PROJECT AND DATA MANAGEMENT PLAN

Name: Austine Igho Efenarua

Student ID: 23096669

USING MACHINE LEARNING TO PREDICT USED CAR PRICES

PROJECT OVERVIEW

1. Summary of the Project Topic and Background:

This project aims to develop a machine learning model to accurately predict the prices of used cars based on various vehicle features such as brand, model, mileage, year, engine type, fuel type, and transmission. The predictions will help car buyers, sellers, and dealerships make informed decisions. The project will involve data preprocessing, exploratory data analysis, feature engineering, model training, evaluation, and result interpretation.

Several studies have explored the application of machine learning techniques to predict used car prices, aiming to improve pricing accuracy and decision-making in the automotive resale market. Gao (2024) evaluated Multiple Linear Regression and Random Forest for predicting second-hand car prices. The study showed that while linear regression helps identify basic pricing factors, Random Forest offered better accuracy by capturing non-linear patterns and variable interactions. Najib (2023) created a comprehensive used car dataset on Kaggle, including features like brand, mileage, fuel type, and transmission. This dataset has become a key resource for testing machine learning models.

O.Abhila Anju et al. (2024) compared several regression and ensemble methods, including Random Forest and XGBoost, and found that ensemble models, especially XGBoost, delivered the best prediction performance. Zhu (2023) explored a range of algorithms such as Linear Regression, Decision Trees, Support Vector Machines (SVM), and Neural Networks, concluding that ensemble and deep learning models generally yield better results than traditional statistical approaches.

2. Research Questions

i. **Exploratory Data Insight:** What insights, trends or patterns can be uncovered from Exploratory Data Analysis (EDA) through used car price prediction?
ii. **Model Performance Comparison:** Which machine learning model gives the most accurate price predictions?

3. Project Objectives

i. To perform exploratory data analysis (EDA) to understand trends and patterns in used car pricing.
ii. To develop and compare three machine learning models, such as Linear Regression, Random Forest and XGBoost).
iii. To evaluate and compare the different Machine learning models for accuracy using standard metrics.
iv. To identify features that most influence car price

PROJECT PLAN - Timeline

Task	Duration	Date Range
Literature Review	1 week	24 May – 31 May 2025
Data Collection	1 week	1 Jun – 7 Jun 2025
Exploratory Data Analysis	1 week	8 Jun – 14 Jun 2025
Feature Engineering	1 week	15 Jun – 21 Jun 2025
Model Building & Tuning	2 weeks	22 Jun – 5 Jul 2025
Model Evaluation	1 week	6 Jul – 12 Jul 2025
Report writing and review	2.5 weeks	13 Jul – 1 Aug 2025

DATA MANAGEMENT PLAN

1. **Dataset Link:** https://www.kaggle.com/datasets/taeefnajib/used-car-price-prediction-dataset?resource=download

2. **Data Collection:**

Data was sourced from a Kaggle dataset. It contains 4,009 records of used car listings, including both categorical variables (like car make, model, fuel type) and continuous variables (like year, mileage, engine size, price).

3. **Metadata:**

i. **Format:** CSV (Comma-Separated Values) file
ii. **Number of Records**: 4,009
iii. **Size**: 2.92 MB
iv. **Type of Data:** Mixed (continuous and categorical data)
v. **Origin:** Public dataset from Kaggle by Ta'eef Najib

4. **Document Control**

i. **GitHub Address**: *https://github.com/megaigho*

ii. **Commit Frequency**: Weekly commits starting 1St June. Tags such as:
"Data_Collection_2025-06-01"
"EDA_Complete_2025-06-14"

iii. **File Naming Convention**:
CarPricePrediction_TaskName_2025-06-01.

iv. **Version Control**:
Branches used for stages (e.g., data-cleaning, model-training)
Merges into main after completion and review of each phase

5. ReadMe File (Contents)

i. **Project Overview**: Summary of the project and its objectives
ii. **Instructions**: Environment setup, required libraries, and how to run the code
iii. **Dataset Description**: Variables, data types, source, and structure
iv. **Model Information**: Algorithms used and performance summaries
v. **Result:** Key findings, EDA insights and best model performance
vi. **Contact Information**: Email or GitHub for queries

6. Security and Storage

Project files, including code and data will be securely stored on Github, with weekly backups on Google drive, share with supervisors and evaluators via Github with appropriate permission.

7. Ethical Requirements

i. **GDPR Compliance**: The dataset is anonymized, no Personal Identifiable Information (PII) and complies with General Data Protection Regulations (GDPR).
ii. **UH Ethical Policies**: The project aligns with University of Hertfordshire's data ethics guidelines.
iii. **Permission to Use the Data**: Dataset is publicly available for educational and research purposes.
iv. **Ethical Data Collection**: Data sourced from reputable, open platforms with proper citation.

8. References

Gao, J. (2024). Second-hand car price prediction based on multiple linear regression and random forest. *Theoretical and Natural Science*, 52(1), pp.31–40. doi:https://doi.org/10.54254/2753-8818/52/2024ch0105.

Najib, T. (2023). *Used Car Price Prediction Dataset*. [online] Kaggle.com. Available at: https://www.kaggle.com/datasets/taeefnajib/used-car-price-prediction-dataset?resource=download [Accessed 5 May 2025].

O.Abhila Anju, M.Yoga, M. Sri Kruthika, M.Manikandan, K.S.Aswin and S.Kishore (2024). Predicting Used Car Prices Using Machine Learning: A Comparative Analysis of Regression and Ensemble Models. *International Research Journal on Advanced Engineering Hub (IRJAEH)*, 2(12), pp.2796–2801. doi:https://doi.org/10.47392/irjaeh.2024.0386.

Zhu, Y. (2023). Prediction of the price of used cars based on machine learning algorithms. *Applied and computational engineering*, 6(1), pp.671–677. doi:https://doi.org/10.54254/2755-2721/6/20230917.