Capstone Final Report

Credit Card Customer Prediction

Prepared by: Daniel Shaw

Introduction

How does your solution detect customers at risk of churning, and what proactive measures can be implemented to retain them?

A manager at the bank is increasingly concerned about the rising number of customers leaving their credit card services, resulting in a loss of business and revenue. In order to address this challenge, the bank wants to predict which customers are at high risk of churning, enabling proactive interventions to retain them. The goal is to develop a predictive model that can accurately identify customers likely to cancel their credit card services in the near future. By doing so, the bank can reach out to these customers with personalized offers, better services, or targeted communications to persuade them to stay and prevent further attrition.

Problem Statement

How should a Credit Card Bank company lower churn rate by the end of the year in order to (a) development a 80% Accurate early warning system predictive model in determining which customers will churn and (b) Provide the bank's management and customer service teams with insights and actionable strategies for effective intervention.

Data Wrangling

Data wrangling, also known as data cleaning or preprocessing, is the process of transforming and preparing raw data into a structured and usable format for analysis. The raw dataset used was from Kaggle and contained

10127 rows and 20 columns. It offered a sufficiently large sample size, which is ideal for developing predictive models. Having a bigger dataset usually means more accurate models and less overfitting, which gives you better insights. It's big enough to spot trends, patterns, and relationships in the data, which are key for understanding customer behavior and churn.

When working with this dataset, I performed several key data wrangling tasks. First, I searched for missing values to ensure completeness, addressing any gaps in the data. I also reviewed and corrected the data types of features to ensure they were appropriately assigned (e.g., converting dates to datetime format or numerical values to integers/floats). Additionally, I examined the dataset for unique values and duplicates, removing any unnecessary redundancies to maintain data integrity. These steps helped to clean and structure the dataset for accurate analysis and reliable insights.

Exploratory Data Analysis

Exploratory Data Analysis (EDA) is a crucial step in understanding the dataset before diving into more complex analysis or modeling. During EDA, I used various techniques are used to explore the underlying structure, patterns, and relationships in the data. I also used data visualizations, such as histograms, pie chart, and Heatmaps to identify trends, outliers, and potential correlations between variables.

During the Exploratory Data Analysis (EDA), I investigated several key relationships between features to uncover valuable insights:

Customer Churn by Age: I analyzed how customer churn varies across different age groups, helping to identify if certain age demographics are more prone to attrition. The average age of customers who are still active with the service is approximately 46.26 years, while the average age of customers who have left the service is approximately 46.66 years.

Distribution of Customer Age: I explored the overall distribution of customer ages to understand the age profile of the customer base shown in figure 1.

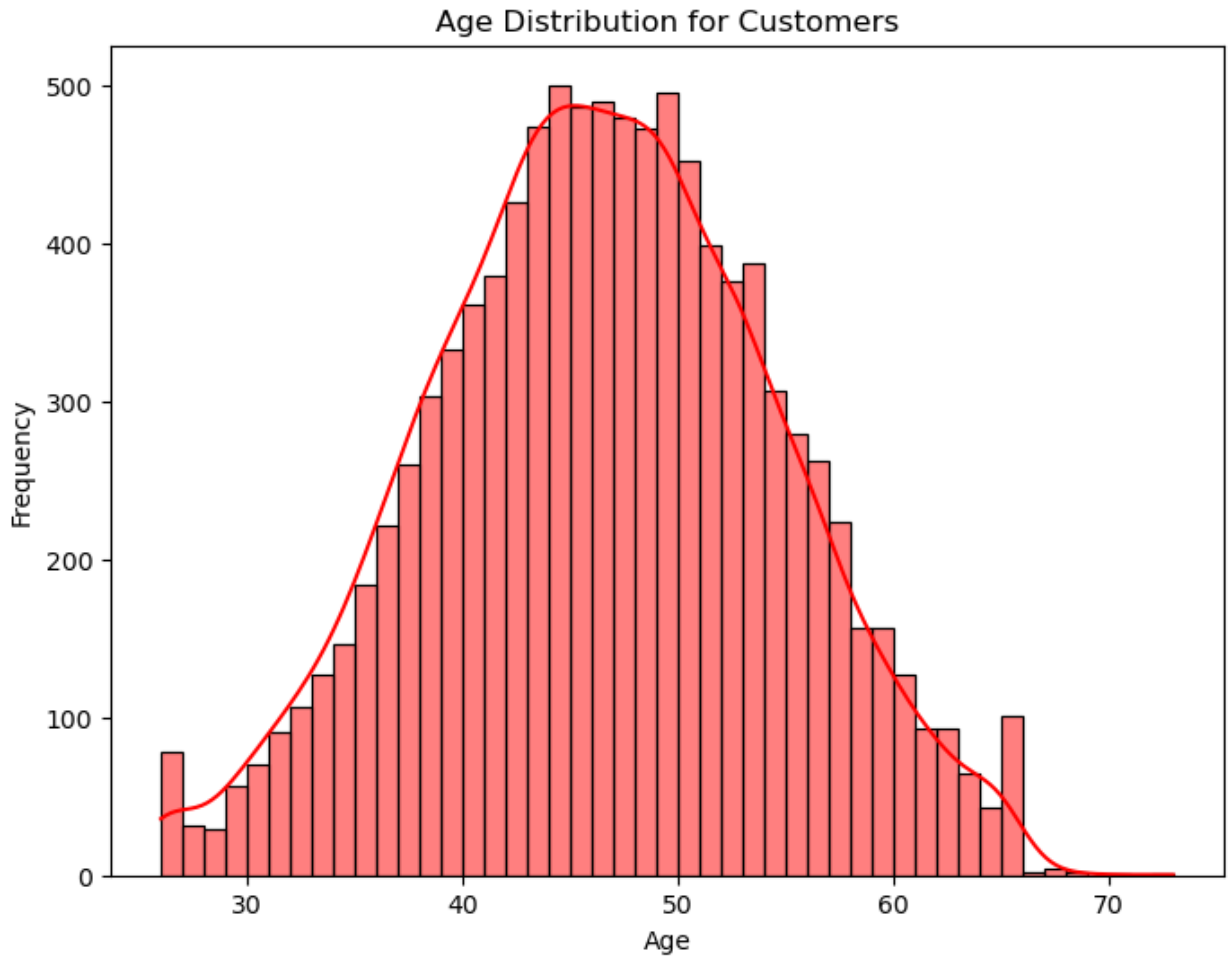


Figure 1.

