REDDIT MACHINE LEARNING POST CLASSIFICATION

Predicting Subreddits From Posts

WELCOME!

Christopher Landschoot

I am a Data Scientist leading the
team hired by Reddit for this project.

BACKGROUND

- The internet has been flooded with data and it is becoming increasingly hard to determine the validity of content.
- Additionally, this data provides opportunities to glean information about whoever creates it.
- As Reddit has continued to increase in size, the amount of data they collect has increased as well.
- Reddit is interested in finding new ways to analyze the data on their site in order to better inform their moderation policies, target marketing, and analyze general trends of their users.

PROBLEM STATEMENT

- Reddit is seeking to create a machine learning model that can accurately classify which subreddit a post originates from solely based on its text.
- To achieve this, Reddit has hired my team of data scientists to create a proof-of-concept model that will be trained on 2 subreddits.
- Success will be measured by achieving:
  - Accuracy of classification > 95%
  - Precision and Recall of classification > 90%
- If successful, financial resources will be allocated to developing this technology to be generalized to the entire site.

THE DATA SCIENCE PROCESS

01

Data Collection

Collecting posts from two subreddits.

02

Exploratory Data Analysis

Analyzing trends in the data.

03

Preprocessing

Natural Language Processing (NLP) of the text data.

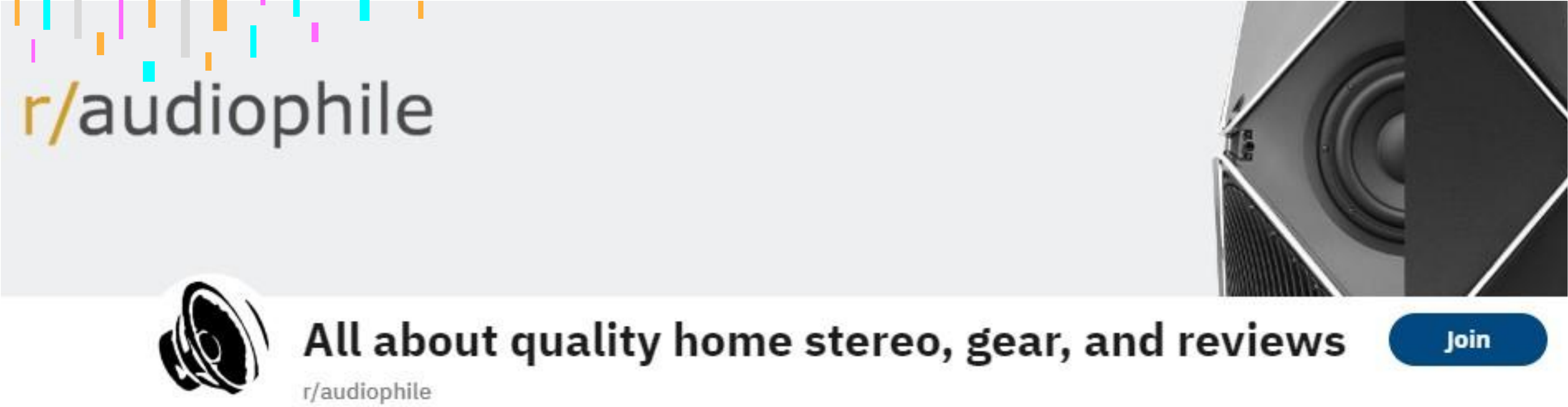
04

Modeling

Optimizing multiple models and ensembling.

