Banking Customer Churn Analysis Report

12/08/2024

GROUP – 3

Banking Customer Churn Analysis Report

Introduction

The banking industry has seen significant growth, partly increased by advances in technology and shifting consumer preferences. As more customers demand personalized services and flexibility, understanding customer churn – when customers leave a bank – has become crucial. High churn rates can lead to major financial losses as acquiring new customers is more expensive than retaining existing ones. To address this, banks need to focus on identifying the reasons behind customer departure and take proactive steps to reduce churn, thus improving customer satisfaction and long-term profitability.

This project aims to predict which customers are likely to leave the bank by analyzing data from 10,000 customers. The key goal is to use data-driven methods to identify customers at risk of churn so that targeted retention strategies can be implemented. By doing so, banks can focus their efforts on high-risk customers and improve retention rates, reducing costs and increasing revenue.

Our dataset, "Bank Customer Churn Prediction," contains a mix of categorical and numerical variables that provide insights into customer behavior. The categorical variables include Geography, Gender, HasCrCard (whether the customer has a credit card), IsActiveMember (whether the customer is actively using the bank's services), and Exited (the target variable showing whether a customer has churned). The numerical variables include CreditScore, Age, Tenure (the number of years a customer has been with the bank), Balance, NumOfProducts (the number of products a customer holds with the bank), and EstimatedSalary.

By performing statistical analyses and building predictive models using tools like SAS, we aim to uncover the factors that influence customer churn. These factors can help the bank create better customer retention strategies tailored to different customer segments.

The dataset used for this project is from Kaggle, available at:
Bank Customer Churn Prediction from Kaggle.

In the next sections, we will dive into the details of our analysis and explain how each factor impacts customer churn. The goal is to provide actionable insights that can guide the bank in reducing churn and boosting customer loyalty.

Research Question

1. Does being an active member have a significant difference in credit scores compared to inactive members?

Compare the average CreditScore of active members (IsActiveMember = 1) and inactive members (IsActiveMember = 0) to determine if the difference is statistically significant.

2. Differences in Balance Across Geography Groups?

Use a one-way ANOVA to test whether the average Balance differs significantly across the geography groups.

3. Is there a relationship between balance and estimated salary?

Use Linear Regression and correlation analysis to examine the relationship between these numerical variables.

4. What are the key factors influencing customer churn?

Perform Logistic Regression to quantify the effect of independent variables (e.g., age, tenure, balance, geography) on the binary outcome (Exited). Use Odds Ratios and p-values to identify the most significant predictors.

4(a). Is there a significant interaction between gender and geography in determining churn?

Include an Interaction Term (e.g., Gender × Geography) in a Logistic Regression model to assess its statistical significance. Use Likelihood Ratio Tests to compare models with and without the interaction term.

ANALYSIS DESCRIPTIONS:

1. Analysis of Credit Score Differences by Active Membership Status

Our analysis examined whether being an active member (IsActiveMember = 1) was associated with significantly different credit scores compared to inactive members (IsActiveMember = 0). Using a two-sample t-test, we identified a small but statistically significant difference in credit scores between the two groups.

The average credit score for active members was 652.9, compared to 648.0 for inactive members. This difference of 4.96 points, while statistically significant, has minimal practical implications, as it represents only a marginal variation in creditworthiness.

Key Findings:Average Credit Scores:

Active Members: 652.9

Inactive Members: 648.0

Difference: 4.96 points ($t = -2.57$, $p = 0.0103$).

Variance Equality:

The equality of variances test ($F = 1.05$, $p = 0.1167$) confirmed that the variances of the two groups were not significantly different, allowing for pooled variance estimation in the t-test.

Statistical Significance:

The difference between the groups was statistically significant at the 0.05 level ($p = 0.0103$).with a 95% confidence.

