Environmental Impact of Online Products

Marium Sarah   
*College of Business & Economics*  
*California State University, East Bay*Hayward, CA marium.sarah@csueastbay.edu  
  
line 2: *dept. name of organization (of Affiliation)*  
line 3: *name of organization (of Affiliation)*line 4: City, Country  
line 5: email address or ORCID  
  
line 1: 4th Given Name Surname  
line 2: *dept. name of organization (of Affiliation)*  
line 3: *name of organization (of Affiliation)*line 4: City, Country  
line 5: email address or ORCI

*Abstract*—A solution to empower customers, to reduce the amount of plastic ending up in landfill and leading to an increase in greenhouse gas emissions, by accessing the environmental impact statistics and brand’s eco-friendly status of online products.

Keywords—environment, sustainability, image processing

# Introduction

Plastic is durable, cheap, and readily used to contain various products. Although plastic is accessible, it is also non-degradable and made to last over 400 years. A staggering 85% of the plastic thrown away in recycle bins end up in landfill, and the process of recycling the other 15% takes 16M tonnes of greenhouse gas emissions which is equivalent to 3.5M vehicles on the road [1]. The goal of this project is to alert customers of the environmental impact of products they purchase from online stores. The focus of this project is to estimate the level of plastic content in Amazon products and analyze the carbon footprint of the products and at-scale-businesses selling on Amazon.

## Project Overview

This project uses web-scraping of amazon products extracting detailed-description of weight and size, image processing to identify the type of plastic in products, statistical analysis and data visualization to compare and analyze the environmental impact of various brands.

## Problem Statement

Plastic is the most persistent non-degradable pollutant on Earth. The level of plastic packaging in products found online is rarely mentioned, thus estimating environmental impact is difficult. Knowing that majority of the plastic thrown away in recycle bins ends up in landfill, over 78% customers (surveyed by GreenPrint) hope to find eco-friendly products however they have a difficult time identifying them [2]. This project will help customers compare similar products based on their plastic content and the manufacturing brand’s environmental impact.

# Execution

Before we proceed to understanding the various data points to help estimate the carbon footprint of products. We will first produce an in-depth understanding of the project execution plan.

## Execution Plan

Fig. 1. highlights the inner workings all the way from the input to the final output of the program. The following are the 4 highlighted feature requirements of this project.

1. Web Scraping Amazon: Selected Product Categories will be scraped from Amazon to treat as input to our program.
2. Plastic Detection Model: Detect if the input Product Image contains Plastic or Not. Using an existing dataset for training and validation. Followed by testing directly on Amazon Product Images.
3. Plastic Level Identification: Algorithm to calculate the carbon footprint based on the volume and quantity provided in the Amazon Product Image Description.
4. An optional 4th existing model will be used to further identify if an image that is not Plastic is Recyclable or not. This will further help users understand that recycling also has a negative impact.
5. 1-3 Execution Requirements
6. Execution Plan Workflow

## Execution Steps

The following table highlights the execution steps that will be involved in the 12-week Capstone Project.

1. Execution Plan

| S.No. | Execution Plan | |
| --- | --- | --- |
| Step | Description |
| **1** | **Web Scraping** | Selection of certain Amazon Product Categories,  Web scraping over 1000 product details using beautiful soup package on Python. |
| **2** | **EDA & Data Cleaning** | Analyse the extracted data based on product type, popularity and any environmental impact mentions in description. Data cleaning to remove products without weight and size provided, and those without proper images provided. |
| **3** | **Partitioning & Pre-processing** | Various Pre-processing of Image data will be conducted to select the best data reduction methods. Including image cropping, edges detection, truncation. Furthermore, filters including Partition membership, discretization, normalization will be tested. Data partitioning will be used for the next step of predictive modelling. |
| **4** | **Predictive Modelling** | Create and compare common image detection predictive models including YOLO, CNN, RCCN. |
| **5** | **Compare Modelling Performance** | Cross-validation will be used along with evaluative metrics including accuracy, precision and recall comparing the model’s performance and choose the best model for our final prediction |
| **6** | **Visualize Product-level results** | Using statistical insights and common formulas of estimating environmental impact we will use various plots and graphs to create a standard easy to understand methodology of the eco-friendly level of any given product. |
| **7** | **Data Visualization Dashboard** | Using Tableau to visualize the environmental impact of various brands. This will include plots and graphs to highlight pattern and compare the performance of brands. |

## Strategy

Given the above execution plan, the strategy in place will involving performing Step 1 (Web Scraping) and Step 2 (Data Cleaning & EDA) in a Waterfall methodology one after the other in the first 2 weeks of the project.

