1 Predicting Donor's Choose

1. What is the primary metric you care about in this task? Be sure to clearly state the question about why this is the case.
2. Is the column `teacher_number_of_previously_posted_projects` a good predictor for the approval of the project? Use both a `KNearestNeighbors` and `LogisticRegression` model. Which model performs better? Can you select the best parameters for each? The best penalty for `LogisticRegression`? Form a pipeline that includes a scaling transformation? Compare this to a `DummyClassifier`?
3. What are the top 8 states in terms of raw number approved? The lowest? Show me with a nice barplot.
4. Are these states different from the number of proportion of applications approved by state? Show me.
5. Does your model improve with the inclusion of the `teacher_prefix` column in a `LogisticRegression` model?
6. What is your best parameter for the training set with these inputs?
7. Construct a feature that is simply `STEM`, which is 1 if a scientific discipline is a part of the `subject_subcategory` column, or 0 if not. Did your model improve?
8. What if you include the `project_grade_category` column? Is your model improved?
9. What if your client only cares about what's happening in New York. Is there a difference in the performance of a `LogisticRegression` model? `KNearestNeighbors`?

2 Incorporating Textual Features

Continuing with the Donor's Choose example, we will examine how to make use of the textual information in columns like the `project_essay_1` column.

2.0.1 Problem

Using the `essay_sample` variable (first essay from first row of our Donor's Choose data), use the basic text strategies to do the following:

- remove any punctuation, if important to nature of word use (! vs. ?) determine a way to account for this.
- make sure all words are lowercase
- choose a few words that you believe to be the most important in the essay. Why did you choose these?

2.0.2 Tokenizing Text

2.0.3 Using the essay as features

2.0.4 `min_df`

```
When building the vocabulary ignore terms that have a document     frequency strictly
lower than the given threshold. This value is also     called cut-off in the
literature.     If float, the parameter represents a proportion of documents, integer
absolute counts.     This parameter is ignored if vocabulary is not None.
```

2.0.5 Stop Words

2.0.6 tf-idf

2.0.7 Problem

Use these features in a `LogisticRegression` model. Create a barplot of the top five and bottom five coefficients of the model and their feature name. What do these mean?

2.0.8 n-Grams

2.0.9 NLP Refresher

REGEX TUTORIAL

https://www.analyticsvidhya.com/blog/2015/06/regular-expression-python/

2.0.10 Scraping

2.0.11 Basic NLP

2.0.12 CountVectorizer

2.0.13 Tfidf

3 Basic Scrape

3.0.1 Kanye's Tweets

3.0.2 Sentiment

3.1 Yelp

3.2 Yelp

3.2.1 Simple Classifier

3.2.2 Automating things