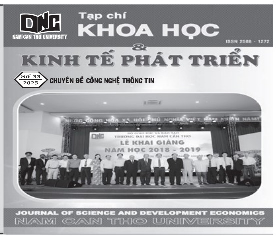
Journal of Science and Development Economics
Nam Can Tho University

Website: jsde.nctu.edu.vn



Predicting customer churn in banking with EKI's algorithms for adapting Vietnamese market

Nguyen My Phung[2], Bui Thi Diem Trinh[2], Nguyen Anh Duy[1*], Ngo Ho Anh Khoi[2]

[1]*Adhightech Ltd., Vietnam*

[2]*Faculty of Information Technology, Nam Can Tho University, Vietnam*

[*]*Corresponding author: Nguyen Anh Duy (email: nguyenanhduy@adhigtechn.com)*

Received: 2/1/2025

Revised: 20/1/2025

Accepted: 10/2/2025

Keywords: *AI, churn, churn prediction, data stream, deep learning, evolving machine learning, machine learning*

Từ khóa: *AI, dự đoán tỷ lệ rời bỏ, học sâu, học máy, học máy đang phát triển, luồng dữ liệu, tỷ lệ rời bỏ*

ABSTRACT

With the rapid advancement of technology, a specific branch of artificial intelligence, known as machine learning, has emerged. This technique has significant potential and plays a crucial role in various industries, including digital transformation, finance, and banking. One of its applications is in identifying potential customers and mitigating risks that may lead to customer attrition. In this article, we will discuss the use of a database, called Churn Modeling, which collects statistical data from banks. We will also explore the application of the BernoulliNB algorithm, combined with the incremental machine learning method, to process streaming data and analyze and predict customer churn rates in banks. The ultimate goal is to provide timely solutions to retain customers. The experimental results demonstrate that this combination yields positive outcomes and has been successfully implemented in the development of a prototype electronic platform.

TÓM TẮT

Với sự phát triển không ngừng của công nghệ, một nhánh của trí tuệ nhân tạo được gọi là học máy đã xuất hiện. Đây là một trong những kỹ thuật có tiềm năng đáng kể và đóng vai trò quan trọng trong các lĩnh vực chuyển đổi số, tài chính, ngân hàng trong việc xác định khách hàng tiềm năng và giảm thiểu rủi ro có thể dẫn đến mất khách hàng. Bài viết này sẽ thảo luận về việc sử dụng cơ sở dữ liệu được thu thập từ dữ liệu thống kê từ các ngân hàng có tên Churn_Modeling, áp dụng thuật toán BernoulliNB kết hợp với phương pháp học máy gia tăng để xử lý dữ liệu phát trực tuyến nhằm *phân tích và dự đoán tỷ lệ rời bỏ của khách hàng tại ngân hàng. Điều này nhằm mục đích cung cấp các giải pháp kịp thời để giữ chân khách hàng. Kết quả thực nghiệm cho thấy sự kết hợp này mang lại kết quả tốt và đã được triển khai trong việc phát triển nền tảng điện tử nguyên mẫu.*

1. INTRODUCTION

In the current developing and highly competitive economy, retaining customers is of utmost importance for banks. Each bank is employing its own strategies and policies to attract customers who may consider discontinuing their services. However, implementing these policies in practice presents another challenge. Fortunately, the recent development of data science has provided solutions to many such issues. This is crucial for banks to address the losses incurred from customers discontinuing service by utilizing the rich data at their disposal to improve the accuracy of their predictions.

This study proposes a model for predicting customer churn, providing a basis for banks to formulate strategies to retain customers who are inclined to stop using their services. Currently, up to 75% of customers have stated they are willing to switch brands if they encounter a poor experience or if issues are not resolved promptly. This is why every business needs to focus on developing a workforce that can resolve problems quickly and professionally. With the increasing customer churn rate, not only in the banking sector but also across various businesses, it is imperative to identify potential customers and offer suitable services to retain them. A high churn rate can adversely affect profits and hinder growth. The churn rate is a critical factor in the telecommunications industry. In most sectors, competition among businesses facilitates users' easy transition from one provider to another [5].

