PROJECT PROPOSAL FOR AI SATURDAYS LAGOS-COHORT 8

BY

CHRISTOPHER ATTAH ADAI

PROJECT TITLE:
Predicting Credit Card Approvals

DATE: 28/10/2023

1.0 Introduction

Commercial banks receive a lot of applications for credit cards. Many of them get rejected for many reasons, like high loan balances, low income levels, or too many inquiries on an individual's credit report, for example. Manually analyzing these applications is mundane, error-prone, and time consuming. Luckily, this task can be automated with the power of machine learning and pretty much every commercial bank does so nowadays. In this project, I will build an automatic credit card approval predictor using machine learning techniques, just like the real banks do. Consequently, churn customers cause the profits of a banking system to decrease. Therefore, there has been increased interest from banking professionals to design an early-warning system to classify a bank's customers into churn or non-churn customers. The system would be able to notify the bank's managers so that they can communicate with customers who are expected to churn to improve their services, which is an appropriate way to keep the customers satisfied with their bank.
Bank churn prediction aims to understand the possibility of customers moving from one bank to another. The reasons for movement include the availability of the latest technology, low interest rates, services offered, and credit card benefits [30]. This study aims to predict churned customers based on credit card and customer information (i.e., age and gender).

2.0 Problem Description

At present, the market is dynamic and highly competitive due to the availability of large numbers of service providers, especially banks, worldwide. One of the main challenges for this sector is the change in customer behavior. Customers are the core of all industries, especially customer-dependent organizations, such as the banking sector, which is responsible for accepting deposits, making investments, and granting loans. Long-term customers are directly connected to the production of profits; hence, banks should avoid losing customers. The Harvard Business Review believes that a 5% defection in customers can lead to an increase in profits for firms of between 25% and 85% [4,5]. Thus, given that customers are the most important assets with strong effects on a bank's profit, there are five essential pillars for the modern banking business: capital, liquidity, risk, assets, and customer management [6,7]. Focusing effectively on the five pillars can ensure that management effectively maximizes the profits of a bank [7,8]. Therefore, customer churn is a fundamental challenge for banks. Customer churn can be divided into two groups: voluntary churn and nonvoluntary churn. Nonvoluntary churn occurs when the bank withdraws services from customers, and it is easy to detect. On the other hand, voluntary churn is more difficult to identify, because it is a conscious decision by a customer to terminate their relationship with a certain bank.

3.0 Existing Solutions

Customer churn prediction utilizing big data is a research area within machine learning technology, which works to classify distinctive types of customers into either churning or non-churning customers [14,19,20]. Many studies in the literature have created multiple prediction models relying on statistical and data mining techniques (machine learning models), such as linear regression, decision trees, random forests, logistic regression, neural networks, support vector machines, and deep neural networks [1,2,5,21,22]. The prediction of credit card

customer churn is not a new field; many researchers have developed various prediction models. Kaya et al. (2018) [23] developed a prediction model which considered the individual transaction records of customers. The model they used mainly used information related to spatiotemporal factors, as well as choice and behavioral trait factors. The results showed that the developed model had more accurate prediction than the traditional models that considered demographics-based features. Moreover, early researchers tried to address the question of how machine learning models could be developed to predict customer churn, as discussed in Miao and Wang (2022) [24]. The authors developed a credit card customer churn prediction model by considering three machine learning approaches: random forest, linear regression, and k-nearest neighbor (KNN). The collected dataset contained 10,000 datum with 21 features, and the model was evaluated using the ROC, AUC, and confusion matrix.

