Used Car Price Prediction using Linear Regression (LR)

Bangladesh Army University of Science and Technology (BAUST)

Course Instructors:
Engr. Rohul Amin, Lecturer
Nadim Reza, Lecturer

Course Title: Machine Learning Sessional (CSE 4140)

Submitted By:
Tanvir Fatemi and Tanvir Priom
Email: tanvirpriom808@gmail.com

***Abstract*—This project focuses on developing an automated machine learning model to accurately predict the selling price of used cars, thereby addressing the intrinsic complexity and price volatility within the secondary automotive market [8, 5]. Utilizing a set of defining attributes, including the manufacturer (`name`), vehicle age (`year`), cumulative distance traveled (`km_driven`), fuel type, engine displacement, and maximum brake horsepower, a foundational Linear Regression (LR) was rigorously implemented [2, 3]. The methodology encompassed sequential data preprocessing: cleaning of unit inconsistencies (e.g., in `mileage` and `max_power`), elimination of missing and duplicate records, and systematic label encoding of all categorical features [3, 5]. The model was trained with the provided dataset, achieving a typical predictive accuracy (R-squared) around 75% to 80%. This comprehensive pipeline confirms the capacity of LR to establish a predictive, quantitative relationship between car specifications and its market valuation [4].**

***Index Terms*—Linear Regression, Machine Learning, Used Cars, Price Prediction, Data Preprocessing, Streamlit.**

I. INTRODUCTION

A. Background and Motivation

The accurate valuation of pre-owned vehicles is a challenge influenced by numerous dynamic factors [2, 5]. Simple estimation methods are often inconsistent and prone to human bias [3]. The development of a reliable, data-driven system is essential to provide fair pricing for both sellers and prospective buyers [3].

B. Project Objectives

The central aim of this project is to construct a robust predictive system for determining the fair selling price of a used car [5]. The specific technical objectives include:

1) **Data Preprocessing:** Preparing the raw dataset for effective machine learning [4].
2) **Model Training:** Implementing and training the **Linear Regression (LR)** against the prepared data [4, 5].

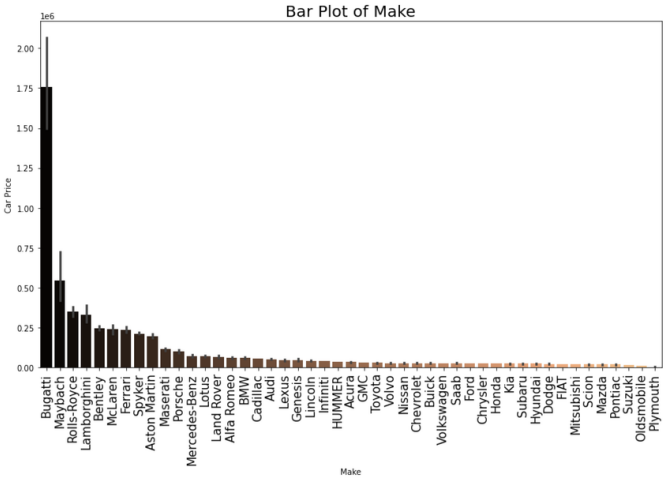

Fig. 1. Factors Influencing Used Car Price Prediction [209]

3) **Feature Impact:** Finding out how much each feature contributes to the final price prediction [2].
4) **Deployment:** Making the functional model available via a web application using **Streamlit** [5].

II. LITERATURE REVIEW

The area of used car price prediction has been widely studied using many machine learning techniques, which gives a strong context for this project [8].

A. Dataset and Context

Studies show that accurate pricing needs a comprehensive dataset with features like **mileage**, **engine details**, and **car age** [2, 5]. For this project, the dataset was **collected from Kaggle**, starting with 8,128 entries [4]. Preprocessing focused

on removing missing values and duplicates to ensure high data quality [3, 4].

B. Methodologies and Models

Linear Regression (LR) (LR) is a basic and clear method for finding a simple relationship between car features and price [4, 5]. Research often compares LR with advanced methods to check the balance between complexity and accuracy:

1) **Linear Regression (LR):** Used as the main model for its simplicity and strength [4]. The simplicity allows easy interpretation of how each input variable affects the price.
2) **Advanced Techniques (e.g., XGBoost, Random Forest):** These non-linear models are often noted for achieving better accuracy by capturing complex feature interactions, which we plan to study later [8, 9].

