MGT - 6203

FINAL REPORT

Reducing Bank Customer Churn and Improving Customer Retention Strategies

NOV ‘22

Team #44:

Sofia Sabirova

Lucia Etchecopar

Ricardo Goncalves

Sebastian Riveros

Background

It is no secret that finding new and retaining existing customers is one of the key aspects of running any successful and profitable business. However, finding new clients usually comes with a considerable cost (time, administration, marketing, etc.), therefore, it might make more sense for companies to focus their efforts on retaining their current clientele. According to a fintech company, FI Works, the estimated cost of a bank acquiring a single customer is $500 (source: *Banking Statistics – FI Works – Sales and Marketing You Can Bank on*), and their expected revenue per year is $300, making them profitable after their 2nd year in the bank. This is why maintaining clients for a long-term is crucial. However, it is not easy to determine which customers are at risk of leaving since there are hundreds of different reasons why a customer might cancel his/her services. This poses a great problem for most companies across different industries, therefore this is why we believe that developing an accurate prediction model that labels which customers are likely to churn is an interesting problem to solve. To further the importance of this issue, according to research, "improving the retention rate by up to 5% can increase a bank's profit by up to 85%" (source: *"Predicting Customers Churning in Banking Industry: A Machine Learning Approach"*, p.1).

Therefore, when a customer leaves before the bank can, at least, recuperate the cost of their recruitment, it constitutes a loss. However, there is more to attrition than a calculated short-term monetary loss. When existing customers tend to leave, it is hard for any business to build a loyal customer base which, ultimately, may influence their overall future earning potential - an impact that may not be quantified right away. Having a loyal customer base contributes to building strong brand recognition, helps attract new customers, and, consequently, withstand challenging business situations. Unsurprisingly, it is an ongoing industry-wide task to prevent clients from churning and retain them for years to come. In our project, we will take on the challenge of building and testing several bank churn predictive models. Based on the results of modeling and additional analyses, we will propose data-informed strategies for retaining customers.

According to the literature, churn can be defined as a sequence of actions (or a lack thereof) by a customer over a certain period (source: *Why Customers Leave & What Can Banks Do? | Tiger Analytics*). The literature search also confirmed our initial methodology plan. In most cases (including our dataset), churn is a binary indicator variable, therefore, classification type of models are the most appropriate for the task. In terms of the specific models, Muneer et al. (source: *Predicting Customers Churning in Banking Industry: A Machine Learning Approach*) suggests several methods that have demonstrated reliable results, such as SVM, KNN, logistic regression, decision tree, and random forest among many others including some hybrid and proprietary algorithms. For the purpose of this study, we decided to focus on applying the basic supervised ML algorithms such as SVM, logistic regression, and random forest decision tree to 1) test which model appears to do a better job predicting churning customers and 2) strengthen our understanding of these foundational methods to build more rigorous predictive models in the future (when we are farther along in our analytics training).

Problem Statement

We are going to dive the project into two main components:

1) Preventing churn of bank customers by building predictive models that will label high-risk clients, thus, enabling the customer relationship management team to take necessary steps to prevent them from churning (predictive modeling)
2) Figure out the most effective bank customer retention strategies which may include but are not limited to product development (making sure the product fits clients' needs), operations (ease of online banking and reliable and responsive customer support, etc.) and, possibly, others (prescriptive modeling)

We plan to solve this by finding the variables contributing to whether the customer stays with the bank or leaves, and use these insights to propose business initiatives that target the flaws that cause the user to churn.

How we did it

Hypotheses. In terms of a single variable impact ("holding all other factors constant"), we hypothesize that:

1) the more products a person is using, the less likely they are to churn;
2) the more balance they have, the less likely they are to churn;
3) the older they are, the less likely they are to churn;
4) active members are less likely to churn.

We want to answer the following ***research questions***:

RQ#1. Which of the models (logistic regression, a support vector machine model, and a random forest classification decision tree) is the best for predicting bank customer churn using the data source?
RQ#2. Which of the following variables are most impactful for predicting attrition and whether the magnitude of these variables' impact is meaningful: customer's credit score, country, gender, age, tenure, existing balance, the number of bank products a customer is using, whether a person has a credit card with a bank or not, level of customer activity, and estimated salary.
RQ#3. Which of the following variables predict whether the customer will sign up or not for a term deposit subscription offered via a direct marketing call: age, job, marital status, education, if the client has credit in default, their balance at the bank, whether the client has a housing loan, whether the call has been made via cell phone or telephone, day of the week of the contact, month of the contact, duration of the contact, number of times the client was conducted over the course of the campaign, number of days that passed by after the client was last contacted from a previous campaign, and number of contacts performed before this campaign and for this client.
RQ#4. Based on the findings in RQ#2 and RQ#3, (1) what are the key characteristics of clients with a high risk for churning and (2) what would be the most effective retention strategies for them, and (3) how can we effectively target new customers that are less likely to churn? We'll provide an answer to this research question in ***Conclusions***.

