CS519 Group Project - Gamma
Stage 3 Report

Indronil Bhattacharjee, Roscoe Hill, and Erica Flores

1. Motivation

E-commerce defines the activity of buying or selling goods through an online, digital platform which has become a critical part of the global economy. Retailers of every size, from small home-based businesses to large corporate entities utilize online selling platforms to increase their reach to new customers. This is particularly important for small businesses, who can avoid the large overhead costs of a brick-and-mortar storefront.

According to the article "38 eCommerce Statistics of 2023" (Forbes Advisor Online):

- By 2026, the e-commerce market is expected to total over $8.1 trillion.
- By 2026, 24% of retail purchases are expected to take place online.
- 20.8% of retail purchases are expected to take place online in 2023.

2. Problem Definition

E-commerce does come with the caveat of trying to sell an item based on listing information and photos to a customer who cannot physically interact with the item before deciding to purchase. A business involved in e-commerce must consider which attributes of their products or services are most critical to translating an online shopping session into a purchase.

3. Machine Learning Tasks (Refined)

3.1 Online Shoppers Purchasing Intention

<u>Classification</u>

We can utilize the 'Revenue' attribute (0 or 1) already incorporated into the dataset, which indicates whether a session ended in a sales transaction, where 0 is the negative class (unsuccessful) and 1 is a positive class (successful).

3.2 Sales of Summer Clothes in E-commerce Wish

<u>Classification</u>

While there is no readily available class label attribute for this dataset, we can create one by converting a numerical attribute into a categorical one. For example, we can convert the 'units_sold' numerical attribute into a category class label that indicates how well the products sell and classify items as Top, Mid, and Bottom Tier products based on units sold.

We could possibly cluster merchants and products based on rating counts to determine which products and business will appear to audiences first.

<u>Regression</u>

How do the other attributes change when products utilize additional paid ad boosts within the website (uses_ad_boosts attribute) or offer express shipping (shipping_is_express attribute).

4. Datasets

4.1 Online Shoppers Purchasing Intention Dataset

4.1.1 Instances and Features

- Number of Instances: 12,330
  - 10,422 Negative Class and 1908 Positive Class
- Number of Features: 18
  - 10 numerical and 8 categorical attributes
  - 'Revenue' attribute used as class label
- Missing Data: None

4.1.2 Attribute Descriptions

Attribute	Description	Data Type
Administrative	Number of page visits	Integer
Administrative_Duration	Duration of visit (seconds)	Integer
Informational	Number of page visits	Integer
Informational_Duration	Duration of visit (seconds)	Integer

ProductRelated	Number of page visits	Integer
ProductRelated_Duration	Duration of visit (seconds)	Continuous
BounceRates	Percentage of visitors who enter the site from that page and then leave ("bounce") without triggering any other requests to the analytics server during that session	Continuous
ExitRates	The percentage that were the last in the session	Continuous
PageValues	Average value for a web page that a user visited before completing an e-commerce transaction	Integer
SpecialDay	Closeness of the site visiting time to a specific special day (e.g. Mother's Day, Valentine's Day)	Integer
Month	Month of visit	Categorical
OperatingSystems	OS of the visitor	Integer
Browser	Browser of visitor	Integer
Region	Geographic Region of Visitor	Integer
TrafficType	Traffic Source	Integer
VisitorType	Returning or New Visitor or Other	Categorical
Weekend	Weekend or Not Weekend	Binary
Revenue	Class Label whether a session ends in a transaction (positive) or not (negative)	Binary

Table 1: Attribute description of Online Shoppers Purchasing Intention dataset

4.1.3 Statistics

Histogram plots were made showing distribution of the values of various feature found in the dataset.

