Bayesian Network's Application to Bank Customer Churn Analysis

41474 - TMBD

V. Ana (33.(3)%)[a], G. Beatriz (33.(3)%)[b], R. Obafemi (33.(3)%)[a]

[a] *Master in Mathematics and Applications*
[b] *Master in Data Science*
[c] *Department of Mathematics University of Aveiro*

November, 2022

Abstract

Bayesian Networks are graphical models that combine knowledge with data to represent the causal probabilistic relationships between a set of variables. The application of this method provides insight into the satisfaction of bank customers. Accordingly, we were able to conclude that there is a higher probability of the customer being an active member and continuing to use the bank's services (86%), followed by the probability of not being active and continuing to use the bank's services (73%). Also, there is a lower probability that the customer is not active and leaves the bank (27%). Finally, the probability of being active and leaving the bank is the lowest (about 14%).

Keywords— churn, Bayesian Network, conditional dependencies.

1. Introduction

The customer churn phenomenon consists in the egress of costumers or subscribers of a business, which results in no more purchases of a given product and interaction with that supplier. It is more expensive to acquire a new customer rather than retaining one already filiated, taking this into account, the phenomenon of customer churn can prove to be a roadblock for an exponentially growing organization [5]. Therefore, the prediction of this event shows itself to be very important, because by acknowledging that a customer intends to churn, there is a possibility of discouraging him from doing so through means such as marketing strategies.

This study approaches a Bayesian Network's application to a churn problem. This problem's dataset [1] is composed by data relative to costumers in ABC Multistate bank, containing 10000 observations. Each observation regards a bank's client, for every single one was collected information on credit score, raging from 350 to 850, country of residence, which includes France, Germany and Spain, gender, with the levels Male and Female, age, which span from 18 to 92, tenure concerns to the number of years the client has an account in ABC bank which can assume values up to 10 years, account balance, corresponding to a numerical variable extends to 251000, number of products from bank varies from 0 to 4. Moreover, there is knowledge about the ownership of a credit card and the activity status of the member of the bank. Since there is no specifications on the levels of credit card and active member variables its assigned, in the case of credit card variable, 1 as to owning a credit card and 0 not owning one. Likewise, for the active member variable, 1 stands for being an active member and 0 being an inactive member. As for churn, 1 indicates that a client has left the bank and 0 indicates that he remained a bank's client.

The aim of this exploration is predicting, based on variables' dependencies, whether a costumer will leave the bank or not, using a Bayesian Network. [11]

2. Methodology

Considering a model with n variables, the joint distribution of these variables is described by 2^n combinations,

$$P(h_i|d_1, d_2, \ldots, d_n) = \frac{P(d_1, d_2, \ldots, d_n|h_i) \cdot P(h_i)}{P(d_1, d_2, \ldots, d_n)}$$

where all the 2^n probabilities must be known.

Given the complexity of this model, it can be approached through the description of the dependence of events by some mathematical structure such as a directed acyclic graph, which is the case of Bayesian Networks. These models describe the probability distribution of a set of random variables by combining conditional independence assumptions with conditional probabilities.

Therefore, a Bayesian Network is a probabilistic graphical model which represents the knowledge about an uncertain domain where each node corresponds to a variable and each edge represents the conditional probability for the corresponding variables. Thence, these models are a representation of causality relationships, where the conditional dependence structure is inferred based on the Bayes theorem

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}.$$

[11] [2]

This method also satisfies the Markov local property, which states that a random variable associated with a node x_i is conditionally independent of its non-descendants have given its parents. This allows simplifying the joint distribution obtained using the chain rule, given by

$$\begin{aligned} & P(x_1, x_2, \ldots, x_n) \\ = & P(x_1) \cdot P(x_2|x_1) \ldots P(x_n|x_1, \ldots, x_{n-1}) \end{aligned}$$

Resulting that the joint distribution of a Bayesian Network is equal to

$$\begin{aligned} P(x_1, x_2, \ldots, x_n) &= \prod_{i=1}^{n} P(x_i|x_1, x_2, \ldots, x_{i-1}) \\ &= \prod_{i=1}^{n} P(x_i|Parents(x_i)) \end{aligned}$$

In larger networks, this property allows a great reduction of the amount of required computation, since generally, most nodes will have few parents relative to the overall size of the network.

