Boosting E-commerce Revenue Through Consumer Purchase Analysis

Hima Vamsi Gonuguntla
MS in ESDS, SEAS
University at Buffalo
Buffalo, New York, US
hgonugun@buffalo.com

Sai Snehitha Ramisetty
MS in ESDS, SEAS
University at Buffalo
Buffalo, New York, US
saisnehi@buffalo.edu

Bhavana Chinnamgari
MS in ESDS, SEAS
University at Buffalo
Buffalo, New York, US
bchinnam@buffalo.edu

Abstract*— The analysis will focus on a detailed e-commerce dataset containing over 55,000 transactions, providing insights into customer demographics, purchasing behaviors, and promotional usage. The goal is to explore how age, gender, and discount usage influence consumer choices, leading to actionable insights for businesses. This study aims to enhance customer segmentation, improve retention strategies, and boost marketing efforts. Ultimately, it will enable eCommerce companies to adopt data-driven strategies that personalize experiences, increase revenue, and foster long-term customer relationships, while deepening understanding of consumer dynamics in the evolving eCommerce landscape.

Keywords*— Customer Behavior, E-commerce, Purchasing Trends, Predictive Modeling, Regression, Classification Algorithms

I. Introduction

Now-a-days digital age, eCommerce has completely transformed how we shop, giving customers access to a wide array of products at their fingertips. Understanding consumer behavior trends has become essential for success in this competitive landscape. This project focuses on analyzing eCommerce transaction data to uncover insights about customer behaviors, spending habits, and the effectiveness of promotional strategies. The aim is to explore the dataset for patterns that can help business leaders develop marketing strategies, enhance product offerings, and boost customer satisfaction, ultimately driving revenue growth.

II. Problem Statement

In this fast-growing world of eCommerce, businesses need to stay in tune with what customers want and how they shop to boost satisfaction and drive revenue. This project focuses on analyzing transactional data to identify customer segments, discount usage, and purchasing methods. By doing this, companies can create more effective marketing campaigns, improve product recommendations, and adjust pricing strategies. Ultimately, these insights will help enhance customer loyalty and maximize profits.

A. Background and Significance of the Problem

E-commerce is rapidly evolving, intensifying competition, and businesses must leverage data to enhance customer experiences to stay relevant. Understanding consumer behavior—what drives purchases, preferred products, and discount usage—is vital for developing effective business strategies. This project aims to uncover the key factors influencing consumer decisions to anticipate future trends.

- **Customer Retention:** Retaining existing customers is often more cost-effective than acquiring new ones. By identifying the factors that foster loyalty or lead to churn, businesses can make informed decisions to improve retention.
- **Personalization:** Customers seek tailored shopping experiences. Analyzing transaction data helps pinpoint what appeals to different customer segments, enabling businesses to provide personalized product recommendations.
- **Maximize Revenue:** Pricing strategies and discounts significantly influence purchasing behavior. By examining how discounts affect purchase decisions, marketers can devise strategies to boost sales revenue.
- **Market Competitiveness:** As the e-commerce landscape expands, predicting trends and adapting to evolving customer needs is essential for staying competitive.

Through comprehensive analysis of transactional data, organizations can better understand customer behavior, optimize marketing strategies, reduce churn, and ultimately drive sales.

B. Contribution and Importance of the Project

This project makes valuable contributions in three key areas of understanding consumer behavior:

1. **Customer Segmentation**: Customers are categorized by demographic factors like age and gender, along with their buying habits. This allows companies to tailor campaigns to specific groups effectively.
2. **Churn Prevention**: By identifying patterns that lead to customer churn, businesses can implement targeted strategies to improve customer retention.
3. **Marketing Optimization**: Analyzing which product categories and discount types drive high sales helps companies refine their marketing and promotional strategies.

Why is this important?

- **Increased Personalization**: Personalized product recommendations enhance customer satisfaction since they stem from individual buying preferences.
- **Strategic Decision Making**: Data-driven insights into purchasing behavior enable businesses to optimize

product offerings and marketing efforts, ensuring the right customers receive appropriate promotions.

- **Revenue Growth**: Optimizing discounts and promotions boosts conversion rates, leading to increased overall revenue.

Ultimately, this project provides actionable insights that help companies enhance customer engagement and satisfaction, improving their bottom line through better integration of transactional data for more personalized marketing.

III. Data Sources

An eCommerce dataset sourced from Kaggle, featuring over 55,000 rows and 13 columns. This comprehensive dataset effectively engaged real-world eCommerce behavior, making it ideal for in-depth analyses like customer segmentation, purchase pattern analysis, and discount effectiveness studies.

- **Source Link:** Kaggle eCommerce Dataset for Data Analysis
- **Row Count:** 55,000 rows
- **Column Count:** 13 columns

Here are the key features included in this dataset for analysis:

#	Column Name	Description	Data Type
1	CID	Unique identifier for each customer, helpful in tracking purchase history.	int64
2	TID	Unique identifier for each transaction, ensuring each purchase is recorded distinctly.	int64
3	Gender	Indicates the gender of the customer (Male/Female), useful for demographic analysis and segmentation.	object
4	Age Group	Segmentation based on predefined age ranges (e.g., 18-25, 26-35), facilitating the study of purchasing behavior.	object
5	Purchase Date	Date when the transaction took place, assisting in identifying seasonal trends and customer buying cycles.	object
6	Product Category	Category of each sold item (e.g., Electronics, Apparel), aiding in product-level analytics and demand forecasting.	object
7	Discount Availed	Binary variable indicating if a discount was utilized (Yes/No), helping to analyze the effect of discounts on buying habits.	object
8	Discount Name	Specific name of the discount offer availed and useful for analyzing promotional effectiveness.	object
9	Discount Amount (INR)	Monetary value of the discount applied, aiding in deriving net transaction amounts and studying discount usage trends.	float64
10	Gross Amount	Total value of a transaction before any discount, providing insights into average order value and spending patterns.	float64
11	Net Amount	Amount remaining after the discount is deducted from the gross amount, reflecting the actual revenue received.	float64
12	Purchase Method	Preferred mode of transaction (e.g., Credit Card, Debit Card, Net Banking), helping to understand customer payment preferences.	object
13	Location	City where the purchase occurred, useful for conducting regional or market-level analyses.	object

IV. Data Cleaning/ Processing

Data cleaning and processing are crucial for effective data analysis, ensuring the dataset is reliable and ready for insights. Here are the key steps taken to prepare the dataset:

Handling Missing Values: The "Discount Name" column had missing values, which were filled with "No Discount" to maintain consistency. This decision preserved data integrity and allowed for more accurate analysis of discount-related metrics, revealing customer behavior without introducing bias.

