UBC Climate-Friendly Label recipe analysis 2023-2024

Student WorkLearn 2023-2024 Report
April 2024

By: Sharon Marfatia, CFFS Data Analyst
BSc. Computer Science (3rd Year)

Supervised by: Dr. Juan D. Martinez
Applied Research Coordinator, Climate Action & Food Systems
SEEDS Sustainability Program

Cover Photo: UBC Brand and Marketing
The Spruce / Nisanova Studio
UBC Brand and Marketing

Disclaimer: UBC SEEDS Sustainability Program provides students with the opportunity to share the findings of their studies, as well as their opinions, conclusions and recommendations with the UBC community. The reader should bear in mind that this is a student research project and is not an official document of UBC. Furthermore, readers should bear in mind that these reports may not reflect the current status of activities at UBC. We urge you to contact the research persons mentioned in a report or the SEEDS Sustainability Program representative about the current status of the subject matter of a report.

SEEDS Sustainability Program

Table of Contents

Practitioners' Summary

Background

Over the last 12 months, I have worked on various aspects of the data analysis work for CFFS. This report outlines key contributions including adding land utilization metric, re-evaluating baselines for GHG emissions, stressed water, and nitrogen loss. Additionally, I made enhancements in automating the process of assigning GHG categories to data, automating the process of standardizing units for non-standardized food items, I developed a machine learning solution to classifying products, evaluated CFF product labels for UBC Food Services for the 2023-24 Winter Semesters, and successfully evaluated products for the first AMS venture with the UBC CFF Label at "The Gallery" restaurant.

Key Advancements

- Incorporated a land use factor into the environmental impact evaluation of recipes to better reflect their environmental footprint.
- Improved the calculation of CFFS baseline values for the categorization of products.
- Developed automated systems for categorizing GHG emissions categories and standardizing unit measurements, reducing manual work and increasing analysis efficiency.
- Successfully launched the label at "The Gallery" in UBC AMS.
- Participated in Sustainability Week at UBC in 2024 to advocate the message of UBC SEEDS and UBC CFFS and educate students on the label.
- Utilized Natural Language Processing (NLP) techniques to categorize food service data, aiding in the identification of sustainability practices and operational efficiencies.

The Next Stages Moving Forward

- Collaborating with more venues.
- Exploring more ways to increase the precision of the label.
- Finding methods to conduct faster deployments.
- Conducting more student-outreach initiatives to teach people on the benefits of eating with a sustainability focused mindset.

Executive Summary

Research Objectives and Methodology

The core objective of Sharon's work was to refine and extend the environmental impact analysis of food services at UBC. This involved incorporating a new land use metric into the assessment framework, re-evaluating baselines for greenhouse gas (GHG) emissions, stress-weighted water use and embodied nitrogen. A multi-faceted approach was employed, leveraging both existing and dynamic data analysis techniques. Key methodologies included:

- Enhancing environmental impact assessments by integrating land use factors with existing metrics.
- Automating the classification of food items into GHG impact categories and standardizing unit measurements to streamline procurement data analysis.
- Deploying machine learning models, specifically focusing on natural language processing (NLP), to classify and analyze UBC Food Services data, with the objective to use UBC Food Services data to predict CF label assignments at other venues where detailed recipe information is unavailable.

Major Findings

The application of these methodologies yielded several significant findings:

- The introduction of a land use factor provided a more comprehensive reflection of food choices' environmental impacts.
- Automation of data categorization and unit standardization substantially improved the efficiency of data analysis processes.
- The NLP-based machine learning model demonstrated a robust capacity for classifying food service data (final test score of 0.7356 = 73.56%), a positive prospect towards deploying the label more effectively.

Significant Conclusions

As we move forward, it is clear that the insights and methodologies developed have the potential to significantly influence both the academic and operational landscapes of food services at UBC and beyond. The enhanced analytical capabilities have not only improved the precision of environmental impact assessments but have also paved the way for more informed decision-making regarding food procurement and sustainability practices.