Despite the simplifications the Bayesian methods allow, they present some downfalls. The Bayesian Network's method does not provide information on how to select a prior, which may generate misleading results. Considering the variables' dependencies, it can produce posterior distributions that are heavily influenced by the priors. Also, networks with a large number of parameters, which, consequently, leads to networks with high complexity, can be computationally expensive [3].

Today Bayesian Networks applications emerge in a variety of areas. Areas such as identifying oil locations, approving medical devices, medical diagnostics, operational risk management, legal profession, email filtering for spam status, skill classification for modern video games, cell phone recognition ([6][9]), document classification and image processing can be highlighted.

According to more recent applications we have applications, for example in environmental sciences, including forestry ([8]), fisheries ([12]), and water resource management ([10]). [4] [7]

3. Problem's Definition

Over time, churn prediction has become a huge part of many modern businesses, mainly due to the performance gains it offers, thus it is now considered a very helpful tool for companies.

This study lays on the central goal of analyzing the churn phenomenon of ABC Multistate bank's customers based on information of 10000 clients [1]. In order to do this analysis, clients' aspects such as gender, age, country of residence, tenure, account balance, number of products from the bank, estimated salary, ownership of a credit card and whether the client is an active member of the bank are considered, therefore these are taken as the problem's variables.

A Bayesian Network is a graphical model that represents a set of variables, denoted by nodes, and their conditional dependencies through a directed acyclic graph. Thus, the network's nodes are established as the variables indicated above.

A network can be divided into layers. Considering that the nodes corresponding to gender, age and country are independent events, these compose the first level of the network, which represents the root nodes.

Having in mind that incoming salary may vary from country to country and between genders, the salary variable is dependent from country and gender variables. Regarding aspects such as tenure and having a credit card, it is reasonable to suppose that they are, in some way, correlated to the age of clients. For its part, tenure and estimated salary may influence account balance. The credit score is a measure to evaluate the probability of an individual repaying loans in a timely manner, hence this variable may depends on account balance, which consequently depends on tenure and age of clients. To analyze if a client is an active member of the bank their credit score is taken into consideration, as the number of products purchased by the customer through the bank and the ownership of a credit card. Lastly, it is understandable to deem the churn variable dependent from the activeness of a client.

4. Research/Work question

As our analysis consists in getting a probabilistic analysis of churn, it is given a brief description of what was done in the code to get to the desired answer.
Beginning with data cleaning, we verified the non existence of missing values and duplicated rows. After that, since the costumer id doesn't show itself as important to the analysis, it will be dropped from our data.

The next goal is to assign probabilities to variables. Once the columns regarding age, credit score, tenure, balance and estimated salary have many unique values we proceed to create bands for each.

The pomegranate distribution "DiscreteDistribution" returns a discrete distribution composed of the variables and their probabilities, assuming that these probabilities sum to one. Applying it to each of the variables within each column, we are able to assign a probability to them. We do this for the columns concerning the country, age and gender since these are the events that we define as "independent", that is, they have no prior conditions that impose them.

To learn the probabilities of the remaining events, there is the necessity of computing the conditional probabilities of variables' dependencies described in section 3. For this effect, we applied the pandas *groupby* function, for each of the relations. In the case where the conditional probabilities were equal to zero, it is imperative to insert them manually into the conditional probability tables, since they did not appear.

The last relation considered yield the probabilistic prediction of churn taking into account all the previous relations. For a better understanding of all the dependencies/relationships in our Bayesian Network, we graph it using a directed acyclic graph, where problem's variables are represented through nodes and variables' dependencies by edges.

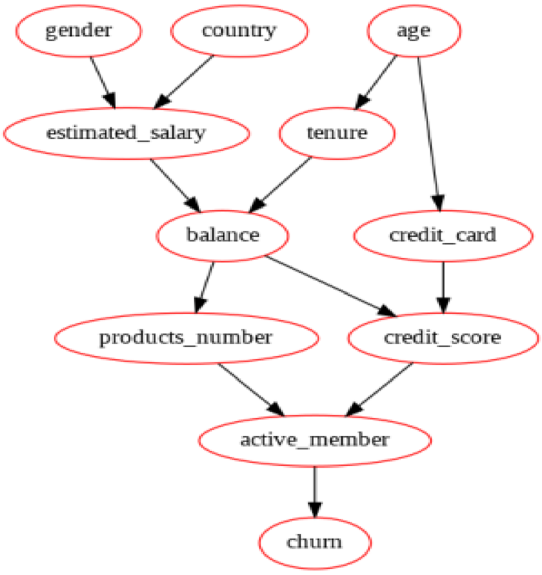

Figure 1: Bayesian Network.

