Table of Contents

Final Report

Group 2: Kira Luo, Rishabh Setty, Yan Wang, Yuan Zhang

Executive Summary

Our study for Dillard leverages machine learning to mitigate Black Friday's high product return rates. By analyzing product features and utilizing SMOTE, K-means clustering, and Random Forest algorithms, we've developed a model to predict customer retention of products. This concise report details our methodology, initial results, and proposes a strategic implementation for the Black Friday sales, aiming to transform Dillard's inventory management and enhance customer satisfaction.

Introduction and Objectives

The onset of Black Friday ushers in a period of heightened retail transactions but also brings the challenge of increased product returns, presenting a dual-edged sword for businesses like Dillard. The ensuing costs—from augmented shipping demands to the administrative load of restocking—necessitate a strategic approach. This project's cornerstone is the application of machine learning techniques to mitigate these post-sale challenges. Our mission is two-fold: to minimize return rates, thus reducing ancillary costs, and to enhance overall customer satisfaction through predictive analytics. We have synthesized a robust dataset capturing diverse product features to conduct a thorough analysis in two phases. Initially, we assessed the data to gauge the frequency and distribution of product returns. The subsequent phase involved adjusting the data's balance using SMOTE, enhancing the fairness and accuracy of our predictive model. Our approach culminated in the creation of machine learning models to discern and predict the return likelihood ('R') versus purchase likelihood ('P') of products. This comprehensive report outlines our analytical journey, showcases the technical underpinnings of our model, and reflects on the broader implications of our research. Ultimately, it provides a tactical blueprint for leveraging data-driven insights to refine product recommendations and bolster Dillard's Black Friday sales outcomes.

Exploratory Data Analysis

The Dillard’s have millions of rows of raw data, which is composed of five tables: deptinfo, strinfo, sksrtnfo, skuinfo, and trnsact. Since all the datasets miss titles, we assign titles to each column of the tables according to the schema given and adjust when necessary. We gathered much basic information about Dillard's POS data from these tables by summary statistics. For example:

- Dillard's has 59 departments in total
- Dillard's has 452 different stores, which are located in 298 different cities and 31 different states
- The POS data has 120,916,896 transactions, and the total amount of transaction charge has the minimum of 0, maximum of 6,017, and average of 24.62.
- The store with id 8402, located in Metairie city, LA 70002, is the busiest store based on their number of transactions (944,982)

In order to better investigate our research topic, we integrated tables we needed by inner joining trnsact, skuinfo, and sksinfo tables. We selected a subset of data to work with, which is the dataframe during the time 2004-11-26, approximately 220732 rows (Table 1 in Appendix shows how we select the desired dataframe). In the analysis of the subset data, we have discovered that a significant portion of retail prices for products falls below $200 (Figure 1 in Appendix). Furthermore, our examination of the Return vs. Purchase Proportion highlights an imbalance in the dataset (Figure 2 in Appendix).

Before starting on feature engineering, our initial focus lies on data cleaning. This involves several steps, including checking data types to ensure consistency, addressing missing data through appropriate strategies, rectifying inconsistent text and typos for data uniformity, and identifying and removing duplicate entries and outliers. These actions lay the groundwork for a more reliable dataset, setting the stage for subsequent processes.

Feature Engineering

Our approach to refining the dataset began with standardizing features critical for model accuracy. Colors were categorized into 15 distinct groups, and sizes into 21 categories, accommodating a broad range of products. Retail Prices and Pack Sizes were scaled to neutralize the bias due to variable magnitudes. We then grouped items by SKUs to calculate return rates, labeling those with rates >=60% as 'likely to be returned' and others as 'more likely to be purchased.' To counteract the imbalanced nature of our data, SMOTE was employed, facilitating a balanced representation between returned and purchased products. Categorical variables underwent One-Hot encoding, converting them into a binary matrix essential for machine learning algorithms to process without the artificial imposition of ordinality. Visual aids such as histograms and pie charts for Retail Prices, Pack Sizes, and standardized categorical variables validated our data manipulation process and offered insights into the distribution and relationship of the variables within our dataset. These measures ensured a robust foundation for our predictive modeling, essential for achieving our goal of reducing product returns for Dillard. Our final variables of interest could be referenced by Table 2 in Appendix.

