EMSE 6574 Final Project

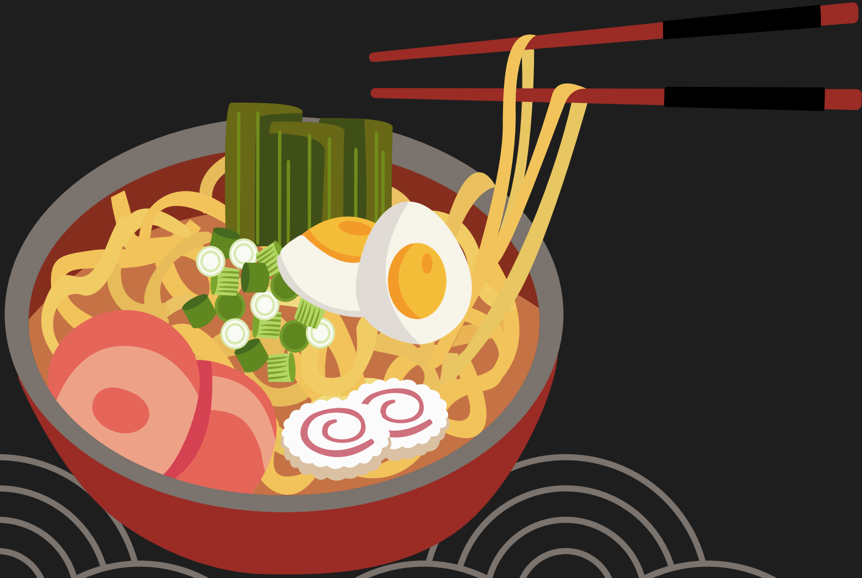
Prediction on Cuisine Type and Calories Count Using Recipe Ingredients

Team Members

- Wisdom Ebirim
- José Garcia
- Michael Salceda
- Kamran Arshad
- Kahang Ngau

01

Introduction



Introduction

- **Goal**

  Created two machine learning models to predict cuisine type and calories based on ingredients

- **Approach**

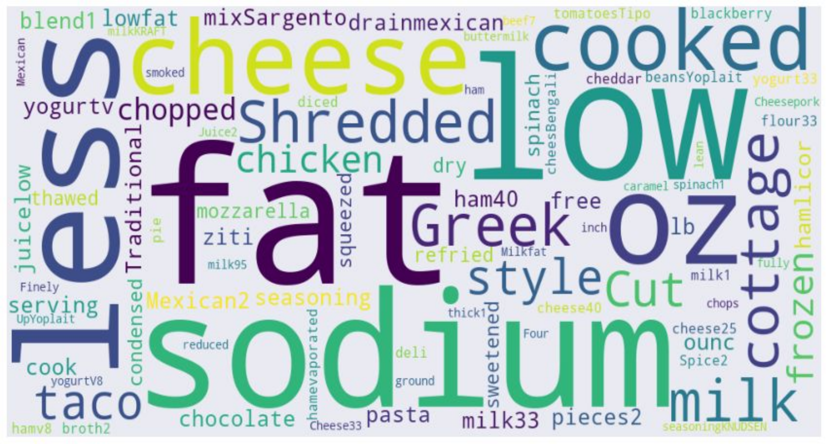
  Seperated the dataset into two parts, Calories model and Cuisine model

  Explored how clean the data is and visualized the basic data statistics (label distribution, ingredient counts, etc.)

  Conducted machine learning techniques and saved the model that can take a list of ingredients and outputs the calories count and cuisine type.

Cuisine Model

- **Dataset**
  - https://www.kaggle.com/c/whats-cooking/data
- **Model Information**
  - The cuisine-type model takes in a list of ingredients, preprocesses the list using the Python spaCy library, and passes it into a scikit-learn pipeline consisting of a term-frequency vectorizer and a support-vector classifier (SVC) model.
- **Output**
  - Accuracy of the model. 79%