- As an ***optional research question RQ#4.1***, we were going to investigate the underlying structure of the bank customer churn using K-means clustering to see the main segments of customers. This could yield some useful information to answer RQ#4.

Also, as an ***added bonus***, we wanted to provide our hypothetical bank "client" a list of promising potential customers with a low risk for churning (see description for Dataset C below).

Overview of Data

We use three data sources to address the problem:

1) Dataset A: for predicting the customer churn for a given Bank A.
2) Dataset B: descriptive information about customers in a given bank B and a log of recent communication with the customers regarding a deposit subscription promotional campaign.

We will use datasets A and B to help us build a profile of a good customer, in terms of churn and acceptance of promotions.

3) Dataset C: This dataset will be used as a source of potential bank customers that would fit our previously defined profile.

EDA. We performed data cleaning on the three datasets and found no missing or invalid data. Still, there were columns that weren't meaningful (for example, identification), and we removed those. We did a correlation matrix in datasets A and B and were able to find some key variables (the ones that were most correlated with the outcome). From that analysis, we were able to identify the most meaningful variables to predict churn (age, active_member, and country) and the outcome of a campaign (duration and type of contact). Then, we continued with a visual exploration of the relationship between dependent and independent variables. Regarding feature engineering, we did different kinds of transformations:

- Create binary dummy variables for the categorical features (for example, gender, country)
- Make "age" categorical instead of continuous (create three categories, young/adult/elder) using equal width binning, and then make that variable binary dummy.
- The bank campaign dataset (B) had a month granularity; we removed it and made it seasonal.
- Similarly, we removed the date column and created a day_of_week because the date is too granular.

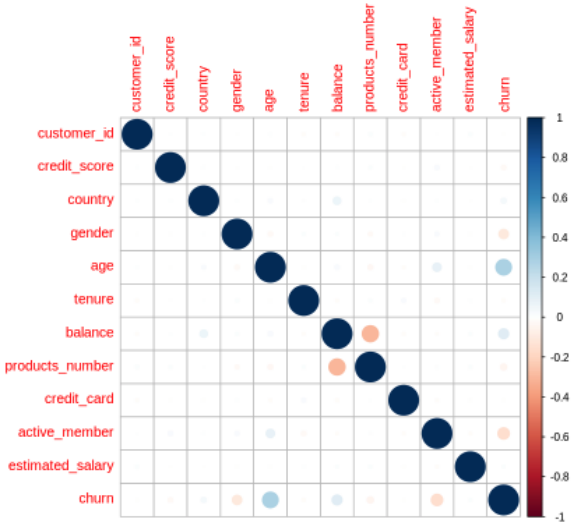

Figure 1 - Correlation matrix dataset A

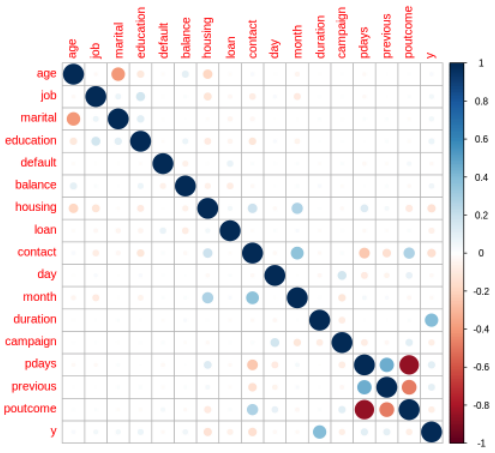

Figure 2 - Correlation matrix dataset B

The models

Research questions # 1 through #3

Logistic Regression (RQ1, 2 and 3). We created two logistic regression models, the first to see if a given customer will churn and the second to predict if the customer will accept a promotion. To begin with, we split each data set between training a test, using 70% for training and the rest for testing. We started creating a model with all variables, but as expected from our previous analyses, there were several variables that weren't meaningful. Using a backward elimination approach, we eliminated the variables that were not significant, and we ended up with our final models.

