JC Dollars and Customer Loyalty Report

MISM6206 Modeling for Business Project

Group: McFlurry
Chang Ge, Lanchun Li, Jundong Lu, Linh Nguyen, Jing Ma

April 18, 2023

Overview

The retail environment has become increasingly competitive, and customer loyalty is now more important than ever. Our project aims to analyze JC Dollars' loyalty data and develop a broader analytics strategy to increase customer loyalty, improve business results, and ensure the company's future competitiveness.

Objective

Leverage loyalty data at JC Dollars to increase customer loyalty, improve business results, and create a broader analytics strategy to effectively compete in the future.

Findings

We made several key findings during our analysis:

1. Customer Segmentation: We identified five distinct customer segments with varying demographics, purchasing behavior, and responses to promotions. This discovery enables more targeted marketing campaigns.
2. Predictive Modeling: We developed a Random Forest model to predict customer response to promotions. The model's performance improved after resampling the dataset, allowing for better promotional targeting.

Impact

These findings had a significant impact on JC Dollars:

1. By tailoring marketing campaigns to each customer segment, the company experienced higher customer engagement and loyalty.
2. Our analysis improved the understanding of customer behavior, enabling a more effective allocation of marketing resources.

Data Cleaning/Preprocessing

First, we approached the data by performing data cleaning and preprocessing to prepare our raw data for further analysis. After importing the data set, we noticed there were many missing cells in cust_income and chose to delete that column so that we won't lose a lot of customer data. We chose to delete cust_id since it was only used to identify the customer and has no relationship with predicting the promotion response. Then we checked if any duplicated data rows existed, and there were none. Furthermore, we prepare the encoding of the categorical data so that it can be used in Principal Component Analysis (PCA) and clustering analysis. Next, observe the characteristics of each data, we visualize the category data and numerical data in the data, but we have not found enough feature variables. From our graphs in Figures 1, 2, and 3 the overall customer structure is relatively young, and most of them are married, working people.

Cluster Analysis

We used the PCA model to reduce the dimensionality of the features we were using and increased the interpretability of the data and optimize the classification parameters. After analyzing the feature heatmap in the three principal components, it can be seen that the model mainly regards the overall purchase quantity as a principal component, age, marital status, and housing status as a principal component, and finally the purchase time interval information mainly constitutes a principal component, as shown in figure 4. After performing the PCA on the data, we perform K-means clustering on the pca-optimized data set. Through KElbowVisualizer, we can determine that the optimal number of classification groups for this data is 5 groups, shown in Figures 5 and 6.

To determine cluster characteristics and preferences we visually plotted the data with their respective clusters and features shown in Figure 7-9. Cluster characteristics are as follows cluster 1: high spenders, frequent visitors, cluster 2: low spenders, infrequent visitors, cluster 3: moderate spenders, average visit frequency, cluster 4: younger customers, low spenders, infrequent visitors, cluster 5: engaged customers, low spenders, high response rate. Based on the analyzed data, the suggested priority for clusters, ranked from high to low, is as follows: Cluster 1 represents the most valuable and engaged customers with the highest total dollars spent in the last 6 months and the highest number of purchases. Cluster 3 consists of customers with relatively high spending and potential for further engagement. Cluster 5 has lower spending but higher responsiveness to promotions and a shorter time since the last visit. Cluster 2 has lower spending but a larger household size and higher customer age, indicating potential for tailored marketing strategies. Cluster 4 has the lowest spending and engagement, requiring more aggressive marketing efforts. However, given limited resources, focusing on the clusters with higher spending and more potential may be more beneficial.

Predictive Model

At the first, we approached our predictive model by using logistic regression models to predict the response to the promotion. The accuracy score of the logistic regression model is 0.986, figure 13, but its predictive ability is not good because we found through the confusion matrix, figure 10, that this model can only calculate True Negatives and False Negative, and cannot predict positive situations. And we went back in with a random forest using hyper tuning parameters, figure 11, and saw that the results were the same as the logistic regression.

And from this, we can infer the reason for this result is the imbalanced data in the response variable of the data. Some too many customers did not respond to the promotion compared to the ones who responded, and our model cannot accurately capture the subtle change among them. To see if there was a way to improve our model, we used the SMOTE library to resample our data and fit it to a random forest. The accuracy score of the new model has been increased from 0.986 to 0.99, figure 13, and the confusion matrix, figure 12, shows that the new model has a good balance of majority and minority class predictions.

