Labelbox

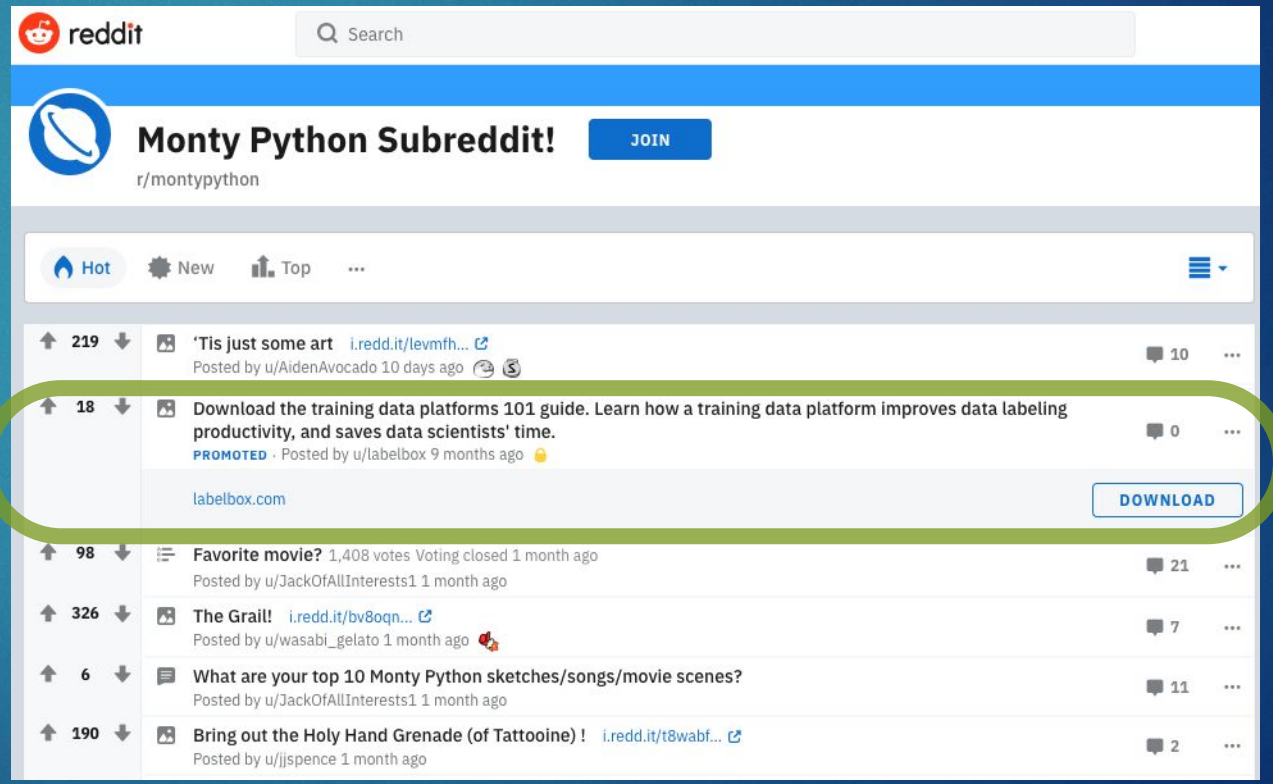
Optimizing Reddit Advertising

ENABLING ONE-TO-ONE MARKETING VIA CLASSIFICATION MODELING

Marketing Challenge

Reddit display ads are a high performing marketing channel for LabelBox, a data science service provider.

- LabelBox does batch marketing on subreddits focused on Python and Data Science.
  - Most campaigns are highly profitable, with Return on Ad Spend (ROAS) of $4 - $7.
- The LabelBox marketing team has discovered ads inadvertently displaying in non-data science subreddits, as shown to the right.
  - These campaigns are proving largely unprofitable, however there has been anecdotal success which has led to some lucrative clients.
- Rather than removing all ads from these "imposter" subreddits, the marketing team wants to individually target reddit users who mis-post data science questions to these subreddits.
  - Competitors are not focused on these users.



< $1 ROAS*

* Return on Ad Spend or ROAS is a common marketing performance measure calculated as:
Net Sales from Campaign / Total Campaign Cost
As a rule of thumb, you need a $3 average ROAS for a marketing channel or campaign to be profitable