Regarding the performance of the churn model, we need to set up a threshold of only 30% (meaning that if the customer has 30% of churn, it will be identified as churn) to get decent results regarding sensitivity (the number of true positives over true positives plus false negatives). This can be explained by the fact that the dataset is unbalanced, meaning that only 20% of customers churn versus 80% that don't, so it's more difficult to identify that 20%. On the other hand, the model is pretty good in specificity and accuracy because of the number of negative instances.

The marketing campaign dataset has the same unbalanced issue, however, there are four times more data, so with a 50% threshold, we get a decent sensitivity (40%), which, of course, can be improved if we decrease the cutoff value.
Using a backward elimination approach, we eliminated the variables that were not significant. We ended up with the following significant variables:

- Churn model: **Age, Gender, Country, Active_member, Balance, Products_number**
- Marketing dataset: **Housing, Loan, Contact, Duration, Campaign, Poutcome, Season, Age, Marital status**

For the **Churn *(RQ2)*** model, based on the logistic regression models and the visual analysis, we can conclude that the less likely people to churn are the younger (less than 30 years old) males, from France and Spain, that are active members, have a higher balance and product number than the mean.
For the **Campaign *(RQ3)*** model, we identified some characteristics of the contact and a client that are associated with a higher probability of acceptance of a promotion. Promotions that are accepted are being offered from cellular calls with longer durations and during summer. Customers who don't have a house or a personal loan, who are single (which is compatible with the age group from the other model), and that have previously accepted a promotion are more likely to accept a new one.

SVM for Predicting Customer Churn (RQ1). We also built several SVM models and tested different values for the hyperparameter C (Lambda). While being relatively accurate (~80% of corrected predicted test cases), the first model's sensitivity was essentially 0% because all of the model's predicted values were non-churners. We then conducted a series of experiments to find the most optimal value of C and the kernel (Table 1).

Table 1. ***SVM models comparison***

SVM Models	Factors	Type	Kernel	C	Scaled Data	Accuracy on Test Data	Sensitivity on Test Data
Model 1	11	C-Classification method	Linear ("vanilladot")	0.1	True	~80%	0%
Models 2 ***(multiple sub-models with different Cs)***	11	C-Classification method	Linear ("vanilladot")	Sequence of Cs tested Start:1; End: 10,001 Step: 2,000. *C=10,000 was tested separately as part of Model 2 submodels.	True	Between ~ 77% and 80%	0% for all tested values of C, with minor improvement for C=8,001 & C=10,001.
Model 3	11	C-Classification method	Non-linear ("rbfdot")	C=10,000 (best sensitivity (~44%) from Models 2)	True	~79%	~49%

To improve the performance of our SVM models, we reviewed some online resources to understand how the model can be optimized. The most straightforward and basic way of optimizing an SVM is changing the C parameter. We tested a sequence of C numbers from 1 to 10,001 with a step of 2,000 and found that while C=10,001 increased the error rate, it also improved sensitivity to a little over 20%. Additional test models with a C=10,000 actually demonstrated better results with a sensitivity of about 44%. Another recommended step in optimizing an SVM is changing the kernel type. We started with a linear kernel but decided to try a non-linear one to see if it produced a model with a better fit. Indeed, our results showed that the combination of hyperparameter C=10,000 and the non-linear kernel "rbfdot" sensitivity increased to ~49%.

Random Forest (RQ1 through RQ3). In general, Random Forest models tend to perform better in terms of accuracy compared to SVMs and Logistic regressions since they consider all possible trees and continue performing iterations until it captures all "nodes". Therefore, the key factors to take into account when running RF models is whether you can improve its performance or not. According to Leo Breiman and Adele Cutler from Berkeley University, the amount of trees that you choose for your model does not overfit it, therefore the key question is to choose the optimal amount of trees to reduce complexity and CPU performance while not sacrificing its predictive power. However, in our case, since the number of variables we had in our datasets as well as the amount of rows we had were considerably small, performance wasn't a concern for our analysis.

