Data 602: Introduction to Data Analysis and Machine Learning

Final Project Executive Summary
Dynamic Fare Prediction in Ride-Hailing Apps

Introduction and Problem Statement:

Dynamic pricing algorithms used by ride-hailing companies, such as Uber and Lyft, take into account real-time factors such as distance, ride-time, time of day, and demand to calculate fares. Such fare pricing models make it difficult for the users to predict fare fluctuations, triggering discontent among customers and an opaque pricing model.

The project aims to build a machine-learning model to forecast ride-hailing fare prices considering historical trip data so that understanding the major trip-related factors that affect fare amounts can serve businesses to better pricing strategies and customers to better estimate their ride costs.

Methodology:

We followed the CRISP-DM framework to complete this project. The steps included data understanding, preprocessing, exploratory analysis, model selection, evaluation, and deployment.

Dataset Overview

- **Source:** Uber Ride Data (2024), Contains ride information such as trip distance, trip duration, start time, and fare. https://data.cityofchicago.org/Transportation/Taxi-Trips-2024-/ajtu-isnz/about_data
- **Records:** Over 7 million observations
- **Key features:** Trip Miles, Trip Duration, Start Hour, Day of Week, and Fare (target)

Data Cleaning & Feature Engineering

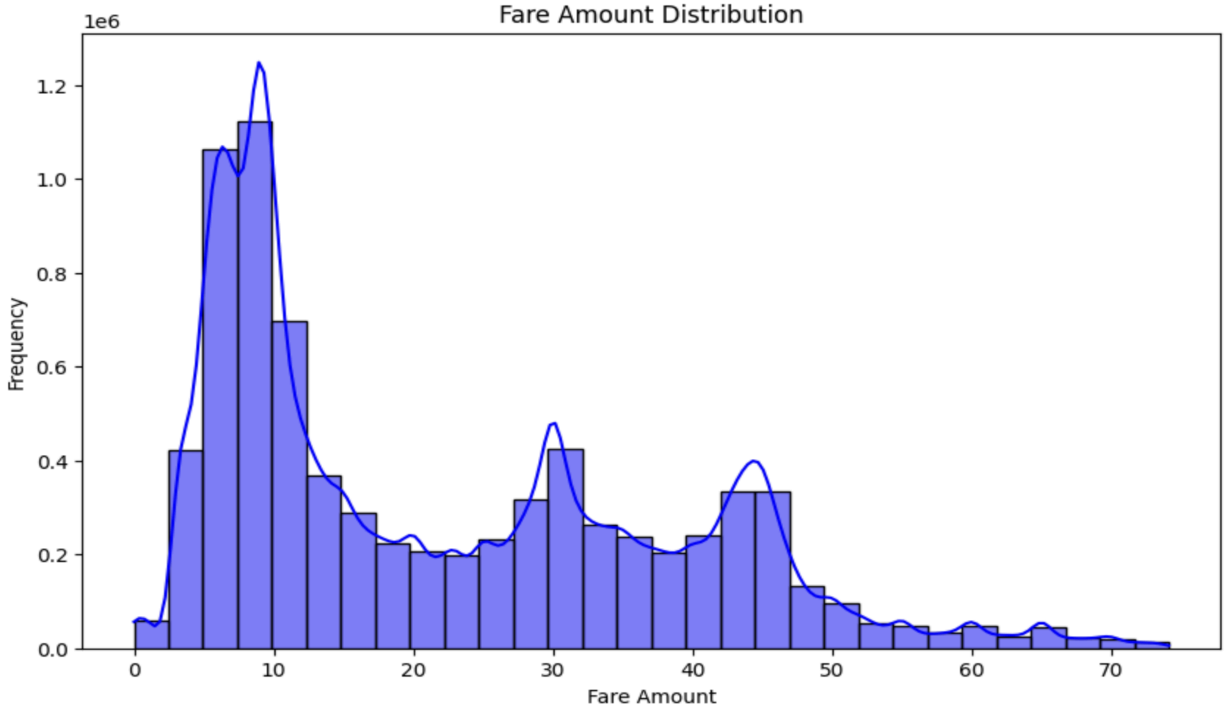
- Removed irrelevant and highly missing columns
- Imputed missing values using mean
- Extracted new features like Trip Duration, Start Hour, and Day of Week from timestamps

Exploratory Data Analysis (EDA):

