Introduction

Given supply disruptions due to recent global events such as the Covid 19 pandemic and the war in Ukraine, Canadian food prices have risen at an alarming rate of 11% year on year in 2022, putting pressure on the average consumer's budget[1]. This combined with the benefits of home cooking[2] has undoubtedly led to many busy working Canadians to cook more at home. However, assuming a busy work life, many adults need a way to ensure whatever they choose to cook is worth the precious time and effort after work. The recipe classifier seeks to address this issue for busy working adults by classifying recipes as worth the time and effort or not worth it, given the different elements available in online food recipes. Natural Language Processing (NLP) and other techniques were applied to 33,691 food recipes gathered from allrecipes.com to train a logistic regression model, achieving a final accuracy of 76% in determining if a recipe is worth it.

Applications of machine learning on food and recipes are not new. Content based recommendation systems are common[3] and even the recently trending large language models[4] have been used to generate text recipes. This project seeks to use machine learning to identify the factors that make a recipe worth it on allrecipes.com (recipes with average rating 4.5 or more), and to help users verify if existing and new recipes are worth their time and effort or not. Building on top of this project, a recommendation system can be built to suggest new recipes to users and text generators can be used to generate meal plans, grocery lists, and even new recipes based on a user's input and preferences.

Data Collection and Preparation

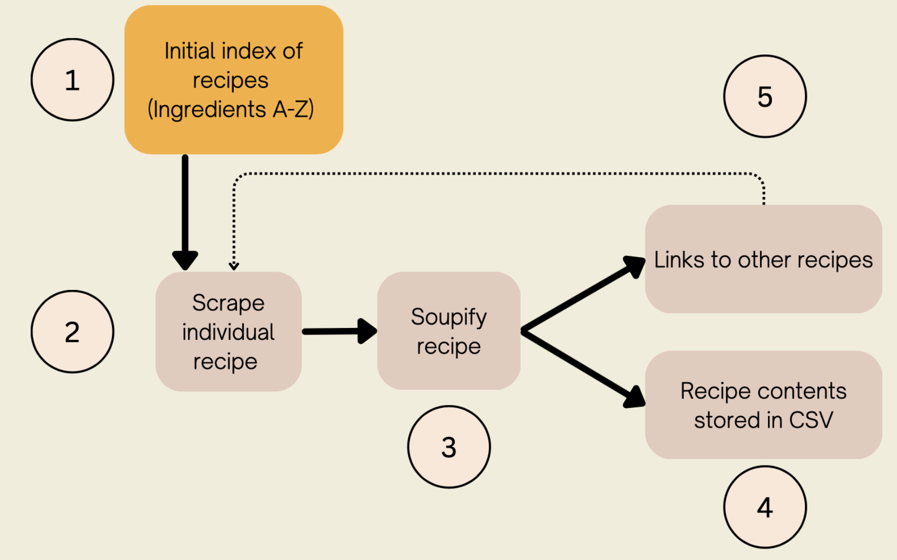
1
Initial index of recipes (Ingredients A-Z)
5
Links to other recipes
2
Scrape individual recipe
Soupify recipe
Recipe contents stored in CSV
3
4

Figure 1: Flowchart of Data Gathering

The data was web-scraped from allrecipes.com, meaning the webpage for each recipe was downloaded and used as data. As there is no complete index listing all the recipes on allrecipes.com, the feedback loop process in Figure 1 was used to iteratively gather data. The data gathering process begins with step 1, where web-links to other recipes were gathered using the 'Ingredients A-Z' landing page. Each of the web-links from step 1 were downloaded in step 2, then converted into a specific format, the BeautifulSoup format,

[1] https://www150.statcan.gc.ca/n1/pub/62f0014m/62f0014m2022014-eng.htm
[2] https://food-guide.canada.ca/en/healthy-eating-recommendations/cook-more-often/
[3] https://foodcombo.com/
[4] https://chef-transformer-chef-transformer-app-cmiy6l.streamlit.app/

in step 3. This conversion was necessary to access specific parts of the recipe website for storage as data and to avoid storing any data not crucial to this project, such as website formatting. The desired contents are then extracted and stored in a comma separated value(CSV) text file in step 4. Finally and most importantly for the feedback loop to work, links to other recipes within the recipe webpage were used to gather more recipes. Steps 2-5 were repeated until sufficient recipes were gathered, which in this case was 40,001 recipes, of which only 33,691 were usable. In terms of data types gathered, numerical data such as nutritional information, textual data such as cooking instructions, and unstructured data such as images were all gathered.

