Journal of Applied Finance & Banking, Vol. 14, No. 4, 2024, 65-109
ISSN: 1792-6580 (print version), 1792-6599(online)
https://doi.org/10.47260/jafb/1445
Scientific Press International Limited

Bank Customer Churn Prediction Using Machine Learning Framework

Rasha Ashraf [1]

Abstract

Using real customer data from a large community bank in the South of the US, this paper analyzes the customer churn prediction problem by constructing and comparing ten machine learning classification models with five sample techniques. Our results show that Random Forest, XG Boost, AdaBoost, and Bagging Meta classifiers dominate others in terms of overall accuracy, F-score, and AUC curve for the test observations. For the four classifiers, the overall accuracy ranges from 87% to 96% across five different sampling methods explored, while the AUC values range between 0.9 to 0.93. Considering overall accuracy and F-Score, AdaBoost with original and MTDF sampling technique dominates others; however, considering the AUC measure, XG Boost and Random Forest perform similarly to AdaBoost, which slightly dominate Bagging Meta across all sampling techniques; although the performance measures for these four classifiers are comparable across all sampling techniques. The paper further presents important features of customer churn behavior as predicted by the model. The diagnostic analysis also provides an insightful comparison between churned and non-churned customers.

JEL classification numbers: C0, C5, C8, G21.
Keywords: Machine learning, Big data, Sampling techniques, Customer churn, Customer retention, Financial services, Community bank.

[1] J. Mack Robinson College of Business, Georgia State University, Atlanta, USA.

Article Info: Received: May 15, 2024. Revised: May 30, 2024.
Published online: June 4, 2024.

1. Introduction

The financial services industry is facing increased competition due to rapid change and the need to adapt to the technology and the penetration of FinTech companies into traditional banking services.[2] With the fierce competition in the financial services industry customer retention or churn is one of the challenging problems facing the industry. Acquiring new customers is five to six times costlier than retaining customers (Athanassopoulos, 2000; Bhattacharya, 1998; Colgate and Danaher, 2000; Rasmusson, 1999). Creating a long-term customer base would be important for service industries as loyal customers are less susceptible to aggressive marketing efforts from competitors, use a multitude of products, less costly to serve, and help spread the goodwill of the company (Ganesh et al., 2000; Hwang et al., 2004; Verbeke et al., 2011). It is well established that focusing on customer satisfaction and service quality helps create a long-term customer base and is one of the most effective ways to build a competitive position in the service industry (Lewis 1993). Vast literature documents customer satisfaction and building long-term relationships are critical for service industries.[3] Relationship marketing focuses on the increased role of quality and satisfaction by integrating marketing, quality, and customer service.[4] As much of the marketing efforts are dedicated to retaining existing customers and creating a long-term customer base, understanding and detecting customers who are likely to leave the company is crucial so that the company can reach out to those at-risk customers and build strategies on minimizing the risk of losing them or reducing the overall churn rate. Companies need reliable and comprehensible churn prediction models to identify the customers who are likely to churn and understand what causes them to leave and how to prevent it. With the explosion of big-data analytics and machine learning models, industries in the service sector are able to deep dive into the customer churn problem and come up with sophisticated prediction models to identify customers who are likely to churn with reliable accuracy and build strategies to mitigate the loss of existing customers and thereby improving on developing long-term customer base. Verbeke et al. (2011) and Amin et al. (2017) provide an extended overview of the literature on the use of data mining and machine learning models in customer churn prediction modeling. Most of the literature is concentrated on the churn problem in the telecommunication industry applying machine learning techniques and evidence of the problem in the financial services industry is scant, especially for US banks, mainly due to a lack of real reliable data that can be used for building robust models and testing and validating them.

The objective of this research is to analyze customer churn or retention problems

---

[2] Erel and Liebersohn, 2020 document that FinTech is disproportionately used in areas with scarcity of banking presence and conclude that FinTech mostly expands the overall supply of financial services, rather than redistributing it. Buchak et al. (2018) show that traditional banks contracted in markets where they faced more regulatory constraints. Gopal and Schnabl (2022) argue that finance companies and FinTech lenders are major suppliers of credit to small businesses and played an important role in the recovery from the 2008 financial crisis.

[3] Colgate et al., 1996; Ganesh et al., 2000; Paulin et al., 1998; Reichheld, 1996; Stum & Thiry, 1991; Zeithaml et al., 1996.

[4] Berry, 1995; Gummesson, 1993; Christopher et al., 1991; Athanassopoulos, 2000.

using data provided by a large community bank in the South of the US and develop machine learning models to address the problem. The Federal Reserve defines a community bank organization as those with less than 10 billion in total assets. Using customer-level cross-sectional data of 47,386 observations for the period of September 2021 to September 2022, the churn model identifies customers who are at risk of leaving the business and send an alert so that bank management can proactively reach out and work with the customer to address the need and possibly refrain from exiting the business. The churn prediction problem belongs to the formulation of traditional classification models to predict categorical variables into the churn and un-churn types. With the explosion of machine learning models, the customer churn prediction problem is getting traction in finding novel ways to address the problem and improving the accuracy of the prediction of customer churn. Mainly the aim of the machine learning model is to identify customers who are likely to leave the bank so the bank can reach out to the customers to address their concerns. Although the problem seems very straight forward, however, due to the imbalanced nature of the data, which arises as the number of customers leaving the bank is relatively much lower than the number of customers staying with the bank, finding a robust model that can identify customers who are at risk of leaving the bank reliably is relatively difficult. Therefore, even if the overall accuracy of the model might be high, the false negative alerts are relatively high with the imbalanced churn data using traditional classification models, as the model classifies customers who are likely to churn as a non-churn category. To address this issue, this research will explore a plethora of machine learning models with various sampling techniques to address the imbalanced data concern and compare the performance and accuracy of the models as measured by overall accuracy, recall, precision, specificity, false positive rate, and false negative rate, F1-score, and ROC measures and identifies models that provide the best performance for the out-of-sample data with regard to the churn problem.

To analyze the churn prediction model, we apply traditional classification models and several advanced machine learning models; such as - Naïve Bayes, Support Vector Machine (SVM), Logistic Regression, Decision Tree, Random Forest, Artificial Neural Network (ANN), and ensemble bagging and boosting techniques such as - Bagging Meta Estimator, AdaBoost, XGBoost, and Gradient Boosting. To address the imbalanced nature of the data, in addition to applying more commonly used sampling techniques such as under-sampling and over-sampling methods, we also apply more recent sampling techniques that are getting attention in the machine learning literature – such as SMOTE and MTDF methods.

Our results show that Random Forest, XG Boost, AdaBoost, and Bagging Meta classifiers dominate others in terms of overall accuracy, F-score, and AUC curve for the test observations. These four classifiers show the highest level of accuracy for the original sample and across the four different sampling methods presented, ranging from 87% to 96% of overall accuracy. The AUC values range between 0.9 to 0.93 for the four classification models. While considering overall accuracy and F1-Score, AdaBoost with original and MTDF sampling technique dominates others; however, considering AUC measure, XG Boost and Random Forest performs similarly with AdaBoost, which slightly dominate Bagging Meta across all

sampling techniques; although the performance measures for these four classifiers are comparable across all sampling techniques. We also provide an out-of-sample comparison of model performance and present and discuss important features selected by the models. Prior to the construction of the machine learning models we perform data cleaning and exploration to make sure the data is ready to provide reliable results. We perform feature selection using Pearson Correlation and Variance Inflation Factor (VIF) which reduced the number of features from 102 to 84, dropping 18 features. We created additional 55 features for age groups, business types, balance deciles, banking relations quartiles, etc, which contributed to total features of 139. We also present a diagnostic analysis that provides an insightful comparison between churned and non-churned customers.

Analysis of the churn problem using machine learning models has been explored extensively in the telecommunication industry.[5] The research in the banking industry is limited and there is no study based on banks in the US, perhaps due to the limitation of data availability. Our study, therefore, brings novel insights by analyzing data from a large US Community Bank in the South and thus allowing us to investigate the churn problem and understand what causes customers to leave the bank and what banks can do to address the issues. The machine learning prediction models will identify customers who are at-risk of leaving the bank, thereby allowing the bank to proactively reach out to address concerns and provide services that would reduce such risks and perhaps help the bank to retain customers.

2. Literature

Companies can improve profit significantly if they are able to prevent customers from not leaving (Reichheld 1996). It is important to establish mechanisms to identify at-risk customers who are likely to leave and take strategic initiatives to respond promptly and take measures to prevent such turnover. With the explosion of data and advanced machine learning techniques companies can develop sophisticated models that will predict with reliable accuracy customer churn behavior.

Among all the machine learning techniques used in the literature, none dominates consistently with regard to the performance of churn model predictions. Studies analyze and compare the effectiveness of prediction results using the techniques: Naïve Bayes, Logistic Regression, Neural Network, Support Vector Machines (SVM), Decision Trees (DT), Random Forest, other ensemble boosting methods such as AdaBoost, Gradient Boosted Machine Tree, Extreme Gradient Boosting (XGBoost). Mozer et al. (2000) and Hwang et al. (2004) apply logistic regression and neural networks to predict churn. While Mozer et al. (2000) find neural networks to perform better, however, Hwang et al. (2004) find in favor of logistic regression. Using major Belgian financial services company data, Lariviere and Van den Poel (2005) examine product purchase and cancellation and profitability outcomes and show that random forest provides a better fit for the estimation and validation sample compared to linear regression and logistic regression models.

[5] Most of the studies come from Computer Science field.

There is a large body of literature investigating churn problems in the emerging market of the telecommunication industry. Using Telkom Indonesia customers data Hanif (2019) documents that the XGBoost algorithm provides better prediction than the Logistic Regression model based on prediction accuracy, specificity, sensitivity, and ROC curve. Ahmad et al. (2019) explore Decision Tree, Random Forest, Gradient Boosting Algorithm, and XG Boost using SyriaTel telecommunication customer data and show that the XGBoost provides the best prediction outcomes. On a public data set in Greece, Vafeiadis et al. (2015) find that SVM-POLY using AdaBoost performed best compared to Artificial Neural Networks (ANN), Support Vector Machines (SVM), Decision Trees (DT), Naïve Bayes, and Logistic Regression.[6] Using a plethora of machine learning models, a researcher can explore what methods work best for the prediction of the churn problem. Moreover, machine learning models address the overfitting problem while minimizing the bias, so the bias-variance tradeoff using the cross-validation techniques assures the performance of the model for out-of-sample data.

One of the main challenges to predicting customer churn arises due to the imbalanced nature of the data where the number of churned customers is much lower than the non-churn category, which causes the false negative rates to be high, that is identifying customers who are likely to leave the bank as a non-churn category. Most commonly sampling techniques used to handle the imbalanced data are under (over) sampling methods, in which the majority (minority) class is eliminated (duplicated) to balance the distribution in the dataset. Using six different European business data, Burez and Van den Poel (2009) show that under-sampling provides improved accuracy, especially when evaluated with AUC. However, the under-sampling technique has potential drawbacks as it discards important data which can lead to low model performance. On the other hand, the over-sampling method can take more time to train the classifier and increase the likelihood of overfitting. Advanced sampling techniques that are getting traction are synthetic minority oversampling technique (SMOTE) (Chawla et al., 2002), mega trend diffusion function (MTDF) (Li et al., 2007), adaptive synthetic sampling approach for imbalanced learning (ADASYN) (He et al. 2008).[7]

We apply various machine learning models and sampling techniques used in the literature to address the churn problem for our bank customer data. To the best of our knowledge, our study is the first to analyze the churn problem on real-life bank customer data in the US.

[6] See Wu (2022) for a detailed literature review on the comparison of the performance of various machine learning models to address the customer churn problem.

[7] See Adnan et al. (2016) for comparisons among six sampling techniques: mega trend diffusion function (MTDF), synthetic minority oversampling technique (SMOTE), adaptive synthetic sampling approach, couples top-N reverse k-nearest neighbor, majority weighted minority oversampling technique, and immune centroids oversampling technique.

3. Data

The data comes from a large community bank in the South of the US. The data has information on bank customer attributes along with the information on banking services that they avail, such as, product information, transaction data, account information, and customer activity with banking services. The data is used to construct machine learning models to address customer retention and churn problem to predict when a customer is at risk of closing an account or ending the banking relationship.

The transaction and account level data are for 13 months period between September 2021 to September 2022. Using the customer ID, which is the unique identifier for each customer, we construct customer-level cross-sectional data from the time series data. The variables age, demography, region, branch location, overdraft limit, etc. are at the customer level. Variables that change over time, such as balance, credit and debit transactions, loan amount, overdraft charge, fees, etc. are converted to customer level taking mean (for balance, credit and debit transactions etc.) or max (overdraft charge, fees and others). The customer-level cross-sectional data has 47,386 observations. All variable definitions are provided in Appendix A1.

A customer is identified as "churned" if the customer has stopped banking relationships during the 13 months of the data period and zero otherwise. For the entire sample, 10.32% of the customers have churned. Table 1, Panel A provides the descriptive statistics of the customer level attributes and Panel B provides account and transaction level attributes. Panel A shows that 65.75% of the customers live in the same zip code as the branch, 20.44% are business customers, and the average age of the customers is 47.83 with 9.12 years of banking relationships. The age demography portrays that 23.42% are Baby Boomers, 18.29% are Gen X, and 20.43% are Millennials. With regards to region, 40.48% lives in Region 1, 49.99% are from Region 2, and the remaining in Region 3. There is a large variability in customer relations with the bank. On average customer has 1.1 checking accounts with a maximum of 46 accounts. All other accounts also show similar variability. About 89.68% of customers have an active banking relationship, with 3.91% inactive, and 5.12% dormant (no contact with the customer for a long period of time). For different banking relations, such as *Account Analysis*, *RDC*, *ACH*, *Wire Transmit Setup*, etc. most show less than 1% of the customers have these relationships. *Overall Banking Relations* is a dummy variable equaling 1 if the customer is enrolled in at least one of the banking relations listed as *Account Analysis*, *RDC*, *ACH*, *Positive Pay*, *Wire*, *Wealth Management*, *Mortgage*, or *Credit Card* or has account related to checking, saving, debit card, loan, time deposit, or safety deposit account; and 0 otherwise. On average customer has 2 banking relations with a maximum of 8 and a minimum of 1 relation.

