# Introduction

Social media marketing is a great way to help launch a startup. It’s cheap and effective -- if the startup has a strategy. This report aims to compare the social media content of three typical companies in different industries to make reasonable recommendations for the promotion of an Internet startup.

A screenshot of a cell phone

Description automatically generated

The name of three companies and their corresponding counts:

A screenshot of a cell phone

Description automatically generated

Descriptions of continuous variables:

A screenshot of a cell phone

Description automatically generated

# 1. Data Preprocessing and Exploration

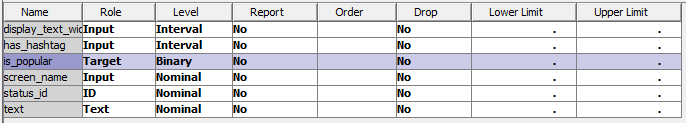
## 1.1 Create a new project.

- Load the training dataset into SAS using File Import node

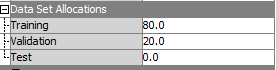
- split the dataset into training and validation sets (80% vs. 20%).

- Set is\_popular as Target for Role and binary for Level after running the File Import note

Here I set the ‘is\_popular’ as Target for role and binary for level.



Split the dataset into training and validation sets (80% vs. 20%):



## 1.2 Preprocess the texts using the **Text Parsing** and **Text Filter** nodes.

In the **Text Filtering** node, set **Frequency Weight** to Log and **Term Weight** to Mutual Information.

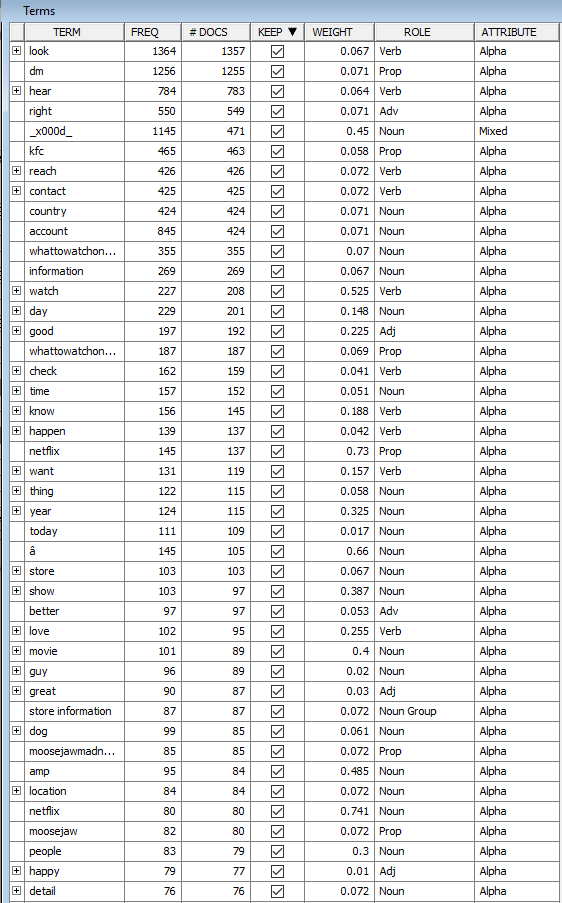
Add Text Parsing and Text Filter nodes:



