Topic Modeling Using Natural Language Processing

Johnny McGregor
General Assembly

Topic Modeling Using Natural Language Processing

Johnny McGregor
General Assembly

Objective

- Build a model that can help a science website identify whether posts to their website is relevant.
- Collect posts from the science and comedy subreddits.
- Use Natural Language Processing to make the text compatible with the model.
  - Represent the words as numbers
- Evaluate how accurate the model is at distinguishing between the two topics.

Data

- Collected using the Reddit API
    - Collected as a JSON file (Java Script Object Notation)
    - Using Python we can convert into a Pandas Dataframe

- 2868 total posts
    - 1344 science
    - 1524 comedy

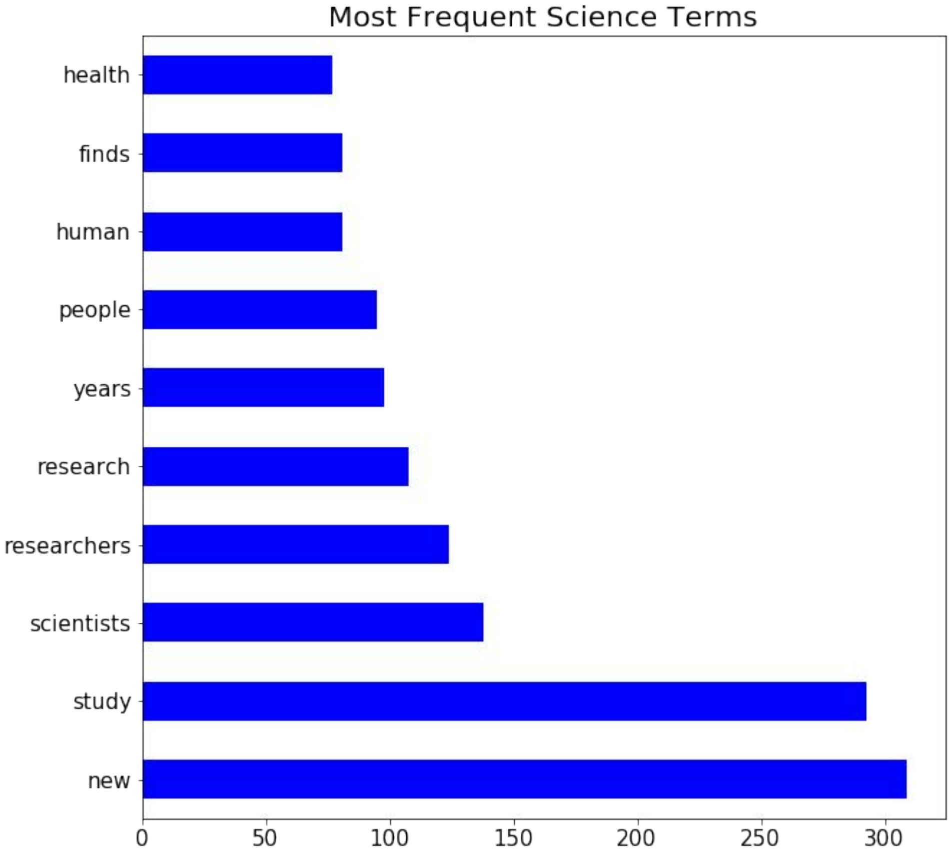
Natural Language Processing

- “Hello, how are you?”
  - hello
  - how
  - are
  - you
- Filter out everything except the letters
  - Comma, questions mark, and quotes are removed above.
  - Each word can become a feature in the model

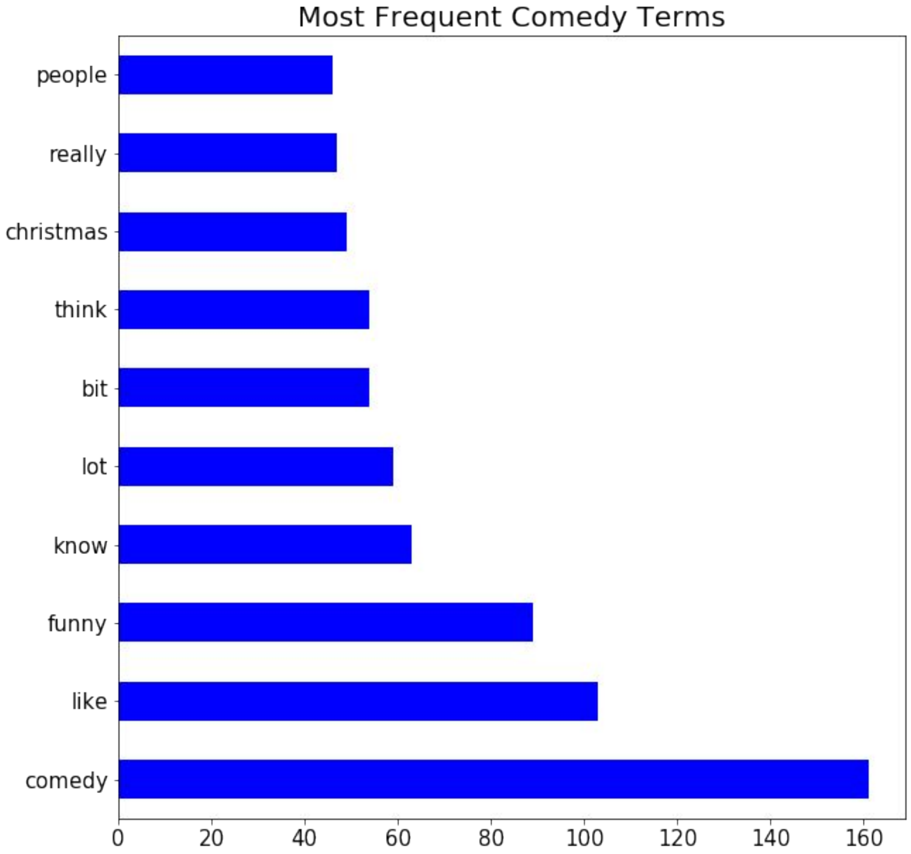
Natural Language Processing

- Word Count
  - Tracks the amount of times a word occurs.
  - Ignores grammar, structure, order
- Frequency
  - Number of times a word is in a document (post) / Total words in document
  - Number of total documents (2868 in this case) / Number of documents that contain the word
- Often a good idea to discard “stop words”
  - Due to high frequency they are usually not meaningful in distinguishing between categories
  - I, by, me, my, of, we you
- Feature Engineering
  - Shortening words to their roots
  - Combining words into pairs/trios

Most Common Words



Most Frequent Science Terms



Most Frequent Comedy Terms

Modeling

- Split up our data
  - 2151 posts used for training data set
  - 717 posts used for test data set
- Fit Training data to the model
  - Set certain parameters like maximum features, maximum document frequency
- Predict whether each post in the test data came from science or comedy subreddit

How Do We Measure Success?

