Prediction of Likely Churning Bank Customers

Anusha Kondle
School of Engineering and Applied Sciences
State University of New York at Buffalo
Buffalo, New York, USA
anushako@buffalo.edu

Dennis M. Kyalo
School of Engineering and Applied Sciences
State University of New York at Buffalo
Buffalo, New York, USA
dkyalo@buffalo.edu

Muhammad Faique
School of Engineering and Applied Sciences
State University of New York at Buffalo
Buffalo, New York, USA
mfaique@buffalo.edu

Sol Jang
School of Engineering and Applied Sciences
State University of New York at Buffalo
Buffalo, New York, USA
soljang@buffalo.edu

Sungjoon Park
School of Engineering and Applied Sciences
State University of New York at Buffalo
Buffalo, New York, USA
spark55@buffalo.edu

Abstract— *Churning from bank percentage has increased drastically in previous years. With the rise of digitalization, especially in the financial sector, banks have to keep working on improving their services to keep their customers. It can be safe to assume that services might be the factors. But what about a person's financial situation or increasing inflation? In this paper, we are trying to find which customers are more likely to churn out and the main factors involved in churning customers from the bank.*

Keywords— churning, linear regression, collinearity, Education Level, age

I. INTRODUCTION

According to McKinsey's Global Banking Report published in 2019: "In the US, the churn rate had risen from 4.2% in 2013 to 5.6% in 2017". The same report also stated that: "Millennials and much younger people tend to churn significantly more than people did in the past."

The question is why people are leaving their banks more than ever? There could be multiple reasons. One among these can be the digital transformation of banks:

Gerrard Schmid, President, and CEO Diebold Nixdorf, wrote in an article for [1] that: "The financial industry is undergoing a sweeping transformation due largely to changing expectations and preferences among customers." Banking sectors have changed due to digital transformation over the years, the concept of a cashless transaction is raising among customers. With so many banking options and digitization, moving money and assets have become easy and it is not very surprising that customers are often changing banks when they see a better option that fits them more appropriately.

So which bank is checking most of the boxes for customers can be a factor? Another factor could be bank policies. For example, Bank of America (BofA) requires you to maintain at least $1500 to avoid penalty fees, contrasting to M&T bank, where there is no such obligation. People who spend most of their earning and students are more likely to choose M&T bank over BOFA.

There could be different internal and external factors; it can be bank policies, an individual's way of managing their finances, or individuals' salary or bank benefits and services that can promote their financial growth.

Our objective is to design a model that predicts which customer is more likely to churn out and which factor has a high impact on churn using banking data that could be replicable for data from any bank.

II. MOTIVATION

A bank’s success is measured by its growth rate which should exceed its churn rate. Customer acquisition costs are 7 times higher than retaining existing ones, which means losing customers can be very financially detrimental. A data driven approach can help banks to proactively take action to prevent attrition before it’s too late. Traditionally, the research in this area were focused on domain specific variables. We are inspired to investigate variables which are backed by data in addition to the domain knowledge and use some classification and regression models to effectively predict attrition.

III. ANALYSIS LOGISTICS

We summarize here the step-by-step approach of how we tackled the project. First, our initial challenge when we started the project was selecting a problem statement that was both stimulating and relevant to the course of study. After a wide search on some real-world problems where we could apply our learnings of statistical learning, we selected the problem of churning customers in the banks.