According to a report by Forrester, in 2023, there will be a significant shift in customer shopping intentions. Specifically, customers are becoming increasingly discerning in their consumption decisions. They demand more in every aspect, including product quality and the quality of the experience. This is attributed to three main reasons: the COVID-19 pandemic, the explosion of technology, and economic downturns [4]. There are many reasons that lead to customer churn, including a lack of quality control, not conducting regular quality checks, imposing rules and regulations on customer interaction, and poor customer service, which are the most common causes of this issue. Customers are the lifeblood of any business, regardless of the industry. Losing customers is a fast track to failure, and it is a significant concern that many business owners are acutely aware of. In a world with abundant choices, standing out is harder than ever [3].

From these issues, this research team aims to experiment with applying the Bernoulli Naive Bayes (BernoulliNB) algorithm. Therefore, the primary objective of this study will also be to apply the classical BernoulliNB algorithm combined with techniques capable of frequently updating and transforming the internal content, to address the problem of customer churn rates in banks, starting from international data and gradually adapting the variables for data inputs from Vietnam.

In the first part, the research team will explore the theoretical foundations and research methods: the methods of measuring and analyzing churn rates developed in the past, the history of databases related to customer churn statistics, the history of algorithms that have been applied to predict customer churn rates in banking, and the methods and metrics for analyzing data. The experimental section will include adjustments to the database, algorithm setup, experimental results, and diagrams of the experimental results. This part will present the methodology for implementing and experimenting with the BernoulliNB algorithm to evaluate and select the most suitable model for predicting customer churn rates. The final step presents the implementation of a test application in the form of an online page to predict customer churn rates for banks. This test page is designed to be user-friendly, easy to navigate, and provides all necessary information for users regarding the content and operation of the customer churn prediction method for banks. Finally, there will be a concluding section.

2. RESEARCH METHODS

Currently, databases are facing the issue of focusing on datasets that do not change over time, primarily because they are trained using classical algorithms (where training is done once and requires retraining from scratch with new incoming data; for example, if data 1 is trained into a model and data 2 arrives, data 1 must retrain from the beginning along with data 2 to create a new model). In contrast, in modern reality, data environments change over time, necessitating continuous training in real time, with model updates. Thus, data learning must occur in continuously changing environments, meaning the experimental method will be machine learning through continuous learning in unstable environments.

Some methods have been applied to these classical algorithms to transform them into continuous learning methods instead of using sliding window techniques to modernize classical machine learning methods. A description of the Sliding Windows approach is as follows: It considers the development of concepts in the most recent non-fixed training data environment, determined by a time window of a specified size (based on time scale or the number of data points). This approach can perform reclassification of the "group" (on data selected by the temporary window) or update models if the online learning method permits. In this case, the "forgetting" (as mentioned above) is automatically managed by this learning method. This type of method typically involves three steps: 1) Detecting concept changes using statistical tests on different windows; 2) If an observable change occurs, selecting recent representative data to adjust models; and 3) Updating models. The window size is predefined by the user. The key point of these methods is to determine the window size. Most methods use a fixed-size window configured for each practical problem. In this way, classical algorithms can be used in dynamic environments, but do not embody incremental learning (not reusing stored data, using only models for improvement). Therefore, the history of the following algorithms focuses on presenting incremental learning algorithms, which have been researched and developed in recent years.

The system calculates the balanced accuracy metrics for the results. Traditional accuracy measures the proportion of correctly classified cases out of the total number, and it is generally reliable. However, this metric can be misleading in cases of severe class imbalance, such as a 90:10 ratio. For example, if 100 cases are tested with 99 cases being diseased and 1 case healthy, the balanced accuracy might appear high even if no meaningful model is present. Therefore, for imbalanced datasets, Balanced Accuracy (BA) is used. The choice of evaluation metrics depends on the problem's objectives and the composition of the dataset. In situations with significant class imbalance, where one class is underrepresented, traditional accuracy becomes unreliable. Therefore, metrics like the area under the ROC curve (AUC) and BA are preferred. Metrics such as balanced accuracy, sensitivity, and specificity are less effective for imbalanced data. For concordance detection, metrics based on the true positive rate/false positive rate, such as balanced accuracy, sensitivity, and F-Score, are appropriate. In contrast, for discordance detection, metrics based on the true negative rate/false negative rate, such as specificity, are suitable, although less common in practice. Sensitivity, balanced accuracy, and F-Score are criticized for ignoring the true negative cell of the confusion matrix and being prone to prediction bias [1]. BA, which includes both the true positive rate and the true negative rate, provides a balanced evaluation, making it suitable for both concordance detection and imbalanced data situations [2]. BA is an important and simple metric for evaluating binary classifiers in the context of class imbalance, where one class is much more prevalent than the other. The formula for balanced accuracy (BA), which provides a practical and optimal evaluation, is:

Balanced Accuracy (BA) = ½ (Specificity + Sensitivity) (1)

At present, the high risk of customer churn among many clients is a serious issue that directly affects profitability and hinders the growth of businesses, particularly in the banking environment. In the process of searching for data related to this topic, many datasets were found. However, the most accessible datasets that lack complete parameters include: the “Bank Turnover Dataset” published by Tarun Sunkaraneni, "Binary Classification with a Bank Churn_Dataset" [6], and "Bank Customer Churn Prediction" [7] shared by Shubham Meshram.

A brief description of the dataset “Binary Classification with a Bank Churn_Dataset” published by Brensh Lytcher reveals that it is a massive dataset divided into two files: Feature_Combinated_test.csv (700MB) and Feature_Combinated_train.csv (1.05 GB). This database features a considerable number of fields, totaling 398. Many parameters remain unverified, and the information describing each parameter is not clearly indicated. Therefore, this dataset is too complex for application in research and development. Next is the dataset "bankchurn," which was recently published by author Phyngyn. This dataset aggregates attributes and analyses collected from various banks. It contains 23 fields, each describing customer characteristics. However, due to its complexity and many fields lacking clear descriptions by the author, this dataset cannot be applied in the experimental process of the problem.

Now, regarding the “Bank Customer Churn Prediction” dataset, this dataset was researched and published by Shubham Meshram in June 2024. It includes 14 fields and 10,000 data samples. This dataset is the most relevant regarding the customer churn rate and is highly rated for usability on Kaggle. Based on the datasets found, the one that can be used for this project must be numerical data, with specific classifications, and the most recent updates. Only the “Bank Customer Churn Prediction” dataset meets these requirements, making it the most suitable choice for this project. For the reasons mentioned in the previous section, the “Bank Customer Churn Prediction” dataset will be chosen as the primary dataset for researching the prediction of customer churn rates in banking. The dataset includes various features and is described as follows:

RowNumber—corresponds to the record (row) number and has no effect on the output.

CustomerId—contains random values and has no effect on customer leaving the bank.

CreditScore—can have an effect on customer churn, since a customer with a higher credit score is less likely to leave the bank. Higher credit scores typically indicate more reliable and financially stable customers.

Geography—a customer’s location can affect their decision to leave the bank. Different regions may have varying levels of competition, economic conditions, and cultural factors that influence customer loyalty.

Gender—it’s interesting to explore whether gender plays a role in a customer leaving the bank. There may be underlying societal or economic reasons that cause different churn rates between genders.

Age—this is certainly relevant, since older customers are less likely to leave their bank than younger ones. Age often correlates with stability and long-term financial planning.

Tenure—refers to the number of years that the customer has been a client of the bank. Normally, older clients are more loyal and less likely to leave a bank. Longer tenure usually indicates higher satisfaction and deeper integration with the bank’s services.

Balance—also a very good indicator of customer churn, as people with a higher balance in their accounts are less likely to leave the bank compared to those with lower balances. A higher balance suggests a greater level of engagement and investment in the bank.

NumOfProducts—refers to the number of products that a customer has purchased through the bank. More products indicate a stronger relationship with the bank, making customers less likely to leave.

HasCrCard—denotes whether or not a customer has a credit card. This column is also relevant, since people with a credit card are less likely to leave the bank. Credit cards often create more touchpoints and dependencies with the bank.

IsActiveMember—active customers are less likely to leave the bank. Activity levels can reflect customer satisfaction and the perceived value of the bank's services.