- Of the 717 posts in the test data set how many were correctly identified?
  - Accuracy
- Of the posts that were about science how many were correctly identified?
  - True Positive Rate
- Of the posts that weren’t about science how many were correctly identified?
  - True Negative Rate
- Of all science predictions how many were correct?
  - Precision

How Do We Measure Success?

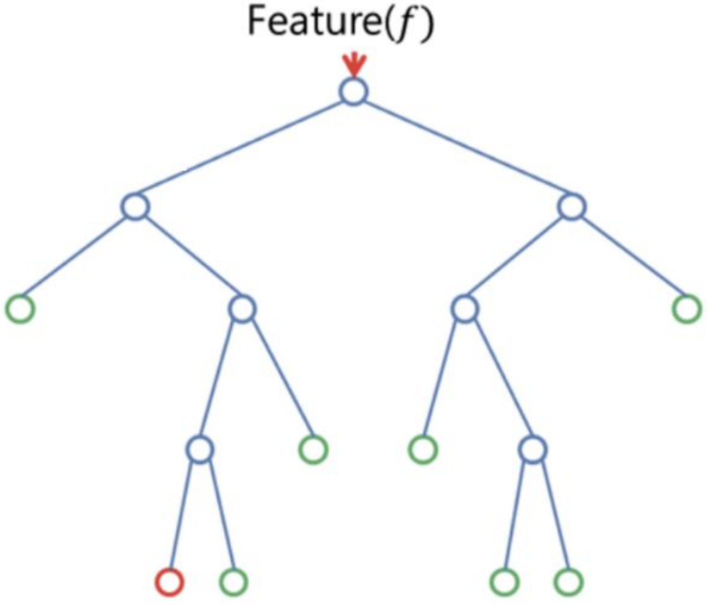
- Accuracy
- Low False Positives Rate (1 - True Positive Rate)
- Baseline Score
  - 1524 comedy posts/ 1344 science posts
  - If we predicted every post to be comedy (or not science) we would be right 53.1% of the time

Logistic Regression

- Makes predictions based on the probability of a post being in either science or not science. 1 or 0.
- Can quantify the likelihood of a 1 or a zero based on each feature. The word study in a post makes it *x* times more likely to be a science post.
    - Three and half times more likely to be classified science If “study” is in the post
    - Probability .78

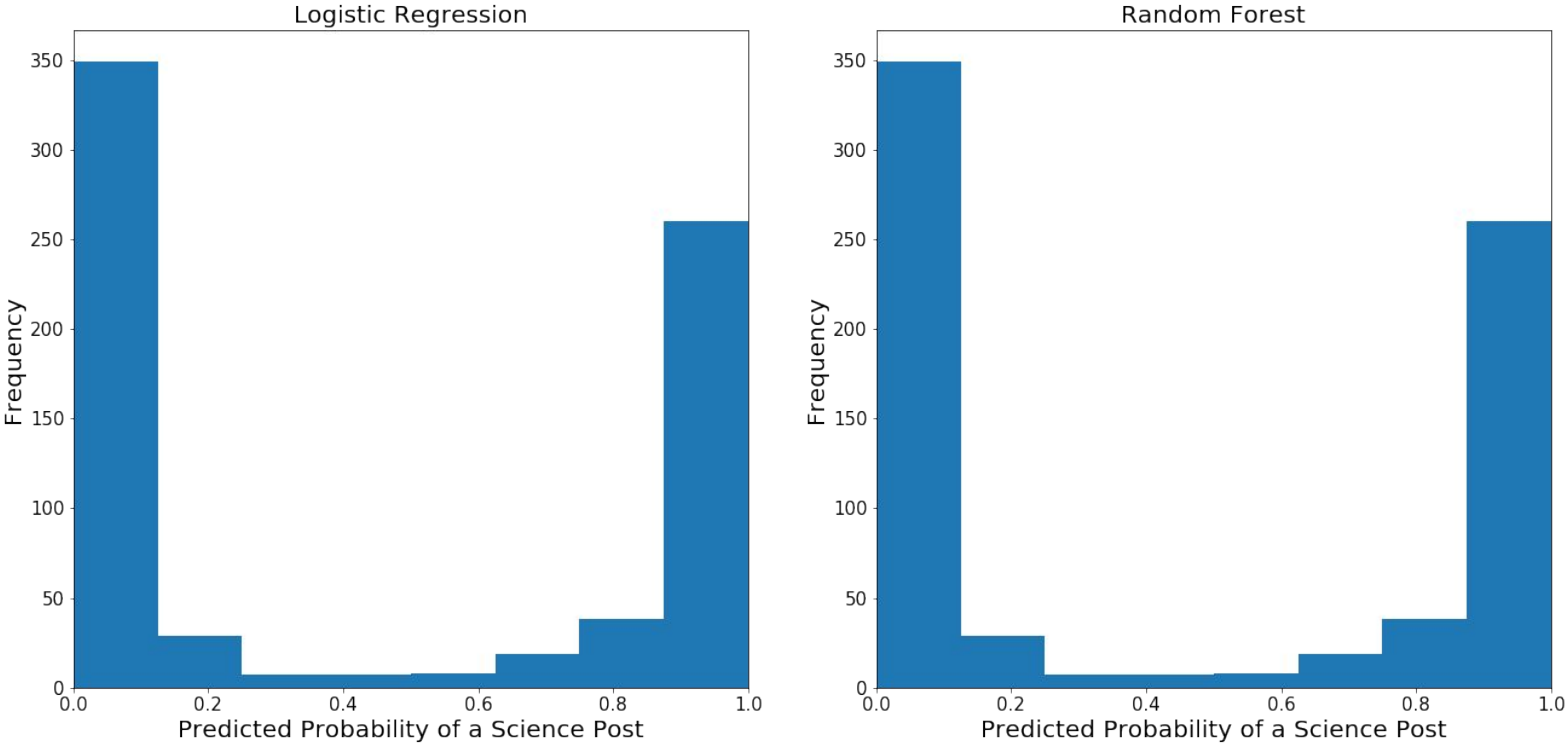
Random Forest

- Splits the data up based on random subset of features
- Can set the depth of the “trees”, maximum amount of features,
- The goal is to separate the two classes as much as possible and this is done many times over with a many Decision Tree Classifier



Decision Tree

Model Comparison



Model Comparison (717 Possible Outcomes)

Logistic Regression:

Accuracy - 97.6

True Negatives - 378 (Rate = 99.2)

True Positives - 322 (Rate = 95.8)

False Negatives - 14

False Positives - 3

Precision = 99.0

Random Forest:

Accuracy - 96.9

True Negatives - 378 (Rate = 99.2)