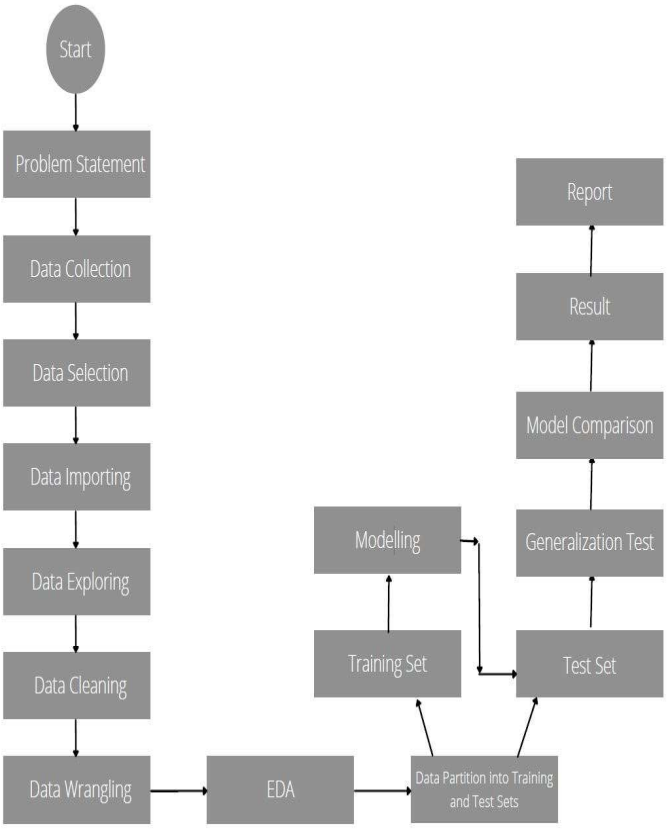


Fig. I Analysis Logistics

The first 6 steps are very important because of the 80/20 rule: 80% of the time spent by a data scientist is on gathering, cleansing, and storing the data.

The first step was to source the data which constitutes data collection and data selection. During this phase, we looked for multiple data sources all through the internet across a variety of repositories. Among all the data sources we collected, we selected the one we found in the Kaggle website due to its resourcefulness and popularity. We imported this data which was available in .csv format into the RStudio.

We then explored this data, in order to understand the data i.e., get an idea of what attributes are available, how many observations are collected and the quality of these observations. This is a very important step because this gives us information on what needs to be done in the next step, i.e., data cleaning such as looking for missing values and replace them with more meaningful data, identifying abnormal data or outliers etc. Our exploration revealed that the data was mostly clean without requiring many modifications in the original data.

As mentioned above, we then proceeded with the data cleaning. Although the data was clean, we removed 2 variables- "Naive_Bayes_Classifier_Attrition_Flag_Card_Category_Contacts_Count_12_mon_Dependent_count_Education_Level_Months_Inactive_12_mon_1" and "Naive_Bayes_Classifier_Attrition_Flag_Card_Category_Contacts_Count_12_mon_Dependent_count_Education_Level_Months_Inactive_12_mon_2" as suggested by Kaggle as they were not reasonable.

In the next phase, we did slight modification of the data. The dependent variable, Attrition_Flag, was a string type data, so we converted "Existing" to 1 and "Attrited" to 0, so that they can be numericals and be consistent with the rest of the data and it is easy for analysis since it now represents binomial data.

Here we just evaluated different variables - both continuous and categorical variables. This can give us some knowledge on these variables and their relevance to the dependent variable. We also checked the correlation between different variables. Multicollinearity would mean that more than one variable is explaining the same thing. So, we checked for multicollinearity and removed the variables that are unnecessary.

Subsequently, we split the data into 75% training set and 25% test set and applied the models - Logistic Regression, Boosting, KNN, Naïve Bayes, Bagging, Random Forest and CART, on the training set to evaluate which model is the best fit for the data.

Although the evaluation metrics might show us which model is the best fit, we want to understand it's performance on unseen data. A model is best generalizable when the test data accuracies are comparable to the 95% confidence interval of the training data accuracies. We compared the evaluation metrics of all the models for both training data set and test data set. We then captured the results in a table and provided our conclusions.

We collated all the data and codes and reported the analysis in this report in a well-structured manner to organize and represent our findings and results.

IV. DATA DESCRIPTION

A. Data Source

The data used in this study is customer data provided by an anonymous bank. This data is being applied to construct models predicting customer churn in various places, including "Kaggle." Even though the source is unclear, considering that this data set is a representative example of predicting customer churn by banks, it is not expected that using this data will cause major problems. Meanwhile, this data set in csv format, named "BankChurners," contains 10127 customer information, and customer information can be identified only by customer number.

B. Target Variable and Features

The target variable for this analysis is "Attrition_Flag," which is "Attrited" if a customer churned and "Existing" otherwise. We encode its observations as 0 for "Existing" and 1 for "Attrited" in order to apply various classifiers, including logistic regression. It appears that about 16% of customers have left the bank, and the rest of 10127 customers, about 84%, are still using the bank. We will split the data so that both types of customers are included in the training data set and the test data set in equal proportions.

In addition, there are 19 features in the customer data, 5 of which are categorical, and the other 14 are continuous. First of all, it is possible to roughly figure out how much influence each categorical feature will have on the decision of whether to leave the bank through the churn ratio by category of the categorical features.

