ZEKELMAN SCHOOL OF BUSINESS & INFORMATION TECHNOLOGY

Data Analytics for Business

ICT IGNITE DATA ANALYSIS

Aarti Anil Zikre (0825897)
Andrews Truman (0824852)
Roshan Khan(0826489)
Vitthlesh Sheth (0825950)
Premkumar Patel (0829257)

Group No: 09

TABLE OF CONTENTS

INTRODUCTION

Many eCommerce retailers and business owners implement free return policies to enhance customer satisfaction and drive revenue. Nevertheless, such policies entail substantial financial implications that can impact business profitability. According to the National Retail Federation (NRF), online consumers return up to 30% of their purchases, emphasizing the importance of accommodating return processes. Despite fostering customer loyalty and maintaining competitiveness, the expenses incurred from returns can be substantial, leading to potential losses. In 2022, the NRF reported an average return rate of 16.5%, amounting to $816 billion in returns. This translates to an average of $165 million in returns for every $1 billion in sales. Moreover, the NRF noted that consumers returned $218 billion worth of online purchases last year, with 10.6% ($23.2 billion) categorized as fraudulent returns.

Why Returns Cost So Much

In assessing the cost of returns, it's vital to consider not only the direct refund expenses incurred by the company but also various hidden costs that significantly contribute to the overall expense of returns.

These hidden costs encompass:

- **Shipping Costs:** This encompasses the expenses associated with managing the entire carrier network, including the logistics of picking up and delivering returns and transporting them to collection warehouses.
- **Customer Service Calls:** This includes the time, effort, and financial resources allocated by customer service teams to handle customer returns, process refunds, and follow up on return inquiries.
- **Warehousing and Refurbishing Costs:** This covers expenses related to leasing warehouse space for storing returned items and employing personnel to assess the condition of returned merchandise and authorize return requests.
- **Value Depreciation of Returned Products:** Certain products, particularly seasonal items, lose value the longer they remain in the return process. This diminishes the potential for resale at full price, often necessitating discounted pricing strategies.
- **Environmental Impact:** Approximately 10% of all returns ultimately end up in landfills, resulting in wasted resources used for packaging, shipping, and repackaging.

INTRODUCTION

What Current Customers Expect When It Comes to Returns

Consumer expectations regarding returns are driven by the desire for convenience and simplicity. They anticipate:

- A hassle-free, no-questions-asked return policy to facilitate easy returns.
- A reasonable timeframe post-purchase to make return decisions and initiate the return process.
- Free returns, eliminating any financial burden on the consumer.
- A streamlined return process, preferably through courier collection services, removing the necessity to visit post offices for returns.
- Full refunds for unsatisfactory merchandise, reflecting their dissatisfaction.
- Preference for receiving refunds in the form of credit or cash, ensuring flexibility.
- Effective and consistent communication at every stage of the return process, promoting transparency and trust.
- Swift resolution and processing of returns, prioritizing efficiency and convenience.

Finding A Balance with Customer Satisfaction and Returns

- Optimizing the returns process is pivotal for enhancing operational efficiency, elevating customer satisfaction, and reducing return-related costs. Analyzing returns data provides valuable insights into process inefficiencies that impact profitability. It's crucial to use this data to refine the approach while meeting customer expectations and safeguarding revenue.
- Start by pinpointing the main cost drivers in the returns process, such as shipping, warehouse operations, or customer service. Once the areas are identified, develop targeted strategies to tackle them effectively. For example, if shipping costs are high, consider teaming up with third-party logistics partners or diversifying carrier options.
- Automation is another game-changer for streamlining returns from start to finish, slashing processing times by up to 90%. However, building internal systems can be costly and time-consuming. That's where specialized returns automation software like Returns Automation comes in handy.

OBJECTIVE

In our pursuit of optimizing returns, we've partnered with ReturnPal, a courier service specializing in facilitating returns from customers to online platforms. Our primary focus in this project is to analyze purchase patterns and identify items with a high likelihood of being returned.

Amidst the rapidly evolving landscape of online commerce, businesses are constantly seeking insights to capitalize on emerging opportunities. This project is dedicated to the realm of online retail, with a specific emphasis on empowering ReturnPal, a company dedicated to simplifying the return process for online shoppers. Our objective is crystal clear: to equip ReturnPal with actionable insights, enabling them not only to navigate the complexities of online shopping but also to proactively seize growth opportunities in the ever-changing marketplace.

