

Online News Popularity

Project Report - Batch 03: Group 08



Amarendar Alwala

Amitesh Eti

Vigneshwaran Kannan

Kalluri Siva Kumar Reddy

Yumin Dong

**Table of Contents**

|  |  |  |
| --- | --- | --- |
| **S.No** | **Header** | **Page No.** |
| **I** | **Introduction** | **2** |
| **II** | **Problem Statement** | **2** |
| **III** | **Dataset Description** | **3** |
| **IV** | **Project diagram(SEMMA)** | **3** |
| **1.** | **Data Pre-Processing**   1. **StatExplore Node** 2. **Multiplot Node** 3. **Filter Node** | **4** |
| **2.** | **Modify**   1. **Skewness** 2. **Binary Conversion of Target** | **6** |
| **3.** | **Data Partition Node** | **7** |
| **4.** | **Prediction Models**   1. **Neural Networks** 2. **Decision Tree** 3. **Logistic Regression** | **7** |
| **5.** | **Model Comparison** | **10** |
| **6.** | **Business Purpose** | **13** |
| **7.** | **Conclusion** | **14** |

**Introduction**

The exponential growth of Internet and web, led to an easy and fast spread of information across the world. Due to the proliferation of online reading, the prediction of online news popularity prior to news’s release is becoming a trendy research topic. This is useful for authors, content providers, advertiser and activists, because they are able to modify an article and the manner of its publication, according to prediction results. In such way, the article published in a particular website will be well propagated among public readers in a short time.

Popularity is measured by considering the number of shares in web and social networks such as Facebook, twitter. News article is very time sensitive in nature. News reader changes their preference all the time. The content of a new article need to match with reader’s preference in order to increase share of market. Using an extensive set of derived features such as keywords, digital media content, earlier popularity of news referenced in the article, we predict the popularity of various news articles.

This project uses the predictive analysis methodology to identify potential patterns in article features of news articles. And using prediction tools is used to find reasons behind popularity of a news article. Our study result will help author, content providers to precisely modify articles using popularity factors before articles’ release.

**Problem Statement**

The major aim of this project is to discover key drivers/factors of online news popularity. Through this project, the following information will be given based on our analysis.

1. Whether a news article will be popular
2. What makes an article become popular

The results will be widely used by various decision-making groups

1. Authors and content providers who want to make article popular
2. Website editor who want to attract more advertisers
3. Activists who want to expand influence of their opinions
4. Advertiser who is looking for a website to promote products

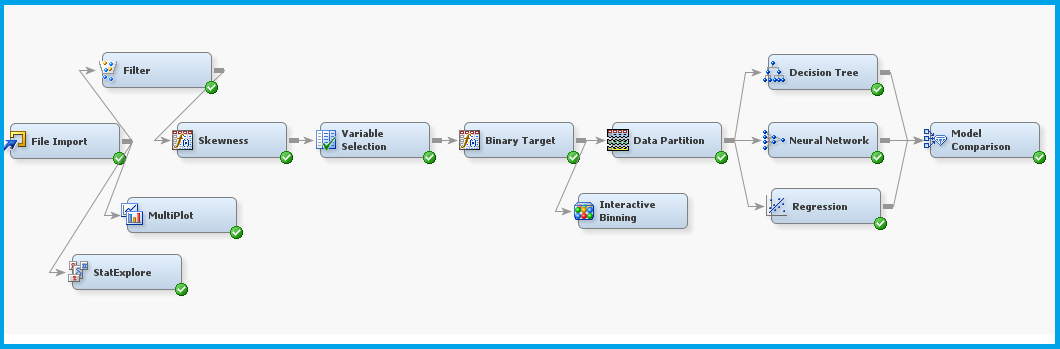
**Dataset Description**

Our project uses second-hand data retrieved from [UCI Machine Learning repository website](https://archive.ics.uci.edu/ml/datasets/Online+News+Popularity). This dataset encapsulates a varied set of features about articles published by Mashable (www.mashable.com) over a period of two years. This multivariate dataset has 39644 instances and covers 61 attributes, including 58 predictive attributes, 2 non predictive and 1 goal (Target) field. The two non-predictive attributes (URL and time delta) of the dataset do not affect the outcomes of analysis.

