Understanding E-commerce: Applying Machine Learning to Inform Business Decisions

Indronil Bhattacharjee, Roscoe Hill, and Erica Flores
Department of Computer Science
New Mexico State University
Las Cruces, NM

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 - Product and Seller Performance

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

4 DATASETS

4.1 Online Shoppers Purchasing Intention

4.1.1 Instances and Features

Number of Instances	**Total**	**Negative**	**Positive**
	12,330	10,422	1,908
Number of Features	**Total**	**Numerical**	**Categorical**
	18	10	8
Class label	Revenue		
Missing Value	None		

4.2 Summer Products with Rating and Performance 2020-08

4.2.1 Instances and Features

Number of Instances	1,574		
Number of Features	**Total**	**Numerical**	**Categorical**
	34	24	10
Missing Value	45 rating attribute data		

4.3 Computed Insight - Success of Active Sellers

4.3.1 Instances and Features

Number of Instances	958		
Number of Features	Total	Numerical	Categorical
	13	12	1
Missing Value	67 urgency_count attribute data		

5 RESULTS

5.1 Summer Wish E-commerce Results - Active Sellers Data

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'.

Correlation Matrix

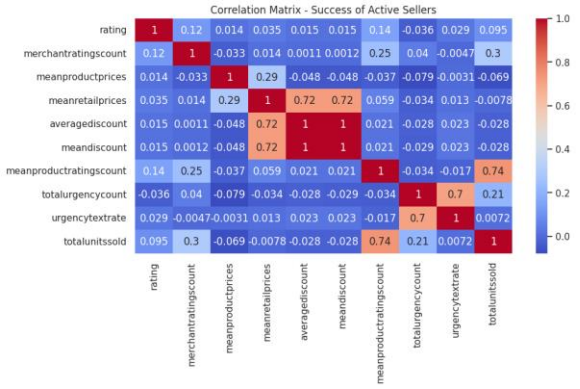

Figure 1. Correlation Matrix of Success of Active Sellers

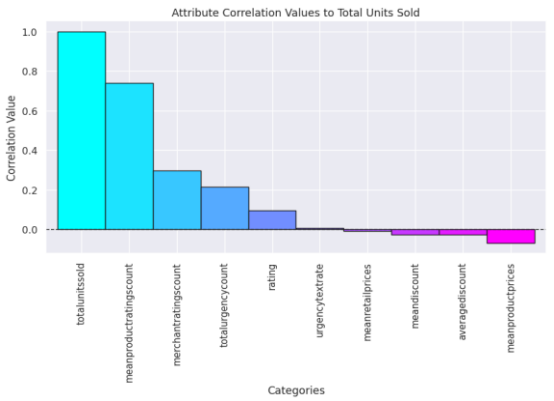

Figure 2. Attribute Correlation of Units Sold

Linear Regression Models

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

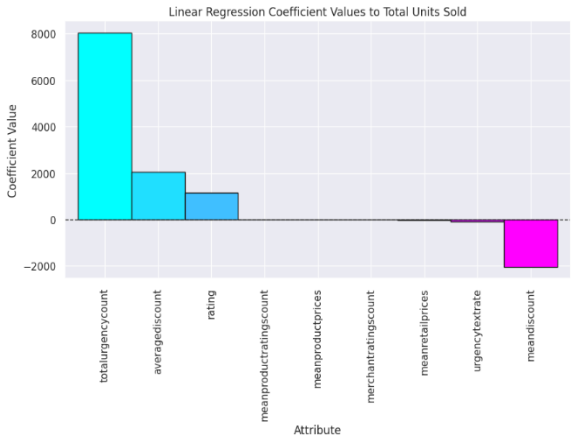

Figure 3. Linear Regression Coefficient Values

5.1.2 Random Forest Regressor Results

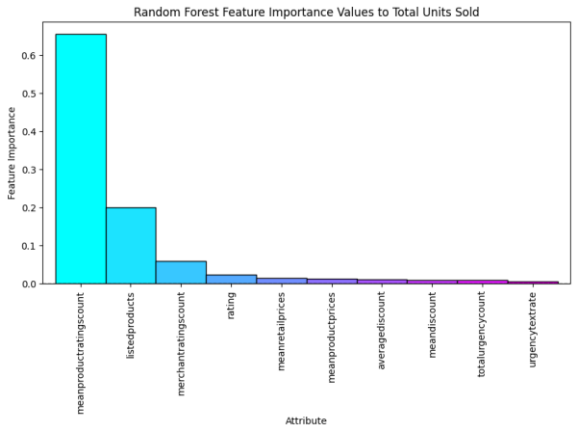

Figure 4. Random Forest Coefficient Values