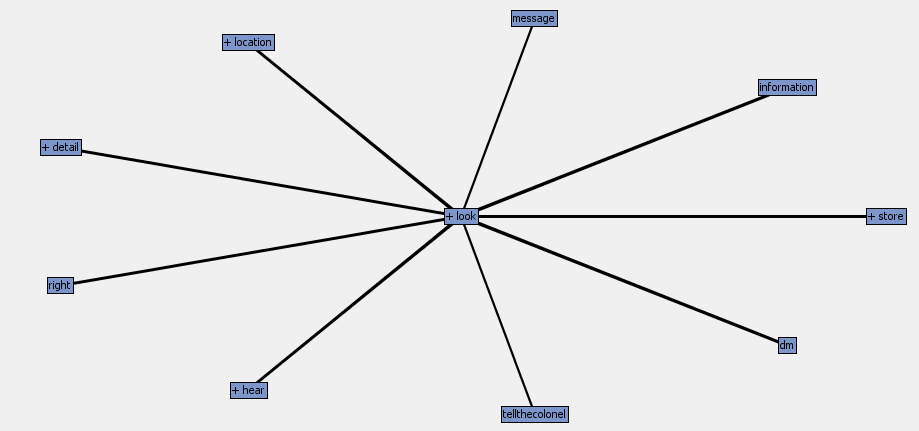
Set Frequency Weight to Log and Term Weight to Mutual Information:



This is a part of terms in the Interactive Filter Viewer:



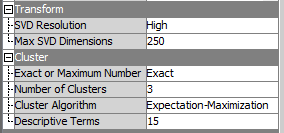
Concept links for the term ‘look’:



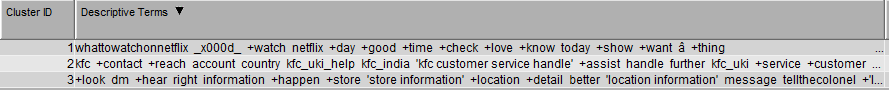
## 1.3 Use the Text Cluster node (SVD) to reduce the dimensionality.

Set SVD resolution as **High**, the maximum number of SVD dimensions as **250**, and the Number of Clusters as **3** (**Exact or Maximum Number** as **Exact**). Use the default for other parameters.

Settings for SVD:



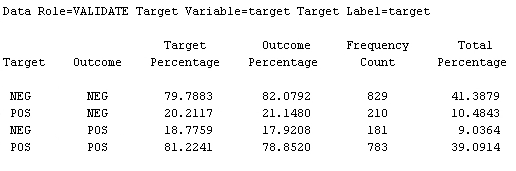
Screenshot of the Descriptive terms of the three clusters:



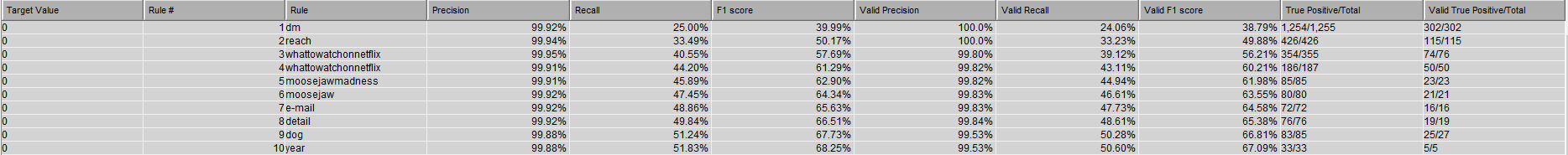
# 2. Model Development

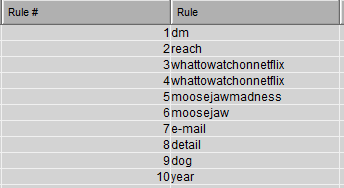
## 2.1 Text Rule Builder.

Build a classifier using the Text Rule Builder node.

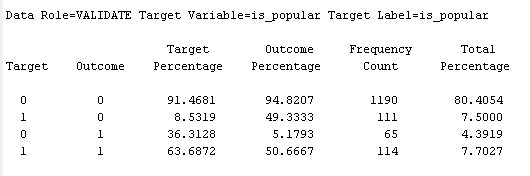


Screenshot for the top 10 rules:





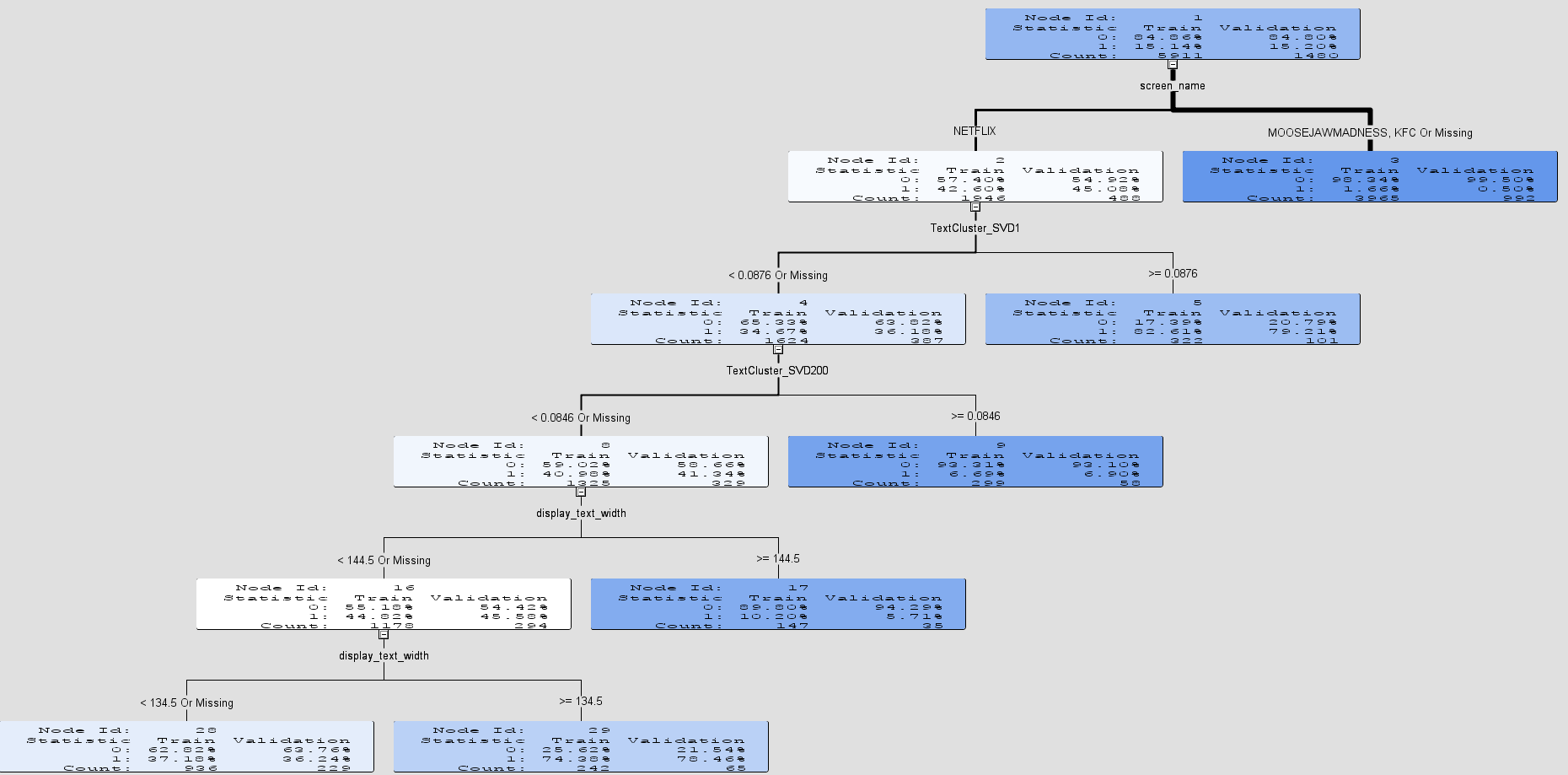
Screenshot of the confusion matrix on the validation set:



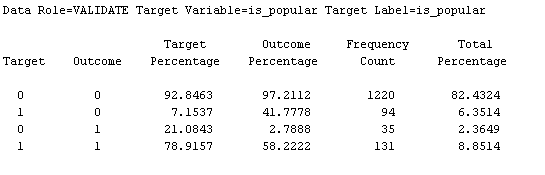
## 2.2 Decision Tree

Build a classifier using the Decision Tree node.

