PROJECT FITNUT - CLASSIFICATION MODELLING VIA NATURAL LANGUAGE PROCESSING

DSI PROJECT 3

PROBLEM STATEMENT

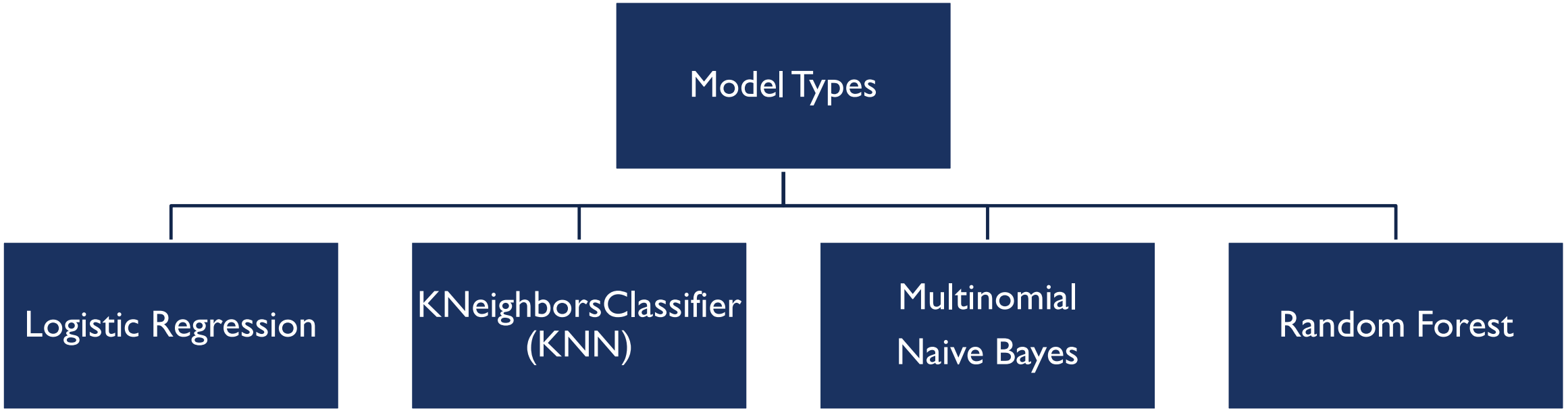
- SURGE - an elite private gym specializing in curated fitness/wellness programmes under the Core Collective group - is exploring a new business unit that focusses on a tailored dual fitness-and-nutrition concept.
- A blanket approach was adopted in downloading 2,000 threads from the bodyweightfitness and EatCheapAndHealthy subreddits. However, the fitness and nutrition portfolios are handled by two different teams in SURGE.
- As the hired Data Science consultant, develop a classification model to determine which of the abovementioned subreddits a post originates from.

