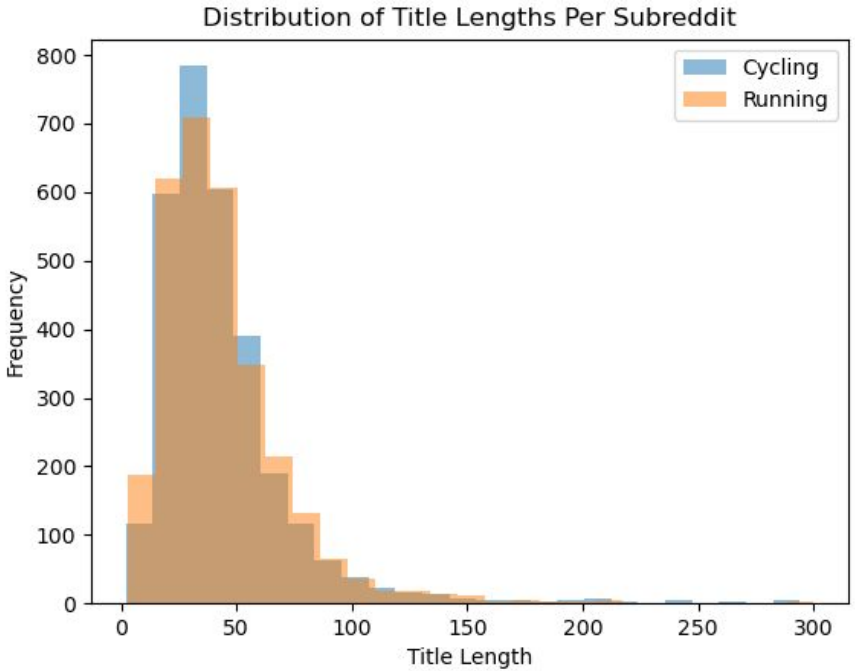
# Classification Modeling for Reddit: Cycling v. Running

By: Jason Lu

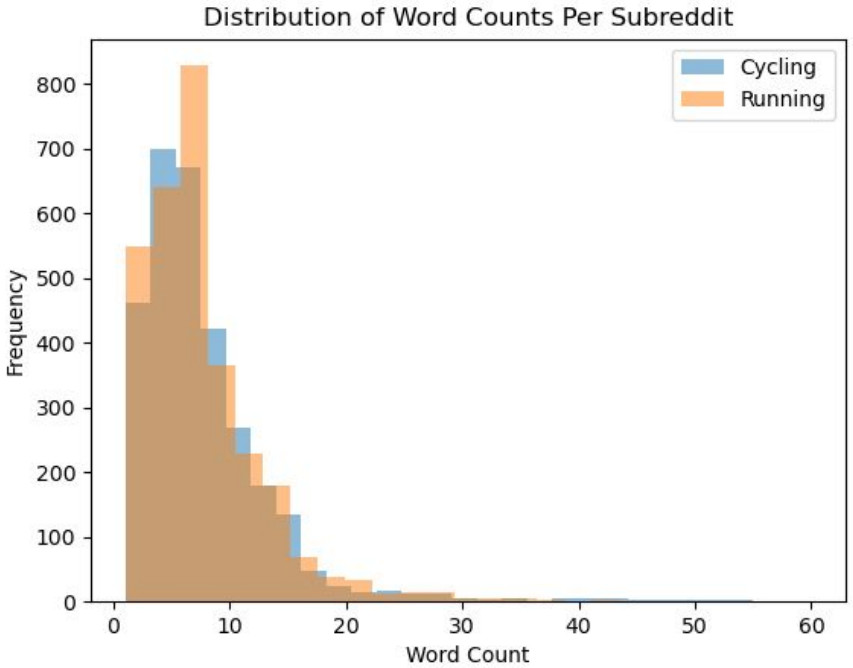
# Introduction

- As a data scientist for a company specialized in fitness tracking
- In order to better meet customer needs, build a classification model
  - distinguish & categorize top two sports: cycling and running