*Please find the attributes file attached below:*



**Project diagram (SEMMA)**

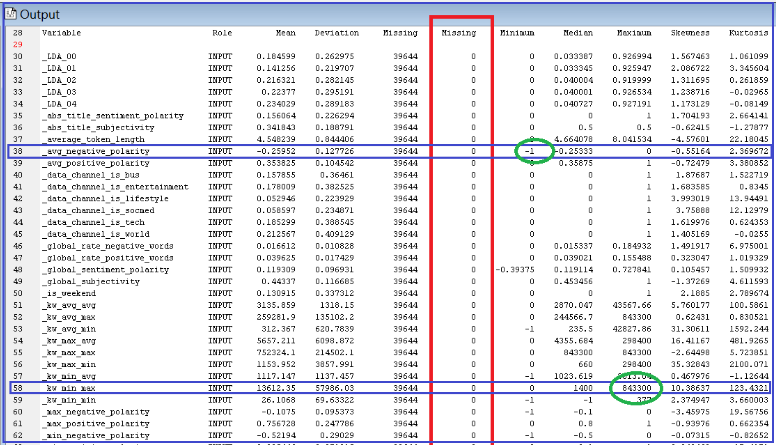


1. **Data Pre-Processing**

The dataset chosen is in csv format, hence File Import node is used to import the dataset.

1. **StatExplore node**

**Missing values result:**



**Key Observations:**

1. The results from StatExplore node showed that there are no missing values in any of the variables.
2. The minimum value for the variable “avg\_negative\_polarity” is -1. The attribute describes the polarity of the negative words. Therefore this is not erroneous and can be retained.
3. The maximum value for the variable “kw\_min\_max” is 843300 and describes the min. shares of the best keyword of the news article. Therefore this is also not erroneous.
4. **Multiplot Node**

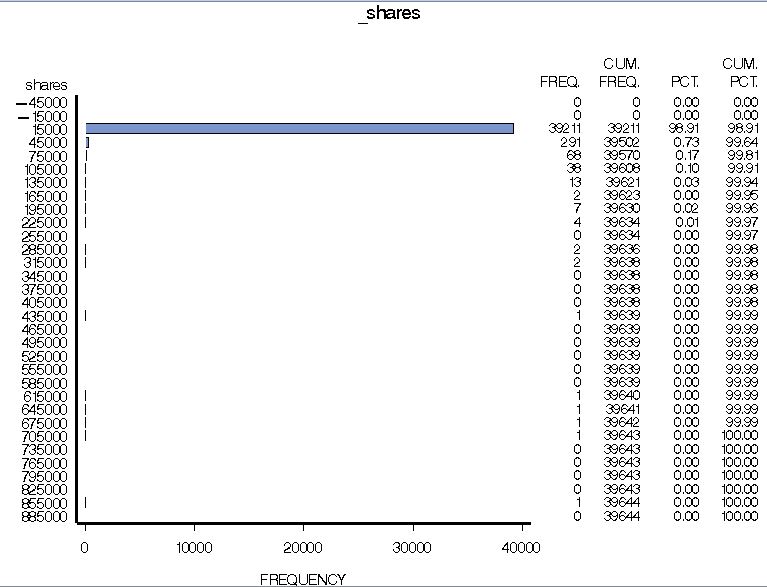
**Key Observations:**

The results from Multiplot node showed that there are no unknown values for any of the variables. After carefully observing the Bar charts of all variables no extreme values are identified.

1. **Filter Node :**

As per the overall data, target variable “\_shares” ranges from 1 to 843300. But nearly 97% of the records are within the 15000 range. Below is the stats of the same.

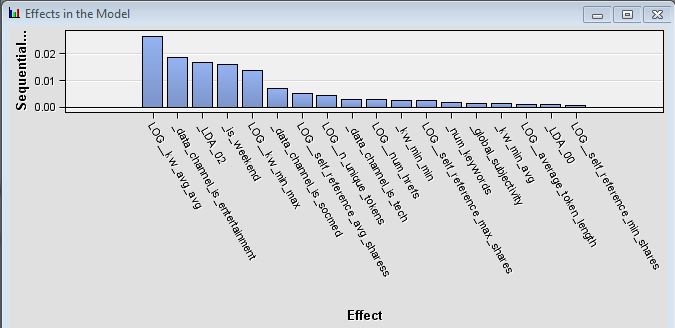
|  |  |  |
| --- | --- | --- |
| Share’s Range | Observations | % Total |
| < =15000 | 38380 | 96.82 |
| > 15000 | 1264 | 3.18 |