Screenshot of the Classification Tree:



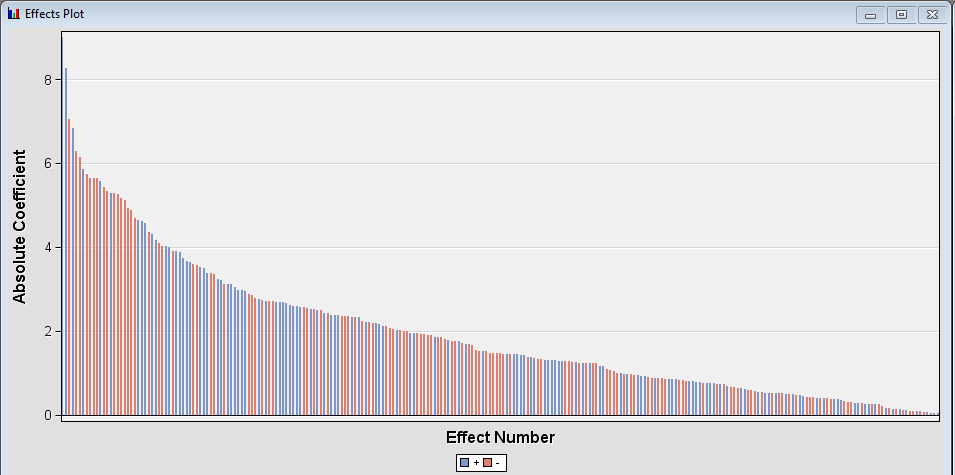
Screenshot of the confusion matrix on the validation set:



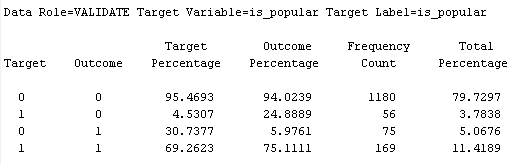
## 2.3 Logistic Regression

Build a classifier using Logistic regression.

Effects Plot:



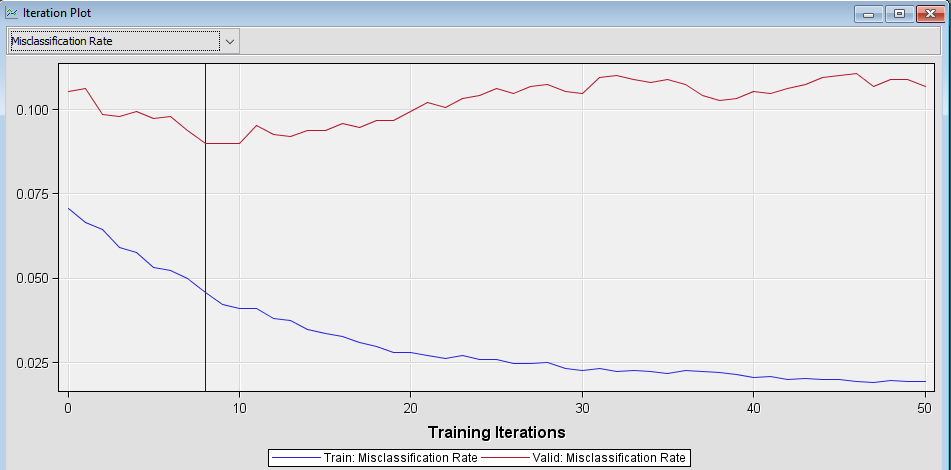
Confusion matrix on the validation set:



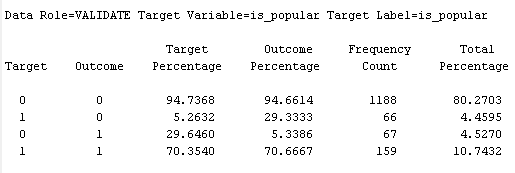
## 2.4 Neural Network.

