Whole-Food Plant-Based vs. Paleo

Identifying the best-performing classification model

by Denise A. Macias

TABLE OF CONTENTS

01

Purpose

02

Data

03

Methodology

04

Best Model Insights

05

Conclusion

01

PURPOSE

Goals

- Lifestyle Eating’s Goal
    - Help people live healthier and happier lives through their diets
    - Build a platform that fosters a supportive community of users for a variety of diets
- Request Goal
    - Roll out the platform with autodetection technology that can detect through a user’s post what diet they’re on
- Phase 1 Goal
    - Identify a classification model that will most accurately detect the diets of the submissions
    - The evaluation metric will be the accuracy scores of the training and datasets

02

DATA

Data Source

- The PlantBasedDiet and Paleo subreddits
- Scraped 5,000 submissions per subreddit through the PushiftAPI
- Features consist of the subreddit name and the title of the submissions

Whole-Food Plant-Based vs. Paleo

- Whole-Food Plant-Based Diet
  - **Focus:** Natural foods that come from plants
  - **Avoid:** Heavily processed and animal-based foods (meat, dairy, eggs & honey)
  - **Main food groups:** Fruits, vegetables, whole grains, legumes
  - **Acceptable foods:** Nuts, seeds, tofu, tempeh, plant-based milks
- Paleolithic Diet
  - **Focus:** Natural foods that were consumed before the Agricultural Revolution (10,000 B.C.) when farming became the primary method of obtaining food
  - **Avoid:** Processed foods, dairy, grains, legumes & carbs that don't come from fruits or vegetables
  - **Main food groups:** Lean meats, fish, fruits, vegetables and nuts

