**OPIM 5604 – Predictive Modeling: Team Project**

**“The work contained and presented here is my work and my work alone.”**

**Team Members:** Harsh Bansal, Dharini Balasubramanian, Jayachandra Balusu, Krupa Bhatt, Yoghendar Bhosrekhar

**INDEX:**

1. [Problem Statement](#Problem_Stetement)
2. [Potential Business Value](#Business_val)
3. [Executive Summary](#Executive_Summary)
4. [Variables Classification / Data Dictionary](#variable_Classification)
5. [Exploratory Data Analysis](#Explore_data)
6. [Variables Recode](#variables_recode)
7. [Outlier Analysis](#outlier_analysis)
8. [Distribution of Variables – Standardization and Transformation](#standardization_Transformation)
9. [Missing Data Pattern Analysis](#Missing_data)
10. [Principal Component Analysis, Correlation](#PCA)
11. [Other Analysis (Predictor Screening)](#Factor_Analysis)
12. [Validation and Test Dataset](#Validation)
13. [Models – For Prediction of Shares](#Continuous_model)
14. [Bin creation for Classification Analysis](#bin)
15. [Models – Classification](#model_classification)
16. [Comparison of Model Results](#comparison)
17. [Problems Faced](#prob_faced)
18. [Inference/Key Learnings and Recommendation](#inf_keylearn)

**Problem Statement**:

This dataset summarizes a heterogeneous set of features about articles published by Mashable website in a period of two years. The goal is to predict the number of shares in social networks (popularity) and also to classify or rank the news.

**Potential** **Business Value**:

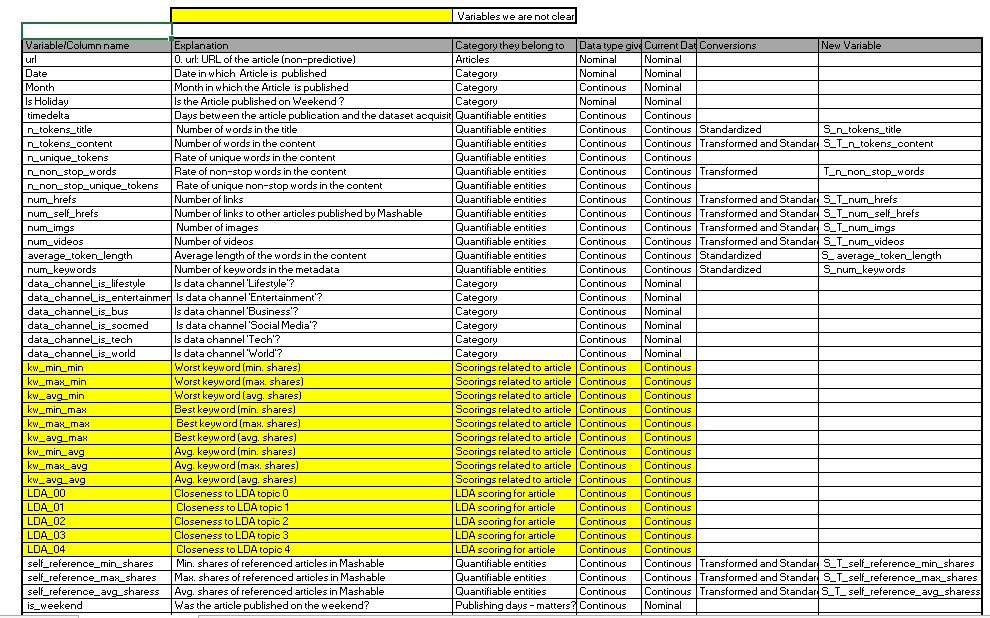
1. Recommendation systems and media advertising
2. Better Search engine optimization for popular articles
3. Better understanding of readers likes and dislikes (sentiment analysis)
4. Real-world outcome prediction E.G. Financial trends
5. Online Marketing
6. Better placement of articles and advertisements in a website

**Executive Summary:** <[<Index](#Index)

1. Considered ~38,000 records – news links and the respective details scrapped from Mashable.com website to predict the popularity of news based on their shares
2. Modified the variable classifications for some variables depending on our understanding of the data dictionary
3. Performed recode and observed the pattern. We noticed that the variable column “Token\_Content” has some values as 0. This could be due to data errors caused while scrapping information online. We decided to determine the impact of these records to our response variable and saw that they have very minimal influence. We deleted these outlier data from our dataset.
4. Performed missing data pattern analysis and noticed that there were no missing data
5. Analyzed the distribution of some of the continuous variables to determine the outliers and the general pattern
6. Performed Principal Component Analysis to understand the correlation between variables. Observed that mostly the variables were least correlated to each other
7. Transformed some variables and created new columns to store those transformations
8. The purpose of some variables such as LDA\_Topic0, LDA\_Topic1, Kw\_Max\_Max, Kw\_Max\_Min etc. have not been explained at all in the data dictionary. Even though we explored the dataset and tried to understand the relationship between these variables and the response variable, we feel that these columns are unreliable and hence performed modeling once excluding these variables and later with them as part of the input to the model
9. We performed a number of exploratory data analysis to understand the data, the relationship between variables in depth
10. Created a validation column (60 % training, 40% validation) for training the model
11. In order to perform classification analysis, we binned the shares and formed another ordinal variable that describes the shares as very low, low to very high
12. We have applied different Modeling techniques, such as Decision tree, neural net, bootstrap forest, least squares, step wise linear regression, logistic regression and also a combination of different below stated model results (ensemble)

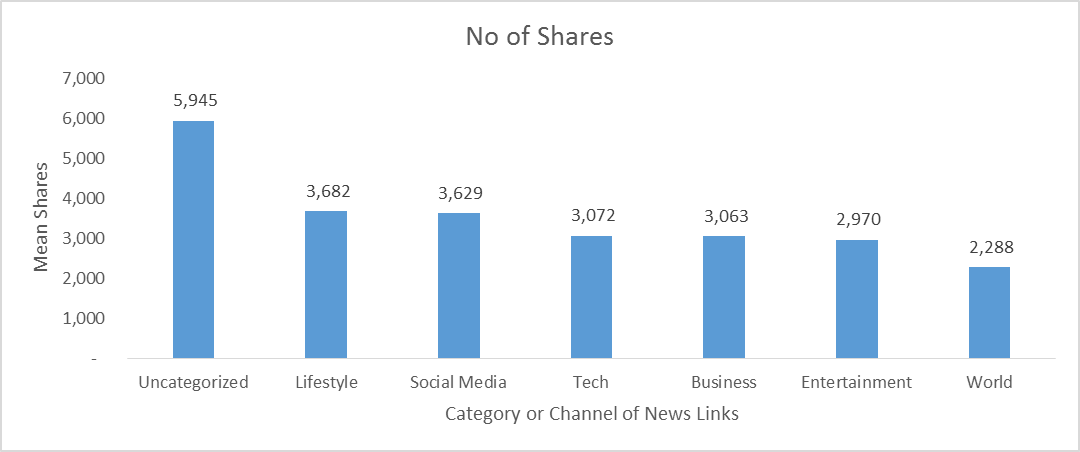
**Variable Classification and Data Dictionary**: <[<Index](#Index)

We scrapped the date of publishing of the articles from each URL and included the date variable in the dataset. Apart from that we tried to analyze/explore the relationship between various variables and wherever necessary changed the classification of columns (continuous to categorical and vice versa), removed outliers, standardized and transformed variables



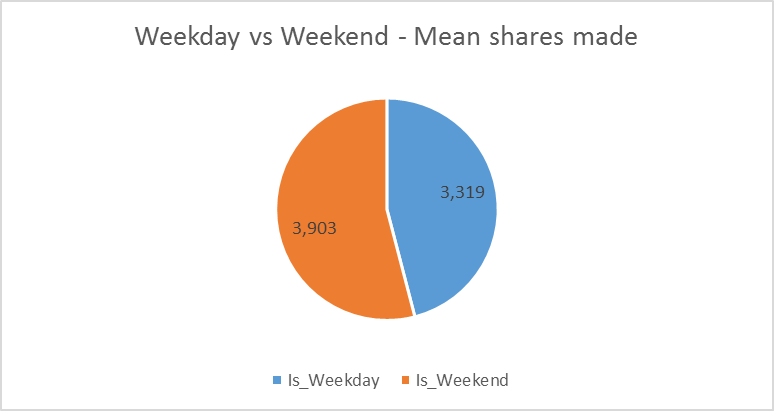
**Exploratory Data Analysis:** <[<Index](#Index)

**Does Category of news affect is popularity?**



Lifestyle related articles have the maximum no of shares followed by Social Media whereas articles belonging to world channel have the lowest no of shares. Also, there are articles which do not belong to any of these categories and have been termed “Uncategorized” and we may have to look at them individually during model creation.

**Does the popularity of an article get affected by the day it got published ?**

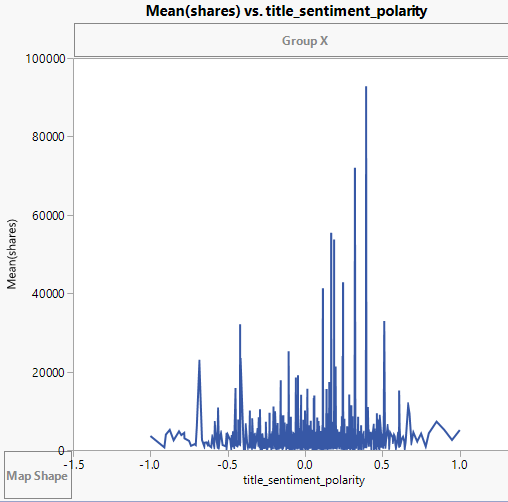


No of Shares are more for articles that are posted on weekends, when compared to articles posted on weekdays.



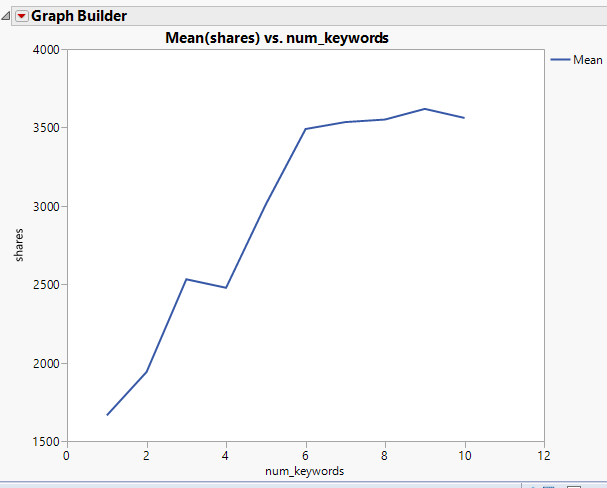
The Distribution of shares across the week shows that articles published on weekends are more popular than weekdays, followed by Mondays and Wednesdays.