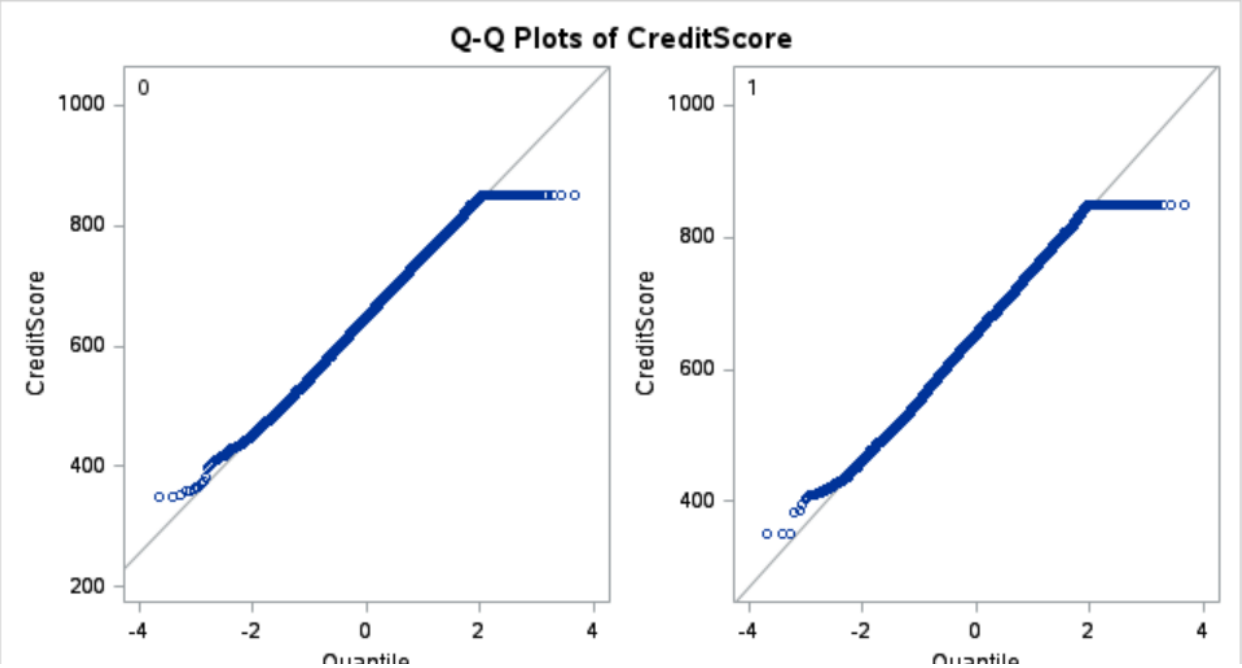
IsActiveMember	Method	N	Mean	Std Dev	Std Err	Minimum	Maximum
0		4849	648.0	97.7252	1.4034	350.0	850.0
1		5151	652.9	95.5804	1.3318	350.0	850.0
Diff (1-2)	Pooled		-4.9606	96.6263	1.9334		
Diff (1-2)	Satterthwaite		-4.9606		1.9347		

IsActiveMember	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
0		648.0	645.2	650.7	97.7252	95.8183	99.7101
1		652.9	650.3	655.5	95.5804	93.7698	97.4628
Diff (1-2)	Pooled	-4.9606	-8.7505	-1.1707	96.6263	95.3055	97.9845
Diff (1-2)	Satterthwaite	-4.9606	-8.7530	-1.1682			

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	9998	-2.57	0.0103
Satterthwaite	Unequal	9930.3	-2.56	0.0104

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	4848	5150	1.05	0.1167

And considering the Q-Q plot which looks good and normal with tightly packed on the line.



Business Implications: The credit score difference of 4.96 points is negligible in terms of real-world creditworthiness assessments. Banks should not consider active membership as a strong determinant of credit scores.

While credit score differences are minimal, the higher average score for active members might reflect broader financial engagement. Encouraging membership programs could indirectly foster financial responsibility among customers.

2. **Analysis of Balance Differences Across Geography Groups**

Our one-way ANOVA analysis assessed whether average account balances significantly differed across three geographic groups: France, Germany, and Spain. The results strongly indicate that geography plays a significant role in determining customer balances, with the differences being highly statistically significant ($F = 958.43$, $p < 0.0001$).

Key Findings:

Geographic Group Differences:

France: Average Balance = 62,092.64, Standard Deviation = 64,133.57

Germany: Average Balance = 119,730.12, Standard Deviation = 27,022.01

Spain: Average Balance = 61,818.15, Standard Deviation = 64,235.56

The results reveal that German customers have significantly higher average balances compared to customers from France and Spain, whose balances are nearly identical.

Model Statistics:

R-Square = 0.1609: The model explains 16.09% of the variance in account balances, suggesting other unexamined factors contribute to balance variability.

Least Square Means for effect Geography:

German customers have significantly higher balances than both French ($p < 0.0001$) and Spanish customers ($p < 0.0001$).

No significant difference between French and Spanish customers ($p = 0.9791$).

