CS519 Group Project - Gamma
Stage 4 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

3.1 Online Shoppers Purchasing Intention

Classification

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).

Classification - Feature Importance

Using a decision tree classifier, we can extract the feature_importance_ values to determine which were most important in training the model to differentiate between a successful and unsuccessful transaction.

3.2 Sales of Summer Clothes in E-commerce Wish

Classification - Product Performance

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.

Regression - Seller Performance

Utilizing the ‘total_units_sold’ attribute as a target variable, we can perform linear regression to predict seller success. We can also determine the coefficient, or feature importance value depending on the model, to determine which were most important for predicting seller performance in terms of total units sold.

4 DATASETS

4.1 Online Shoppers Purchasing Intention

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

Online Shoppers Purchasing Intention - 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.

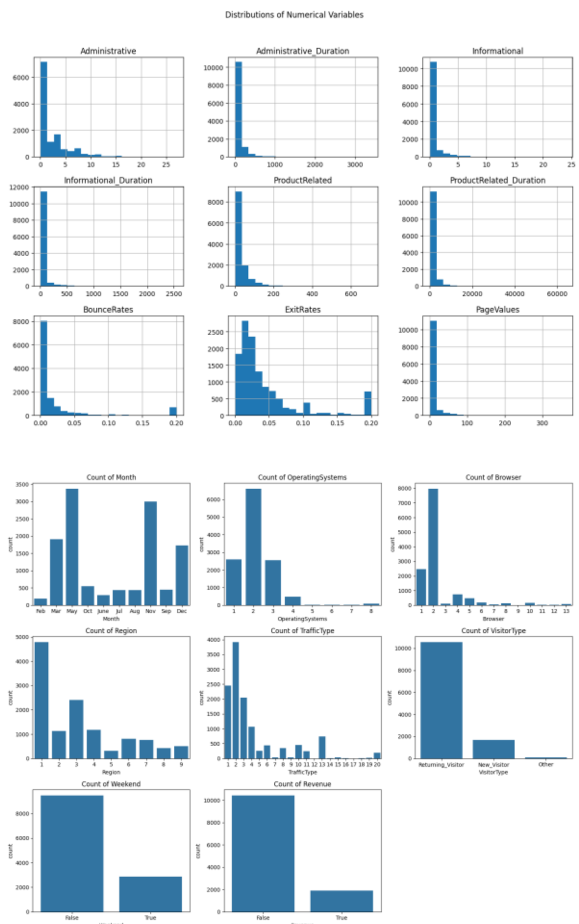

Figure 1. Distribution of variables

4.2 Summer Products with Rating and Performance 2020-08 Dataset

4.2.1 Instances and Features

Summer Products with Rating and Performance 2020-08 Data

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

Sales of Summer Clothes in E-commerce Wish - 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 2: Attribute description of Summer Products with Rating and Performance dataset

4.3 Computed Insight - Success of Active Sellers

4.3.1 Instances and Features

Computed Insight Success of Active Sellers Data Instances and Features

- Number of Instances: 958
- Number of Features: 13
    - 12 numerical and 1 categorical attributes
- Missing Data: 567 products missing urgency_count attribute data
    - Filled with zeros

4.3.2 Attribute Descriptions

Success of Active Sellers in E-commerce Wish - Attribute Descriptions		
Attribute	**Description**	**Data Type**
merchantid	Unique merchant (seller) ID	Categorical
listedproducts	Number of listed products	Integer
totalunitssold	Total units sold	Integer
meanunitssoldperproduct	Means units sold per product	Integer
rating	Seller rating	Continuous
merchantratingscount	Count of merchant ratings	Integer
meanproductprices	Mean product prices	Continuous
meanretailprices	Mean retail prices	Continuous
averagediscount	Average discount offered	Integer
meandiscount	Mean discount	Integer
meanproductratingscount	Mean count of product ratings	Integer
totalurgencycount	Total number of urgency messages	
urgencytextrate	Rate of urgency messages	Integer

Table 3: Attribute description of Success of Active Sellers

5 RESULTS

5.1 Summer Wish E-commerce Results

<u>Active Sellers Data</u>
The Summer Wish E-commerce dataset primarily includes product data, but also provides a dataset with insight into the merchant's performance, which is analyzed to determine the attributes that contribute most to being a successful seller in terms of 'total units sold'.