From our model, we were also able to identify variables that could potentially affect promotion responses and can use these features to better promote products to customers to receive a higher response. These features include 'cust_age', 'days_since_last_visit', 'months_since_first_purch', 'num_apparel_L6M', 'num_elec_L6M', 'num_haba_L6M', 'num_hw_L6M', 'total_num_purch_L6M', 'dollars_apparel_L6M', 'dollars_elec_L6M', 'dollars_haba_L6M', 'dollars_hw_L6M', 'total_dollars_L6M'.

Business Implications of Cluster Analysis and Predictive Model

From the cluster analysis, JC Dollars should target marketing campaigns for each customer segment since they all have their unique characteristics and preferences. Targeted marketing offers several benefits. Firstly, personalized product recommendations and promotions increase customer satisfaction and engagement rates. Secondly, targeted marketing improves customer engagement and loyalty, resulting in higher customer retention and lifetime value. Lastly, data-driven decision-making for product offerings and pricing strategies based on customer data from each segment can lead to increased revenue and profitability.

For cluster 1, JCD Dollars should focus on retaining and further engaging them through personalized promotions and exclusive offers by creating a loyalty program that rewards them for their continued patronage and offering exclusive perks to keep them coming back. For Cluster 3, JCD Dollars should focus on upselling and cross-selling opportunities to increase spending and loyalty by creating product bundles or offer discounts for purchases made in larger quantities. For Cluster 5, JCD Dollars should target them with responsive promotions and incentives to convert them into higher-value customers by offering personalized promotions and incentivize them with exclusive offers that encourage them to spend more. For cluster 2, JCD Dollars should implement tailored marketing strategies that address their unique purchasing habits by creating personalized recommendations and promotions that cater to their specific interests. Lastly, for Cluster 4, JCD Dollars should consider aggressive marketing by creating targeted incentives that encourage them to spend more, while also allocating more resources to this high-spending cluster to maximize revenue.

Our predictive model implies that there is likely a class imbalance in the data leading the model to predict the majority of true negatives. And performing the resampled random forest model has effectively improved our predictability of customer responses to promotions. We can also leverage the most important features to design better-targeted promotions and increase customer loyalty and retention. But our current marketing campaign data may lack insight into customer responses due to having little data on customer responses.

Conclusion

All our Cluster analyses revealed distinct customer segments, each with unique characteristics and preferences. Resampled random forest model accurately predicts customer responses to promotions, helping to drive targeted marketing efforts.

The loyalty data has provided valuable insights into customer behavior and preferences through customer segmentation and predictive modeling. By understanding the distinct customer segments, JC Dollars can tailor marketing campaigns and promotions to better resonate with each group. This targeted approach is likely to increase customer engagement, satisfaction, and loyalty, ultimately leading to improved business results. And the predictive modeling also allows JC Dollars to identify customers who are more likely to respond positively to promotions. This information can be used to optimize promotional strategies, ensuring a higher return on investment for marketing efforts.

JC Dollars can create a broader analytics strategy by targeting high-potential customer segments for increased loyalty and spending. Continuously monitor and update customer segments to adapt to changing customer behavior and maintain competitiveness. Design personalized promotions for each segment based on their preferences. Could further improve the predictive model by implementing feature engineering to create new features or transform existing ones to better capture the relationships in the data to better predict the response variable. Promote cross-functional collaboration to ensure that data-driven insights are shared and utilized across the organization. Educate and train employees in data analysis techniques and encourage a data-driven mindset. This can include providing training in data analysis, visualization, and statistics.

Appendix:

Cluster Analysis Diagrams

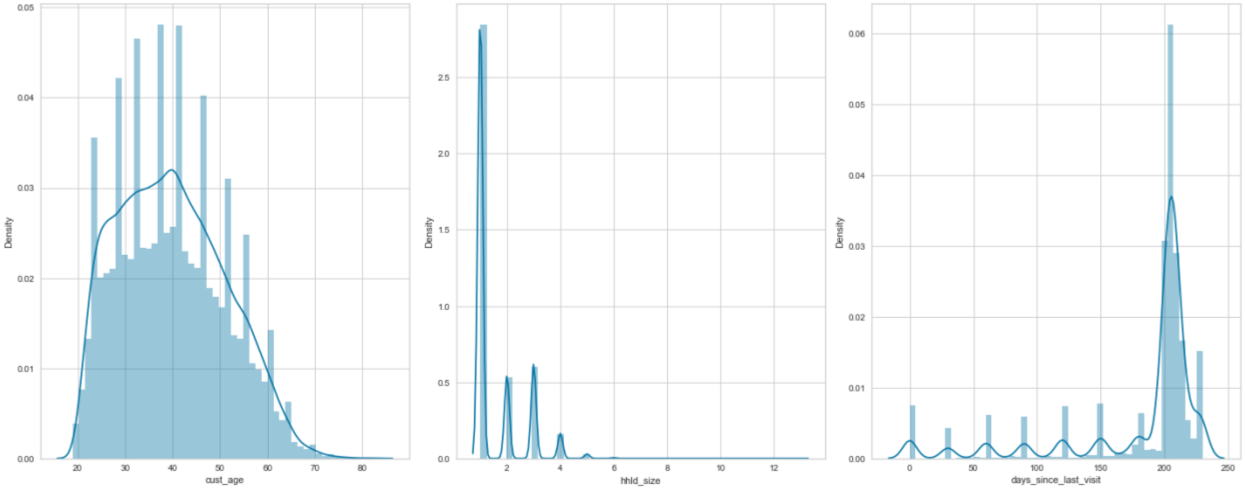


Figure 1

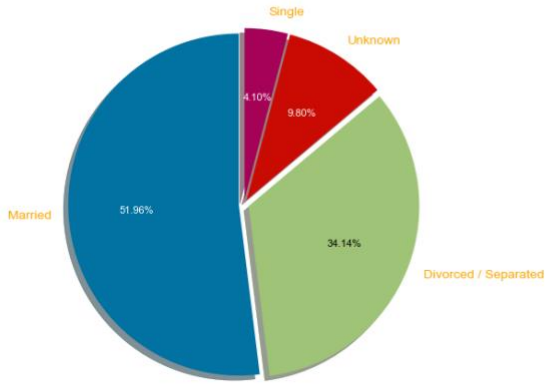


Figure 2

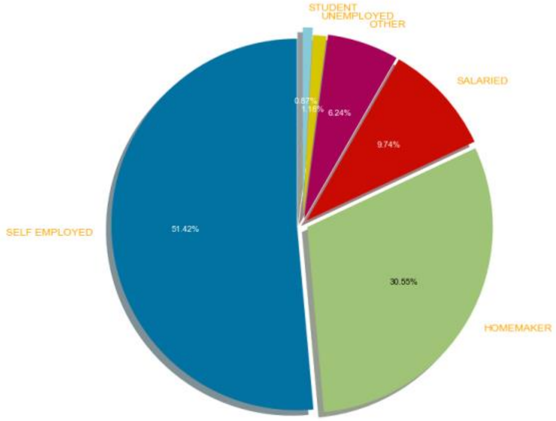


Figure 3

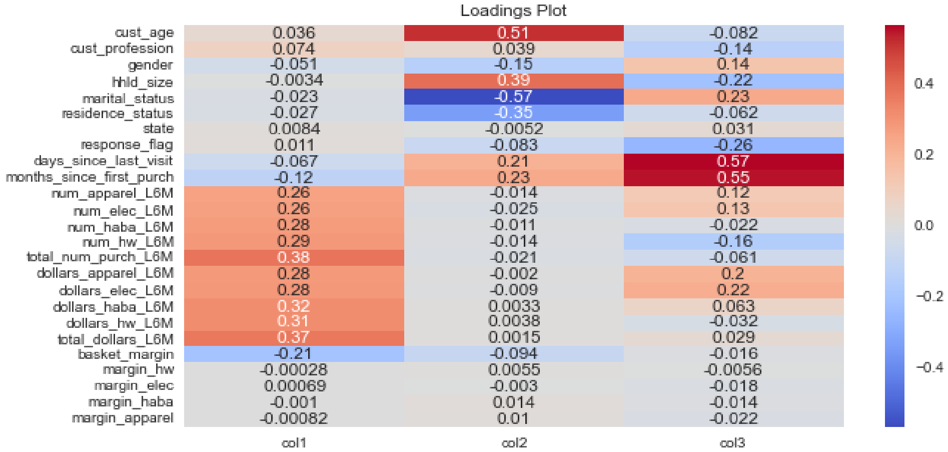


Figure 4 Heatmap of PCA features

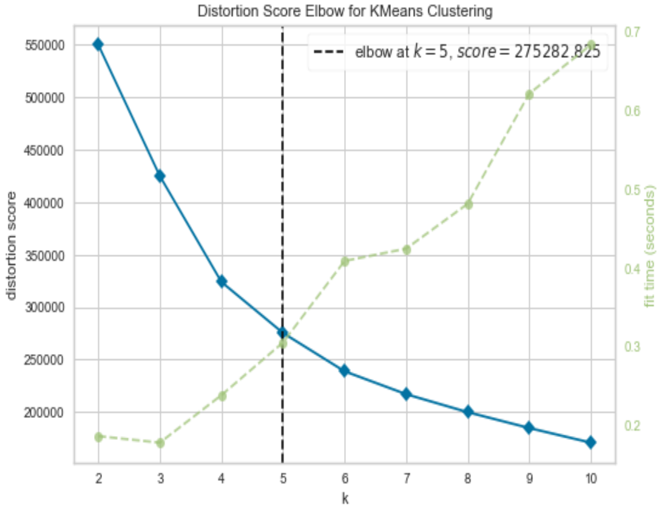


Figure 5 Elbow plot of Kmeans analysis

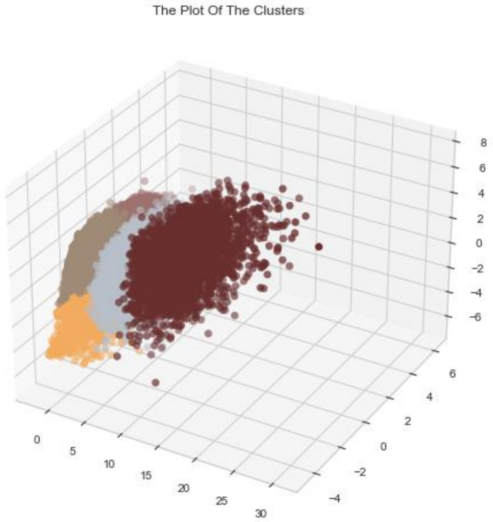


Figure 6 PCA Kmeans analysis