For the **Churn model (*RQ1 and RQ2*),** our Random forest model showed us that the top 5 variables were: products_number, balance, credit_score, estimated_salary_year and estimated_salary. When we compared the RF model considering all variables vs. top 5 variables, we saw a slight decrease in performance of our model, 83% accuracy vs 79% accuracy. Since our total number of variables is fairly small, it is better to use the RF model with all the variables. However, if our dataset had 800+ variables, then we could consider sacrificing accuracy to reduce the number of variables, and therefore improving processing performance of the model. Additionally, something important to consider is that while we had a high specificity (~96%), our sensitivity (~26%) was very poor. To prevent customers from churning, it would be ideal for a model to predict accurately which customers are most likely to churn (sensitivity), such that initiatives can be put in place to focus on those customers that are predicted to churn. Having a small sample size of customer churning in our dataset might explain why our sensitivity is so low. Capturing more data points would improve our model.

In the case of the **Campaign** model ***(RQ3)***, we observed a similar pattern. Running the model for all variables produced better results than running it only against the significant variables (duration, balance, p_days and weekdays). We ran into the same issue of our model not being able to predict "effective campaigns" (low sensitivity), while doing a fantastic job of predicting "ineffective campaigns". Again, this intuitively makes sense because usually marketing campaigns tackle a huge audience, and it is known that a lot of individuals will disregard the campaign.

Table 2. ***Churn prediction models comparison (Dataset A)***

Item	Logistic Regression	SVM	Random Forest
Confusion Matrices	Confusion Matrix and Statistics Reference Prediction 0 1 0 2010 362 1 383 245 Accuracy : 0.7517 95% CI : (0.7358, 0.76 No Information Rate : 0.7977 P-Value [Acc > NIR] : 1.0000 Kappa : 0.2405 Mcnemar's Test P-Value : 0.4637 Sensitivity : 0.40362 Specificity : 0.83995 Pos Pred Value : 0.39013 Neg Pred Value : 0.84739 Prevalence : 0.20233 Detection Rate : 0.08167 Detection Prevalence : 0.20933 Balanced Accuracy : 0.62179 'Positive' Class : 1	Confusion Matrix and Statistics Reference Prediction 0 1 0 2068 312 1 325 295 Accuracy : 0.7877 95% CI : (0.7726, 0.80 No Information Rate : 0.7977 P-Value [Acc > NIR] : 0.9165 Kappa : 0.3474 Mcnemar's Test P-Value : 0.6345 Sensitivity : 0.48600 Specificity : 0.86419 Pos Pred Value : 0.47581 Neg Pred Value : 0.86891 Prevalence : 0.20233 Detection Rate : 0.09833 Detection Prevalence : 0.20667 Balanced Accuracy : 0.67509 'Positive' Class : 1	Confusion Matrix and Statistics Reference Prediction 0 1 0 2335 445 1 58 162 Accuracy : 0.8323 95% CI : (0.8185, 0.8455) No Information Rate : 0.7977 P-Value [Acc > NIR] : 0.0000007634 Kappa : 0.3184 Mcnemar's Test P-Value : < 0.00000000000000 Sensitivity : 0.26689 Specificity : 0.97576 Pos Pred Value : 0.73636 Neg Pred Value : 0.83993 Prevalence : 0.20233 Detection Rate : 0.05400 Detection Prevalence : 0.07333 Balanced Accuracy : 0.62132 'Positive' Class : 1
Sensitivity	40.3%	48.6%	26.7%
Specificity	84.0%	86.9%	97.6%
Accuracy	75.2%	78.8%	83.3%
Findings	The less likely people to churn are the: -younger males (less than 30 years old) -from France and Spain -that are active members -with a higher balance and higher number of products than the mean	N/A	-Fairly straightforward to run Random Forest models on datasets with limited variables
Pros	-Great performance in specificity and accuracy because of the number of negative instances	-Best sensitivity of the three models	-Best accuracy of the three models
Cons	-Threshold of 30% (to get decent results in sensitivity) Because the dataset is unbalanced (20% of customers churn versus 80% that don't, so it's more difficult to identify that 20%)	-It works as a “black-box” algorithm because we do not know which variables contribute the most to the response	-Sensitivity is low and there are no straightforward ways to improve it by manipulating model parameters. Increasing the number of data points for customers that churned should improve it

