Assignment # 1

Name: Muhammad Awais

Roll Number: i222390

Project Overview

This project explores the application of machine learning techniques to a large real-world dataset. The dataset, sourced from NYC Open Data, consists of the 2015 Yellow Taxi Trip Data, containing over one million records. The project follows a structured machine learning pipeline, including data collection, exploratory data analysis, preprocessing, feature engineering, model selection, hyperparameter tuning, and final evaluation. Additionally, extra credit components such as a deep learning model and a deployment snippet have been incorporated to enhance the overall solution.

Workflow Summary

Dataset Selection & Understanding

A publicly available dataset containing over one million records was chosen. The dataset was thoroughly examined to understand its domain, the target variable, key features, and overall record structure.

Exploratory Data Analysis (EDA)

Various data visualizations, including histograms, scatter plots, and correlation matrices, were employed to explore feature distributions and relationships. Missing values and outliers were identified and handled accordingly.

Data Preprocessing & Feature Engineering

To ensure high-quality inputs for modeling, missing values were imputed, categorical features were one-hot encoded, and numerical features were standardized. A streamlined preprocessing pipeline was built for efficiency.

Model Selection & Training

Several regression models, including Linear Regression, Decision Tree, Random Forest, and

Gradient Boosting, were trained. Cross-validation techniques were used to assess and compare their performance.

Hyperparameter Tuning

To enhance model performance, GridSearchCV was applied for hyperparameter optimization, ensuring the best model parameters were selected.

Final Model Evaluation

The optimized model was tested on a holdout dataset, with key performance metrics calculated to validate its effectiveness.

Extra Credit Enhancements

To extend the project, a deep learning model using Keras/TensorFlow was implemented. Furthermore, a deployment snippet was developed using Streamlit, showcasing real-world application potential.

Dataset Details

The dataset used in this project is the 2015 Yellow Taxi Trip Data, sourced from NYC Open Data. It falls under the transportation domain and is used to solve a regression problem. With over one million records, it provides valuable insights into taxi trips, including ride durations, distances, fare amounts, and other attributes essential for predictive modeling.

Conclusion

This project effectively demonstrates the complete machine learning workflow, from raw data analysis to advanced model implementation. Through careful data preprocessing, thorough exploratory analysis, and robust model selection, an optimal predictive model was built. Further enhancements, including a deep learning model and deployment setup, highlight the project's real-world applicability. Overall, this assignment not only meets academic requirements but also serves as a strong foundation for practical machine learning solutions.