An investigation into r/keto and r/zerocarb

Sujay Jangam, GA-DSI 27

Context

My Nutrition Clinic

- There has been a rise in the popularity of Ketogenic Diets, and Zerocarb or Carnivore diets over the years.

- There are many studies to show that these diets work well

- However this diet is not without its risks

Context

- The rebound
  - Too difficult
  - ‘End of diet’
  - Bingeing

Problem Statement.

My Nutrition Clinic

- Our team wants to build a classification model to classify our clients into each group based on subreddit data, and minimize false positives especially when it comes to the **Zerocarb** group.

    - We will be focusing on text data, i.e. subreddit post title text, as well as subreddit post text.
    - The main metric of focus will be the maximizing the precision score which directly reduces the percentage of False Positives. (True Positives/Predicted Positives)

Data Scraping

My Nutrition Clinic

- Custom Class to scrape posts from ‘r/keto’ (3M) and ‘r/zerocarb’(120k)
  - Conditions:
  - Not video
  - Not image
  - Not removed
  - Not deleted
  - Not empty

- 10, 000 rows of data (with conditions)

- ‘r/zerocarb’ ran out of posts after pulling 9, 935 posts

Data Cleaning

My Nutrition Clinic

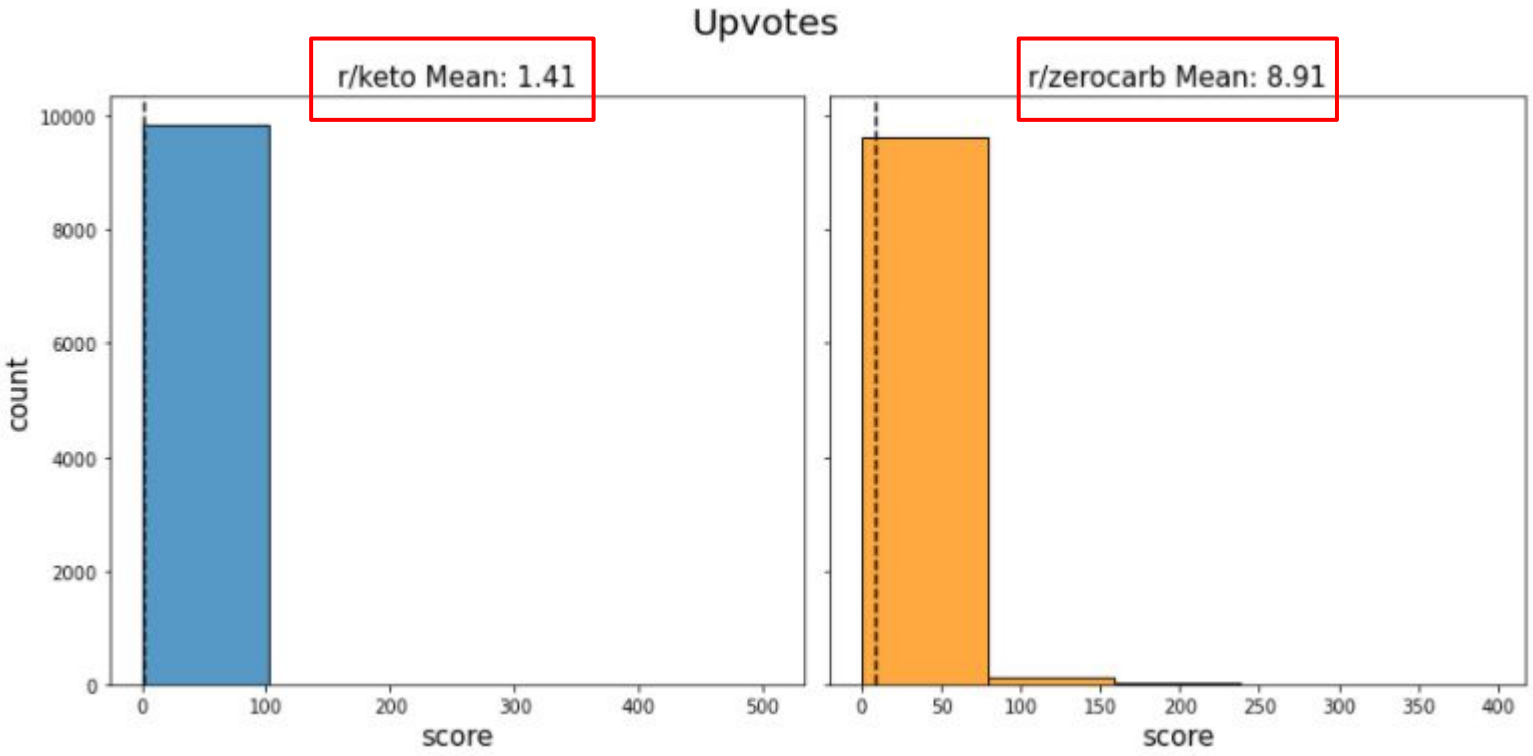
- Custom Class resulted in
  - No null or missing values
  - Only text posts