Moreover, retail banking churn prediction was considered by Bharathi et al. (2020) [26]; the sample covered 602 young adult bank customers. To validate the developed model, different machine learning models were used, including ridge classifier cross-validation, the k-nearest neighbor classifier, decision tree, logistic regression, support vector classifier (SVC), and linear SVC. The results showed that the extra-tree classifier model had the highest performance compared with other models. The research showed that the top features to develop a prediction model were the absence of mobile banking, zero-interest personal loans, zero balance, and other services online. Saias et al. (2022) [27] developed a churn risk prediction model for customers of cloud service providers. The aim of the study was to create an alert system to avoid losing cloud customers based on neural network model AdaBoost and random forest. The results found that random forest outperformed other models. Moreover, many researchers developed a prediction system for different churned customers in various fields such as the telecommunication industries [28], and the E-commerce industry [29]. Bank churn prediction aims to understand the possibility of customers moving from one bank to another. The reasons for movement include the availability of the latest technology, low interest rates, services offered, and credit card benefits [30]. This study aims to predict churned customers based on credit card and customer information (i.e., age and gender). Meanwhile, researchers have attempted to find credit card fraud detection using machine learning models [31,32] or by developing optimization algorithms with machine learning models [33].

4.0 Proposed Method

Most of the prediction models so far have focused on developing prediction models for various problems in the banking system with a little interest in credit card customers. Therefore, I tried to fill the gaps of the previous studies in the field. The contributions of this project are as follows:

1. Prediction models developed based on forwarding different numbers of independent variables.
2. The capability of the prediction models was validated based on two-step clustering and k-nearest neighbors.
3. The capabilities of the machine learning models to predict credit card customer churn in banks were predicted.
4. The top features for developing a credit card churn prediction method were determined.

The primary aims of this proposal are as follows:
1. To use different independent variables in building a prediction model based on two-step clustering and k-nearest neighbors.
2. To select the appropriate machine learning models with top features for predicting churn customers.

Various studies have developed models for predicting customer churn without utilizing significant variables. To overcome this issue, it has been suggested that categorical variables are merged into one variable. Therefore, this research gap prompted the me to find an appropriate model for predicting customer churn. The primary step for developing a customer churn prediction model is to collect, analyze, and clean the dataset. Poorly cleaned data are unable to establish a relationship between input and output variables; in turn, this affects the performance of the prediction model. Therefore, the cleaned dataset can be applied in three models to build customer churn prediction models. The methods aim to select input variables depending on different independent variables that are selected by feeding all the independent variables in the dataset (continuous and categorical), selecting variables based on two-step clustering and logistic regression (continuous and cluster number variable), and selecting variables based on a feature-selection method. The outputs of the three models are applied in various machine learning models, including random forest, neural network, CR-Tree, C5 tree, Bayesian network, CHAID tree, support vector machine, quest tree, multinomial logistic regression, and a linear regression model. For brevity, only the top five machine learning models were considered in this study. This section is divided into subsections: data collection, developed prediction models, machine learning models, and performance metrics.

4.0.1. Data Collection:

The dataset used in this project is the Credit Card Approval dataset from the UCI Machine Learning Repository. Customers have the option to choose one of four credit card types: blue, silver, gold, or platinum. When customers decide to change their bank, they are recorded as churn customers. The dataset contains the following data: a churn value (dependent), age, gender, number of dependents, education level, marital status, income category, product variable (type of credit card—blue, silver, gold, platinum), period of relationship with bank, total number of products held by the customer, number of months inactive of the last 12 months, number of contacts in the last 12 months, credit limit on the credit card, total revolving balance on the credit card, open to buy credit line (average of last 12 months), change in transaction amount (Q4 over Q1), total transaction amount (last 12 months), total transaction count (last 12 months), change in transaction count (Q4 over Q1), and average card utilization ratio.

Firstly, I will divide the dataset into categorical and continuous variables as independent variables and one dependent variable (churn customers). Next, I will analyze the variables using different statistical metrics including min, max, variance, standard deviation, chi square (for categorical variable), and correlation analysis (for continuous variable). If The initial analysis will show that the linear relationship between variables does not exist; therefore, a nonlinear model was applied to develop a prediction model for customer churn.

