ADVANCES IN DATA SCIENCE AND ARCHITECTURE

FINAL PROJECT

OPTIMIZING RIDE-HAILING FARES WITH PREDICTIVE ANALYTICS

Kumar Mehul (002761391)
Krutik Kanakia (002787847)
Rhea Bajpai (002702927)
Tanuj Verma (002726506)

Project Overview

INTRODUCTION

In the competitive landscape of ride-hailing services, strategic fare pricing is critical for business success and customer satisfaction. This project taps into the potential of data science to create a predictive model for taxi fares, empowering ride-sharing companies to optimize their pricing models dynamically.

ABOUT THE DATASET

The dataset is a rich compilation of ride details from a ride-sharing service, with each entry capturing the essence of urban commutes. The key attributes include fare amount, pickup and drop-off coordinates, passenger count, and datetime information—each serving as a pillar for constructing our predictive analysis.
Below is a description of each column contained within the dataset:

- **key**: Unique identifier for each trip, derived from ride timestamps or generated hashes.
- **fare_amount**: Total fare charged, serving as the target variable for prediction.
- **pickup_datetime:** Start date and time of the ride, key for analyzing fare trends.
- **pickup_longitude:** Longitude of the pickup point, crucial for location-based analysis.
- **pickup_latitude:** Latitude of the pickup point, determines the ride's starting location.
- **dropoff_longitude:** Longitude of the drop-off point, aids in calculating travel distances.
- **dropoff_latitude:** Latitude of the drop-off point, completes the route data for analyses.
- **passenger_count:** Count of passengers, affects fare due to different pricing for passenger numbers.

Dataset Kaggle link: https://www.kaggle.com/datasets/yasserh/uber-fares-dataset

OBJECTIVES

The overarching goal of this project is to accurately predict taxi fares to aid ride-sharing companies in:

- Refining their pricing strategies.
- Improving the overall customer experience.
- Achieving operational excellence.

Specific objectives encompass data preprocessing to handle geospatial and temporal information, extensive exploratory data analysis to unveil underlying patterns, visualization to represent complex relationships, and rigorous model evaluation to ascertain predictive accuracy.

SIGNIFICANCE OF THE PROJECT

This endeavor is significant for enabling data-driven decision-making, leading to:

- **Business Optimization**: Enhanced dynamic fare management for increased profitability and competitive advantage.
- **Customer Satisfaction:** Greater transparency in fare calculation boosts customer trust and retention.
- **Operational Efficiency:** Insights from the model inform better resource allocation and route optimization.

METHODOLOGY

Data Preprocessing: Transformed geospatial data for precise mapping and analysis.

Exploratory Data Analysis (EDA): Conducted in-depth statistical reviews to understand fare distribution and key contributing factors.

Visualization: Employed a variety of graphical representations to showcase insights and foster a deeper understanding of the data.