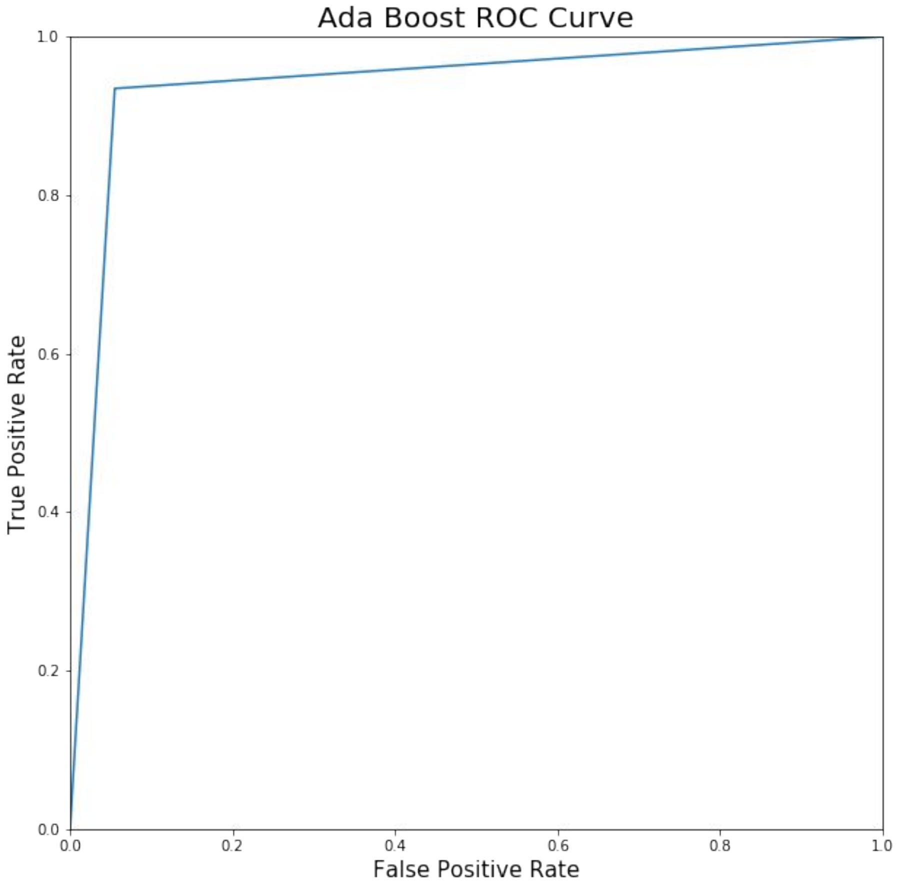
True Positives - 317   (Rate = 94.3)

False Negatives - 19

False Positives - 3

Precision = 99.0

Ada Boost ROC Curve



True Negatives: 363

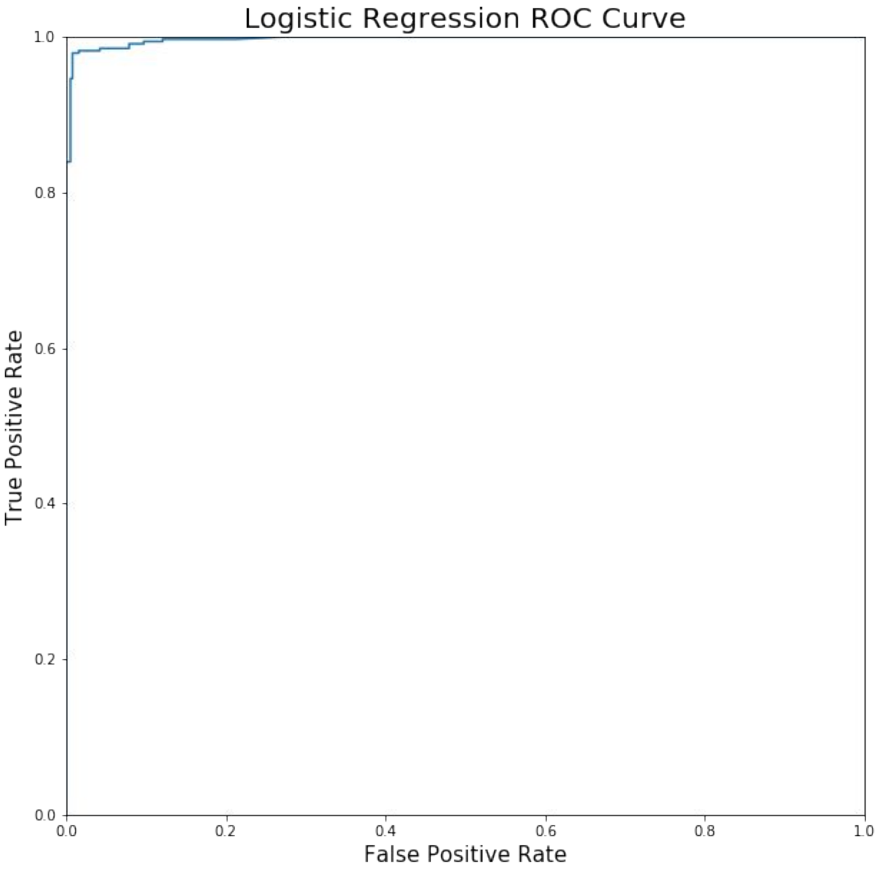
False Positives: 18

False Negatives: 22

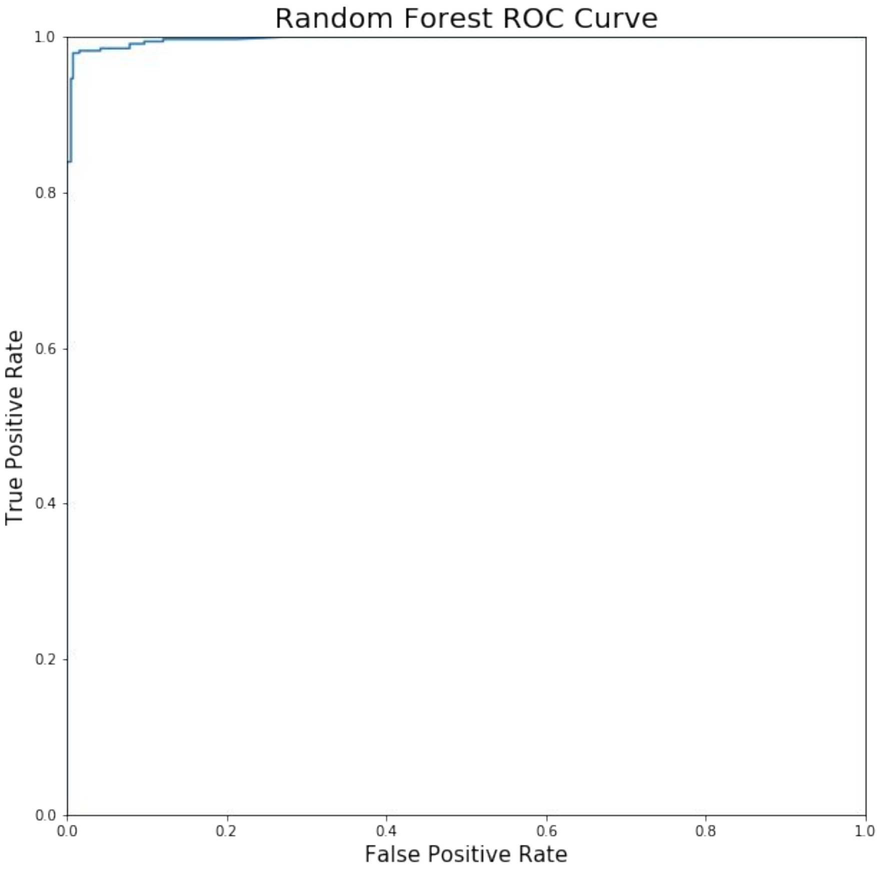
True Positives: 314

Adjust Probability threshold lower to lessen false positives, adjust higher for false negatives

