# Prediction of Used Cars' Prices by Using SAS EM

**Jaideep A Muley**

**Student, MS in Business Analytics**
**Spears School of Business**
**Oklahoma State University**

ABSTRACT

**Prediction of Used Cars' Prices by Using SAS EM**

The used car market has grown tremendously over the last few years. According to the research, the US used car market has grown by 68% since the sub prime crisis and is expected to grow for the next few years. Online market places and advancement of technology in auto industry are some of the major contributors for this growth. Franchises of used car firms and giant online marketplaces are leveraging this growth. The aim of this paper is to analyze the market trend of used car industry and find out what are the factors that are important decide the price of a used car and finally predict the price of a used car. With the help of SAS Enterprise miner I have used statistical methods such as Transformations, Decision Trees, and Regression to identify the target variable.

## Introduction

The used car dataset was obtained from Kaggle. Kaggle is an online data provider which provides data for data enthusiast and conducts data science competition for students and professionals. This particular data set was obtained by scrapping EBay's used car buy sell portal. The selected dataset has more than 300,000+ data points with more than 15 attributes with 40 different brands of car.

Following is the attributes of the dataset before cleaning and selecting variable for analysis

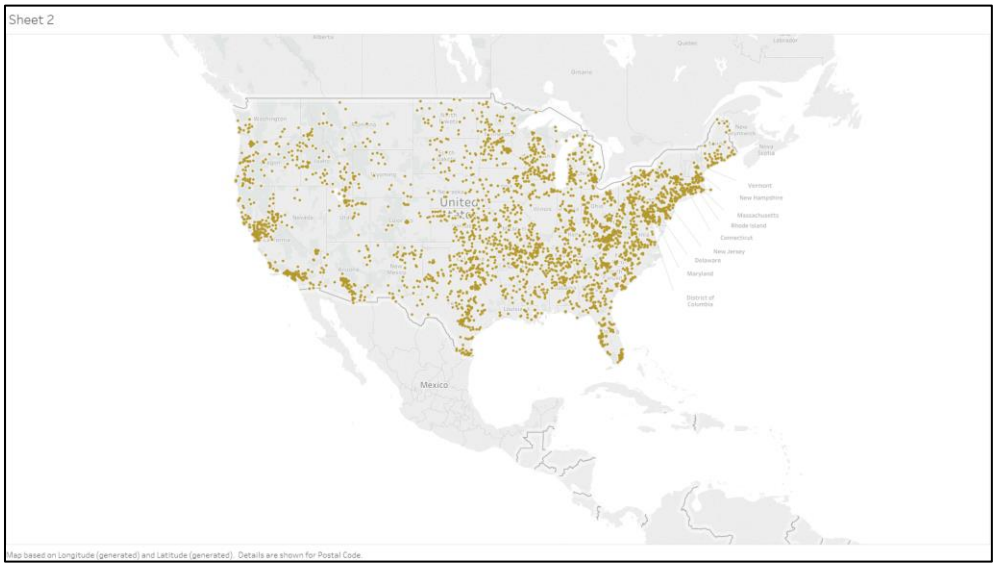
| **Variable** | **Description** |
|---|---|
| Date Crawled | When this ad was first crawled, all field-values are taken from this date |
| Name | "Name" of the car |
| Price | The price on the ad to sell the car |
| Vehicle Type | Sedan/Coupe/Limo/SUV |
| Year of Registration | at which year the car was first registered |
| Gearbox Type | Automatic/Manual |
| Power in PS | Power of the car in PS |
| Model | Name of the Model |
| KM Reading (KM) | A- Less than 75k, B-75k-100k, C-100k-125k, D-125k and more |
| Month of Registration | which month the car was first registered |
| Fuel Type | Gas/Diesel/Hybrid/Electric |
| Brand | Multiple Brands |
| Damage repaired? | if the car has a damage which is not repaired yet (1-Yes; 0-No) |
| Date Created | the date for which the ad at ebay was created |
| Last Seen | when the crawler saw this ad last online |