**Do readers prefer news with positive or negative intonation?**

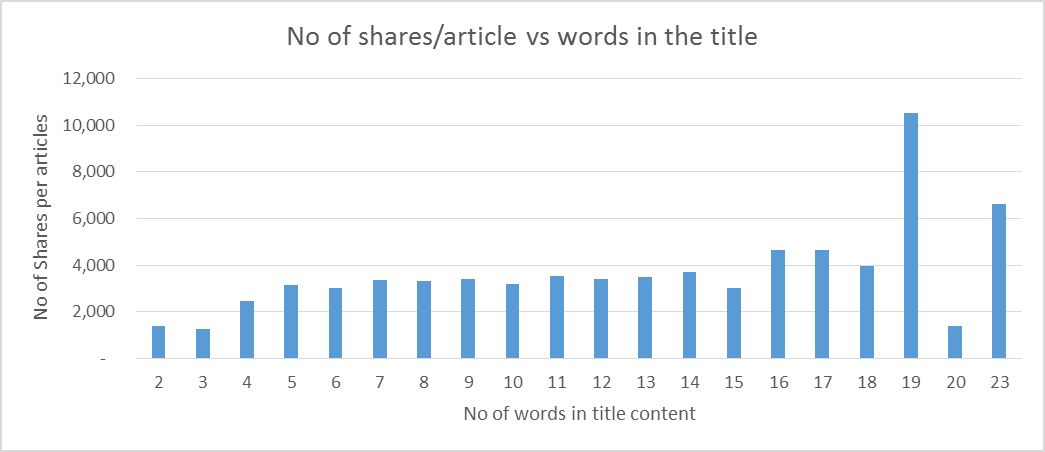


In general, articles with less negative intonation has more shares

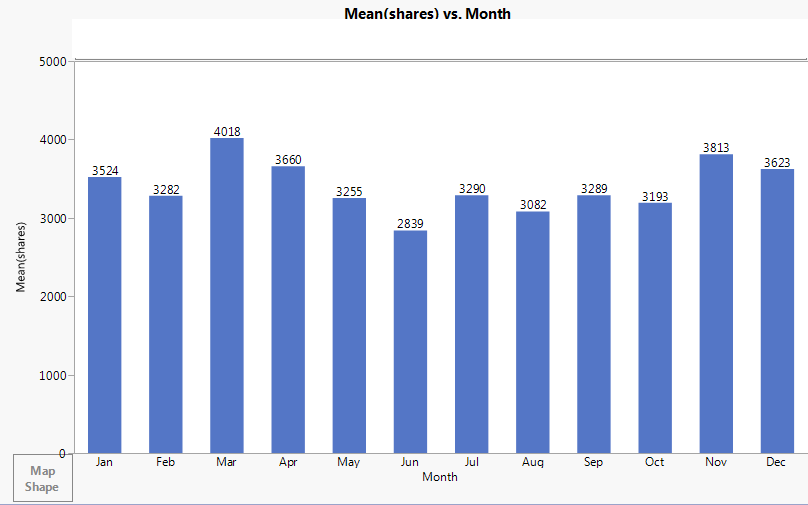
**Does it matter whether links, keywords are mentioned in the article?**



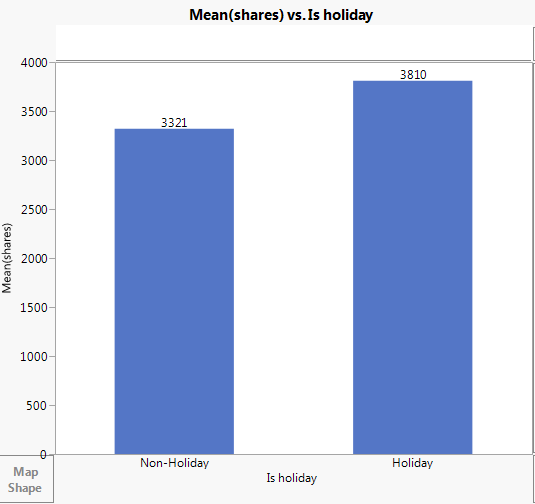
More the number of keywords, more popular the article is – i.e. the shares are more



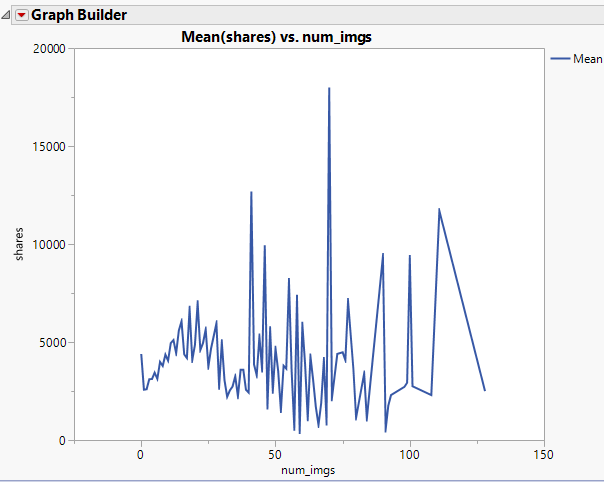
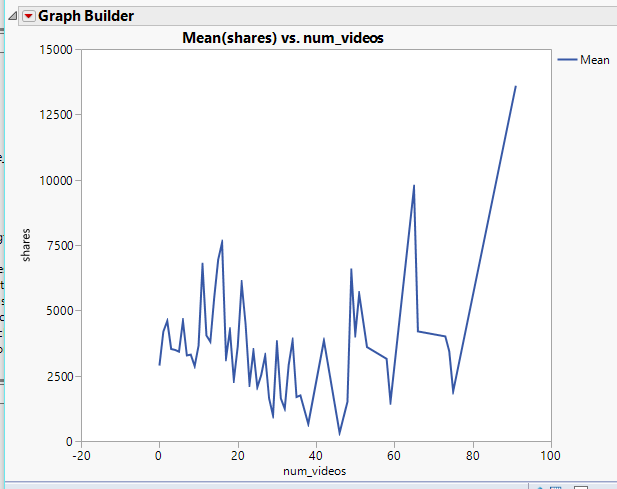
Maximum no of shares have been made to articles containing 19 words. On an average, articles having words from 7 -14 in the title has similar sharing pattern



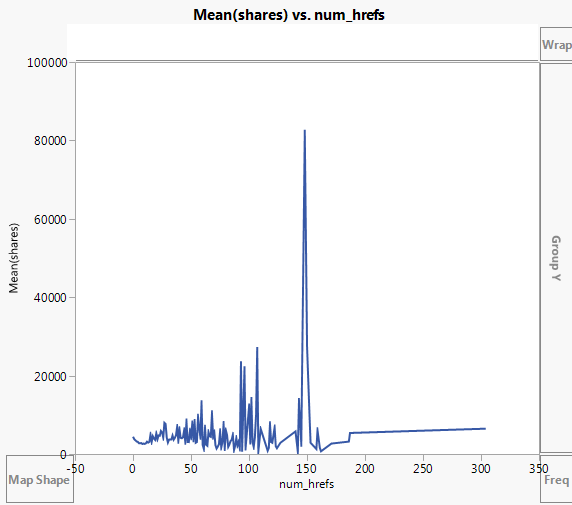
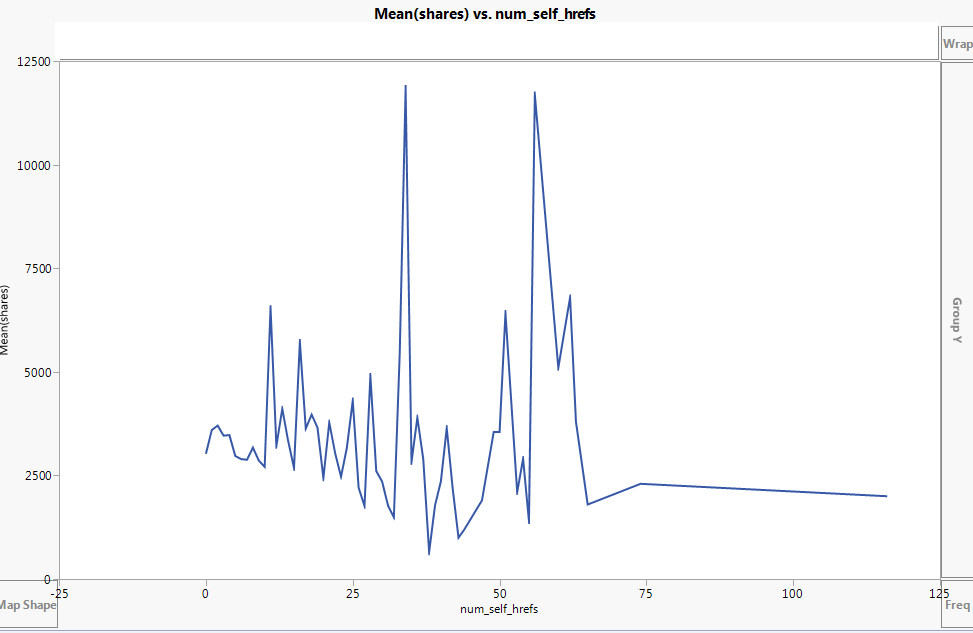
Articles published in March are having more shares followed by November and those published in June are having least number of shares.



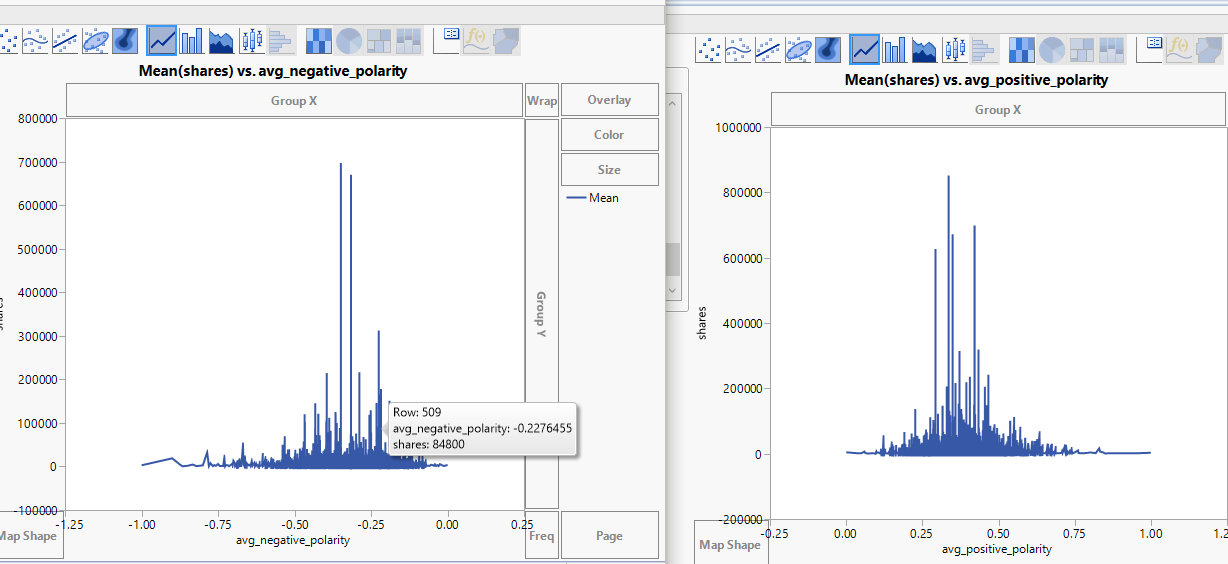
Numbers of shares are more for articles published on a holiday compared to a non-holiday. (Holidays include weekends and public holidays)



In general, more the no of videos, more popular the article is in the mashable.com website

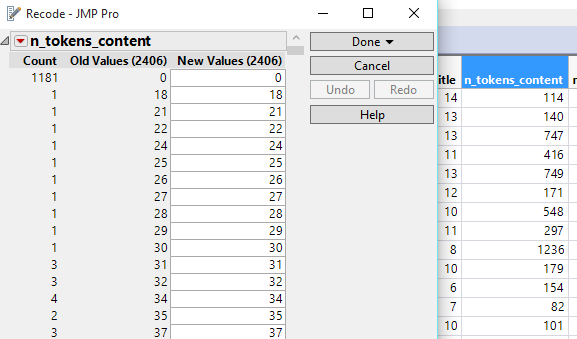
We can conclude from here that the reference links have very less impact on the popularity of an article. Articles with less references seems very volatile where in articles with around 200+ references do not show much variation at all