reddit

Batch Marketing

Python
r/Python

Data Science
r/datascience

Quick python tips
r/pythontips

Python Education
r/learnpython

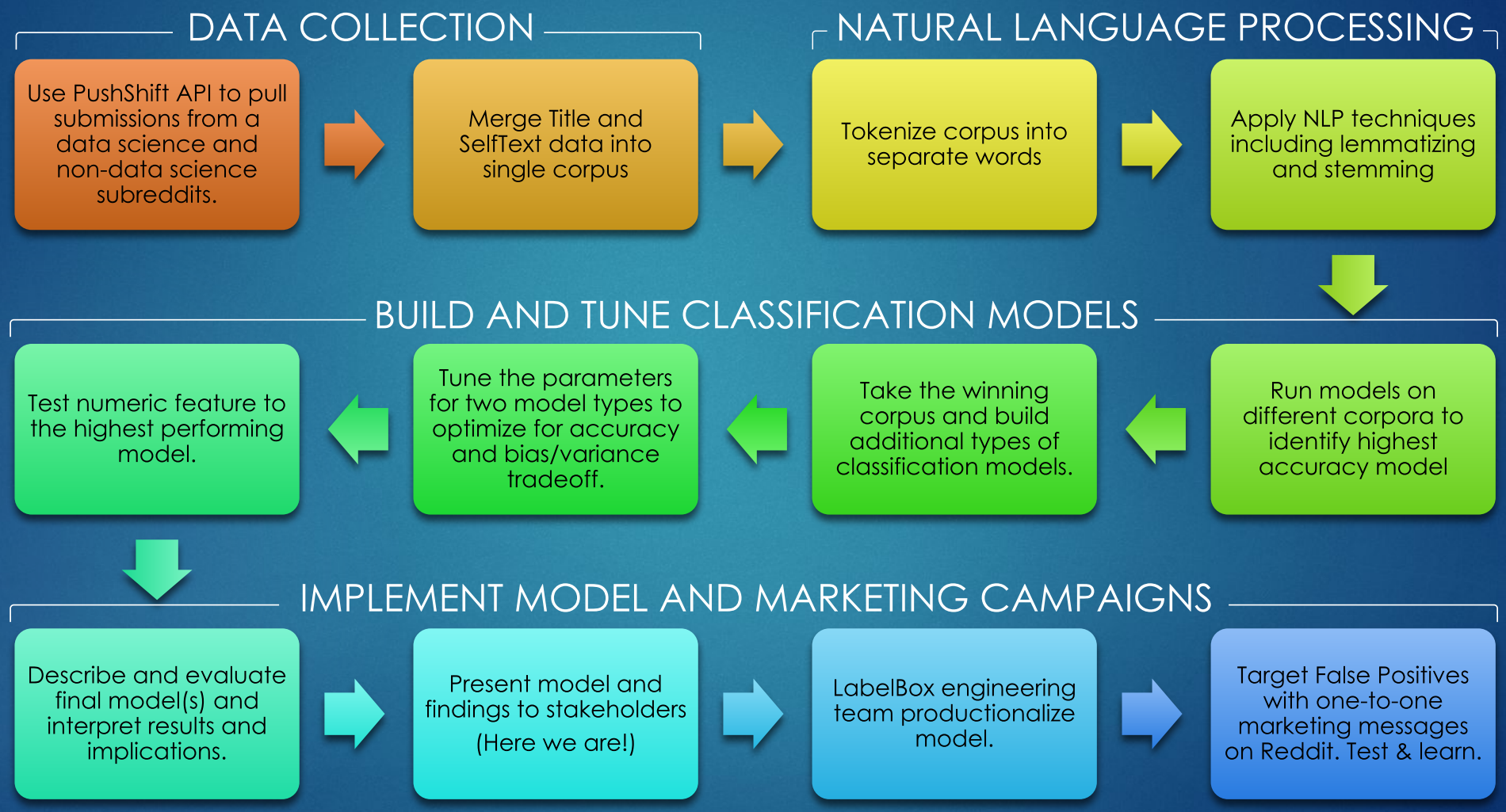
One-on-One Targeting

Monty Python Subreddit!
r/montypython

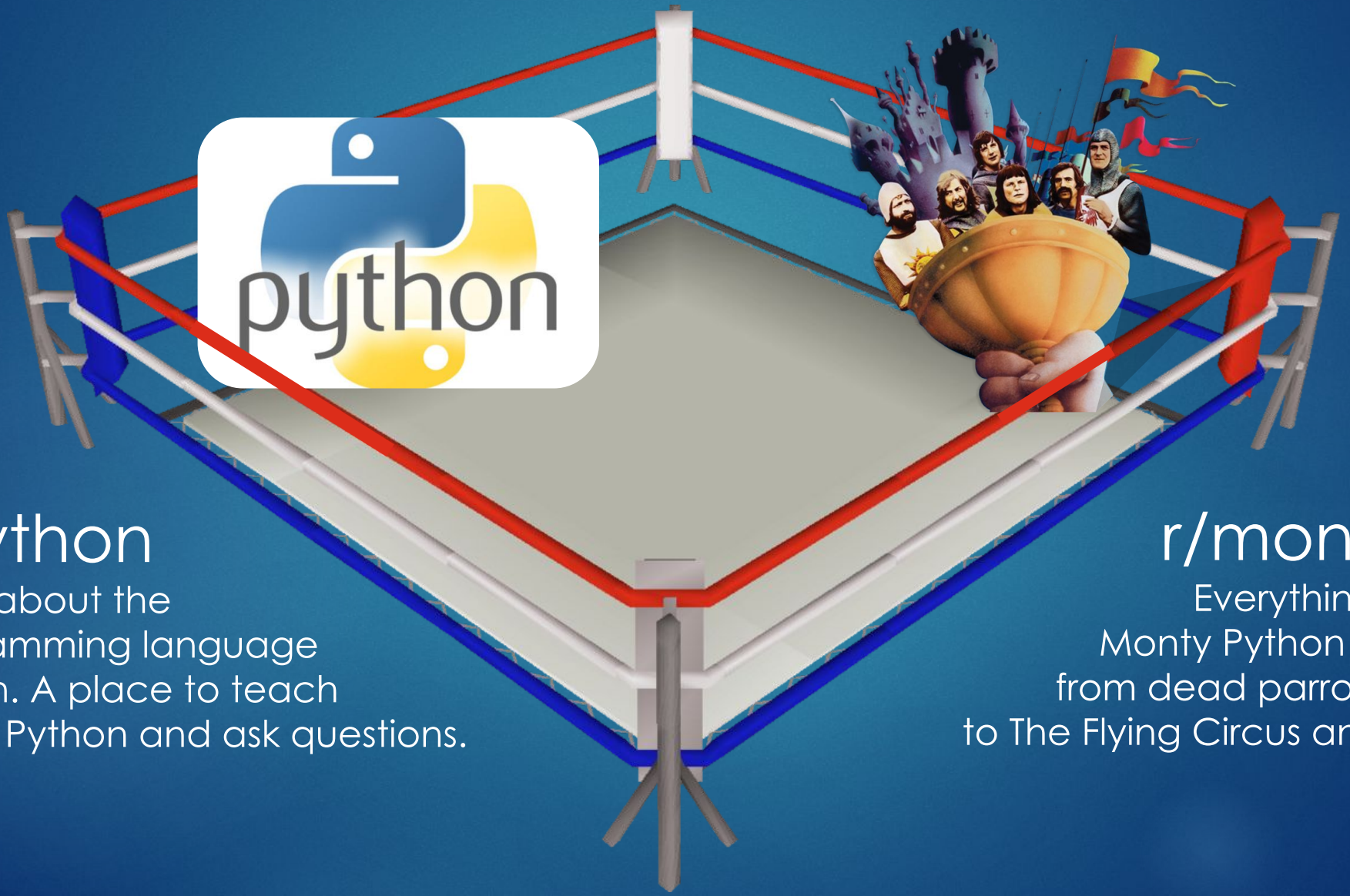
Ball Python
r/ballpython

Everything Science
r/EverythingScience

The Process


DATA COLLECTION
Use PushShift API to pull submissions from a data science and non-data science subreddits.
Merge Title and SelfText data into single corpus
NATURAL LANGUAGE PROCESSING
Tokenize corpus into separate words
Apply NLP techniques including lemmatizing and stemming
BUILD AND TUNE CLASSIFICATION MODELS
Run models on different corpora to identify highest accuracy model
Take the winning corpus and build additional types of classification models.
Tune the parameters for two model types to optimize for accuracy and bias/variance tradeoff.
Test numeric feature to the highest performing model.
IMPLEMENT MODEL AND MARKETING CAMPAIGNS
Describe and evaluate final model(s) and interpret results and implications.
Present model and findings to stakeholders (Here we are!)
LabelBox engineering team productionalize model.
Target False Positives with one-to-one marketing messages on Reddit. Test & learn.

Let's Meet our Subreddits!



