A. I used  as parameters the **NewDesk**, **SectionName** and **Subsection** Name (filled from each other). In **WordCount** I replaced entries with 0 with the word count from the max(abstract/headline).

B. Custom parameters: **Hour**, **WeekDay**, **IsTheHeadlineQuestion**, **Common unpopular Articles** (like daily clip report, today in politics etc).

C. I noticed that unpopular articles have headlines:

* 1. From *Games* section the "**Variety**:..."
  2. From *Arts* the "**Recap**:..." about "**Gotham**", "**Sleepy**" and "**Knick**" (Recap: was popular on Arts)
  3. From *Health*the "**diabetes|death|disease|vegetarian**"
  4. From *Opinion* the "**friday night|joe on wnyc|reading now**"

And I created custom parameters for those.

D. I created **Bags of Words** from Headlines and Abstracts for the categories *Dealbook*, *Other*, *Opinion*, *Technology* and *U.S.* without the common unpopular articles from B and C.

* 1. Using as stopwords the stopwords('SMART') and some common words as weekdays and extracted dictionaries.
  2. I used SMARTspec = 'bnn' (binary occurrences of terms)
  3. These dictionaries were used on the corpus of all the Headines and Abstracts to create DTMs for all the articles.

Unfortunately I didn't have enough time to use cross validation to fine tune the parameters of xgboost and the sparsity of the DTMs .

My final submission (eta = 0.01, gamma = 1, max\_depth = 9, sub\_sample = 0.8) gave me 0.93529 on the Public LeaderBoard (86 position) and 0.91297 on the Private (21 position).