Unstructured Data Analysis to Improve Digital Eligibility of E-Commerce Listings

Suneet Abraham, Akshay Deshmukh, Anish Jasti,
Sharan Shirodkar, Nikhitha Siddi, Matthew A. Lanham
Purdue University, Daniels School of Business
abraha46@purdue.edu, deshmu25@purdue.edu; jastia@purdue.edu;
sshirodk@purdue.edu, nsiddi@purdue.edu; lanhamm@purdue.edu

BUSINESS PROBLEM FRAMING

Our team is collaborating with a nationwide supercenter chain that mandates vendors to furnish product details along with at least one image for digital eligibility on their website. However, when vendors fail to comply with these requirements, numerous products are left unlisted on the client's website, causing an uneven customer experience.

The grocery chain's website only features products that meet specific requirements, and unfortunately, only 33% of the products in their portfolio fulfill these criteria, resulting in a loss of revenue.

The supercenter chain faces a limitation where vendors are unable to verify whether their product descriptions meet the required standards or not. Consequently, vendors tend to submit inadequate descriptions to pass the initial digital eligibility screening. To ensure uniformity and quality, there is a need to categorize descriptions appropriately.

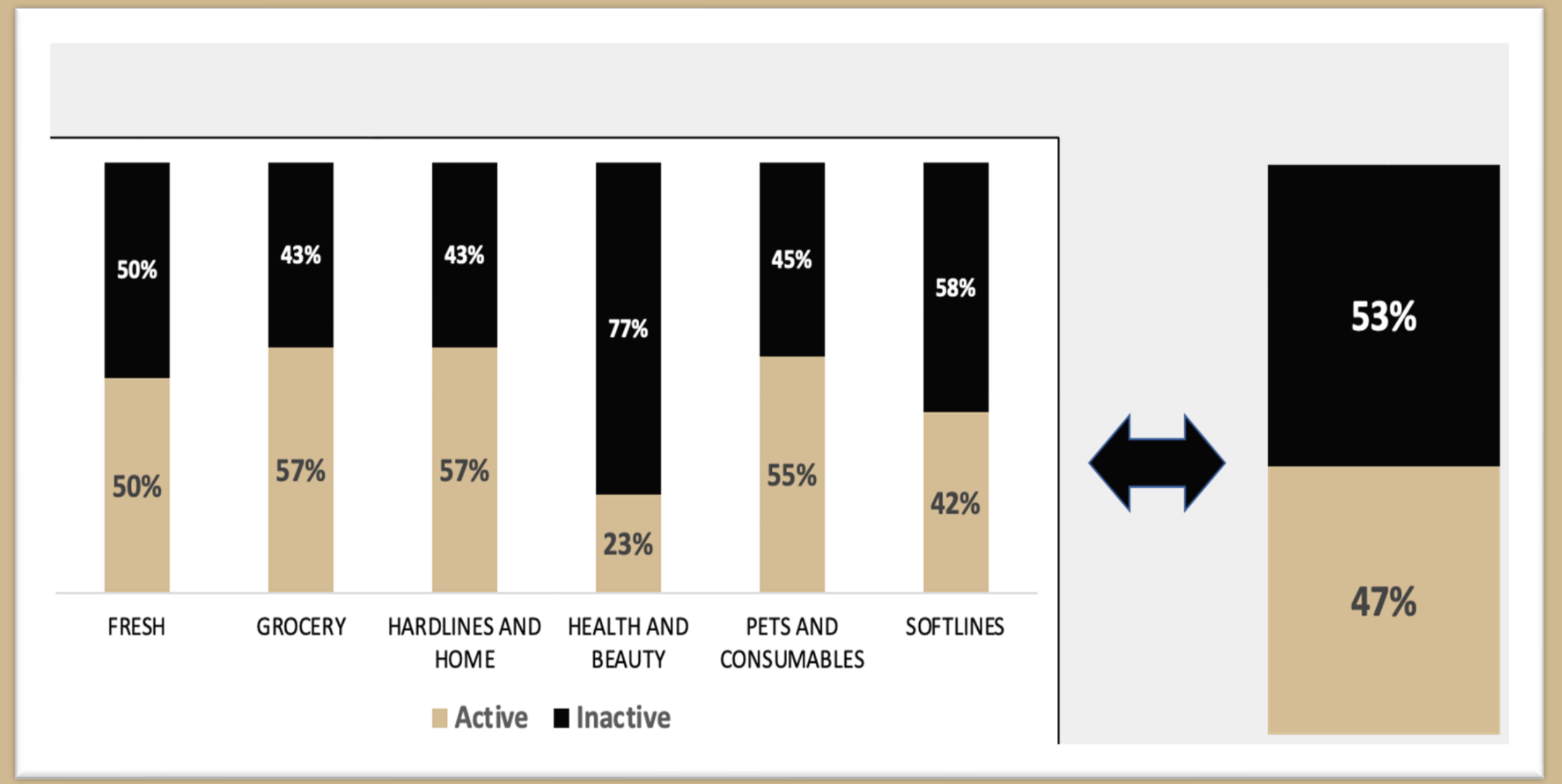


Fig. 1 Category-wise and overall split of eligible and ineligible products