- 2 custom functions for text pre-processing
  - Remove URLS
  - Remove special characters (e.g. :P)

EDA

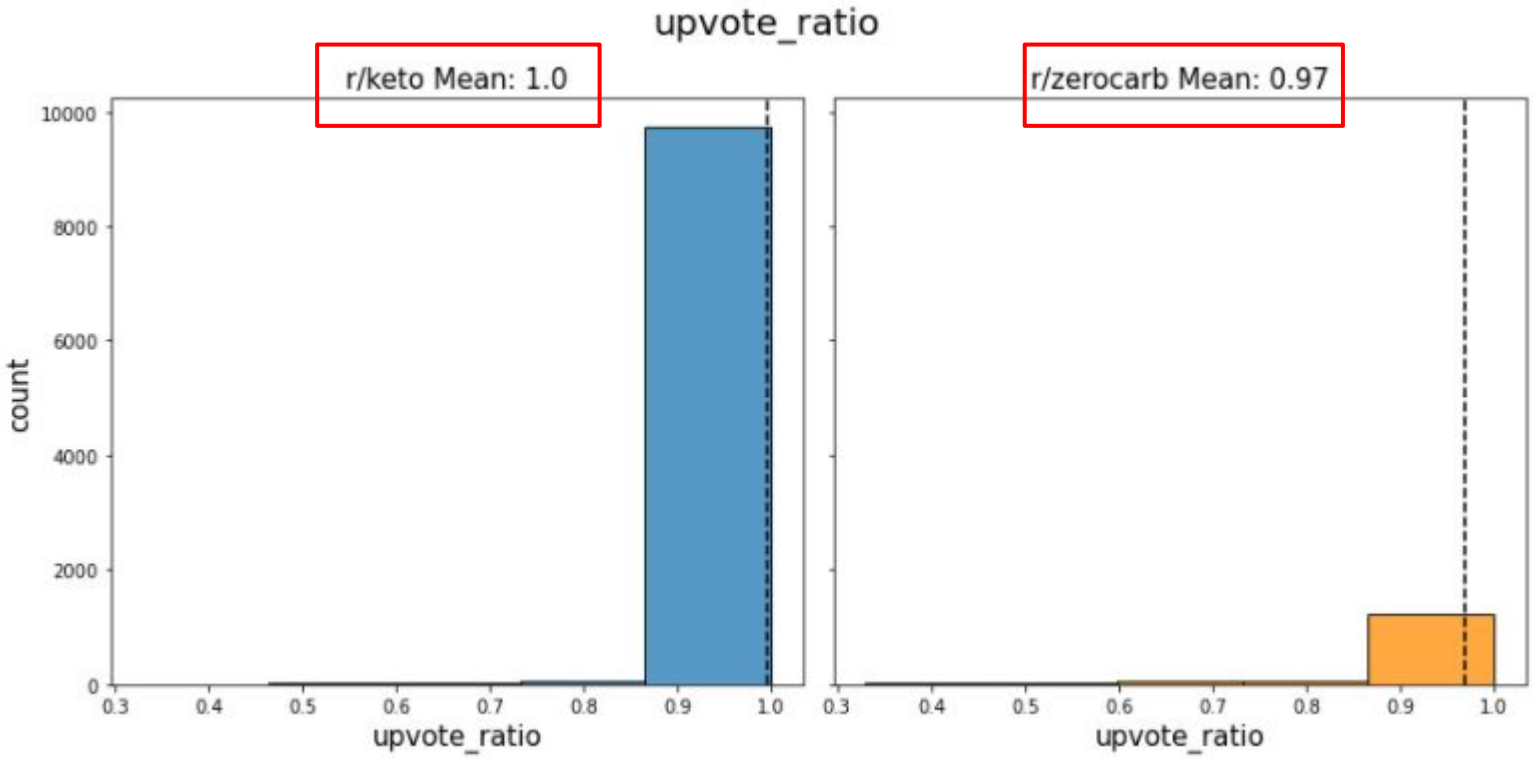
Data Cleaning and EDA

My Nutrition Clinic

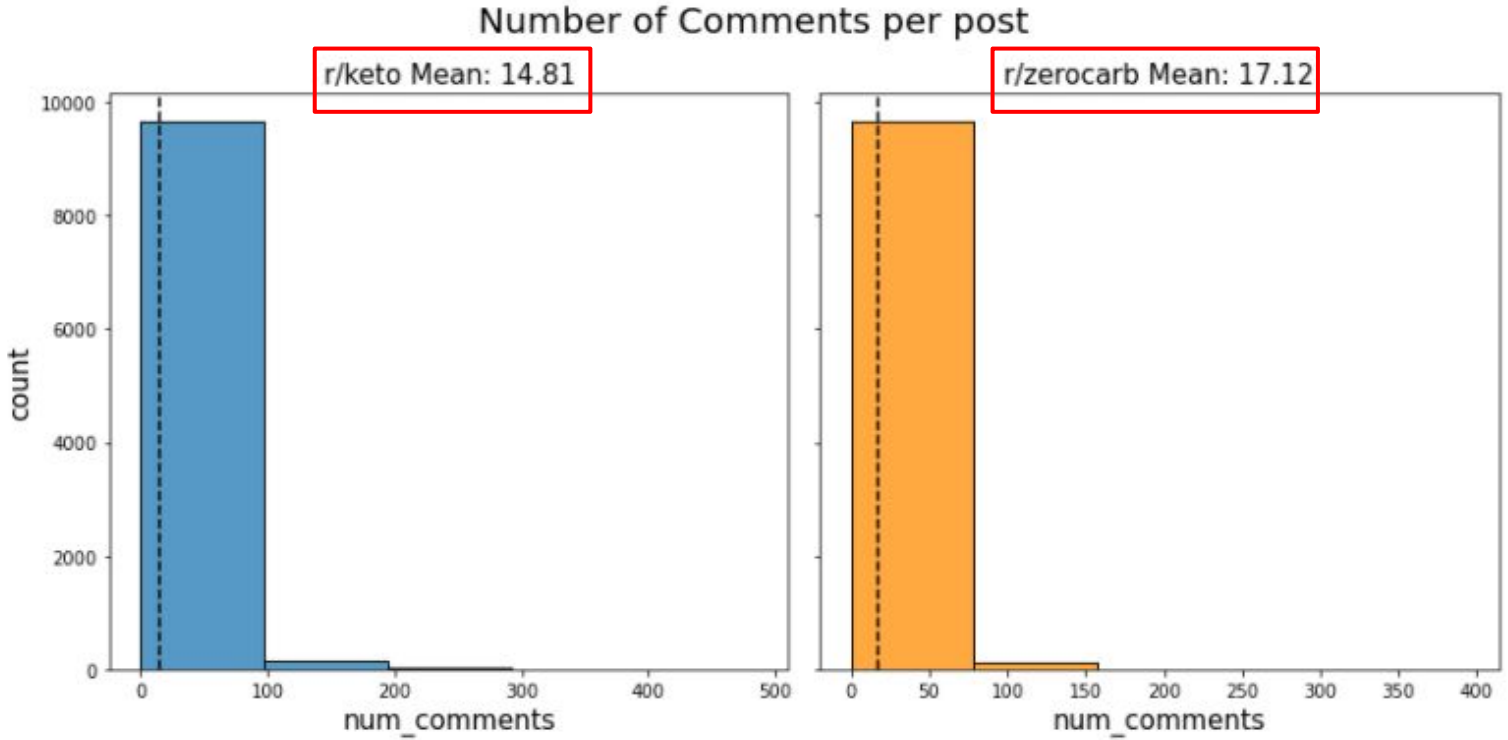


Data Cleaning and EDA

