Predicting and Reducing Bank Customer Churn

Alex Hamedaninia

MS Data Science, Bellevue University

DSC 680 Applied Data Science

Professor Iranitalab

April 30, 2024

Business Problem

Can we reduce customer churn by predicting which customers are at risk of churning, and take action to reduce future customers churning?

This project seeks to answer this question by conducting predictive analysis on customer bank churn data, obtained from Kaggle. Through the development of robust predictive models, the objective is to pinpoint customers most susceptible to churning and devise effective strategies to mitigate this risk, thereby enhancing customer retention and reducing overall company cost.

Background

Customer churn, or customer turnover or customer attrition, refers to the act of customers dissolving their relationship to a company by no longer using the companies' products or services. When speaking about banks, this specifically refers to clients closing their accounts, ceasing to use banking services, or transferring their business to another financial institution. Company costs increase significantly when a customer churns, as not only is obtaining new customers expensive, but they also lose the business of their lost customer, with the customer potentially starting business with a competing company.

However, there is a way to mitigate this loss. This is through customer churn prediction. By identifying customers at risk of leaving, banks can implement targeted retention strategies to minimize revenue losses and acquisition costs. Churn prediction also enables banks to proactively address underlying issues such as customer dissatisfaction or competitive pressures, thereby enhancing customer satisfaction and loyalty. This proactive approach not only preserves customer revenue, but also strengthens a company's relationship with its current customers, ultimately improving the bank's competitive position in the market. Predicting churn helps banks

optimize resource allocation by spending less to maintain current customers rather than spend more to obtain new clients, and it improves the overall operational efficiency of the company. Overall, churn prediction plays a vital role in sustaining the current client base, and succeeding as a bank in an increasingly competitive and dynamic financial landscape.

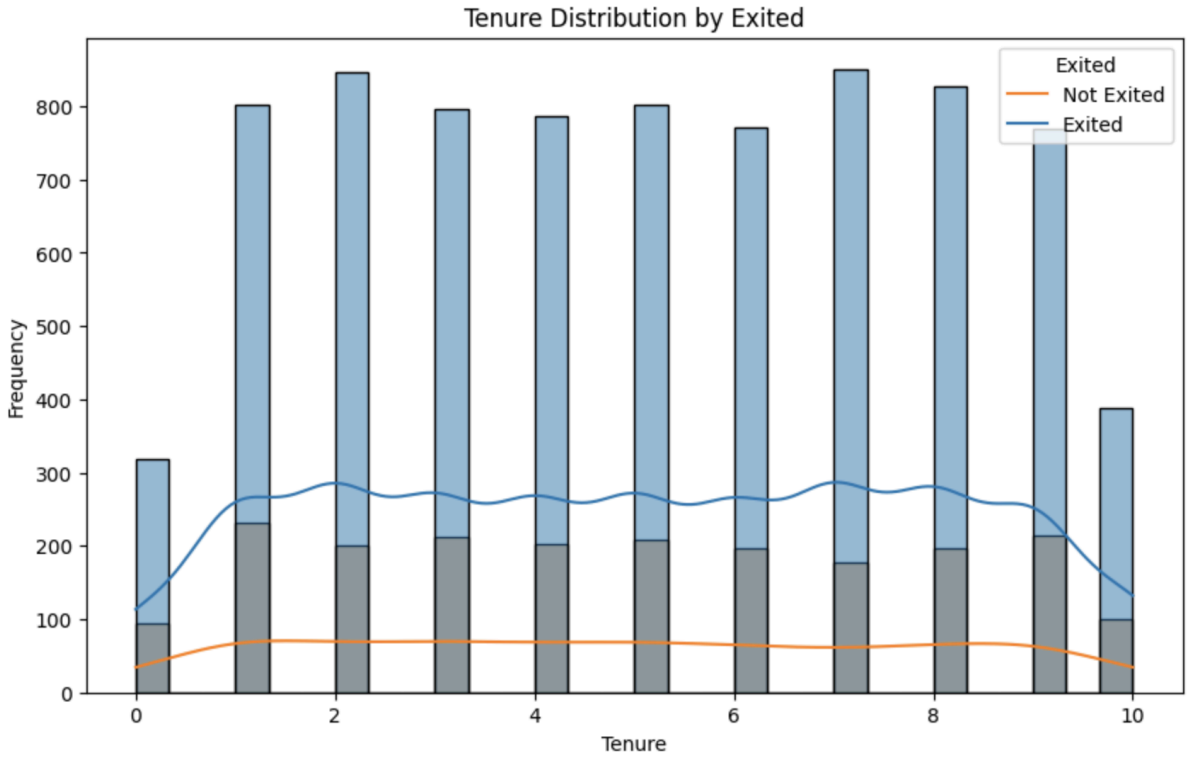
Data Explanation

The dataset utilized in this project was obtained from Kaggle titled “Bank Customer Churn”. This dataset contains 18 variables, which includes geography, gender, age, tenure, credit score, balance, complaints, satisfaction score after complaint has been resolved, a binary indicator if the customer has left the company or not, and other variables that may prove to be useful to this analysis.