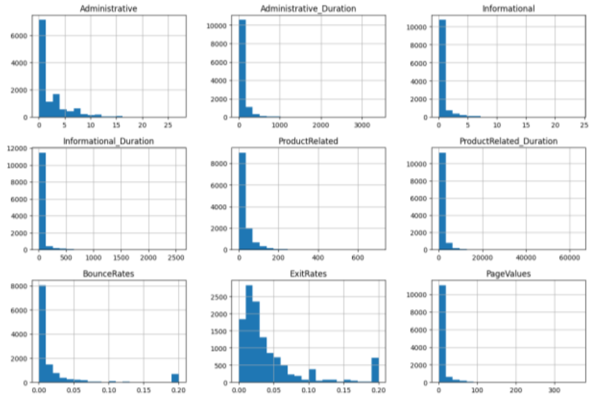

Figure 1. Distribution of numerical variables

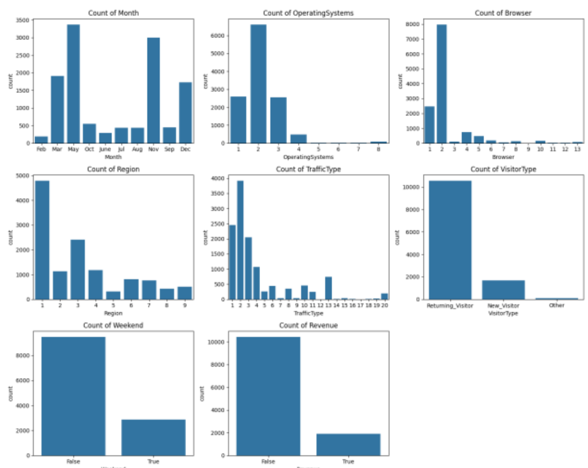

Figure 2. Distribution of categorical variables

4.1.4 Preliminary Data Analysis

Timing is a critical component of ensuring successful sales and understanding when the 'busy' and 'slow' seasons occur would benefit any business. Based on this quick analysis, we can see that the busiest times for this particular business peak in May and November. This may be due to the occurrence of celebrations and holidays, where graduations, Mother's and Father's Day occur during the late Spring peak and the peak of the winter holiday shopping season in late Fall.

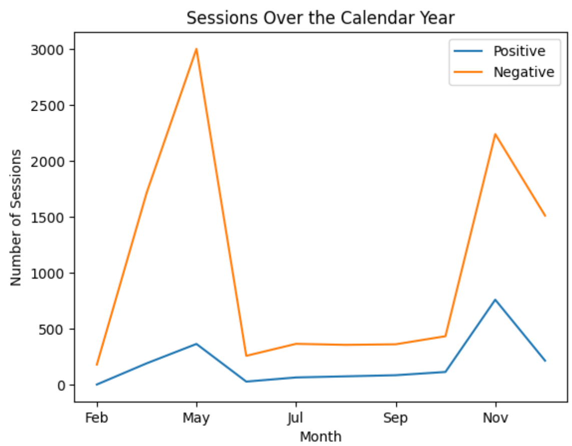

Figure 3: Session over the calendar year

We also performed a correlation analysis, allowing us to evaluate the relationship between two variables in a dataset. Values obtained in the matrix indicate the direction of the correlation (positive or negative) and the strength of the correlation, indicated by the heatmap colors.

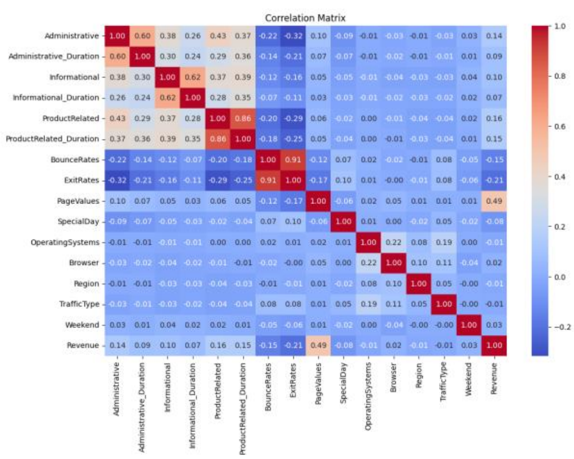

Figure 4: Correlation matrix

4.1.5 Preliminary Machine Learning Task for Online Shoppers Purchasing Intention

- Classification

Utilizing the 'Revenue' attribute, we can classify sessions as either successful (positive class) or unsuccessful (negative class) and here we compare several different models. The hyperparameter settings are summarized in the table below and the confusion matrixes are also included.

Model	Parameters	Time Elapsed	Test Acc	Train Acc	CrossVal Score
Perceptron	max_iter=40, eta=0.1	0.01s	0.88	0.88	0.85
Non-Linear (RBF)	C=1.0, gamma=0.10	2.11s	0.89	0.85	0.85
KNN	n_neighbors=10	0.02s	0.89	0.94	0.86