Converting Data Types: The "Purchase Date" column was initially in string format, complicating date-based analysis. By converting it to datetime format, we could easily extract specific components like month and year, enabling trend analysis and enhancing our understanding of customer purchasing patterns over time.

Creating New Features: For better analysis, we derived new features from the "Purchase Date" column, creating "Year" and "Month" columns. This allowed for more detailed trend analysis, revealing seasonality in customer purchases.

Validating Categorical Data: We ensured that only valid entries were retained in the "Age Group" column, removing any inconsistencies. This step improved segmentation accuracy and the overall quality of age-related analyses.

Duplicate Removal: Duplicates were identified and removed to ensure the dataset was clean, allowing for more reliable analysis.

Normalizing Text Data: We standardized the "Product Category" and "Purchase Method" fields by converting text to lowercase and removing extra spaces. This consistency improved the accuracy of grouping and filtering operations.

Encoding Categorical Variables: Categorical variables were converted to numeric format using One-Hot Encoding, enabling better interpretation by machine learning algorithms without assuming any ordinal relationships.

Handling Outliers: Outliers, particularly in the "Gross Amount" and "Net Amount" columns, were addressed using the interquartile range (IQR) method to avoid skewed insights in financial analysis.

Calculating Derived Metrics: We created a derived metric called "Discount Percentage," which provides the percentage discount on transactions, offering insights into the value of discounts for different purchases.

Statistics Summary: Out of 55,000 customers, 27,585 did not use discounts, while 27,415 did, which shows a balanced distribution. The average gross amount spent was approximately 3,013, with a median of 2,954, and the net amount averaged 2,876, shows some outliers due to refunds. These metrics help analysis of trends in discount usage.

Reindexing and Final Cleanup: After cleaning, the dataset was reindexed to maintain a consistent order, ensuring that the data remained organized and ready for analysis.

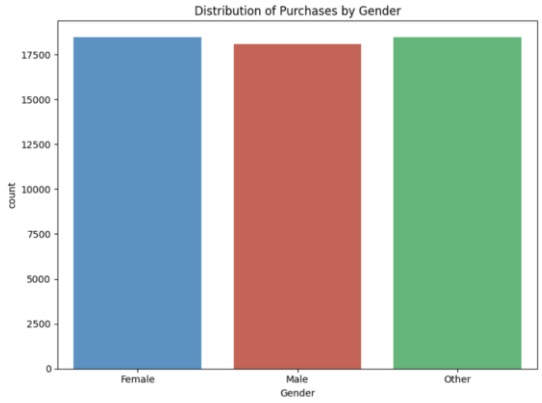
Overall, these essential cleaning and processing steps significantly enhanced the quality of the eCommerce dataset,

making it more usable for driving data-driven decisions in the business.

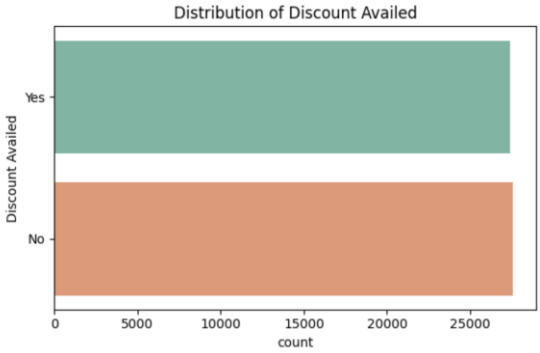
EXPLORATORY DATA ANALYSIS (EDA)

Exploratory Data Analysis (EDA) is a crucial step in data analysis, designed to uncover patterns, spot anomalies, and test hypotheses through visual and statistical methods. By using these techniques, EDA reveals important trends and relationships within the data. In this project, the focus will be on analyzing customer purchasing behavior to generate actionable insights that inform strategic decisions and enhance marketing efforts.

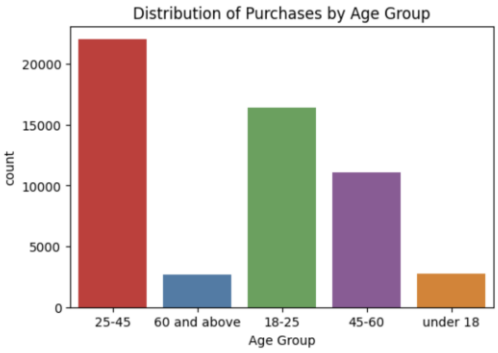
Purchases by Gender: This analysis groups data by gender and counts the number of purchases for each group. A bar plot illustrates the differences in purchasing habits between male and female customers, offering valuable insights for gender-targeted marketing strategies.



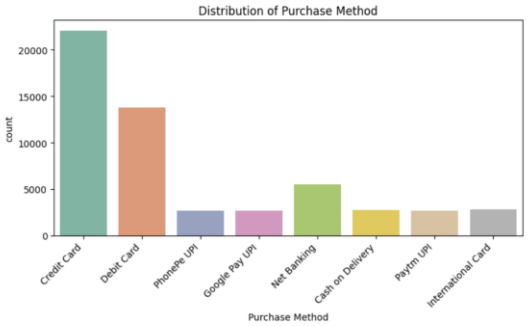
Discount Usage: This analysis looks at how many discounts are utilized, providing insights that can help to improve revenue generation.



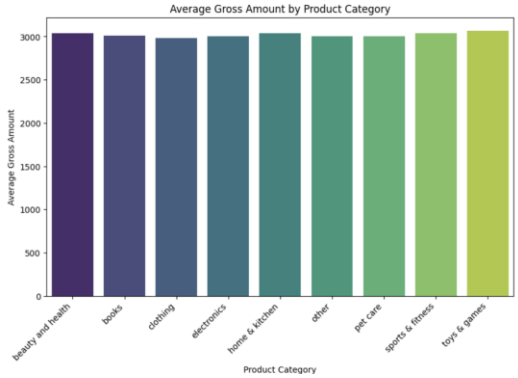
Purchases by Age Group: Data is categorized by age group, with a bar plot displaying purchase counts for each category. This helps identify the most active buying age groups, allowing for getting more marketing and product strategies.