01

Data Collection



All about quality home stereo, gear, and reviews

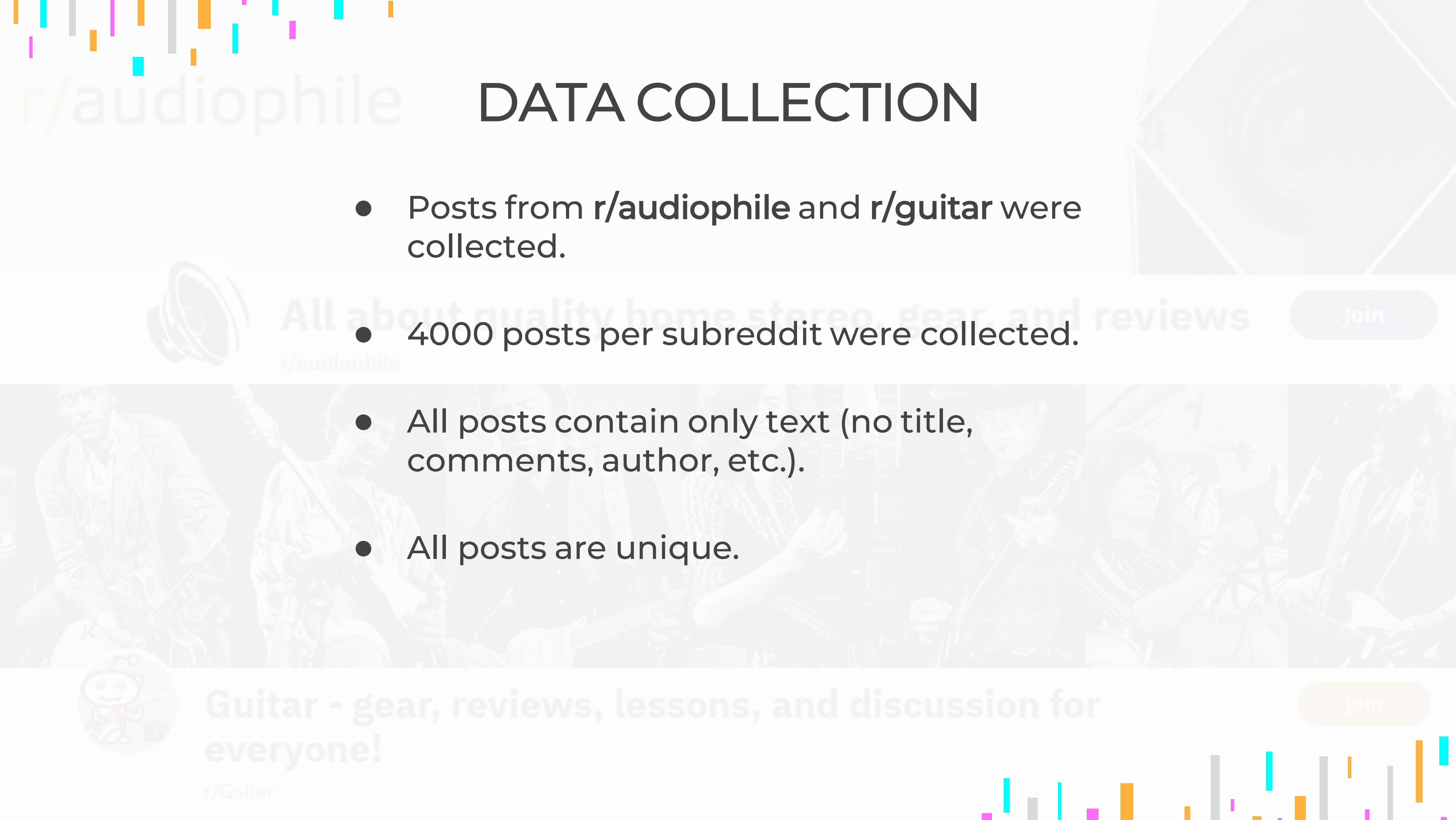
r/audiophile

Guitar - gear, reviews, lessons, and discussion for everyone!

r/Guitar

DATA COLLECTION

- Posts from **r/audiophile** and **r/guitar** were collected.
- 4000 posts per subreddit were collected.
- All posts contain only text (no title, comments, author, etc.).
- All posts are unique.



r/audiophile

All about quality home stereo, gear, and reviews

Join

r/audiophile

Guitar - gear, reviews, lessons, and discussion for everyone!