Not all variables mentioned above were used for the analysis, whereas some new variables were created considering some assumptions for the ease of analysis. Those variables are mentioned below:

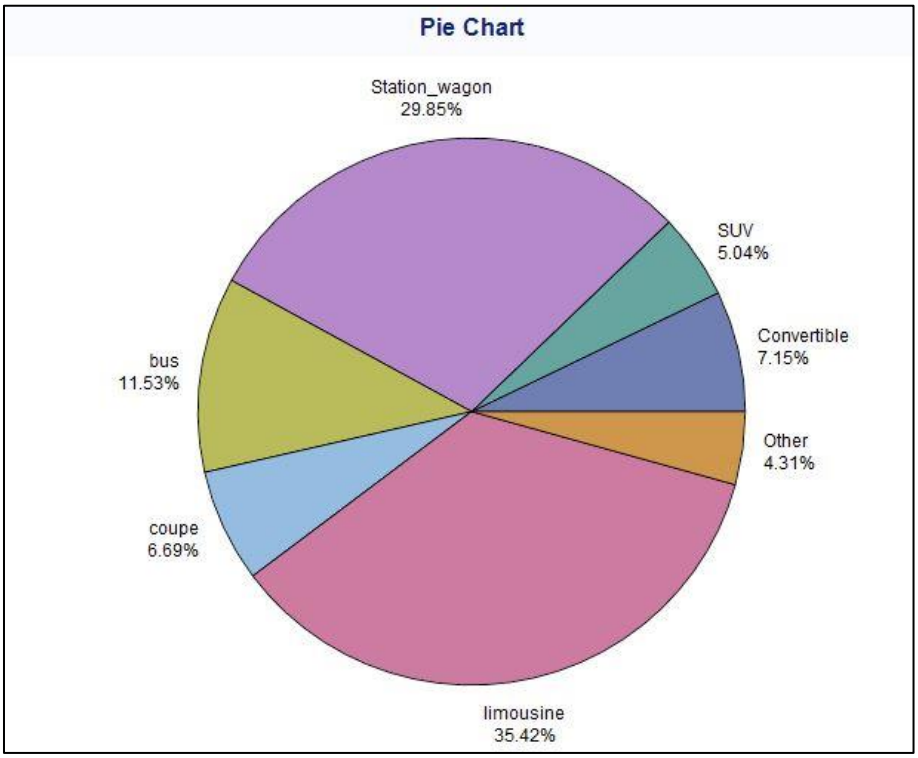
| **New Variables** | |
|---|---|
| **Variable** | **Description** |
| Age | 2017 minus Year of Registration |
| Country | From the brand, country of the car was decided |
| Days | Last seen - Date Crawled |

## Descriptive Analysis of the Data

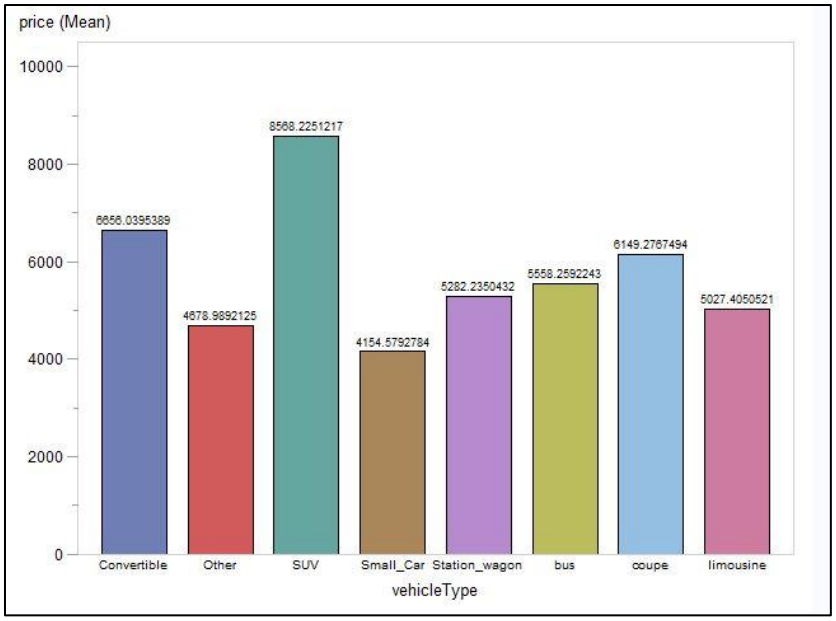
The following diagram shows the density of listed on the portal State wise, It can be seen that density is more on the eastern coast as compared to oter states.