Fig. II Comparing Two Distributions (Gender, Number of Dependents and Education Levels) by Attrition

Fig. II show each distribution of the respective category and by attrition. Here, a categorical feature with a large difference within a categorical feature can be highly expected to be a useful for classification training. According to these figures, it can be seen that the visually significant difference between two distributions by attrition is found only in 'Gender'. While there exists about 14.4%p gap between the churning share by male and female customers, the proportions of male and female customers who have not decided to churn are close each other. To the contrary, in the rest categorical features in Figure I and II, both two distributions of each categorical feature are almost the same each other.

While categorical variables are mainly related to customers' personal information, all continuous variables except for "Customer_Age" are related to customers' banking service usage history. Eight of these continuous features show a noticeable difference in distribution according to attrition or not. In Fig. III and IV, there are Eight boxplots of the features which show different distributions by the values of the target variable, "Attrition_Flag."

This study has not decided whether the important categorical and continuous variables revealed in the data will be included in the model. However, in the future, we will decide whether these variables are important based on domain knowledge and some proper statistical criteria and whether to include them as features in the training model of this study.

Fig. III Visually Notable Continuous Features

Fig. IV Visually Notable Continuous Features

V. Feature Selection

Multicollinearity shows the intercorrelation between variables in a dataset. To check for the presence of multicollinearity, we generated a correlation matrix that illustrates the correlation between the predictors.

TABLE I. CORRELATION MATRIX

Correlation Coefficients	***Credit_Limit***	***Avg_Open_To_Buy***	***Total_Trans_Amt***	***Total_Trans_Ct***
Credit_Limit	1.0000	0.9960	0.1717	0.0759
Avg_Open_To_Buy	0.9960	1.0000	0.1659	0.0709
Total_Trans_Amt	0.1717	0.1659	1.0000	0.8072
Total_Trans_Ct	0.0759	0.0709	0.8072	1.0000

From Table I above, we opted to highlight the strongly correlated variables. The analysis shows Credit_Limit and Avg_Open_To_Buy had a strong correlation of 0.9960, whereas Total_Trans_Amt and Total_Trans_Ct had a high correlation of 0.8072. As a result, the Avg_Open_To_Buy and Total_Trans_Ct variables were removed to reduce the element of multicollinearity.

The dataset contained many variables; thus, we excluded some of them based on their relevance in predicting the Attrition_flag, the dependent variable. To perform the elimination, we applied the recursive feature selection approach to the train set to determine the variable significance of the predictors. This strategy starts by fitting a model and then iteratively excludes the weakest features until a predetermined number of features that strongly predict the dependent variable is attained.

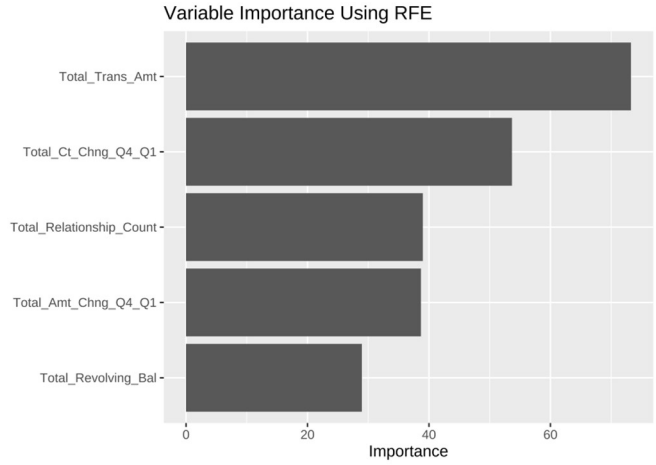


Fig. V Variable Importance Using RFE

The five best predictors in the dataset include Total Transaction Count for the last 12 months (Total_Trans_Amt), Change in Transaction Count (Q4 over Q1) (Total_Ct_Chng_Q4_Q1), Total number of products held by the customer (Total_Relationship_Count), Change in Transaction Amount (Q4 over Q1) (Total_Amt_Chng_Q4_Q1) and the Total Revolving Balance on the Credit Card

(Total_Revolving_Bal). Fig. V shows the variables' relative importance in predicting the Attrition_Flag (dependent) variable.

