

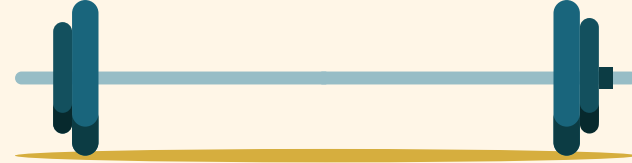
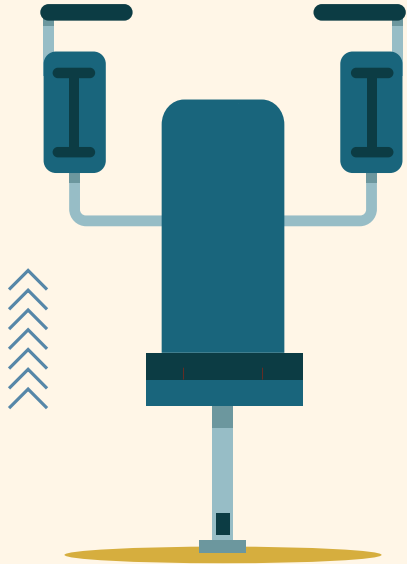
# Reddit Classification: Powerlifting vs. Bodybuilding

by Kevin Trinh



# Problem Statement

A new local gym called Global Fitness wants to host powerlifting and bodybuilding events to attract more long-term customers and increase gym membership sales. In order to properly advertise to powerlifters and bodybuilders for their respective competitions, the marketing team at Global Fitness is looking at subreddits for powerlifting and bodybuilding to understand the nuances of each topic. However, because there are so many posts, they are looking for a data scientist to figure out the best ways to utilize these subreddits for marketing purposes.





# Table of Contents



**01**

**Data Collection  
and Cleaning**

**02**

**Exploratory Data  
Analysis**

**03**

**Models**

**04**

**Results**

**05**

**Conclusions and  
Recommendations**

**06**

**Limitations and  
Future Studies**



# Data Collection and Cleaning

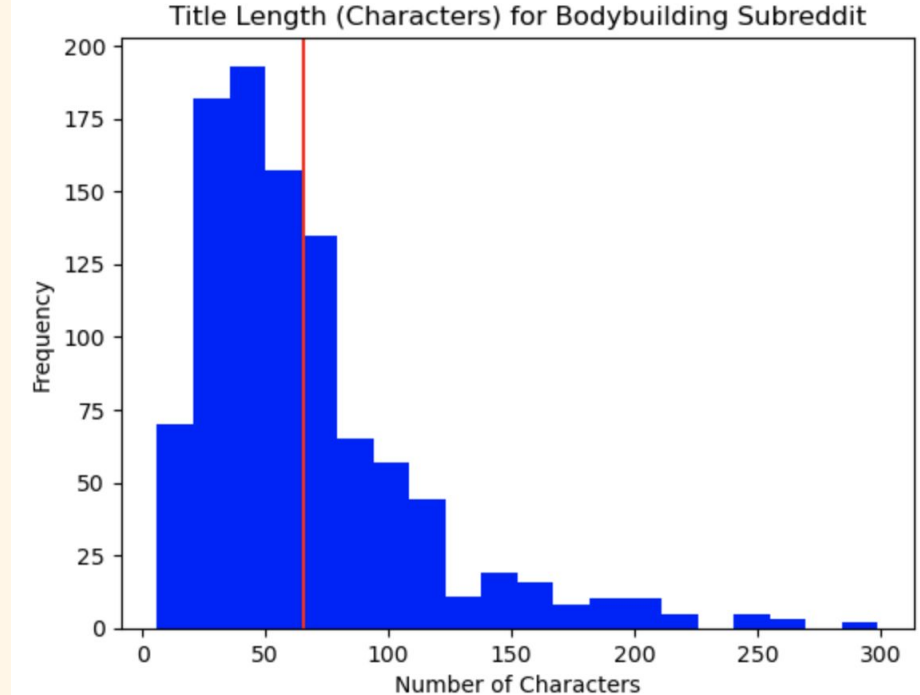
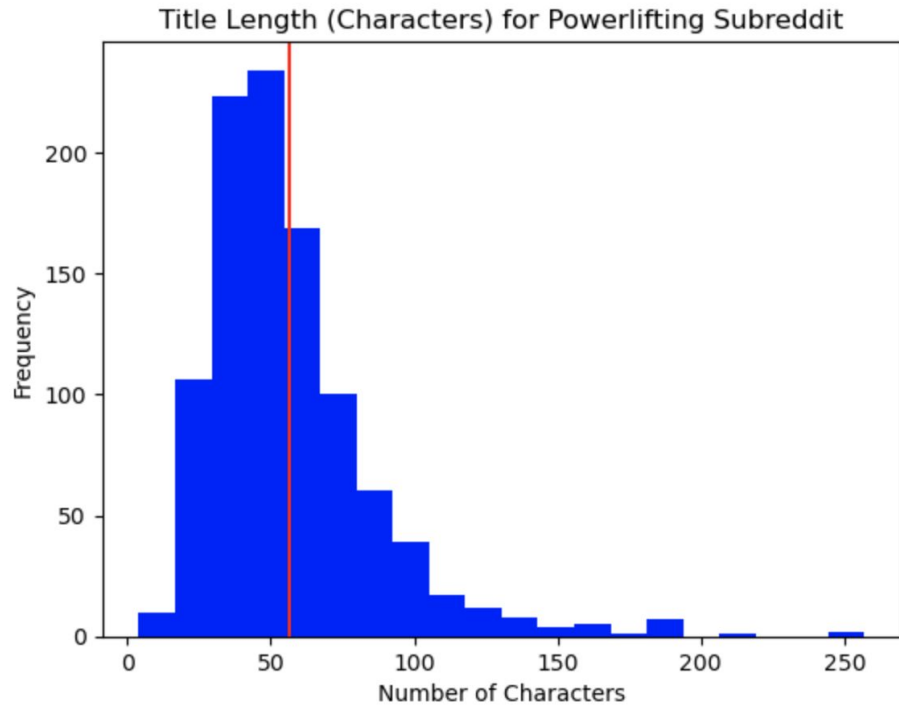


