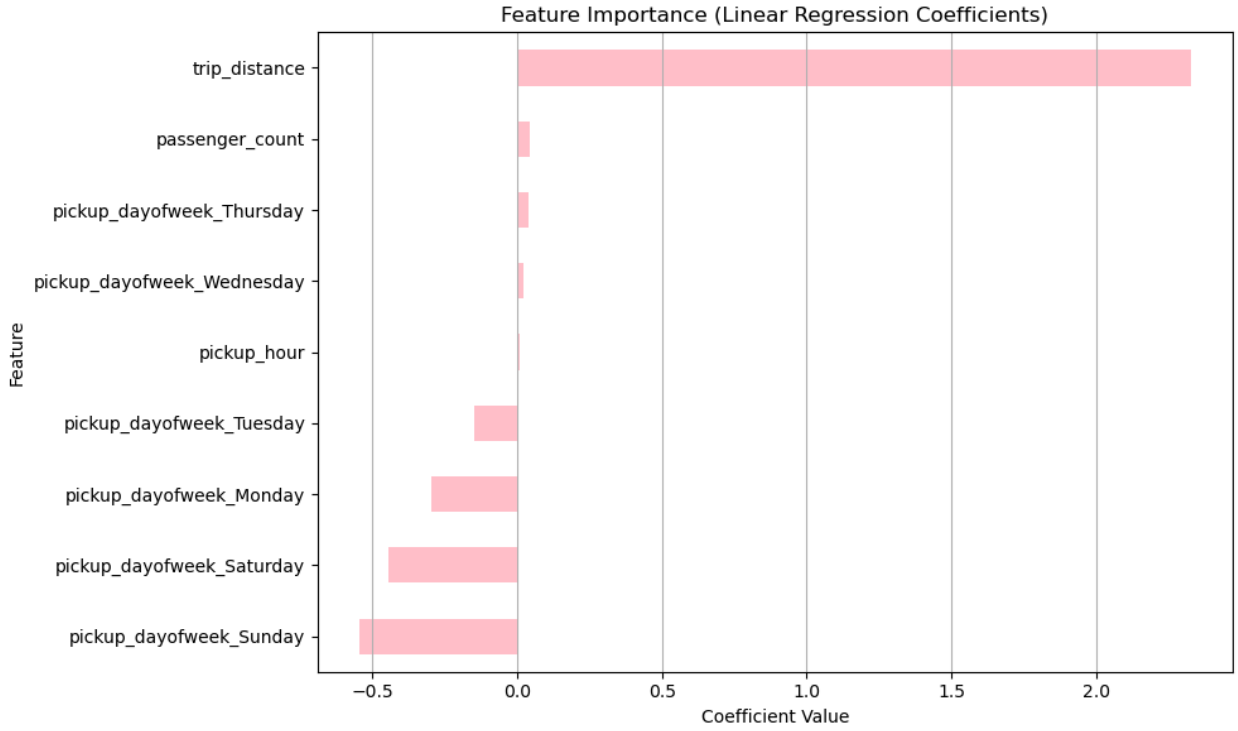
Data Visualizations

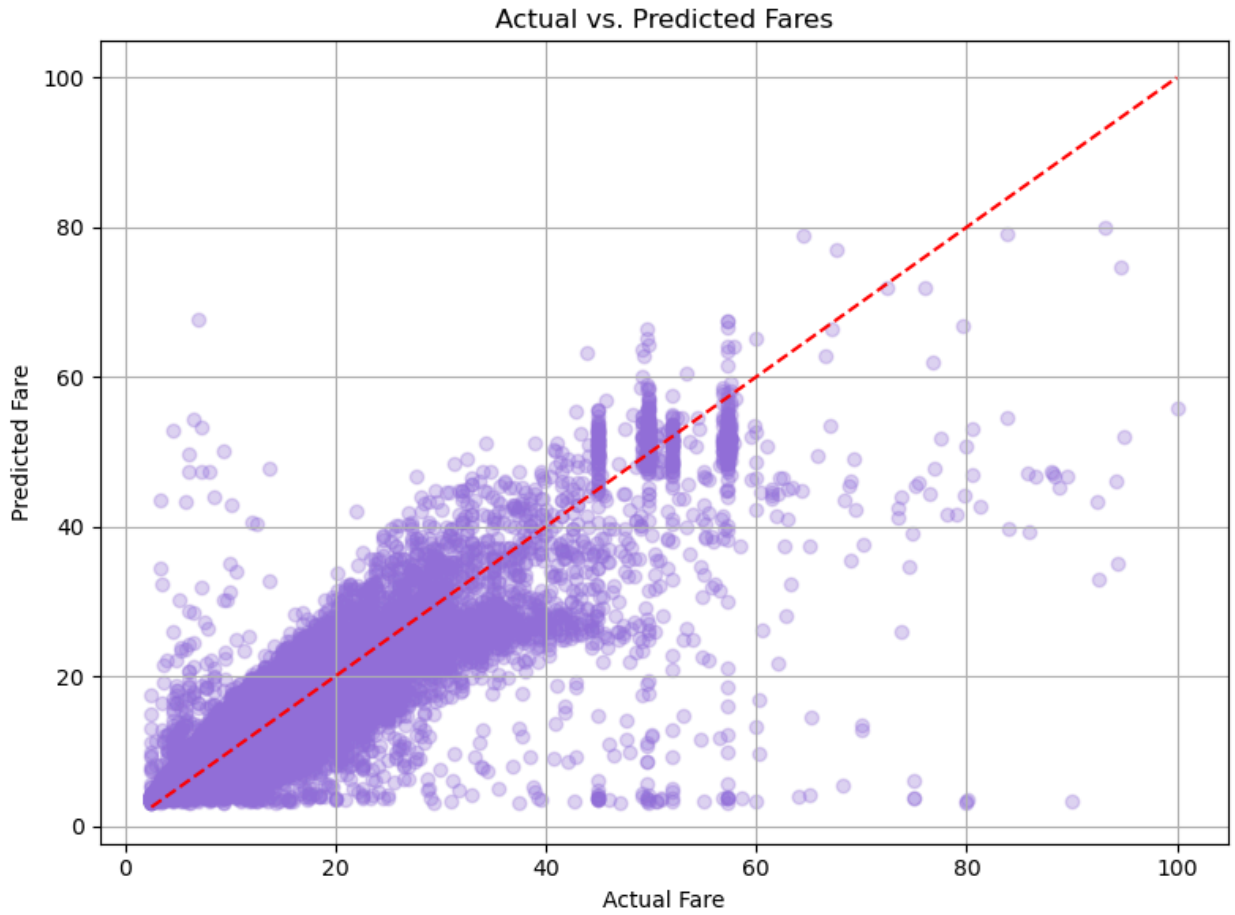
Marisa Thomas, Keaton Magnuson, Emily Szolnoki, Klodje Toure