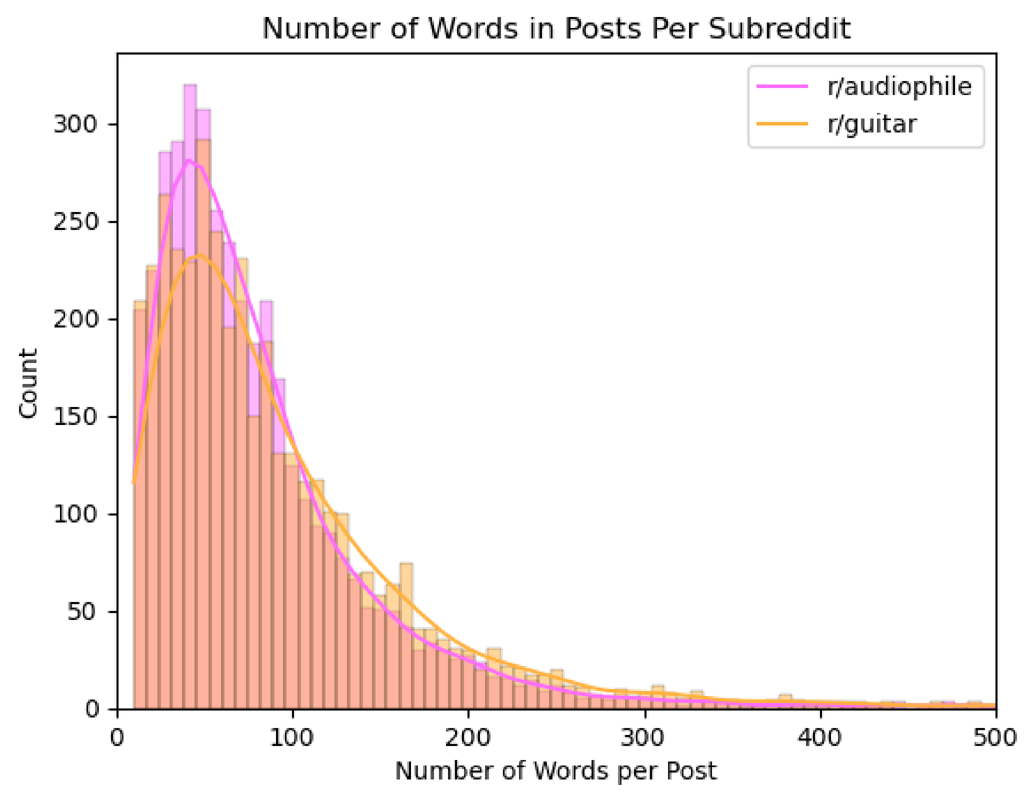
Join

r/Guitar

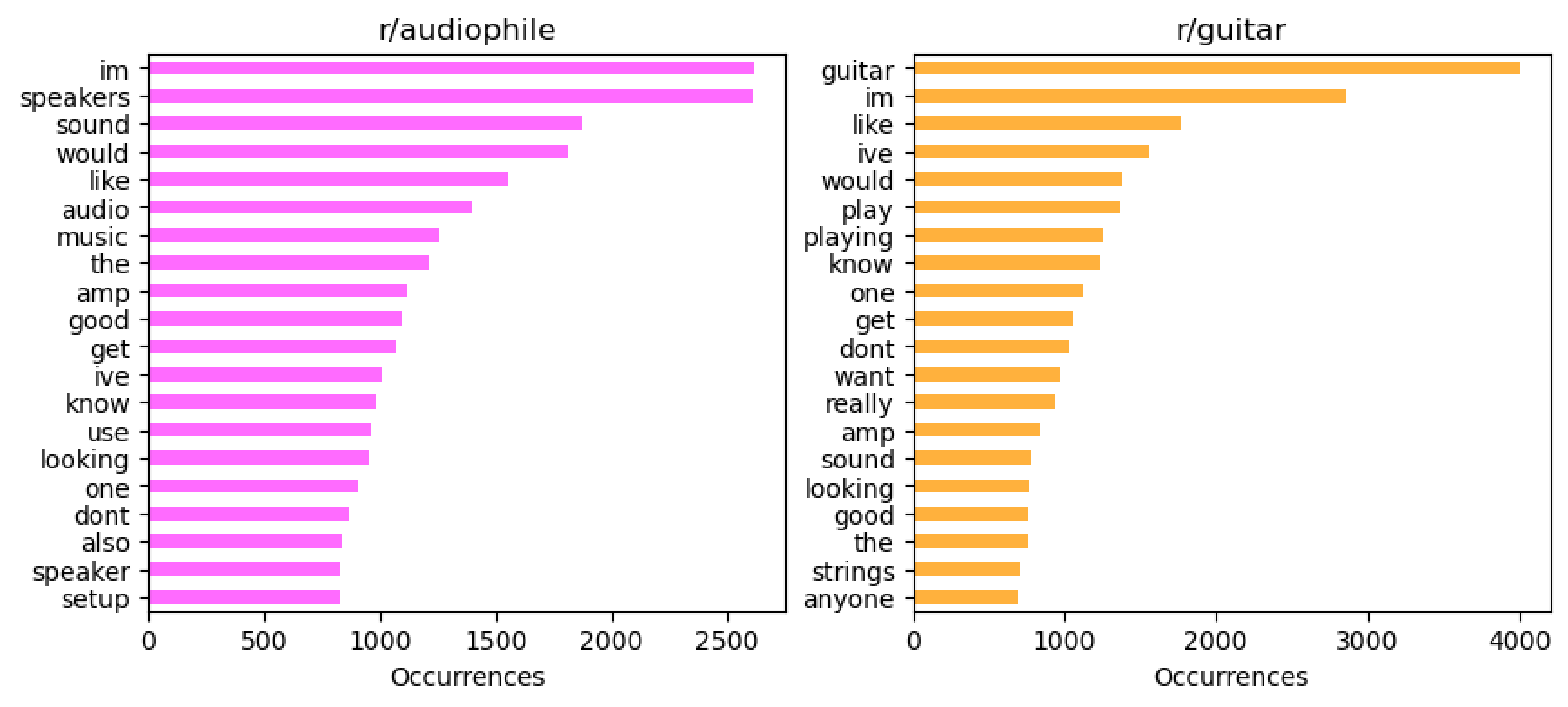
02

Exploratory Data Analysis

POST DISTRIBUTION



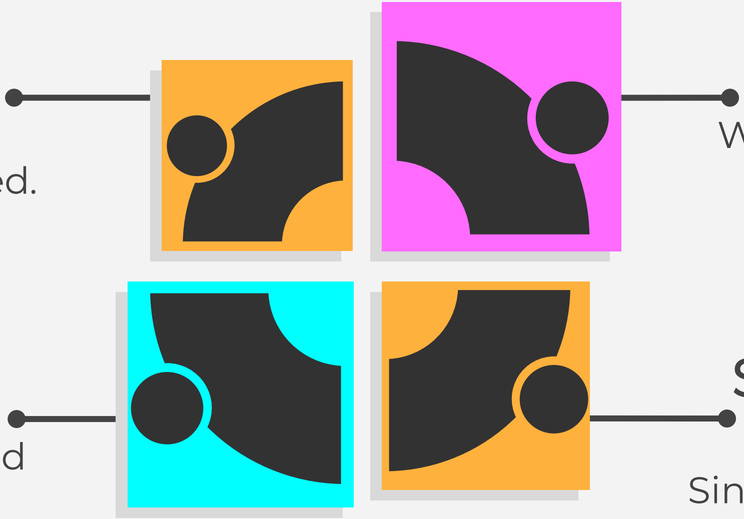
MOST COMMON WORDS



03

Natural Language Processing

NATURAL LANGUAGE PROCESSING



Stopwords Removed

Common English words were removed.

Stemming Words

Words were reduced to their base form.

Lemmatizing Words

Words were reduced to their base form.

Commonly Shared Words Removed.

Single words, bigrams, and trigrams.

04

Modeling

MODELS

- Multinomial Naïve Bayes
- Logistic Regression
- Random Forest
- Extra Trees
- Gradient Boosting
- XGBoost
- Support Vector Machine

MODELS

Multinomial Naïve Bayes

96% Accurate, Best Recall

96% Accurate

Logistic Regression

Random Forest

Extra Trees

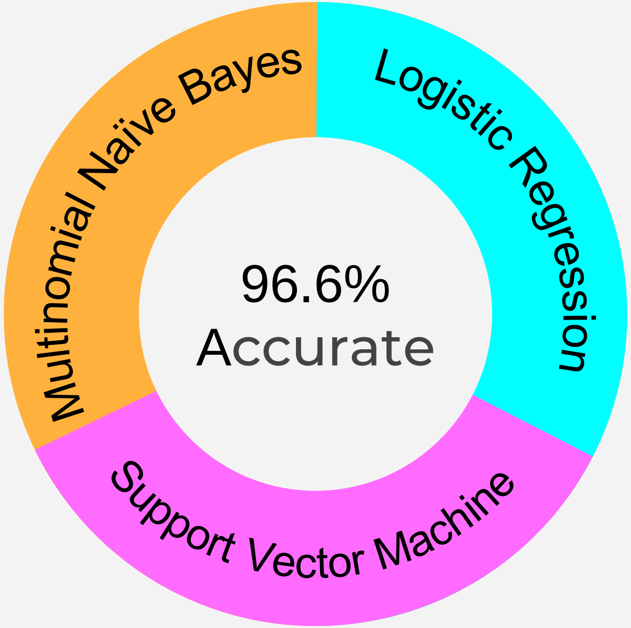
Gradient Boosting

XGBoost

Support Vector Machine

96.2% Accurate, Best Precision

Ensembled Model



CONCLUSIONS

- Text data was processed using natural language processing techniques.
  - Stop words were removed.
  - Text was stemmed (reduced to base form).
  - Shared most common words were removed.
- Seven separate models were tested. The best 3 (Multinomial Naïve Bayes, Logistic Regression, SVM) were combined into a final ensembled model.
- Ensembled Model
  - Accuracy: 96.6% (>95% Success Metric)
  - Precision: 96.5% (>90% Success Metric)
  - Recall: 96.7% (>90% Success Metric)

RECOMMENDATIONS

- This prototype model exceeded the success metrics defined by Reddit.
- Reddit should allocate financial resources to further develop this technology for the entire site.
- Next Steps:
  - Introduce sentiment analysis.
  - Introduce additional data to the model, including title of post, comments, author, upvotes, etc.
  - Test other NLP methods (other than bag-of-words).

THANKS!

Any Questions?