**CS 7280 - Project Proposal**
**Analysis and Prediction on Online News Popularity**

Gabriel Bakiewicz
Darshan Patel

**Problem Description:**

Social Media has become an epicenter of online news distribution and consumption. More than half of the social media site users have shared news stories, images or videos and nearly as many have discussed the news on social media [1]. As a result, there is an increased interest in identifying the articles that will receive a significant amount of user attention.

The Online News Popularity Data Set has a set of features extracted from articles published by Mashable in a period of two years [2]. The primary goal is to predict the number of shares on social media. It is also possible to explore a different approach, turning the problem into classification by defining classes on intervals of the number of shares.

**Data Summary:**

There are 61 attributes in the dataset: 58 are predictive, 2 are non-predictive and 1 is the goal attribute. They represent information extracted directly from online news articles, like unique words in text and number of images, and meta features, like text polarity and closeness to LDA of topics. Some of the features are binary and represent a one hot encoding of a categorical feature (there are two: “week day” and “data channel”, encoded using 8 and 6 attributes respectively). Also, there are 18 attributes which are different summary (Avg, Min, Max) values of 6 underlying attributes such as polarity of positive/negative words and shares of worst/average/best keyword.

An exploratory data analysis will show if there are missing values (there should not be, according to the dataset description), outliers and/or zero variance attributes. The measure of pairwise correlation can give a hint on some attributes that can be removed or combined.

**Methods:**

For the problem of predicting the number of shares, Multivariate Regression models will be used. General linear tests will be used to check for linear relationship between predictor and the target; together with feature elimination on the data analysis step, it should bring down the number of predictors significantly.

After selecting a reduced number of attributes, we estimate the regression coefficients and analyze the residual plots to look for violations of the assumptions of the model. Then the models can be used for inference in new observations and to create confidence intervals.

If the classification approach is chosen, Multinomial Logistic Regression will be used in a similar way as described above.

**Reference:**

[1] http://kng.ht/1R5VqIE
[2] K. Fernandes, P. Vinagre and P. Cortez. A Proactive Intelligent Decision Support System for Predicting the Popularity of Online News. Proceedings of the 17th EPIA 2015 - Portuguese Conference on Artificial Intelligence, September, Coimbra, Portugal