Build a classifier using the Neural Network node.

Iteration Plot for Misclassification Rate:



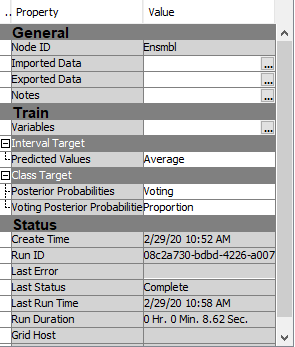
Confusion matrix on the validation set:



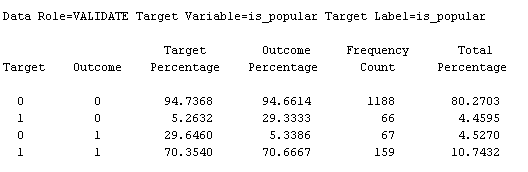
## 2.5 Ensemble Model.

Build an ensemble model based on all the above four models.

Under the property Class Target, set Posterior Probabilities to **Voting** and Voting Posterior Probabilities to **Proportion**. The Ensemble Model is particularly error prone in SAS Enterprise Miner. Sometimes, the **Score** node may not work if the Ensemble Model is the best model. In that case, link the second best model (instead of the Model Comparison node) directly to the **Score** node.



Confusion matrix on the validation set:

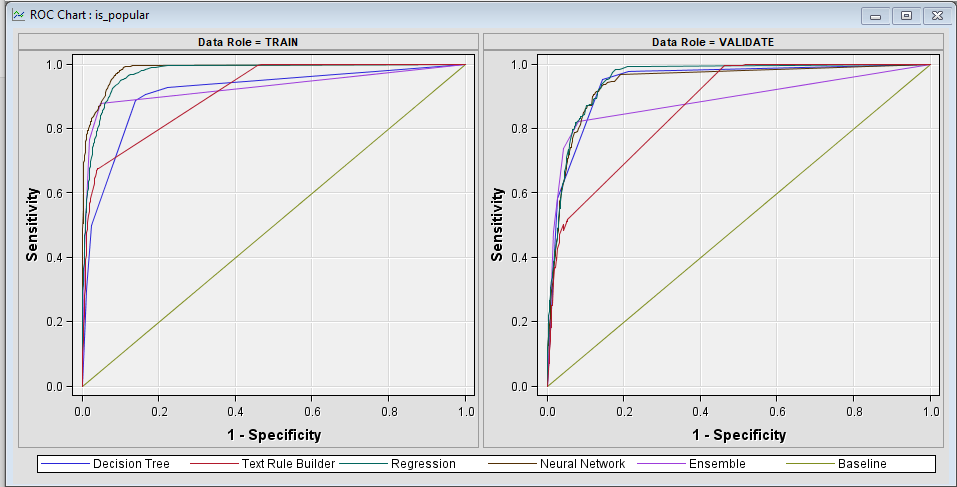


## 2.6 Model Comparison

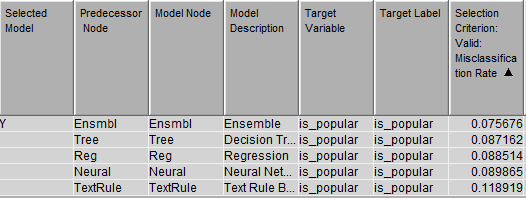
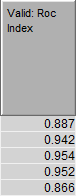
Logistic regression performs the best in terms of ROC on the validation set

Ensemble model performs the best in terms of misclassification rate on the validation set

ROC Chart:



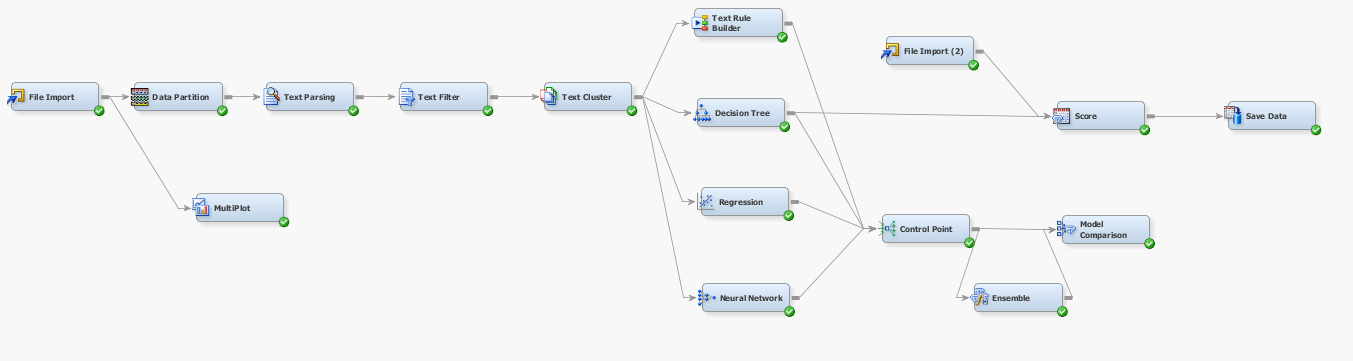
ROCs and Misclassification rates for all models:

2.7 Import the score dataset and generate predictions on the score set

Based on the best model and save the predictions on the score set as an Excel file. Finally, provide the screenshots of your final diagram and save it as an XML file. (2 points)

The Ensemble Model performs the best, so we link the second best model, Decision Tree, directly to the Score node. The screenshots of the final diagram:



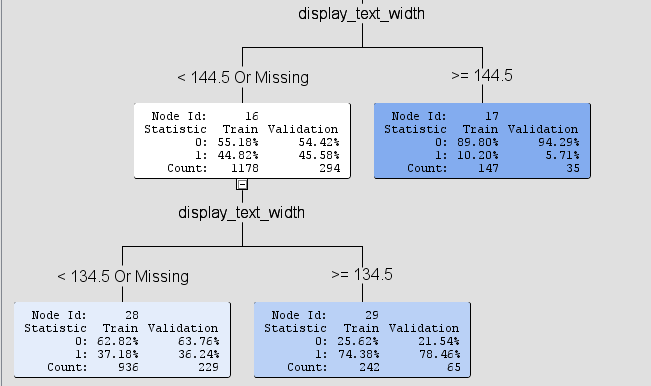
Tips: a) In the property panel of the File Import node for score dataset, make sure the **Role** of the data is set to “Score”. b) In the prediction file, the column with the binary predictions is called “EM\_Classification”. The column with the predicted probabilities is called “EM\_EventProbability”. c) You may zoom out a diagram using the tool at the bottom right corner.

**3. Business Insights (3 Points)**

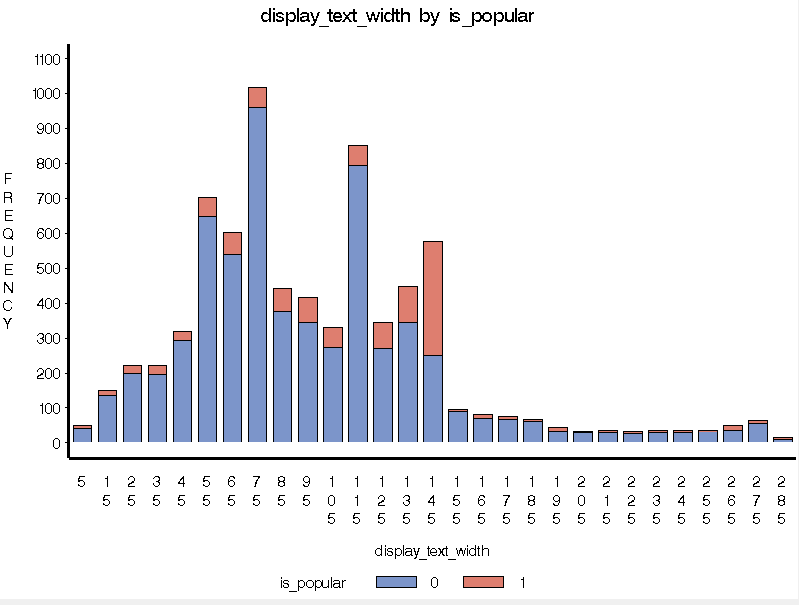
Based on your analysis of the dataset, what do you think the brands can do to improve the engagements (i.e., retweets) on their tweets? Please provide at least three recommendations and each of them should be supported by a specific finding from the data analysis (provide a screenshot for the finding when available).