Table 1: Descriptive Statistics

Panel A: Customer Level Variables					
	Mean	Median	Max	Min	Standard Deviation
Customer Attributes					
Churn %	10.32				
Near Branch (lives in the same zip code) %	65.75				
Business Customer %	20.44				
Age	47.83	49.00	112.00	0.00	21.04
Customer Number of Years	9.12	6.00	45.00	0.00	9.19
Generation (%)					
GI	0.28				
Silent	7.00				
Boomer	23.42				
Gen X	18.29				
Millennial	20.43				
Gen Z	9.95				
Regions (%)					
Region 1 %	40.48				
Region 2 %	49.99				
Region 3%	10.11				
% of Customer with Accounts					
Checking Accounts %	92.83				
Debit Cards %	63.86				
Savings Accounts %	17.91				
Time Deposits %	3.99				
Safety Deposit Boxes %	3.65				
Loans %	8.50				
Number of Accounts the Customer holds with the bank (#)					
Checking Accounts	1.10	1.00	46.00	0.00	0.71
Debit Cards	0.73	1.00	23.00	0.00	0.68
Savings Accounts	0.19	0.00	23.00	0.00	0.42
Time Deposits	0.07	0.00	37.00	0.00	0.58
Safety Deposit Boxes	0.04	0.00	5.00	0.00	0.21
Loans	0.12	0.00	74.00	0.00	0.62
Customer Accounts Attributes ($)					
Overdraft Limit	526.28	500	1,500	0	487.90
Credit Card Limit	109.89	0.00	150,000	0.00	1,935
Wealth Management Market Value	2,397	0.00	7,796,283	0.00	66,964
Last Mortgage Loan Amount	3,990	0.00	1,575,000		36,919
Total Balance of Loans	16,831	0.00	11,640,256	0.00	178,831
Customer Loan Delinquency					
Number of Days Past Due Loans	1.47	0.00	999	0.00	34.63
Number of Times Late at least 30 Days	0.25	0.00	242	0.00	3.51
Number of Times Late at least 60 Days	0.10	0.00	143	0.00	1.91
Number of Times Late at least 90 Days	0.06	0.00	109	0.00	1.48

Account Status					
Active %	89.68	100	100	0.00	29.57
Dormant % (no contact with the customer for a long period)	5.12	0.00	100	0.00	21.98
Escheated %	0.06	0.00	100	0.00	2.45
Frozen % (a temporary block is placed)	0.00	0.00	0.00	0.00	0.00
Inactive % (no activity, shorter than dormant)	3.91	0.00	100	0.00	18.15
New %	0.97	0.00	100	0.00	9.64
Limited % (more restrictive than Frozen)	0.21	0.00	100	0.00	4.38
To be Closed %	0.05	0.00	100	0.00	2.04
Banking Relations					
Account Analysis %	1.21				
RDC (Remote Deposit Capture) %	0.76				
ACH (Automatic Clearing House) Originator %	0.57				
Positive Pay (Fraud Prevention Setup) %	0.03				
Wire Transmit Setup %	0.32				
Wealth Management %	0.91				
Mortgage Customer %	1.65				
Hold Credit Card %	0.99				
Overall Banking Relations	1.97	2.00	8.00	1.00	0.796
Panel B: Account and Transaction Level Variables					
Variables	Mean	Median	Max	Min	Standard Deviation
Fees Charged to Customers					
Overdraft fees YTD $_t$	84.32	0.00	23808	0.00	457.49
Return fees YTD $_t$	10.32	0.00	22880	0.00	139.57
Overdraft Charged QTD $_t$	0.765	0.00	49	0.00	2.526
Overdraft Waived QTD $_t$	1.022	0.00	71	0.00	3.032
Service Charges $_t$	3.068	0.00	4800	0.00	53.03
Overdraft and Return Fees $_t$	51.55	0.00	12896	0.00	231.93
Transaction Fees $_t$	10.21	0.00	254102	0.00	1174.11
Total Fees Charged $_t$	63.97	0.00	254139	0.00	1198.65
Account Balance					
Balance_t	35,905	1,628	99,539,173	-33,014	530,364
$\Delta\text{Balance}_{t-1}$	21.56	0.161	213,077	-1.00	1150.28
$\Delta\text{Balance}_{t-2}$	27.21	0.292	212,230	-1.00	1156.49
Credit and Debit Transactions					
$\$CT_t$	18,624	1,183	57,456,058		452,119
$\Delta\$CT_{t-1}$	176.97	0.0981	2,732,308	-1.00	13673
$\Delta\$CT_{t-2}$	132.50	0.114	43,305	-1.00	6281.41
$\#CT_t$	3.82	1.69	1342.61	0	13.94
$\Delta\#CT_{t-1}$	0.156	0.077	48.25	-1.0	0.381
$\Delta\#CT_{t-2}$	0.185	0.077	56.61	-1.0	0.451

$\$DT_t$	18,263	1,136	57,405,837	0.00	455,909
$\Delta\$DT_{t-1}$	196.38	0.084	38,625	-0.982	35401.26
$\Delta\$DT_{t-2}$	43.38	0.105	476,194	-0.989	3145.22
$\#DT_t$	17.55	6.69	2770.77	0.00	30.03
$\Delta\#DT_{t-1}$	0.211	0.077	57.50	-1.00	0.849
$\Delta\#DT_{t-2}$	0.268	0.077	47.00	-1.00	0.927
OLB Activity Last 90 Days					
Money Management 90 Day Active (%)	3.40	0.00	100	0.00	13.28
SMS 90 Day Active (%)	10.71	0.00	100	0.0	30.22
App 90 Day Active (%)	31.99	0.00	100	0.00	45.16
Tablet 90 Day Active (%)	1.16	0.00	100	0.00	9.72
VRU 90 Days (%)	3.94	0.00	100	0.00	18.10
OLB 90 Day Active (%)	41.12	0.00	100	0.00	48.12
Deposits and Transactions Last 3-months					
Deposits Count (#)	3.03	0.667	4178.49	0.00	23.30
Deposits ($)	21,325.75	400.00	113,238,523	0.00	561,825.89
Mobile Deposits Count (#)	0.277	0.00	156.18	0.00	2.00
Mobile Deposits ($)	153.82	0.00	163,908	0.00	1,708.09
ACH Deposits Count (#)	18,269	907	46,970,702	0.00	324,259
ACH Deposits ($)	18,269	908	46,970,702	0.00	324,259
POS Debit Count (#)	24.29	2.00	598.25	0.00	42.37
POS Debit ($)	1032.28	78.86	24552	0.00	1940.36
Check Card Transaction Count (#)	23.21	2.61	598.00	0.00	41.30
Check Card Transactions ($)	1225.32	124.46	162,342	0.00	2970.81
RDC Deposits Count (#)	0.267	0.00	982.46	0.00	7.925
RDC Deposits ($)	5,282	0.00	42,625,584	0.00	244,481
Time Deposit Balance	2,662	0.00	2,950,746	0.00	33,097
Interest Paid					
Interest Paid YTD ($)	14.96	0.00	82,497	0.00	454.37
Interest Accrued but Not Paid	0.35	0.00	436.56	0.00	4.30
Interest Paid Last 3-months	10.10	0.00	67223.50	0.00	343.27
Interest Rate (%)	0.01	0.00	0.75	0.00	0.03

The table provides descriptive statistics of the sample customer data that comes from a large community bank in the US for the period September 2021 to September 2022. The number of observations is 47,386, which is the number of customers in the bank for the sample period. Panel A provides descriptive statistics of customer level variables and Panel B provides descriptive statistics of account and transaction level variables which are obtained by taking mean or maximum values over the 13 months period depending on the variable under consideration.

Panel B provides descriptive statistics of account and transaction level variables that vary over time. These variables are constructed at the cross-sectional customer level by taking the mean or maximum of the attributes (shown in Appendix A1 Panel B). *Overdraft fees YTD* has a mean of 84.32, a median of 0, a standard deviation of 457.49, and a maximum of 23,808. All other variables show similar large variability. For example, *Balance* has a mean of about $36K, with a standard deviation of $530K and a maximum $99 million. The dollar amount of credit transactions $\$CT_t$ has a mean $18K, standard deviation $452K, maximum $57 million. *Deposits* shows a mean of $21K, with a standard deviation of $562K and a maximum $113 million. All other account and transaction level variables show similar large variability.

3.1 Feature Selection

The preliminary feature selection is done by looking into the correlation of variables. We construct pairwise correlation of variables that are within the same category, such as balance, prior month balance, prior two-months balance and if the correlation of variables is close to 0.75 we include only one of the predictors. Table 2, Panel A shows $Balance_t$ is highly correlated with $Balance_{t-1}$, $Balance_{t-2}$, *Average Balance Past 12 months*, and *Average Balance Past 3 months*. It has lower insignificant correlation with the changes in balance measures $\Delta Balance_{t-1}$ and $\Delta Balance_{t-2}$; however, the latter two are highly correlated. So, we include Balance_t and $\Delta\text{Balance}_{t-1}$ in the model to predict the churn rate.

The Panel B shows that credit transaction variable $\$CT_t$ is highly correlated with $\$CT_{t-1}$, $\$CT_{t-2}$. It has lower insignificant correlation with the changes in balance measures $\Delta\$CT_{t-1}$ and $\Delta\$CT_{t-2}$. So, we select $\$CT_t$, $\Delta\$CT_{t-1}$ and $\Delta\$CT_{t-2}$ as features for churn model prediction. Similarly, untabulated results show that number of credit transactions $\#CT_t$ is highly correlated with $\#CT_{t-1}$, $\#CT_{t-2}$ and it has lower insignificant correlation with the changes in balance measures $\Delta\#CT_{t-1}$ and $\Delta\#CT_{t-2}$. So, we include $\#CT_t$, $\Delta\#CT_{t-1}$, and $\Delta\#CT_{t-2}$ in the model to predict the churn rate. Similarly, from debit transaction analysis we include $\$DT_t$, $\Delta\$DT_{t-1}$, and $\Delta\$DT_{t-2}$ and from the analysis for number of debit transactions, we include $\#DT_t$ and $\Delta\#DT_{t-1}$, and $\Delta\#DT_{t-2}$ in the model to predict the churn rate.

Table 2: Correlation of Features

Panel A: Correlation of Balance Accounts							
	$Balance_t$	$Balance_{t-1}$	$Balance_{t-2}$	$\Delta Balance_{t-1}$	$\Delta Balance_{t-2}$	Average Balance Past 12 months	Average Balance Past 3 months
$Balance_t$	1.00						
$Balance_{t-1}$	0.934 (0.00)	1.00					
$Balance_{t-2}$	0.912 (0.00)	0.932 (0.00)	1.00				
$\Delta Balance_{t-1}$	0.001 (0.59)	-0.0003 (0.80)	-0.0003 (0.82)	1.00			
$\Delta Balance_{t-2}$	0.001 (0.56)	-0.0002 (0.89)	-0.0004 (0.75)	0.904 (0.00)	1.00		
Average Balance Past 12 months	0.945 (0.00)	0.948 (0.00)	0.947 (0.00)	0.000 (0.99)	0.0001 (0.98)	1.00	
Average Balance Past 3 months	0.973 (0.00)	0.971 (0.00)	0.953 (0.00)	0.0001 (0.97)	0.0002 (0.91)	0.973 (0.00)	1.00

Panel B: Correlation of $Credit Transaction ($CT)						
	$\$CT_t$	$\#CT_t$	$\$CT_{t-1}$	$\$CT_{t-2}$	$\Delta\$CT_{t-1}$	$\Delta\$CT_{t-2}$
$\$CT_t$	1.00					
$\#CT_t$	0.491 (0.00)	1.00				
$\$CT_{t-1}$	0.926 (0.00)	0.487 (0.00)	1.00			
$\$CT_{t-2}$	0.912 (0.00)	0.488 (0.00)	0.921 (0.00)	1.00		
$\Delta\$CT_{t-1}$	0.002 (0.10)	-0.001 (0.93)	-0.001 (0.93)	-0.000 (0.99)	1.00	
$\Delta\$CT_{t-2}$	0.001 (0.29)	-0.001 (0.98)	-0.001 (0.99)	-0.001 (0.88)	0.0128 (0.00)	1.00

The table provides correlation of features related to balance accounts. All variable definitions are provided in Appendix A1.

We perform similar exercise for all other variables within the same category to see the correlation and following are the highlights of the analysis. Among *"Deposits and Transactions Last 3-months"* variables, correlation among variables *Deposits* and *Deposits Count* is 0.82, *POS Debits* and *POS Debit Count* correlation is 0.8, and *RDC Deposits* and *RDC Deposits Count* is 0.77. Among *Customer Loan Delinquency* variables, correlation among *Number of Times Late at least 60 Days* and *Number of Times Late at least 90 Days* is very high i.e., 0.96 and correlation among *Last Mortgage Loan Amount* and *Mortgage Customer* is 0.85.

Table 3: Multicollinearity Check Using VIF (Variance Inflation Factor) Scores

VIF Scores after removing features	
Variable	**VIF**
Last Mortgage Loan Amount	3.538153
Mortgage Customer	3.288177
Deposits	3.288177
Deposits Count	3.084946
POS Debits Count	2.821267
POS Debits	2.799281
RDC Deposits	2.783408
RDC Deposits Count	2.447022
Number_of_Times_Late_90_Days_Loans	2.265269
Number_of_Times_Late_30_Days_Loans	2.265247
Status Code A	2.159148
Status Code D	2.120654
Credit Dollar Amt of Transactions	1.600062

The table presents VIF analysis after removing features with highest VIF scores within each category of highly correlated features.

For the above mentioned correlated variables (with correlation over 0.75) we perform VIF analysis to make the final selection of which variable among each category to drop. Exploring Variance Inflation Factor (VIF) feature elimination method allows us to further identify features that are highly correlated. VIF measures the extent to which an independent variable is explained by other independent variables in the data. For each independent variable, VIF score is obtained by regressing the variable using OLS on all other features.[8] Literature (see Cheng et al., 2022 and Gomez et al., 2020) suggests that VIF measures over 10 is considered to be presence of multicollinearity. Machine learning community suggests 5-10 as the cutoff point.[9] Using unablated results, we drop one variable from each category whose VIF value is highest and greater than 5, i.e., Debit Dollar Amt of Transaction (27.061606) and Number of Times Late 60 Days Loans (25.456273). After removing these variables, we perform VIF analysis again and find that all variables VIF values are below 5 shown in Table 3.

Finally, before feeding into our models, all variables are standardized by min-max standardization method, where the minimum value of the variable is converted to 0 and maximum to 1 and all others take the value between 0 and 1.

[8] See Gomez et al. (2020) for detailed overview of the VIF process.

[9] See Analytics Vidhya and Analytics Explorer by S&P Global

3.2 Data Exploration and Churn Rate

Before constructing the machine learning models we provide a diagnostic analysis of the data to obtain insights from the trends related to customer churn. Figure 1A shows churn rate across three regions. Almost 50% of the customers are from Region 2, with 40% living in Region 1, and 10% are from Region 3. However, Region 3 shows much higher churn rate (11.8%) with Region 1&2 churn rate being at 10% (10.35%). Figure 1B presents churn rates across bank branches and shows that for most branches churn rates are proportional to customer population. Branches 2, 8, and 28 have very low churn rates, and branch 12 shows the highest churn rate. Bank could identify the customer attributes and bank services to see what may be contributing to such low (high) churn rates.

