WEB API & CLASSIFICATION

Problem Statement

- To build different classification models that can accurately classify between subreddit topics based on text posts

- As a reddit moderator, it is useful to find posts that are posted in the wrong subreddit.

- subreddits chosen r/datascience & r/learnprogramming

  Number of rows in subreddit datascience : 827
  Number of rows in subreddit learnprogramming : 976
 Baseline accuracy: 0.5413200221852468

Modelling process

- Scrape subreddit data

- Clean text, lemmatize and remove stop words

- Convert processed text to vectors using count vectorizor or tf-idf

- Perform grid search over all parameters for all models

- Evaluate models base on accuracy score

- Find out frequently occurring words that are classified correctly and wrongly

# Model Results for r/datascience and r/learnprogramming

Models	Test Score	TP	FP	TN	FN	Precision	Sensitivity	F1
Logistic Regression	0.907	233	31	176	11	0.882	0.955	0.917
K-Nearest Neighbours	0.894	233	37	170	11	0.863	0.955	0.907
Naive Bayes	0.883	238	47	160	6	0.835	0.976	0.900
Random Forest	0.896	232	35	172	12	0.879	0.951	0.914
Extra Trees	0.880	226	36	171	18	0.863	0.927	0.894
Ada Boost	0.858	218	38	169	26	0.852	0.893	0.872

