Predicting Customer Churn in Electronic Banking

Marcin Szmydt(✉)

Department of Information Systems, Poznan University of Economics and Business, Aleja Niepodległości 10, Poznan, Poland
marcin.szmydt@ue.poznan.pl

Abstract. The following paper is an outline of the current author's research on the churn prediction in electronic banking. The research is based on real anonymised data of 4 million clients from one of the biggest Polish banks. Access to real data in such scale is a substantial strength of the study, as many researchers often do use only small data sample from a short period. Even though current research is still preliminary and ongoing, unlimited access to these data provides a great environment for further work. The study strongly connects with real business goals and trends in the banking industry as the author is also a practitioner. Described research focuses on methods for predicting customers who are likely to leave electronic banking. It contributes especially in further classification of an electronic churn and a broader definition of customer churn in general. Recommended solutions should contribute to the increase in the number of digital customers in the bank.

Keywords: Banking · Churn prediction · Electronic banking

1 Introduction

In almost every business retaining existing customers is undoubtedly cheaper than gaining new ones. In the past, banking used to be a sector with relatively low churn rate. Each customer was treated individually and offered a customised service. After years of intense competition, customers exchanged highly personalised service for a lower price, higher anonymity and reduced variety [11]. Right now, in the age of digitalisation, it is easier than ever to open a new bank account and transfer all assets without even leaving home. This situation forced banks to become more interested in the subject of customer retention. However, before taking efficient ways of retaining existing clients, it is necessary to predict those who are about to leave [1].

Furthermore, with digital transformation of the banking industry, it is also essential for banks to keep customers in electronic channels like Internet or mobile [2]. Reason for this is that digital customer is more profitable than the traditional one. There is no need for a physical branch nor a staff that is providing services to this client. The digital customer is also more willing to use

W. Abramowicz and A. Paschke (Eds.): BIS 2018 Workshops, LNBIP 339, pp. 687–696, 2019.
https://doi.org/10.1007/978-3-030-04849-5_58

additional services available in electronic or mobile banking and is susceptible to e-marketing [13]. Thanks to this trend, a new issue arose - *digital churn*. Predicting this phenomen means identifying customers who are about to leave using remote services. Leaving electronic channels may indicate that customer has opened another account in competitors bank or is just not satisfied with a banking portal or mobile application any more. An increasing number of banks do not charge customers for running accounts therefore most customers do not close bank accounts any more, they just stop using them. This is the reason why predicting churn from electronic channels might be more valuable than just predicting closing all bank accounts and products - as the churn is described in most research papers in existing literature. It is also vital to discover churn reasons of a given customer to take appropriate actions in time [7]. To the best of author's knowledge, there are no articles nor publications with the main focus on the prediction of electronic banking churn with the further classification of an electronic churn type to match adequate retention campaigns.

The above motivation gives the way to outline the main research problem analysed in this paper, which is: *How to predict customers who are leaving electronic channels and how to retain them?* To solve this problem, the following research questions need to be addressed:

Q1: Who are the customers who are leaving electronic channels and what are the reasons for their churn? (electronic churn groups)
Q2: What are the existing methods of customer churn prediction in banking?
Q3: How efficiently apply existing methods to electronic channel churn prediction and further electronic churn classification?

2 Research Methodology

Having in mind empirical and theoretical aspects of the problem, proper research methodology must be chosen. The methodology behind this research follows the principles of Design Science Research (DSR) with guidelines designed by Hevner et al. in [4]. Additionally, some recommendations provided by Webster and Watson in [14] has also been used.

DSR methodology consists of the following phases: *Awareness of Problem, Suggestion, Development, Evaluation*, and *Conclusion*. After a broad analysis of the banking sector with the main focus on electronic channels and digitalisation of the banking industry, the problem was formulated in Sect. 1. In the next stage, it was extended to a form of research questions (also Sect. 1). The literature review concerning an existing churn prediction methods in banking has been described in Sect. 3. Suggested methods to address presented issues has been provided in Sect. 4. In the same section, some evaluation tools for this methods also were proposed. Conclusions and further actions of this ongoing research are described in Sect. 5.

Following the Design Science Research Guidelines, the artifact in the form of a process has been developed. Business needs described in Sect. 1 justify research problem. In Sects. 1 and 4 environment analysis has been performed and some

proposals of empirical evaluation methods were included in Sect. 4. Even though the study is still incomplete, research rigour is enforced to yield best results in current and further work (Fig. 1).