Figure 1A: Churn Rate across Regions

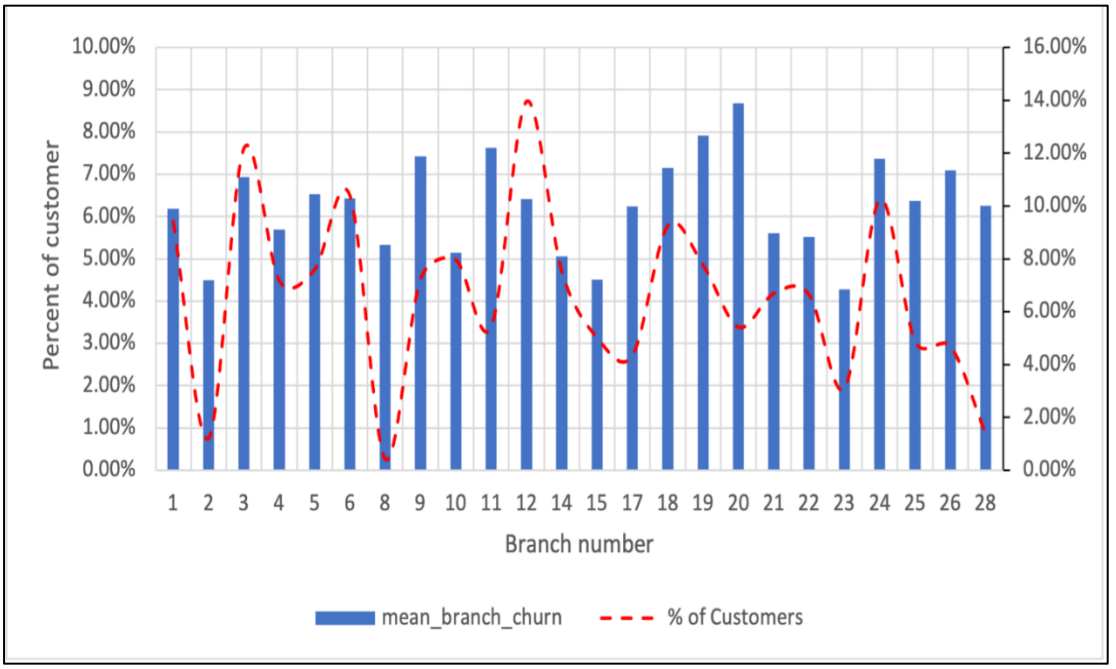


Figure 1B: Churn Rate across Branches

Bank customers are either business or personal category. Slightly higher than 20% are business customers and rest are personal category, but both display about 10% churn rates. Figure 2A presents churn rates across different age demography groups with business customers and Figure 2B presents different business categories with personal customers. Excluding the GI population, which has the highest churn rate of 17%, across all other age demography groups, Millennial show highest churn rates of 14.8%, with Gen Z at 11.4% and Boomer and Gen X are at 7.3% and 9.3% rates. Comparatively higher churn rates for Millennial and Gen Z portrays that the traditional banking relationship may be shifting and bank need to cater towards the changing needs for these generations to have a more long lasting and loyal customer base. Figure 2B displays that Estate business customers have the highest churn rate even though they represent only about 1% of the customer population. Among other business types, Sole Proprietorship shows 13% churn rate and Corporation, Partnership, and Trust show about 9% churn rates. Understanding what causes business customers to leave the bank would be important to develop long term relationship. It is possible that businesses cease to exist that is contributing to churn. Isolating business entity that leave voluntarily would be important to understand the customer need for banking relations and take initiative to develop long lasting relationship.

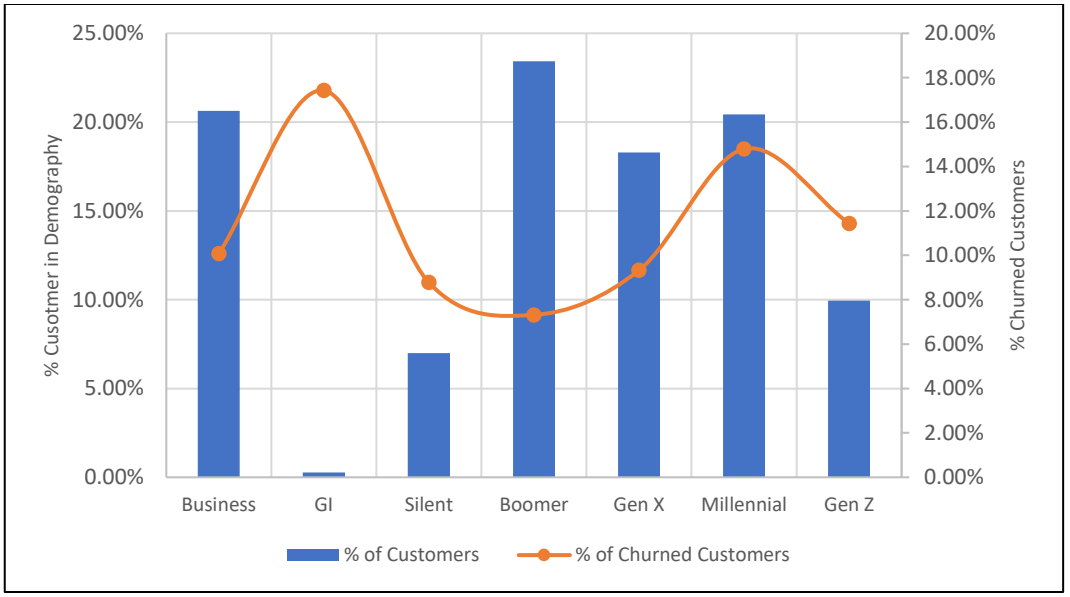


Figure 2A: Demography and Churn Rate

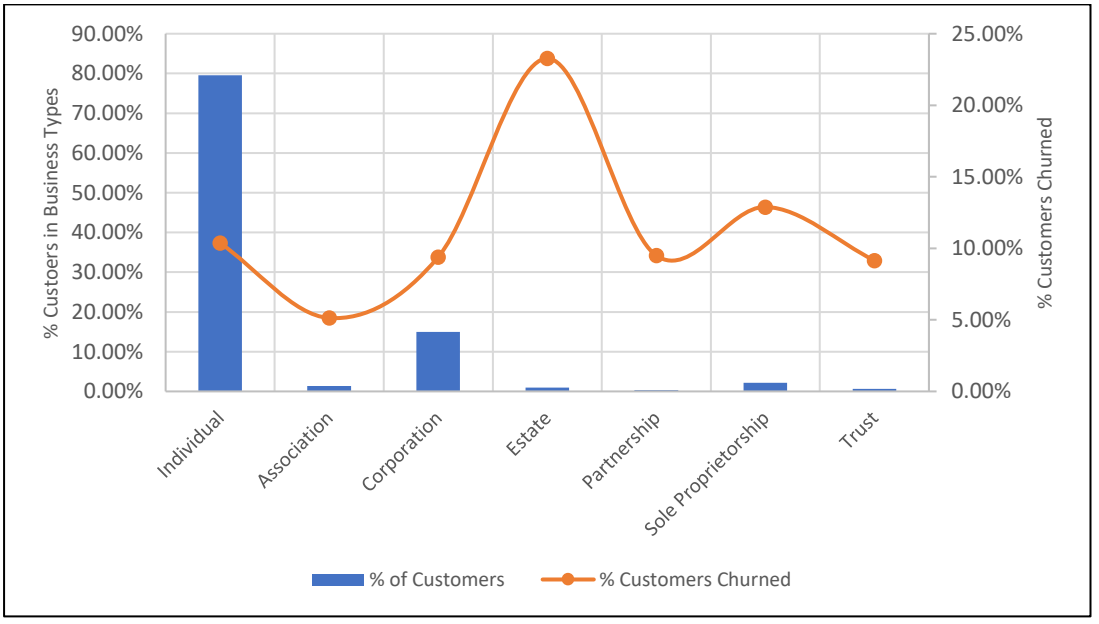


Figure 2B: Business Types and Churn Rate

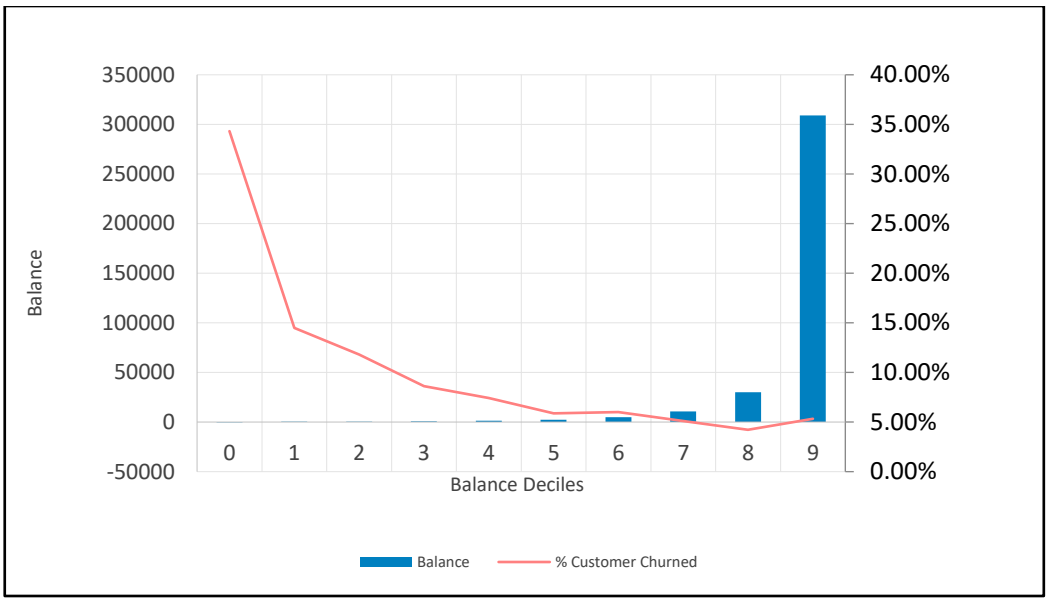


Figure 3A: Total Balance Deciles and Churn Rate

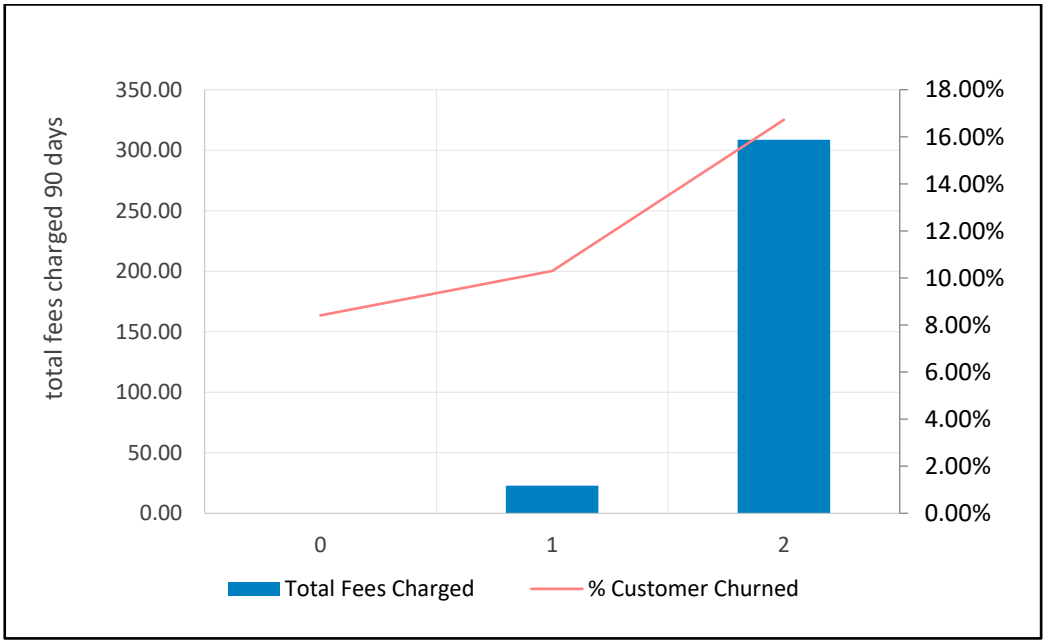


Figure 3B: Total Fee Charged Terciles and Churn Rate

We posit that customer who maintain deep relationship with the bank is less likely to leave. Account balance, credit and debit transactions, accounts and deposits of various forms, loan arrangements, credit and overdraft limits are likely to be indicators of deep relations with the bank and customers with more such relations are less likely to leave. On the other hand, customers who are charged higher fees in terms of service charge, overdraft, and other transaction fees may be more inclined to leave. It is possible that these customers are looking for a better deal with regard to the fee structure and leave voluntarily or they reach to point in their financial state that they are unable to maintain a banking relationship any longer. Identifying voluntary and involuntary churn customers will be important to address the churn issues to cater to need for each category. For financially struggling customers what role bank can take to mitigate financial constraints would be important avenue to explore for the bank to keep such customers financially afloat and continue to remain banking relations.

To observe effect of total account balance (for all accounts) on churn rate we construct deciles based on balance. Figure 3A shows churn rate declines steeply with balance deciles, indicating customers who have deep banking relations are less likely to leave. To observe the effect of fees charged to customers we construct *Total Fees Charged*, which is sum of *Service Charges*, *Overdraft and Return Fees*, and *Transaction Fees*. Based on *Total Fees Charged* customers are ranked into terciles to observe churn rate in each group. Figure 3B shows that churn rate increases steeply for higher tercile fee customers, indicating that customers who are charged higher fees are likely to leave the bank.