List of Figures

List of Abbreviations

- AWS: Amazon Web Services
- BERT: Bidirectional Encoder Representations from Transformers
- CFFS: Climate-Friendly Food Systems
- CSV: Comma-separated values
- GCP: Google Cloud Platform
- GHG: Green House Gases
- GPT: Generative Pretrained Transformer
- ML: Machine Learning
- NaN: Not a Number
- NLP: Natural Language Processing
- OC: Optimum Control
- TF-IDF: Term Frequency-Inverse Document Frequency
- UBC FS: UBC Food Services

Introduction

I am excited to highlight our ongoing efforts to foster a sustainable food system on campus, a key component of our CAP 2030 Action. Initiated in 2021, our strategy encompasses a broad range of activities, from sourcing to waste management, aimed at reducing our food system's greenhouse gas (GHG) emissions by 50% by 2030 in alignment with the Paris Agreement.

This work represents a continuously evolving applied research project designed to enhance resiliency and sustainability across our food systems. Through initiatives such as the campus-wide Climate-Friendly Food System (CFFS) label and strategies for food waste prevention and recovery, UBC SEEDS and UBC CFFS is actively advancing climate-friendly food practices that benefit our community and the planet. In this report, I will discuss some advancements by UBC CFFS from May 2023 to April 2024.

Enhancement of Environmental Impact Assessment through Land Use Integration

The research aimed at refining the environmental impact assessment methodology used within the University of British Columbia (UBC) Food Services' operations has yielded significant advancements. Prior to this project, the assessment of recipes at various UBC food joints—Totem, Open Kitchen, and Gather—relied on a weighted average method that considered three environmental attributes: greenhouse gas (GHG) emissions, nitrogen footprint, and stress-weighted water use.

Incorporation of Land Usage Factor:

This section describes the integration of a fourth attribute into the environmental impact assessment model: the **land use factor**. This inclusion ensures a more accurate and holistic reflection of the environmental footprint associated with food choices.

Methodological Adjustments

The extension of the methodology involved significant modifications to the existing analytical workflows. Specifically, functions and methods within the python-based analysis, both in Jupyter Notebook and PyCharm environments, were adjusted to accommodate the additional land use metric. This process entailed an examination and integration of land use factors for specific food categories under evaluation, highlighting the project's commitment to precision in environmental impact assessments.

Impact

The revised environmental impact assessment methodology, now inclusive of the land use factor, is an advancement in the analysis of sustainable food products at UBC. This methodological enhancement aligns with broader sustainability goals, providing a valuable framework for informed decision-making in food service procurement and menu design, with potential applicability beyond the UBC context.

In conclusion, the integration of the land usage factor into the environmental impact assessment for UBC food products represents a significant step forward in the detailed understanding and management of food-related environmental impacts.

Revaluation of Environmental Impact Assessment Baselines for RED, YELLOW, GREEN Classification

This classification is key for comprehensive analysis of environmental impact data, integrating a thorough understanding of greenhouse gas (GHG) emissions, nitrogen footprint, stress-weighted water use, and the newly incorporated land use factor.

Methodological Adjustments

The code for this change can be found in the directory: *"baseline_calc"* and within the Jupyter Notebook file in this directory, it outlines a process designed for the analysis and processing of food recipes' environmental impact data. Here's a concise explanation of the notebook's workflow:

1. The script reads a CSV file, containing data on food recipes, selecting specific columns related to the food items and their environmental impact attributes.
2. Another CSV file, *Recipes Footprints.csv*, is loaded, which includes detailed environmental impact metrics (e.g., nitrogen lost, stress-weighted water use, land use per 100g) for various products by their IDs.
3. The notebook filters the recipes to include only those with a weight greater than or equal to 6 grams, focusing on significant ingredients.
4. A loop is then executed to match the product IDs in the recipes with those in the environmental metrics dataset. When matches are found, the script updates the recipe dataset with the corresponding nitrogen loss, stress-weighted water use, and land use metrics from the environmental impact dataset.
5. This matching and updating process enriches the original recipe data with precise environmental impact metrics, enabling a more comprehensive analysis of each recipe's environmental footprint.
6. Finally, the dataset is saved to a new CSV file, consolidating the results for further analysis or reporting purposes.

