Churn Prediction of Bank Customers

Andreea Ioana Chiriac[1]
[1] *The Bucharest University of Economic Studies, Bucharest, Romania.*
E-mail: chiriacandreea15@stud.ase.ro

Please cite this paper as:
Chiriac, A., 2021. Churn Prediction of Bank Customers. In: R. Pamfilie, V. Dinu, L. Tăchiciu, D. Pleșea, C. Vasiliu eds. 2021. *7th BASIQ International Conference on New Trends in Sustainable Business and Consumption.* Foggia, Italy, 3-5 June 2021. Bucharest: ASE, pp. 232-237
DOI: 10.24818/BASIQ/2021/07/030

Abstract

The term "churn" occurs in the situation of subscription products and symbolizes that customers are cancelling a service. Churn also applies to services and products that clients are reaching out with over a significant period of time. The main objective of this paper is to highlight the importance of churn prediction because in the last years, has been registered a large number of discretionary online services for customers and that means churn is a consistent and a perpetual problem that needs to be addressed. Churn prevention allows companies to elaborate loyalty programs and retention campaigns to keep their customers. Usually, the service companies concentrate on purchasing, but it is important in order to achieve success to minimize churn. If churn is not approached in a continued and proactive way, the service will not accomplish its full potential. In this research paper the main aim is to establish the determining factors of bank customers who decided to leave the bank or not. I use a binary logistic regression because it is essential to identify what leads a client towards the decision to leave the company. Binary logistic regression estimates the probability of an event occurring, in this case the probability that bank customers will leave the bank or not. The software used in order to express my findings about this topic is IBM SPSS Statistics.

Keywords

Churn, statistics, binary logistic regression, bank customers, data analysis.
DOI: 10.24818/BASIQ/2021/07/030

Introduction

What is Churn? We called "churn" when a client cancels a subscription or decides to give up using a service. Usually, the service companies concentrate on purchasing, but it is important in order to achieve success to minimize churn. If churn is not approached in a continued and proactive way, the service will not accomplish its full potential. The word "churn" came from the term "churn rate", which means the ratio of customers moving out in a particular interval of time. This action conducts to the client or user population changing in the long run. Initially, the word meant "to move about vigorously" (like in churning butter), but from a business point of view, churn can be used as a verb ("the customer is churning") and also can be used as a noun ("the customer is a churn"). On the other hand, the customers that are not churning from a service can be also considered in a positive sense. When this is happening, it is named customer retention.

Customer retention means maintaining customers using a service and also extending and renewing their subscription. So, customer retention is the opposite of churn. Decreasing churn is analog to increasing customer retention. When an objective is declared as retaining more clients for a longer period of time, then additionally saving clients that are at risk of churning, there should also be a priority on keeping the customers involved. Usually, there is the chance of upselling more advanced versions of the services for more money to the most involved clients. The important goals for services with continuous customer interactions are: boost engagement, saving churns and upsells. The difference between these is a matter of attention and not a difference in the intention.

Even though there are a lot of products and services with regular and frequent customers, there is one set of techniques for using data in order to fight churn and also increase retention, engagement and upsell. If there is a company that creates a subscription product, then it is recommended: a product or a service is provided and uses on a recurrent basis, customers need to interact with the product/service, clients may have subscriptions in order to receive a product or a service that cost money. Subscriptions can be ended or canceled, which is recognized as churn. If there are no subscriptions, a client churns when the product is not used anymore. The prices, timing and payments for customers and subscriptions (if any) are stored in a database, usually a transactional one. Also, when clients use a product or a service, it is stored in a data warehouse.

Essentially, directors have been expanding their view and reflection from believing that customers are gained and dedicated for a long period of time to the impression that suppliers re-win their clients' businesses daily. In the long term, this intends prioritizing and planning investments to secure long lasting customers retention and achievement, rather than just hoping that it will take place.

There are five primary strategies used by the companies in order to reduce churn: targeting purchasing, product improvement, customer relationship, engagement campaigns and reducing the prices or changing the subscription terms. All of these procedures are efficient if they are data driven.

