E-Commerce Consumer Analysis and Fraud Behavior Detection

Farzin Gholamrezae[1]

***Abstract*— E-commerce purchases represented over 17% of global retail sales in 2024. This sales yield valuable data that can reveal customer trends, forecast sales, and market gaps. This paper applies supervised and unsupervised machine learning techniques to explore consumer behavior within e-commerce transactions on a generated data set replicating Walmart.com.**

I. Introduction

In 2024, over 17% of retail purchases made used e-commerce platforms [5]. In the US alone, $1.119 trillion dollars was spent online in 2023, a 7.6% growth year over year, and accounting for 15.3% of total sales [2], [3]. Online retail sales also provide the accrual of accompanying data. This data is highly valuable in that it can garner insights into customer behavior, predict sales trends, identify market gaps for new products, and be used to improve customers' online experiences. However, not all online shoppers are good actors. Over $48 billion dollars were globally lost to e-commerce fraud in 2023, with the US taking the lion's share of the losses, accounting for 42% of value [1].

The main objective of this research project is to explore which machine learning techniques best identify and predict consumer behaviors and sales trends using both supervised and unsupervised machine learning applications. This aim is accomplished by analyzing the dataset to answer the following research questions: (i) Which supervised machine learning algorithms are most effective at predicting consumer behavior trends? (ii) How can unsupervised clustering methods be leveraged to target new consumer groups? (iii) How can explorations of fraud detection be done within the dataset, which was reserved as reach aim.

II. Data

The corpus was chosen from Kaggle, selected specifically for it's potential to be further divided and analyzed by subcategories due to it's large number of categorical variables with respect to age, gender, product categories, occupation, and cities [4]. The generated dataset contains 10 variables and approximately 550K rows, which also include user IDs, product IDs, length of city residence, marital status, and receipt sums [4].

Data Preprocessing & Exploration: The data was cleaned to correct for potential nulls, missing values, to remove non-alphanumeric characters, and correct for data type inconsistencies. Examination of unique customer IDs showed consistency across city, age, occupation, and other variables ensuring reliability for downstream analysis. The data was examined to identify outliers and distributions of categorical and continuous variables via a series of barplots, boxplots, pairplots, and countplots.

Exploration of distributions revealed 3,631 unique product IDs across 18 product categories. Receipt sums ranged from $12-$23,961. Ages skew towards 26-35 year olds (40%), who produce 39.9% of all sales. The gender ratio skews towards males 75.3%, potentially influencing the most popular product categories and sales trends. City B amassed 42% of all sales within the dataset. Then, bivariate and multivariate analysis was performed to guide the decision making process in trend analysis variable selection. Figure 1 below displays the correlation matrix used to guide variable selection.

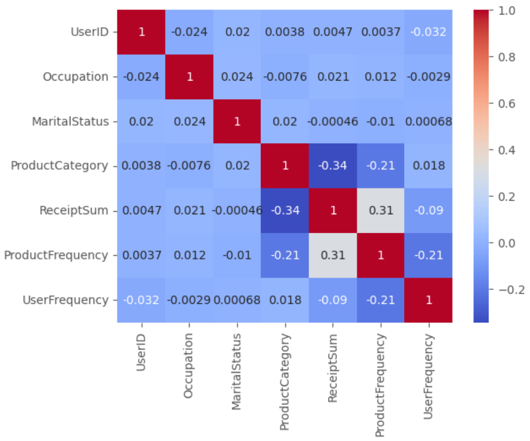

Fig. 1. Correlation Matrix

III. Method

A. Supervised Learning

Early model testing and data exploration led to the selection of the receipt sum and product categories as the target variables. Prior to processing, data was split into training, validation, and test sets using a 70/20/10 split.

1) Linear Regression: The linear regression model was evaluated using one-hot encoding and label encoding, with one-hot encoding performing .5% better. Scaler testing was performed, including minmax, robust, and standard scalers. Standard and robust scalers demonstrated equal performance with a R^2 score of .121 and MSE of 22,062,071.26. The low accuracy and high MSE score indicated linear regression was not a suitable model, which aligns with the highly categorical nature of the variables. PCA was applied for exploratory analysis with results yielding a slight decrease in performance.

2) Logistic Regression: The logistic model was evaluated using one-hot encoding and label encoding, with one-hot encoding performing better. Scaler testing, including minmax, robust, and standard scalers, resulted in the standard scalers performing best. The initial binary categorization value was determined using the median value of the receipt sum. Initial results were .64 across accuracy, precision, recall and F1 scores. After a series of iterative tests, and a change of the binary classifier from median to mean, the results improved an average of 14.5% across all metrics. Final results of .78 across accuracy, precision, recall and F1 scores were achieved. Application of PCA showed minor improvements when dimensionality was reduced and began to degrade beyond a 0.90 variance threshold, showing the model was already well-fit to the original feature space.

