

Title: A Machine Learning Approach to Predict the Features and Price of Used Cars

Team members: Farah Ahmed, Priyanka Singh, Megan Tiet, Musa Mustapha

Group Name: PredictiveWizards

Our group project focuses on creating and evaluating a machine learning model to predict the prices of secondhand cars. To achieve this, we've gathered a comprehensive dataset tailored for tasks like price prediction, data analysis, and model development.

It's important to understand that this dataset captures a specific moment in time, giving us a snapshot of used car listings available at that time. While it's a valuable resource for our goal, there may be missing information and some small inconsistencies. To make sure our dataset is suitable for our research and modeling work, we're starting our project by careful pre-processing followed by data analysis and visualization.

We believe that this dataset can benefit the wider data science community by advancing our understanding of automotive pricing models. Our primary goal is to use machine learning techniques to build a model that can estimate the prices of older cars accurately. In doing so, we hope to provide practical insights and tools that can assist both buyers and sellers in making better-informed decisions in the used car market, ultimately enhancing our collective knowledge of this industry.

The dataset we will be using for the project is a public dataset available in Kaggle.

Data Source: <https://www.kaggle.com/datasets/tugberkkaran/used-car-listings-features-and-prices-carscom>

Original Data Source: <https://www.cars.com/>

The dataset contains a variety of attributes, encompassing details such as the car's brand, model, production year, mileage, engine specifications, fuel type, and numerous other factors. The dataset was obtained through web scraping from cars.com, a well-recognized online platform dedicated to facilitating the trade of pre-owned vehicles. The web scraping process involved the extraction of comprehensive information from a diverse range of car listings, covering essential features like the car's brand, model, production year, mileage, engine specifications, and more. It's important to highlight that this dataset represents a specific moment in time, providing us with a snapshot of the used car listings available on the platform during that particular period.

Q1. How effectively we can develop an accurate machine learning model for predicting used car prices based on the available features, and what is the optimal algorithm and feature selection approach to achieve this?

Q2. Which features have the most significant impact on the price of a used car? Can we identify the key features that influence car pricing decisions the most?

Q3. Can data preprocessing techniques, such as transformation, standardization, or normalization, enhance the prediction performance of machine learning models? What evaluation metrics should be used to assess the performance of the price prediction models?

Q4. To what extent do the models built on this dataset generalize to new, unseen data?