To initiate the data cleaning process, several steps are necessary to ensure the data is suitable for modeling. This includes removing any unnecessary variables from the DataFrame, converting categorical variables to dummy variables, and checking for any missing values. First, any unnecessary variables should be removed from the DataFrame. This includes the variable *RowNumber*, as this is redundant to the indexing of the pandas DataFrame. In addition to this, the variables *CustomerID* and *Surname* are also removed to protect customer privacy. Since many modeling algorithms require numerical input, any categorical variables present must be converted to dummy variables. This transformation involves converting the two categorical features, *Location* and *Card Type* to dummy variables. This creates a new binary indicator variable for each category present under the respective variable, indicating with a 1 if that location or card type applies to the observation, and 0 otherwise. This process ensures that categorical information is appropriately represented for analysis. Lastly, a check for any missing

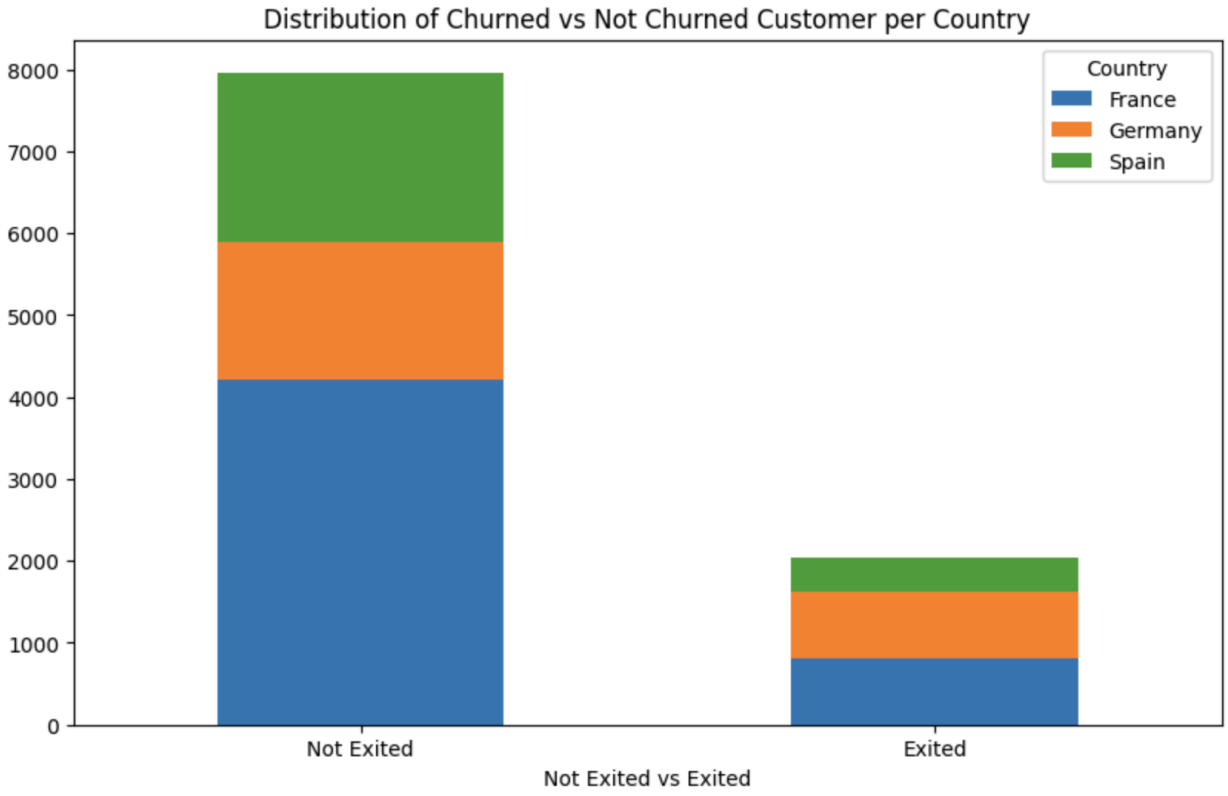
values is conducted, fortunately revealing an absence of such instances for this dataset, subsequently streamlining the analysis process. The dataset is now primed for robust modeling and analysis, beginning with exploratory data analysis.

The dataset can be explored through univariate and bivariate graphical analysis to gain a better understanding of the data. The different variables will be thoroughly examined uniquely and comparatively to one another. This will give a deeper understanding to how the different variables may react to one another, and in future modeling, which features have a bigger weight in affecting the modeling and prediction process. First is a stacked bar plot displaying the distribution of the tenure of each customer, and displaying if they had exited or not.



