Data Mining for Customer Segmentation and Attrition Prediction to improve Marketing Efforts

Identifying Which Customers are Most Important to Keep Attrition Rates Low

Project Report

December 9, 2024

Introduction

In the world of banking and financial services, customer segmentation and attrition analysis are vital components of business strategy for companies. These two components go hand in hand because if you can accurately analyze one, it will give you insight on how to improve the other. Understanding customer behavior and trends allows for businesses to be proactive instead of reactive to retain high-value customers and fine tune their services.

There were two main goals of this project. The first was to group customers into six distinct segments based on their demographics and spending habits. This segmentation is meant to uncover patterns that the business can then use to improve marketing strategies. The second goal was to create a predictive model that accurately forecasts whether a customer is likely to churn based on their "customer profile."

Team Members

This project was completed individually by Dedric McCoy.

Data Set

For this project, I used a pre-made comma separated value data set named "BankChurners." In total, there are 10,127 rows and 20 columns used throughout the analysis of the data.

CLIENTNUM	Attrition_Flag	Customer_Age	Gender	Dependent_count	Education_Level	Marital_Status	Income_Category	Card_Category
768805383	Existing Customer	45	M	3	High School	Married	$60K - $80K	Blue
818770008	Existing Customer	49	F	5	Graduate	Single	Less than $40K	Blue
713982108	Existing Customer	51	M	3	Graduate	Married	$80K - $120K	Blue
769911858	Existing Customer	40	F	4	High School	Unknown	Less than $40K	Blue
709106358	Existing Customer	40	M	3	Uneducated	Married	$60K - $80K	Blue
713061558	Existing Customer	44	M	2	Graduate	Married	$40K - $60K	Blue
810347208	Existing Customer	51	M	4	Unknown	Married	$120K +	Gold
818906208	Existing Customer	32	M	0	High School	Unknown	$60K - $80K	Silver
710930508	Existing Customer	37	M	3	Uneducated	Single	$60K - $80K	Blue
719661558	Existing Customer	48	M	2	Graduate	Single	$80K - $120K	Blue

Figure 1: Table of the first 5 rows of the dataset before data processing.

The data was easily downloaded from Kaggle after setting up an account. As you can see, some of the columns in the raw data set were categorical, so I had to begin thinking about ways to convert them for analysis.

Methodology

This project was split into three phases: data preprocessing, exploratory data analysis (EDA), and model development and evaluation. Python was the programming language used and packages within Python include: PySpark, PySpark ML, SparkSQL, Pandas, Matplotlib, Scikit-learn, and Seaborn.

Data Preprocessing

As mentioned earlier, the raw data set included values in certain columns that were categorical instead of numerical. To fix this, I used feature encoding. Specifically, the *StringIndexer* module. I assigned each value in those specific columns to a specific number.

	Columns					
Index	Attrition Flag	Gender	Education Level	Marital Status	Income Category	Card Category
0	Existing Customer	Female	Graduate	Married	< $40K	Blue
1	Attrited Customer	Male	High School	Single	$40K - $60K	Silver
2	N/A	N/A	Unknown	Unknown	$80K - $120K	Gold
3	N/A	N/A	Uneducated	Divorced	$60K - $80K	Platinum
4	N/A	N/A	College	N/A	Unknown	N/A
5	N/A	N/A	Post-Grad	N/A	>= $120K	N/A
6	N/A	N/A	Doctorate	N/A	N/A	N/A

Table 1: Encoded columns with their associated values.

Another data preprocessing technique used was scaling. Some values in the data set were significantly larger than others which would have likely skewed the results. For instance, the values in the *Credit_Limit* column are higher in comparison to *Total_Relationship_Count*. This data discrepancy, unadjusted, could have led to the *Credit_Limit* column being more significant in the analysis than it was. The *StandardScaler* module was applied to standardize columns like this.

