AOS C204 Final Project

Classifying Healthcare Procurement Data using Machine Learning

Kelli Fletcher

Professor Alex Lozinski

University of California, Los Angeles

6 December 2024

Introduction

The primary duty of healthcare providers is first and foremost "to do no harm" according to the Hippocratic oath, but if we continue to disregard environmental consequences in the delivery of patient care: are we really living up to this standard? Scope 3 emissions are difficult to quantify, but they "account for about 80 percent of health sector emissions and [are] generated largely from supply chain expenses"(Agbafe et al., 2024). According to the Greenhouse Gas Protocol, scope 1 emissions are classified as *direct* emissions (i.e., direct emissions from owned or controlled sources) and scope 2 (i.e., indirect emissions from the generation of purchased energy consumed by the reporting company) and 3 as *indirect* emissions. Scope 3 emissions are more specifically defined as "occur[ring] from sources owned or controlled by other entities in the value chain (e.g., materials suppliers)" (*Corporate Value Chain (Scope 3) Accounting and Reporting Standard*, n.d.). The current industry standard for measuring scope 3 emissions is through US Environmentally-Extended Input-Output (USEEIO) modeling, which "is an environmental-economic model of US goods and services that can be used for life cycle assessment, footprinting, national prioritization, and related applications" (Ingwersen et al., 2022). USEEIO takes economic information and converts it to Greenhouse Gas (GHG) emissions based on US commodity categories, essentially creating a model that uses dollars as inputs and generates GHG emissions as outputs. This approach is inherently flawed due to the weak correlation between product price and product emissions intensity, but with thousands of products in a businesses supply chain, individual Life Cycle Assessments (LCA) to calculate emissions for each product are extremely cumbersome and unrealistic.

The aim of this project is to investigate the use of machine learning models to improve the identification of high emissions intensity products based on product text descriptions and quantities from procurement data. There have been few studies utilizing machine learning for measuring business emissions, even fewer studies targeting scope 3 measurement, and none that have involved the US healthcare industry. Machine learning classification offers a new approach not only for healthcare, but for all businesses that wish to focus their efforts on scope 3 reductions.

Data

The dataset used in this study was provided by University of San Francisco (UCSF) Medical Center, a 600 bed public hospital located in San Francisco, California. They are a level 1 trauma center and provide both tertiary and quaternary care to their patients (*UCSF Medical Center at Parnassus, Mount Zion, Mission Bay | Department of Medicine*, n.d.). The dataset consists of procurement data from one fiscal year, not inclusive of a year affected by the Covid-19 pandemic. The variables included that are of interest for this study include: item description, yearly total quantity, yearly total spent, and commodity type. The sample size is approximately 25,000 unique products.

