MODULE 7 – Final Research Paper

Mauro Maich

Colorado State University Global

MIS543-1: Enterprise Performance Management

Winter 2021 8 Week Session D

Instructor: Dr. Lisa Bryan

Due Date: 6/5/2022

Table of Contents

ABSTRACT

Retail banking is a hyper-competitive industry currently being disrupted and transformed by technology. Banking transactions were historically done in person up until the 2000s when internet broadband access made digital banking via a website or phone app the norm (Futures, 2017). With the proliferation of digital banking, products such as checking and savings accounts and credit cards became heavily commoditized, making it easier for customers to leave their current bank for another. This is generally known as churn, and banks must understand and minimize it because the cost of acquiring a new customer is usually much higher than retaining an existing one. Being able to identify existing customers who are likely to churn soon enables a bank to target these customers with marketing campaigns to prevent their defection to a competitor.

As the banking industry continues to be disrupted by non-traditional providers and fintech startups offering banking services, established banks must understand and retain their existing customers. Understanding which characteristics make an existing customer more likely to close their accounts and move their business to another bank before it happens would give the bank a chance to retain that customer relationship and address any incentives the customer may have to switch banks.

INTRODUCTION

This research project will use an anonymized dataset from real banking customers to build a predictive churn model. This model will help estimate the probability of a customer leaving the current bank for a competitor in the near term, to enable early detection and retention of the client. While marketing initiatives aimed at new customer acquisition are necessary, customer retention and churn management are critical for any business to maintain healthy profitability levels. If churn outpaces new customer acquisition, the marketing efforts to bring new customers to the bank would be futile, hence customer churn management is essential for any bank seeking to create value and retain its customers.

Because large banks in the United States generally have millions of clients across different geographical reasons, it becomes challenging to follow a consistent approach to customer retention in each region. It would be extremely hard and likely ineffective to ask each local branch manager across the country to implement the same customer churn prevention methodology. A centralized approach using statistical analysis models to identify existing customers with a high propensity to defect to a competitor is necessary to be successful in preventing customer churn.

OBJECTIVES

The goal is to identify which customer variables have more influence in predicting a customer leaving a bank for another and to build a predictive model that can be used by any bank to identify possible defectors based on it. Because the financial services industry is "largely undifferentiated and characterized by fierce competition among homogenous players for customer acquisition", an accurate predictive attrition model would be a strategic advantage to any bank (Ahn et al., 2019, p. 569). Furthermore, understanding the reason for a customer

leaving their current bank could help the company make improvements to its operations to prevent other clients to defect. Long-term, satisfied customers are not only valuable because of the existing relationship but also because they can be great ambassadors for the bank (Raj & Azad, 2020).

This research aims to produce insights and a working model that can be applied in the banking industry to ascertain which customers are likely to churn and help prevent attrition. Using different statistical classification and regression techniques to build a predictive model can enable target marketing initiatives that can result in lower attrition rates. Furthermore, this research would review different modeling techniques for goodness of fit to understand which statistical analysis method and software tool can produce the most accurate predictive model. Methods such as decision trees, logistic regression, and random forest would be considered and the results analyzed to find a superior model. Further research would be needed to understand which customers are worth retaining from a profitability perspective, as this research will center on the propensity to churn only and not on which customers the bank should concentrate on retaining.

OVERVIEW OF STUDY

The plan to analyze the dataset consists of sequential steps, starting with a more general, summary analysis, getting to a more specific analysis of each variable, and finally using predictive methods to find the best fitting model to predict future customer churn. The initial analysis will be performed using MS Excel to do a first exploration of the dataset using pivot tables to investigate the count of the binary variables, such as Gender = Male / Female, Exited = 0/1, IsActiveMember= 0/1, and others. This will help the research get grounded on the structure of the data and start understanding which variables should be included in the analysis. A close

follow-up will be uploading the dataset to SAS Studio and using the Summary Statistics task to get statistical measures of the variables such as mean, standard deviation, min/max, and total count (N).

Next, move to Python Jupyter Notebook to understand each unique values for each variable and drop any variable that may not add value to the analysis. Some initial pie and bar charts can be used to show comparisons such as the number of customers that churned ("Exited" variable) vs. not, as well as select scatter plots and correlation matrix charts to get preliminary insights into any potential relationships between variables. Histograms can also be used to understand the distribution of each variable and check for normalcy, skewness, and kurtosis.

