**An Analysis of Bank Churn**

Daniel G. Barstow

Department of Computer Science and Information Systems

Elmhurst University

Dr. Jim Kulich

18 October 2022

**Relevant Figures**

|  |  |
| --- | --- |
| churn (0 or 1) | 0 |
| credit\_score | 652 |
| age | 37 |
| tenure | 5 |
| balance | €97,198.54 |
| products\_number (1,2,3, or 4) | 1 |
| credit\_card (0 or 1) | 1 |
| active\_member (0 or 1) | 1 |
| estimated\_salary | €100,193 |
| country (France, Spain, Germany) | France |
| gender (Male or Female) | Male |

Figure 1 – Median of each feature in the dataset.

Chart, waterfall chart

Description automatically generated

Figure 2 – Lower triangular correlation matrix of features in the data.

|  |  |
| --- | --- |
| H0 : pwomen = pmen  Ha : pwomen ≠ pmen | p-value = 2.2e-16 < 0.05  Reject H0 |
| H0 : pactive = pInactive  Ha : pactive ≠ pInactive | p-value = 2.2e-16 < 0.05  Reject H0 |
| H0 : pcredit\_card = pno\_credit\_card  Ha : pcredit\_card ≠ pno\_credit\_card | p-value = 0.4754 > 0.05  Fail to reject H0 |
| H0 : pGermany = pSpain = pFrance  Ha : at least one of the proportions is different | p-value = 2.2e-16 < 0.05  Reject H0 |
| H0 : p1 = p2 = p3 = p4  Ha : at least one of the proportions is different | p-value = 2.2e-16 < 0.05  Reject H0 |
| H0 : p1 = p2 = p3 = … = p10  Ha : at least one of the proportions is different | p-value = 0.1776 > 0.05  Fail to reject H0 |
| H0 : p350 = p400 = p450 = … = p850  Ha : at least one of the proportions is different | p-value = 1.5e-13 < 0.05  Reject H0 |
| H0 : p18 = p24 = p30 = … = p90  Ha : at least one of the proportions is different | p-value = 2.2e-16 < 0.05  Reject H0 |
| H0 : p€0 = p€16k = p€33k = … = p€234k  Ha : at least one of the proportions is different | p-value = 2.2e-16 < 0.05  Reject H0 |
| H0 : p€11.58 = p€16k = p€33k = … = p€183k  Ha : at least one of the proportions is different | p-value = 0.8276 > 0.05  Fail to reject H0 |

Figure 3 – Null, Alternative hypotheses, and p-values for proportions tests about churn.

|  |  |
| --- | --- |
| H0 : pwomen = pmen  Ha : pwomen ≠ pmen | p-value = 0.02417 < 0.05  Reject H0 |
| H0 : pcredit\_card = pno\_credit\_card  Ha : pcredit\_card ≠ pno\_credit\_card | p-value = 0.2354 > 0.05  Fail to reject H0 |
| H0 : pGermany = pSpain = pFrance  Ha : at least one of the proportions is different | p-value = 0.07049 > 0.05  Fail to reject H0 |
| H0 : p1 = p2 = p3 = p4  Ha : at least one of the proportions is different | p-value = 0.0006447 < 0.05  Reject H0 |
| H0 : p1 = p2 = p3 = … = p10  Ha : at least one of the proportions is different | p-value = 0.1716 > 0.05  Fail to reject H0 |
| H0 : p350 = p400 = p450 = … = p850  Ha : at least one of the proportions is different | p-value = 0.1232 > 0.05  Fail to reject H0 |
| H0 : p18 = p24 = p30 = … = p90  Ha : at least one of the proportions is different | p-value = 2.2e-16 < 0.05  Reject H0 |
| H0 : p€0 = p€16k = p€33k = … = p€234k  Ha : at least one of the proportions is different | p-value = 0.6793 > 0.05  Fail to reject H0 |
| H0 : p€11.58 = p€16k = p€33k = … = p€183k  Ha : at least one of the proportions is different | p-value = 0.1258 > 0.05  Fail to reject H0 |

Figure 4 - Null, Alternative hypotheses, and p-values for proportions tests about banking activity.

Graphical user interface, application

Description automatically generated

Figure 5 – A dashboard in Power BI of the relevant features related to churn.

Graphical user interface, application

Description automatically generated

Figure 6 – A dashboard in Power BI of the relevant features related to banking activity.