The performance summary indicates that both linear regression and random forest models were evaluated for predicting sales of active sellers. Linear regression, despite its simplicity, yielded suboptimal results with a high MSE on both the training and testing sets, indicating not a very good fit to the data. In contrast, the random forest model, with a specified maximum depth of 10, exhibited significantly improved performance, achieving substantially lower MSE values and higher R2 scores, suggesting better predictive capability and capturing more variance in the data. The longer training time of the random forest

model is a trade-off for its superior performance compared to the linear regression model.

Summer Wish Dataset - Active Sellers - Regression Performance Summary								
Model	**Parameters**	**Training time**	**MSE Train**	**MSE Test**	**R^2Train**	**R^2Test**	**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

5.2 Summer Wish E-commerce Results - 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	Top Tier: • Train: 0.36 • Test: 1.00
	Mid Tier • Train: 0.52 • Test: 0.83	Mid Tier • Train: 0.52 • Test: 0.83
	Bottom Tier • Train: 0.84 • Test: 1.00	Bottom Tier • Train: 0.84 • Test: 1.00
Non-Linear SVM (RBF Kernel)	Top Tier: • Train: 1.00 • Test: 1.00	Top Tier: • Train: 0.36 • Test: 1.00
	Mid Tier • Train: 1.00 • Test: 1.00	Mid Tier • Train: 0.52 • Test: 0.83
	Bottom Tier • Train: 1.00 • Test: 1.00	Bottom Tier • Train: 0.84 • Test: 0.84
K Nearest Neighbor	Top Tier: • Train: 1.00 • Test: 1.00	Top Tier: • Train: 0.36 • Test: 1.00
	Mid Tier • Train: 1.00 • Test: 1.00	Mid Tier • Train: 0.52 • Test: 0.83
	Bottom Tier • Train: 1.00 • Test: 1.00	Bottom Tier • Train: 0.84 • Test: 1.00