Model Results (Graph)
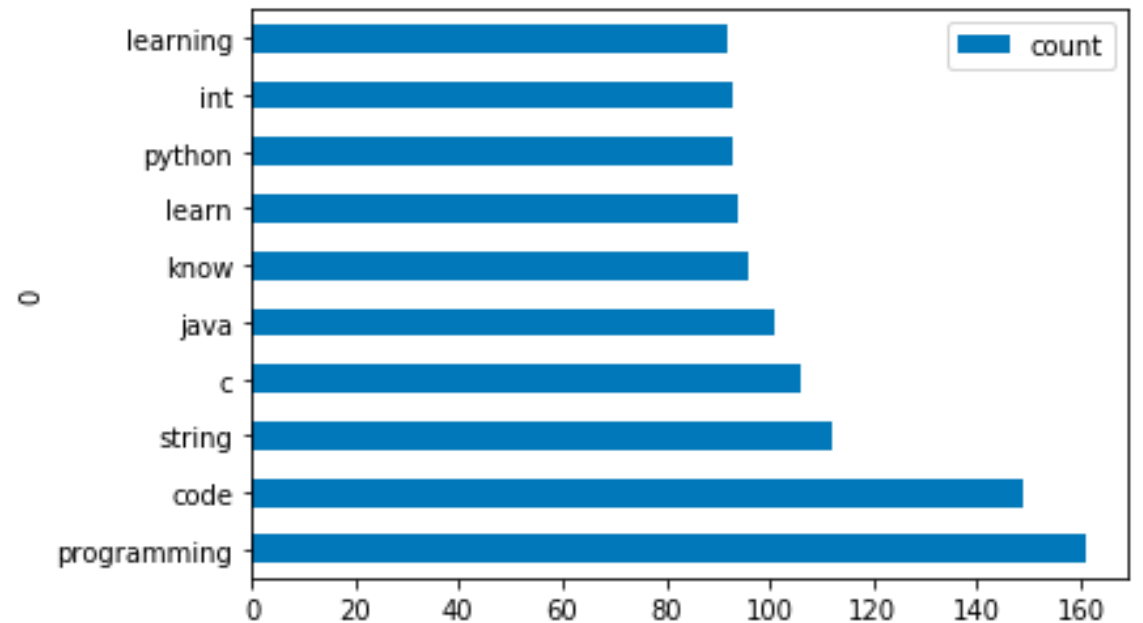

r/datascience & r/learnprogramming
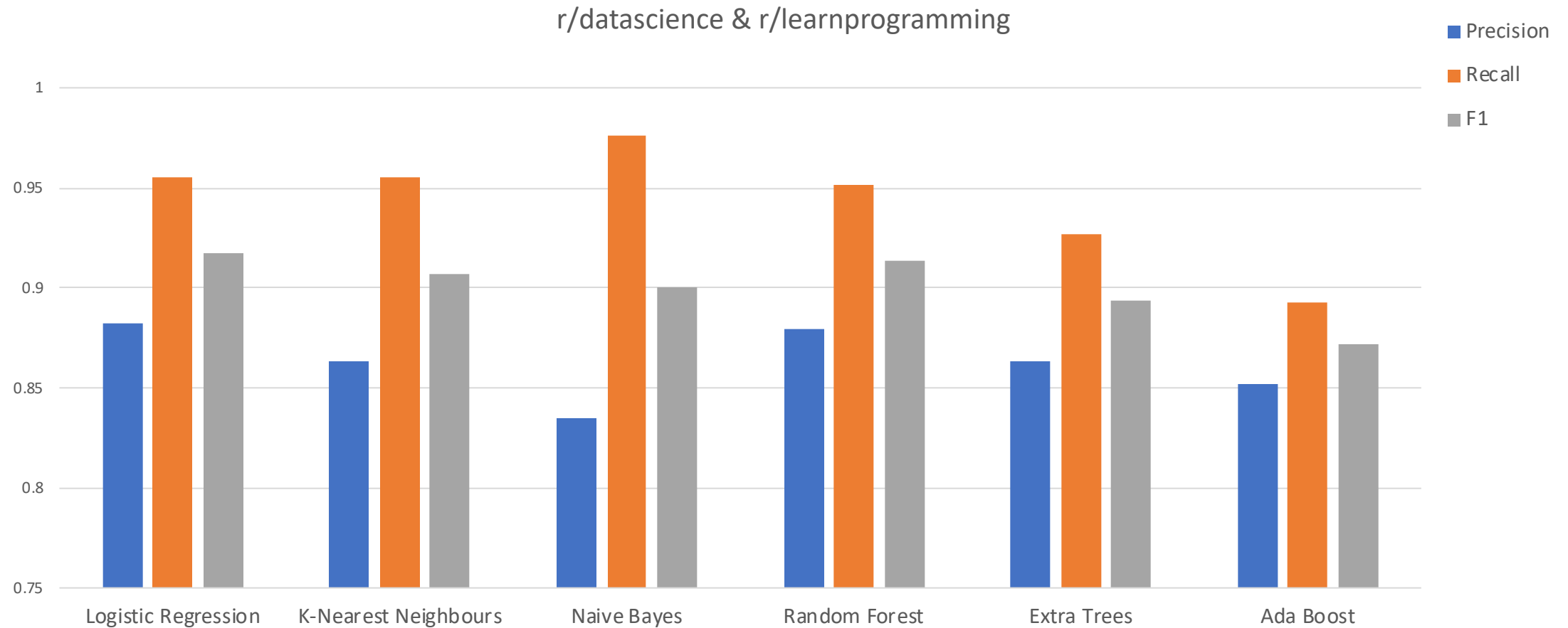
Top words used that are correctly classified
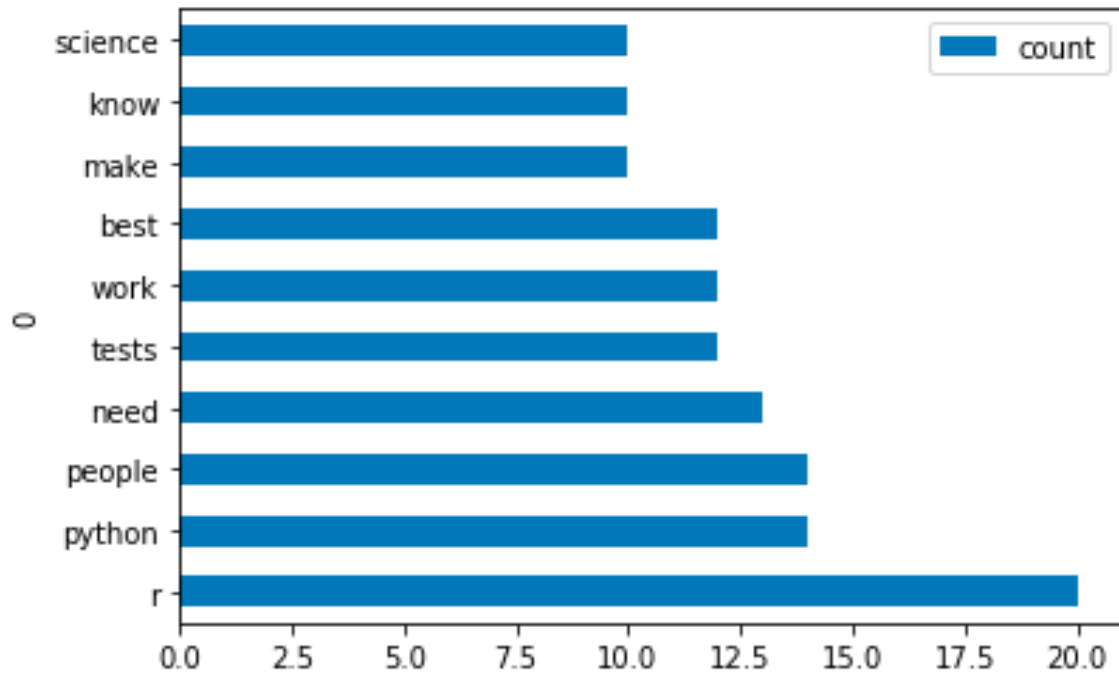