One of the best method to be aware of customer churn is not to assume about it in terms of maintaining. Instead, it is better to consider churn as a learning moment and also as an occasion to find out how to prevent it. So, this type of unfavorable phenomenon can lead to a increased level of proactive involvement for the remaining clients and also the future ones. Occasionally, there is a "blind churn" and that means that there are clients that left the company even if this was totally unexpected. In this case, the company/provider need to discover the root cause. Sometimes, this type of churn appears in the situation where the left customers choose a competitor.

There are a lot of arguments behind leaving a company. Some of the frequent causes that determine customer churn are weak customer service, not getting value in the products and services, the deficit of communication and the lack of client loyalty. The initial step in order to maintain the customer is to check client churn over time. If the churn rate is usually increased, then it is recommended to dedicate some resources in order to enhance customer retention.

Churn rate formula is defined in equation 1:

$$Churn\ rate = \frac{Number\ of\ Churned\ Customers}{Start\ Customers} \tag{1}$$

Retention rate formula is defined in equation 2:

$$Retention\ Rate = \frac{Number\ of\ Retained\ Customers}{Start\ Customers} \tag{2}$$

Also, there is the following relationship between the rates:

$$100\ \% = Churn\ Rate + Retention\ Rate \tag{3}$$

In order to ensure that the client retention rate is improved, the top priority should be understanding the customers' needs. This can be done by reaching and surveying the clients that already churned. Another solution is contacting the existing customers and asking them about their requirements. For example, a data analytics approach would be to have a look into the data, in order to check how the clients services call logs are handled, how long their wait time was and also if their problems and concerns were solved. Performing this type of analysis on these data points can expose the issues that a company is facing in retaining their existing clients.

Review of the scientific literature

It is well-known that retaining clients with an increased churn risk is one of the hardest challenges (Miguéis, et al., 2012) because nowadays there are a large number of service and products providers, and customers have a lot of options to churn. Usually, clients tend to compare their providers with others and this leads to churning (Balle, et al., 2013). According to P. Kotler in 1994, the price of convincing a client not to churn to the opponent is 16 times less than the price of finding and determining contact with a new client. Also, the cost of convincing new clients is 5 to 6 times more than for maintaining the existing ones.

According to studies, it is approximated that a service supplier can increase their returns by between 25% and 85% by decreasing the customer churn rate by 5% (Reichheld and Sasser, 1990).

Churn affects businesses everywhere around the world and churn rates fluctuate often. Mobile phone companies in Europe have churn rates between 20% and 38%. Wireless business could improve their earnings by almost 10% if they took steps in order to reduce churn.

Research methodology

The database that I used in my research paper contains information about Churn Prediction of bank customers and was extracted from kaggle.ro. The dataset consists of approximately 10 000 records.

In order to establish the determining factors of bank customers who decided to leave the bank or not, I used a binary logistic regression because it is essential to identify what leads a client towards the decision to leave the company. Churn prevention provides companies to elaborate loyalty programs and retention campaigns to keep their customers. The dataset was imported in IBM SPSS Statistics and all the variables were coded accordingly.

The independent variables that were in used in my analysis are: Credit Score, Tenure, Balance, Number of Products and Estimated Salary and the binary dichotomous dependent variable is Exited. Credit Score can have an effect on customer churn, which means that a customer with an increased credit score is less likely to leave the bank. The variable “Tenure” relate to number of years that the bank customer has been a client. Typically, older clients are less likely to leave the bank. Another very good indicator of customer churn is balance because people with a higher balance in their accounts are less likely to leave the bank compared with those with lower balances. The variable “Number of Products” is related to the number of products that customer has bought through the bank. Like balance, estimated salary is also a good indicator because people with lower salaries are more likely to leave the bank compared with those with higher salaries. The binary variables “Exited” describes if a customer left the bank or not and was coded with 0 and 1. (1-Yes, 0-No).

Results and discussion

As it is already known, it is more expensive to sign in a new client than keeping an existing one. Binary logistic regression estimates the probability of an event occurring, in this case the probability that bank customers will leave the bank or not.

Table no. 1. Classification Table

	Observed		Predicted		
Step 0	**National Standards**		**Exited**		**Percentage correct**
	Exited		No	Yes	
		No	7963	0	100.0
		Yes	2037	0	.0
	Overall Percentage				79.6

Source: Author own research results

According to the classification table (table no. 1), the model always assumes "no" because there are more customers who do not leave the bank compared to those who leave. (7963 compared with 2037). The overall percentage tells us that this approach to prediction is correct with 79.6%, which is a good approximation.

