Predictive Modelling of Car Prices for Market Entry Strategy in the US Auto Industry- Major Project

-By

Sanjay R

23m0002@iitb.ac.in

Krutanic-Machine Learning Course

Problem Statement:

A Chinese automobile company Geely Auto aspires to enter the US market by setting up their manufacturing unit there and producing cars locally to give competition to their US and European counterparts. They have contracted an automobile consulting company to understand the factors on which the pricing of cars depends. Specifically, they want to understand the factors affecting the pricing of cars in the American market, since those may be very different from the Chinese market. The company wants to know:

Which variables are significant in predicting the price of a car?

How well those variables describe the price of a car?

Models Tested

- **Linear Regression:** Simple and interpretable, it models the direct relationship between features and target prices, providing a clear baseline.
- **Ridge Regression:** Adds regularization to prevent overfitting by penalizing large coefficients, useful for handling multicollinearity.
- **Lasso Regression:** Encourages sparsity by shrinking some feature coefficients to zero, helping with feature selection and reducing overfitting.
- **Random Forest Regressor:** Combines multiple decision trees for accurate predictions and handles complex, non-linear relationships well.
- **Support Vector Regressor (SVR):** Uses kernels to manage non-linear relationships and high-dimensional data, though it needs careful tuning.
- **XGBoost Regressor:** An advanced boosting technique that builds models sequentially to improve accuracy and handle complex data patterns effectively.

Performance of models

Performance	MSE in Price	R2 Score	Accuracy	Precision	Recall	F1 Score
Linear Regression	12081874.0312	0.8470	0.8780	0.6875	1	0.8148

Ridge Regression	11689788.6653	0.8519	0.8537	0.6471	1	0.7857
Lasso Regression	12074507.2070	0.8470	0.8780	0.6875	1	0.9796
Random Forest Regressor	3448402.0310	0.9563	0.9024	0.7333	1	0.8462
Support Vector Regressor	16540710.8674	0.7905	0.8780	0.6875	1	0.9796
XGBoost Regressor	5688432.6889	0.9279	0.9024	0.7692	0.9091	0.8333

The following are the best models for predicting car prices in this dataset: Random Forest Regressor and XGBoost Regressor.

