Car Price Prediction

Sri Sudheera Chitipolu
S538422@nwmissouri.edu

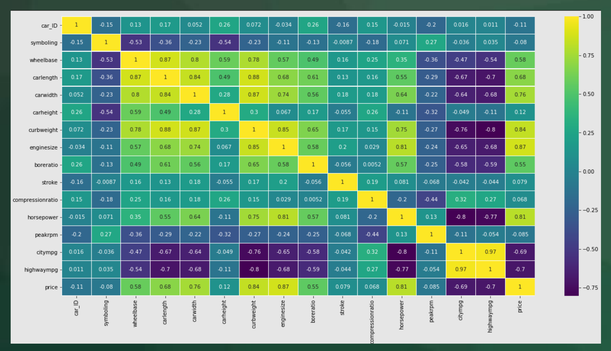
Introduction

For understanding pricing dynamics of the new market in the different cars for business growth, we will predict the car's prices depending on different independent variables. Several factors, including mileage, make, model, etc. can influence the actual worth of a car. From the perspective of a seller, it is also a dilemma to price a car appropriately. Based on existing data, the aim is to use machine learning algorithms to develop models for predicting car prices.

Dataset and Pre-Processing

For this project, we are using the dataset of car prices of different cars from United States, available on Kaggle[1]. The features available in this dataset are Mileage, Car Name, Model, Fuel type, Aspiration, etc.

- Drop Unnecessary Columns: Car ID, car name does not affect the price.
- Created Dummies: Because of non numerical variables in dataset created dummies on categorical data.



Interesting Relationship: There is huge correlation between price and engine size and next price and car weight, car length, car weight, wheel base and boreratio.

But price and engine size are weekly correlated with highwaympg.

From heat map we can observe that citympg, highwaympg, peakrmp, symboling and car Id are negatively correlated to price. So we can consider other features to predict.

Methodology

We utilized several classic and state-of-the-art methods, including ensemble learning techniques, with a 70-30 split on test and training data.

Linear Regression: Linear Regression model is used on dataset and evaluated the model with mean square error and R square and also improved the model performance by choosing other regressions like Ridge Regression, Lasso Regression, Elastic Net Regression

Classification: Decision Tree and SVM with different kernels are used on dataset and evaluated model with cross validation and also performed evaluation on test set using SVM model with different kernels. Used SVM Kernels are

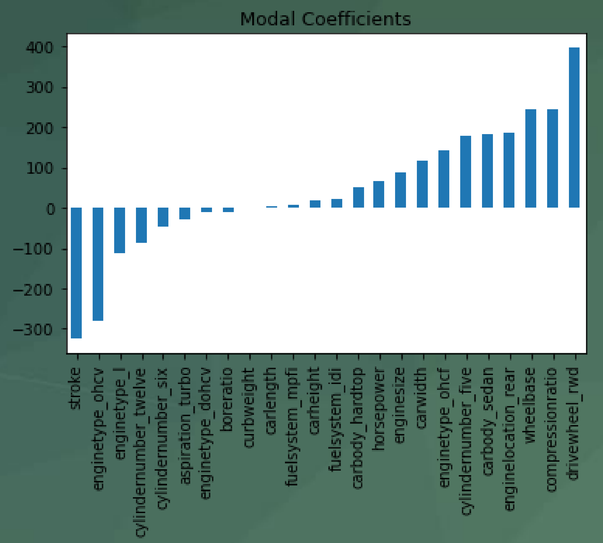
- Linear, RBF, Polynomial, Sigmoid

Random Forest To account for the large number of features in the dataset used one of the ensemble learnings called random forest with random forest regressor.

Anomalous Data: Performed all outliners detectors methods to know about all anomalous data. Methods are: Isolation Forest, Minimum Covariance Determinant, Local Outlier Factor, One-Class SVM

Results

Learning Algorithm	R Square
Linear Regression	85%
Ridge Regression	82%
Lasso Regression	86%
Elastic Net Regression	76%



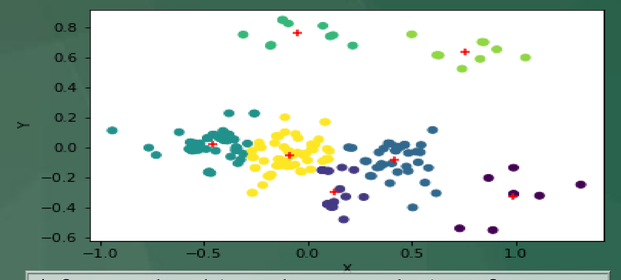
Elastic net regression is working very well, because R square value decreased and feature coefficient values also performed good way rather than focusing only one feature, now most of the features of dataset are important.

Decision Tree	99%
Random Forest	98%

Among all kernels Linear kernel having best results followed by polynomial kernel. Also in test evaluation linear and polynomial kernel gave best accuracy values.

So overall linear kernel SVM gave best accuracy values.

Linear	RBF	Polynomial	Sigmoid
Accuracy is 0.9230769230769231	Accuracy is 0.7972027972027972	Accuracy is 0.8321678321678322	Accuracy is 0.20279720279720279
Precision is 0.9235617323852616	Precision is 0.8094985985674246	Precision is 0.8404354266423232	Precision is 0.14564777327935222
Sensitivity is 0.9230769230769231	Sensitivity is 0.7972027972027972	Sensitivity is 0.8321678321678322	Sensitivity is 0.20279720279720279
F1 is 0.9228563567787569	F1 is 0.7795675084883028	F1 is 0.8307458507275668	F1 is 0.16863228304875902



After rechecking the correlation for clustering, found car weight and engine size are most correlated features through PCA with 70% sum of variance ration. Got 7 clusters. There are 13 outliners which are removed using local outliner detector by which we got 21% AMSE – linear Reg.

Future Work

We plan to utilize Artificial Neural Networks to improve our prediction performance while avoiding overfitting. In addition to this, we shall tune Decision-Tree parameters like the number of trees, depth, etc. and sub-sample data points.

References

1. https://www.kaggle.com/hellbuoy/car-price-prediction

2. N. Monburinon et al "Prediction of prices for used car by using regression models," ICBIR 2018, Bangkok, 2018, pp. 115-119.