<u>Correlation Matrix</u>

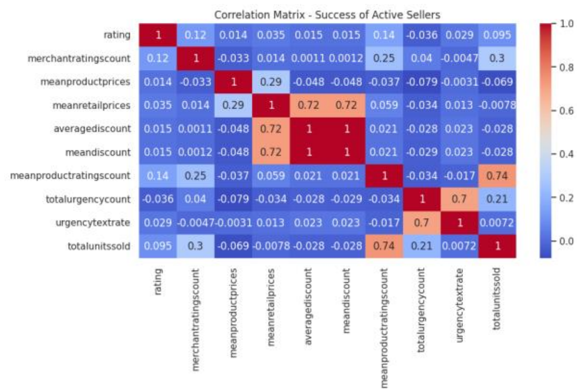

Figure 2. Correlation Matrix of Success of Active Sellers

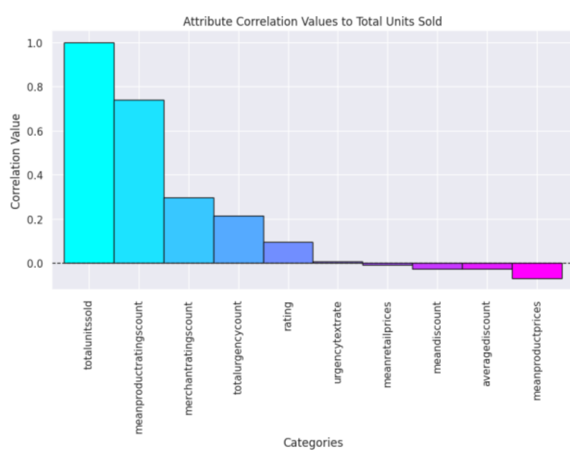

Figure 3. Attribute Correlation of Units Sold

<u>Linear Regression Models</u>
For this dataset, the 'TotalUnitsSold' feature was used as the target variable to predict based on all of the other features measured.

Two different regression models were used, a linear regressor and a random forest regressor.

5.1.1 Linear Regressor Results

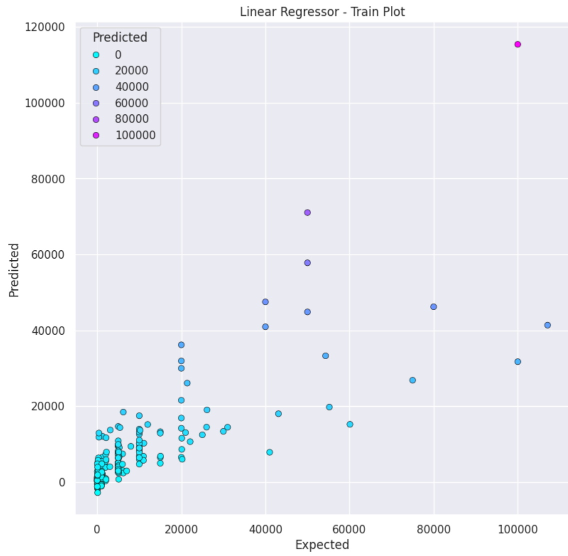

Figure 4. Linear Regression Train Plot

Fitting time: 0.00 seconds
MSE train: 67125961.667, **test:** 81614006.257
R^2 train: 0.654, **test:** 0.652
The R^2 score: 0.651
The MSE: 81614006.25
Coefficients:

	Attribute	Coefficient
0	rating	1143.77432
1	merchantratingscount	0.009708
2	meanproductprices	4.934802
3	meanretailprices	-27.927045
4	averagediscount	2047.384084
5	meandiscount	-2043.620307
6	meanproductratingscount	5.412371
7	totalurgencycount	8045.077184
8	urgencytextrate	-98.793026