ROC Curves

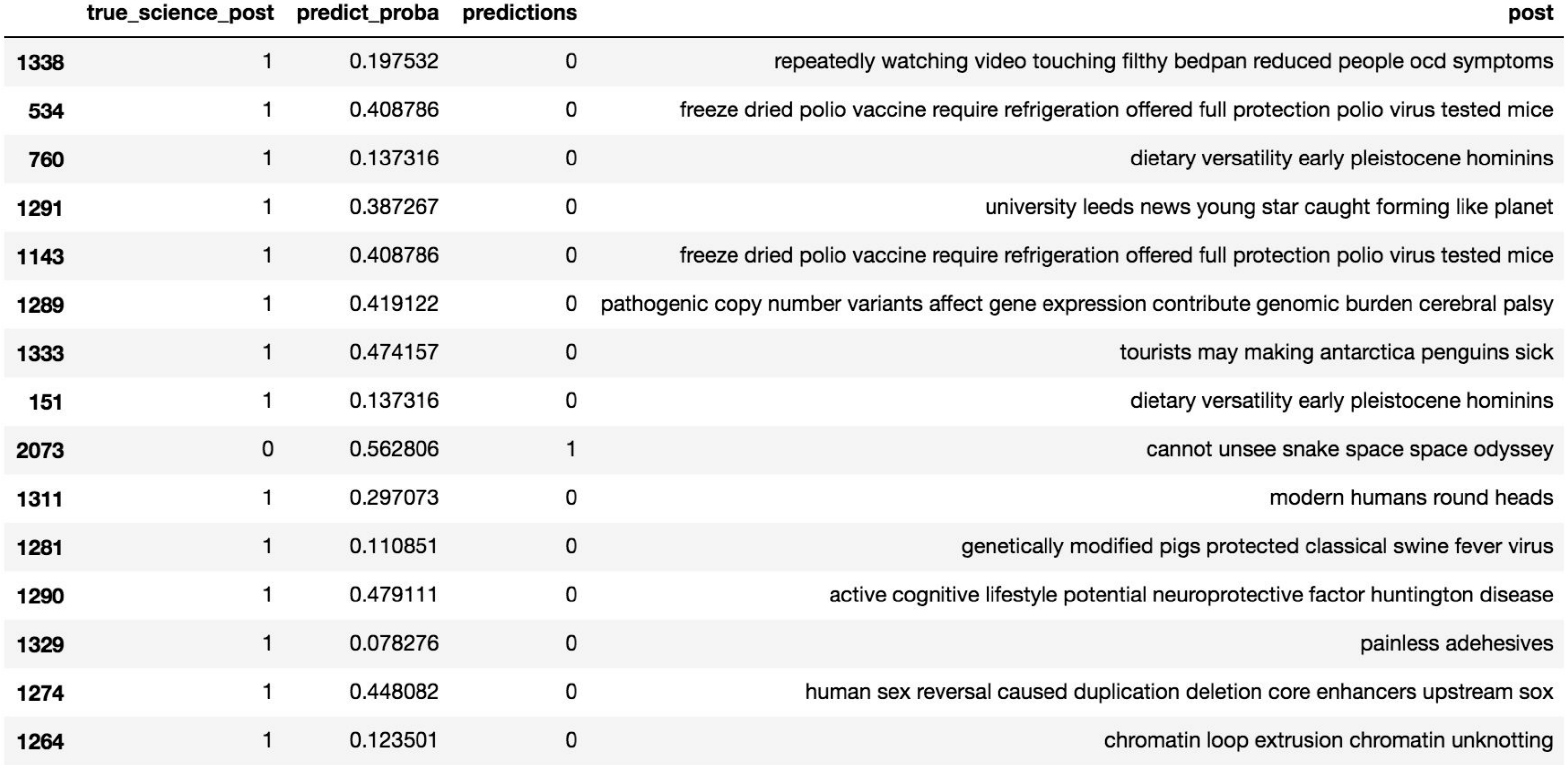


Logistic Regression ROC Curve



Random Forest ROC Curve

Misclassifications



	true_science_post	predict_proba	predictions	post
1338	1	0.197532	0	repeatedly watching video touching filthy bedpan reduced people ocd symptoms
534	1	0.408786	0	freeze dried polio vaccine require refrigeration offered full protection polio virus tested mice
760	1	0.137316	0	dietary versatility early pleistocene hominins
1291	1	0.387267	0	university leeds news young star caught forming like planet
1143	1	0.408786	0	freeze dried polio vaccine require refrigeration offered full protection polio virus tested mice
1289	1	0.419122	0	pathogenic copy number variants affect gene expression contribute genomic burden cerebral palsy
1333	1	0.474157	0	tourists may making antarctica penguins sick
151	1	0.137316	0	dietary versatility early pleistocene hominins
2073	0	0.562806	1	cannot unsee snake space space odyssey
1311	1	0.297073	0	modern humans round heads
1281	1	0.110851	0	genetically modified pigs protected classical swine fever virus
1290	1	0.479111	0	active cognitive lifestyle potential neuroprotective factor huntington disease
1329	1	0.078276	0	painless adehesives
1274	1	0.448082	0	human sex reversal caused duplication deletion core enhancers upstream sox
1264	1	0.123501	0	chromatin loop extrusion chromatin unknotting

Recommendations

- If you want to eliminate posts incorrectly being classified as science, increase the threshold higher than .5
- Random Forest over Logistic Regression
- Look at some common words from misclassified posts to better train the model, and possibly flag inappropriate posts

Objective

- Build a model that can help a science website identify whether posts to their website is relevant.
- Collect posts from the science and comedy subreddits.
- Use Natural Language Processing to make the text compatible with the model.
    - Represent the words as numbers
- Evaluate how accurate the model is at distinguishing between the two topics.

