Predicting Customer Attrition in Claims and Direct Billing Services of Health Insurance

Trung-Duy Nguyen[1], Thanh-Hieu Bui[1*]

[1]Department of Business Information Technology, University of Economics Ho Chi Minh City, Ho Chi Minh City, Vietnam
hieubt@ueh.edu.vn

Abstract. This paper proposes an approach for predicting the likelihood of customers discontinuing their use of claims and direct billing services in health insurance. Focused on delineating customer profiles and segmentation based on service usage behavior, we leverage a comprehensive dataset detailing customers' insurance participation history. We utilize advanced machine learning techniques including K-Means clustering to analyze this data effectively. Additionally, four predictive models including Logistic Regression, Random Forest, k-nearest Neighbor and XGBoost are rigorously tested to ascertain their predictive accuracy and reliability. The critical aspect of this study involves employing various evaluation methodologies to determine which model offers the most accurate prediction. The experimental result shows that the Random Forest model stands out, delivering the highest performance index among the evaluated models. This model's ability to handle complex data patterns makes it particularly effective for this application. Moreover, the study extends beyond mere prediction. It classifies customers into four distinct segments, thereby providing a nuanced understanding of different customer groups. This classification is pivotal for developing targeted and effective customer retention strategies for each segment, underscoring the practical implications of the research for health insurance companies.

Keywords: Claims, Direct Billing Service, Health Insurance, Customer Segmentation, Machine Learning.

T. A. Trinh et al. (eds.), *Proceedings of the 2nd International Conference - Resilience by Technology and Design (RTD 2024)*, Advances in Intelligent Systems Research 186,
https://doi.org/10.2991/978-94-6463-583-6_23

1 Introduction

In the business environment, the habit of using modern technological devices at work, leading to health problems such as back pain and eye strain, has partly helped increase the demand for insurance to protect employees' health. For insurance companies specializing in claims and direct billing services, changing the insurance usage habits of the corporate customer group who do not buy insurance requires significant financial investment and understanding of the traditional market, contributing to the transformation of their insurance usage perspective. Meanwhile, retaining insurance-using customers is the foundation for developing the health insurance market, enhancing the value of services, thereby creating a difference in customer approach. The transformation of corporate customer needs is a factor that requires better solutions in long-term strategies; otherwise, it will be a barrier to the insurance industry. Importantly, factors affecting the probability of customers continuing to use insurance services are about the expectations of quality and reliability from insurance companies. Moreover, as mentioned earlier, retaining customers can minimize costs in maintaining customer relationships. The reason is that customers who are about to leave are more easily persuaded by benefits favorable to them (Tamaddoni et al., 2010). The issue raised is whether there is a method to help businesses understand customer behavior better after they have used their insurance services and from that understanding, can predict the likelihood of discontinuing the use of claims and direct billing services? A system that helps predict the likelihood of discontinuing service use specifically for health insurance customers could be a solution to the problem of businesses.

Customer Relationship Management (CRM) is a customer care process using data integration from various sources to analyze user behavior data and have a plan to care for and maintain increasingly good relationships with customers (Bardicchia, 2020). Among them, the customer behavior factor consists of service usage actions quantified and represented by various variables. These quantified behavioral variables are commonly used, such as the number of medical examinations, frequency of service use, etc. These attributes often come with more complex variables like the

total insurance cost divided by the types of common medical treatment costs. Evidence shows that these variables are statistically significant and improve the model's prediction (Bach et al., 2021; Buckinx & Poel, 2005). It is also known that higher customer expenditures lead to a desire to continue being customers, and the service portfolios used previously affect the customer's decision to stay. A plausible explanation for leaving based on previously used services is that the satisfaction level when using a type of service is low and unrelated to the high cost or low quality of the used service (Mozer, 2000). There are four questions that CRM methods can be used to answer (Oliveira, 2012): (1) Identifying “customers": which occasional buyers have the potential to become permanent customers of the business? (2) Attracting customers: how to convert occasional buyers into customers? (3) Developing customers: How can customers bring value to the business? (4) Retaining customers: how to make customers stay with the company? The last step of CRM is the purpose of this paper, that is, it will help businesses predict potential customers and increase the likelihood of retaining them. On the other hand, improving customer loyalty is to create profits for the company. Within the scope of this study with the research data source and the process of building the experimental model collected from Fullerton Health Group in the period from 2019 to 2022, the main goal of the study is to propose a system predicting the likelihood of abandoning the use of claims and direct billing services by customers in the health insurance field using data mining model methods K-Means, Logistic Regression, Random Forest, k-NN, and XGBoost.