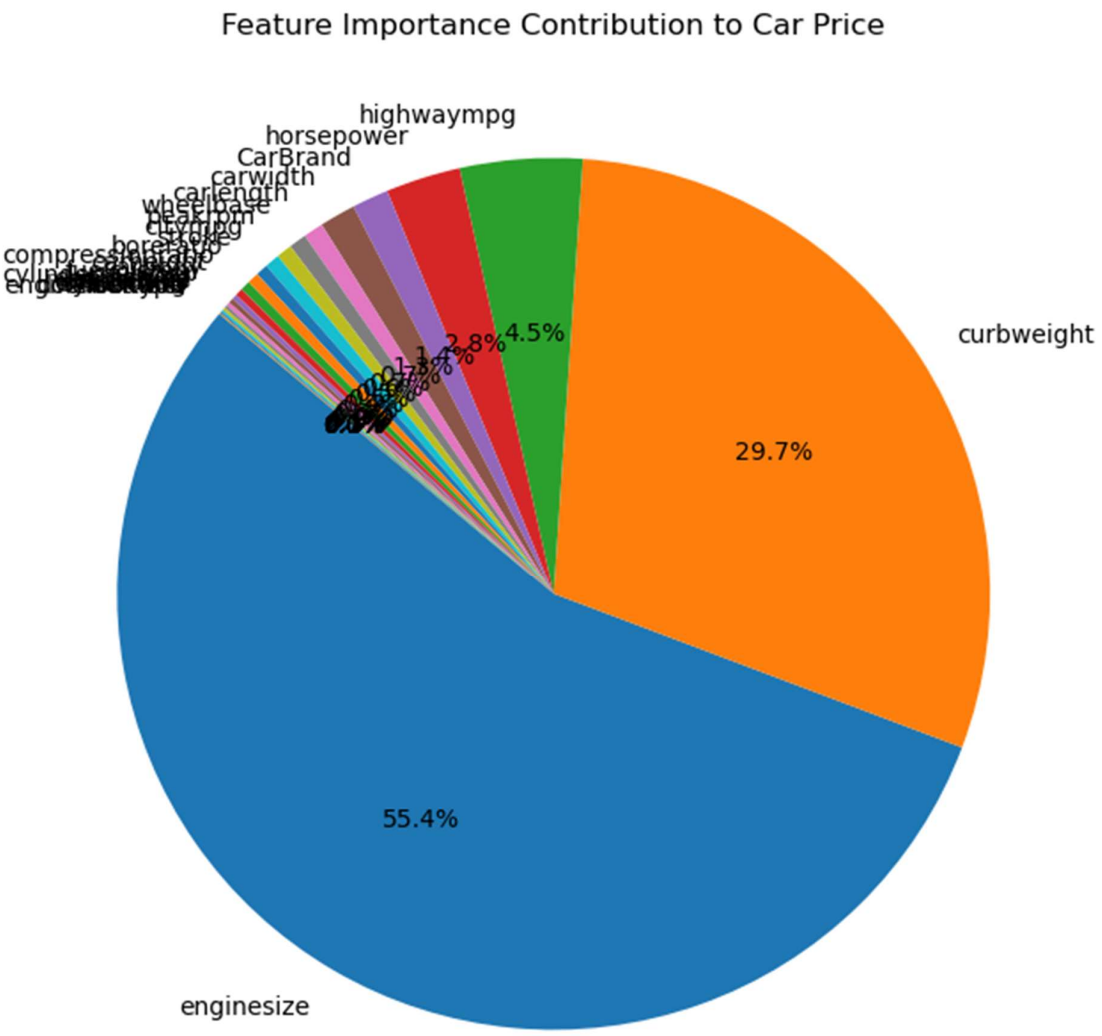
- Random Forest Regressor had the lowest Mean Squared Error (3,448,402) and highest R^2 Score (0.9563), indicating that it captures the variability of the data well and with high precision. Moreover, it has a higher F1 score and accuracy which implies that it balances well between recall and precision thus making it a robust model to handle various data patterns.
- XGBoost Regressor also performed well with low Mean Squared Error of 5,688,432 and high R^2 Score of 0.9279. This model's strong performance is associated with its boosting technique which allows for handling complicated relationships through iterative improvement of weak learners.
- However, other models such as Linear Regression, Ridge Regression and Lasso Regression performed quite well but were defeated by ensemble methods (Random Forest and XGBoost). Support Vector Regressor on the other hand had decent accuracy and F1 score though its Mean Squared Error was still higher than others, suggesting that it may not be an appropriate fit for this dataset.

Challenges faced:

- One of the first challenges was to ensure that the data is clean, well-formatted, and all the categorical variables are encoded properly. Besides, there is mixed data type and missing values, scaling features for some of the models that make the preprocessing even more complex.
- Car prices depend on complex interactions between features, while linear models can't describe them well. This required running with models such as Random Forest and XGBoost, more complicated yet efficient.
- A significant challenge lied with MLP Regressor (Neural Network); the model made that much use of computational power, whereby in the absence of hyperparameter tuning, the model's performance was meager. Finally, it was not taken into consideration because it was taking too long for processing, thus resulting in the hang of Jupyter Notebook.

Conclusion:

- The variables significant with the Price of the car is evaluated with the feature importance contribution of the variable with the price, which is taken from the random forest regression fit. From the below image, it can be observed that the engine size affects heavily followed by the curb weight.



- In an ensemble, engine size and curb weight describe about 85.1% of the price of a car. The pie chart overall describes how well the variables describes the price of a car.