Northeastern University

Final Presentation

Jatin Satija, Utkarsh Kothari, Pruthvi Patel, Zeel Rajendrakumar Patel,

Ayush Hemeshbhai Patel

Master of Professional Studies in Analytics, Northeastern University

ALY6040: Data Mining Applications

Prof. Shahram Sattar

May 13, 2024

Uber And Lyft dataset

1.606 Best model with the lowest RMSE for predicting price

100% Maximum accuracy achieved in predicting cab types

About the Dataset

- The dataset is a comparison between Uber and Lyft rides in Boston, MA.
- It contains various details about each ride, including id, timestamp, hour, day, month, datetime, timezone, source, destination, cab type, product id, name, price, distance, surge multiplier, latitude, and longitude.
- The dataset also includes weather data for each hour, providing a short summary of the weather, temperature, apparent temperature, humidity, wind speed, wind gust, visibility, dew point, pressure, wind bearing, cloud cover, UV index, ozone, sunrise time, sunset time, moon phase, and more.

Reason for choosing this dataset

- The dataset provides comprehensive features for price prediction and cab type (Uber or Lyft) prediction.
- Includes ride details and hourly weather data, allowing analysis of various factors on ride prices.
- Comparison between Uber and Lyft offers insights into user preferences for these services.

Introduction

Project Objective

Objective 1

Dynamic Pricing Strategy
Predicting prices can aid in developing dynamic pricing strategies. This can optimize revenue and ensure competitive pricing in the market.

Objective 2

Customer Preference Analysis
Predicting the cab type (Uber or Lyft) can provide insights into customer preferences. This can inform marketing strategies, improve customer targeting, and enhance the overall customer experience.

Objective 03

Demand Forecasting
By predicting ride prices, we can better anticipate demand. This can help in managing supply efficiently, leading to improved service and customer satisfaction.

Data Set Overview

Rows 6,93,071

Cols 57

Summary

```
[ ]  # Get updated list of numerical & categorical features
     categorical_data = list(ride_data_copy.select_dtypes(include=['object','category']).columns)
     numerical_data = list(set(ride_data_copy.select_dtypes(exclude=['datetime64']).columns) - set(categorical_data))

     # Show descriptive statistics
     ride_data_copy[numerical_data].describe()
```

	cloudCover	apparentTemperatureMin	sunriseTime	sunsetTime	temperatureHigh	windBearing	precipIntensityMax	hour
count	693071.000000	693071.000000	6.930710e+05	6.930710e+05	693071.000000	693071.000000	693071.000000	693071.000000
mean	0.686502	29.731002	1.544027e+09	1.544060e+09	45.040982	220.055853	0.037374	11.619137
std	0.358534	7.110494	6.911393e+05	6.906634e+05	5.996541	99.102736	0.055214	6.948114
min	0.000000	11.810000	1.543147e+09	1.543181e+09	32.680000	2.000000	0.000000	0.000000
25%	0.370000	27.760000	1.543406e+09	1.543440e+09	42.570000	124.000000	0.000000	6.000000
50%	0.820000	30.130000	1.543752e+09	1.543785e+09	44.680000	258.000000	0.000400	12.000000
75%	1.000000	35.710000	1.544789e+09	1.544822e+09	46.910000	303.000000	0.091600	18.000000
max	1.000000	40.050000	1.545135e+09	1.545168e+09	57.870000	356.000000	0.145900	23.000000

8 rows × 46 columns

Price Prediction Models

The **pricing prediction algorithm** in the Uber and Lyft datasets helps to accurately estimate ride fares, which benefits both **service providers and passengers**. It helps providers optimise their pricing strategies and revenue management. It helps customers plan and budget their trips, improving their overall experience and trust in ride-sharing services.