Table 2: Preliminary Classification Comparison

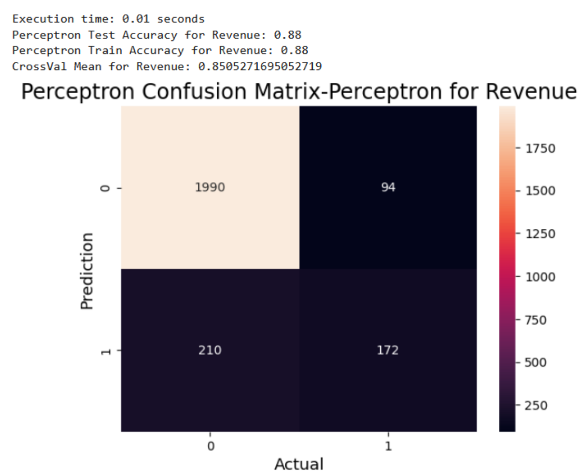

Figure 5: Perceptron confusion matrix

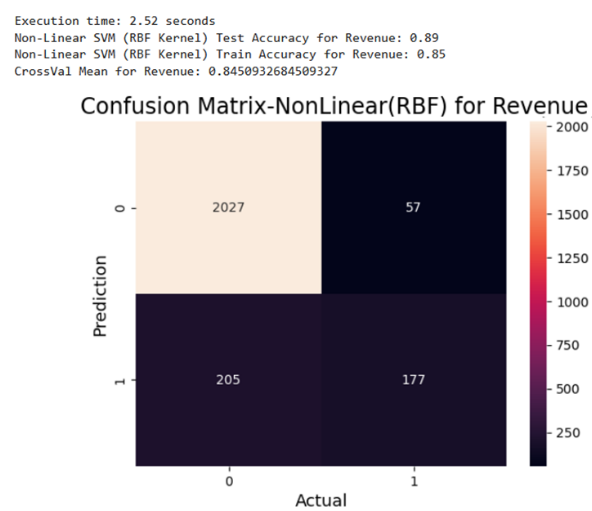

Figure 6: Non Linear (RBF) confusion matrix

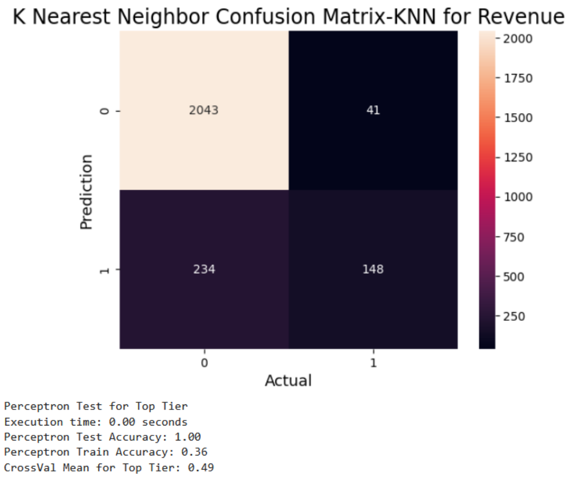

Figure 7: KNN confusion matrix

4.2 Summer Products with Rating and Performance 2020-08 Dataset

4.2.1 Instances and Features

- Number of Instances: 1,574
- Number of Features: 34
  - 24 numerical and 19 categorical attributes
- Missing Data: 45 products missing rating attribute data

4.2.2 Attribute Descriptions

Attribute	Description	Data Type
title	Title for localized for European countries	Categorical
title_orig	Original English title of the product	Categorical
price	price you would pay to get the product	Continuous
retail_price	reference price for similar articles on the market, or in other stores/places. Used by the seller to indicate a regular value or the price before discount.	Integer
currency_buyer	currency of the prices	Categorical
units_sold	Number of units sold. Lower bound approximation by steps	Integer
uses_ad_boosts	Whether the seller paid to boost his product within the platform	Binary
rating	Mean product rating	Continuous
rating_count	Total number of ratings of the product	Integer
rating_five_count	Number of 5-star ratings	Integer
rating_four_count	Number of 4-star ratings	Integer
rating_three_count	Number of 3-star ratings	Integer
rating_two_count	Number of 2-star ratings	Integer
rating_one_count	Number of 1-star ratings	Integer
badges_count	Number of badges the product or the seller have	Integer
badge_local_product	A badge that denotes the product is a local product. Conditions may vary (being produced locally, or something else). Some people may prefer buying local products rather than. 1 means Yes, has the badge	Binary
badge_product_quality	Badge awarded when many buyers consistently gave good evaluations 1 means Yes, has the badge	Binary