My Nutrition Clinic

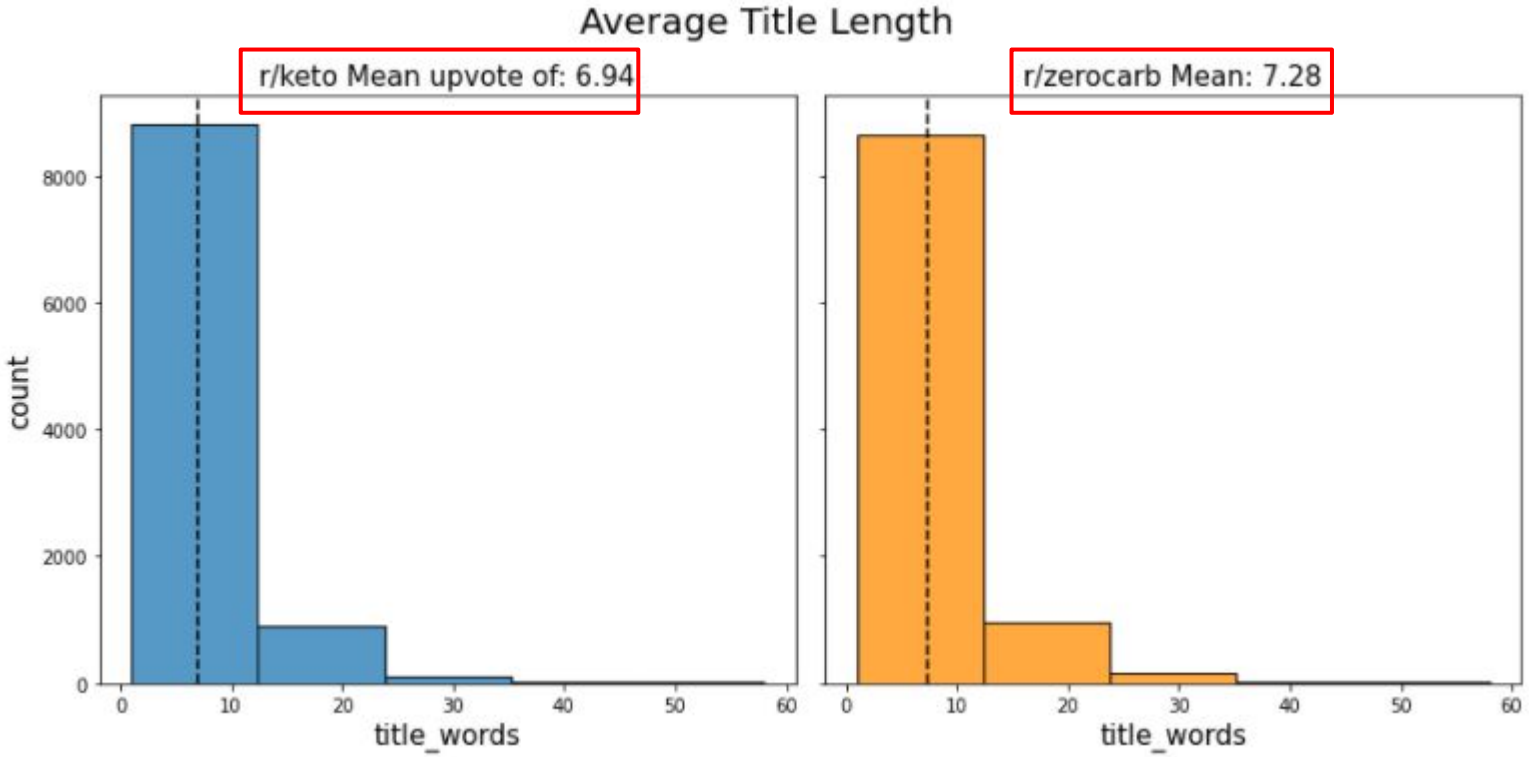


Data Cleaning and EDA



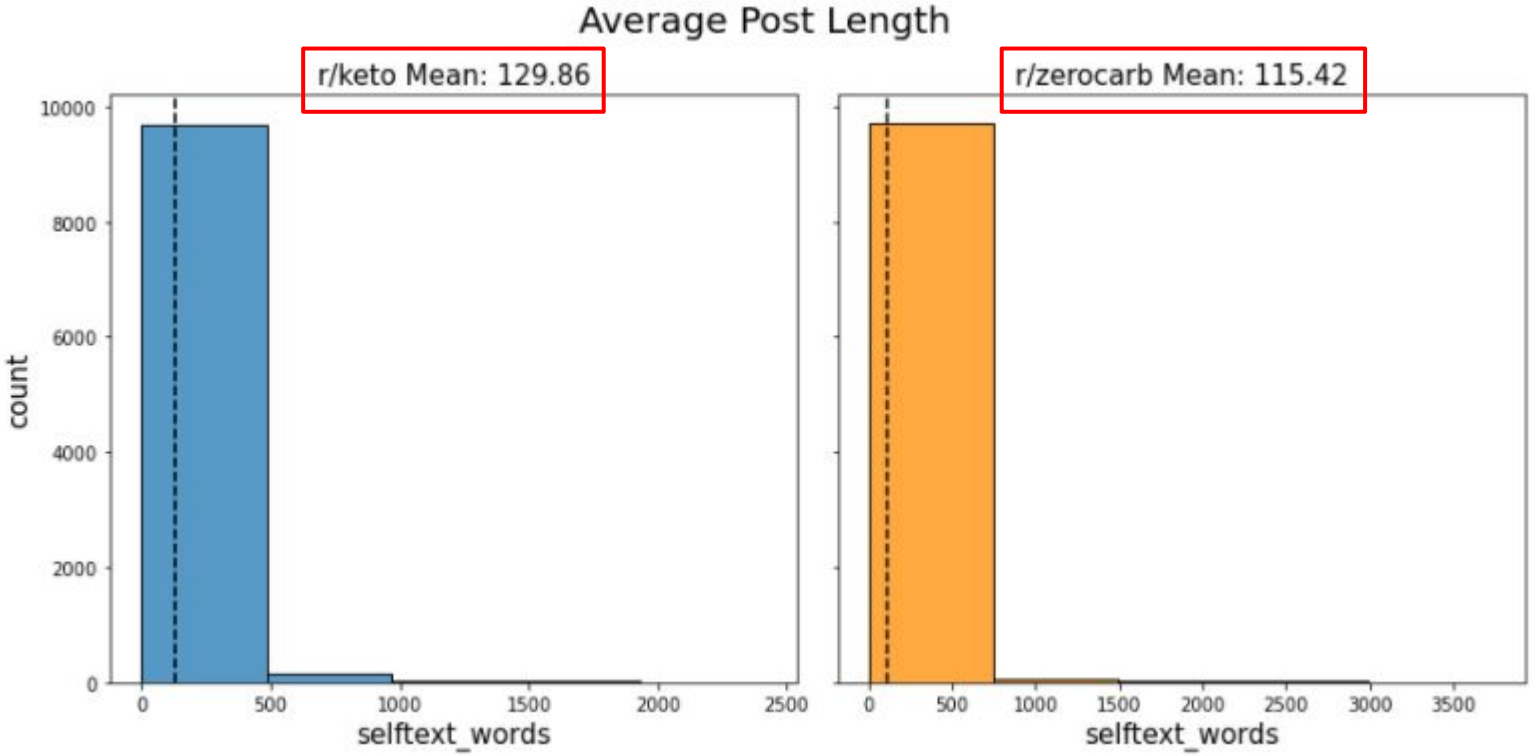
Data Cleaning and EDA

My Nutrition Clinic



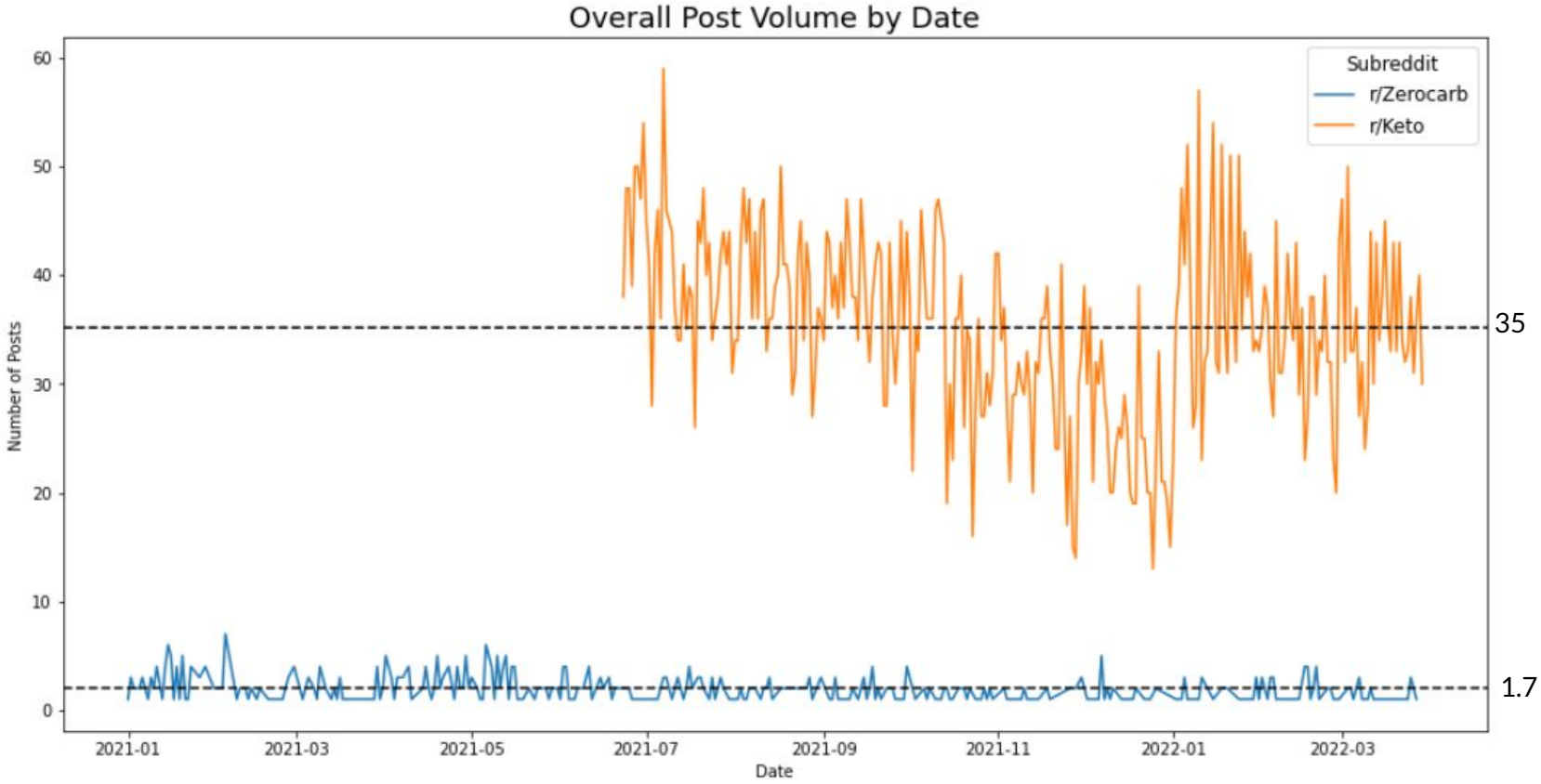
Data Cleaning and EDA

My Nutrition Clinic



