Predicting Prices for Used Cars

Julios Fotiou
Andreas Hadjoullis
Computer Science
University of Cyprus

I. Goals

The automotive market has a wide range of car prices depending on factors such as brand, age, mileage, and vehicle type. Predicting car prices accurately is valuable for both buyers and sellers. This project aims to build a predictive model using structured automotive data.

Our dataset is derived from Kaggle in which we have a comprehensive collection of valuable data about used cars, and provides insight into how the cars are being sold, what price they are being sold for, and all the details about their condition.

This project focuses on predicting car prices using machine learning models trained on our real-world dataset. To improve prediction quality, extreme values outside the practical range [1,000 – 200,000] are excluded. Various regression models are evaluated using standardized preprocessing and feature selection pipelines. We have also tested various classification algorithms on our dataset, after splitting our dataset into bins, decided by their price.

The main goals of the project are:

- Understand our dataset.
- Clean and preprocess raw car listing data.
- Select relevant features and remove noise.
- Train and compare multiple regression and classification models.
- Evaluate model performance using robust statistical metrics.

II. Approach

A. Exploratory Data Analysis (EDA)

First of all we need to understand our dataset. Before doing that however, we get rid of all rows/instances in which the price does not belong in the range of [1,000 – 200,000]. Our rezoning behind this choice, is that we only find our model practical for values that lie in said range, and all other rows would make it harder for the models to predict accurately prices in our desired target range. After removing said rows, we end up with 288,023 rows instead of the original 371,528.

We start by checking for features that have little to no deviation. For example 'offerType' only has two unique values, and one of them appears four times. Meaning this feature has no impact on the target of our dataset. The same applies for 'seller' and 'nrOfPictures'. So we remove the aforementioned features after also removing the insignificantly few rows that had a different value. We are now only working with 17 features/columns.

Some features, such as, 'vehicleType', 'gearbox', 'model', 'fuelType' and 'notRepairedDamage', have a significant number of missing values into the tens of thousands.

name	model	brand
Golf_3_1.6	golf	volkswagen
A5_Sportback_2.7_Tdi	NaN	audi
Jeep_Grand_Cherokee_"Overland"	grand	jeep
GOLF_4_1_4__3TÜRER	golf	volkswagen
Skoda_Fabia_1.4_TDI_PD_Classic	fabia	skoda
BMW_316i___e36_Limousine___Bastlerfahrzeug__Export	3er	bmw
Peugeot_206_CC_110_Platinum	2_reihe	peugeot
VW_Derby_Bj_80__Scheunenfund	andere	volkswagen
Ford_C___Max_Titanium_1_0_L_EcoBoost	c_max	ford
VW_Golf_4_5_tuerig_zu_verkaufen_mit_Anhaengerkupplung	golf	volkswagen

TABLE I: Sample of car name, model, and brand from dataset

1) Car Name: Our first difficulty comes when dealing with the feature 'name'. As shown from the Table I, the

III. Milestones

Break down the key stages or checkpoints in your project timeline. For example:

- Literature review
- Prototype implementation
- Testing and iteration
- Final deployment

IV. Experimental Setup

Describe the experimental environment:

- Hardware and software used
- Datasets or simulations
- Parameters or configurations

V. Results and Evaluation

Present your findings:

- Quantitative results in tables/graphs
- Qualitative insights
- Comparison with baseline or existing methods

VI. Discussion

Interpret your results. What do they mean? Any surprising findings?

VII. Conclusion and Future Work

Summarize your project outcomes and propose what could be improved or continued in future research.