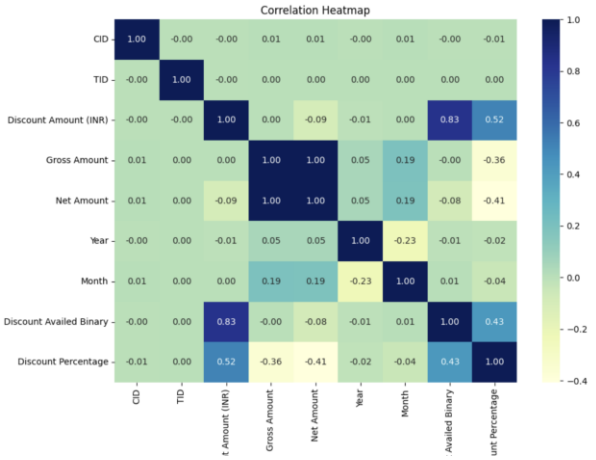
Purchase Methods Distribution: A count plot visualizes the various purchasing methods customers use. By using a clear color palette, this analysis highlights customer preferences, building businesses on where to focus their marketing efforts.



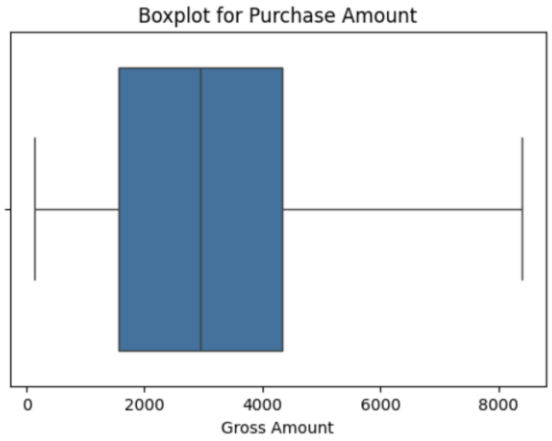
Average Gross Amount by Product Category: The dataset is grouped by product category to calculate the average gross amount spent. A bar plot showcases which categories generate higher revenue, aiding inventory and promotional decisions.



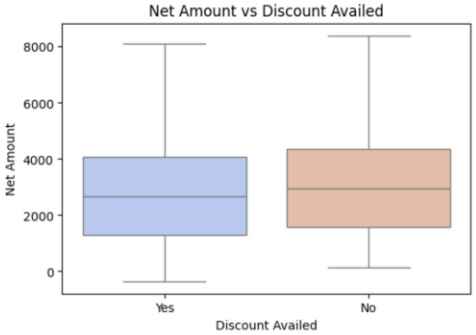
Correlation Matrix for Numerical Variables: This correlation matrix visualizes the relationship between "Gross Amount" and "Discount Amount." Resulting heatmap reveals significant relations that could inform predictive modeling.



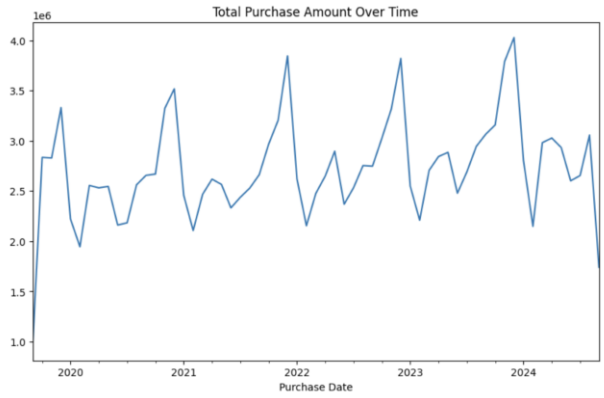
Outliers in Purchase Amount: A box plot examines purchase amounts for potential outliers, revealing extremes that could indicate unusual spending patterns or data entry errors.



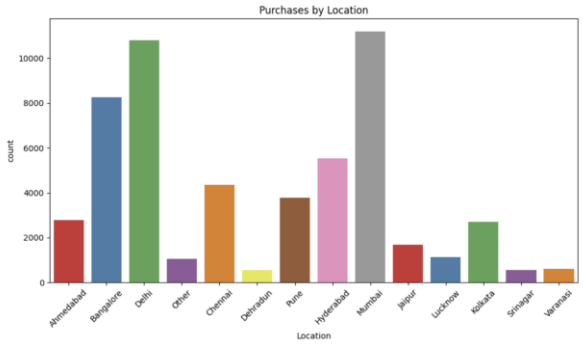
Purchase Amount vs. Discounts Used: A box plot compares purchase amounts based on discount usage, focussing on spending habits when discounts are applied.



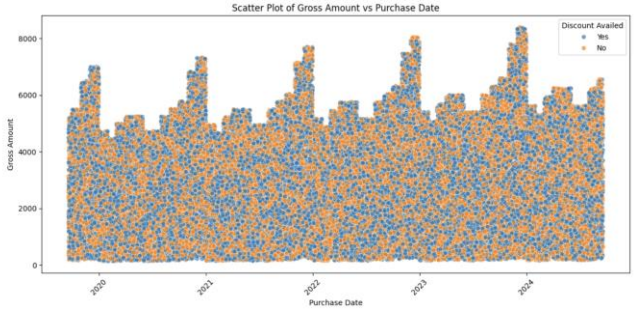
Analyzing Purchases Over Time: The "Purchase Date" is converted to a datetime format to analyze purchasing trends. Resampling data allows for visualization in a line graph, helping to identify seasonal patterns in consumer behavior.



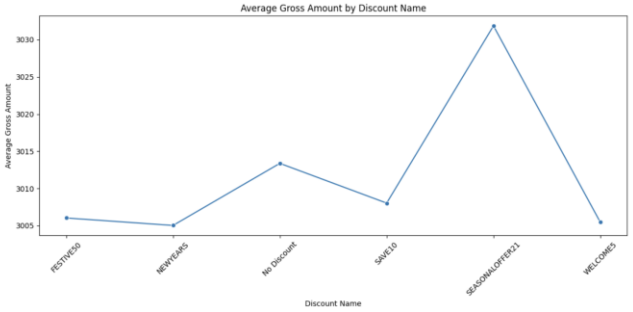
Purchases by Location: Grouping the dataset by location provides a bar plot showing total purchases from each area. This results in local marketing efforts and resource allocation.