DATA SOURCES

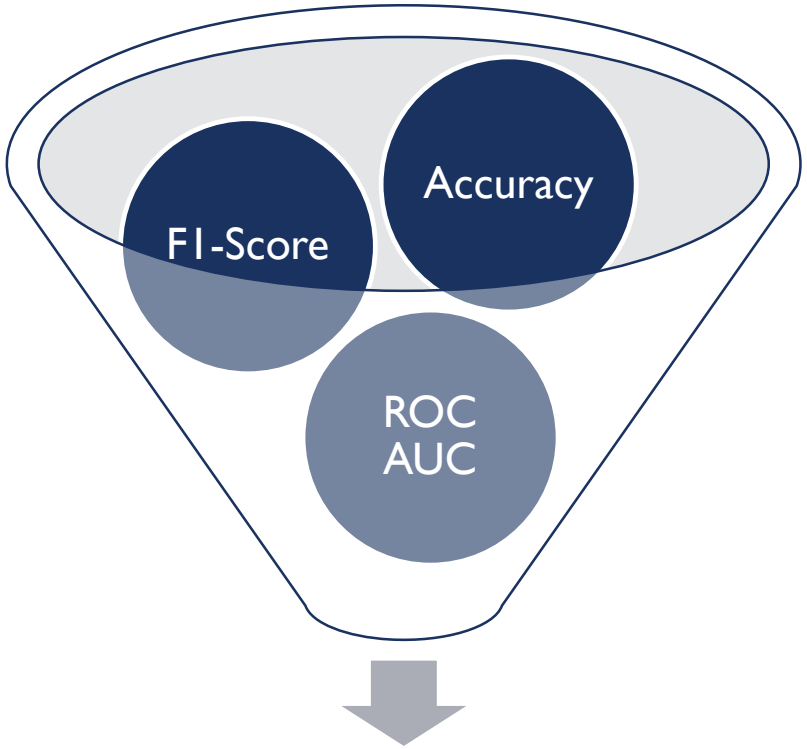
bodyweightfitness

EatCheapAndHealthy

MODEL FRAMEWORK



SCORING METRICS



SCORING METRICS

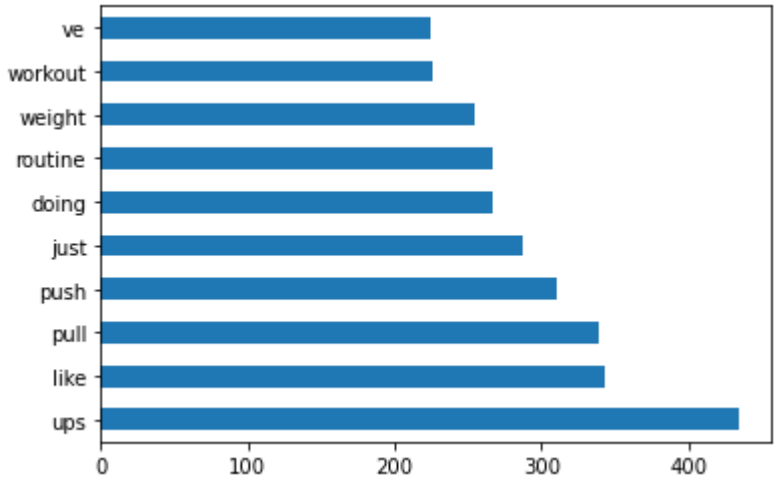
i) Accuracy: Most intuitive indicator which gives the ratio of correct predictions to total predictions

ii) F1-score: Not only gives weight to the percentage of true positives over the total positives in the data but serves as an indicator for confidence of predicted positives too; seeks to optimize both precision and recall simultaneously

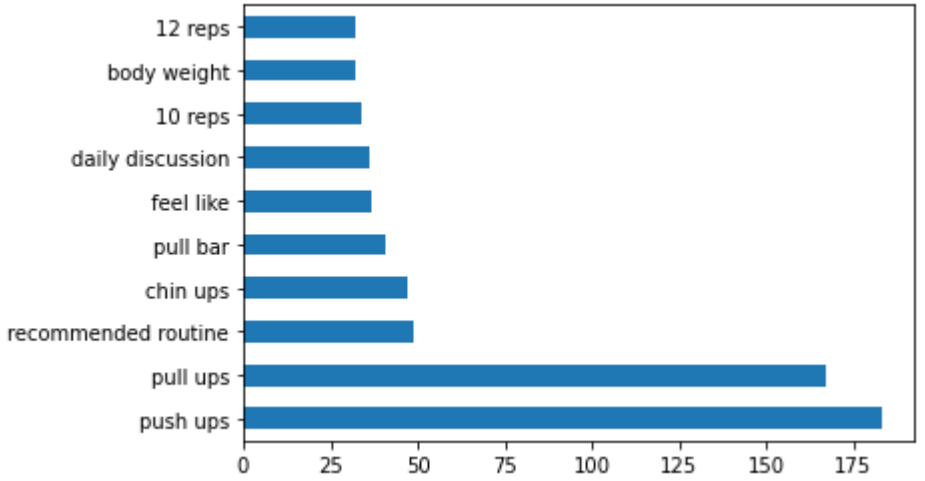
iii) ROC AUC Score: Shows a model's effectiveness in minimizing minimizes false positives and false negatives, as well as the ability to distinguish between binary classes

EDA (FITNESS)