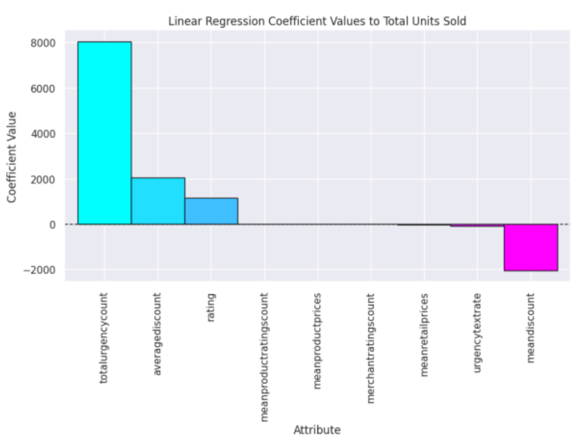

Figure 5. Linear Regression Coefficient Values

5.1.2 Random Forest Regressor Results

Fitting time: 0.97 seconds
MSE train: 7075697.592, test:36126721.495
R^2 train: 0.964, test:0.846
The R2 score: 0.845
The MSE: 36126721.494

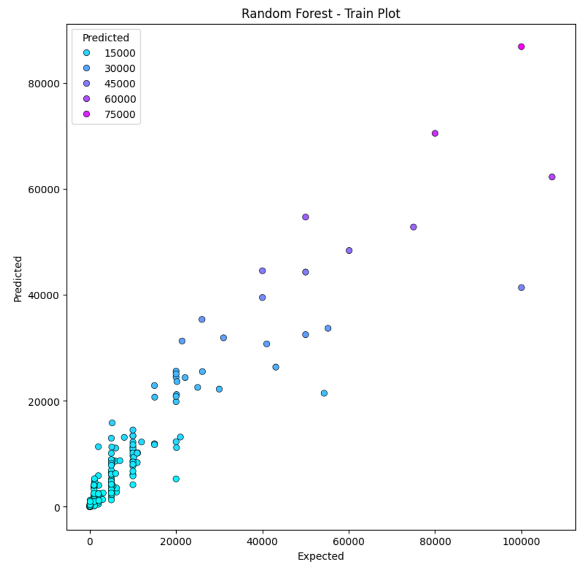

Figure 6. Random Forest Train Plot

Feature Importances:

	Attribute	Feature Importance
0	listedproducts	0.200331
1	rating	0.023183
2	merchantratingscount	0.058443
3	meanproductprices	0.011912
4	meanretailprices	0.015035
5	averagediscount	0.010815
6	meandiscount	0.009523
7	meanproductratingscount	0.655319
8	totalurgencycount	0.009080
9	urgencytextrate	0.006359

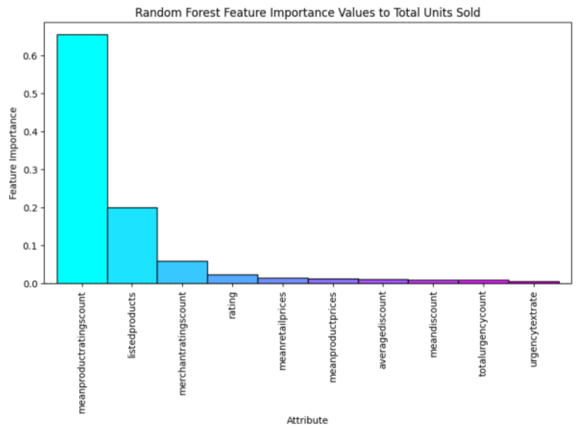

Figure 7. Random Forest Coefficient Values

Summer Wish Dataset - Active Sellers - Regression Performance Summary								
Model	**Parameters**	**Time Elapsed**	**MSE Train**	**MSE Test**	**R^2 Train**	**R^2 Test**	**MSE**	**R^2**
Linear Regression	Default	0.00s	67125961.66	8	0.65	0.65	81614006.25	0.65
Random Forest	max_depth=10	0.97s	7075697.59	36126721.49	0.96	0.84	36126721.49	0.845

Table 4: Summer Wish Dataset - Active Sellers - Regression Performance Summary

5.2 Product Performance Data

5.2 1 Classification

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.