This project explores factors that influence Uber fare prices. We used a Linear Regression model to predict fare amounts based on trip distance, time of day, passenger count, and day of the week. The dataset, consisting of over 193,000 cleaned Uber trip records in New York City, was preprocessed and used to develop predictive visual insights.

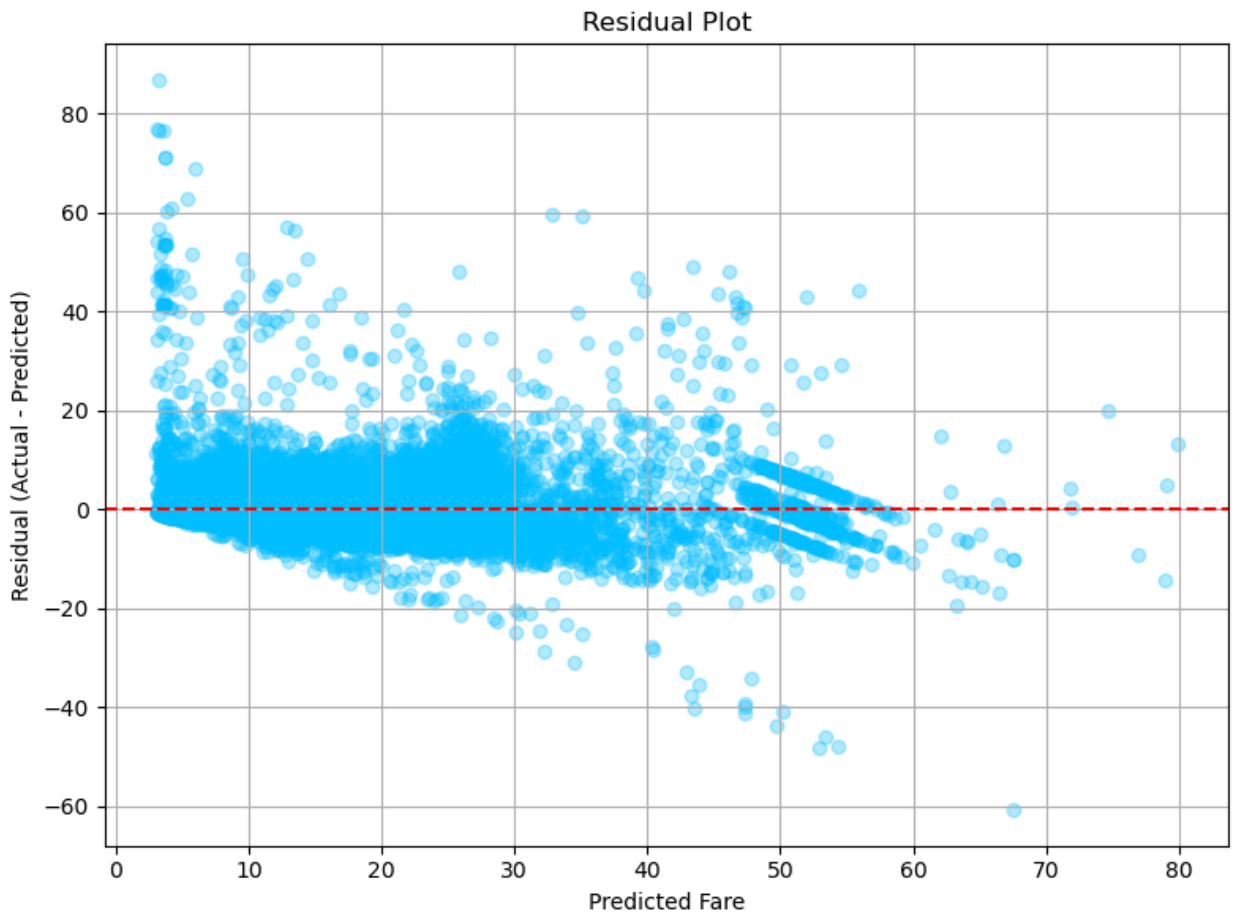


This bar chart visualizes the coefficients learned by the Linear Regression model. Each bar shows the impact of that feature on the predicted Uber fare amount. In this graph we can see that Trip Distance has the largest positive effect. The longer bar means it's the most important feature. For every 1 mile increase in distance, the fare goes up by approximately $2.33. Day of the Week affects pricing relative to Friday. Sunday has the most negative effect as fares are around 55 cents cheaper than on Fridays. Saturday, Monday and Tuesday also reduce fares, but to a lesser extent. Thursday and Wednesday slightly increase fares. Pickup Hour and Passenger Count have minimal influence. This suggests that time of day and number of

passengers don’t significantly impact fare in a linear way. This helps confirm that trip distance is the dominant predictor of fare. It also shows subtle pricing differences by day, which could inform dynamic pricing strategies.



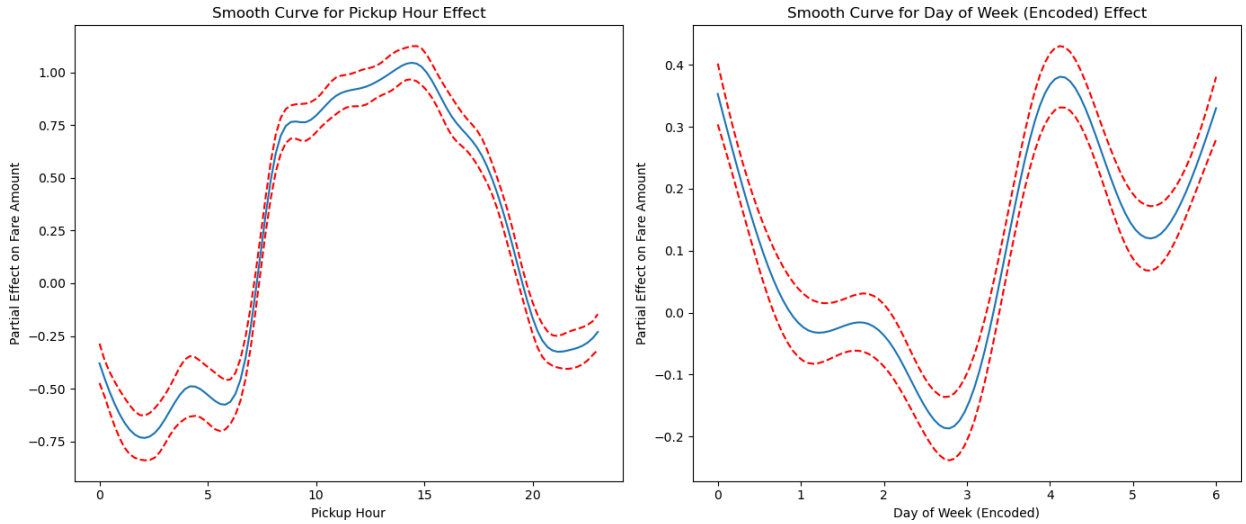
This scatter plot compares the actual fare amounts to the predicted fares from the Linear Regression model. The red dashed line represents a perfect prediction, where the predicted fare equals the actual fare. Most points cluster closer to the line, especially for fares under $40, indicating that the model performs well on typical