Impact

The incorporation of RED, YELLOW, and GREEN thresholds in the environmental impact assessment methodology provides a metric to decide which category a food product comes under.

In conclusion, by simplifying complex environmental data into an easily understandable, color-coded system, it enables more straightforward communication and education about the ecological implications of food choices to stakeholders.

‘

Enhancements in GHG Categorization and Unit Standardization (Procurement Data Work)

The *July31_Reorganizing_categories*.ipynb and *Assigning_categoryID_from_Items_Assigned_List.ipynb* notebooks under the directory, Categorizing_IDs_to_GHG_IDs illustrate significant advancements in automating the categorization of food items into greenhouse gas (GHG) impact categories and standardizing measurement units for procurement data analysis. The following section is an overview of the process.

Methodological Adjustments

July31_Reorganizing_categories.ipynb:

1. This notebook automates the assignment of GHG categories to food items by analyzing their descriptions.
2. It employs conditional logic to detect keywords in food descriptions, mapping each item to an appropriate GHG category.
3. This method significantly reduces manual intervention, previously required for category assignment, by leveraging pattern recognition in text data through multiple if-statements that search for key words indicating a specific GHG category.
4. The process includes handling exceptions and edge cases, ensuring a high degree of accuracy in categorization.
5. The notebook also implements checks for NaN values, indicating items that may require manual review, thus maintaining data integrity.

Assigning_categoryID_from_Items_Assigned_List.ipynb:

1. This notebook focuses on linking food items with predefined GHG categories by matching item IDs to a list of assigned categories.
2. It automates the process of associating each food item with a specific GHG impact category, based on a list that consolidates all the manual labour of previous iterations.
3. The script reads and processes input data to dynamically assign category IDs, eliminating the need for manual item-by-item categorization.
4. Items not found in the predefined list are flagged for review, ensuring comprehensive coverage and accuracy.
5. The automation of category ID assignment streamlines the workflow for environmental impact assessment, making it more efficient and scalable.
6. This process is crucial for large datasets, where manual categorization would be impractical and error-prone.

Impact

By automating this aspect of data preparation, the notebook enhances the efficiency of environmental impact analyses in the food supply chain.

Enhancing Sustainability insights through NLP: A Machine Learning Approach to Classifying UBC Food Products

Reasoning

The purpose of creating a ML application is largely focused towards creating expansion in the delivery of the CFFS label across different food joints in UBC. Given the large amounts of datasets it will become increasingly challenging to do so for many restaurants. One potential solution is a more automated workflow which can be achieved through machine learning models.

I developed this machine learning model on Google Colab Notebook and used T4 GPU to run the application. The *CFFS_NLP_model.ipynb* notebook details the deployment of a machine learning model to analyze UBC Food Services data for the 2023-2024 study term, focusing on natural language processing (NLP) techniques to enhance understanding and categorization within the dataset. The following contains a summary of the code and its process.