4.0.2. Developing Customers Churn Prediction Models

To develop a prediction model, I intend to use three methods to control the number of independent variables used in the prediction model. First, all independent variables will be directed to one of the applied machine learning models, which is referred to as Model 1. Next, I intend to improve the independent variables by applying a two-step clustering method to the categorical variables only. Afterwards, the continuous variables with the cluster values will be forwarded to each one of the machine learning models. To make the model more realistic, a logistic regression approach will be used to predict the cluster number based on the categorical variables. This model will be denoted as Model 2. Model 2 will be divided into two phases: the clustering phase and the prediction phase. In the clustering phase, the dataset will be divided into groups using two-step clustering, and the continuous variables with a group variable were used to build a prediction model using the neural network.

In the clustering model, the categorical variables will be used to divide the customers into a specific number of groups. The purpose of this step is aimed to minimize the number of input variables forwarded to the machine learning model, in addition to simplifying the meta data of the customers. Moreover, the k-nearest neighbor model will be used to ensure that the developed model was suitable for the online scenario and to avoid repeating the clustering analysis for future data. The k-nearest neighbors model uses customers' categorical meta data as input and the cluster number from the two-step clustering step as output.

Furthermore, in the prediction phase, one of the machine learning models will be used to build a prediction model by receiving the inputs from the two-step clustering step and the continuous meta data of the customers. To build a prediction model which depends on machine learning, the dataset will be divided into training, validating, and testing data. The training data will be used to learn the network from the previous information about the churn and nonchurn customers. The test data were used to test the capability of the machine learning model in predicting the future churn customers. Based on previous research, the best percentages for building training, validating, and testing data for a dataset were found to be 70%, 15%, and 15%, respectively. The proposed Model 2 based on k-nearest neighbors and two-step clustering. Two-step clustering is a tool designed to handle the nature of the data and to find main insights. The differences between the two-step method and other clustering models include the following: it can use both categorical and continuous variables; then, it can automatically choose the appropriate number of clusters. Grouping data using the two-step clustering method involves an initial use of the distance measure to divide the data into groups; then, probabilistic approach is applied to select the optimal group. For Model 3, all the independent variables, including the categorical and continuous variables, will be forwarded to the feature-selection method to select the features that were related to the churn customers. The feature-selection method ranked the features as important, marginal, and unimportant. Only the important variables were forwarded to the machine learning models to be used in building the prediction models.

4.0.3. Machine Learning Models

This project will adopt three methods for selecting independent variables, which will aim to understand the most suitable model for improving the performance of the prediction model based on machine learning models. The project will apply approximately ten or more machine learning models: random forest, neural network, CR-Tree, C5 tree, Bayesian network, CHAID tree, support vector machine, quest tree, multinomial logistic regression, and a linear regression model.

4.0.4. Performance Metrics

To design a predictor, the dataset will be divided into training, validating, and testing datasets, with cutoff percentages of 70%, 15%, and 15% for training, testing, and validating the data, respectively. The performance of churn customers prediction models can be evaluated by using classification parameter variables: recall, precision, accuracy, false omission rate, and F1 score. To find the performance metrics, a confusion matrix will be generated first using the output of the classification results.
The confusion matrix contains four parts: true positive (TP), true negative (TN), false positive (FP), and false negative (FN).

The definitions of the confusion matrix are:

1. True positives (TP): Number of churn customers correctly predicted as true churn.
2. True negatives (TN): Number of non-churn customers correctly predicted as non-churn.
3. False positives (FP): Number of churn customers incorrectly predicted as non-churn.
4. False negatives (FN): Number of non-churn customers incorrectly predicted as churn.

To calculate performance metrics, the following equations are used:

$$Accuracy = TP + TN\ TP + FP + FN + TN \quad \text{.....(1)}$$

$$precision = TP\ TP + FP \quad \text{.....(2)}$$

$$Recall = TP\ TP + FN\ False\ Omission\ rate \quad \text{.....(3)}$$

Conclusion

Building this credit card predictor, is expected to tackle some of the most widely-known preprocessing steps such as scaling, label encoding, and missing value imputation.