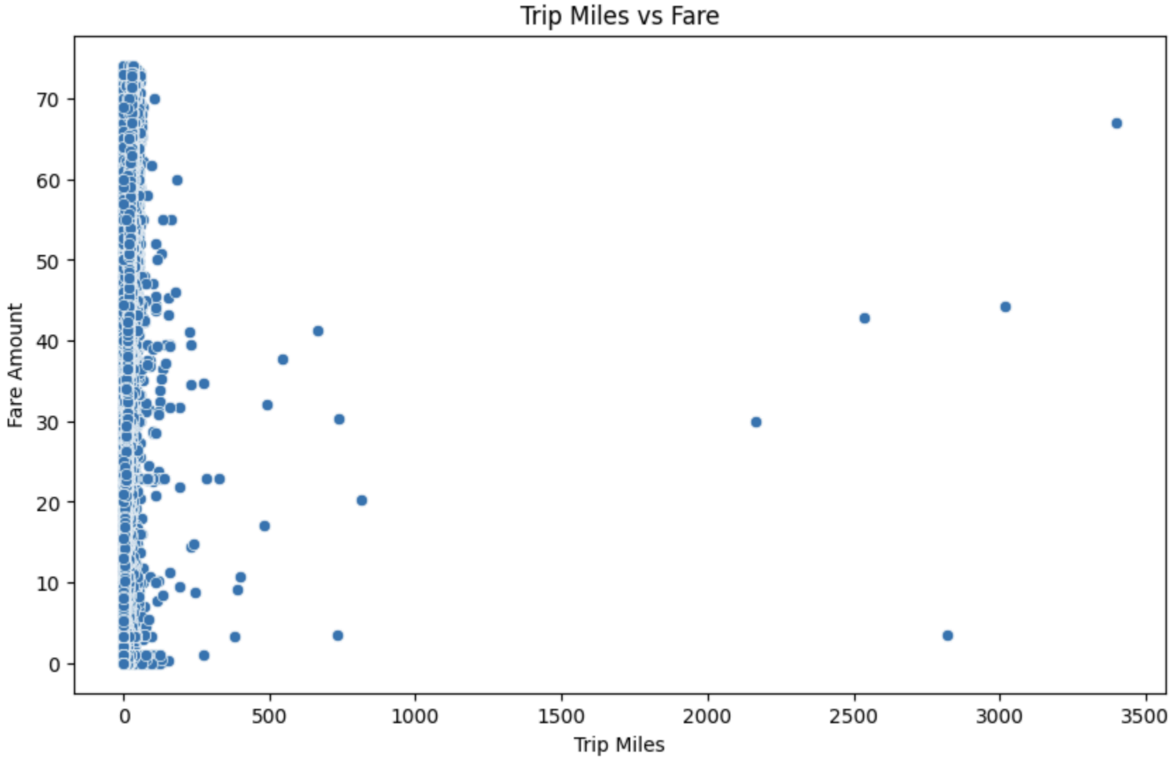
EDA helped uncover patterns in fare amounts and evaluate feature correlations.

- **Fare Distribution:** Right-skewed, with peaks around $10, $30, and $50
- **Outliers:** Some records had very high mileage but unusually low fares, indicating possible data errors
- **Correlations:**
    - Trip Miles and Fare: 0.86 (strong positive correlation)
    - Trip Duration and Fare: 0.47 (moderate)
    - Time-based features (e.g., Start Hour) had weak correlation with fare