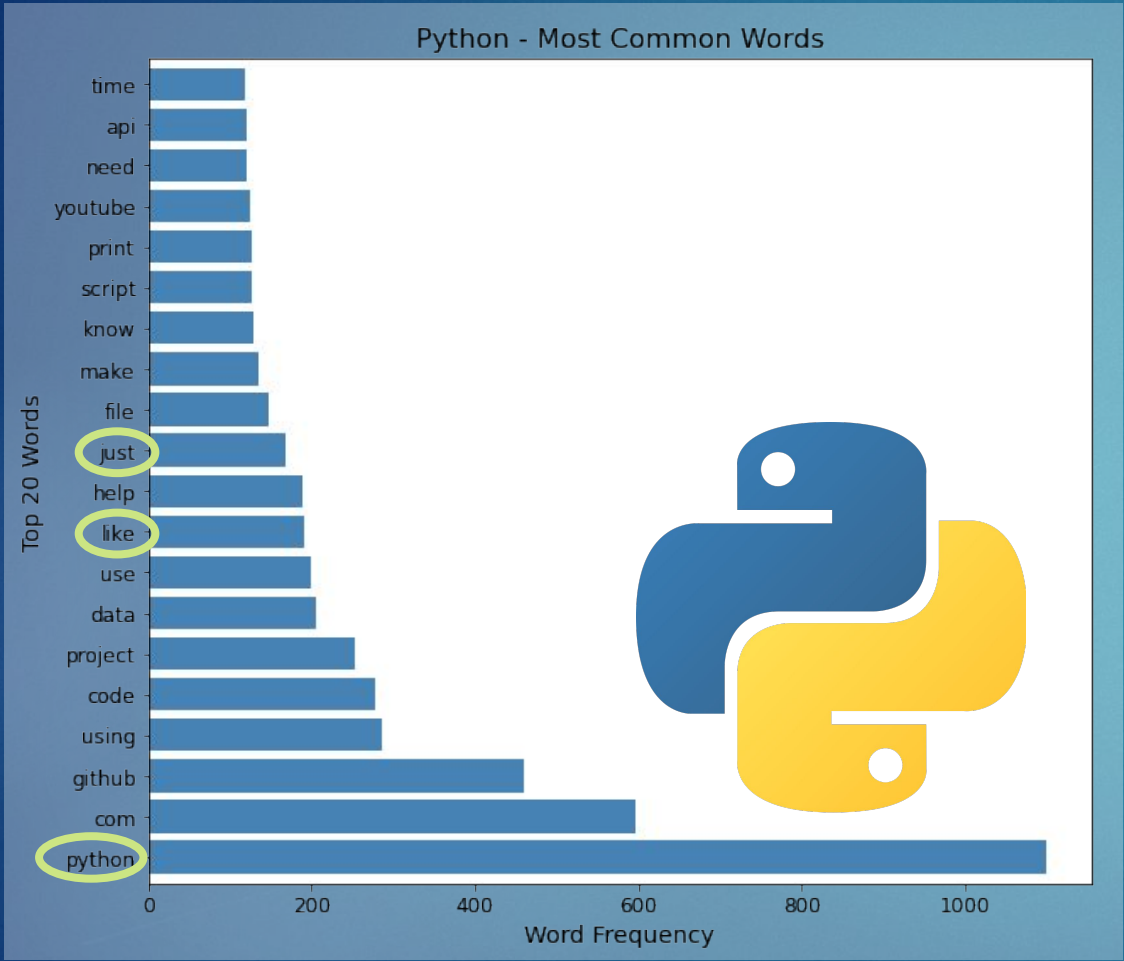
r/Python

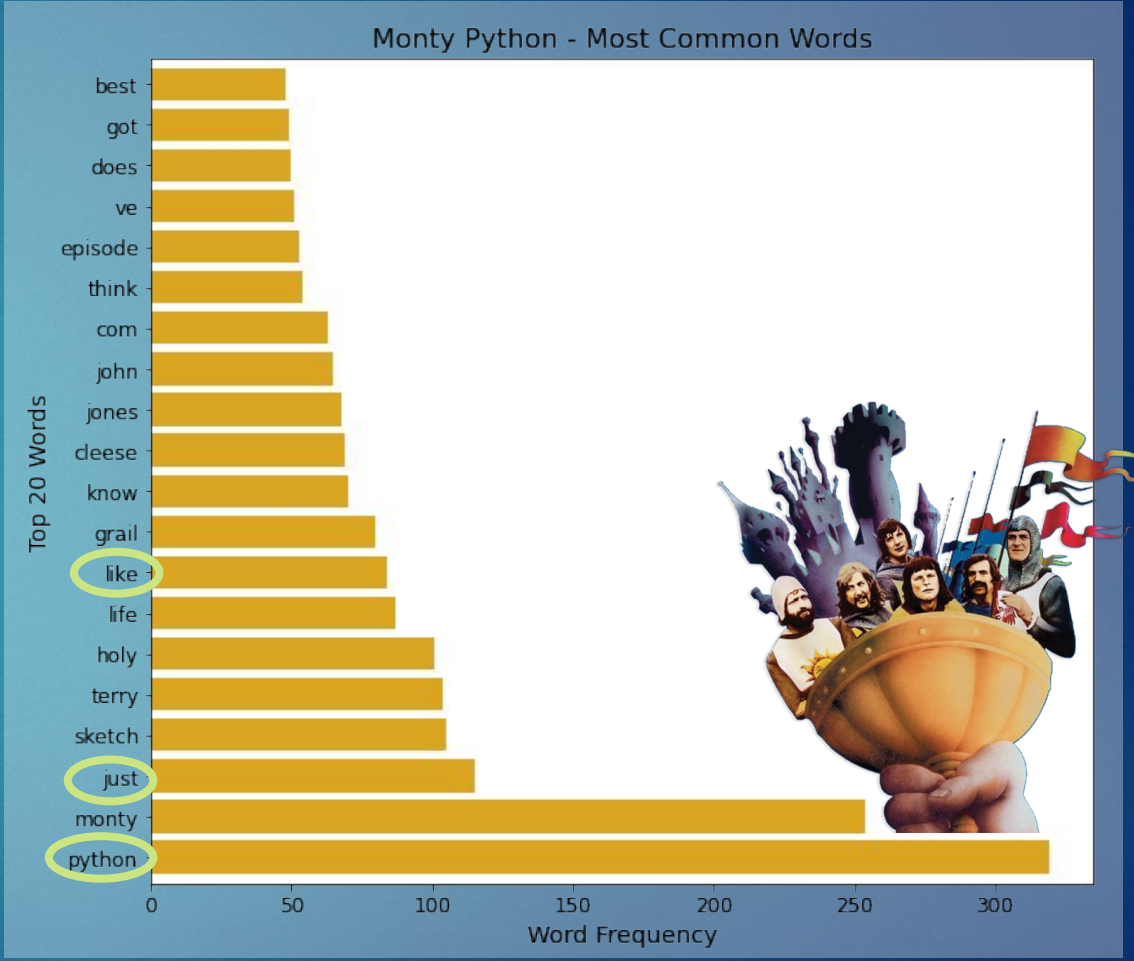
News about the programming language Python. A place to teach others Python and ask questions.