****

**Key Observations:**

1. By filtering the observations >15000 i.e. 3.18% of overall data, “\_shares” upper limit is decreased to 15000 against the initial value 843300.
2. This helps us in better prediction model as the range of shares variable is minimized.
3. **Modify**
4. **Transform Variables node – Skewness**
5. From the Stat Explore node we have observed both positive and negative skewness for few variables.
6. The variables with skewness greater than absolute value of 1 have been transformed using Log transformation.
7. Hence the skewness of target variable is minimized for better Model Prediction.
8. **Variable Selection node**

**TOP 18 Variables Results:**

****

**Key Observations:**

1. The results from the Variable Selection node showed the top 18 variables which will contribute the best in predicting the target variable.
2. R square selection criterion with minimum value of 0.0005 is used to arrive at the top variables.
3. **Transform variable node – Target variable conversion to binary**
4. In the Transform variables node, the target variable which was interval type, is converted to binary using formulas.

Formula: LOG\_shares >= 7.24

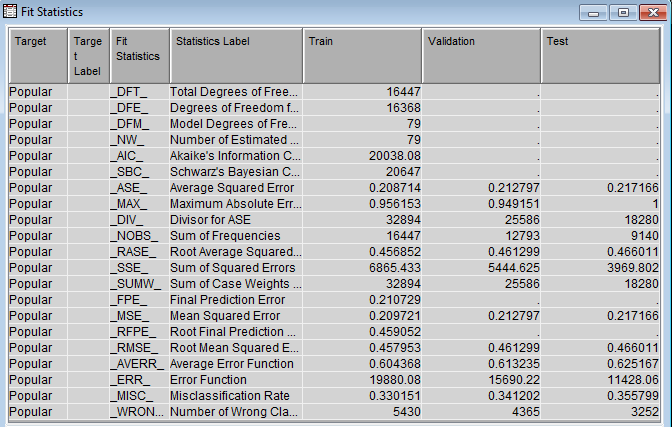
1. Using the Interactive binning node, two equal target variable bins are created with 7.24 cut-off.
2. **Data Partition Node**

**Key Observations:**

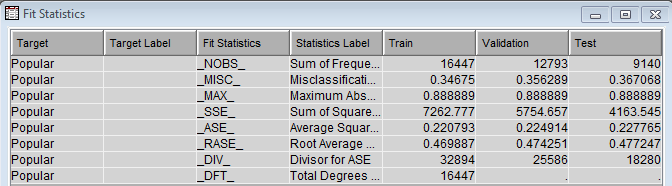
The data set consists of 38380 rows of data. This node provides mutually exclusive datasets and data is partitioned into training data, validation data and test data as below:

|  |  |
| --- | --- |
| Type | No. of Observations |
| Train (60%) | 23028 |
| Validate (20%) | 7676 |
| Test (20%) | 7676 |

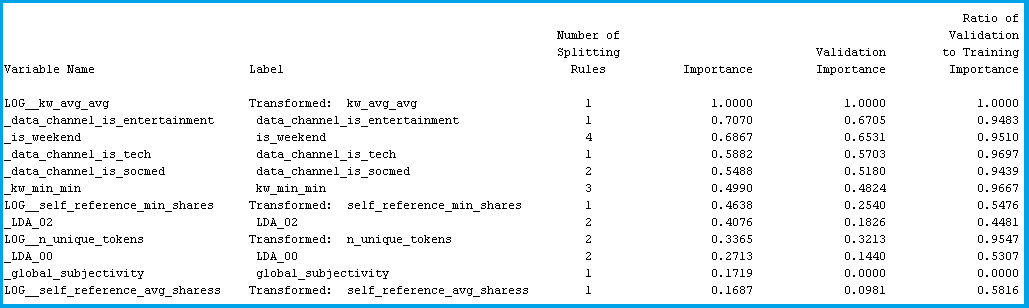
1. **Prediction Models**
2. **Neural Network :**
3. The fit statistics for the Neural Network shows the Average Squared Error for validation data is 0.221797
4. Misclassification Rate is 0.341202 and Akaike’s Information Criterion for training data as 20038.08.
5. The number of estimated weights chosen is 79. The network architecture used is Multilayer perceptron-a feedforward artificial neural network that maps sets of input data onto a set of appropriate outputs.