Unigrams

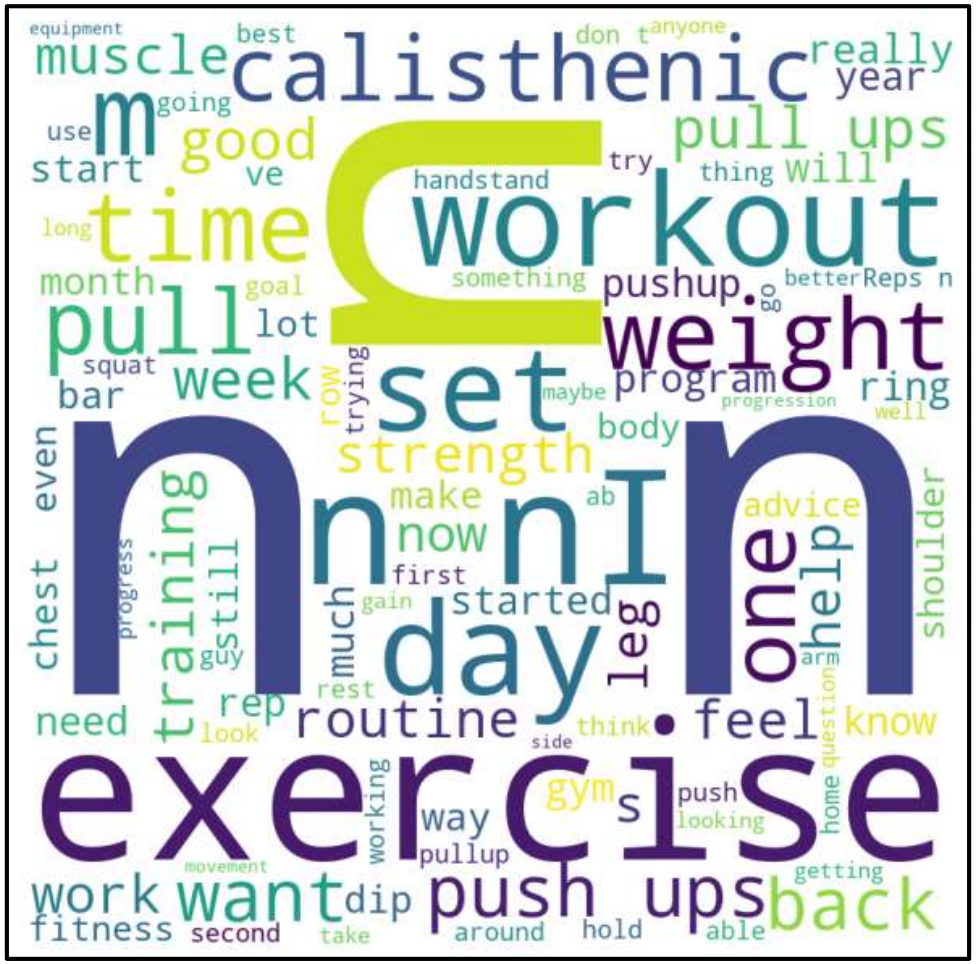


Bigrams



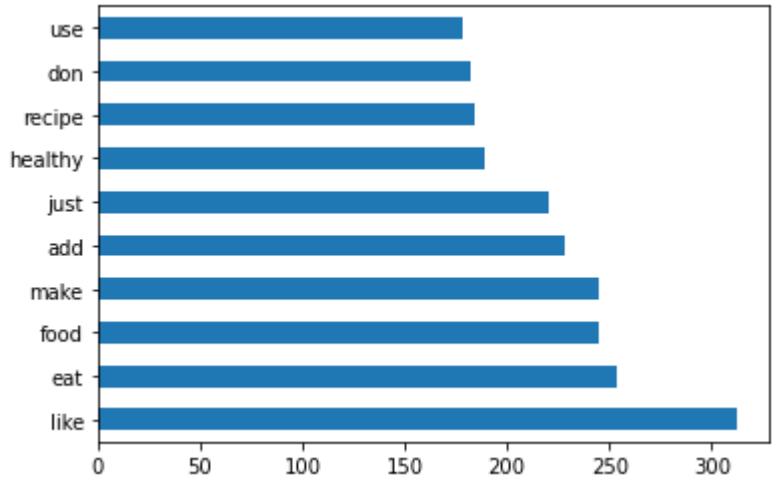
- Bigrams generally more meaningful

WORDCLOUD (FITNESS)