Table no. 2. Variables in Equation

		B	S.E.	Wald	df	Sig.	Exp (B)
Step 0	**Constant**	-1.363	.025	3014.868	1	.000	.256

Source: Author own research results

The variables in the table of the equation show us the coefficient for the constant (). According to the table, the model with this constant has a statistically significant predictor of the result, because Sig = 0.000. The model has a high accuracy of almost 80%.

Table no. 3. Omnibus Test of Model Coefficients

		Chi-Square	df	Sig.
Step 1	**Step**	155.038	5	.000
	Block	155.038	5	.000
	Model	155.038	5	.000

Source: Author own research results

Omnibus tests of the model coefficients (table no. 3) are used to check if the new model (with explanatory variables included) is an improvement of the basic model. The Chi-Square test was used to see if there was a significant difference between the -2Log likelihood of the base model and the new model. In this case, Chi-Square=155.038 and Sig=000, which means that the null hypothesis is rejected. Because Chi-Square is significant, means that the new model is significantly better. The "Model" row always compares the new model with the original one. The Step and Block rows are important only if the explanatory variables are added to the model in a gradual or hierarchical manner. If the model was built in stages, then these rows would compare -2 Log likelihood of the newest model with the previous version to determine if each new set of explanatory variables determined improvements or not. In this case, I added all five explanatory variables in a single block and therefore there is only one step. This means that the Chi-square values are the same for step, block and model. Sig values are equal to 0.000 which indicates improved model accuracy when the explanatory variables are added.

However, the most important of all the results is the table Variables in the table of equations. This table needs to be studied very closely, as it is at the heart of the answer to our questions about the common association of Credit Score, Tenure, Balance, Number of Products and Estimated Salary.

Table no. 4. Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp (B)
Step 1	**CreditScore**	-.001	.000	7.697	1	.006	.999
	Tenure	-.011	.009	1.583	1	.208	.989
	Balance	.000	.000	118.870	1	.000	1.000
	NumOfProducts	-.058	.045	1.676	1	.195	.943
	EstimatedSalary	.000	.000	1.218	1	.270	1.000
	Constant	-1.190	.197	36.337	1	.000	.304

Source: Author own research results

This table provides the regression coefficient (B), Wald statistics (to test statistical significance), and (Exp (B)) for each variable category.

The variables “Credit Score” and “Balance” are statistically significant because Sig is less than 0.05. Wald for Credit Score is equal with 7.69 and for balance is equal with 118.87.

The classification chart answers a similar question as the classification table which is “How accurate is our model in classifying individual cases”? However, the rating chart provides some finer details. This graph shows the frequency of categorizations for different predicted probabilities and whether there were "yes" or "no" categorizations. This provides a useful visual guide to how accurate the model is by showing how many times the model would predict a "yes" result based on the predicted calculated probability when in fact the participant's result was "no".

```
             Step number: 1

             Observed Groups and Predicted Probabilities

    1600 +                                                                                                     +
         I                                                                                                     I
         I                                                                                                     I
F        I              Y                                                                                      I
R   1200 +              Y                                                                                      +
E        I              N                                                                                      I
Q        I              N                                                                                      I
U        I             YNY                                                                                     I
E    800 +             NNY      YYY                                                                            +
N        I             NNN     YYNYY                                                                           I
C        I             NNN    YNNNNY                                                                           I
Y        I             NNN    YNNNNNY                                                                          I
     400 +             NNNY   NNNNNNN                                                                          +
         I             NNNN  NNNNNNNNN                                                                         I
         I            NNNNNYYNNNNNNNNNN                                                                        I
         I            NNNNNNNNNNNNNNNNNNN                                                                      I
Predicted ---------+---------+---------+---------+---------+---------+---------+---------+---------+----------
  Prob:   0        .1        .2        .3        .4        .5        .6        .7        .8        .9         1
  Group:  NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYY

          Predicted Probability is of Membership for Yes
          The Cut Value is .50
          Symbols: N - No
                   Y - Yes
          Each Symbol Represents 100 Cases.
```

Figure no. 1. Observed Groups and Predicted Probabilities

Source: Author own research results

It is observed according to figure 1 that the model is good for predicting the result for individual cases, because there are a lot of observations towards the left and right ends of the graph. Such a diagram shows that where the event took place (on the 0.5th place - indicated by the letter Y on the graph), the predicted probability was also high and that where the event did not occur (letter N on the graph) the predicted probability was low. The graph shows that there are quite a few cases with the letter Y on the graph, which means that the model predicts a 80% probability that the event will happen.