5.2.3 Correlation Matrix

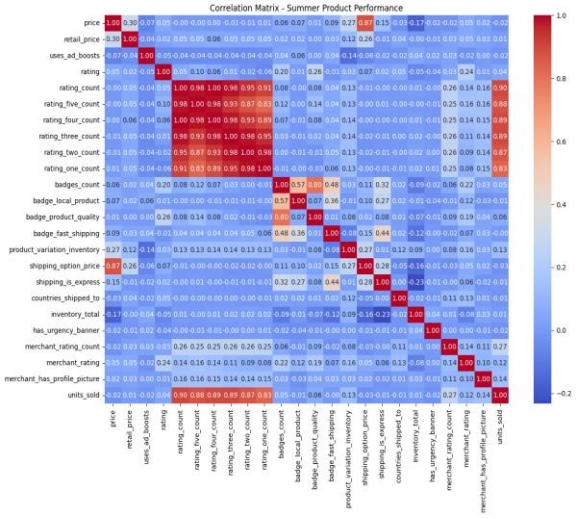

Figure 5. Correlation Matrix of Product Performance

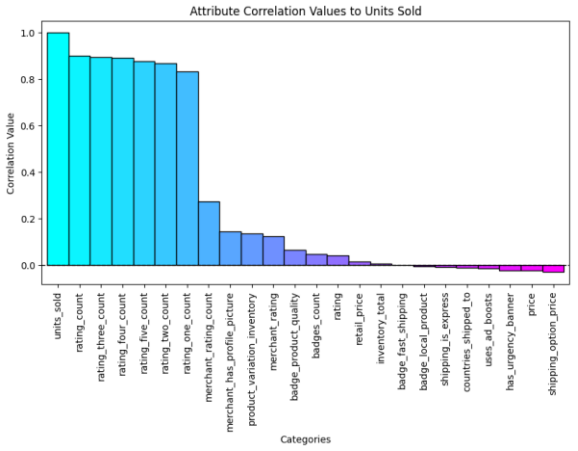

Figure 6. Attribute Correlation of Units Sold

Regression Analysis

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

5.2.3 Linear Regression Model

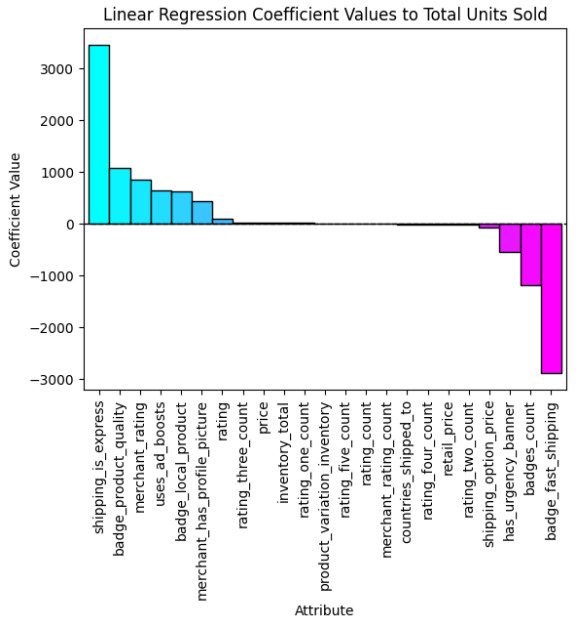

Figure 7. Linear Regression Coefficient Values

5.2.4 Random Forest Regression

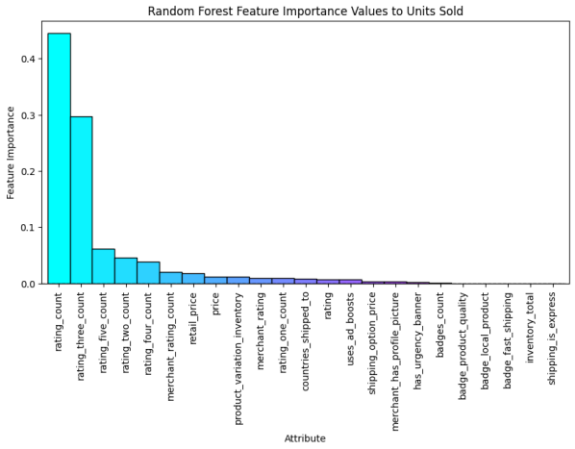

Figure 8. Random Forest Coefficient Values

Summer Wish Dataset - Products - Regression Performance Summary								
Model	**Parameters**	**Training time**	**MSE Train**	**MSE Test**	**R2 Train**	**R2 Test**	**MSE**	**R2**
Linear Regression	Default	0.01s	14591653.2	2.01E+06	0.783	0.848	2.01E+07	0.84
Random Forest	max_depth=20	1.22s	2423655.66	23057334.96	0.964	0.825	2.31E+07	0.825

The performance summary indicates that both linear regression and random forest models were evaluated on the Summer Wish Dataset for predicting product sales. Linear regression and the random forest model, with a specified maximum depth of 20, exhibited significantly similar type of performance, achieving substantially nearly same MSE values and similar R2 scores, suggesting better predictive capability and capturing more variance in the data. The longer training time of the random forest model is a trade-off for its superior performance compared to the linear regression model.

5.3 Online Shopper Intention Results

Classification Comparison					
Model	**Parameters**	**Training time**	**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