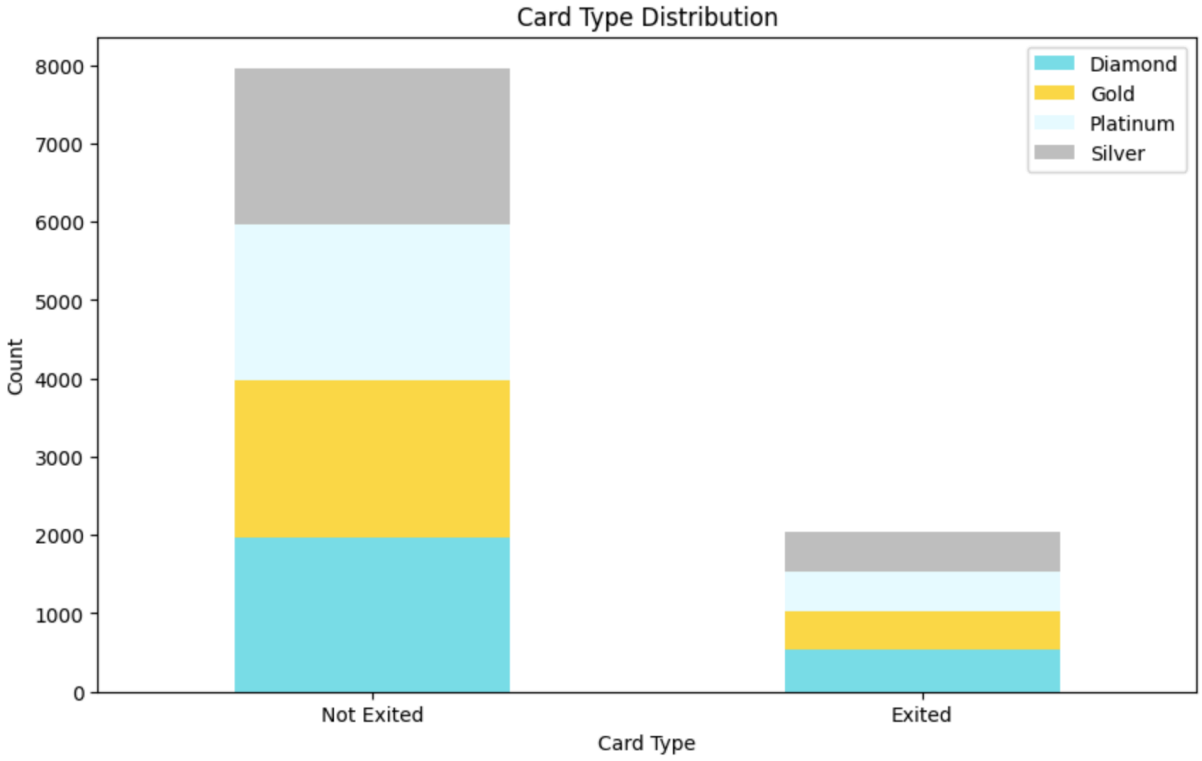
The first notable aspect of this chart is that the distribution of tenure for customers who churned is quite evenly distributed from new customers to customers who had been with the company for about 10 years. There is a small dip for brand new customers and those who have

been with the company for 10 years, but it is quite even throughout. This suggests that the length of time a customer has been with a company does not significantly impact the likelihood of them churning alone. Next graph is a distribution of customers per country, separated whether they customers churned or not.



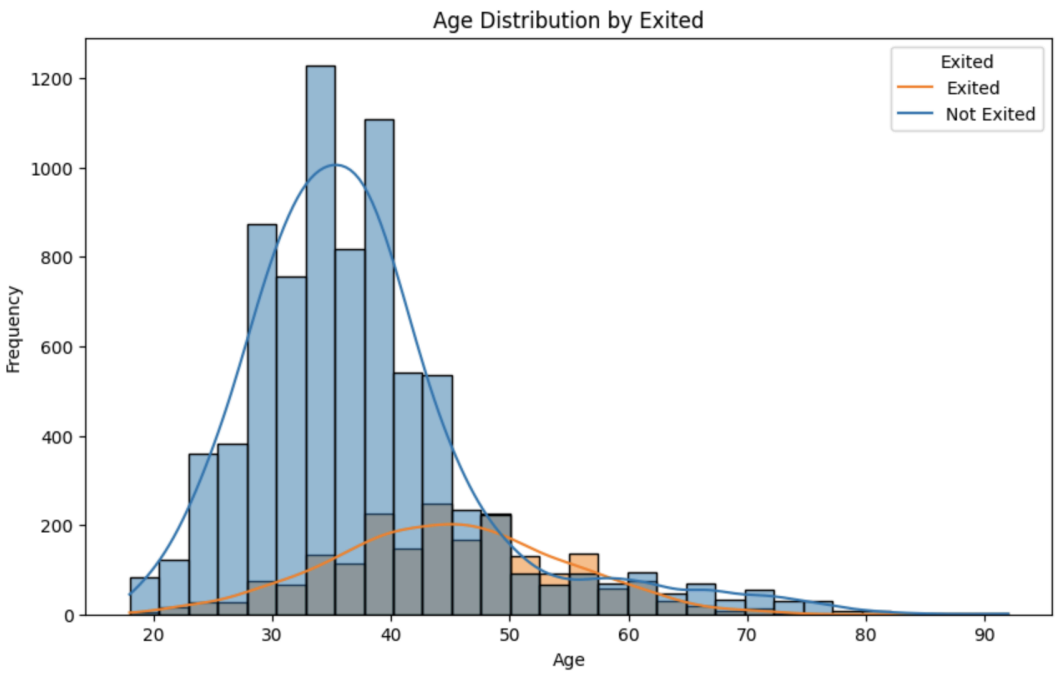
Most of the customers are from France, and the amount of customers churning are proportional to the amount not churning for each country, with the exception of Germany. This country appears to be half the size of its original amount. Upon analyzing the churn proportions across different countries, it is evident that the ratio of non-churned to churned customers varies. Specifically, in Germany, this ratio stands at approximately 48%, while in France and Spain, it hovers around 20% each. This indicates that customers from Spain are more likely to churn, which will be important for the analysis later on.

The next graph to look at is a distribution of card types for both churned and customers who are still with the company..

