BELLEVUE UNIVERSITY

Used Vehicles Price Prediction

DSC 550, FALL 2020

RAMAKRISHNA DANDA

Introduction: -

Used cars market in USA is approximately about 88billion per year (1), of which portions of sales happens on the popular online market platforms like craigslist, truecar and other sites. On most instances, seller would post a vehicle based on the perception, knowledge they have on their vehicle and condition to post a price without having any insights and factors effecting the price.

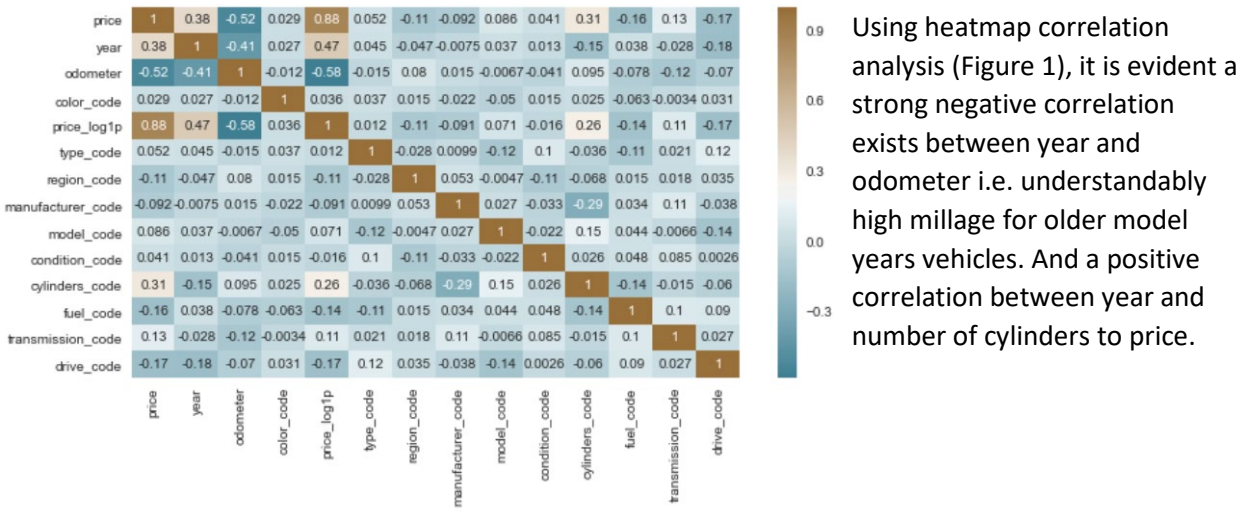
Proposal: -

It would be beneficial to have an option for buyers and sellers to know the fair price of a particular vehicle based on various factors of year, millage, region, color, and model of vehicle effecting the price. This project is to find correlation of multiple features that effect the price of a vehicle and predicting price of the used cars based on the data posted to craigslist vehicles for years 2010 to 2015.

Analysis of data & findings: -

Data is sourced from Kaggle which has many other features like 'url', 'region_url', 'title_status', 'vin', 'county','size', 'image_url', 'lat','long'. Since these are no helpful to the use case, it was removed from the overall data set along with null values. Changed datatypes from non-integers to numeric data. Since most vehicles are driven by gas powered, removed the outliers of diesel and electric vehicles using histogram analysis. Also, found the condition of the vehicles to be mostly spread in excellent, good and like new and removed all other outlier values from the dataset. Missing odometer values are filled by median values and type categorical values as most frequently occurred in the dataset. Since the price is not normally distributed, the entire data set is normalized into natural logarithmic values. Observation is most vehicles sold are made by Chevrolet and Ford, thus indicating the customers of these are not willing to hold these cars and selling them more often. Another indication is most of the white colored vehicles are indicated as excellent condition, followed by black and silver. There is a chance that most cars in used cars market is the indication of the popular colors being bought as brand new. The other observation is that Orlando has the highest excellent and good cars listed followed by Washington DC.

Figure 1



Using heatmap correlation analysis (Figure 1), it is evident a strong negative correlation exists between year and odometer i.e. understandably high millage for older model years vehicles. And a positive correlation between year and number of cylinders to price.

Model Implementation steps & observations: -

First the entire categorical variables are encoded for color, type, region & cylinder while flattening the values of price to keep them all in the same normalized format, before splitting the data in to test and train. That is fit to Random Forest Regression algorithm using all the available 88 features that formed from 15 original features using one hot encoding.

