Documentation Outline for "AI-Based Used Car Price Prediction"

Submitted By: Muhammad Mehdi Mesam

Submitted To: Sir Rasikh Ali

Roll No: SU92-BSAIM-F23-**004**

Section: BSAI(**3A**)

Introduction

Context

Machine learning has become a key tool for automating decision-making in many industries. In the rapidly growing used car market, accurately predicting car prices is essential. This helps buyers, sellers, and online marketplaces build trust and make informed decisions.

Relevance

A reliable price prediction model can guide users in assessing a car's value based on key details like mileage, age, and engine specifications. This ensures better financial planning and transparent transactions for all parties involved.

Highlight

The goal of this project is to create a machine learning model that efficiently predicts the price of used cars based on various features, offering accuracy and reliability.

Objective

Main Goal

To create a reliable machine learning system that accurately predicts used car prices using different regression models.

Specific Objectives

- Clean and prepare the dataset for analysis.
- Identify the key factors that influence car prices.
- Test and compare multiple machine learning models to find the most effective one.
- Adjust model settings (hyperparameters) to improve accuracy and performance.

Dataset Description

Source

The dataset used in this project comes from [insert source, e.g., Kaggle, a proprietary database, etc.].

Features

The dataset includes the following key columns:

- **Make**: The car's manufacturer (e.g., Toyota, Ford).
- **Year**: The year the car was manufactured.
- **Mileage**: The total distance the car has been driven (in kilometers).
- **Engine**: Engine details, such as displacement in cubic centimeters (cc).
- **Price**: The car's price (target variable, in USD).

Key Characteristics

- **Size**: The dataset contains approximately [e.g., 10,000 rows and 15 columns].
- **Challenges**: Includes issues such as missing values and outliers that required cleaning during preprocessing.

Project Workflow

Steps Overview

The project followed these main steps:

1. **Data Collection and Exploration**: Gathering and understanding the dataset.
2. **Data Preprocessing and Cleaning**: Fixing missing values, removing outliers, and preparing the data.
3. **Feature Engineering**: Selecting and creating important features for the model.
4. **Model Selection and Training**: Testing different models and training them on the data.
5. **Hyperparameter Tuning**: Optimizing model settings for better performance.
6. **Evaluation and Results**: Measuring model accuracy and comparing performance.

Methodology

Data Preprocessing

- **Handling Missing Values**: Filled or removed missing data to ensure a complete dataset.
- **Outlier Detection**: Identified and managed outliers using methods like the IQR rule.
- **Encoding Categorical Data**: Converted categorical features into numerical ones using techniques like one-hot encoding.

Feature Engineering

- **Feature Scaling**: Standardized numerical features using StandardScaler for better model performance.
- **Feature Selection**: Used Lasso regression to identify the most important features.

Modeling Approach

- **Algorithms**: Tested multiple models including Linear Regression, Random Forest, Gradient Boosting, and LightGBM.
- **Data Splitting**: Divided the data into training and testing sets to evaluate model performance.
- **Cross-Validation**: Applied cross-validation to minimize overfitting and ensure model reliability.

Technical Overview

Tools and Libraries

- Python, NumPy, Pandas, Matplotlib, Seaborn, Scikit-learn, LightGBM.

Algorithms

- **Linear Regression**: Used as a baseline model.
- **Decision Tree & Random Forest**: Captured non-linear patterns.
- **Gradient Boosting (LightGBM)**: Provided detailed and accurate predictions.

Tuning

- Optimized model performance with GridSearchCV for Random Forest and LightGBM.

CONCLUSION

Summary:

- Briefly restate the problem and approach.
- Key finding: *"LightGBM performed best with an R^2 score of 0.92 and RMSE of 1,500."*

Impact:

- Discuss how the model can be used in real life, e.g., *"This model can help car dealerships estimate prices more accurately."*

Future Work

- *Improvements*:
  - Collecting more diverse datasets for better generalization.
  - Incorporating additional features like fuel type, location, or previous owners.
  - Exploring deep learning techniques (e.g., neural networks).
- *Deployment*:
  - Building a web-based or mobile app to allow users to input features and receive price predictions.