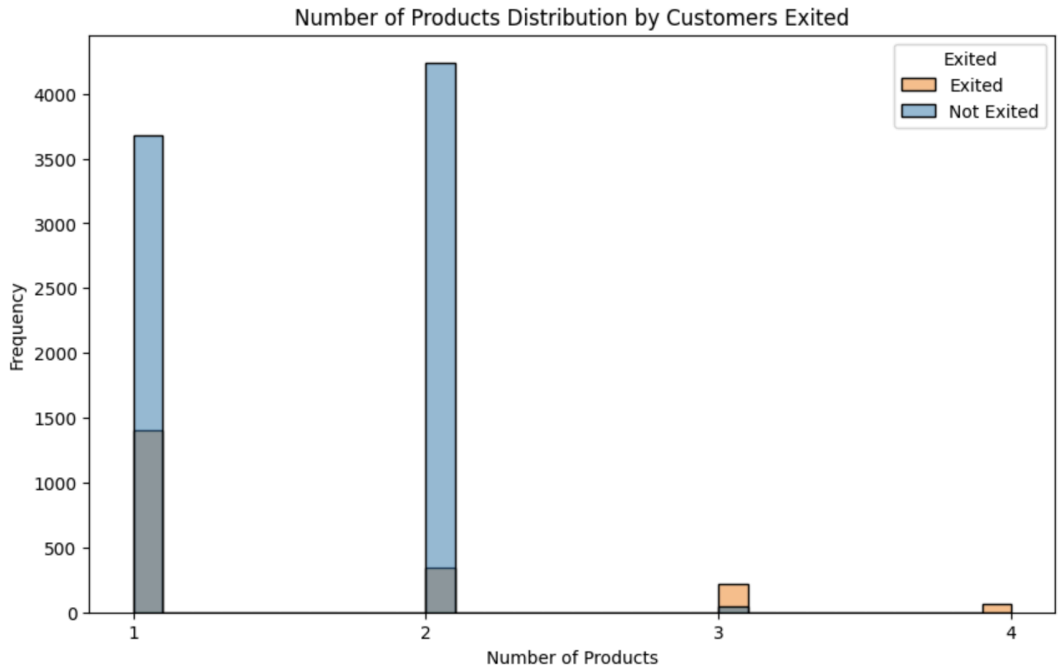


It is interesting to note that each one seems perfectly proportional from non-churned to churned. It suffices to say that the card type does not play a significant factor into customers deciding to churn or not.

The next graph here is an age distribution, a stacked bar chart distinguishing those who exited the company or not. In this case, it is apparent to note that those on the younger side seem less likely to leave, but between the ages of 45 to 60, these customers are at a higher risk of churning. This may be due to nearing retirement age and thus a change in finances. Whatever the case may be, this will be important to note for future analyses.



In the last graph here, we have the distribution of the number of products, again separated by those exited or not. It seems 2 is the optimal number of products for a customer to have. If they have more, they are incredibly likely to churn, and if they only have 1, it seems to be half of the customers churn.



Methods

Now the data is ready for modeling. As this is a binary classification problem, a logistic regression modeling technique will be utilized for its efficiency and accuracy. To begin the process, the data is split from its features and the target variable, *Exited.* Following the initial split, the data is further split into training and test sets with a 70/30 ratio to facilitate model training and evaluation. The features are standardized using StandardScaler to enhance the model's performance. The logistic regression model is initialized using scikit-learn's LogisticRegression class. With the model trained on the training data, its performance is evaluated on the testing data, computing metrics such as accuracy, precision, recall, and F1-score. The results classification report and confusion matrix provided comprehensive insight into the model's predictive capabilities.

Analysis

The model yielded an impressive accuracy of 99.87%. The model demonstrates exceptional performance in predicting customer churn. The classification report further enhances this finding, revealing high precision, recall, and F1-score values for both churned and non-churned customers. Specifically, the model achieved a precision of 99% and recall of 100% for churned customers, and 100% for both precision and recall for non-churned customers. This indicates the model's ability to effectively identify both positive and negative instances of churn. The confusion matrix corroborates these findings, with minimal misclassifications observed. This remarkable accuracy and precision scores underscores the robustness and reliability of this logistic regression model in predicting customer churn.

With the model's success in its high accuracy in predicting customer churn, the next step is reducing the chance of customer churn. Through the exploratory data analysis, there were a few aspects uncovered that may lead customers to churn. First, it was found that customers from Germany had a more than double chance of churning, as opposed to the customers from France or Spain. This may be due to economic uncertainty in Germany, influencing higher churn rates, or the rapid digitalization and adoption of online banking in Germany. More data, such as customer's reasons for complaining and leaving, are needed to gain an understanding of how to mitigate this issue.

```
Classification Report:
              precision    recall  f1-score   support

           0       1.00      1.00      1.00      2416
           1       0.99      1.00      1.00       584

    accuracy                           1.00      3000
   macro avg       1.00      1.00      1.00      3000
weighted avg       1.00      1.00      1.00      3000


Confusion Matrix:
[[2413    3]
 [   1  583]]
```

Another case of higher churn uncovered was for those between 50 and 60 years of age, who seem more prone to churn. This may be due to upcoming retirement plans causing a shift in needs for these customers. This can result in customers seeking other banking companies who can provide a service more suitable. Again in this case, more data would be needed, such as a customer's reason for deciding to leave the company.

