Predicting Bank Client Churn

Dominique R. Grimes

November 18, 2023

Table of Contents

Predicting Bank Client Churn

Profit is the lifeline of an organization. It is the source that keeps the employees paid, the lights on, investing in the future and, ultimately, success. An organization's clients are the direct link to profits. If a client's loyalty is gained, then profits will follow. It sounds simple, but anyone contributing towards the success of a business will quickly find themselves navigating the nuances of winning and achieving client satisfaction. Once that goal is obtained, the journey has just begun. The task at hand then becomes maintaining the relationship and bringing more value to clients over the competition. If the team falls short of meeting that objective, the result is a lost client and reduced profits.

Problem Statement

It is unrealistic to expect a 100% client retention rate in any organization. However, it is essential to notice downward trends in retention rates and take action to save clients and profits. Often, teammates spend time and resources guessing at what went wrong reactively. Instead, data science can be leveraged to apply a predictive analytics model focused on bank client churn. The model will identify potential client churn proactively to anticipate a client's departure. This allows teammates to act and save the client relationship prior to a loss, which minimizes attrition and maximizes profits.

Benefits

A client churn prediction model can benefit any company that has clients, but the model would have to be tailored to the organization's unique features and client behaviors. The advantages of building a churn prediction model include but are not limited to revenue prediction, client acquisition cost reduction, improved client experience, enhanced market share, better resource allocation, data-driven decision making, and customer lifetime value optimization (S., 2023). The primary target audience of this tool will be bank executives responsible for profit and loss sheets, the executive suite focused on strategic objectives, and the front-line teammates responsible for client relationships. The insights and trends that the

model identifies will be instrumental for managing client relationships. The model's features will identify the driving issues for client retention hidden within the data and provide guidance on where to pivot in current sales or service approaches.

The Data

The idea for the project came from an article that focused on predictive modeling practice ideas (Daivi, 2023). The dataset for this project was obtained from Kaggle and labeled Predicting Churn for Bank Customers (Adam, 2018). There are thirteen features and ten-thousand rows available for analysis. The features consist of CustomerID, Surname, CreditScore, Geography, Gender, Age, Tenure, Balance, NumberofProducts, HasCrCard, IsActiveMember, EstimatedSalary, and Exited. Each row represents an individual bank client. One concern is that the dataset has limited features and rows available which may impact the robustness of the prediction results. Bank data is generally limited due to regulatory controls which restricts additional datasets available for supplementation.

The dataset captures historical attrition behavior per client up to a certain point in time. The goal is to find key relationships and patterns between the historical client attributes to predict a potential client loss within a defined lead time and level of confidence. This will allow the ability to make decisions based on the identified potential losses and ultimately reduce risk for the organization.

Methods and Results

Client churn prediction requires a supervised, classification model. This type of model is chosen since the dataset is labeled, and there is a pre-determined, clear outcome. The simplest and most obvious approach is a binary classification model such as Logistic Regression, Naïve Bayes Classifier, Tree-based, and Random Forest (Daivi, 2023). The binary classification model will separate the data into two defined classes, whether the client will exit or not exit the bank's services (Avcontentteam, 2023). Additional models to explore are Adaptive Boosting, Support Vector classifier, and Neural Networks such as Multi-layer Perceptron classifier.