5.2.2 Ensemble Methods vs Single Algorithms

Summer Wish Dataset - Product Performance - Classification Performance Summary		
Model	**Original Accuracy**	**Bagging Accuracy**
Perceptron	Top Tier: • Train: 0.36 • Test: 1.00 Mid Tier • Train: 0.52 • Test: 0.83 Bottom Tier • Train: 0.84 • Test: 1.00	Top Tier: • Train: 0.36 • Test: 1.00 Mid Tier • Train: 0.52 • Test: 0.83 Bottom Tier • Train: 0.84 • Test: 1.00
Non-Linear SVM (RBF Kernel)	Top Tier: • Train: 1.00 • Test: 1.00 Mid Tier • Train: 1.00 • Test: 1.00 Bottom Tier • Train: 1.00 • Test: 1.00	Top Tier: • Train: 0.36 • Test: 1.00 Mid Tier • Train: 0.52 • Test: 0.83 Bottom Tier • Train: 0.84 • Test: 0.84
K Nearest Neighbor	Top Tier: • Train: 1.00 • Test: 1.00 Mid Tier • Train: 1.00 • Test: 1.00 Bottom Tier • Train:	Top Tier: • Train: 0.36 • Test: 1.00 Mid Tier • Train: 0.52 • Test: 0.83 Bottom Tier • Train:

	1.00 • Test: 1.00	0.84 • Test: 1.00

Table 5: Summer Wish Dataset - Product Performance - Classification Performance Summary

Figure 8. Distribution of Numerical Variables for Summer Products Performance Data

5.3 Online Shopper Intention Results

Classification Comparison					
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
Random Forest	n_estimators=100	1.61	0.90	1.00	0.89

Table 6: Classification Comparison for Online Shopper Intention

5.3.1 Random Forest Classification

```
Execution time: 1.61 seconds
Random Forest Train Accuracy: 1.00
Random Forest Test Accuracy: 0.90
CrossVal Mean: 0.8967558799675588
```

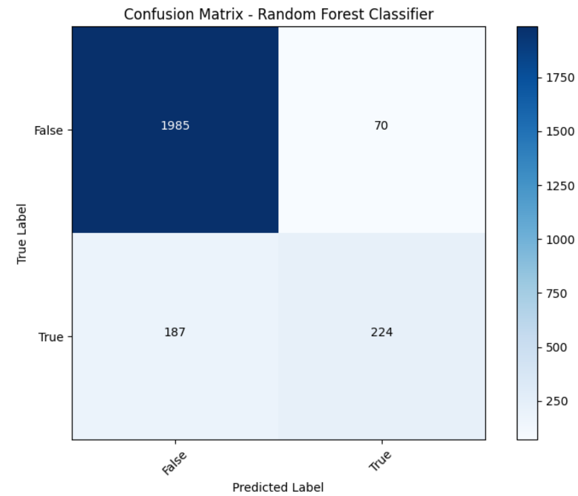

Figure 9. Confusion Matrix – Random Forest

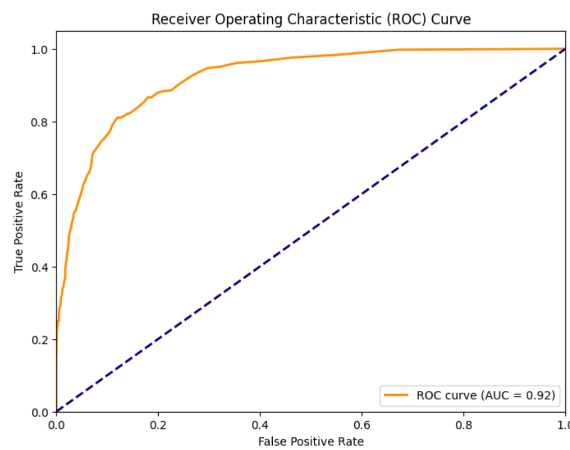

Figure 10. ROC Curve for Random Forest

5.3.2 Feature Importance for Classification

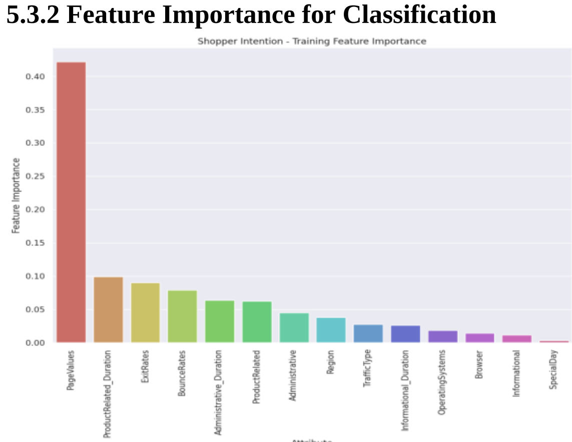

Figure 10. ROC Curve for Random Forest

Utilizing the ‘feature_importances_’ attribute of a Decision Tree Classifier, we were able to determine which features were most important for training the classifier to distinguish between a successful transaction and a non-successful one.