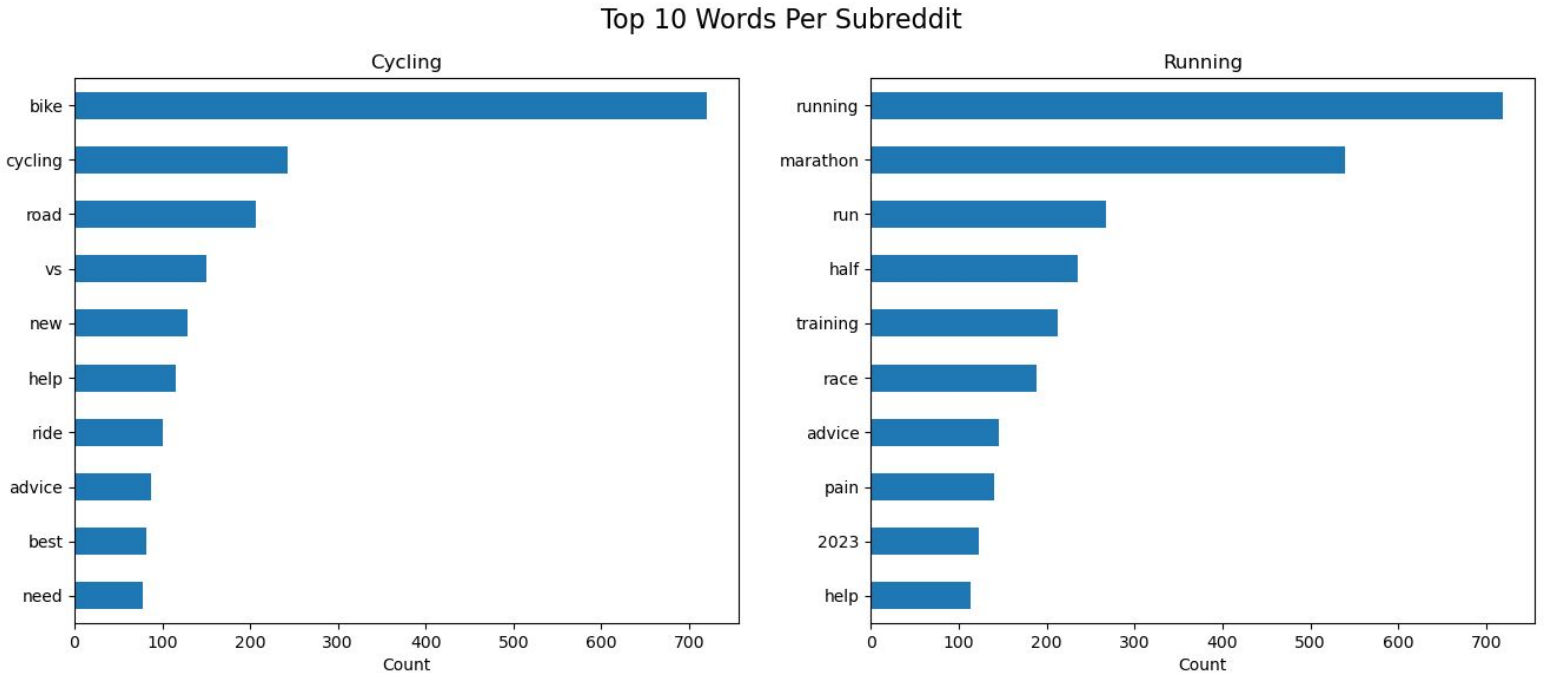
# Title Lengths



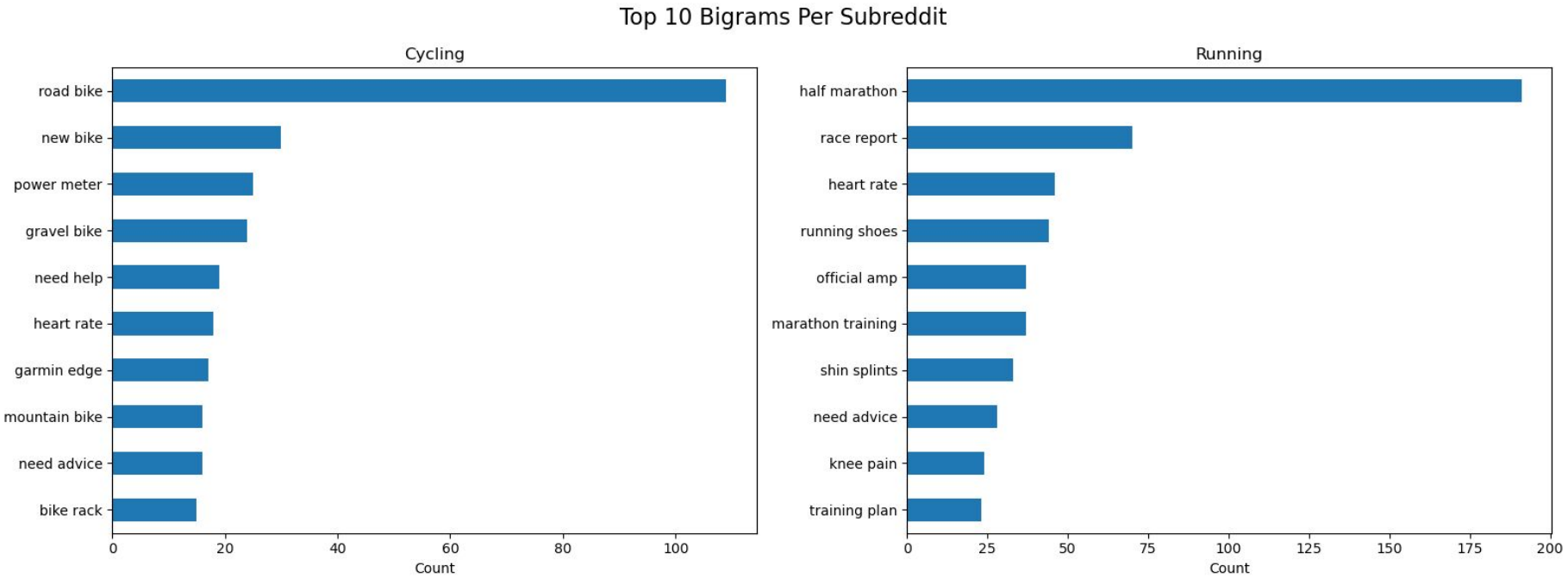
# Word Counts


Distribution of Word Counts Per Subreddit
Cycling
Running
Frequency
800
700
600
500
400
300
200
100
0
0
10
20
30
40
50
60
Word Count