**Recommendation 1 Control number of characters within a range.**

To improve retweets, brands should carefully plan the length of the tweets by controlling the display text width, that is, the number of characters. As shown in the following branches of the decision tree, 44.82% of tweets with less than 144.5 characters (or missing) are popular, while 10.21% tweets with more than 144.5 characters (or missing) are popular. Moreover, within the range [0,144.5), 37.18% of tweets with less than 134.5 characters (or missing) are popular, 74.38% tweets with number of characters in [134.5,144.5) are popular. Therefore, keeping the number of characters within [134.5,144.5) is a strategy to increase the possibility of retweeting. In practice, too short tweets may contain too little information to retweet, and too long tweets may make readers impatient, reducing the retweet rat

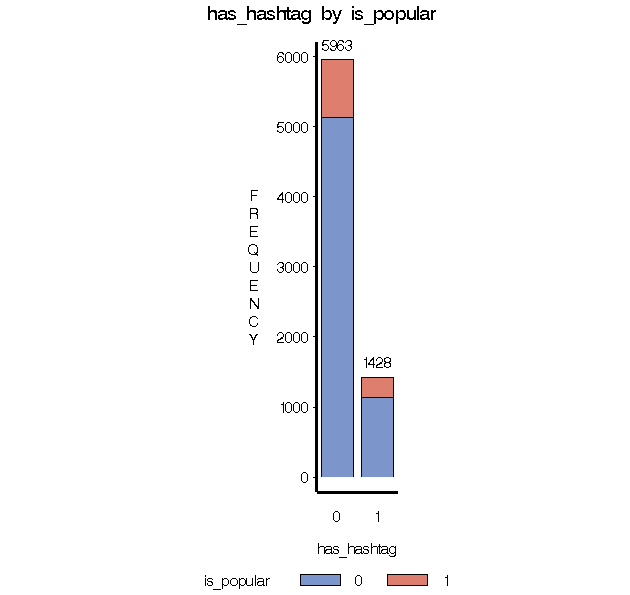


To support this recommendation, the graph below shows the popularity of tweets in terms of different display text width. We can that tweets with they display text width from 135 to 145 have larger proportions of popular tweets. Therefore, it is demonstrated that brands can try to keep their text widths of tweets within this range.