**r/powerlifting + r/bodybuilding**

- Collected data from the top 1000 posts of all time from each subreddit via the Reddit API
- Created dataframes for each and converted them to CSV files.
- Cleaned datasets, engineered features, and combined all data into one dataset.



# Exploratory Data Analysis



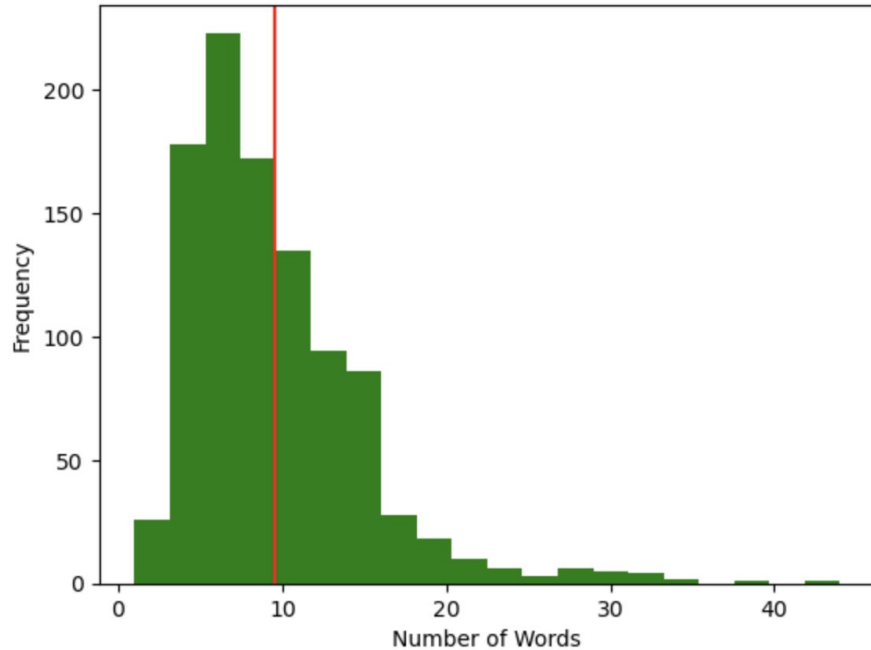
Average of 57 characters for Powerlifting and 66 characters for Bodybuilding in post titles