1. Linear Regression

2. Decision Tree Regressor

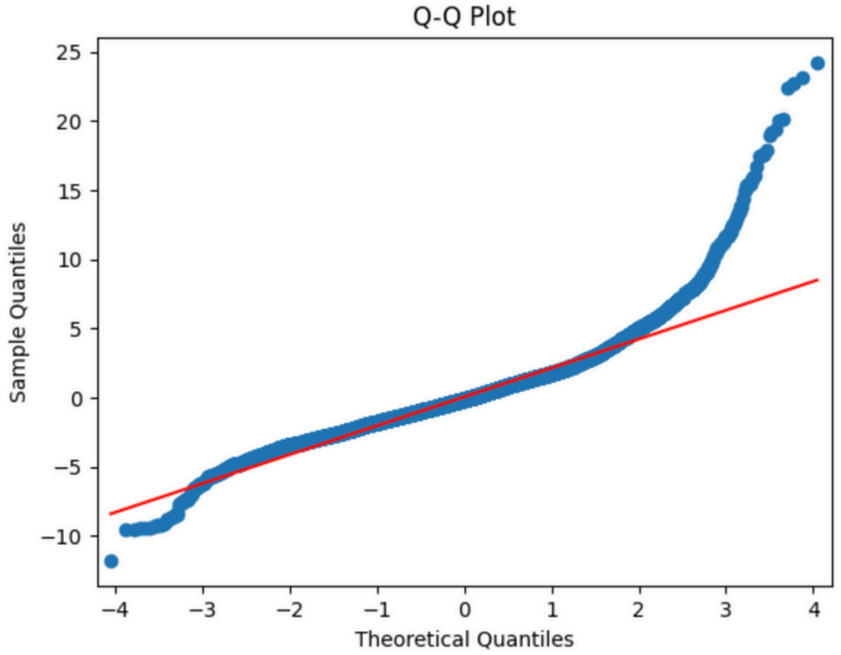
Linear Regression

RMSE: 2.088 **R2: 0.946**

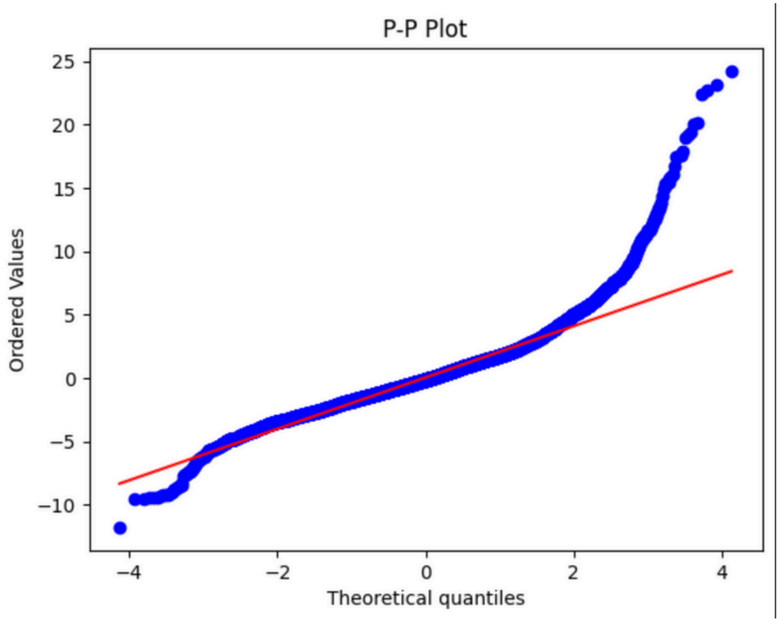
- Linear Regression is a statistical model used to predict a dependent variable (price) based on various independent variables (like distance, and service type), assuming a linear relationship between them.
- Performance metrics: RMSE and R^2 score.
- R^2 value of **0.946** indicates a **94.6%** variance in price explained by features.
- RMSE value of **2.087** shows average prediction error is about **$2.087**.
- The model demonstrates strong predictive ability and accurate price estimation.