Link to the ML Project:
https://colab.research.google.com/drive/1pckGYAkNr7-rkkefSF6GWSJlBQZN9T--?usp=sharing

Methodologies

1. The model employs a text classification approach, leveraging NLP to categorize textual data related to food services, aiming to identify patterns and insights that could inform sustainability practices.
2. It utilizes a pipeline combining TF-IDF (Term Frequency-Inverse Document Frequency) vectorization and a logistic regression classifier to process and categorize the text data efficiently.
3. TF-IDF vectorization is applied to transform the textual data into a format that the machine learning model can work with, emphasizing words that are important for classification but not common across all documents.
4. The logistic regression model, chosen for its effectiveness in binary and multiclass classification problems, is configured with hyperparameters to optimize its performance.
5. Hyperparameter tuning is conducted through a grid search approach, evaluating 18 different combinations across parameters like regularization strength (C) and TF-IDF settings (max_df and ngram_range).
6. The model's performance is rigorously evaluated using a 5-fold cross-validation strategy, ensuring the reliability of the results by training and testing on different subsets of the data.
7. The model achieves a best cross-validation score of 0.702, indicating its ability to accurately categorize data 70.2% of the time, with a final test score of 0.7356 = **73.56%**, showcasing its robustness and effectiveness in classifying unseen data.

This model exemplifies the application of machine learning to enhance the analytical capabilities of UBC Food Services, offering insights that can drive sustainability and operational efficiency by enabling the CFFS label to be generated for more products in various locations in a timely manner. There is huge potential for growth and improvement through continuous growth of training data input.

Constructing an NLP Pipeline in Azure ML Studio

Creating a Natural Language Processing (NLP) pipeline on Azure involved using Azure Machine Learning Studio to define, design, and deploy a series of data processing steps that enable text analysis at scale. The provided image is a screenshot of such an NLP pipeline configured within Azure Machine Learning Studio for processing the Totem 23-24 Dataset. Below is a brief summary of the process depicted in the image:

1. The process begins with the "Totem23-24 Data" module, where the raw dataset is input for analysis. Text preprocessing is performed using the "Preprocess Text" module, which likely includes standard NLP tasks like tokenization, stop word removal, and possibly lemmatization or stemming.
2. Following preprocessing, the "Split Data" module divides the dataset into training and test sets to prepare for model training and evaluation, ensuring a fair assessment of the model's performance.
3. "Feature Hashing" is then applied to transform textual data into a numerical format that machine learning algorithms can process; it converts the n-grams (word pairs) into a vector of numbers.
4. The "Extract N-Gram Features from Text" module is responsible for creating a vocabulary and extracting both unigram (single word) and bigram (two-word sequences) features, which are essential for capturing context in text data.
5. A "Multiclass Logistic Regression" module indicates that a logistic regression algorithm is used for categorizing the text into multiple categories or classes.
6. The "Train Model" module combines the feature data with a logistic regression algorithm to build a predictive model.

The entire pipeline is designed for modularity and reproducibility, allowing for easy adjustments and iterations to improve model performance or to repurpose the pipeline for other NLP tasks. Similar pipelines can be developed by UBC students taking up this position if they have a keen interest in advancing the ML research for CFFS.

