The Diet Chase

By Hung Boon, Kenny, Stanven

General Info

- Different types of diets has been rising in popularity over the past years
- Healthy eating coupled with balanced lifestyle
- Norm that is commonly promoted as a foundation to a successful, healthy and happy life
- Increasingly important for Food Manufacturers and Dieticians to understand market interests

About Us

NutriYes

Core Business: Meal subscription service

- Cooked and packed meals
- Raw Ingredients w/ recipe

Objective

- Classify any texts into different categories of diet using classification models that can achieve minimum 95% accuracy in prediction
- Develop menu to capture a wider audience
- Extract insights to better engage customers

How can we process any food or diet related texts?

Are there valuable insights that we can retrieve from the words?

Data Collection & Cleaning

Data Source:
Reddit

Subreddit:

- Keto
- Vegan

Data Cleaning:

- Kept NSFW
- Kept numbers
- Remove Moderator posts
- Remove Duplicates
- Remove Links
- Remove generic words

Model Selection

Data Collection	Data Cleaning EDA	Data Preprocessing	Modelling	Production Model Evaluation
Extract posts from r/vegan and r/keto subreddits using reddit's API	Study unstructured data formats Remove duplicated, advertisement, moderators posts Identify key information columns	Concat key columns for Preprocessing Simplify word features with lemmatizer, Porter Stemmer and Stop Words libraries	Setup training and testing splits dataset Explore Count & TF-IDF vectorizing Explore Naive Bayes, KNN and Logistic Regression Models Construct & Run Gridsearch pipelines to optimize hyperparameters	Compare accuracy & variance results of transformers-estimator combinations. Determine Production model and parameters Review Production Model metrics (Accuracy, Specificity, Precision, Sensitivity, ROC-AUC) Review Keyword Importance

Model Selection

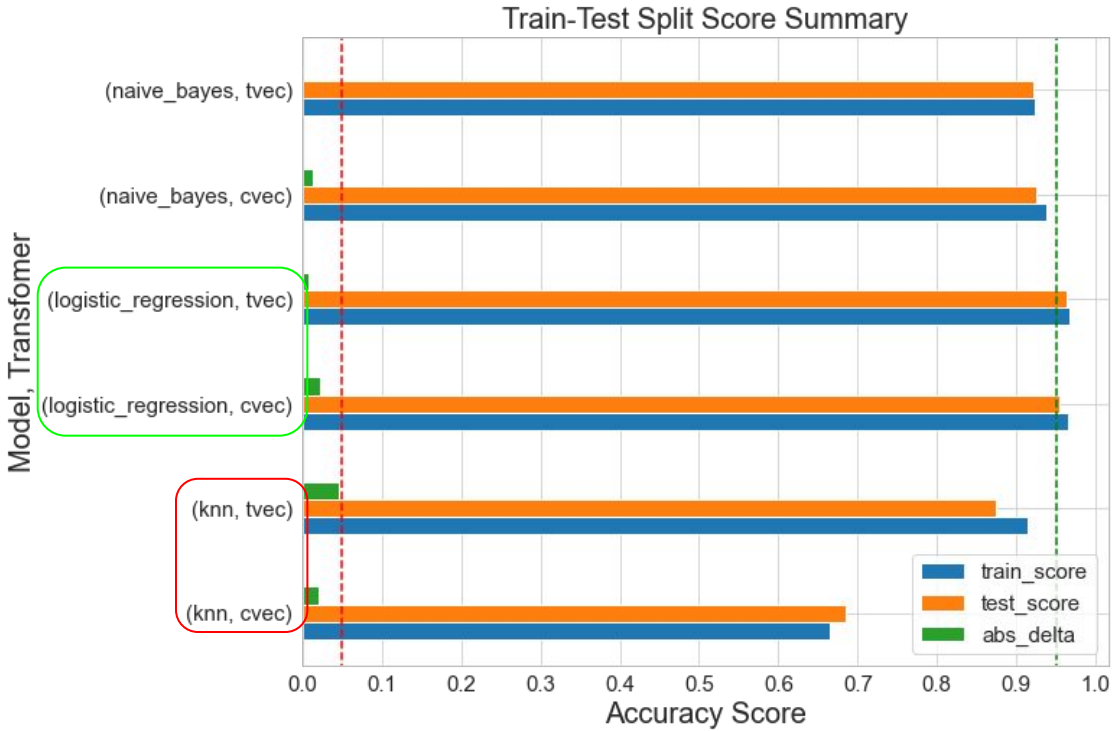
6 combinations of Text transformers and Estimators were explored.