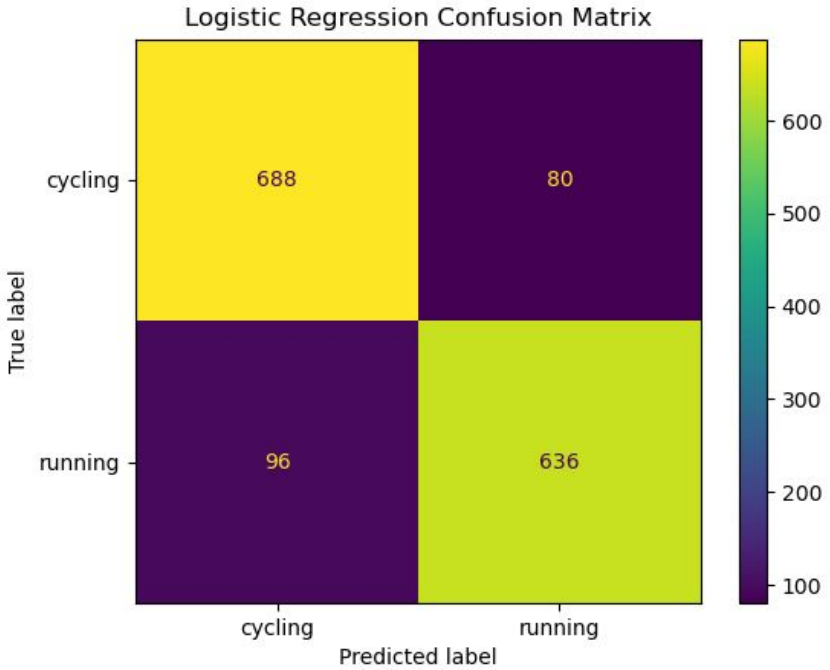
# Top 10 Words


Top 10 Words Per Subreddit
Cycling
bike
cycling
road
vs
new
help
ride
advice
best
need
0
100
200
300
400
500
600
700
Count
Running
running
marathon
run
half
training
race
advice
pain
2023
help
0
100
200
300
400
500
600
700
Count

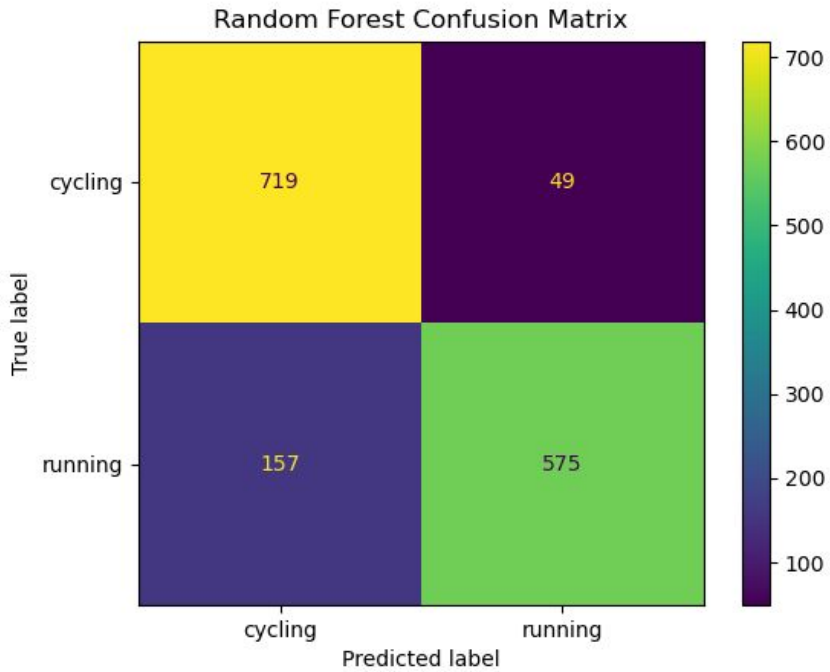
# Top 10 Bigrams (cons. words)