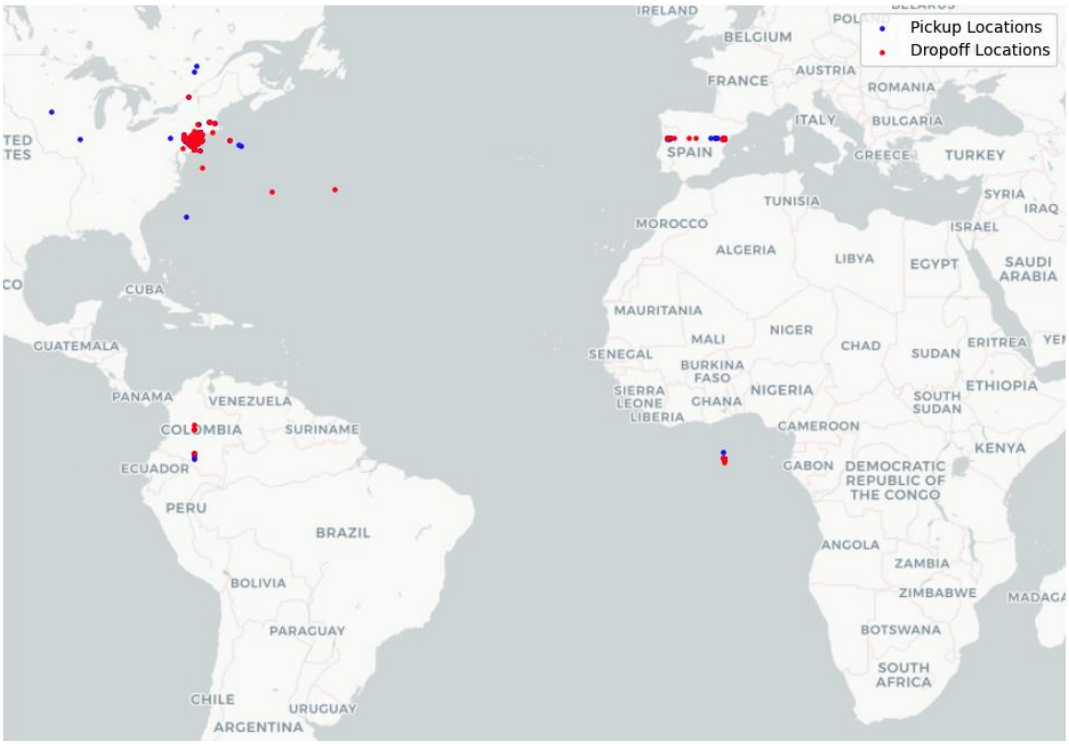
Model Evaluation: Assessed multiple regression models using key performance indicators to identify the most accurate predictive model.

BROWNIE POINTS ACHIEVED

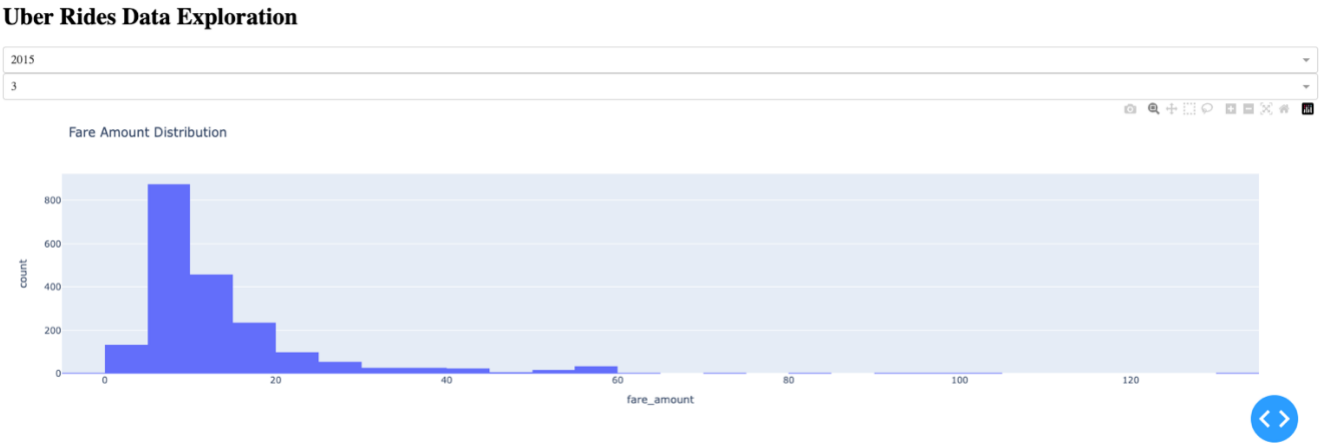
- **Novel Engineering:** We split the date column into multiple columns like Year, Month, Weekday, Hour, Monthly_Quarter and Hourly_Segments. We do this to enrich our dataset into valuable segments for clearer analysis.