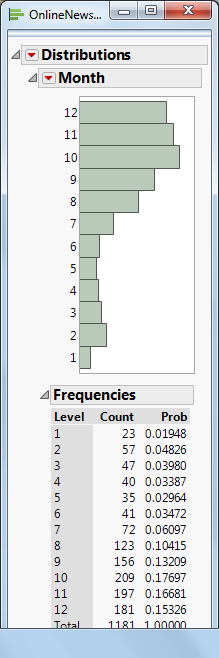
In general, Positive articles have more popularity as compares with negative articles having the same impact

**Variables Recode:** <[<Index](#Index)

Perform recode of columns to understand the distribution of values, for each variable. Select each **column -> cols -> utilities -> recode**



There are 1181 values missing for the variable - "No of words in the Article"   
Since article without any words seems less probable, we have decided to delete these rows from our dataset. We can also impute by replace the 0 values with the mean value for this variable



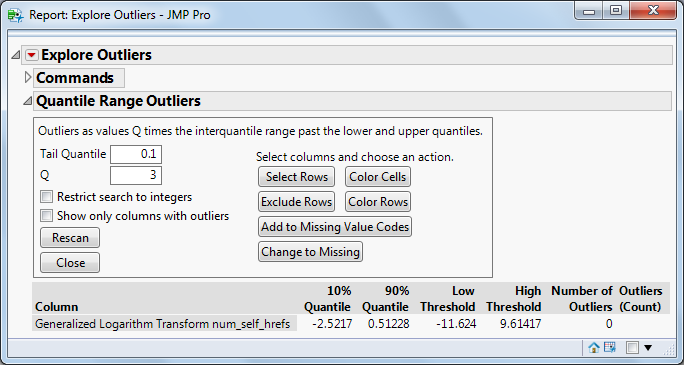
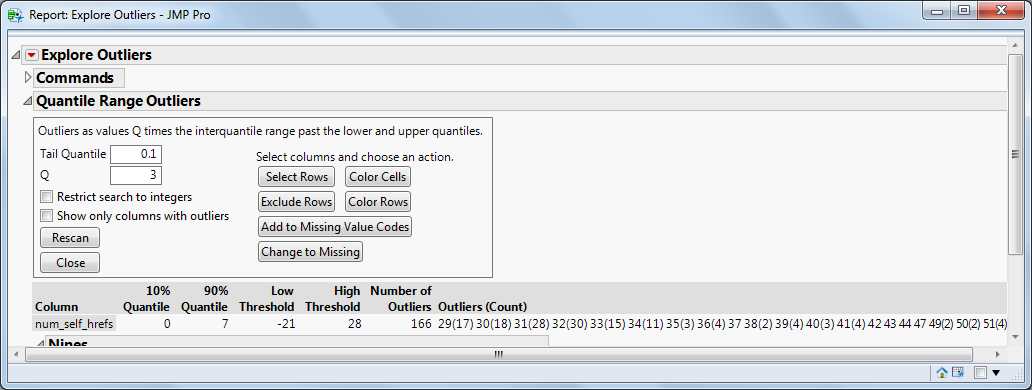
We determined the months when the article content was recorded as 0 and noticed that mostly it happened in the months from Aug – Dec

**Outlier Analysis:** <[<Index](#Index)

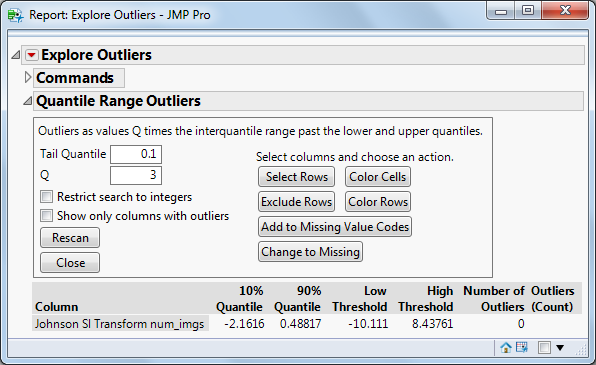
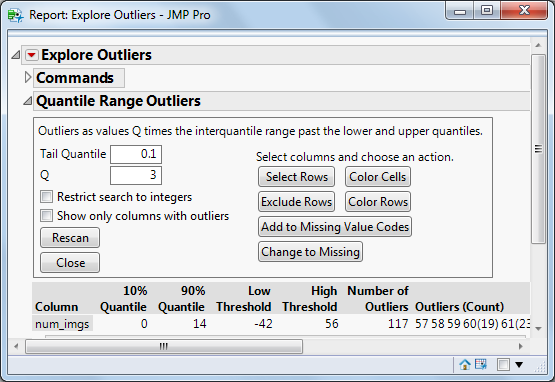
**Checked cols -> modeling utilities -> Explore Outliers**

Tried different outlier methods to understand the distribution of data within a range

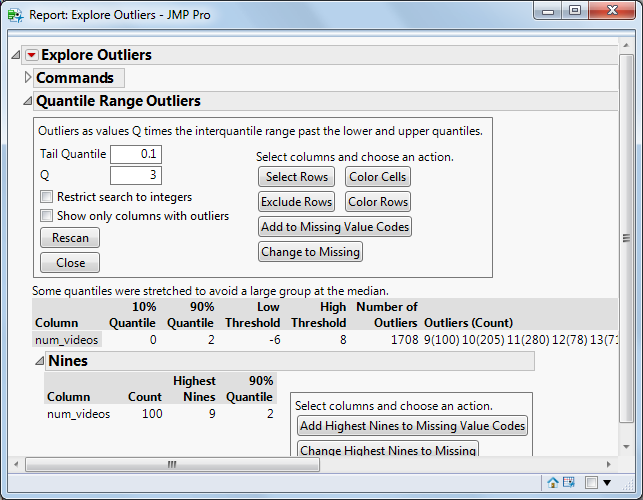
After determining the outliers, we transformed the variables to handle the outliers and where needed standardized them. As can be seen below, comparison of variables with respect to outliers, before and after transformation

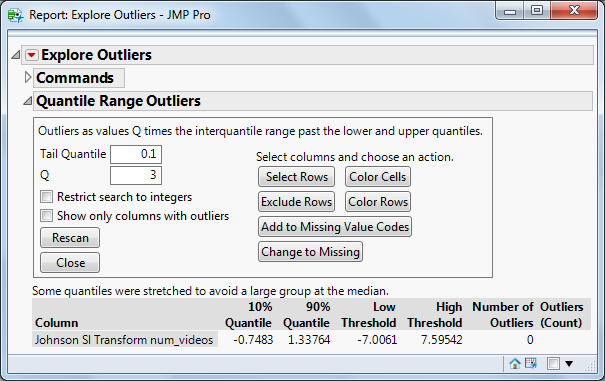


Comparison of variable – num\_self\_href before and after outlier treatment



Comparison of variable – num\_imgs before and after the outlier treatment





Comparison of variable – num\_videos before and after the outlier treatment

**Distribution of Variables – Standardization and Transformation**: <[<Index](#Index)

Right click the red button next to the graph **distribution -> Continuous fit -> All,**

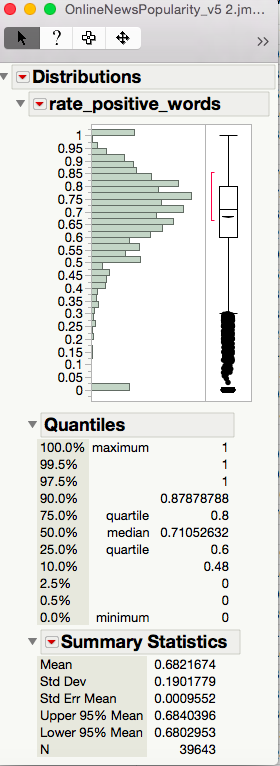
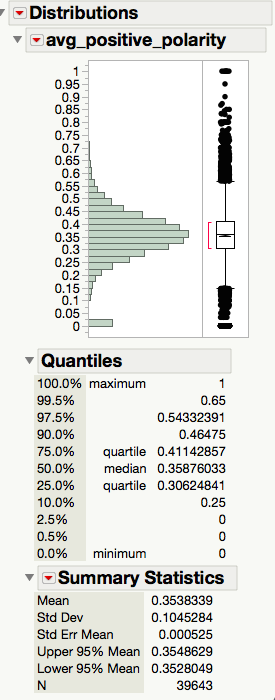
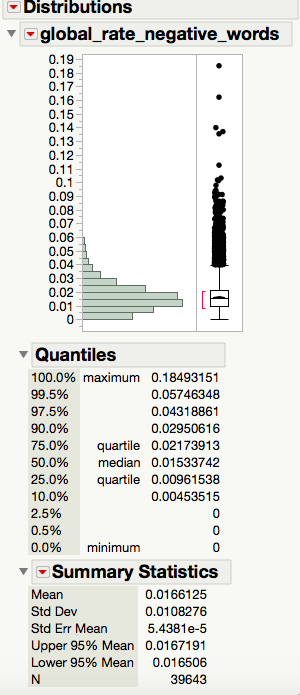
To find the Best fit for the distribution. Johnson SI appears as the best fit for the variable. Save the transformed variable

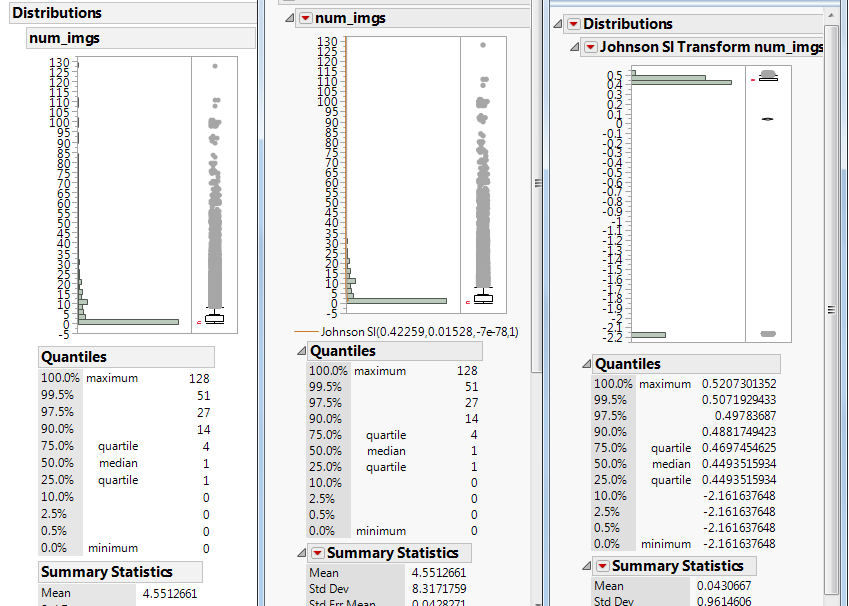
Similarly analyzed different transformation fits for the variables and then selected the most suitable one. Even clustering can be utilized to create buckets for the continuous variables. But, currently, we were able to find a suitable fit using the above method.

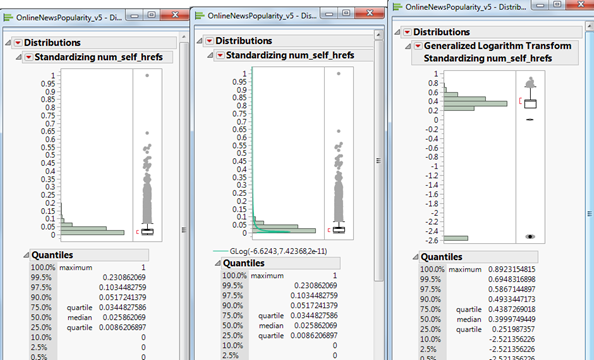
When some variables are in different scales when compared to the rest of the continuous variables. This needs to be handled, else the model will give weightage to values in higher scales.