VI. PERFORMANCE COMPARISON

A. Comparison in terms of Generalizability

In the last step of this study, we compare the performance of several classifiers and finally suggest the best-performing model. However, in a situation where only feature selection for logistic regression is currently decided, this study verifies the absolute performance of this classifier and whether it can be generalized to data other than the BankChurner data used in the training data set. The elements of the confusion matrix are used as performance indicators.

To forecast the attrition_flag, we used the following classification models on the train set: Boosting, K-NN, Logistic Regression, Naïve Bayes, CART, Bagging, and Random Forest. Afterward, we compared their accuracies, sensitivity, and specificity.

TABLE II. CLASSIFICATION METRICS ON THE TRAINING SET

	Accuracy			***Sensitivity***	***Specificity***
		Lower	***Upper***		
Boosting	0.9516	0.9465	0.9563	0.9821	0.7920
K-NN	0.9138	0.9072	0.9200	0.9633	0.6552
Logistic Regression	0.8820	0.8746	0.8892	0.9818	0.3611
Naïve Bayes	0.9093	0.9026	0.9157	0.9837	0.5209
CART	0.8856	0.8782	0.8927	0.9540	0.5283
Bagging	0.9992	0.9983	0.9997	0.9995	0.9975
Random Forest	1.000	0.9995	1.000	1.000	1.000

Table II shows the classification metrics of each model as applied to the training data set, along with the 95% confidence interval of their accuracies. It is evident that Random Forest had the highest accuracy of 100%, while logistic regression had the least accuracy of 88.2%.

TABLE III. CLASSIFICATION METRICS ON THE TEST SET

	Accuracy	***Sensitivity***	***Specificity***
Boosting	0.9451	0.9774	0.7759
K-NN	0.9000	0.9515	0.6305
Logistic Regression	0.8862	0.9859	0.3645
Naïve Bayes	0.9127	0.9868	0.5246
CART	0.8807	0.9482	0.5271
Bagging	0.9320	0.9661	0.7537
Random Forest	0.9388	0.9689	0.7808

After that, we applied the algorithms on the unseen test set, and the above table shows the classification metrics. Comparing the test data accuracies to the 95% confidence interval of the training data accuracy, we can confidently conclude that Boosting, KNN, Logistic regression, Naïve Bayes, and CART models were generalizable as their test accuracies lay within the 95% confidence interval of the train accuracies. Bagging and Random Forest, on the other hand, were non-generalizable despite having high accuracies on the training dataset. Finally, we can deduce that Boosting had the highest test accuracy of 94.508% in predicting the Attrition_Flag and was also generalizable.

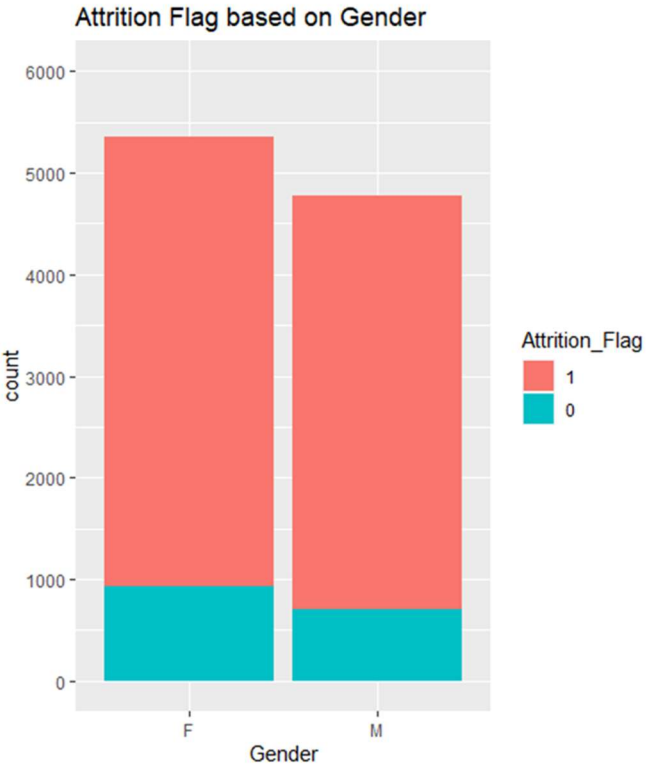
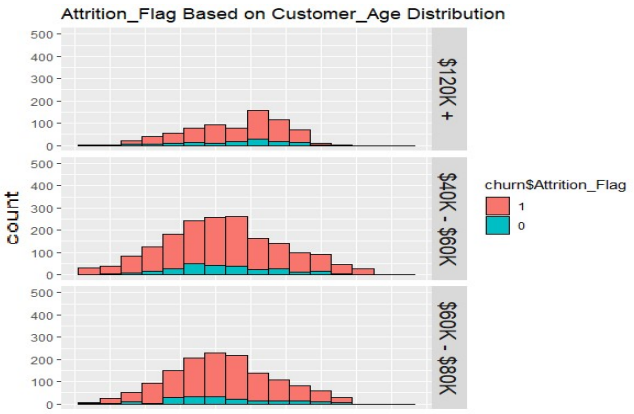