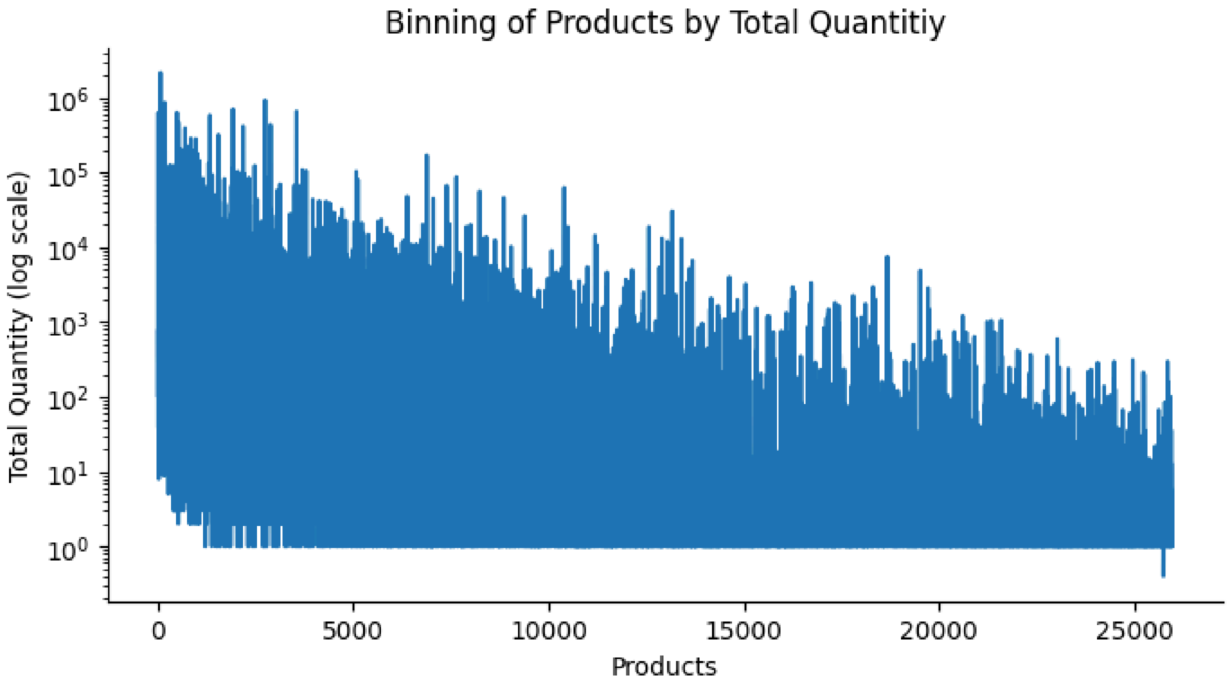


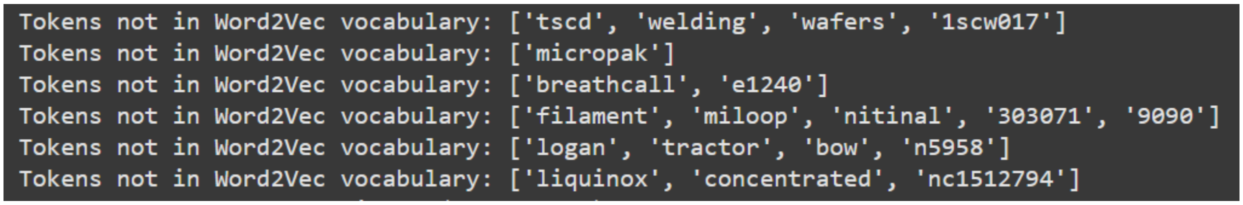
Figure 1. Log10 distribution of 25,000 products by quantity.

Modeling

The objective of this study was to attempt to predict whether healthcare items were low, medium, or high emissions products based on their text description, annual quantities, and annual cost. The main machine learning approaches utilized in this study were **K-means** and **Decision Tree Classification**. The workflow of this study began with preprocessing text data, followed by k-means clustering, cosine similarity calculation, decision tree model training, fine tuning, visualization and concluded with a confusion matrix model evaluation.

Preprocessing

After loading the dataset into Google Collab, item descriptions (categorical), yearly total quantity (continuous), department (categorical), and yearly total cost (continuous) were extracted as relevant columns to later be combined as a feature set. Continuous data such as yearly total cost and yearly total quantity were scaled using MinMaxScaler. Wordtokenizer was utilized to vectorize the item descriptions into tokens, which were then cleaned to convert the text to lowercase, and remove stopwords, spaces, and the catalog number at the end of the description. Empty descriptions were filtered out. The cleaned tokens were used to train the Word2Vec model, which then generated 100 vector embeddings to represent the text descriptions. There were 3 cases of manual tagging for keywords such as “disposable”, “implant”, and “single-use” to capture specific attributes. The categorical and continuous features were then combined into a single feature set array. The output shapes of the combined features were then generated to be, (25959, 106). Below is a list of words that were excluded, because they are not in the Word2Vec vocabulary.



```
Tokens not in Word2Vec vocabulary: ['tscd', 'welding', 'wafers', '1scw017']
Tokens not in Word2Vec vocabulary: ['micropak']
Tokens not in Word2Vec vocabulary: ['breathcall', 'e1240']
Tokens not in Word2Vec vocabulary: ['filament', 'miloop', 'nitinal', '303071', '9090']
Tokens not in Word2Vec vocabulary: ['logan', 'tractor', 'bow', 'n5958']
Tokens not in Word2Vec vocabulary: ['liquinox', 'concentrated', 'nc1512794']
```

K-Means Clustering

K-Means clustering was applied to the combined feature set and tasked to group the data into 3 corresponding clusters, “low”, “medium”, and “high”. K-Means works iteratively to assign data to clusters, minimizing intra-cluster variance. In this study 10 random initializations and a maximum of 300 iterations were performed. Once the model was fit, the final labels are added into a new ‘cluster’ column for use in the Decision Tree later. Cluster performance was evaluated using the silhouette score, measuring separation and cohesion between clusters.