Our analysis of the provided data involves assessing the distribution of digitally eligible and ineligible products across various categories, as shown in Figure 1. Our findings indicate that 53% of the post EDA data is comprised of products that are ineligible to be listed on the retailer's website due to incomplete product information. The ultimate objective of our business is to reduce the percentage of inactive products.

ANALYTICS PROBLEM FRAMING

We have identified two analytical problem goals to address the business problem –

- The first goal is to develop an algorithm to score the descriptions on a scale of 100 to categorize them as 'Good' or 'Poor'. Once we have identified the bad descriptions, we aim to correct them. The parameters we use to make this decision are readability, sentiment, length and relevance of the description. We assume only the above parameters to be pivotal in the segregation.
- The second goal involves using Optical Character Recognition (OCR) to extract text from product images and using language model (OpenAI's ChatGPT API) to generate descriptions for them.

Our success metrics are to increase digital eligible products and the percentage of 'Good' descriptions. We will focus solely on images with text and not consider object identification within the scope of this problem.

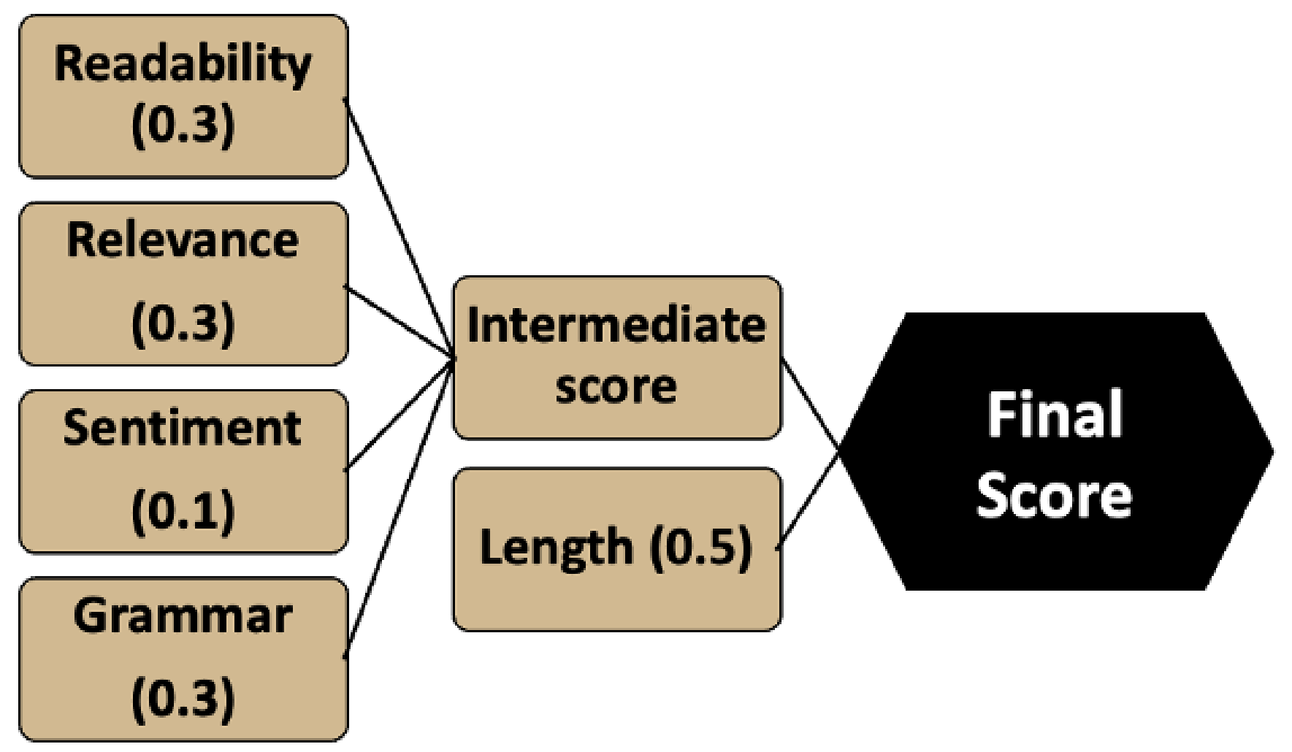

Readability (0.3)
Relevance (0.3)
Sentiment (0.1)
Grammar (0.3)
Intermediate score
Length (0.5)
Final Score