badge_fast_shipping	Badge awarded when this product's order is consistently shipped rapidly	Binary
tags	tags set by the seller	Categorical
product_color	Product's main color	Categorical
product_variation_size_id	One of the available size variation for this product	Categorical
product_variation_inventory	Inventory the seller has. Max allowed quantity is 50	Integer
shipping_option_name	Name of shipping option	Categorical
shipping_option_price	shipping price	Continuous
shipping_is_express	whether the shipping is express or not. 1 for True	Binary
countries_shipped_to	Number of countries this product is shipped to	Integer
inventory_total	Total inventory for all the product's variations (size/color variations for instance)	Integer
has_urgency_banner	whether there was an urgency banner with an urgency	Binary
urgency_text	A text banner that appear over some products in the search results.	Binary
origin_country	Country of Origin	Categorical
merchant_title	Merchant's displayed name (show in the UI as the seller's shop name)	Categorical
merchant_name	Merchant's canonical name. A name not shown publicly. Used by the website under the hood as a canonical name. Easier to process since all lowercase without white space	Categorical
merchant_info_subtitle	The subtitle text as shown on a seller's info section to the user.	Categorical
merchant_rating_count	Number of ratings of this seller	Integer
merchant_rating	merchant's rating	Continuous
merchant_id	merchant unique id	Categorical
merchant_has_profile_picture	Convenience boolean that says whether there is a `merchant_profile_picture` url	Binary
merchant_profile_picture	Custom profile picture of the seller (if the seller has one). Empty otherwise.	URL
product_url	url to the product page	URL
product_picture	Url to product picture	URL
product_id	product identifier.	Categorical
theme	the search term used in the search bar of the website to get these search results.	Categorical
crawl_month	Metadata info	Date

Table 3: Attribute description of Summer Products with Rating and Performance dataset

4.2.3 Statistics

Histogram plots were made showing distribution of the values of some important features found in the dataset.

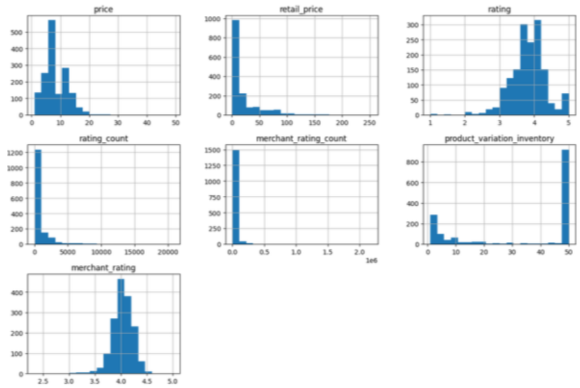

Figure 8. Distribution of some important variables

4.2.4 Preliminary Data Analysis

Collecting data on e-commerce is more about basic product listings. It's necessary to find data that helps with computing whether a product sells "well" or does not. By showcasing graphs of different columns from our dataset we can see that not all products have the same rating compared to others. Similarly the merchants also follow this trend. This can impact customer activity and sales of certain products. When listing products on an e-commerce application, it's imperative that the highest rated products with the highest rated merchants are shown first to customers.

4.2.5 Preliminary Machine Learning Task

Utilizing the units_sold' attribute, we can classify how well a product sells by categorizing the data into class labels. Where the Bottom Tier consists of products that are below 100 units sold, Mid Tier consists of the number of units sold between 100 and 5000, and Top Tier being any product with over 5000 units sold. Here we compare several different models. The hyperparameter settings are summarized in the table below and a sample of the confusion matrixes are also included.

Preliminary Classification Comparison					
Model	Parameters	Time	Test Acc Top Tier	Train Acc Mid Tier	CrossVal Score
Perceptron	max_iter=40, eta=0.1	Top Tier: 0.00 Mid Tier: 0.00 Bottom Tier: 0.00	Top Tier: 1.00 Mid Tier: 0.83 Bottom Tier: 1.00	Top Tier: 0.36 Mid Tier: 0.52 Bottom Tier: 0.84	Top Tier: 0.49 Mid Tier: 0.52 Bottom Tier: 0.95
Non-Linear (RBF)	C=1.0, gamma =0.10	Top Tier: 0.00-0.001 Mid Tier: 0.00-0.001 Bottom Tier: 0.00-0.001	Top Tier: 1.00 Mid Tier: 1.00 Bottom Tier: 1.00	Top Tier: 1.00 Mid Tier: 1.00 Bottom Tier: 1.00	Top Tier: 1.00 Mid Tier: 1.00 Bottom Tier: 1.00
KNN	n_neighbors= 10	Top Tier: 0.00 Mid Tier: 0.00 Bottom Tier: 0.00	Top Tier: 1.00 Mid Tier: 1.00 Bottom Tier: 1.00	Top Tier: 1.00 Mid Tier: 1.00 Bottom Tier: 1.00	Top Tier: 1.00 Mid Tier: 1.00 Bottom Tier: 1.00