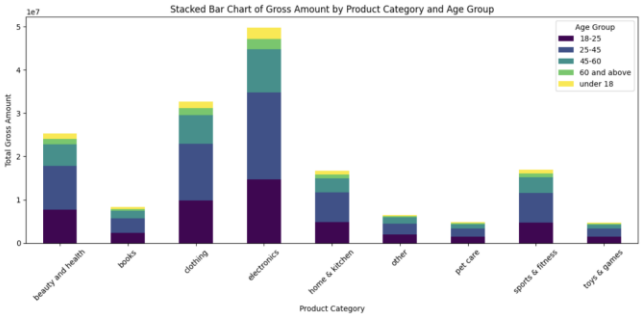
Gross Amount vs. Purchase Date: It shows the relationship between "Gross Amount" and "Purchase Date," revealing trends in spending habits over time.



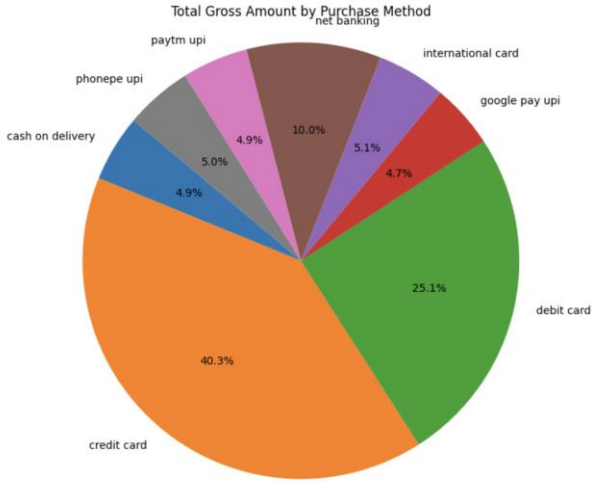
Average Gross Amount by Discount Name: Data is grouped by "Discount Name" to calculate average gross amounts for each discount type. This plot helps identify which discounts drive revenue, aiding in better tailoring of discount offerings.



Gross Amount by Product Category and Age Group: Data is aggregated by product category and age group, with a stacked bar chart showing how different age groups contribute to revenue across product categories.



Total Gross Amount by Purchase Method: This chart displays the share of gross amounts from various purchase methods, helping inform strategic decisions on channel investments and marketing focus.



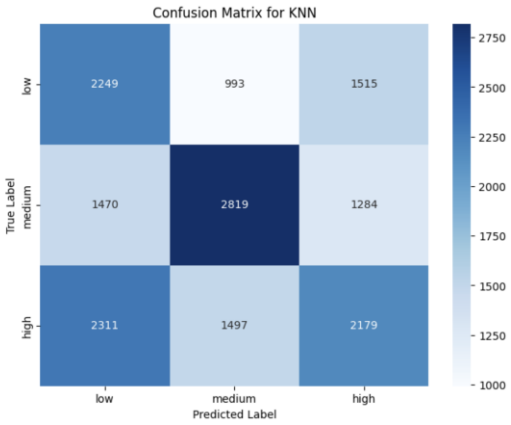
ALGORITHMS AND VISUALIZATIONS

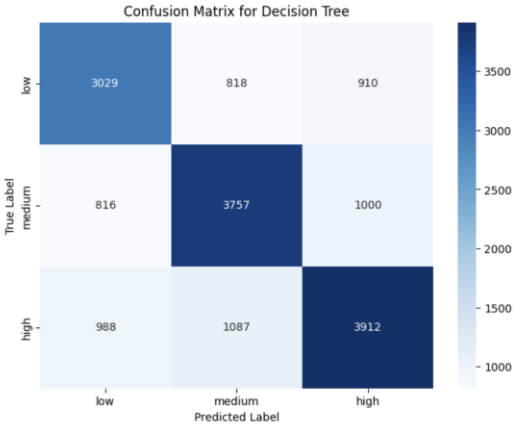
In this project, we used some algorithms for the prediction and classification. The models involve *Linear Regression, K-Nearest Neighbors (KNN), Decision Tree, Random Forest, Support Vector Machine (SVM), and XGBoost*. Each model was trained and tested to get its performance based on metrics such as MSE, RMSE, and R^2 scores.

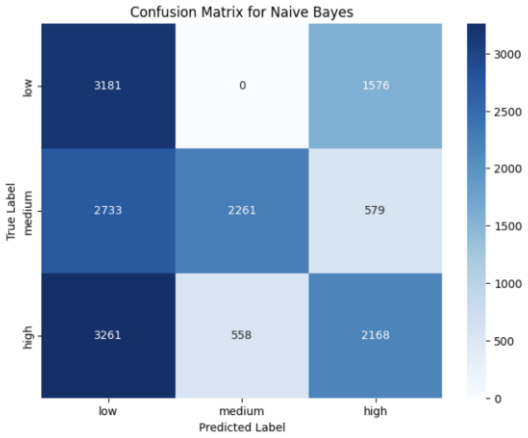
In addition to that, we have implemented different classification algorithms like *Logistic Regression and Gaussian Naive Bayes* for the prediction of spending categories of customers from the transactional data. These models were compared on the basis of accuracy, F1 score, precision, and recall. We have plotted confusion matrix, as visual insight into the performance of each model, indicating how well a model has classified target categories correctly.

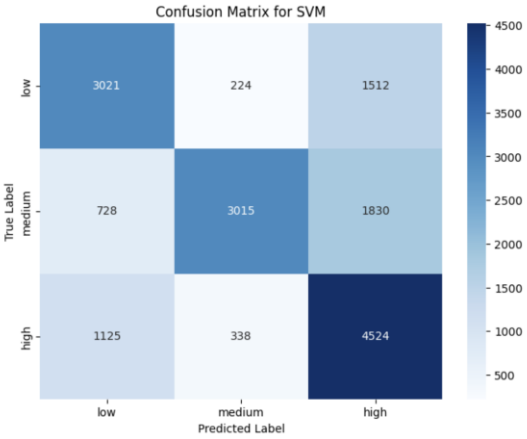
```python
models = {
    "KNN": KNeighborsClassifier(),
    "Decision Tree": DecisionTreeClassifier(),
    "SVM": SVC(),
    "Logistic Regression": LogisticRegression(),
    "Random Forest": RandomForestClassifier(),
    "Naive Bayes": GaussianNB(),
    "XGBoost": xgb.XGBClassifier(use_label_encoder=False, eval_metric='mlogloss')
}

# training and testing models
for model_name, model in models.items():
    print(f"\n{model_name}")
    model.fit(X_train_scaled, y_train)
    y_pred = model.predict(X_test_scaled)

    # Calculation of metrics for analysis
    accuracy = accuracy_score(y_test, y_pred)
    f1 = f1_score(y_test, y_pred, average='weighted')
    precision = precision_score(y_test, y_pred, average='weighted')
    recall = recall_score(y_test, y_pred, average='weighted')

    print(f" Accuracy: {accuracy:.4f},\n F1 Score: {f1:.4f},\n Precision: {precision:.4f},\n Recall: {recall:.4f}")

    cm = confusion_matrix(y_test, y_pred)

    # Plotting Confusion Matrix for each model
    plt.figure(figsize=(8, 6))
    sns.heatmap(cm, annot=True, fmt='d', cmap='Blues', xticklabels=labels, yticklabels=labels)
    plt.title(f'Confusion Matrix for {model_name}')
    plt.xlabel('Predicted Label')
    plt.ylabel('True Label')
    plt.show()
```