The shape of the distribution resembles a bell, with most of the data points concentrated around the mean and fewer data points as you move further away from the mean.

Customer Churn by Gender: I examined the churn rate across male and female customers to determine if gender plays a role in customer retention.

Out[13]:

	Gender	Attrition_Flag	Frequency
0	F	Attrited Customer	930
1	F	Existing Customer	4428
2	M	Attrited Customer	697
3	M	Existing Customer	4072

Figure 2.

Gender and Attrition Flag: I further analyzed the relationship between gender and the attrition flag to explore if there are any significant patterns or differences in churn rates by gender.

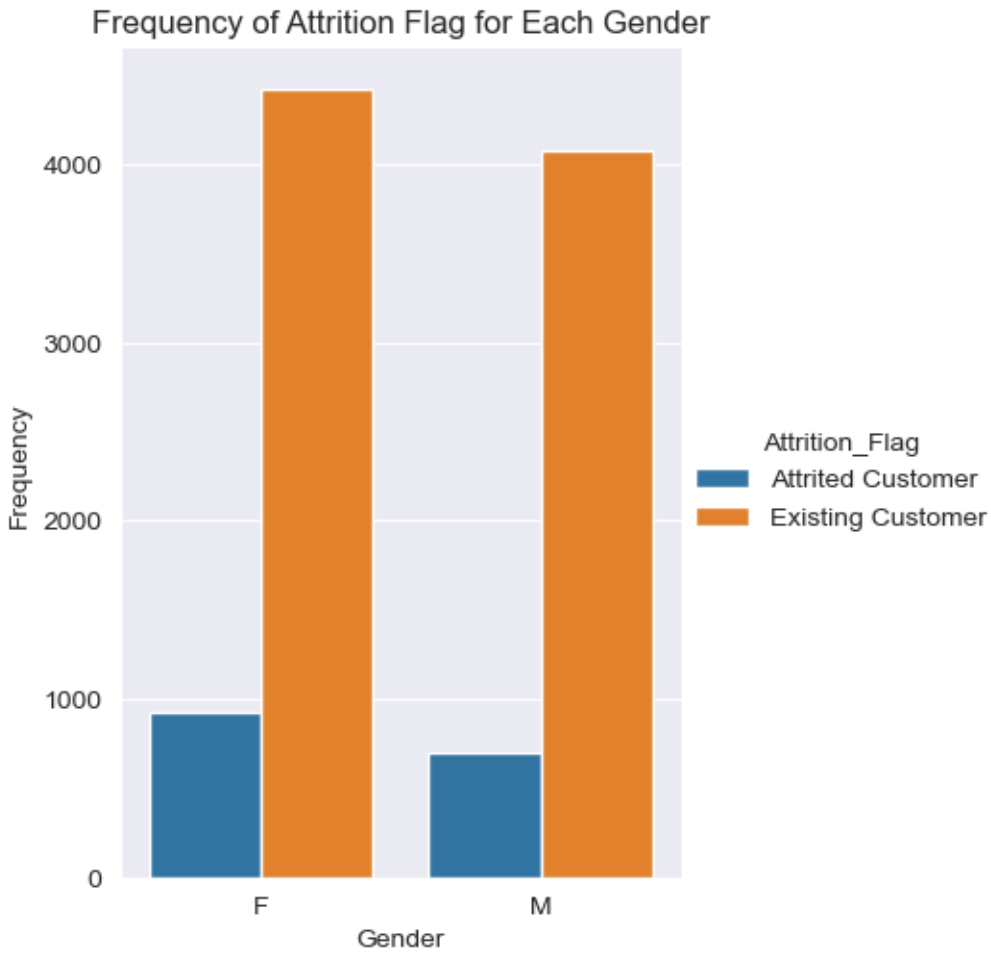


Figure 3.

In Figure 3. More female customers are leaving the service, but the total number of female customers is still higher. This could suggest that while females may have a higher tendency to churn there is still a larger initial pool of female customers to begin with, so their total numbers are still higher.

Marital Status and Dependent Count: I explored how marital status and the number of dependents impact customer behavior and potential churn.