Cosine Similarity

Cosine similarity was used to measure the similarity of text descriptions from the Word2Vec embeddings. Each Word2Vec embedding was normalized, then a cosine similarity min, max, mean, and standard deviation help verify the quality and consistency of the embeddings and their pairwise relationships. MinMaxScalar was used to normalize the scores with a range of 0 to 1, to enable cross comparisons. This process allows the categorical data of text descriptions to be represented numerically and further analyzed with continuous data such as quantities and cost.

Validate Emissions Bins

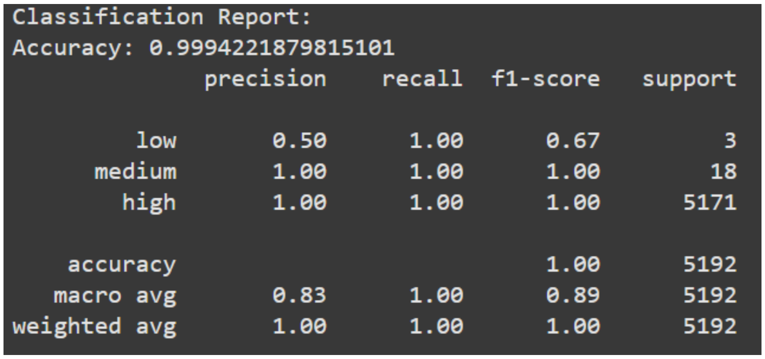
An emissions score was calculated for each product by combining weighted contributions of quantitative features and binary indicators for product attributes.

```
# Compute the emissions score with Word2Vec contribution
proxy['EMISSIONS'] = (
    0.25 * proxy['YEARLY_TOTAL_QTY_RCV'] +
    0.1 * proxy['YEARLY_TOTAL_SPENT'] +
    0.2 * proxy['is_disposable'] +
    0.2 * proxy['is_single_use'] +
    0.2 * proxy['is_implant'] +
    0.05 * proxy['word2vec_similarity_scaled']  # Add Word2Vec 
contribution
)
```

This emissions score is used to categorize products into 3 bins, “low", “medium”, and “high” emissions. The thresholds for these bins were determined from trial and error. LabelEncoder was used to encode the bins numerically. A histogram was created to show binning distribution, as seen in Figure 3.

Train Decision Tree Classifier

DecisionTreeClassifier was used to predict emissions categories (low, medium, high) based on the combined feature set proxy[‘EMISSIONS’]. An 80-20 training/test split was implemented on the dataset. Initially the maximum depth was set at 5, but was later evaluated to be optimized at 3. The model’s accuracy was evaluated and detailed in the classification report, which includes precision, recall, and F-1 scores for each of the 3 categories.



```
Classification Report:
Accuracy: 0.9994221879815101
              precision    recall  f1-score   support

         low       0.50      1.00      0.67         3
      medium       1.00      1.00      1.00        18
        high       1.00      1.00      1.00      5171

    accuracy                           1.00      5192
   macro avg       0.83      1.00      0.89      5192
weighted avg       1.00      1.00      1.00      5192
```

Figure 2. Classification report of the Decision Tree Model.

Emissions Distribution

The emissions score distribution (Figure 3) and feature contributions were visualized in histograms, to reveal distribution and frequency across the defined bins. Feature contributions were shown visually to highlight their relative weights, shown in Figure 4. These weights were chosen based on background knowledge of Life Cycle Assessment calculated emissions and healthcare products.