Data

- Collected using the Reddit API
    - Collected as a JSON file (Java Script Object Notation)
    - Using Python we can convert into a Pandas Dataframe

- 2868 total posts
    - 1344 science
    - 1524 comedy

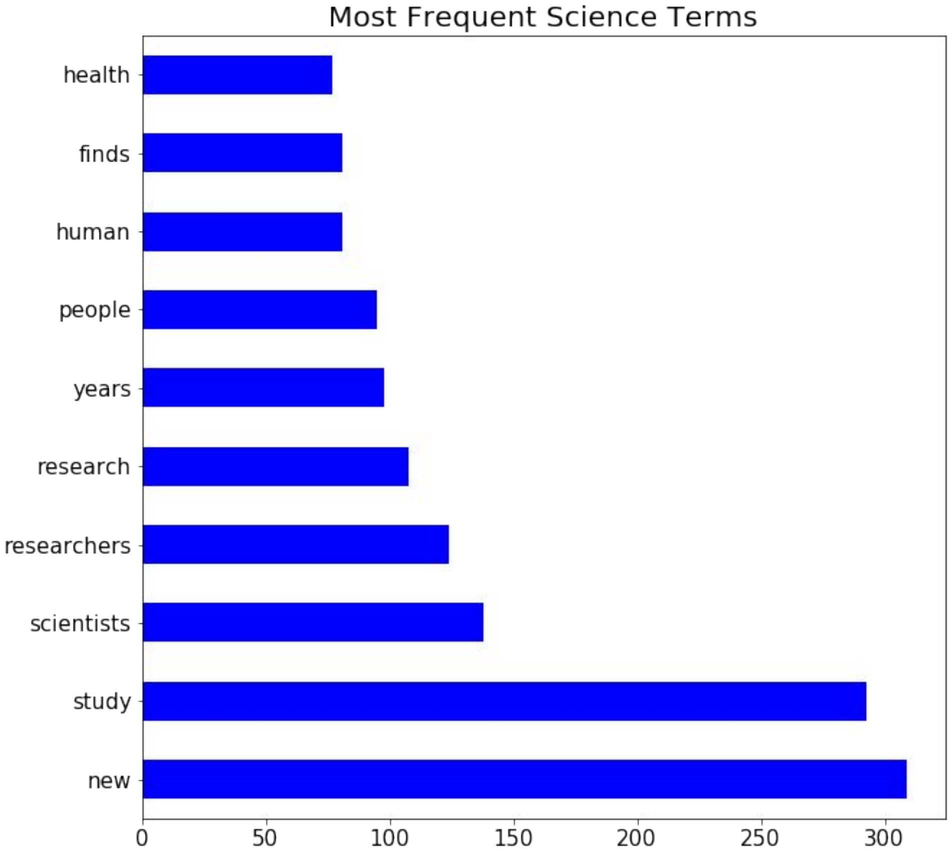
Natural Language Processing

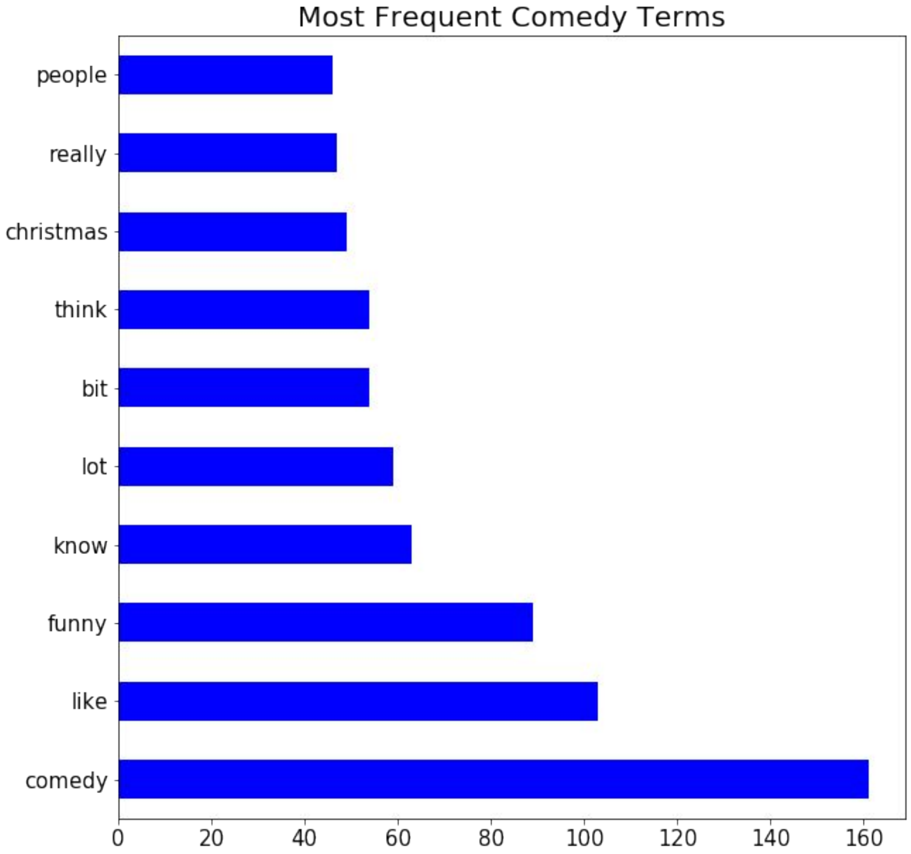
- “Hello, how are you?”
  - hello
  - how
  - are
  - you
- Filter out everything except the letters
  - Comma, questions mark, and quotes are removed above.
  - Each word can become a feature in the model

Natural Language Processing

- Word Count
  - Tracks the amount of times a word occurs.
  - Ignores grammar, structure, order
- Frequency
  - Number of times a word is in a document (post) / Total words in document
  - Number of total documents (2868 in this case) / Number of documents that contain the word
- Often a good idea to discard “stop words”
  - Due to high frequency they are usually not meaningful in distinguishing between categories
  - I, by, me, my, of, we you
- Feature Engineering
  - Shortening words to their roots
  - Combining words into pairs/trios

Most Common Words





Modeling

- Split up our data
    - 2151 posts used for training data set
    - 717 posts used for test data set
- Fit Training data to the model
    - Set certain parameters like maximum features, maximum document frequency
- Predict whether each post in the test data came from science or comedy subreddit

How Do We Measure Success?

- Of the 717 posts in the test data set how many were correctly identified?
  - Accuracy
- Of the posts that were about science how many were correctly identified?
  - True Positive Rate
- Of the posts that weren’t about science how many were correctly identified?
  - True Negative Rate
- Of all science predictions how many were correct?
  - Precision

How Do We Measure Success?