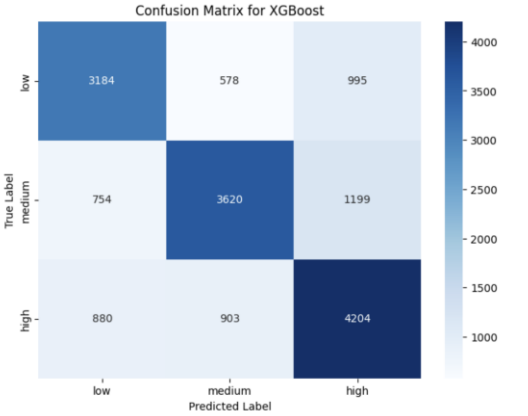
Here are the confusion matrices for the ML models

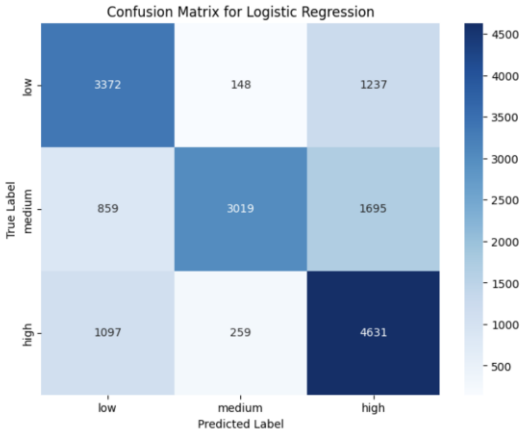


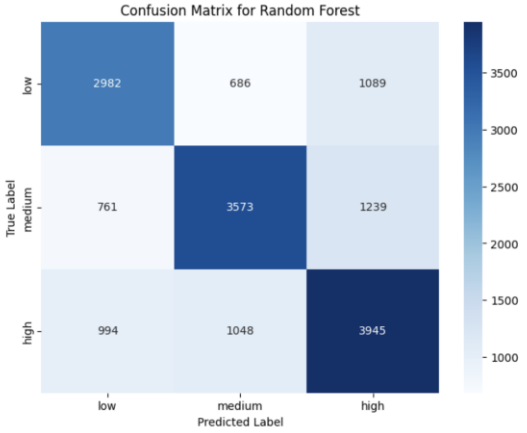

Confusion Matrix for Decision Tree
True Label
low
medium
high
3029
818
910
816
3757
1000
988
1087
3912
low
medium
high
Predicted Label
3500
3000
2500
2000
1500
1000


Confusion Matrix for Naive Bayes
True Label
low
medium
high
3181
0
1576
2733
2261
579
3261
558
2168
low
medium
high
Predicted Label
3000
2500
2000
1500
1000
500
0


Confusion Matrix for SVM
True Label
low
medium
high
3021
224
1512
728
3015
1830
1125
338
4524
low
medium
high
Predicted Label
4500
4000
3500
3000
2500
2000
1500
1000
500


Confusion Matrix for XGBoost
True Label
low
medium
high
3184
578
995
754
3620
1199
880
903
4204
low
medium
high
Predicted Label
4000
3500
3000
2500
2000
1500
1000


Confusion Matrix for Logistic Regression
True Label
low
medium
high
3372
148
1237
859
3019
1695
1097
259
4631
low
medium
high
Predicted Label
4500
4000
3500
3000
2500
2000
1500
1000
500


Confusion Matrix for Random Forest
True Label
low
medium
high
2982
686
1089
761
3573
1239
994
1048
3945
low
medium
high
Predicted Label
3500
3000
2500
2000
1500
1000

EXPLANATION AND ANALYSIS

A. Regression models

Linear Regression

It is an algorithm to model the relationship between a dependent variable and one or more independent variables by fitting a linear equation to the observed data. The approach was preferred because of its simplicity, interpretability of the results, so useful at the beginning of the analysis. However, it did require tuning in order to optimize performance, mostly concerning regularization parameters in order not to overfit. Upon evaluation, Linear Regression returned an MSE of 2.280448M and an RMSE of 1.51012k with an R^2 score of 0.2274. Therefore as much as the model does have considerable predictive power, it explains only about 22.74% of the variance in the target variable, hence a call for more complex models to better capture the underlying relationship in the data.

K-Nearest Neighbors (KNN)

K-Nearest Neighbors is an instance-based learning algorithm in which the classification of data points depends on the closeness of their nearest neighbors. KNN is opted in this one due to the intuitively easy approach it offers and its excellent performance in multi-class classification problems. The tuning involved finding the value of k, or the number of neighbors to consider, which can have large impacts on the accuracy of the model. However, KNN's performance on this dataset turned out to be quite poor, with an MSE of 2.537822M, RMSE of 1.593k, and R^2 score of 0.1402. These metrics indicate that KNN cannot model the spending amount variation quite well and hence may not capture complex patterns in context.

Decision Tree

This algorithm creates a tree-like model to predict output by means of partitioning data records based on their feature values. Desirable not only due to intuitive interpretability for this model, it can handle different data types. Tuning was done with respect to tree depth and a minimum number of samples required for a split to regularize overfitting. Even after that, this model showed an MSE of 2.889581M, an RMSE of 1.699k, and R^2 score of 0.0210. These results mean that the predictive ability of this model is rather low, with the maximum percentage of variance explained by no more than 2.10%, hence arguing for the need to pursue more robust modeling techniques.

Random Forest

Random Forest is a kind of learning where in multiple decision trees are combined to reduce overfitting and improve predictive performance. This algorithm was chosen for its strong reliability in dealing with datasets that are complicated, including interactional effects across features. This tuning involved optimizing the number of trees along with the maximum depth each tree will have. With that, this model had yielded an MSE of 1.447472M, an RMSE of 1.203k, and an R^2 score of 0.5096. Set of improvements was evidence of a leap in performance, which makes Random Forest effective at capturing more variance in the target variable and this was effective for spending amount prediction.