r/montypython

Everything to do with the Monty Python comedy troupe from dead parrots and silly walks to The Flying Circus and The Holy Grail

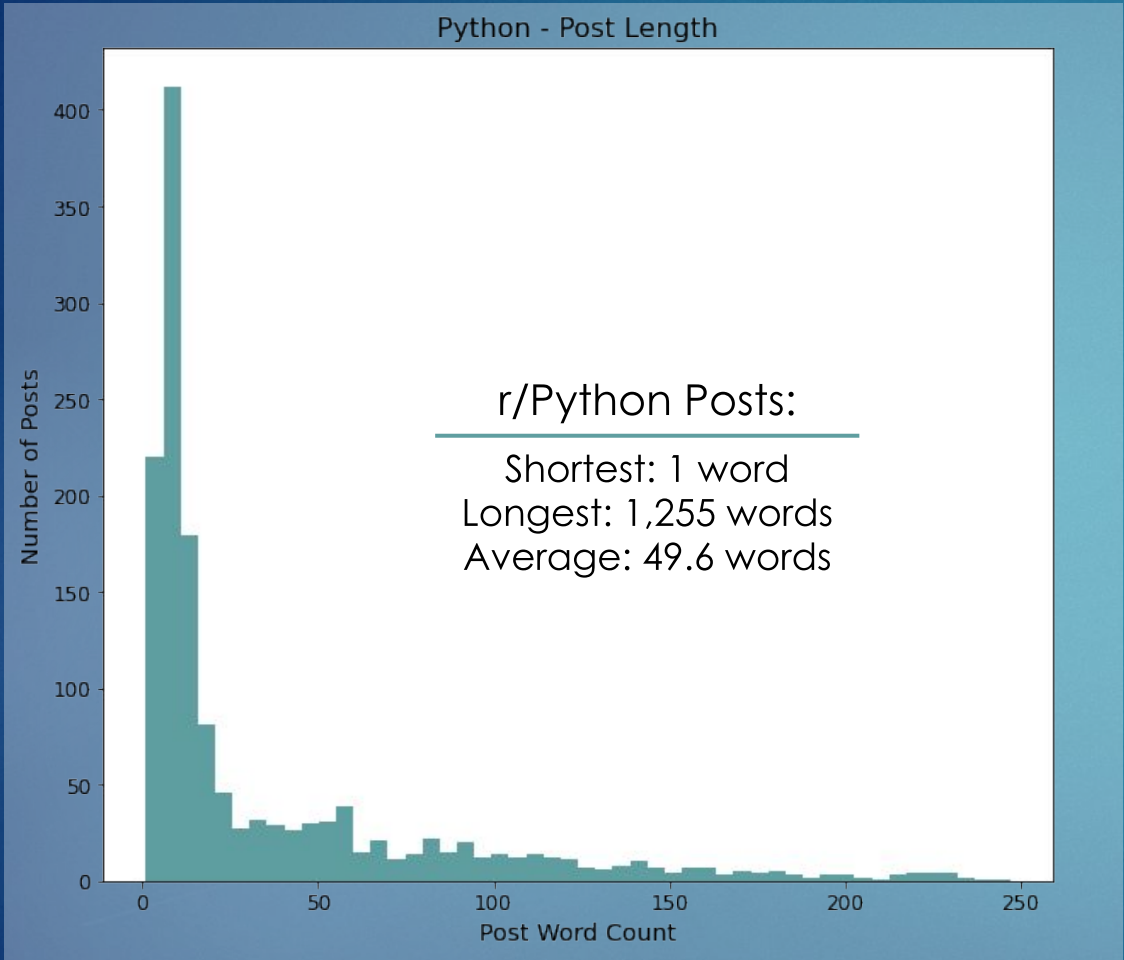
NLP at Work: Word Frequency

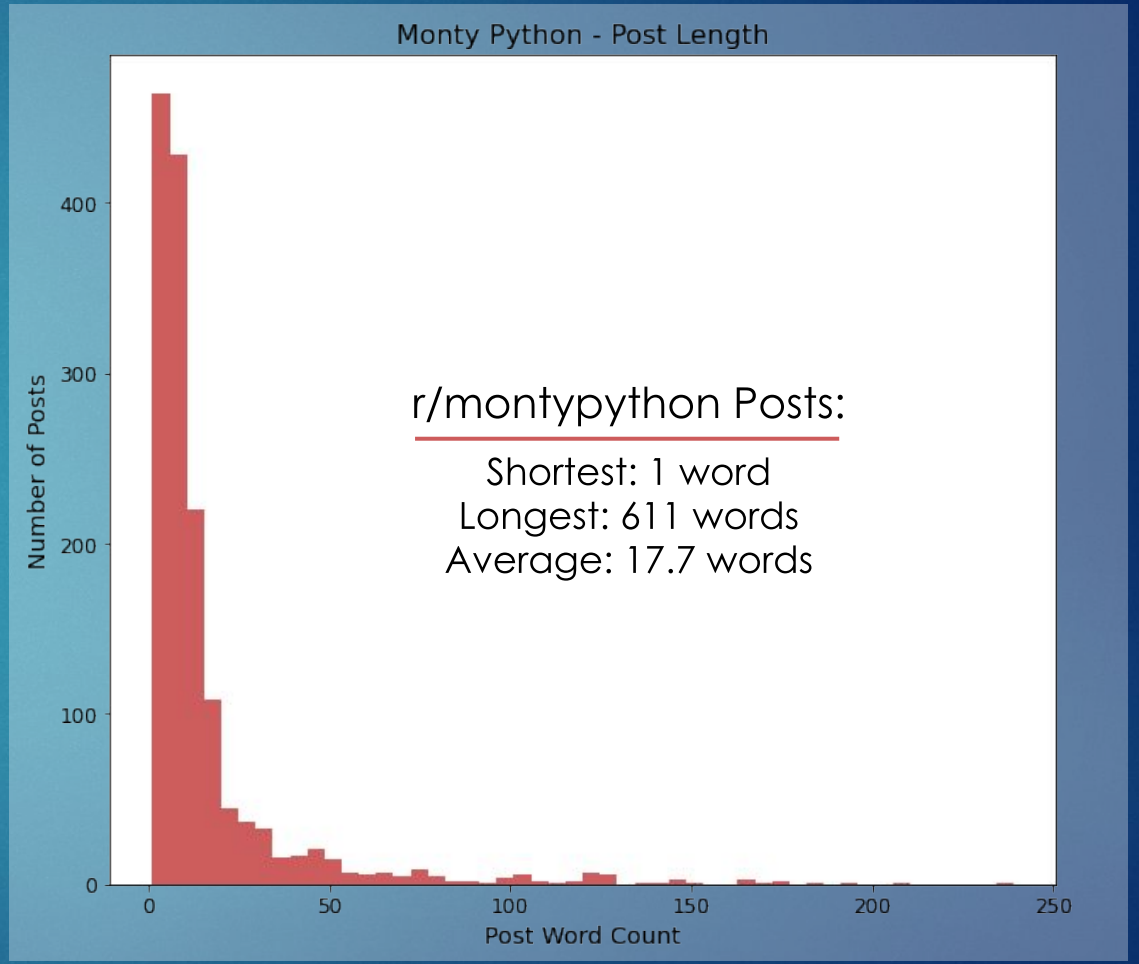




Results: 17 of the top 20 words in the subreddits are unique, so a bag of words approach could have strong predictive power.

NLP at Work: Post Length