A good model should score high on accuracy (correct classification) and low variance (score difference between different datasets)

All combinations were able to achieve low variance < 1%

TF-IDF word transformer + Logistic Regression combination delivers the highest accuracy of >95%

Production Model
TF-IDF + Logistic Regression



Model Selection

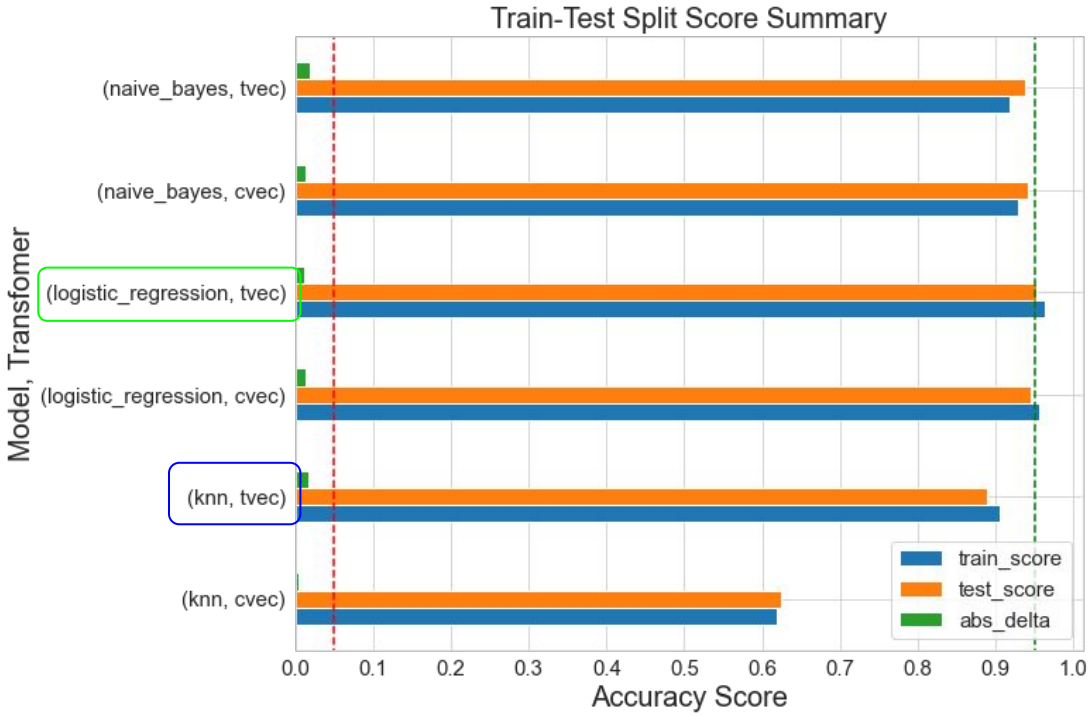
Count vs TF-IDF vectorizing

It is interesting to note that knn with TF-IDF vectorizer scored significantly better close to 0.9 accuracy as compared to Count Vectorizer.