Finally, the more involved step of fitting different models to the data using SAS Enterprise Guide and the Jupyter Notebook including Decision Trees, Multiple Variable Regression, and ANN. This is the initial intention, but other methodologies such as support vector machine (SVM) and Random Forests may be used if deemed necessary to find the model with the highest possible predictive accuracy. The ultimate goal is to arrive at a model that can be used in a real-life scenario to input variables such as age, gender, credit score, tenure, etc., and get a churn propensity score for each customer.

RESEARCH QUESTIONS AND HYPOTHESES

The goal of this project is to answer the question: what factors contribute most to banking customer churn? Furthermore, what is the optimal combination of factors to predict banking customer churn? Finding answers to these questions will enable financial institutions to take proactive action to prevent profitable customer defections to a competitor, which in turn will have a positive impact on the bottom line. This research can be particularly useful to large incumbent banking institutions that are at risk of losing customers to fintech startups, online

banks, and other non-traditional financial services providers.

Another important aspect of retaining profitable banking customers is that it is six times less expensive than acquiring a new customer (Amin et al., 2019). In today's ultra-competitive financial services environment, retaining the right customers is critical to improving the bottom line, and these research questions bear the potential to equip incumbent banking institutions with an effective tool to prevent churn.

Hypotheses

Null Hypothesis: All factors contribute equally to bank customers' churn propensity

Alternative Hypothesis: Certain variables contribute more to bank customers' propensity to churn.

The first hypothesis explores if there are certain independent factors, such as demographic or current product mix variables, that may impact an existing customer's propensity to move their accounts to a competitor. While there could be a reasonable assumption that this is always the case, using statistical analysis tests to prove or disprove this hypothesis is critical to understanding if the existing dataset has any predictive value. For example, it could be reasonably presumed a priori that customers of pre-retirement age, living in zip codes with high median incomes will be highly coveted by competitor financial institutions, and thus become the target of a high volume of sophisticated, omnichannel marketing efforts. This could make these "prime" customers more likely to defect to a competitor than customers with low bank balances who are just starting in their financial lives. However, this research will take the approach of first validating or invalidating that the variables contained in the existing dataset indeed bear predictive value, and then move to the second hypothesis.

Null Hypothesis: No specific combination of factors has predictive power over banking

customers' churn

Alternative Hypothesis: A certain combination of factors has predictive power over banking customers' churn

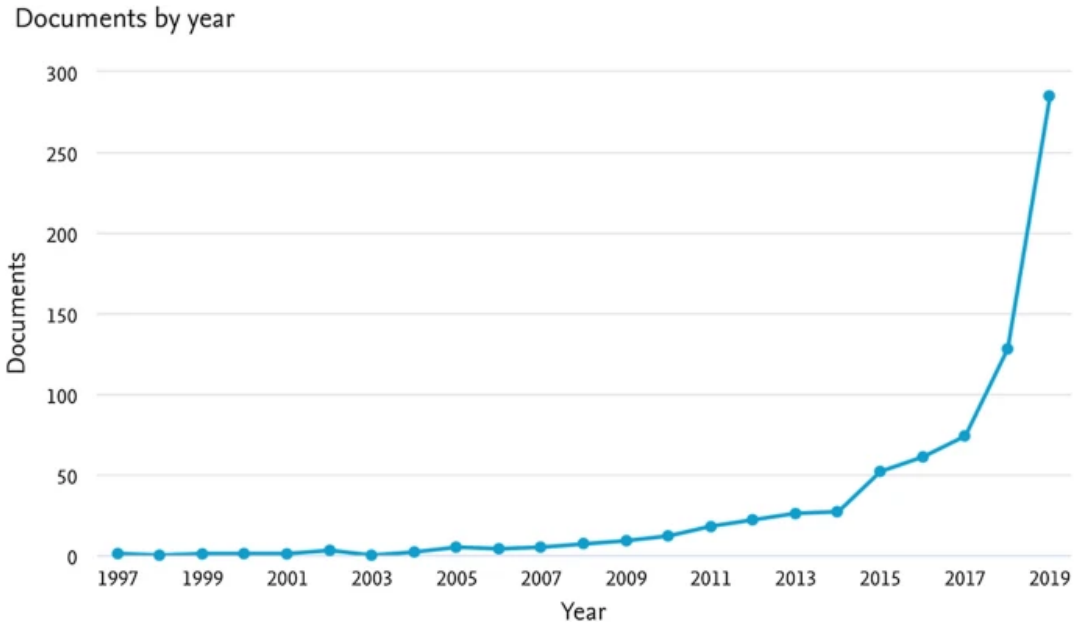
The second hypothesis builds on the outcome of the first hypothesis test by further looking into the potential predictive power of a combination of variables, as opposed to testing individual variables. While some individual variables may have more predictive value than others, finding out if a specific combination of variables can increase the probability of a customer defecting to a competitor may deliver a more effective predictive model.