Table 3. ***Marketing campaign models comparison (Dataset B)***

Item	Logistic Regression	Random Forest
Confusion Matrices	Confusion Matrix and Statistics Reference Prediction 0 1 0 11317 848 1 584 815 Accuracy : 0.8944 95% CI : (0.8891, 0.8995) No Information Rate : 0.8774 P-Value [Acc > NIR] : 0.000000000366632 Kappa : 0.4733 Mcnemar's Test P-Value : 0.000000000003653 Sensitivity : 0.49008 Specificity : 0.95093 Pos Pred Value : 0.58256 Neg Pred Value : 0.93029 Prevalence : 0.12260 Detection Rate : 0.06009 Detection Prevalence : 0.10314 Balanced Accuracy : 0.72050 'Positive' Class : 1	Confusion Matrix and Statistics Reference Prediction 0 1 0 11589 1070 1 312 593 Accuracy : 0.8981 95% CI : (0.8929, 0.9032) No Information Rate : 0.8774 P-Value [Acc > NIR] : 0.00000000000002499 Kappa : 0.4109 Mcnemar's Test P-Value : < 0.00000000000000022 Sensitivity : 0.35658 Specificity : 0.97378 Pos Pred Value : 0.65525 Neg Pred Value : 0.91548 Prevalence : 0.12260 Detection Rate : 0.04372 Detection Prevalence : 0.06672 Balanced Accuracy : 0.66518 'Positive' Class : 1
Sensitivity	49%	36%
Specificity	95%	97%
Accuracy	89%	90%
Findings	Promotions are more likely to be accepted whe: -are being offered from cellular calls with longer durations and during summer. -people that don't have a house or a personal loan -who are single -and that have previously accepted a promotion	-Fairly straightforward to run Random Forest models on datasets with limited variables
Pros	-Best sensitivity of the two models	-Very good model for determining which customers will NOT accept the marketing campaign
Cons	-Threshold of 50% (to get decent results in sensitivity) Because the dataset is unbalanced	-Again, a small quantity of “positive” responses in our dataset caused a low sensitivity

Research Question #4.1

Outlier detection. To study the underlying data structure for the purpose of customer segmentation, we analyzed the key 6 numeric variables (credit score, age, tenure, balance, number of products they use, and estimated salary) among all bank customers (whether they are labeled as churners or not). Our review identified that variables age and credit score include a significant number of observations that are below the first quartile plus 1.5 times the interquartile range (Q1+1.5*IQR) or above the third quartile plus 1.5 times the interquartile range (Q3+1.5*IQR). Since K-means is not robust to outliers, we decided that it was beneficial to remove these observations. We also found that one observation in the variable "products used" is outside the upper whisker of the boxplot. However, due to the fact that it is a single observation within a 10,000 case dataset, we did not think it would have any measurable impact on the accuracy of clustering. In total, we removed about 4.3% of original observations before attempting to cluster our data.

Determining an optimal number of clusters. Based on the elbow chart (Figure 3, left), the most optimal number of clusters is 4 since the within sum of squares (wss) has leveled off once it reached 4 clusters. With a larger number of clusters, the decrease in total wss appears negligible. Another way to determine the optimal number of clusters is to look at the gap statistic. The gap statistic indicates the difference in parameters between clusters vs the difference found in the unclustered distribution. Therefore, the higher that statistic, the larger the difference between clusters and nonclustered distribution. From the gap statistic chart (Figure 3, right), it appears that the gap statistic is at its highest at 10 clusters, however, similarly to the elbow

chart, an increase in the gap stat starts to level off at 4 clusters. Therefore, we consider 4 clusters to be the most optimal number of groupings, confirmed by the two visualizations.

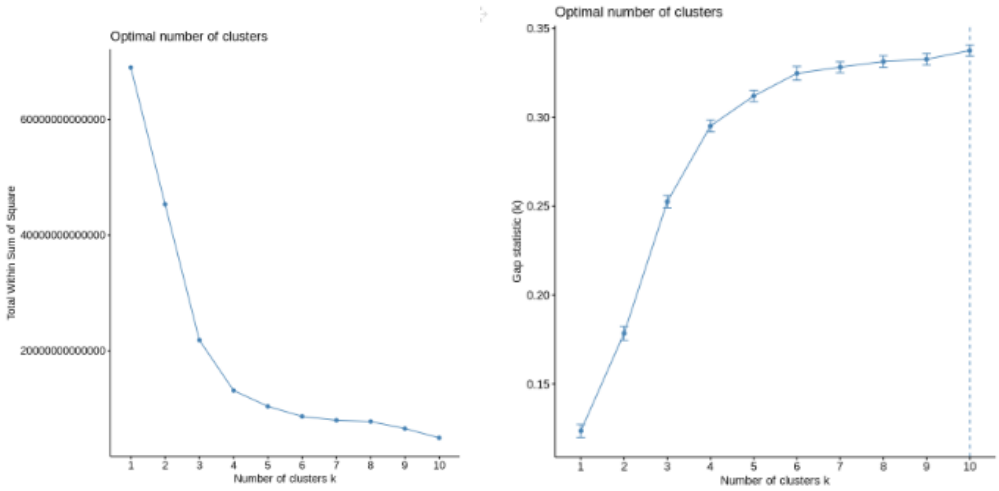

Figure 3 - Elbow chart and Gap statistic for determining the optimal number of clusters