Fig. VI Gender Proportion by Bank Churning

From the Fig. VI, we can deduce that 83% of the females and 85% of the males are churning. Clearly, males are more likely to churn than females.

Based on the distribution in the Fig. VII, number of the customers aged between 40-50 is highest category compared to the other categories. Also, income wise we noticed that the customers who earn less than 40,000 dollars a year are more likely to churn than any other category. This might seem intuitive given the unstable financial condition of low income categories.



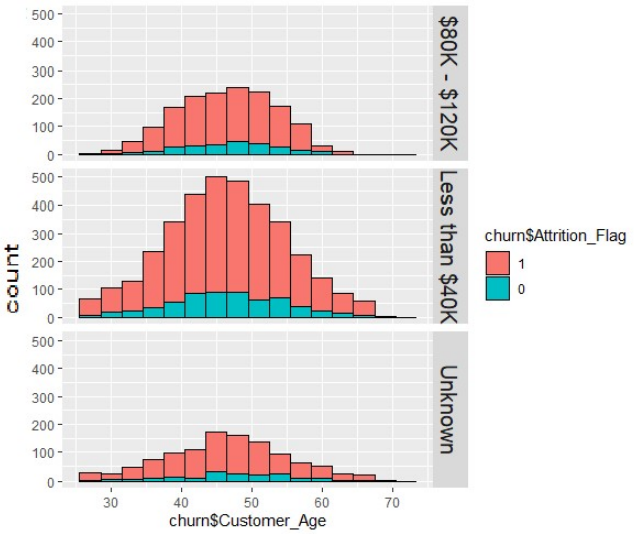


Fig. VII Age Distributions by Bank Churning

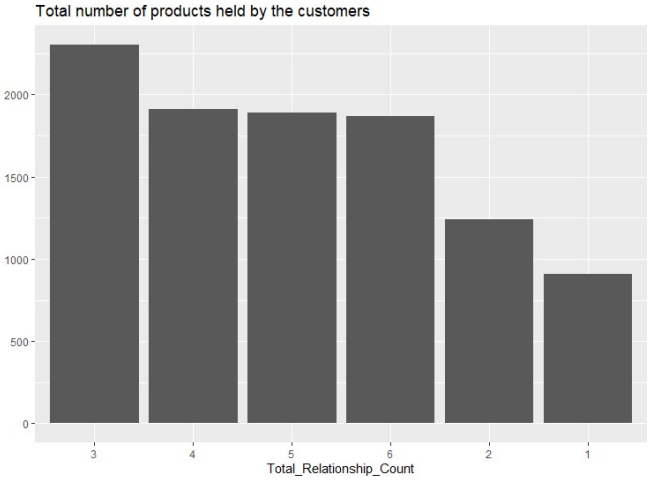


Fig. VIII Total Number of Products Held by the Customers

The number of customers holding three products are the highest and number of customers holding one product are the lowest. See Fig. VIII.

Let's look at the table in more detail. Bagging is an abbreviation for 'Bootstrap Aggregating'. In other words, the Weak Tree is trained with different data each time through sampling with replacement, and the results are derived through aggregating multiple Weak Trees. The reason for using out-of-bag concept when using Random Forest, which is a representative concept of bagging, is because of this random sampling. Thanks to random sampling and ensemble, both Bagging and Random Forest are able to show best performance. On the one hand, Boosting also use sampling with replacement. However, there are some changes. That is, it gives weight to Whole Origin Data. Weighting for specific data increases the probability that specific data is sampled.

In random forest and bagging, each weak tree is independently trained. It can learn in parallel, and the learning speed is much faster. On the other hand, boosting depends on each weak tree. The more well-learned the weak tree, the more weight it gives. So, Boosting makes model learning much better by weighting misclassified data to highlight difficult-to-classify cases. Therefore, boosting generally performs better than bagging. However, in our experiment, the results were not quite consistent with the theory. Insufficient number of data could be one possible reason for this.