2 Literature Review

The proposed study aims to explore various customer segmentation models to better understand customer behavior, such as the K-Means, Logistic Regression, Random Forest, and k-NN models. Additionally, to achieve the ultimate goal of the research, several machine learning models are deployed to predict customer churn. Subsequently, the model results are evaluated based on performance metrics including Accuracy, Recall, Precision, and F1-score.

2.1 Customer Segmentation Machine Learning Models

K-Means is one of the most commonly used clustering methods to partition n vectors into k clusters based on their features, where $k < n$. The algorithm starts by randomly selecting k initial centroids, then assigns vectors to the nearest centroid using a distance measure, such as Euclidean distance, and recalculates new centroids as the basis for the assigned data vectors. This process is repeated iteratively until the vectors no longer change clusters between iterations (Khadka, 2019). However, in the K-Means algorithm, the number of clusters is chosen arbitrarily, leading to unreliable clustering results if the number of clusters is inaccurately chosen. This raises the fundamental question: How to select the correct number of expected clusters (Dick & Basu, 1994). In this study, two methods were used to determine the optimal number of clusters for K-Means: the Elbow method and the Silhouette score. In this research, the sum of squared errors (SSE) and the average Silhouette coefficient (S) are expressed in the respective formulas below:

$$SSE = \sum_{i=1}^{k} \sum_{x_j \epsilon c_j} \left\| y_i - c_j \right\|^2 \tag{1}$$

$$S_j = \frac{b_j - a_j}{\max(a_j, b_j)} \tag{2}$$

This research applied the K-Means technique with different values of k and then plotted the curves of SSE and the average Silhouette score according to the number of clusters to analyze both curves and determine the optimal number of clusters. The optimal number of clusters can be found in a dataset by finding the number of clusters where there is an elbow, peak, or inflection point in the graph of the evaluation when represented by the number of clusters. The K-Means clustering algorithm divides the input dataset of N rows into a smaller number of subsets, where k is always less than N. The parameter k is chosen randomly. It calculates the mean value for each cluster and determines the distance between them. This process is repeated until the required number of clusters is achieved. The distance is calculated according to the formula:

$$j = \sum_{i=1}^{m}\sum_{k=1}^{k} w_{ik}\|x_i - \mu_k\|^2 \tag{3}$$

2.2 Machine Learning Model for Predicting Usage and Churn Behavior

Logistic Regression Model. Logistic regression (LR) is a classical classification method (Khadka, 2019). It predicts the probability that an observation belongs to a given class. Renjith (2015) predicted customer churn using logistic regression and proposed a strategy for personalized customer retention using machine learning. The model can be represented as $P(Y|X)$ in the form of conditional probability distribution. Considering a binary classification problem, where X is an n-dimensional vector and Y (label) is 0 or 1. The mathematical expression for the predicted outcome is:

$$P(Y = 1|X) = \frac{\exp(w^T x + b)}{1 + \exp(w^T x + b)} \tag{4}$$

$$P(Y = 0|X) = \frac{1}{1 + \exp(w^T x + b)} \tag{5}$$

In which the real values of the linear regression model $z = (w^T x + b)$ are transformed into values in the range [0, 1] by the sigmoid function. In other words, the outcome for a particular observed sample X as calculated by the equation above is the probability that sample X belongs to a certain class (0 or 1). We can classify them by setting a threshold. Generally, the values calculated by the sigmoid function are assigned to class 1 if they are greater than or equal to 0.5, and to class 0 otherwise. The advantage of logistic regression is that the results are easy to interpret and can be applied to both continuous and categorical variables.