LITERATURE REVIEW

According to research, there has been a recent spike in customer churn prevention literature published in the banking industry, driven by the concern of losing customers to new fintech startups and non-traditional banking (de Lima Lemos et al., 2022)

Figure 1

Number of publications in journals and conferences from 2003 to 2019 with the co-occurrence of the terms machine learning and churn either in the title, abstract, or keyword list



In another published study Ahn, Y., Kim, D., & Lee, D.-J. (2019) named "Customer attrition analysis in the securities industry: A large-scale field study in Korea", the authors apply the ubiquitous recency, frequency, and monetary (RFM) framework to analyze real data from banking customers in Korea. In their conclusion, they found that there is a need for more sophisticated churn research that also incorporates factors such as attention and duration (Raj & Azad, 2020). The contention is very insightful because a customer's decision to do business with a financial services provider to purchase an intangible service such as a brokerage or deposit account is very complex and plagued with qualitative intricacies. (Ahn et al., 2019 574)

In the recently published article entitled "Propension to customer churn in a financial institution: a machine learning approach" by de Lima Lemos, R. A., Silva, T. C., & Tabak, B. M., 2022 the authors use a three-step approach to create a churn prediction model. Firstly, they compare a variety of supervised algorithms to ascertain which one is a better fit for the extensive data set under study. Secondly, they compiled a large dataset of customer-level transactional data from a real financial institution in Brazil, and finally, they used the variables in the dataset to create a model of customer churn prediction, as well as investigate which attributes had the most predictive power. (de Lima Lemos et al., 2022). This research will take the concept further and attempt to understand if variables used as a group, instead of individually, can increase the predictive power of the model.

RESEARCH DESIGN

Methods

This research will focus on the deductive method of hypothesis testing, using a quantitative approach focused on the narrow scope of the hypothesis; namely, can a group of customer

characteristics help the bank predict which customers are more likely to move their bank assets to a competitor? The quantitative analysis will be developed using statistical analysis tools such as python (Jupyter Notebook) and SAS Enterprise Guide, leveraging different features to explore, visualize, and understand the dataset, as well as leveraging the many open source libraries (numpy, matplotlib, seaborn, etc.) and SAS "tasks and utilities" perform statistical tests.

Some of the exploratory tools considered for this research are summary statistics, a correlation matrix of all relevant variables, bar charts and histograms of variable distribution, and scatterplots to visualize potential relationships among variables. Doing exploratory analysis using visualization tools upfront will help the researcher become familiar with and internalize the data and provide a first glance at any potential relationships among variables, as well as gain a deeper understanding of the dataset.

As far as the modeling techniques, the initial intent is to use Decision Trees, Multivariable Regression, Logistic Regression, and Gradient Boosting, and then compare which technique offers the best fitting model for the churn prediction problem. Decision Trees are particularly attractive to analyze binary variables and "…frequently used for classification and prediction" problems, and yet can sometimes suffer from lack of performance and robustness (Kim & Lee, 2022, p. 1336).

Limitations

One of the main challenges with published research about customer churn in financial services is the fiduciary relationship that generally exists between customers and the institutions related to the flow of information between the parties (McKechnie, 1992). Most transactions executed by financial institutions on behalf of their customer are considered covered by the fiduciary standard, that is, a legal responsibility to act only in the best interest of the customer,

including safeguarding their information. While the fiduciary standard is a sound and sensible policy implemented to protect the client, it makes obtaining real transactional information from customers very challenging. The Churn_Modelling.csv file available on Kaggle.com contains anonymized data of real bank customers. The data is anonymized because of the sensitive nature of personal customer information held by the bank, but it does contain relevant information for this research as it was originally from real customers.

Ethical Considerations

Because this research will be centered on a data set of anonymized real customers from a banking institution, it is important to be aware of and manage any potential subjectivities and inherent biases intrinsic to working with demographic information such as gender, age, country of residence, or even credit scores. The choice of which bank to do business with appears to be largely commoditized, with previous research highlighting criteria such as "…dependability and size of the institution, location, convenience and ease of transactions, professionalism of bank personnel and availability of loans" (McKechnie, 1992, p. 5). While it would appear as though there is a low risk of researcher bias given the quantitative approach and commodities nature of financial services, there is always a potential for insensitivity towards issues such as gender, age, and geographical location. To control for these potential biases, this research will include literature on consumer behavior and ethics in financial services, paying special attention to avoiding assumptions based on demographic characteristics.