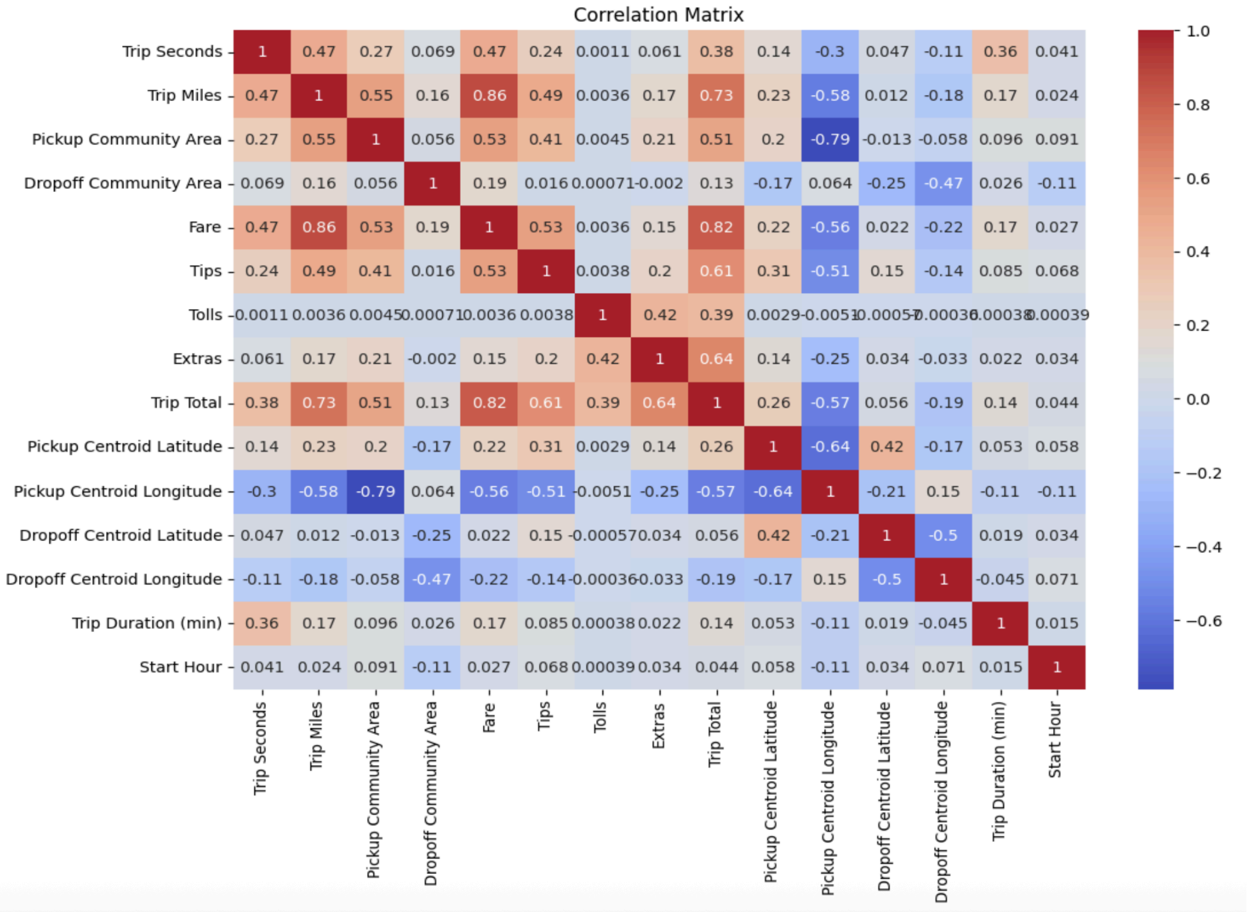
Histogram or KDE plot of fare distribution:



Scatter plot – Trip Miles vs. Fare



Correlation heatmap or bar chart of correlation values



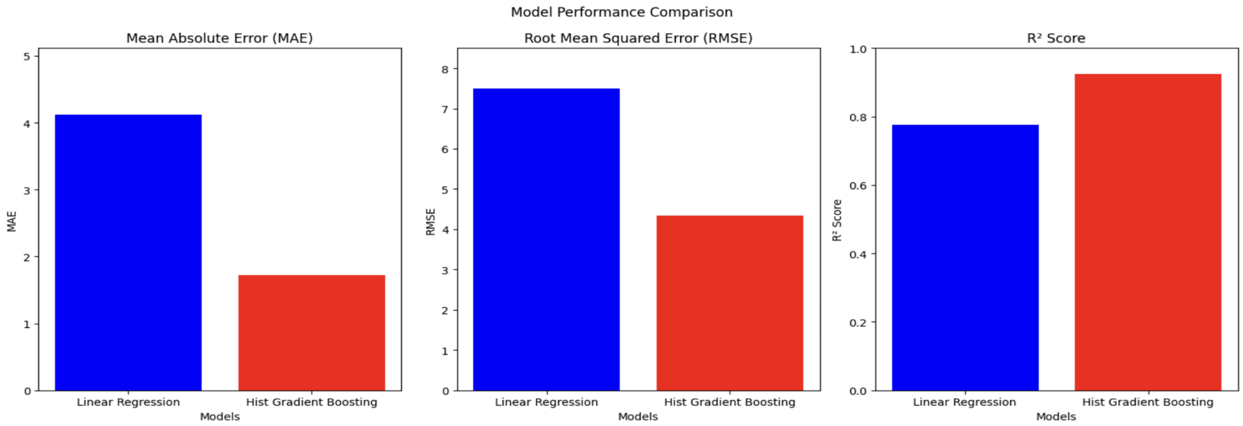
Model Development and Evaluation:

Two models were trained and evaluated:

Model	MAE	RMSE	R² Score
Linear Regression	4.12	7.50	0.775
Histogram Gradient Boosting (HGB)	1.72	4.34	0.925

HGB significantly outperformed linear regression with higher accuracy and lower error margins and was chosen as the final model for deployment.