Table 4: Preliminary Classification Comparison

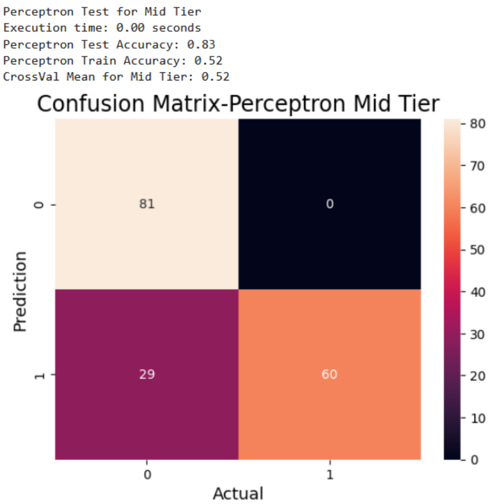

Figure 9: Perceptron Mid Tier confusion matrix

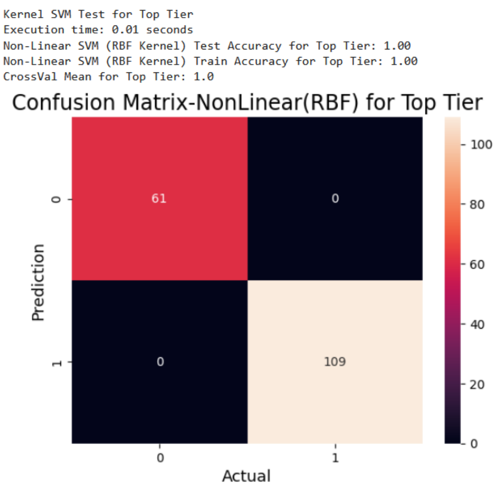

Figure 10: SVM Top Tier confusion matrix

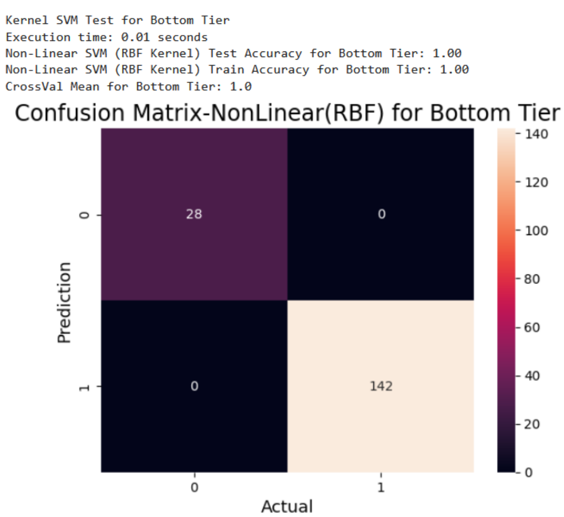

Figure 11: SVM Bottom Tier confusion matrix

4.2.6 Dataset Justification

We believe that both of these datasets have a desirable high number of instances and a mix of numerical and categorical attributes that allow for multiple avenues of analysis. We can perform a wide range of machine learning tasks with these datasets, including classification, regression, and clustering to provide recommendations for attributes that may contribute most to successful E-commerce transactions.

4.2.7 Initial Solution:

Our proposed solution is to incorporate more ML tasks and algorithms to answer the original problem: To consider which attributes of their products or services are most critical to translating an online shopping session into a purchase. We have already made good strides with our preliminary classifications and hope to expand more into regression. There are many attributes that both our datasets provide us that we can dive into. We hope to find more relationships that will produce an ideal solution to any e-commerce business.

References

Dataset Sources

1. Online Shoppers Purchasing Intention Data Set
https://archive.ics.uci.edu/dataset/468/online+shoppers+purchasing+intention+dataset

2. Sales of Summer Clothes in E-commerce Wish
https://www.kaggle.com/datasets/jmmvutu/summer-products-and-sales-in-ecommerce-wish/data?select=summer-products-with-rating-and-performance_2020-08.csv

Literature Sources

[1] Akarsh654, Machine Learning Project, 2020, GitHub Repository
https://github.com/Akarsh654/Machine-Learning-Projects/blob/master/Linear%20Regression/Ecommerce/Ecommerce%20Project.ipynb

[2] Baluch, A. “38 eCommece Statistics of 2023”, Forbes Advisor, 8 February 2023,
https://www.forbes.com/advisor/business/ecommerce-statistics/#sources_section

[3] Sakar, C. O., Polat, S. O., Katircioglu, M., & Kastro, Y. (2018). Real-time prediction of online shoppers' purchasing intention using multilayer perceptron and LSTM recurrent neural networks. In Neural Computing and Applications (Vol. 31, Issue 10, pp. 6893–6908). Springer Science and Business Media LLC. https://doi.org/10.1007/s00521-018-3523-0