According to research, "…information conveyed by non-directly observable information can add explanatory power for estimating risk attrition" (Tang et al., 2014, p. 630), which could open the possibility of using derived variables beyond what is directly contained in the data set under consideration. That would be outside of the scope of this research, thus reducing the possibility

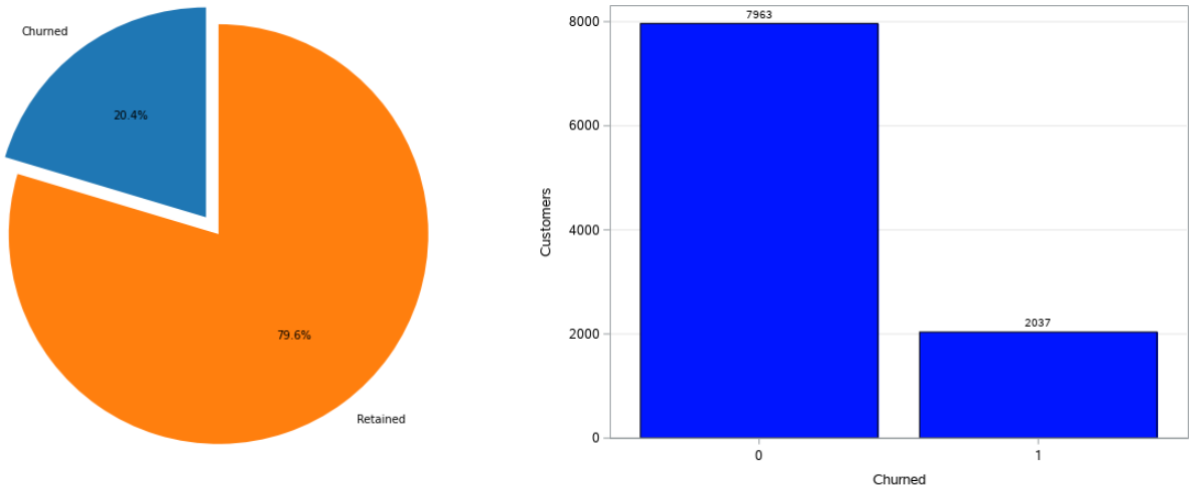
of researcher bias. Instead, this research will be centered on the variables contained in the data set and the statistical modeling described above.

FINDINGS

Establishing a baseline for this research is key. Understanding what is the percentage of customers that churned, or left for a competitor, shows that a little over 20% of customers left the bank for a competitor. This baseline gives us a starting point to learn about the actual churn rate, and seek to understand the factors affecting churning, trends, predictability, and ultimately what can be done to retain customers that are profitable to the bank. The figures below provide simple visuals showing that 20% of customers left the bank, which equates to about a 20% churn rate.

Figure 1

Percentage of customers and actual number of customers who left for a competitor



Having a baseline understanding of the current churn percentage, the next step of this research is looking into the statistics of the variables in the dataset to start diving deeper into the population of this bank's customers. The variables "CustomerID", "RowNumber", and

"Surname" are dropped from the dataset because they are random and it is clear that do not possess any valuable for predictive analysis. Some of the interesting preliminary findings in the variables that are left are that the average customer is about 38 years old, has a credit score of 650, holds 1.5 products with the bank, and has an estimated salary of $100,000. About 70% of existing customers hold a credit card and their average tenure with the bank is 5 years. The table below provides a consolidated visual of the variables included in the analysis which shows a view into the type of customer in this bank, and the histograms provide a visual of the distribution.

