Analyzing and Classifying Customer Returns with Machine Learning Models

MLDS 400 - Team 6

Muhammet Ali Buyuknacar, Zhiwei Gu, Jialong Li, Siyan Li

Executive Summary

Customer returns pose a great challenge to department stores as they can result in high costs. Through meticulous data analysis and advanced modeling techniques, we developed a robust system that efficiently identifies patterns and trends in customer returns. By leveraging the predictive capabilities of our SVM model, department stores can optimize their inventory management strategies, ultimately minimizing the occurrence of returns and resulting in a Return on Investment (ROI) rate of 11.81%. Further enhancements to this model involve the incorporation of customer segmentation and basket content analysis, allowing for a more personalized and customer-centric approach to inventory management. By tailoring customer returns with the help of this model, Dillard's could not only increase profit but also foster long-term customer satisfaction and loyalty.

Introduction

In light of the growing competition, it becomes paramount to cultivate a deep understanding of customer behaviors. A critical concern in this regard is the occurrence of customer returns, a phenomenon associated with reduced revenue and customer dissatisfaction. This project centers on the crucial task of predicting these returns, offering companies the opportunity to strategically align their inventory management to mitigate both the frequency and subsequent financial repercussions of return incidents. To achieve this, we advocated for a machine-learning approach that leverages store location and product information for accurate return predictions. This methodology could provide valuable insights for optimizing operational efficiency and reinforcing customer-centric frameworks for Dillard's.

Data Cleaning & EDA

We were given 5 datasets including STRINFO, SKSTINFO, SKUINFO, TRNSACT, DEPTINFO. In order to handle the large dataset especially for TRNSACT, we made use of an SQL server to merge these datasets. During the uploading process, we checked the order of columns and excluded columns without clear meanings. In the SKUINFO table, certain data entries contained extra commas, resulting in additional columns upon dataframe loading. Given their negligible proportion, we removed rows with such anomalies. Furthermore, rows with null values in COST and RETAIL were omitted, as these variables play a crucial role in ROI analysis. Subsequently, all datasets were aggregated into a comprehensive dataset.

The subsequent step involved grouping purchases and returns to match each return with its corresponding purchase record and prevent duplications. Employing SQL, we utilized self-joins and unions to streamline this process. Matching returns with purchases was based on SKU, INTERID, and a SALEDATE difference within a 30-day window, aligning with the return policy. Rows with all zeros in INTERID were discarded, ensuring each return correlated with precisely one purchase. This meticulous process yielded 20,767,217 rows for purchases and 1,097,188 rows for returns.

For the dependent variable, we recorded "P" from the STYPE column as 1 and "R" as 0, resulting in 19,101,307 purchases and 1,011,886 returns.

Feature Engineering

For the COLOR and SIZE columns from the SKUINFO table, we grouped them to create new features. The COLOR column initially included 71,322 distinct colors, posing analytical challenges. Employing a color dictionary, we grouped these into 12 major categories, treating the remaining ones as the 13th group (Figure 1). A similar categorization was applied to the SIZE column, differentiating sizes into large, small, one, and other categories. These results were then transformed into dummy variables (Figure 2).

For SALEDATE column, since we believe customer behavior may differ during holidays or weekends, we created dummy variables for whether a date is a holiday (holiday_flag) or not and whether a date is a weekend or not (weekend_flag). Additional features included the calculation of discounts based on AMT / (ORGPRICE * QUANTITY). Rows with 0 in ORIPRICE, signifying an unreasonable scenario, were expunged to prevent infinite discount rates. A dummy variable column denoting whether an item is discounted was also introduced.

As it can be seen in Figure 3, to understand the difference between regions, we would like to consider socio-economic factors including GDP, population, poverty rate, and median income. By associating each zip code with these factors, we applied clustering to condense variables in two groups. The resulting cluster number was transformed into a dummy variable, with 36483 zip codes in group 0 and 1092 zip codes in group 1, facilitating its integration into the analysis.