Data Validity & Exploratory Data Analysis

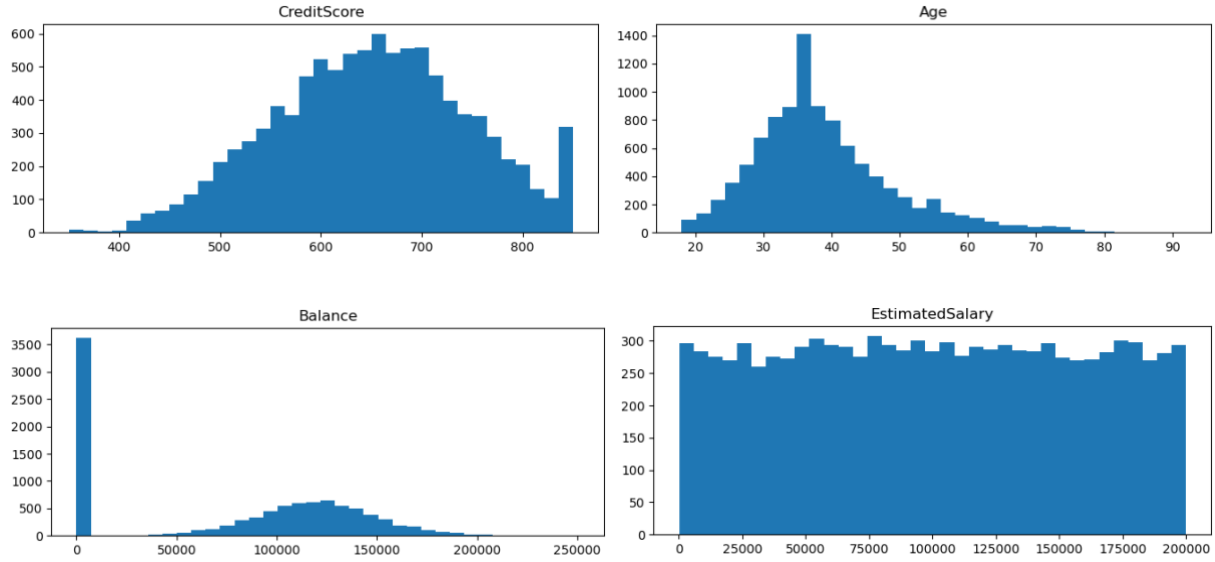
I initiated checking the validity of the data by looking for comparable alternative client churn datasets. I verified that other related datasets had similar features. I also checked for null values which revealed no missing data.

I observed the distributions of the continuous variables CreditScore, Age, Balance, and EstimatedSalary by reviewing summary statistics as well as visualizing box plots and histograms. I identified that three features have skewed distributions. CreditScore and Balance have negative skew and Age has a positive skew. Two of those features, CreditScore and Age have outliers and the outliers in the dataset are determined to be legitimate. I also observed the value counts of the categorical and binomial variables. For all features, the values make sense and don't seem to have input errors.

I completed Pearson correlations on continuous variables and Point-biserial correlations between the binomial features and each continuous feature. The continuous variables did not have any notable correlations. The only moderate correlation to point out from the categorical variables to the continuous variables is the positive correlation of .4 between Geography in Germany with binomial feature Balance. This may only be relevant if Geography or Balance are prominent features used in the model.

Figure 1

Histograms of continuous features



Note. CreditScore is bimodal with negative skew. Mean (650) and Median (652) are relatively consistent; however, the mode is the maximum credit score value of 850 which could be the cause in the skewed data. These values may need to be looked at separately than the rest of the credit score. Age is unimodal with a positive skew and is leptokurtic with a very sharp peak. Median and mode are both 37 with mean slightly to the right at approximately 39. Balance looks to have a relatively normal distribution if 0 is excluded from the balance value. The mean value of approximately 76,500 is impacted by the mode value of 0. With zero included, the median balance is approximately 97,000 and the distribution is bi-modal and negatively skewed. EstimatedSalary is relatively uniform in distribution. The mean and median are approximately 100,000 with a minimum close to 0 and maximum of nearly 200,000. However, the mode is approximately 25,000.

Data Preparation

I adjusted the Age variable with log transformation which helped correct the positive skew. However, log transformation did not help the skewness of CreditScore. For the Balance feature, there are a large amount of zero balances skewing the distribution. To optimize the feature's potential contribution to the model, I'm exploring five different options. To accomplish this, I copied the data frame with the transformed Age variable into four additional data frames. The optional data frames for modeling and evaluation are:

1. No change to the Balance variable.
2. Drop all zero Balance rows. This reduces the data frame from 10,000 rows to 6,383, which is a significant loss of data.
3. Convert all zero balances to one. Then complete log transformation of the values.
4. Create a new feature that computes the ratio of Balance to Salary called Bal_Sal. Then, drop the Balance and Salary columns due to collinearity.
5. Adjust the Balance feature to a binomial feature called Zero_Bal. Change all zeros to ones and all values greater than one to zero.

In addition to adjusting features for skewness, a very important data adjustment is balancing the target variable, Exited. I account for this by creating a balanced sampling of exited versus not-exited bank clients for each data frame. I took the total number of Exited in each optional data frame and made the sampling approximately a fifty-to-fifty binomial split. For example, data frame option 1 has a total 2,037 Exited values. I picked all values of the exited class one, then filled the remaining selected total sample of 4,000 with exited class zero. The final value counts for option 1 are 1,963 for class zero and 1,963 for class one. For evaluation to select the strongest balance feature options, I fit and tested the model with stratified training and test sets with a 70 to 30 percent split.