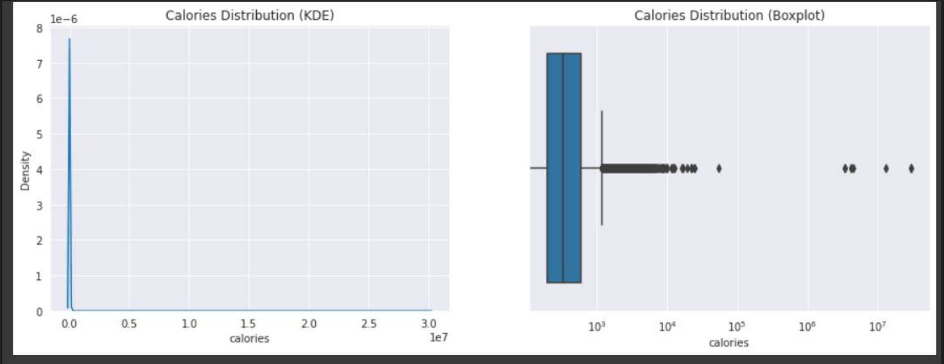
Calories Model

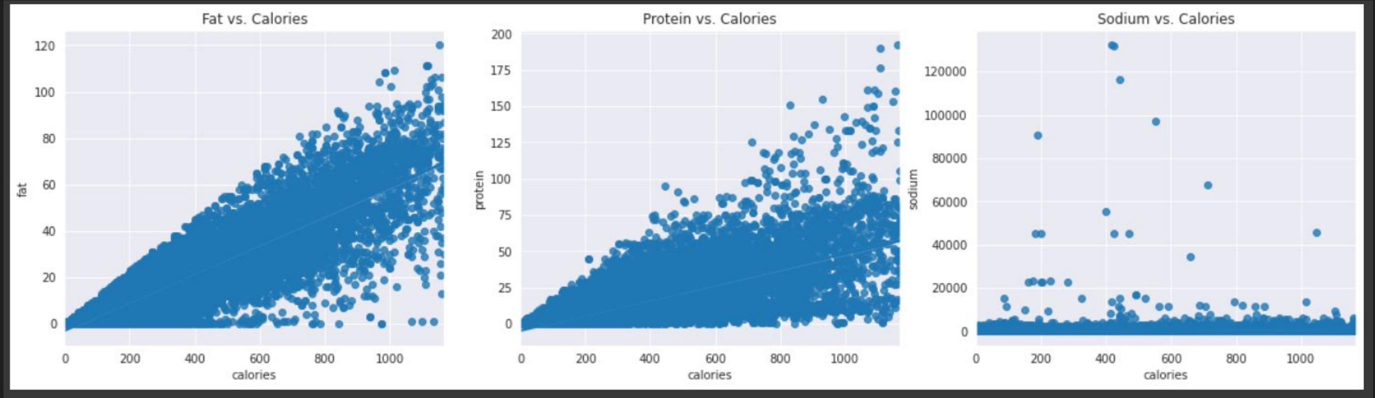
- **Dataset**
  - https://www.kaggle.com/hugodarwood/epirecipes
- **Model Information**
  - For this particular model, we focus on two main columns in this data: the target label (calories) and the ingredients list for each recipe. Because we want this model to use the same inputs as the cuisine-type model, we want to only use ingredients.
- **Output**
  - RMSE: 196.02772585200677
  - Random Forest Model

02

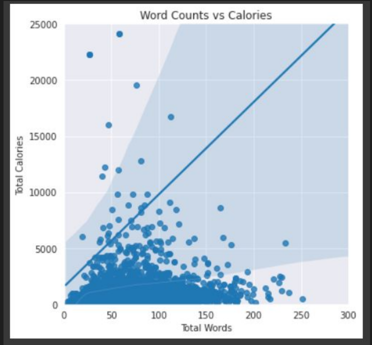
Methodology

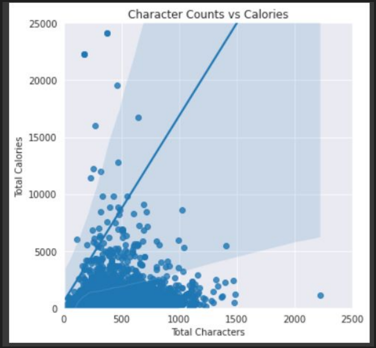
Exploratory Data Analysis - Calories Model