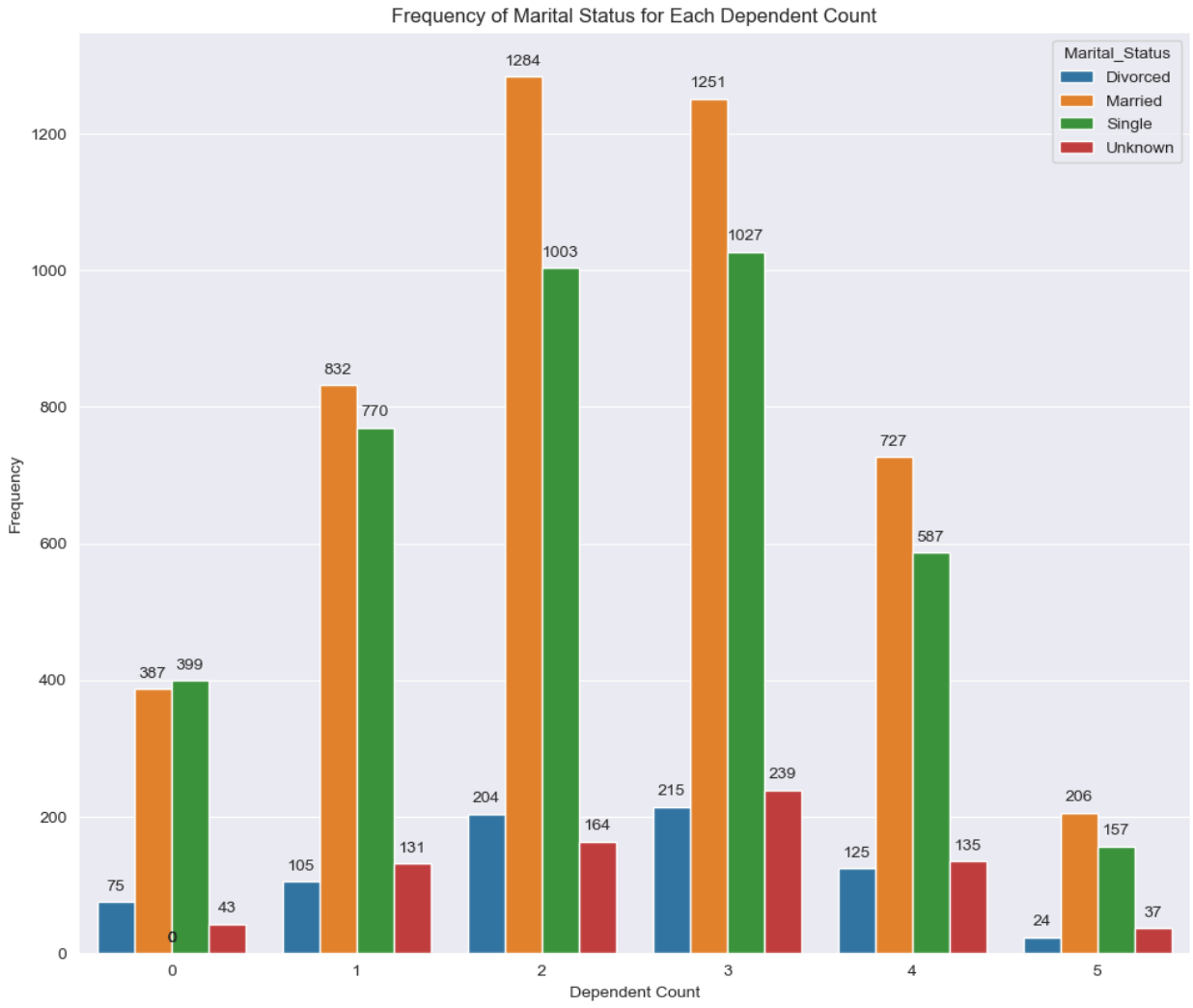


Figure 5.

In the distribution shown in Figure 5. married individuals have the highest frequency for each dependent count.

Education Level and Attrition Flag: I looked at the relationship between customers' education levels and their likelihood of churning, which could indicate the role of education in retention.

Out[17]:

	Education_Level	Attrition_Flag	Frequency
0	College	Attrited Customer	154
1	College	Existing Customer	859
2	Doctorate	Attrited Customer	95
3	Doctorate	Existing Customer	356
4	Graduate	Attrited Customer	487
5	Graduate	Existing Customer	2641
6	High School	Attrited Customer	306
7	High School	Existing Customer	1707
8	Post-Graduate	Attrited Customer	92
9	Post-Graduate	Existing Customer	424
10	Uneducated	Attrited Customer	237
11	Uneducated	Existing Customer	1250
12	Unknown	Attrited Customer	256
13	Unknown	Existing Customer	1263

Figure 6.

From the data shown in Figure 6. More Graduates customers are leaving the service, but the total number of Graduates customers is still higher. This could suggest that while Graduates may have a higher tendency to churn, there is still a larger initial pool of Graduates customers to begin with, so their total numbers are still higher.

Education Level and Income Category: I investigated how education level correlates with income category to identify if higher educational attainment correlates with higher income brackets.