Analyze -> Distribution, after observing the dispersion of values. Right click on the red hot button -> **Save -> standardized or** we can also standardize variables by creating a new column with the formula –

(xi – min (in that col))/ (max (in that col) – min (in that col)).



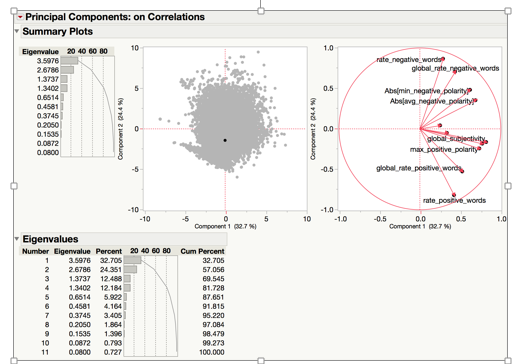


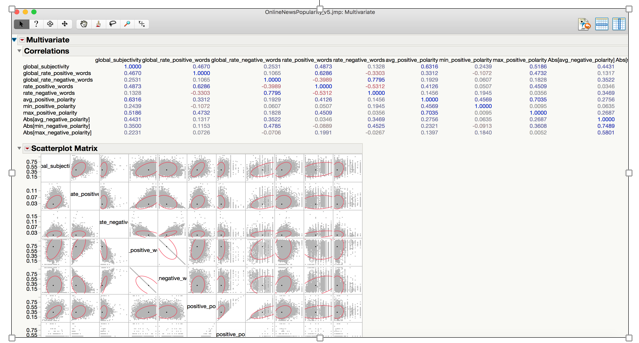


**Principal Component Analysis and Correlation Analysis:** <[<Index](#Index)

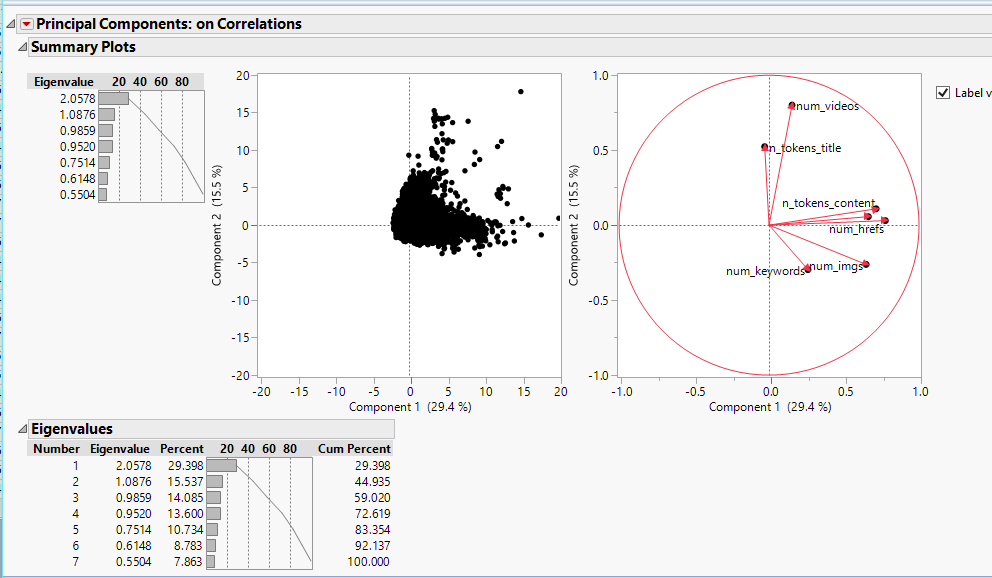
**Analyze -> Multivariate Methods -> Principal Components**

**Analyze -> Multivariate -> Multivariate**

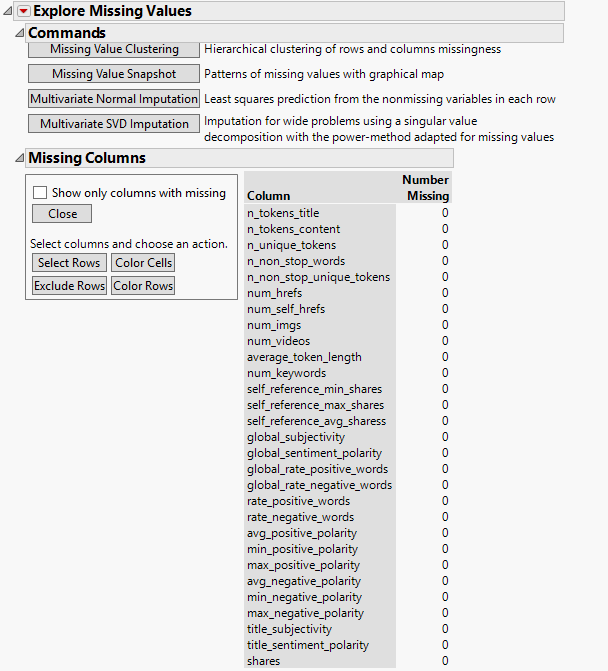




There are variables, which have negative impact on each other, but in general we can see that variables considered are very less correlated to each other. PCA analysis shows that there is very less multicollinearity among the independent factors.



**Missing Pattern Analysis:** <[<Index](#Index)



Even though the current dataset has outliers, wrongly entered data, there are no missing values present in the dataset

**Other Analysis**: <[<Index](#Index)

Performed Predictor Screening, to understand the impact of variables on the shares. We see that variables such as global\_subjectivity, max\_positive\_polarity etc seem to have an high impact.

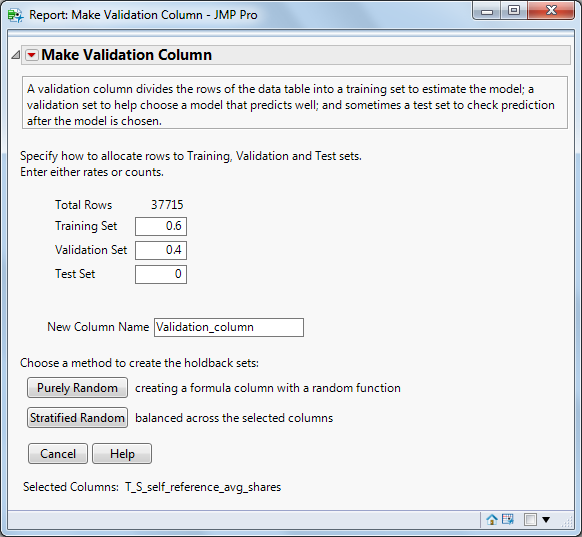


**Validation Column Creation:** <[<Index](#Index)

In order to create a validation column,

**Cols -> Modeling Utilities -> Make Validation Column**

Select Purely Random



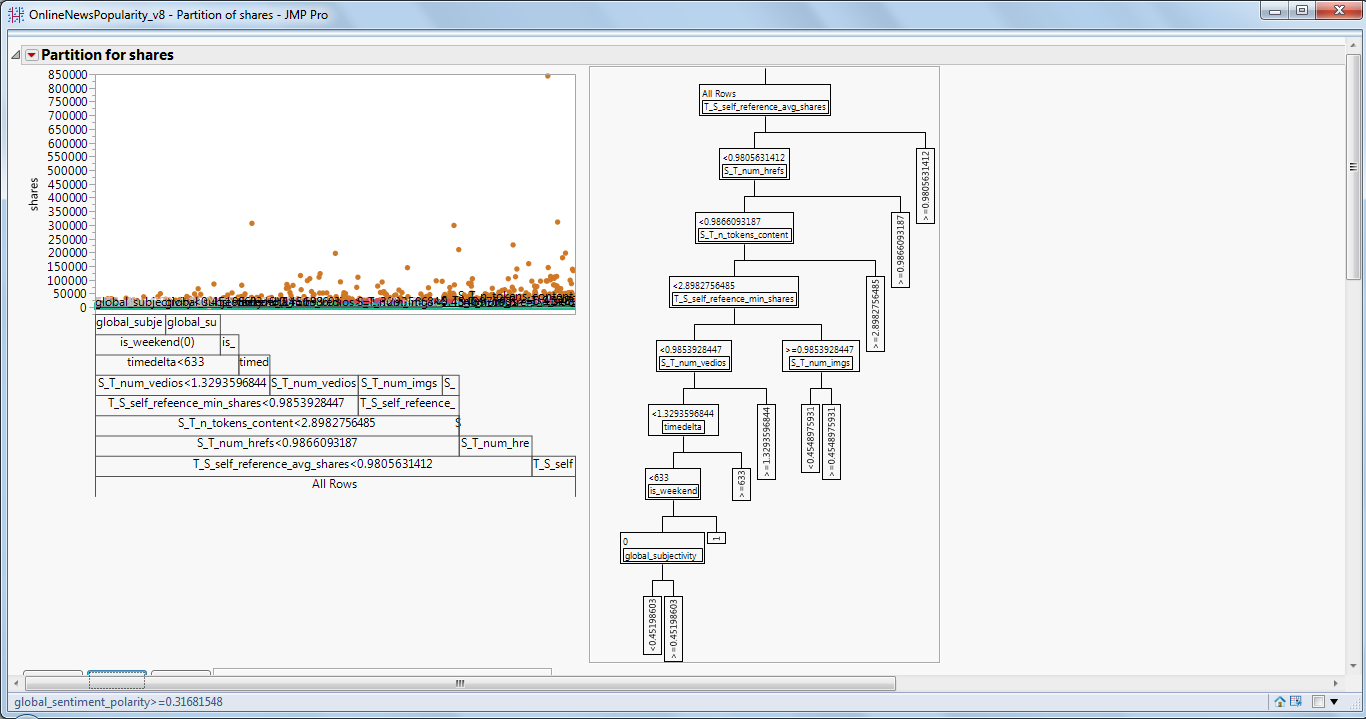
**Models – For Prediction of Shares** <[<Index](#Index)

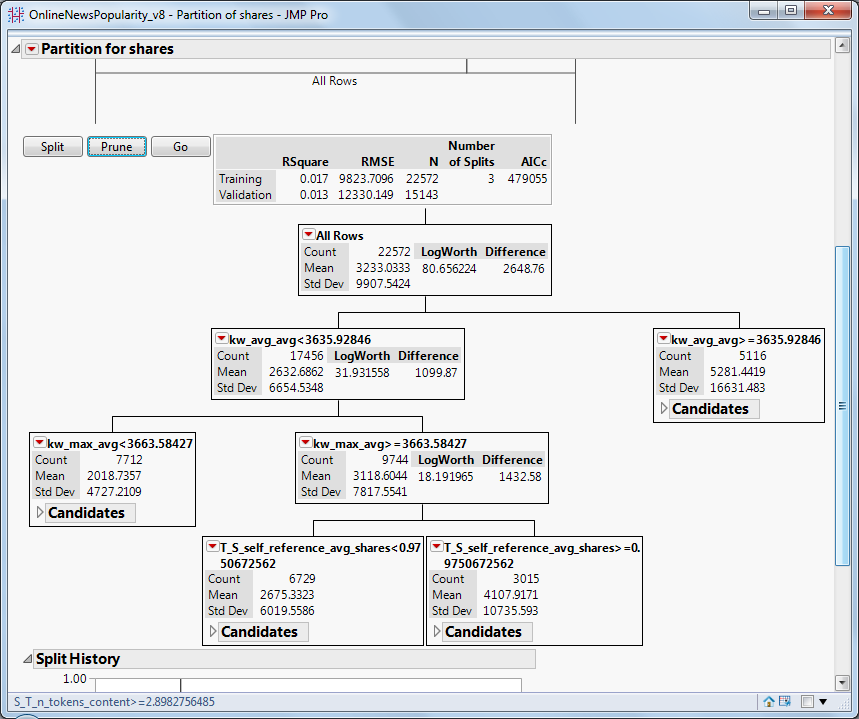
1. **Decision Tress:**

Analyze -> Modeling -> Partition (For Decision Tree).Put Y, Response variable as -> “Shares”