Linear Regression cont'd

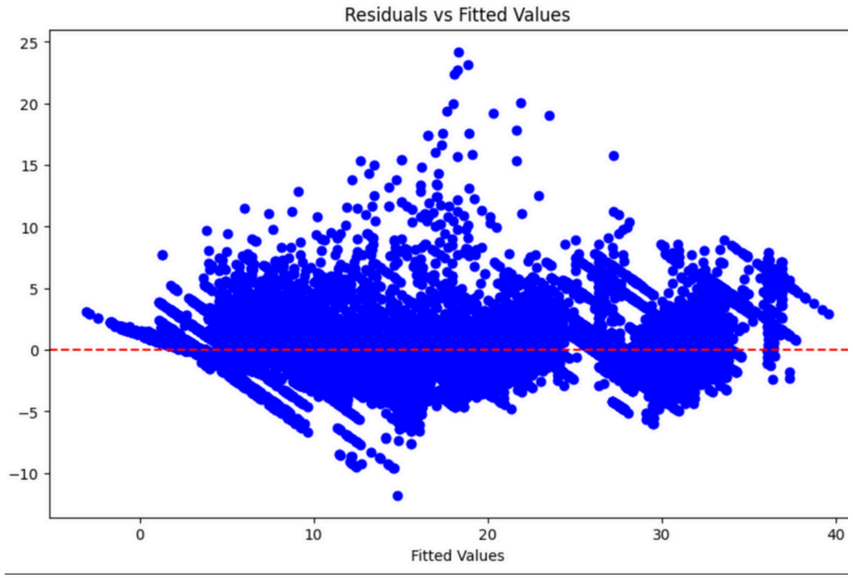
Visualisations



This indicates a strong resemblance between the dataset's quantiles and those of the theoretical distribution, suggesting a good fit between the two distributions.



This suggests a strong similarity between the empirical cumulative distribution function (ECDF) of the dataset and the cumulative distribution function (CDF) of the theoretical distribution, suggesting a good fit.



Since the residuals are randomly scattered, it means the model's predictions are consistent across all values, indicating a good fit for the data