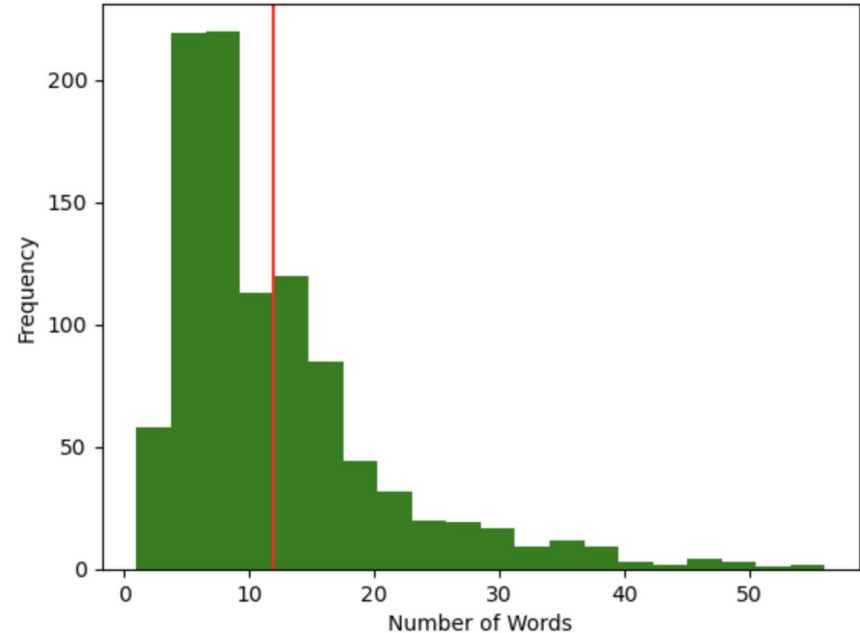
# Exploratory Data Analysis



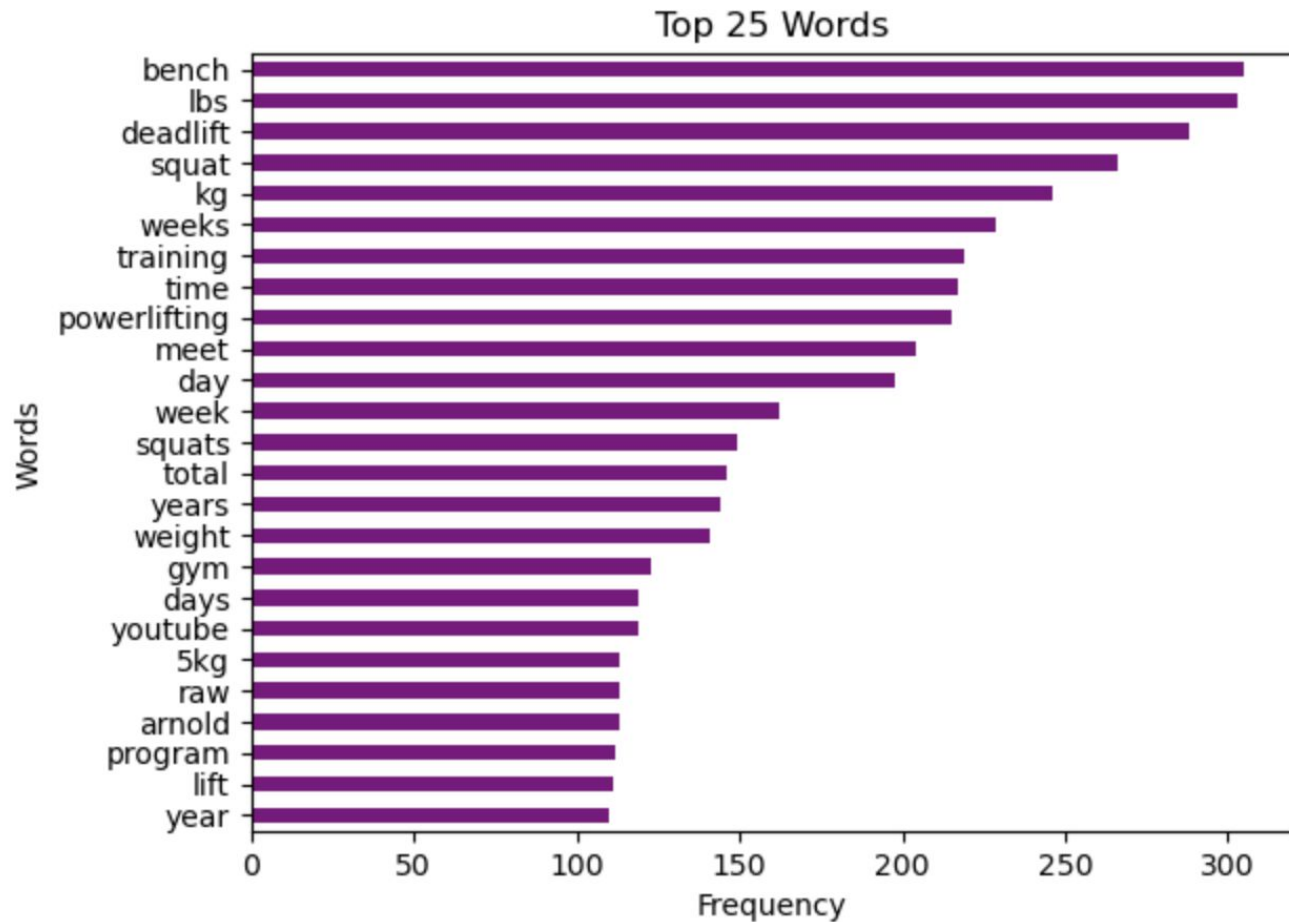
Title Length (Words) for Powerlifting Subreddit