Followed by utilizing a Sprint-based Agile Methodology focused on choosing the best Predictive Models (Steps 3-5), each sprint will start with creating the model, applying different combinations of filters and pre-processing techniques and evaluating the model’s performance against previous models tested.

Finally Step 6 and Step 7 compiling the Data Analytics and visualization will conclude the project completion. This will be completed in the final 2 weeks of the project.

# Model Dataset

To develop a strong Image Detection Model (referring to Fig. 1. 2nd Requirement – Plastic Detection Model) we first need to research and identify datasets with a proper differentiation of Plastic and Non-Plastic products.

A majority of relevant Plastic detection datasets are developed with a focus on image detection solutions for waste identification, with images mainly coming from Waste Facilities, Landfills and Litter in cities and beaches. An elimination process is required to ensure only datasets with unused new product images are considered rather than Waste or Litter. The following table are the top 5 researched datasets relevant to this project.

1. Researched Datasets

| Dataset | Plastic and Non-Plastic Datasets | | |
| --- | --- | --- | --- |
| Images | Classification | Comments |
| 1. Glassense-  Vision [8] | 500 | banknotes, cereals, medicines, cans, sauces, bottles, deodrants | **Dataset created to develop a model to help visually impaired people identify objects that are harder to recognize by touch or sound.**  **Pros:** Clear background images similar to Amazon Test Product Images, Includes variety of brands and products **Cons:** Smaller dataset |
| 2.Drinking Waste Classification [7] | 25.4k | Aluminium, cans, glass, bottles, Plastic bottles, Milk bottles | **Pictures taken manually using 12MP Phone Camera**  **Pros:** Clear background, Good variation of Plastic Images, 2 Existing Models available for Experimentation, Bigger Dataset  **Cons:** Includes used products and trash images |
| 3.Waste Classification Data [5] | 26k | Organic (12.6k), Recyclable (10k),  Non-Recyclable (3k) | **Purpose:** **Reduce toxic waste ending in Landfill. Scraped from Google Search**  **Pros:** Only dataset including Organic images. Non-trash images  **Cons:** Vague Categories, Includes irrelevant products, All Recyclable Material included (Paper, Glass, etc) |
| 4.Trashnet [4] | 2k | plastic (482)/ trash, paper, metal, glass, cardboard (1.6k) | Pictures manually taken in room lighting using a 13MP Camera.  **Pros:** Includes Metal and Cardboard, 75% Precision model included for experimentation  **Cons:** Includes various angles of same products, Mixed with trash and used product images. Uses parts of Trashnet datasets. |
| 5.Trashbox [3] | 17k | Plastic, e-waste, medical waste, paper, metal, glass, cardboard | Images scraped from the web. Uses Quantum Transfer learning to classify images with 80% accuracy  Pros: A well distributed dataset. A high precision Learning based model available for experimentation.  Cons: Model focused on fine-grained classification. |

Considering all the above datasets, manual elimination and selection of relevant categories is done to ensure the right sample training dataset is created with a good distribution of categories to ensure a realistic representation of the population.

The following file consists of the final datasets.

The Y prediction output value is fixed The final modifications are as follows:

* + - 1. Glassense-Vision: This dataset was removed due to the small size and repetitive images as the other datasets.
      2. Trashbox: This dataset consists of relevant Medical products including Syringes, Masks, Gloves and Medicine. These were appropriately separated as Plastic and Non Plastic
      3. Waste Classification Dataset: This dataset consisted of a recyclable category. A manual separation was done to seclude the Recyclable Plastic material from the Recyclable Non-Plastic Material.
      4. Trashnet & Drinking Waste Classification: Non Plastic and Plastic data is already separated in these datasets, hence no modifications were required.