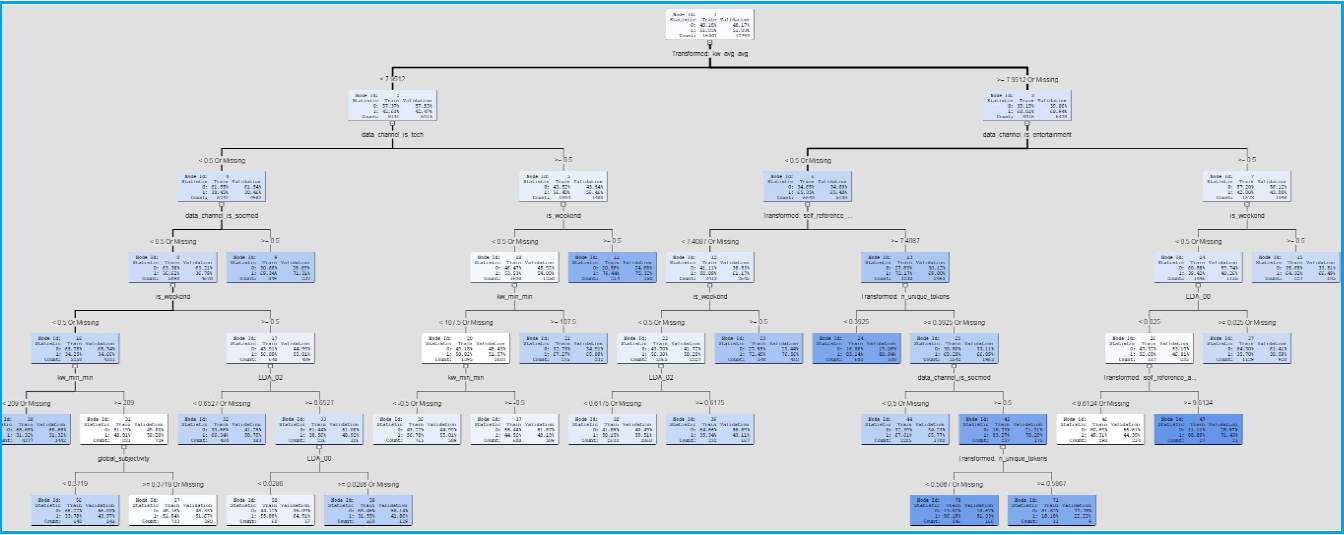
1. **Decision Tree :**
2. There are totally 5 leafs in the tree and Maximum depth of the tree = 6.
3. The fit statistics for the Neural Network shows the Average Squared Error for validation data is 0.220793
4. Misclassification Rate for the decision Tree is 0.356289



**Top Significant variables**:



**Decision Tree Diagram:**

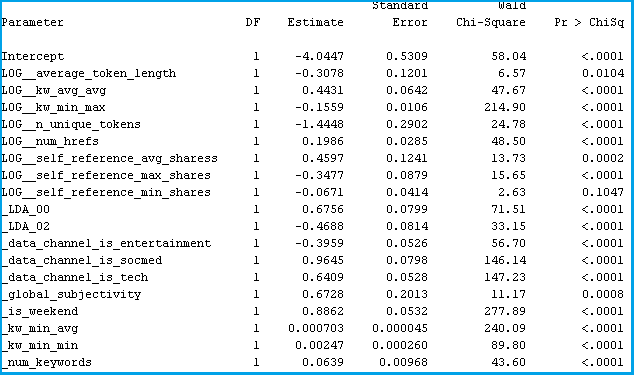


1. **Regression**

The target variable in our dataset is converted to Binary, we tested the dataset using the Logistic Regression model. The dataset was tested using all the three selection models:

|  |  |
| --- | --- |
| **Model type** | **Misclassification Rate** |
| Forward Regression | 0.3532 |
| Backward Regression | **0.3527** |
| Stepwise Regression | 0.3531 |
| None | 0.3528 |

From the above table, the misclassification rate for backward regression is least and hence it is selected.