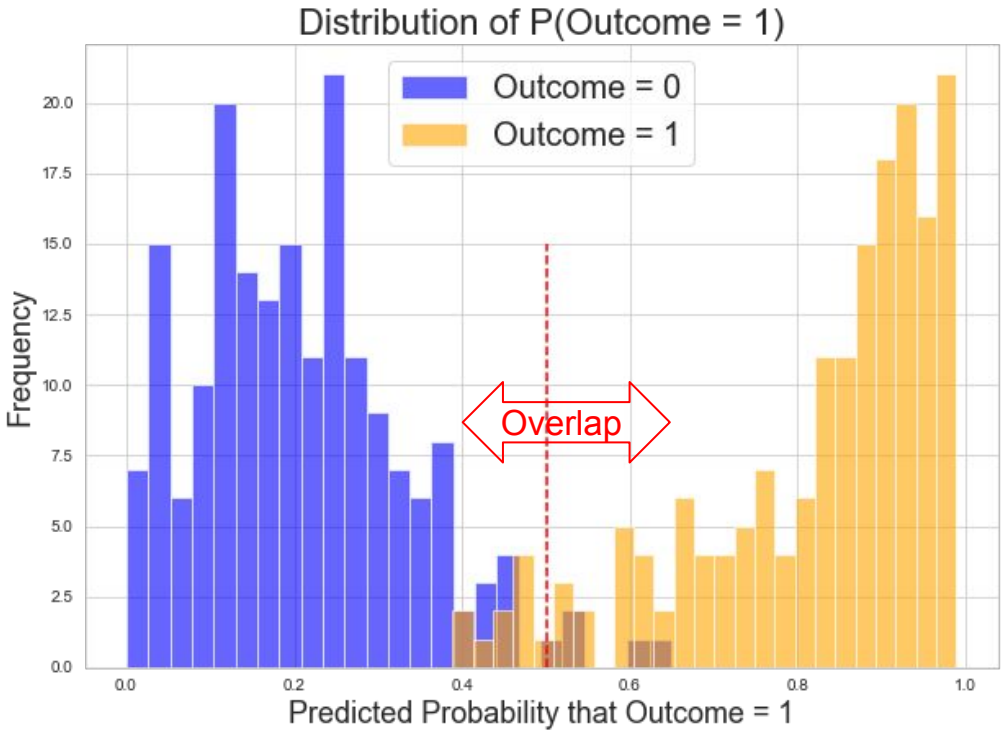
Although we do not notice any significant score improvements from TF-IDF vectorizer on Naive Bayes & Logistics Regression models, we believe that applying TF-IDF should make our production model more robust when classifying unseen posts.

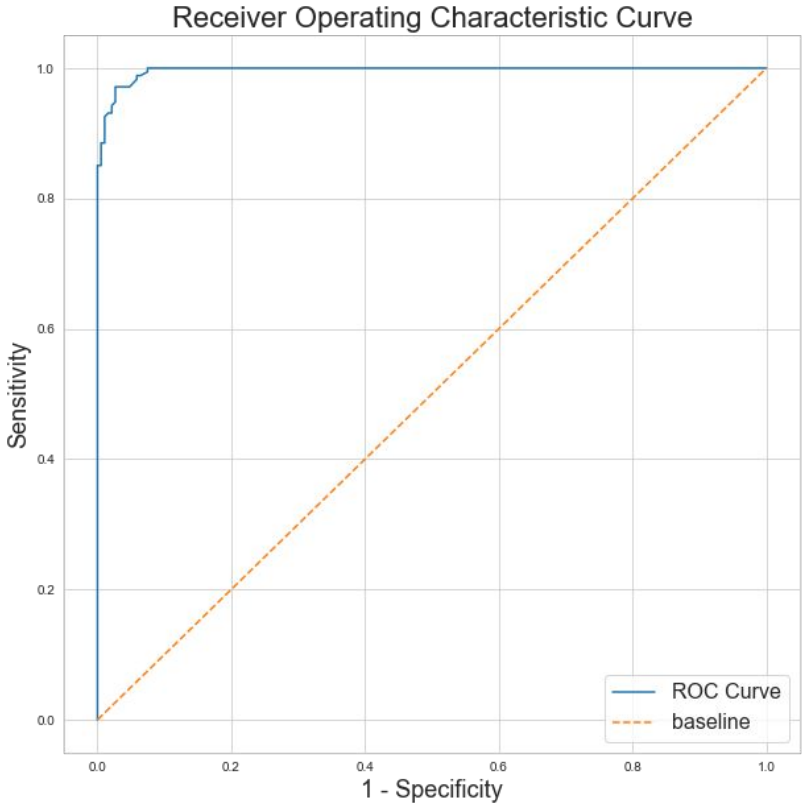
PRODUCTION MODEL

Logistic Regression with TF-IDF vectorizer



Production Model Evaluation (ROC-AUC: 0.9965)