Validation variable as -> “Validation\_Column”. The rest of the variables as X, Factor

Decision tree without kW, LDA (Target=shares)



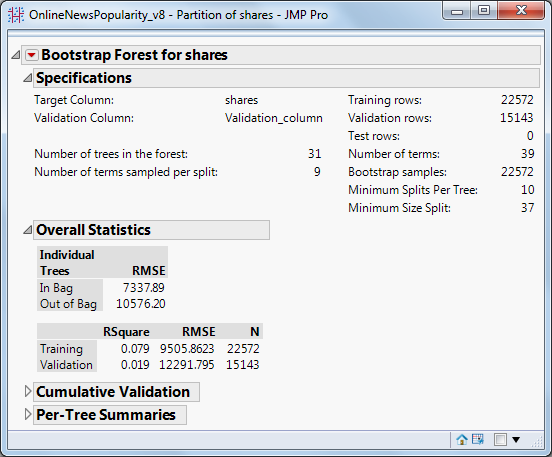


1. **Bootstrap forest :**

Analyze -> Modeling -> Partition (Bootstrap forest).Put Y, Response variable as -> “Shares”

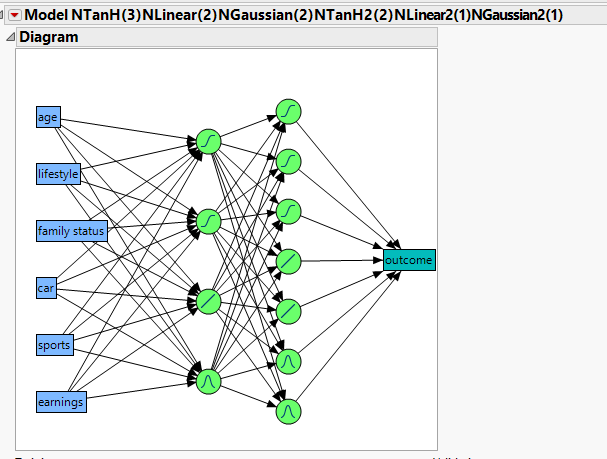
Validation variable as -> “Validation\_Column”. The rest of the variables as X, Factor

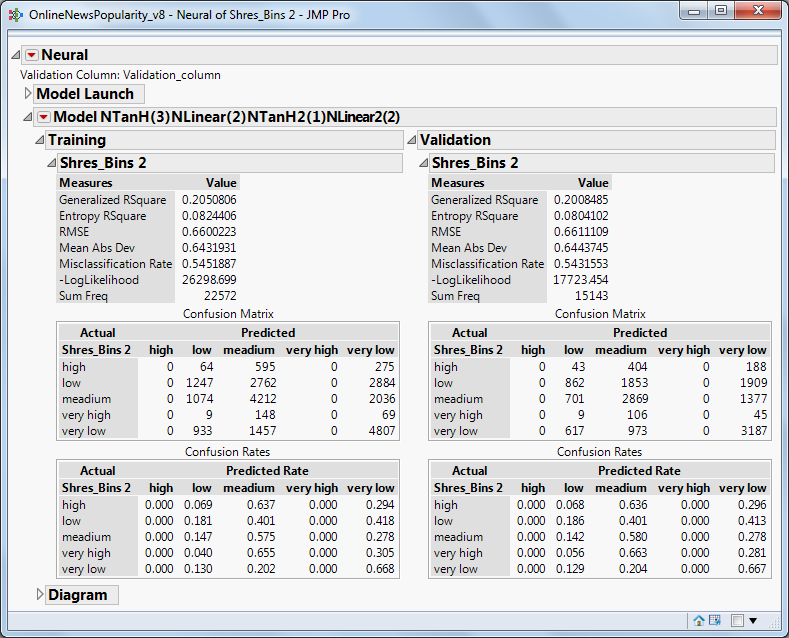
Bootstrap forest without kW, LDA (Target=shares)



1. **Neural Net**:

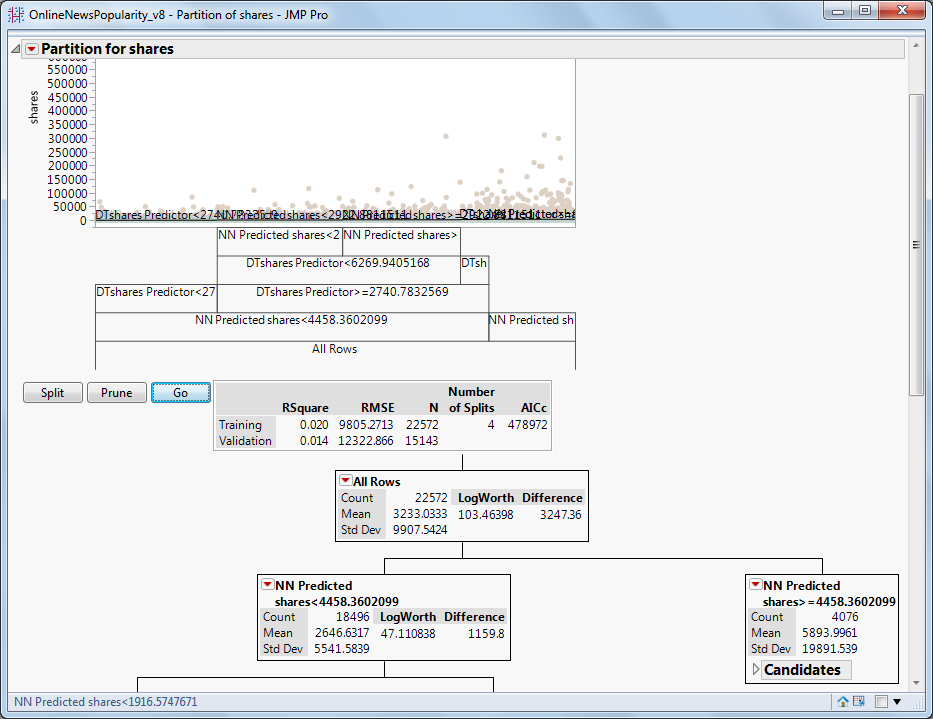
**Analyze -> Modeling -> Neural (For Neural Net).**  Put Y, Response variable as -> “Shares” or “Bins”. Validation variable as -> “Validation”



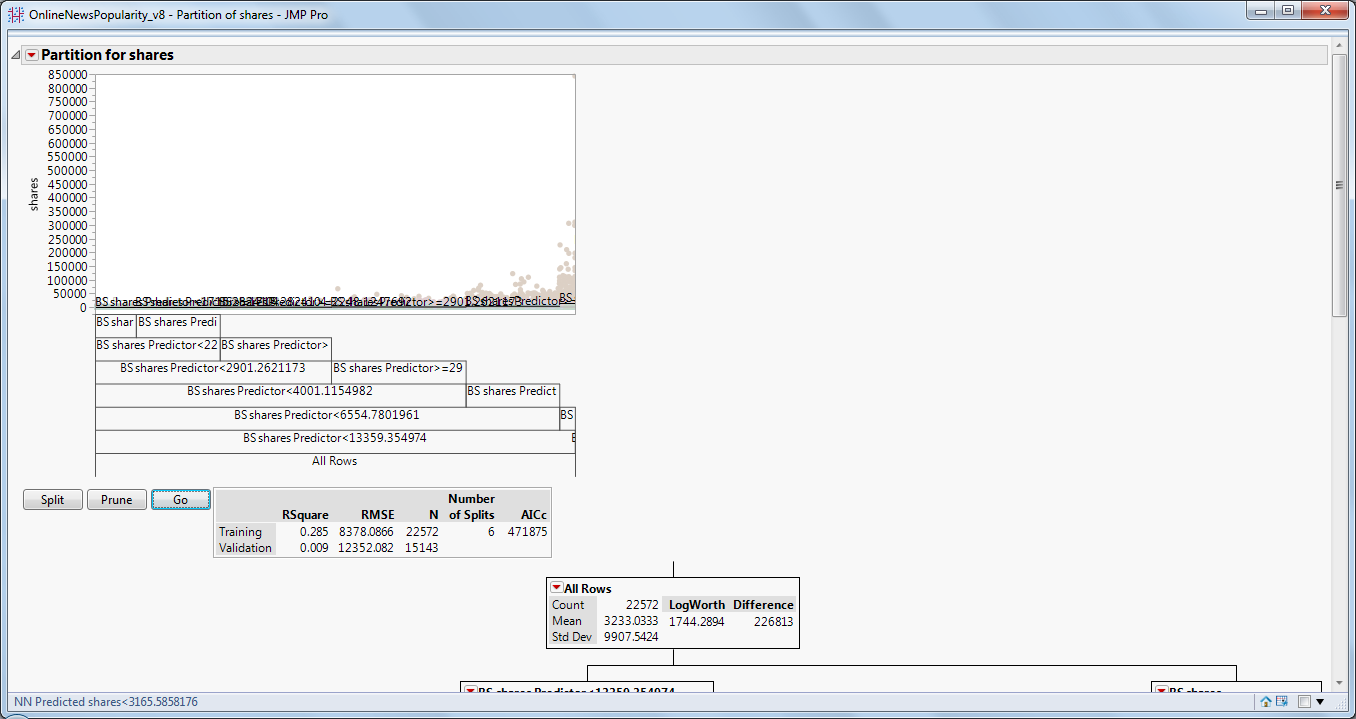


1. **Decision tree Ensemble model:**

Input: NN shares, DT shares

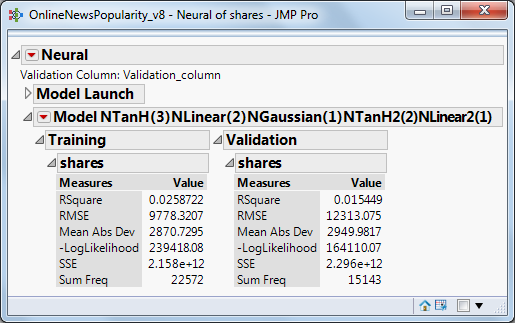


Input: NN shares, BS shares

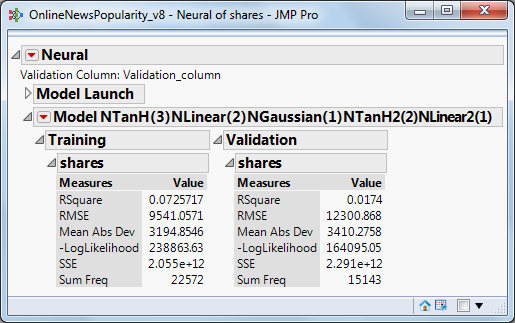


1. **Neural Net Ensemble model**

Input: NN shares, DT shares



Input: NN shares, BS shares



1. **Stepwise – Linear Regression:**

We performed stepwise linear regression (Both forward and backward), to determine the most important set of variable that have the highest impact on the response variable (shares).