Other feature transformations include converting categorical features into dummy variables in preparation for modeling as well as feature scaling. Feature scaling with both Standard Scalar and Min Max Scalar are utilized. The final data preparation step includes splitting the selected balance feature option data frames into train, validate, test sets with the target feature stratified. I reserved 10% of the data for validation, and I split the remaining 90% as 15% test and 85% training sets.

Modeling

I chose Random Forest Classifier as the first model since normalizing the data isn't required. This will be beneficial for some of the features where log transformations did not fully correct the skewness of that data. The Random Forest model is also an ensemble which accounts for overfitting and generally has high accuracy (Great Learning Team, 2023). I used this model to fit each of the data frames created for the Balance feature options for performance comparison. Once the optimal Balance feature option is selected, the additional models are fit with and without scaling for evaluation.

Evaluation

One of the most popular methods to assess the performance of classification models is a confusion matrix. A confusion matrix consists of four quadrants that reflect the amount of true positive, false positive, false negative, and true negative results. Evaluating and interpreting the results depends on the scenario. For example, to predict cancer results the desired outcome is to minimize false negatives to correctly identify all patients that have cancer. For detecting emails that are spam, the converse is true. This scenario requires minimizing false positives so that valid, important emails are not caught in a spam filter.

The classification model calculates the probability that a customer falls in one category or another. One way to set the optimal barrier between categories is the Receiver Operating Characteristic (ROC) Curve. To accomplish this, a graph line is plotted to show the true positive

rate against the false positive rate at each decision threshold. This leads to the evaluation metric called Area Under the ROC Curve (AUC). The validity of the predictions is determined by the height of the curve in conjunction with additional evaluation metrics. (Confusion, n.d.). I reviewed confusion matrices, classification reports (including precision, recall, f1-score, and accuracy scores), ROC curves, and ROC AUC scores to evaluate the models.

Interpreting Results of Balance Feature Options. In reviewing the evaluations metrics for the Random Forest Options without scaling, the strongest options that that meet the objectives of the bank client churn model are 1 and 5. I determined this since options 1 and 5 have the highest ROC AUC and Recall scores. Option 1 is showing a stronger recall score; however, I'll continue fitting the additional models with both options.

Table 1

Evaluation Metrics Summary of Balance Feature Options

Option	Precision	Recall	F1-score	Accuracy	AUC
one	.77	.78	.77	.77	.85
two	.76	.71	.74	.75	.82
three	.81	.73	.77	.78	.85
four	.78	.72	.75	.76	.84
five	.77	.74	.76	.76	.85

Interpreting Results of Additional Models. The evaluation metrics for additional models without scaling shows that Random Forest and AdaBoost are the strongest models for further evaluation as shown in Table 2. The other five models reflected scores between the upper 40s and lower 70s. I evaluated the models with Standard and Min Max scaling tools to determine optimal model selection. Overall, the confusion matrices for option 5 had stronger Recall scores and ROC AUC scores than option 1. This means that option 5 model was more effective at potentially overlooking clients that planned on leaving the bank.

Table 2

Top Evaluation Metrics Summary of Additional Models

Models	Precision	Recall	F1-score	Accuracy	ROC AUC	False Neg
1 Random Forest (RF)	.77	.72	.74	.75	.749	75
1 RF (Standard Scaler)	.78	.74	.76	.77	.768	69
1 RF (Min Max Scaler)	.78	.72	.75	.76	.762	73
1 AdaBoost	.77	.72	.74	.75	.751	74

Models	Precision	Recall	F1-score	Accuracy	ROC AUC	False Neg
5 Random Forest (RF)	.79	.83	.81	.80	.801	44
5 RF (Standard Scaler)	.80	.83	.81	.81	.809	46
5 RF (Min Max Scaler)	.81	.82	.81	.81	.813	47
5 AdaBoost	.77	.79	.78	.78	.775	55

Model Selection

All three Random Forest Classification models for option 5 have very close evaluation scores. However, I'm selecting the Random Forest model with Standard Scalar due to the strong recall score along with higher precision, accuracy, and ROC AUC over no scaling being utilized. When exposing the model to a small validation dataset, the model performed consistently with evaluation scores slightly below 80. For the selected model, the top five contributing features are Age_log (.23), Zero_bal (.13), EstimatedSalary (.12), CreditScore (.12), and NumberOfProducts_2 (.09).