Modeling

Model 1: Logistic regression

Logistic Regression is a statistical model that models the probability of an event taking place by having the log-odds for the event be a linear combination of one or more independent variables. Our response variable here is Predict P or R (binary: P vs R) and our predictor variable matrix is composed of retail price, style, standardized_color, standardized_size, packsize, vendor, brand. The confusion matrix result of the logistic regression indicates that the model predicted purchased items more accurately than class 'R' (second row), with 190 true positives and 53 true negatives, accompanied with 67.18% accuracy and 78% precision for purchase class (Model 1 in Appendix). Yet, it does not perform well on the returned class with only 49% precision.

Model 2: K-means Clustering + Logistic Regression

We integrated K-Means clustering as a form of feature engineering and then used a logistic regression to further predict the binary response variable. Clustering assignments were used as additional features, hypothesizing that certain clusters may correspond to higher or lower likelihoods of returns. We experimented with different cluster counts, ranging from 3 to 15. Since the highest accuracy of 67.68% is achieved by k = 12 (Model 2). We choose k = 12 here because all three numbers are not large enough to affect the accuracy and the more clusters we add, the easier it is for the algorithm to reduce the distance between points and centroids, reducing the within variability. K-means Clustering model's accuracy improved with the inclusion of cluster assignments. This suggests that our feature engineering successfully captured additional patterns within the data that were indicative of return likelihood.

Model 3: Random Forest

Our model in Appendix Model 3 was initialized with 100 estimators and a fixed random state to ensure reproducibility. We fitted the model using our resampled training data, which was balanced using SMOTE to address the initial skew towards purchased products. Upon training, the model's performance was evaluated on a separate test set. The accuracy metric stood at approximately 63.6%, a respectable figure considering the complexity of consumer behavior. The precision for predicting purchases (P) was notably higher than returns (R), indicating a stronger confidence when predicting the more frequent class. The classification report provided deeper insights, with a precision of 0.75 and a recall of 0.70 for purchases, while returns had a precision of 0.44 and a recall of 0.50. The F1-score, a harmonic mean of precision and recall, was consistent with these findings.

Model Selection

By comparing the accuracy matrices of the three models, we choose model 2: K-means Clustering + Logistic Regression as our final model to adopt. The reasons are listed as follows. We observed that Model 1 and Model 2 demonstrated quite similar precision, recall, and F1 scores. However, Model 2 emerged as the superior performer, achieving the highest accuracy of 67.68% when the number of clusters was set to 12. Although the differences in accuracy among the models are marginal, if this applies to more than 1000 different SKUs in the future, this slight improvement could make a huge impact.

Return on Investment (ROI) Analysis

In our analysis of Black Friday sales, the baseline scenario indicates an offline revenue of $1,885,912.21 with a 56% lift rate, assuming a 50% success rate in classifying purchased items. However, with the implementation of our model, the lift rate increases to 78%, resulting in a higher revenue of $2,942,023.05. For returned items, the baseline scenario, assuming a 50% success rate in classifying returns, yields a revenue of $1,885,912.21. Interestingly, the implementation of our model maintains the same revenue, indicating that improvements in the classification of returned items were not significant. Investments in labor costs, including data scientists, a full-stack engineer, and sales training, amount to $466,027. Additionally, computing costs, based on AWS usage, total $119.81. Open-source tools are utilized for added cost efficiency. Considering these factors, the total Return on Investment (ROI) stands at $589,964, with an impressive rate of 227%. This suggests that the implementation of our model significantly enhances the overall performance and revenue generation during Black Friday sales. Sources for our analysis include reputable statistics on apparel retail sales, department store opportunities, and retail marketing insights. Additionally, financial results from Dillard's Inc. provide valuable context for the department store sector.

Risks and Mitigations

The success of ROI projections is heavily dependent on the accuracy of the implemented model, posing a Model Accuracy Risk. Inaccuracies in classifying purchased and returned items can result in suboptimal decision-making, directly impacting the expected revenue lift. Furthermore, Market Dynamics Risk introduces the influence of external factors such as economic conditions, consumer sentiment, and competitive strategies on Black Friday sales. Unforeseen market dynamics or shifts in consumer behavior may lead to deviations in actual transaction counts and success rates from the initially projected ROI.