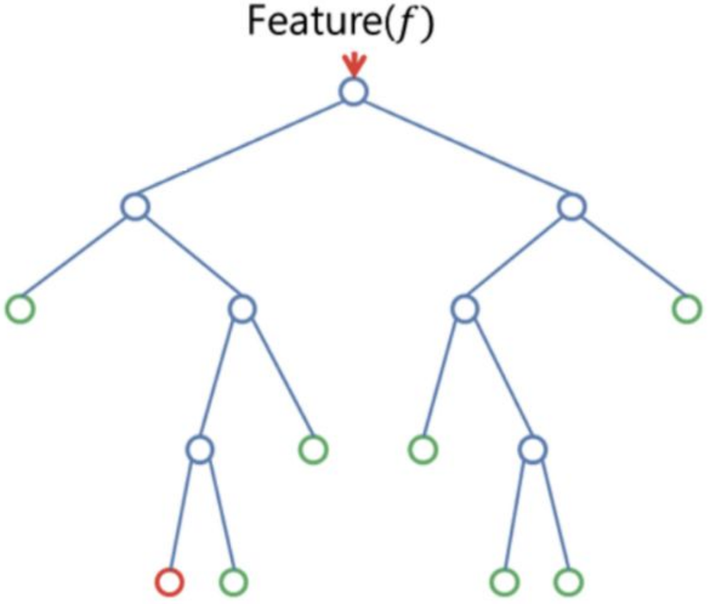
- Accuracy
- Low False Positives Rate (1 - True Positive Rate)
- Baseline Score
    - 1524 comedy posts/ 1344 science posts
    - If we predicted every post to be comedy (or not science) we would be right 53.1% of the time

Logistic Regression

- Makes predictions based on the probability of a post being in either science or not science. 1 or 0.
- Can quantify the likelihood of a 1 or a zero based on each feature. The word study in a post makes it *x* times more likely to be a science post.
  - Three and half times more likely to be classified science If “study” is in the post
  - Probability .78

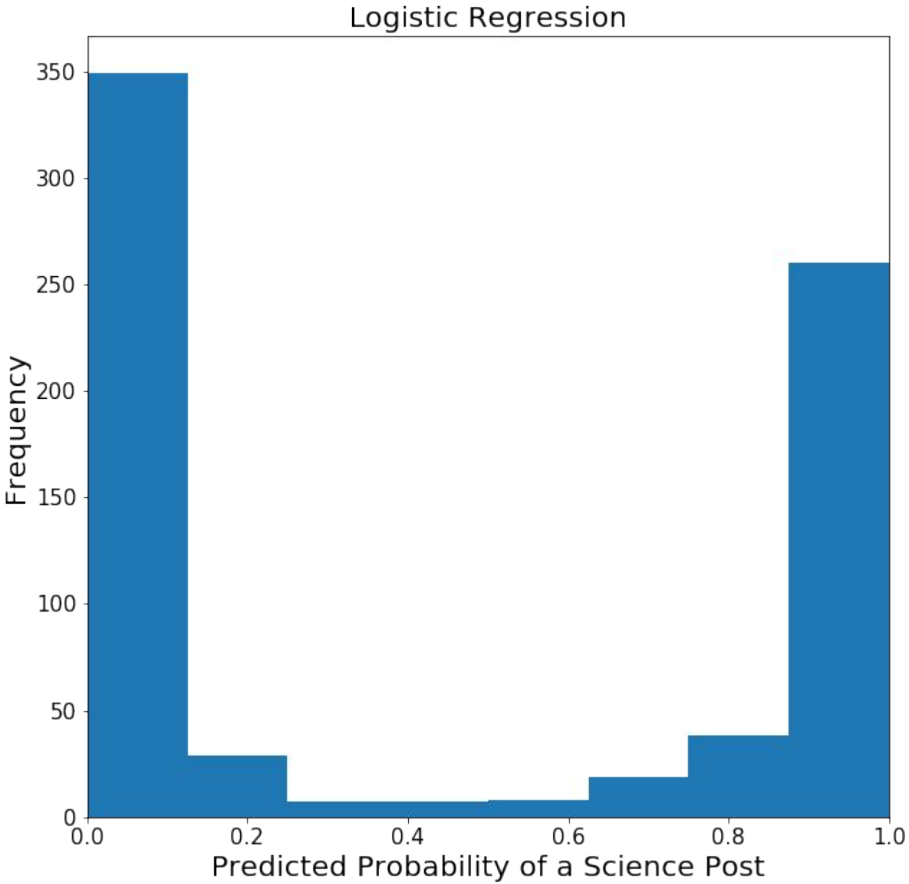
Random Forest

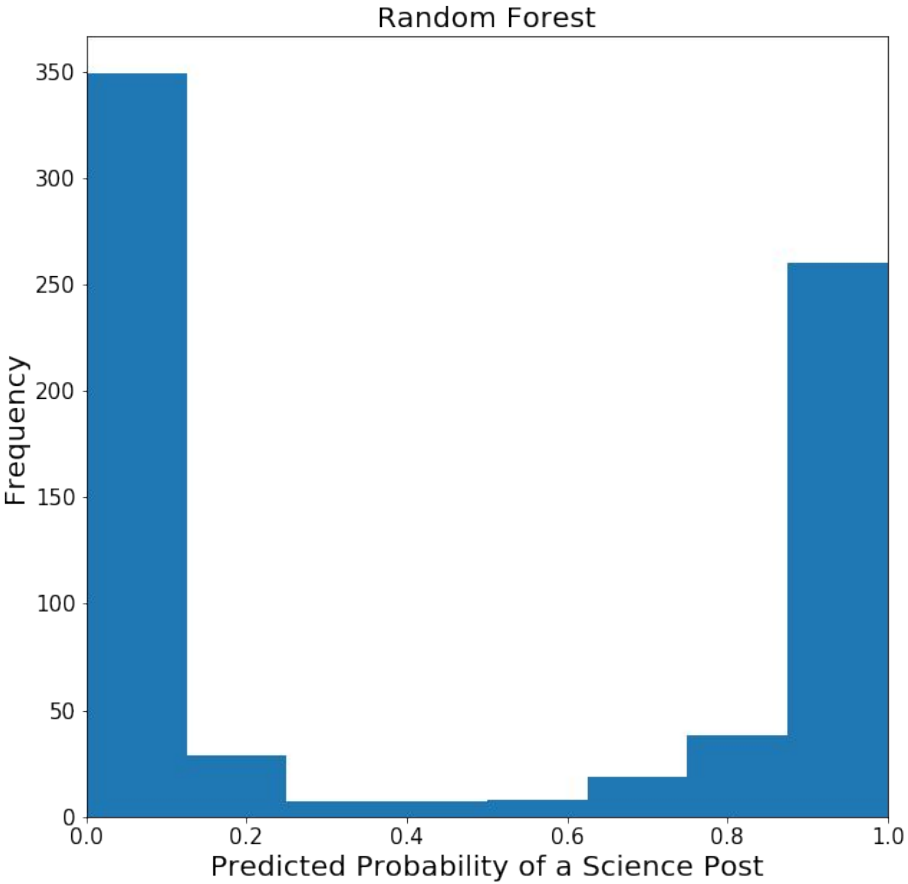
- Splits the data up based on random subset of features
- Can set the depth of the “trees”, maximum amount of features,
- The goal is to separate the two classes as much as possible and this is done many times over with a many Decision Tree Classifier