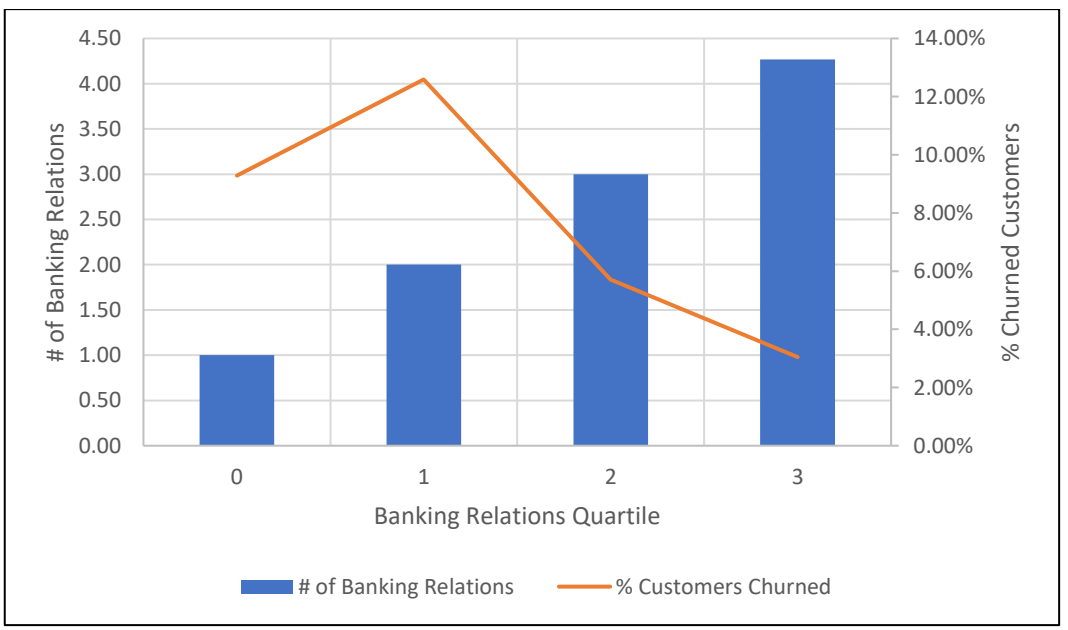


Figure 4: Banking Relations Quartile and Churn Rate

Customers who have strong banking ties are perhaps less likely to leave. Based on the attribute *Overall Banking Relations*, which captures customers banking relations in various forms (defined earlier), customers are ranked into quartiles to observe how churn rate varies across deep banking relations. Figure 4 shows that customers in the highest quartile with more than 4 banking relations on average has the lowest churn rate of about 3%, whereas customers with two banking relations on average has the highest churn rate over 12.6%, while the lowest quartile customers with one banking relation has churn rate of 9.3%. Understanding why customers in the second quartile group with more than 2 banking relations leave the bank more than the lowest quartile group would be important. Also, while customers with highest banking relations have lowest churn rate, however it will be insightful to explore what triggers them to leave the bank. Are these mainly voluntary churned customers? What bank could to do to retain these customers, what services these customers are looking for - exploring these questions would be important to develop understanding of bank's customer relations perspectives.

3.3 Difference in Customer Attributes: Churned vs. Non-Churned Customers

Table 4 presents the one-sample t-test for the churned and non-churned customers for different customer, account, and transaction level attributes. Panel A shows difference in customer level attributes and Panel B shows difference in account and transaction level attributes. As expected, non-churned customers has longer banking relations of average 9.5 years compared to 6 years for churned customers. Percentage of business customers across both categories are about the same. As discussed in Figure 2A, that millennials and Gen Zs are more likely to churn. There is not much difference with the number of accounts the customer holds with the bank, although shows statistical difference between the attributes. All attributes under *Customer Account Attributes* show significantly higher values for non-churned customers as compared to churned category. Notable ones include, non-churned customers has mortgage loan amount of $4.3K on average as opposed to $1.3K for churned customers and the total balance of loans is about $17.8K for non-churned customers and $8.1K for churned category and the differences are significant. *Customer Loan Delinquency* measures in terms of payment delays do not seem to have any significant difference between the two groups. For *Account Status* attributes, there is no significant difference in %Active account for churned vs. non-churned customers, while the %Dormant shows that non-churned customers have higher percentage of dormant customers (5.2%) as opposed to under churned category (4.19%). This implies that just being in dormant status not likely to indicate the customer is at risk of leaving the bank. However, percentage of limited accounts is significantly higher for churned category. *Banking Relations* attributes show that non-churned customers have significantly higher banking relations as compared to the churned category.

Table 4: Difference in Features: Churned vs. Non-Churned Customers

Panel A: Customer Level Features			
	Non-Churned N=42,496	**Churned N=4890**	**p-value for diff.**
Customer Attributes			
Near Branch (lives in the same zip code) %	65.94	64.15	0.01
Age	48.32	43.70	0.00
Number of Months of Data	12.2	7.79	0.00
Customer Number of Years	9.48	5.98	0.00
Business Customer %	20.48	20.08	0.52
Generation (%)			
GI	0.26	0.47	0.00
Silent	7.12	5.95	0.00
Boomer	24.22	16.61	0.00
Gen X	18.49	16.52	0.00
Millennial	19.41	29.28	0.00
Gen Z	9.82	11.02	0.00
Regions (%)			
Region 1 %	40.46	40.59	0.86
Region 2 %	49.59	47.85	0.02
Region 3%	9.95	11.55	0.00
Number of Accounts the Customer holds with the bank (#)			
Checking Accounts	1.12	1.01	0.00
Debit Cards	0.73	0.77	0.00
Savings Accounts	0.20	0.11	0.00
Time Deposits	0.08	0.03	0.00
Safety Boxes	0.04	0.02	0.00
Loans	0.12	0.05	0.00
Customer Accounts Attributes ($)			
Overdraft Limit	534.3	456.7	0.00
Credit Card Limit	117.5	44.02	0.01
Wealth Management Market Value	2,576.1	843.6	0.09
Last Mortgage Loan Amount	4303.2	1264.3	0.00
Total Balance of Loans	17,840	8,062	0.00
Customer Loan Delinquency			
Number of Days Past Due Loans	1.52	1.00	0.32
Number of Times Late at least 30 Days	0.25	0.19	0.23
Number of Times Late at least 60 Days	0.10	0.09	0.83
Number of Times Late at least 90 Days	0.06	0.06	0.83
Account Status			
Active %	89.66	89.92	0.56
Dormant % (no contact with the customer for a long period)	5.22	4.19	0.00
Escheated %	0.07	0.00	0.07
Frozen % (a temporary block is placed)	0.00	0.00	
Inactive % (no activity, shorter than dormant)	3.95	3.57	0.17
New %	0.95	1.11	0.26
Limited % (more restrictive than Frozen)	0.12	0.98	0.00

To be Closed %	0.03	0.23	0.00
Banking Relations			
Account Analysis %	1.26	0.82	0.00
RDC (Remote Deposit Capture) %	0.80	0.41	0.00
ACH (Automatic Clearing House) Originator %	0.60	0.33	0.02
Positive Pay (Fraud Prevention Setup) %	0.03	0.02	0.69
Wire Transmit Setup %	0.33	0.23	0.21
Wealth Management %	0.96	0.43	0.00
Mortgage Customer %	1.77	0.61	0.00
Hold Credit Card %	1.05	0.43	0.00
Overall Banking Relations #	1.98	1.87	0.00
Panel B: Account and Transaction Level Features			
Variables	**Non-Churned N=42,496**	**Churned N=4890**	**p-value for diff.**
Fees Charged to Customers			
Overdraft fees YTD $_t$	76.03	156.4	0.00
Return fees YTD $_t$	8.09	29.71	0.00
Overdraft Charged QTD $_t$	0.689	1.426	0.00
Overdraft Waived QTD $_t$	0.878	2.277	0.00
Service Charges $_t$	3.163	2.247	0.12
Overdraft and Return Fees $_t$	45.89	100.7	0.00
Transaction Fees $_t$	5.13	54.38	0.00
Total Fees Charged $_t$	53.32	156.6	0.00
Account Balance			
Balance_t	37,290	23,871	0.00
$\Delta\text{Balance}_{t-1}$	20.29	32.57	0.50
$\Delta\text{Balance}_{t-2}$	27.80	22.05	0.47
Credit and Debit Transactions			
$\$CT_t$	20,202	4,916	0.00
$\Delta\$CT_{t-1}$	193.6	32.75	0.02
$\Delta\$CT_{t-2}$	117.3	265.0	0.53
$\#CT_t$	4.08	1.59	0.00
$\Delta\#CT_{t-1}$	0.158	0.139	0.00
$\Delta\#CT_{t-2}$	0.185	0.189	0.41
$\$DT_t$	19,822	4,721	0.00
$\Delta\$DT_{t-1}$	216.7	20.0	0.28
$\Delta\$DT_{t-2}$	41.95	55.78	0.74
$\#DT_t$	18.60	8.37	0.00
$\Delta\#DT_{t-1}$	0.202	0.289	0.00
$\Delta\#DT_{t-2}$	0.258	0.339	0.00
OLB Activity Last 90 Days			
Money Management 90 Day Active	0.032	0.053	0.00
SMS 90 Day Active	0.112	0.066	0.00
App 90 Day Active	0.321	0.313	0.25
Tablet 90 Day Active	0.012	0.007	0.00
VRU 90 Days	0.038	0.049	0.00
OLB 90 Day Active (%)	41.50	37.85	0.00

Deposits and Transactions Last 3-months			
Deposits Count (#)	3.182	1.687	0.00
Deposits ($)	22,680	9,559	0.00
Mobile Deposits Count (#)	0.297	0.109	0.00
Mobile Deposits ($)	165.6	51.29	0.00
ACH Deposits Count (#)	8.025	2.880	0.00
ACH Deposits ($)	19,948	3,676	0.00
POS Debit Count (#)	25.14	17.00	0.00
POS Debit ($)	1,072.7	681.5	0.00
Check Card Transaction Count (#)	24.06	15.85	0.00
Check Card Transactions ($)	1274.9	794.4	0.00
RDC Deposits Count (#)	0.294	0.030	0.00
RDC Deposits ($)	5868.0	191.0	0.00
Time Deposit Balance	2874.2	819.9	0.00
Interest Paid			
Interest Paid YTD ($)	16.04	5.49	0.00
Interest Accrued but Not Paid	0.377	0.114	0.00
Interest Paid Last 3-months	10.97	2.53	0.00
Interest Rate (%)	0.012	0.006	0.00

The table provides one sample t-Test for the selected features between churned and non-churned customers. Panel A compares customer level features and Panel B compares account and transaction level features. All variable definitions are provided in Appendix A1.

For accounts and transaction level variables, *Fees Charged to Customers* are significantly higher for churned category. Although balance is higher for non-churned group, however, the changes in balance accounts are not significantly different. Dollar amount and number of transactions for both credit and debit accounts are significantly higher for non-churned category.

Although most of OLB Activity measures are higher for non-churned group, however the money management in last 90 days is higher for churned category. It is possibly capturing customers getting ready to leave the bank. Attributes under *Deposits and Transactions Last 3-months* and *Interest Paid* show significantly higher values for non-churned group. Overall banking activity and relations are much higher for the non-churned group compared to churned category.

4. Methodology

The research explores the churn prediction models that are known in the machine learning community. We explore machine learning models for classification problem that are widely used in the literature. We start with simple classification models such as Naïve Bayes and Support Vector Machine, then perform traditional Logistic Regression, then onto machine learning models such as, Decision Tree, Random Forest, Artificial Neural Networks (ANN), and ensemble bagging and boosting methods such as XGBoost, Bagging Meta Estimator, AdaBoost, and Gradient Boosting. The sample data set is split into training and test observations where model is constructed on the training sample and model performance is

validated in the test data set. In addition, during the model construction using the training data, cross-validation technique tackles the bias-variance tradeoff, addressing the overfitting problem while minimizing the bias thus assuring out-of-sample model performance.

To evaluate classifiers performance, we first start with confusion matrix as presented in Appendix A1, then define overall accuracy, precision, recall, specificity and F-Measure and misclassification errors Type-I and Type-II errors. We also use ROC and AUC curve for performance measurement of the classification problem at various threshold settings. Due to imbalanced nature of the data due to lower proportion of churn observations than non-churned category, the models provide lower true positive rate or conversely higher false negative rate, i.e., model incorrectly assigns an individual who churns to the non-churn category. This implies that Type-II error rate is likely to be higher as the model is likely to falsely predict the positive class (churned) labels to be negative (non-churned). Type-I Error, which happens when the model falsely classifies the negative class (non-churned) labels to be positive (churned), is likely to be lower in this problem, since non-churned category is more prevalent and model is less likely to identify non-churned customers to be churned category. For banks that are trying to identify customers who are likely to leave, detecting churn customers and improving the true positive rate (identifying customers who are at risk of leaving the bank as churned category) is essential for overall performance. The research will explore various ways to handle the imbalanced data, such as - over-sampling technique by increasing the number of minority class members in the training data set and under-sampling method which will reduce the number of majority class observations. The research will further explore other over-sampling methods used in the literature to improve the performance of the models, such as SMOTE and MTDF.

Below we describe (popular definition in the machine learning community) the machine learning models that we use for churn prediction: [10]

4.1 Naïve Bayes

Naïve Bayes methods are a set of supervised learning algorithms based on applying Bayes' theorem with the "naive" assumption of conditional independence between every pair of features given the value of the class variable. Naive Bayes classifiers assume that the value of a particular feature is independent of the value of any other feature, given the class variable. Naive Bayes learners and classifiers can be extremely fast compared to more sophisticated methods and achieved good results on the churn prediction problems in telecommunications industry (Kirui et al., 2013).

[10] Model descriptions are obtained from the following website: Scikit-Learn Documentation, Towards Data Science, IBM Topics,

4.2 Support Vector Machine (SVM)

Support vector machines (SVMs) are supervised learning models with associated learning algorithms that analyze data and recognize patterns for classification and regression analysis. SVMs are one of the most robust prediction methods, being based on statistical learning frameworks. SVM is an extension of the support vector classifier that results from enlarging the feature space in a specific way; using kernel the algorithm tries to find the optimal hyperplane which can be used to classify new data points. In churn prediction analysis Hur et al. (2005) find SVM outperforms other learning methods such as Decision Tree and Artificial Neural Network.

4.3 Logistic Regression

Logistic Regression is used for prediction of binary or categorical variables. It estimates probability of event occurring by estimating log-odds for the event (churn) based on linear combination selected features. Some of the work in churn predictions applying Logistic Regression are Eiben et al. (1998) Mozer et al.(2000), Buckinx and Van den Poel (2005) Neslin et al (2006), among others.

4.4 Decision Trees (DT)

A Decision Tree is a non-parametric supervised learning algorithm, which is utilized for both classification and regression tasks. It has a hierarchical, tree structure, which consists of a root node, branches, internal nodes and leaf nodes. It involves stratifying or segmenting the predictor space into a number of simple regions. For a given observation, prediction is made based on the mean of the training observations in the region to which it belongs. Tree based methods are simple and useful for interpretation. Au et al. (2003) and Wei and Chiu (2002) applied Decision Tree for analyzing churn problem in wireless telecommunications industry.

4.5 Random Forest

Random forests, a supervised machine learning algorithm, is an homogeneous ensemble learning method that constructs a multiple of decision trees at training time and can be used for regression and classification problems. For classification models each tree predicts a class and takes the majority of the vote for classification. For the churn problem, some of the notable works that include Random Forest are Buckinx and Van den Poel (2005), Lariviere and Van den Poel (2005), Kumar and Ravi (2008), among many others.

4.6 Artificial Neural Network

Artificial Neural Network (ANN) simulates decision making like human brain, which is composed of millions of neurons that process information and send signals. It is part of deep learning algorithms. ANN works with node layers in form of input, hidden, and output layers and obtains complex relationship between inputs and outputs to discover a pattern. Each node or artificial neuron is connected to another

and has an associate weight and threshold. Data is passed to the next layer of the network if the node is above the threshold value. Au et al. (2003) and Mozer et al. (2000) find in favor of ANN compared to other models for churn prediction problem.