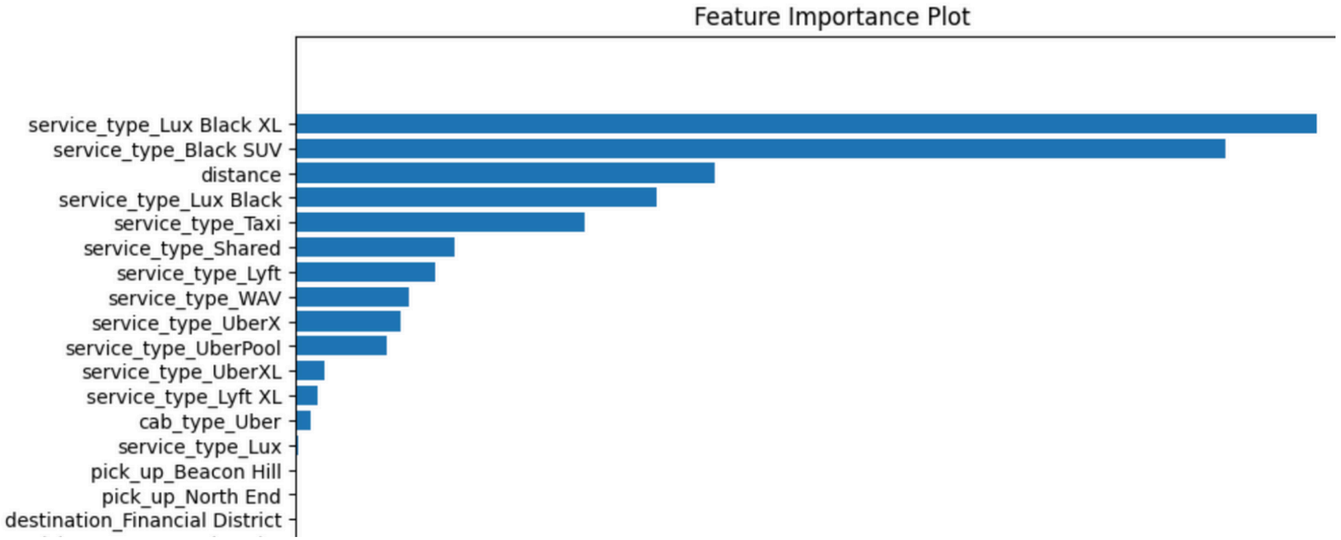
Decision Tree Regressor

RMSE: 1.606 **R2: 0.968**

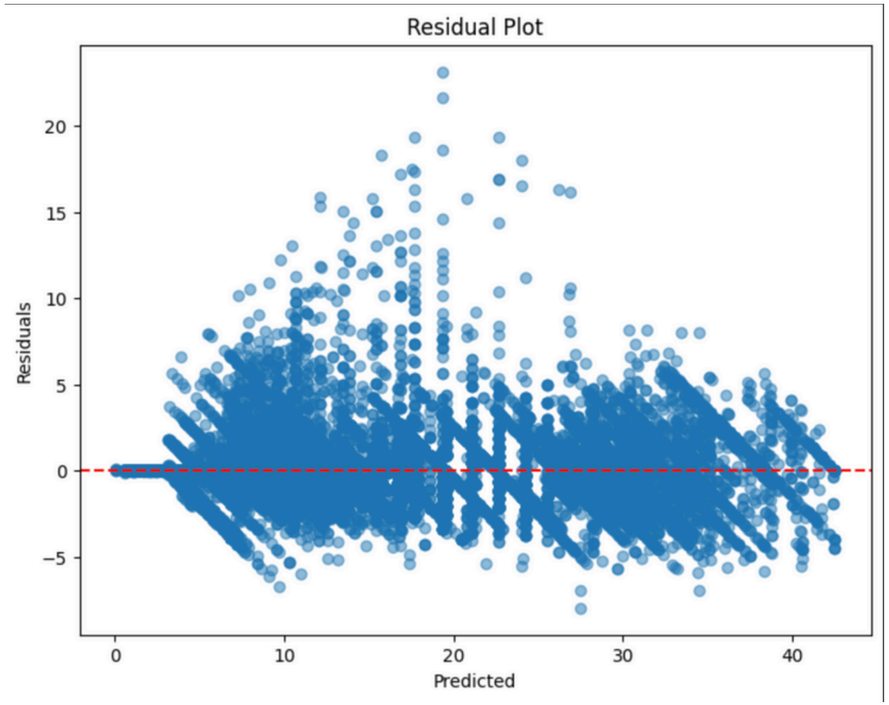
- A Decision Tree Regressor is a machine learning algorithm that builds a tree-like model to predict continuous target variables based on input features, splitting the data into subsets to make sequential decisions, aiming to minimize prediction error.
- Decision Tree Regressor's hyperparameters were fine-tuned using **Grid Search Cross-Validation.**
- A model was trained and tested with these parameters.
- Achieved R^2 score: **0.968 approximately** and RMSE: **1.606 approximately**.
- Significant improvement in prediction accuracy was found.

Decision Tree Regressor cont'd

Visualisations



This helps identify important factors influencing the price by demonstrating which features have the greatest influence on the model's predictions. This helps by concentrating on significant features, which helps with feature selection, understanding the dataset, and enhancing model performance.