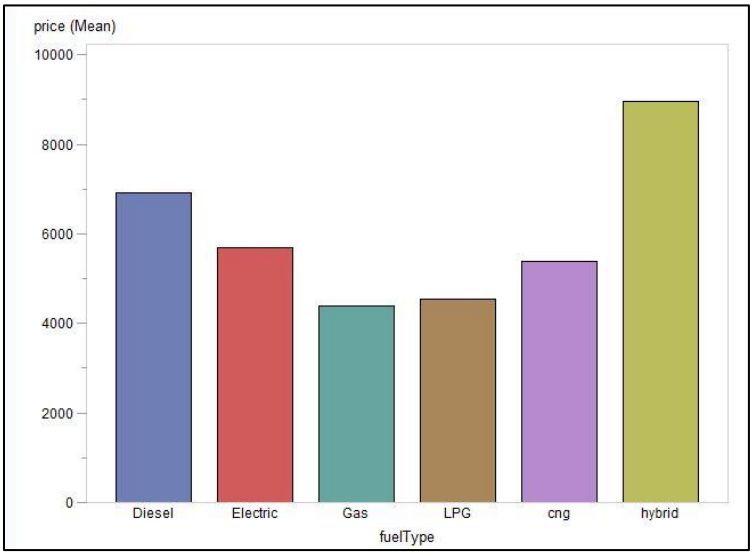

Following pie cart shows the distribution of cars in the dataset according to the type of the car




The following visualization shows the frequency of different types of cars. It can be seen that type ‘SUV’ has the highest price followed by the Coupe.




The following visualization shows the bar graph of type of fuel type vs average price of car. Average price of hybrid car is the highest while average price of the gas segment is the lowest in the given dataset

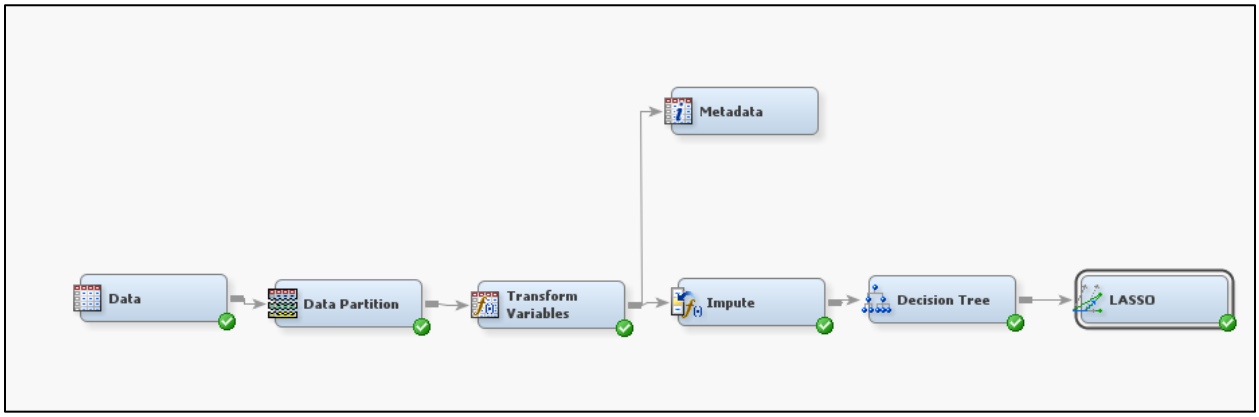



The following two visualizations shows the correlation matrix of all the variables with each other. It can be observed that the target variable 'Price' has the highest positive correlation with the variable 'PowerPs' while Price has the highest negative correlation with Age.

```
In [10]: corr
Out[10]:
                     price   powerPS  notRepairedDamage      days       Age
price              1.000000  0.445492          -0.196557  0.135526 -0.461373
powerPS            0.445492  1.000000          -0.021768  0.053979 -0.042662
notRepairedDamage -0.196557 -0.021768           1.000000 -0.052423  0.064323
days               0.135526  0.053979          -0.052423  1.000000 -0.006607
Age               -0.461373 -0.042662           0.064323 -0.006607  1.000000
```

