Project 3: Advanced Runner Marketing Retargeting using NLP

Amanda Walsh

Introduction

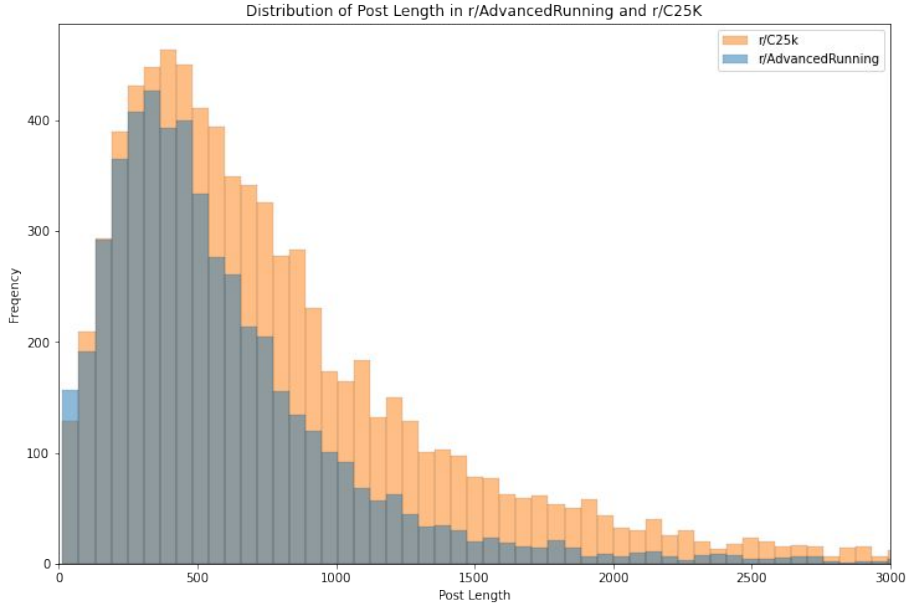
Objective: A running enthusiast mobile application company for all levels of runners is looking to launch a retargeting marketing campaign to convert Advanced Runners who have downloaded the application to being paid members. When the app is downloaded, users select their skill level. In an effort to understand their customer's needs, “Couch-to-5k (C25k)" and “Advanced Running” subreddits were analyzed and modeled using Natural Language Processing (NLP), Random Forest and Logistic Regression Classification Techniques

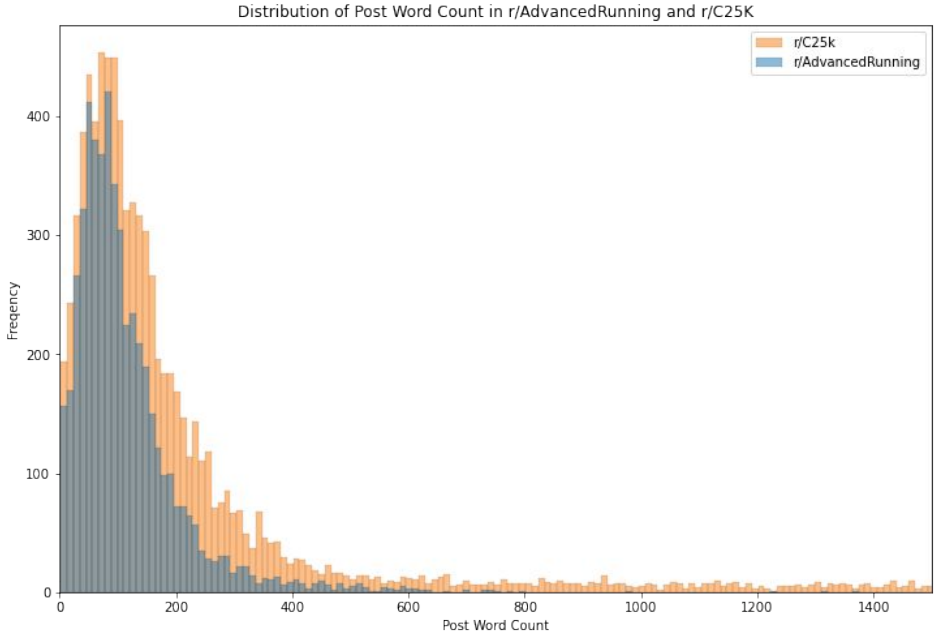
Goal: This project aims to identify the type of language being used between beginner and advanced runners in order to provide marketing re-targeting ad recommendations for the advanced runners customer segmentation.

