Report: Car Price Prediction Project

Prepared by: Afees A A
Date:27-May-2023

Introduction

This project predicts car prices using regression analysis. The dataset, 'cars_price.csv', has 206 samples and 25 features. Our goal is to build an accurate regression model for car price estimation.

To achieve this, we preprocess the data, handle missing values, and encode categorical variables. Different machine learning techniques like linear regression and random forest are explored. Model performance is optimized through parameter tuning.

The 'final_model' is evaluated using metrics such as MSE, MAE, and R2-score. Feature importance analysis is conducted to identify key predictors of car prices.

Through this project, we aim to develop an accurate regression model for car price prediction and gain insights into important factors influencing car prices.

Data Loading and Preprocessing:

Initially, I imported the necessary libraries required for our project and then loaded our data set “cars_price.csv”. I have used only google colab for coding purposes.I have checked the dimensions of the data and conducted a descriptive statistical analysis.

It was found that there are few Null values present in the data set. I took the sum of all Null values of each column and performed Data cleaning on most of the columns. The feature- “Normalized-Losses” had a great number of Null values. As such, I have dropped this feature since this will not add any value to the result and may even affect the final model.

Subsequently, I have dropped the rows with Null values in the price feature, since it is our dependent variable. Furthermore, I have corrected the data types of some of the features as they were incorrectly represented. Then I cleaned the “Make” feature, since it was not properly representing the car company name.

Exploratory Data Analysis

I have visualized all the features against the price factor in order to identify the correlation between the features. I have identified that some of the features do not influence the price of the car in reality, so I have not considered them for model building.
Furthermore, I have observed many insights by analyzing each factor which are noted as insights along with each of the analysis performed in the project.
Some of the information cannot be further explored or the final decision cannot be made because of a minimal number of data points. I have mentioned them as notes in my project.

Feature Engineering and Data Splitting

Derived a new feature from “car company”, as we made an insight above that we can split the car company name into different price ranges.Like Low Range, Medium Range, High Range cars.

Subsequently, I created a new data frame with all the useful features. I have then created dummy variables for all categorical features. I proceeded with feature scaling of numerical data.

I then selected features and labels for model testing and training. Then I have splitted the data into training and testing sets at 80 and 20% respectively.

Machine Learning Techniques

I have performed various machine learning algorithms on the data. These include Linear Regression, Decision Tree Regression, Random Forest and XGBooster. I compared the performances of all the models using RMSE and R2 Score values. It was found that the highest performance was given by XGBoost Model around 94.5% with least RMSE value.However, Decision tree and Random Forest Model performance were also good.
So we can use any of these three Models for predicting the price of a car in the future.

Model Tuning and Building 'final_model'

I have selected the XGboost Model as the final model and performed model tuning action using GridSearchCV function. It was observed that MAE and MSE value decreased. Lower MAE indicates that the model's predictions are closer to the actual values on average.Lower MSE indicates that the model's predictions are not only closer to the actual values but also have smaller overall errors.While A lower RMSE (Root Mean Squared Error) indicates better model performance in terms of the average magnitude of the prediction errors. R2 value for testing data increased while remained mostly unaffected for training data. A higher R2 indicates better model performance in explaining the variability of the target variable.

Results and Interpretation

It was observed that the XGBoost Model performed the best compared to the other models. With tuning the parameters, it was identified that the performance was improved even better.

It was observed that MAE and MSE value decreased. Lower MAE indicates that the model's predictions are closer to the actual values on average.Lower MSE indicates that the model's predictions are not only closer to the actual values but also have smaller overall errors.While A lower RMSE (Root Mean Squared Error) indicates better model performance in terms of the average magnitude of the prediction errors. R2 value for testing data increased while remained mostly unaffected for training data. A higher R2 indicates better model performance in explaining the variability of the target variable.

Final Model	Training Score	Testing Score	MSE	MAE	RMSE
XGBoost	0.989279	0.957326	5.220986e+06	1372.678747	2284.947627

Conclusion

- First we did the Basic Understanding of Data
- Then we performed Data Cleaning to make the raw data more usable while analyzing.
- Then we performed Exploratory Data Analysis to generate insights from the data.
- Then we performed Data Preprocessing to make data suitable for model training & testing.
- Then we trained our model using different Machine Learning Algorithms.
- In the end we came with 94.5% accuracy which was given by XGBoost Model.The accuracy of the model in predicting the car price is measured with RMSE, RMSE of the test dataset is 2589.
- The model was further tuned to perform better. We have 95.7% accuracy with MSE, MAE and RMSE values less than the previous values.
- So we can use this Final model for predicting the price of a car in future.