Modeling

Logistic Regression

Logistics regression is a commonly used model with high interpretability. To rebalance the samples in the training set, we applied undersampling on the majority class to reduce the size of the dataset, and the ratio between the majority and minority classes is set to 1:1. Based on VIF, we removed columns including "color_group_other", "size_group_other", "QUANTITY", "holiday_flag", "ORGPRICE", "AMT". Grid search is used to find the best hyperparameters. For the training set, the accuracy is 0.606 and the F1 score is 0.604. For the test set, the accuracy is 0.601 and the F1 score is 0.741. All coefficients have p-values less than 0.05, which means high significance.

Tree-Based Algorithms

Compared with logistic regression, tree models could better capture complex relationships between variables and deal with nonlinear cases. We applied three different types of trees, including Decision Tree, Random Forest, and Gradient Boosting Tree. Each model is tested on multiple different maximum depths.

Support Vector Machine

SVM usually performs well for binary classification tasks with high dimensions and can model different types of relationships. We use random oversampling to balance our data set. By testing on the validation sets, we figured out the best ratio of Purchase to Return was about 3:1. Similarly to what we did for the logistic regression model, we dropped "color_group_other", "size_group_other", "QUANTITY", "holiday_flag", "ORGPRICE", "AMT". The best kernel is ‘rbf’, and it achieves an F1 score of 0.94.

Model Benchmarking

According to the model benchmark table as can be seen in Table 1, tree-based algorithms have the highest F1 score among our model experimentations. However, when we look at the confusion matrices, we see that the models are quite optimistic as they simply predict everything as a “Purchase”. Therefore, we have decided to use the Support Vector Machine algorithm which has almost as good an F1-score of 0.94 with more balanced predictions as the confusion matrix of the model can be seen in Table 2.

In conclusion, our final strategy is to use the Support Vector Machine (SVM) model with the “rbf” kernel.

ROI Analysis

We calculated our investment cost based on the discounted salary and cloud infrastructure costs as can be seen in Tables 3 through 6, which has a total of $201,268.82. As our model allows the company to have better inventory management that would reduce the operating expenses, we based our calculations on Selling, General, and Administrative Expenses for 2022 and discounted it to 2006, which is operating expenses minus the depreciation and amortization costs. We analyze the return on investment by simulating three different scenarios which are optimistic, most likely, and pessimistic and our ROI values are 39.76 %, 11.81%, -16.14% respectively. On average, our expected return on investment is 11.81% indicating that this project is worth investing (Table 7).

Risks

While the risks associated with our classification model are relatively minimal, they primarily center around the potential for inaccurate predictions. Inaccuracies may result in customers being unable to purchase desired items or an excess of stocked products, both of which can lead to a decrease in revenue and an increase in costs. To mitigate these risks, the implementation of more sophisticated models can enhance our understanding of user behavior and improve the precision of predictions.

Conclusion

In an ideal scenario, this model serves as a valuable tool to complement inventory management practices. When evaluating a new product for purchase from a vendor, the company can input its information, including store location and product information such as color, size, price, into the model. By assessing the likelihood of the product being returned, the company can make informed decisions about its suitability for the store or adjust inventory levels accordingly.

To further enhance the model's capabilities, it is recommended to incorporate additional customer-related information, conduct a more in-depth analysis of each customer segment, and refine predictions to offer a more personalized approach. Additionally, exploring the contents of customers' shopping baskets, such as the quantity and types of products, may also offer valuable insights. This iterative process can contribute to the effectiveness of the model in minimizing return rates while maximizing overall profitability.

REFERENCES

Bureau, U. C. (2021, October 8). *State and County Poverty Estimates for 2005*. Census.gov. https://www.census.gov/data/datasets/2005/demo/saipe/2005-state-and-county.html

Glassdoor, *Company Salaries*
https://www.glassdoor.com/Salaries/index.htm

Google Cloud, *Pricing per product*.
https://cloud.google.com/pricing/list

Ofer, D. (2018, March 18). *US Zip Codes to County State to FIPS Crosswalk*. Kaggle. https://www.kaggle.com/datasets/danofer/zipcodes-county-fips-crosswalk/

U.S. Bureau of Economic Analysis, "CAGDP1 County and MSA gross domestic product (GDP) summary" https://apps.bea.gov/

U.S. Bureau of Labor Statistics, *CPI Inflation Calculator*
https://www.bls.gov/data/inflation_calculator.htm