Data Science Workflow

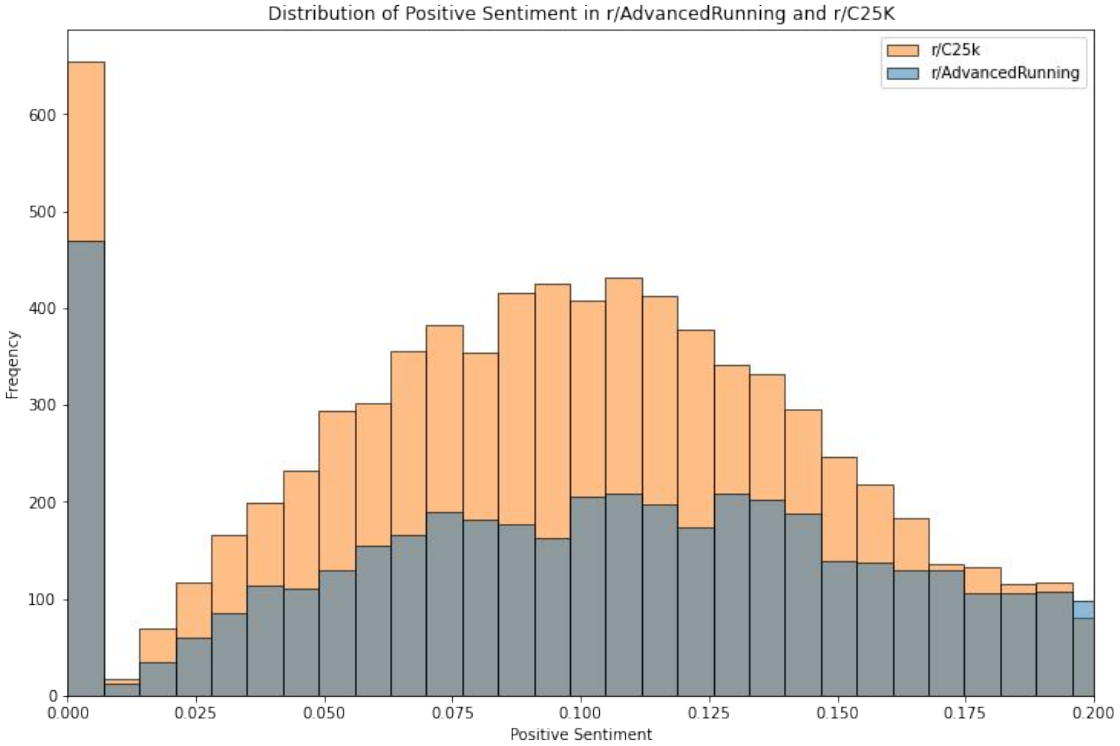
1. Web-scraping
2. Initial Raw Data Cleaning
3. Feature Engineering/Sentiment Analysis
4. Cleaning/Pre-Processing
5. Modeling
6. Conclusions and Recommendations

Feature Engineering





Sentiment Analysis



- **Positive**:
  - r/Advanced Mean: 0.10
  - r/C25k Mean: 0.12
- **Negative:**
  - r/Advanced Mean: 0.046
  - r/c25k Mean: 0.068
- **Neutral:**
  - r/Advanced Mean: 0.84
  - r/c25k Mean: 0.81