**Top 9 significant variables:**

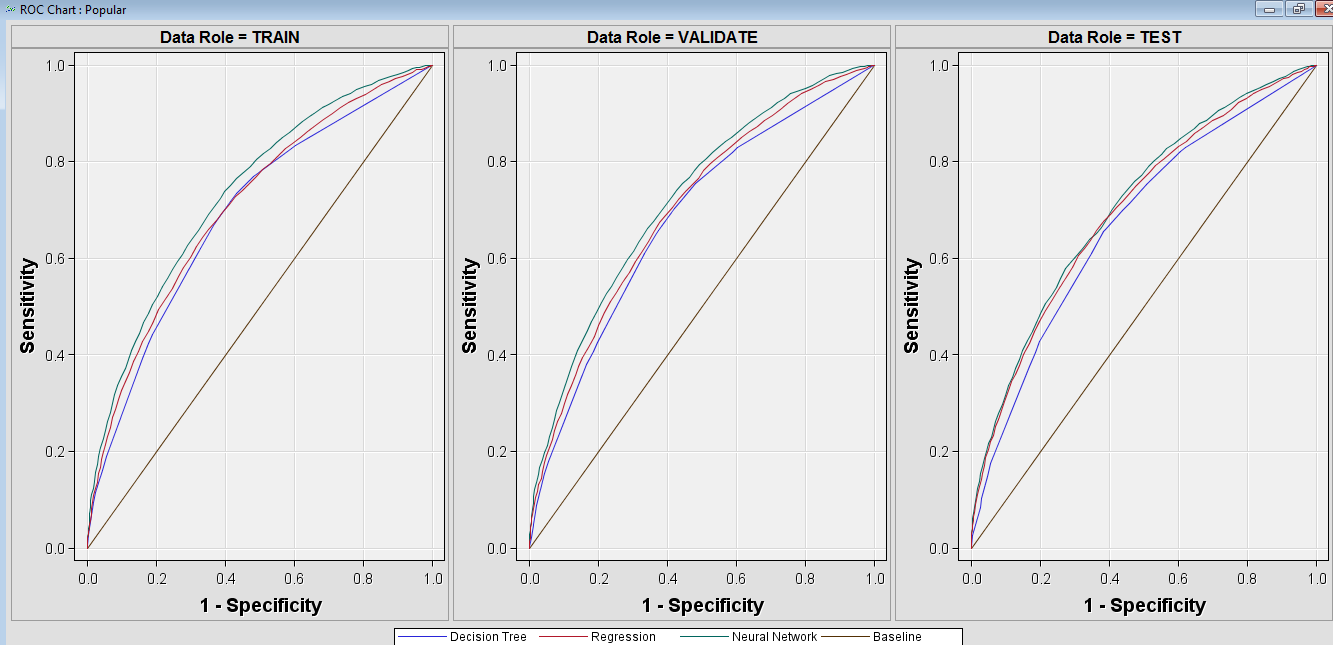
1. LOG\_kw\_avg\_avg
2. \_global\_subjectivity
3. \_data\_channel\_is\_socmed
4. \_is\_weekend
5. \_data\_channel\_is\_tech
6. LOG\_n\_unique\_tokens
7. \_LDa\_00
8. LOG\_self\_reference\_avg\_shares
9. LOG\_self\_refrence\_max\_shares
10. **Model Comparison**

Receiver Operating Characteristic curve (ROC) is a plot of the true positive against the false negative rate for the different possible cut point of a diagnostic test.

The closer the curve follows the left-hand border and then the top border of the ROC

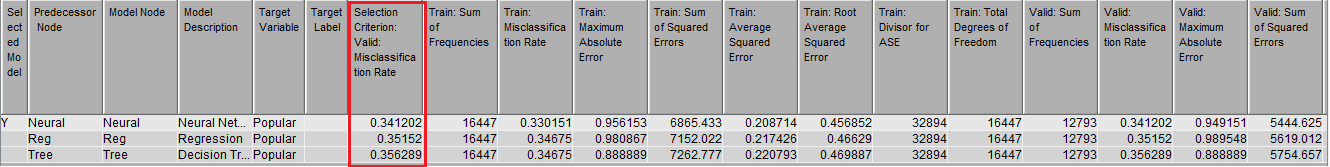
Space, the more accurate the test. The area under the curve is a measure of test accuracy.

From the below ROC, neural network and Regression are doing fairly good in comparison with decision tree.