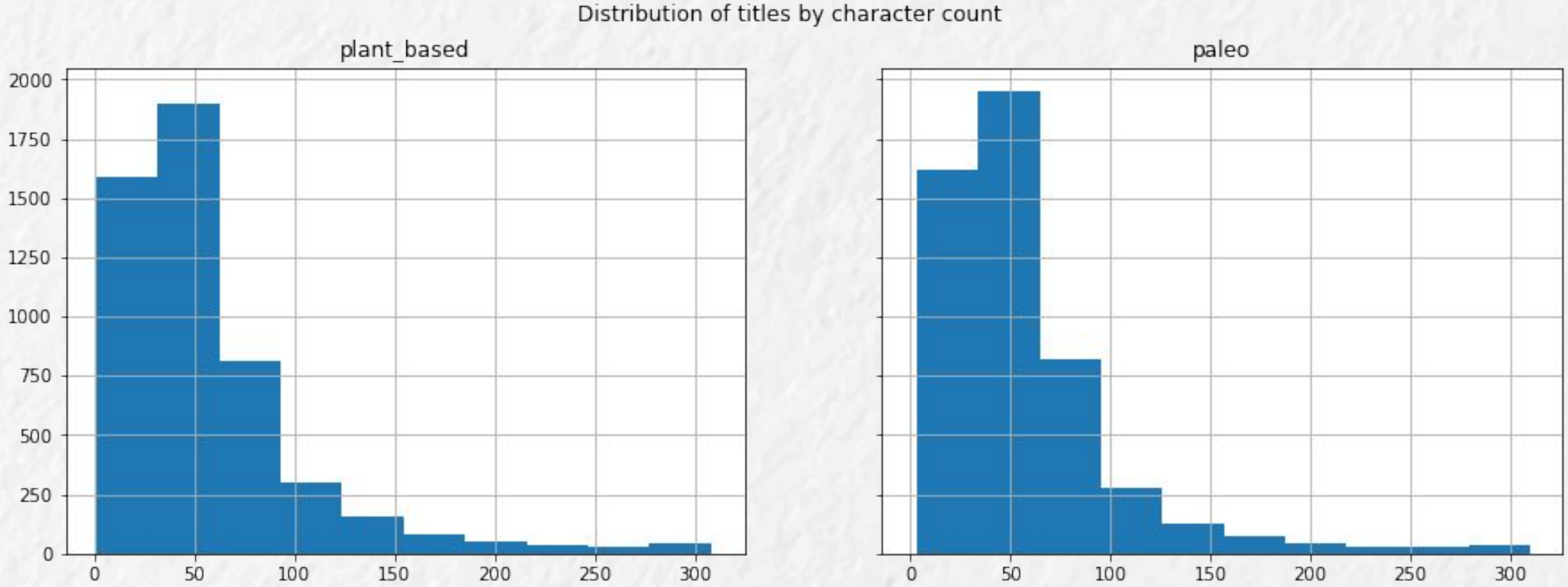
03

METHODOLOGY

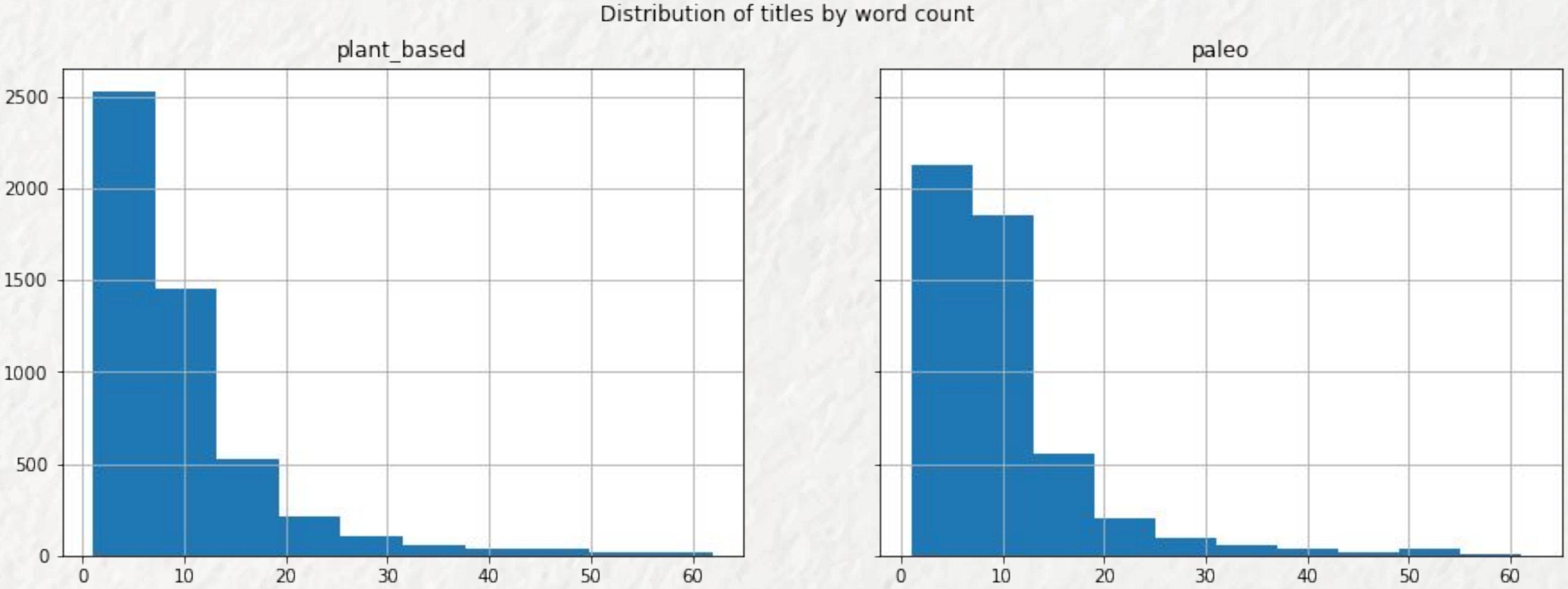
Data Cleaning

- Dropped the selftext and created_utc features
- Removed URLs, digits, punctuation, special characters and emojis from the titles