5. Results

To better visualization, the results are organized into tables, in each all the conditional probabilities are listed, rounded to the hundredth.
For the independent variables which, as mentioned above, are the variables concerning the country, age and gender, the following results were obtained

Country	Probability
France	0.501
Germany	0.251
Spain	0.248

Table 1: Probabilities for the independent variable country.

Gender	Probability
Female	0.454
Male	0.546

Table 2: Probabilities for the independent variable gender.

Age	Probability
15-34	0.368
35-54	0.544
55-74	0.083
75-95	0.005

Table 3: Probabilities for the independent variable age.

Due to its size, the tables where are presented all the conditional probabilities for the remaining relations are shown in the appendix. From these, by analyzing the conditional probabilities yielded, we can see what is more or less likely to happen in each of the relations we imposed.

For the first relation, between the country, gender, and estimated salary, given in Table A.4, we see that the conditional probabilities are quite similar. However, the sequence of events that is more likely to happen is the client being from Spain, female, and having an estimated salary in the range of 100k-150k ($P \simeq 0.281$), while the one that would be less presumable to happen is the client being from Spain and female earning an estimated salary of 50k-100k ($P \simeq 0.234$). Also, it is possible to observe that this relation depends only on variables that we consider independent.

In the second relation, between age and tenure, given by Table A.5, we conclude the sequence of events most probable to happen is the client having an age between 75-95 years and tenure between 5-10 ($P \simeq 0.630$). On the other hand, is less frequent to have a customer among the 75 to 95 years old who is a bank's client for over 0-4 years ($P \simeq 0.370$). Again, this relation depends only on an independent variable.

In the relation between tenure, estimated salary, and account balance, given by Table A.6, we have that the sequence of events most likely to happen is the client having a tenure between 0-4, an estimated salary between 50k-100k and an account balance between 100k-150k ($P \simeq 0.395$). What is less likely to happen is that the client has a tenure between 0-4, an estimated salary between 0-50k and an account balance between 200k-255k. In this table more conditional probabilities appear with this value since we round all our values to hundredths. However, looking at the values with more decimal places, in our code, we see that this is in fact the smallest conditional probability value in this relationship ($P \simeq 0.0008857$). We should also take note that we are calculating the conditional probability of account balance based on the estimated salary. However, the estimated salary is not an independent variable and, therefore, it will have the contribution of the independent variables country and age.

In the fourth relation, between account balance and products number, given by Table A.7, we have that the sequence of events most likely to happen is that the client has an account balance between 0-50k and a total of 2 bank products, with a probability, $P \simeq 0.710$. What is less likely to happen is that the client has an account balance between 200k-250k, with a probability of zero. Again, we must keep in mind the previous dependencies, i.e. product number depends on the account balance. However, account balance is not an independent variable, in fact, it derives from two dependent variables as well: tenure, which depends on age, and estimated salary, which depends on gender and country.

Considering now the relation between account balance and credit score, given by Table A.8, noticed that a client having an account balance among 200k-255k and a credit score of 650-749 is the sequence of events with higher probability. For its part, a client having an account balance in the range of 200k-255k and a credit score of 350-449 has a null probability of occurring.
Again, credit score depends on balance, which, as we have seen, is not an independent variable, having the dependencies described above. These dependencies allow us to make a mental visualization of how our network looks.

In the sixth relation, between age and credit card, given by Table A.9, we have that the sequence of events most likely to happen is that the customer is between 55-74 years old and has a credit card. Although there are more conditional probabilities with the same value, looking at our code with more decimal places we see that this is in fact the highest probability value ($P \simeq 0.7101449$). What is less likely to happen, on the other hand, is that the customer is between 55-74 years old and does not own a credit card. Again, we note that there are more probabilities with the same value, however, looking at the code for values with more decimal places, we see that this is, in fact, the smallest conditional probability value ($P \simeq 0.2898551$).
Finally, for the credit card, credit score, products number and active member's relationship, given the Table A.10, we find that the sequence most likely to happen is that the customer has no credit card, has a credit score between 350k-449k, has 3 products in the bank and is not an active member, with probability $P = 1$. We have more than one relation with zero probability. Thus, the least likely sequences to happen are the customer not having a credit card, having a credit score between 350k-449k, having 3 products in the bank and being an active member, or not having a credit card, having a credit score between 350k-449k, having 4 products in the bank and not being, or being an active member, or not having a credit card, having a credit score between 450k-549k, having 4 products in the bank and being, or not being an active

member. Active membership depends on all the previous variables in the chain, as it is related to product number and credit score.