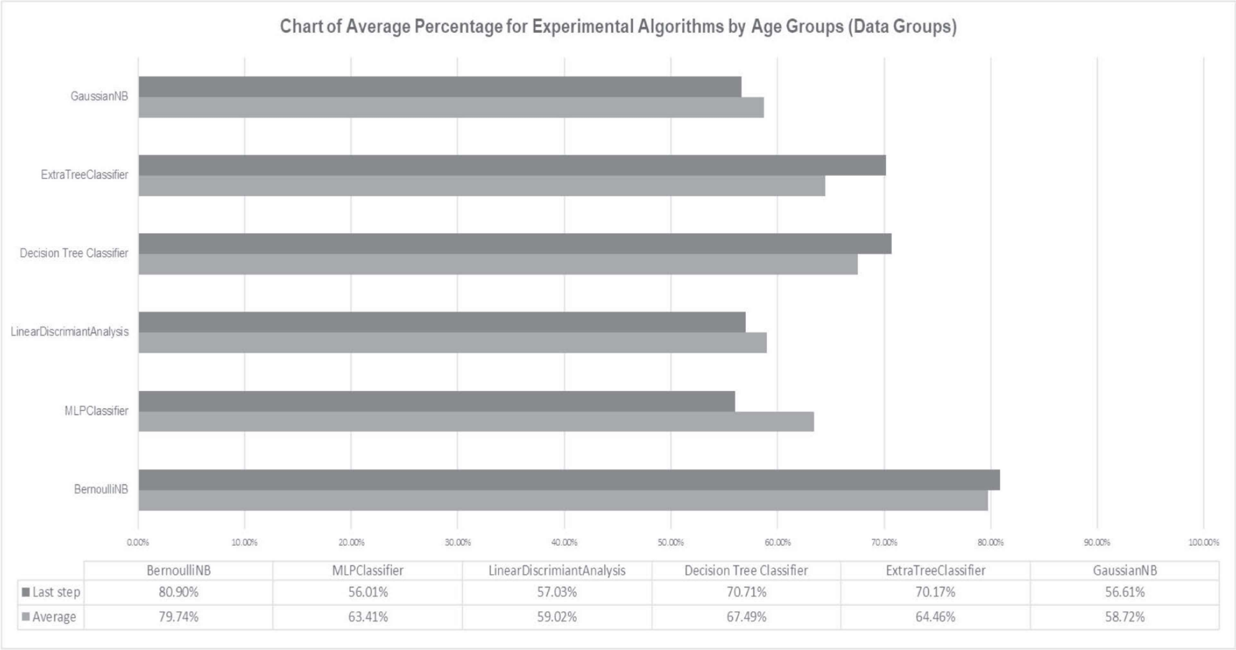
EstimatedSalary—as with balance, people with lower salaries are more likely to leave the bank compared to those with higher salaries. Higher salaries may correlate with greater financial stability and satisfaction with banking services.

Class—whether or not the customer left the bank. This is the target variable used to determine customer churn.

Regarding the methodology used during the experiment, the dataset for this experiment consists of two parts: the training data and the testing data. The training dataset includes 7,000 samples (accounting for 70% of the original data), while the testing dataset consists of 3,000 samples (accounting for 30% of the original data). The positions of these data points will change in each experiment; each time, they will be randomly shuffled and reshuffled after training.

3. RESULTS AND DISCUSSION

Using artificial intelligence methods, specifically six algorithms including BernoulliNB, MLPClassifier, Linear Discriminant Analysis, Decision Tree Classifier, ExtraTreeClassifier, and GaussianNB, combined with the sliding window method. By using a chart to compare the average results of the nine algorithms, this approach provides the fairest representation of data fidelity when comparing results from the algorithms. The average experimental results of the algorithms are presented as Figure 1:



	BernoulliNB	MLPClassifier	LinearDiscrimiantAnalysis	Decision Tree Classifier	ExtraTreeClassifier	GaussianNB
■ Last step	80.90%	56.01%	57.03%	70.71%	70.17%	56.61%
■ Average	79.74%	63.41%	59.02%	67.49%	64.46%	58.72%

Figure 1. Chart of average percentage for experimental algorithms by age groups (Data groups)

Looking at Figure 1, using the method of comparing the average results and the final results of the 6 algorithms, we see that there are differences among the algorithms. With the average value, the chart compares the average percentage of the algorithms and shows a clear difference in performance among them. Some algorithms perform better than others on this dataset. To gain a better understanding of the performance of each algorithm, we will divide them into two groups: a high-performance group and a low-performance group, based on the average value. The high-performance group includes algorithms with high performance, such as BernoulliNB with 79.74%, which has the highest average performance among all the algorithms, indicating very good and stable classification capability. Next is the Decision Tree Classifier, which is also a high-performance algorithm with 67.49%, indicating that this model is also very effective in classification. Finally, the Extra Tree Classifier has a value close to that of the Decision Tree with an average of 64.46%, which is also an effective and stable classification model.