Exploratory Data Analysis
Title Character Counts

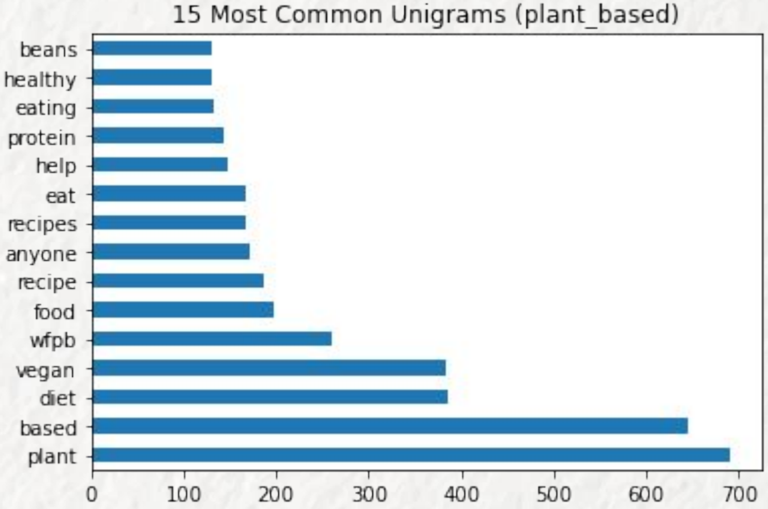


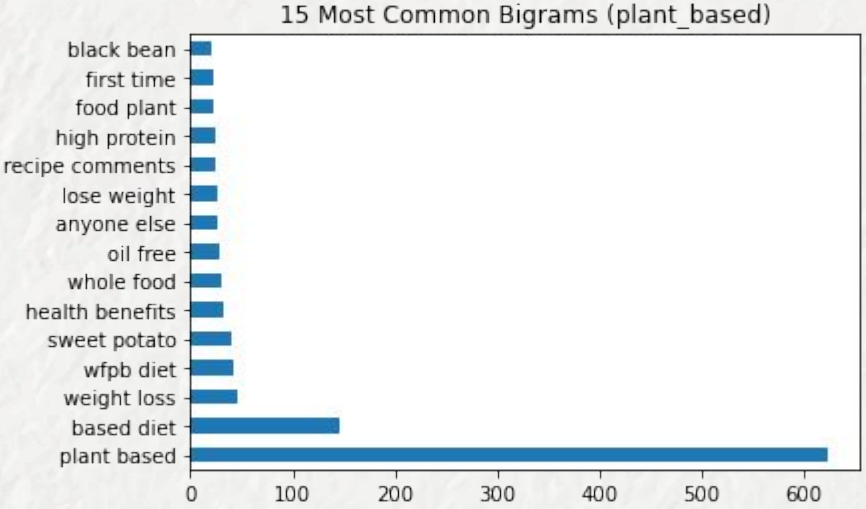
Exploratory Data Analysis
Title Word Counts