Conclusion

The overall objective of this project is to identify as many clients as possible that are planning to leave the bank. Out of 400 validation clients, the selected model correctly identified 157 (.79 accuracy) that will exit if no action is taken to mend the relationship. Model improvement will be focused on minimizing the false negatives (42 clients) since these are missed opportunities to salvage client relationships.

More analysis will need to be completed to determine the actual impact and value the model may bring to the organization since the sampling was balanced to account for the uneven binomial variable, Exited. However, the model shows promise and should be implemented if the value surpasses the cost of the model implementation and maintenance.

Potential Model Enhancements

Although the model adds value to be deployed as is, there are a few more methods to explore that may enhance the model further. Hyperparameter tuning can be taken into consideration with adjusting parameters such as max leaf nodes and max depth. Since the sample size is small, a learning curve can be utilized to determine if it is worthwhile to increase the size of the dataset. It is notable that several models could be explored further with penalties or weights applied. One final adjustment to explore is managing the bimodal distribution of the Credit Score feature.

Ethical Implications

An ethical implication to consider when creating a churn prediction model is selection bias. It is essential to verify that the sampling utilized to create the model is representative of the entire client population. If not, the results may be biased (S., 2023). Churn models that are biased have the potential to cause disproportionate discrimination towards certain groups of clients. This leads to unfair banking practices which are heavily regulated. Once way to minimize sampling bias is applying stratified random sampling.

Once the model is in production, it's essential to be cognizant of classes or groups of clients getting more or less of a particular treatment than others because of the model. Analysis and monitoring will need to be continued after deployment to recognize trends in socioeconomic factors and demographics for unintentional bias in banking practices.

In addition, the data of bank clients is highly sensitive. Data protection regarding privacy must be top priority. Model success is highly dependent on the quality and quantity of data available. Often data is gathered without the explicit consent of the consumer. More regulations

have been implemented over the past several years due to data collection and privacy infringement. The client needs to have control over what data is being collected, how it is used, and how it is shared.

What I learned

I learned about the strengths and weaknesses of several supervised, classification models. I discovered which evaluation metrics are relevant to these types of models. More importantly, I was able to apply how to evaluate the models through understanding several scoring metrics and connect the relation to confusion matrices and ROC curves. I gained a better understanding of accounting for collinearity between features, balancing target features, splitting datasets with stratification, and how to prepare and research different feature options to optimize model performance.

References

Adam. (2018, October). Predicting Churn for Bank Customers, Version 1. Retrieved September 6, 2023, from https://www.kaggle.com/datasets/adammaus/predicting-churn-for-bank-customers

Akula, Gowtham. (2023, September 16). Unveiling the Future: Machine Learning's Power in Predicting Customer Churn in Subscription-Based Enterprises. *LinkedIn.* https://www.linkedin.com/pulse/unveiling-future-machine-learnings-power-predicting-customer-akula#:~:text=Ethical%20Considerations%20in%20Churn%20Prediction,groups%2C%20leading%20to%20unfair%20practices.

Avcontentteam. (2023, June 27). Classification vs. Clustering- Which One is Right for Your Data? *Analytics Vidhya.* https://www.analyticsvidhya.com/blog/2023/05/classification-vs-clustering/?utm_source=related_WP&utm_medium=https://www.analyticsvidhya.com/blog/2018/05/24-ultimate-data-science-projects-to-boost-your-knowledge-and-skills/

Confusion Matrix - Machine Learning Interview Questions. (n.d). *Algo Daily.* Retrieved September 9, 2023, from https://algodaily.com/lessons/ml-interview-questions/confusion-matrix#:~:text=A%20confusion%20matrix%20is%20a,each%20of%20the%20possible%20classes

Daivi. (2023, September 1). Top 5 Predictive Financial Modeling Project Ideas for Practice. *Project Pro*. https://www.projectpro.io/article/predictive-financial-modeling-projects/611

Great Learning Team. (2023, June 13). Random forest Algorithm in Machine Learning: An Overview. *Great Learning.* https://www.mygreatlearning.com/blog/random-forest-algorithm/

S., Dr. Nagaraj. (2023, February 8). Disentangling Customer Churn and its Challenges. *LinkedIn.* https://www.linkedin.com/pulse/disentangling-customer-churn-its-challenges-dr-nagaraj-s-/