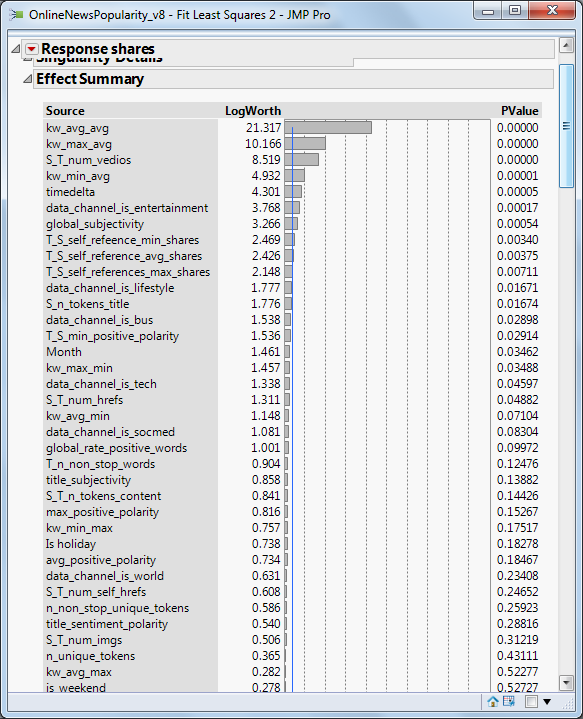
We looked at the R square value and also the intercept value to understand the overall effect of the variables and the accuracy of the prediction

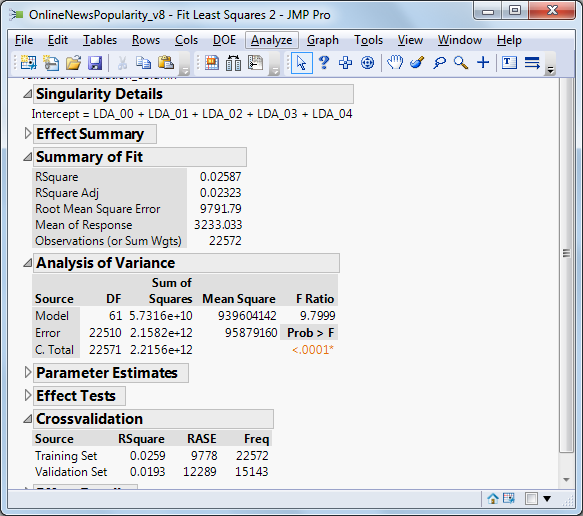


1. **Least Square:**

Analyze -> Fit Model (Least squares).Put Y, Response variable as -> “Shares”

Validation variable as -> “Validation\_Column”. The rest of the variables as construct mode effects in Add, Least Square without kW, LDA (Target=shares)





**Bin Creation**: <[<Index](#Index)

Created a column that bins the articles into categories from very low – very high depending on the total no of shares for that article. We saw that mostly the articles belonged to the lower bins indicating that the dataset was not evenly distributed and skewed towards lower popularity



**Models – Classification** <[<Index](#Index)

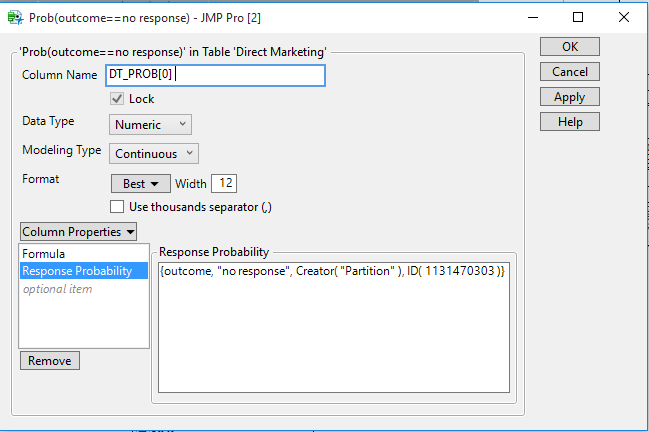
Here the response variable is the newly created ordinal variable that describes the shares as being very low – very high

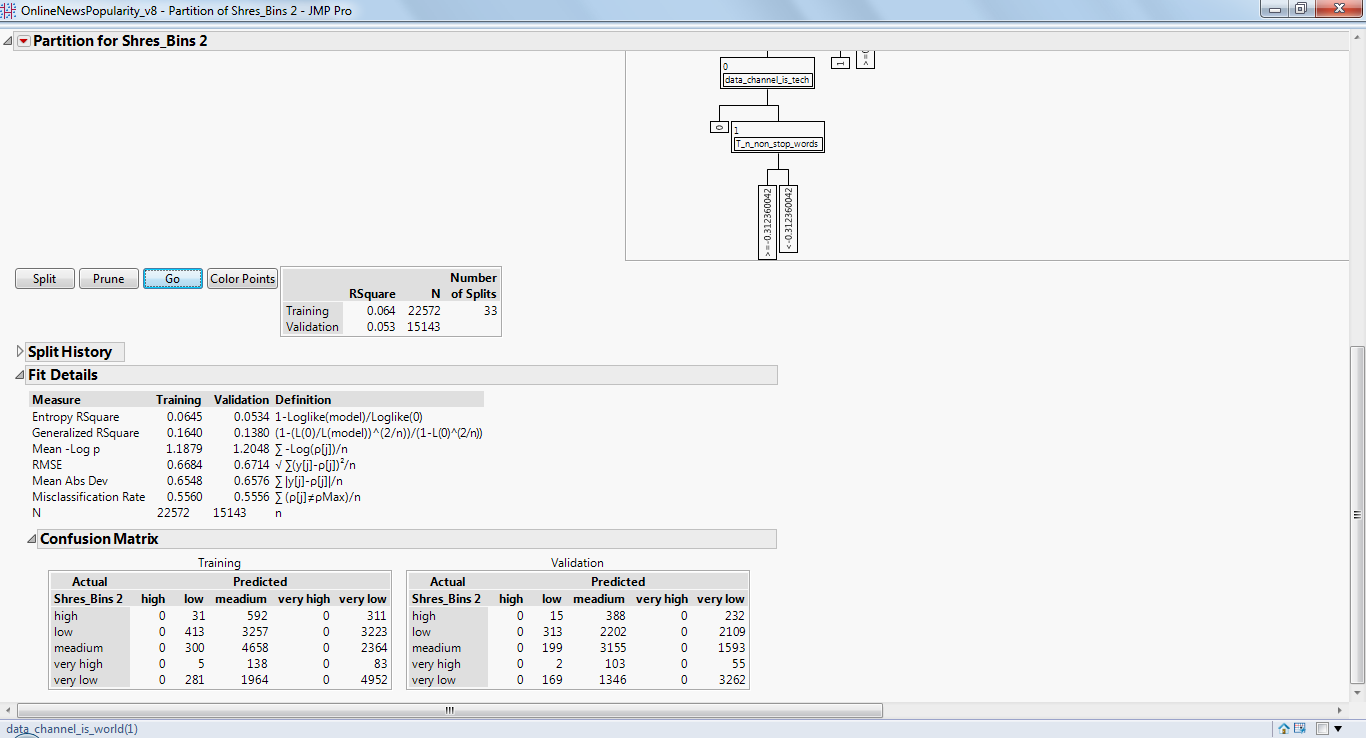
1. **Decision tree:**

Confusion Matrix: provides confusion statistics for both the training and validation data

From the report’s red triangle menu, select **Save Columns > Save Prediction Formula**.

To save the predicted values in a new column. Rename the column Prob (outcome == low) as DT\_PROB [low] and Prob (outcome == high) as DT\_PROB [high] and the same for all the probabilities

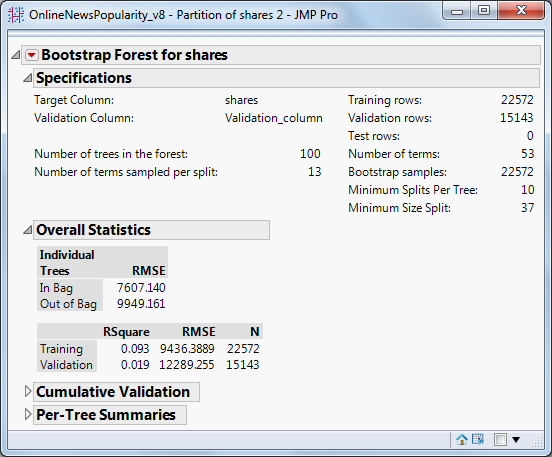


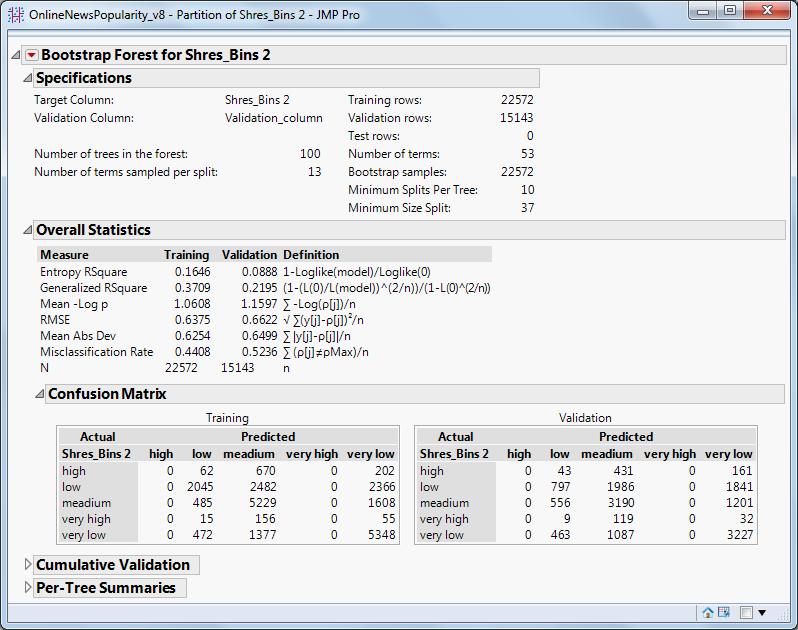


1. **Boot strap forest :**

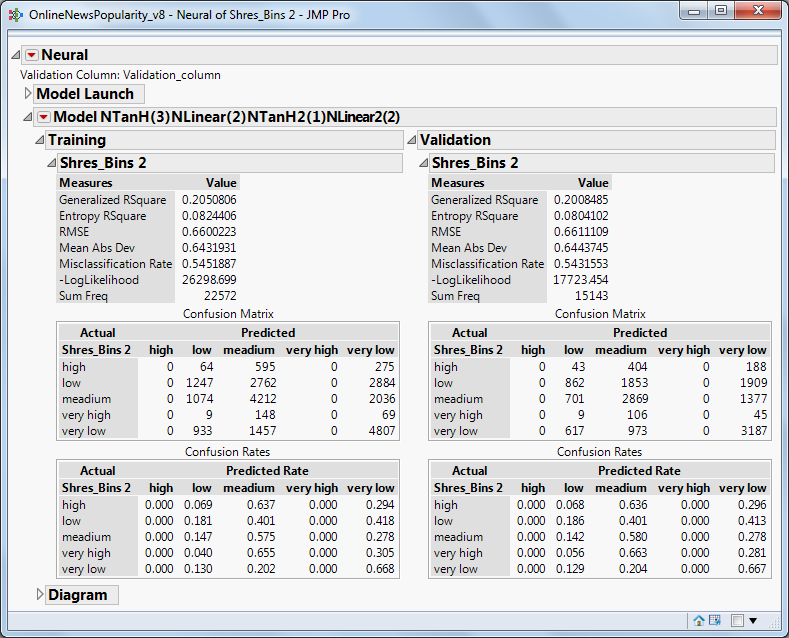
Analyze -> Modeling -> Partition (Bootstrap forest).Put Y, Response variable as -> “Bins”

Validation variable as -> “Validation\_Column”. The rest of the variables as X, Factor