Yahoo! (2023, December 8). *Dillard's, Inc. (DDS) Income Statement*. Yahoo! Finance.
https://finance.yahoo.com/quote/DDS/financials?p=DDS

APPENDIX

Table 1. Model Benchmark Table

Model	F1 Score
Logistic Regression	0.75
Decision Tree	0.97
Random Forest	0.97
Gradient Boosting Tree	0.97
Support Vector Machine	0.94

Table 2. Confusion Matrix of the Final Model SVM

	Predicted Purchase	Predicted Return
Actual Purchase	76,234	5,065
Actual Return	4,373	695

Table 3. Median Salaries

Role	**2023 Median Annual Salary**	**2006 Discounted Median Annual Salary**
Data Scientist	$156,137.00	$103,492.89
Machine Learning Engineer	$151,158.00	$100,192.6
Data Engineer	$124,820.00	$82,734.92

Table 4. Salary Costs

Number of Data Scientists	2
Number of Machine Learning Engineers	1
Number of Data Engineers	1
Project Start Date	2006
Project Duration in Months	6
Project Duration in Years	0.5
Salary Cost (2023)	$ 294,126.00
Salary Cost (Discounted, 2006)	$ 194,956.67

Table 5. Google Cloud Pricing List

Category	vCPUs	Memory	Price per hour
e2-highcpu-16	16	16 GB	$0.46

Table 6. Computing Cost

Computing Hours	4320
Cloud Computing Cost (2023)	$ 1,979.60
Cloud Computing Cost (Discounted, 2006)	$ 1,312.15
Cloud Migration Cost	$ 5,000

Table 7. ROI Analysis Table

Investment Cost	$ 201,268.82
Selling, General and Administrative Expenses (2022)	$ 1,697,500
Selling, General and Administrative Expenses (Discounted, 2006)	$ 1,125,166
% Savings in Operating Expenses (Optimistic)	25
Net Return (Optimistic)	$ 281,291.45
% Savings in Operating Expenses (Most Likely)	20
Net Return (Most Likely)	$ 225,033.16
% Savings in Operating Expenses (Pessimistic)	15
Net Return (Pessimistic)	$ 168,774.87
Return on Investment (Optimistic)	39.76 %
Return on Investment (Most Likely)	11.81 %
Return on Investment (Pessimistic)	-16.14 %
Expected Return on Investment (ROI)	11.81 %