Random Forest Model. The Random Forest is widely used in actual customer prediction and can be applied as a fundamental model for loss prediction (Neslin et al., 2006). The dataset on customer churn contains many attributes of customers, and not all of these attribute variables are beneficial for predicting service abandonment.

The redundancy of irrelevant variables in the dataset may hinder the predictive performance of the model. Therefore, it is necessary to select relevant attributes. Random Forest is an effective attribute selection algorithm with relatively high classification accuracy, good noise resistance, and high generalization ability. Therefore, Random Forest is widely used in business management, economics, finance, biological sciences, and other fields. The number of variables in this study is relatively high at 17, so the Random Forest algorithm is very suitable for selecting characteristic attributes. The Random Forest algorithm is one of the well-known machine learning methods for classification problems. The main idea is to apply Random Forest on the recommendation system to extract the top relevant products based on user preferences. The classification technique in RF includes the following stages:

$$h(x) = \frac{1}{P}\sum_{n}^{P=1} h(\gamma, \lambda_P) \tag{6}$$

Random Forest (Khadka, 2019) is a type of ensemble learning technique used for classification, regression, and other tasks by setting up a large number of decision trees during training and then generating an output class for classification problems or predicting values for regression problems. The Gini index can be formally expressed as follows:

$$Gini = 1 - \sum_{x=1}^{c} (p_x)^2 \tag{7}$$

In which, p_x represents the probability of an item being assigned to a specific category. The following formula for entropy can be used to calculate the information gain:

$$entropy = \sum_{x=1}^{c} -p_x * log_2(p_x) \tag{8}$$

In this formula, p represents probability and it is a function of entropy.

XGBoost Model. XGBoost is one of the powerful boosting algorithms in machine learning systems. This algorithm can predict, classify, and optimize defined systems with high accuracy based on data structure. The proposed services in this system contain classification results and predictions combined by applying collaborative filtering techniques. It helps in more efficient proposal and prediction. Ji et al. (2021) proposed a feature selection algorithm based on XGBoost. The algorithm selected features from two perspectives to identify the most important features for predicting customer churn rate while eliminating redundant features. Experimental results show that this method has good predictive performance. To determine the probability of prediction for the provided services, we need to obtain real values, defined as follows:

$$[P_{\alpha}, P_{1-\alpha}] \tag{9}$$

The prediction capability of XGBoost is determined by the equation:

$$\phi_i = \theta_2 z_i + \theta_1 + \lambda_i, i = 1, \dots, M \tag{10}$$

Where the probabilities of observations z_i and λ_i follow independent normal distributions. The proposed probability also encompasses the error rate in the prediction process. To reduce the error rate, it requires increasing the evaluation coefficient of θ_i.

k-NN Model. In statistics, the k-NN (k-Nearest Neighbors) algorithm is a non-parametric supervised learning method developed by Evelyn Fix and Joseph Hodges in 1951, later extended by Thomas Cover. It is used for classification and regression. In both cases, the input consists of k nearest training samples in the dataset. The output depends on whether k-NN is used for classification or regression. In k-NN classification, the output is the class membership. An object is classified based on the majority similarity of its neighboring data points and is assigned to the most common class among its k nearest neighbors (where k is a positive integer, typically small). If k = 1, then the object is simply assigned to the class of its nearest neighbor. In k-NN regression, the output is the attribute value of the object. This value is the average of the attribute values of its k nearest neighbors. The neighbors are taken from a set of objects with known class labels (for k-NN classification) or attribute values of the objects (for k-NN regression). This can be seen as the training set for the algorithm,

although it does not require an explicit training step. k-NN has some strong consistent results. As the amount of data approaches infinity, the k-NN algorithm for binary classification problems is guaranteed to have an error rate no more than twice the error rate of Naive Bayes classification. Various speed improvements for k-NN can be achieved by using proximity graphs. For multi-class k-NN classification, Cover and Hart (1967) proved that the limiting error rate lies within the bound:

$$R^* \leq R_{kNN} \leq R^*\left(2 - \frac{MR^*}{M-1}\right) \tag{11}$$

2.3 Evaluation Methods for Predictive Models

This section of the research considers the precision of the performance evaluation metrics of the proposed model. In reality, models are prone to errors and misclassifications are always present. Comparing predicted values with actual values is a stepping stone for evaluating model performance. All the performance evaluation methods for classification problems include the Confusion Matrix which is a matrix used to determine how many data points are correctly classified into a category. The confusion matrix will tell us which category has the highest correct classification rate and which category's data is often misclassified into another as follows:

True Positive: The total number of positive data points correctly predicted as positive compared to the actual number of positive data points.

False Positive: The total number of positive data points incorrectly predicted compared to the actual number of positive data points.

True Negative: The total number of negative data points correctly predicted as negative compared to the actual number of negative data points.

False Negative: The total number of negative data points incorrectly predicted compared to the actual number of negative data points.

Table 1. Confusion Matrix

	Predicted Positive	Predicted Negative

Actually Positive	True Positive (TP)	True Negative (FN)	P = TP + FN
Actually Negative	False Positive (FP)	False Negative (TN)	N = FP + TN
	P' = TP + FP	N' = FN + TN	P + N = P' + N'

Source: Burez and Van den Poel, 2009

Accuracy: is the percentage of samples correctly predicted out of all the predicted classifications made by the model. Mathematically, it can be defined as:

$$Accuracy = \frac{T_P + T_n}{T_p + F_p + T_n + F_n} \tag{12}$$

Precision: estimates the total number of positive samples correctly identified by the algorithm out of the total number of positive samples predicted. It can be calculated as follows:

$$Precision = \frac{T_p}{T_p + F_p} \tag{13}$$

Recall: estimates the total number of positive samples correctly identified by the algorithm out of the total number of actual positive samples. Mathematically, it can be calculated as follows:

$$Recall = \frac{T_p}{T_p + F_n} \tag{14}$$

F1-score: is calculated by taking the harmonic mean of Precision and Recall. It can be defined as:

$$F1 - score = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{15}$$

3 Methodology and Data

After presenting the relevant theoretical background, this section of the study introduces the proposed research model and experimental results. This includes modeling experiments starting from data collection, preprocessing, implementing the K-Means machine learning model for customer clustering, and machine learning models predicting customer churn. The obtained results consist of customer clustering datasets and predictions of customer churn, which compensate for business health insurance co-payment and warranty fees.

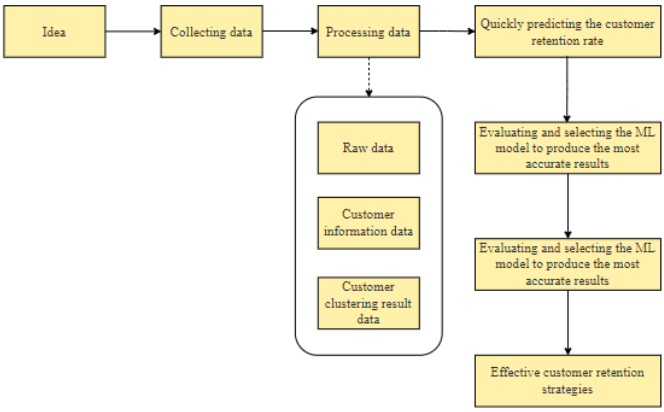

Fig. 1. The overview research model

3.1 Idea Formation

Combining medical treatment history and customer profiles to predict the likelihood or rate of customer service usage in the health insurance sector.

3.2 Data Collection

Medical treatment data, compensation information, and hospital fee guarantees for over 800 thousand customers in the Fullerton Health Group system were collected.