Characteristics of the four clusters of all bank churn data. We ended up with 4 clusters of the following sizes and proportions of the original dataset: 1843 (19.3%), 1795 (18.8%), 2939 (30.7%), and 2996 (31.3%). In terms of the means of the clusters for the key variables, there's some variation in the number of products used, balance and estimated salary. However, it is not a clear cut differentiation among the 4 groups of clients. We plotted clusters against the first two principal components on the axes, and did notice that there was a lot of overlap among the four groups (see Figure 4).

Figure 4 - The first 2 principal components of the factors plotted for four clusters of the churn bank dataset

We also reviewed the clusters' mean statistics. When comparing the cluster means and the mean across the dataset, we noticed that these clusters are well in line with the mean for credit score, age, and tenure. They do, however, seem to differ from the overall, ungrouped data in terms of balance, number of products, and estimated salary. Clusters 1 & 2 mean balance (less than 3,000) is way below the mean balance across the dataset (76,535). Clusters 3 & 4, on average, use a slightly lower number of products as the overall, unclustered, dataset (1.37 and 1.4 vs 1.5 respectively). Clusters 1 & 2, on average, use a larger number of products (~1.8) compared to the overall dataset (1.5 products). Salary-wise, clusters 1 and 3 have salaries on the lower spectrum, i.e., less than the first quartile, or 25% of observations. Clusters 2 and 4, on average, have salaries above the 3rd quartile, or 75% of the dataset. Then, we wanted to see the percentage of "churners" within each of the 4 clusters. As shown in Table 6, "churners" are mostly concentrated in clusters 3 and 4.

Table 6. Characteristics of the Four Groups of Bank Customers (Dataset A)

Cluster	Mean						Churn %	Size relative to the dataset w/o outliers
	Credit Score	Age	Tenure	Balance	Number of Products Used	Estimated Salary		
1	649.5	37.3	5.1	2,352.06	1.8	50,169	13.8	1843 (19.3%)
2	649.0	37.0	5.1	2,560.53	1.8	149,616	14.9	1795 (18.8%)
3	651.6	38.0	4.9	122,160	1.4	50,575.6	23.5	2939 (30.7%)
4	651.8	37.8	5.0	121,734	1.4	149,849	23.9	2996 (31.3%)

Churn Status and Cluster Assignment Dependence. We also wanted to see if this seemingly disproportionate concentration of churners in clusters 3 and 4 is a coincidence or if there's truly a statistical dependence between the churn status and cluster status. To check for the presence of any statistically significant relationship, we ran a chi-square test which showed that, indeed, there is some dependence (see figure 7 in Appendix). Therefore, it is likely that clusters 3 and 4 could be shed more light on the "average" churner than clusters 1 and 2. Overall, it appears that individuals assigned to clusters 3 and 4 tend to use a lower number of products than other clusters, and they have a much higher balance than other clusters. However, they are strikingly different in terms of their estimated salary with cluster 4 being the cluster of high earners. Table 8 (Appendix) provides an overview of the main differences among the four clusters. One interesting observation is that clusters that are high risk for churn customers tend to have a higher balance and use fewer products.

Clustering within the Group of Churners. In addition to investigating the underlying structure of the overall bank customer dataset A, we also wanted to conduct a more granular level of analysis of churn customers only and see if we could identify distinct groups of high risk clients. We followed the same procedure to find the optimal number of clusters which also turned out to be 4. The groups' mean statistics are presented in Table 9. As we can see, while the credit scores are, on average, similar to the mean of the overall dataset, these groups incorporate slightly older individuals. An average customer in these groups is around 44 years old vs 38 years old in the overall dataset. While the mean tenure for clusters 2, 3, and 4 is about 5 years, it is slightly less than the average across the entire sample. *Profiles of the Most and Least "Risky" Churn Customers.* At a more granular level, our most "problematic" group of customers are women in their mid-forties living in Germany who have higher balances (see Table 10). Among these women there are two subgroups - higher earners and lower earners. Among churners, customers in Spain, overall, appear the least likely to churn with French customers representing the mid-range risk.