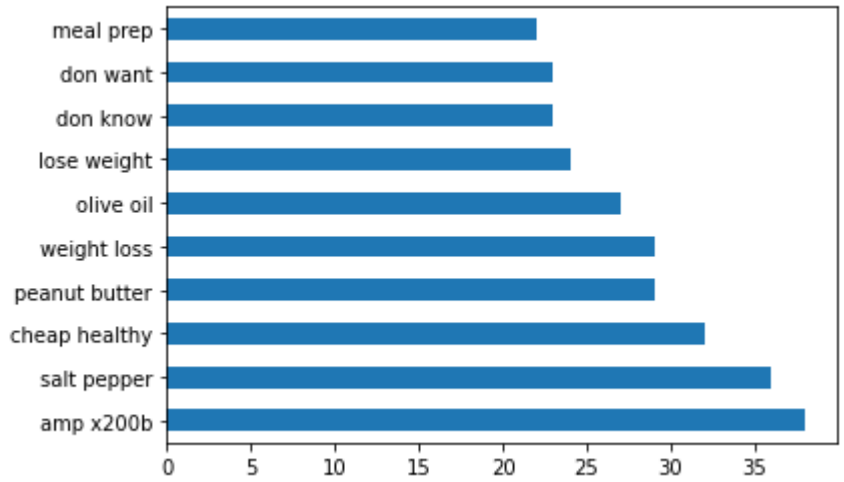


EDA (NUTRITION)

