Churn Prediction in Retail Banking Customers

Adi Srikanth • Andre Chen • David Roth

Jin Ishizuka • Lev Paciorkowski

NYU Center for Data Science

December 21, 2021

Abstract

Customer "churn" refers to the rate at which...

I. Introduction

In this analysis, we look at a dataset of 10,000 retail banking customers and try to predict whether an individual customer is likely to "churn", that is, if they are likely to close their account and cease their financial relationship with their bank.

II. Data Summary

OUR dataset contains information on customer churn at a retail bank, specifically features/predictor variables and a target variable (0 if the customer churned, 1 if the customer is still at the bank). The features include general information about the customers including demographics, socioeconomic information, and the level of engagement each customer has with the bank. Our goal is to study the relationships between the customer predictor variables and their probability of churning to understand what leads to customer churn at this bank.

We determined that investigating this dataset would be broadly useful as practice for future roles as data scientists in industry. Specifically, a very common use case in both consumer and enterprise-facing companies is understanding leading indicators of churn, as churn contributes to significant revenue erosion and needs to be proactively addressed. This dataset offers a great opportunity for us to practice not only the technical skills involved in determining predictors of customer churn, but also the more business-oriented aspects of asking suitable questions and presenting a cohesive conclusion that could benefit our hypothetical company.
The data is broken down as follows:

Rows: Each row represents an individual customer of the bank, providing information about that customer and whether or not they churned, which we assume in this context means they closed their account with the bank and stopped doing business with them altogether.

Columns: The 14th and last column of the dataset ("Exited") is a binary target variable that indicates whether a customer churned (0) or not (1). The first 13 columns of the dataset include information about each individual customer, of which a large subset were identified as potentially useful predictor variables and are described in detail below. We omitted columns like "CustomerId" and "Surname" from the columns listed below, as these columns likely have no relationship with a customer's likelihood of churning:

- *CreditScore*: The customer's credit score on a scale from 350 (worst) to 850 (best)
- *Geography*: The customer's country of resi-

dence (dataset includes France, Germany, and Spain)
- *Gender*: Male or Female
- *Age*: Age of customer $(18 - 92)$
- *Tenure*: How long the customer has been with the bank, in years.
- *Balance*: Customer balance to date. (Note: customers who have closed their accounts with the bnk may still have uncollected balances)
- *NumOfProducts*: The number of bank products used by the customer.
- *HasCrCard*: Whether the customer has a credit card (1) or not (0)
- *IsActiveMember*: Binary variable indicating if customer's activity is above the bank's minimim threshold for "Active".
- *Estimated Salary*: Customer's estimated yearly salary.

III. Hypothesis Testing

Do customers with varying levels of engagement with the bank have significantly different probabilities of churn? What does this tell us about the measures of engagement which matter the most?

OUR ultimate goal is to determine indicators of churn that the bank can act on, so that we could then recommend specific areas the bank could target for improvement in their interactions with customers. We accomplish this by defining customer profiles based on three measures of engagement (*NumOfProducts*, *IsActiveMember* and *Tenure*), and then implement hypothesis testing to see if churn proportion is significantly different from one profile to another.

"Product-Driven" Customers

We find that only about 3% of customers have more than two products with the bank, and of them, 85.9% churned compared to just 18.2% of non Product-Driven customers.

$$H_0 : P(churn)_{\text{product-driven}} = P(churn)_{\text{non-product-driven}}$$
$$H_1 : P(churn)_{\text{product-driven}} > P(churn)_{\text{non-product-driven}}$$

"Active" Customers

Customers are evenly split between Active and not Active (as determined by the bank), and only 14.3% of Active customers churned, compared to 26.9% of non-Active customers.

$$H_0 : P(churn)_{\text{active}} = P(churn)_{\text{non-active}}$$
$$H_1 : P(churn)_{\text{active}} > P(churn)_{\text{non-active}}$$

Table 1: *Significance Results*

	Test	p-value
0	*Product-Drivenness*	$2.9e-196$
1	*"Active"-ness*	$3.0e-55$
2	*Tenure*	$1.2e-02$