Support Vector Machine

Support Vector Machine is another powerful algorithm that identifies the optimal hyperplane to separate the classes of data. This model was selected based on its strength in high-dimensional spaces, other than tracing nonlinear boundaries using the right kernel settings. Model tuning in this regard focused much on the choice of kernel functions and the regulation parameters. Still, the model performed poorly with a mean squared error of 2.707824M and a root mean square error of 1.645k, besides an R^2 score of 0.0826. These results mean that poor SVM could hardly give meaningful predictions of the target variable due to its assumptions, which do not likely fit well with the structure of the dataset.

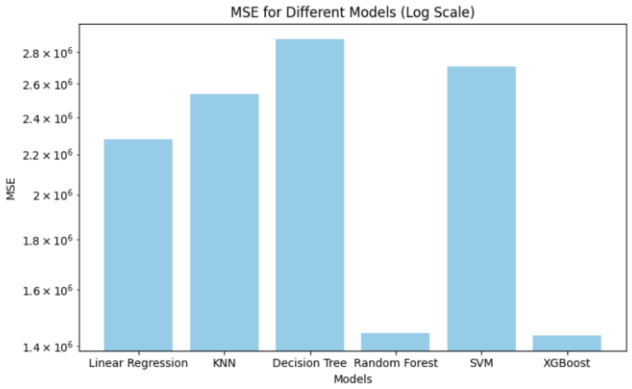
XGBoost

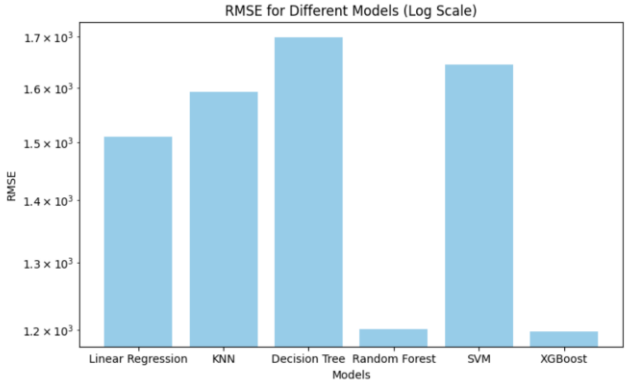
XGBoost is an advanced ensemble learning algorithm that builds on decision trees to make predictions using boosting techniques. This model was chosen due to its high performance, speed, and efficiency in terms of handling missing data. This included tuning parameters such as learning rate and max tree depth to bring out the best of the model. XGBoost yielded an MSE of 1.433795M, an RMSE of 1.197k, and an R^2 score of 0.5142, thus turning out to be the best performing model out of those tested. These results will show that XGBoost captures the nuances in the spending categories well and will provide great insight into potential targeted marketing strategies, adding value to business decisions in general.

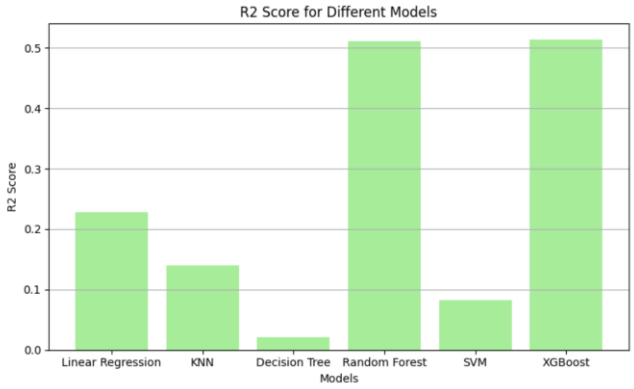
Intelligence gained from the models

Summary of results:

- Linear Regression: MSE: 2280448.0915, RMSE: 1510.1153, R^2 Score: 0.2274
- KNN: MSE: 2537822.0416, RMSE: 1593.0543, R^2 Score: 0.1402
- Decision Tree: MSE: 2889581.3988, RMSE: 1699.8769, R^2 Score: 0.0210
- Random Forest: MSE: 1447472.6768, RMSE: 1203.1096, R^2 Score: 0.5096
- SVM: MSE: 2707824.6081, RMSE: 1645.5469, R^2 Score: 0.0826
- XGBoost: MSE: 1433795.9946, RMSE: 1197.4122, R^2 Score: 0.5142







Overall Analysis:

Among all, the best one is XGBoost, which has an MSE of 1.433795M and an RMSE of 1.197k with a significant R^2 score of 0.5142, thus explaining more than 51% in the variance of the spending amounts. It really takes advantage of its ensemble approach by boosting to capture some very complex relationships in the data that other models like Linear Regression and KNN. Random Forest also goes well, but XGBoost outruns it a bit because of better handling of details and missing data.

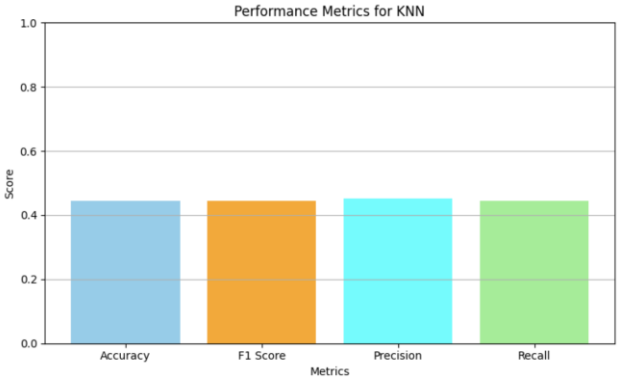
A model like XGBoost, which gives very good results with the problem statement with prediction spending behavior for focused marketing and other business decisions will give further insight about how spending is done, hence further improving the segmentation of customers and, by extension, better decisions on promotions and what to offer. Its performance directly impact the ability to predict future spending and help in tailoring business strategy toward customer needs.

B. Classification models

K-Nearest Neighbors

KNN is a simple, instance-based learning algorithm that classifies data points by the majority vote of their nearest neighbors. It works particularly well for problems that have irregular boundaries between decisions. In this analysis, KNN was selected because it is easy to use and suitable for multi-class classification. This model required tuning of one main parameter, k - representing the number of neighbors to consider.

