TinkerWeek.py

Machine Learning

Project Bash Report

By
Midhun Chandran
midhunchandran511@gmail.com

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

Based on various market surveys, the consulting firm has gathered a large data set of different types of cars across the American market.

Business Goal

We are required to model the price of cars with the available independent variables. It will be used by the management to understand how exactly the prices vary with the independent variables. They can accordingly manipulate the design of the cars, the business strategy, etc. to meet certain price levels. Further, the model will be a good way for management to understand the pricing dynamics of a new market.

Steps involved in experimental evaluation

1. Analysis of the given dataset
2. Cleaning the data and making necessary changes
3. Visualising the data(Numerical & Categorical)
4. Data preprocessing
5. Trying out preferred algorithms and evaluating the results
6. Determining the most fit algorithm and further working to improve the results

Summary

With the preliminary analysis of the given dataset I could conclude that there were 24 independent variables and the dependent variable, price.The data was clean, there was no missing values.I could gather preliminary insights and get the hang of data by visualising it. I used matplotlib and seaborn for it. I drew inferences by visualising the relations of **numerical data**('symboling', 'wheelbase', 'carlength', 'car width', 'carheight', 'curb weight', 'enginesize', 'boreratio', 'stroke', 'compression ratio', 'horsepower', 'peak rpm', 'city mpg', 'highway mpg') with price with the help of pairplot from seaborn and visualised the relations of **Categorical data**('CarName', 'fuel type', aspiration', carbody', 'drive wheel', 'engine location', 'engine type', 'fuel system', 'door number', 'cylinder number') using box plot and histograms from the same.

After finding out relative importance between variables I didn't actually weed the un-important ones out, instead I planned on trying to fit them and make predictions with the algorithms I had in mind and then by further evaluations and iterations remove the unnecessary variables.

In the data preprocessing stage, I used label encoding to encode the categorical variables and used minmax scaler for scaling numerical variables.Even though some of the categorical variables weren't exactly ordinal, I used label encoding because I had a plan to use the ensemble algorithms.The very nature of ensemble algorithms helps them to deal with the categorical variables so using label encoder helps us save disk space and also you won't be able to extract variable importance id you one hot encode it.
Though I did try algorithms like SVR where I had to one-hot encode the data.

I tried out the following algorithms and upon evaluation it gave the following :

1. Random Forest Regressor - `Mean Absolute Error: 1546.252445376344

Mean Squared Error: 5621769.098518753
Root Mean Squared Error: 2371.0270134519246
r2_score: 0.9148698826613778

2.Decision Tree Regressor('mse') - Mean Absolute Error: 1900.39
Mean Squared Error: 9213132.88
Root Mean Squared Error: 3035.31
r2_score: 0.86

3.Linear Regression - RMSE = 3786.57
Accuracy = 78.29 %

Ridge Regression - RMSE = 3738.44
Accuracy = 78.84 %

4.Support Vector Regression('rbf') - Mean Absolute Error: 2644.416
Mean Squared Error: 13936459.48
Root Mean Squared Error: 3733.15
r2_score: 0.86

End analysis

From the initial evaluation of all the algorithms I tried I chose Random Forest Regressor and decided to improve its accuracy by selecting the most important variables.
From further analysis I found out that the variables:

'wheelbase', 'carheight', 'peak rpm', 'carlength', 'horsepower', 'car width', ' highway mpg', 'curb weight' , 'enginesize', 'stroke', 'city mpg'

are the most important and their relative importance scores being:

Variable: enginesize Importance: 0.649
Variable: curbweight Importance: 0.202

Variable: highwaympg Importance: 0.042
Variable: horsepower Importance: 0.023
Variable: carwidth Importance: 0.019
Variable: carlength Importance: 0.018
Variable: citympg Importance: 0.011
Variable: wheelbase Importance: 0.01
Variable: peakrpm Importance: 0.01
Variable: carheight Importance: 0.008
Variable: stroke Importance: 0.008

The final Evaluation results of **Random Forest Regression** with only important variables:

Mean Absolute Error : 1516.21
Mean Squared Error : 5347250.95
Root Mean Squared Error : 2312.41
R2_score : 0.92