Dependent Variable: Balance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	6.2636257E12	3.1318129E12	958.43	<.0001
Error	9997	3.2666843E13	3267664559		
Corrected Total	9999	3.8930468E13			

R-Square	Coeff Var	Root MSE	Balance Mean
0.160893	74.73730	57163.49	76485.89

Source	DF	Type I SS	Mean Square	F Value	Pr > F
Geography	2	6.2636257E12	3.1318129E12	958.43	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Geography	2	6.2636257E12	3.1318129E12	958.43	<.0001

Least Squares Means
Adjustment for Multiple Comparisons: Tukey-Kramer

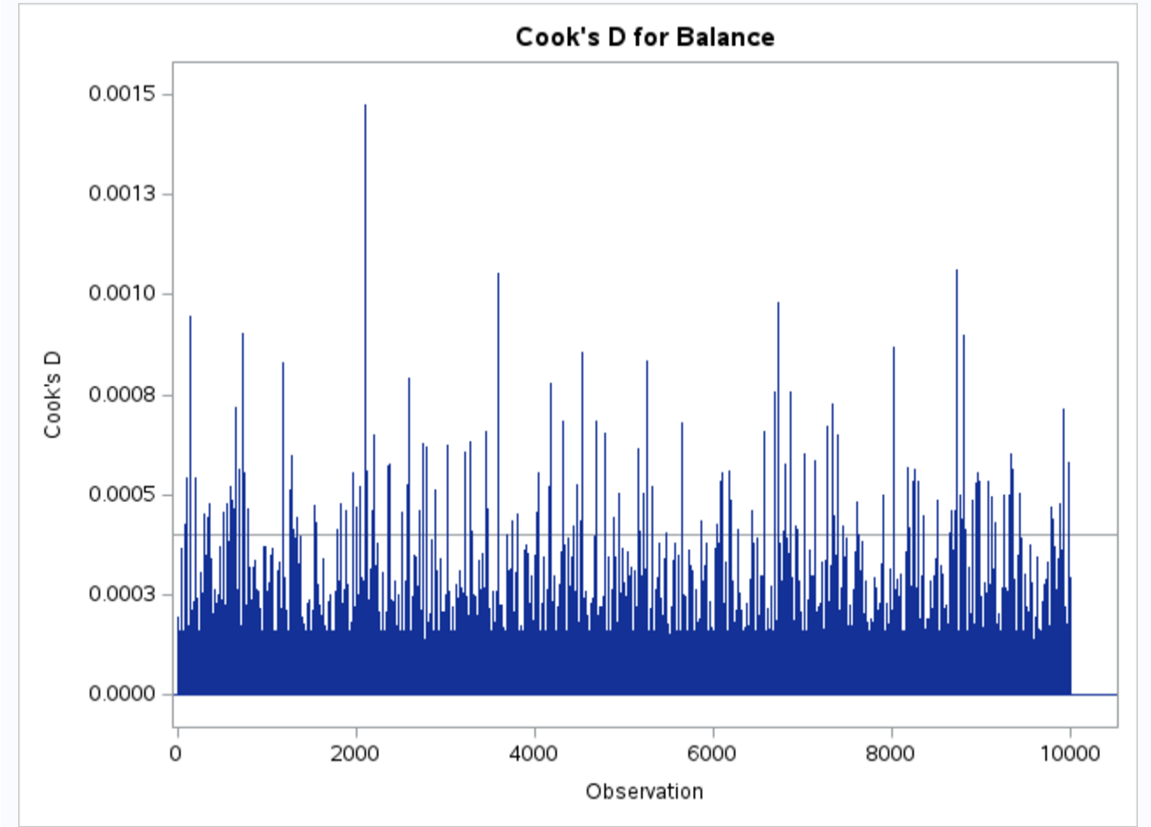
Geography	Balance LSMEAN	LSMEAN Number
France	62092.637	1
Germany	119730.116	2
Spain	61818.148	3

Least Squares Means for effect Geography
Pr > |t| for H0: LSMean(i)=LSMean(j)

Dependent Variable: Balance

i/j	1	2	3
1		<.0001	0.9791
2	<.0001		<.0001
3	0.9791	<.0001	

Cook's D graphs for balances has many outliers but that doesn't effect model because **there all below "1".**



Business Implications:

French and Spanish customers, with their comparable balances, could benefit from targeted campaigns to increase savings or promote higher-value account types.

The high variability within the French and Spanish groups highlights the need for personalized financial products. Understanding individual financial behaviors could improve engagement and profitability.