Decision Tree

Model Comparison





Model Comparison (717 Possible Outcomes)

Logistic Regression:

Accuracy - 97.6

True Negatives - 378 (Rate = 99.2)

True Positives - 322 (Rate = 95.8)

False Negatives - 14

False Positives - 3

Precision = 99.0

Random Forest:

Accuracy - 96.9

True Negatives - 378 (Rate = 99.2)

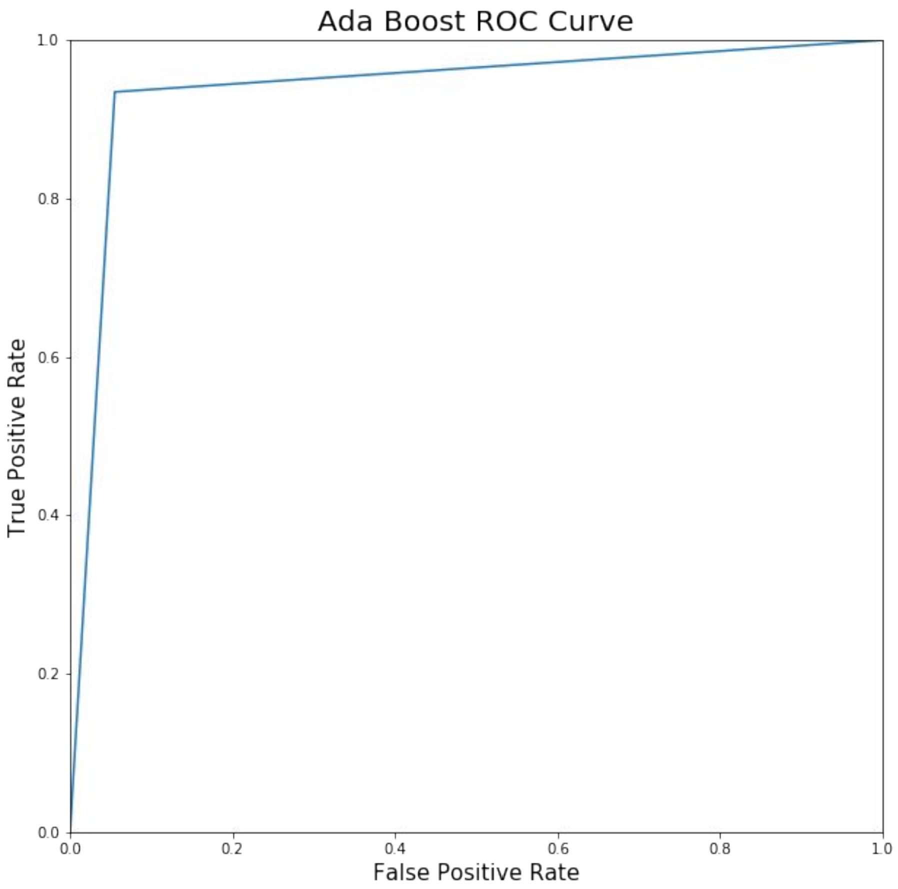
True Positives - 317   (Rate = 94.3)