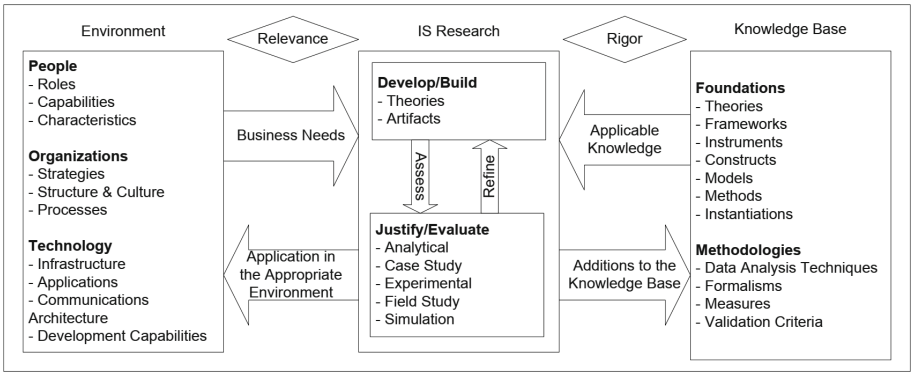

Fig. 1. Design science research with guidelines designed by Hevner et al. [4]

3 Literature Review

3.1 Searching Process

Comprehensive literature analysis has been conducted to analyse the current state of the art of research in the area of churn prediction in banking. Key recommendations provided by Webster and Watson in [14] and Hevner et al. [4] has been used as a guideline to enhance the rigour of the research process. The chosen research databases consisted of Springer, IEEExplore, ACM Digital Library, ResearchGate and ScienceDirect because they covered research papers from the fields of Information Systems, IT, Engineering and Economics. Additionally, following guidelines of Webster and Watson [14] Google Scholar and Mendeley tools were used to provide relevant articles that were not found in the databases listed above. By this means, a literature search was not limited to only top journals and databases.

The search was conducted by using the keywords and phrases: “Churn Prediction”; “Customer Attrition Prediction”; “Graph-based Churn Prevention”; “Predicting Customer Churn”; “Customer Churn Model”; “Customer Churn Analysis” and combining them with the domain keywords: “banking”, “banks”; “electronic banking”; “electronic channels”; “financial sector”; “finance” and “fintech”. Backwards and forward search has also been comprehensively performed to find relevant papers. The main goal of this literature review was to summarise existing knowledge in churn prediction with a primary focus on the banking domain.

Initial number of publications related to churn prediction and churn models consisted of 1427 items. After filtering papers related to the banking industry, a total number of publications decreased to 150. Not being able to describe all of them, the author tried to present the most representative cases which can be found in Table 1 in a form of summarized results.

The literature has been sorted historically with a description of their key attributes of its approaches. The first attribute in the table describes the amount of real data sample acquired from the bank or other financial institution. It is an important factor because it emphasises the utility of a provided solution and simulates real-life environment available for practitioners. The second attribute provides information on duration of the time period for which these data refer to. Short periods may not contain seasonality and may not have been enough to predict other periods efficiently. Number and types of features may strongly affect prediction performance. Therefore next columns describe the initial available number of features and dataset feature types used during training of prediction model. Further characteristics describe methodologies used for building these models.

3.2 Related Work

This section provides information about different approaches in literature related to the prediction of a customer churn with a primary focus on the banking sector.

To tackle the problem of churn prediction not only does it need to be chosen an efficient algorithm, but also it is vital to acquire proper dataset of relevant features and validate the methodology on a real customer data. [9] used logistic regression to predict customer churn on data retrieved from a Finnish bank. Their dataset included typical features of customers like socio-demographics, aggregated account transactions, banking products and services. Class imbalance problem has been addressed by using down-sizing (under-sampling) methods which reduced the number of non-churners to prevent classifying all customers into one group. Constructed models were evaluated mostly using lift curve and counting the number of correct predictions. Despite the simplicity of this approach, it predicted churners relatively well.

On the other hand, [12] worked on a data of 5000 customers aged in the range of 18 to 80 years old from Croatian bank chosen by random sampling. Class imbalance problem in this study was solved by choosing precisely half of the data that included churners and the other half with non-churners. Input data set contained similar variables to [9]. All of the variables were measured at five different points in time (t0 to t4). The fuzzy C-Means algorithm has been applied to predict customer churn. According to their research fuzzy-based methods performed better than the classical ones.