Conclusion

Predicting which customers are at risk of churning can be a major benefit to a company, as it can reduce future costs of acquiring new customers to replace the loss. This project embarked on a comprehensive exploration of customer churn prediction within the banking sector. Leveraging machine learning techniques, particularly logistic regression, we were able to develop a robust model capable of accurately predicting customer churn with high precision and

recall. Through meticulous data preprocessing and data exploration, we optimized the model's performance to an impressive accuracy of 99.87%.

Subsequent exploratory data analysis uncovered key insights into factors influencing customer churn, such as their location. We found that the customers based in Germany have a nearly 50% chance of churning, as compared to France's and Spain's 20% chance of churning. This finding underscores the significance of regional dynamics and the need for specified retention strategies to address the underlying drivers of churn effectively. We also found that the age group of 45-60 are more likely to churn, and we speculated this may be due to upcoming retirement financial plans not aligning with the bank's current options, and we propose investigating the complaints of these customers to prepare a financial plan more suited to their needs. Additionally, it was shown that optimally, customers should have about 1-2 products, and having 3 or more leads to a higher likelihood of churning. Overall, the goal of this project was successful in accurately predicting customer churn, which will henceforth enhance customer retention efforts, fortifying customer relationships, and ultimately driving sustained business growth in an increasingly competitive landscape.

Assumptions

While undertaking this project, some assumptions have to be made about the data and the results. In regards to the data, there is an assumption of data quality and reliability of the dataset used for analysis. Any biases or incoherences within the data will impact the findings and the validity of those findings. It is assumed the data is complete, accurate, and representative of the population. Additionally, when it comes to modeling utilizing linear regression, the model assumes a linear relationship between the variables, and that little to no multicollinearity exists

between any of the variables. By acknowledging and carefully considering these assumptions, stakeholders can better interpret the results of the analysis and make informed decisions based on the results of the analysis.

Limitations

While this project and its results can be utilized to great ends, there are limitations to be aware of. While the data was able to give an understanding of which customers are at risk of churning, and give an insight as to possibilities as to why, more data is required to gain a more in depth understanding as to why these customers are churning. The data showed that 99% of the customers who complained ended up exiting the company, so any data that contains these complaints can give insight as to why these customers are leaving.

While the data displayed some correlations between the data, it cannot establish causality. The observed relationships may be influenced by unmeasured or confounding variables not accounted for in the analysis. Additionally, a logistic regression modeling algorithm, while suitable for a binary classification problem like this one, has limitations in capturing complex nonlinear relationships between the predictor and target variables. More sophisticated machine learning algorithms may be required to capture these nonlinear relationships. By acknowledging these limitations, stakeholders can review the results of this project more cautiously and consider potential biases when making business decisions using the results of the analysis.

Challenges

There were a few challenges presented when going through this project. While logistic regression provides interpretable coefficients, the model's interpretability may be limited for

more complex relationships or interactions between variables. Communicating the findings requires effective visualization and communication strategies. Additionally, translating the insights gained into actionable strategies for reducing customer churn presented challenges. Without an understanding of why these customers left the company, the only insights that are able to be gained were from exploring the data and identifying trends from those who did choose to leave the company.

Future Uses

There are many future uses and additional applications for this project. Having insight into which customers are at risk of churning allows banks some time to coordinate a plan in an attempt to retain these customers. Knowing which customers are at risk of churning, and comparing their data to insights observed from this analysis, can allow companies a chance to retain these customers, reducing overall company costs and increasing relationships and trust with current clients.

While the insights gleaned from this project gave insight as to which of these customers are at risk of churning, a next step for this project is gaining insight as to why these customers left. When a customer decides to part with the banking institution, it is important to request for them to fill out a survey asking why they chose to leave the company. With data containing reasons why a customer decided to take their business elsewhere, this project can further utilize that data to understand why a customer churns, and take action to reduce the chances of it happening again.

Recommendations

There are a few recommendations for this project to further advance the findings gained from it. While knowing which customers are at risk of churning, it is also important to understand why customers churn. To gain a better understanding of why customers choose to churn from the company, investing in data collections methods and processes to not only improve the quality and reliability of the dataset, but asking customers as to why they choose to leave the company will allow for this project to take one step further and determine actions for the company to take to actively reduce customer churn.

A few additional recommendations is exploring additional features or transformations on the data that may capture meaningful relationships with customer churn. Consider incorporating additional external data sources such as economic indicators, customer feedback, or market trends to enrich the dataset and enhance predictive performance. Additionally, experiment with alternative modeling approaches beyond logistic regression, such as random forests, gradient boosting, or neural networks, to identify the most effective algorithm for predicting customer churn. Evaluate model performance beyond what was chosen here, such as R^2 score or ROC-AUC score for evaluation. One last step to take before implementing this into the company is ensuring the model performs just as well with new input of data, ensuring the data was not overfitted to the model. This will ensure that the model is robust and ready to be implemented into the company.

Implementation Plan

It is important to establish an implementation plan after the project is complete. This can include, but is not limited to, communicating the findings of the project to stakeholders so they are able to make informed business decisions for when and how they would like to implement

this model into the day to day functioning of the company. This also includes after this model has been tested thoroughly to ensure robustness and reliability. It is important to define the objective as identifying which customers are at risk of churning. Then the model is ready to be deployed into production environments, such as online banking portals, customer service platforms, or marketing automation systems, to enable real-time predictions and decision-making.