Unigrams

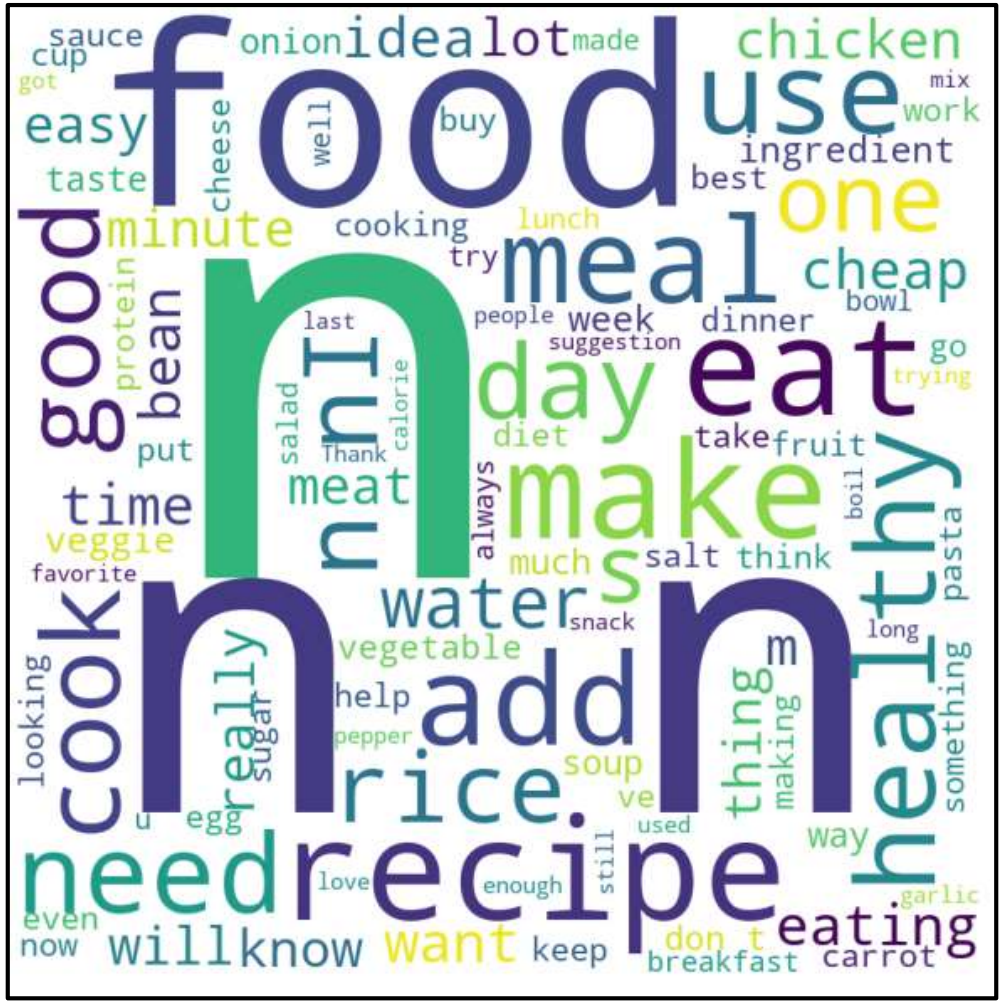


Bigrams

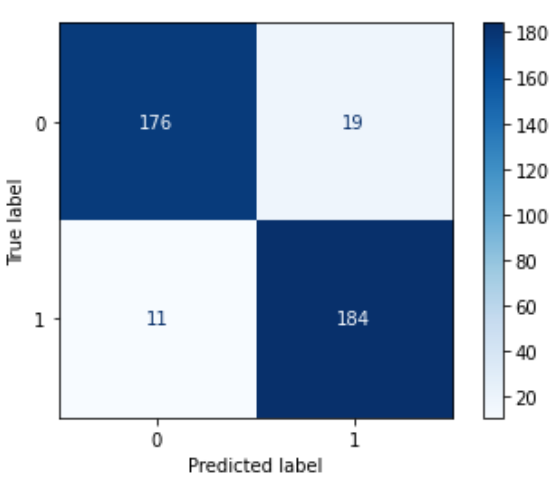


- Expected unigrams and bigrams based on topic

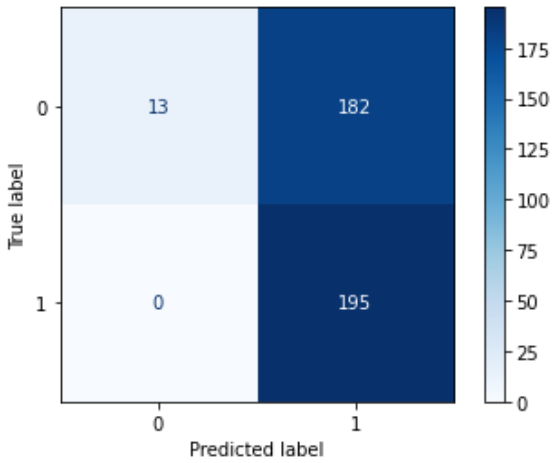
WORDCLOUD (NUTRITION)



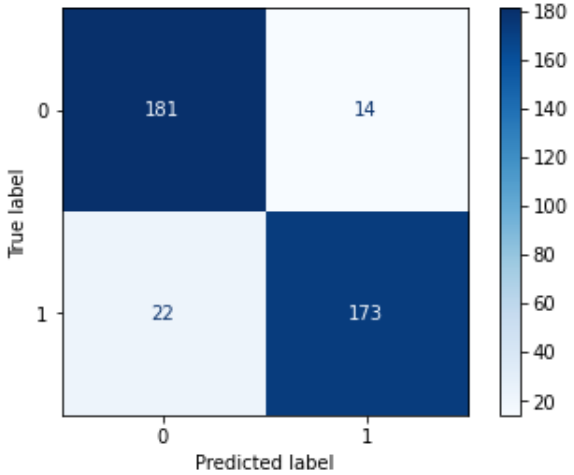
CONFUSION MATRIX



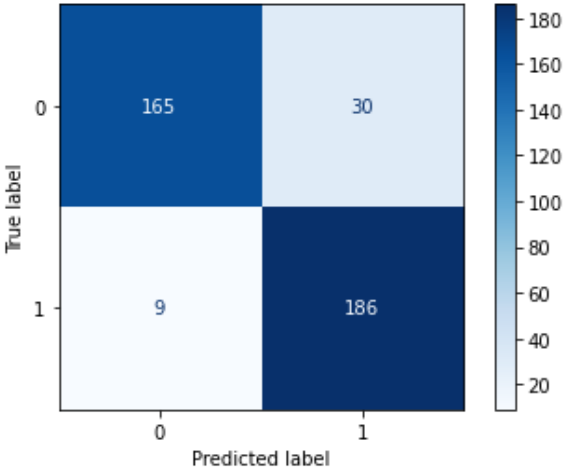
Logistic Regression



KNeighborsClassifier (KNN)



Multinomial Naive Bayes



Random Forest

TP = True Positives (Post predicted as belonging to bodyweightfitness subreddit and indeed belonging to bodyweightfitness subreddit)
TN = True Negatives (Post predicted as belonging to EatCheapAndHealthy subreddit and inteed belonging to EatCheapAndHealthy subreddit)
FP = False Positives (Post predicted as belonging to bodyweightfitness subreddit but actually under EatCheapAndHealthy subreddit)
FN = False Negatives (Post predicted as belonging to EatCheapAndHealthy subreddit but actually under bodyweightfitness subreddit)