	fare_amount	pickup_longitude	pickup_latitude	dropoff_longitude	dropoff_latitude	passenger_count	year	weekday	Monthly_Quarter	Hourly_Segments	Distance
0	7.5	-73.999817	40.738354	-73.999512	40.723217	1	2015	3	Q2	H5	1681.11
1	7.7	-73.994355	40.728225	-73.994710	40.750325	1	2009	4	Q3	H6	2454.36
2	12.9	-74.005043	40.740770	-73.962565	40.772647	1	2009	0	Q3	H6	5039.60
3	5.3	-73.976124	40.790844	-73.965316	40.803349	3	2009	4	Q2	H3	1661.44
4	16.0	-73.925023	40.744085	-73.973082	40.761247	5	2014	3	Q3	H5	4483.73
5	4.9	-73.969019	40.755910	-73.969019	40.755910	1	2011	5	Q1	H1	0.00
6	24.5	-73.961447	40.693965	-73.871195	40.774297	5	2014	6	Q4	H2	11734.67
7	2.5	0.000000	0.000000	0.000000	0.000000	1	2012	1	Q4	H4	0.00
8	9.7	-73.975187	40.745767	-74.002720	40.743537	1	2012	4	Q1	H3	2338.56
9	12.5	-74.001065	40.741787	-73.963040	40.775012	1	2012	3	Q1	H5	4891.12

- **Innovative Visualization Techniques:** Using folium for geographic analysis visualization.

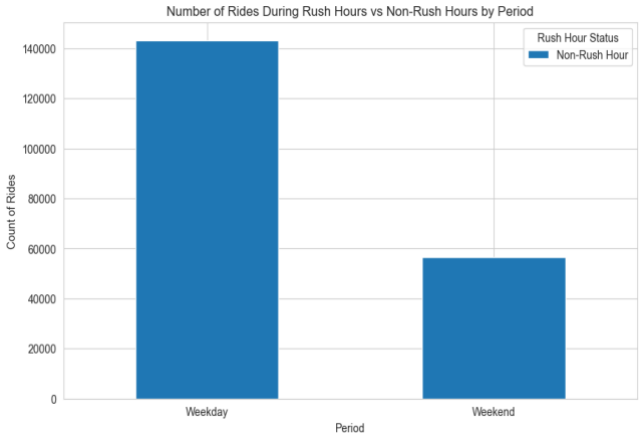


- **Interactive Dashboards:** Created dynamic, user-interactive dashboards for real-time data exploration using dash.