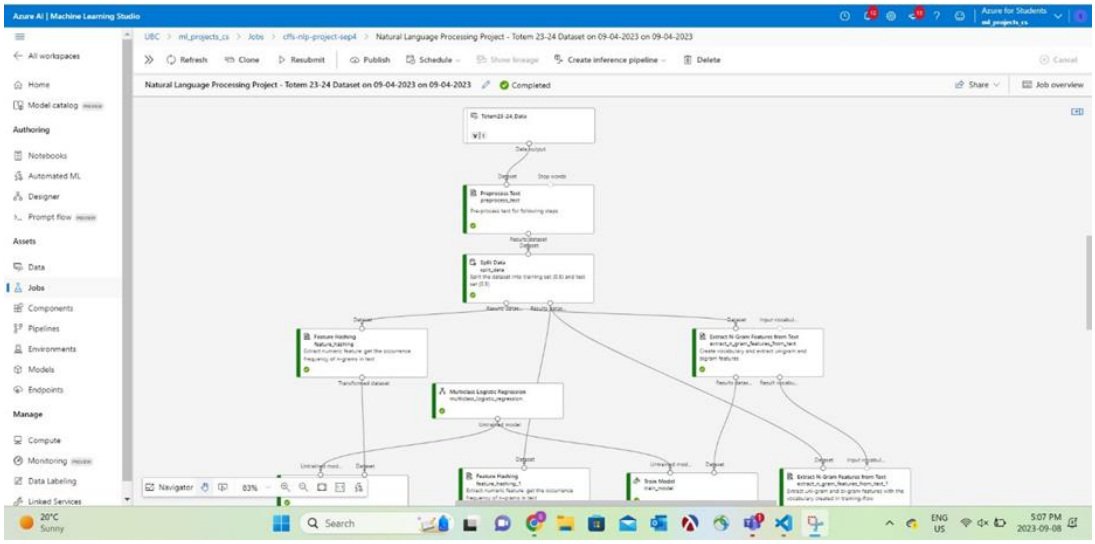

Figure 1—*Microsoft Azure Natural Language Pipeline* – The pipeline uses Totem 23-24 Data, splits the data, feature hashing, and multiclass logistic regression model.

Discussions for ML continuation

There were two paths explored for the ML integration – an Azure based pipeline and a Google Colab Python programming version. The Azure path had to be discontinued due to cost requirement to run and store the service pipeline. Additionally, some expert knowledge support would be beneficial to drive this forward as it involves an advanced cloud computing service. The second method was more practical and applicable due to the limited time and resources available. Similar ML pipelines can be developed by UBC students taking up this position if they have a keen interest in advancing the ML research for CFFS.

Continuing Collaboration and Analysis for UBC Food Services

The ongoing project with UBC Food Services sustained its focus on the analysis of products from the three All Access Dining (AAD) halls – Gather, Totem, and Open Kitchen. We continued the lines of communication and collaboration with the UBC Food Services IT team. This partnership has facilitated an exchange of data and insights, ensuring that analytical efforts are both informed and supportive of operational goals.

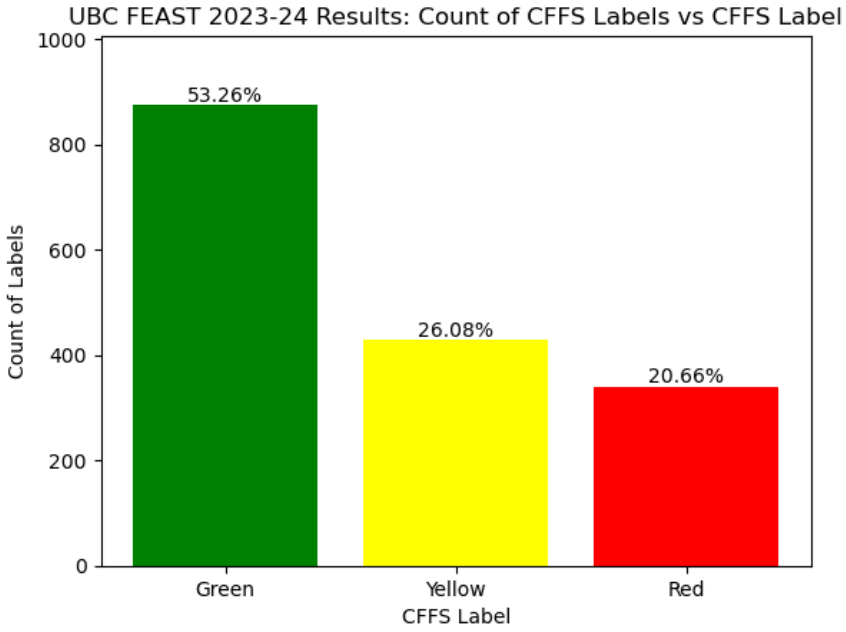

Figure 2—UBC FEAST 2023-24 CFFS Labels Results – *The green bar represents the number of products labelled "GREEN". The yellow bar represents the number of products labelled "YELLOW". The red bar represents the number of products labelled "RED". The grey bar represents the total number of produts. The percentage above each bar represents the percentage of each label count with respect to the total label count.*