trip prices. Predictions for higher fare trips tend to fall below the line, meaning the model often underestimates expensive rides. Some horizontal groupings suggest fixed-rate trips, like airport rides, which the linear model may not fully capture. Overall, the model provides a good approximation for most fares, but struggles with outliers and very high priced trips.



This residual plot displays the difference between actual and predicted fare amounts for each trip. The red dashed line at 0 represents perfect predictions. The closer the points are to this line, the more accurate the model.

Most residuals are clustered around zero, especially for typical fare ranges, which shows the model is generally well calibrated. Positive residuals, above the line, indicate underpredictions, where actual fares were higher than predicted. Negative residuals, below the line, indicate overpredictions. There's some increasing spread for lower predicted fares and a tapering effect for high predicted fares, which suggest there is possible nonlinearity or variance in lower cost rides and the model struggles more with outliers and extreme values.

The Generalized Additive Model (GAM) is a regression method that allows for nonlinear relationships between predictors and the outcome. In our case, GAM was used to model Uber fare_amount based on predictors like pickup_hour, pickup_dayofweek, trip_distance, and passenger_count. Unlike linear regression, which assumes a straight line effect for each variable, GAM uses smoothing functions to capture more complex, curved patterns in the data. It is especially useful for time-based features where pricing may rise or fall at different rates during the day or across the week.



