

Return vs Purchase Prediction for Black Friday Sales Analysis

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

Table of contents

01

Problem
Overview

02

EDA

03

Feature
Engineering

04

Modeling

05

Model
Selection

06

ROI

The background is a dark purple color. It is decorated with various abstract geometric shapes and lines in different colors. These include straight lines in shades of pink, blue, green, and orange. There are also curved lines and shapes that resemble paper clips or loops in colors like dark blue, light blue, purple, and green. The overall aesthetic is modern and graphic.

01

Problem Overview

Minimize Return Rate by ML

Problem:

Black Friday is the biggest retail sales day in the United States. Yet, it accompanied with a spike in return rate, which increases shipping cost, administrative costs, and many other costs for businesses. Our project aims to minimize the return rate after Black Friday for Dillard.

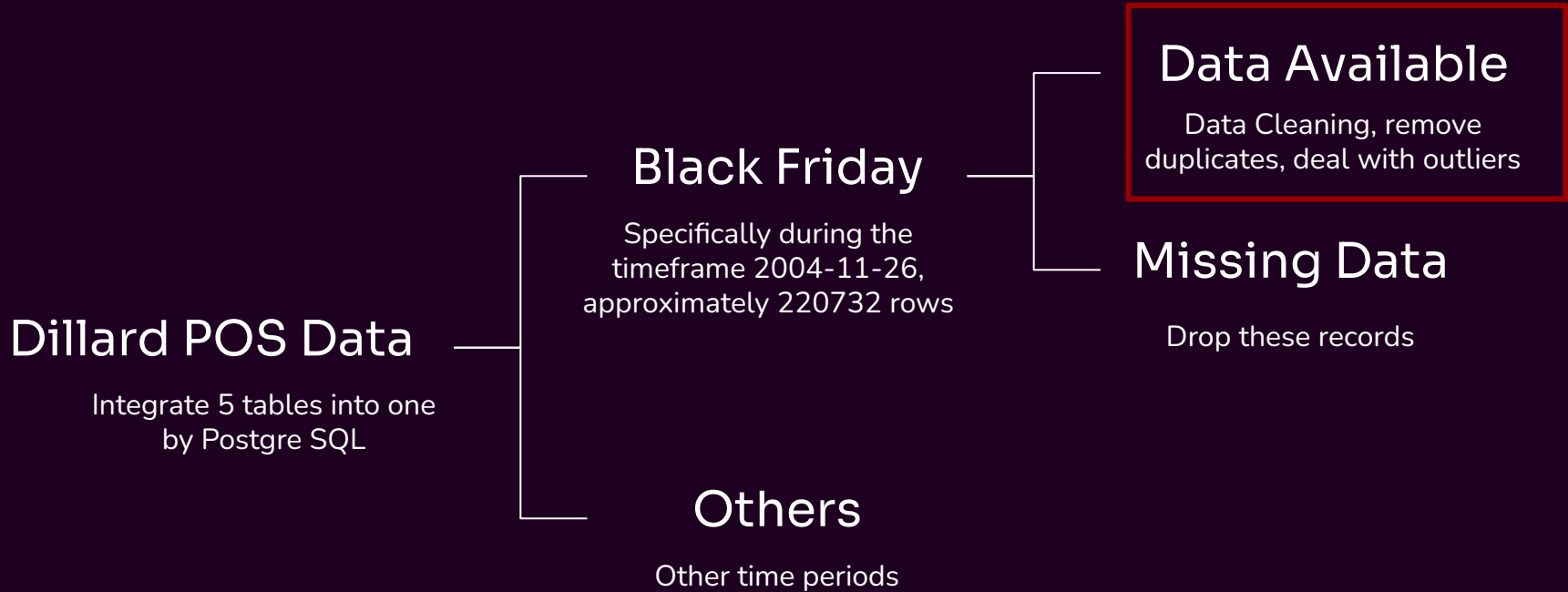
Research Topic:

How to recommend customers with the right types of products to minimize the return rate after Black Fridays?

Approach:

Design a machine learning model that could predict products with which features would be purchased or returned. During the Black Friday Sales, let the website recommend customer more with the purchased products instead of probably returned ones.