Exploratory Data Analysis

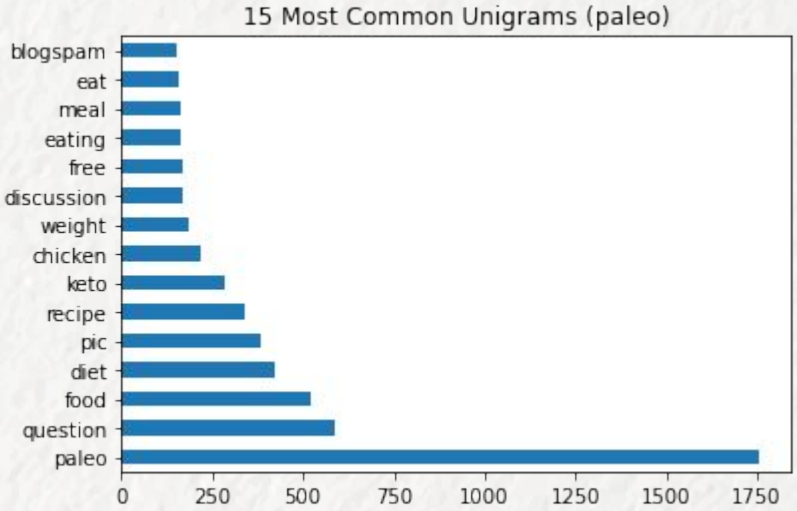
Plant-Based Unigrams & Bigrams

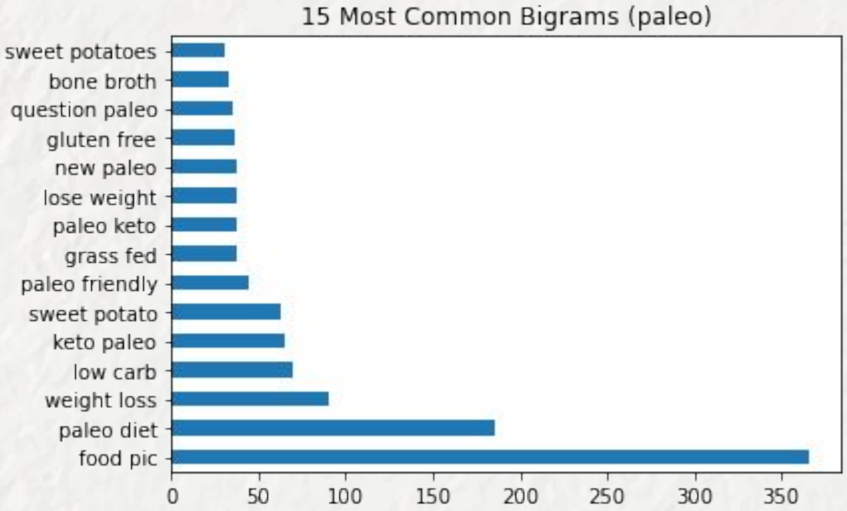




Exploratory Data Analysis

Paleo Unigrams & Bigrams





Model Exploration

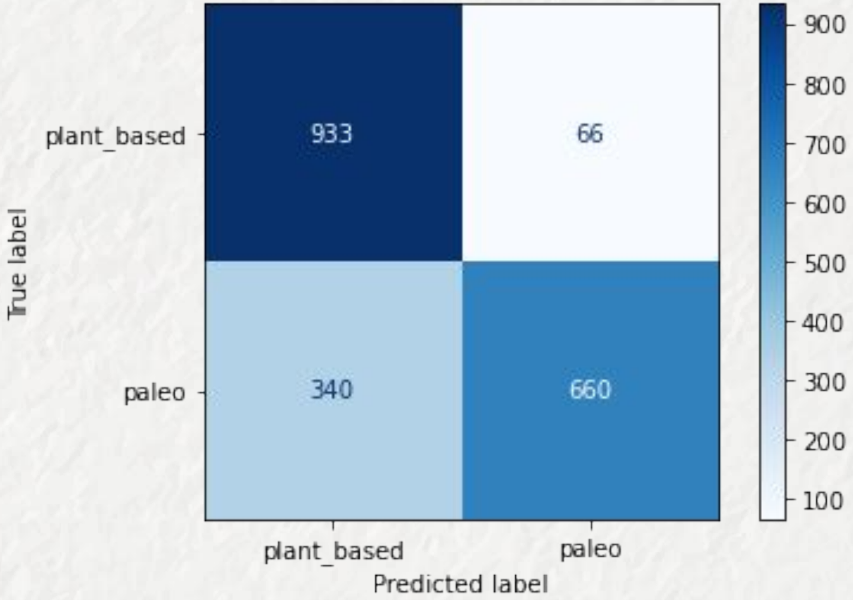
	Random Forest	Logistic Regression	k-Nearest Neighbors	AdaBoost	Gradient Boosting	XGBoost	SVC	Bernoulli NB	Multinomial NB
Training	0.9897	0.9014	0.6488	0.8038	0.8028	0.8551	0.9691	0.9016	0.9076
Testing	0.8084	0.8204	0.5757	0.7938	0.7873	0.8079	0.8244	0.8119	0.7963
Difference	0.1813	0.081	0.0731	0.01	0.0155	0.0472	0.1447	0.0897	0.1113
Training	0.9019	0.9419	0.8986	0.8164	0.8925	0.8124	0.949	0.9578	0.9667
Testing	0.8009	0.8219	0.6983	0.7953	0.8024	0.7968	0.8189	0.8254	0.8124
Difference	0.101	0.12	0.2874	0.0211	0.0901	0.0156	0.1301	0.1324	0.1543

04

Best Model Insights

Insights

- True negative (plant-based) accuracy: 93%
- True positive (paleo) accuracy: 66%
- **Optimizes for true negatives**



05

CONCLUSION

Recommendations & Next Steps

- Keep the XGBoost classification model in mind as a best-performer in subsequent phases
- With the plant-based and paleo datasets:
  - Identify words that could be causing the paleo submissions to be mistaken for plant-based submissions
  - Pull more submissions from each subreddit and run the model on larger datasets
- Try the model on other diet datasets and evaluate performance

THANKS!

Questions?

Denise A. Macias
Data Science Candidate @ General Assembly
deniseamacias1@gmail.com

CREDITS: This presentation template was created by **Slidesgo,** including icons by **Flaticon**, and infographics & images by **Freepik**

Please keep this slide for attribution