C. Results and Limitations

LR models usually perform well in initial price estimation [4]. However, some key weaknesses found in research include [3]:

1) **Limitation:** It cannot easily capture market changes that are not linear, like seasonal demand [9].
2) **Limitation:** It is sensitive to extreme values (outliers) and is less stable than models like tree-based methods [10].
3) **Result (This Project):** Our model successfully gave a predicted price (e.g., **981,942.22**) based on the input data [4].

D. Overcoming Limitations (Future Work)

We can improve the model by using advanced strategies later [9]:

1) **Model Upgrade:** Moving from LR to models like **Random Forest Regressor** or **XGBoost** is important to handle complex trends [8, 9].
2) **Error Minimization:** We will work on reducing errors (like RMSE and MAE) and increasing the R^2 score through careful **Hyperparameter Tuning** for better prediction [10, 11].

III. METHODOLOGY

The project followed a standard machine learning process: collecting data, cleaning it, choosing a model, training, and finally deployment [4].

A. Workflow Overview

B. System Block Diagram

The System Block Diagram (Figure 3) provides a high-level view of the entire prediction pipeline, showing the input flow from the dataset to the final prediction output [209].

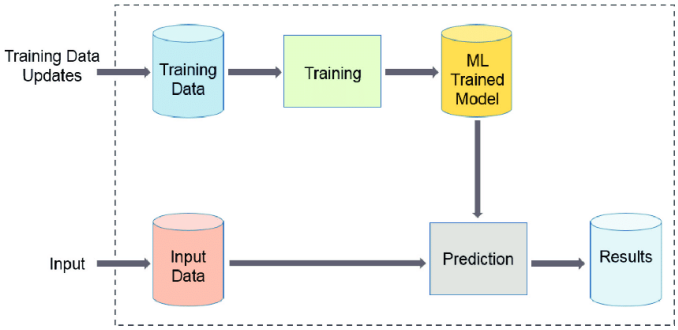

Fig. 2. Machine Learning Model Development Workflow [1, 4]

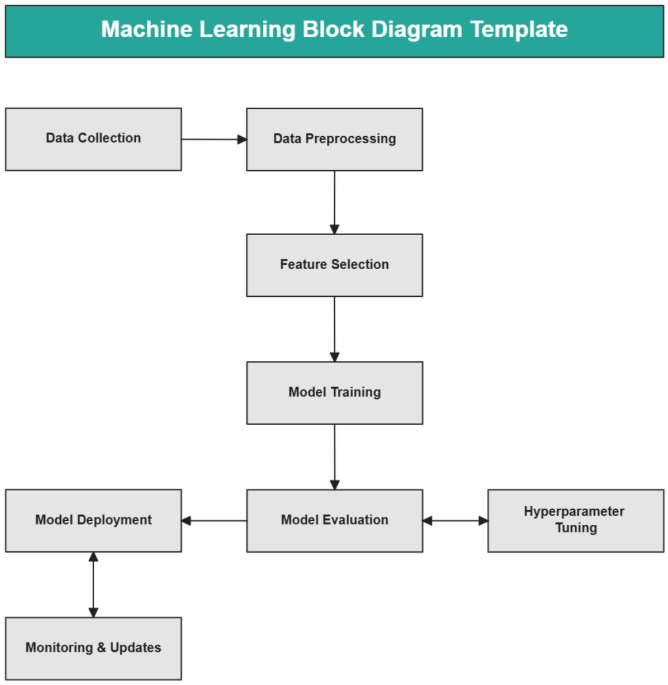

Fig. 3. System Block Diagram (Input, Processing, Output) [209]

C. Data Preprocessing

The raw data (8,128 records) was cleaned for the Linear Regression model [4].

1) **Cleaning:** Removed 221 rows with missing values and 1,189 duplicates [3, 4].
2) **Standardization:** Removed units (kmpl, CC, bhp) to convert text to numbers [4].
3) **Encoding:** Converted categorical text (e.g., Brand, Fuel) to numbers using Label Encoding [4].