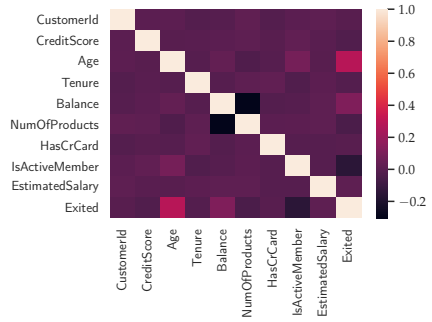

Figure 1: *Correlations among predictor variables.*

"Established" Customers

We find that Customers with tenures longer than 1 year churned at a rate of 20%, compared to new customers, who churned at a rate of 22.6%.

$$H_0 : P(churn)_{\text{established}} = P(churn)_{\text{new}}$$
$$H_1 : P(churn)_{\text{established}} > P(churn)_{\text{new}}$$

Results

In general, Product-Driveness and "Active"-ness are highly significant predictors of churn at $\alpha = 0.005$, while Tenure is not (See Table 1). Although Product-Driven customers are quite rare, they have a much greater tendency to churn compared to mainstream customers who use only one or two products. Perhaps this could be an indication of so-called credit card "hopping", but deeper investigation would be required to evaluate this. Based on these results, the bank can be confident in how it assesses "Active" customers, but might consider interventions to disincentivize customers from becoming Product-Driven. In contrast, a customer's Tenure should not be a central focus.

Table 2: *Boruta Example*

A	*A_copy*	*B*	*B_copy*
1	5	22	30
5	6	25	21
6	7	21	25
7	1	30	22

IV. Feature Selection

In order to remove excess "noise" from our model, we embark on feature selection. Specifically, we implement the feature selection process known as the Boruta algorithm. Boruta functions by creating a copy of each variable in the dataset, randomizing the copy, and attaching it back to the dataset. For example, if a dataset has original columns A and B, Boruta takes that dataset and produces a new dataset with columns *A*, *A_Copy*, *B*, and *B_Copy*, where *A_Copy* and *B_Copy* are shuffled versions of A and B, respectively (consider the copy columns essentially a random number generator sampling from the range of their original columns). Table 2 illustrates this example.

Next, Boruta takes all of the "copy" columns and selects the copy column that was the best predictor of the dependent variable. If any of the original columns is a worse indicator of the dependent variable than the best copy column, that original column is removed. This entire process is repeated to ensure statistical significance. The foundational idea is that no variable in the data set is a worse predictor than a random column of data – each variable must have a non-random contribution towards predicting the dependent variable.

Ultimately, we whittle our original set of variables down to seven variables: age, estimated salary, credit score, country of origin, number of products owned, member status, and remaining balance. We use these variables in the next step of our analysis, model implementation, tuning, and evaluation.

V. Model Implementation, Tuning, & Evaluation

To build an effective model using our dataset, we take a rapid prototyping approach by optimizing and comparing various supervised classification models in their ability to predict churn. Below we detail steps in our generalized approach to assessing each model, including train/test split, hyperparameter tuning, and model evaluation:

i. Evaluation Metric

Considering the costly impacts of both false positives (wasting resources on customers not at risk of churn) and false negatives (losing customers), a reasonable metric that helps us balance between the two in model optimization is the F1 score. Specifically, we look at the F1 score for our "positive" predictions (overall ability to predict churn) as well as the macro-averaged F1 (unweighted average of F1 scores on predicting churn and retention, respectively) as a secondary metric.

ii. Train/Test Split

We used an 80/20 split for training/testing (holdout). The 20% holdout set was used only to compute a final evaluation metric for a given model after optimizing on the training set.

iii. Standardization

For models that require standardization (e.g. Logistic Regression, SVM, KNN), we center the data and scale to unit variance.

iv. Hyperparameter Tuning

For each model, we identify its most important hyperparameters and define a reasonable search space for each parameter. We use grid search or randomized search to approximate an optimal combination of hyperparameters based on F1 score.

v. Final Evaluation

For each optimized model, we compute its F1 score by comparing its predictions on the test set features against the true values of the test set target variable. F1 scores for five different models are reported in Table 3

Table 3: *Model Performance*

Model	F1	Macro Averaged F1
Logistic Regresion	TBD	TBD
K-Nearest Neighbors	0.55	0.73
Support Vector Machine	0.54	0.73
Random Forest	0.6	0.75
XGBoost	0.57	0.75

VI. Results

VII. Discussion

i. Subsection One
ii. Subsection Two