Denise Dodd

Predicting Used Car Prices

Proposal

TOPIC

This project will predict used car prices. Based on these predictions, a user will be able to input their desired car characteristics and return a predicted price.

BUSINESS PROBLEM

Recently, I made my first ever purchase of a car. I had no idea what I was doing and left feeling quite demoralized by the experience. I had no concept of what a car costs or how different features affected the sale price. Understanding these data points would have helped me focus my car search and empowered me to negotiate for a fair price. I will be approaching this project from the perspective of an individual car buyer setting a target budget and selecting which features they would like in a car. However, predicting car prices can also be beneficial to car brokers, dealerships, and individual car sellers.

DATASET

For this study, I will be utilizing a dataset titled "Car Price" (Brar, 2024). This is a dataset with details of over 8,000 used cars. Variables include numerical variables of varying units of measurement including selling price, kilometers driven, engine power, number of seats and the production year of the car. It also includes categorical variables such as the type of fuel the car takes, what kind of transmission that car has, what the prior ownership status of the car was, and if the car is being sold via a dealer or an individual.

METHODS

After reviewing and cleaning the data, I will assign the target variable to be "selling_price" and the features variables to be the rest of the dataframe. Training and testing variables will be created using an 80/20 split of the features and target data. A correlation matrix will be created showing how strongly the numeric variables are correlated to selling price and other numeric data. Due to the varying numeric units of measurement, I will use a pipeline to scale the numerical features variables. Also included in the pipeline, I will create dummies of the categorical features variables. The transformed data will be used to fit several predictive models and make predictions. I will move forward with the model that is most efficient based on RMSE and R-Squared metrics. Regression plots and residual plots will visualize the efficiency of the chosen model. I will extract the top features that impact the model. Once a model has been trained and proven effective, I will demonstrate how a car buyer can input data for their ideal car and receive an anticipated price for the described car. The user can use the knowledge gained from the correlation matrix and features importance to adjust the details of their desired car to raise or lower the anticipated selling price.

ETHICAL CONSIDERATIONS

If a broker or dealer is deploying this model, they will have a dataset of cars where they facilitated the sale. However, an individual buyer or seller will not have access to this data and will have to ensure that they are obtaining the data needed to train their model in an ethical manner. They may be able to obtain a dataset of recently sold cars in their area from a data repository. It is important that the information in their data set is ethically sourced and does not contain private information pertaining to the buyers and sellers.

CHALLENGES/ISSUES

I made a purposeful decision to challenge myself on this project by using pipelines. I have studied pipelines but have never implemented them into one of my projects. I anticipate some learning curves along the way, but I am eager to discover how pipelines can make my coding more efficient and less redundant.

Another challenge is that the units of measurement in the dataset are unclear. There are some variables which are labeled as being in kilometers. Based on the data, I can deduce that the data in the "engine" column is displayed in liters and the data in the "max_power" column is displayed in kilowatts, but this is not clearly stated. The variable that I cannot determine a unit of measurement from is the "selling_price" variable. If I knew the original unit of measurement, I could convert it to dollars or another desired unit of currency. Without knowing the original currency, I cannot make a conversion. As noted in the "Ethical Considerations" section of this proposal, a dealer or broker would have their own set of data when running this project and they would know the unit of currency in their dataset. An individual buyer or seller would have to source data and be conscious that they are using consistent units of measurement throughout their dataset.

REFERENCES

The below articles complement each other and add context to my study. One article discusses how factors pertaining to an individual car impact the selling price, and the other article details how outside factors can change the overall market for used cars.

1) [Just What Factors Into The Value Of Your Used Car?](#) (D'Allegro, 2021) - This article details how unique characteristics of each car impact the value of the car. Variables included in my dataset such as mileage, transmission type, and model are discussed. It will be interesting to see if these factors affect the price of the cars in my dataset in accordance with the findings in the article.
2) [Average Used Car Price Down 4% Since Last Year](#) (Kelly Blue Book, 2024) - This article discusses how outside factors not pertaining to a specific car impact the overall used car market. These are factors not included in my dataset such as inflation, supply and demand, and emerging trends such as electric vehicles. If I am unable to train my models to a degree of accuracy that I am comfortable with, the information in this article might inform me of additional data points that I will want to incorporate into my study.

CITATION

Brar, S. (2024, March 28). *Car price prediction dataset*. Kaggle.
https://www.kaggle.com/datasets/sukhmandeepsinghbrar/car-price-prediction-dataset

D'Allegro, J. (2021, October 19). *Just what factors into the value of your used car?*.
Investopedia. https://www.investopedia.com/articles/investing/090314/just-what-factors-value-your-used-car.asp

Tucker, S. (2024, February 19). *Average used car price down 4% since last year - Kelley Blue Book*. Kelley Blue Book. https://www.kbb.com/car-news/averaged-used-car-price-down-4-since-last-year/