Table 9. Characteristics of the Four Groups of Churn Only Bank Customers (Dataset A)

Clusters of Churn Customers only	Mean					
	Credit Score	Age	Tenure	Balance	Number of Products Used	Estimated Salary
1	645.1	43.7	5.1	123,441.0	1.5	150,283.13
2	650.6	43.8	4.9	123,769.2	1.4	49,175.13
3	651.7	44.1	4.6	4,203.3	1.5	54,950.39
4	644.3	44.1	4.9	5,892.5	1.5	154,113.27

Table 10. Country and Gender Distribution by Churn Only Clusters

Clusters of Churn Customers only	France		Germany		Spain		Grand Total
	Female	Male	Female	Male	Female	Male	
1	5.9%	5.2%	11.1%	8.5%	3.8%	2.3%	**36.8%**
2	5.4%	5.0%	10.8%	9.6%	2.3%	2.5%	**35.7%**
3	6.0%	3.4%	0.0%	0.0%	2.8%	2.0%	**14.1%**
4	5.2%	3.8%	0.1%	0.1%	2.5%	1.8%	**13.4%**
Grand Total	**22.5%**	**17.4%**	**22.0%**	**18.1%**	**11.4%**	**8.6%**	**100.0%**

Added Bonus - Low Risk for Churn Customer Database

From the models above, we learned a lot about customers of two different banks. We know the main characteristics of customers that don't churn and also we know about characteristics of people that tend to accept promotional offers, and some conditions in which those offers were presented. We can use that information to build a database of potential new customers that a bank (our hypothetical client) can contact. As we have learned in this project, it is much more expensive to recruit a new customer than retain an existing one, so we will be only selecting potential customers that we know with a certain degree of confidence that will be good customers. Therefore, we are only selecting individuals from our adult's dataset C who fit the following criteria: men, younger than 44 years old, and single (we will consider never married or separated/divorced as single).

Challenges

One of the main challenges we faced at the start of the project was finding at least 2 related datasets, per project's requirements. Initially, we set out to search for datasets that can be merged on the basis of subjects (i.e., the same observations). We weren't able to find such datasets that also met other requirements, however, we located datasets that compliment each other very well content-wise. The two main datasets used for modeling in this project deal with information related to factors that may potentially impact customer interest in specific bank's services. In terms of the analyses, one of the challenges we faced was a long run-time (~40 minutes) of the SVM models looping through different C values which makes it harder to fine tune the models on the fly (modifying them would cause the need to rerun them). If we had more time as a team, it would have been beneficial to look into the possibility of parallel execution.

Conclusions

Pros and Cons of the Main Models. We built three main models: Logistic Regression, SVM, and Random Forest. While even the "best" sensitivity-wise performing model, SVM, is only good at predicting churn in 49% of cases, we have learned a great deal of information about these three different types of models and the importance of analyzing the same problem using various methodological tools. For example, (1) Logistic Regression is great for probability outcomes like churn; it also can tell us which specific variables have a greater impact on the outcome. We can use it to tease out the exact drivers of the outcome variable such as churn.(2) SVM, on the other hand, works as a "black-box" algorithm. Its output does not provide the relative importance of the factors influencing the outcome. However, it showed the best sensitivity of the three models.(3) Random Forest performed better for accuracy since we can test as many variables and dataset combinations (random sampling with replacement) as we want to. It also can show us which variables are the most impactful in the model. The drawback is that we cannot get a metric that estimates the magnitude of importance of these variables.

Choice of the "Best" Model. Our end goal was to (1) build a predicting model to see if a customer is likely to churn and (2) find the variables with the higher impact, taking into account the brief description in the previous paragraph of the three models, we

consider that the best choice is Logistic Regression since it is 75% accurate and 40% sensitive, and can tell us which specific variables have the most impact as well as provide a quantifiable measure of impact (e.g. odds/probability). Based on the logistic regression results, variables such as country, whether the user is an active member or not, account balance, and the number of used products are the most influential for predicting churn. If our goal was prediction only and we didn't need to know the exact variables impacting churn, we could choose the SVM model is ~87% accurate and ~49% sensitive. We also found it interesting that the credit score variable did not appear statistically important in the logistic regression and SVM models, but has shown to be impactful, based on results of a random forest on a decision tree.