Results: Post length varies a fair amount between r/Python and /montypython, so this could be an interesting numeric value to add to the text features in the models.

Modeling at Work:

Step 1: Choose text corpus for modeling

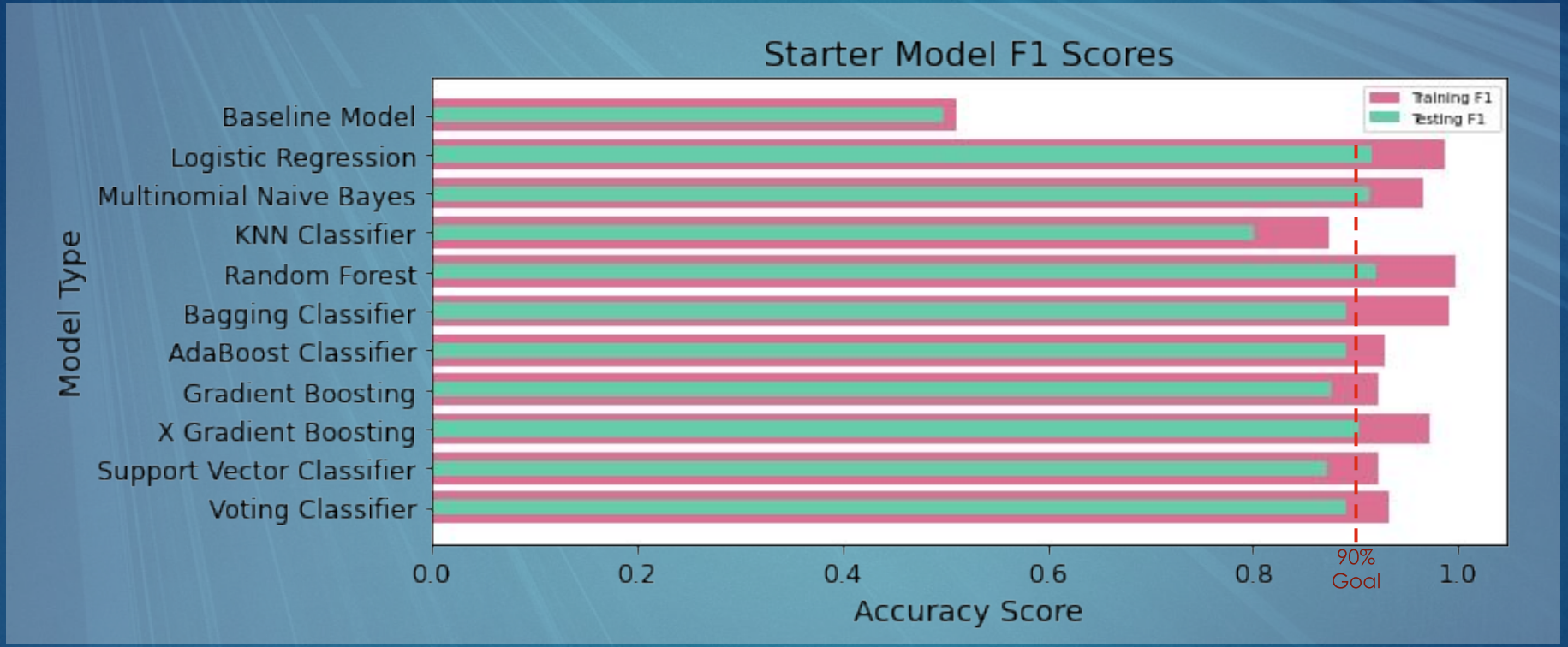
	Logistic Regression		Naïve Bayes		AdaBoost	
Corpus	Training	Testing	Training	Testing	Training	Testing
1 Original	98.7%	90.3%	96.9%	91.5%	92.8%	88.2%
2 Lemmatized	98.6%	90.7%	96.7%	91.8%	92.7%	87.7%
3 Stemmed	98.7%	91.6%	96.6%	91.3%	92.9%	89.1%