Pre-Processing

- **Cleaning (Pre Train-Test-Split)**
  - **NLTK**
  - Regex to remove special characters
  - Post was tokenized and lower-cased
  - Part-of-speech tags were added to tokens
  - Tokens were then lemmatized
    - Precision vs. Recall
- **Pre-Processing (Post Train-Test-Split)**
  - TFID-Vectorizer on lemmatized column, removing stop words
  - Standard Scaler on numerical columns

Modeling

- Random Forest with BayesSearchCV
- Balanced Random Forest with BayesSearchCV
- PCA and Logistic Regression with BayesSearchCV
- **Balanced Logistic Regression with BayesSearchCV - Best Model**

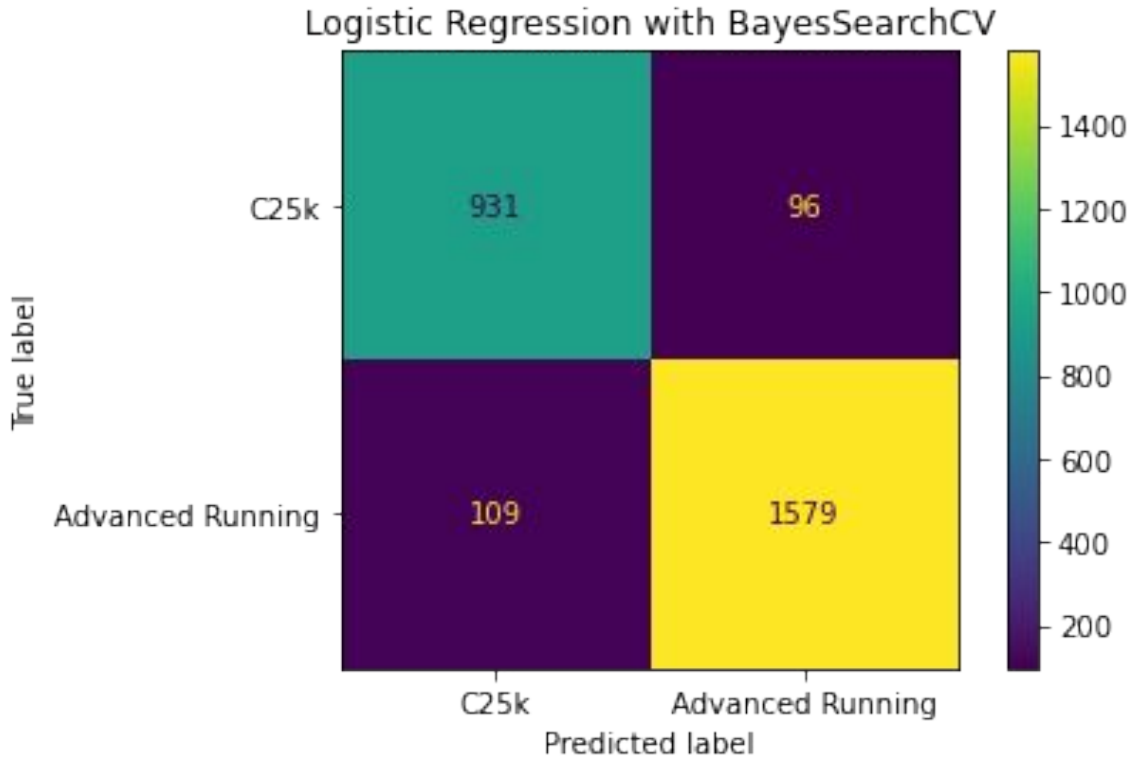
Model Results

Model	Training Accuracy	Testing Accuracy	Balanced Testing Accuracy	Precision
Random Forest	.7462	.7366	.7639	.8964
Balanced Random Forest	.7634	.7506	.7749	.8983
PCA, Logistic Regression	.8873	.8939	.8928	.9294
Balanced Logistic Regression	.9241	.9244	.9209	.9427