**Abstract**

An exploratory data analysis was performed with data from a European bank to determine relevant factors contributing to churn and frequent banking activity. Features were engineered and visuals were created using Power BI and statistical tests were conducted using R. Two target demographics were identified to reduce churn: young women from Germany with low credit scores who had an account balance of €0 and owned just 1 financial product, and middle-aged women from Germany with average credit scores, account balances around €100,000, and owned 3 or 4 financial products. It was further found that older men who owned 2 financial products were most likely to be active in their banking activity, thus unlikely to churn. Suggestions for increasing engagement among these customers were given to reduce churn.

**Background**

Broadly speaking, banks profit by taking their customers’ money in the form of low interest saving accounts so that they can invest or loan it out at a higher interest rate. The banks are happy since they have a stockpile of cash to do with what they please, and customers are happy because they can rest easy knowing that their money is safe. Churn refers to a decrease in banking transactions or a loss of customers who use the bank [1]. Since the cost associated with sourcing new customers is higher than the cost of retaining customers, banks need to worry about keeping the customers they already have [2]. This means banks want to reduce churn as much as possible.

**Methodology**

The data used for this project was collected from Kaggle’s dataset library [3]. It contains 12 features: customer\_id, credit\_score, country, gender, age, tenure, balance, products\_number, credit\_card, active\_member, estimated\_salary, and churn. The customer\_id feature went unused in the analysis because it was determined that no information could be discerned from it—it’s merely an identifier for a database. There are 3 values for the country feature: France, Spain, and Germany. The tenure feature is the number of years the person has been at that particular bank. The products\_number feature is the number of financial products the person has purchased from the bank. The credit\_card, active\_member, and churn features are binary variables telling whether the person has a credit card or not, whether they make frequent transactions or not, and whether or not they churned.

The dataset did not contain any missing values, thus did not require any cleaning. Power BI was used for feature engineering and for creating visualizations. Since age, balance, credit\_score, and estimated\_salary, were all continuous features, they were binned to help identify outliers. Credit\_score ranged from 350 to 850 and was binned in increments of 50. Age ranged from 18 to 92 and was binned in increments of 6 years. Balance ranged from €0 to roughly €234,000 and was binned into 15 groups in increments of roughly €16,000. Estimated\_salary ranged from €11.58 to roughly €183,000 and was binned into 12 groups in increments of roughly €16,000.

Basic descriptive statistics were calculated using the Python library Pandas in Jupyter notebook. The median for each feature is shown in fig. 1. A heatmap correlation matrix was created using the Python library Seaborn to determine any correlation between the features. Nearly all of the features were weakly correlated with Pearson coefficients in the range (-0.3, 0.29). The strongest positive correlation was between age and churn with a value of r = 0.29. The strongest negative correlation was between products\_number and balance with a value of r = -0.3. Churn and active\_member were weakly negatively correlated with a value of r = -0.16. The correlation matrix generated is shown in fig. 2.

Hypothesis tests were carried out using R [4]. Since the features were all categorical in nature, proportions tests were used to determine if there were significant differences in the proportions of people in each category with respect to churn and then banking activity. A z-test was used for the 2-category tests and ANOVA was used for 3 or more category tests. A 95% confidence interval was used, meaning that a p-value greater than 0.05 indicated that there was enough evidence to reject the null hypothesis.

A table of the null and alternative hypotheses along with their p-values for the churn proportions tests are shown in fig. 3. A similar table for the activity proportions are shown in fig. 4. Power BI was used to create visuals incorporating the features deemed significant by the hypothesis tests. A dashboard of the relevant features related to churn is shown in fig. 5. A similar dashboard of relevant features related to banking activity is shown in fig. 6.

**Results and Discussion**

Through this analysis, significant differences in the proportions of churn were found based on gender, banking activity, country of residence, number of financial products owned, credit score, age, and account balance. Owning a credit card, tenure at the bank, and estimated salary were deemed not to have any bearing on churn. The data suggests that: women are more likely to churn than men, people from Germany are more likely to churn than people from France or Spain, active bankers are less likely to churn than inactive bankers, people who own 2 financial products are less likely to churn, people with low to average credit scores are more likely to churn, people between the ages of 40 and 60 are more likely to churn, and people who have an account balance of either €0 or around €100,000 are more likely to churn.

Some of these findings are intuitive. Since churn is defined by customer attrition, it would make sense that active bankers are less likely to churn than inactive ones. Lower credit scores being more likely to churn also makes sense because credit score is a measure of the likelihood to repay debts. If someone can’t repay their debts, then they’re more likely to declare bankruptcy which affects the churn rate. People who have €0 in their account have the largest amount of churn across all balances which also makes sense because banks usually charge fees for not carrying a balance and can lead to overdraft fees if they spend more than they have. This is likely true even with the sampling bias of more people having €0 than any other amount.