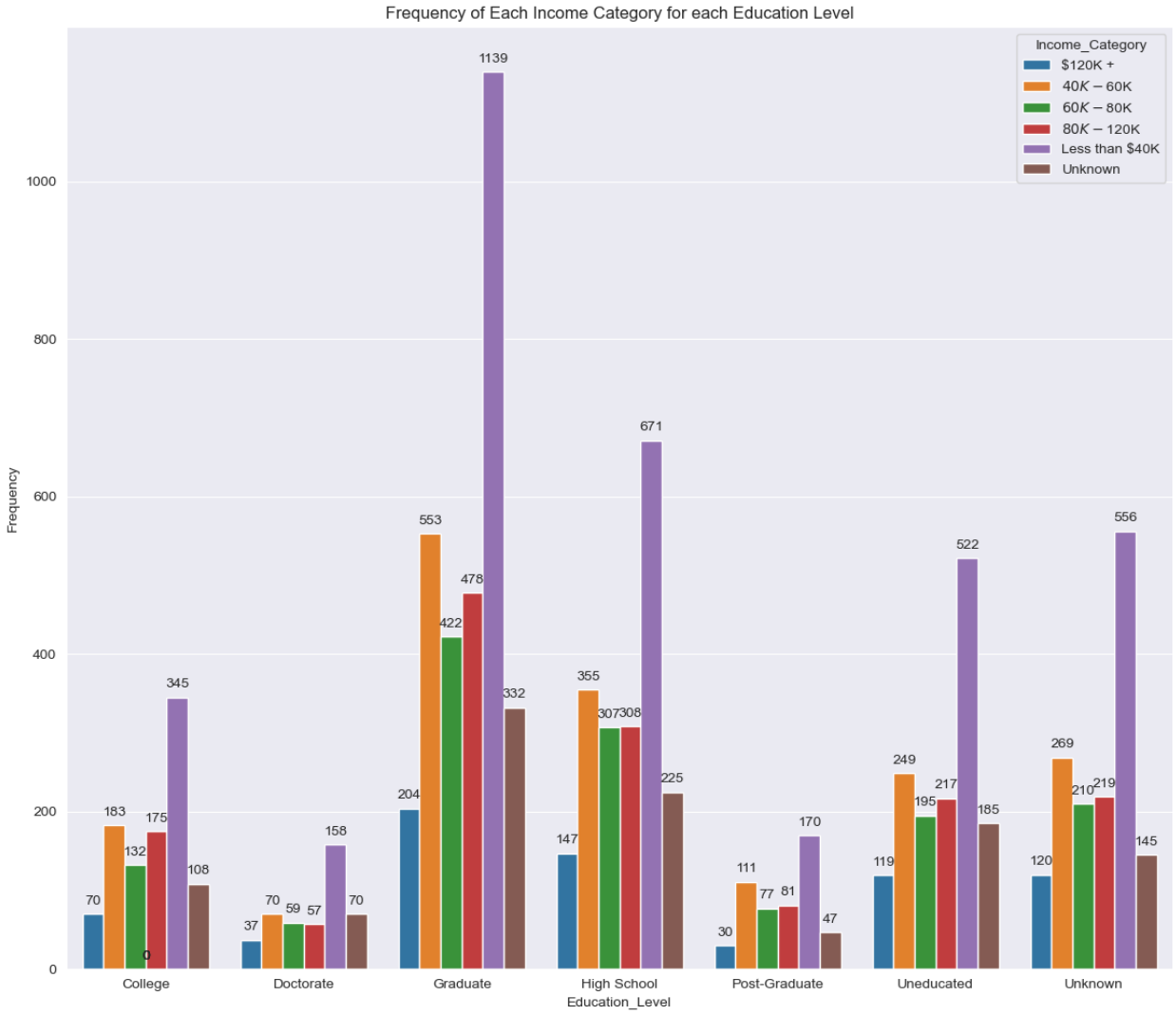


Figure 7.

In Figure 7. Shown above customers making less than 40k make up the majority of customers in each education level, it may suggest that income may not be strongly correlated with education in our dataset

Income Category and Attrition Flag: I analyzed how different income categories influence the likelihood of customer churn, to see if income level plays a significant role in retention.

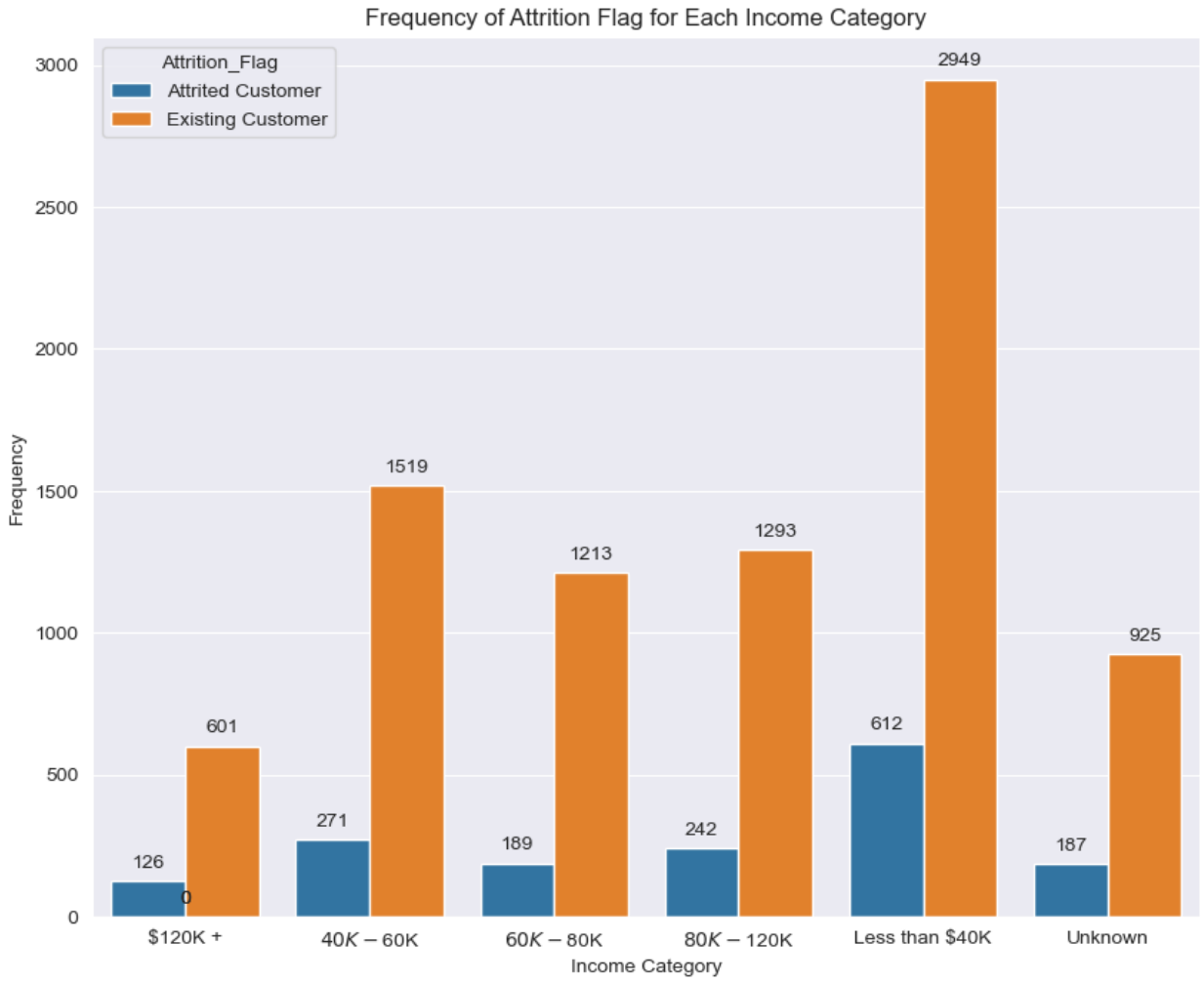


Figure 8.