Once the model has been deployed, it would be sufficient to implement automated alerts and notifications triggered by the churn prediction model to proactively identify customers at risk of churning. This would notify relevant stakeholders, such as account managers or customer service representatives, to take timely action. It is important to establish mechanisms for continuous monitoring and optimization of the churn prediction model and retention strategies. This includes tracking model performance metrics, customer feedback, and business outcomes to identify areas for improvement and refinement. Once the model is ready for use, it is important to provide training and education to relevant staff members on the use of the churn prediction model and retention strategies. Ensure that these employees understand how to interpret model predictions and take appropriate actions to retain customers effectively. Taking each of these steps will allow for seamless deployment of the model in the production environment and allow for staff to effectively interact with customers at risk of churn.

Ethical Assessment

It is important to maintain a code of ethics when working with customer data, especially when dealing with personal information relating to the customers. It is vitally important to adhere to any data privacy and protections relevant to the country and state the model is deployed in. This can be accomplished by implementing measure such as data encryption,

anonymization, and access controls to safeguard information. Additionally, it is important to obtain informed consent from customers by updating the User Terms and Conditions agreement, having customers allow the company to utilize their information in such a way. When dealing with customer data, it is also important to ensure there are no unintentional biases in the results that may disproportionately impact certain demographic groups. By prioritizing ethical considerations, the organization can maintain and build trust with their clients to maintain their data security.

10 Questions

1. How was customer churn defined and measured in this project?

Customer churn was defined by the binary indicator variable *Exited,* which had a 1 if the customer churned, and 0 if they were still with the company.

2. What preprocessing techniques were applied to the dataset to ensure data quality and prepare it for modeling?

I removed the redundant variable *RowNumber*, as well as any personal information variables such as *CustomerID* and *Surname*. I then converted the categorical variables *Location* and *Card_Type* to dummy variables.

3. What machine learning algorithms were considered for predicting customer churn, and why was logistic regression ultimately chosen as the primary modeling approach?

Random forest modeling was also considered for its feature selection, but ultimately logistic regression was chosen after checking its near perfect accuracy score.

4. How was class imbalance addressed in the dataset, and what techniques were employed to ensure reliable predictions for both churned and non-churned customers?

The distribution of the data was checked graphically to ensure a mostly normal distribution, ensuring the data was balanced and to ensure reliable predictions for both churned and non-churned customers.

5. What features were identified as significant predictors of customer churn, and how were they selected or engineered from the original dataset?

The features identified as significant predictors of churn were chosen through the exploratory data analysis, and were identified through the graphs made.

6. How was the performance of the churn prediction model evaluated, and what metrics were used to assess its effectiveness in predicting churn accurately?

Accuracy score, precision, F1-score, recall, and a confusion matrix were utilized to assess the effectiveness in predicting churn accurately.

7. What insights were gained from the exploratory data analysis, particularly regarding regional differences in churn rates and potential drivers of churn among customers from different countries?

We were able to determine that customers from Spain and France had a churn rate of approximately 20%, but customers from Germany have a churn rate of 48%. More information is required to understand why this is.

8. What ethical considerations were taken into account throughout the project, and how were privacy, fairness, transparency, and accountability upheld in the collection and use of customer data?

Customer data protection was taken into account with the removal of customer ID and surnames in the model creation. Additionally, understanding the limitations and assumptions as clearly depicted in the write up are essential for stakeholders to understand.

9. How were the findings and recommendations from the churn prediction project communicated to stakeholders, and what actions were taken based on the model's predictions to reduce customer churn?

The findings and recommendations are communicated to the stakeholders through this PowerPoint presentation at a later date, and the actions are yet to be determined.

10. What are the limitations and future directions of the churn prediction project, and what steps will be taken to address them and further improve the effectiveness of customer retention strategies in the banking sector?

In this analysis we were only able to understand so much as to why these customers are churning. More information regarding what the customers complained about and why they decided to leave the company is required, and conducting sentiment analysis using natural language processing to understand where the company can improve to lessen the risk of these customers churning.