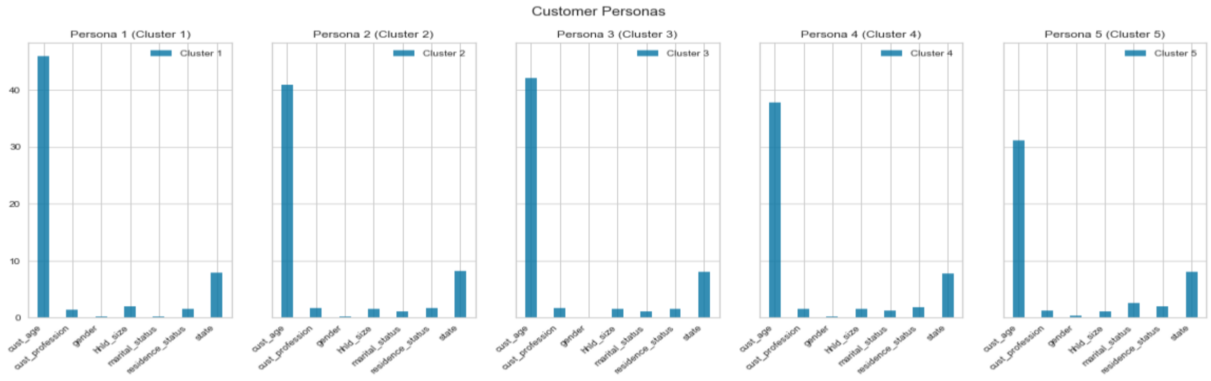


Figure 7 cluster characteristics

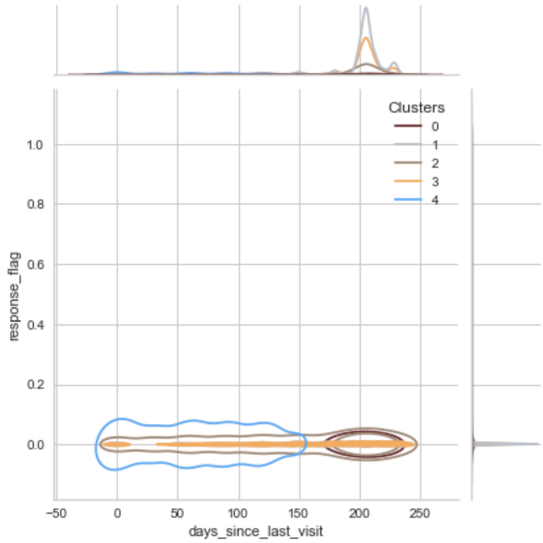


Figure 8 response flag vs. days since last visit

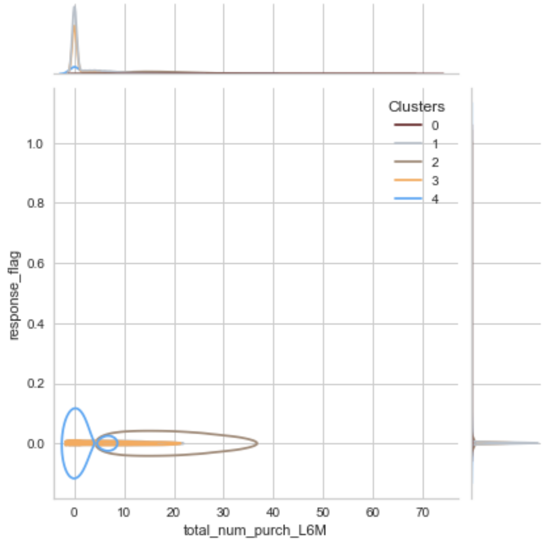


Figure 9 response flag vs. total number purchase in last 6m

Predictive Model Diagrams

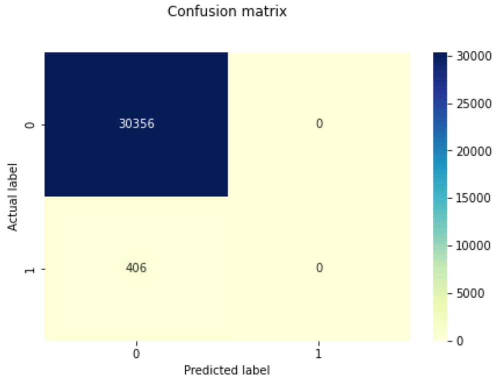


Figure 10. Logistic Regression Confusion Matrix

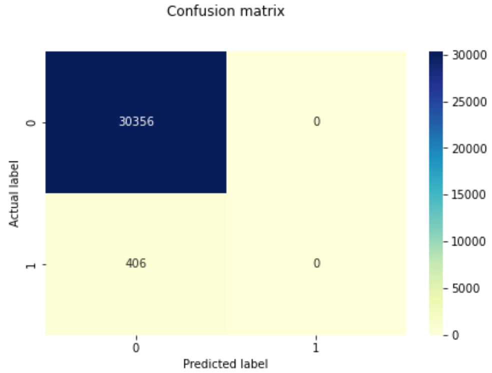