1. **Obtain the Data**
2. **Clean the Data**
3. **Exploratory Data Analysis**
4. **Modeling**

BUSINESS PROBLEM

eCommerce retailers and business owners face significant financial implications due to high return rates, impacting profitability. The challenge lies in effectively managing return processes to minimize losses and maintain competitiveness in the online market.

Client

ReturnPal

Objective

Our objective is to gain insights into predicting which purchases are likely to be returned and identifying the categories prone to returns. By understanding purchase patterns and return trends, we aim to optimize our return management processes and minimize associated costs.

RETURNPAL

OBTAINING THE DATA

The initial step in addressing our business problem involves acquiring pertinent data. Numerous accessible data sources on the internet offer valuable insights, with our dataset sourced from Reddit. Comprising 20,000 rows and 15 columns, our dataset includes essential details such as product ID, purchase date, product description, product category, and return status. This comprehensive dataset serves as the foundation for our analysis, enabling us to delve into purchase patterns and identify potential returns. Below is a snapshot of our raw dataset. Through meticulous analysis of this data, we aim to extract actionable insights to optimize our return management processes and enhance overall business performance.

ca982eca8304150849735ffef9	2016-03-25 22:59:23 +0000	http://www.flipkart.com/alisha-solid-women-s-c...	Alisha Solid Women's Cycling Shorts	["Clothing >> Women's Clothing >> Lingerie, Sl...	SRTEH2FF9KEDEFGF	999.0
3d550aaa89d34c77bd39a5e48	2016-03-25 22:59:23 +0000	http://www.flipkart.com/fabhomedecor-fabric-do...	FabHomeDecor Fabric Double Sofa Bed	["Furniture >> Living Room Furniture >> Sofa B...	SBEEH3QGU7MFYJFY	32157.0
5dcbc041b6ae5e6a32717d01b	2016-03-25 22:59:23 +0000	http://www.flipkart.com/aw-bellies/p/itmeh4grg...	AW Bellies	["Footwear >> Women's Footwear >> Ballerinas >...	SHOEH4GRSUBJGZXE	999.0
acd0c664e3de26e97e5571454	2016-03-25 22:59:23 +0000	http://www.flipkart.com/alisha-solid-women-s-c...	Alisha Solid Women's Cycling Shorts	["Clothing >> Women's Clothing >> Lingerie, Sl...	SRTEH2F6HUZMQ6SJ	699.0
a42ee6bef5ac7cea3fb5cfbee7	2016-03-25	http://www.flipkart.com/sicons-all-	Sicons All Purpose Arnica	["Pet Supplies >> Grooming >> Skin &	PSOEH3ZYDMSYARJ5	220.0