4.7 Ensemble Boosting and Bagging methods

Ensemble Bagging and Boosting methods improves performance by combining predictions output of multiple single weak learners. Bagging Meta-Estimator combines multiple base estimators (weak learners) to create a robust model by creating multiple bootstrap samples from the training data and train base models independently on each of these samples. For each iterations the boosting algorithm change the weight of the training data distributions based on miss-classification. Three different boosting methods are widely used in the customer churn literature, and they are, AdaBoost, Gradient Boosted Machine Tree, Extreme Gradient Boosting (XGBoost). Recently XGBoost algorithm is getting lot of attention in the literature addressing the churn problem (Hanif, 2019, Ahmad et al. 2019).

4.8 Handling Imbalanced Data

One of the main challenges to predict customer churn arises due to the imbalanced nature of the data where number of churned customers are much lower than the non-churn category, which causes the false negative rates to be high, that is identifying customers who are likely to leave the bank as a non-churn category. We explore four different sampling techniques to handle the imbalanced nature of the data: over sampling, under-sampling, SMOTE, and MTDF. The over-sampling technique is used by randomly increasing the number of minority class to match number of instances in the majority class in the training data set, whereas in the under-sampling method the number of majority class observations is reduced. The under-sampling method can help mitigate the impact of class imbalance by reducing the dominance of the majority class; however, this may lead to information loss from the majority class, which can potentially affect model's general performance. SMOTE (Synthetic Minority Oversampling Technique) is an algorithm that expands the data by creating synthetic data points of the minority class instead of over-sampling with replacement. The algorithm identifies k-nearest neighbors for each minority class sample and synthetic samples are generated by interpolating the feature vector of the sample and its nearest neighbor. The new synthetic samples are then combined with the original dataset to train a model that can better handle imbalanced classes. The other oversampling technique used is Megatrend Diffusion Function (MTDF). MTDF diffuses the information of the data and based on a triangular membership function it estimates attribute domain of sample data and generates virtual sample within the domain.[11]

[11] See Chawla et al. (2002) for the details of the SMOTE procedure, Li et al. (2007) for MTDF procedure. See Amin et al. (2016) for comparison of SMOTE and MTDF in customer churn prediction model.

5. Results

The sample data is split into train and test data sets based on 80:20 criteria. The models are constructed using train data set and the model performances are measured and compared using test data. We build ten classification models and test and validate the models with four sampling techniques discussed in the prior section, in addition to the original test data. We compare model performances across various metrics such as overall accuracy, precision, recall, specificity, F1 score, AUC-ROC scores.

5.1 Comparisons of Model Performances

Table 5 presents performance metrics of ten classification methods that we apply for five different sampling techniques. The results show the overall accuracy of the models, which measures the percentage of correct predictions of churned and non-churned categories for the test data set, range from 23% to 96%. Across the ten classification methods presented, *Random Forest*, *XG Boost*, *AdaBoost,* and *Bagging Meta* show highest level of accuracy across the five different sampling methods presented, ranging from 87% to 96% of overall accuracy. Across the five sampling techniques, all methods expect for under-sampling provide 94% accuracy for *Random Forest*. For the *XG-Boost* method - original and MTDF dominate with 94% overall accuracy, for the *AdaBoost* classifier - original and MTDF dominate with 96% accuracy, and for the *Bagging Meta Estimator* - original, over-sampling, and MTDF dominate with 94% overall accuracy.

In addition to improving the overall accuracy, the goal of the classification methods is to improve accuracy measures such as Specificity, Recall, and Precision. Specificity measures the proportion of actual negatives (non-churned category) that are correctly identified as such. Since there are more non-churned customers, Specificity is likely to be high for the churn problem. Therefore, the False Positive Rate (FPR), which is also known as misclassification error of Type-I category, that occurs when the model falsely classifies the negative class (non-churned) class to be positive (churned), is likely to be low. From the bank's strategic perspective for the churn problem, FPR is not the primary concern as misclassifying non-churn as churned type will only prompt the bank to reach out to customers who are not actually at risk of leaving the bank and which eventually will makea path for developing deep banking relations. However, lowering FPR is important as resources are limited and targeting likely churn customers with some degree of confidence would be important from strategic point of view. The results in Table 5 show that Specificity measures vary from 87% to 99% for *Random Forest, XG Boost*, *AdaBoost*, and *Bagging Meta* which are the four models with best overall accuracy, therefore the FPR ranges from 1% to 13%, suggesting that the non-churned customers being identified as churn category is likely to be low.

Table 5: Performance of Machine Learning Models

Model	Data	Accuracy	Specificity (TNR)	Recall (TPR)	Precision	F1-score	AUC-ROC
Naïve Bayes	Original	0.23	0.15	0.94	0.11	0.2	0.62
	RUS	0.27	0.19	0.93	0.12	0.21	0.69
	ROS	0.23	0.15	0.95	0.11	0.2	0.61
	SMOTE	0.25	0.18	0.92	0.11	0.2	0.58
	MTDF	0.89	0.99	0	0.08	0.01	0.45
SVM Classifier	Original	0.92	0.99	0.24	0.87	0.38	0.84
	RUS	0.88	0.91	0.61	0.45	0.52	0.87
	ROS	0.89	0.91	0.67	0.48	0.56	0.88
	SMOTE	0.90	0.92	0.67	0.49	0.57	0.88
	MTDF	0.92	0.99	0.24	0.86	0.38	0.84
Logistic Regression	Original	0.9	1.00	0.07	0.69	0.13	0.77
	RUS	0.69	0.69	0.7	0.2	0.31	0.78
	ROS	0.7	0.70	0.7	0.21	0.32	0.77
	SMOTE	0.71	0.71	0.7	0.22	0.33	0.77
	MTDF	0.9	1.00	0.06	0.67	0.11	0.76
ANN	Original	0.88	0.95	0.31	0.39	0.34	0.62
	RUS	0.67	0.67	0.68	19	0.3	0.67
	ROS	0.83	0.88	0.4	0.28	0.33	0.64
	SMOTE	0.86	0.91	0.34	0.31	0.33	0.62
	MTDF	0.87	0.93	0.3	0.34	0.32	0.61
Decision Tree Classifier	Original	0.92	0.95	0.61	0.58	0.6	0.78
	RUS	0.81	0.81	0.78	0.32	0.45	0.8
	ROS	0.91	0.95	0.56	0.58	0.57	0.76
	SMOTE	0.9	0.93	0.62	0.49	0.55	0.77
	MTDF	0.91	0.95	0.6	0.57	0.59	0.77
Random Forest Classifier	Original	0.94	0.99	0.46	0.89	0.6	0.93
	RUS	0.87	0.87	0.81	0.42	0.55	0.92
	ROS	0.94	0.99	0.5	0.85	0.63	0.93
	SMOTE	0.94	0.98	0.58	0.73	0.65	0.93
	MTDF	0.94	0.99	0.46	0.89	0.61	0.93

XGBoost Classifier	Original	0.94	0.99	0.51	0.87	0.64	0.92
	RUS	0.88	0.89	0.81	0.45	0.58	0.93
	ROS	0.9	0.92	0.78	0.52	0.62	0.93
	SMOTE	0.89	0.91	0.72	0.47	0.57	0.92
	MTDF	0.94	0.99	0.5	0.87	0.64	0.91
Gradient Boosting	Original	0.91	1.00	0.16	0.93	0.27	0.83
	RUS	0.8	0.71	0.72	0.3	0.42	0.85
	ROS	0.72	0.71	0.81	0.24	0.37	0.86
	SMOTE	0.82	0.84	0.61	0.31	0.41	0.82
	MTDF	0.91	1.00	0.17	0.95	0.28	0.8
Ada Boost	Original	0.96	0.99	0.66	0.90	0.76	0.93
	RUS	0.89	0.90	0.82	0.48	0.61	0.92
	ROS	0.90	0.90	0.81	0.49	0.62	0.93
	SMOTE	0.94	0.96	0.73	0.72	0.73	0.92
	MTDF	0.96	0.99	0.66	0.90	0.76	0.93
Bagging Meta Estimator	Original	0.94	0.99	0.55	0.84	0.66	0.9
	RUS	0.87	0.88	0.79	0.43	0.56	0.91
	ROS	0.94	0.98	0.59	0.77	0.67	0.9
	SMOTE	0.93	0.97	0.6	0.72	0.65	0.9
	MTDF	0.94	0.99	0.55	0.85	0.67	0.9

The table provides performance measures of nine machine learning models using 4 sampling techniques in addition to the original test data. The machine learning models are: Naïve Bayes, KNN Classifier, Logistic Regression, Artificial Neural Networks (ANN), Decision Tree Classifier, Random Forest Classifier, XG Boost Classifier, Gradient Boosting, and Bagging Meta Estimator. Four sampling techniques used are: Random Under-sampling (RUS), Random Over-sampling (ROS), Synthetic Minority Oversampling Technique (SMOTE), and Megatrend Diffusion Function (MTDF). The performance measures reported are: overall Accuracy, Specificity, Recall, Precision, F1-score, and AUC-ROC measures. The description of the performance measures is provided in Appendix A2.

Recall, which is also known as True Positive Rate (TPR), measures the proportion of correct prediction of actual positive class (churn category). Due to imbalanced nature of the data it is difficult to identify customers who are likely to churn accurately. From the bank's business strategy perspective, it wants to lower the False Negative Rate (FNR), which is falsely identifying churn customers as non-churn category, the misclassification that is also known as Type-II error. Improving on Recall measure will reduce the FNR and therefore reduce the Type-II error. The results in Table 5 shows that the recall ranges from 46% to 82% for the four classification methods *Random Forest*, *XG Boost*, *AdaBoost*, and *Bagging Meta*, which have the highest overall accuracy and for all four methods the under-sampling technique provides the best recall measure (79% - 82%), however, the overall accuracy is lower for this sampling method. One thing to note is that, although Naïve Bayes shows over 90% Recall measure, however, the overall accuracy for the model is very low which is about 20% and therefore will not be a suitable model for the churn problem.

Precision measures the proportion of correct predictions out of all positive predicted classes. The range of Precision scores is between 42% to 90% for the four estimators discussed above. One thing to note is that if we decrease the false negative (select more positives), recall always increases, but precision may increase or decrease. It is difficult to compare two models with low precision and high recall or vice versa. So, to make them comparable, we use F1-Score which is a composite measure of Recall and Precision at the same time and it is the harmonic mean of the two scores. It can have a maximum score of 1 (perfect precision and recall) and a minimum of 0. Across the four above mentioned classifiers *AdaBoost* with original and MTDF sampling provides the best F1-Score of 76% with Precision of 90%, Recall of 66%, and 96% overall accuracy. The next selections of models and sampling methods are: *AdaBoost* with SMOTE; *Bagging Meta Estimator* with MTDF, over-sampling, and original sample; *XG-Boost* with original and MTDF; and *Random Forest* with SMOTE; overall accuracy of all of which is 94% with F1-Scores range from 64% to 73%. Figure 5 displays the overall accuracy scores and Figure 6 shows the F1-Score for the models and displays that the AdaBoost, *Bagging Meta*, *XG-Boost*, and *Random Forest* dominate in performance.

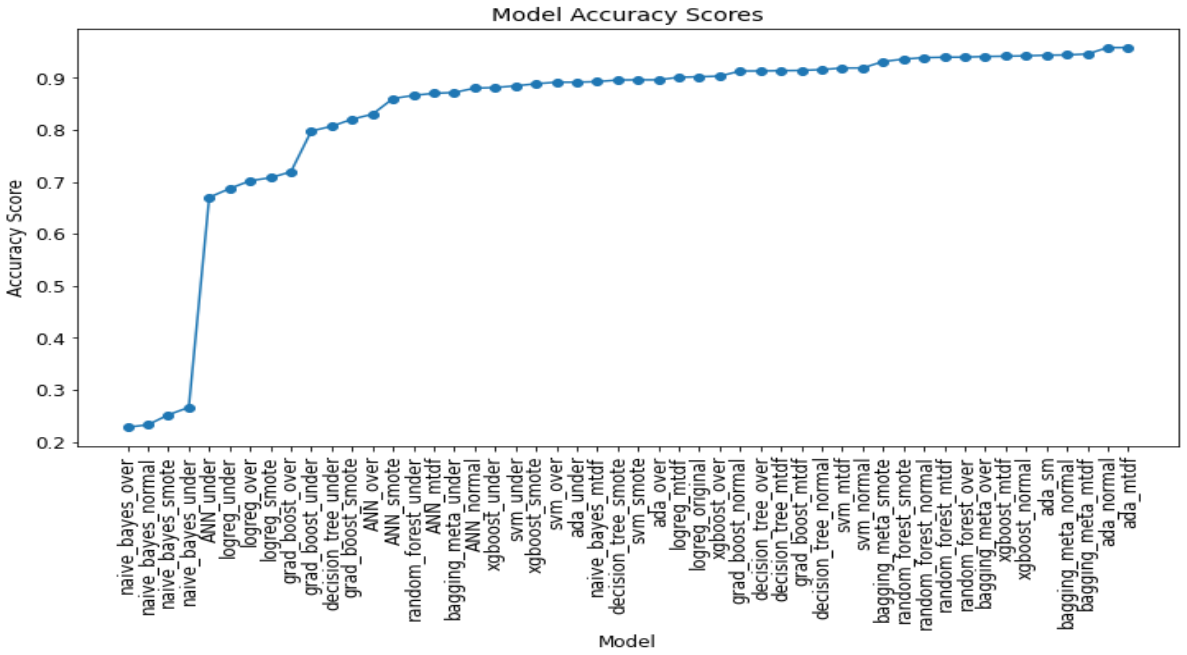


Figure 5: Comparisons of Overall Accuracy Scores for the Ten Machine Learning Models across Five Samples

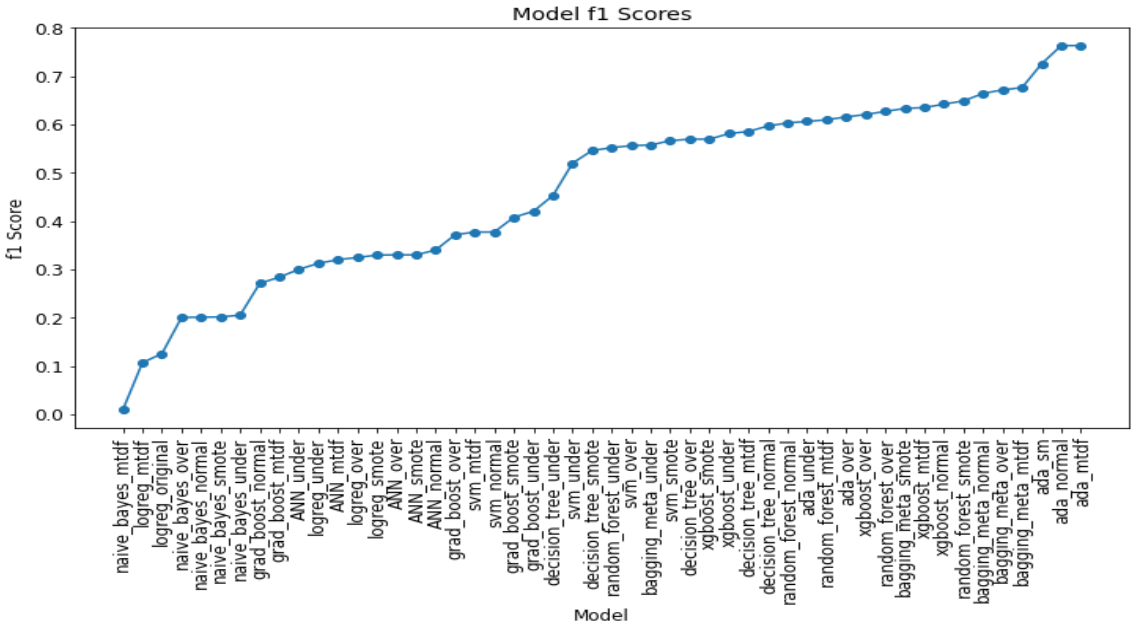


Figure 6: Comparisons of F1-Scores for the Nine Machine Learning Models across Five Samples