The final dataset consists of 31k Images, out of which 32% are Plastic images (Fig. 1). The non-plastic images consists of 66% of Organic Fruits and Vegetables, moving forward removing 50% of fruits and vegetables will help ensure a balanced dataset.

Chart, pie chart

Description automatically generated

1. Dataset Distribution

# Test Data Collection & Web Scraping

The purpose of the test dataset is to apply the Plastic Level Algorithm to estimate the Carbon footprint (referring to Fig. 1. 3nd Requirement – Plastic Level Algorithm) after an image has been classified as Plastic by the Plastic Detection ML Model. To build this algorithm we first need to create an appropriate dataset of products with detailed information of size and volume to make the right carbon footprint calculation.

## Data Source Selection

Amazon being the world’s largest retailer was chosen to web scrape. On Amazon there are multiple categories ranging from “Beauty & Personal Care”, “Amazon Home”, “Groceries”, “Pharmacies” and more. Selecting the Groceries category for this project is the ideal options due to the following reasons:

1. Plastic containers and packaging in supermarkets and groceries is the largest contributor to plastic waste [6]. This solution can further aid physical supermarkets and groceries to help customers be more sustainable when picking the right product.

2. The Groceries category is the most diverse in terms of sub-categories with a variety of products sold with detailed information by Amazon Fresh and Whole Foods. The 17 Categories of goods are provided in (Fig. 3.)

3. Grocery items tend to have a lot of detail in terms of size, weight, quantity and volume. This will be useful in estimating the carbon footprint of products.

Given all the online grocery stores available including Walmart, 7Eleven, Instacart, and Amazon. Instacart and Amazon are the most detailed in terms product details. (1) Instacart provides the weight for every single product, with further details displaying the exact labels and calories given on each product. (2) Amazon provides the size or weight as provided by the vendor, with the option to allow customers to choose between various selections of quantity. As a result, Amazon and Instacart will be used to extract product details for the test dataset.

1. Sub-categories
2. Amazon Grocery Subcategories

## Webscraping

Webscraping becomes challenging with websites that practice internal regulations to protect their data, this makes it more difficult to scrape.

Beautifulsoup and Selenium are two of the most commonly used packages for Webscraping. Running multiple URL page requests on Amazon using beautifulsoup resulted in pages that were interpreting the visitor as bots, making it difficult to scrape (Fig. 4).

1. Amazon URL Blocks
2. Amazon URL Blocks

Hence an alternative to “beautifulsoup” is using an API that avoids proxies, bots and CAPTCHA messages that block request. ScraperAPI provides a free trial with 5000 URL requests per month.

From each category available, products were extracted from each page until 300 products were successfully retrieved. The Snacks & Sweets category had a staggering 44% share of grocery products available for purchase, followed by Beverages at 37% (Fig. 5). The Amazon Test Dataset produced consists of 4,907 products with a roughly equal distribution across the 17 grocery categories.

Chart, pie chart

Description automatically generated

1. Amazon Categories Share
2. Category Level Variety

# Data Description

The test data produced will be extracted from the Amazon website (referring to Fig. 1. 1st Requirement – Web Scraping Amazon) from specific categories focused on common household items that are purchased frequently. The extracted dataset consists of the following fields:

* **Product Name**
* **Image Link**– Link to image to extract pixels
* **Units-**Number of units within the product
* **Size-**The size ofitem
* **Item dimensions:** Given in inches LxWxH
* **Unit count:** The number of units/items in the given product
* **Item Weight:** The weight of the product
* **Item Volume:** For liquid items, the volume of the contentgiven in the packaging
* **About this item:** A detailed description about the specifications of the item on sale
* **No. of Ratings:** Total number of ratings given to this item
* **Ratings:** The overall rating on the item (scale of 1-5)

# Exploratory Data Analysis

xxx

# Data Preprocessing

This section covers the Data Preprocessing techniques adapted by the various papers cited in this paper.