Data Preprocessing





02

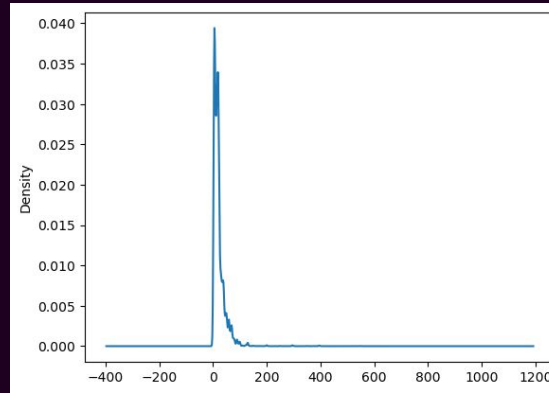
EDA

Data Insights

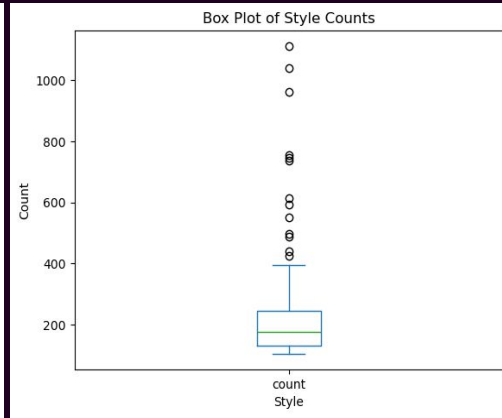
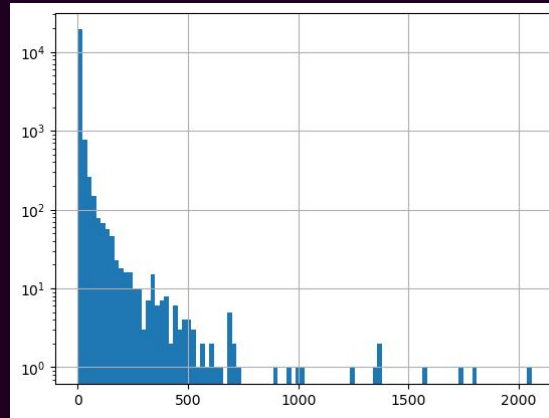
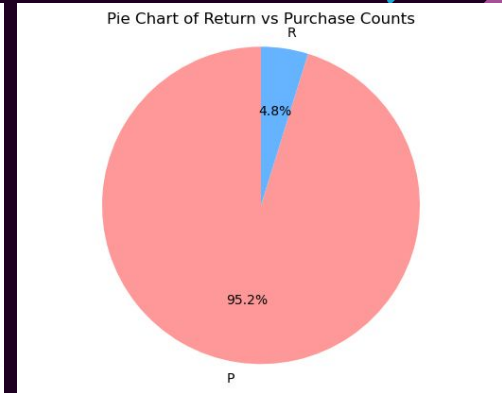
During Black Friday, among all transactions:

1. Most retail prices of products are in the range of below 200 dollars
2. Most popular colors are Black, White, Multi, Navy, Silver, Red, and Pink among all.
3. Return vs Purchase Proportion indicates an imbalanced dataset
4. Deal with right skewed styles distribution by setting thresholds

Retail Price Density Plot



Pie Chart of Return vs Purchase Prop



Style bar chart and box and whisker plot

Exploratory Data Analysis

Remedies to dirty Data:

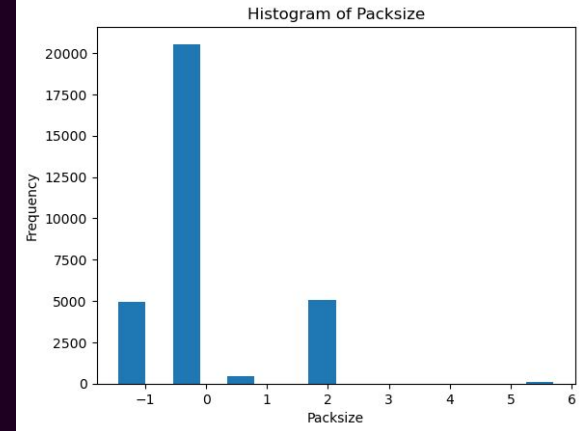
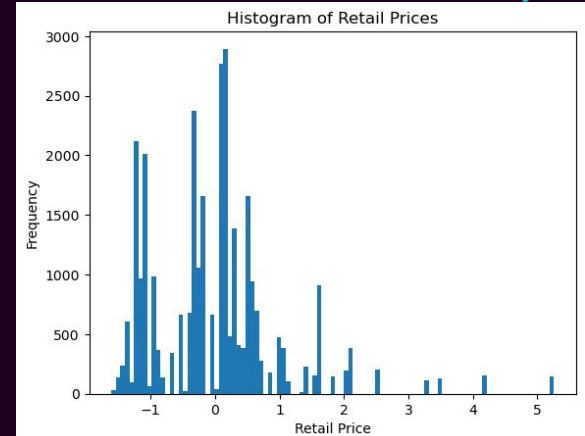
Step 1: Check the data types