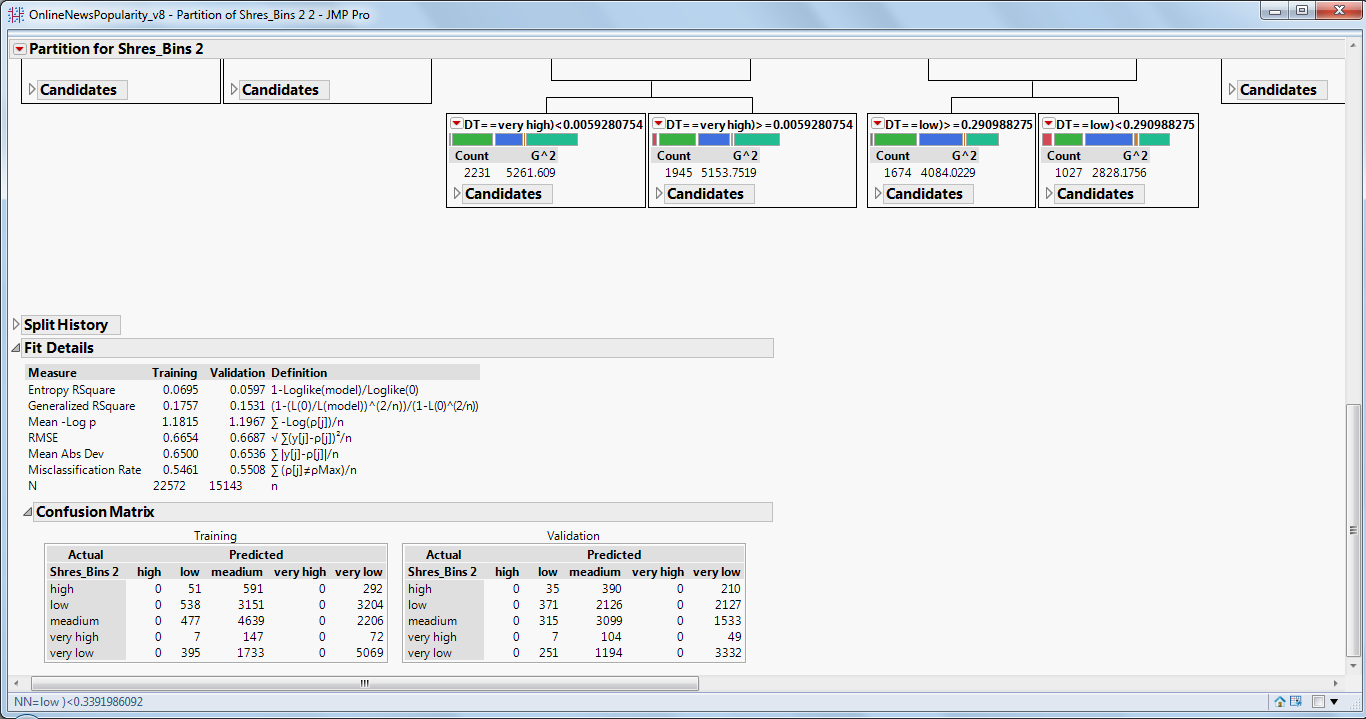


1. **Neural Net:**

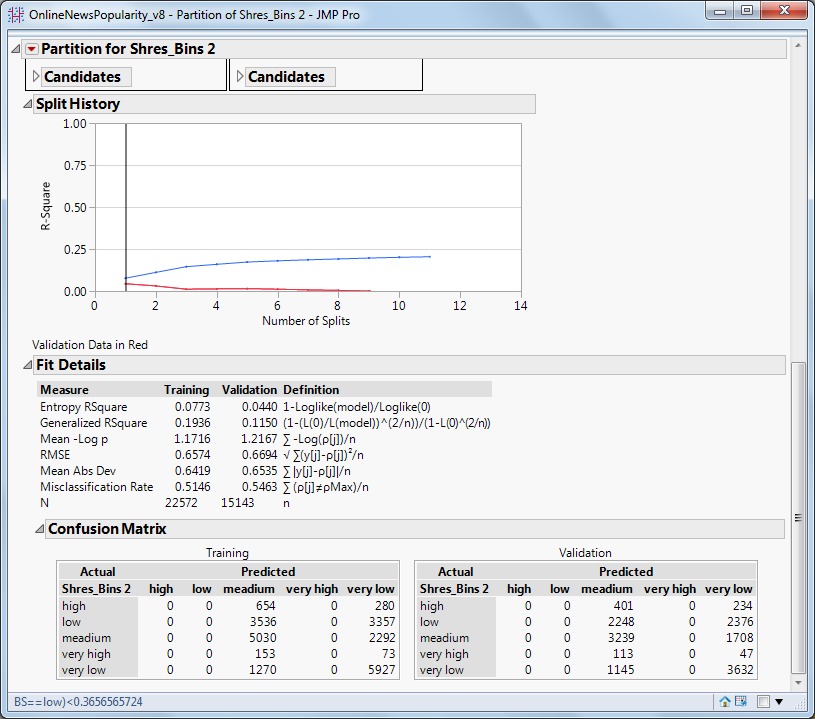


1. **Decision Tree Ensemble**:

Input: NN bins, DT bins

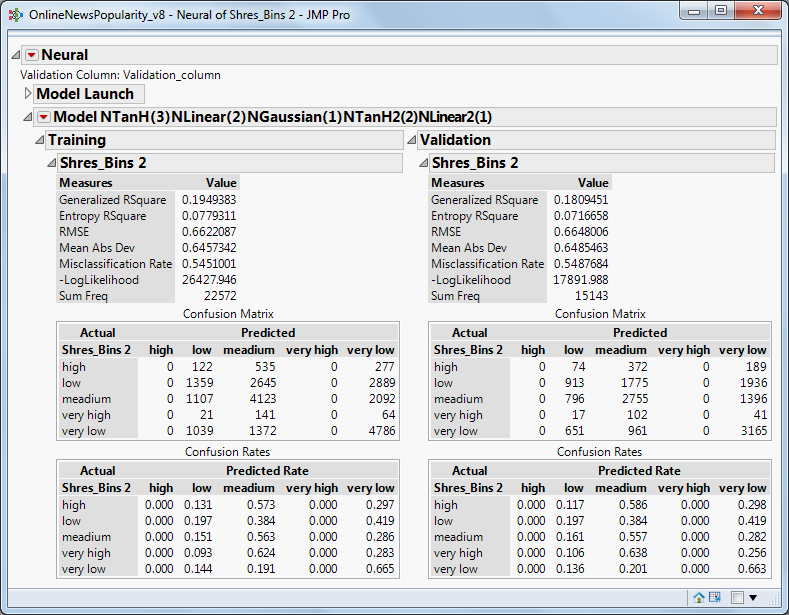


Input: NN bins, BS bins

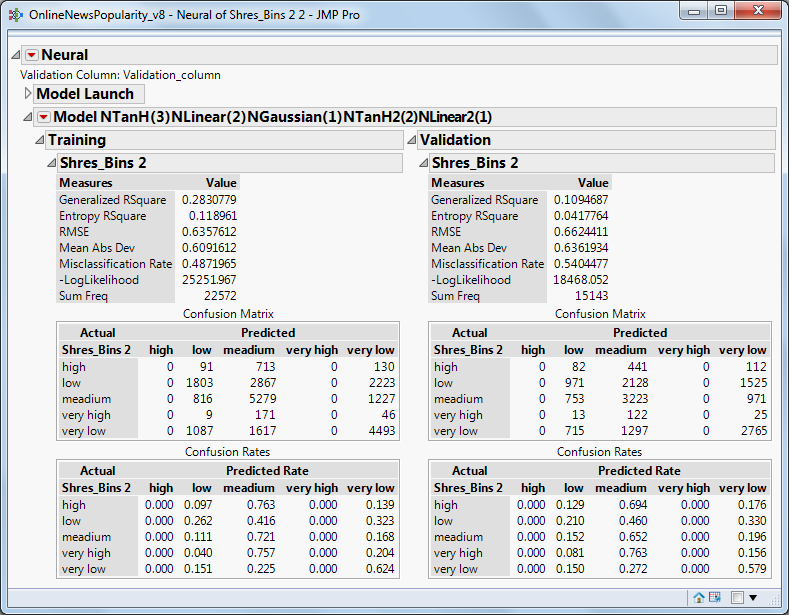


1. **Neural Net Ensemble:**

Input: NN bins, DT bins



Input: NN bins, BS bins

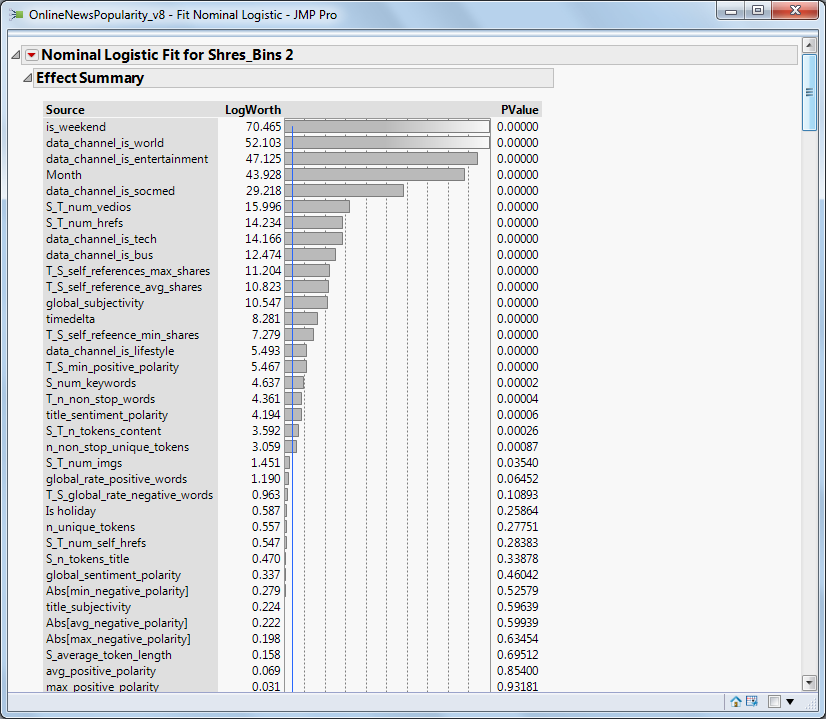


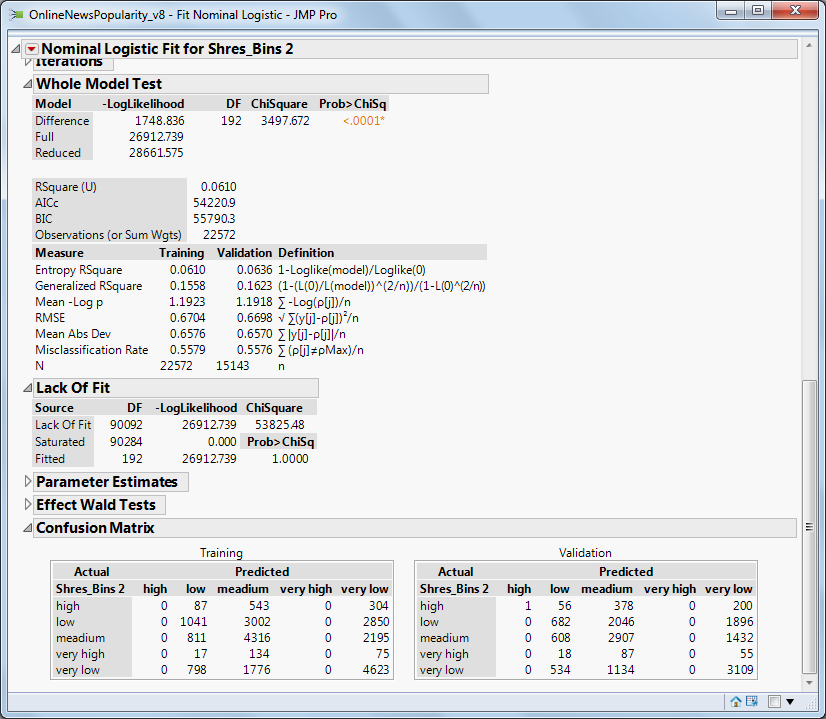
1. **Logistic Regression**:

Analyze -> Fit Model (Nominal Logistic).Put Y, Response variable as -> “Shares\_Bins”

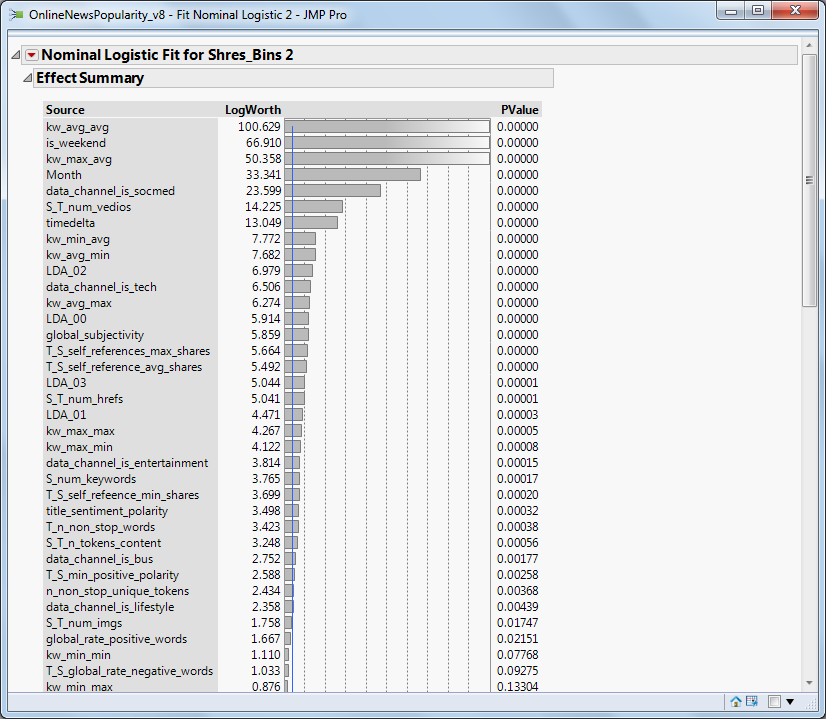
Validation variable as -> “Validation\_Column”. The rest of the variables as construct mode effects in Add

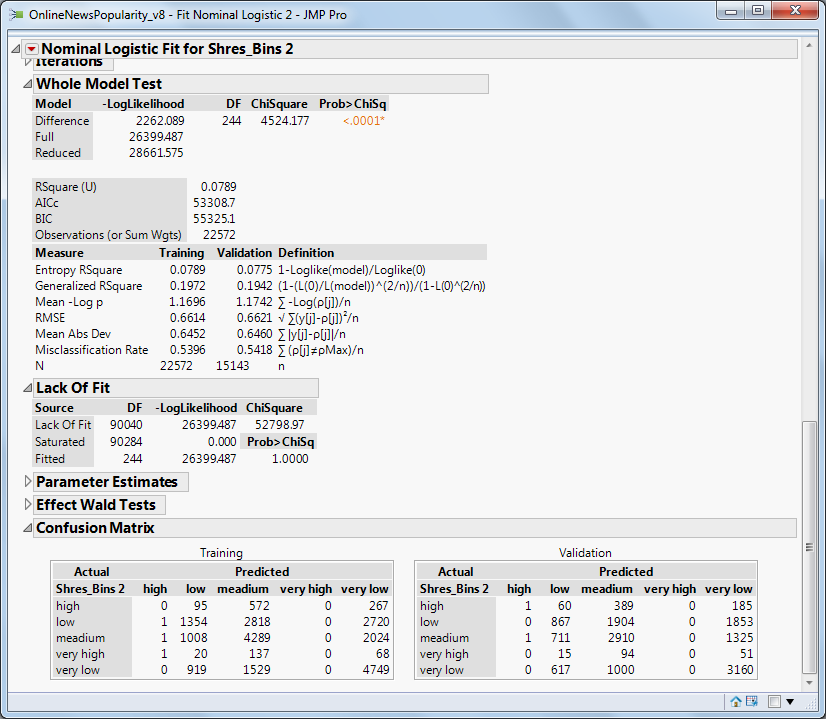
a) Without kW, LDA





b) With KW, LDA

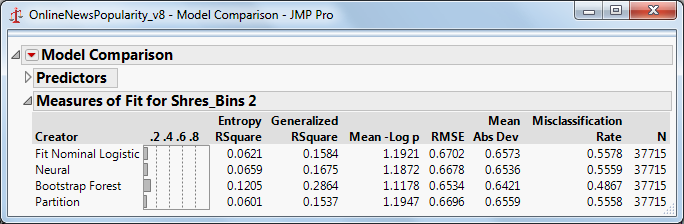


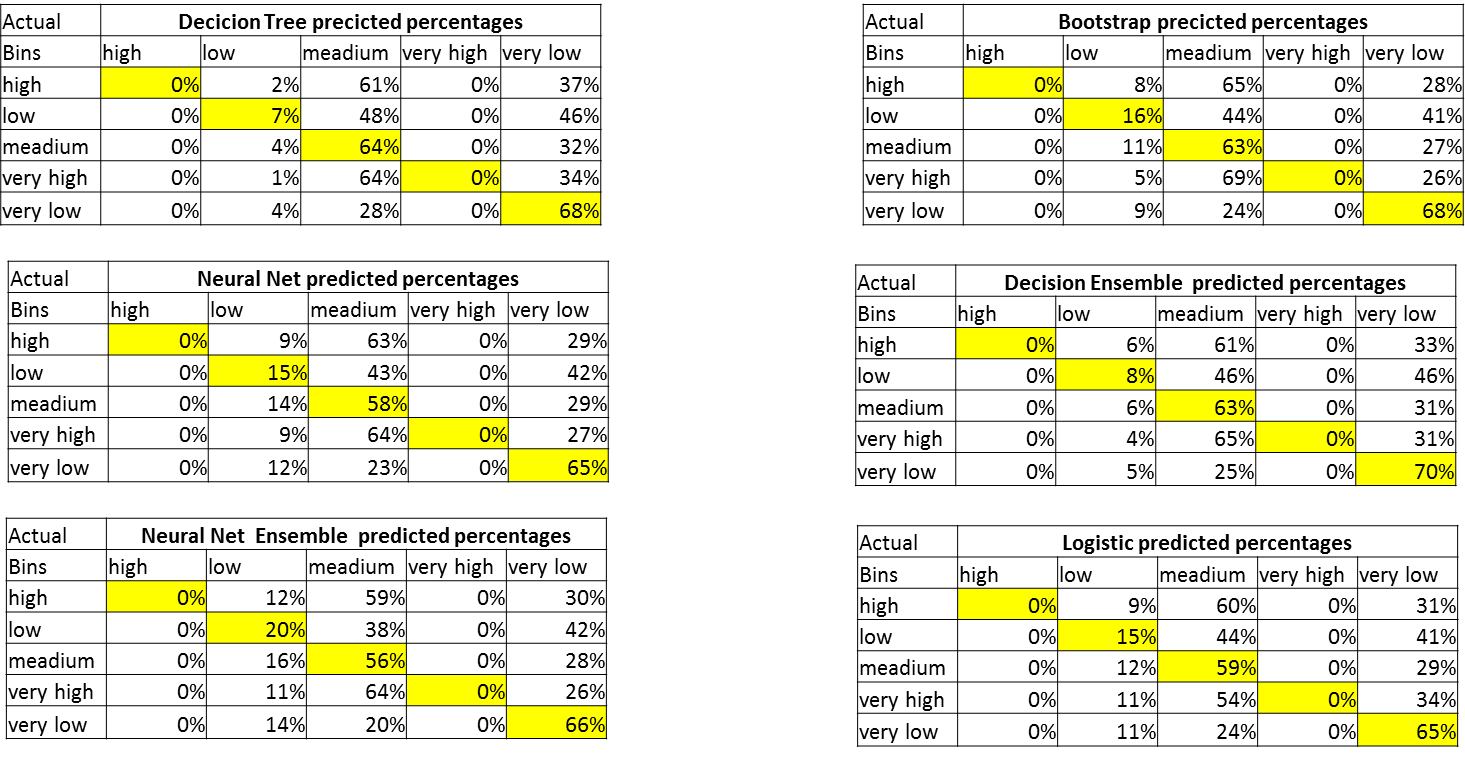


**Comparison of Model Results:** <[<Index](#Index)







****

**Model Summarization:** <[<Index](#Index)

1. Results from the different models are almost similar to each other for both training and validation data, as can be seen from the comparison snapshot above
2. From the models and the outputs we realized that predicting continuous variables is not suitable for this particular dataset especially when we are not very clear about the variable definitions
3. A decision ensemble and bootstrap are by the better among the rest even with very poor results. Even though all the models are not predicting well (as can be seen- the wrongly predicted values are more), it is very difficult to explain the results of a neural net and the ensemble models in general are very complicated
4. The main disadvantage of this neural network model is that the results are not easily interpretable, since there are intermediate layers rather than a direct path from the X variables to the Y variables

**Problems Faced during the entire project work**: <[<Index](#Index)

1. PCA Analysis was not effective as there was very less correlation between variables
2. There were many instances where the variables were not properly scrapped from the website and was wrongly populated in the dataset
3. The definitions of many variables such as LDA Topics and KW\_words were not clear. The available data dictionary did not instill confidence in terms of reliability for the model
4. There exists a high skewness in the popularity distribution (i.e., large number of unpopular articles and very few popular articles) will lead to a bias in the prediction process carried out
5. The scoring and creation of sentiment, polarity related variables remained a black box
6. We are trying to predict the popularity at a specific future time and are oblivious to how the popularity is spread over the entire lifetime of an article (Very low temporal data)
7. We are not aware of the demographics of the users who share these articles or in general the users who dominate this website where the articles are posted (the type of news categories they like/dislike, what are they influenced by?)

**Inference/Key Learnings and Recommendation:** <[<Index](#Index)

1. All Modelling techniques are not equally effective
2. Realized that a problem needs to be approached as scenario based
3. Generic relations might not be true in Actual ( e.g. : More Video, More popularity)
4. We realized that prediction of continuous variables is not a good approach. Rather binning the shares and measuring/ranking the articles in terms of ordinal variable – low, medium, high yielded better results
5. Realized that while predicting continuous variables, it is better to provide a buffer range but was unable to do due to lesser business knowledge
6. Given a popular and an unpopular article, the difference between the predicted popularity values is not of high importance as long as we can correctly rank the popular article above the unpopular article
7. All Modelling techniques are not equally effective
8. Realized that a problem needs to be approached as scenario based
9. Generic relations might not be true in Actual ( e.g. : More Video, More popularity)
10. We realized that prediction of continuous variables is not a good approach. Rather binning the shares and measuring/ranking the articles in terms of ordinal variable – low, medium, high yielded better results
11. Realized that while predicting continuous variables, it is better to provide a buffer range but was unable to do due to lesser business knowledge
12. Given a popular and an unpopular article, the difference between the predicted popularity values is not of high importance as long as we can correctly rank the popular article above the unpopular article
13. We will yield better results in classifying news article as popular or unpopular. We need to keep in mind that misclassification have different costs associated with it
14. Also, there were sizeable no. of outliers in the dataset which may end up skewing the data if we try to predict for continuous variables whereas classification lessens the error