## Methodology

Below is the complete node diagram of the model that was used for this project. Each node along with its purpose and utility is mentioned below:




### Data:

This node imports the cleaned dataset into SAS EM environment. Following are the roles and levels of the variables selected for the analysis.

| Name | Role | Level |
|---|---|---|
| Age | **Input** | **Interval** |
| Country | **Input** | **Nominal** |
| days | **Input** | **Interval** |
| fuelType | **Input** | **Nominal** |
| gearbox | **Input** | **Nominal** |
| kilometer | **Input** | **Nominal** |
| name | **Rejected** | **Nominal** |
| notRepairedDam | **Input** | **Interval** |
| powerPS | **Input** | **Interval** |
| price | **Target** | **Interval** |
| vehicleType | **Input** | **Nominal** |

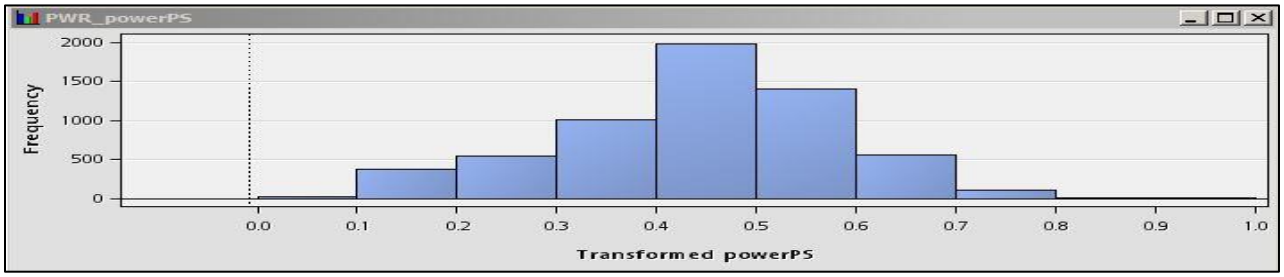
### Data Partition

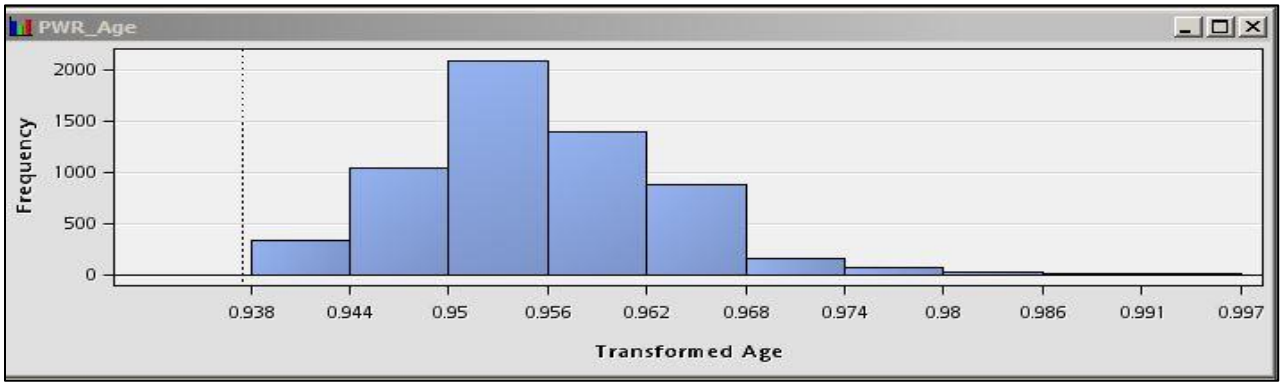
Data partition node divides the data into Training and validation. The ration used for the analysis was 70% training and 30% validation. No other changes were made in this node.