The regression equation is written:

Log(p/1-p) = -1.190 -.001 * Credit Score -.011 * Tenure -.058 * Number of Products.

Conclusions

In conclusion, Churn is not the case when a predictive model can help to acquire the main goal, which is churn reduction, because churn-reducing interventions build upon knowing the causes and effect for churn. In order to reduce churn, the data analyst should create great client metrics. The metrics may have some aspects like: efficiently known by the business, certainly correlated with churn and retention, increasing engagement by segmenting the customers in a way that is favorable for addressed interventions and also useful in different functions of the business (e.g., marketing, support, product). In order to establish the determining factors of bank customers who decided to leave the bank or not, I used a binary logistic regression because it is essential to identify what leads a client towards the decision to leave the company. As it is already known, it is more expensive to sign in a new client than keeping an existing one. Binary logistic regression estimates the probability of an event occurring, in this case the probability that bank customers will leave the bank or not. There are more customers who do not leave the bank compared to those who leave. (7963 compared with 2037). The overall percentage tells us that this approach to prediction is correct with 79.6%. The model with this constant has a statistically significant predictor of the result, because Sig = 0.000. The model has a high accuracy of almost 80%.

References

Amal, M., Aksoy, M.S. and Alzahran, R., 2014. A Survey on Data Mining Techniques In Customer Churn Analysis For Telecom Industry. *Int. Journal of Engineering Research and Applications*, 4(5), pp.165-171.

Balle, B., Casas. B., Catarineu, A., Gavalda, R. and Manzano-Macho D., 2013. The Architecture of a Churn Prediction System Based on Stream Mining. In *Artificial Intelligence Research and Development: Proceedings of the 16th International Conference of the Catalan Association for Artificial Intelligence*, vol. 256, Article number: 157.

Davis, J., 2017. *Measuring Marketing.* [online] Available at: <https://learning.oreilly.com/library/view/measuring-marketing/9781501507229/> [Accessed 14 Apr 2021]

Gold, C., 2020. *Fighting Churn with Data*. [online] Available at: <https://learning.oreilly.com/library/view/-/9781617296529/?ar> [Accessed 1 Apr. 2021].

Hwang, Y.H., 2019. *Hands-On Data Science for Marketing.* [online] Available at: <https://learning.oreilly.com/library/view/hands-on-data-science/9781789346343/> [Accessed 15 Apr. 2021]

Kotler, P., 1994. *Marketing Management: Analysis, Planning, Implementation, and Control*. New Jersey: Prentice-Hal.

Lemmens, A., 2013. *Managing Churn to Maximize Profit.* [online] Available at: <https://pubsonline.informs.org/doi/abs/10.1287/mksc.2020.1229> [Accessed 15 Apr. 2021]

Mehta, N., Pickens, A., and Martinez, M., 2020. *The Customer Success Economy*. [online] Available at: <https://learning.oreilly.com/library/view/the-customer-success/9781119572763/> [Accessed 14 Apr. 2021].

Miguei, V.L., Van den Poel, D., Camanho, S. and Falcao e Cunha, J., 2012. Modeling partial customer churn: On the value of first product-category purchase sequences. *Expert Syst. Appl*., 39(12), pp.11250–11256.

Reichheld, F. and Sasser, W.E., 1990. Zero Defections: Quality Comes to Services. *Harv. Bus. Rev*., 68(5), pp.105–11.

Vaidyanathan, A. and Ranago, R., 2020. *The Customer Success Professional's Handbook*. [online] Available at: <https://learning.oreilly.com/library/view/the-customer-success/9781119624615/> [Accessed 15 Apr. 2021].