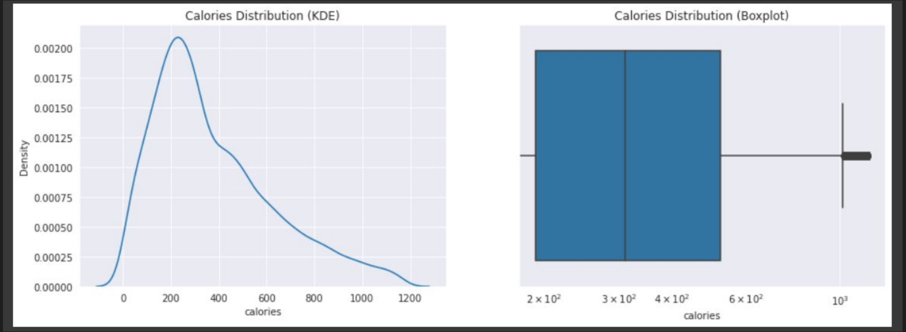
Words vs Characters





Feature Engineering - Calories Model

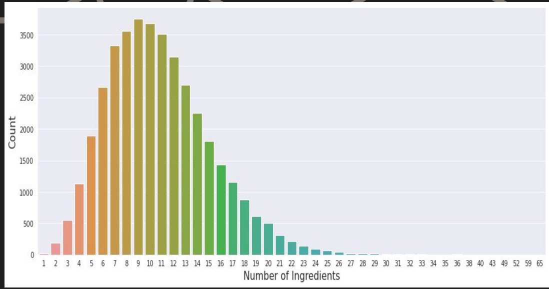
- Removed strange calories values (outliners) from the calories column
- Ingredient column:
  - Punctuation
  - Stop words
  - Lemmatizing



	calories	fat	protein	sodium	ingredients	total_words	total_char	ingredients_processed
0	426.0	7.0	30.0	559.0	4 cups low-sodium vegetable or chicken stock\n...	90	544	4 cup low sodium vegetable chicken stock 1 cup...
1	403.0	23.0	18.0	1439.0	1 1/2 cups whipping cream\n2 medium onions, ch...	123	767	1 1 2 cup whip cream 2 medium onion chop 5 tea...
2	165.0	7.0	6.0	165.0	1 fennel bulb (sometimes called anise), stalks...	39	243	1 fennel bulb call anise stalk discard bulb cu...
3	547.0	32.0	20.0	452.0	1 12-ounce package frozen spinach soufflé, tha...	33	212	1 12 ounce package frozen spinach soufflé thaw...
4	948.0	79.0	19.0	1042.0	2 1/2 cups (lightly packed) fresh basil leaves...	51	331	2 1 2 cup lightly packed fresh basil leave 1 c...

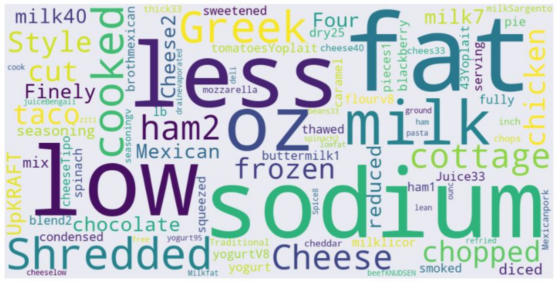
Exploratory Data Analysis - Cuisine Model

- Outliers were identified based in the number of ingredients.



- Recipes with more than 30 ingredients were identified in: Mexican, Indian, and Italian cuisines.

- Weird:
    - Symbols: ® â ç è é í î ú ’ € ™"
    - Ingredients: mi', 'mi', 'v8', 'v8
    - Regionalisms: 'american cheese slices'
    - Upper cases, apostrophes, hyphens, numbers, and units

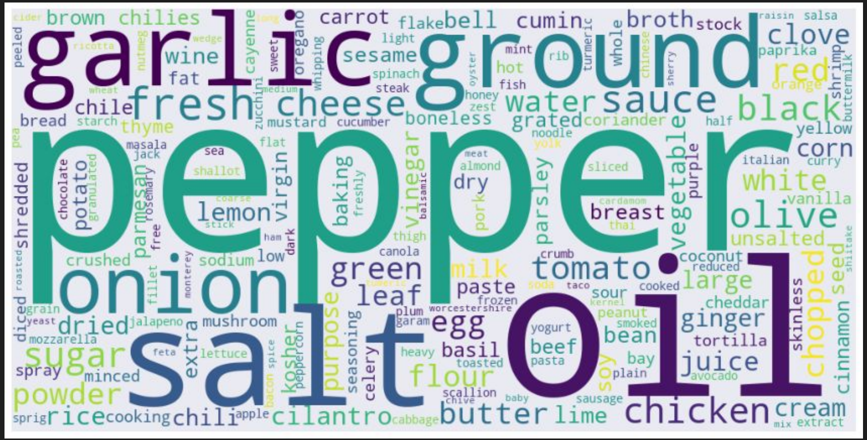


Feature Engineering - Cuisine Model

- Removed outliers, alpha numeric instances, and weird symbols.

- Punctuation, Stop words, Lemmatization.