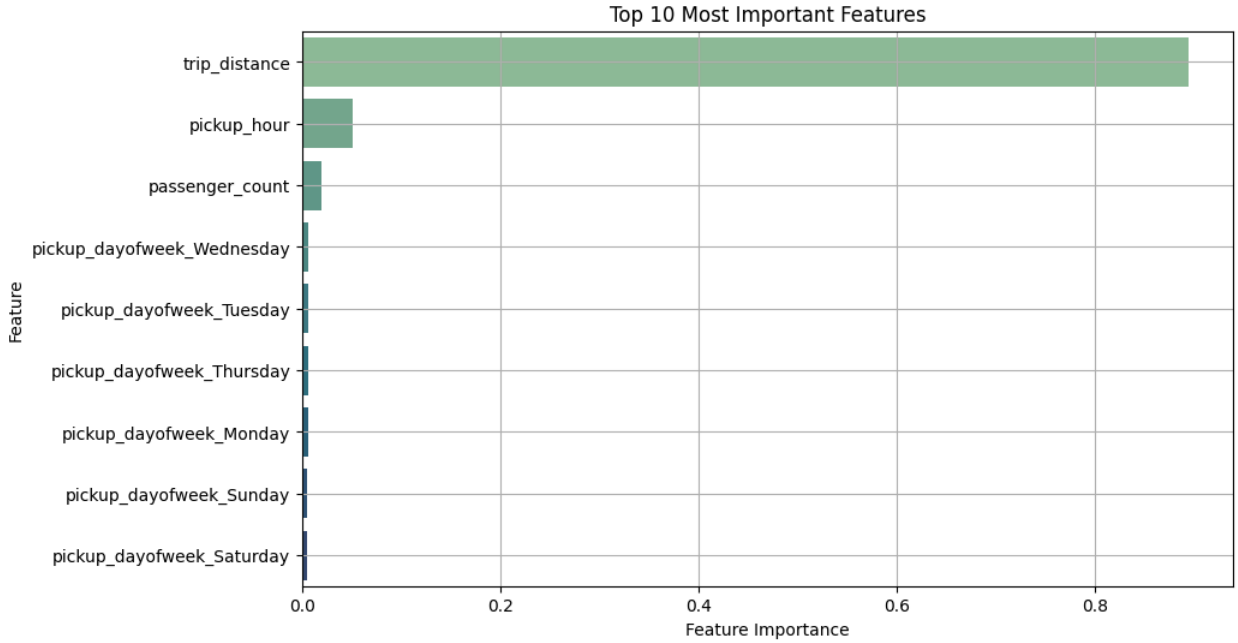
Looking at this graph, fares tend to increase during peak pickup_hour periods, especially between 8 AM and 6 PM, when demand is higher due to commuting patterns and daytime activity. The GAM plot for pickup_hour reveals a sharp rise in predicted fare amounts in the early morning hours, peaking around midday, then gradually tapering off into the evening. This reflects real world demand surges, such as early morning airport rides or afternoon business travel.

The GAM curve for *pickup_dayofweek* shows that fare amounts are generally higher toward the end of the week, particularly on Fridays and weekends, likely due to increased leisure and nightlife activity. In contrast, midweek days like Tuesday and Wednesday show dips in fare

impact, indicating less demand. By using smoothed curves, the GAM model captures nonlinear time based fare patterns more effectively than traditional linear regression.

The Random Forest Regressor is a learning method that builds multiple decision trees and averages their predictions to improve accuracy and reduce overfitting. Unlike Linear Regression, which models a single linear relationship, Random Forest captures nonlinear interactions between variables and naturally handles categorical data after one-hot encoding

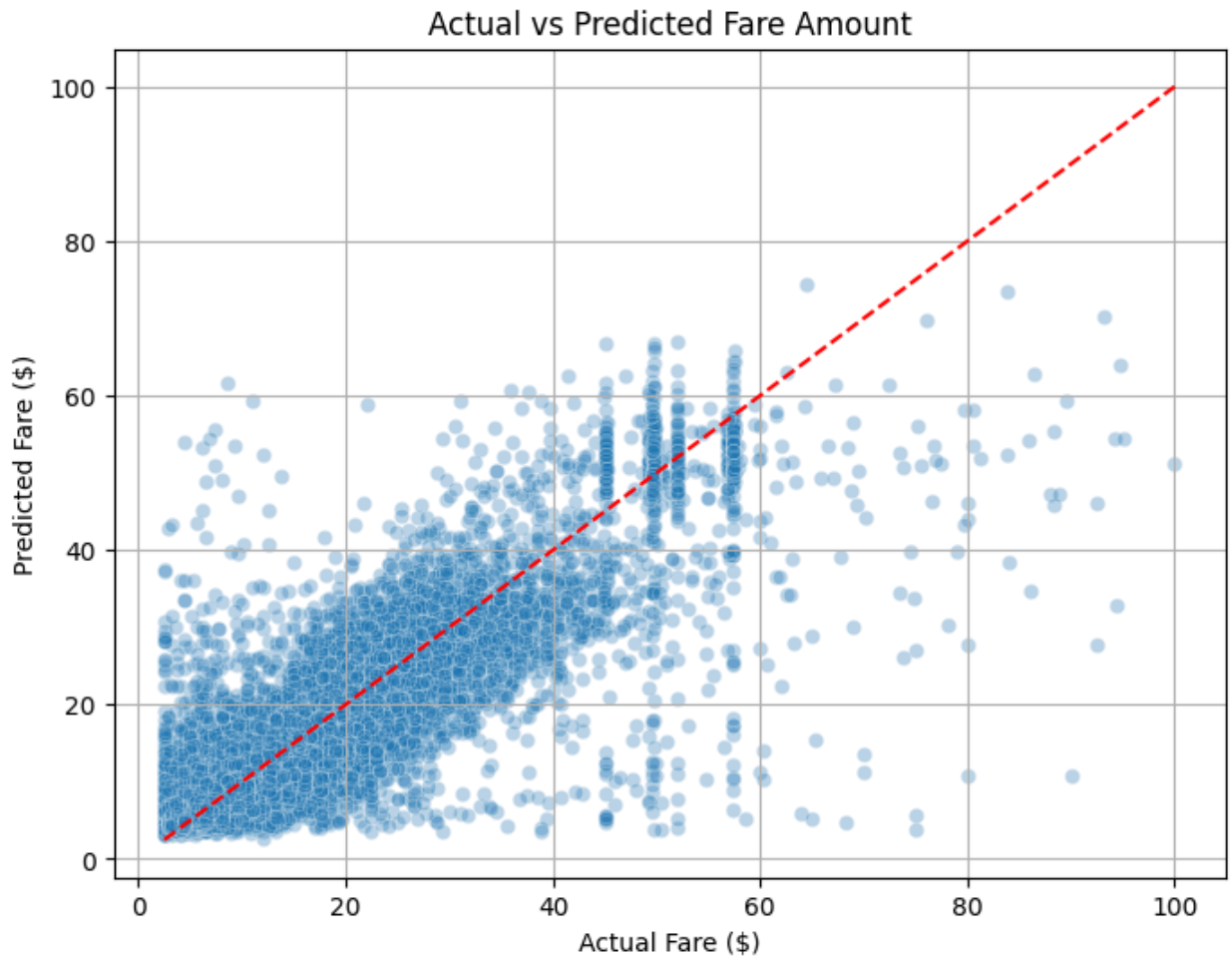
In this model, we trained the Random Forest using the same key features from our earlier analysis: *trip_distance, pickup_hour, passenger_count,* and *pickup_dayofweek*



