University of Sri Jayewardenepura

Faculty of Computing

Project Proposal – June 2025

Car Price Prediction Using Machine Learning

Group 11

FC211034 - N.D. Samararathne Kodikara

FC211013 - N.W.V. Tharindu Pabasara

FC211025 - W.M.M.C.B. Wijesundara

1. Project Overview

Determining the fair market value of used vehicles is essential for facilitating informed transactions between buyers and sellers in the automotive sector. The proposed project utilizes a publicly available dataset of used car listings to examine the key factors that influence vehicle pricing. The project aims to develop a predictive model capable of estimating car prices based on relevant vehicle attributes. The findings will contribute to more informed decision-making and pricing strategies in the used car market.

2. Dataset Description

2.1 Dataset Overview

The dataset used in this project is the Car Price Prediction Challenge Dataset, publicly available on Kaggle: https://www.kaggle.com/datasets/deepcontractor/car-price-prediction-challenge. Each instance in the dataset represents a single used car listing, containing various attributes related to the vehicle's specifications, condition, and usage, along with the associated selling price as the target variable**.**

2.2 Dataset Structure

- The dataset comprises 19,237 used car listings, each representing a single vehicle.
- All instances are of a single type: individual car profiles with feature attributes and a continuous target variable (price)**.**
- Each instance consists of 17 features and 1 target variable.

2.3 Details of categorical and numerical variables

In this section, the dataset's variables are categorized as categorical or numerical based on their type. The variable names are presented exactly as in the original dataset, with brief descriptions and possible values provided where applicable.

2.3.1 Numerical variables

1. **ID** – A unique identifier assigned to each car listing.

2. **Price** – The target variable representing the vehicle's market price. (in USD)
3. **Prod. Year** – Year the vehicle was manufactured. Treated as a numerical variable, as newer cars typically have higher prices, showing a continuous and linear relationship.
4. **Airbag** – The total number of airbags equipped in the vehicle.
5. **Levy** – Indicates whether there is a legal claim or lien on the vehicle due to unpaid taxes or debts. This means the car may be subject to seizure or restrictions until the outstanding obligations are settled. (in USD)

2.3.2 Categorical variables

6. **Manufacturer** – Who originally built the car (ex: Honda, Ford).
7. **Model** – Model of the car (ex: RX 450, CHR, Elantra, etc.).
8. **Category** – Classifies type of the vehicle (ex: jeep, sedan, micro bus).
9. **Cylinder** – Refers to the number of engine cylinders. Although recorded as a float, it represents discrete categories (e.g., 3, 4, 6, 8 cylinders) and is therefore typically treated as a categorical variable in analysis.
10. **Leather interior** – Indicates if the car has leather seats (Yes or No).
11. **Fuel type** – Specifies the vehicle fuel type (Petrol, Diesel, Hybrid, CNG, etc.)**.**
12. **Engine volume** – Describes the engine size in liters, sometimes including additional descriptors like "Turbo" (e.g., 2.0, 2.2 Turbo).
13. **Mileage** – Number of miles covered by the car (in km).
14. **Gear box type** – What type is the gear box (Automatic, Manual, Variator, etc.)
15. **Drive wheel** – Specifies which wheels are powered (Front, Rear, 4X4).
16. **Doors** – Number of doors on the vehicle. (The categories are 2-3, 4-5, 5>).
17. **Wheel** – Indicates whether the vehicle has a left-hand drive or right-hand drive.
18. **Color** – Color of the car.

3. Objectives of the project

3.1 Main Objective

To develop a predictive model for used car prices to help buyers and sellers make informed decisions.

3.2 Specific Objectives

- To analyze the used car dataset for key trends and patterns affecting prices.
- To perform data cleaning and preprocessing to ensure data quality and consistency.
- To conduct feature engineering by transforming, extracting, or creating new features to enhance model performance.
- To assess how features such as mileage fuel type impact vehicle value.
- To develop and evaluate machine learning models for accurate price prediction.
- To optimize model performance through feature selection and hyperparameter tuning.
- To provide insights that support buyers and sellers in making data-driven pricing decisions

4. Problems Addressed by the Predictive Model

The used car market commonly experiences challenges related to price uncertainty and information asymmetry between buyers and sellers. Buyers often find it difficult to determine whether a vehicle's asking price reflects its true market value, which can lead to overpaying or missed opportunities. Conversely, sellers may struggle to set competitive and realistic prices due to lack of comprehensive market data, resulting in prolonged sales or undervaluation of their vehicles.

The proposed predictive model aims to address several key problems faced by stakeholders:

- Assisting buyers in accurately estimating the fair market value of used cars based on key vehicle attributes.
- Helping sellers determine competitive and realistic selling prices to optimize sales outcomes.
- Reducing information gaps by providing transparent, data-driven pricing estimates that benefit both parties.

By solving these problems, the predictive model seeks to enhance trust and efficiency in the used car marketplace.

5. Summery

The proposed project aims to develop a machine learning-based predictive model to estimate the fair market value of used cars. The study examines important vehicle properties of both numerical and categorical variables using a publicly accessible Kaggle dataset with over 19,000 car listings. These variables will be cleaned, engineered, and analyzed to create an accurate and dependable model. The project will address common challenges in the used car market, such as pricing uncertainty and information asymmetry by providing data-driven insights. The ultimate goal is to assist both buyers and sellers in making informed decisions and improve transparency and efficiency in pricing practices.