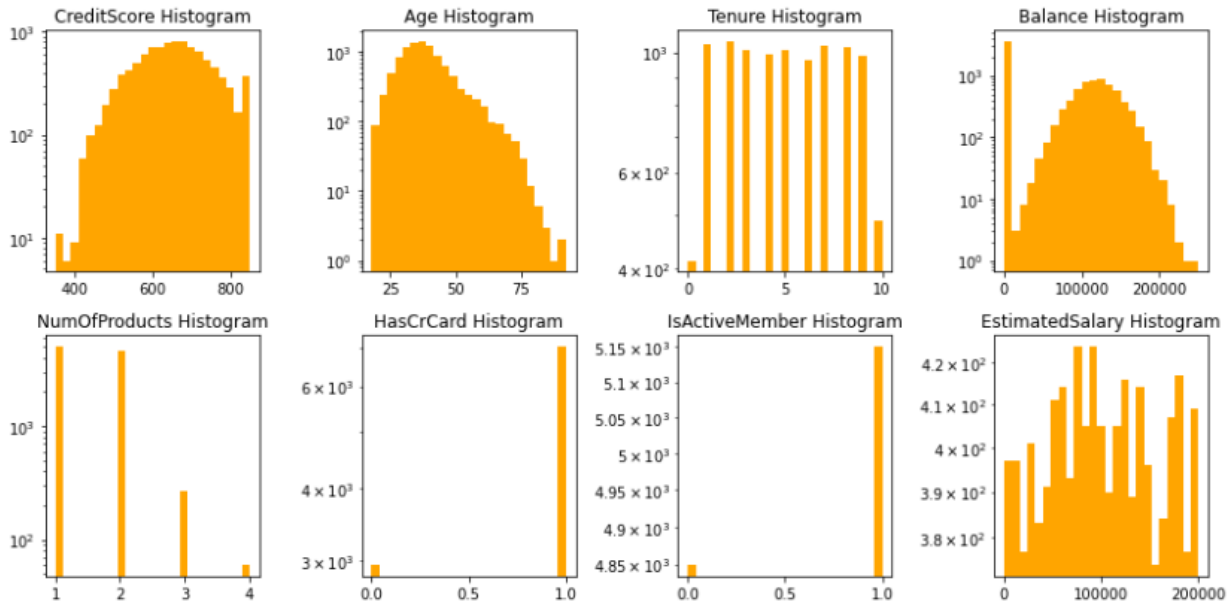
Figure 2

Summary statistics for the Churn data set

	count	mean	std	min	25%	50%	75%	max
CreditScore	10000.0	650.528800	96.653299	350.00	584.00	652.000	718.0000	850.00
Age	10000.0	38.921800	10.487806	18.00	32.00	37.000	44.0000	92.00
Tenure	10000.0	5.012800	2.892174	0.00	3.00	5.000	7.0000	10.00
Balance	10000.0	76485.889288	62397.405202	0.00	0.00	97198.540	127644.2400	250898.09
NumOfProducts	10000.0	1.530200	0.581654	1.00	1.00	1.000	2.0000	4.00
HasCrCard	10000.0	0.705500	0.455840	0.00	0.00	1.000	1.0000	1.00
IsActiveMember	10000.0	0.515100	0.499797	0.00	0.00	1.000	1.0000	1.00
EstimatedSalary	10000.0	100090.239881	57510.492818	11.58	51002.11	100193.915	149388.2475	199992.48
Exited	10000.0	0.203700	0.402769	0.00	0.00	0.000	0.0000	1.00

Figure 3

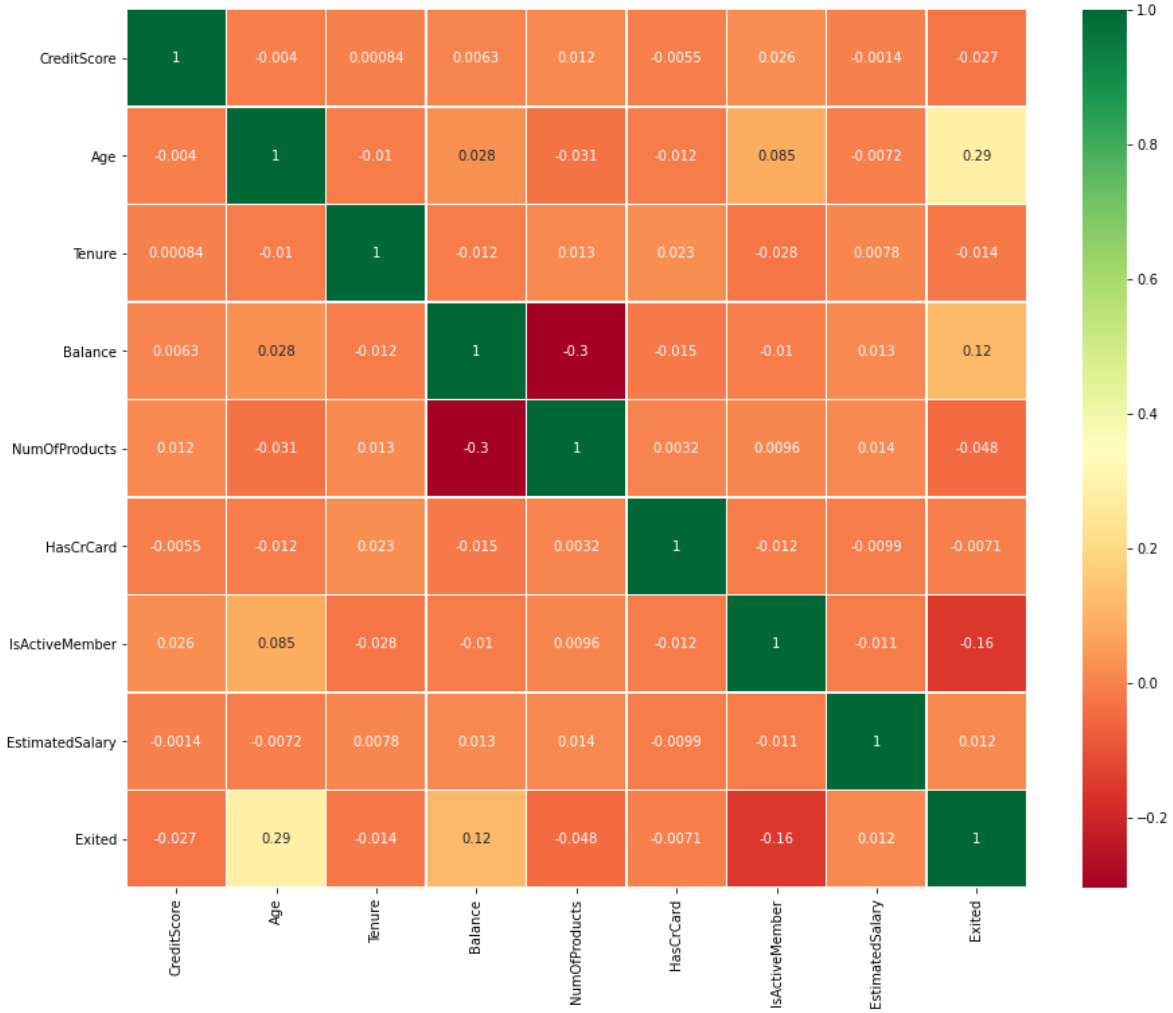
Histrogram of variables under investigation in the churn data set