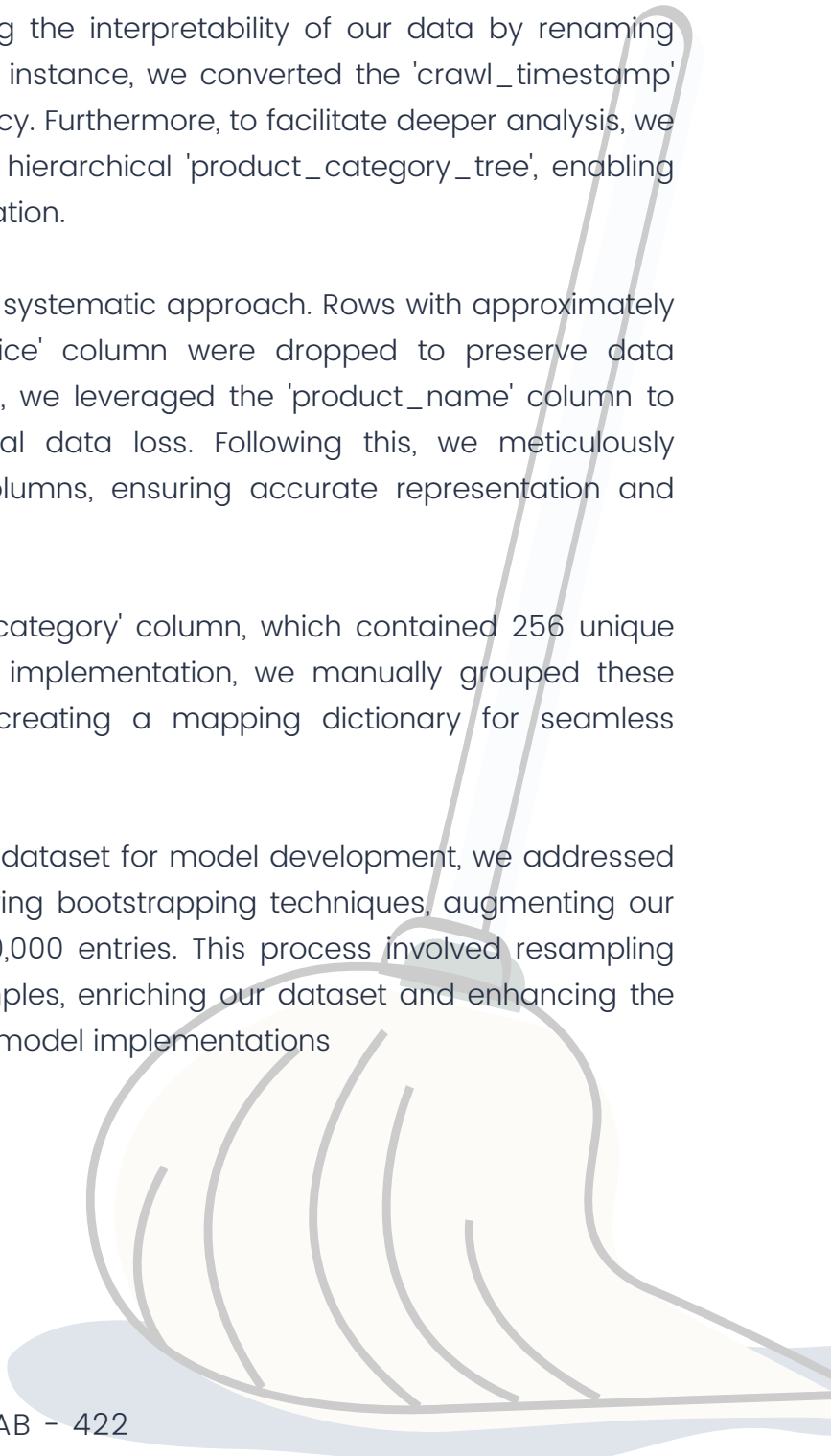
CLEANING THE DATA

Our data cleaning process commenced with the removal of redundant columns, including product URL, specifications, and image links, to streamline our dataset. Subsequently, we focused on enhancing the interpretability of our data by renaming columns to more descriptive labels. For instance, we converted the 'crawl_timestamp' column to 'date' for clarity and consistency. Furthermore, to facilitate deeper analysis, we derived subcategory columns from the hierarchical 'product_category_tree', enabling us to explore granular product categorization.

Addressing missing data, we adopted a systematic approach. Rows with approximately 80 missing entries in the 'product_price' column were dropped to preserve data integrity. For missing 'brand' information, we leveraged the 'product_name' column to extract brand names, ensuring minimal data loss. Following this, we meticulously standardized data types across all columns, ensuring accurate representation and efficient analysis.

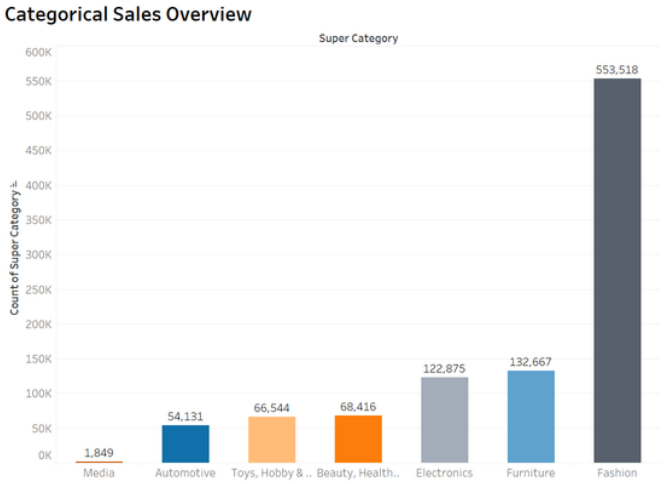
A significant challenge arose with the 'category' column, which contained 256 unique items. To simplify analysis and model implementation, we manually grouped these items into seven distinct categories, creating a mapping dictionary for seamless integration into our dataset.