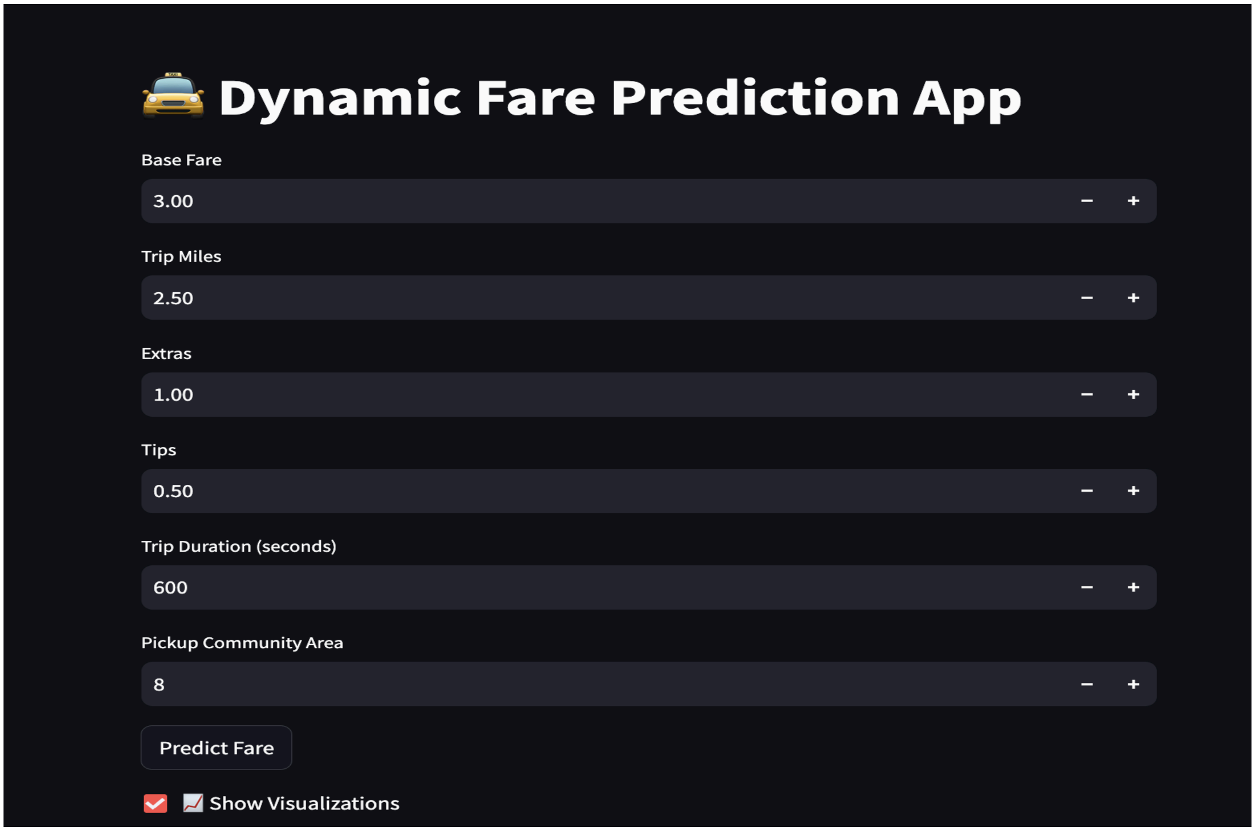
Bar chart of Model performance comparison



Deployment:

The final model was deployed using Streamlit, enabling real-time predictions via a user-friendly web interface.

- **User Inputs:** Trip distance, start hour, and day of the week
- **Outputs:** Predicted fare in real-time
- **Why Streamlit?** Lightweight, interactive, and simple to integrate with Python ML models



🚖 Dynamic Fare Prediction App

Base Fare

3.00

Trip Miles

2.50

Extras

1.00

Tips

0.50

Trip Duration (seconds)

600

Pickup Community Area

8

☑ 📈 Show Visualizations

Conclusion:

This project successfully developed and deployed a fare prediction model using Histogram-Based Gradient Boosting, achieving an R^2 score of 0.925, which indicates strong predictive performance.

The analysis confirmed that Trip Miles is the most impactful feature, while time-related factors like Start Hour and Day of Week showed minimal influence. The Streamlit app enabled real-time fare predictions, providing an intuitive and interactive user experience.

Key Takeaways & Next Steps

- Trip Duration moderately influenced fares but was secondary to distance.
- Data cleaning was critical due to outliers like low-fare long trips.
- The deployed model is lightweight and suitable for user-facing tools.
- Incorporate external data (e.g., weather, traffic) to improve accuracy.
- Enable real-time retraining using streaming data.
- Experiment with deep learning for further performance gains.