- Bag of words with Countvectorizer methodology

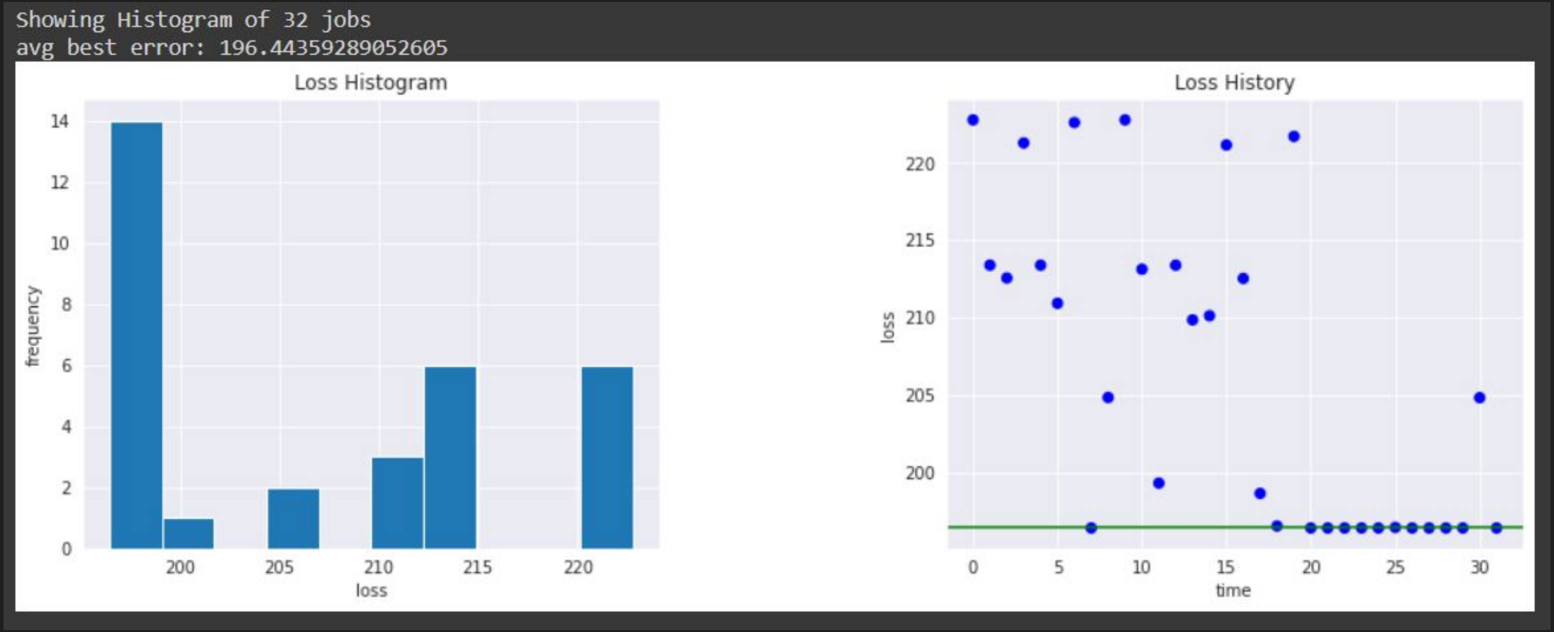


03
Modeling & Results

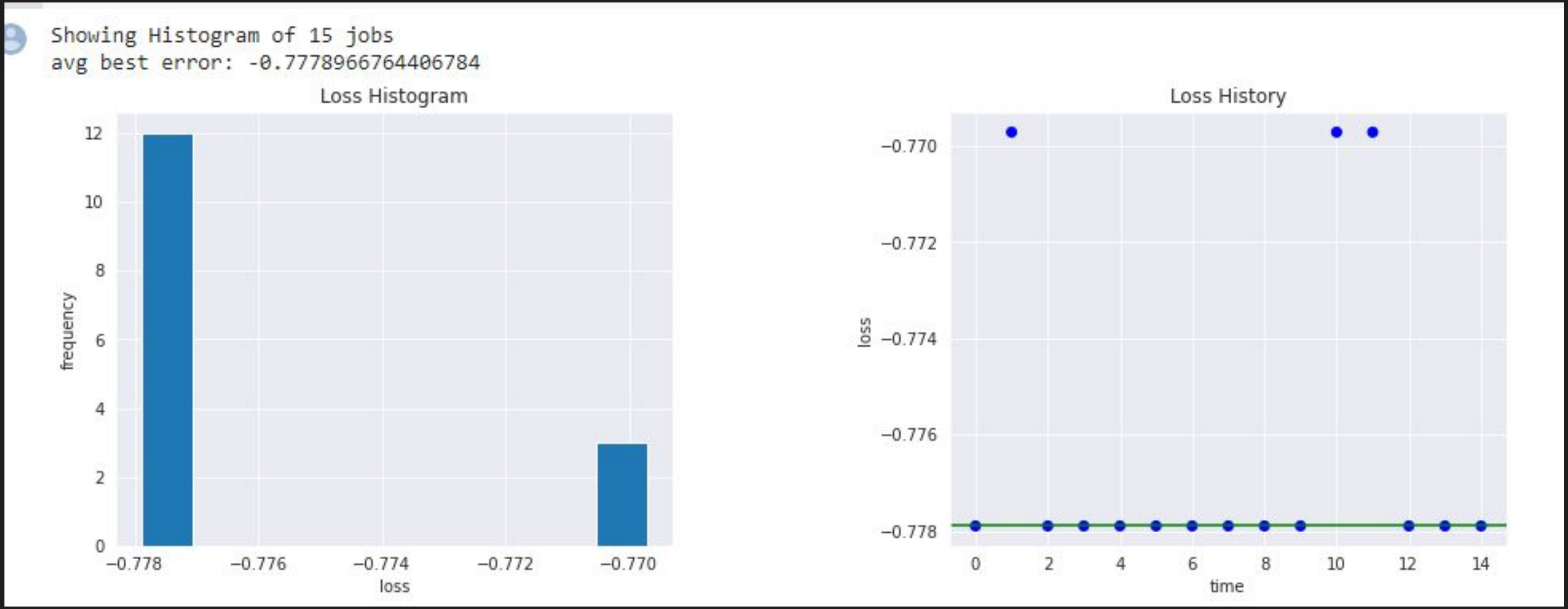
Calories Model Results

	Description	RMSE (Calories)	Avg. Runtime (min)
Baseline	Term-Frequency + Elastic-Net Regressor	256.02	4.60
Iteration 1	Term-Frequency + Random Forest Regressor	196.02	5.53
Iteration 2	TF-IDF + Random Forest Regressor	200.19	5.93
Iteration 3	Term-Frequency + Linear Regressor	218.09	0.39
Iteration 4	Term-Frequency + Passive Aggressive Regressor	210.18	0.07

Random Forest - Final Calories Model



Cuisine Model



04

Conclusion

Conclusion

Calories Model

- The passive regressor, TF-IDF, and Linear Regression models did not perform as well as our Random Forest regressor
- Based on the results, model Random Forest regressor with term frequency did the best, beating our baseline RMSE by 60 (256.02 vs 196.02)