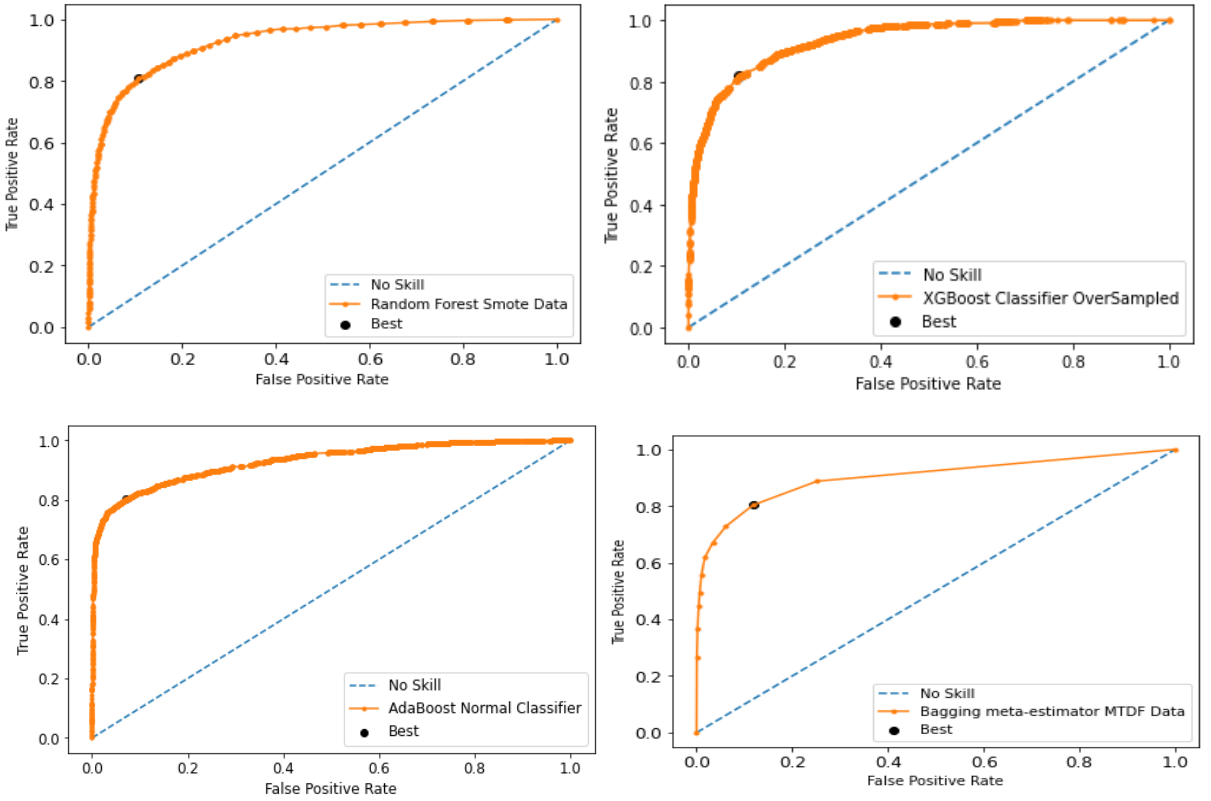


Figure 7: AUC – ROC Curves for Top-Four Models

Figure 7 shows the ROC curves for the four churn prediction models, *Random Forest*, *XG Boost*, *AdaBoost*, and *Bagging Meta*. The classification problem requires identifying data into distinct categories whereas the model results provide probability of an occurrence of a certain class. Therefore, a cut-off value needs to be defined to identify the desired class based on the probability of the model results. The Receiver Operating Characteristic (ROC) curve evaluates the performance of the classification problem by plotting the true positive rate (Sensitivity) vs, false positive rate (1-Specificity) for different cut-off values (Swets, 1988). Each point in the ROC curve represents performance of the model for a particular threshold value and closer the ROC curve towards the top left corner the higher the overall accuracy, with perfect (100%) sensitivity and specificity representing by the upper left corner. The Area Under the Curve (AUC) is an accepted traditional performance metric for a ROC curve (Duda, Hart, & Stork, 2001; Bradley, 1997; Lee,2000), with AUC of 100% indicating a perfect classifier. Table 5 shows that AUC values for the four classifiers vary from 0.62 to 0.93 and the best AUC values are for *Random Forest*, *XG Boost*, *AdaBoost*, and *Bagging Meta*, providing 0.9 to 0.93 AUC values.

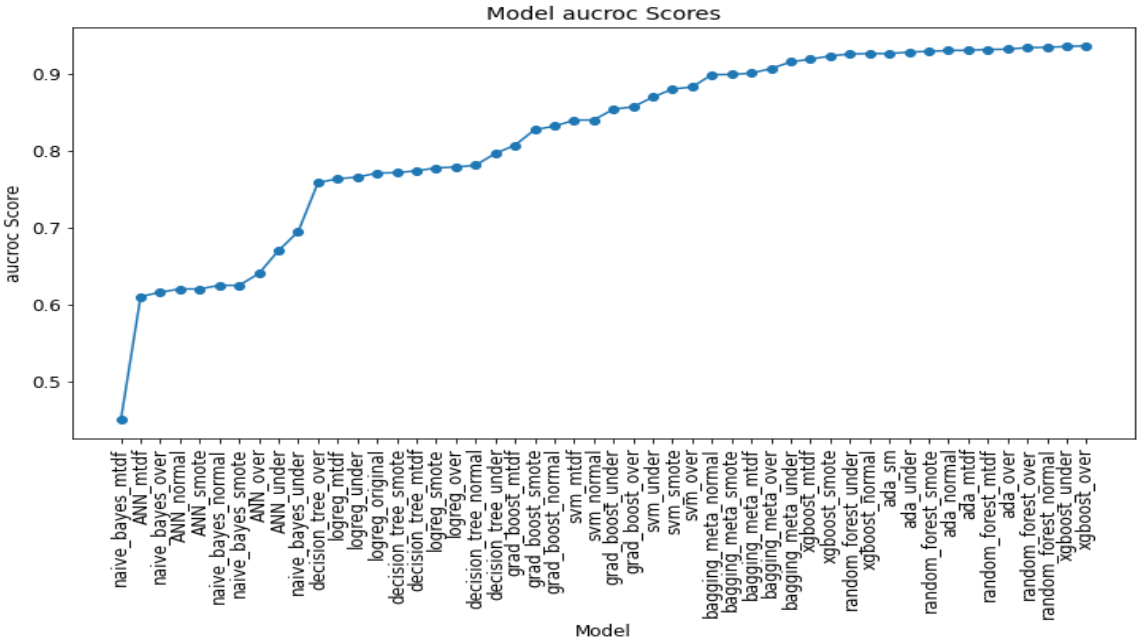


Figure 8: Comparisons of AUC-ROC Scores for the Ten Machine Learning Models across Five Samples

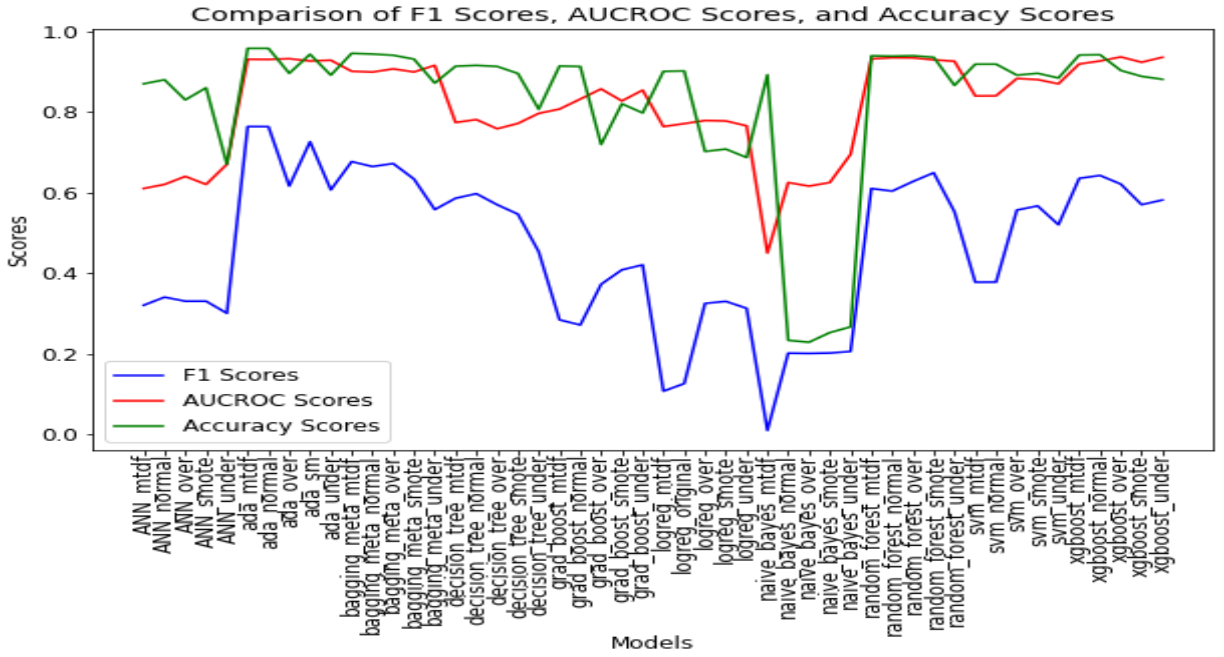


Figure 9: Comparisons of F1-Scores, AUC-ROC Scores, Overall Accuracy Scores for the Ten Machine Learning Models across Five Samples

Figure 8 displays the AUC measures for the ten classifiers and for five different sampling methods presented in Table 5. Considering AUC measure, *XG Boost*, *Random Forest*, and *AdaBoost* are very similar and slightly dominate *Bagging Meta*, but they all are comparable. Figure 9 displays the F1-scores, AUC-ROC scores, and overall accuracy for the ten machine learning models across five sampling techniques and reiterates the findings.

5.2 Significant Features

Figure 10 presents most significant features for the churn prediction model based on the *AdaBoost* classifier with original test data. We see that the number of years a customer has been with the bank is an important predictor of churn. In addition, amount and change in balance, and checking, debit and credit transactions are important predictors of customer churn. Moreover, overdraft waived, ACH deposit, and fees charged seem to contribute in the classification.

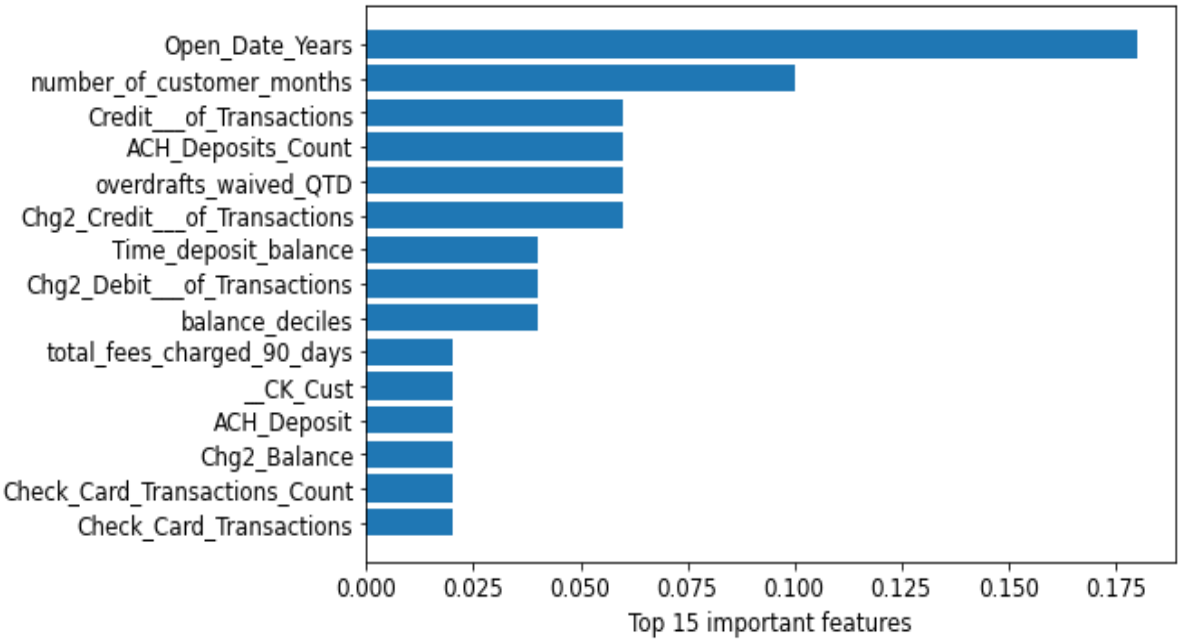


Figure 10: Most Signiant Features for the Ada Boost Classifier on Original Data

These features highlight that deep banking relations reduce the probability of customer churn. Customers with higher balance, more debit and credit transactions, and changes in accounts are less likely to leave the bank, which are reliable indicators of deep banking relationship. Also, longer tenure customers are less likely to leave indicating that the banking relationship could be very sticky such that once a customer gets comfortable with the service being provided, she may not be

looking for an alternate banking relationship. Therefore, developing long-term relation would be key for customer retention. Customers whose overdraft fees are waived are less likely to leave, indicating that when customers have overdraft protection they feel the bank is looking after them and are likely to stick with the bank. Customers who have higher ACH transactions are less likely to churn. The ACH system is the primary system that agencies use for electronic funds transfer (EFT) so higher engagements on this attribute indicate that the customer is relying on the bank for electronic transfer of funds and that could be related to salary, business, or any other personal electronic transactions.

