

Final project

Ban210BAN210NAA Predictive Analytics



SUBMITTED TO: DR. GULSUN DEMIREZEN

BY: SAI MOHAN TIRUNAGARI

ID: 147150213

Auto\_mpg data set

**BAN 210 Predictive analysis – Fuel efficiency predictive model**

This final project aims to build predictive models on the auto\_mpg data set compare which model is best suited to predict the fuel efficiency of a car.

**Data Dictionary**

The auto\_mpg data is obtained from the UCI machine learning repository. "The data concerns city-cycle fuel consumption in miles per gallon, to be predicted in terms of 3 multivalued discrete and 5 continuous attributes." (Quinlan, 1993)

“This dataset was taken from the StatLib library which is maintained at Carnegie Mellon University. The dataset was used in the 1983 American Statistical Association Exposition.”

Using the variables, the goal is to find how thy effect the target variable and how it can be useful to predict the fuel efficiency.

|  |  |  |  |
| --- | --- | --- | --- |
| **Variable** | **Type** | **Measurement** | **Description in SAS Enterprise** |
| mpg | Numeric | Continuous | Target Variable |
| cylinders | Numeric | Multi-Valued  Discrete | Input |
| displacement | Numeric | Continuous | Input |
| horsepower | Numeric | Continuous | Input |
| weight | Numeric | Continuous |  |
| acceleration | Numeric | Continuous | Input |
| model\_year | Numeric | Multi-Valued  Discrete |  |
| origin | Numeric | Multi-Valued  Discrete | Input |
| car\_name | Char | Nominal | Input |

Summary of statistics

Table

Description automatically generated

We can see form the above snip that there are no missing values in the data set and the target variable is “mpg”.

**Exploratory Data Analysis**

Chart, bar chart

Description automatically generated

The graph generated using the “statexplore” node, we can see significance of variables or influence of the have on mpg variable. The graph illustrates that the variables “weight”, “displacement”, “cylinders” have high significance and “car\_name” has the least. Hence, we can ignore the variable ‘’car\_name” to conduct analysis.

Chart, bar chart

Description automatically generated

The above graph shows the positive and negative correlations between the variable and “mpg”.

**Graph node**

Graphical user interface, bar chart

Description automatically generated with medium confidence

Chart, histogram

Description automatically generated

The snap snip gives us the information of frequency distribution of each variable. We can see that the variable acceleration is normally distributed, but weight and cylinders ae not normally distributed. Hence, we transform them before conducting logistic regression.

**Data partition**

The data set is divided into 2 parts:

Training – 80%

Test – 20%

Table

Description automatically generated

**Transformation node**

Chart, bar chart

Description automatically generated

The variables are transformed using log10 transformation technique.

A screenshot of a computer

Description automatically generated with low confidence

**Variable selection node**

Graphical user interface, table

Description automatically generated

In this node, we check for variables with less significance. From the above snip we can see that variables acceleration, car\_name and horsepower have less significance because of small R-square value.

**Model Comparison**

**Logestic** **regression**

Table

Description automatically generated

The above image shows the results of the linear regression. This model has an average squared error of 6.935433. For a regression lesser the error, the better it performs.

**Decision tree**

The below picture is of decision tree model. It has 2 branches and 6 levels of depth. The subtrees undergo evaluation by average square error and cross validated. We can see the error value is 6.408945 also and the model also gives the information that the carriable cylinder has the highest significance compared to all other variables.

Word, timeline

Description automatically generated

**Neural Network**

Table

Description automatically generated

Form the above image, it is depicted that the average squared error for neural network model is 0.567727 for the variable mpg, which is the target variable.

**Model comparison**

Graphical user interface, application

Description automatically generated

Chart

Description automatically generated

After comparing the results it can be deduced that neural network model performs or best fits as a prediction model because it has the lowest Average squared error, which is 0.567727, compared to logistic regression model with error value 6.9354 and decision tree with 6.408945.

Diagram

Description automatically generated

Above is the picture of all the nodes used for building the models and comparing them in SAS enterprise miner.

**Declaration**

I, Sai Mohan Tirunagari declare that the attached assignment is my own work in accordance with the Seneca Academic Policy. I have not copied any part of this assignment, manually or electronically, from any other source including web sites, unless specified as references. I have not distributed my work to other students.