Executive Summary
Predicting Popularity of Online News

Online news has increasingly become the preferred channel for many to receive their news content. The popularity of a news item is often measured by such things as likes, the number of comments, and the number of times it has been shared. The goal of our research was to use several exploratory and predictive modeling techniques to build a model that would predict the number of shares of online news articles as well as identify those features that contribute most significantly to a news item's popularity.

The dataset "Online News Popularity" is publicly available by the University of California, Irvine from their machine learning repository. The data represent online articles published by the news website Mashable during a two-year period from January 2013 to January 2015. Included are 61 attributes related to the 39,797 articles collected during that period.

They type of information that was extracted from each news article fell into several categories: number of words, number of links in the article, amount of digital media (photos & videos), the day of the week, keyword analysis, and natural language processing scores that evaluated various angles of positivity and negativity.

The results of our work will help authors determine the most effective elements that influence whether their article gets seen and read by many or by few. In our current climate, online news spreads via shares. There is a lot of focus on figuring out how to increase the number of shares for news articles.

Our initial exploration and analysis was focused on finding the key features that contribute to popularity as well as on discovering latent meaning in the data.

When we predict something, we want to find items or variables that influence and help predict the outcome. On the other hand, we don't want those items to influence each other or it will confuse the model. In our case, there were many items from our group of predictors that influenced each other. Before we could proceed, we had to transform them. We used a process called principal component analysis or PCA. This grouped together predictors in a way that had them not confuse the final model. Another benefit of this technique is that it reveals patterns and information "below the surface" that is often very useful.

Using exploratory factor analysis, we established and ranked five general concepts that this dataset is measuring: size and positivity of words, positivity of content, popularity and number of links to other Mashable articles, popularity of keywords, and size of the content of the article. These five concepts helped us understand what questions are most important in predicting popularity of the articles.

Canonical Correlation Analysis (CCA) is concerned with assessing the relationship between two sets of variables, i.e. can variable group A be said to determine variable group B, and the reverse. Unlike standard regression which is used to predict a single variable, CCA attempts to predict multiple dependent variables from multiple independent variables. For the purposes of this study, the results from the PCA were used as the groups of variables.

Correspondence analysis was found to be useful to determine any correlation between our categorical variables, day of the week and Mashable data channel (article category). The idea behind this was to discover, within the dataset, the best day to publish a specific article category. Future application of this result can be beneficial for bloggers or those in the digital marketing industry to maximize their website popularity by posting subject-specific articles on the optimal day.

All the techniques were used to choose the best, most predictive combination of predictors that were included in a linear regression model. However, we found that the linear model was not an effective nor predictive model.

At this point, we chose to focus on popularity, which really was the underlying goal of predicting shares. Since linear regression was not successful, we thought classification models may be a better option. Our team used the methods decision tree and random forest that help predict outcomes by sorting them into groups or classes. To make this work, we had to divide up the shares into smaller groups called bins. Then the model was used to predict which bin the article would fall into.

The findings from our research revealed that strong positive and negative words have the most influence on whether people will share an article with their network. This knowledge can be used by journalists and bloggers to write more effectively and have a broader reach.

Future work that needs to be done is to get a better understanding of the types of positive and negative words that have the most impact. Regarding this data set, study of the attributes for articles with numbers of shares outside the normal range would be worthwhile. The median number of share for this dataset was 1400 shares. Of future interest are those articles with shares in the hundreds of thousands.

Our main conclusion is that emotion is key to an online news article being shared by many. Not surprisingly, news articles with very strong positive or negative messages get shared the most. With this knowledge, articles can be reviewed for sentiment and those not showing strong emotion, either positive or negative, can be rewritten to elicit a stronger emotional response from the reader. The danger here, however, is that in an effort to increase a news item’s popularity, emotional bias is introduced, clouding the facts.