On the other hand, the low-performance group consists of algorithms with lower average values, such as MLPClassifier: 63.41%. The performance of MLPClassifier is not high and decreases in the final step, indicating that this model may not be stable. The Linear Discriminant Analysis has an average performance of 59.02%, suggesting that this model is not entirely suitable for this dataset and may need further improvement. GaussianNB has a lower performance compared to the other algorithms, with an average of 58.72%, indicating that this algorithm may not be the best choice for this dataset. Based on the average performance, BernoulliNB is considered the preferred algorithm for this dataset with the highest average performance (79.74%). However, the stability in the final steps of the 6 algorithms also needs to be considered to make the best choice. Specifically, as shown in Figure 2.

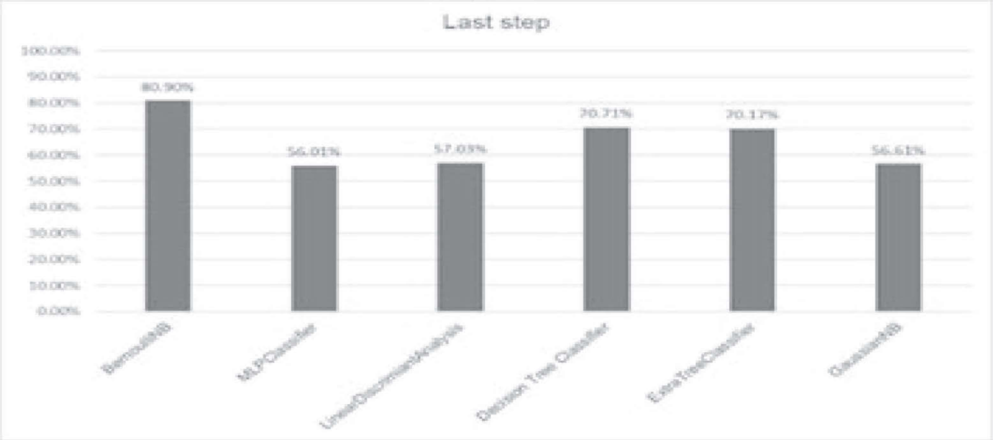


Figure 2. The last step result chart of the 6 algorithms

From Figure 2, we can see a clear difference in performance among the algorithms in the final step of the training process. The BernoulliNB algorithm has the highest final step performance at 80.90%, indicating very good and stable classification capabilities. It is followed by two algorithms at around 70%, including Decision Tree with 70.71% and Extra Tree with 70.17%. Algorithms such as LDA with 57.03%, GaussianNB with 56.61%, and MLPClassifier with 56.01% exhibit relatively low final step performance, averaging just above 50%, suggesting that they may not be as suitable as the high-performance algorithms for this dataset. Based on the final step performance, BernoulliNB is considered the preferred algorithm for this dataset, with average performances of 79.74% and a final step performance of 80.9%. This represents a strong and stable choice for data classification. In addition to averaging the results of the algorithms, another approach such as comparing the experimental model results by age provides a more comprehensive and detailed overview, helping us visually assess to reach the most accurate conclusion about the experimental model results by age, as depicted in Figure 3:

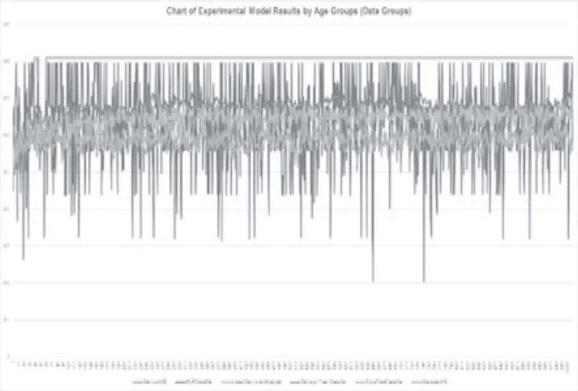


Figure 3. BA progression chart of the six algorithms

With data visualized from the Figure 3 chart, we notice that all 6 algorithms start with low scores in the initial steps. This is commonly seen in most algorithms due to the initial amount of data being insufficient to capture the important characteristics of the data. However, after the initial steps, the algorithms begin to stabilize. Some algorithms, such as BernoulliNB and Decision Tree Classifier, show rapid stabilization, while others, such as MLPClassifier and GaussianNB, still experience significant fluctuations. Among these, the most noticeable fluctuation is seen in the MLPClassifier algorithm, which starts low and increases but maintains an average level (around 0.56), particularly with considerable data volatility in the early and middle training phases. Therefore, this is not the optimal algorithm to apply to the problem.