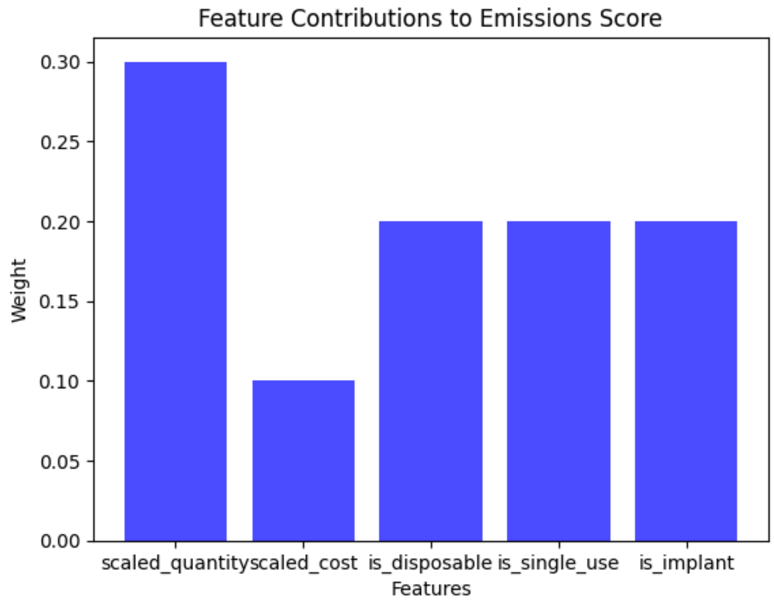


Figure 4. Feature contributions by weight.

Visualization

The Decision Tree Classifier was visualized using plot_tree, showing how features contribute to the prediction of emissions binning categories (low, medium, high). The visual representation (Figure 5) provides insights into the mechanisms of the model and features importance for classification.

Additionally, a t-SNE (t-distributed Stochastic Neighbor Embedding) visualization (Figure 6) was generated to see the relationship between data points in 2D. The scaled Word2Vec similarity scores and scaled quantities were used as input features.

Fine Tuning

Utilizing GridSearchCV, hyperparameter tuning for a DecisionTreeClassifier was performed to identify optimum model performance. The grid search evaluates all combinations for maximum tree depth, minimum samples required to split an internal node, and minimum samples required to be a leaf node with a 3-fold cross-validation, scoring based on accuracy. Fine tuning allows for better generalization and performance.

Model Evaluation

A confusion matrix (Figure 7) was generated to evaluate the performance of the Decision Tree Classifier model in predicting emissions categories (low, medium, high). By using the y_test and y_pred labels, we can see the number of correct predictions for each binning category. It provides a clear summary of the model's strengths and weaknesses.

Results

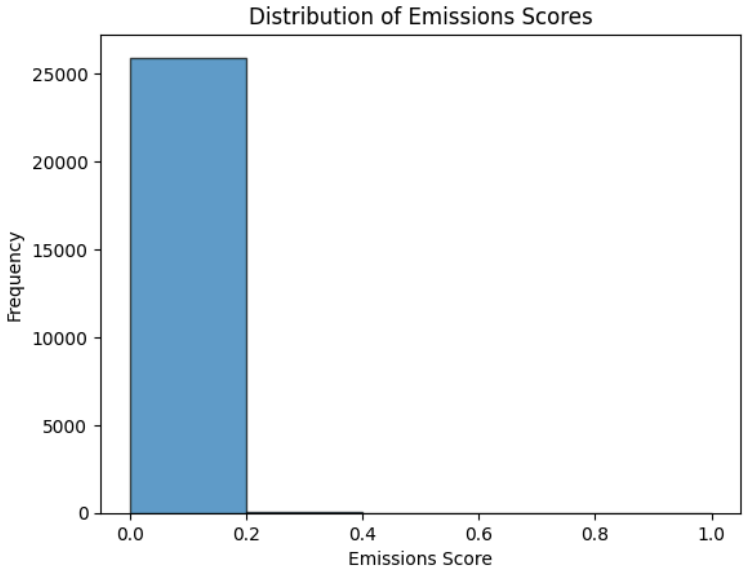


Figure 3. Emissions score distribution based on frequency.