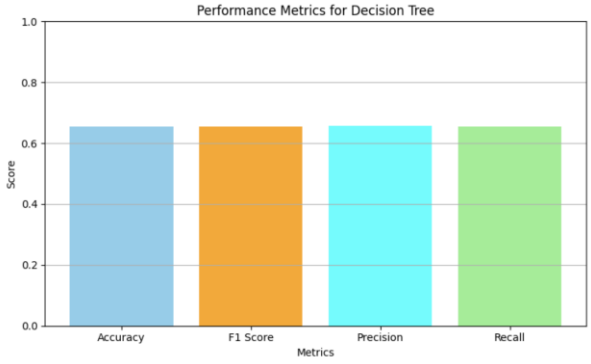
When applied to the dataset, KNN reached an accuracy of 0.4441 only, with an F1 score of 0.4444, precision of 0.4507, and recall of 0.4441. This shows that even though KNN is a good starting point, in this particular application, it struggled to effectively classify spending categories. By considering this result, one might explains that KNN does not suit the data and requires more complex models which would capture the relationship deeper.



Decision Tree Classifier

This Classifier works by creating a model that looks like a tree. Feature decisions are shown in the nodes, and their outcomes as different branches. It had become a reasonable option for the type of analysis because of its interpretability and its capability to work with various data types. The tuning process was considered more towards preventing overfitting by optimizing the hyperparameters on tree depth and the minimum number of samples required for splitting an internal node.

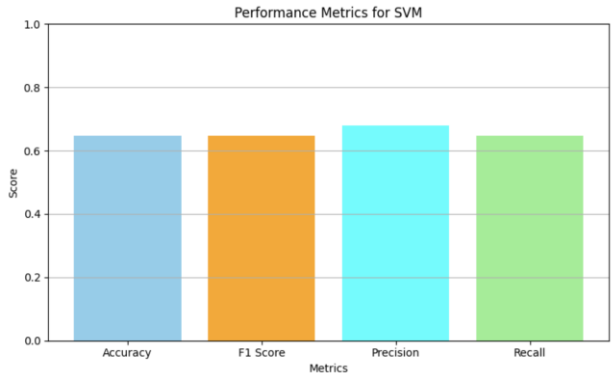
For Decision Tree, the accuracy came to 0.6556, the F1 score was 0.6557, precision was 0.6559, and recall was valued at 0.6556. Thus, the result actually showed the inclusions of reasonably good accuracy in classifying spending categories by the Decision Tree. Hence, Decision Tree was pretty effective in this scenario. Based on the key features, such insights have been able to reveal key features affecting spending behavior.



Support Vector Classifier

SVC is reputed for finding that optimal hyperplane that separates classes effectively. It had been selected because of the ability to handle high-dimensional data and the creation of nonlinear decision boundaries. Only major tuning was in the selection of the kernel type and regularization parameters that helped in improving the performance of the model.

Accuracy of 0.6472, F1 score of 0.6476, precision of 0.6796, and a recall of 0.6472 are retrieved from the SVC. These are not bad results since they are not the lowest. Therefore, still, the insight into customer segmentation based on the way customers spend their money is given by SVC. The model's result showed there was a need for more analysis, therefore the checking of the other algorithms that could yield a better result was needed.



Logistic Regression

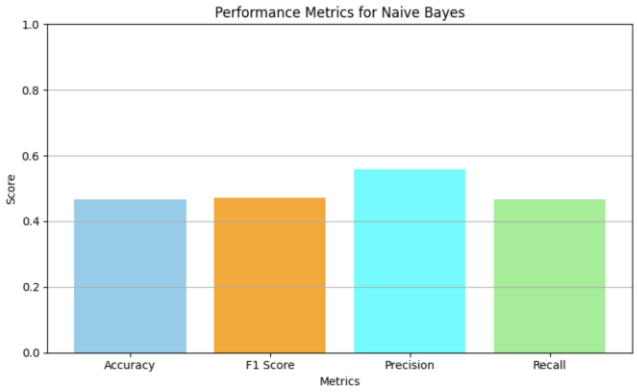
Logistic Regression is a methodology that can be used for making a prediction of categorical outcomes given a set of predictor variables. It is simple and interpretative. Although primarily a binary classifier, it does adapt to multi-class problems. Tuning was involved in adjusting the regularization parameters that resulted in an improved performance.

Logistic Regression yielded an accuracy of 0.6755, an F1 score of 0.6749, a precision of 0.7102, and a recall of 0.6755. This model performed reasonably well by those metrics in spending category classification, lending sharp insight into which features most impacted the variance in customer behavior. Its results really underlined the importance of certain demographic factors in their purchasing decisions.

Gaussian Naive Bayes

Gaussian Naive Bayes tells independence among predictors and thus can be a fast option for multi-class classification problems. It is also efficient and effective to make initial assessments with not much tuning.

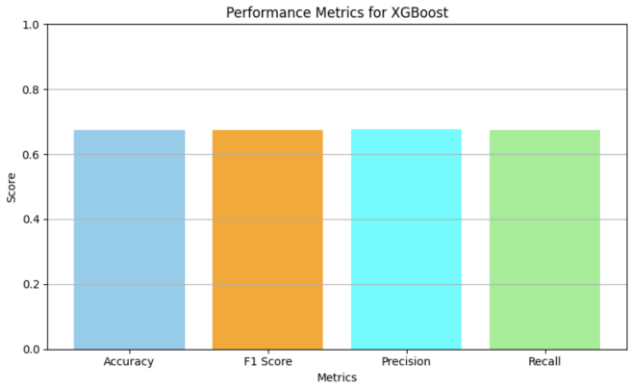
It achieved an accuracy of 0.4664, with the F1 score of 0.4715, precision of 0.5590, and recall of 0.4664. Such results do mean that this model failed to classify spending categories well. On the other hand, they were giving some insights into customers' spending behavior. This finding shows that more sophisticated models should be used to enhance their performance in classification.



XGBoost

XGBoost is an ensemble algorithm that makes predictions by committee, taking the predictions of many base learners and, in turn, making more accurate classifications. The high speed and handling of missing data with a far greater degree of accuracy make this model ideal for the analysis at hand.

The XGBoost algorithm returned accuracy of 0.6746, an F1 score of 0.6747, precision of 0.6761, and recall of 0.6746 on the given dataset. Based on the results from XGBoost, performing equally well as Logistic Regression captures each detail of spending categories. The insight one may get from this model will serve targeted marketing strategy and improve overall business decisions.