Regarding the remaining 5 algorithms, we clearly observe the difference between BernoulliNB and the other 4 algorithms; even with a low starting performance, BernoulliNB quickly rises and stabilizes at a high level in the middle and at the end of the training process with an accuracy of 80.9%. Following that are 2 algorithms that achieve around 70%: Decision Tree and Extra Tree, which score 70.71% and 70.17%, respectively. Decision Tree also has high and stable performance after the initial phase. Extra Tree demonstrates performance and stability similar to Decision Tree, making it another reliable choice, especially when handling large datasets with many features.

The other two algorithms only achieve around 50-60%: LDA at 57.03% and GaussianNB at 56.61%. LDA operates at an average level with relatively stable performance after the initial phase. However, further improvements are needed to enhance performance and stability in more complex scenarios. GaussianNB has average performance and high sensitivity to fluctuations in the data, making it not a preferred choice for complex problems. Linear Discriminant Analysis and GaussianNB perform at an average level, with LDA being more stable but in need of improvement, while GaussianNB is sensitive to data fluctuations and not an optimal choice.

In tasks requiring high performance and stability, the 3 algorithms BernoulliNB, Decision Tree, and Extra Tree Classifier would be the best options. However, it is clear that Bernoulli is the best algorithm among those analyzed, with high and stable performance. This is the preferred choice for classification tasks with binary data.

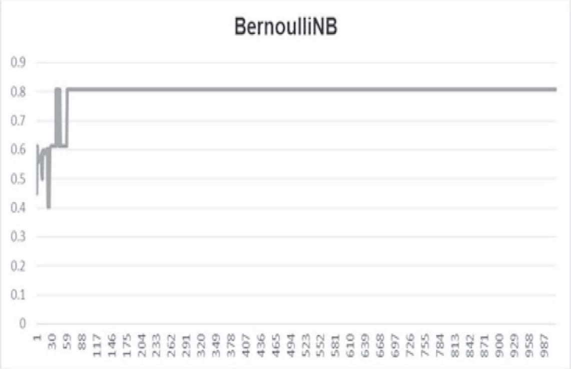


Figure 4. Experimental model result chart by age (BernoulliNB)

In summary, the BernoulliNB algorithm has many advantages and stable performance with an accuracy of 80%. This indicates that BernoulliNB could be a good and useful choice for many prediction and classification problems, especially with the data that the research team is investigating.

System setup:

From the perspective of real-world application, creating the necessary dataset for learning and prediction in the experimental environment in Vietnam is crucial. In Vietnam, due to the lack of related databases, immediately establishing a prediction system is not feasible. The initialization of a dataset for this topic, aimed at conducting learning processes and building a prediction system, requires substantial resources, which could take years or even decades. Currently, implementing this solution remains challenging. A more cost-effective approach, which has been applied in similar cases, is to use continuous learning models. Instead of waiting for a large amount of data to proceed with the learning process and build the prediction system, we can use a small amount of data to improve the model by continuously adjusting the basic concepts and gradually shifting the initial basic concept closer to a new basic concept (based on the Vietnamese dataset). This process is called "concept drift," and the model will be continuously improved by adding more accurate new data (data on Vietnam's land and crops). This method allows the prediction system to be used immediately and gradually improved through small errors in the model, rather than waiting a long time before the improved model can be utilized.

Based on the final results presented in the previous section, the Bagging algorithm was selected to address the problem. The application will include features such as prediction functionality, running classical algorithms, a list of processed models, system configuration, and login. It will be installed on a web environment and will be divided into main functions: the algorithm installer (administrator or developer) and the diagnostician (user), described by the use case diagram below:

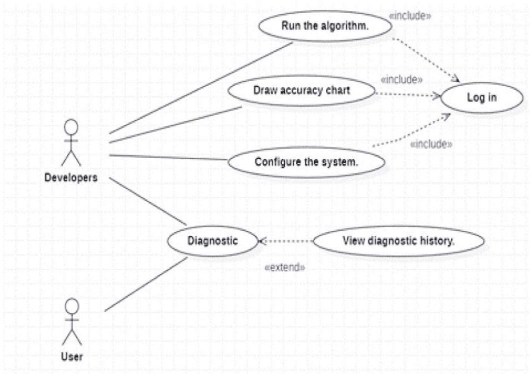
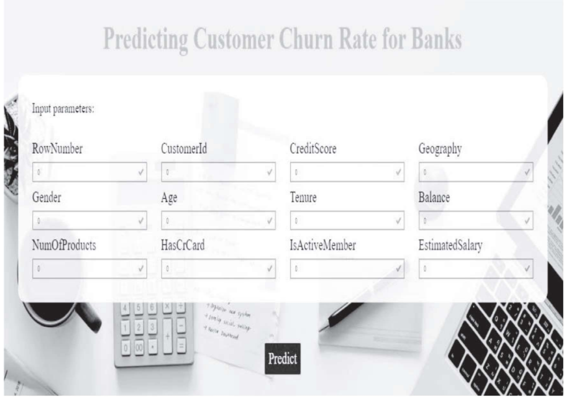


Figure 5. Use case diagram

Download the file "setup.zip," and after extracting it, you will find the following key files and folders: SETUP, DB, APP, INSTALL.bat, RunServer.bat, requirements.txt. Install the Python program by running the 'python-3.9.9-amd64.exe' file located in the SETUP folder. Install the necessary libraries to run the program by executing the CaiThuVien.bat file. Running the Remove.bat file will delete all program data. The database file is located in the 'DB' folder and is named Data.db, which can be opened using the 'DB Browser for SQLite.exe' tool located in 'DB\DB Browser for SQLite'. To change the administrator account, edit the file '\APP\static\dataUser.csv'.

To start the program, run the 'RunServer.bat' file or open the command line and run the command 'manage.py runserver'. The default server port is 8000, which can be changed by using the command ‘manage.py runserver <port>’. When the command line displays ‘Starting development server at http://127.0.0.1:8000/’, you can access the main application page at ‘http://127.0.0.1:8000/’ (Figure 6).



Predicting Customer Churn Rate for Banks

Input parameters:

RowNumber: 0
CustomerId: 0
CreditScore: 0
Geography: 0
Gender: 0
Age: 0
Tenure: 0
Balance: 0
NumOfProducts: 0
HasCrCard: 0
IsActiveMember: 0
EstimatedSalary: 0

Figure 6. Website index

4. CONCLUSION

After completing the research and drafting the report, a comprehensive evaluation of the results has been achieved. The report is structured with clear content, providing specific explanations of the data, charts, and algorithms used. From a system perspective, the program successfully implements the classical BernoulliNB algorithm for the topic. The model's training and prediction processes address the challenge of processing dynamic and variable data, a limitation that many other algorithms face. Future development of the topic will involve updating data through surveys and incorporating medical knowledge to create an optimized and accurate dataset aligned with real-world needs. Additionally, applying web-based platform techniques to smart electronic devices will enhance quality for banks and businesses. The application’s deployment is expected to yield numerous benefits, significantly boosting the market economy and fostering a healthy competitive environment for Vietnam. This system will play a vital role in promoting social progress and alleviating factors contributing to economic downturns.

REFERENCES

[1] Powers, D. M. (2020). *Evaluation: From precision, recall, and F-measure to ROC, informedness, markedness, and correlation.* arXiv preprint arXiv:2010.16061.

[2] Khoi, N. H. A. (2015). *Méthodes de classifications dynamiques et incrémentales: Application à la numérisation cognitive d'images de documents* (Doctoral dissertation). Tours.

[3] Long, A. T. (2024). *8 Reasons Why Your Business Loses Customers.* Retrieved from https://www.bitrix24.vn/articles/8-ly-do-khien-doanh-nghiep-cua-ban-mat-khach-hang.php.

[4] GAPONE. (2023). *Trends in 'Recreating' Customer Experiences in the Digital Age.* Retrieved from https://gapone.vn/trai-nghiem-khach-hang-trong-thoi-dai-so/.

[5] Minh Lan. (2019). *What is the Customer Churn Rate?* Retrieved from https://vietnambiz.vn/ti-le-khach-hang-roi bo-customer-churn-rate-la-gi-20191121181913287.htm.

[6] Brensh Lytcher. (2024). *Binary Classification with a Bank Churn Dataset.* Retrieved from https://www.kaggle.com/datasets/brenshlytcher/binary-classification-with-a-bank-churn-dataset.

[7] Muhammad Husban. (2024). *Bank Customer Churn Prediction.* Retrieved from https://www.kaggle.com/datasets/muhammadhusban/bank-churn.