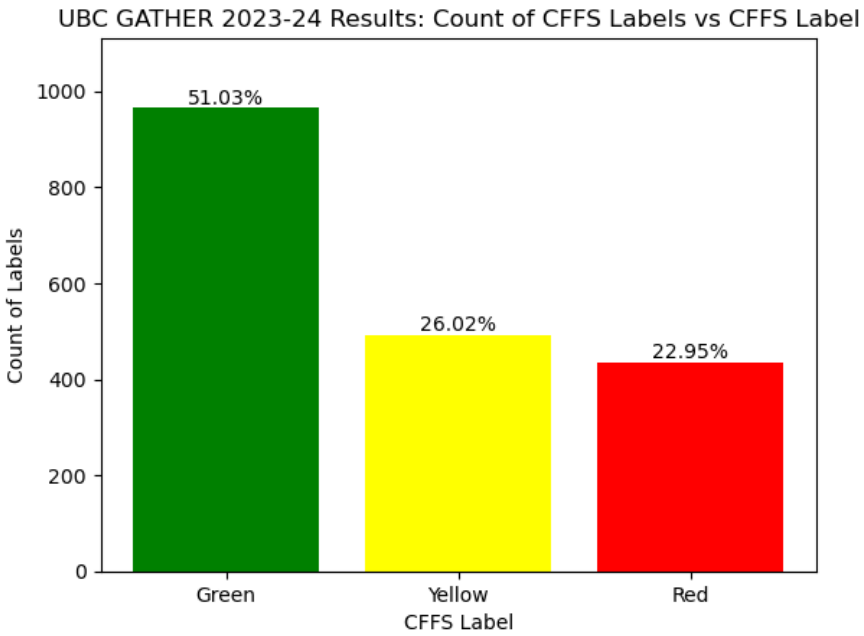

Figure 3—UBC GATHER 2023-24 CFFS Labels Results – *The green bar represents the number of products labelled "GREEN". The yellow bar represents the number of products labelled "YELLOW". The red bar represents the number of products labelled "RED". The grey bar represents the total number of produts.The percentage above each bar represents the percentage of each label count with respect to the total label count.*