Process: Set each corpus as X variable in several base models (default params) and look at Accuracy Score to determine which corpus performs strongest.

Results: Stemmed corpus was the strongest performer in two out of three models (green indicates strongest performing corpus in each model/dataset).

Modeling at Work:

Step 2: First Round Models



Process: Fit the"winning" Stemmed corpus to a variety of "starter" classification models using default parameters, to get a baseline training and testing F1 scores.

Results: Four models achieved 90% or higher F1 on testing data, even before any tuning. We will move forward with Logistic Regression and Naïve Bayes to try to reduce some variance.

Modeling at Work :

Step 3: Tune Logistic Regression & Naïve Bayes models

Perform various transformations and tune hyperparameters for both models to fine tune the balance between Precision, Recall and Bias/Variance tradeoff, while maintaining high accuracy.

Logistic Regression Model

Features		Text Numeric
Techniques		GridSearch Pipeline FeatureUnion
Transformers		CountVectorizer TFIDFTransformer
Hyperparameters	Count Vectorizer	Stop Words Max Features Min Document Freq. Max Document Freq. N-Gram Ranges
	Tf-idf Transformer	Use IDF reweighting Normalization
	Logistic Regression	Solver Penalty

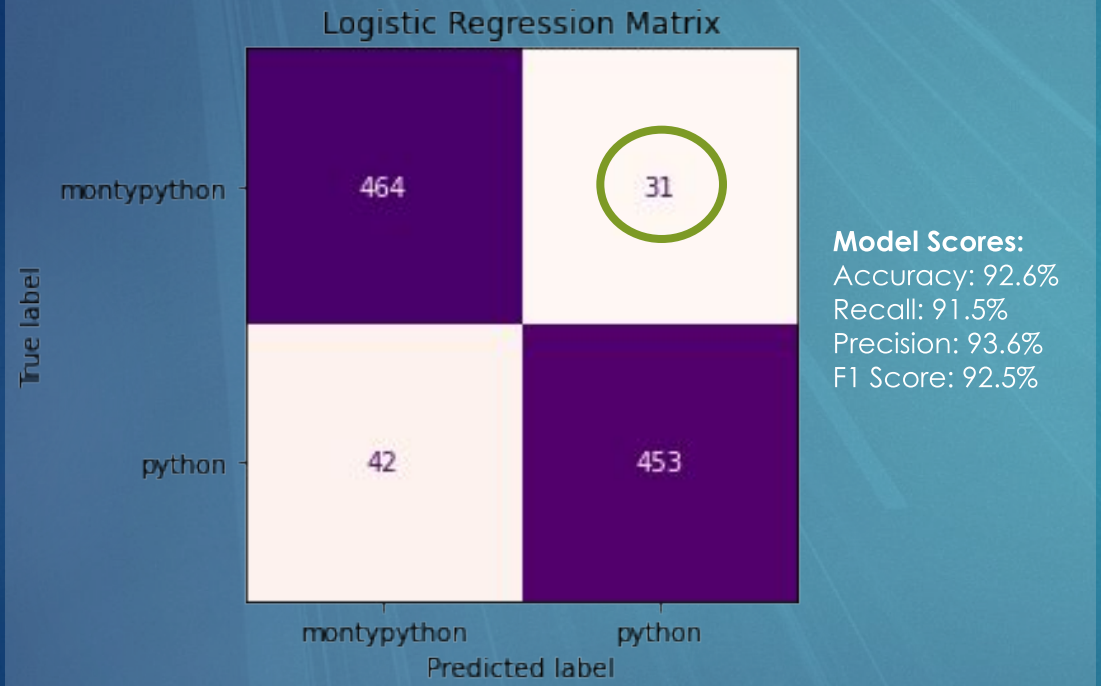
Multinomial Naïve Bayes

Features		Text
Techniques		GridSearch Pipeline
Transformers		CountVectorizer TFIDFTransformer
Hyperparameters	Count Vectorizer	Stop Words Max Features Min Document Freq. Max Document Freq. N-Gram Ranges
	Tf-idf Transformer	Use IDF reweighting Normalization
	Naïve Bayes	Alpha