False Negatives - 19

False Positives - 3

Precision = 99.0

Ada Boost ROC Curve



True Negatives: 363

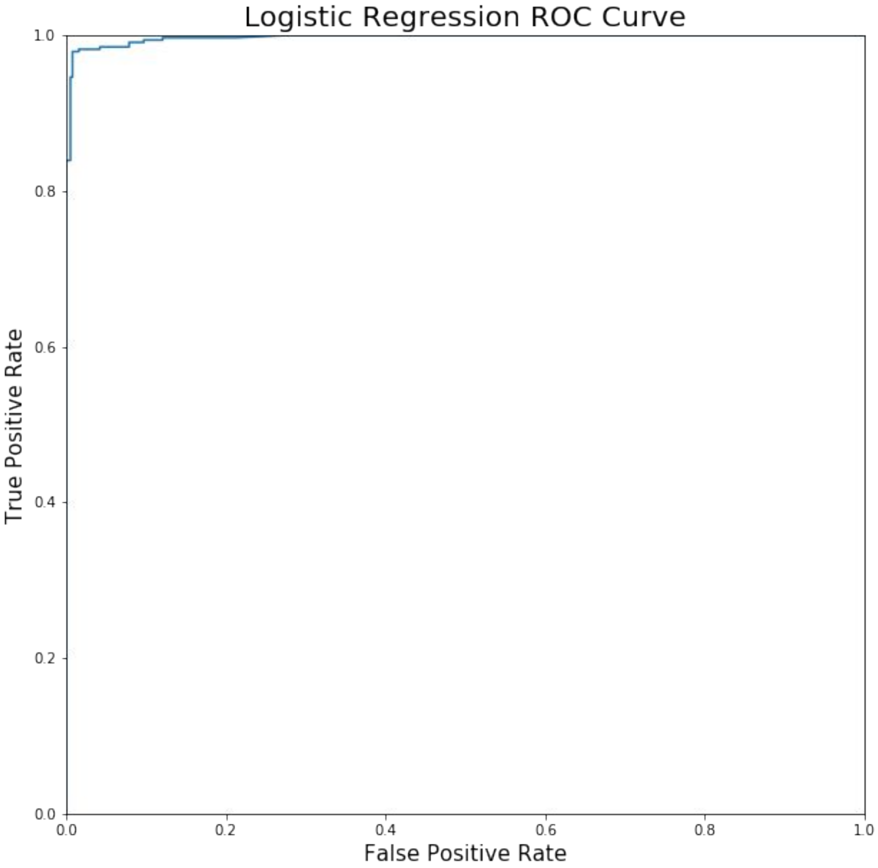
False Positives: 18

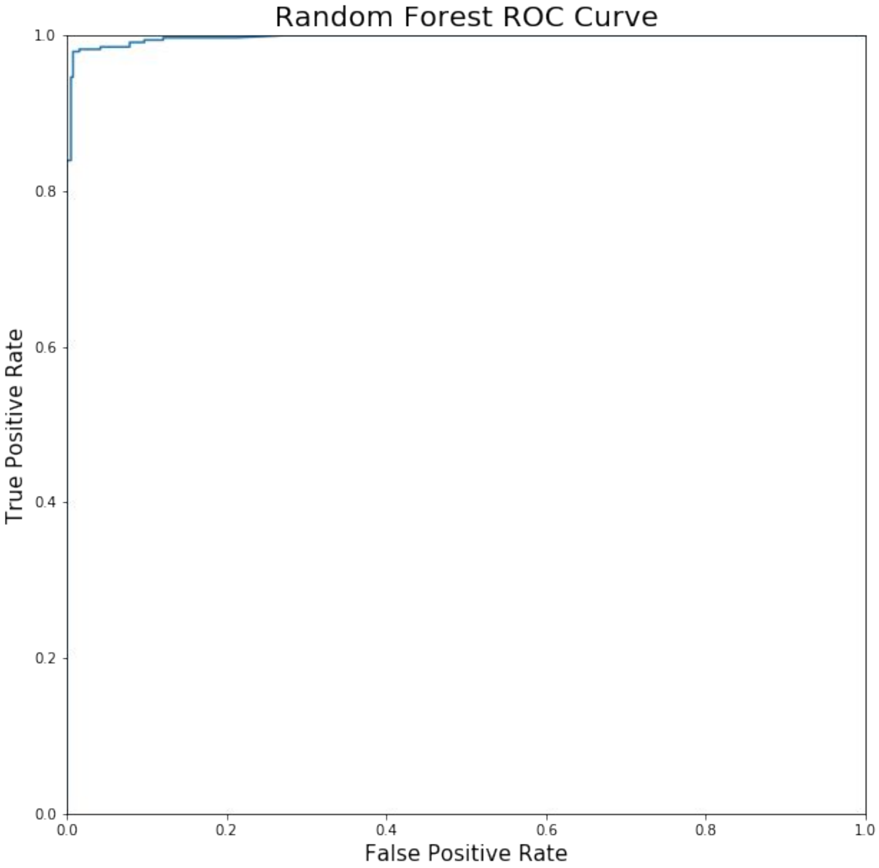
False Negatives: 22

True Positives: 314

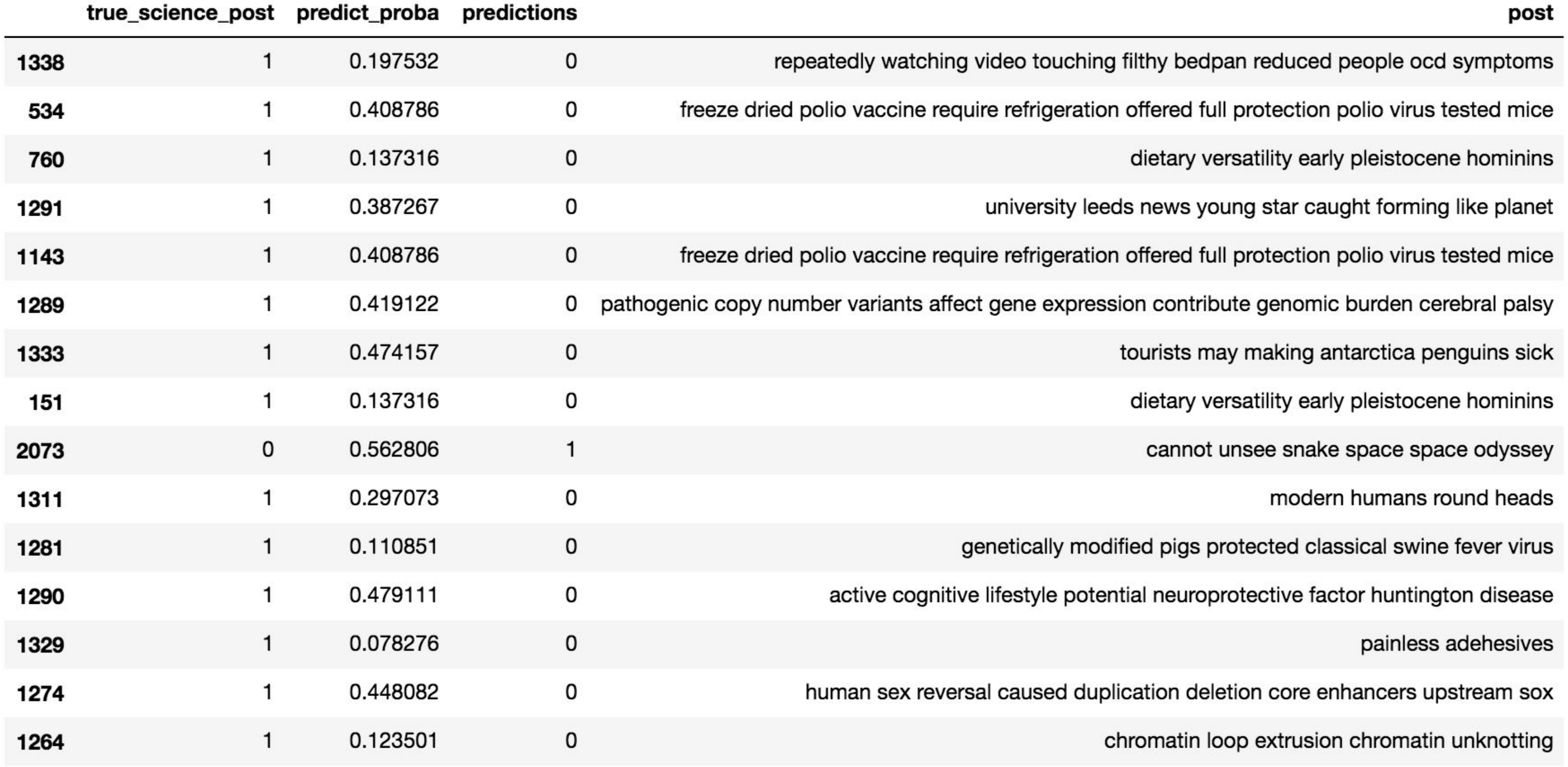
Adjust Probability threshold lower to lessen false positives, adjust higher for false negatives

ROC Curves


Logistic Regression ROC Curve
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.2
0.4
0.6
0.8
1.0


Random Forest ROC Curve
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.2
0.4
0.6
0.8
1.0

Misclassifications



	true_science_post	predict_proba	predictions	post
1338	1	0.197532	0	repeatedly watching video touching filthy bedpan reduced people ocd symptoms
534	1	0.408786	0	freeze dried polio vaccine require refrigeration offered full protection polio virus tested mice
760	1	0.137316	0	dietary versatility early pleistocene hominins
1291	1	0.387267	0	university leeds news young star caught forming like planet
1143	1	0.408786	0	freeze dried polio vaccine require refrigeration offered full protection polio virus tested mice
1289	1	0.419122	0	pathogenic copy number variants affect gene expression contribute genomic burden cerebral palsy
1333	1	0.474157	0	tourists may making antarctica penguins sick
151	1	0.137316	0	dietary versatility early pleistocene hominins
2073	0	0.562806	1	cannot unsee snake space space odyssey
1311	1	0.297073	0	modern humans round heads
1281	1	0.110851	0	genetically modified pigs protected classical swine fever virus
1290	1	0.479111	0	active cognitive lifestyle potential neuroprotective factor huntington disease
1329	1	0.078276	0	painless adehesives
1274	1	0.448082	0	human sex reversal caused duplication deletion core enhancers upstream sox
1264	1	0.123501	0	chromatin loop extrusion chromatin unknotting

Recommendations

- If you want to eliminate posts incorrectly being classified as science, increase the threshold higher than .5
- Random Forest over Logistic Regression
- Look at some common words from misclassified posts to better train the model, and possibly flag inappropriate posts