Creating a correlation matrix of the variables under study we can already see that some variables show a higher degree of correlation with the "Exited" variable that denotes a customer who churned, leaving for a competitor bank. A customer's age, balance, and estimated salary show a positive correlation, while the remaining variables show a negative correlation. While this correlation isn't necessarily an indication of causation, it does provide some interesting insights worth exploring, which may help shed light on the null hypothesis which says that all factors have an equal effect on customer churn.

Figure 4

Correlation matrix heat map of the variables under study



The next step is to build different models and test them on the churn.csv data set to ascertain which one is superior with respect to its predictive power. For this research, three models are applied to the data set and compared for predictive accuracy: Decision Tree, Gradient Boosting, and Multiple Linear Regression. To build and deploy these models the academic version of SAS Enterprise Miner is used for its simple drag and drop capabilities as well as its processing power. After importing the file and dropping the random variables (Surname, CustomerID, and RowNumber), the binary “Exited” variable is selected as the target variable.

Figure 5

Variables under consideration and Exited as target variable

(none) | not | Equal to | Apply | Reset

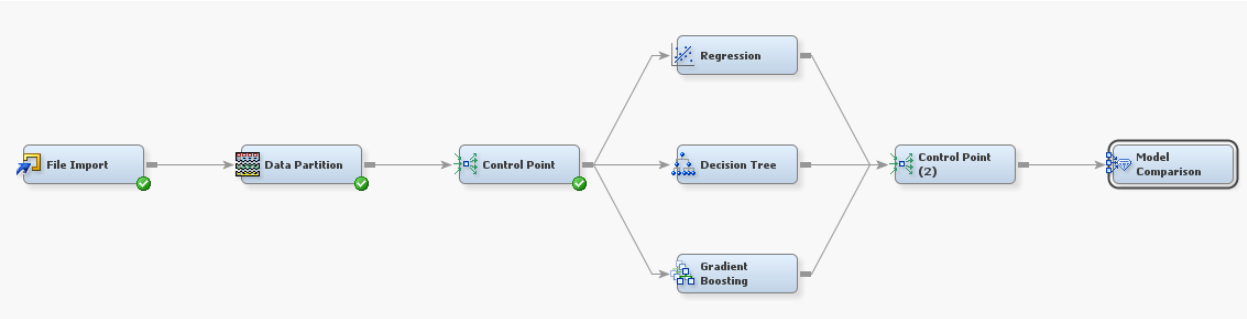
Columns: Label | Mining | Basic | Statistics

Name	Role	Level	Report	Order	Drop	Lower Limit	Upper Limit
Age	Input	Interval	No		No	.	.
Balance	Input	Interval	No		No	.	.
CreditScore	Input	Interval	No		No	.	.
CustomerId	Input	Interval	No		Yes	.	.
EstimatedSalary	Input	Interval	No		No	.	.
Exited	Target	Interval	No		No	.	.
Gender	Input	Nominal	No		No	.	.
Geography	Input	Nominal	No		No	.	.
HasCrCard	Input	Interval	No		No	.	.
IsActiveMember	Input	Interval	No		No	.	.
NumOfProducts	Input	Interval	No		No	.	.
RowNumber	Input	Interval	No		Yes	.	.
Surname	Input	Nominal	No		Yes	.	.
Tenure	Input	Interval	No		No	.	.