Exploratory Data Analysis (EDA) – Correlation, Clustering, and Visualizations

Taking all columns in the data set, I used correlation analysis to identify any potential relationships between columns. After conducting this analysis, I found that many of the correlations identified in the heat map were self-explanatory. For instance, the highest correlation of 0.81 was between *Total_Trans_Ct* and *Total_Trans_Amt*. However, one relationship that stuck out was *Credit_Limit* and *Card_Category* with a correlation of 0.49. The most issued card was "Blue." Beating the second most issued card by 8,881. This tells me that the other cards are being heavily under utilized and could potentially use a revamp to compete with the "Blue" card. The average credit limit of those with the "Blue" card could be used as a baseline of where the other cards should be to increase product usability.

After analyzing the correlation heat map, I was then able to select the columns for clustering analysis. Those columns were: *Total_Revolving_Bal, Total_Trans_Amt, Total_Trans_Ct, Customer_Age,* and *Credit_Limit*. I made sure to standardize the data as mentioned above to ensure proper scaling. Kmeans was then used to cluster the customers into 6 segments.

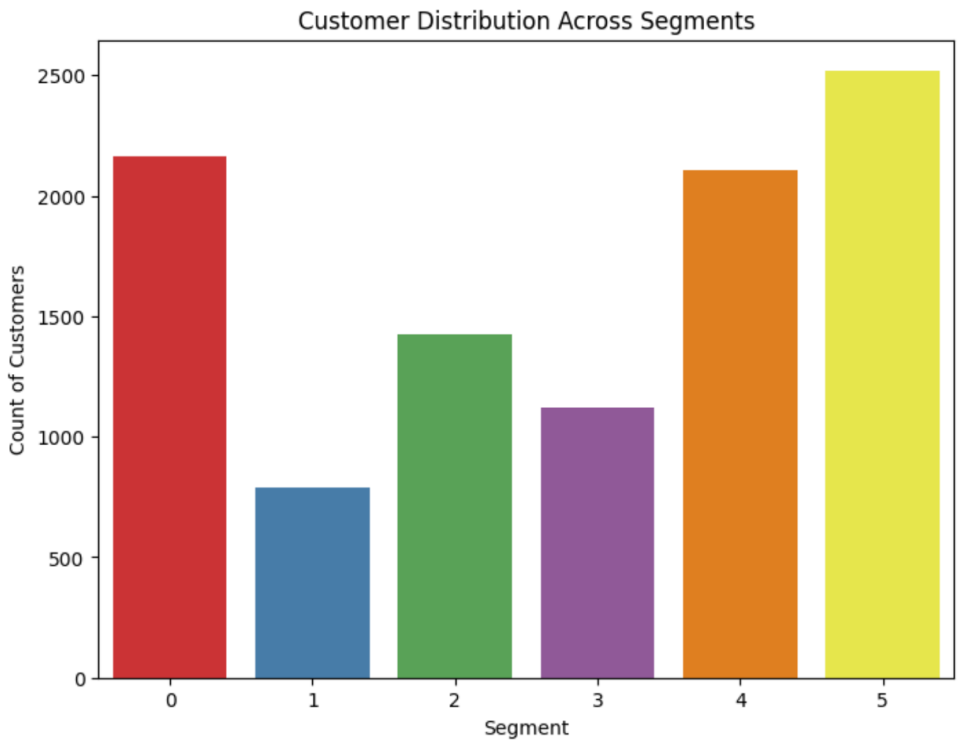

Figure 2: Bar chart showing the count of customers in each cluster.

Model Development and Evaluation

To predict customer attrition, I compared two models: a random forest model and logistic regression model. I included all columns except *CLIENTNUM,* as this is an identification column and has no relationship to any of the other columns. I split the data into training (80%) and test (20%) sets and began training the models. Once the models were trained, I assessed the predicted attrition value to the actual *Attrition_Flag* column. The linear regression model showed to be 91% accurate while the random forest model only showed to be 88% accurate. Since the logistic regression model was more accurate, I chose to utilize this model. I then went on to analyze which features (columns) were the most important in the linear regression model. The top three columns associated with *Attrition_Flag* were: *Total_Trans_Ct, Total_Ct_Chng_Q4_Q1, and Total_Revolving_Bal*. The bar chart showing the importance of all columns is below in Figure 3.