Rough set approach and flow network graph are another techniques that can be applied in predicting churners. [8] utilised these methods to a sample data of 21 000 customers from a commercial bank in Taiwan. Their data set also included demographics, psychographic and transactional features. Initially, their dataset consisted of 43 variables, but only 16 variables were selected to training

phase by their experts from university marketing department and managers from the banking industry. A drawback of this study is that there is only one-month period data taken into account when training models.

Other interesting studies on churn prediction by [1] suggest that customer social network analysis can significantly improve predictive accuracy. It introduces a concept that customers do not act independently and their decisions are highly influenced by others. Therefore customer network features (i.e. a degree of centrality, betweenness centrality and density) have been added to their training data. The dataset contained real records of 244 787 customers from a European financial services company. Beside network features, it also included socio-demographic and home banking behaviour. Random forests have been applied to train model for churn prediction. Their research revealed that contextual network features are even more important than socio-demographics and account features. This approach has also been explored and applied in prior studies from the telecom industry [5].

Further research in this area [3] included Support Vector Machine (SVM) analysis and naïïve Bayes tree (NBTree). A dataset of 14 814 customers used in this study was obtained from a Chilean bank that suffered from an increasing number of credit card churners. Two groups of features were available for each customer: socio-demographic and product behaviour. Accordingly to their study, by considering sensitivity alone, hybrid model of SVM + NBTree using Support vectors with corresponding predicted target values with reduced features and balanced by SMOTE (Synthetic minority over-sampling technique) yielded the best value of sensitivity.

[10] worked with data from one of the major Nigerian banks by using K-Means clustering, decision trees and JRip algorithm. The raw dataset included 1,048,576 customer records described with eleven features including socio-demographics, account information and transactional behaviour. However, only 4958 records were chosen for analysis after data cleaning and preparation. Furthermore, after sampling and variable selection only 500 customers with four attributes were included in the final data set for model training. A lot of missing values like age or customer type were a substantial problem of this research therefore only a small fraction of a whole database was used in the construction of a prediction model.

Current research seems to validate the view that simple techniques like decision trees are still widely used for churn prediction. [6] explored data using the CRISP methodology and built decision tree model for electronic banking customer churn prediction. This paper is one of few that touched the subject of electronic banking churn prediction. They randomly sampled 4383 customers of e-banking services from the bank's database with features related to customer dissatisfaction, service usage and socio-demographics. Their decision tree brings a lot of knowledge in the area of electronic churn and may be an object to draw some conclusions on reasons for customer churn.

Table 1. Summary of related literature on churn prediction in banking domain

Year	Researchers	Real data sample	Time period	Initial number of features	Dataset features	Method
2006	Teemu Mutanen, Jussi Ahola, Sami Nousiainen	151 000 customers	12 points in time (with intervals of 3 months during 34 months)	75	Socio-demographics, products, services, transactional data	logistic regression
2008	Dzulijana Popovic, Zagrebacka Banka, Bojana Dalbelo Basic	5 000 customers (choosen by random sampling from bank's database)	Does not say ("client population in 2005")	73	Socio-demographics, products, financial data, bad-behaviour	Fuzzy C-Means Clustering, Canonical Discriminant Analysis, K-means clustering
2011	Chiun Sin Lin, Gwo Hshiung Tzeng, Yang Chieh Chin	21 000 customers	1 month	43	Socio-demographics, transactional data	Rough Set Approach, Flow Network Graph
2012	Dries F. Benoit, Dirk Van Den Poel	244 787 customers	7 months	31	Socio-demographics, customer social network features, electronic banking behaviour	Random forests, Social Network Analysis (SNA)
2014	M.A.H. Farquad, Vadlamani Ravi, S. Bapi Raju	14 814 customers	Does not say ("data obtained in 2004")	22	Socio-demographic, behavioural data	Support Vector Machine (SVM), naive Bayes tree (NBTree)
2015	A.O. Oyeniyi, A.B. Adeyemo	1 048 576 customer records (model trained on only 500 records)	Does not say	11	Socio-demographics	K-Means clustering, Decision tree, JRip algorithm
2016	Abbas Keramati, Hajar Ghaneei, Seyed Mohammad Mirmoham-madi	4383 customers (choosen by random sampling from bank's database)	2 years	11	Socio-demographics, customer dissatisfaction, service usage	The CRISP Methodology, decision tree

3.3 Literature Discussion and Author's Contribution

The literature review shows that most researchers are limited to data and characteristics provided by the bank. Although the majority of prior research has applied a wide range of prediction tools, little attention has been paid to determine what customer features lead to the churn intention. Furthermore, in most literature, churn is defined as a closure of a given product (like account or credit card) and little attention is also paid to those customers who do not formally close their products, but just stop using them. Widely applied in the telecom industry - customer social network features - has been used only in one paper referring to the financial domain. There are only few publications

related to electronic banking churn. According to author's knowledge, no previous research has further classified electronic churners into different churn groups to match appropriate retention campaigns. This approach also partly addresses issues with churning without formally closing products. Few attention is also paid to actions after proper predicting of a customer who is about to leave and which retention methods yield best results. Trying to fill some of the existing gaps in the literature, below research has been conducted.

