Nutrition vs Keto

Subreddit Classification



Data Science Process

01 BUSINESS Problem and Introduction

02 DATA CLEANING and Pre-processing

03 EXPLORATORY Data Analysis

04 MODEL SELECTION and Insights

Business Problem:

Can we automate the classification of post so that only appropriate topics are posted for each subreddits?

Intro to our subreddits

Nutrition

A subreddit for the discussion of nutrition science. Macronutrients, micronutrients, vitamins, diets, and nutrition news are among the many topics discussed.

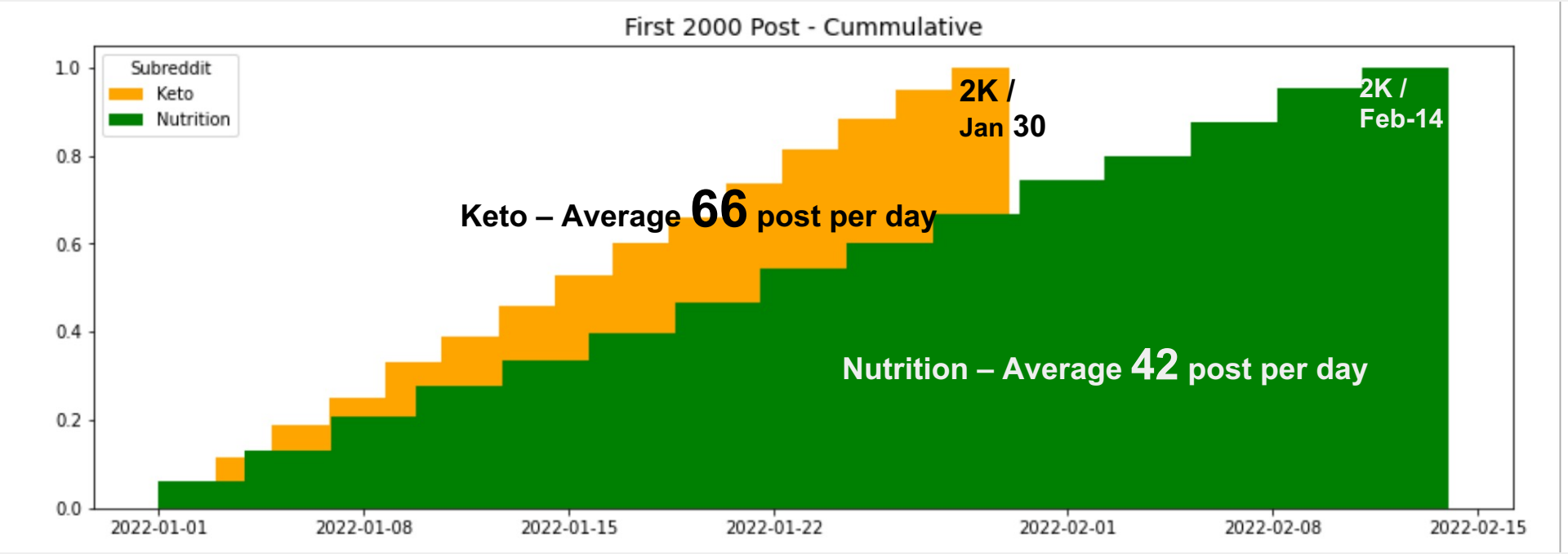
2.3M Members

Keto

The Ketogenic Diet is a low carbohydrate method of eating. Place to share thoughts, ideas, benefits, and experiences around eating within a Ketogenic lifestyle. Helping people with diabetes, epilepsy, autoimmune disorders, acid reflux, inflammation, hormonal imbalances, and a number of other issues, every day.

2.9M Members

Data Collection (Data from Jan to Feb '22)



First 2000 post from Jan 1, 2022:

Since Keto is more active and have more members, it reached the 2K post on Jan 30, 2022. Nutrition's reached 2K

Cleanup and Pre-processing

Cleanup Author:

Keto- >AutoModerator, Basic_Site5449 **1.5%**

Nutrition >AutoModerator Disastrous-Drop-8085 **2.0%**

Remove Duplicate:

Title **2.0%**

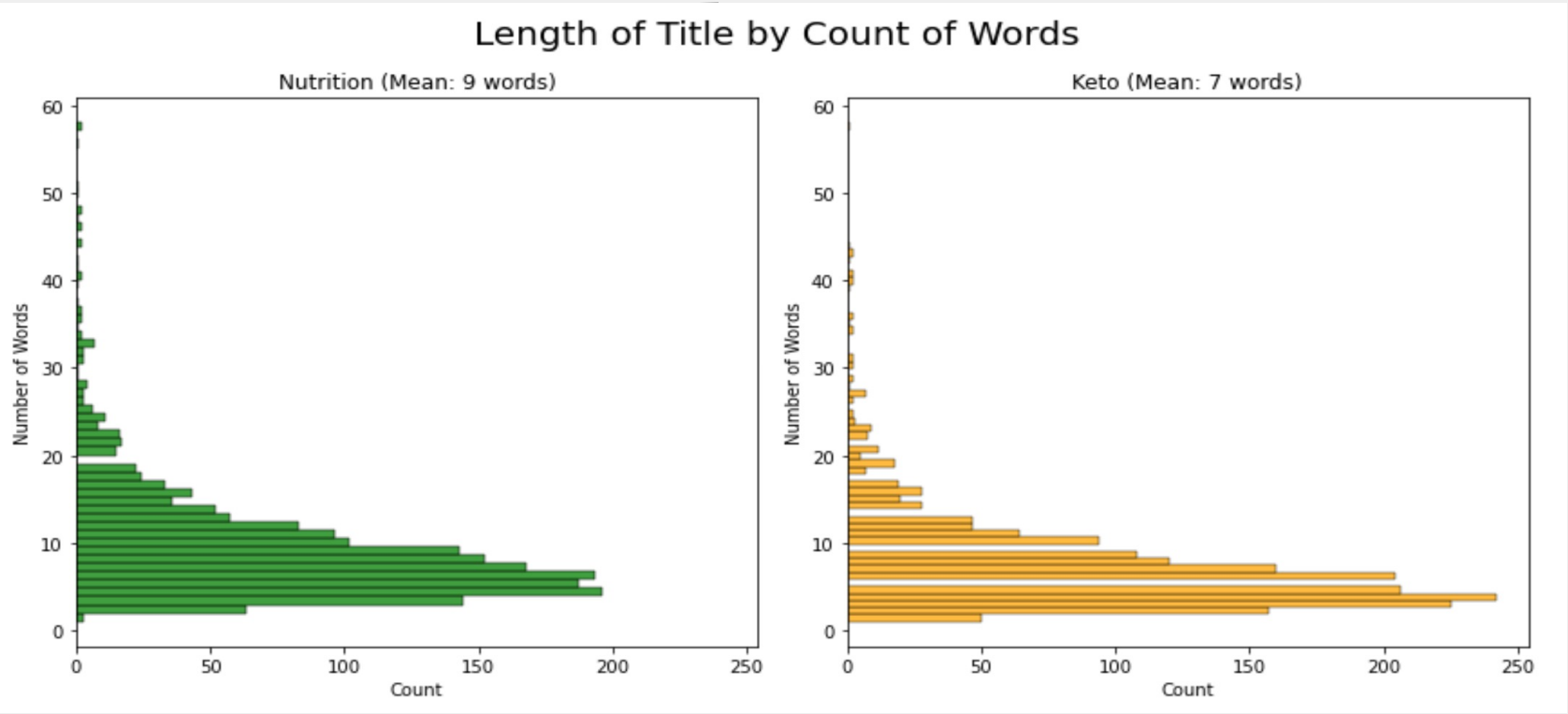
- **Tokenizing** (lower case and removed punctuations)
- **Stemming** (transforms a word into its root form, *e.g. eating/eats to eat, losing/loses to lose, why to whi, does to doe*)



Feature(s)?