Our result, a probabilistic view of churn, is represented in Table A.11. From the table displayed we can perceive that being an active member and stay as client's bank is the most probable event ($\simeq 86\%$), which was to be expected. The probability of being inactive and still be a client of this bank is $\simeq 73\%$. Further, the probability of being an active member and leaving the bank is ($\simeq 14\%$) and the probability of being inactive and leaving is $\simeq 27\%$.

6. Conclusion

The Bayesian Networks have vast applications, within which applications in the financial area. In this project it was possible to model a Bayesian Network that yielded a probabilistic prediction of churn for ABC Multistate bank, taking into account certain characteristics of its current customers, provided by our dataset.

Thus, seven relations were proposed in order to study the conditional dependencies of our variables, resulting in a probability of churn of 14% if the customer is active, and 27% if the customer is not active.

Modeling with a Bayesian network allowed a simplification of the complexity inherent in this type of analysis and also proved to be a structure easy interpreted, due to the intuitive visualization of cause-effect relationships.

In the elaboration of this study, we faced some limitations. Because of the variety of some variables' values, we grouped them considering some ranges, this may have influenced the outcome generated for the conditional probabilities, as the setting of the lower and upper bounds influences the frequencies of each class. Besides, due to the specificity of this work, we are not able to make a comparison with related work in the literature. Here, we have obtained a probabilistic view of churn, taking into account very particular dependencies given by our dataset. Thus, we are unable to find any work with this type of dependency in order to make a comparison that does justice to our results.

References

[1] *Bank customer churn dataset*, https://www.kaggle.com/datasets/gauravtopre/bank-customer-churn-dataset (visited on 10/28/2022).

[2] Ben-Gal, ≪Bayesian Networks≫, Encyclopedia of Statistics in Quality and Reability (2007).

[3] J. Berger, *Statistical Decision Theory and Bayesian Analysis* (Springer-Verlag, 1985).

[4] J. F. Carriger, M. Glendell, and S. J. Moe, ≪Increased Use of Bayesian Network Models Has Improved Environmental Risk Assessments≫, (2020).

[5] *Customer churn: definition, rate, analysis and prediction*, https://www.questionpro.com/blog/customer-churn/ (visited on 11/02/2022).

[6] N. M. Fenton N, *Risk assessment and decision analysis with Bayesian networks* (Taylor and Francis, 2013).

[7] V. Lendave, ≪A Guide to Infering With Bayesian Network≫, Developers Corner (2021).

[8] B. G. Marcot, R. S. Holthausen, M. G. Raphael, M. M. Rowland, and M. J. Wisdom, ≪Using Bayesian belief networks to evaluate fish and wildlife population viability under land management alternatives from an environmental impact statement≫, Ecol Manage (2001).

[9] M. D. Pearl J, *The book of why: The new science of cause effect.* (Basic Books, 2018).

[10] T. Saloranta, B. D. N. Moe SJ, H. O. Eggestad, D. N. Kuikka SBarton, T. Saloranta, S. J. Moe, H. O. Eggestad, and S. Kuikka, ≪Bayesian belief networks as a meta-modelling tool in integrated river basin management: Pros and cons in evaluating nutrient abatement decisions under uncertainty in a Norwegian river basin≫, Ecol Econ (2008).

[11] D. Soni, ≪Introduction to Bayesian Networks≫, (2018).

[12] Uusiatlo, S. Kuikka, P. Kauppila, P. Soderkultalahti, and S. Back, ≪Assessing the roles of environmental factors in coastal fish production in the northern Baltic Sea: A Bayesian network application≫, Integr Environ Assess Manage (2012).