4 Research Progress and Solution Design Proposal

This section provides a brief description of analysis and gives an outline of the proposed method. Data mining on records from last three years from corporate data warehouse has been performed to get an overview of the bank's customers. After this analysis, to answer the first research question about who are the customers that are leaving electronic channels the following cases have been identified:

Group 1: Customers who are leaving electronic channels and then closing all banking products. This group of customers contains classical bank churners who are widely and frequently described in the literature. They are formally closing all banking products like accounts, credit cards and online services. Most likely, they want to end their relationship with the bank permanently.

Group 2: Customers who are leaving electronic channels and keep using banking products in offline channels. This group of customers stop using electronic channels or significantly reduce the intensity of usage of these channels. Reasons for this behaviour may vary from security concerns to switching primary bank but still using these products as a secondary option.

Group 3: Customers who are leaving electronic channels and stop using banking products (without formally closing products). These clients cease their activity in electronic as well as offline channels. In case of current accounts, as data mining of cash flows revealed, most frequently they transfer all their money to accounts in other banks in their last few months of electronic activity. However, formally they still have a bank account or other products that can be used in future. This is the biggest group but yet hard to identify during first stages. It is difficult to discover whether they decided to stop their activity permanently or just for a short period.

The proposed process of predicting and classifying electronic banking churners consists of seven stages presented in Fig. 2. First one focuses on acquiring customer data from banks databases. These sources may vary from traditional data warehouses to data retrieved from system logs and big data environment. During the second stage, data is being cleaned and prepared to use in the next stages. Later, proper methods like backward, forward feature selection or recursive feature elimination should be applied. This stage focuses on selecting best features and data sources for model training phase. After features selection, prediction models are being trained. During this stage, a variety of models created

by different methodologies may be constructed. Next stage evaluates these models to choose a model with best prediction efficiency. Methods like the Area Under the ROC Curve or similar tools should be applied to evaluate the performance of the constructed models. After having a database of potential electronic churners, classification methods might be applied to further categorise these customers into three groups of electronic churners as described above. After having a database with potential churners divided into these groups, different retention methods might be applied to keep them in bank and electronic channels. This process should be carried out regularly in cycles to adapt to constantly changing environment and to keep retaining customers in electronic channels.

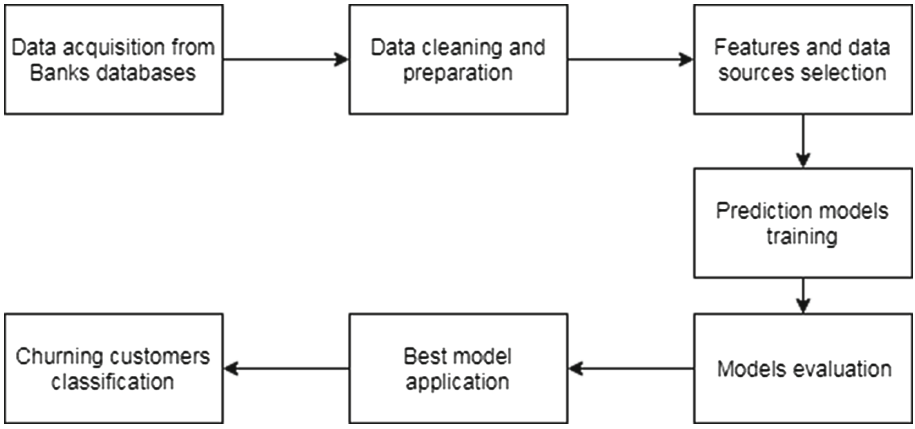

Fig. 2. Seven stages of the proposed process

5 Conclusions and Further Work

The above research reveals that electronic churn in banking industry might be an important aspect that is worth taking a closer look. The problem of electronic churn is broader then regular churn which is often described in the literature. As the customers may leave electronic channels and it may or may not mean that they churned from the bank. This approach will also reach customers that will be inactive or unsatisfied with their current electronic services. Having a proper prediction tool, banks may more successfully keep customers from churning in general and keep them in profitable electronic channels. This might be helpful to achieve bank's goals to incrase the number of a "digital customers" in the age of digital transformation of the banking industry. With the ability of further classification of electronic churners by their churn context, different retention campaigns may be used for different groups. This strategy should yield better results than just one campaign for all electronic churners.