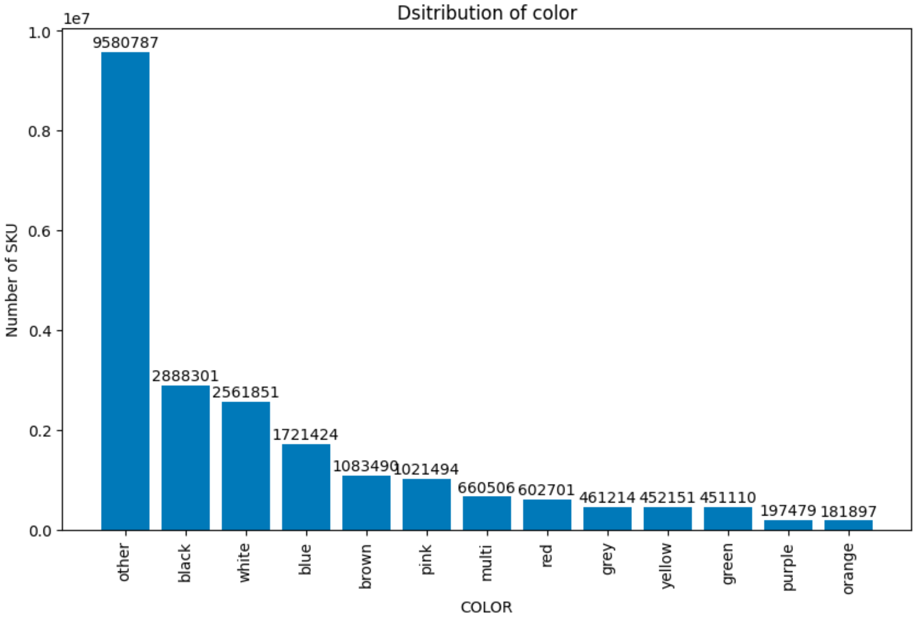

Figure 1: Color Grouping

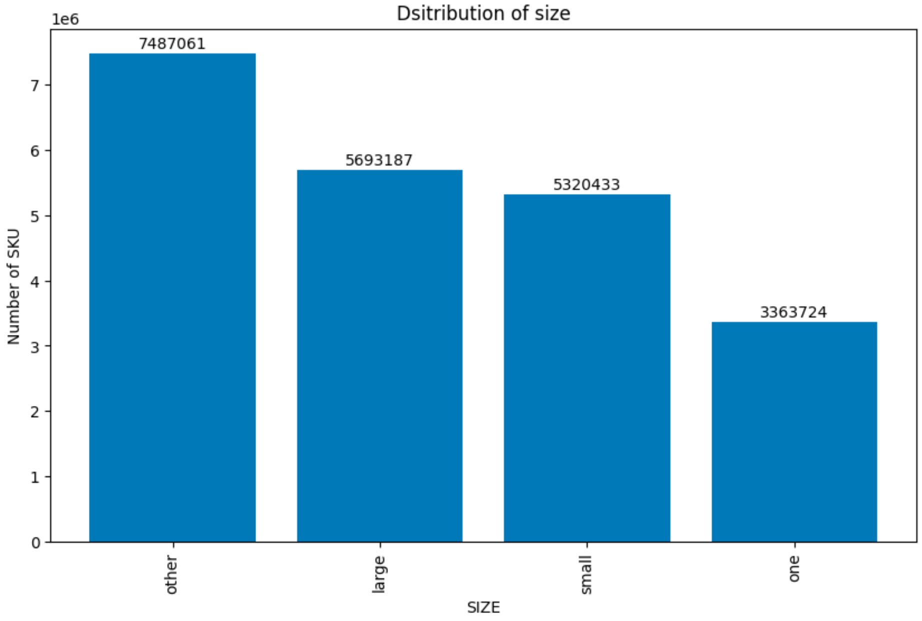

Figure 2: Size Grouping

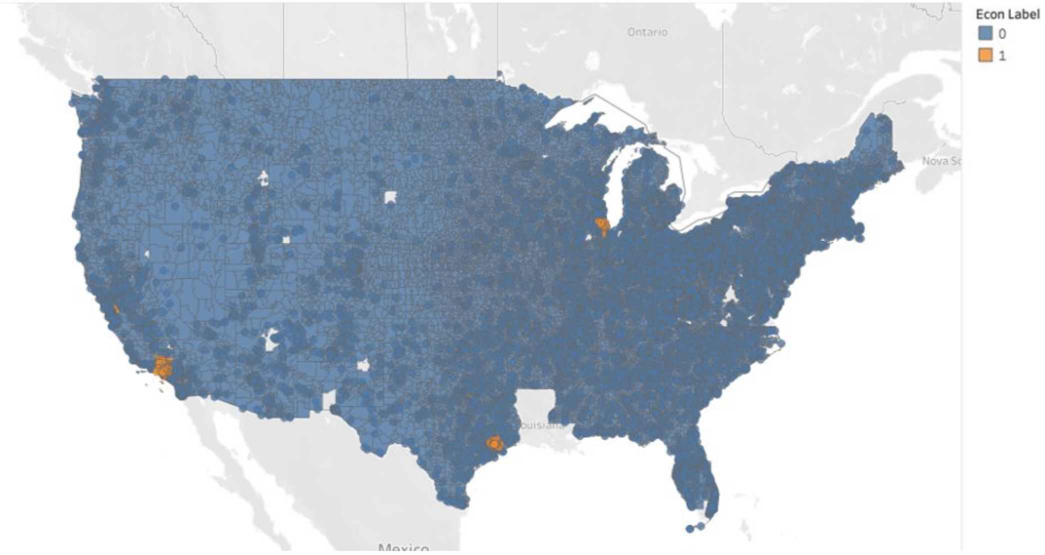

Figure 3: Socio-economic Factors Analysis by ZIP Codes