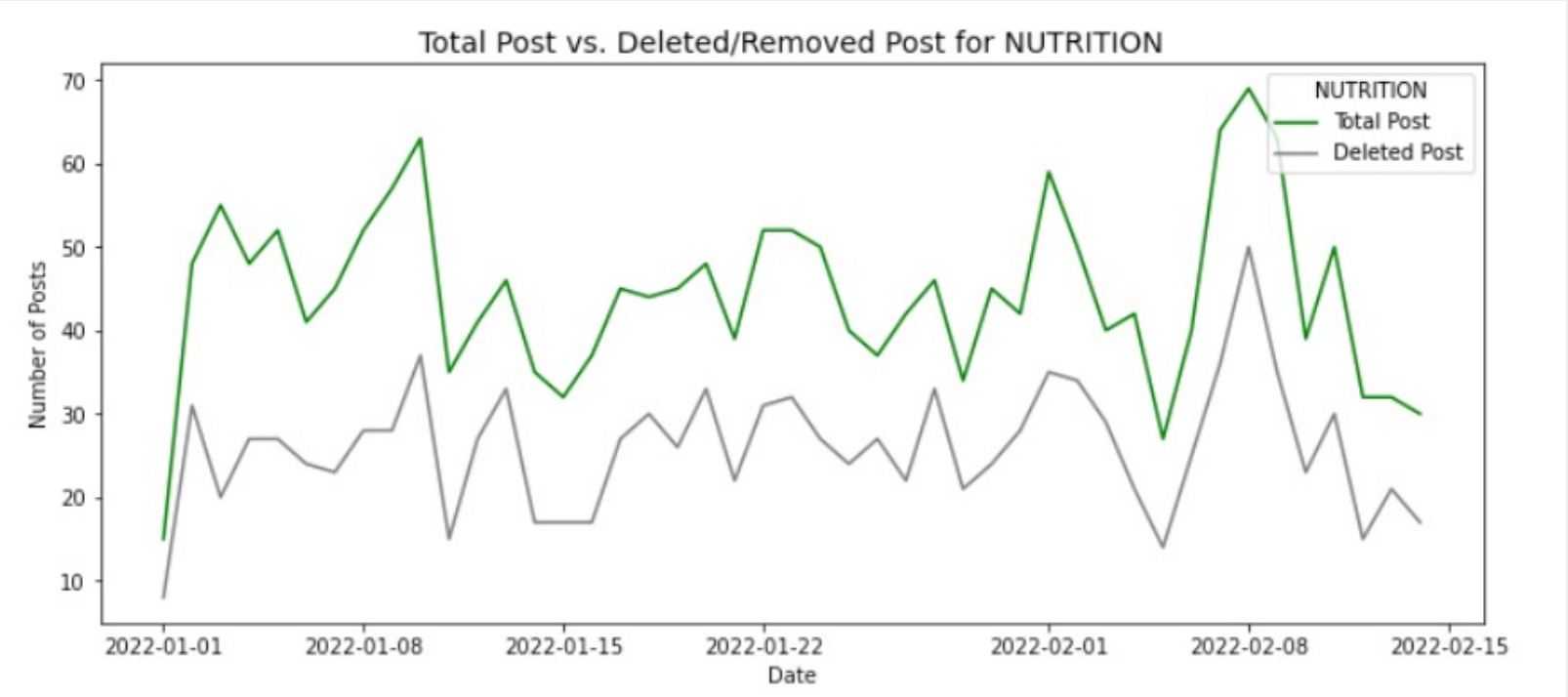
Title or Self-text,
... Or both?

Exploration/ Title Analysis



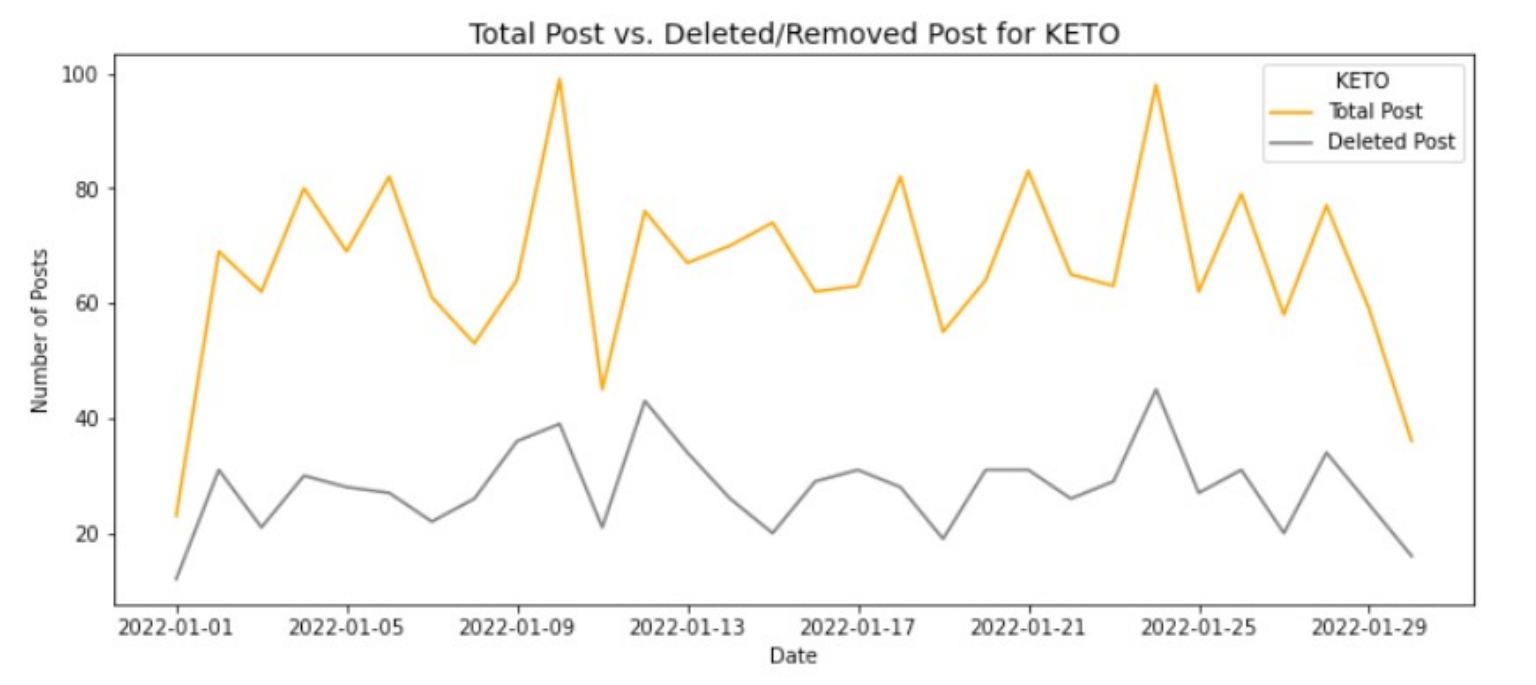
Nutrition has slightly higher average count of words, they have (almost) similar distributions

