Majdi Mariano

MGMT 6160

Case Assignment 3

In this assignment we will evaluate a data set that includes a large list of attributes that could potentially enable us to predict the target feature, in this case it is ‘shares. The data includes over 30,000 articles posted by Mashable, in which the number of shares that each tuple (article) had received is listed along with 58 predictive features and 1 non-predictive feature.

In this case, due to the variety in shares per Mashable article, we initially binarized the data set by converting all share values greater than or equal to 1400 into 1 and all those less than 1400 into 0. 1400 being the 50th percentile of all values in the data set. Following this set we had set up other measures to accurately evaluate the data to measure the predictability of certain features in the data. These measures included creating groups of similar features to assess the correlation between each grouped feature and the target feature. These groups included: Token Features, Channel Features, Kew Word Features, Self-Reference Features, Day of the Week Features, LDA Features, Global Features and Emotional Features. In choosing which features would remain in the group it was important to note that many of the features were redundant. Many were taken as fractions of a combination of other features, due to this we had only grouped those features which did not include rates or did not have relationships to two or more features, this makes readability easier.

It is very important to first see the correlation of our features to our target because it will give us early insight into some features that could provide predictability in the future of our analysis. In this case it is crucial to see the low correlations between almost all predictive features measure to that of our target feature. Majority if not all the predictive variables have a correlation value below 0.1, showing us quite early on that it is possible these features will not be able to give much insight into what causes the greatest number of shares.

Below are several heat maps depicting the correlation between grouped attributes to that of our target feature. It is important to not the legend and change in maximum and minimum values. Most dark blocks are assumed to be large numbers but, in this case, they are the maximum correlation relative to the others in the heatmap – again most of which are below 0.1. Though even with small correlating values it is safe to assume that positive correlations, especially when implemented on together could have a positive effect on the number of shares an article would receive. In this case increasing the rate of positive words in an article that has to do with tech or social media, while also posting on the weekend could be an example of measure taken that has the ability to increase the number of shares that an article would receive. Though in order to better understand how accurate this would be, we must further evaluate the predictability using applications such as KNN, Logistic regression and Decision Trees.

Chart, histogram

Description automatically generatedChart

Description automatically generatedChart

Description automatically generated with low confidenceTimeline

Description automatically generatedChart

Description automatically generated

Prior to performing a KNN analysis we split the data into ‘x’ and ‘y’. ‘x’ containing all the predictive features in the data set excluding the target feature (shares) and non-predictive features (url and timedelta). ‘y’ contains the target data (shares). Following the separation of our data we standardized it using two techniques: Standard scalar technique and MinMax scalar. This was only don’t to see if one method would prove to be more accurate than the other upon evaluation. We set the n\_neighbors value to 3 and in both cases, using both standardization techniques, the cross-value score, rounded to the nearest 2 decimals, came out to 0.59. A score of 0.59 is relatively low in the grand scheme of things and shows predictability like that of a coin-toss, and not a data set with 58 predictive features. Before making any conclusions, we performed 1 more predictability analysis to see if the fault lied only in the KNN analysis.

In the final predictive analysis, we performed a train test split with a test size of 60% in order to use the values in a Decision Tree Classifier. The resulting accuracy score came out to a belittling 0.58. Furthermore, a cross-value score of 0.55 shows a lower score than the KNN analysis. Following the underwhelming scores of both the KNN and DTC analysis, we provided a chart showing the most important features through Logistic Regression and Linear Regression in which the top 20 most important features are shown below. Logistic regression being on the left and Linear Regression on the right.

Text

Description automatically generatedText

Description automatically generated

In understanding these predictive measurements, and why the predictability is so low it is important to understand what the features represent and how these features interact with each other. In this specific data set, there are many features which provide redundancy in the data, this increases the noise in the data allowing for accuracy to be, as we have seen, very low. Upon analyzing and reducing the data it is clear, that if we were CEO and forced to decide on how to improve predictability, the data set must be split into similar features, similar to grouping that we had done previously. All ‘rates’, ‘subjectivity’ or ‘polarity’ features should be separated from or put onto an alternative data set. These features give very misleading data into predictive applications overall reducing the efficiency and accuracy. If we were to implement any such measure from what was gathered, we would solely make changes based off the correlation charts shown in the beginning of this paper. In implementing these changes together, it may have the best positive net effect on the amount of shares that an article receives, though the level at which it works will be determined after it is already implemented.