Step2: Deal with missing data

Step 3: Deal with Inconsistent Text & Typos

Step 4: Remove duplicate and Outliers

Histogram of Retail Prices after scaling



Histogram of Packsize after scaling



03

Feature Engineering



Variables of Interest

We select a subset of variables that would be useful for building our models

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

Data Manipulation

Remedies to dirty Data:

Step 5: Standardize color into 15 categories and size into 21 categories (include shoes & clothes)

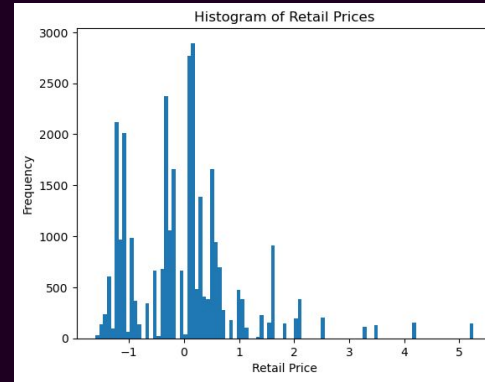
Step 6: Scale Retail Price and Pack Size

Step 7: Group data according to SKU, calculate the return rate for each SKU, products with return rate $\geq 60\%$ are classified as “predicted to be returned” as R and others as “more likely to be purchased” as P

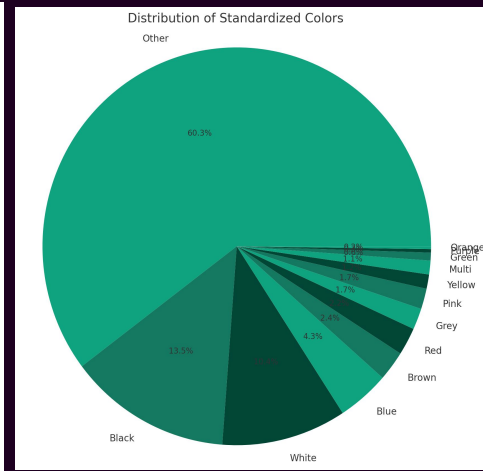
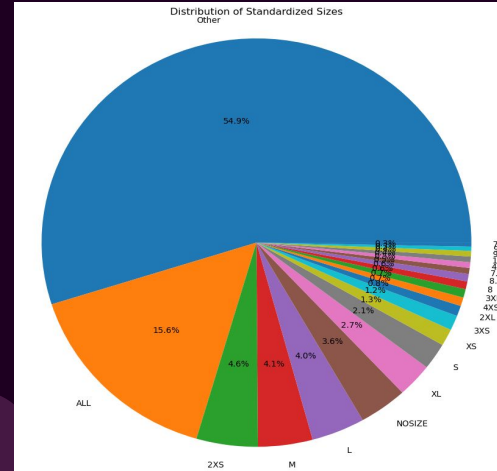
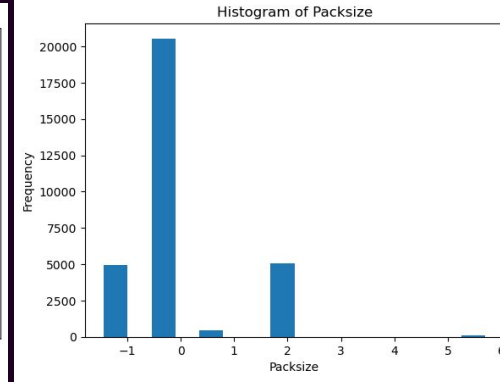
Step 8: using SMOTE to deal with imbalanced dataset between return vs purchase proportion

Step 9: One Hot encoding for categorical values

Histogram of Retail Prices after scaling



Histogram of Packsize after scaling



The background is a dark purple color. It is decorated with various abstract geometric shapes and lines in different colors. In the top left, there is a blue squiggly line and a brown paperclip-like shape. In the top right, there are blue and purple diagonal lines and a thin orange line. In the bottom left, there are dark blue and light green shapes. In the bottom right, there are dark blue diagonal lines, a bright pink diagonal line, and a green squiggly line.