Additionally, there is a Technology Implementation Risk associated with the seamless integration of the model with existing systems. Technical issues, compatibility problems, or delays in implementation can hinder the model's performance, thereby affecting the expected ROI. The quality and integrity of transactional data used for model training are critical factors, constituting Data Quality and Integrity Risk. Incomplete or inaccurate data may result in biased predictions, impacting the reliability of ROI projections. Lastly, Competitive Response Risk emphasizes the importance of understanding and mitigating potential responses from competitors who may adopt similar models. If competitors implement comparable strategies, the anticipated revenue lift may be diluted due to increased competition, underscoring the need for strategic planning to sustain the projected ROI.

Addressing the limitations in the return prediction accuracy of our K-means model, we propose a multifaceted mitigation strategy. This includes the integration of additional predictive variables that may influence return behavior, such as customer reviews and seasonal trends. Additionally, we should adopt advanced machine learning techniques like ensemble learning to enhance the model's robustness. Regular updates to the training dataset will ensure the model adapts to evolving consumer patterns, while cross-validation methods will be employed to rigorously assess the model's performance and prevent overfitting. These measures aim to refine the model's predictive accuracy, ensuring a more balanced and comprehensive approach to forecasting Black Friday sales outcomes.

Conclusion

Our K-means model has significantly improved the identification of purchased items during Black Friday sales, as reflected by a 227% ROI. However, the model's impact on returns classification was nominal, suggesting a need for further model optimization. Despite this, the substantial increase in revenue underscores the model's value in enhancing sales performance. Future efforts will concentrate on improving return predictions, with the goal of achieving a comprehensive enhancement of Black Friday revenue streams for Dillard's.

Reference

ROI Analysis Link:
https://docs.google.com/spreadsheets/d/1iM0TvwdE0M-nmOqlamIUMMtnQBbar4u6/edit?usp=sharing&ouid=117001893924717839918&rtpof=true&sd=true