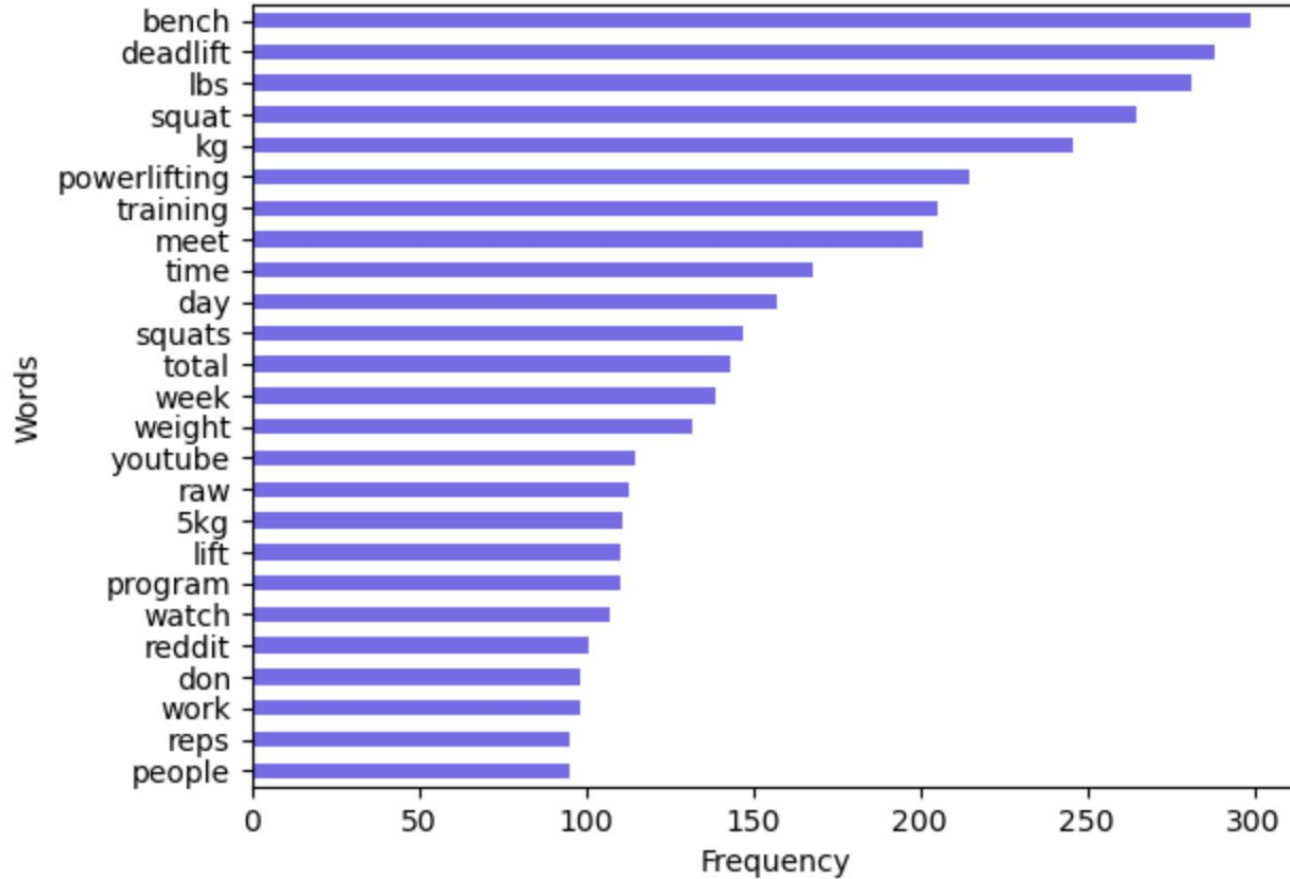
Title Length (Words) for Bodybuilding Subreddit