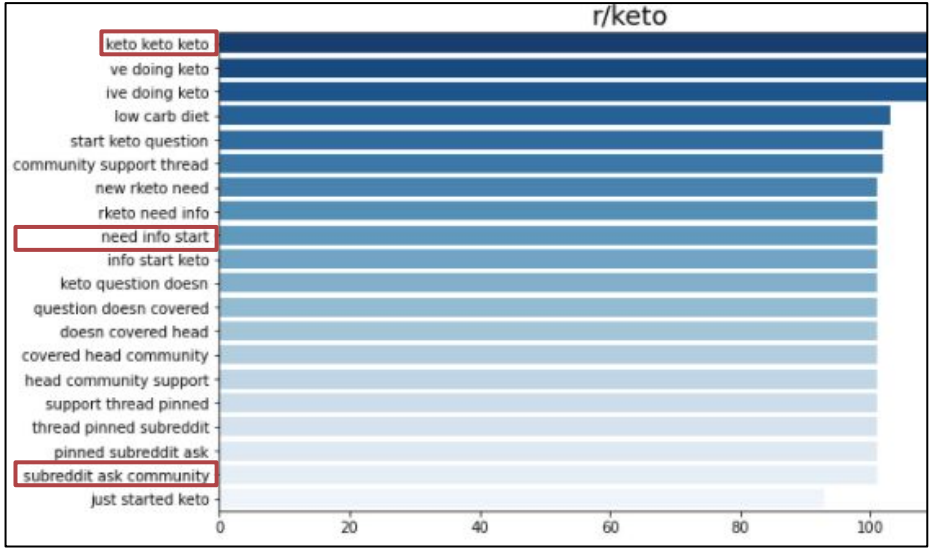
Data Cleaning and EDA

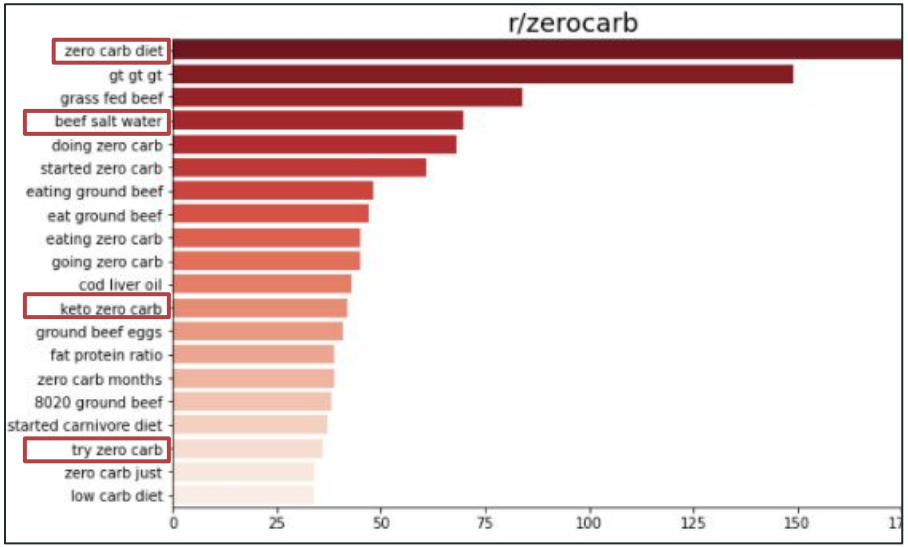
My Nutrition Clinic



Top Trigrams

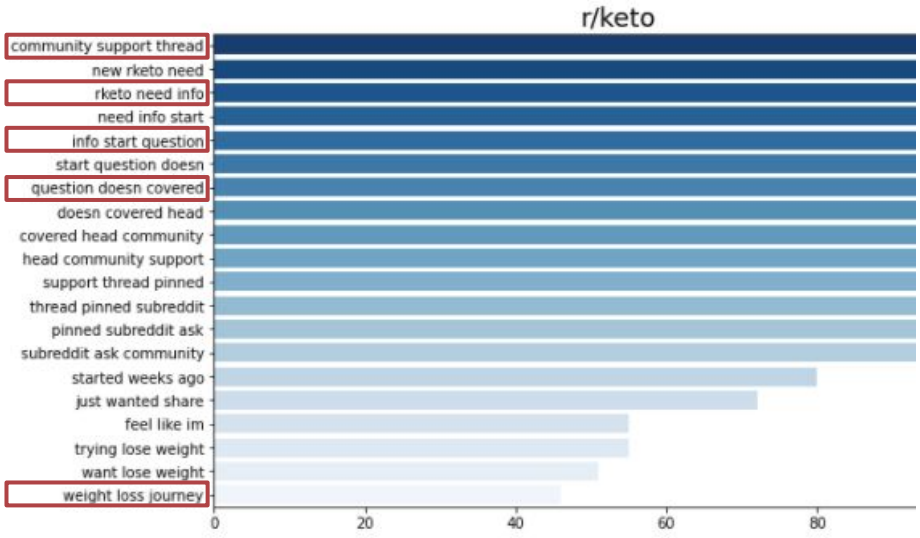
My Nutrition Clinic

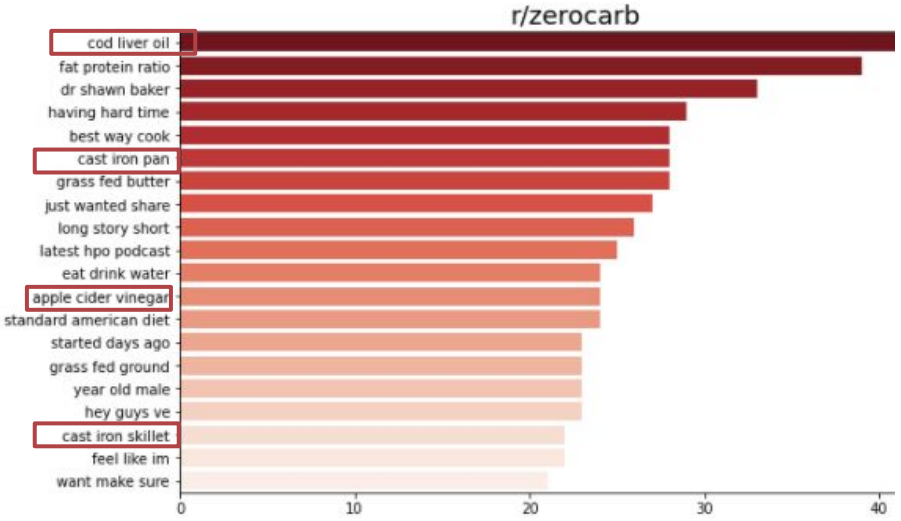