Appendix A. Conditional probabilities

Appendix A.1. Country, Gender and Estimated Salary relation

Country	Gender	Estimated Salary	Probability
France	Female	100k-150k	0.261
France	Female	50k-100k	0.254
France	Female	0-50k	0.248
France	Female	150k-200k	0.237
France	Male	50k-100k	0.262
France	Male	150k-200k	0.251
France	Male	100k-150k	0.246
France	Male	0-50k	0,240
Germany	Female	100k-150k	0.265
Germany	Female	150k-200k	0.261
Germany	Female	0-50k	0.238
Germany	Female	50k-100k	0.236
Germany	Male	50k-100k	0.254
Germany	Male	0-50k	0.252
Germany	Male	150k-200k	0.248
Germany	Male	100k-150k	0.246
Spain	Female	100k-150k	0.281
Spain	Female	150k-200k	0.243
Spain	Female	0-50k	0.242
Spain	Female	50k-100k	0.234
Spain	Male	50k-100k	0.265
Spain	Male	0-50k	0.254
Spain	Male	100k-150k	0.246
Spain	Male	150k-200k	0.235

Table A.4: Conditional probabilities for the relation: estimated salary, country of residence and gender.

Appendix A.2. Age and Tenure

Age	Tenure	Probability
15-34	5-10	0.551
15-34	0-4	0.449
35-54	5-10	0.551
35-54	0-4	0.449
55-74	5-10	0.540
55-74	0-4	0.460
75-95	5-10	0.630
75-95	0-4	0.370

Table A.5: Conditional probabilities for the relationship: tenure and age.

Appendix A.3. Tenure, Estimated Salary and Account Balance

Tenure	Estimated Salary	Account Balance	Probability
0-4	0-50k	100k-150k	0.372
0-4	0-50k	0-50k	0.367
0-4	0-50k	50k-100k	0.163
0-4	0-50k	150k-200k	0.097
0-4	0-50k	200k-255k	0.001
0-4	100k-150k	100k-150k	0.390
0-4	100k-150k	0-50k	0.358
0-4	100k-150k	50k-100k	0.158
0-4	100k-150k	150k-200k	0.090
0-4	100k-150k	200k-255k	0.004
0-4	150k-200k	100k-150k	0.387
0-4	150k-200k	0-50k	0.358
0-4	150k-200k	50k-100k	0.156
0-4	150k-200k	150k-200k	0.093
0-4	150k-200k	200k-255k	0.005
0-4	50k-100k	100k-150k	0.395
0-4	50k-100k	0-50k	0.363
0-4	50k-100k	50k-100k	0.152
0-4	50k-100k	150k-200k	0.083
0-4	50k-100k	200k-255k	0.006
5-10	0-50k	0-50k	0.390
5-10	0-50k	100k-150k	0.375
5-10	0-50k	50k-100k	0.133
5-10	0-50k	150k-200k	0.099
5-10	0-50k	200k-255k	0.003
5-10	100k-150k	100k-150k	0.378
5-10	100k-150k	0-50k	0.365
5-10	100k-150k	50k-100k	0.154
5-10	100k-150k	150k-200k	0.099
5-10	100k-150k	200k-255k	0.004
5-10	150k-200k	100k-150k	0.391
5-10	150k-200k	0-50k	0.368
5-10	150k-200k	50k-100k	0.145
5-10	150k-200k	150k-200k	0.093
5-10	150k-200k	200k-255k	0.002
5-10	50k-100k	0-50k	0.380
5-10	50k-100k	100k-150k	0.377
5-10	50k-100k	50k-100k	0.150
5-10	50k-100k	150k-200k	0.091
5-10	50k-100k	200k-255k	0.001

Table A.6: Conditional probabilities for the relation: tenure, estimated salary and account balance.

Appendix A.4. Account Balance and Products Number

Account Balance	Products Number	Probability
0-50k	1	0.258
0-50k	2	0.710
0-50k	3	0.028
0-50k	4	0.004
50k-100k	1	0.645
50k-100k	2	0.329
50k-100k	3	0.023
50k-100k	4	0.004
100k-150k	1	0.662
100k-150k	2	0.302
100k-150k	3	0.027
100k-150k	4	0.009
150k-200k	1	0.637
150k-200k	2	0.329
150k-200k	3	0.027
150k-200k	4	0.006
200k-250k	1	0.706
200k-250k	2	0.265
200k-250k	3	0.029
200k-250k	2	0.0

Table A.7: Conditional probabilities for the relation: account balance and products number.