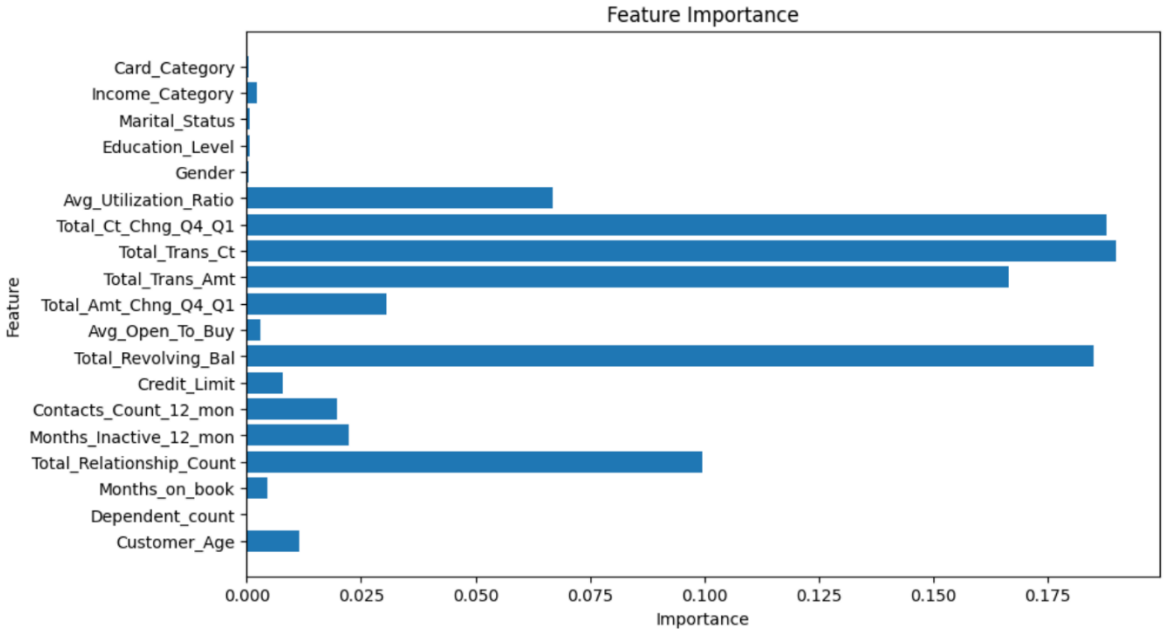

Figure 3: Importance of each column in the attrition predictive model.

Evaluation

To further show the accuracy of the predictive model, I used the receiver operating characteristic (ROC) curve to show the area under curve (AUC) which compares the true positive rate to the false positive rate. This value was 0.93 with the maximum being 1.0.