Appendix

Table 1. Dataframe Selection

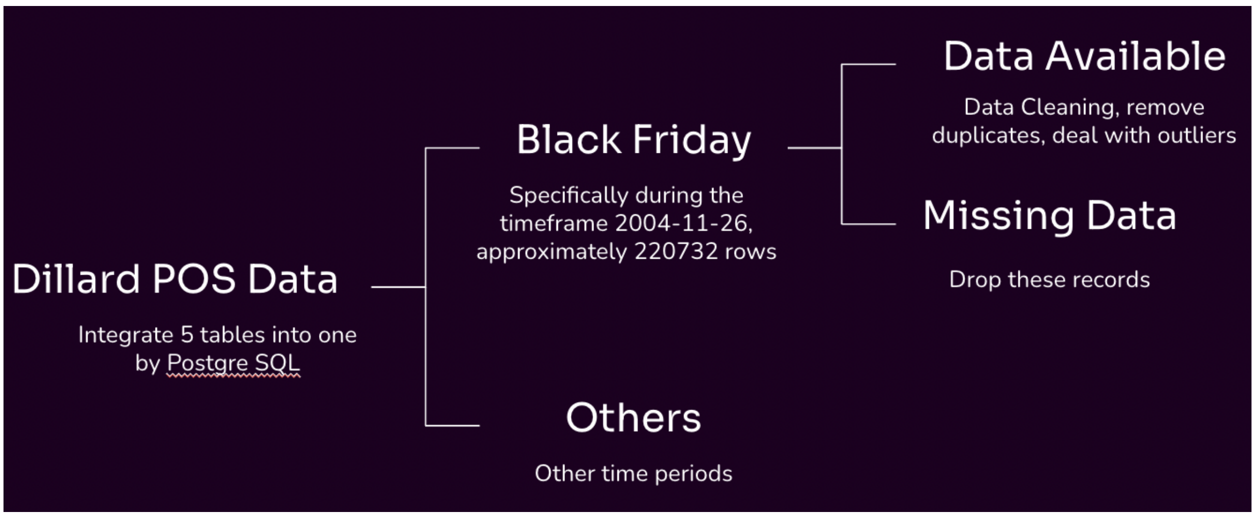

Figure 1. Density of Retail Price

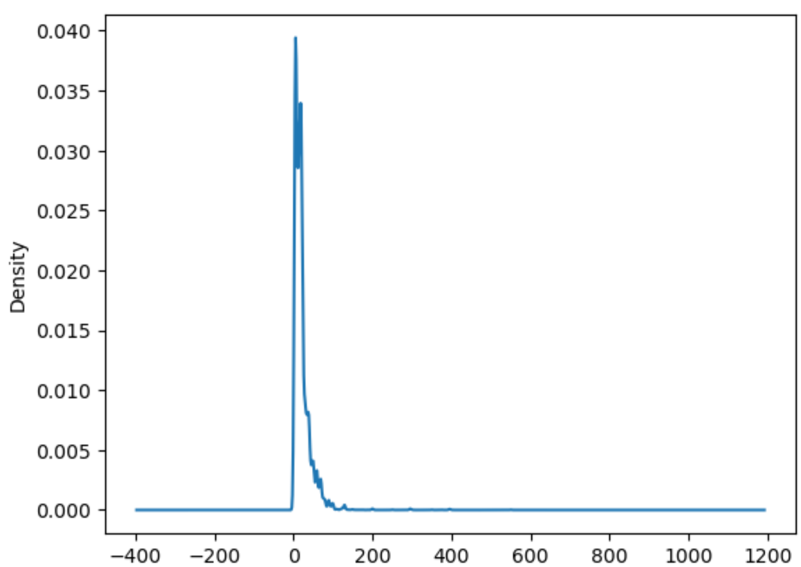

Figure 2. Return vs. Purchase Counts

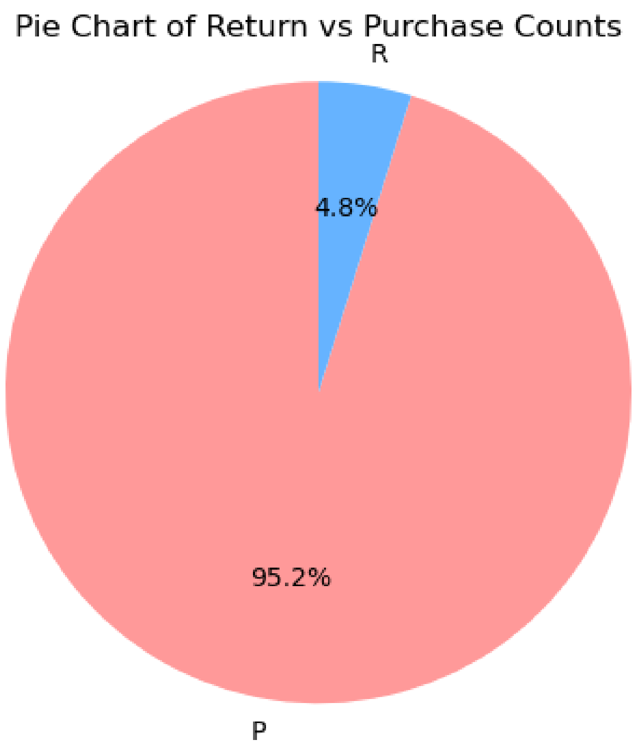

Table 2. Variables of Interest

Response Variable	Binary: Purchase or Return; column name “Predict P or R”
Predictor Variables	All of the followings
SKU	Stock Keeping Unit number of the stock item
Style	Categorical, the specific style of the stock item
Retail	Numerical, the selling price of Dillard’s products
standardized_color	Categorical, the color of the stock item after categorization

standardized_size	Categorical, the size of the stock item after categorization
vendor	Categorical, the vendor number of the stock item
brand	Categorical, the brand name of the stock item
Packsize	Numerical, the quantity of item per pack

Model 1. Logistic regression

```
              precision    recall  f1-score   support

           P       0.78      0.71      0.75       266
           R       0.49      0.58      0.53       127

    accuracy                           0.67       393
   macro avg       0.64      0.65      0.64       393
weighted avg       0.69      0.67      0.68       393

Accuracy: 0.6717557251908397
```

Model 2. K-means clustering + logistic regression

```
Number of Clusters: 12
              precision    recall  f1-score   support

           P       0.78      0.72      0.75       266
           R       0.50      0.58      0.54       127

    accuracy                           0.68       393
   macro avg       0.64      0.65      0.64       393
weighted avg       0.69      0.68      0.68       393

Accuracy: 0.6768447837150128
```

Model 3. Random Forest

```
Accuracy: 0.6361323155216285
Confusion Matrix:
[[187  79]
 [ 64  63]]
Classification Report:
              precision    recall  f1-score   support

           P       0.75      0.70      0.72       266
           R       0.44      0.50      0.47       127

    accuracy                           0.64       393
   macro avg       0.59      0.60      0.60       393
weighted avg       0.65      0.64      0.64       393
```