## Relevant Data Preprocessing Techniques

1. Resize & Rescale:

It is vital to make sure all RGB Images are the same size, rescaling allows us to reduce the number of pixels and ensure all images have the same number of parameters going forward.

2. Thresholding:

In Thresholding all the pixels with an intensity greater than threshold are converted to 1, values less than the threshold are converted to 0. This method converts a colored RGB Image into a black and white binary image.

3. Normalization:

This is a vital step in preprocessing of Images. It refers to rescaling the pixels to ensure they lie within a defined range. The three normalization methods are as follows:

Pixel Normalization: Scale to values ranging 0-1

Pixel Centering: Scale pixel values to have a zero mean

Pixel Standardization: Scale pixel values to have a zero mean and unit variance.

4. Geometric Transformation:

This method allows to eliminate any form of geometric distortion that occurs within images. This involves rotation, scaling, distortion and undistortion. This process involves spatial transformation and grey level interpolation.

# Model and Methods

This section covers existing models developed for the researched datasets followed by a proposed methodology to create a new multi-class product plastic detection model.

## Existing Models

1. *Support Vector Machine:*

The TrashNet model consists of 2.5k images, the models used for this dataset classifies trash into 6 different categories. The SVM was first used as a simplistic model compared to others, it uses less memory and classifies images by drawing a decision boundary in a multidimensional data [2]. The TrashNet dataset was experimented on 2 Models SVM and CNN [2]. The author reported that SVM is a much better modem with the best accuracy at 63% and CNN with a low accuracy of only 22%. This low score with a complex model is attributed to the CNN’s learning rate being too aggressive leading it fluctuate erratically and not decrease at a consistent rate.

*2.* *Convolutional Neural Network:*

An 11 Layer Torch2 CNN similar to AlexNet was used, AlexNet is known to support the overfitting problem using Data Augmentation and Dropout, this is useful considering the size of dataset is small [2].

*3.* *ResNet:*

The layers within this CNN convolves images inputted and extracts important features along the way to help classify the image. ResNet is different as it protects the CNN from the vanishing gradient problem. Considering the small dataset size, data augmentation methods using shearing and scaling was done to augment the dataset. [9]

5. *Inception-v3 Transfer Learning Model:*

A learning that avoids training from nothing for larger datasets. In transfer learning featured learned from a base dataset are then reused to a second target dataset. This works well when the features are general, using the TrashNet dataset a 92% accuracy was obtained. [10]

6. *Quantum Transfer Learning Model:*

Quantum Computing advancements has allowed to solve complex computation problems resulting from Neural Networks in lesser time and cost. This model trains transfer learning models more effectively and process data in an optimal way. Using QTL reduced the training time by 27% and improved the accuracy by 11% [3]