6. Conclusion

Customer churn is a critical and challenging problem facing financial services industry as customers are seeking alternate banking relations in the rapidly changing competitive and changing environment due to entrants of new FinTech companies and changes in the technology space that are shaping the industry. Banks and financial institutions are relying on big-data analysis using machine learning models to explore the churn problem to gain insightful predictions on customer churn behavior. Using real customer banking relations data from a large community bank in the US, we analyze the churn problem by providing a diagnostic analysis and constructing ten machine learning models and compare performances across five sampling techniques. Among the ten machine learning models that are explored, *Random Forest*, *XGBoost*, *AdaBoost*, and *Bagging Meta Estimator* dominate in performance across various measures and sampling methods. Using the F1-Score and overall accuracy AdaBoost with original and MTDF sampling technique dominate, while *Bagging Meta* with MTDF and over-sampling, *Random Forest* with SMOTE, and *XGBoost* with MTDF provide very close performance measures. Considering AUC measures AdaBoost, *XGBoost*, and *Random Forest* slightly dominate *Bagging Meta* for all sampling techniques, although the performance of these four classifiers across the various sampling techniques are very similar.

The results suggest that overall banking activity and relations are much higher for the non-churned group compared to churned category. Non-churned customers have been with the bank longer and have higher balance, more credit and debit transactions in dollar and numbers, higher loans, more overdraft limit protection, are charged less, get more waiver, and use various banking services compared to churn category. Although, customers with highest banking relations have lowest churn rate, however it is not completely absent and it will be important for the bank to identify what contributes toward their exit decision – is it voluntary or involuntary? It is possible that businesses cease to exist or an individual face extreme financial difficulty that is contributing to churn. What a bank can do to make sure that financially constraints customers remain afloat and address the need of customers who are looking for a different service would be important to mitigate the customer churn behavior and develop long lasting relationship with the bank. Furthermore, comparatively higher churn rates for Millennial and Gen Z portrays

that the traditional banking relationship may be shifting and bank need to cater to the changing needs for these generations to have a more long lasting and loyal customer base. Overall creating high-touchpoints would be important to understand customer needs and address the concerns that are contributing to churn behavior.

Acknowledgements

The author gratefully acknowledges the support of Research Grant ($19.14K) from the large Community Bank at South that provided the data and Georgia FinTech Academy and GRA work by Hiteshchandra Sai and Charan Yellanki. The author assumes responsibility for any errors.

References

[1] Ahmad, A. K., Jafar, A., and Aljoumaa, K., 2019. Customer Churn Prediction in Telecom Using Machine Leaning in Big-Data Platform. *Journal of Big Data* 6(28), 1-24.

[2] Amin, A., Anwar, S., Adnan, A., Nawaz, M., Howard, N., Qadir, J., Hawalah, A., Hussain, A., 2016. Comparing Oversampling Techniques to Handle the Class Imbalance Problem: A Customer Churn Prediction Case Study. *IEEE Access* 4, 7940-7957.

[3] Amin, A., Anwar, S., Adnan, A., Nawaz, M., Alawfi, K., Hussain, A., and Huang, K., 2017. Customer Churn Prediction in the Telecommunication Sector Using a Rough Set Approach. *Neurocomputing* 237, 242-254.

[4] Athanassopoulos, A., 2000, Customer Satisfaction Cus to Support Market Segmentation and Explain Switch Behavior. *Journal of Business Research* 47 (3), 191-207.

[5] Au, W., Chan, K., and Yao, X., 2003. A Novel Evolutionary Data Mining Algorithm with Applications to Churn Prediction. *IEEE Transactions on Evolutionary Computation* 7(6), 532-545.

[6] Berry, L., 1995. Relationship Marketing of Services - Growing Interest, Emerging Perspectives. *Journal of Academy of Marketing Science* 23(4), 236-245.

[7] Bhattacharya, C., 1998. When Customers are Members: Customer Retention in Paid Membership Contexts. *Journal of the Academy of Marketing Science* 26(1), 31-44.

[8] Bradley, A. P., 1997. The Use of the Area Under the ROC Curve in the Evaluation of Machine Learning Algorithms. *Pattern Recognition*,30(6), 1145–1159.

[9] Buchak, G., Matvos, G., Piskorski, T., Seru, A., 2018. Fintech, Regulatory Arbitrage, and the Rise of Shadow Banks. *Journal of Financial Economics* 130, 453–483.

[10] Buckinx, W., and Van den Poel, D., 2005. Customer Base Analysis: Partial Defection of Behaviorally Loyal Clients in a Non-contractual FMCG retain setting. *European Journal of Operations Research* 164(1), 252-268.

[11] Burez, D., Van den Poel, D., 2009. Handing Class Imbalance in Customer Churn Prediction. *Expert Systems with Applications* 36(3), 4626-4636.
[12] Chawla, N. V., Bowyer, K.W., Hall, L.O., and Kegelmeyer, W.P., 2002. SMOTE: Synthetic Minority Over-Sampling Technique. *Journal of Artificial Intelligence Research* 16(1), 321-357.
[13] Cheng, J., Sun, J., Yao, K., Xu, M., and Cao, Y., 2022. A Variable Selection Method Based on Mutual Information and Variance Inflation Factor. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, 268 120652. https://doi.org/10.1016/j.saa.2021.120652
[14] Christopher, M., Payne, A. F., and Ballantyne, D., 1991, Relationship Marketing: Bring Quality, Customer Service, and Marketing Together, *Butterworth -Heinemman*, Oxford, UK.
[15] Colgate, M.R. and Danaher, P.J. (2000). Implementing a Customer Relationship Strategy: The Asymmetric Impact of Poor versus Excellent Execution. *Journal of the Academy of Marketing Science*, 28, 375-387. https://doi.org/10.1177/0092070300283006
[16] Colgate, M., Stewart, K., & Kinsella, R. (1996). Customer Defection: A Study of the Student Market in Ireland. *International Journal of Bank Marketing*, 14(3), 23–29.
[17] Duda, R., Hart, P., & Stork, D., 2001. Pattern Classification. *Wiley-Interscience*.
[18] Eiben, A., Koudijs, A., and Slisser, F., 1998. Genetic Modeling of Customer Retention. *Lecture Notes in Computer Science* 1391, 178-186.
[19] Erel, I., Liebersohn, J., 2020. Does Fintech Substitute for Banks? Evidence from the Paycheck Protection Program. National Bureau of Economic Research.
[20] Ganesh, J., Arnold, M., & Reynolds, K., 2000. Understanding the Customer Base of Service Providers: An Examination of the Differences between Switchers and Stayers. *Journal of Marketing*, 64(3), 65–87.
[21] Gómez, R.S., Sánchez, A.R., Garcia, C.G., and Pérez, J.G., 2020. The VIF and MSE in Raise Regression. *Mathematics* 8 (4) (2020) 605. https://doi.org/10.3390/math8040605
[22] Gopal, M., Schnabl, P., 2022. The rise of finance companies and fintech lenders in small business lending. *The Review of Financial Studies* 35(11), 4859-4901.
[23] Gummesson, E., 1993. Quality Management in Service Organizations. ISQA, Stockholm University, Sweden.
[24] Hanif I., 2019. Implementing Extreme Gradient Boosting (XGBoost) Classifier to Improve Customer Churn Prediction. *International Conference of Statistics and Analysis 2019*, August, Bogor, Indonesia
[25] He, H., Bai, Y., Garcia, E. A., and Li, S., 2008. ADASYN: Adaptive Synthetic Sampling Approach for Imbalanced Learning. *IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence*, 1322-1328.

[26] Hur, Y., Lim, S., 2005. Customer Churning Prediction Using Support Vector Machines in Online Auto Insurance Service. *Advances in Neural Network – ISSN 2005 Springer*, 928-933

[27] Hwang, H., Jung, T., and Suh, E., 2004. An LTV Model and Customer Segmentation Based on Customer Value: A Case Study on the Wireless Telecommunication Industry. *Expert Systems with Applications* 26(2), 181-188.

[28] Kirui, C., Hong, L., Cheruiyot, W., and Kirui, H., 2013. Predicting Customer Churn in Mobile Telephony Industry Using Probabilistic Classifiers in Data Mining. *International Journal of Computer Science* 10(2), 165-172.

[29] Kumar, D., and Ravi, V., 2008. Predicting Credit Card Customer Churn in Banks Using Data Mining. *International Journal of Data Analysis Techniques and Strategies* 1(1), 4-28.

[30] Lariviere, B., and Van den Poel, D. 2005. Predicting Customer Retention and Profitability by Using Random Forest and Regression Forest Techniques. *Expert Systems with Applications* 29(2), 472-484.

[31] Lee, S. 2000. Noisy Replication in Skewed Binary Classification. *Computational Statistics and Data Analysis*, 34.

[32] Lewis, B., 1993. Service Quality: Recent Development in Financial Services. *International Journal of Bank Marketing* 11, 19-25.

[33] Li, D.-C., Wu, C.-S., Tsai, T.-I., Lina, Y.-S., 2007. Using Mega-Trend Diffusion and Artificial Samples in Small Data Set Learning for Early Flexible Manufacturing System Scheduling Knowledge. *Computers and Operations Research* 34(4), 966-982.

[34] Mozer, M.C., Wolniewicz, R., Grimes, D., Johnson, E., Kaushansky, H., 2000. Predicting Subscriber Dissatisfaction and Improving Retention in the Wireless Telecommunications Industry. *IEEE Transactions on Neural Network* 11(3), 690-696.

[35] Neslin, S., Gupta, S., Kamakura, W., Lu, J., and Mason, C., 2006. Detection Defection: Measuring and Understanding the Predictive Accuracy of Customer Churn Models. *Journal of Marketing Research* 43(2), 204-211.

[36] Paulin, M., Perrien, J., Ferguson, R., Salazar, A., and Seruya, L., 1998. External Effectiveness of Commercial Banking in Canada and Mexico. *International Journal of Bank Marketing* 16(1), 24-31.

[37] Rasmusson, E., 1999. Complaints Can Build Relationships. *Sales and Marketing Management* 151(9), 89-90.

[38] Reichheld, F. F., 1996. Learning from Customer Defections. *Harvard Business Review* 74(2), 56-69.

[39] Stum, D. L. and Thiry, A., 1991. Building Customer Loyalty. *Training and Development Journal* 45, April, 34-36.

[40] Swets, J., 1988. Measuring the Accuracy of Diagnostic Systems. *Science*, 240, 1285–1293.

[41] Verbeke, W., Martens, D., Mues, Chrsitophe, Baesens, B., 2011. Building Comprehensible Customer Churn Prediction Models with Advanced Rule Induction Techniques. *Experts Systems with Applications* 38, 2354 – 2364.
[42] Vafeiadis, T., Diamantaras, K.I., Sarigiannidis, G., Chatzisavvas, K. Ch., 2015. A Comparison of Machine Learning Techniques for Customer Churn Prediction. *Simulation Modelling Practice and Theory* 55, 1-9.
[43] Wei, C., and Chiu, I, 2002. Turning Telecommunications Call Details to Churn Prediction: A Data Mining Approach. *Expert Systems with Applications* 23, 103-112.
[44] Zeithaml, V., Berry, L., and Parasuraman, A., 1996. The Behavioral Consequences of Service Quality. Journal of Marketing 60(2), 31-46.

Appendix

A1. Variable Definitions

Variable construction from monthly observation data to customer level data.

Panel A: Customer Level Variables	
Variables	**Definition**
Customer Attributes	
Near Branch (lives in the same zip code) %	A flag equaling 1 if the customer lives in the same ZIP code as a branch, and 0 otherwise
Age	The customer's age in years. This field is NULL for business customers
Number of Months of Data	The number of months the customer has data
Customer Number of Years	The number of years customer has relationship with the bank.
Business Customer %	Percentage of customers who are business entity
Generation (%)	A categorization based on the customer's year of birth.
GI	1901-1926
Silent	1927-1945
Boomer	1946-1964
Gen X	1965-1980
Millennial	1980-2000
Gen Z	Born after 2000
Regions (%)	
Region 1 %	Percentage of customers belong to bank branches in Region 1.
Region 2 %	Percentage of customers belong to bank branches in Region 2.
Region 3 %	Percentage of customers belong to bank branches in Region 3.
Number of Accounts the Customer holds with the bank (#)	
Checking Accounts	The number of checking accounts that the customer has with the bank.
Debit Cards	The number of debit cards that the customer has with the bank.
Savings Accounts	The number of savings accounts that the customer has with the bank.
Time Deposits	The number of time deposits that the customer has with the bank.
Safety Boxes	The number of safe deposit boxes that the customer has with the bank.
Loans	The number of loans that the customer has with the bank.
Customer Accounts Attributes ($)	
Overdraft Limit	The number of dollars a customer is allowed to go below zero balance before incurring an overdraft fee.
Credit Card Limit	The limit on the customer's credit cards, if they have any credit cards.
Wealth Management Market Value	The value of all customer assets managed by the bank Wealth Management.

Last Mortgage Loan Amount	The amount of the customer's last mortgage financed with the bank Mortgage.
Total Balance of Loans	The balance of all bank loans held by the customer.
Customer Loan Delinquency	
Number of Days Past Due Loans	The number of days past due a customer is on their loan payments.
Number of Times Late at least 30 Days	The number of times a customer has been at least 30 days late on a loan payment, ever.
Number of Times Late at least 60 Days	The number of times a customer has been at least 60 days late on a loan payment, ever.
Number of Times Late at least 90 Days	The number of times a customer has been at least 90 days late on a loan payment, ever.
Account Status	
Active %	The proportion of the customer's deposit accounts that are in the status of Active.
Dormant % (no contact with the customer for a long period)	The proportion of the customer's deposit accounts that are in the status of Dormant (no contact with customer for a long period of time).
Escheated %	The proportion of the customer's deposit accounts that are in the status of Escheated (funds to be remitted to the state).
Frozen % (a temporary block is placed)	The proportion of the customer's deposit accounts that are in the status of Frozen (a temporary block has been placed on the customer).
Inactive % (no activity, shorter than dormant)	The proportion of the customer's deposit accounts that are in the status of Inactive (no activity for a period of time, shorter than Dormant).
New %	The proportion of the customer's deposit accounts that are in the status of New (customer has just been opened and has not yet been funded).
Limited % (more restrictive than Frozen)	The proportion of the customer's deposit accounts that are in the status of Limited (a block has been placed on the customer, more restrictive than Frozen).
To be Closed %	The proportion of the customer's deposit accounts that are in the status of To be closed (customer is about to be closed).
Banking Relations	
Account Analysis	A flag equaling 1 if the customer is enrolled in customer Analysis (a pricing system for high-activity customers), and 0 otherwise.
RDC (Remote Deposit Capture)	A flag equaling 1 if the customer is set up for remote deposit capture (RDC), and 0 otherwise.
ACH (Automatic Clearing House) Originator	A flag equaling 1 if the customer is set up as an Automated Clearing House (ACH) originator, and 0 otherwise.
Positive Pay (Fraud Prevention Setup)	A flag equaling 1 if the customer is set up to use Positive Pay (a US treasury check fraud-prevention system), and 0 otherwise.