### Transform Variables:

After analyzing the data initially, it was found that interval variables such as PowerPS, Kilometer and Age which are being used for the analysis are skewed and do not follow normal distribution. In order to achieve the optimal results, all the predictors should be close to normal distribution. To perform this transformation, Transform node is used. 'PowerPs' was transformed using PWR transformation and Age was transformed using SQRT transformation
The transformed variables are shown in following figures, it can be observed that most of them are close to normal distribution now

| Source | Method | Variable Name |
|---|---|---|
| Input | Original | Age |
| Input | Original | powerPS |
| Output | Computed | PWR_powerPS |
| Output | Computed | SQRT_Age |



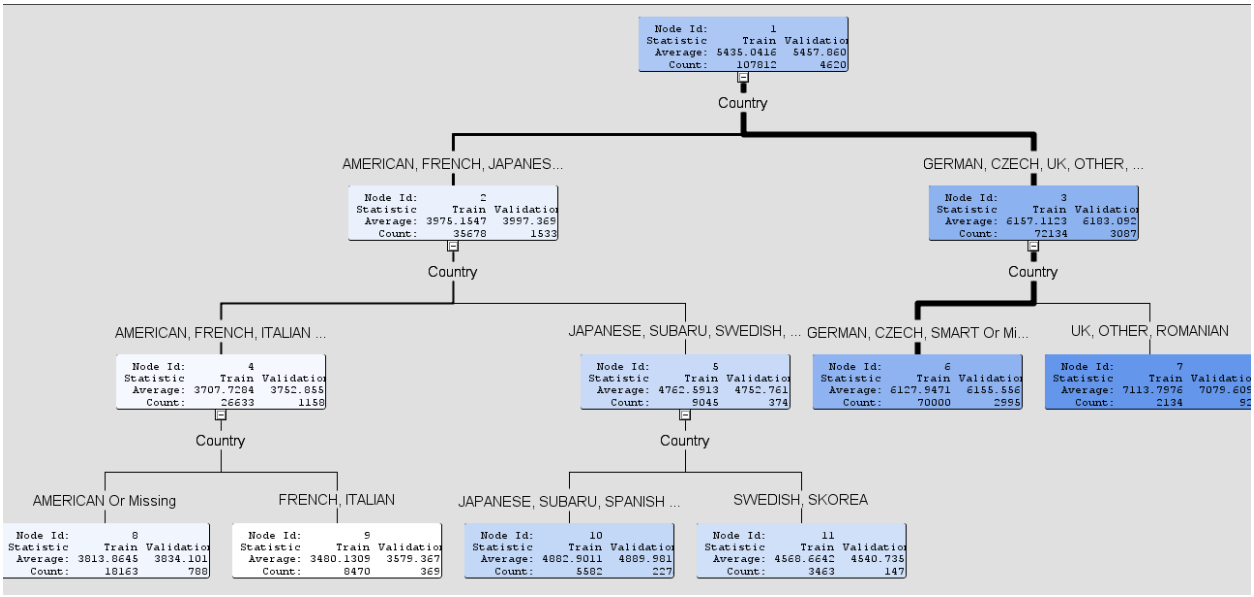




Impute Node

The data used for this analysis contained missing values for class variables. The target was to predict the interval variable which required regression analysis. To overcome the issue of missing values, impute node was used where missing data points of class variables were imputed by respective medians.