The described analysis is only a brief indication of a problem and the research is still ongoing. The article presents a general idea for research in this field and

its vision. Further work will consist of carefully application each stage of this solution in the production environment, monitoring and comparison of best prediction models, especially those with nonstandard features like social influence or behavioural characteristics. This might also lead to further data feature types prediction efficiency analysis. These actions may improve further models performance and precision. After having an ability to precisely predict churners, further retention actions must be taken. Different retention strategies may be applied for specific groups of potential churners (as described in Sect. 4) to improve their success rates. These strategies should also be part of the research with a primary goal of finding the best-performing ones. It is also interesting question how many months before actual churn these actions should be initialised. Psychological and behavioural aspects should also be taken into account during retention strategy planning. After some empirical tests, machine learning and other prediction techniques may also be applied to identify best performing retention strategies.

References

1. Benoit, D.F., Van Den Poel, D.: Improving customer retention in financial services using kinship network information. Expert Syst. Appl. **39**(13), 11435–11442 (2012). https://doi.org/10.1016/j.eswa.2012.04.016
2. Cuesta, C., Ruesta, M., Tuesta, D., Urbiola, P.: The digital transformation of the banking industry. BBVA Research (2015). https://www.bbvaresearch.com/wp-content/uploads/2015/08/EN_Observatorio_Banca_Digital_vf3.pdf
3. Farquad, M.A., Ravi, V., Raju, S.B.: Churn prediction using comprehensible support vector machine: an analytical CRM application. Appl. Soft Comput. J. **19**, 31–40 (2014). https://doi.org/10.1016/j.asoc.2014.01.031
4. Hevner, A., March, S., Park, J., Ram, S.: Design science in information systems research. MIS Q.: Manage. Inf. Syst. **28**(1), 75–105 (2004)
5. Hill, S., Provost, F., Volinsky, C.: Network-based marketing: identifying likely adopters via consumer networks. Stat. Sci. **21**(2), 256–276 (2006). https://doi.org/10.1214/088342306000000222. http://projecteuclid.org/euclid.ss/1154979826
6. Keramati, A., Ghaneei, H., Mirmohammadi, S.M.: Developing a prediction model for customer churn from electronic banking services using data mining. Financ. Innov. **2**(1), 10 (2016). https://doi.org/10.1186/s40854-016-0029-6
7. Liébana-Cabanillas, F., Nogueras, R., Herrera, L.J., Guillén, A.: Analysing user trust in electronic banking using data mining methods. Expert Syst. Appl. **40**(14), 5439–5447 (2013)
8. Lin, C.S., Tzeng, G.H., Chin, Y.C.: Combined rough set theory and ow network graph to predict customer churn in credit card accounts. Expert Syst. Appl. **38**(1), 8–15 (2011). https://doi.org/10.1016/j.eswa.2010.05.039
9. Mutanen, T., Ahola, J., Nousiainen, S.: Customer churn prediction-a case study in retail banking. In: Proceedings of the ECML/PKDD Workshop on Practical Data Mining, pp. 13–19 (2006)
10. Oyeniyi, A.O., Adeyemo, A.B.: Customer churn analysis in banking sector using data mining techniques. Afr. J. Comput. ICT **8**(3), 165–174 (2015)
11. Peppard, J.: Customer relationship management (CRM) in financial services. Eur. Manage. J. **18**(3), 312–327 (2000)

12. Popović, D., Banka, Z., Bašić, B.D.: Churn prediction model in retail banking using fuzzy C-means algorithm. Informatica **33**, 243–247 (2009). http://wen.ijs.si/ojs-2.4.3/index.php/informatica/article/viewFile/242/239
13. Sumra, S.H., Manzoor, M.K., Sumra, H.H., Abbas, M.: The impact of e-banking on the profitability of banks: a study of Pakistani banks. J. Public Adm. Gov. **1**(1), 31–38 (2011)
14. Webster, J., Watson, R.T.: Analyzing the past to prepare for the future: writing a literature review. MIS Q. **26**(2), xiii–xxiii (2002)