Author Contributions

Formal analysis, H.A.-N.; investigation, D.A.-N., N.A.-R. and H.A.-N.; methodology, D.A.-N., N.A.-R. and H.A.-N.; writing—original draft preparation, D.A.-N., N.A.-R. and H.A.-N.; writing—review and editing, D.A.-N., N.A.-R. and H.A.-N.

References

1. Jagadeesan, A.P. Bank customer retention prediction and customer ranking based on deep neural networks.
Int. J. Sci. Dev. Res. 2020, 5, 444–449. [Google Scholar]
2. Amuda, K.A.; Adeyemo, A.B. Customers churn prediction in financial institution using artificial neural network.
arXiv 2019, arXiv:1912.11346. [Google Scholar]
3. Kim, S.; Shin, K.-S.; Park, K. An application of support vector machines for customer churn analysis: Credit
card case. In Proceedings of the International Conference on Natural Computation, Changsha, China, 27–29
August 2005; Springer: Berlin/Heidelberg, Germany; pp. 636–647. [Google Scholar]
4. Kumar, D.A.; Ravi, V. Predicting credit card customer churn in banks using data mining. Int. J. Data Anal.
Tech. Strateg. 2008, 1, 4–28. [Google Scholar] [CrossRef]
5. Keramati, A.; Ghaneei, H.; Mirmohammadi, S.M. Developing a prediction model for customer churn from
electronic banking services using data mining. Financ. Innov. 2016, 2, 10. [Google Scholar] [CrossRef][Gree
n Version]
6. Bastan, M.; Akbarpour, S.; Ahmadvand, A. Business dynamics of iranian commercial banks. In Proceedings of
the 34th International Conference of the System Dynamics Society, Delft, The Netherlands, 17–21 July 2016.
[Google Scholar]
7. Bastan, M.; Bagheri Mazrae, M.; Ahmadvand, A. Dynamics of banking soundness based on CAMELS Rating
system. In Proceedings of the 34th International Conference of the System Dynamics Society, Delft, The
Netherlands, 17–21 July 2016. [Google Scholar]
8. Iranmanesh, S.H.; Hamid, M.; Bastan, M.; Hamed Shakouri, G.; Nasiri, M.M. Customer churn prediction using
artificial neural network: An analytical CRM application. In Proceedings of the International Conference on
Industrial Engineering and Operations Management, Bangkok, Thailand, 5–7 March 2019; pp. 23–26. [Google
Scholar]
9. Domingos, E.; Ojeme, B.; Daramola, O. Experimental analysis of hyperparameters for deep learning-based