In Figure 7. Shown above more customers earning less than 40k are leaving the service, but the total number of customers earning less than 40k is still higher. This could suggest that while customers earning less than 40k may have a higher tendency to churn, there is still a larger initial pool of customers earning less than 40k customers to begin with, so their total numbers are still higher.

Distribution of Card Category: I analyzed the percentage of how many customers hold each type of credit card offered by the bank.

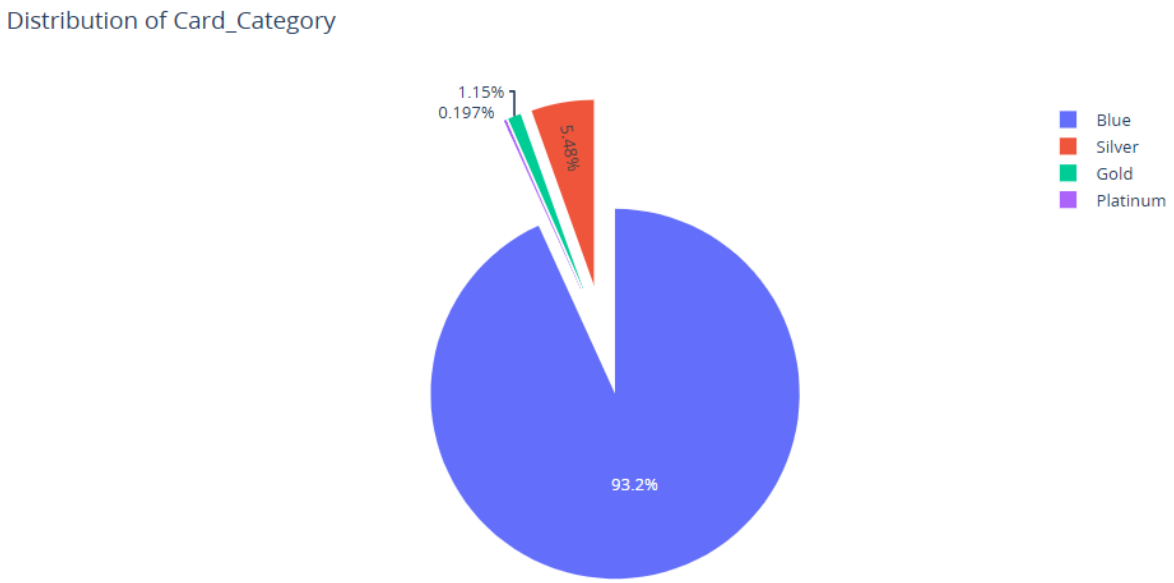
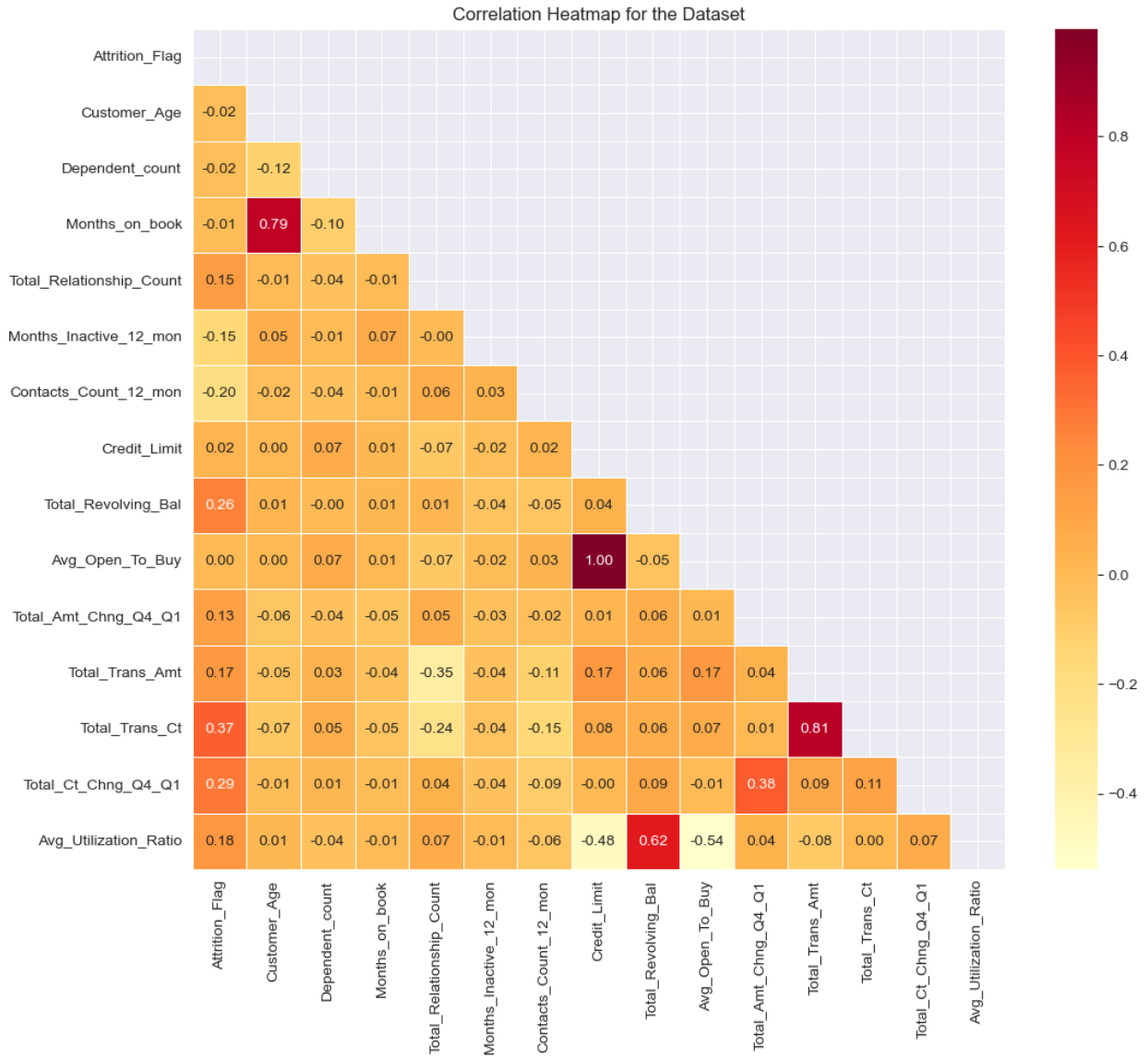


Figure 9.