- We then saved this pipeline as a JOBLIB file to be used for a Streamlit app for future prediction

Cuisine Model

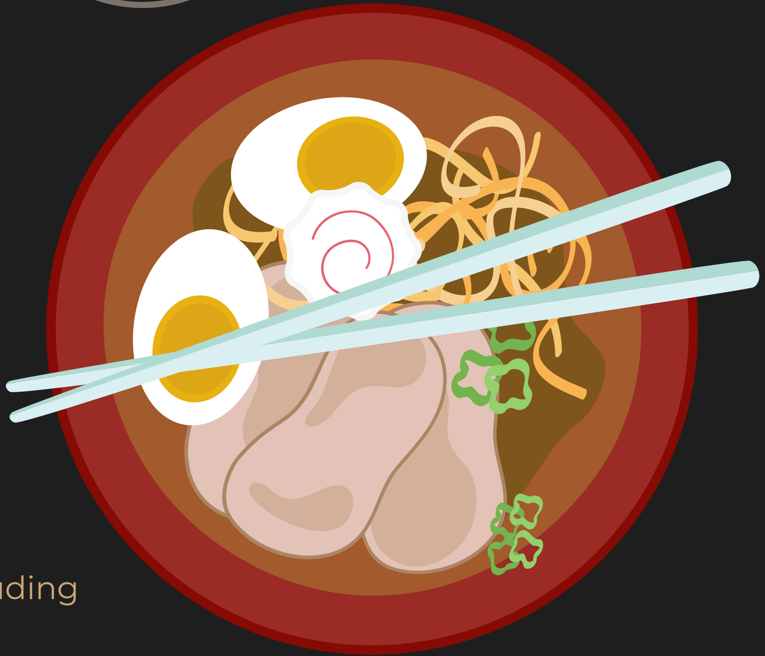
- The cuisine-type model takes in a list of ingredients, preprocesses the list using the Python spaCy library, and passes it into a scikit-learn pipeline consisting of a term-frequency vectorizer and a support-vector classifier (SVC) model.
- The metric we used to evaluate is 'accuracy' with it's highest value of 77% using SVC. After hyper parameters optimization, the accuracy improved to 79 %
- We then saved this pipeline as a JOBLIB file to be used for a Streamlit app for future prediction

Streamlit App

- https://share.streamlit.io/msalceda/emse-6574-final-project/main/final_project_app.py

Thanks!

Any questions?



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