<h1><b>Introduction </b></h1>

The purpose of this document is to satisfy the Capstone Project for completion of the Data Science Workshop through SlideRule. The Data Science Workshop is a collection of courses related to the field with the intention of providing a good foundation in Data Science.

For the Capstone Project I have elected to participate in a Kaggle Competition related to the Analytics Edge course offered through edX by MIT. This competition used data about New York Times blots articles and the purpose was to create a model which would predict if an article would be popular or not based on the variables provided. Once submitted the models were evaluated using AUC, which is a commonly used metric for binary classification problems. It provides the proportion of time that the model predicted the correct value. This competition ran for three weeks from April 14, 2015 – May 4, 2015.

<h1><b>Data Set Outline</b></h1>

Two datasets were provided, a training set with 6,532 articles and a test set with 1,870 articles. The dependent variable is <b>Popular</b> which is binary and set to 1 if an article had 25 or more comments in its online comment section. There were 8 independent variables in the datasets and are as follows.

1. NewDesk = the New York Times desk that produced the story
2. SectionName = section the article appeared in
3. SubsectionName = subsection the article appeared in
4. Headline = title of the article
5. Snippet = small portion of article text
6. Abstract = summary of blog article, written by the New York Times
7. WordCount = number of words in the article
8. PubDate = Publication Date
9. UniqueID = unique identifier for each article

(source: https://www.kaggle.com/c/15-071x-the-analytics-edge-competition-spring-2015/data)

<h1><b>Exploratory Data Analysis </b></h1>

The purpose of Exploratory Data Analysis (EDA) is to allow the analyst to understand the data set and start to gauge what may or may not be related. During this phase the analyst will often create plots, calculate statistics and understand which variables could be manipulated for use in the model.

<h2><b>Summary Data Set Information </b></h2>

To start it is important to get a sense of what the underlying data looks like, first looking at the percentage of the records within the training set that are noted as being popular. This gives us a baseline accuracy that we can work from, if we assume that all predictions are popular.

Scatterplots

Correlation Matrix

Word cloud?

Bucket word count? Graphically – popular by number or words

<h1><b>Model Ideation/Creation</b></h1>

Simple logistic model (no text)

Text with Logistics

Show dendrogram for CART model

Final Model Used and Results

Conclusions

How I did

What else would I do? Advanced algorithm – Bayesian, k-fold cross validation within the random forest

Additional factorization – change hours to time of day – morning, afternoon, evening, night

Log of Word Count

Determination of the most important words

1. one model made with the function bayesglm()
2. one model made with the randomForest()

sharing your code.