In Figure 9. 93.2% of customer have the basic “blue” credit card. With the high number of customers earning less than $40k may explain why the majority of cardholders have chosen the Blue card, which falls into the lowest category.

For the numerical columns a correlation heatmap is a powerful tool for quickly visualizing the relationships between numerical variables in your dataset. It can help identify patterns in customer behavior, such as spending habits, credit utilization, and balances shown in Figure 10. below.



Box-and-Whisker plots were also used to analyze and get a graphical representation of the distribution of a dataset. I was able to see the central tendency, spread, and outliers in the data. I saw some outliner that can skew statistical analyses and machine learning models, leading to inaccurate predictions or misleading results. One example shown in the below Figure 11. Credit Limit

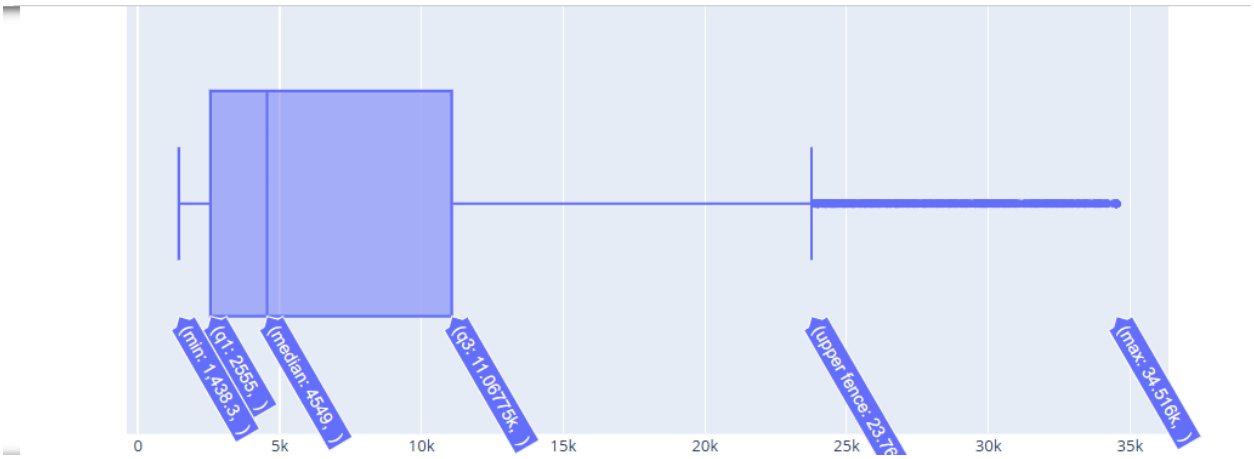


Figure 11. Credit Limit

Correcting these outlines helps the model learn patterns more effectively.

Pre-processing

The steps taken to clean and prepare raw data for analysis or model training are crucial because raw data often contains inconsistencies. Preprocessing transforms the data into a format that is suitable for analysis and model training.

Before training a machine learning model, it's essential to split the data into training and test sets. This ensures that the model is tested on unseen data. The initial step I took was to apply one-hot encoding, which generates dummy variables by converting categorical features into binary columns (0 or 1). This is crucial when working with categorical variables in machine learning, as it transforms them into a format suitable for models. One-hot encoding creates a new column for each unique category in the original feature, with each row receiving a value of 1 in the relevant column if it belongs to that category, and 0 if it does not.

The next step was to standardize and scale our data. Standardizing your data ensures fair treatment of all features, speeds up training for certain

algorithms, and prevents any one feature from dominating the model due to its scale.

Model Selection

When selecting a model for predicting churn, it's important to consider various factors such as performance, interpretability, computational efficiency, and the nature of the data you're working with. For this project I used three popular models: Logistic Regression, Support Vector Machines (SVM), and Naive Bayes.

Logistic regression provides easily interpretable coefficients that can help you understand the relationship between the features and the target variable.

```
                                    Accuracy  Precision    Recall  \
LogisticRegression(solver='liblinear')  0.89297   0.914491  0.961666

                                    F1 Score   ROC AUC  \
LogisticRegression(solver='liblinear')  0.937486  0.906856

                                    Cross-Validation Accuracy
LogisticRegression(solver='liblinear')                 0.908756
```