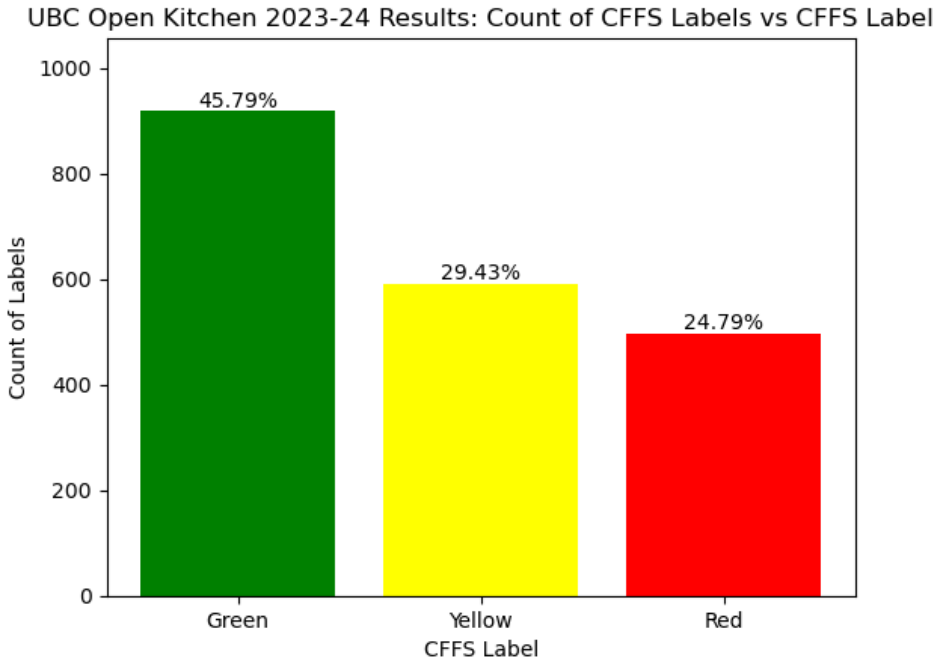

Figure 4—UBC OPEN KITCHEN 2023-24 CFFS Labels Results – *The green bar represents the number of products labelled "GREEN". The yellow bar represents the number of products labelled "YELLOW". The red bar represents the number of products labelled "RED". The grey bar represents the total number of produts.The percentage above each bar represents the percentage of each label count with respect to the total label count.*

Successful Launch of AMS Collaboration at “The Gallery” with UBC CFFS

The code is in path CFFS-S23/CFFS-22-23/AMS_2023/RECIPE_PROCESSES_2023_2024 and it leverages the analysis process initially developed for UBC Food Services for most of the workflow. However, the first part of the analysis and the data extraction and separation into Preps, Items, and Ingredients differs. Here is a summary:

1. Preps are set by finding data that are both in the Product and Ingredient columns.
2. Items are set by finding all data that are not preps and not products.
3. The data that are not preps and items and are also in the Product column are evaluated as the final products that need to be given an assigned CFF label.

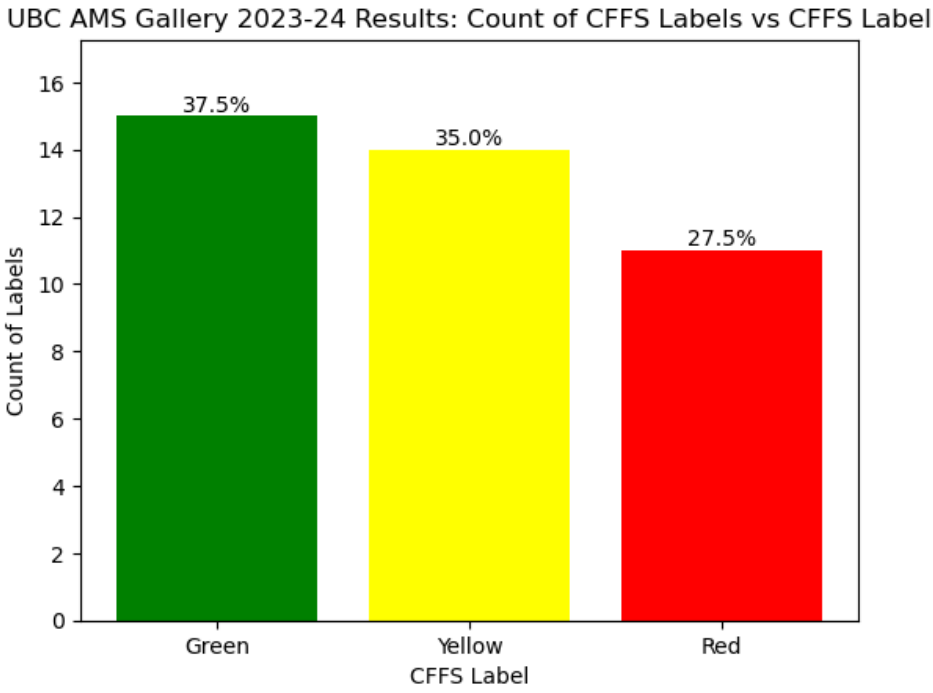

Figure 5—UBC AMS GALLERY 2023-24 CFFS Labels Results – *The green bar represents the number of products labelled “GREEN”. The yellow bar represents the number of products labelled “YELLOW”. The red bar represents the number of products labelled “RED”. The grey bar represents the total number of produts.*

Recommendations

Recommendations for Action

Immediate Actions:

1. **Expand Data:** Encourage stakeholders to provide more data for food joints in UBC so that the label can expand from three restaurants in UBC Food Services and The Gallery at AMS.