B. *Unsupervised Learning*

1) K-Means: In order to instantiate the K-Means algorithm, first the number of clusters was determined using the elbow method and silhouette score. The elbow method provided support at 8 or 9 clusters, with diminishing returns beyond that range. The silhouette score leaned towards K=14 achieving a score of .65, with results only improving slightly to .7 at 20 clusters which may indicate overfitting. Figure 2 presents the results for the K-Means algorithm at K=14.

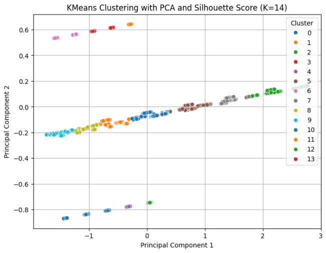

Fig. 2. K-Means Clustering Results

2) DBSCAN: In plotting heatmaps DBSCAN revealed trade-offs between clustering quality and data coverage. Silhouette scores peaked meaningfully at 0.87 with minimum samples in the range of 23–25, with corresponding 24–25 clusters. Here, 96% of transactions were clustered indicating strong model performance. At higher epsilon values cluster counts depreciated, creating concerns for overlapping groupings or potential under-segmentation of the data. With this in mind, the final DBSCAN implementation used minimum samples = 23 and epsilon = 0.03, achieving a 96.3% clustering rate, creating 24 clusters, and a silhouette score of 0.87. The results from the heatmaps explored in DBSCAN encouraged the implementation of HDBSCAN to reduce sensitivity to fixed parameter tuning.

3) HDBSCAN: Finally, HDBSCAN implementation shown in figure 3 used minimum samples = 120 and minimum cluster size = 40, achieving a 89.4% clustering rate, creating 14 clusters, and a silhouette score of 0.89.

IV. Results and Future Work

1) Results: Towards Aim (i), logistic regression was selected. Scores of .78 across accuracy, precision, recall and F1 scores were yielded, with no improvements in applying PCA. This improved upon the regression tree models where random forest models produced a high MAE and MSE, noted above, and R^2 value of 0.652. High MSE indicates further analysis is needed in ways to optimize the models for accuracy via methods such as hyperparameter tuning and regularization techniques. With an R^2 score of .124 and high MSE, linear regression was not considered to be a suitable model for the data. In Aim (ii), PCA-reduced KMEANS, DBSCAN, and HDBSCAN clustered relationships between the receipt sum and product category variables. This selection was informed by prior model and correlation testing, which indicated significant inter-dependencies across the features. The resulting clusters were well-formed, with consistent cluster counts observed across models, validated through the elbow method, silhouette scores, and heatmap visualizations. Interpretation of the models proved more challenging, as scaling and PCA-reduced visualizations introduced complexity in interpreting outputs via the original feature space.

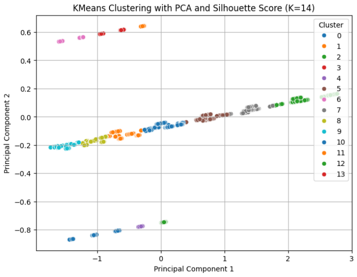

Fig. 3. PCA-Reduced DBSCAN

For Aim (iii), a stretch goal, the development of clustering models laid the groundwork for next steps, such as outlier and anomaly detection.

2) Future Work: Further interpretation of the unsupervised clustering outputs is necessary, particularly in mapping cluster structures back to original features. Future efforts will also focus on advancing outlier detection techniques and developing fraud detection models using these unsupervised techniques.

References

[1] M. Identity, "5 industry best practices to preventing ecommerce fraud in 2024," Ekata, 18-Oct-2023. [Online]. Available: https://ekata.com/resource/5-industry-best-practices-to-preventing-ecommerce-fraud-in-2024/. [Accessed: 18-Sep-2024].

[2] A. Haleem, "US ecommerce sales reached $1.119 trillion in 2023," Digital Commerce 360, 26-Feb-2024. [Online]. Available: https://www.digitalcommerce360.com/article/us-ecommerce-sales/. [Accessed: 19-Sep-2024].

[3] "US E-Commerce Sales as Percent of Retail Sales quarterly trends: Quarterly E-commerce report," Ycharts.com. [Online]. Available: https://ycharts.com/indicators/us_ecommerce_sales_as_percent_retail_sales. [Accessed: 19-Sep-2024].

[4] V. Devaraj, "e-Commerce (Walmart) Sales Dataset." 30-May-2024.

[5] Statista Research Department, "E-commerce as share of total global retail sales 2015–2027," Statista, 2024. [Online]. Available: https://www.statista.com/statistics/534123/e-commerce-share-of-retail-sales-worldwide/. [Accessed: May 2, 2025].