Self-Text Analysis



59% of the self-text was removed/deleted

Self-Text Analysis

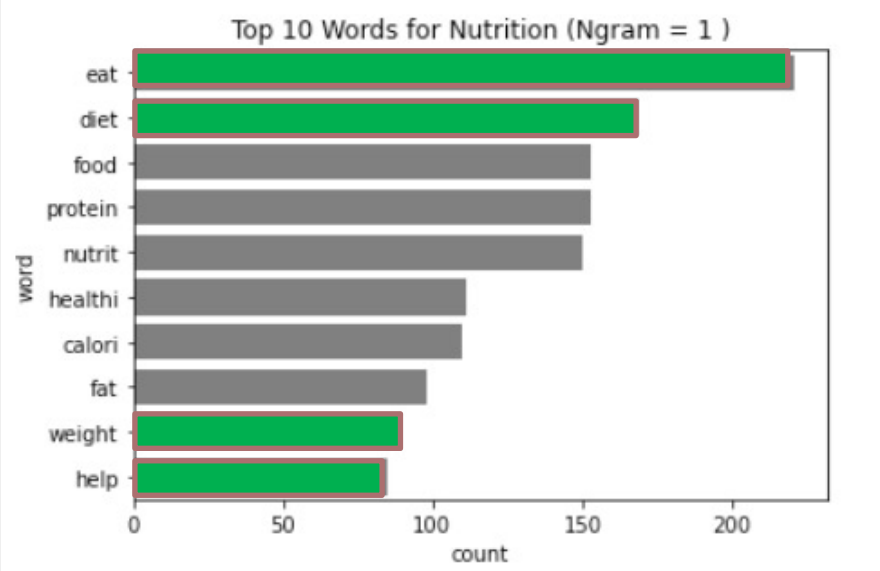


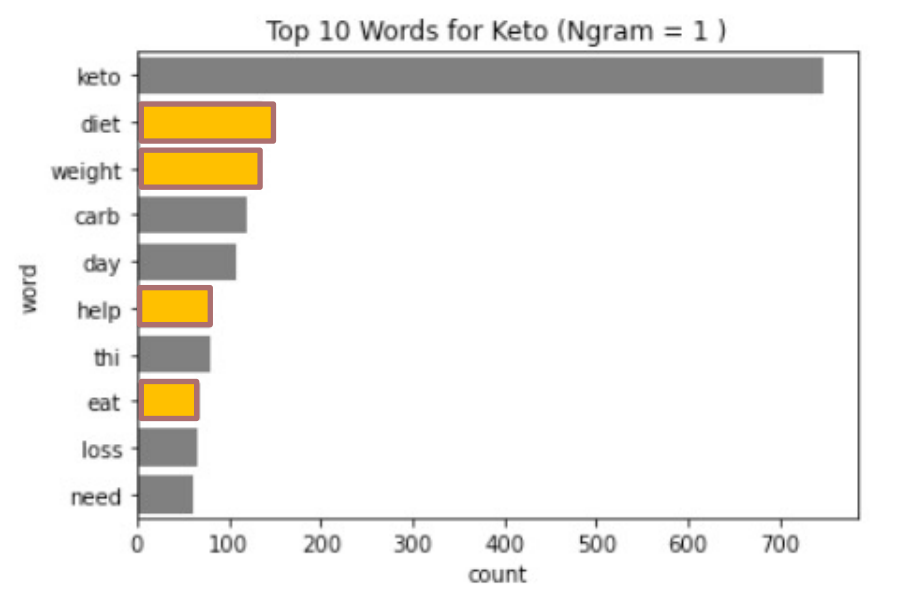
41% of the self-text was removed/deleted

Feature: Title

We use **CountVectorizer** to extract feature from ***Title..*** to transform ***Title*** into a bag of words (in a simple term), or a vector on the basis of the frequency or count of each word that occurs in the entire ***Title*** (as a technical description)