Top Trigrams - Custom Stop Words

My Nutrition Clinic





Modeling

Modeling

My Nutrition Clinic

- 3 different models were picked for this process:
  - Logistic Regression
  - Multinomial Naive Bayes
  - Random Forest Trees

	model_name	model	vectorizer	test_score	precision	is_tuned
3	tvec_nb	nb	tvec	0.827402	0.873543	False
6	tvec_lr_gs	lr	tvec	0.847229	0.859408	True
1	tvec_lr	lr	tvec	0.847483	0.856096	False
2	cvec_nb	nb	cvec	0.830452	0.845251	False
9	tvec_nb_gs	nb	tvec	0.838332	0.843799	True

Final Model

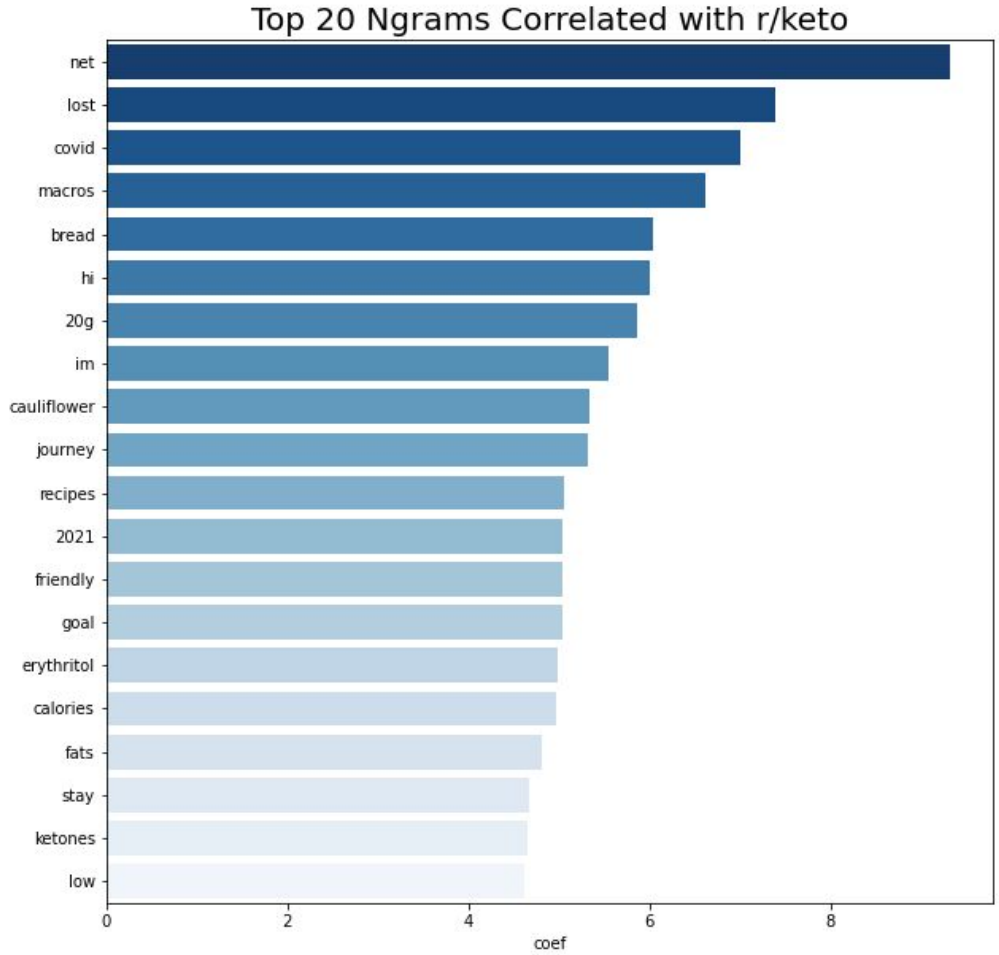
	model_name	model	vectorizer	test_score	precision	is_tuned
17	tvec_lr_gs_V6	lr	tvec	0.840366	0.85471	True
18	tvec_lr_gs_V7	lr	tvec	0.840366	0.85471	True
19	baseline_model	nb	tvec	0.813670	0.86547	False