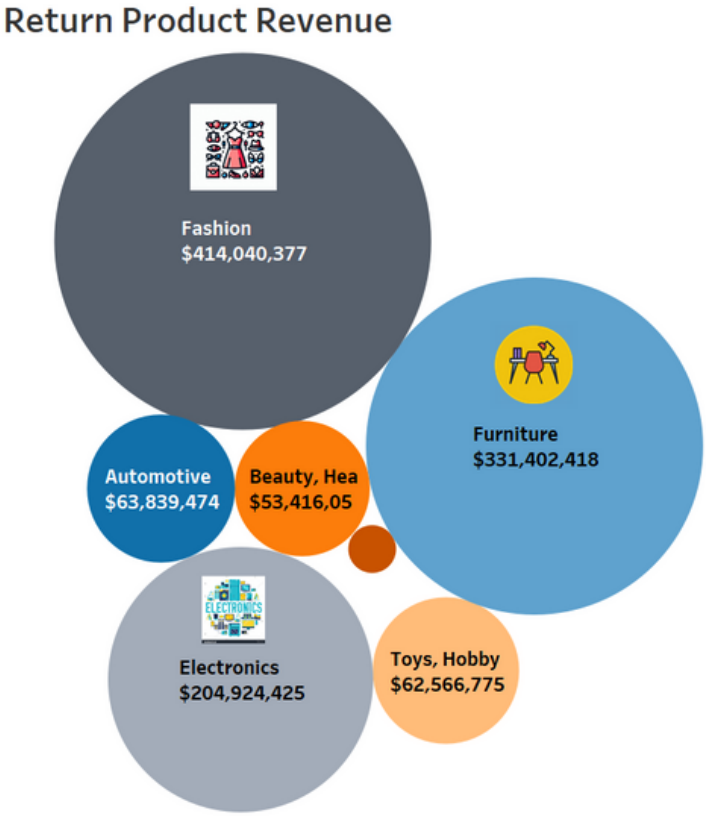
Recognizing the importance of a robust dataset for model development, we addressed the issue of insufficient rows by employing bootstrapping techniques, augmenting our dataset to a comprehensive size of 100,000 entries. This process involved resampling existing data to generate synthetic samples, enriching our dataset and enhancing the robustness of subsequent analyses and model implementations



EXPLORATORY DATA ANALYSIS

The graph represents sales counts in different super categories, with Fashion having the highest sales count and Media having the lowest