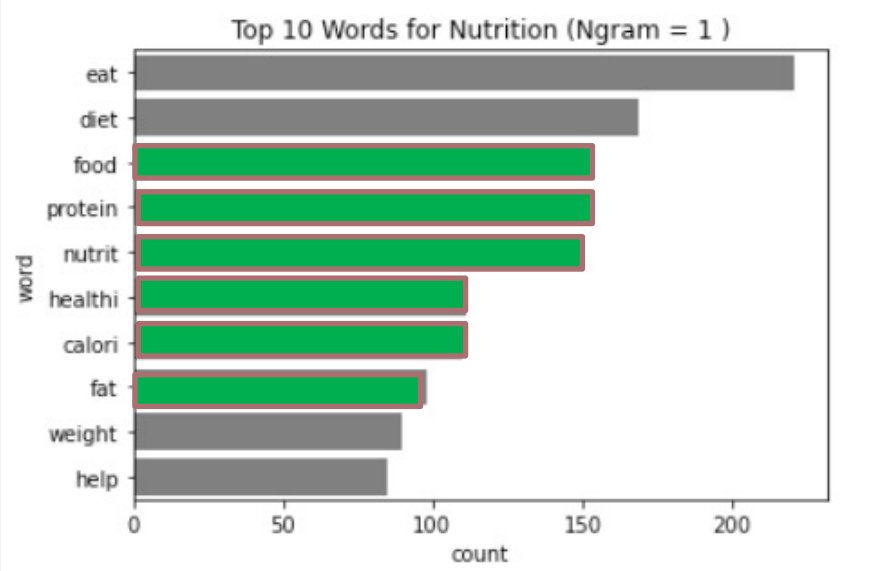
Word Analysis
Top 10 words

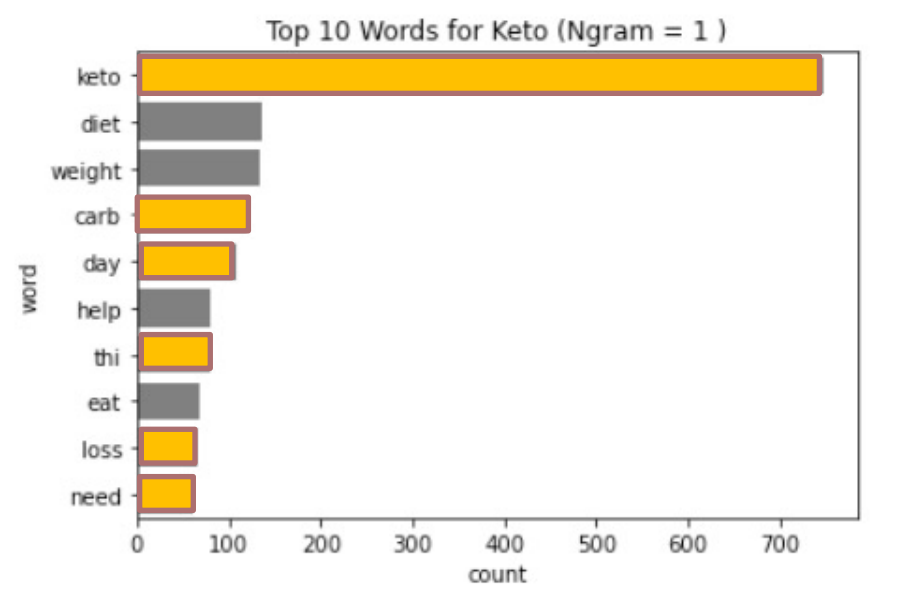




Similarities

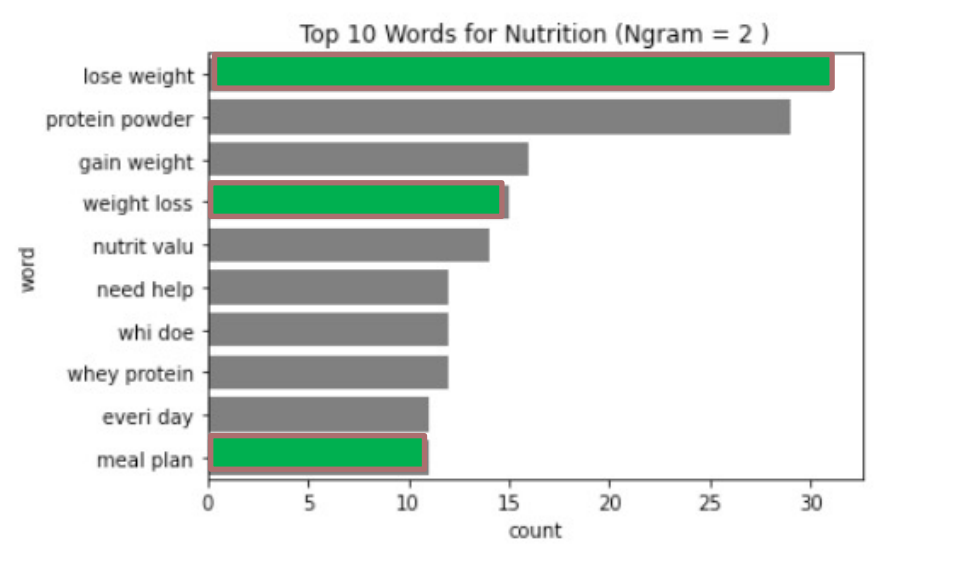
Word Analysis
Top 10 words

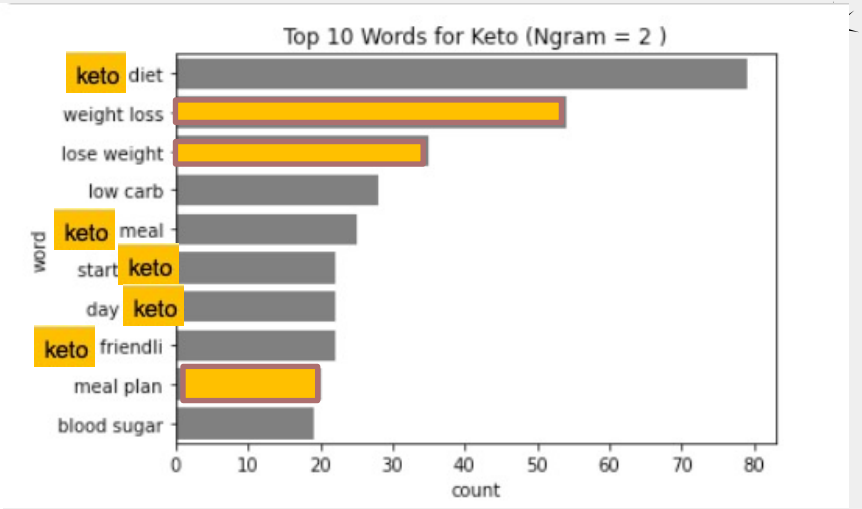




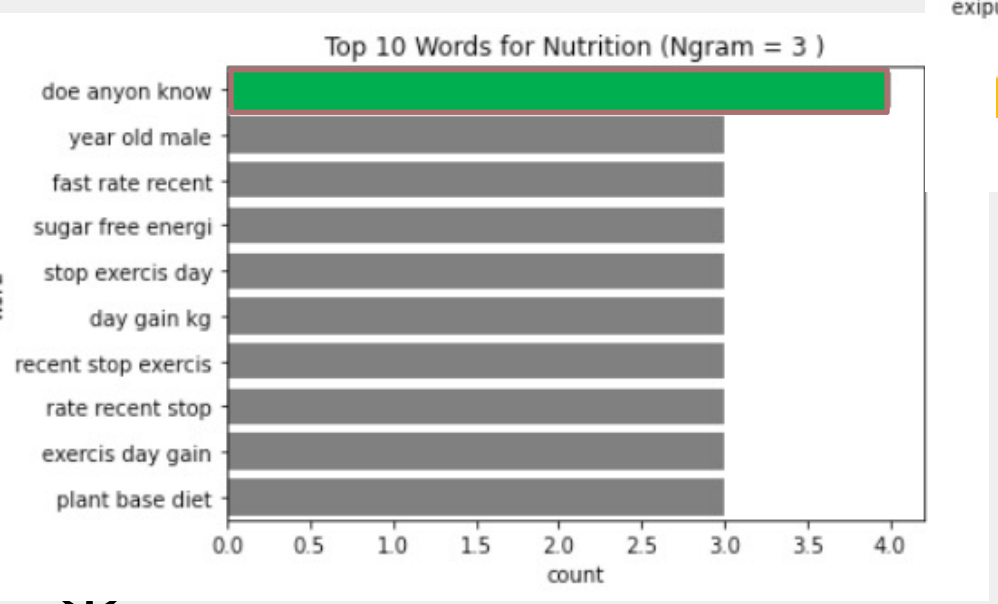
Differences

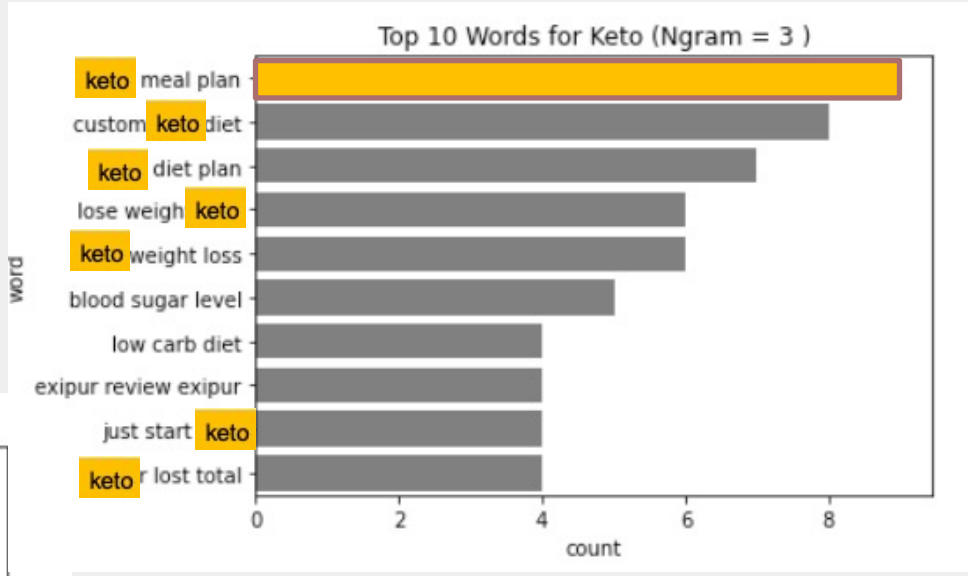
Word Analysis (BiGram)