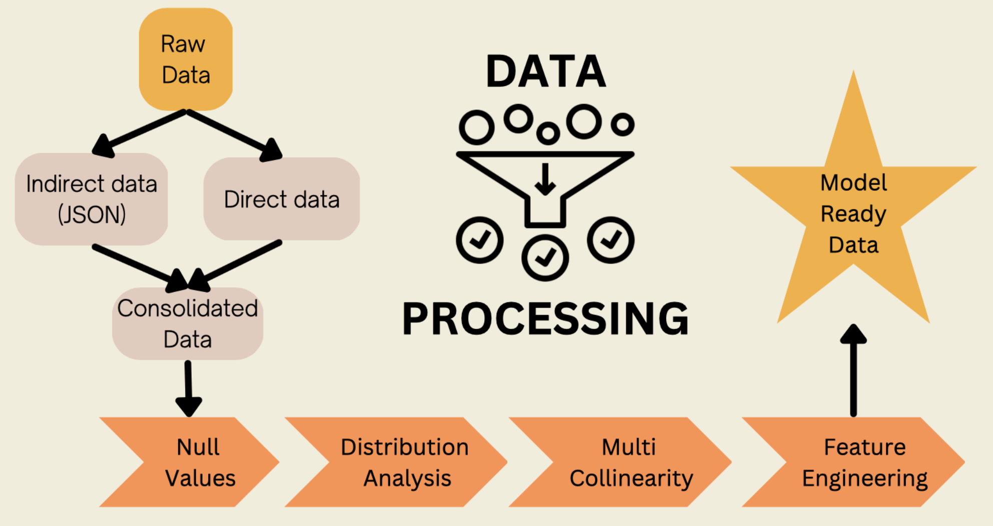

Figure 2: Flowchart of Data Processing

The gathered data, labelled as direct data, was discovered to contain more data packaged within it (JSON dictionaries), which contained overlapping data. For example, the title of a recipe was available in both the direct and indirect data. To avoid the problem of conflicting information from the dual sources, each field of interest was cross examined between direct and indirect data, with the final consolidated data being a standard for future data gathering. The consolidated data was further analysed in terms of null values, outlier values and overall distribution. During this step, outliers were identified but not removed. Instead, outliers and sharp skewages were dealt with by taking the log of the data. Then, columns were cross examined for multicollinearity. This step was essential to prevent a duplicate representation of data when modelling. Then, some features such as ‘word count per instruction’ were created from existing features to reduce the number of columns and enrich its ability to predict the target feature. Finally, the processed data was ready for modelling.

Analysis and Modelling

Many columns were observed to have sharp skewage, as observed in the left subplot in the figure below for sodium content within recipes. The sharp peak at the mean of 575 mg shows that most recipes are between the mean and median of 350 mg. The peak is made sharp due to the existence of outliers, such as recipes that involve brining with salt or pickling recipes. To avoid removing outliers, the log of data was taken resulting in the subplot on the right, where the sharp peak has been mellowed out into a better defined hill where the mean and median are roughly the same. Visually, this makes it easier to identify trends and differentiate recipes and this applies at a mathematical level when training the models.