Fig. 2 Parameters for description scoring

DATA

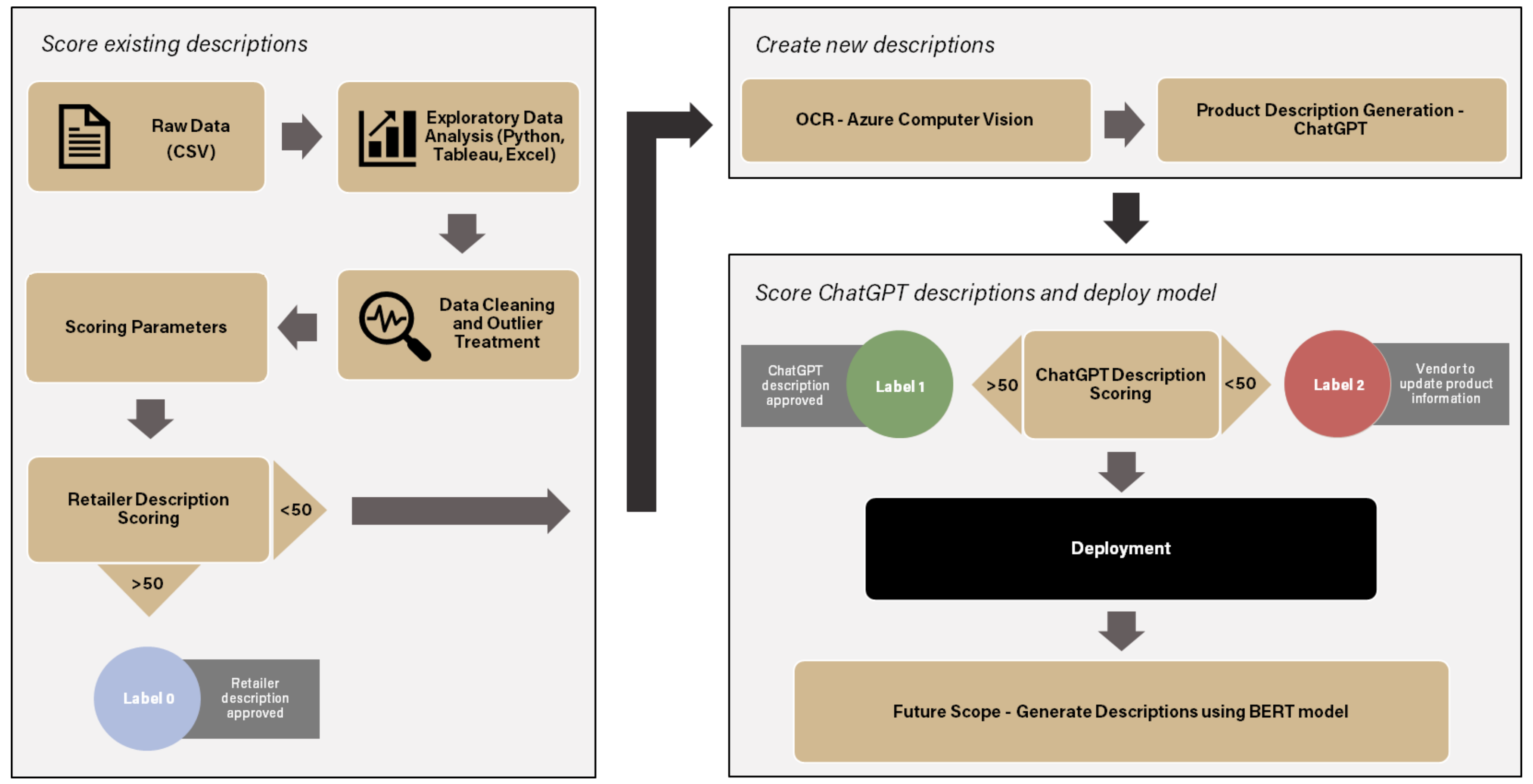
Column	Description
ItemSku	Stock Keeping Unit
UPCTypeName	Universal Product Code Type
ProductName	Name of the product for internal purpose
MarketingName	Name of product listed on website
MarketingDetails	Description of the product listed on website
ItemDocumentValue	URL to the product image

Table 1 Data dictionary

The data is confidential and is provided to us by the supercenter chain. Preprocessing involved removing non readable characters, excluding records with missing product information, and narrowing down the selection of columns to the ones required for the project.