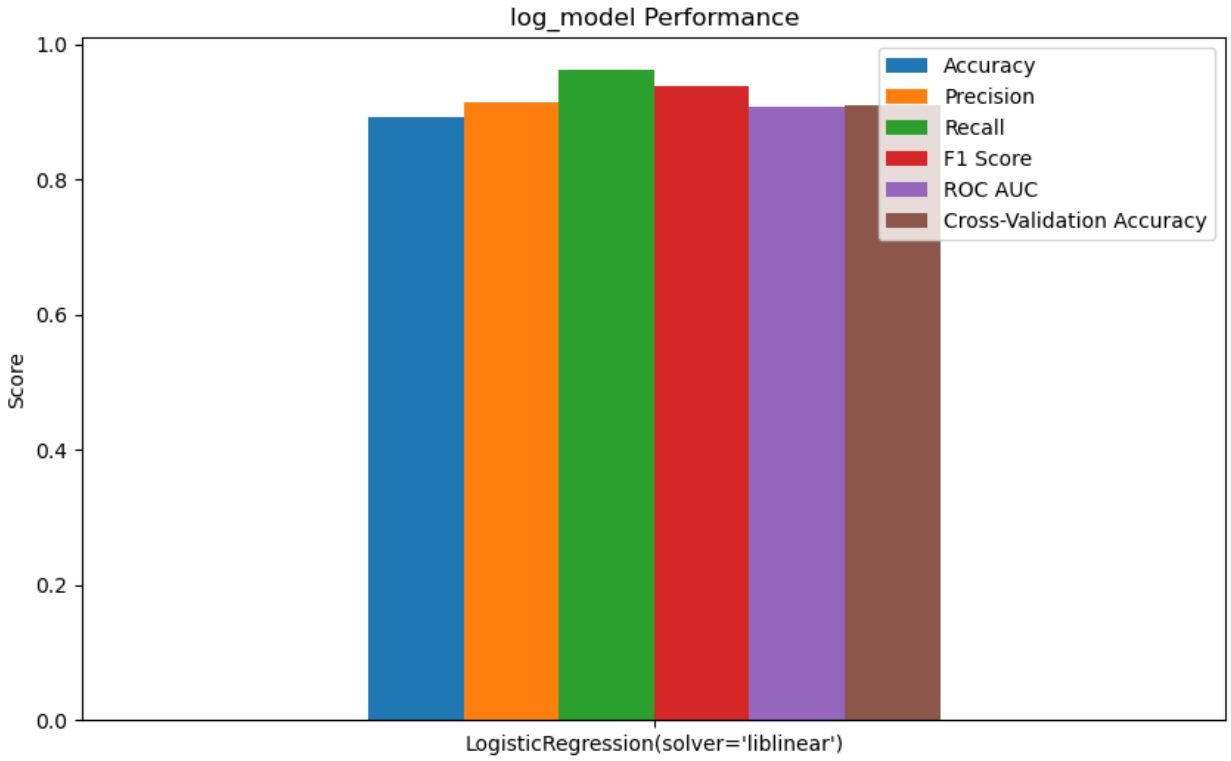


Figure 12.

The Logistic Regression model appears to perform well across several metrics. It shows high precision, recall, and F1 score, indicating that it is effective at both identifying positive instances and minimizing errors. The high ROC AUC and cross-validation accuracy suggest that the model generalizes well and is stable across different datasets.

Another model used is Support Vector Classification (SVC), which classifies data into distinct categories by identifying the optimal separating hyperplane.

	Accuracy	Precision	Recall	F1 Score	ROC AUC	\
SVC(probability=True)	0.907583	0.918113	0.976337	0.94633	0.939136	

	Cross-Validation Accuracy
SVC(probability=True)	0.913496

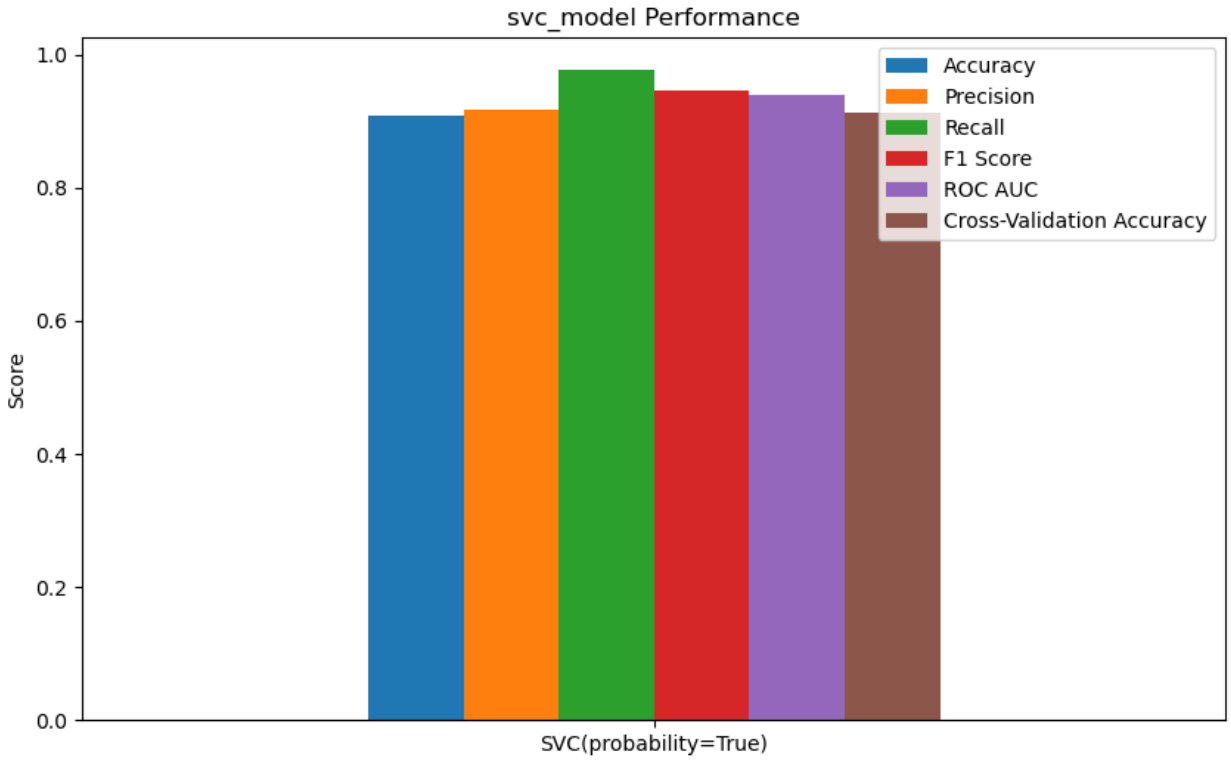


Figure 13.