The bar chart above shows the features that influence fare prediction the most. As we observe, we notice that *trip_distance* is by far the most important feature. This aligns with our hypotheses. It contributes to over 90% of the decision-making power in this model.

This confirms that the length of a ride is the strongest driver of fare amount.

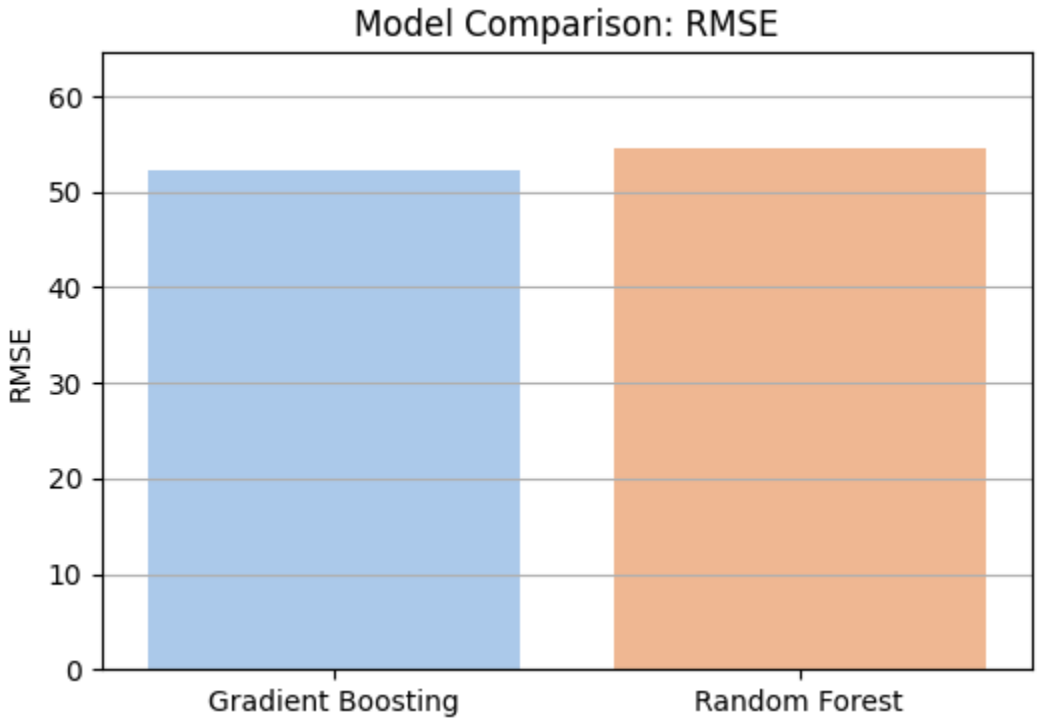
Pickup_hour still has some effect on it as well. It captures that there are also some time-based demand patterns that exist at Uber.



This scatter plot compares the model's predicted fare amounts to the actual fares from the test set. Most of the points cluster around the red diagonal line, which indicates accurate predictions.

This model performs well for the majority of fares under $60, but we do notice increased variance and under-prediction for higher fares. This could be due to unmodeled factors like traffic conditions or surge pricing.