Modeling Results:

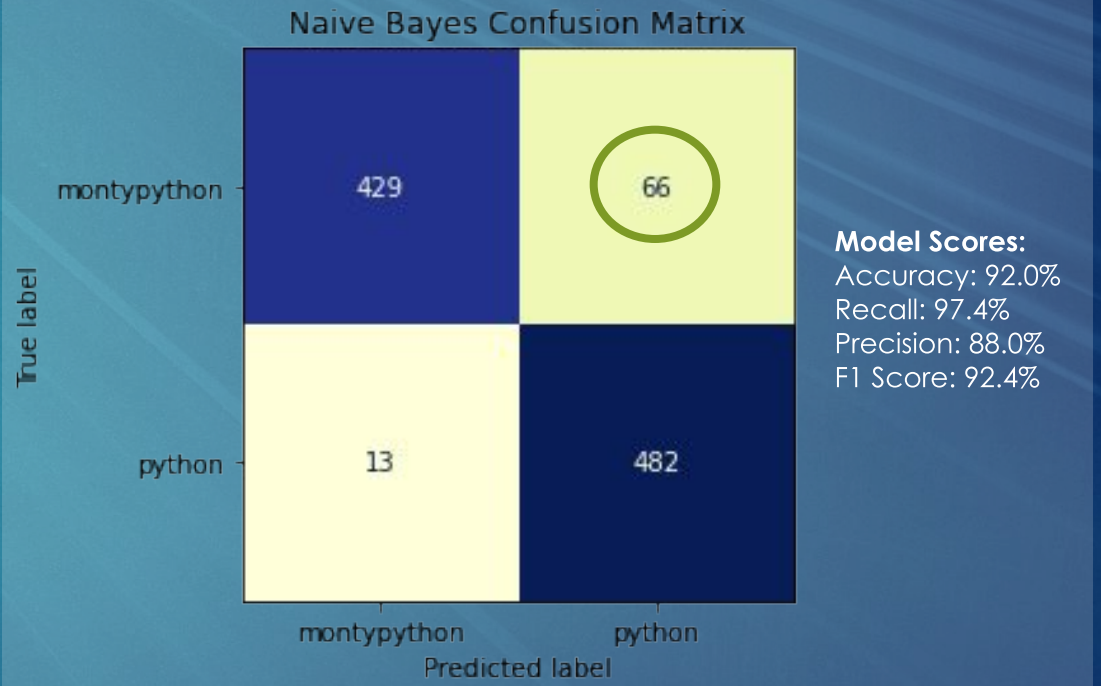
Confusion Matrices and Key Performance Metrics:

Logistic Regression Model



Model Scores:
Accuracy: 92.6%
Recall: 91.5%
Precision: 93.6%
F1 Score: 92.5%

Multinomial Naïve Bayes



Model Scores:
Accuracy: 92.0%
Recall: 97.4%
Precision: 88.0%
F1 Score: 92.4%

Results: Each model is has identified posts that were made in the montypython subreddit which, based on the language in the post, the model believes *should* be in python. These "false positives" will become the target audience for one-to-one marketing messages.

Conclusions

Both Logistic Regression and Naïve Bayes models are performing strongly based on key metrics:

- Accuracy on testing data > 90%
- F1 Scores for testing data ~ 92.5%

Where the models differ is in Precision and Recall:

Model	Score	Result
Logistic Regression	↑ Precision	↓ False Positives
	↓ Recall	↑ False Negatives
Naïve Bayes	↓ Precision	↑ False Positives
	↑ Recall	↓ False Negatives

Because False Positives will become the target audience for one-to-one targeting, marketing objectives will drive final model selection:

- Naïve Bayes = Increase reach, lower ROAS
- Logistic Regression = Reduced reach, higher ROAS

Next Steps

Recommendations:

1. Perform A/B Testing on False Positives from each model to determine effectiveness and ROAS from each before making final decision.
2. Execute A/B testing of marketing tactics to determine ideal one-to-one marketing approach and measure ongoing return on ad spend (ROAS), cost to acquire (CTA) and other relevant performance metrics.

Next Steps:

1. Expand scope of model to include additional training subreddits (those relevant to data science, machine learning, python) and additional "imposter" subreddits (where users may mistakenly post data science questions).
2. Rerun data science process to ensure models continue to display strong performance and meet objectives for marketing.
3. LabelBox engineering team to productionalize final model(s).