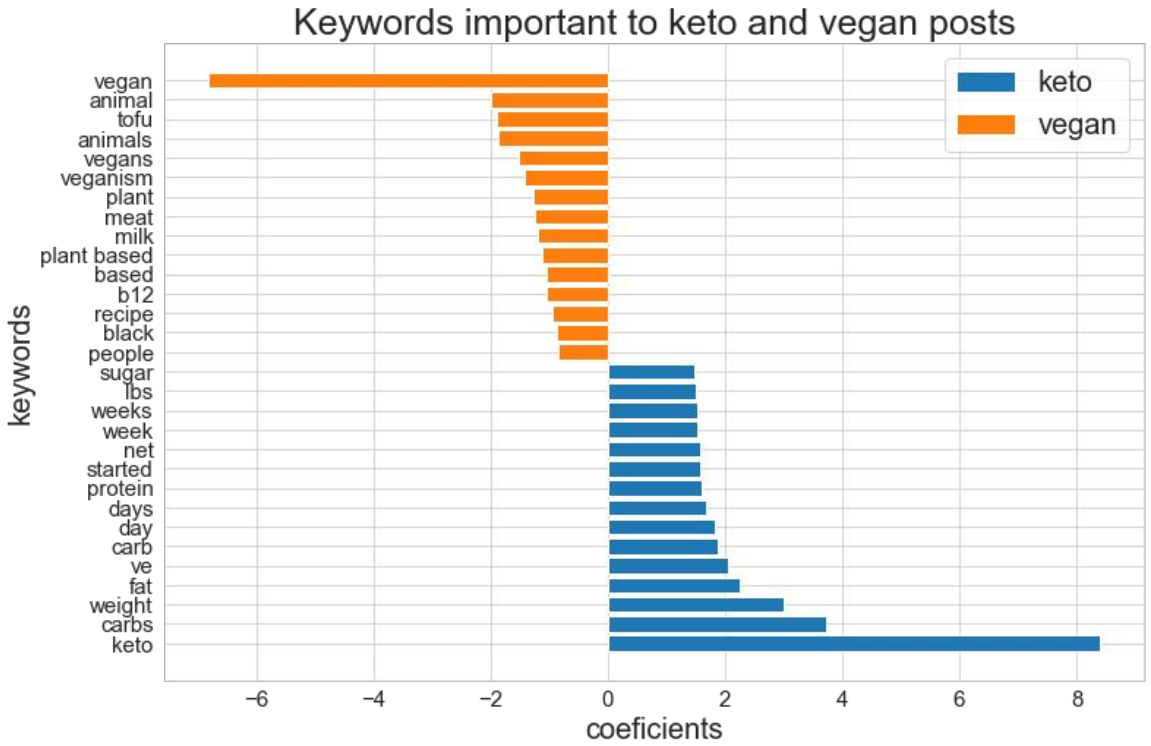


Production Model Evaluation

- Production model is able to classify correctly with > 95% accuracy rate
- Model is specific and precise to both vegan & keto classes > 93% rate
- Combined Type 1 & 2 error rate is < 5%
    - Type 1 error: True vegan post wrongly classified as keto.
    - Type 2 error: True keto post wrongly classified as vegan
- Overlap between Positive and Negative class is < 5% of sample posts
- ROC-AUC score is almost close to perfection of 1 indicating that model is both highly Specific and Sensitive (high degree of separability)

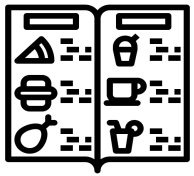
	precision	recall	f1-score	support
vegan	0.95	0.97	0.96	187
keto	0.97	0.94	0.96	174
accuracy			0.96	361
macro avg	0.96	0.96	0.96	361
weighted avg	0.96	0.96	0.96	361

Keyword Importance



- Model is able to generate keyword rankings based on it's regression coefficients
- The more positive coefficient means that keyword is likely keto class related (class 1)
  - As keto word count increases by 1, the post is $e^{8} \cong 3000$ times as likely to be classified into the keto class
- The more negative coefficient mean that keyword is likely to be vegan related (class 0)
  - As vegan word count increases by 1, the post is only $e^{-7} \cong 0.0009$ times as likely to be classified as keto class
- **Important information for product webpage SEO (Search Engine Optimization)**

Conclusions



Strong POC:

- 96% Accuracy

Motivation:

- Vegan: Ethical Reasons
- Keto: Weight Loss

Preference of content:

- Vegan: Practical Advice
- Keto: Success Stories/progress updates

Recommendations

Invest more resources:

- Generalise to identify other dietary trends
- Menu planning

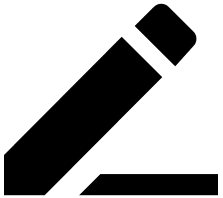
Vegan -> Education:

- Science of the diet
- Updates on change that Veganism has brought
- Tips and Things to avoid

Keto -> Strong Community:

- Feature progress updates
- Tip and recipes
- Free flowing 2 way communication

Next Steps



- Reddit Comments

- Sentiment analysis

Any Questions?