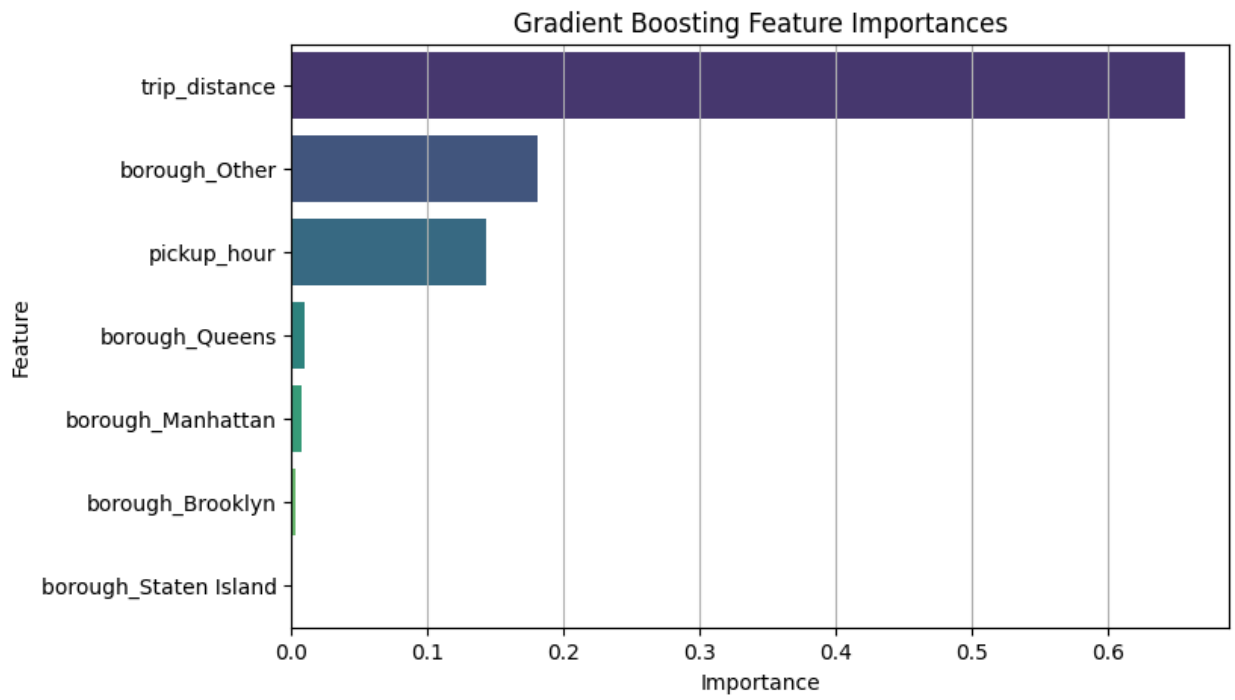
Gradient Boosting was chosen for its ability to capture complex interactions between continuous and categorical variables. It builds a series of trees where each one corrects the mistakes of the previous, leading to strong predictive performance on structured data.



This bar chart compares the Root Mean Squared Error (RMSE) for two models, Gradient Boosting and Random. RMSE measures the average size of the prediction error. Lower values indicate more accurate models.

On the left, we see that Gradient Boosting is slightly shorter, meaning it had a lower RMSE and performed better. On the right, Random Forest has a slightly higher error, suggesting it was less precise in predicting fare-per-mile.

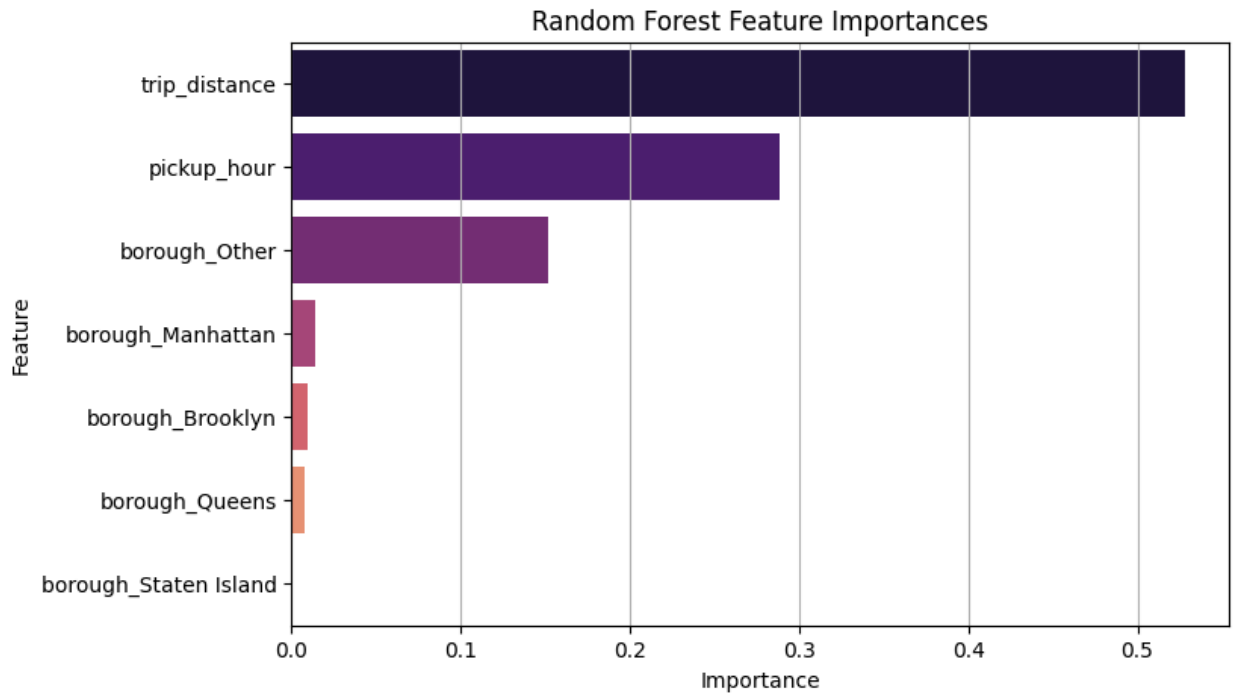
This visual shows that Gradient Boosting was the most accurate.