Decision Tree

One of the class variables Country which had 15 levels which signifies the country to which particular brand belongs. Had it used as it is, it would have complicated the model and would have made the interpretation of the model difficult. To overcome this problem, decision tree was used to reduce the levels of the class variable and bundle them into less categories. The following snapshot of the result shows that the 15 levels were grouped under 4 groups after running Decision Tree. These new variables were automatically named as Node1, Node2……




## Regression

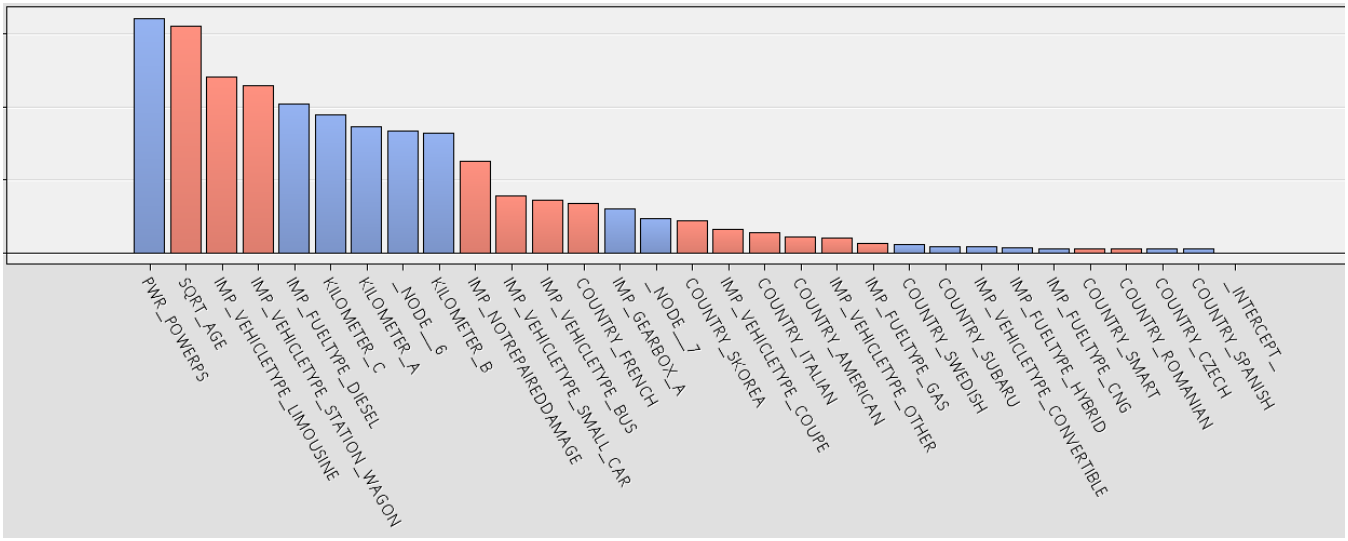
A regression node was attached to decision tree node in order to run the model. LASSO method was selected for the analysis. LASSO (Least Absolute Shrinkage and Selection Operator) penalizes the absolute size of the regression coefficients. In addition, it is capable of reducing the variability and improving the accuracy of linear regression models. The assumptions of this regression is same as least squared regression except normality is not to be assumed it shrinks coefficients to zero (exactly zero), which certainly helps in feature selection. This is a regularization method and uses 'L1' regularization. If a group of predictors are highly correlated, LASSO picks only one of them and shrinks the others to zero

## Results

The following results were obtained after regression analysis. Root Average Squared Error for Training data was USD 2739 while Root Average Squared Error was validation data was around USD 2740.