04

Modeling



Here are 3 models we built:

Among 1965 unique SKUs, we test which models predict best for return vs purchase:

Model 1: Logistic Reg

It 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.


Model 2: K-means+Logistic

Step 1: using K-means clustering, it is a method of vector quantization, originally from signal processing, that aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster.

Step2: Then, we use logistic regression.

Model 3: Random Forest

It is an ensemble learning method for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time.



Logistic Regression

Y: Predict P or R (binary: P vs R)

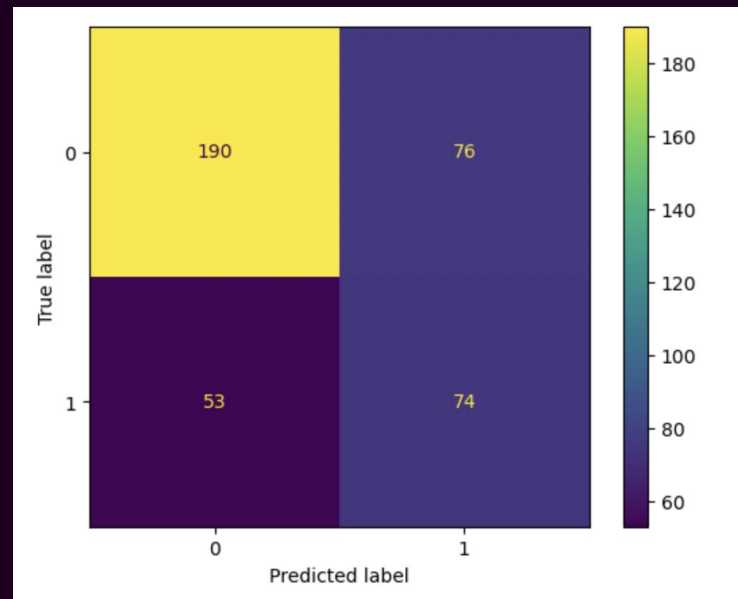
X matrix: retail price, style, standardized_color, standardized_size, packsize, vendor, brand

Key Takeaway:

- 67.18% Accuracy
- 78% precision for purchase class
- Yet, 49% precision for return classification

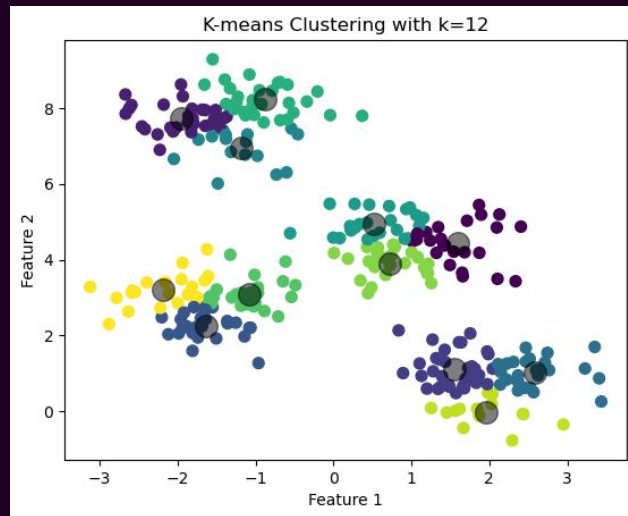
	precision	recall	f1-score	support
P	0.78	0.71	0.75	266
R	0.49	0.58	0.53	127
accuracy			0.67	393
macro avg	0.64	0.65	0.64	393
weighted avg	0.69	0.67	0.68	393

Accuracy: 0.6717557251908397



K-means Clustering + Logistic

- This analysis evaluates clustering performance from 3 to 15 clusters, and it performs best when $k = 12$
- We choose $k=12$ here because this number is 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.
- Key Takeaway: Highest accuracy of 67.68% when $k=12$