Feature Importance plots help us understand which variables are driving the model's predictions. This provides insight into the factors that most influence fare-per-mile.

Trip Distance is continuing to be the most important feature. This reflects how shorter trips tend to inflate fare-per-mile due to base fare. *Borough_Other* and *pickup_hour* follow in importance, indicating that both location and time of day meaningfully influence fare efficiency. The other boroughs like Queens, Manhattan, and Staten island have very low importance individually.

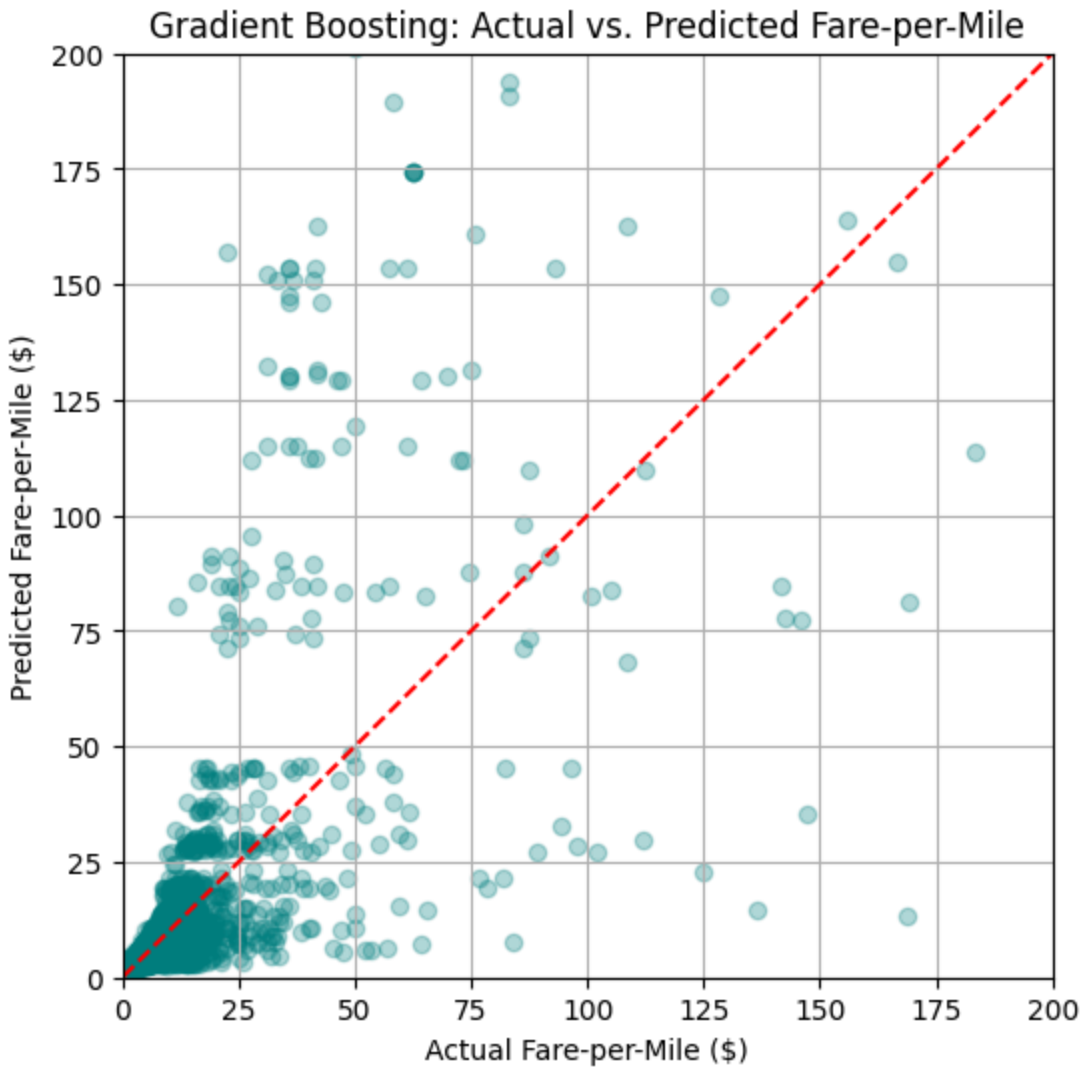
This visual reinforces that distance and timing dominate fare-per-mile variation, while geographic effects are present but more subtle.