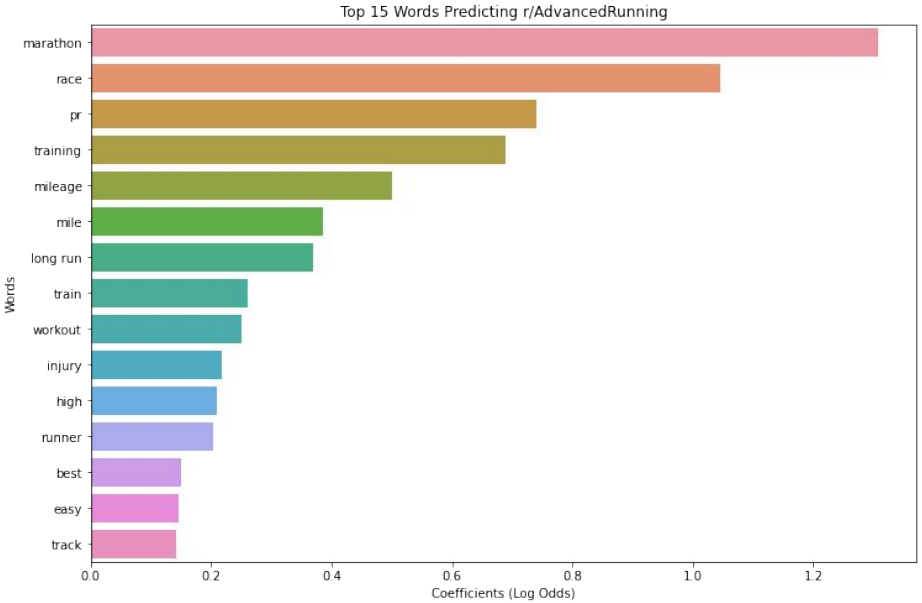
- Logistic Regression accounted for 92.09% variance in r/AdvancedRunning

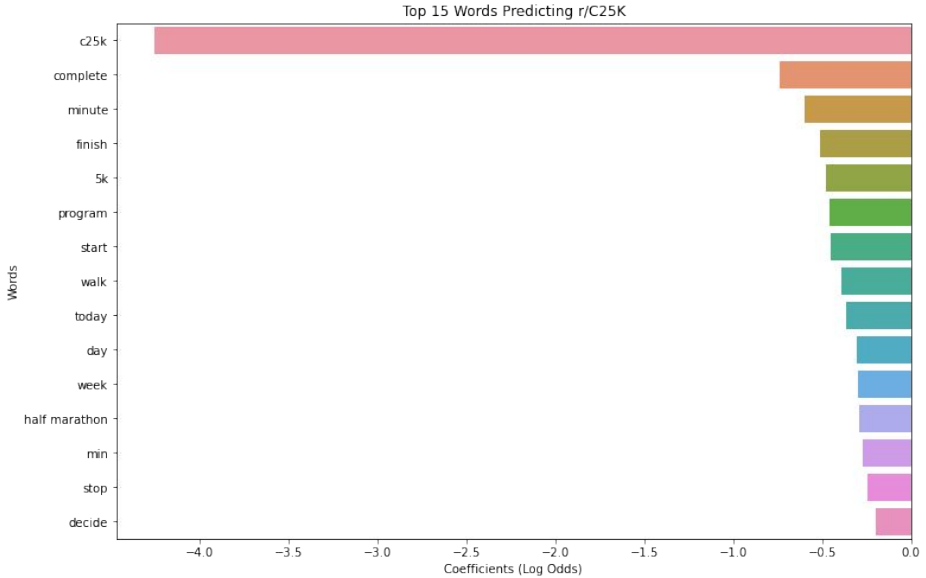
Confusion Matrix

- **Precision:** Minimizing False Positives



Top Predictive Words





Conclusions and Recommendations

- Best Model: Balanced Logistic Regression Model
  - 92.09% BAC
- Competitive words focused around improvement and optimization of race performance were predictive of r/AdvancedRunning
- Starting/completing theme over optimization when predicting r/C25K
- Re-targeting content for Advanced Runners:
  - Tips for longer runs
  - How to get a PR in your next race
  - How to prevent and deal with injury

Thank you!