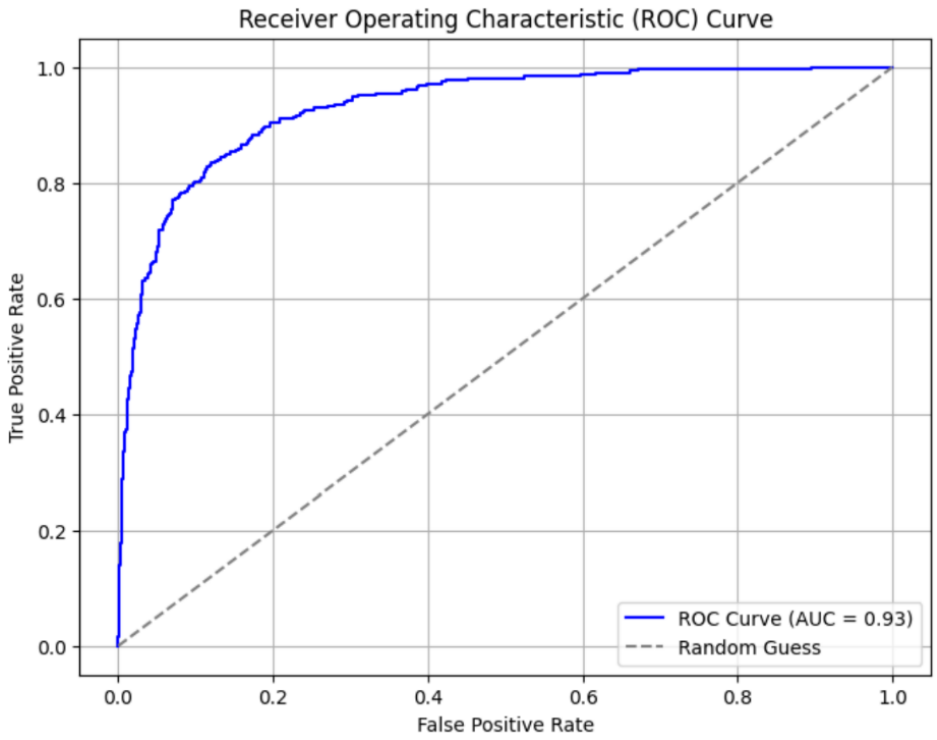

Figure 4: ROC-AUC Chart

To summarize each of the six cluster profiles, I viewed the descriptive statistics and created box plots to show the distribution of data by each featured column. From this analysis, I

created profile summaries, as well as assigning priority for marketing purposes. They are listed below:

Cluster 0: Low monthly spenders, many small transactions, middle aged customers, low credit limit, likely members rebuilding credit or middle-income customers (Priority #4)

Cluster 1: High card users, highest monthly spending and transaction counts, middle aged, high credit limit, likely customers with very good credit history and significant financial flexibility (Priority #1)

Cluster 2: Highest revolving balance, low transaction count, older aged customers, likely conservative spenders and limited engagement with credit cards (Priority #6)

Cluster 3: High credit limit, moderate use of cards, middle aged, likely middle-income with a balanced and responsible approach to managing credit cards (Priority #2)

Cluster 4: Younger customers, low credit limit, low transaction amount, likely customers starting to build credit with low income (Priority #5)

Cluster 5: Older aged, low credit limits, high transaction frequency, likely financially stable individuals with moderate credit card activity and engagement (Priority #3)

To assess the impact of feature selection, I assigned a baseline and alternative feature. My baseline consisted of the features used for the clustering analysis above and my alternative feature consisted of Credit_Limit, Avg_Open_To_Buy, and Total_Ct_Chng_Q4_Q1. I stated my null hypothesis as: the baseline feature having clustering cohesion greater than or equal to the alternative set. My alternative hypothesis was that the inclusion of the alternative features would improve the clustering cohesion overall. To test the hypothesis, I used a paired t-test. The results were: t-statistic of -70.60 and p-value of 2.4122. This means that the null hypothesis was rejected. Using alternative features improved the clustering of customers. This was expected though, because two of the features in the alternative baseline were in the top 3 of feature importance for my logistic regression model.

Milestones

Milestone #1

Week 3: Finalize the dataset selection and begin preparation for cleaning, standardizing, encoding, etc.

Milestone #2

Week 4-7: Begin installation and experimentation of necessary python packages. Complete segmentations analysis, visualizations, and predictive machine learning model.

Milestone #3

Week 8 - Finals: Evaluate model performance on the validation set, compare it to the pre-trained model, and perform final analysis on segmentations.

Conclusion

Using segmentation and predictive models can help give businesses a clear understanding of what type of business approach they should use when reaching their target audience. This project grouped customers into six segments and provided a reliable predictive model for customer attrition. When determining which marketing strategies to use, the business can then refer to these segments to understand what type of audience they should target to have the most positive effect on their business.

References

What is Customer Attrition & Customer Attrition Rate | Optimove. (2024, February 19). Optimove. https://www.optimove.com/resources/learning-center/customer-attrition

Customer Segmentation Analysis: Definition & Methods - Qualtrics. (2023, August 25). Qualtrics. https://www.qualtrics.com/experience-management/brand/customer-segmentation/

Alam, M. (2024, September 26). *What is Customer Segmentation? Definition, Models, Analysis, Strategy and Examples*. IdeaScale. https://ideascale.com/blog/what-is-customer-segmentation/

Staff, P. (2024, November 22). *What is Customer Attrition and How Do You Stop It*. Podium. https://www.podium.com/article/customer-attrition/