This chart displays how much each feature contributed to each Random Forest model's ability to predict fare-per-mile.

Trip distance is again the most influential feature, but less dominant compared to the Gradient Boosting model. Pickup hour plays a more significant role in Random Forest than in Gradient Boosting, reflecting Random Forest's sensitivity to time-based variations. *Borough_Other* is the third most important, indicating that location does affect fare efficiency but isn't as impactful as distance or time. Individual boroughs continued to contribute only minor gains.

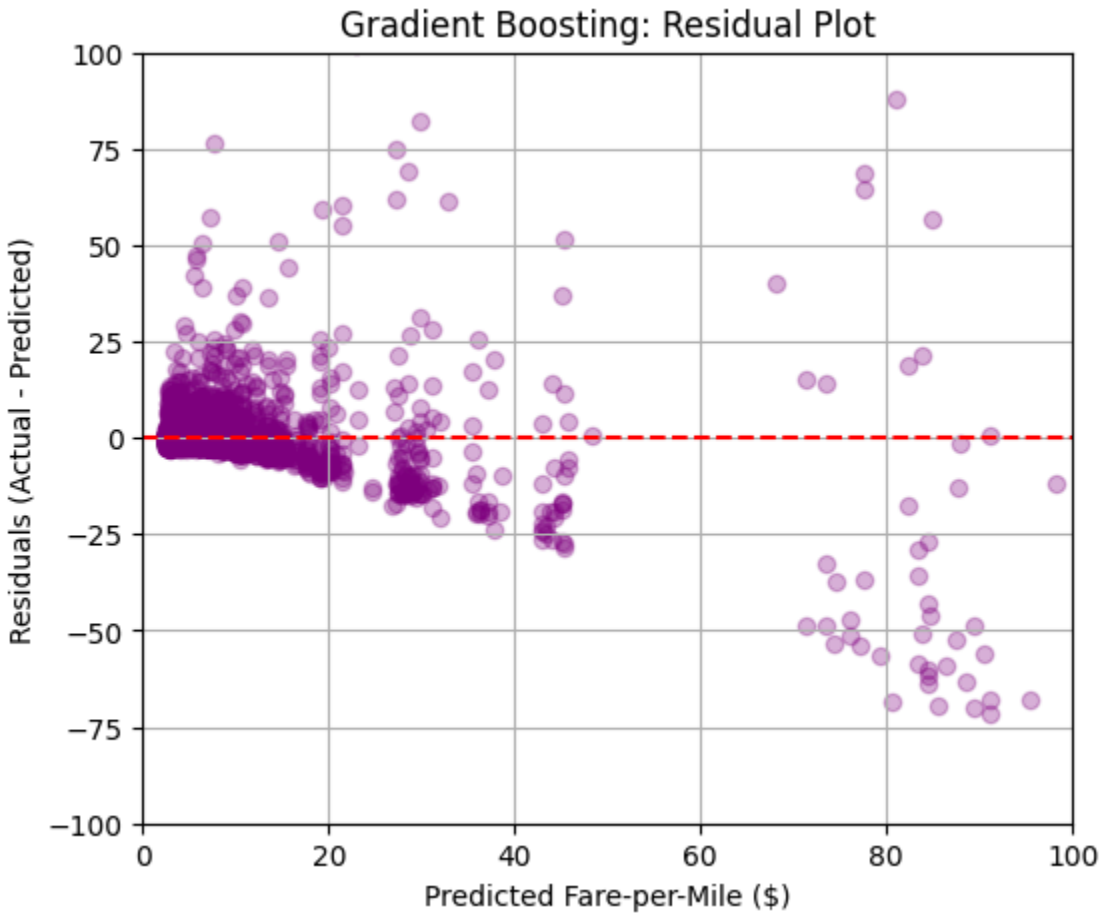
Overall, this visualization suggests that Random Forest relies more evenly on time and geography than Gradient Boosting, but both models agree that trip distance is the most critical factor.



This plot allows us to visually assess how well the Gradient Boosting model fits the data across fare-per-mile values. Points close to the diagonal line indicate good predictions.

Most points are clustered in the lower-left region, where fares-per-mile are lower and predictions are more accurate. As fares increase, the model's predictions tend to underestimate. Many points fall below the red line, indicating actual fares are higher than predicted. The spread widens for larger values, showing greater error and variability in high fare-per-mile trips. A few extreme outliers show predictions far from actuals, which more than likely reflect unusual or short trips with very high base fare impact.

Gradient Boosting predicts common, reasonably priced trips well, but struggles with extreme or outlier fare-per-mile values. This could be because of the unpredictable nature of short trips, surcharges, or borough specific events.



The Residual Plot helps us check for model bias and whether prediction errors are consistent across the range of fare-per-mile predictions. Ideally, residuals should be centred around zero with no clear patterns.

Most residuals cluster near zero for predictions under $20, indicating good accuracy for typical trips. There's a clear pattern of underprediction at higher fare-per-mile values. Many residuals

fell above the red line, meaning actual fares were higher than predicted. Similarly, there's overprediction at the far end where the model overshot a few extreme values.

The Gradient Boosting model performs consistently for the majority of trips, but its residuals highlight growing uncertainty at higher fare-per-mile predictions. This indicates the model could benefit from better handling of outliers or more advanced tuning for extreme trip types.