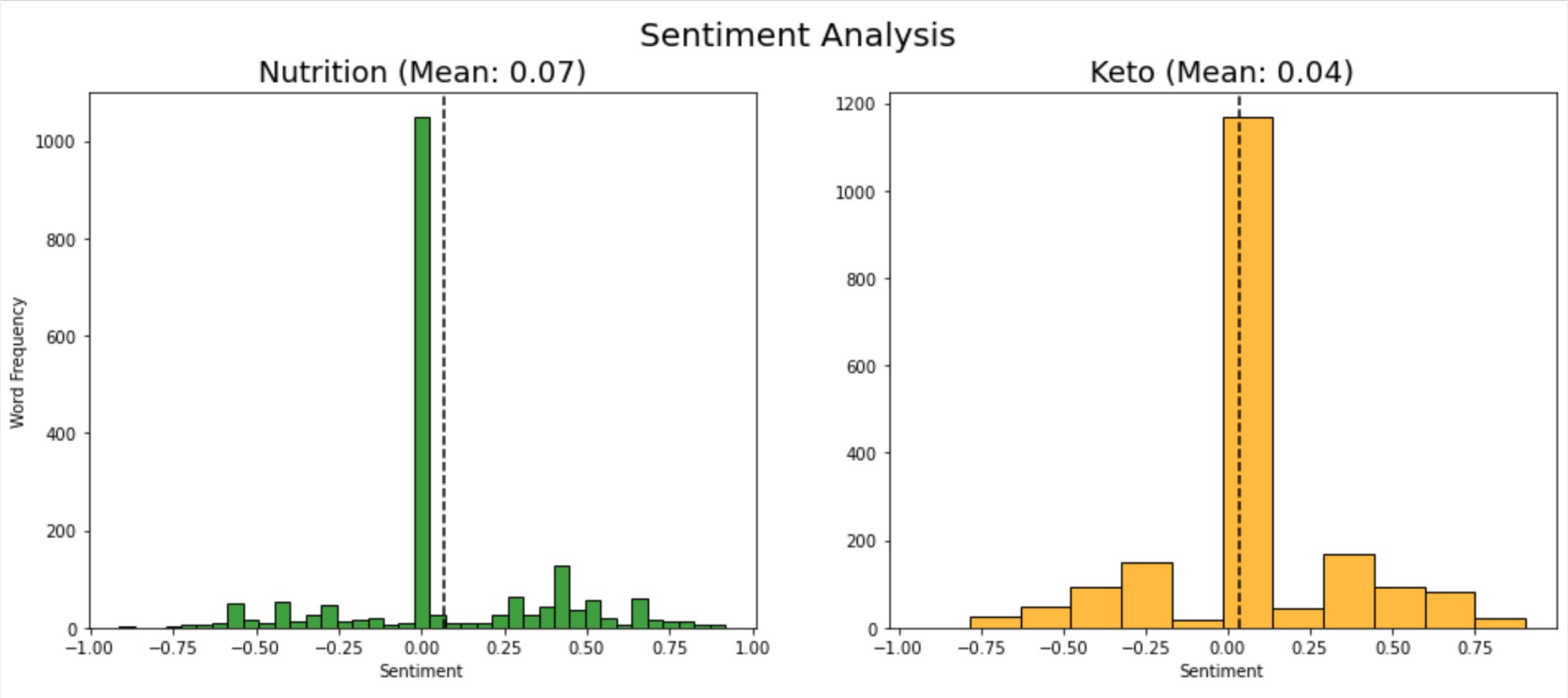


Word Analysis (TriGram)





Exploration / Sentiment Analysis



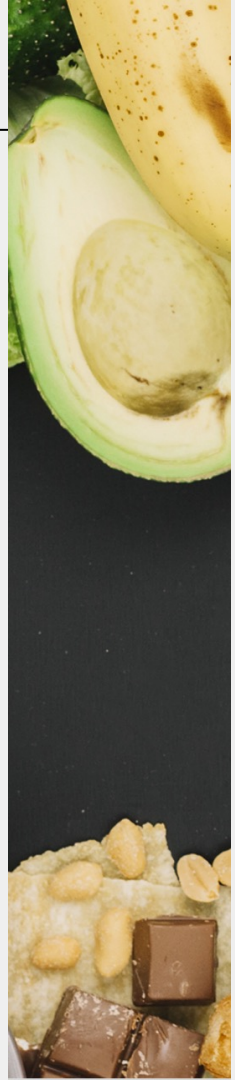
Nutrition has more positive sentiments than Keto

Nutrition

80%

*'What are some of the **best healthy** foods to eat to boost appetite and sustain energy?'*

'Best liver care for high alcohol consumption'



-70%

“I screwed up and ate 40 to 60 grams of saturated fat everyday for 2-3 years while bulking. Have I caused ireveisable damage to my body?”

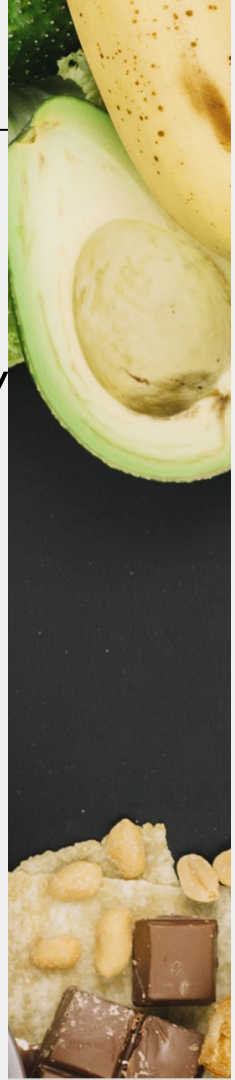
"I'm Scared My Diet Is Going To Kill Me Early"

Keto

80%

"That feeling of bliss when you cook a new recipe/dish for the first time and it's so good it brings a tear to your eye..."

"The best fish I've ever eaten!! What is your best food?"



-80%

'Do You Make These Four Common Diet Mistakes That Sabotage Your Health and Stall Fat Loss?'