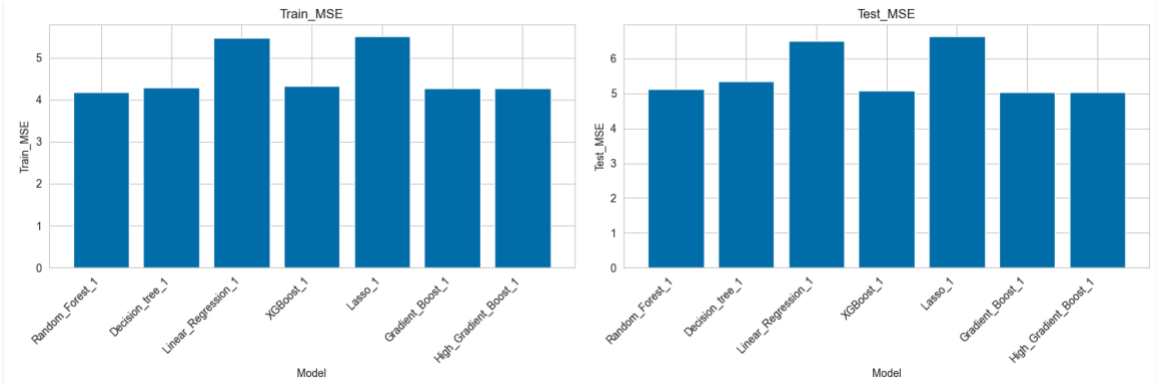


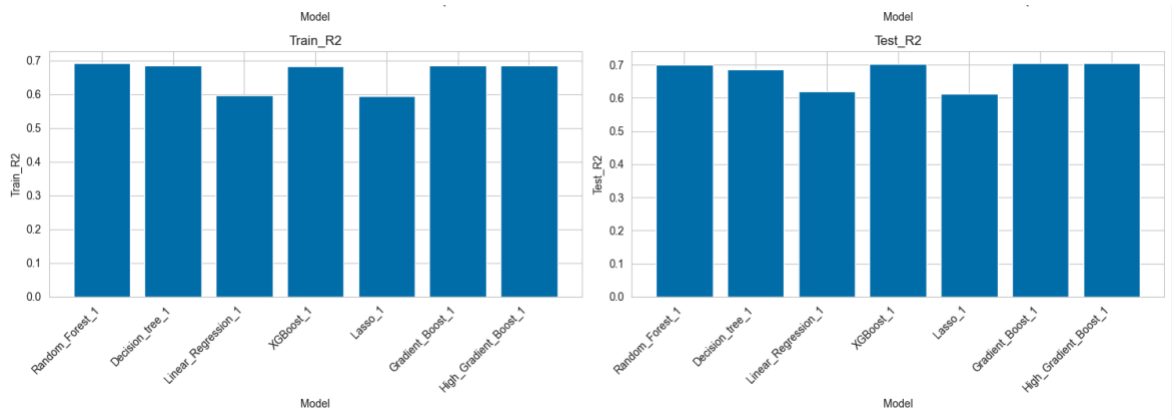
Created an interactive dashboard where you can view the passenger count and fare amount comparison year and weekday wise.

- **Novel Feature Engineering:** Devised new features like rush-hour and fare efficiency for nuanced insights.



- **Algorithmic Innovation:** `RandomizedSearchCV` and `GridSearchCV` are both methods used for hyperparameter tuning in machine learning models





The **`HistGradientBoostingRegressor`** and **`GradientBoostingRegressor`** exhibit superior performance in terms of test R-squared values, indicating their strong generalization capabilities. This suggests that for this dataset, boosting methods are beneficial due to their ability to iteratively correct errors from previous models.

- **Creative Problem Framing:** By viewing our data through fresh lenses, we aim to reimagine the analytics landscape:

1.Dynamic Pricing: Transition from mere analysis to real-time prediction of fare efficiency, incorporating factors like traffic, events, and surge pricing.

2. Passenger Centricity: Refocus analysis on what drives passenger satisfaction, potentially incorporating comfort, safety, and sharing opportunities into our feature set.

3.Sustainability: Redirect our analytical efforts to measure and promote rides with lower environmental impacts, motivating the use of eco-friendlier transportation options.

CONCLUSION

Based on the analysis of the model performance, we can conclude the following:

1.**Boosting Methods Outperform Other Models**: The `HistGradientBoostingRegressor` and `GradientBoostingRegressor` exhibit superior performance in terms of test R-squared values, indicating their strong generalization capabilities. This suggests that for this dataset, boosting methods are beneficial due to their ability to iteratively correct errors from previous models.

2. **Overfitting in Complex Models**: There is a significant disparity between training and testing R-squared values for the `DecisionTreeRegressor`, which suggests overfitting. Although decision trees can capture complex patterns in the training data, they may not generalize well to unseen data. Ensemble techniques, as seen with the `Random Forest` and boosting algorithms, help overcome this by averaging multiple trees.

3. **Consistency Between Train and Test Performance**: Models like `XGBoost` and `Gradient Boosting` display a smaller gap between training and testing metrics, demonstrating a good balance between bias and variance. Such consistency indicates that these models are well-tuned to the problem at hand and are neither overfitting nor underfitting excessively.

4.**Simpler Models Have Merit**: Despite lower performance metrics, simpler models such as `Linear Regression` have their advantages, particularly in interpretability and faster prediction times. In situations where the explain ability of a model is crucial, or when computational resources are limited, these models can be a good compromise, given their reasonable performance and efficiency.

These insights can guide future iterations of the modeling process, prompting considerations such as additional feature engineering, alternative model selection, or further hyperparameter tuning to enhance predictive performance.