The model's predictions are consistent across all values because the residuals are randomly distributed, which suggests a good fit for the data.

Cab Type Classifier

The **Cab Type Classifier** is designed to accurately classify rides as Uber or Lyft based on ride-related parameters such as service type, price information, and other relevant variables. This classification enables **service providers** to modify their offers, advertising, and pricing strategies for each platform. Additionally, it enables **consumers** to make informed selections about which ride-sharing service to choose, taking into account aspects such as cost, availability, and preferred service features.

1. Logistic Regression

2. Random Forest Regression

3. Linear Discriminant Analysis

Logistic Regression

Accuracy: 92.33%

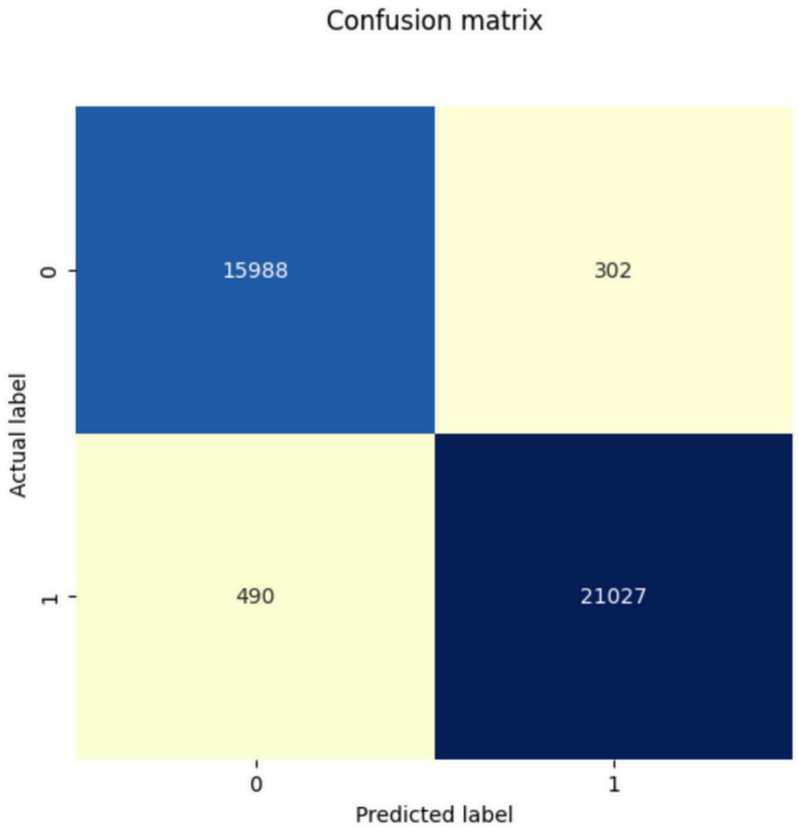
Precision: 93.46% **Recall: 92.33%.** **F1: 92.36%**

.