Intelligence gained from the models

Summary of results:

- KNN: Accuracy: 0.4441, F1 Score: 0.4444, Precision: 0.4507, Recall: 0.4441
- Decision Tree: Accuracy: 0.6556, F1 Score: 0.6557, Precision: 0.6559, Recall: 0.6556
- SVC: Accuracy: 0.6472, F1 Score: 0.6476, Precision: 0.6796, Recall: 0.6472
- Logistic Regression: Accuracy: 0.6755, F1 Score: 0.6749, Precision: 0.7102, Recall: 0.6755
- Naive Bayes: Accuracy: 0.4664, F1 Score: 0.4715, Precision: 0.5590, Recall: 0.4664

- XGBoost: Accuracy: 0.6746, F1 Score: 0.6747, Precision: 0.6761, Recall: 0.6746

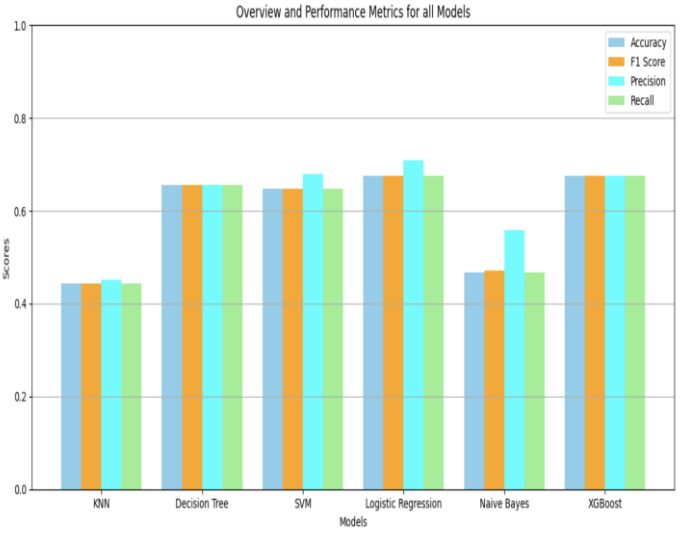
Overall Analysis:
Based on above results, Logistic Regression and XGBoost has the the best results, at accuracies of 0.6755 and 0.6746, respectively. Logistic Regression performed better on precision, it issued fewer false positive classifications, which are highly critical in applications like targeted marketing, where precision forms the key to identifying such essential segments of customers. The Decision Tree also performed fairly well with an accuracy of 0.6556 for interpretability.

Logistic Regression and XGBoost provide actionable insights into which features-not just demographic ones-drive spending, given the problem statement of the project, which is analyzing customer spending behavior to inform marketing strategies. Because of their high effectiveness in the classification of spending categories, more accurate segmentation, targeted marketing, and promotional strategies will be able to maximize business revenues. In practical applications, the use of such models would help to highlight the most valued information about the customers.

CONCLUSION

As E-Commerce data is useful for many companies and it's a trending in the market from a long time, based on the overall analysis of the project results and metrics obtained provides great insight into customer purchasing behavior and some very interesting trends across the board in demographics, product categories, and use of discount. Analysis showed how effective marketing strategies could be designed based on customer segmentation using demographic and transactional data. Also delivered deep into various classification algorithms and did prediction on customer spend classification by categorizing transactions into net amount brackets. After preprocessing the data, models were applied: K-Nearest Neighbors, Decision Tree, Support Vector Classifier, Logistic Regression, Naive Bayes, and XGBoost etc.

Each of these has been evaluated on accuracy, F1 score, precision, and recall. Among these models, the most powerful approaches are represented by Logistic Regression, XGBoost, and Decision Tree because they showed better overall accuracy with more balanced metrics. These models provided a clear view of customer purchasing patterns, which served as the basis for further optimization and using the outputs within strategic decision-making for marketing and inventory management. This tends to increase the marketing sales based on the fitted algorithm and the metrics which is useful to boost the marketing based on different features by considering which helps companies and other stakeholders to gain profit with the analysis of this project.



REFERENCES

[1] A. Maurya, S. Pratap, P. Pratap and A. Dwivedi, "Analysis of Behavioural Data of Customer for the E-Commerce Platform by using Machine Learning Approach," 2023 International Conference on Computational Intelligence and Sustainable Engineering Solutions (CISES), Greater Noida, India, 2023, pp. 801-806, doi: 10.1109/CISES58720.2023.10183475.

[2] M. Gull and A. Pervaiz, "Customer Behavior Analysis Towards Online Shopping using Data Mining," 2018 5th International Multi-Topic ICT Conference (IMTIC), Jamshoro, Pakistan, 2018, pp. 1-5, doi: 10.1109/IMTIC.2018.8467262.

[3] F. P. Chamorro-Zapana, H. E. Chumpitaz-Caycho, E. N. Espinoza-Gamboa, M. A. Espinoza-Cruz and F. Cordova-Buiza, "Application of big data for analyzing consumer behavior in e-commerce companies," 2023 IEEE 6th International Conference on Big Data and Artificial Intelligence (BDAI), Jiaxing, China, 2023, pp. 30-34, doi: 10.1109/BDAI59165.2023.10256889.

[4] B. Arivazhagan, S. Pandikumar, S. B. Sethupandian and R. S. Subramanian, "Pattern Discovery and Analysis of Customer Buying Behavior Using Association Rules Mining Algorithm in E-Commerce," 2022 First International Conference on Electrical, Electronics, Information and Communication Technologies (ICEEICT), Trichy, India, 2022, pp. 1-5, doi: 10.1109/ICEEICT53079.2022.9768473.

[5] J. Panduro-Ramirez, "Machine Learning-Based Customer Behavior Analysis for E-commerce Platforms," 2024 International Conference on Advances in Computing, Communication and Applied Informatics (ACCAI), Chennai, India, 2024, pp. 1-5, doi: 10.1109/ACCAI61061.2024.10602204.

[6] https://www.researchgate.net/publication/374774848_Research_on_Consumer_Behavior_Prediction_Based_on_E-commerce_Data_Analysis

[7] "Analysis of E-Commerce Marketing Strategy Based on Xgboost Algorithm". https://www.researchgate.net/publication/371160409_Analysis_of_E-Commerce_Marketing_Strategy_Based_on_Xgboost_Algorithm

[8] F. Guo and H. -L. Qin, "The Analysis of Customer Churns in e-Commerce Based on Decision Tree," 2015 International Conference on Computer Science and Applications (CSA), Wuhan, China, 2015, pp. 199-203, doi: 10.1109/CSA.2015.74.