Average of 10 words for Powerlifting and 12 words for Bodybuilding in post titles

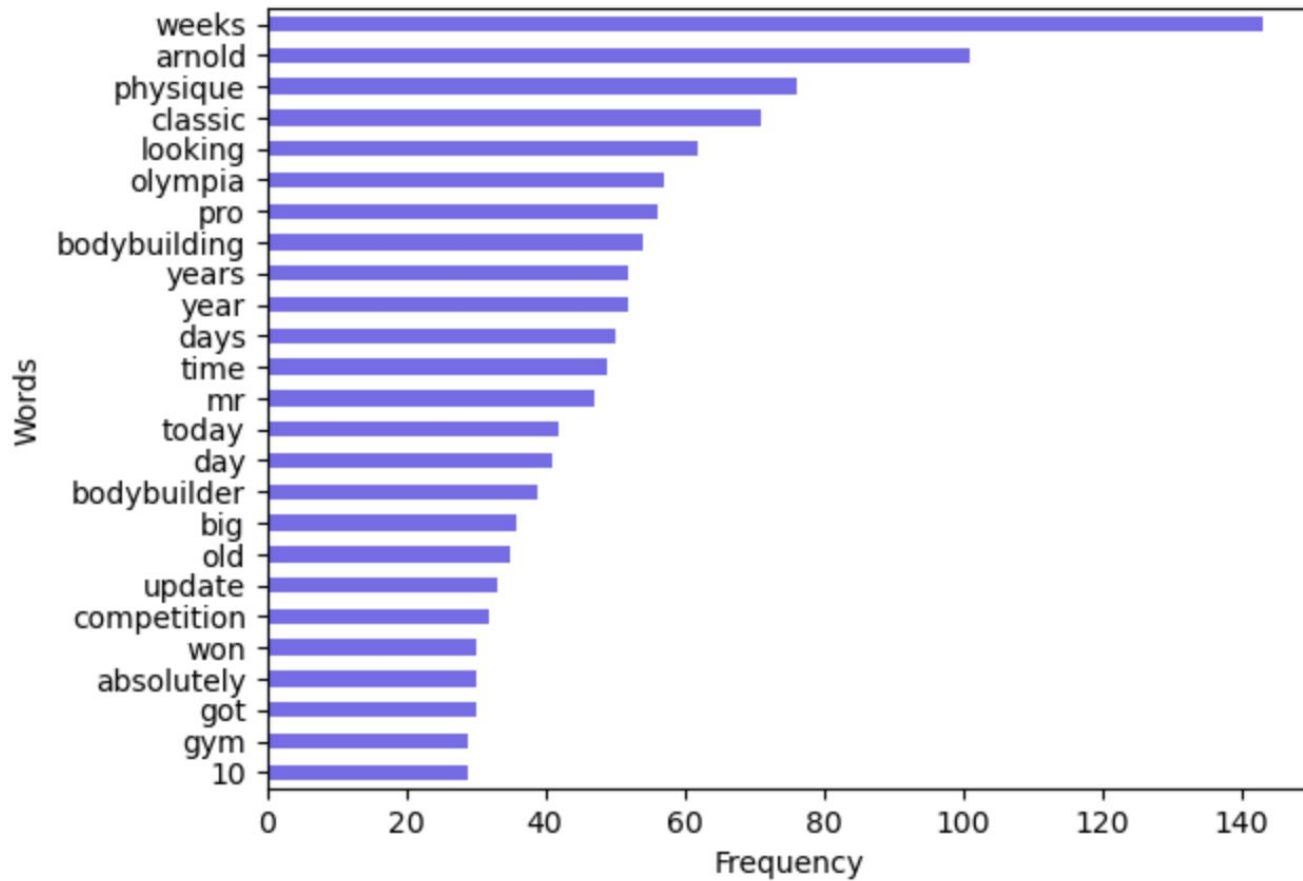


Top 25 Words (Powerlifting)