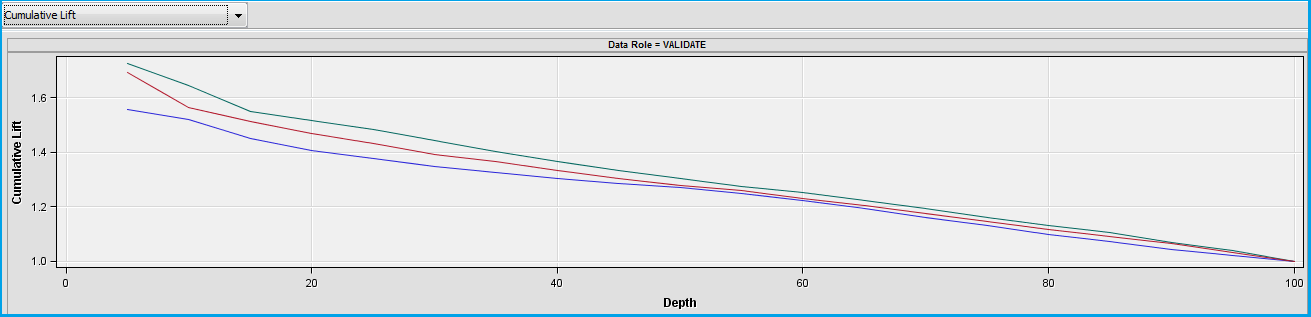
**Misclassification rate for Different models:**

|  |  |
| --- | --- |
| **Model type** | **Misclassification Rate** |
| Neural Network | **0.341202** |
| Regression | 0.35152 |
| Decision Tree | 0.356289 |



From the above table, the Misclassification rate for the neural network is 0.341202 which is least among three models. Therefore neural network best suits this dataset.

**Cumulative Lift:**

****

* High value of cumulative lift suggest that the model is doing good job separating the popular and non-popular articles
* It can be seen that the cumulative lift is around 1.75 in first decile for Neural Network. The entire graph shows Neural Networks as a better fit.

**Confusion Matrix:**

To calculate the accuracy rate for each model, we draw confusion matrix to determine the number of true positives and true negatives.

**Neural Network:**

|  |  |  |  |
| --- | --- | --- | --- |
|  | | **Model Prediction** | |
| Non-popular | Popular |
| **Actual** | Non-Popular | 9072 | 4780 |
| Popular | 5015 | 10373 |

Accuracy rate = (9072+10373) / (9072+4780+5015+10373)

= 66.5%

**Regression:**

|  |  |  |  |
| --- | --- | --- | --- |
|  | | **Model Prediction** | |
| Non-popular | Popular |
| **Actual** | Non-Popular | 8791 | 4904 |
| Popular | 5296 | 10249 |

Accuracy rate = (8791+10249) / (8791+4904+5296+10249)

= 65.11%

**Decision Tree:**

|  |  |  |  |
| --- | --- | --- | --- |
|  | | **Model Prediction** | |
| Non-popular | Popular |
| **Actual** | Non-Popular | 8196 | 4370 |
| Popular | 6521 | 10783 |

Accuracy rate = (8196+10783) / (8196+4370+6521+10783)

= 63.53%

From the above calculations, we observe that the accuracy rate is higher for **Neural Network model**. Hence we can conclude that **Neural Network** is most accurate for the given dataset.

1. **Usefulness of Model from Business Perspective:**
2. **Authors:** Mashable authors and content providers can use this report to modify articles, using a result report from variable selection analysis. For example, what the most suitable is changed by the author is to increase the average number of key words in their article prior to its publication.
3. **Website Authority:** Marshable website editors can also use this result to change the manner of its publication before it publishes, for example publishing the article in weekend. And editors can encourage more advertisers to put their ads in this page where this article displays, because this article is going to popular according to this predicting report.
4. **Advertisers:** Advertiser who needs a page in Marshable website to show their products or services can use their predicting report to know which the best page to put ads, in order to earn the profitability.
5. **Activists:** Meanwhile activists may consider publishing their articles at weekend which is a factor closely associated with the popularity of their article, in order to have more reader and expand influence of their opinions.
6. **Conclusion:**
7. After identify missing value and unknown value, transforming skewness of a variable, we discover there are 18 variables closely related to the popularity.
8. Adopting binary target node, we define LOG\_share bigger or equal to 7.24 as popularity.
9. Decision tree, neural network, and Regression model based on data partition are used to find the best predictive model for this dataset
10. According to misclassification rate and cumulative lift, Confusion matrix tests the accuracy of each model is determined.
11. As a result, neural network is tested as the most efficient model, with the lowest 34.12% misclassification rate, the highest 66.5% accuracy rate.