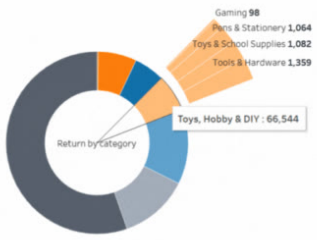


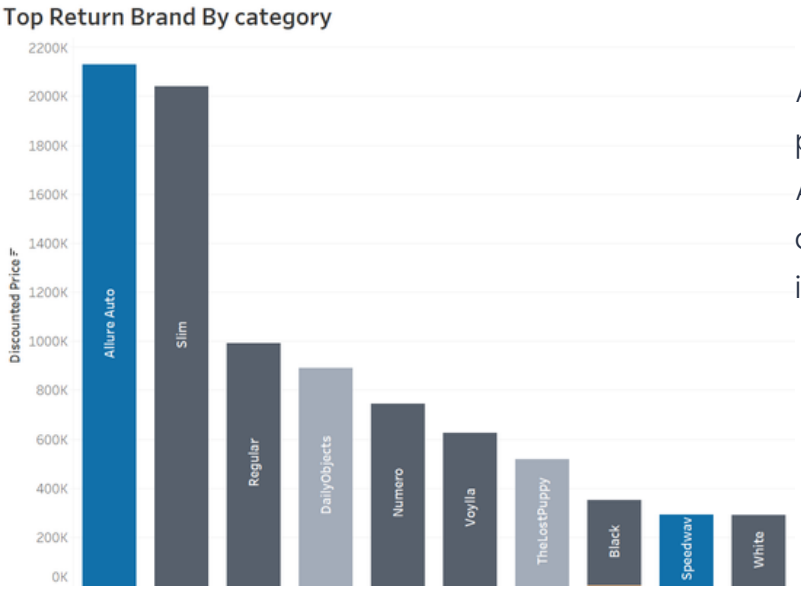


Returning products can indeed have a negative impact on a company, affecting both its revenue and reputation. Customers often feel dissatisfied and inconvenienced by the return process. This can result in a significant loss of revenue for the company. Fashion items tend to have the highest return rates, followed by electronics and furniture. Automotive and beauty & health products typically contribute smaller amounts to return-related revenue losses. For example, the revenue of a fashion organization may be impacted by $414,040,377 due to returns.

EXPLORATORY DATA ANALYSIS

The drill-down donut chart provides insight into which subcategories contribute to returns within the super category. For instance, within the overall fashion category, there were 553,518 returns. The top three main subcategories contributing to this total are clothing, jewelry, and bags, with contributions of 39,933, 16,352, and 1,735 returns respectively.





According to our data, the most popular brand in terms of returns is Allure Auto, which belongs to the automotive sector and returned 1,513 items.

MODELING

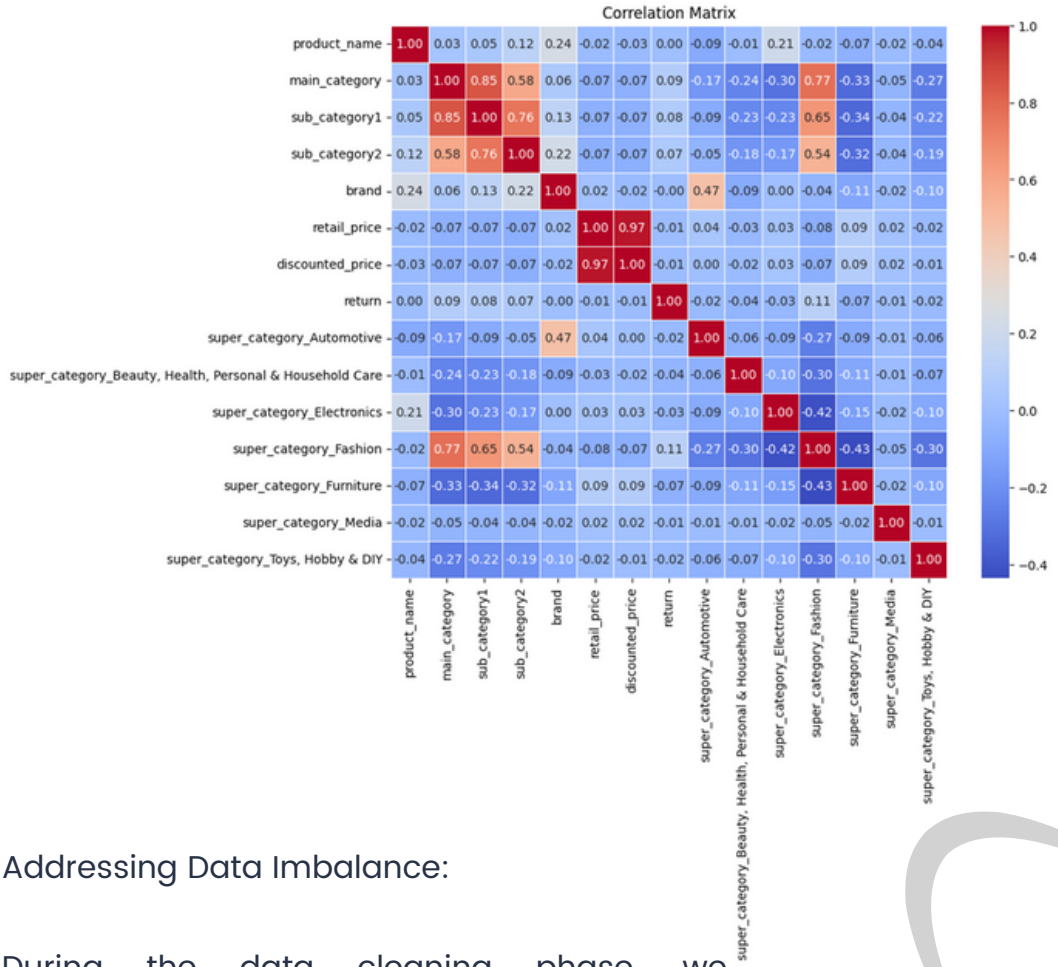
As we delve into category prediction, a crucial step involves transforming categorical variables into numerical ones. To achieve this, we employed encoding techniques tailored to our dataset's characteristics. For columns with more than 50 unique values, frequency encoding was applied to capture the frequency of each category, preserving valuable information without introducing excessive dimensionality. Conversely, for the super category column representing broader product categories, one-hot encoding was employed to create binary variables for each category, facilitating model interpretation and analysis.