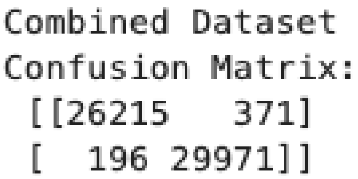
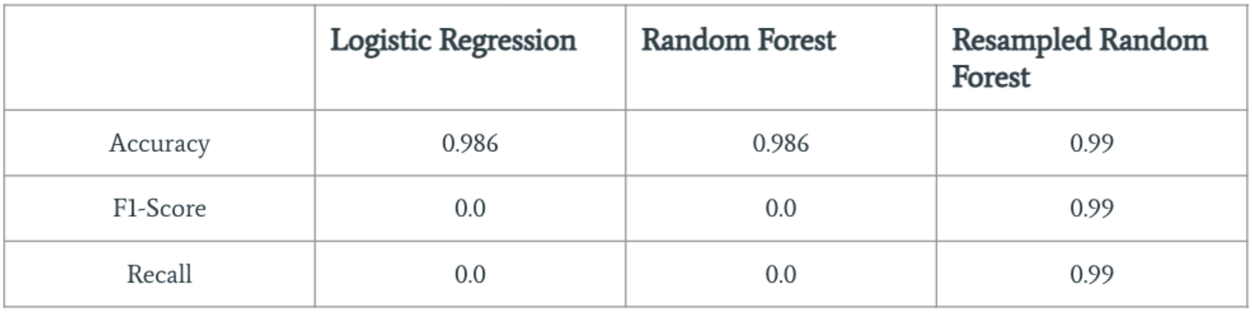


Figure 11 Random Forest Confusion Matrix



```
Combined Dataset
Confusion Matrix:
 [[26215   371]
 [  196 29971]]
```

Figure 12 Random Forest with Resampling Confusion Matrix



	Logistic Regression	Random Forest	Resampled Random Forest
Accuracy	0.986	0.986	0.99
F1-Score	0.0	0.0	0.99
Recall	0.0	0.0	0.99

Figure 13 Three predictive model performances