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					

Random Forest

Accuracy: 63.6%

Classification Report:

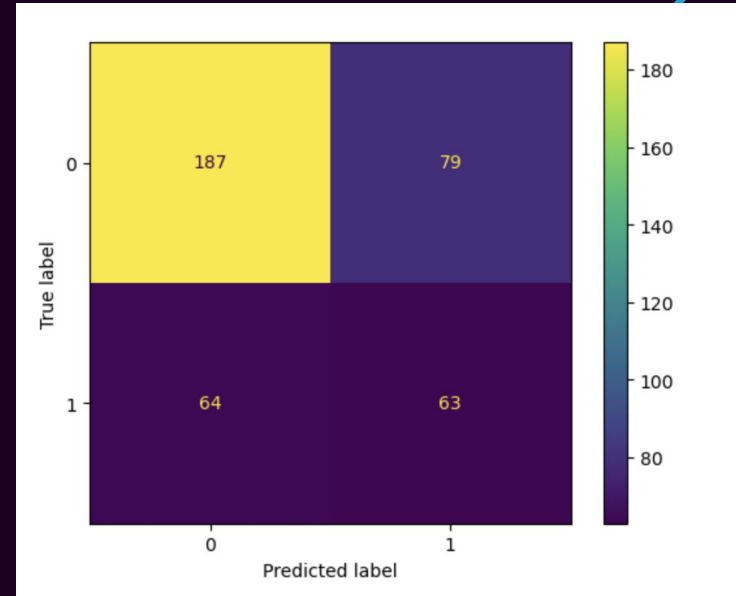
- Precision (False): 0.44, Recall: 0.50, F1-Score: 0.47
- Precision (True): 0.75, Recall: 0.7, F1-Score: 0.72

Additional Metrics:

- Macro Avg: Precision 0.59, Recall 0.6, F1-Score 0.6
- Weighted Avg: Precision 0.65, Recall 0.64, F1-Score 0.64

Key Takeaway:

The model excels in identifying positive instances (Purchased class) but shows room for improvement in handling negative instances (Returned class). This is because the proportional number of items that being purchased and returned. The percentage of purchased is significantly higher than the number of returned.



Accuracy: 0.6361323155216285

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

The background is a dark purple color. It features several abstract, colorful elements: a green paperclip-like shape in the top left, a blue diagonal line, a red curved line, a dark blue paperclip-like shape, a pink paperclip-like shape, a light blue curved line, a grey paperclip-like shape, and a red diagonal line. These elements are scattered across the left and bottom portions of the slide.

05

Model Selection



Compare Models by Accuracy Matrix

Measure models' performances on the test set by the following measurements:

Accuracy

measures the number of correct predictions made by a model in relation to the total number of predictions made

Precision


refers to the number of true positives divided by the total number of positive predictions (i.e., the number of true positives plus the number of false positives).

Recall

the percentage of data samples that a machine learning model correctly identifies as belonging to a class of interest—the “positive class”—out of the total samples for that class

F1 Score

a measure of the harmonic mean of precision and recall. The F1 score integrates precision and recall into a single metric to gain a better understanding of model performance



Reasons for selecting Model 2

Highest Accuracy

Among all three models, Model 1 and Model 2 got pretty similar precision, recall and F1 score. But model 2 wins with highest accuracy of our K means model performed the best based on our accuracy of 67.68% when k=12.

Although models only differ slightly in accuracy, but if this applies to more than 1000 sku products in the future, this slight improvement could make huge impact.

Models	Accuracy
Logistics Only	67.18%
K-means + Logistics	67.68%
Random Forest	63.6%



06

ROI

Black Friday Sales ROI

Returns in Purchased Items:

- Baseline Model:
 - Rate of Purchasing: 50%
 - Revenue Generated: \$1,885,912.21
- With Our Model:
 - Rate of Purchasing: 78%
 - Revenue Generated: \$2,942,023.05
 - Lift Rate: 56%

Investments:

- Labor Costs: \$466,027
- Computing Costs: \$119.81

= \$589,964

Total ROI rate = Net Return/ Cost of Investments = 227%

*Assumptions:

- Offline Revenue Share: 55%
- Transaction Count on Black Friday: 500,000
- All items transactions are in-store



Thank you