churn prediction in the banking sector. Computation 2021, 9, 34. [Google Scholar] [CrossRef]
10. Chen, S.C.; Huang, M.Y. Constructing credit auditing and control & management model with data mining
technique. Expert Syst. Appl. 2011, 38, 5359–5365. [Google Scholar]
11. Hadden, J.; Tiwari, A.; Roy, R.; Ruta, D. Computer assisted customer churn management: State-of-the-art and
future trends. Comput. Oper. Res. 2007, 34, 2902–2917. [Google Scholar] [CrossRef]
12. Risselada, H.; Verhoef, P.C.; Bijmolt, T.H. Staying power of churn prediction models. J. Interact. Mark. 2010,
24, 198–208. [Google Scholar] [CrossRef]
13. Kim, H.S.; Yoon, C.H. Determinants of subscriber churn and customer loyalty in the Korean mobile telephony
market. Telecommun. Policy 2004, 28, 751–765. [Google Scholar] [CrossRef]
14. Xia, G.; He, Q. The research of online shopping customer churn prediction based on integrated learning. In
Proceedings of the 2018 International Conference on Mechanical, Electronic, Control and Automation
Engineering (MECAE 2018), Qingdao, China, 30–31 March 2018; pp. 30–31. [Google Scholar]
15. Olaniyi, A.S.; Olaolu, A.M.; Jimada-Ojuolape, B.; Kayode, S.Y. Customer churn prediction in banking industry
using K-means and support vector machine algorithms. Int. J. Multidiscip. Sci. Adv. Technol. 2020, 1, 48–54.
[Google Scholar]
16. Nie, G.; Rowe, W.; Zhang, L.; Tian, Y.; Shi, Y. Credit card churn forecasting by logistic regression and decision
tree. Expert Syst. Appl. 2011, 38, 15273–15285. [Google Scholar] [CrossRef]
17. Seng, J.L.; Chen, T.C. An analytic approach to select data mining for business decision. Expert Syst. Appl.
2010, 37, 8042–8057. [Google Scholar] [CrossRef]
18. Tsai, C.F.; Lu, Y.H. Customer churn prediction by hybrid neural networks. Expert Syst. Appl. 2009, 36, 12547–
12553. [Google Scholar] [CrossRef]
19. Rahman, M.; Kumar, V. Machine learning based customer churn prediction in banking. In Proceedings of the
2020 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA),
Coimbatore, India, 5–7 November 2020; IEEE: Piscataway, NJ, USA; pp. 1196–1201. [Google Scholar]
20. Khodabandehlou, S.; Rahman, M.Z. Comparison of supervised machine learning techniques for customer
churn prediction based on analysis of customer behaviour. J. Syst. Inf. Technol. 2017, 19, 65–93. [Google
Scholar] [CrossRef]

21. Miguéis, V.L.; Van den Poel, D.; Camanho, A.S.; e Cunha, J.F. Modeling partial customer churn: On the value
of first product-category purchase sequences. Expert Syst. Appl. 2012, 39, 11250–11256. [Google Scholar]
[CrossRef]
22. Kolajo, T.; Adeyemo, A.B. Data Mining technique for predicting telecommunications industry customer churn
using both descriptive and predictive algorithms. Comput. Inf. Syst. Dev. Inform. J. 2012, 3, 27–34. [Google
Scholar]
23. Kaya, E.; Dong, X.; Suhara, Y.; Balcisoy, S.; Bozkaya, B. Behavioral attributes and financial churn prediction.
EPJ Data Sci. 2018, 7, 41. [Google Scholar] [CrossRef][Green Version]
24. Miao, X.; Wang, H. Customer churn prediction on credit card services using random forest method. In
Proceedings of the 2022 7th International Conference on Financial Innovation and Economic Development
(ICFIED 2022), Online, 14–16 January 2022; Atlantis Press: Paris, France; pp. 649–656. [Google Scholar]
25. de Lima Lemos, R.A.; Silva, T.C.; Tabak, B.M. Propension to customer churn in a financial institution: A
machine learning approach. Neural Comput. Appl. 2022, 4, 11751–11768. [Google Scholar] [CrossRef]
26. Bharathi, S.V.; Pramod, D.; Raman, R. An ensemble model for predicting retail banking churn in the youth
segment of customers. Data 2022, 7, 61. [Google Scholar] [CrossRef]
27. Saias, J.; Rato, L.; Gonçalves, T. An approach to churn prediction for cloud services recommendation and
user retention. Information 2022, 13, 227. [Google Scholar] [CrossRef]
28. Thakkar, H.K.; Desai, A.; Ghosh, S.; Singh, P.; Sharma, G. Clairvoyant: AdaBoost with cost-enabled costsensitive classifier for customer churn prediction. Comput. Intell. Neurosci. 2022, 2022, 9028580. [Google
Scholar] [CrossRef] [PubMed]
29. Xiahou, X.; Harada, Y. Customer churn prediction using AdaBoost classifier and BP neural network techniques
in the E-commerce industry. Am. J. Ind. Bus. Manag. 2022, 12, 277–293. [Google Scholar] [CrossRef]
30. Nie, G.; Wang, G.; Zhang, P.; Tian, Y.; Shi, Y. Finding the hidden pattern of credit card holder's churn: A case
of china. In Proceedings of the International Conference on Computational Science, Vancouver, BC, Canada,
29–31 August 2009; Springer: Berlin/Heidelberg, Germany; pp. 561–569. [Google Scholar]
31. Kulatilleke, G.K. Challenges and complexities in machine learning based credit card fraud detection. arXiv

