# Final Report: Predicting Ride Fare Amounts

Title: Predicting Ride Fare Amounts Using Machine Learning Models

1. Introduction

The goal of this project is to develop a predictive model for ride fare amounts, leveraging historical ride data. Accurate fare predictions will help optimize pricing, improve customer satisfaction, and streamline driver operations.

2. Data Exploration and Preprocessing

- Dataset Overview: The dataset contains 200,000 entries with features such as pickup/dropoff locations, time, and passenger count.
- Missing Values: Minor missing data in latitude/longitude was handled by removing affected rows.
- Feature Engineering:
  - Derived the distance between pickup and dropoff points using the Haversine formula.
  - Extracted time-based features (hour, day of the week, month) from the pickup time to capture temporal patterns.

3. Modeling Process

Three regression models were trained:

- Linear Regression
- Decision Tree Regressor
- Random Forest Regressor

The dataset was split into training (80%) and testing (20%) sets. After model training, the Random

Forest model showed the best performance, and further fine-tuning was done using RandomizedSearchCV to improve its parameters.

4. Model Performance Comparison

| Model | MSE | MAE | $R^2$ |
|---|---|---|---|
| Linear Regression | 103.85 | 6.06 | 0.001 |
| Decision Tree Regressor | 54.39 | 3.09 | 0.477 |
| Random Forest Regressor | 33.13 | 2.39 | 0.681 |
| Fine-Tuned Random Forest | 29.94 | 2.31 | 0.712 |

The fine-tuned Random Forest model, with an $R^2$ score of 0.712, showed the highest predictive accuracy, explaining 71% of the variance in ride fares. This model significantly outperformed both the Linear Regression and Decision Tree models in terms of accuracy.

5. Feature Importance

The most important features influencing fare predictions were:

- Distance (90% importance)
- Dropoff longitude and latitude

Distance was the most critical feature for fare prediction, as expected, while time-related features and passenger count had lower predictive power.

6. Business Implications

- Dynamic Pricing Strategy: Use the model's predictions to implement surge pricing during peak hours or in high-demand areas, maximizing revenue and balancing supply-demand.

- Driver Allocation: Deploy more drivers in areas and times where higher fare predictions occur, optimizing service availability and reducing customer wait times.

- Revenue Growth: Predictive insights from the model can help optimize pricing strategies and improve profitability during high-demand periods.

7. Model Limitations and Future Work

- Overfitting Concern: The slight difference between the training and test $R^2$ (0.826 vs. 0.712) suggests a potential for overfitting, which could be mitigated by collecting more diverse data.

- Additional Features: Incorporating features like traffic conditions, weather, and special events could further enhance the model's predictive capability.

- Advanced Techniques: Future iterations could explore Gradient Boosting or XGBoost to further improve accuracy.

8. Conclusion

The fine-tuned Random Forest model offers a robust and accurate solution for predicting ride fares, with an $R^2$ of 0.712. This predictive capability will enable the company to implement dynamic pricing, optimize driver allocation, and improve customer satisfaction, contributing directly to revenue growth.