Name: Hussain Ibrahim Almakrami ID: 201849180

ICS 474: Big Data Analytics

Project Visualization Document

Name: Hussain Ibrahim Almakrami ID: 201849180

1. Project Overview

Q: What is the goal of the project?

Answer:

The goal of the project is to analyze Uber trip data to uncover patterns and build a machine learning model to classify trips into distance categories (Short, Medium, Long). The analysis includes data preprocessing, visualization, and modeling using Python and Jupyter Notebook.

2. Dataset Overview

Q: What is the source and context of your dataset?

Answer:

The dataset consists of Uber trip logs, providing details like trip distance, purpose, and type (Business/Personal). The analysis aims to understand patterns in trip purposes and classify trips based on distance categories.

Q: What are the features in the dataset?

Answer:

Start Date: The time when the trip started.

End Date: The time when the trip ended.

Category: Indicates whether the trip was for Business or Personal purposes.

Start: Location where the trip began.

Stop: Location where the trip ended.

Miles: Distance traveled during the trip.

Purpose: The reason for the trip (e.g., Meeting, Errands).

3. Data Understanding and Exploration

Q1. What is the dataset's purpose?

Answer:

The dataset is designed to provide insights into Uber trips and classify trips into Short, Medium, or Long distance categories.

Name: Hussain Ibrahim Almakrami ID: 201849180

Q2. What features are present in the dataset?

Answer:

Start Date and End Date (object → converted to datetime).

Category (encoded for modeling).

Start and Stop (textual locations).

Miles (numeric, key predictor for classification).

Purpose (encoded for insights and modeling).

Q3. What is the structure of the dataset?

Answer:

The dataset contains 1156 rows and 7 columns, representing individual trip details.

Q4. Are there missing values or duplicates in the dataset?

Answer:

Missing Values: Found in the Purpose column. Rows with missing values were dropped.

Duplicates: No duplicate rows were detected.

Q5. What is the statistical summary of the dataset?

Answer:

Mean Miles: 11.2

Median Miles: 6.4

Max Miles: 310.3

Standard Deviation: 22.98

Q6. How are the features distributed?

Answer:

Most trips are under 20 miles, showing a skewed distribution.

Q7. What are the relationships between features and the target variable?

Answer:

Miles is strongly correlated with the target variable (Distance Category), making it a critical feature.

Q8. Are there any outliers?

Answer:

Outliers in Miles were detected but retained to ensure the model can handle extreme values.

4. Data Preprocessing

Q9. How were missing data handled?

Answer:

Rows with missing values in the Purpose column were dropped.

Q10. How were categorical variables encoded?

Answer:

Category: Encoded using Label Encoding.

Purpose: Encoded using Label Encoding.

Q11. Was feature scaling applied?

Answer:

Miles was normalized using Min-Max Scaling to bring values into a range of 0 to 1.

Q12. How were features selected for the model?

Answer:

All features were retained for modeling based on their importance scores obtained from the Random Forest model.

5. Modeling

Q13. Which algorithm was selected and why?

Answer:

The Random Forest Classifier was chosen for its ability to handle categorical and numerical data effectively, and for its robustness to overfitting.

Q14. How was the data split into training and testing sets?

Answer:

The dataset was split into 80% training and 20% testing sets using train_test_split.

Q15. How was the model trained?

Answer:

The Random Forest Classifier was trained with default hyperparameters. GridSearchCV was later used for hyperparameter tuning.

Q16. How was the model evaluated?

Answer:

Metrics like accuracy, precision, recall, and F1-score were used to evaluate model performance. The model achieved high accuracy.

Q17. What was the performance on the testing set?

Answer:

The model achieved a classification accuracy of 100% on the testing set, as shown in the classification report.

Q18. How was the model improved?

Answer:

Hyperparameter tuning was performed using GridSearchCV to identify the best combination of parameters.

Q19. How was the model validated?

Answer:

Cross-validation was performed with 5 folds to ensure the model generalizes well to unseen data.

Name: Hussain Ibrahim Almakrami ID: 201849180

Q20. Which model was selected and why?

Answer:

The Random Forest Classifier was selected as the final model for its simplicity, interpretability, and excellent performance metrics.