3. Balance and Salary Relationship

Contrary to what might be expected, we found practically no relationship between customers' account balances and their estimated salaries. The correlation was negligibly small (r = 0.0128) and not statistically significant. The linear regression analysis confirmed this lack of relationship, explaining only 0.02% of the variance in account balances. This surprising finding suggests that high-earning customers don't necessarily maintain higher balances of 81.58% (Coeff var), and EstimatedSalary standard deviation of 57,510.49.

Pearson Correlation Coefficients, N = 10000
Prob > |r| under H0: Rho=0

	Balance	EstimatedSalary
Balance	1.00000	0.01280 0.2007
EstimatedSalary	0.01280 0.2007	1.00000

Root MSE	62395	R-Square	0.0002
Dependent Mean	76486	Adj R-Sq	0.0001
Coeff Var	81.57768		

The MEANS Procedure

Variable	N	Mean	Std Dev	Minimum	Maximum
CreditScore	10000	650.5288000	96.6532987	350.0000000	850.0000000
Age	10000	38.9218000	10.4878065	18.0000000	92.0000000
Balance	10000	76485.89	62397.41	0	250898.09
EstimatedSalary	10000	100090.24	57510.49	11.5800000	199992.48

4. Analysis of Customer Churn Factors

Our logistic regression analysis revealed compelling insights into customer churn patterns at the bank. The model demonstrated strong statistical significance and good predictive power, correctly classifying 76.7% of cases. The most striking finding was the impact of age on customer churn - for each year increase in age, customers became 7.5% more likely to leave the bank. This suggests that older customers require special attention in retention strategies.

Customer engagement emerged as another crucial factor. Non-active members were nearly three times more likely to churn compared to active members, highlighting the critical importance of maintaining customer engagement. Geographic differences also played a significant role, with German customers showing twice the likelihood of churning compared to Spanish customers, while French customers showed no significant difference from Spanish customers.

Class Level Information			
Class	Value	Design Variables	
Geography	France	1	0
	Germany	0	1
	Spain	0	0
Gender	Female	1	
	Male	0	
HasCrCard	0	1	
	1	0	
IsActiveMember	0	1	
	1	0	

Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-5.0057	0.2553	384.3397	<.0001
CreditScore		1	-0.00067	0.000280	5.6832	0.0171
Geography	France	1	-0.0352	0.0706	0.2486	0.6181
Geography	Germany	1	0.7395	0.0788	88.0361	<.0001
Gender	Female	1	0.5285	0.0545	94.0706	<.0001
Age		1	0.0727	0.00258	796.9204	<.0001
Tenure		1	-0.0159	0.00935	2.9067	0.0882
Balance		1	2.637E-6	5.142E-7	26.3001	<.0001
NumOfProducts		1	-0.1015	0.0471	4.6393	0.0312
HasCrCard	0	1	0.0447	0.0593	0.5668	0.4515
IsActiveMember	0	1	1.0754	0.0577	347.5682	<.0001
EstimatedSalary		1	4.807E-7	4.737E-7	1.0299	0.3102

Gender differences were notable, with female customers being 1.7 times more likely to leave the bank than male customers. While account balance showed statistical significance, its practical impact was minimal. Similarly, credit score had a statistically significant but small negative effect on churn probability. The number of products held by customers had a modest impact, with each additional product slightly reducing churn probability.

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
CreditScore	0.999	0.999	1.000
Geography France vs Spain	0.965	0.841	1.109
Geography Germany vs Spain	2.095	1.795	2.445
Gender Female vs Male	1.696	1.525	1.888
Age	1.075	1.070	1.081
Tenure	0.984	0.966	1.002
Balance	1.000	1.000	1.000
NumOfProducts	0.903	0.824	0.991
HasCrCard 0 vs 1	1.046	0.931	1.175
IsActiveMember 0 vs 1	2.931	2.618	3.282
EstimatedSalary	1.000	1.000	1.000

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	76.7	Somers' D	0.535
Percent Discordant	23.3	Gamma	0.535
Percent Tied	0.0	Tau-a	0.174
Pairs	16220631	c	0.767

Key findings from the analysis:

- Age is the strongest predictor ($\chi^2 = 796.92$, $p < .0001$) with an odds ratio of 1.075, indicating that for each year increase in age, the odds of churning increase by 7.5%
- Active Membership has a substantial impact ($\chi^2 = 347.57$, $p < .0001$). Non-active members are 2.93 times more likely to churn
- Geography shows significant variation ($\chi^2 = 145.20$, $p < .0001$). German customers are 2.095 times more likely to churn compared to Spanish customers
- Gender is significant ($\chi^2 = 94.07$, $p < .0001$) with females being 1.696 times more likely to churn than males
- Balance shows significance ($\chi^2 = 26.30$, $p < .0001$) but with minimal practical effect
- Credit Score has a small negative effect ($\chi^2 = 5.68$, $p = 0.0171$)
- Number of Products shows modest significance ($\chi^2 = 4.64$, $p = 0.0312$)
- Tenure, HasCrCard, and EstimatedSalary were not significant predictors ($p > 0.05$)

Type 3 Analysis of Effects			
Effect	DF	Wald Chi-Square	Pr > ChiSq
CreditScore	1	5.6832	0.0171
Geography	2	145.2001	<.0001
Gender	1	94.0706	<.0001
Age	1	796.9204	<.0001
Tenure	1	2.9067	0.0882
Balance	1	26.3001	<.0001
NumOfProducts	1	4.6393	0.0312
HasCrCard	1	0.5668	0.4515
IsActiveMember	1	347.5682	<.0001
EstimatedSalary	1	1.0299	0.3102

4(a). Is there a significant interaction analysis between gender and geography in determining the churn?

The analysis of how gender and geography jointly affect churn revealed interesting patterns. While both gender ($\chi^2 = 28.35$, $p < .0001$) and geographic location ($\chi^2 = 155.43$, $p < .0001$) independently influence churn rates significantly, there was no meaningful interaction between these factors ($\chi^2 = 1.33$, $p = 0.5133$). This suggests that the effect of gender on churn remains consistent across different geographic regions, and vice versa. The model showed moderate predictive ability, correctly classifying 63.2% ($c = 0.632$) of cases. However, simpler model without interaction might be more appropriate.

Joint Tests			
Effect	DF	Wald Chi-Square	Pr > ChiSq
Geography	2	155.4338	<.0001
Gender	1	28.3537	<.0001
Geography*Gender	2	1.3339	0.5133

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	54.7	Somers' D	0.265
Percent Discordant	28.2	Gamma	0.320
Percent Tied	17.1	Tau-a	0.086
Pairs	16220631	c	0.632

Customer Demographics Overview

The bank's customer base shows interesting demographic patterns. The average customer is in their late thirties (mean age 38.92 years), with substantial variation in age (SD = 10.49 years). Credit scores cluster around 650, with moderate variation. Account balances average €76,486 but show considerable variation (SD = €62,397), while estimated salaries center around €100,090 with similar variability.

The MEANS Procedure

Variable	N	Mean	Std Dev	Minimum	Maximum
CreditScore	10000	650.5288000	96.6532987	350.0000000	850.0000000
Age	10000	38.9218000	10.4878065	18.0000000	92.0000000
Balance	10000	76485.89	62397.41	0	250898.09
EstimatedSalary	10000	100090.24	57510.49	11.5800000	199992.48

The customer base is slightly skewed toward male customers (54.57%) and is geographically concentrated in France (50.14%), with Germany and Spain each representing about a quarter of customers. The overall churn rate of 20.37% indicates that roughly one in five customers leaves the bank, highlighting the importance of effective retention strategies.

The FREQ Procedure

Geography	Frequency	Percent	Cumulative Frequency	Cumulative Percent
France	5014	50.14	5014	50.14
Germany	2509	25.09	7523	75.23
Spain	2477	24.77	10000	100.00

Gender	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Female	4543	45.43	4543	45.43
Male	5457	54.57	10000	100.00

Exited	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	7963	79.63	7963	79.63
1	2037	20.37	10000	100.00

These findings provide valuable insights for developing targeted customer retention strategies, particularly focusing on older customers, inactive members, and specific geographic regions where churn rates are higher.

Conclusion

The analysis reveals several critical insights that can inform strategic decision-making in customer retention efforts. Most notably, age emerges as a paramount factor in churn prediction, while customer engagement levels and geographic location also play substantial roles. The surprising disconnect between salary levels and account balances, combined with the inability to predict credit scores from available variables, suggests that traditional assumptions about customer behavior may need reassessment. Moving forward, these findings indicate that retention strategies should prioritize age-specific engagement programs, particularly targeting older customers, while also accounting for regional variations in customer behavior. Additionally, the marked difference in churn rates between active and non-active members underscores the need for proactive engagement initiatives. By implementing

targeted interventions based on these insights, the bank can work toward reducing its 20.37% churn rate and strengthening customer relationships across all segments of its diverse customer base.