The SVC with probability estimation performs exceptionally well across all metrics, achieving the highest recall, F1 score, and cross-validation accuracy. It outperforms Logistic Regression in all aspects, especially in recall, indicating that it is better at identifying positive cases while maintaining a good balance between precision and recall. This model would be a strong candidate if maximizing recall and the balance between precision and recall is a priority.

The final model created for this project is Gaussian Naive Bayes, which is mainly used for classification tasks. It can be adapted to handle multi-class problems and is based on probabilistic principles.

	Accuracy	Precision	Recall	F1 Score	ROC AUC	\
GaussianNB()	0.859795	0.908837	0.924752	0.916725	0.850081	

	Cross-Validation Accuracy
GaussianNB()	0.876103

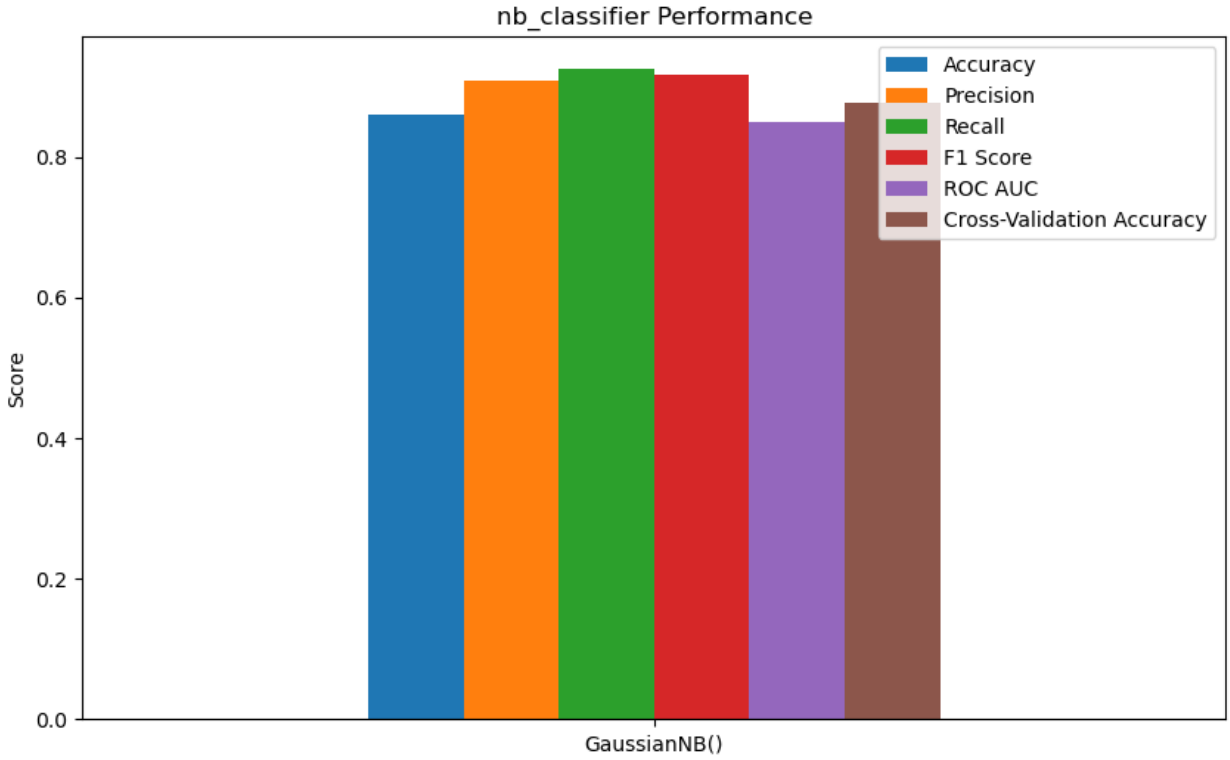


Figure 14.

The Gaussian Naive Bayes (GaussianNB) model performs reasonably well, especially in terms of precision and recall, but it falls short in comparison to SVC and Logistic Regression in terms of accuracy, ROC AUC, and cross-validation accuracy. The model is still effective, particularly in situations where a probabilistic approach with independent features is appropriate, but for this specific dataset, the SVC with probability estimation and Logistic

Regression models are likely the better choices in terms of overall performance.

Conclusion

The SVC with probability estimation is the top-performing model overall. It really excels in recall, with an impressive score of 97.6%, meaning it's great at spotting positive cases and minimizing false negatives. The model also does well with an F1 score of 94.6%, showing a solid balance between precision and recall. Plus, with a cross-validation accuracy of 91.4%, it shows it works well across different data splits.

Logistic Regression is a strong, reliable model that performs well across key metrics, making it a good option. Meanwhile, Gaussian Naive Bayes doesn't perform as well here, suggesting it might not be the best fit for this dataset, though it could still be useful in other cases where assumptions about feature independence are different.

In the end, the SVC model is the best choice when you want to maximize recall and get a well-rounded performance across all the metrics.

Takeaways

Given that the majority of churned customers are female graduates earning less than $40k per year, I recommend three strategies to reduce churn. First, offer more budget-friendly financial products, such as low-fee accounts or savings programs, specifically designed for customers earning less than $40k. Second, provide premium financial products or services tailored to graduates, such as exclusive investment or savings plans, which may be more appealing to this educated demographic. Lastly, introduce tiered loyalty programs that reward customers for upgrading to higher-tier credit cards. By offering incremental benefits like better rewards and lower fees, customers will be encouraged to move up the tiers, making higher-tier cards more attractive.

In closing, we successfully developed over an 80% accurate early warning system predictive model to identify customers at risk of churning. This model equips the bank's management and customer service teams with valuable insights and actionable strategies to intervene effectively and reduce churn.

Future Research

Future research based on this Banks Churn dataset could focus on Feature Engineering and Selection. One potential avenue for future work would be to focus on refining the feature engineering process. Incorporating additional customer behavior data, such as transaction history, customer support interactions, or browsing patterns, could enhance model performance. Research into automated feature selection techniques could also further improve the predictive power of the model.