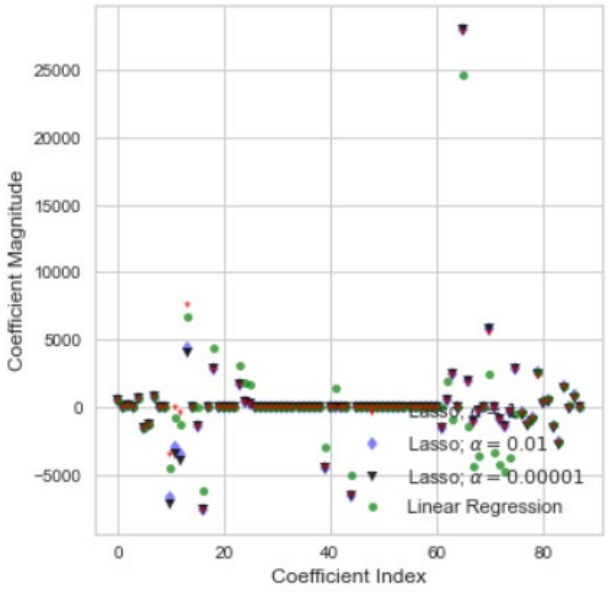
Observation is, ***Random forest*** model gave a R^2 value of 0.689. Resulted in the mean absolute error 2894, while the score is at 71%. So, a price prediction that is varying of about $2,894 for each price it is predicting per vehicle. Up on applying the PCA on the scaled data that reduced 88 features in to 3, but it produced an absolute error of $7,278, while the score went down by -1.54.

Conclusion on Random forest model is, not to apply PCA and go with base features to get better prediction of price for the given used cars dataset.

Tried with ***Lasso regression*** model on finding the relationship between a scalar response of price and with explanatory variables gave a detail break down on each feature of vehicle that is affecting the price by the year. i.e. for each additional year (latest model year) to a vehicle a $3,506 value is added to the price listed, which is the biggest factorial change determining the price variance. Followed by odometer reading, i.e. for each additional (mean – std) value of 11k millage, a negative -$3674 is attributed to the price of the vehicle. Number of cylinders of a vehicle has an effect of $3,900 influence on each unit of cylinders it goes up. It is not intuitive to interpret the region in which the vehicle is sold, and there is a negative -1,000 dollars. All these factors are for the mean price of the vehicles at $32,968. (Note: - took good and excellent condition vehicles for the entire analysis)

Lasso model gave a R^2 value of 0.997 with a score of 73%, which is slightly higher than the random forest R^2 value of 0.68 and 71% score. Since Lasso gave a detail breakdown of factors effecting the price, I prefer to use Lasso over Random forest.

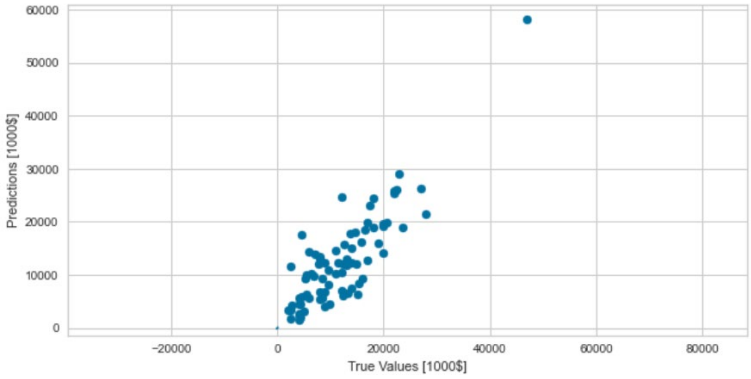
Figure 2



To further explore, fed the data to ***Ridge regression*** algorithm with varying degree of alpha (hyper parameters). Training score for alpha=0.01 is 0.62, while Test score for alpha =0.01 is 0.57. Comparing this to Lasso and LR, it is evident Lasso is giving better prediction power even after varying degree of hyper parameters c(green) vs red circle dots of Lasso (Figure 2), which is evenly distributed across the 88 features at 0.73 score.

Thus, giving more reasons to use Lasso model for the given used vehicle data set.

Figure 3



Exploring further behavior using Neural network, fed data to ***Keras Sequential*** and is produced a R^2 score of 0.63, which pretty much in line with ***Scikit Learn Standard Scalar*** and Ridge described above. Figure 3 is an indicative of prediction to values are linear up to $30k value of price and their predicted values using Keras model.

Conclusion: -

On going through different models exploring data, scoring and R^2, I find it is Lasso model is a good fit for this data set. It has performed much better both in error rates, prediction values and in performance of code, with very little changes to 12 features and not needing to encode a lot of categorical variables comparing to other models. Lasso has also given a way to identifying features causing the variance in the final price, which is also one of the original hypotheses made.

1)https://www.ibisworld.com/industry-statistics/market-size/used-car-dealers-united-states

2)https://www.kaggle.com/austinreese/craigslist-carstrucks-data