Kaggle 3 Bloggers Summary Questions (C2-9)

Will Gleave, swg8jq

Kanika Dawar, kd2hr

Andrew Dahbura, amd6ua (Github coordinator)

GITHUB Link : https://github.com/andrewdahbura/sys6018-competition-blogger-characteristics

Who might care about this problem and why?

In this problem we worked on predicting the age of the user based on certain attributes given to us, the main one being *text* that the user wrote.

Demographics (age, gender), Intent (negative/positive towards a topic), and Interests (topic) of the users function as a valuable categorization tool for businesses. They are extremely useful for websites and businesses to gauge their audiences and better serve their needs. By understanding consumers, any business and website can increase marketing efforts to target those most likely to buy and customize products and content targeted uniquely to each sector of the intended audience.

What made this problem challenging?

The biggest roadblock revolved around the size of the data, which took large amounts of time to load as well as run for various regression models such as OLS, Lasso and Ridge regression. Ideally cross-validation would be used to tune multiple parameters such as ideal lambda for L1 and L2 regularization as well as ideal number of features to deal with sparsity of the tfidf matrix. Additionally, for LDA the ideal number of topics would have been solved through the validation approach. All of these methods suffered severe performance speeds due to the large size of the feature set and the original post. Running random forest for feature importance was also not possible given our computational capabilities and the size of the data.

With regards to technical approach, this was the first time that any of us have done text analysis that generated data frames with more than 100 features, which made data exploration involving visualizations and traditional plots also quite difficult. The challenge to narrow down the feature space proved a difficult but worthwhile task. Finally, extensive care was required in addressing various technical approaches and inclusion of various predictors such as sentiment analysis, topic modeling and TF-IDF analysis.

What other problems resemble this problem?

In context of NLP:

1. Topic extraction from the text
2. Sentiment of the text
3. Demographic Prediction (gender/income level/education level/occupation)
4. Interest Area of the text
5. Intention to buy if it’s a consumer centric website/blog post
6. Customer satisfaction ratings