Report on Price Prediction model for Used Cars

Summary

Visualising the data, I can see that the 3 types of car that sell the most are SUV, Sedans and Pickups, so one might consider aligning your stocks to just those 3 types. Interestingly, whatever region you are in does not change that.

Having analysed the used car data we have available, I have built a model that has a MSE of 58,209,825. That has root means squared error of $7,629. The average price of car is $17,954, the model's percentage error is approximately plus or minus 42.49%. This is not a great model, so it needs a few more iterations at least to perform better.

The current model tells me that manufacturer and fuel type are the 2 most significant factors.

Findings

Firstly I cleaned that data as best I could, preserving 70% of the original data. I then used OHE technique to convert the categorical features into numerical. Except for the condition feature, which I ranked. I also combined US States into regions. All of this left me However, in order to be able to work with the data, I could only sample 20% of this data.

However, in order to accomplish this model I did have to perform a PCA transformation to 80%. The models I tested were

1. Linear Regression
2. Polynomial on data PCA to 80% variance (Best model with order on 3)
3. Ridge Model on original data, and then Ridge model having done a SFA on the Polynomial Data PCA data
4. Lasso on original data, and then Lasso model having done a SFA on the Polynomial Data PCA data

While I have not done exhaustive modelling, it is clear that there is not much difference between the Linear, Ridge, and Lasso models on the data. And when I did Ridge and Lasso models on the Polynomial data, the results were not as good. The Mean Squared errors for all these models were

Simple Linear=65.712,815

Ridge Regression with Optimised Alpha=65,740,841

Ridge Regression with Optimised Alpha on the polynomial data=67,254,331

Lasso Regression with Optimised Alpha=65,667,587

Lasso Regression with Optimised Alpha on the polynomial data=67,253,241

The Polynomial Model showed some promise when degree was of order 2.

The Polynomial Model on training data=56,103,940

The Polynomial Model on testing data =58,209,825

So I think it is the most promising model to explore going forward, with a bigger sample size, and more feature selection, possibly without PCA. If I was to improve this model this is the model I would continue to work on.

Finally the Lasso models tells me that below are the 5 most signicant features. That might indicate to me that manufacturer and fuel type are the 2 most significant factore, but also the very highest end manufactured cars may be skewing the data, and I might consider taking them out of future models.

manufacturer_ferrari

manufacturer_aston-martin

fuel_hybrid

fuel_gas

manufacturer_porsche