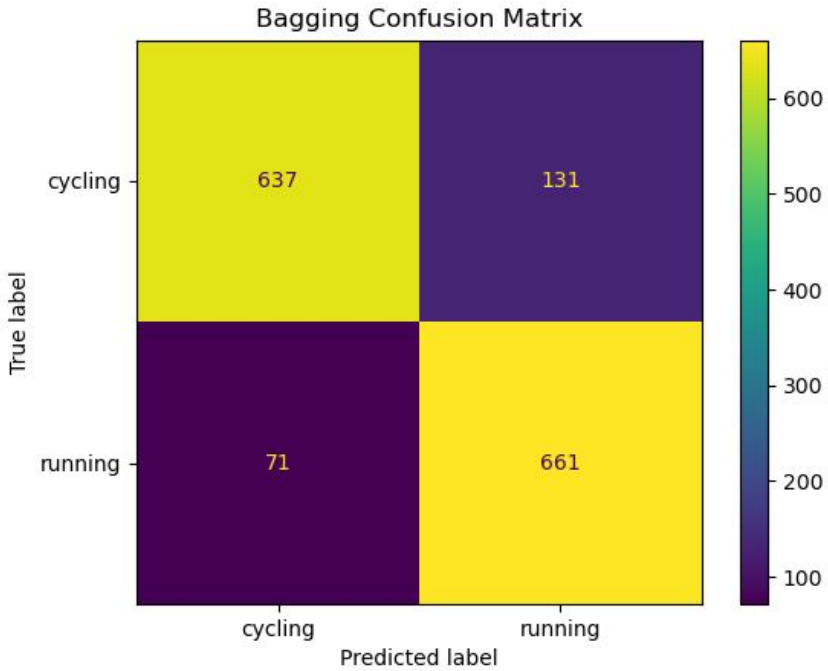
# Logistic Regression



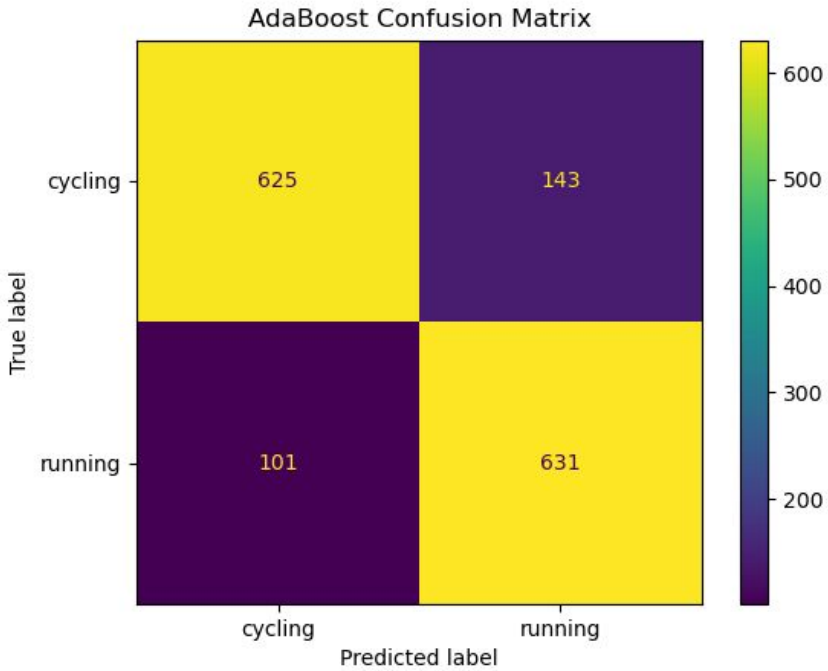
# Random Forest



# Bagging Model



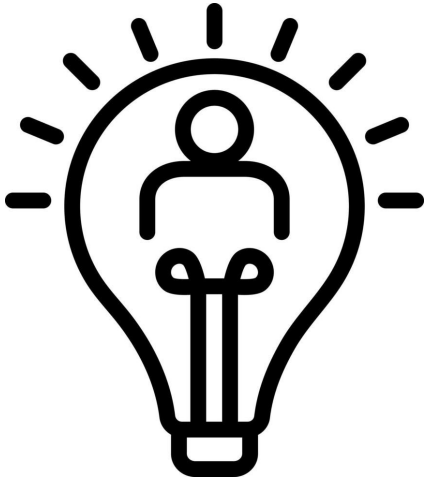
# AdaBoost Model



# Model Performance Summary

| **Model Name** | Recall | Precision | F1 | Accuracy |
|---|---|---|---|---|
| Logistic Regression | 0.8958 | 0.8775 | 0.8865 | 0.8827 |
| Random Forest | 0.9362 | 0.8207 | 0.8747 | 0.8627 |
| Bagging | 0.8294 | 0.8997 | 0.8631 | 0.8653 |
| AdaBoost | 0.8294 | 0.8997 | 0.8631 | 0.8653 |

# Conclusions



- Highest performing model:
  - Logistic Regression with CountVect.
  - 88.2% accuracy
- From our exploratory analysis
  - Cycling = focus on equipment
  - Running = focus on individual

# Recommendations

- Add specifically related stop words
- Utilize bigger, more varied dataset
- Additional features
- Future models incorporating not only title
  - Text, Comments

# Thanks for listening!