'I want to stick to keto so bad, but I keep finding myself cheating'

MODEL SELECTION

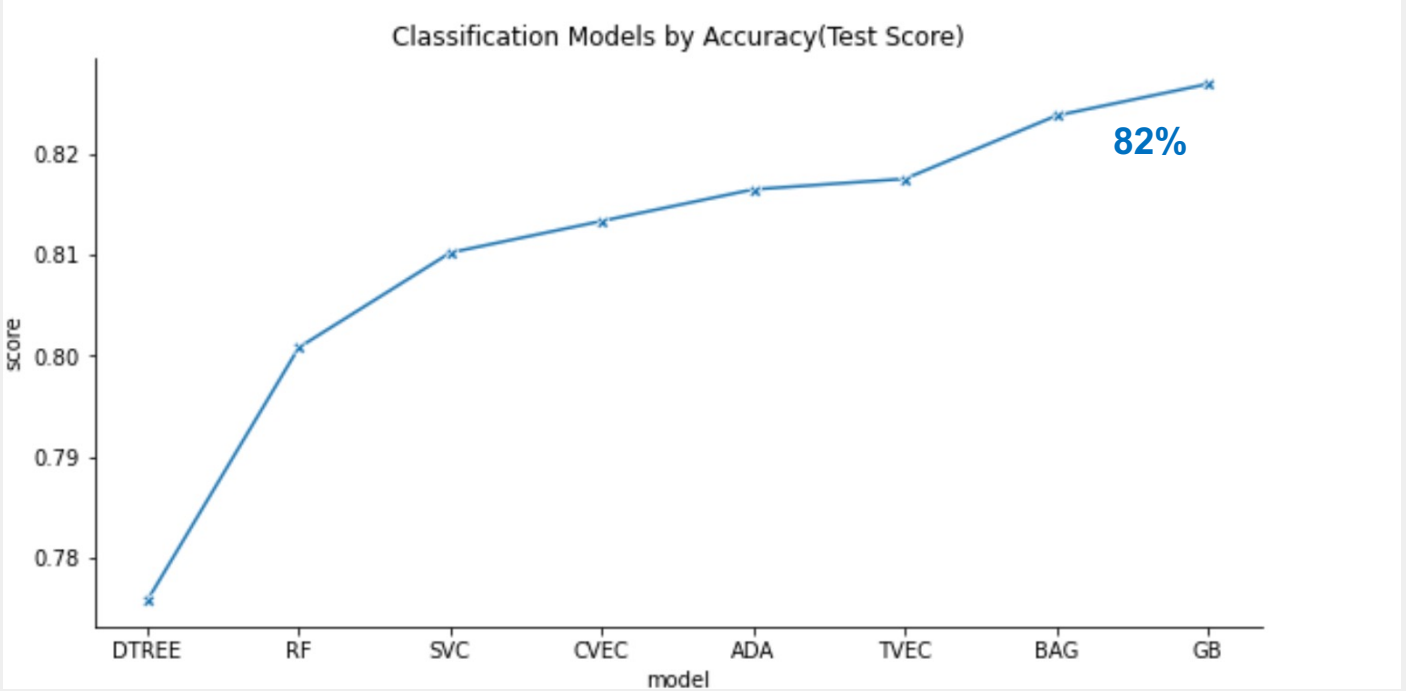
Our baseline model:

50%

CLASSIFICATION MODELS

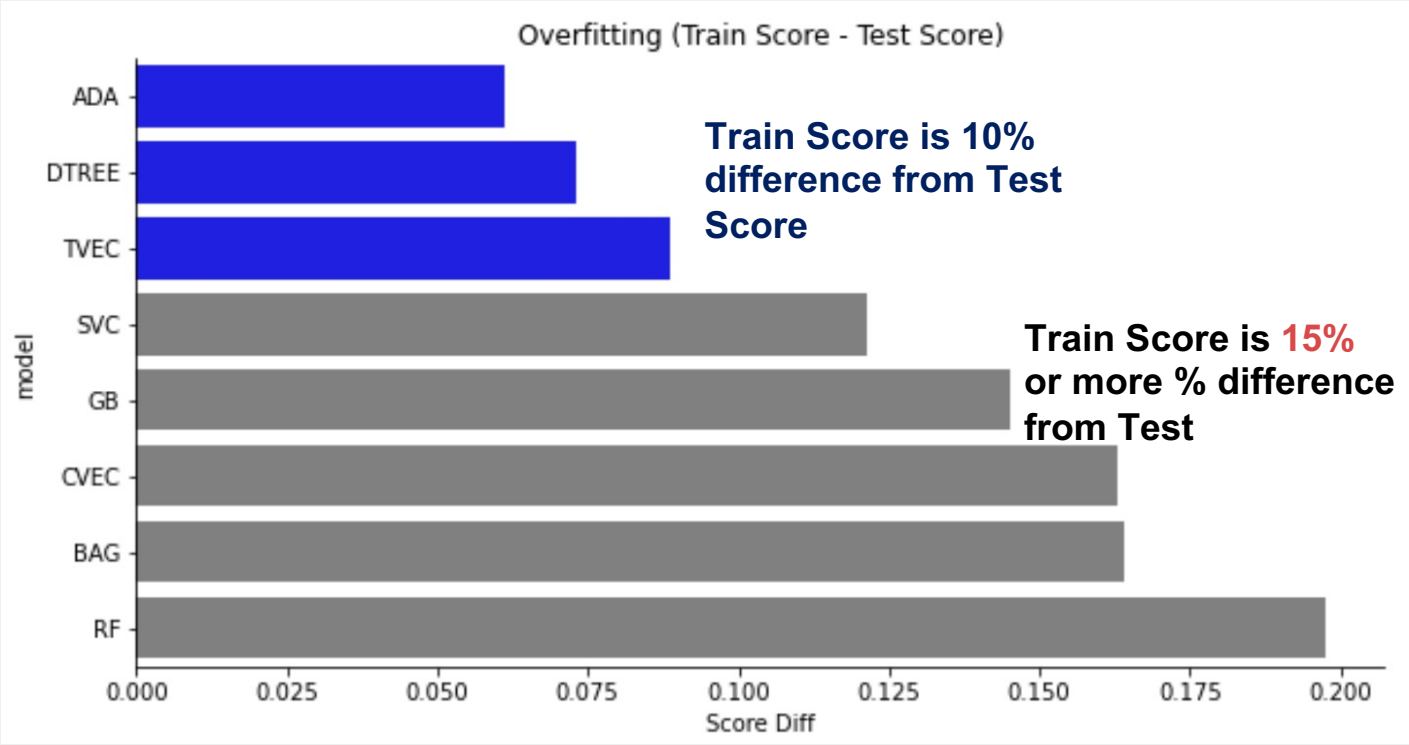
8 Models:

DecisionTree
RandomForest
SVC
CountVectorizer
AdaBoost
TFIDF Vectorizer
BaggingClassifier
GradientBoost



Challenges:

1. Overfitting problem



2. Time to run GridSearch using hyperparameters: *Boosting is slowest amongst all models, as it builds the model in sequential way (compared to Bagging where it trains in parallel or independently) so tuning the hyperparameters can take longer time for Boosting*