2022, arXiv:2208.10943. [Google Scholar]
32. Alfaiz, N.S.; Fati, S.M. Enhanced credit card fraud detection model using machine learning. Electronics 2022,
11, 662. [Google Scholar] [CrossRef]
33. Jovanovic, D.; Antonijevic, M.; Stankovic, M.; Zivkovic, M.; Tanaskovic, M.; Bacanin, N. Tuning machine
learning models using a group search firefly algorithm for credit card fraud detection. Mathematics 2022, 10,
2272. [Google Scholar] [CrossRef]
34. Al-Najjar, D.; Assous, H.F.; Al-Najjar, H.; Al-Rousan, N. Ramadan effect and indices movement estimation: A
case study from eight Arab countries. J. Islam. Mark. 2022. ahead-of-print. [Google Scholar] [CrossRef]
35. Al-Najjar, D.; Al-Najjar, H.; Al-Rousan, N. Evaluation of the prediction of COVID-19 recovered and unrecovered
cases using symptoms and patient's meta data based on support vector machine, neural network, CHAID and
QUEST Models. Eur. Rev. Med. Pharmacol. Sci. 2021, 25, 5556–5560. [Google Scholar]
36. Al-Rousan, N.; Al-Najjar, H.; Alomari, O. Assessment of predicting hourly global solar radiation in Jordan
based on Rules, Trees, Meta, Lazy and Function prediction methods. Sustain. Energy Technol. Assess. 2021,
44, 100923. [Google Scholar] [CrossRef]
37. Al-Najjar, D.; Al-Najjar, H.; Al-Rousan, N.; Assous, H.F. Developing Machine Learning Techniques to
Investigate the Impact of Air Quality Indices on Tadawul Exchange Index. Complexity 2022, 2022, 1–12.
[Google Scholar] [CrossRef]
38. Al-Najjar, H.; Al-Rousan, N. A classifier prediction model to predict the status of Coronavirus COVID-19
patients in South Korea. Eur. Rev. Med. Pharmacol. Sci. 2020, 24, 3400–3403. [Google Scholar]
39. Rajamohamed, R.; Manokaran, J. Improved credit card churn prediction based on rough clustering and
supervised learning techniques. Clust. Comput. 2018, 21, 65–77. [Google Scholar] [CrossRef]
40. AL-Rousan, N.; Mat Isa, N.A.; Mat Desa, M.K.; AL-Najjar, H. Integration of logistic regression and multilayer
perceptron for intelligent single and dual axis solar tracking systems. Int. J. Intell. Syst. 2021, 36, 5605–5669.
[Google Scholar] [CrossRef]
41. Al-Najjar, H.; Alhady, S.S.N.; Saleh, J.M. Improving a run time job prediction model for distributed computing
based on two level predictions. In Proceedings of the 10th International Conference on Robotics, Vision,

Signal Processing and Power Applications, Pulau Pinang, Malaysia, 14–15 August 2018; Springer: Singapore;
pp. 35–41. [Google Scholar]
42. Al-Najjar, H.; Alhady, S.S.N.; Mohamad-Saleh, J.; Al-Rousan, N. Scheduling of workflow jobs based on
twostep clustering and lowest job weight. Concurr. Comput. Pract. Exp. 2021, 33, e6336. [Google Scholar]
[CrossRef]