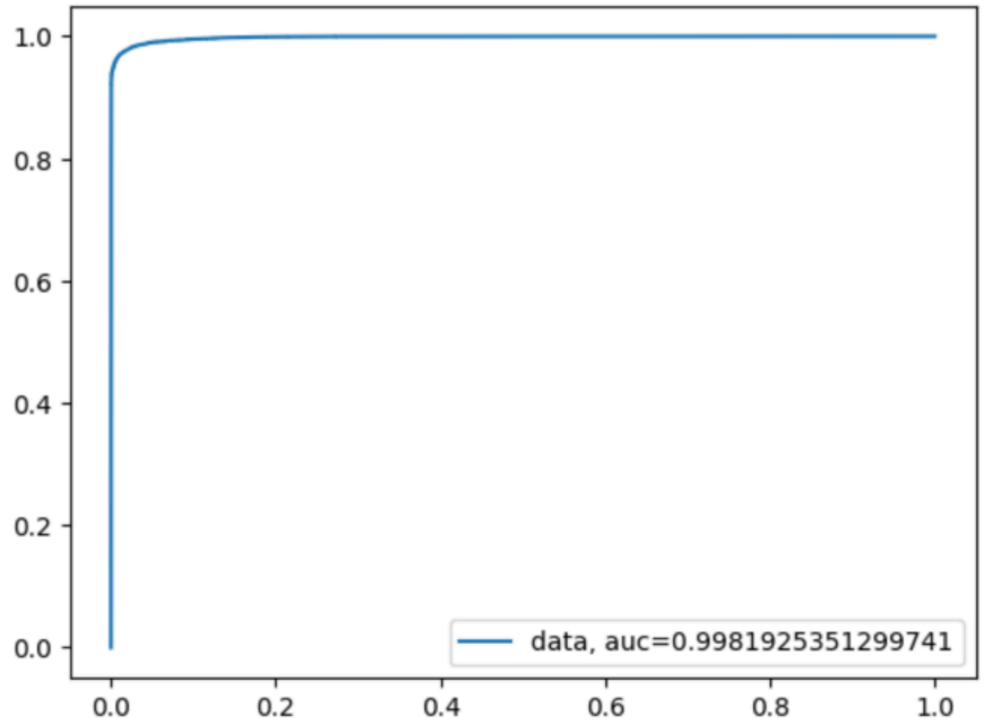
- Logistic Regression is a statistical method used for binary classification, where it models the probability of an event occurring based on input features, typically using the logistic function to map predictions to probabilities.
- It was used to predict cab types.
- The model was trained and tested after encoding categorical variables.
- Achieved accuracy of **92.33%,** demonstrating accurate cab type classification.
- Performance metrics: **Precision (93.46%), Recall (92.33%), F1-score (92.36%)**.
- AUC value of 0.9982, indicating effective differentiation between cab types.

Logistic Regression Cont'd

Visualisations

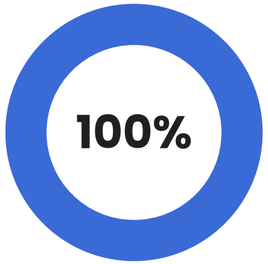


This confusion matrix displays the classification results of our model. Diagonal values represent correct predictions, while off-diagonal values indicate misclassifications.



The AUC (Area Under the Curve) value of 0.998 indicates the model's excellent ability to distinguish between the positive and negative classes. A higher AUC value closer to 1 suggests better discrimination power, with 1 indicating perfect classification.