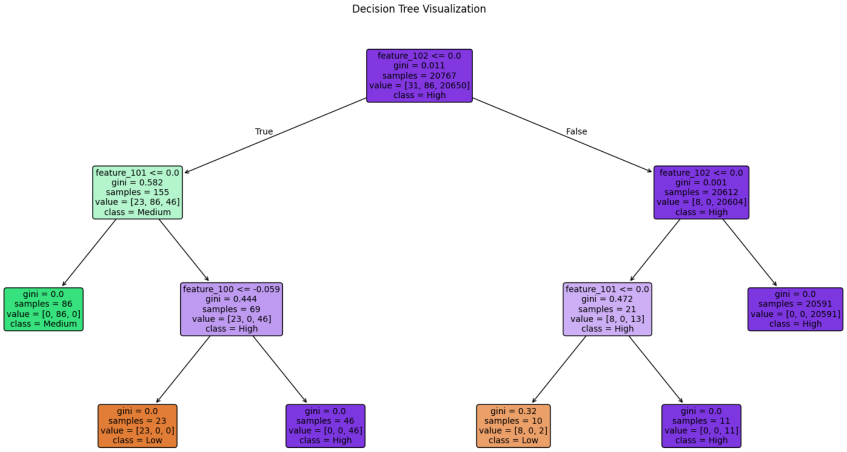


Figure 5. Decision Tree Model with maximum depth of 3.

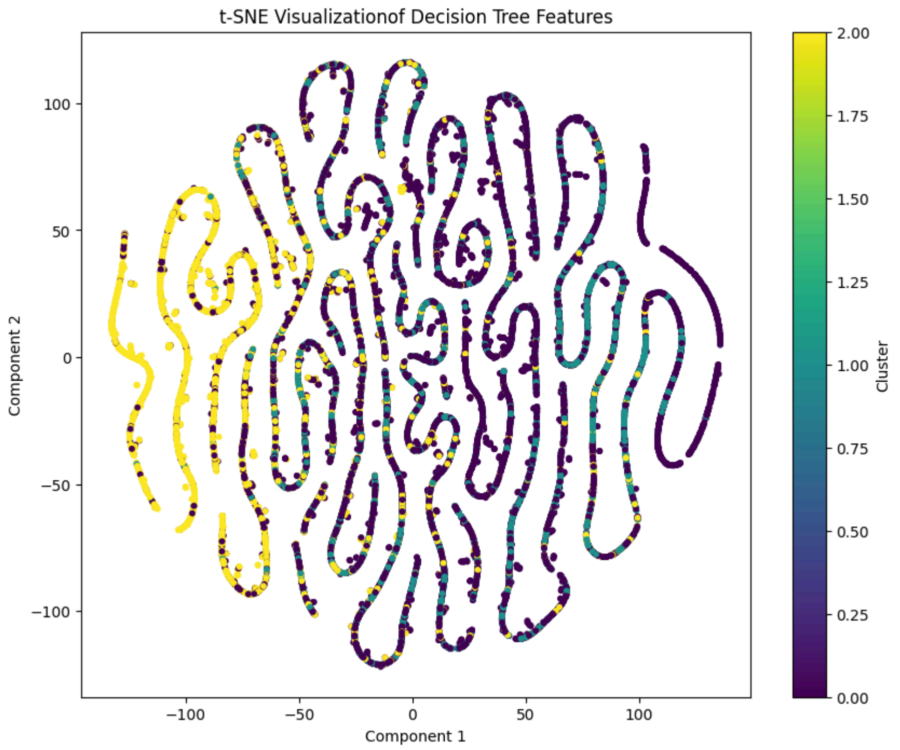


Figure 6. t-SNE visualization of the scaled Word2Vec similarity scores and scaled quantities in 2D.