- Penalty Type: l2
- Inverse Regularization Strength: 6
- Max_features: None
- Max_df: 0.3
- Min_df: 2
- Ngram_range: (1, 2)

Error Analysis

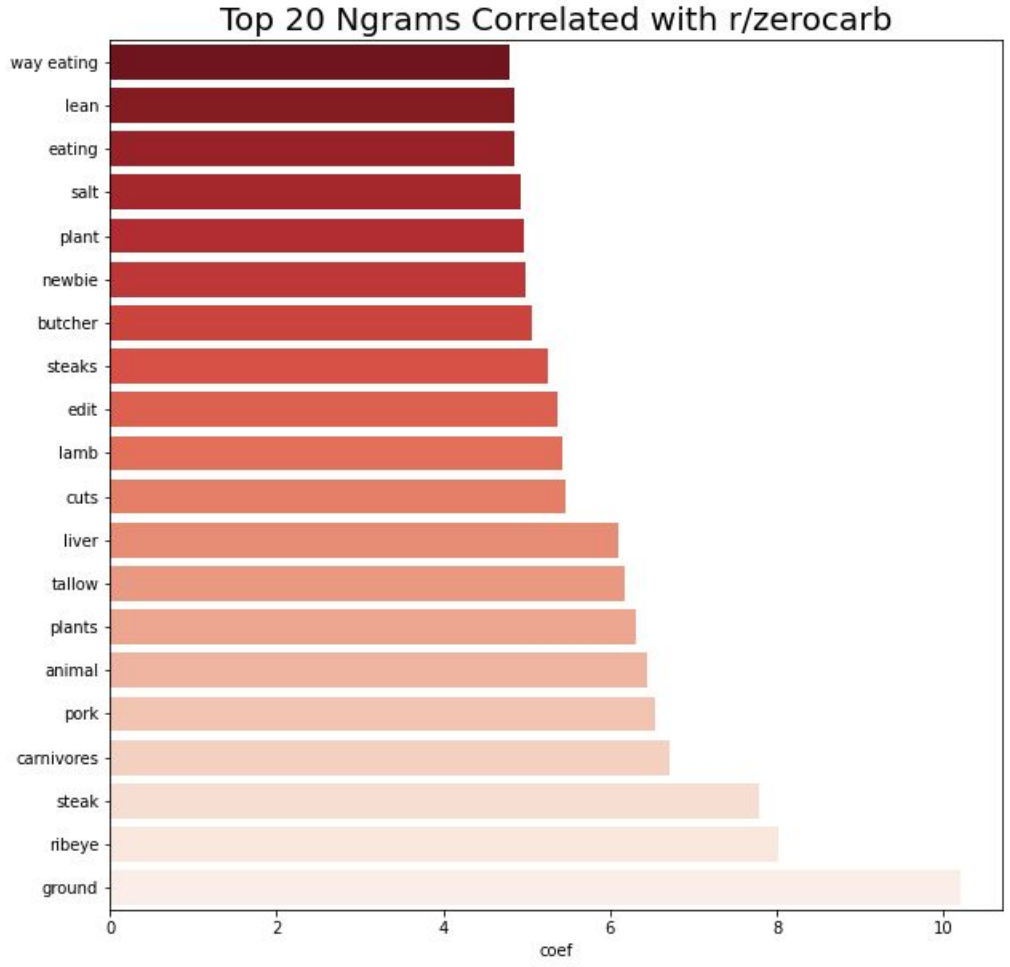
r/keto

- Diet Terminology
- Goals
- Weight Loss
- Progress
- Covid



r/keto

- Ground Beef
- Animal
- Pork
- Organs
- Newbie
- Way Eating
- Cuts
- Butcher
- 30 Days



Limitations, Improvements and Recommendations

Model Limitations

My Nutrition Clinic

- Goal: Reduce false positives (i.e. wrongly classify clients as **Zerocarb** to avoid potentially hazardous rebound
- Model grasps specificity of **Zerocarb** well, i.e. animal products, organs, meat etc.
- Similarity between **Keto** and **Zerocarb**
- The model can be thrown off by specific words like ‘bone’, and mentions of ‘discussions of where to buy meat’

Model Improvements

My Nutrition Clinic

- Test other models: Support Vector Machines, deep learning models
- Adjust the threshold for the classification of **Zerocarb**, making this an imbalanced classification.
- Inclusion of other features in our model, other than just text, sentiment analysis for example

Model Shortcomings and Improvements.

My Nutrition Clinic

- Our current model performs 2% better than our baseline model on accuracy.

- Borderline classification e.g. 51% vs 49%, classifiying those as `keto`. False Negative preferred to False Positive

- Review clients put on **Zerocarb** within one month. Might not be too late to move them to `keto` diets and prevent a rebound from happening. (if required)

Thank You!
QnA