5.4 Ensemble methods

Online Shopping Dataset

Model	Original Accuracy	Bagging Accuracy
Perceptron	Train: 0.88 Test: 0.88	Train: 0.89 Test: 0.89
Non-Linear SVM (RBF Kernel)	Train: 0.85 Test: 0.89	Train: 0.85 Test: 0.89
K Nearest Neighbor	Train: 0.85 Test: 0.89	Train: 0.85 Test: 0.89

Table 7: Classification Comparison for Online Shopping Dataset

5.4.1 Analysis

For improvements in solutions, the team decided to implement and test ensemble algorithms on our current models. We wanted to see the outcome of possibly combining multiple individual models to improve predictive accuracy. We decided to utilize Bagging and have the base estimator be the model's original run for comparison. **What we concluded from our results is that because the accuracy did not change or even decrease for certain models and features, ensemble algorithms aren’t the answer to everything.** This helps not only us but other teams when trying to decide which avenues to take to improve their data analysis.

6 RESULT SUMMARY AND BUSINESS RECOMMENDATIONS

Below, we have summarized the findings from our machine-learning tasks and provided the associated business recommendations based on their results.

6.1 Online Shoppers Intention Dataset

The following recommendations relate to the success of an e-commerce transaction, based on the most important features of products for training the classification algorithm.

Task	Result Summary	Business Recommendations
Decision Tree Classifier **(Determine attributes most important for training the classifier)**	The top three attributes for classification importance are: 1. Page Values 2. Product Related Duration 3. Exit Rates	Page Value - Focus on promoting pages with high-value Product Duration- Maximize the amount of information available to increase duration time Exit Rate- Encourage customers to continue interacting on the page with possible similar products or incentives

Table 8: Result Summary and Business Recommendations for Online Shoppers Intention

6.2 Summer Wish Dataset

The following recommendations relate to the success of sellers, based on the most important features that are associated with more total units sold.

Task	Result Summary	Business Recommendations
Linear Regression	The top three attributes for positive coefficients are: 1.Count of Urgency Messages 2. Average Discount 3. Product Rating	Urgency Text Count- Consider adding urgency messages to products that currently do not display one Discount-Consider running short periods of sales or providing coupons for past customers Ratings - Encourage customers to leave

		ratings on products
Random Forest Regression	The top three attributes for feature importance are: 1. Mean Product Rating Count 2. Count of Listed Products 3. Count of Merchant Ratings	Ratings Count - Encourage customers to leave ratings for products and merchants Listings Count- Merchants should consider maximizing the amount of listed products they offer
Correlation Matrix	The top three features with a positive correlation are: 1. Mean Product Rating Count 2. Count of Merchant Ratings 3. Count of Urgency Messages	Ratings Count- Encourage customers to leave ratings for products and merchants Urgency Text Count- Consider adding urgency messages to products that currently do not display one

Table 9: Result Summary And Business Recommendations for Summer Wish Dataset

7 CONCLUSIONS

7.1 Online Shoppers Intention Dataset

From the feature importance analysis done on this dataset, we found that some of the top attributes that contributed to training the classification of successful and unsuccessful e-commerce transactions generally involved keeping the buyer engaged on the page. Keeping the buyer on your website is critical to working towards a successful transaction. We were able to compare multiple single algorithm classifiers, as well as determine that bagging did not affect accuracy.

7.2 Summer Wish Dataset

It was an interesting finding that based on the best-performing regression model, the number of product and merchant ratings appeared to be more important than the quality of the ratings. The linear regression model also shows the importance for the count of urgency messages that pop up on product pages. The correlation matrix strengthens the relationship between the number of merchant and product ratings and urgency ratings to the number of units sold by individual sellers.

8 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.

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