5.3.1 Random Forest Classification

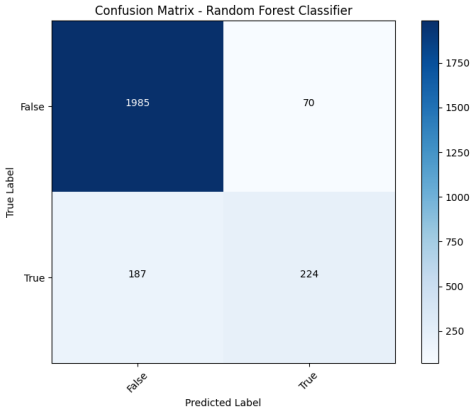

Figure 9. Confusion Matrix – Random Forest

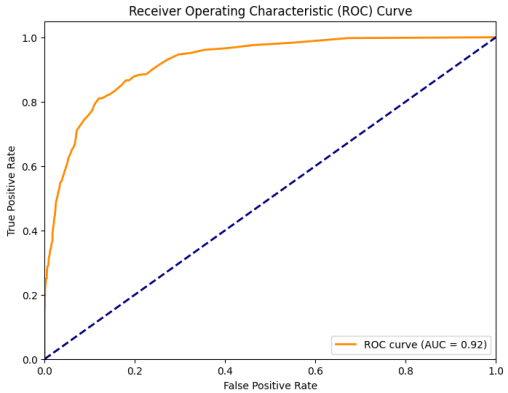

Figure 10. ROC Curve for Random Forest

5.3.2 Feature Importance for Classification

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.

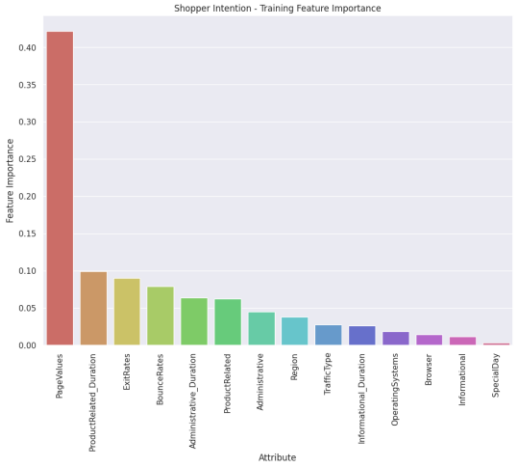

Figure 12. Feature Importance to classification

5.4 Ensemble methods

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

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. **We concluded 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

6.2 Summer Wish Dataset - Product Performance Data

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

Task	Result Summary	Business Recommendations
Linear Regression	The top three attributes for positive coefficients are: 1.Express shipping 2. Product Quality Badge 3. Merchant Rating	**Shipping-** Offer express shipping on as many items as possible
		Quality Badge- Increase listings with these badges
		Merchant Rating- Encourage customers to leave reviews/ratings
Random Forest Regression	The top three attributes for feature importance are: **1.** Rating Count **2.** Rating Count (2-5) **3.** Merchant Rating Count	**Ratings -** Encourage customers to leave ratings with incentives, merchant ratings help with product sales
Correlation Matrix	The top three features with a positive correlation are: **1.** Rating Count **2.** Merchant Rating Count **3.** Merchant Profile Picture	**Ratings -** Ratings for products and merchants both help
		Profile- Encourage merchants to upload a profile photo

6.3 Summer Wish Dataset - Active Sellers Data

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 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

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 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.

7.3 E-Commerce Conclusions

The importance of the total number of ratings for products and merchants was a trend seen in both datasets.We believe that the high number of ratings indicates that the product has been tested and reviewed by many other shoppers and, therefore may be a reliable purchase. Establish a trustworthy, quality storefront that can ship items quickly and incentivize customers to leave reviews and ratings to draw in more future customers.

8 DATASET JUSTIFICATION

The chosen datasets have a desirable high number of instances and a mix of numerical and categorical attributes that allow for multiple avenues of analysis. Through machine learning algorithms, we can provide recommendations that may contribute to successful E-commerce transactions.

REFERENCES

[1] Sakar, C. O., & Polat, S. (2018, October). Online Shoppers Purchasing Intention Data Set. Retrieved May 9, 2024 from https://archive.ics.uci .edu/dataset/468/online+shoppers+purchasing+intention+dataset

[2] Vu, J. (2020, August). Sales of Summer Clothes in E-commerce Wish (summer-products-with-rating-and-performance), Version 1. Retrieved May 9, 2024 from https://www.kaggle.com/datasets/jmmvutu/summer-products-and-sales-in-ecommerce-wish/ data?select=summer-products-with-rating-and-performance_2020-08.csv

[3] Vu, J. (2020, August). Computed Insight: Success of Active Sellers, Version 1. Retrieved Retrieved May 9, 2024 from https://www.kaggle.com /datasets/jmm vutu/ summer-products-and-sales-in-ecommerce-wish/data?select=computed_insight_ success_of_ active_sellers.csv

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

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