	product_name	main_category	sub_category1	sub_category2	brand	retail_price	discounted_price	return	super_category_Automotive	supe
0	0.000272	0.137954	0.006176	0.006127	0.002659	4800.0	2650.0	0	0.0	
1	0.000319	0.137954	0.039177	0.023148	0.000747	1500.0	669.0	0	0.0	
2	0.000044	0.333291	0.208816	0.026003	0.000044	2499.0	999.0	0	0.0	
3	0.000301	0.333291	0.096015	0.048993	0.000350	999.0	519.0	0	0.0	
4	0.000055	0.333291	0.208816	0.105956	0.000055	799.0	399.0	0	0.0	

ice	return	super_category_Automotive	super_category_Beauty, Health, Personal & Household Care	super_category_Electronics	super_category_Fashion	super_category_Furniture	super_catego
0.0	0	0.0	0.0	0.0	1.0	0.0	
9.0	0	0.0	0.0	0.0	1.0	0.0	
9.0	0	0.0	0.0	0.0	1.0	0.0	
9.0	0	0.0	0.0	0.0	1.0	0.0	
9.0	0	0.0	0.0	0.0	1.0	0.0	

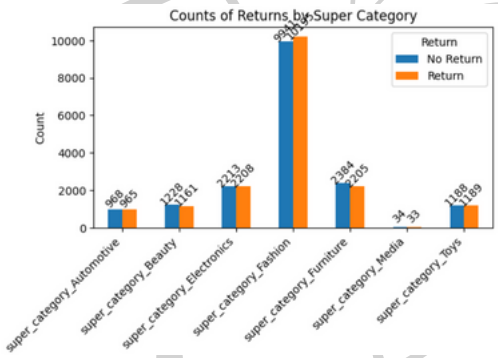
Before proceeding with model implementation, we conducted a thorough examination of variable relationships using a correlation matrix. This analysis provided valuable insights into the interplay between our target variable whether an item is returned and the input features. Notably, we observed positive correlations, particularly between subcategory 1 and the main category, suggesting potential dependencies that could influence return predictions. Additionally, correlations were evident among categorical variables, underscoring the importance of understanding how these variables interact in our predictive model. It provides insights into how variables co-vary with each other, indicating whether they move in the same direction (positive correlation) or opposite directions (negative correlation).

MODELING



Addressing Data Imbalance:

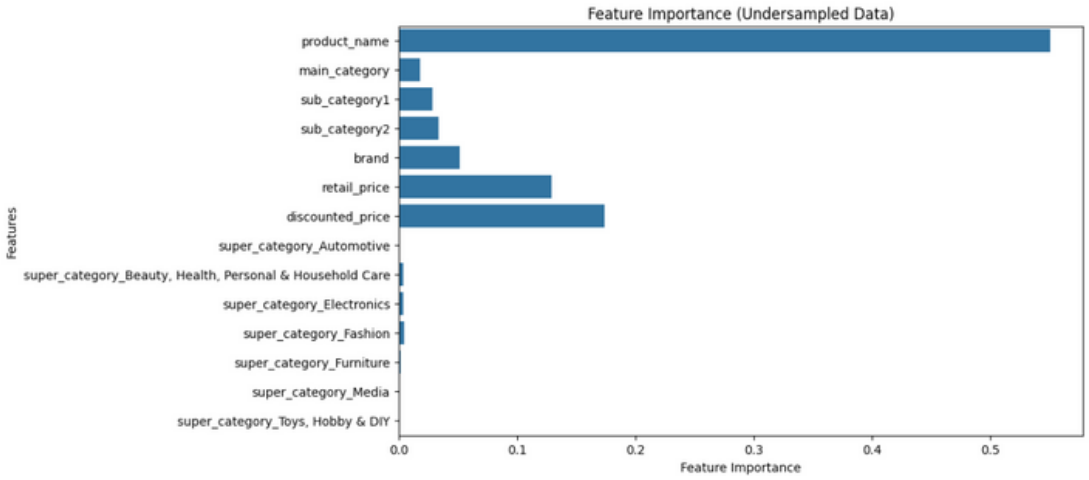
During the data cleaning phase, we encountered a significant imbalance in the distribution of the target variable, where the number of returned items was much lower than non-returned items. To mitigate this issue, we employed the undersampling technique. This involved randomly removing instances from the majority class (non-returned items) to balance the class distribution, ensuring that the model is not biased towards predicting the majority class.



Modeling

Feature Engineering:

In addition to the steps mentioned earlier, we performed feature engineering to enhance the predictive power of our model. This involved creating new features or transforming existing ones to better capture the underlying patterns in the data. For example, we derived new features from existing ones, such as extracting temporal features from the purchase date or engineering interaction features between categorical variables.



SVM Model:

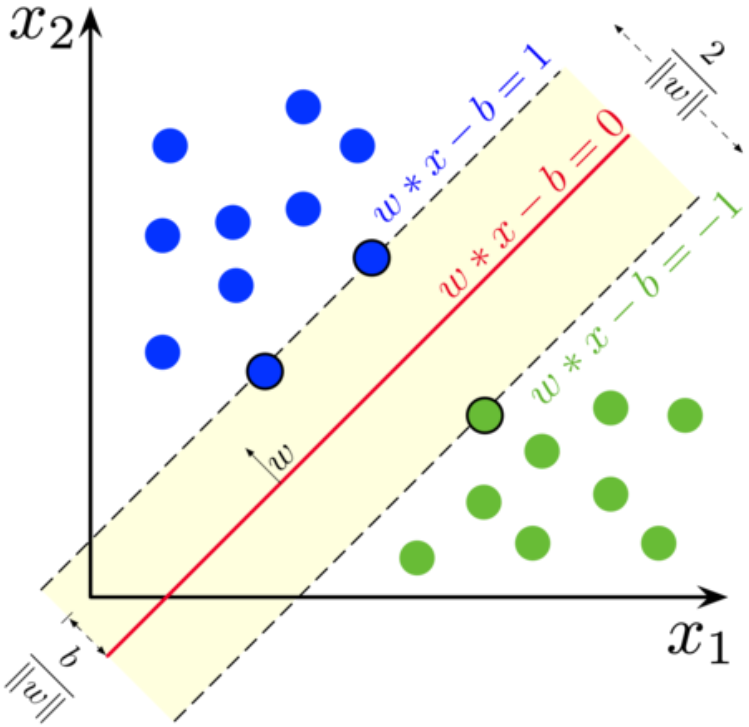
After addressing data imbalance and performing feature engineering, we implemented the Support Vector Machine (SVM) algorithm for classification. SVM is well-suited for binary classification tasks and offers robust performance in high-dimensional spaces. However, due to the nature of our dataset and the complexity of the problem, we observed a moderate accuracy score of 51% with the SVM model.

```
Accuracy with SVM: 0.5071697062508701

Classification Report:
              precision    recall  f1-score   support

           0       0.50      0.44      0.47      3544
           1       0.51      0.57      0.54      3639
```

MODELING



Model Evaluation Metrics:

Upon evaluating the SVM model, we computed several performance metrics to assess its effectiveness in predicting returns. The accuracy of the model was found to be 51%, indicating the proportion of correctly classified instances out of the total. Additionally, we calculated the F1 score, which considers both precision and recall, yielding a value of 51 for returned items. Furthermore, the recall score for returned items was 57%, indicating the proportion of actual returned items that were correctly identified by the model.