Classification Models – Ensemble Technique

GradientBoost Classifier

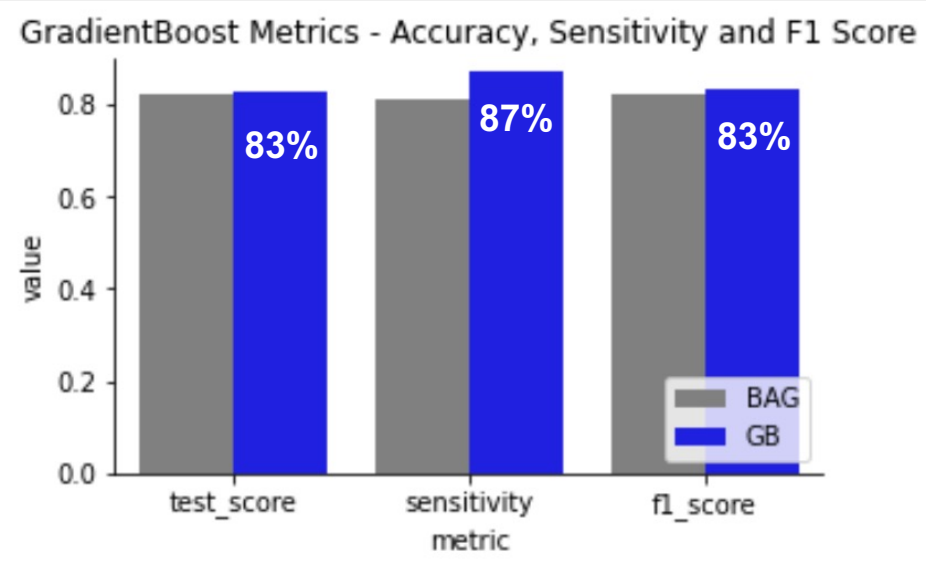
Boosting *is a method of converting weak learners into strong learners. Boosting takes a weak base learner and tries to make it a strong learner by retraining it on the misclassified samples.*

In ***gradient boosting****, it trains many model sequentially. Each new model gradually minimizes the loss function of the whole system using Gradient Descent method.*

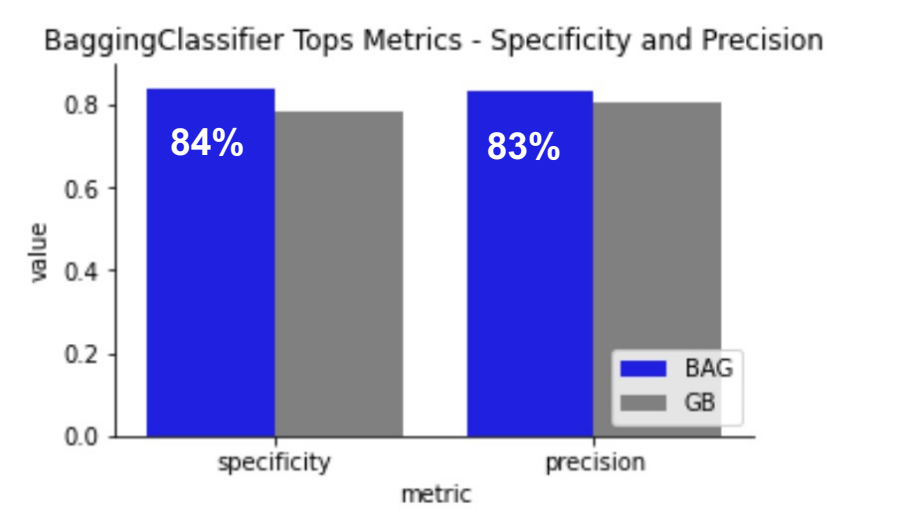
Bagging Classifier

Bootstrap - *means random resampling of data and we control whether samples and features are drawn with or without replacement.*

Other Metrics :