The next step is to partition the data into training and validation sets, using a 60/40 proportion respectively. This will allow for training and validation for each one of the three models in this study. A control point node is included for convenience so that multiple models can be run at once.

Figure 6

Decision Tree, Gradient Boosting, and Multiple Linear Regression methods in the Diagram



The decision tree model is set with a maximum branch, leaf size, and depth of 3 for consistency. The gradient boosting technique uses a partitioning algorithm for optimal partitioning of the data for a single target variable, in this case, “Exited”, which denotes customers who left the bank (SAS Institute Inc, 2018). The regression model is set with stepwise variable selection, so the variables with the lowest predictive valuer are eliminated to arrive at the best possible model.

The results of each model are listed in Table 1, showing the top variables in each model are Age, Number of Products, Balance, IsActiveMember, and Geography. Age is the only variable that is included in the top 3 in all models, making it a top candidate for further exploration of a potential relationship between age and propensity to churn.

Table 1

Results comparison of Gradient Boosting, Decision Tree, and Regression models

Gradient Boosting

Variable Importance

Obs	NAME	LABEL	NRULES	IMPORTANCE	VIMPORTANCE	RATIO
1	Age		104	1.00000	0.88481	0.88481
2	NumOfProducts		35	0.86880	1.00000	1.15101
3	Balance		121	0.61710	0.27873	0.45167
4	IsActiveMember		37	0.59491	0.47655	0.80105
5	Geography		31	0.40363	0.24172	0.59887
6	EstimatedSalary		78	0.40234	0.08521	0.21179
7	CreditScore		77	0.40161	0.14685	0.36565
8	Gender		15	0.18513	0.11496	0.62099
9	Tenure		20	0.18418	0.01094	0.05938

Decision Tree

Variable Importance

Variable Name	Label	Number of Splitting Rules	Importance	Validation Importance	Ratio of Validation to Training Importance
NumOfProducts		1	1.0000	1.0000	1.0000
Age		3	0.8910	0.8126	0.9120
IsActiveMember		3	0.5107	0.4590	0.8987
Balance		3	0.2261	0.2047	0.9055
Geography		1	0.2105	0.2197	1.0440

Regression

Summary of Stepwise Selection

Step		Effect Entered	DF	Number In	F Value	Pr > F
1	•	Age	1	1	496.60	<.0001
2	•	IsActiveMember	1	2	228.97	<.0001
3	•	Geography	2	3	98.27	<.0001
4		Gender	1	4	70.64	<.0001
5		Balance	1	5	19.18	<.0001
6		NumOfProducts	1	6	3.94	0.0471
7		Tenure	1	7	3.90	0.0484

The final step is to compare the three models under consideration for fit. The Model Comparison node, as shown in Table 1, provides fit statistics to select a "Champion Model" for the target variable Exited. The Gradient Boosting model shows the lowers Valid: Average Squared Error (VASE) score at 0.10510 making it the best model fit for our data set.

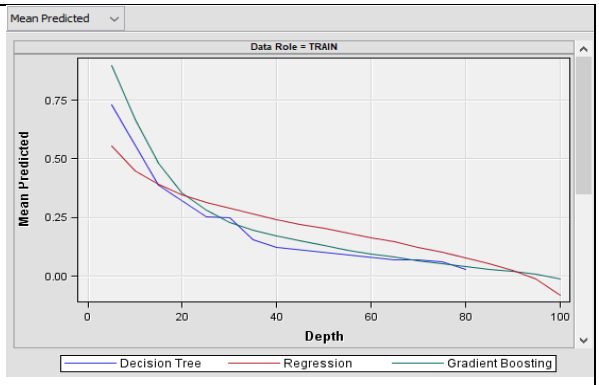
Figure 7

Model selection results Valid: Average Squared Error scores and Mean Predicted chart

Fit Statistics

Model Selection based on Valid: Average Squared Error (_VASE_)

Selected Model	Model Node	Model Description	Valid: Average Squared Error	Train: Average Squared Error
Y	Boost	Gradient Boosting	0.10510	0.08902
	Tree	Decision Tree	0.11289	0.11155
	Reg	Regression	0.13987	0.13661



CONCLUSION

The Gradient Boosting model is the best fitting for the churn.csv data set, and the top variables in that model are Age, Number of Products, and Balance. With this information, we reject both null hypotheses because certain variables appear to have more incidence over churning customers, and the combination of variables in the Gradient Boosting model seems to have predictive power over which customer characteristics make an existing customer more likely to defect.

