Social Media Analytics Homework #1 Write-Up

Group Members:
Matt Barrett, Tim Lai, Christine Mulcahy, Elena Reynolds, Brett Scroggins

Part 1:

The purpose of this portion of the assignment was to identify predictors of social media members with high influence and quantify that influence to a financial value. By quantifying the financial value of the identified influencers; an indirect means of quantifying the financial value of identifying the highly influential is also derived.

The dataset was pulled from the corresponding Kaggle page, and highly correlated variables were identified. This process was aided by the visualization of a heatmap. Once these highly correlated variables were determined, a smaller dataset was created using only variables that provided uniquely useful information.

The next step was to transform the data into a way that can be best utilized by the eventual models. Based upon the information in the dataset, the most useful transformation was determined to be taking the difference between the two individuals in each entry. This provided the differences in followers, list count, mentions, retweets, and posts between individuals as inputs to the generated model.

The data was then split into training and test sets using **33%** of the entries as the test set. Three models were then generated: a logistic regression, a random forest, and AdaBoost. The mean accuracies between the random forest and AdaBoost models were comparable, but the **AdaBoost model** was slightly higher at 76.42%. The AdaBoost model was then implemented in identifying the importance of the features in the data. The most useful features proved to be **list count difference, follower count difference, and retweets received difference**.

The expected profit of each retailer marketing strategy was calculated. Using the difference between the expected value of each of these options, the increase in value by identifying the influencers was determined to be **$10.58 Million** dollars. This increase equates to a lift of **1.394**; or an increase in expected profit of 39.4%. This was then compared to the value of having a 'perfect model' in which perfect information on influencers is known. The expected value of having a perfect model was calculated to be **$13.44 Million** dollars over the no analytics model, an increase of 50.1%. This indicates that having the realistic analytics provides most of the potential profit of perfect information.

** Note: Please see Jupyter Notebook for code and calculations.

Part 2:

The purpose of the second portion of this assignment was to apply learned information from the first part to identify leading influencers on Twitter. Five thousand tweets related to the state of the union address were compiled using the Twitter API. From these tweets a dataset was built that included each Twitter user that was included in these 5,000 tweets: users who wrote those tweets and any users that were retweeted in those tweets. Additionally, the needed information of list count, follower count, and retweets was added to the dataframe.

Using this created dataset and the relative value of these variables determined in part 1, the Twitter users were ranked by their estimated influence value. The values of list count, follower count, and retweets were normalized, and weighted by coefficients calculated by each variable's relative importance. However, there was one major issue discovered at this project stage.

In many cases the Twitter user was added to the dataset via being retweeted and not having a direct tweet in the dataset. This resulted in many instances where the number of followers and list count was not known. For prominent users, these values were manually found and entered, but for others this was determined to not be highly valuable. If a future major project was to arrive at this dilemma, a series of Twitter API queries could be written to wholly complete these disparities.

The resulting 10 most influential Twitter users were identified: MSNBC, ACLU, elizabethforma, glamourmag, NYDailyNews, ajplus, AC360, tomcolicchio, TomiLahren, and B75434425. These results match fairly closely with expectations with some noteworthy differences. The highest score of influence was attained by a news outlet. Next was the American Civil Liberties Union, followed by the Twitter account of Sen. Elizabeth Warren. Though these Twitter users are not news outlets, they are outspoken critics of President Trump with passionate followers that support their opinions.

Upon review of some of the surprising users, Glamour Magazine as the next influencer is an interesting case. At first, Glamour Magazine does not seem associated with politics, however, from current events we have hypothesized that perhaps many of Glamour Magazine's female readers paid additional attention to the address. The Glamour Magazine account therefore may have been a central point of influence to its readers on conveying thoughts (ie: related to the recent women's marches, etc.)

Additionally, the seventh account on the list was that of a chef and family man. While this stands out as wildly unexpected, it shows the nature of social media viral events. If a tweet is sent that resounds with readers, any opinion can climb to the highest tier of influence.

One final conclusion drawn from this portion of the assignment was the lack of lukewarm sources from being highly influential. This points to the nature of stronger opinions carrying much farther influence than those with a more neutral tone. Qualitatively, this makes sense since people often use social media to follow others with similar sentiments and use it as an echo chamber for reinforcing their opinions.