Top 25 Words (Bodybuilding)



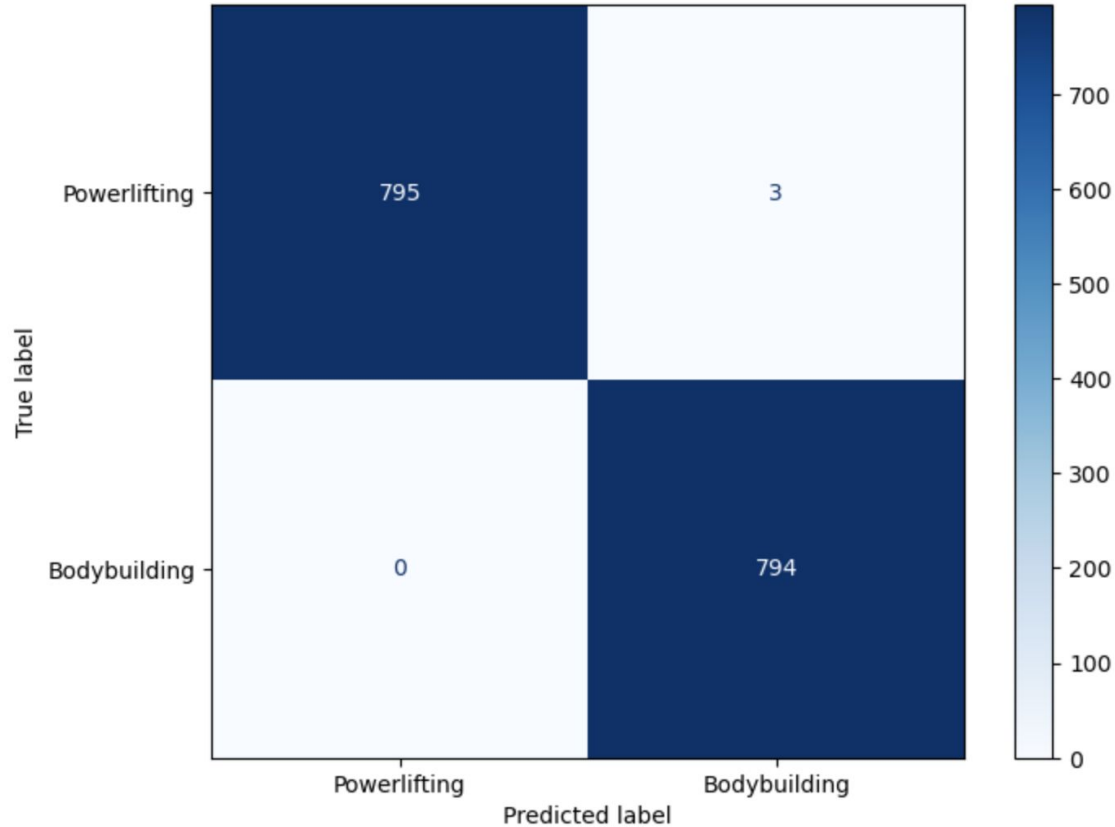


# Models

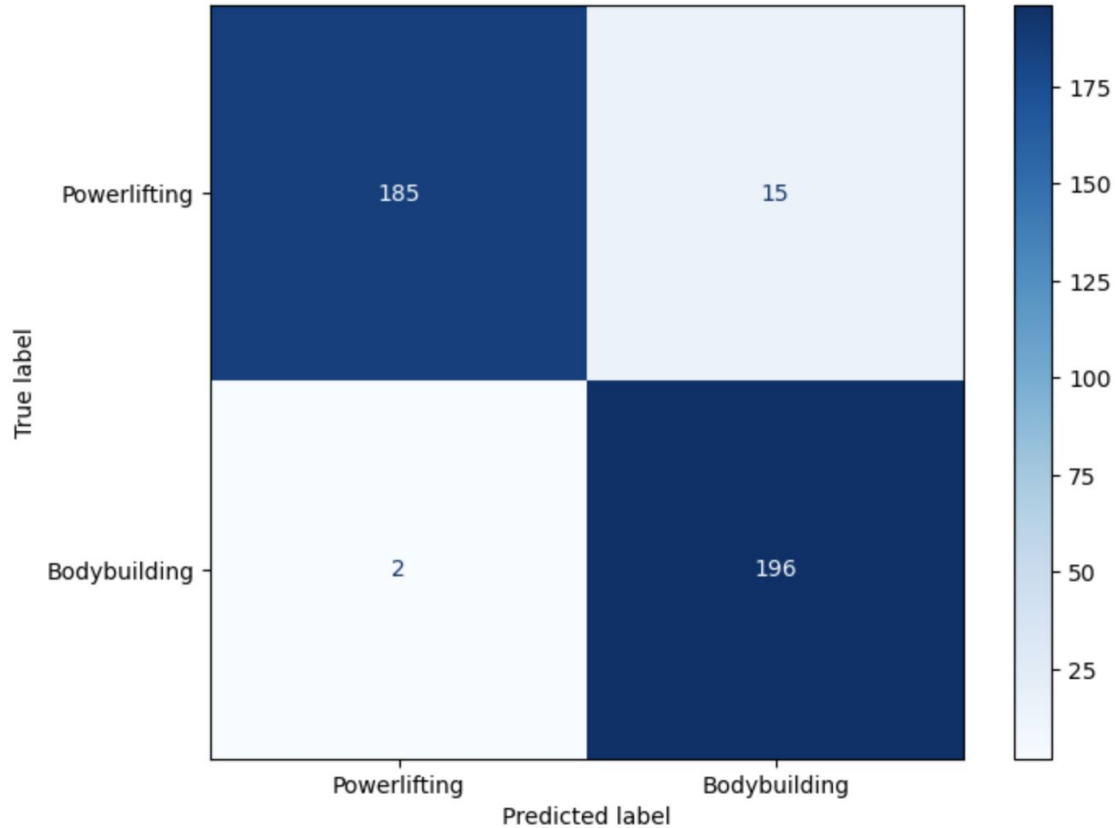


- **Logistic Regression**
  - **K-Nearest Neighbors**
  - **Random Forest**
  - **Support Vector Machines**
  - **Naive Bayes**
- 
- 

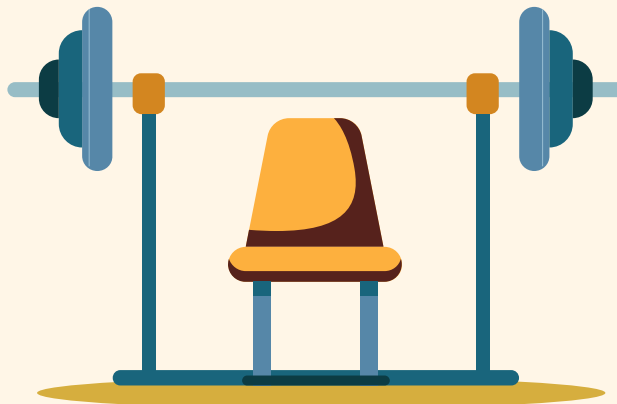
# Results (Training Data)



# Results (Testing Data)



Train Score: 99.8%  
Test Score: 95.7%





# Conclusions and Recommendations

- Words that held the most weight for Powerlifting:
  - “deadlift”
  - “squat”
  - “pulls”
  - “squats”
  - “bench”
- Words that held the most weight for Bodybuilding:
  - “weeks”
  - “arnold”
  - “olympia”
  - “physique”
  - “looking”
- Recommend Global Fitness to use those words and their associations in any form of advertising media to attract powerlifters and bodybuilders to their respective events.



# Limitations and Future Studies

- Most models showed a little bit of overfitting as the training data tended to usually score higher than the testing data.
- Future studies should involve further modification of the training set and continued regularization of the models for less overfit results.
- Most of the baseline models scored fairly high in terms of accuracy without tuning hyperparameters.
- Looking at other sources of data would be valuable, especially when targeting new customers.



# Sources

- Reddit API PRAW Documentation:  
<https://praw.readthedocs.io/en/stable/index.html>
- Powerlifting subreddit:  
<https://www.reddit.com/r/powerlifting/top/?t=all>
- Bodybuilding subreddit:  
<https://www.reddit.com/r/bodybuilding/top/?t=all>





Thank  
You!