Retaining profitable customers is a key competency for all financial services institutions in today's tech-enabled hyper-competitive environment. The gradient boosting model and individual customer characteristics identified in this study can inform marketing campaigns to retain, engage, and win back customers who may be considering leaving the bank or have left and may consider returning.

RECOMMENDATIONS

Based on the specific findings of this research, the bank should concentrate on further exploring the relationship between the top variables in the grading boosting model and the customer propensity to churn. For example, looking into which age group is more likely to churn, the balance held, and the number of products the customer has with the bank could yield valuable insights leading to the creation of a list of customers that could be proactively engaged before moving their accounts to a competitor (Park & Yoon, 2022).

Furthermore, understanding which customers the bank wants to retain would add another layer of value to any churn predicting model. Studies show that sometimes marketing campaigns to retain or entice clients towards a certain product or service can backfire and have the opposite effect, incentivizing the client to defect to a competitor instead

Retaining customers who are profitable to the bank should be the ultimate goal, while at the same time not including in the retention effort customers who are not, and have no prospect of becoming, profitable. This type of analysis can help the bank gain a competitive edge and focus scarce marketing dollars on retaining the right clients. Finally, it is important to acknowledge the limitations of this research, and look deeper into the four prediction categories that appear in the research literature: customer behavior, customer perceptions, customer demographics, and macroenvironment (van den Poel & Larivière, 2004, p. 215).

Going beyond churn analysis and customer retention and using a holistic customer-base management approach should be a priority for the bank, including understanding the life cycle of an existing customer, which transactions may be a predictor or certain actions such as churning to a competitor, purchasing a new product, opening a new account, or make a referral (Valendin et al., 2022).

REFERENCES

Amin, A., Shah, B., Khattak, A. M., Lopes Moreira, F. J., Ali, G., Rocha, A., & Anwar, S. (2019). Cross-company customer churn prediction in telecommunication: A comparison of data transformation methods. *International Journal of Information Management, 46*, 304–319. https://doi.org/10.1016/j.ijinfomgt.2018.08.015

Kim, S., & Lee, H. (2022). Customer churn prediction in Influencer Commerce: An application of decision trees. *Procedia Computer Science*, *199*, 1332–1339. https://doi.org/10.1016/j.procs.2022.01.169

McKechnie, S. (1992). Consumer buying behaviour in financial services: an overview. *International Journal of Bank Marketing*, *10*(5), 5–39. https://doi.org/10.1108/02652329210016803

Tang, L., Thomas, L., Fletcher, M., Pan, J., & Marshall, A. (2014). Assessing the impact of derived behavior information on customer attrition in the Financial Service Industry. *European Journal of Operational Research, 236(2),* 624–633. https://doi.org/10.1016/j.ejor.2014.01.004

Ahn, Y., Kim, D., & Lee, D.-J. (2019). Customer attrition analysis in the securities industry: A large-scale field study in Korea. *International Journal of Bank Marketing*, *38*(3), 561–577. https://doi.org/10.1108/ijbm-04-2019-0151

de Lima Lemos, R. A., Silva, T. C., & Tabak, B. M. (2022, March 6). *Propension to customer churn in a financial institution: A machine learning approach - neural computing and applications*. SpringerLink. https://link.springer.com/article/10.1007/s00521-022-07067-x

SAS Institute Inc. (2018). *Getting Started with SAS® Enterprise Miner™ 15.1*.

van den Poel, D., & Larivière, B. (2004). Customer attrition analysis for financial services using proportional hazard models. *European Journal of Operational Research, 157*(1), 196–217. https://doi.org/10.1016/s0377-2217(03)00069-9

Park, C. H., & Yoon, T. J. (2022). The dark side of up-selling promotions: Evidence from an analysis of cross-brand purchase behavior. *Journal of Retailing*. https://doi.org/10.1016/j.jretai.2022.03.005

Valendin, J., Reutterer, T., Platzer, M., & Kalcher, K. (2022). Customer base analysis with recurrent neural networks. *International Journal of Research in Marketing*. https://doi.org/10.1016/j.ijresmar.2022.02.007