Business Recommendations Based on the Findings. Taking into account the variables that we found were the most important, we want to give business recommendations regarding maximizing our findings and reducing the churn rate: (1) users that are younger males, live in Spain and France, and have higher balance (more money in their accounts) are less likely to churn, therefore, there's an opportunity to understand what is happening in those markets that the bank is succeeding in and apply best practices to the ones where business isn't doing as well; the marketing team can run a market segmentation analysis similarly to what we have done to see gaps in high-risk churn countries: demographic analysis (age, gender, income), market capillary (room to grow as a company and expand to new cities), (2) active members are less likely to churn, so it's important to follow the engagement that actual users are having with the products of the bank; digital products are easier to analyze (since all data is already available in the company), so the digital product team can give insights here in terms of the usability of the bank's app to identify users that are having poor engagement and try to understand if there's an opportunity to improve the app (better communicate information, customer support chat, enhance the purchasing process, etc.) (3) users with higher number of products are less likely to churn, so the business team can use the actual user base to understand which customers are likely to cross-sell (sell products complimentary to those a customer already owns), to increase the usage of all the products of the bank and expand the value proposition to the user.

In addition, as we have learned through the K-means cluster analysis of the churn bank dataset A, the most "risky" customers are women in their mid-forties living in Germany who have higher balances. Among these women there are two subgroups - high earners and low earners. Males with the same characteristics who live in Germany are the second most likely to churn group of clients. Our findings from the analysis of the term deposit subscription campaign indicate that promotional offers are more likely to be accepted if the bank initiated a call on the customer's cell phone during the summer and was able to keep the customer engaged for longer. Our findings also showed that customers who don't have a house or a personal loan, who are single and that have previously accepted a promotional offer are more likely to accept a new one. Therefore, to retain a higher percentage of customers in Germany, the bank might want to initiate "live customer care" calls (not automated), during the summer season, and provide recommendations for new products to single women and men in Germany with higher bank balances who have a low credit load (fewer number of loans or none). Since those who have already accepted promotional offers in the past are more likely to accept them again, the bank might want to reward clients accepted promotions by, for example, providing them with a higher cash back rate, higher yield percentages for saving certificates, etc.

Long Term Goals. If we had more time in this class, we would investigate additional ways that can help increase sensitivity, such as running a principal component analysis on the factor variables and using a 2- or 3-factor model instead of a multiple-one and trying to use cross-validation for training the model instead of the 70/30 train/test split. Cross-validation could yield in helping the algorithm understand the "diversity" in the data, which, we think, might be of critical importance for a dataset that appears as imbalanced as the bank customer churn dataset with 80% of non-churning customers and 20% of churners. One of the problems we faced in our project was class imbalance, as most of the instances in our bank don't churn. This is a known issue in the industry, and some potential solutions can be applied. It would have been nice to try some of these solutions, for example, resample the training set with oversample minority class, undersample majority class or generating synthetic samples; or doing for k-fold validation as a method of splitting the data.

Citations

[1] Muneer, Amgad & Ali, Rao & Alghamdi, Amal & Mohd Taib, Shakirah & Almaghthwi, Ahmed & Ghaleb, Ebrahim. (2022). *Predicting customers churning in the banking industry: A machine learning approach.* Indonesian Journal of Electrical Engineering and Computer Science. 26. 539-549. 10.11591/ijeecs.v26.i1.pp539-549.

[2] *Banking Statistics – FI Works – Sales and Marketing You Can Bank on*, link

[3] *Why Customers Leave & What Can Banks Do?* | Tiger Analytics, link

[4] *Applying Random Forest on Customer Churn Data* | by Akhil Sharma, link

Appendix

Figure 7. Chi-squared Test Output on the Contingency Table "Churn vs Cluster Assignment"

```
        Pearson's Chi-squared test

data:  churn_cluster_all_contingency
X-squared = 122.72, df = 3, p-value < 0.00000000000000022
```

Table 8. Summary of Distinct Characteristics of the Four Main Groups of Bank Customers (Dataset A)

Cluster	Churn Risk	Balance	Number of Products Used	Estimated Salary
1	Lower	Lower	Higher	Lower
2	Lower	Lower	Higher	Higher
3	Higher	Higher	Lower	Lower
4	Higher	Higher	Lower	Higher

Note: Comparisons ("lower"/"higher") of churn risk, balance, number of products used and estimated salary are relative to the overall dataset A statistics.