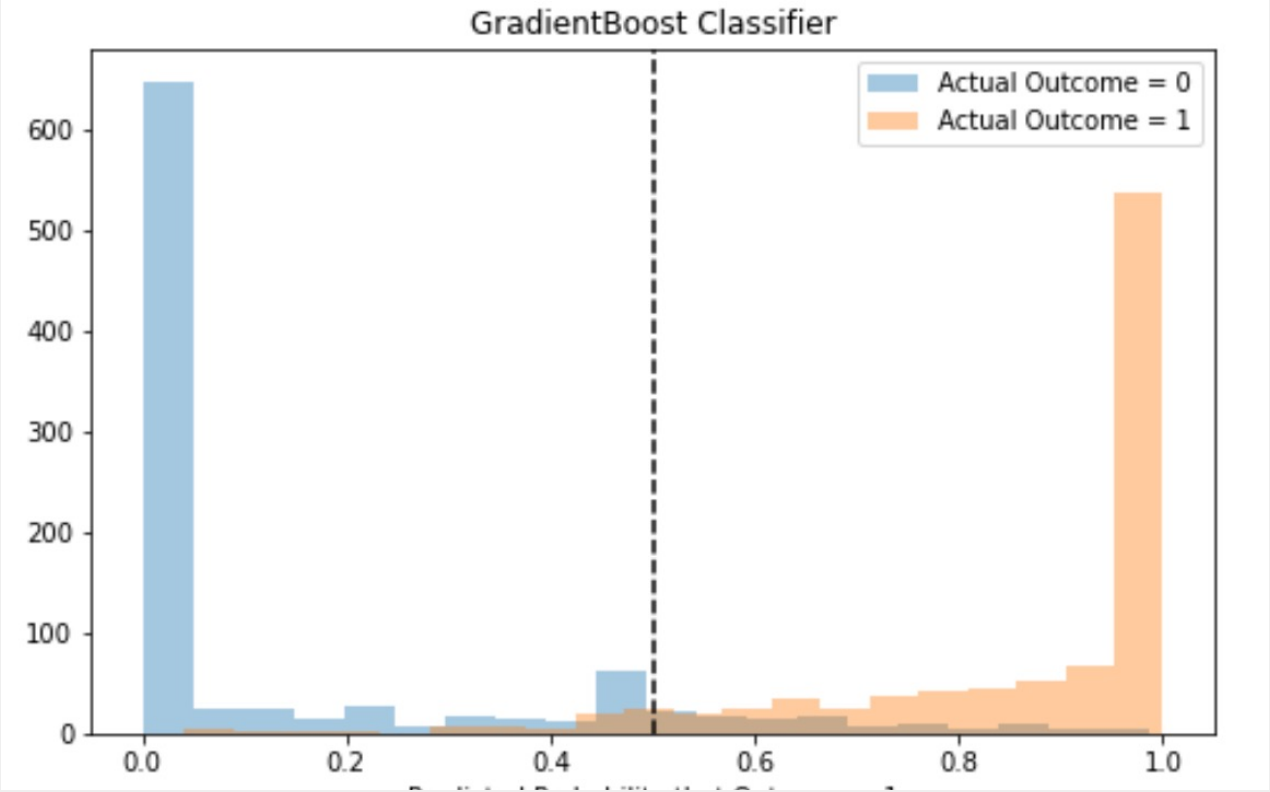


GradientBoost wins Accuracy, Sensitivity and F1 Score

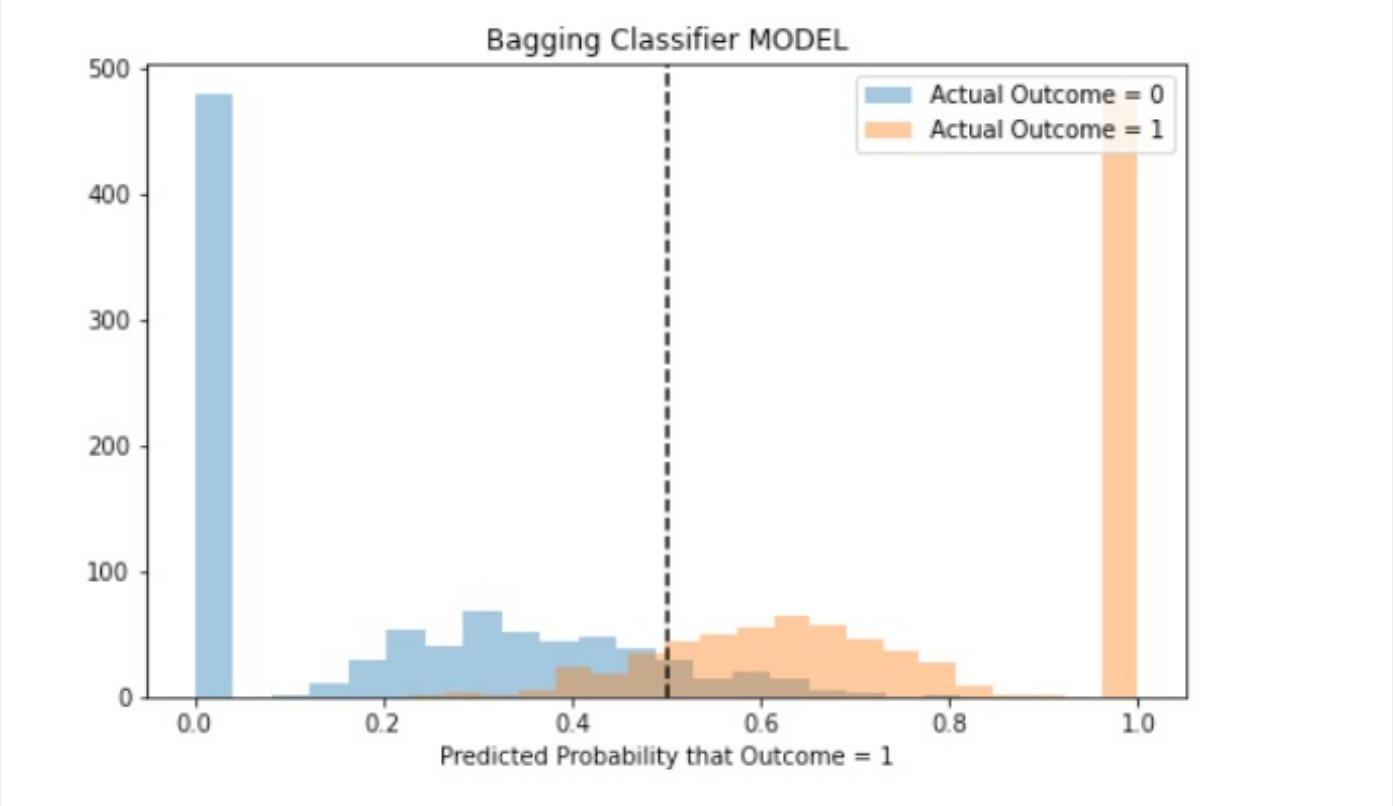


BaggingClassifier wins Specificity and Precision

Distribution of Probabilities – GradientBoost



Distribution of Probabilities – Bagging Classifier



Selected Model: TFIDF Vectorizer

Reason: We selected ***GradientBoost Classifier*** as our best model amongst all models for this classification project, as it is highest not just in Accuracy but also in F1-Score, as we want to balance Sensitivity and Specificity.

There is a trade-off between getting the best result and the interpretability of the models, as this is more complex model and we cannot look into the coefficient of which words are highly correlated to each subreddits.

Best Params:

learning_rate: 0.1,
max_depth: 4,
n_estimators: 600,
random_state: 0

Model Insight: Misclassified Post

False Negative

	title	actual	predict
	Not sure if I can ask this here but…	Nutrition	Keto
	USNews asked 27 nationally recognised experts to rate diets based on short/long-term weight loss, nutritional completeness and heart disease/diabetics risk factors. The Mediterranean diet came in first place and while Keto came in last	Nutrition	Keto
	What are the benefits of the keto diet plan	Nutrition	Keto
	Can 40gms of MonkFruit sweeteners cause Insulin & Energy spikes & crashes?	Nutrition	Keto
	FREE KETO COOKBOOK	Nutrition	Keto
	Is the keto diet good for your body or not?	Nutrition	Keto
	Why do I gain so much weight after a night out and it takes me 5 days to get back to where my weight was?	Nutrition	Keto
	Do keto diet work?	Nutrition	Keto
	Keto Diet and its health benefits	Nutrition	Keto
	Keto Diet or pronounced as Dieto Keta in some European countries	Nutrition	Keto
	Is the high amount of fat and low amount of vegetables in a keto diet unhealthy in the long run?	Nutrition	Keto

Model Insight: Misclassified Post

False Positive

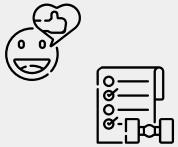
	title	actual	predict
	Healthy Protein bar with no Soy?	Keto	Nutrition
	Vitamin c	Keto	Nutrition
	do you think it's unhealthy to eat 3-4 cups of cheese a day instead of red meat?	Keto	Nutrition
	The best fish I've ever eaten!! What is your best food?	Keto	Nutrition
	Micronutrition tool that could help you improve your nutrition (free)	Keto	Nutrition
	Inaccurate nutrition facts	Keto	Nutrition
	Ravi Speaks:-EXTRA SUGAR INTAKE CAUSES OVERWEIGHT & DISEASED CONDITIONS.	Keto	Nutrition
	Carb manager: I am using the app to keep track of macros. I am doing 5/25/70 but I noticed today that the macros calculator is showing 30g carbs, 148g protein, and 184g fats. With 30 daily carbs.	Keto	Nutrition
	Qualified Medical Nutrition Therapy in the USA	Keto	Nutrition
	Great Value chia seeds nutritional info vs other sources of nutritional info	Keto	Nutrition

To answer our Business Problem:

Yes, we have selected GradientBoost Classifier as our model that can classify posts from two different subreddits based on the TITLE with 83% Accuracy.

Recommendation:

Tuning of hyperparameters to overcome overfitting



Include other features like self-text and probably sentiment analysis score might improve our metrics.

Include images or videos in our analysis for more accurate prediction (which requires more knowledge on different ML domains)

A PICTURE IS WORTH A THOUSAND WORDS -- #WORDCLOUD


slidesgo

Thank you!

Rebellon Carina

Data Analyst | Data Engineer |
Aspiring Data Scientist