**STA 160 - Midterm Project**

Automobile Report

**Yutian Yang (915327218)**

# **I. Introduction:**

* 1. **Background**

Automobiles are being widely used today. It has been through a rapid stage of development in the late 20th century. This dataset was obtained from the UCI machine learning repository website included many important features of different automobiles. The creator, Jeffrey C. Schlimmer collected these automobile data from the1985 Model Import Car and Truck Specifications, 1985 Ward's Automotive Yearbook, Personal Auto Manuals, and Insurance Collision Report from the Insurance Institute for Highway Safety to construct this useful dataset for my following analysis.

* 1. **Goal**

In this research, I am interested in finding the relationship between the different features of automobile and their make or brand. In the meantime, I would also like to find out what are some of the outstanding aspects each make of automobile has shown. Different techniques would be applied under various analysis through data visualization, correlation analysis, and classification. I will perform and interpret these techniques in the following report.

# **II. Dataset:**

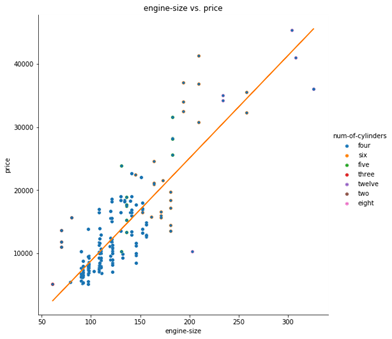
**2.1 Data Summary**

This data set consists of three types of entities: The specification of an auto in terms of various characteristics. Its assigned insurance risk rating which corresponds to the degree to which the auto is riskier than its price indicates. Cars are initially assigned a risk factor symbol associated with its price. Then, if it is riskier (or less), this symbol is adjusted by moving it up (or down) the scale. Actuaries call this process "symboling". A value of +3 indicates that the auto is risky, -3 that it is probably very safe. Its normalized losses in use as compared to other cars, which is the relative average loss payment per insured vehicle year. This value is normalized for all autos within a particular size classification (two-door small, station wagons, sports/speciality, etc...), and represents the average loss per car per year (<https://archive.ics.uci.edu/ml/datasets/Automobile>). Since this dataset contains many missing values, I manually imputed these missing values with the following method. First, I divided these missing values into two groups: numerical and categorical. For numerical missing values, I replaced them with the median value of the column they are in. As for categorical missing value, I replace them with the most common value of the column they are in.

**Attributes:**

There are 26 attributes in this dataset, including "make", "fuel-type", "aspiration", "num-of-doors", "body-style", "drive-wheels", "engine-location", "engine-type", "num-of-cylinders", "fuel-system", "symboling", "normalized-losses", "wheel-base", "length", "width", "height", "curb-weight", "engine-size", "bore", "stroke", "compression-ratio", "horsepower", "peak-rpm", "city-mpg", "highway-mpg", and "price"

**2.2 Data Visualization**

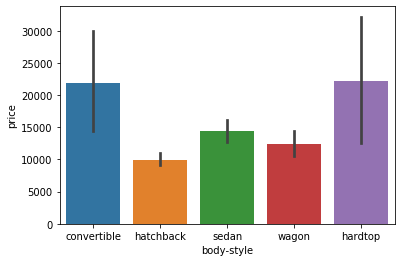
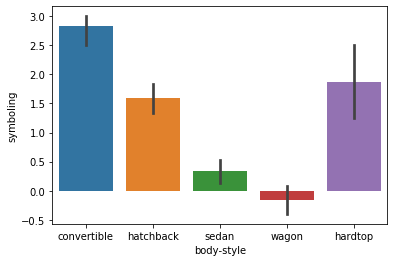


In this step, I was interested in finding the relationship between each feature. I generated a correlation plot for all numerical variables. I noticed that curb weight and length have the overall highest correlation among other variables. Focusing on the correlation between price and other variables, we can see that the larger engine size the car has, the car is very likely to have a higher price. Also, generally, the bigger the engine size, the car is likely to have a greater number of cylinders. Other important factors affecting price would include width, engine size, length, wheelbase, and horsepower. In conclusion, the bigger and more powerful your car is, it is more likely to be more expensive.

| **make** | **avg\_price** |
| --- | --- |
| **jaguar** | 34600.00 |
| **mercedes-benz** | 33647.00 |
| **porsche** | 27179.40 |
| **bmw** | 26118.75 |

| **make** | **avg\_engine\_size** |
| --- | --- |
| **jaguar** | 280.666667 |
| **mercedes-benz** | 226.500000 |
| **porsche** | 187.200000 |
| **bmw** | 166.875000 |

The above table shows the top 4 make on average price and average engine size, we can notice that the order of the make is the same on both table, which also proves that price and engine size are highly correlated.