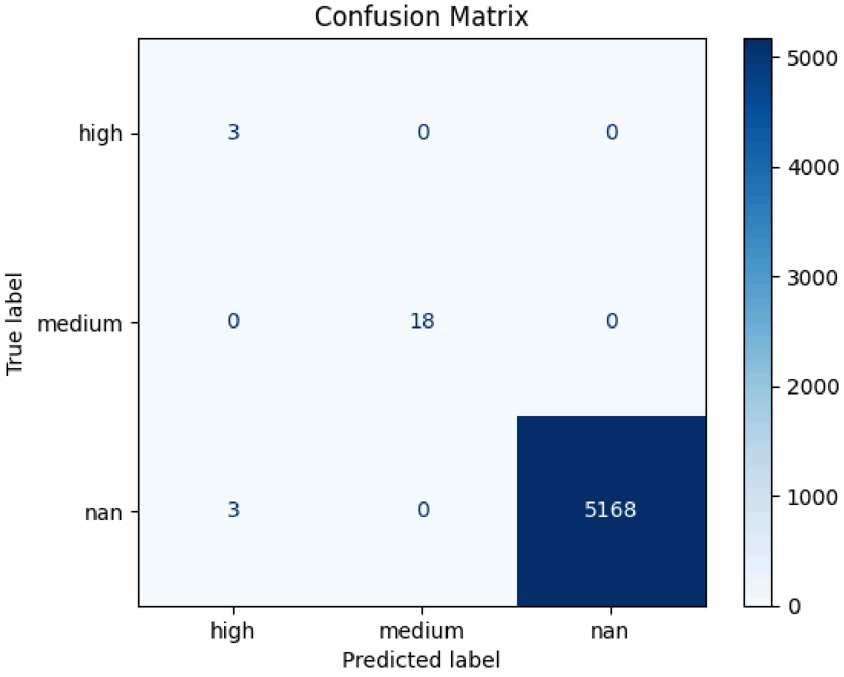


Figure 7. Confusion matrix of Decision Tree Model.

Discussion

From the figures and tables included above it has been made increasingly clear that utilizing procurement data to predict emissions categorization is a complex endeavor that will require much more precise feature engineering. From the K-Means clustering a silhouette score of 0.283 is indicative of weak clustering and suggests that the text descriptions do not naturally form well-defined groups, which lead to large overlap in emissions categorization. The Word2Vec similarity matrix showed an average of 0.316∓0.115, which also concludes a low degree of similarity from the text descriptions. Due to the diversity in the product descriptions and the naming conventions unique to healthcare, there is no shortage of noise in the data, as shown in the t-SNE visual (Figure 6).

The distribution of emissions bins is heavily skewed in favor of ‘medium’, with 104 products in this category. There were 34 products in the ‘high’ and 0 in the ‘low’ categories respectively. The imbalance suggests an issue in either the emissions weighting formula or the binning thresholds (which were chosen with trial and error), leading to unreliable and biased predictions. Although the decision tree shows an almost perfect accuracy score of 0.99, the confusion matrix is overwhelmingly ‘NaN’, indicating overfitting due to class imbalance.

Conclusion

In conclusion, the machine learning approaches used in this study, although unsuccessful in categorizing emissions data, were successful in forging the first steps in a new direction for product emissions predictions. Life Cycle Assessments are cumbersome and require expertise that can be out of reach for some organizations, but machine learning models could be the answer. Emissions disclosure and reporting is increasing every year and new methods are needed to enhance efficiency in identifying areas for reductions, and improve accuracy in measuring and reporting.

Future improvements to continue this study are as follows:

- Address emissions class imbalance by refining the binning thresholds and adjust the decision tree weighting.
- Improve clustering with different algorithms that might work better for healthcare data, and enhance feature scaling and selection for cluster separation.
- Refining feature engineering.
- Create a robust set of labeled training data that is more precise and representative of actual emissions.

References

Agbafe, V. C., Singh, H., & Cerceo, E. (2024). Comprehensive SEC Disclosure Rules Can Reduce Health Care Emissions. *Health Affairs Forefront*. https://doi.org/10.1377/forefront.20240814.180078

Corporate Value Chain (Scope 3) Accounting and Reporting Standard. (n.d.). https://ghgprotocol.org/sites/default/files/standards/Corporate-Value-Chain-Accounting-Reporing-Standard_041613_2.pdf

Ingwersen, W. W., Li, M., Young, B., Vendries, J., & Birney, C. (2022). USEEIO v2.0, The US Environmentally-Extended Input-Output Model v2.0. *Scientific Data*, *9*(1), 194. https://doi.org/10.1038/s41597-022-01293-7

UCSF Medical Center at Parnassus, Mount Zion, Mission Bay | Department of Medicine. (n.d.). Retrieved September 4, 2024, from https://medicine.ucsf.edu/about/locations/ucsf-medical-center-parnassus-mount-zion-mission-bay