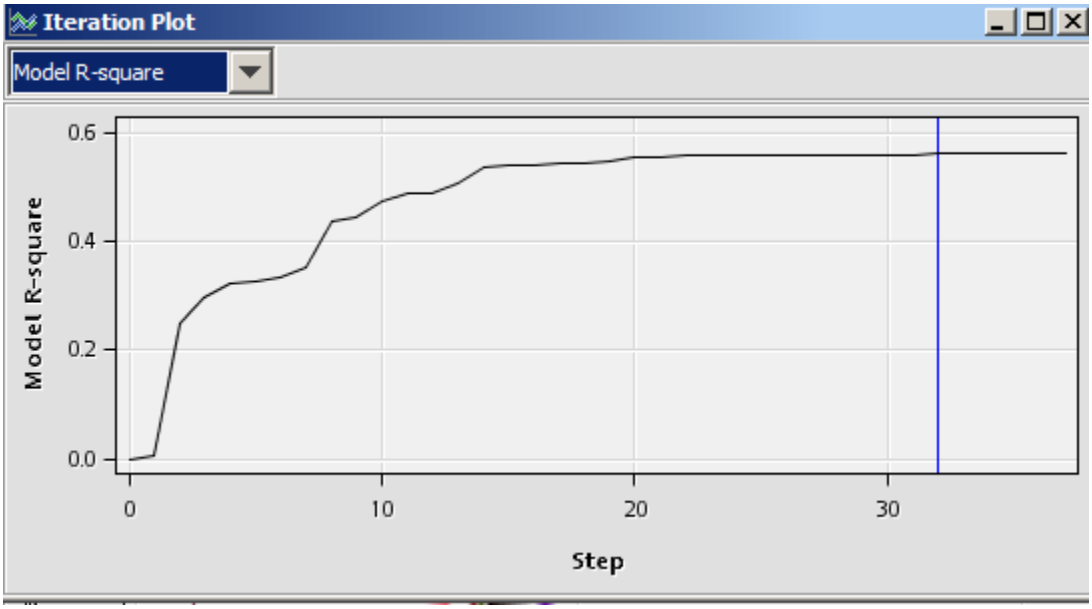
| | | | | | |
|---|---|---|---|---|---|
| price | | _ASE_ | Average Squared Error | 7503466 | 7504076 |
| price | | _DIV_ | Divisor for ASE | 107812 | 46205 |
| price | | _MAX_ | Maximum Absolute Error | 18294.16 | 18029.11 |
| price | | _NOBS_ | Sum of Frequencies | 107812 | 46205 |
| price | | _RASE_ | Root Average Squared Error | 2739.246 | 2739.357 |
| price | | _SSE_ | Sum of Squared Errors | 8.09E11 | 3.467E11 |

The following snapshot shows the effects plot of the variables. Blue bars signify positive effect while Red ones signify negative effects. It can be observed from following table that the height of Age is the highest which has the highest negative effect on the target variable




The following tables show the top 15 selected variables by the model . They are arranged according to their Standardized estimates (highest to lowest).

| Effect |
|---|
| PWR_POWERPS |
| SQRT_AGE |
| IMP_VEHICLETYPE_LIMOUSINE |
| IMP_VEHICLETYPE_STATION_WAGON |
| IMP_FUELTYPE_DIESEL |
| KILOMETER_C |
| KILOMETER_A |
| _NODE__6 |
| KILOMETER_B |
| IMP_NOTREPAIREDDAMAGE |
| IMP_VEHICLETYPE_SMALL_CAR |
| IMP_VEHICLETYPE_BUS |
| COUNTRY_FRENCH |
| IMP_GEARBOX_A |
| _NODE__7 |
| COUNTRY_SKOREA |
| IMP_VEHICLETYPE_COUPE |




R square of 56% was achieved with this model

## Conclusion

- While considering pre-used cars, power of the car play an important role deciding the value of the car.
- Brands from Germany, Czech and UK have higher resale value as compared to other country's brands

## Future Scope

- In future, estimated time of inventory can be calculated with available data, i.e. we can predict when the listed car will be sold off from the day it was advertised on the portal by knowing start and end date of the advertise

## Acknowledgement

I would like to thank Dr Miriam McGaugh, Clinical Professor, Marketing Department at Oklahoma State University for guiding and supporting me through this project.

## References

- ISLR Statistical Learning
- Practical Business Analytics Using SAS: A Hands-on Guide

## Contact Information

Your comments and questions are valued and encouraged.

Contact the author at:
Jaideep A Muley
Oklahoma State University
jaideep.muley@okstate.edu
405-780-3832

Jaideep Muley is a graduate student enrolled in Business Analytics at the Spears School of Business, Oklahoma State University. He holds a Bachelor's degree in Engineering followed by an MBA in Marketing. He has worked as a Sales Tool Analyst Intern with Perterbilt Motors Company, TX.



*This the 15th day of September 2017.*