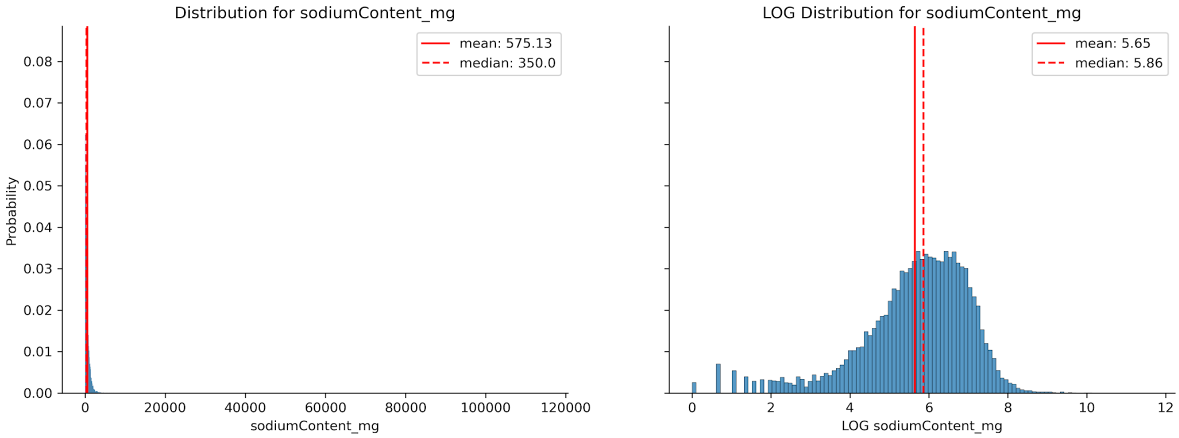

Figure 3: Taking the logarithm to deal with outliers and skewed distributions

Before training the model, the processed data was split into test and remainder sets, where the test set is reserved only for final model evaluation. After the split, null values in the remainder dataset were imputed using a K-Nearest Neighbour imputer and text columns were vectorized using Term Frequency-Inverse Document Frequency (TF-IDF). This step was necessary to process natural language into numerical columns for modelling. Several models were trained and analyzed using accuracy, false positives, and false negatives. After thorough investigation of the factors within the model that affect predictions, ultimately a logistic regression model was chosen as the top factors influencing predictions were a balanced variety of features. Referring to the figure below, the top 5 features affecting a recipe being ‘worth it’ were if the reviews contained the words ‘delicious’, ‘perfect’, ‘love’, if it was a drink recipe and if it contained ‘whole’ ingredients. Further studies are required to understand why these features were important.

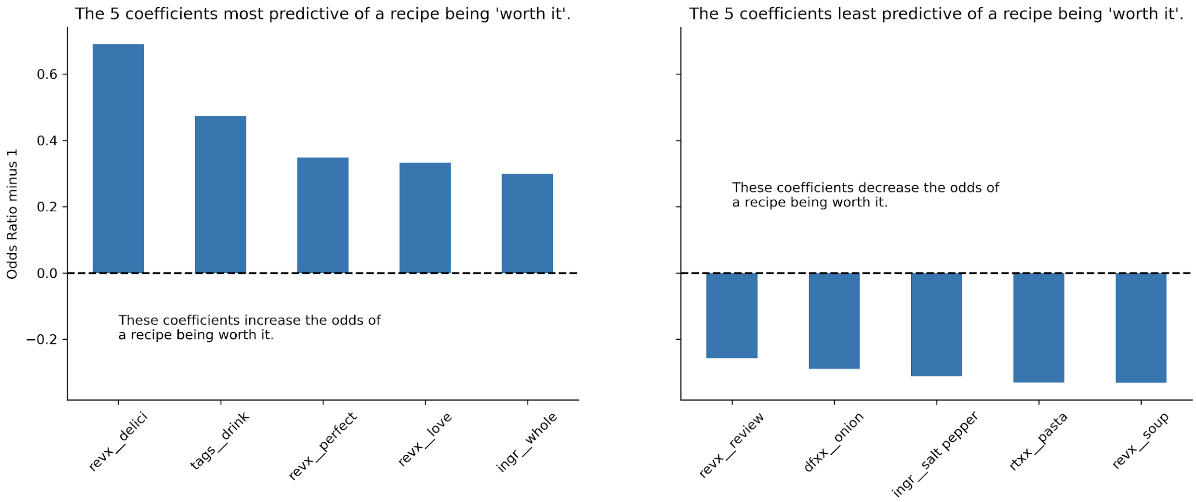

Figure 4: top Coefficients (factors) affecting the final model

Conclusion

Overall, the final model provided a test accuracy of 76%, 18% higher than the baseline of a user guessing all recipes as ‘worth it’. While continuing to improve accuracy, the project has also yielded insight into which elements and what contents in a recipe make a recipe worth it. A recommender system can be built by measuring the similarities between the processed features, which can then be built into a recipe, grocery list, and meal-plan generator in the future, further simplifying home cooking for working adults.