ItemSku is the primary key for the data. ItemDocumentValue is the image URL for the products, and MarketingDetails is the column for the product description, which we use for our scoring algorithm.

METHODOLOGY

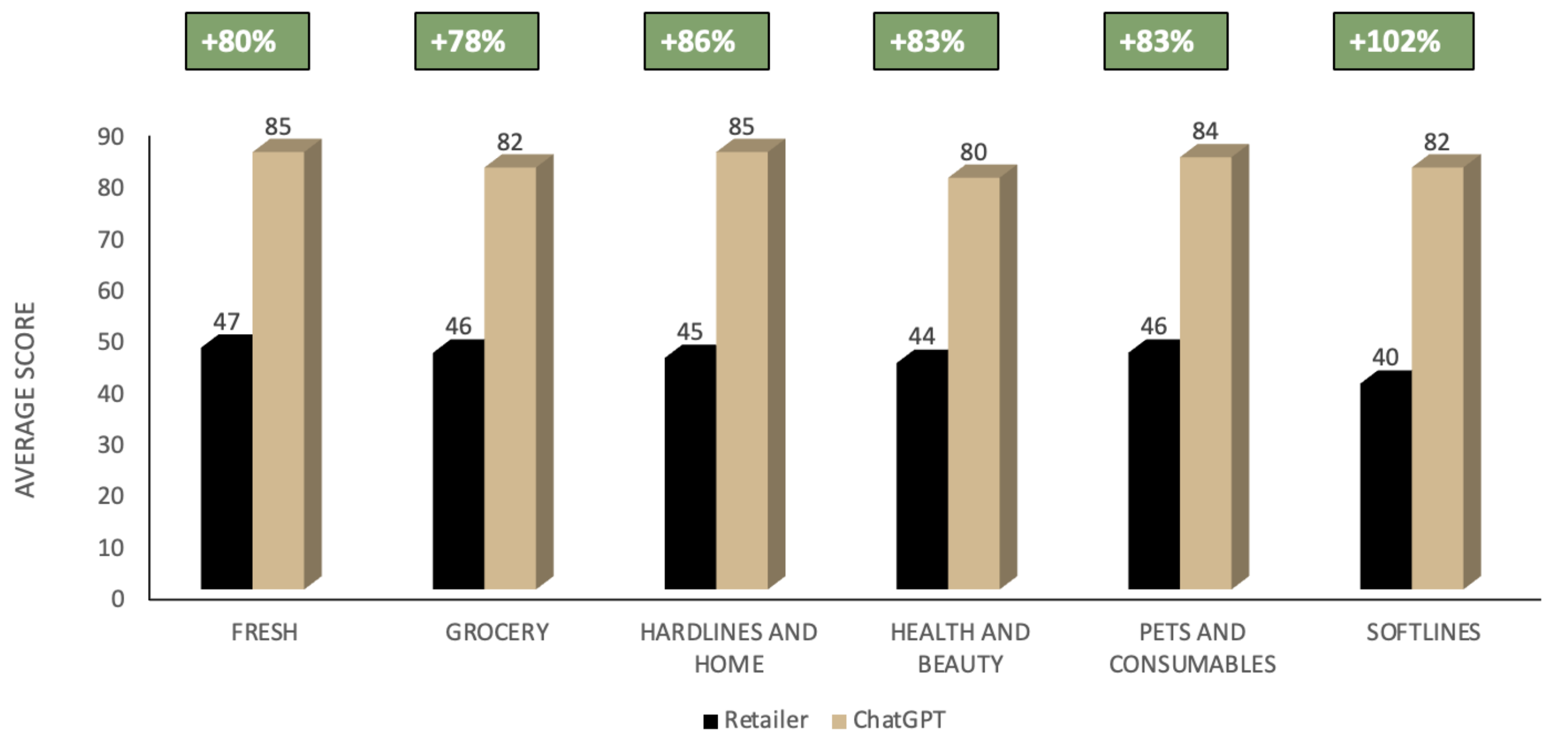

Score existing descriptions
Raw Data (CSV)
Exploratory Data Analysis (Python, Tableau, Excel)
Data Cleaning and Outlier Treatment
Scoring Parameters
Retailer Description Scoring
<50
>50
Label 0
Retailer description approved