2. **Implement Findings:** Start integrating the model's insights into the menu planning processes at UBC (analyzing the number of products to sell based on their environmental impact). This could involve increasing the quantity of food products classified as GREEN.
3. **Timelines for data submissions:** Improve communication on how soon data will be made available by UBC Food Services and UBC AMS. In the past there has been confusion over the timelines which have led to delays and then time-bound analysis windows. The exact product requirements for the current term should also be clarified to avoid extra work.

Mid to Long-term Actions:

1. **Collaborate with Tech Teams:** Work with UBC's IT department or external tech consultants to integrate the machine learning model directly into UBC Food Services' operational software systems. This would automate the categorization process, making it real-time and more efficient.

2. **Sustainability Workshops:** Organize workshops or seminars for staff and students, focusing on understanding and utilizing machine learning insights for sustainable food choices. This will raise awareness and ensure the practical application of research findings.

Recommendations for Future Research

Short-term Research:

1. **Algorithm Comparison:** Conduct research comparing different machine learning algorithms or NLP techniques to find the most efficient and accurate method for classifying UBC Food Services data. This could involve exploring beyond logistic regression and TF-IDF vectorization, considering neural networks or deep learning approaches for image recognition of products that will be able to assign the food label based on predicted ingredients and its quantities. This expansion may require the guidance of an industry professional as it goes beyond the scope of a single undergraduate student.

Long-term Research:

1. **Advanced Natural Language Processing (NLP) Techniques:** Explore the application of more advanced NLP techniques and technologies, such as BERT (Bidirectional Encoder Representations from Transformers) or GPT (Generative Pretrained Transformer), to enhance the model's understanding and categorization of textual data. This could significantly improve the precision of CFFS labels derived from the data.
2. **AWS, and Other Platforms:** Experiment with constructing NLP pipelines in other cloud platforms like AWS to compare efficiency, cost, and ease of use. This could help in identifying the most suitable platform for scaling up the research and its applications.

Recommended timeline:

It takes about 2-4 weeks to get database accessibility, availability of data, and finalize the extraction process of most venues. This process can be faster if the data is readily available, and the latest information is stored on Optimum Control (OC). This data will be accessible through RadminViewer and you will need to coordinate with UBC FS IT team to gain access to their system. Running the code and making manual adjustments due to unstandardized data (data that cannot easily be converted to

grams or millilitres) can take 3-4 weeks for a beginning WorkLearn student position with limited hours per week. In between this work process it is important to communicate with your supervisor and keep them informed and updated on your progress and any issues that arise from the OC side, code bugs, or unexpected results.

By taking these actionable steps and pursuing further research, UBC CFFS may significantly advance its commitment to sustainability.

Conclusion

The work described above has been pillared by work done by Silvia and Jenny. In this iteration, my focus was to preserve the milestones achieved and improve the analysis process by maximizing efficiency and quality in the results. As a brief summary of my work, I added a land use factor in the calculations, automated the earlier process of manually assigning categories to ingredients, re-evaluated the baseline calculations for GREEN, YELLOW, and RED, created a new analysis workflow for AMS stemmed from the UBC FS iteration to launch the venture with “The Gallery”, provided labels for UBC FS data in the 2023 Summer and 2023 Winter terms, and finally set up a Machine Learning NLP solution on Google Colab and Microsoft Azure. Our concerted efforts have been towards ensuring the successful deployment of an automated data extraction and environmental impact assessment process, thoughtfully engineered to improve both the precision and efficiency of our evaluations.

Acknowledgements

I would like to extend my deepest gratitude to Dr. Juan D. Martinez and the entire UBC CFFS-AT for their support and collaboration throughout this enriching term. My experience has been profoundly enriched by this project, and I am incredibly thankful for the chance to make a meaningful impact alongside such esteemed mentors and peers.