Sources

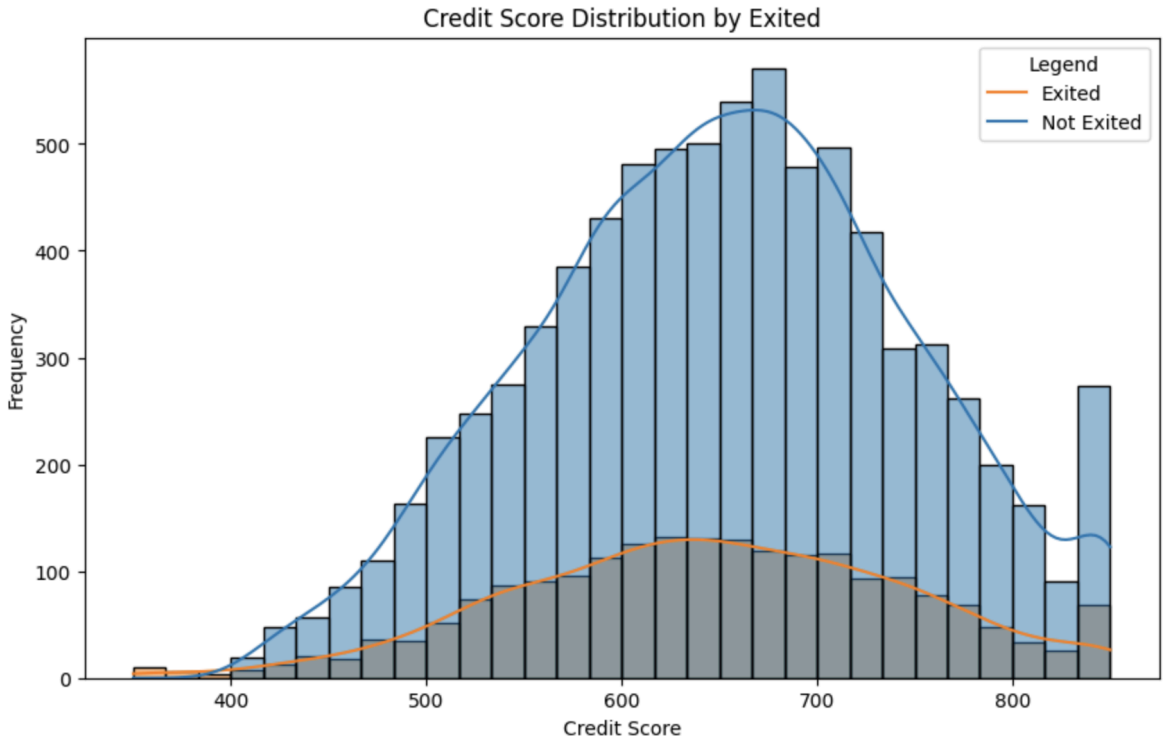
Radheshyam Kollipara. (2022, April). Bank Customer Churn, Version 1. Retrieved April 11, 2024 from https://www.kaggle.com/datasets/radheshyamkollipara/bank-customer-churn/data