Create new descriptions
OCR - Azure Computer Vision
Product Description Generation - ChatGPT

Score ChatGPT descriptions and deploy model
ChatGPT description approved
Label 1
>50
ChatGPT Description Scoring
<50
Label 2
Vendor to update product information
Deployment
Future Scope - Generate Descriptions using BERT model

Fig. 3 Methodology

MODEL BUILDING



Overall Improvement
83%

Avg ChatGPT scores	Avg Retailer scores
83	45

Fig. 4 % Change (category-wise) in description scores

Our model shows an average of 83% improvement in the description score. There are still considerable cases mostly in case of 'Fresh' category, where the images do not have labels (no text) and it limits the ability of our model to generate a good quality description. This can be improved by encouraging the vendors to provide accurate descriptions

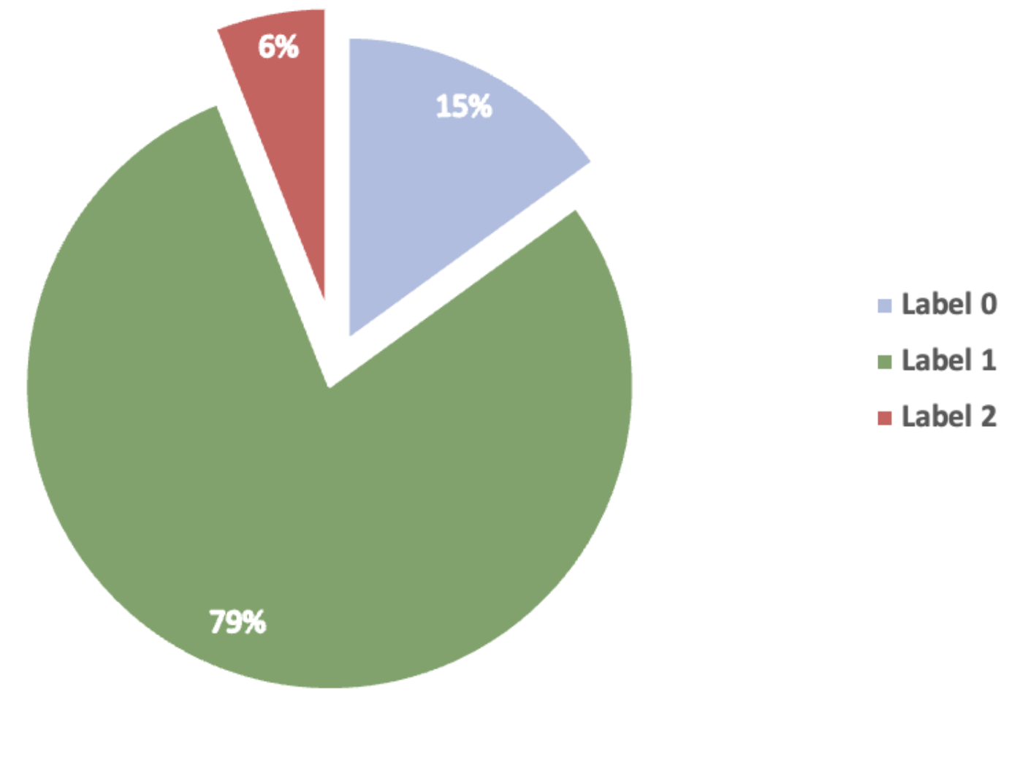


Fig. 5 Split of different labels

Our model generated high-quality descriptions for 79% of the products in our dataset. For 15% of the products, such as fresh produce without labels, the retailer's descriptions were deemed more accurate and were used instead. The remaining 6% of the products had poor descriptions both from the retailer and our model and require correction by the vendor.