Wire Transmit Setup	A flag equaling 1 if the customer is set up to transmit wires, and 0 otherwise.
Wealth Management	A flag equaling 1 if the customer has a wealth management account with the bank, and 0 otherwise.
Mortgage Customer	A flag equaling 1 if the customer has taken out a mortgage with the bank since 1/1/2020, and 0 otherwise.
Hold Credit Card	A flag equaling 1 if the customer holds a bank credit card, and 0 otherwise.
Overall Banking Relations	A flag equaling 1 if the customer is enrolled in at least one the banking relations listed in this group (Account Analysis, RDC, ACH, Positive Pay, Wire, Wealth Management, Mortgage, or Credit Card or has account related to checking, saving, debit card, loan, time deposit or safety deposit account) and 0 otherwise.

Panel B: Account and Transaction Level Variables

Variables	**Definition at monthly data**	**Construction of variable at customer level**
Fees Charged to Customers		
Overdraft fees YTD $_t$	The amount of fees incurred for overdraft from the beginning of the year till the given month *t*.	Max
Return fees YTD $_t$	The amount of fees incurred for returned deposited items (e.g. bounced checks) from the beginning of the year till the given month *t*.	Max
Overdraft Charged QTD $_t$	The number of overdraft events in three months prior to a given month *t* that resulted in a fee being charged.	Max
Overdraft Waived QTD $_t$	The number of overdraft events in three months prior to a given month *t* for which the overdraft fee was waived.	Max
Service Charges $_t$	Service charge fees charged by the bank in three months prior to a given month *t*.	Max
Overdraft and Return Fees $_t$	Overdraft and return fees charged in three months prior to a given month *t*. Includes some other related fees.	Max
Transaction Fees $_t$	Transaction fees charged by the bank in three months prior to a given month *t*. Includes ACH, RDC and wire transaction fees as well as stop payments.	Max
Total Fees Charged $_t$	Sum of Service Charges, Overdraft and Return Fees, and Transaction Fees in three months prior to a given month *t*	Max
Account Balance		
Balance_t	The current balance of all customer accounts, i.e. the amount of funds in all accounts, as of the month *t*.	Mean
$\%\Delta\text{Balance}_{t-1}$	The percentage change in balance over one month prior to a given month *t*	Mean

$\%\Delta Balance_{t-2}$	The percentage change in balance over two months prior to a given month *t*	Mean
Credit and Debit Transactions		
$\$CT_t$	The dollar amount of all credit transactions for the customer in the month *t*.	Mean
$\%\Delta\$CT_{t-1}$	The percentage change in the dollar amount of credit transactions between the month *t* and (*t-1*)	Mean
$\%\Delta\$CT_{t-2}$	The percentage change in the dollar amount of credit transactions between the month *t* and (*t-2*)	Mean
$\#CT_t$	The number of credit transactions made by this customer in the month *t*.	Mean
$\%\Delta\#CT_{t-1}$	The percentage change in the number of credit transactions between the month *t* and (*t-1*).	Mean
$\%\Delta\#CT_{t-2}$	The percentage change in the number of credit transactions between the month *t* and (*t-2*).	Mean
$\$DT_t$	The dollar amount of all debit transactions for the customer in the month *t*.	Mean
$\%\Delta\$DT_{t-1}$	The percentage change in the dollar amount of debit transactions between the month *t* and (*t-1*)	Mean
$\%\Delta\$DT_{t-2}$	The percentage change in the dollar amount of debit transactions between the month *t* and (*t-2*)	Mean
$\#DT_t$	The number of debit transactions made by this customer in the month *t*.	Mean
$\%\Delta\#DT_{t-1}$	The percentage change in the number of debit transactions between the month *t* and (*t-1*).	Mean
$\%\Delta\#DT_{t-2}$	The percentage change in the number of debit transactions between the month *t* and (*t-2*).	Mean
OLB Activity Last 90 Days		
Money Management 90 Day Active	A flag equaling 1 if the customer has used Money Management in the three months prior to a given month *t*, and 0 otherwise.	Mean
SMS 90 Day Active	A flag equaling 1 if the customer has used SMS Banking in the three months prior to a given month *t*, and 0 otherwise.	Mean
App 90 Day Active	A flag equaling 1 if the customer has used the Mobile App in the three months prior to a given month *t*, and 0 otherwise.	Mean
Tablet 90 Day Active	A flag equaling 1 if the customer has used the Mobile App in the three months prior to a given month *t*, and 0 otherwise.	Mean
VRU 90 Days	A flag equaling 1 if the customer has used voice banking (the Voice Response Unit, or VRU) in the three months prior to a given month *t*, and 0 otherwise.	Mean
OLB 90 Day Active	A flag equaling 1 if the customer has used online banking in the three months prior to a given month *t*, and 0 otherwise.	Mean

Deposits and Transactions Last 3-months		
Deposits Count (#)	The number of deposit transactions made in the three months prior to a given month *t.*	Mean
Deposits ($)	The dollar amount of deposits paid to the customer in the three months prior to a given month *t.*	Mean
Mobile Deposits Count (#)	The number of mobile deposit transactions (check scanning apps) made in the three months prior to a given month *t.*	Mean
Mobile Deposits ($)	The dollar amount of mobile deposits (check scanning apps) paid to the customer in the three months prior to a given month *t.*	Mean
ACH Deposits Count (#)	The number of ACH deposit transactions (electronic transfers) made in the three months prior to a given month *t.*	Mean
ACH Deposits ($)	The dollar amount of ACH deposits paid to the customer in the three months prior to a given month *t.*	Mean
POS Debit Count (#)	The number of POS (point of sale) Debit transactions made in the three months prior to a given month *t.*	Mean
POS Debit ($)	The dollar amount of POS (point of sale) debits paid to the customer in the three months prior to a given month *t.*	Mean
Check Card Transaction Count (#)	The number of check card transactions made in the three months prior to a given month *t.*	Mean
Check Card Transactions ($)	The dollar amount of check card transactions paid to the customer in the three months prior to a given month *t.*	Mean
RDC Deposits Count (#)	The number of RDC deposit transactions made in the three months prior to a given month *t.*	Mean
RDC Deposits ($)	The dollar amount of RDC deposits paid to the customer in the three months prior to a given month *t.*	Mean
Time Deposit Balance	The balance of all time deposits held by the customer	Mean
Interest Paid		
Interest Paid YTD ($)	Interest paid to the customer from the beginning of the year till the given month *t.*	Max
Interest Accrued but Not Paid	Interest currently accrued to the customer but not yet paid out. This is because interest accrues constantly but is usually paid out in intervals (e.g. once a month).	Mean
Interest Paid Last 3-months	Interest paid to the customer in the three months prior to a given month *t.*	Mean
Interest Rate (%)	The average interest rate of the customer's accounts	Mean

A2. Performance Evaluation Measures

To evaluate classifiers performance, we first start with confusion matrix as presented in Table A1 then define overall accuracy, recall, precision, specificity and F-Measure and misclassification errors Type-I and Type-II errors. We also use ROC and AUC curve for performance measurement of the classification problem at various threshold settings.

		Predicted Class	
		Churned	Non-churned
Actual Class	Churned	TP	FN
	Non-churned	FP	TN

TP: True Positive; TN: True Negative; FP: False Positive; FN: False Negative

Total = TP+TN+FP+FN

Fraction of correct predictions:

$\frac{TN+TP}{Total}$

Recall: Out of all the positive classes (True), how much we predicted correctly. It should be high as possible (also known as **true positive rate (TPR)**).

$$Recall = \frac{TP}{TP+FN}$$

Precision: Out of all the positive classes we have predicted, how many are actually positive (True).

$$Precision = \frac{TP}{TP+FP}$$

Specificity (**true negative rate (TNR))**: measures the proportion of actual negatives that are correctly identified as such. It is the opposite of recall.

$$Specificity = \frac{TN}{TN+FP}$$

False Positive Rate (FPR)): measures the proportion of actual negatives that are incorrectly identified as positive. **FPR=1-TNR**

$$FPR = \frac{FP}{TN+FP}$$

F-Measure: A composite measure of precision and recall and can be interpreted as a weighted average of precision and recall.

$$F - measure = \frac{2*Recall*Precision}{Recall+Precision}$$

Misclassification Error:

Two types of errors:

Type-I Error: The model falsely classifies the negative class labels to be positive

False positive: Incorrectly assigns an individual who does not churn to the churn category

$$Type - I\ Error = 1 - Specificity = \frac{FP}{FP+TN}$$

Type-II Error: the model falsely predicted the positive class labels to be negative

False Negative: Incorrectly assigns an individual who churns to the non-churn category

$$Type - II\ Error = 1 - Recall = \frac{FN}{TP+FN}$$

Due to imbalanced nature of the data, Type-II error is more likely to happen as churned customers are more likely to be identified as non-churned category.

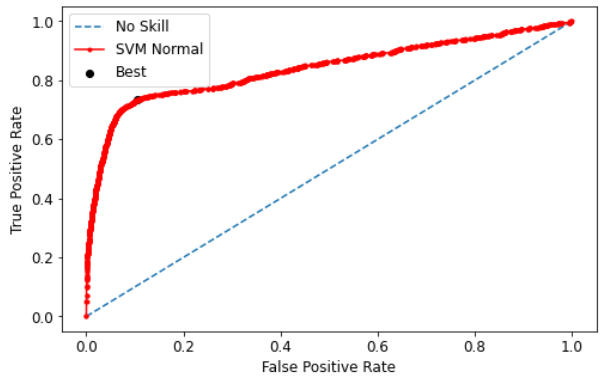
Type-I error is less likely to happen since non-churned category is more prevalent and model is less likely to identify non-churned customers to be churned category.

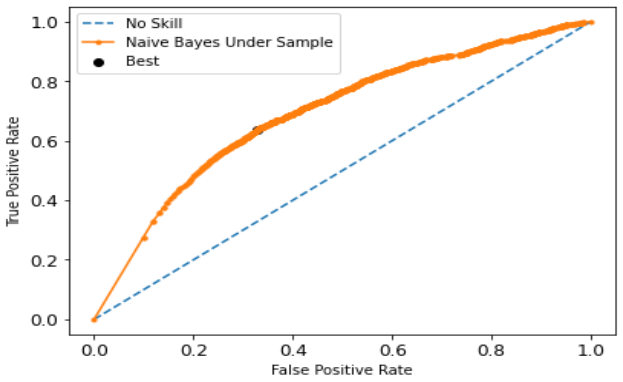
ROC and AUC curve:
ROC curve shows performance of classification model at all classification thresholds. The graph plots True Positive Rate and False Positive Rate. AUC measures the area underneath of the ROC curve and provides an aggregate measure of performance across all possible classification thresholds.

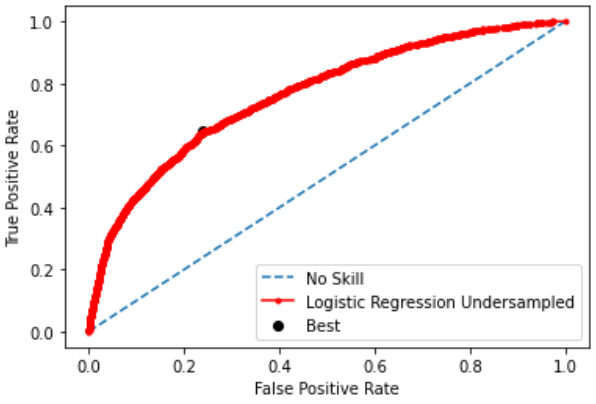
A3. Variable Selections:

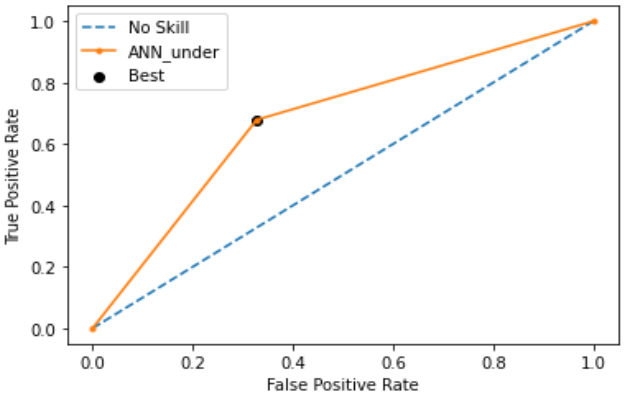
Here are the list variables that are dropped based on correlation and VIF Analysis

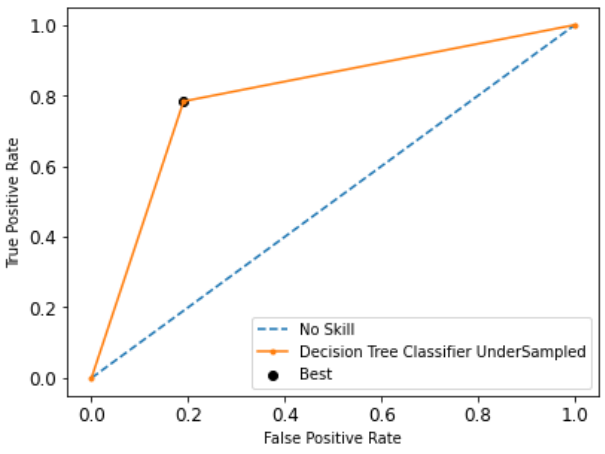
1) 't-1.Debit_#_of_Transactions',
2) 't-2.Debit_#_of_Transactions',
3) 't-2.Credit_#_of_Transactions',
4) 't-1.Credit_#_of_Transactions',
5) 't-2.Debit_Dollar_Amt_of_Transactions',
6) 't-1.Debit_Dollar_Amt_of_Transactions',
7) 't-2.Credit_Dollar_Amt_of_Transactions',
8) 't-1.Credit_Dollar_Amt_of_Transactions',
9) 't-2.Balance',
10) ‘t-1.Balance’,
11) ‘CK_SV_average_12_month_balance’,
12) 'CK_SV_average_balance_QTD'
13) 'Deposits_Count',
14) 'POS_Debits_Count',
15) 'RDC_Deposits_Count',
16) 'Mortgage_Customer'
17) 'Debit_Dollar_Amt_of_Transactions',
18) 'Number_of_Times_Late_60_Days_Loans'

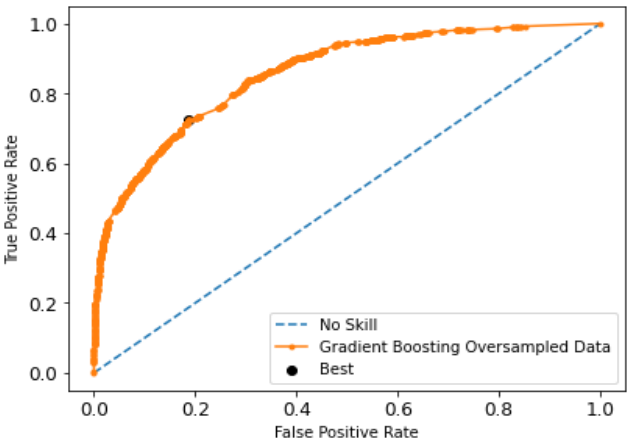












A4: AUC – ROC Curves for the rest of the models except for top-four