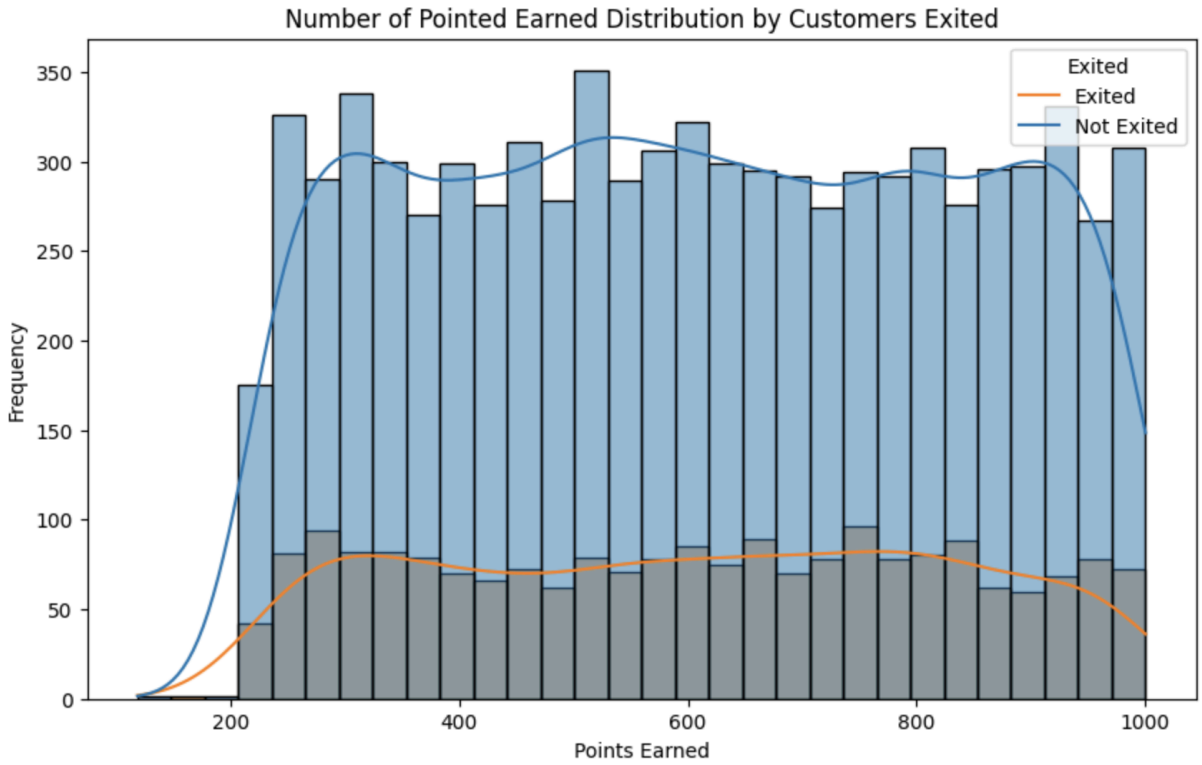
Appendix

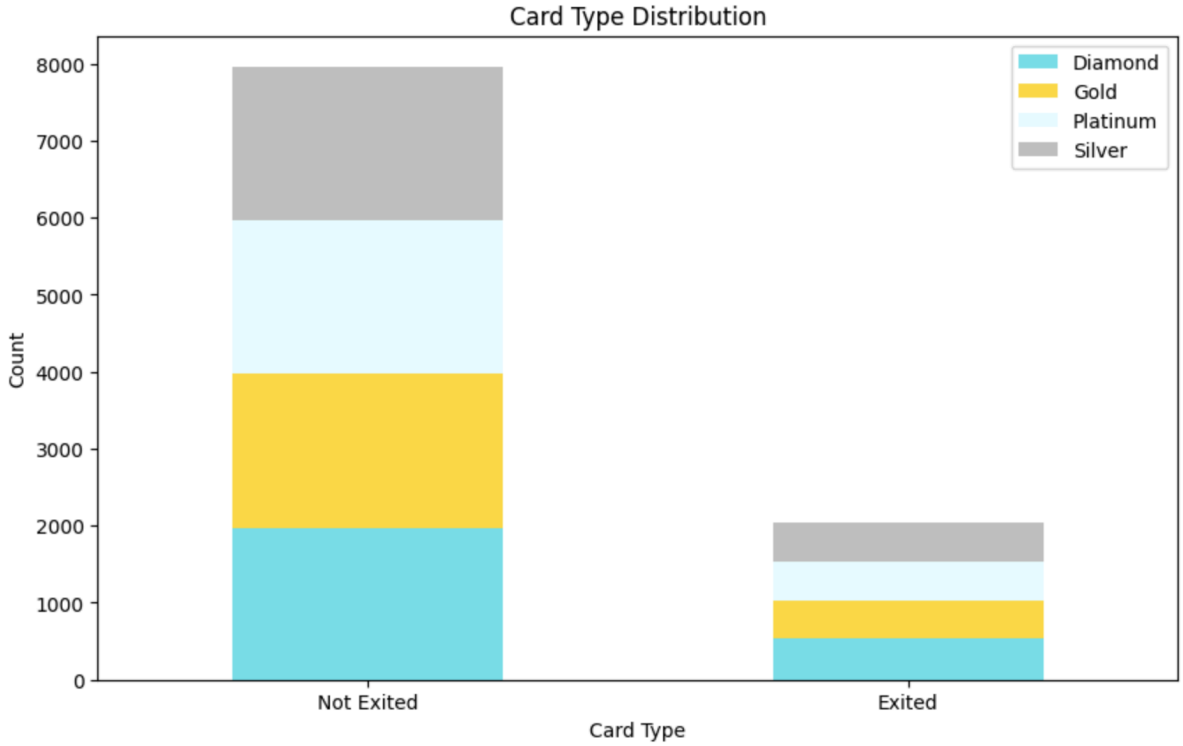
Exited	Complain	Satisfaction_Score	Points_Earned	France	Germany	Spain	Diamond	Gold	Platinum	Silver
00.000000	10000.000000	10000.000000	10000.000000	10000.000000	10000.000000	10000.000000	10000.000000	10000.00000	10000.000000	10000.000000
0.203800	0.204400	3.013800	606.515100	0.501400	0.250900	0.247700	0.250700	0.25020	0.249500	0.249600
0.402842	0.403283	1.405919	225.924839	0.500023	0.433553	0.431698	0.433438	0.43315	0.432745	0.432803
0.000000	0.000000	1.000000	119.000000	0.000000	0.000000	0.000000	0.000000	0.00000	0.000000	0.000000
0.000000	0.000000	2.000000	410.000000	0.000000	0.000000	0.000000	0.000000	0.00000	0.000000	0.000000
0.000000	0.000000	3.000000	605.000000	1.000000	0.000000	0.000000	0.000000	0.00000	0.000000	0.000000
0.000000	0.000000	4.000000	801.000000	1.000000	1.000000	0.000000	1.000000	1.00000	0.000000	0.000000
1.000000	1.000000	5.000000	1000.000000	1.000000	1.000000	1.000000	1.000000	1.00000	1.000000	1.000000

```
df.describe()
```

	CreditScore	Gender	Age	Tenure	Balance	NumOfProducts	HasCrCard	IsActiveMember	EstimatedSalary
count	10000.000000	10000.000000	10000.000000	10000.000000	10000.000000	10000.000000	10000.00000	10000.000000	10000.000000
mean	650.528800	0.545700	38.921800	5.012800	76485.889288	1.530200	0.70550	0.515100	100090.239881
std	96.653299	0.497932	10.487806	2.892174	62397.405202	0.581654	0.45584	0.499797	57510.492818
min	350.000000	0.000000	18.000000	0.000000	0.000000	1.000000	0.00000	0.000000	11.580000
25%	584.000000	0.000000	32.000000	3.000000	0.000000	1.000000	0.00000	0.000000	51002.110000
50%	652.000000	1.000000	37.000000	5.000000	97198.540000	1.000000	1.00000	1.000000	100193.915000
75%	718.000000	1.000000	44.000000	7.000000	127644.240000	2.000000	1.00000	1.000000	149388.247500
max	850.000000	1.000000	92.000000	10.000000	250898.090000	4.000000	1.00000	1.000000	199992.480000




Number of Pointed Earned Distribution by Customers Exited
Exited
Exited
Not Exited
Frequency
Points Earned
0
50
100
150
200
250
300
350
200
400
600
800
1000


Card Type Distribution
Diamond
Gold
Platinum
Silver
Count
Card Type
0
1000
2000
3000
4000
5000
6000
7000
8000
Not Exited
Exited