Model 1 - Highest LB AUC ~ .932

1. Data cleaning of NewsDesk & SectionName by using knowledge of NYT web layout schema. Getting rid of Subsection altogether.
2. Log(WordCount + 1)
3. Feature Engineering of Weekday, Hour, IsWeekEnd, IsWorkHours.
4. Feature Engineering of Emotionally charged topic words IsSex, IsReligion, IsPolitics.
5. Feature Engineering of Punctuation that would invite comments/sharing - IsQuestionMark
6. Bag of Words on Headline and Abstract
7. Modified Sparsity to give Headline more importance than Abstract
8. Used untrained GLM to extract Sig Vars
9. CV, Tuned using caret (KNN, GLM, SVM, RF, GBM & CART)
10. Tossed CART initially. Tossed KNN next.
11. Experimented with weighting eventually using xRF = .29 yGLM = .12 zGBM = .30 WSVM=.29

Model 2 - LB AUC of ~.930

1. Data cleaning of NewsDesk & SectionName by using knowledge of NYT web layout schema. Getting rid of Subsection altogether.
2. Log(WordCount + 1)
3. Feature Engineering of Weekday, Hour, IsWeekEnd, IsWorkHours.
4. Read this <https://static1.squarespace.com/static/5353b838e4b0e68461b517cf/t/5385049ae4b01dc8cdbaea6f/1401226394038/what-makes-online-content-viral.pdf>
5. Based on above, I concluded that features based on Section, Day, Hour, Log(WordCount+1) and Sentiment explained Popularity.
6. Feature Engineering of Punctuation that would invite comments/sharing - IsQuestionMark
7. Used Lexicon of Emotional Valence and scored words from -5 (Very Neg) to +5 (Very Pos) thus arriving a Sentiment Score for Headline and Abstract
8. CV, Tuned models with just 15 vars using caret (KNN, GLM, SVM, RF, GBM)
9. Used 55% GBM & 45% RF since both models offered highest AUC and lowest sdAUC.

I strongly suspect my decision to data clean the NewsDesk and SectionName dropped my Private LB score significantly.

Attached is my code and a box plot of what my model AUC looked like.