**Recommendation 2 Use hashtags properly.**

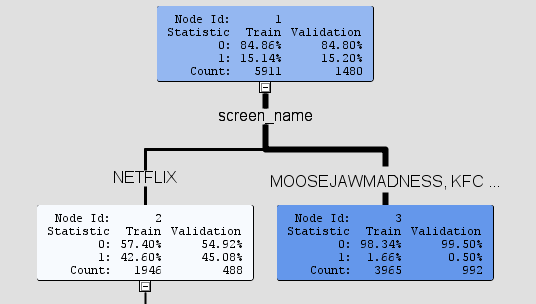
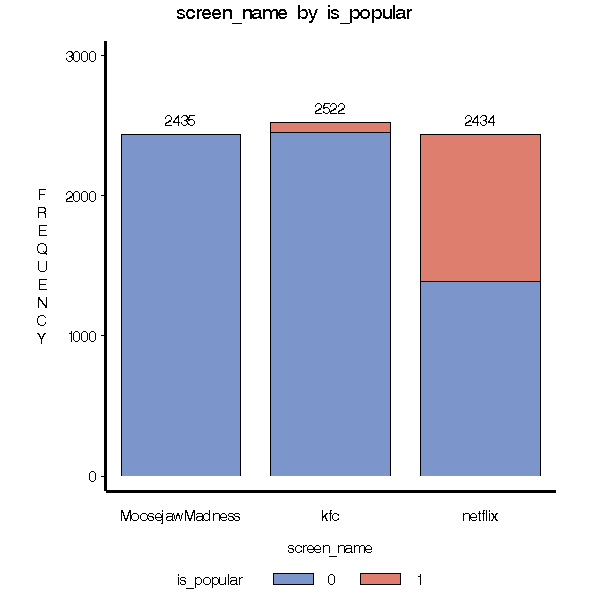
Tweets with hashtags are more likely to be popular. Among the training data, as shown below, 829/5963 = 13.90% of the tweets without hashtags were retweeted, but 291/1428 = 20.38% of the tweets with hashtags were retweeted. Generally, using hangtags can improve the probability of retweeting. In practice, people use the hashtags can make their tweets categorized and help them show more easily in Twitter search. However, when using hashtags, too many hashtags are not desired, and those used should be relevant and precise, and easy to remember and search. For instance, brands can use hashtags when referring specific topics, and it’s also recommended to use no more than 5 hashtags per tweet because too many hashtags makes people lose focus.

 A screenshot of a cell phone

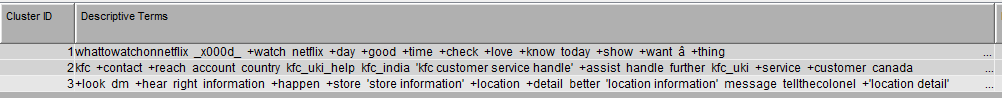
Description automatically generated

**Recommendation 3 Improve the content with dynamic, positive and attracting words.**

Based on the decision tree, tweets from Netflix have a probability of 42.60% to be popular, but those from the other two brands have a probability of 1.66% to be popular. This is also shown in the histogram below. That is, the popularity of different brands varies, so the brands with lower popularity on Twitter can learn from the popular brands in terms of their strategies.

For example, looking at the clasuter of three brands, from the words we can detect that words used in tweets of Netflix such as ‘today’, ‘day’, ‘time’ are dynamic, which shows the content of Netflix may change frequently, which can bring new information to readers with updated tweets. Also, ‘good’ and ‘love’ are words making readers positive, which are more likely to make the information accpted and retweeted. However, in the other two brands, words such as ‘country’, ‘canada’, ‘location’ are not such dynamic, which may be less attracting and provide less new information to readers. This may because of the different properties of different brands: Netflix can have new videos everyday, but it’s difficult for KFC and Moosejaw to do so. To improve this situation, it is recommended that KFC and Moosejaw can spread more recent offers and more positive and attrracting content to provide used with the latest news. For example, based on the results of Text Rule Builder, words like ‘final season’, ‘limited’, ‘deserve’ may become better choices. Also, from the brand themselves, KFC and Moosejaw can try to launch new products and promotions as your budget allows, and promote them through social media. From the data, some tweets of Moosejaw are not directly related to their brands, for instance, “I'm working from home today which is why it took me just over 5 hours to write this tweet.” Such tweets don't contribute to the brand promotion, so personally, I’ll recommend them to avoid such content in Twitter.



**What to submit**

1. The training and score datasets if you collected data on your own (.xlsx)

2. The saved diagram (.xml) and your predictions on the score dataset (.xlsx)

3. The Enterprise Miner project folder (please zip it).

4. This document. please named it as "Assignment2\_[net\_id].docx"