxi-dataset