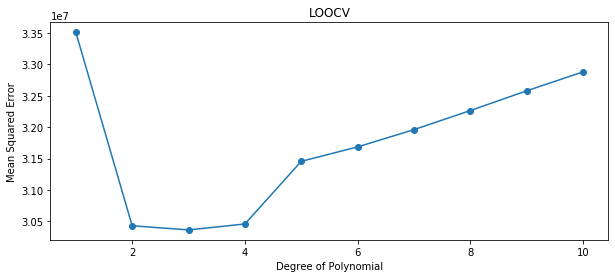
 

These two figures above show the relationship among price, body style, and car’s safety (symboling). From the first figure, we can see that convertible and hardtop are the most expensive type of automobile, but they are also the top 2 kinds of the automobile that is most unsafe. Wagon cars have the second lowest average price but are the safest automobile among other types. Noticing that on the right figure, convertible only varies within a small range, which means most of the convertible do not have a high safety feature.

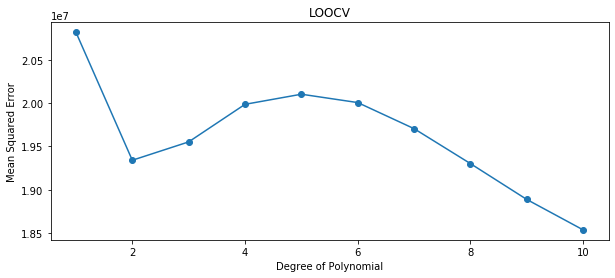
# **III. Main Analysis:**

**3.1 Leave One Out Cross Validation**

In this section, I conduct the LOOCV to analyze the data and the significance portance of the predicting variable as a 1-dimension example for analysis. Since the variable “curb-weight” and “length” have the top 2 average largest overall correlation with other variables, I picked them individually as the predicting variable for the polynomial model for cross validation.



In the above figure, “length” is used as predicting variable. The figure shows that the elbow point of this polynomial function is degree 2, which means the model will be most efficient at polynomial degree equals to 2. Under this model, the R square equals to 0.524 which means 52.4 percent of the data is explained in this model, which is not a relatively high number. All variables under this model are significant since the p-value of the parameter are equals to 0.



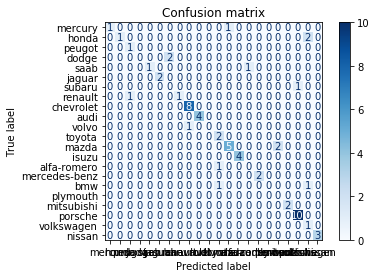
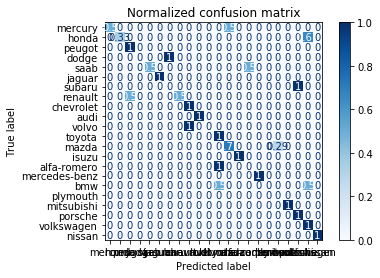
In this figure, “curb-weight” is used as predicting variable. The figure shows at degree 10, the model will have the lowest mean square error, which means the model will be most efficient at polynomial degree equals to 10. Under this model, the R square equals to 0.718 which means 71.8 percent of the data is explained in this model, which is better than the previous model. However, since the polynomial degree is 10, there is a potential possibility of overfitting the data. Besides that, all variables under this model are significant since the p-value of the parameter are equals to 0.

**3.2 Grid Search with Random Forest**

Considering this dataset contains multiple labels. Grid search with random forest method was implemented in order to perform better classification. Since this data have many categorical variables, a numeric encoding was preformed to transfer these categorical variables into numerical variables to prepare for the following model implementing. I limited the max features from 1 to all 25 features, the number of estimators to be either 3, 10, or 30, and the model can either use bootstrap or not when classifying.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **bootstrap** | **max\_features** | **n\_estimators** | **accuracy** |
| **83** | False | 3 | 30 | 0.839524 |
| **19** | True | 7 | 10 | 0.833333 |
| **104** | False | 10 | 30 | 0.832857 |

Above table shows the grid search result under the random forest method. It indicates that when bootstrap is False, and we limited the max features of the model to be 3 and choose the number of the estimators to be 30, the classification model will give the final prediction accuracy of 0. 839524 on the training dataset, which is an acceptable accuracy. I then implemented this model to run through the test dataset, I gathered the result and arranged them into the following confusion matrix.



The two figures above show the classification result. The left figure shows the actual classification result. On the right figure shows classification result that has been normalized. If the number in the blocks on the left diagonal is closer to 1, means it is correctly classified. Calculated from the table and figure indicates, the accuracy of this model which grid search suggests is optimal turns out to be 0.7742, which is about 6.5 percent lower than predicting the training dataset.

# **IV. Conclusion:**

This study provides multiple ways to predict and analysis the data under different circumstances. Through data visualization, we concluded that the bigger and more powerful the car is, the more likely it is going to be expensive. Also, comparing the barplot among “price”, “body-style”, and “symboling”, we conclude that while convertible and hardtop are the most expensive type of car to purchase, they are also the least safe car regarding to the insurance risk rating. For 1-dimensional analysis, I implemented the LOOCV with polynomial model to predict the price of automobile. The result suggests when polynomial degree equals to 10 when using “curb-weight” as predicting variable, the model will perform most efficiently. For K-dimensional analysis, I conducted grid search with random forest method and multiclass classification. The grid search with random forest provides a good accuracy with training dataset. The accuracy is about 6.5% lower when passing the model through the testing dataset.