Other findings require some domain knowledge. In Europe, one can only get deposit insurance for up to €100,000 [5]. This is likely the reason why the distribution for account balance and churn is centered around €100,000 and why most churn occurs around this amount. People want to safeguard their money, so this is the threshold at which they decide to spread their money out across multiple banks. This could also be why people from ages 40 to 60 are more likely to churn, because they have lived long enough to accumulate enough money to want to spread it out and decrease their susceptibility to risk. Another less intuitive finding is that people who own 2 financial products are less likely to churn than people who have 1, 3, or 4 financial products. This could be because people only have a checking account are constantly depositing and withdrawing money which is more precarious than people who have both a checking and savings account. As for the people who have 3 or 4 financial products, they could have home equity loans, personal loans, or mortgages. These products carry more risk than a simple checking or savings account. If the economy tanked like it did in 2008, these people would look to refinance or risk churning from the bank, as many did then.

Further, significant differences in the proportions of activity were found based on gender, the number of financial products owned, and age. Whereas owning a credit card, residing in a particular country, tenure at the bank, credit score, account balance, and estimated salary were deemed to not have any bearing on banking activity. Since it was determined that women are more likely to churn than men, it makes sense that men are more active in their banking activity than women because of the way churn is defined. The data suggests that the older people get, the more active they become in their banking activity. This makes sense because as people reach retirement age, they rely more on the money they have accumulated over the course of their lives. The data also suggests that people with 3 or 4 financial products become less active in their banking activity. This could be because of the type of product they buy. Certificates of deposit (COD) are products which are similar to savings accounts in that they accrue interest but do not allow for withdrawals, which would lead to a decrease in banking activity as the person waits for the COD to mature.

In summary, there are 2 types of customers who were identified to be most likely to churn and be less active in their banking. The first is young women from Germany with small account balances who have only checking accounts and low credit scores. The second is middle-aged women from Germany with average credit scores who own 3 or 4 financial products and have account balances of around €100,000. Based on this analysis, targeting these customers should lead to increased activity and reduced churn.

**Suggestions for Improvement**

One suggestion for reducing churn would be to increase engagement among people who are inactive bankers with targeted marketing materials or surveys due to the inverse relationship between activity and churn. Another suggestion would be to engage people with lower credit scores by offering financial education courses to make them feel like the bank is invested in their success. Doing this could be mutually beneficial because as the bank seeks to reduce churn by increasing customer engagement, customers’ credit scores could increase, allowing them to become more financially literate and make better choices. Further, since people with 3 or 4 financial products are less likely to be active, executives will need a cost-benefit analysis to determine if they should allow for some churn to see if the revenue these customers generate is proportional to the cost of sourcing new customers.

**Acknowledgements**

I would like to thank Dr. Jim Kulich for his advice regarding the hypothesis tests, Stephen Gnidovec for his feedback on the visualizations, Alex Martinez for his suggestion to focus on banking activity in addition to churn, and Brian Tammaro for his advice and kind words of encouragement during discussions.

**References**

[1] <https://www.investopedia.com/terms/c/churnrate.asp>

[2] [https://www.forbes.com/sites/forbestechcouncil/2020/05/27/why-retaining-customers-for- banks-is-as-important-as-winning-new-ones/?sh=574fcc5d3f98](https://www.forbes.com/sites/forbestechcouncil/2020/05/27/why-retaining-customers-for-%09banks-is-as-important-as-winning-new-ones/?sh=574fcc5d3f98)

[3] <https://www.kaggle.com/datasets/gauravtopre/bank-customer-churn-dataset>

[4] [https://sphweb.bumc.bu.edu/otlt/MPH- Modules/BS/R/R6\_CategoricalDataAnalysis/R6\_CategoricalDataAnalysis6.html](https://sphweb.bumc.bu.edu/otlt/MPH-%20%20%20%20%20%20%09Modules/BS/R/R6_CategoricalDataAnalysis/R6_CategoricalDataAnalysis6.html)

[5] [https://www.eba.europa.eu/regulation-and-policy/recovery-and-resolution/deposit-guarantee- schemes-data](https://www.eba.europa.eu/regulation-and-policy/recovery-and-resolution/deposit-guarantee-%09schemes-data)