PURDUE UNIVERSITY

Mitchell E. Daniels, Jr. School of Business

DEPLOYMENT & LIFE CYCLE MANAGEMENT

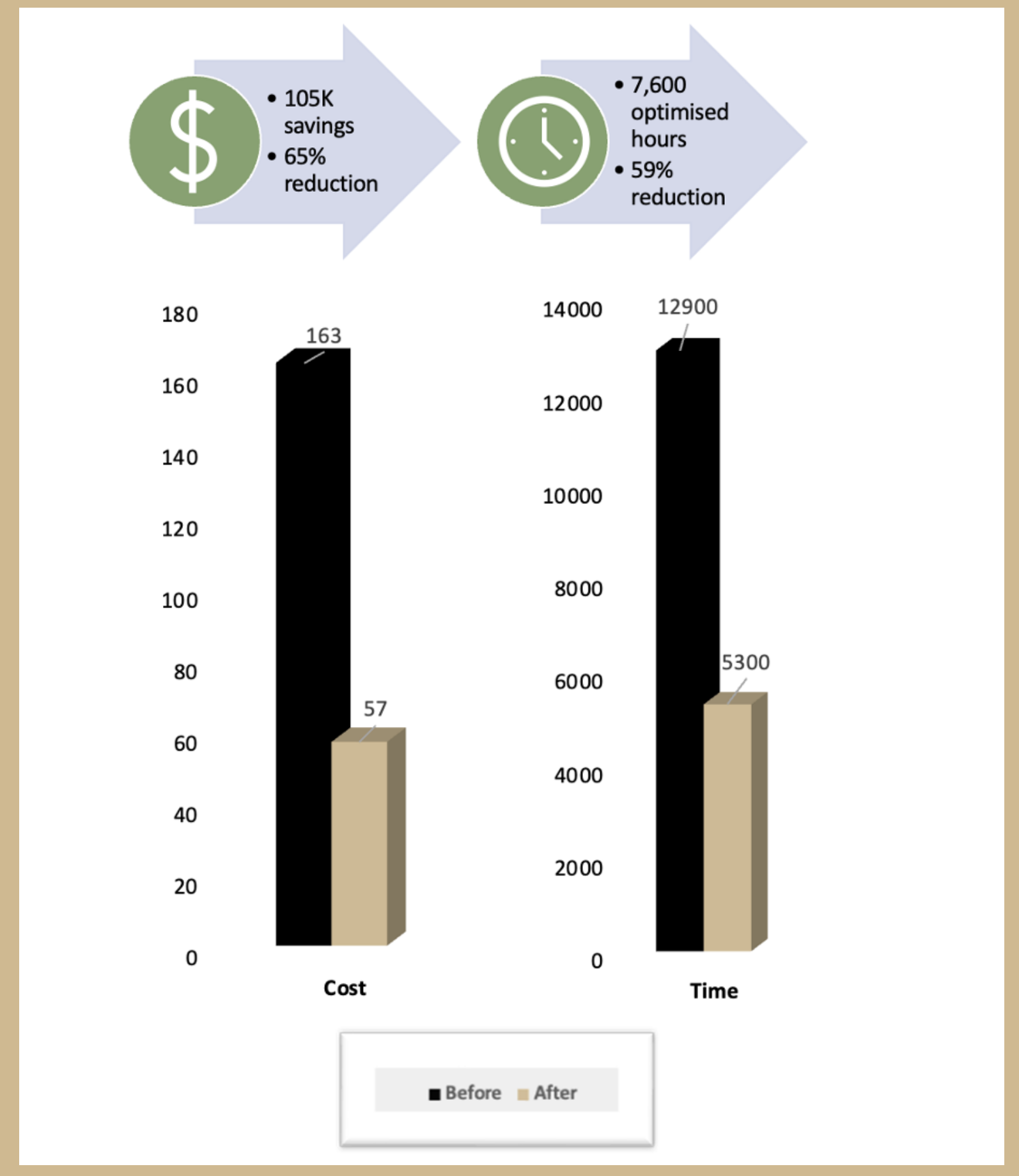


Fig. 6 Benefit analysis of the project, taken for a million products

Conclusion –

- The project used OCR, language model, and scoring algorithm to successfully improve the quality of product descriptions

Findings –

- 53% of products from subset of data were digitally ineligible for website listing
- 83% improvement in product description quality identified after using language model
- Potential savings of 65% and 59% in costs and time respectively
- Limited improvement in Fresh category due to lack of textual information

Future Scope –

- Increase revenue and improve customer experience by improving digital eligibility
- Scoring algorithm can be used to improve search engine optimization for product listings
- Success of the project can lead to collaborating between industry partners and academia

ACKNOWLEDGEMENTS

We would like to thank our industry partner for their guidance and support on this project as well as the Purdue MS BAIM program for partially funding this work.

informs BUSINESS ANALYTICS CONFERENCE