Appendix A.5. Account Balance and Credit Score

Account Balance	Credit Score	Probability
0-50k	650-749	0.352
0-50k	550-649	0.335
0-50k	750-850	0.154
0-50k	450-549	0.139
0-50k	350-449	0.020
100k-150k	650-749	0.338
100k-150k	550-649	0.334
100k-150k	750-850	0.169
100k-150k	450-549	0.141
100k-150k	350-449	0.017
150k-200k	650-749	0.358
150k-200k	550-649	0.314
150k-200k	750-850	0.172
150k-200k	450-549	0.137
150k-200k	350-449	0.018
200k-255k	650-749	0.441
200k-255k	550-649	0.265
200k-255k	450-549	0.206
200k-255k	750-850	0.088
50k-100k	650-749	0.355
50k-100k	550-649	0.324
50k-100k	750-850	0.160
50k-100k	450-549	0.143
50k-100k	350-449	0.017
200k-255k	350-449	0.0

Table A.8: Conditional probabilities for the relation: account balance and credit score.

Appendix A.6. Age and Credit Card

Age	Credit Card	Probability
15-34	1	0.710
15-34	0	0.290
35-54	1	0.702
35-54	0	0.298
55-74	1	0.710
55-74	0	0.290
75-95	1	0.667
75-95	0	0.333

Table A.9: Conditional probabilities for the relation: age and credit card.

Appendix A.7. Credit Card, Credit Score, Products Number and Active Member

Credit Card	Credit Score	Products Number	Active Member	Probability
0	350-449	1	0	0.481
0	350-449	1	1	0.519
0	350-449	2	0	0.6
0	350-449	2	1	0.4
0	350-449	3	0	1.0
0	350-449	3	1	0.0
0	350-449	4	0	0.0
0	350-449	4	1	0.0
0	450-549	1	0	0.477
0	450-549	1	1	0.523
0	450-549	2	0	0.5
0	450-549	2	1	0.5
0	450-549	3	0	0.5
0	450-549	3	1	0.5
0	450-549	4	0	0.0
0	450-549	4	1	0.0
0	550-649	1	0	0.477
0	550-649	1	1	0.523
0	550-649	2	0	0.469
0	550-649	2	1	0.531
0	550-649	3	0	0.448
0	550-649	3	1	0.552
0	550-649	4	0	0.545
0	550-649	4	1	0.454
0	650-749	1	0	0.479
0	650-749	1	1	0.521
0	650-749	2	0	0.470
0	650-749	2	1	0.530
0	650-749	3	0	0.607
0	650-749	3	1	0.393
0	650-749	4	0	0.333
0	650-749	4	1	0.667
0	750-850	1	0	0.470
0	750-850	1	1	0.530
0	750-850	2	0	0.442
0	750-850	2	1	0.558
0	750-850	3	0	0.625
0	750-850	3	1	0.375
0	750-850	4	0	0.5
0	750-850	4	1	0.5
1	350-449	1	0	0.6
1	350-449	1	1	0.4
1	350-449	2	0	0.482
1	350-449	2	1	0.518
1	350-449	3	0	0.667
1	350-449	3	1	0.333
1	350-449	4	0	0.667
1	350-449	4	1	0.333

Credit Card	Credit Score	Products Number	Active Member	Probability
1	450-549	1	0	0.524
1	450-549	1	1	0.476
1	450-549	2	0	0.480
1	450-549	2	1	0.520
1	450-549	3	0	0.643
1	450-549	3	1	0.357
1	450-549	4	0	0.25
1	450-549	4	1	0.75
1	550-649	1	0	0.506
1	550-649	1	1	0.494
1	550-649	2	0	0.475
1	550-649	2	1	0.525
1	550-649	3	0	0.621
1	550-649	3	1	0.379
1	550-649	4	0	0.467
1	550-649	4	1	0.533
1	650-749	1	0	0.483
1	650-749	1	1	0.517
1	650-749	2	0	0.455
1	650-749	2	1	0.545
1	650-749	3	0	0.529
1	650-749	3	1	0.471
1	650-749	4	0	0.727
1	650-749	4	1	0.272
1	750-850	1	0	0.516
1	750-850	1	1	0.484
1	750-850	2	0	0.454
1	750-850	2	1	0.545
1	750-850	3	0	0.606
1	750-850	3	1	0.394
1	750-850	4	0	0.5
1	750-850	4	1	0.5

Table A.10: Conditional probabilities for the relation: credit card, credit score, products number and active member.

Appendix A.8. Active Member and Churn

Active Member	Churn	Probability
0.0	0.0	0.731
0.0	1.0	0.269
1.0	0.0	0.857
1.0	1.0	0.143

Table A.11: Conditional probabilities for the relation: active member and churn.