Random Forest Classifier



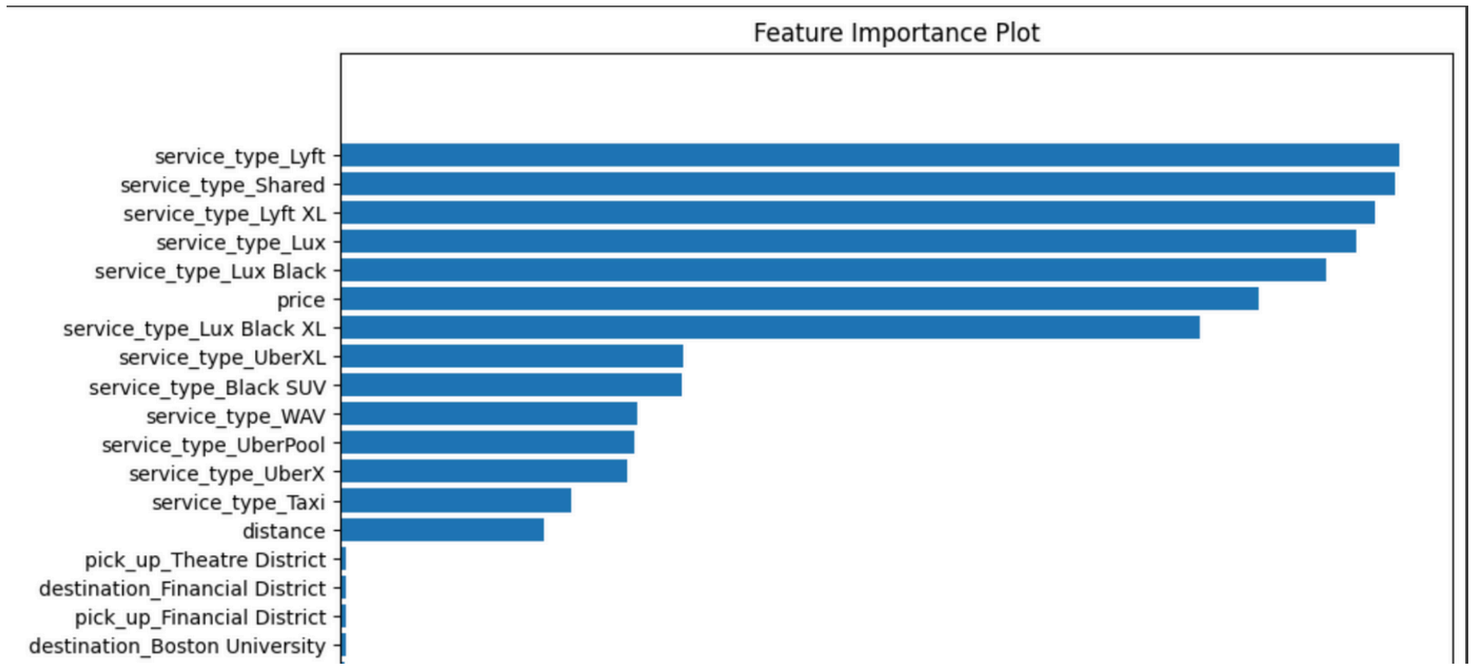
Accuracy: 100%

Precision: 100% **Recall: 100%** **F1: 100%**

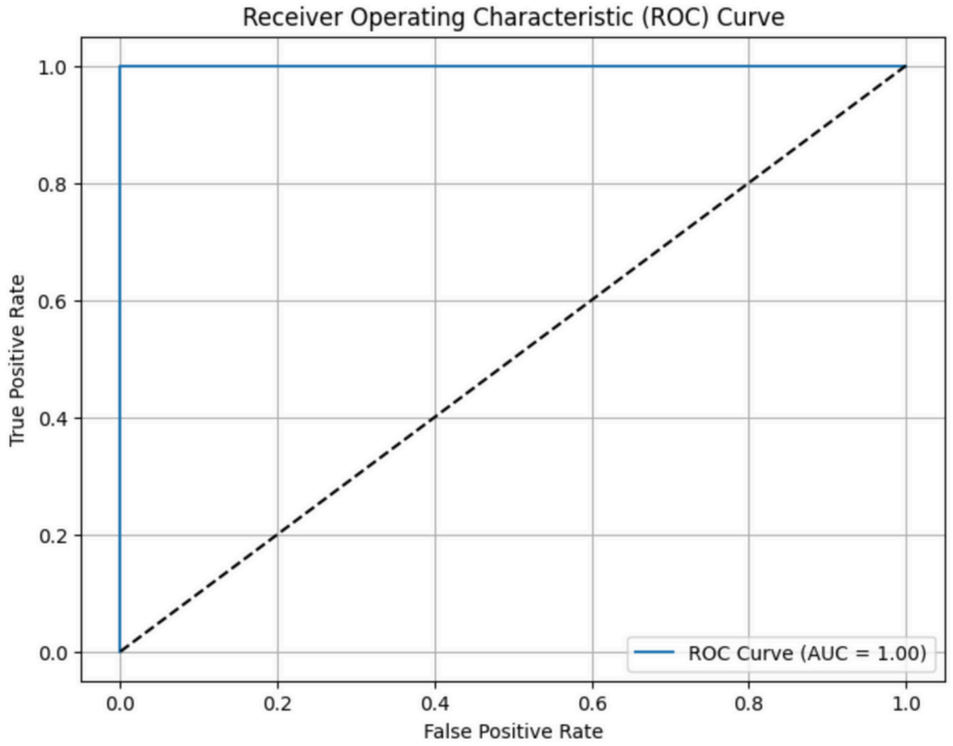
- Random Forest Classifier is a powerful machine learning model that combines many decision trees to make predictions. It picks the most common prediction from all the trees, making it great for classification tasks and handling lots of features.
- It was used to predict cab types.
- The model was trained and tested on the divided dataset.
- Achieved perfect **accuracy of 100%** on both training and testing data.
- Perfect scores **(100%)** for precision, recall, and F1-score metrics.
- Demonstrates robustness and dependability in classifying cab types.