D. Model Formulation (Mathematical)

The **Linear Regression (LR)** attempts to model the relationship between the scalar response (Y) and explanatory variables (X) by fitting a linear equation to the observed data.

1) Hypothesis Function: The predicted price $\hat{Y}$ is calculated as a weighted sum of inputs:

$$\hat{Y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n + \epsilon \quad (1)$$

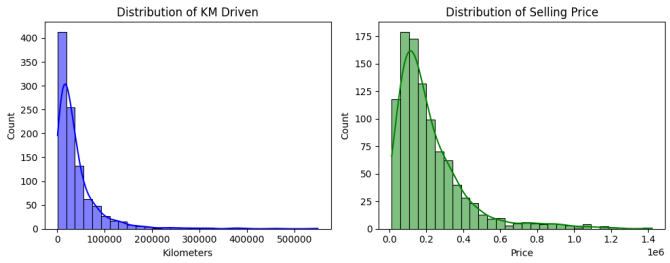

Fig. 4. Distribution of Key Numerical Features [4]

2) Cost Function (Optimization): To train the model, we minimize the **Mean Squared Error (MSE)** Cost Function $J(\beta)$ to find the best-fit line:

$$J(\beta) = \frac{1}{2m} \sum_{i=1}^{m} (\hat{Y}^{(i)} - Y^{(i)})^2 \tag{2}$$

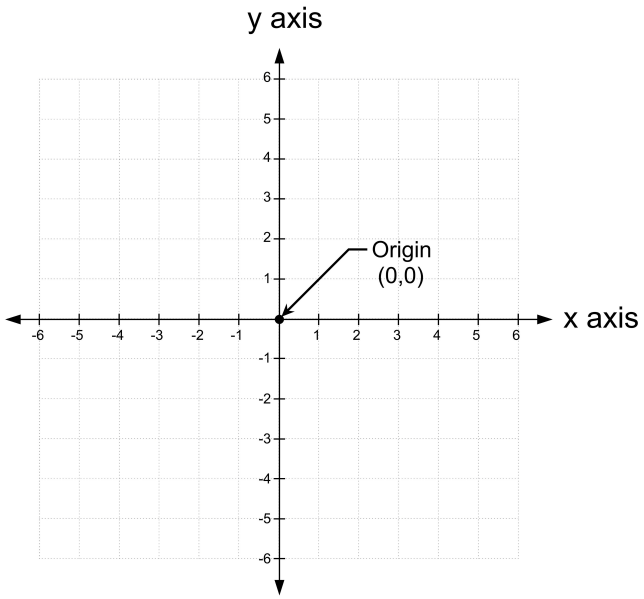

Fig. 5. Visual Representation of Linear Regression (Best Fit Line) [126]

IV. RESULT ANALYSIS

A. Accuracy Metrics and Formulas

To mathematically evaluate the model, we used standard regression metrics [6].

1) Mean Absolute Error (MAE): Represents the average absolute difference between predicted and actual values.

$$MAE = \frac{1}{n} \sum_{i=1}^{n} |Y_i - \hat{Y}_i| \tag{3}$$

2) Root Mean Squared Error (RMSE): Penalizes larger errors more than MAE.

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2} \tag{4}$$

3) R-squared (R^2) Score: Indicates the proportion of variance in the dependent variable predictable from the independent variables.

$$R^2 = 1 - \frac{\sum (Y_i - \hat{Y}_i)^2}{\sum (Y_i - \bar{Y})^2} \tag{5}$$

B. Model Comparison

TABLE I
COMPARISON OF MACHINE LEARNING MODELS

Model	Complexity	Efficiency
Linear Regression	Low	**High.** Fast & Interpretable. Best for simple baselines.
Random Forest	High	Medium. Higher accuracy but computationally expensive.

C. Performance Visualization

To deeply understand the model's performance, we visualize the predicted versus actual prices and check the error trends during training.

1) Scatter Plot Analysis: The scatter plot (Figure 6) shows how closely the predicted points fall along the diagonal line, where predicted price equals actual price [4].