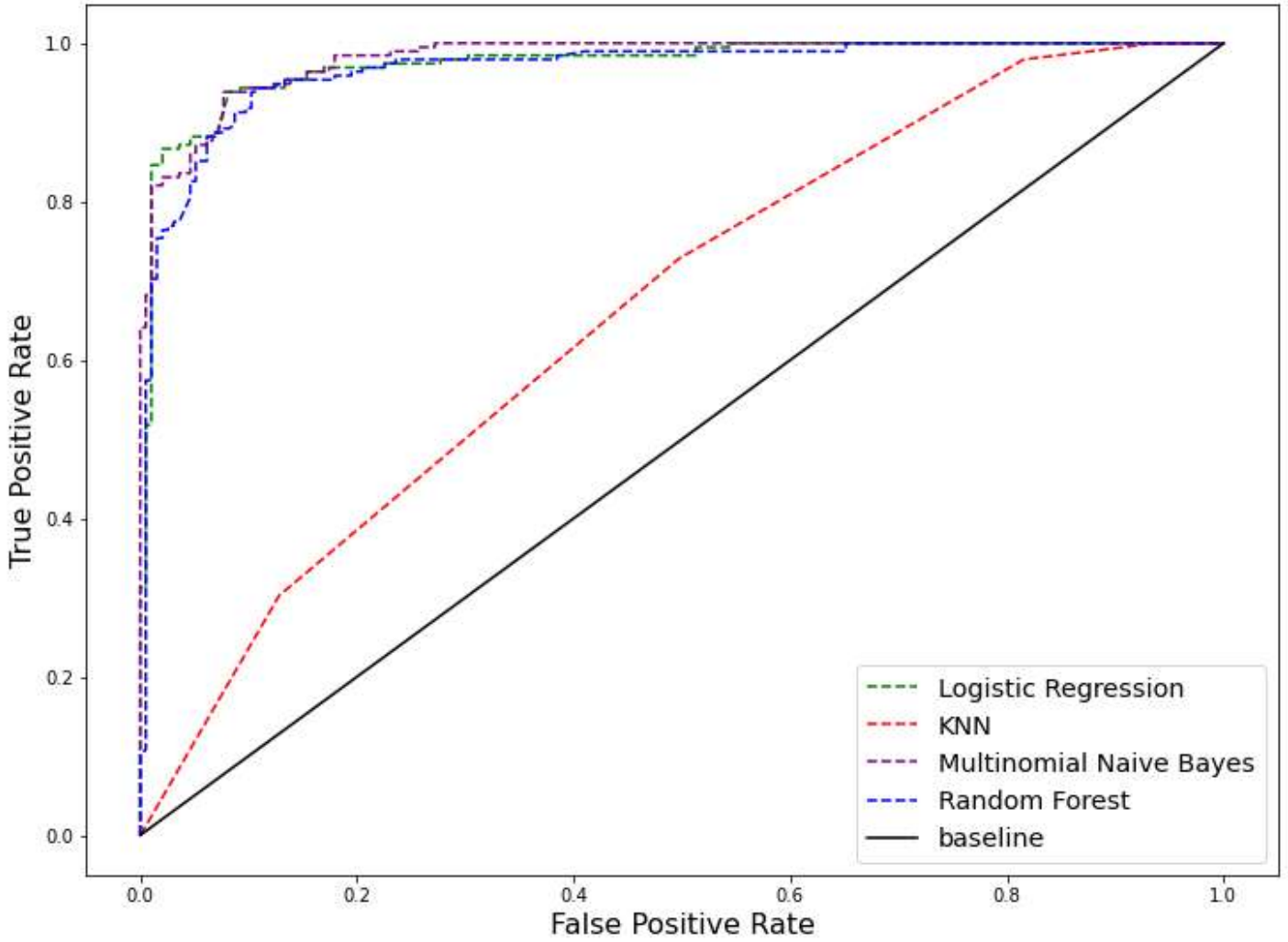
MODEL EVALUATION

	Train Accuracy Score	Test Accuracy Score	F1 Score	ROC AUC Score
Logistic Regression	0.9301	0.9231	0.9246	0.9738
KNeighborsClassifier	0.7936	0.5333	0.6818	0.6645
Multinomial Naive Bayes	0.9244	0.9077	0.9058	0.9804
Random Forest	0.8962	0.9000	0.9051	0.9665

Production model parameters:

- **Model type: Logistic Regression**
- Max number of features: 2,500
- Stop words used: English
- Best NGram: Unigram
- LogReg Penalty: Ridge
- LogReg Penalty Strength: Minimal

ROC CURVES


True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
Logistic Regression
KNN
Multinomial Naive Bayes
Random Forest
baseline

RECOMMENDATIONS

- SURGE may confidently adopt the chosen production model for the fitness and nutrition teams to predict parent subreddits, as the model achieved the established benchmark of 90%.
- Generally, sentiments for both bodyweightfitness and EatCheapAndHealthy subreddits are positive, and information drawn from the posts will likely be indicative of prevailing and upcoming trends.

THANK YOU