## Proposed Methodology

xxx

# Model Evaluation

xxx

# Results

xxx

# Conclusion

xxx

##### References

1. Margaret Osborne, “At Least 85 Percent of U.S. Plastic Wate went o Landfills in 2021,” Smithsonian Magazine, May 9, 2022. *(*[*Link*](https://www.smithsonianmag.com/smart-news/the-us-recycled-just-5-percent-of-its-plastic-in-2021-180980052/)*)*
2. Michele Koch and Zach Harris, “GreenPrint Survey Finds Consumers Want to Buy Eco-Friendly Products, but Don’t Know How to Identify Them,” Business Wire, March 22, 2021 ([Link](https://www.businesswire.com/news/home/20210322005061/en/GreenPrint-Survey-Finds-Consumers-Want-to-Buy-Eco-Friendly-Products-but-Don%E2%80%99t-Know-How-to-Identify-Them))
3. Nikhil Venkat Kumsetty, “Trashbox: Trash Detection and Classification using Quantum Transfer Learning,” March 2022 ([Link](https://github.com/nikhilvenkatkumsetty/TrashBox))
4. Mindy Yang and Gary Thung, “Classification of Trash for Recyclability Status,” Stanford University, May 2016 ([Link](https://github.com/garythung/trashnet))
5. Sapal and Sashaank Sekar, “Waste Classification Data,” Kaggle ([Link](https://www.kaggle.com/datasets/sapal6/waste-classification-data-v2))
6. Jodi Helmer, “Are Supermarkets doing enough to reduce Plastic Waste?,” Food Print, December 2021 ([Link](https://foodprint.org/blog/plastic-waste/#:~:text=In%20the%20US%2C%20food%20containers,comes%20to%20tackling%20the%20problem))
7. Arkadiy Serezhkin, “Drinking Waste Classification,”, University College of London, December 2020 ([Link](https://www.kaggle.com/datasets/arkadiyhacks/drinking-waste-classification))
8. Joan Sosa Garcia and Dibris Unige and Francesca Odone, “Glassense-Vision,” SI4 Life Ligurian Regional Hub ([Link](http://www.slipguru.unige.it/Data/glassense_vision/))
9. O. Adedji and Z. Wang, “Intelligent waste classification system using deep learning convolutional neural network,” Procedia Manugfacturing, March 2019 ([Link](https://reader.elsevier.com/reader/sd/pii/S2351978919307231?token=0747C724E5C4BB4201DA7D3561F19ED0E722F4543CA103CB78B6B5906C927ED06FB5D37371BAE238CB78544083A0E80E&originRegion=us-east-1&originCreation=20221003140342))
10. F. A. Azis, H. Suhaimi, and E. Abas, “Waste classification using convolutional neural network,” in Proceedings of the 2020 International Conference on Information Technology and Computer Communications, ser. ITCC, 2020
11. I. S. Jacobs and C. P. Bean, “Fine particles, thin films and exchange anisotropy,” in Magnetism, vol. III, G. T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271–350.
12. K. Elissa, “Title of paper if known,” unpublished.
13. R. Nicole, “Title of paper with only first word capitalized,” J. Name Stand. Abbrev., in press.
14. Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, “Electron spectroscopy studies on magneto-optical media and plastic substrate interface,” IEEE Transl. J. Magn. Japan, vol. 2, pp. 740–741, August 1987 [Digests 9th01, 1982].
15. M. Young, The Technical Writer’s Handbook. Mill Valley, CA: University Science, 1989.
16. I. S. Jacobs and C. P. Bean, “Fine particles, thin films and exchange anisotropy,” in Magnetism, vol. III, G. T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271–350.Tagawa, “Electron spectroscopy studies on magneto-optical media and plastic substrate interface,” IEEE Transl. J. Magn. Japan, vol. 2, pp. 740–741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].
17. M. Young, The Technical Writer’s Handbook. Mill Valley, CA: University Science, 1989.
18. I. S. Jacobs and C. P. Bean, “Fine particles, thin films and exchange anisotropy,” in Magnetism, vol. III, G. T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271–350.
19. K. Elissa, “Title of paper if known,” unpublished.
20. R. Nicole, “Title of paper with only first word capitalized,” J. Name Stand. Abbrev., in press.
21. Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, “Electron spectroscopy studies on magneto-optical media and plastic substrate interface,” IEEE Transl. J. Magn. Japan, vol. 2, pp. 740–741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].
22. M. Young, The Technical Writer’s Handbook. Mill Valley, CA: University Science, 1989.
23. Annual Conf. Magnetics Japan, p. 301, 1982].
24. M. Young, The Technical Writer’s Handbook. Mill Valley, CA: University Science, 1989.
25. Annual Conf. Magnetics Japan, p. 301, 1982].
26. M. Young, The Technical Writer’s Handbook. Mill Valley, CA: University Science, 1989.
27. Annual Conf. Magnetics Japan, p. 301, 1982].
28. M. Young, The Technical Writer’s Handbook. Mill Valley, CA: University Science, 1989.

**IEEE conference templates contain guidance text for composing and formatting conference papers. Please ensure that all template text is removed from your conference paper prior to submission to the conference. Failure to remove template text from your paper may result in your paper not being published.**