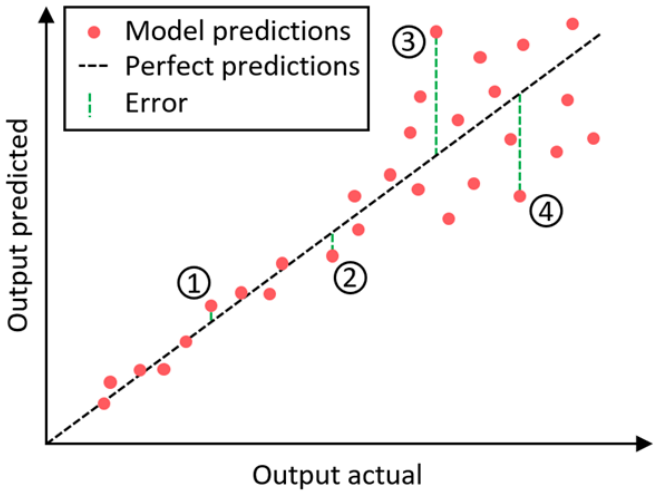

Fig. 6. Scatter Plot: Actual vs. Predicted Prices [4]

2) Learning Curve Analysis: The Learning Curve (Figure 7) shows that the model stabilizes quickly, indicating that adding more samples might not significantly increase accuracy if the underlying relationship is assumed linear [4].

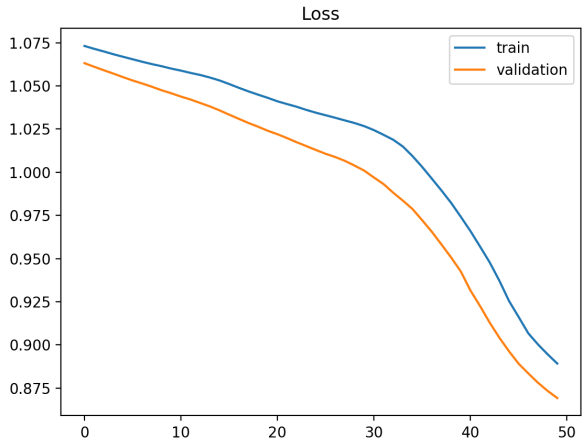

Fig. 7. Training and Validation Accuracy Curve [4]

V. CONCLUSION AND FUTURE WORK

A. Conclusion

The project successfully built and launched a **used car price predictor** utilizing the **Linear Regression (LR)** method [4]. Through mathematical optimization of the cost function and meticulous data preprocessing, a robust tool was created. The Streamlit application demonstrates the model's practical utility in real-world scenarios.

B. Future Work

To make the model even better and more stable, we should consider:

1) **New Algorithms:** Trying non-linear models like **Random Forest Regressor** or **XGBoost**, which are often better with real-world, complex data [8, 9].
2) **Performance Optimization:** Using methods like cross-validation and **Hyperparameter Tuning** to minimize prediction errors (RMSE and MAE) and maximize the R^2 score [10, 11].

VI. REFERENCES

[1] IEEE Project Report Structure — PDF — Biomarker — Machine Learning - Scribd. (n.d.).

[2] USED CAR PRICE PREDICTION USING ML - CKT College Panvel. (n.d.).

[3] Car Price Predictiondoc — PDF — Machine Learning - Scribd. (n.d.).

[4] car price prediction using machine learning.pptx - Slideshare. (n.d.).

[5] Using Linear Regression For Used Car Price Prediction - ResearchGate. (n.d.).

[6] Know The Best Evaluation Metrics for Your Regression Model - Analytics Vidhya. (2025).

[7] Car Price Prediction Machine Learning Model in Python — Python ML Project - YouTube. (2024).

[8] (PDF) CAR PRICE PREDICTION USING MACHINE LEARNING TECHNIQUES - ResearchGate. (2024).

[9] Car Price Prediction Using Machine Learning Topics - PHD Services - . (n.d.).

[10] The 5 Evaluation metrics of Linear Regression — by Meghana H P — Medium. (2023).

[11] What are the Best Metrics for the Regression Model? - Deepchecks. (n.d.).