Next, let's look at Logistic Regression which showed the worst performance in our experiment. In order for logistic regression to show higher performance, there are several prerequisites which must be followed. Binary logistic regression requires that the output should be binary. In other words, the result must be either 1 or 0. Of course, this assumption was not violated.

However, there was no clear guarantee for the following assumptions. First,it should contain only meaningful variables. Next, independent variables must be essentially independent of each other. That is, there should be little or no multicollinearity. Logistic regression should only be applied to large-sized samples.

We selected all variables carefully as much as possible considering their importance, but this process was not perfect. The multicollinearity problem was also not completely eliminated. Not only that, like the previous models, the data we used for the project is not a large-scale sample. In other words, there are many constraints for logistic regression to show good performance, but our dataset cannot satisfy all these consumptions. Thay's why logistic regression showed the worst performance between all these models.

B. Strength and weakness of this analysis

This algorithm works well for this data, but in the real world, the data may not be as ideal. This algorithm still needs to be tested and challenged.

VII. CONCLUSION

Due to the nature of the data, i.e., binomial trait (0 for existing and 1 for attrited), we initially assumed that logistic regression is the most suitable model for the initial analysis of the dataset. We considered statistical significance for our initial analysis to identify the features that can best explain the model and accurately predict the attrition of the customers. We also evaluated categorical and continuous variables to understand the relationships between different features and attrition. Our regression model gave us 13 statistically significant variables. The categorical variables, including 'Income category' and 'card category,' appeared to affect the attrition significantly. As per the 5% level of significane criterion, these variables are consistent with our belief(purely based on instinct) that these categorial variables should be directly correlated to customer attrition.

The other continuous variables, which were statistically significant in our logistic regression, are also consistent with the result of RFE since most of these features, such as Total_Trans_Amt,Total_Ct_Chng_Q4_Q1,Total_Relationship _Count, Total_Revolving_Bal, and Credit_Limit, are among the top six features as per RFE. We may say that Total Transaction Count for the last 12 months (Total_Trans_Amt), Change in Transaction Count (Q4 over Q1)

(Total_Ct_Chng_Q4_Q1), Total number of products held by the customer (Total_Relationship_Count), Change in Transaction Amount (Q4 over Q1) (Total_Amt_Chng_Q4_Q1) and the Total Revolving Balance on the Credit Card (Total_Revolving_Bal) have the highest impact on the churning rates of a customer.

Our prediction model produced an accuracy of 88.20% on the training data set, which is a fairly good prediction. An accuracy of 88.62% on the test set shows that the model is very good for generalizability since both training and test accuracy are similar, and the test accuracy lies in the 95% confidence interval.

Although this model seems to be very generalizable, we were curious to see how the other models such as Boosting, KNN, Naïve Bayes, Bagging, Random Forest and CART would perform. We repeated the process for all these other models and we found out that Random Forest had the highest accuracy of 100%, while logistic regression had the least accuracy of 88.2%. So every other model performed even better than logistic regression. However, when we checked the generalizability using test set data, we realized that even though random forest has highest accuracy, Boosting was more successful both in terms of accuracy of training set and generalizability of the test set.

VIII. BROADER IMPACT

We designed a model that predicts which customer is more likely to churn out and which factor has a high impact on churn using banking data that could be replicable for data from any bank. Therefore, we believe that there is a huge impact of this analysis on the banking sector and this project gives important insights that can be implemented by a bank.

These very useful insights can encourage banks to open a new division of data analytics that is purely dedicated to data based projects. Observations from this project will help different banks to define new bank policies that will further enforce prevention of more such problems. While Data Analytics projects are important to save a lot of money for banks, investment on these projects are a huge investment in itself. Creating a model using similar approach will help banks in realizing financial impact.

REFERENCES

[1] https://internationalbanker.com/

[2] https://medium.com/@noah.fintech/creating-a-banking-customer-churn-model-1a2d0850f071

[3] https://www.qualtrics.com/blog/customer-churn-banking/

[4] https://www.acuitykp.com/blog/a-data-driven-approach-to-reduce-churn-in-financial-institutions/

[5] https://www.researchgate.net/publication/342424673_Prediction_of_Customer_Churn_in_Banking_Industry