Random Forest Classifier cont'd

Visualisations



This shows which features have the most impact on the model's predictions, helping identify key factors driving the classification. This aids in feature selection, understanding the dataset, and improving model performance by focusing on influential features.



Perfect classification performance is indicated by an AUC (Area Under the Curve) value of 1, meaning the model achieves a true positive rate of 1 (100%) and a false positive rate of 0 (0%). This implies that the model accurately distinguishes between positive and negative examples.

Linear Discriminant Analysis

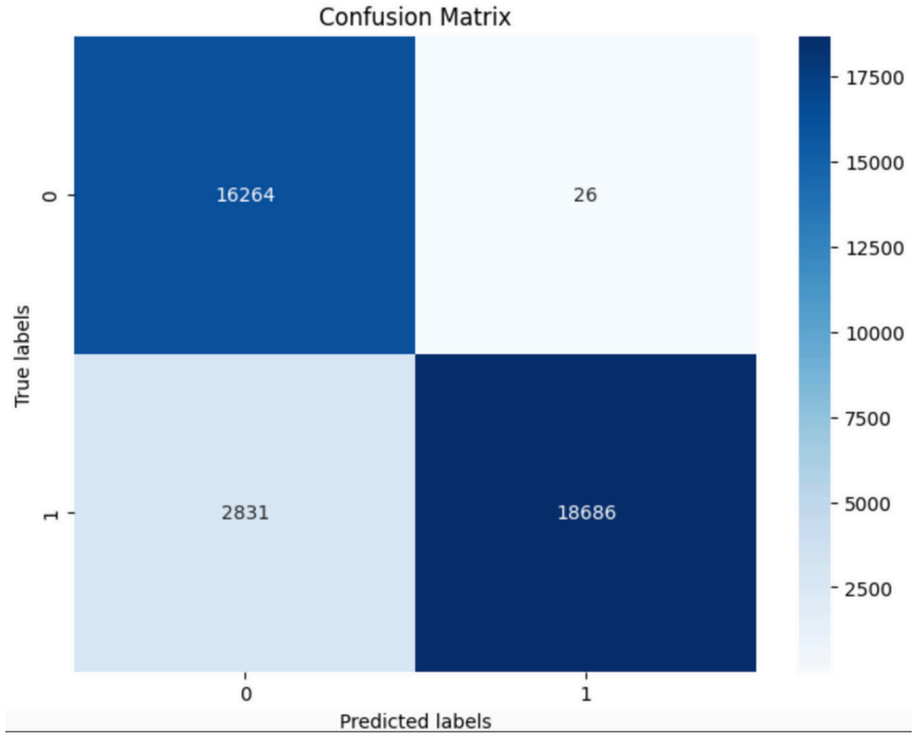
Accuracy: 92.3%

Precision: 93.4% **Recall: 92.3%** **F1: 92.3%**

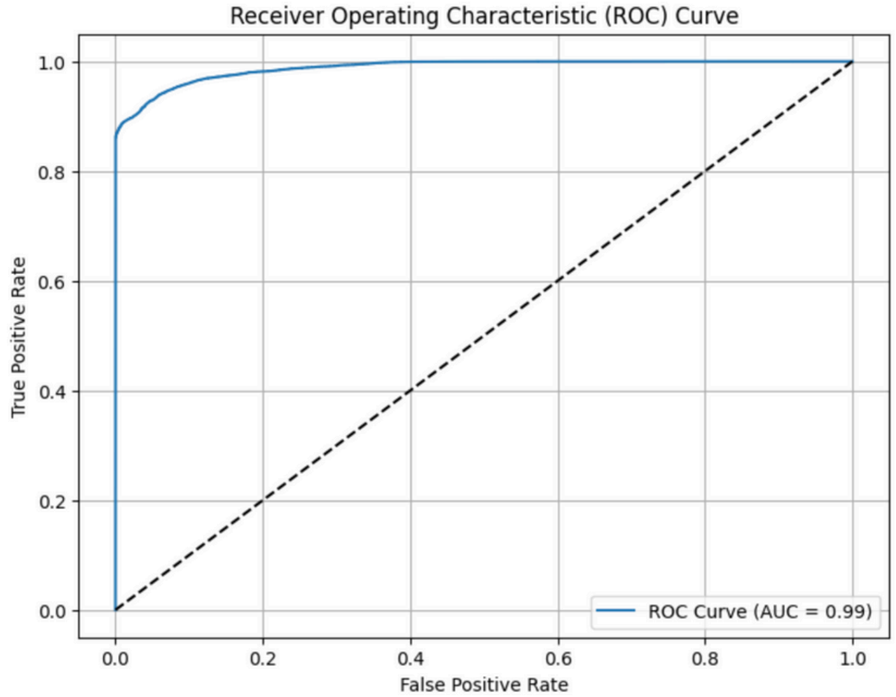
- Linear Discriminant Analysis (LDA) is a technique in statistics and machine learning that helps find a straight line or plane (linear combination) in the data that best separates different groups. It's like drawing a line between different groups of points on a graph to classify them accurately.
- Linear Discriminant Analysis (LDA) is used for cab-type categorization.
- The model fitted to the training set and predictions made on the test set.
- Grid search was used for hyperparameter tuning, achieving the best score of **92.29%.**
- Optimized LDA model fitted to training data.
- Achieved accuracy of **92.33%**, precision of **93.46%,** recall of **92.33%**, and F1-score of **92.36%**.

Linear Discriminant Analysis cont'd

Visualisations



The classification output from our model is shown in this confusion matrix. Correct predictions are represented by diagonal values, and incorrect classifications are indicated by off-diagonal values.



The model effectively distinguishes between the positive and negative classes, as evidenced by the AUC (Area Under the Curve) value of 0.99. Since 1 denotes perfect classification, an AUC value that is closer to 1 indicates better discrimination power.

Recommendation

1 Price Prediction Model

Decision Tree Regressor

- **Real-Time Adjustment:** Implement a system that adjusts ride prices in real time based on the predicted prices. This could involve increasing prices during peak demand times and lowering them when demand is low.

- **Market Analysis:** Regularly analyze market trends and competitor pricing strategies to ensure your pricing remains competitive.

2 Cab Type Classifier Model

Random Forest Classifier

- **Competitive Advantage:** Understanding the factors that influence cab type choice can provide a competitive edge, allowing the business to adapt and stay ahead in the market.

- **Customer Retention:** By addressing the factors that matter to customers, the business can improve customer satisfaction and loyalty, leading to higher retention rates.

Conclusion

In our analysis, we found that the Decision Tree Regressor and Random Forest Classifier were the best models for predicting ride prices and cab types respectively. These models provided us with valuable insights that can be leveraged to enhance our cab service business.

Implementing a Real-Time Adjustment system based on the predicted prices can help us manage supply and demand more efficiently. This could involve increasing prices during peak demand times and lowering them when demand is low.

Regular Market Analysis can ensure our pricing remains competitive. By analyzing market trends and competitor pricing strategies, we can make informed decisions about our own pricing strategies.

Understanding the factors that influence cab type choice can provide us with a Competitive Advantage. This allows our business to adapt and stay ahead in the market.

By addressing the factors that matter to customers, we can improve customer satisfaction and loyalty, leading to higher Customer Retention rates.

References

Uber and Lyft Dataset Boston, MA. (n.d.).Kaggle.
https://www.kaggle.com/datasets/brllrb/uber-and-lyft-dataset-boston-ma

Thank You