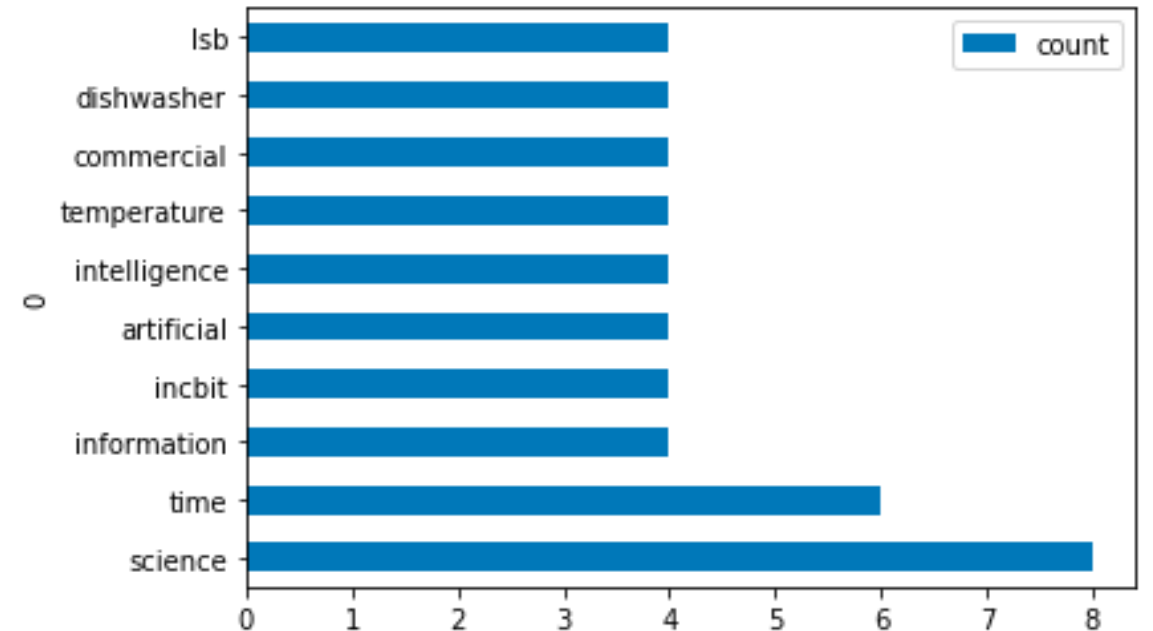r/datascience
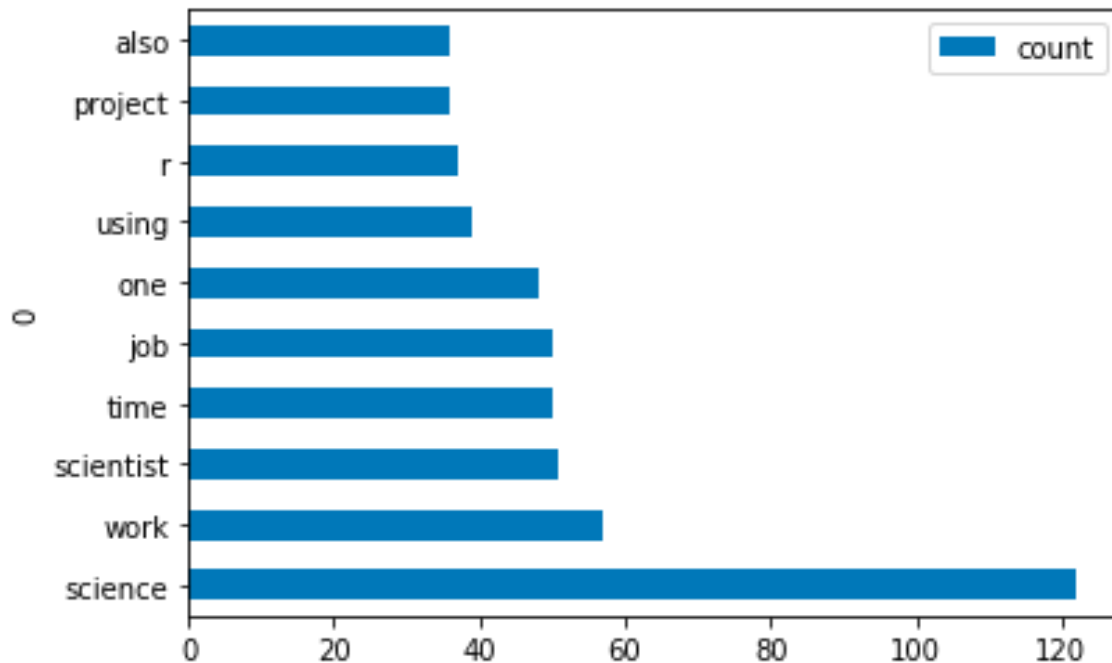
r/learnprogramming

Top words used that are misclassified

r/datascience

r/learnprogramming

Conclusion & Inferences

- There is no one model that can work well for all subreddit

- Best model is logistic regression

- Best parameters for count vectorizer is 1 gram tokenize with 2500 max features, using tfidf transform

- As a moderator it is good to find out those post with words that are misclassified when using the model