3.3 Data Processing

Medical treatment and compensation/guarantee data were extracted, processed, and transformed into the initial raw data table for subsequent analysis steps. Based on the raw data table, this research processed and summarized information about each customer's medical treatment visits to healthcare facilities and submitted compensation claims. Customer information data was then extracted into 2 columns: Total number of medical visits and Average value of each compensation and hospital fee guarantee record (ACC) for clustering customers using K-Means. Total number of medical visits records the number of medical visits typically logged to track healthcare service usage and calculate claims and direct billing records. Average value of each compensation and hospital fee guarantee record (ACC – Average Claims

Cost) is calculated by dividing the total compensation or hospital fee guarantee amount by the number of compensation or hospital fee guarantee records within a certain period. This helps measure the average value of each medical record. The ACC index is an important tool for insurance companies, policy managers, and risk managers in the health insurance field to assess and forecast compensation or hospital fee guarantee costs, thereby making decisions about pricing, contributions, and insurance premiums.

3.4 Predictive Data

The research combines three types of data to quickly predict customer decisions as follows. Raw data table utilizes customer service usage history data for claims and direct billing as input prediction data. Customer information data table gathers basic information about customer profiles using the service. Customer clustering result data table extracts the value evaluation results of customers according to their clusters.

3.5 Evaluation and Model Selection for Most Accurate Results

After obtaining prediction results, the study evaluates the results through evaluation metrics and selects the optimal model for predicting the likelihood of service abandonment. Ultimately, the research proposes a prediction model with the optimal algorithm through visually representing cluster results, predictions, and summarizing some analyses on the research issue.

3.6 Utilizing Prediction Results

Data Collection. The dataset collected from the Fullerton Health Corporation system is based on criteria including:

Personal information: Including age, gender, address, occupation, marital status, and other customer factors.

Medical service usage history: Records the number of medical visits, surgeries, tests, imaging diagnoses, and other medical services that customers have used in the past.

Compensation and hospital fee guarantee history: Includes information on the number of compensation and hospital fee guarantee records that customers have, total compensation value, contributions, and related details.

Basic information about the dataset: Typically, customer factor data stored in the company's system is readily available for model building. This also ensures that the model validation requirements annually comply with the company's standards without the need for manual data collection. This includes collecting personal information, insurance policy information, medical history: Including information about diseases, hospitals visited, treatment methods, and medical test results, the process of using medical services, and information on claims from customers.

Data Preprocessing. Data preprocessing is an important step in data mining, especially in prediction models using machine learning algorithms. We import necessary Python libraries as follows.

NumPy: This is a basic package for scientific computing in Python. It also supports additional large arrays and matrices, multidimensional.

Matplotlib: Python's library for 2D graph plotting, with this library, we need to import a sub-library Pyplot. This library is used to draw any type of chart in Python.

Pandas: One of the famous Python libraries used to import and manage datasets. It is an open-source data analysis and manipulation library.

First, the dataset of customer insurance service usage is uploaded to Google Drive, then opens Google Colab and connects it to Google Drive to download the data to Google Colab. read_csv() function: To import the dataset, the read_csv() function from the pandas library is used. After uploading the dataset, the (dataset_name].head() function is used to preview the first 5 rows of the dataset. The next step in data preprocessing is to handle missing data in the dataset. The original dataset contains some missing data that could pose a significant problem for the research's machine learning model. Therefore, it is necessary to handle missing values in the dataset. There are two main ways to handle missing data:

By deleting specific rows: The first method is commonly used to handle null values. With this method, only specific rows or columns containing null values are deleted, but it is not very efficient, and deleting data may lead to loss of information that will not yield accurate results.

By calculating the mean or median value: With this method, the research calculates the mean or median value of the column or row containing any missing values and places that value in place of the missing value. This strategy is useful for quantitative data such as age, salary, year. For qualitative data types, the most frequent occurrence value can be used.

Editing and deleting some columns and data rows: This step may or may not be necessary in the preprocessing process, however, due to the nature of transactional data having nearly 1 billion rows of data, deleting redundant columns and repeated data rows will help streamline the data and the data running process will be faster. On the other hand, the data column with appointment_time (time of customer medical visits) and client_session (time of customer compensation claim or guarantee request submitted to the company) is the reason for the most rows. Therefore, the study considers simplifying the dataset and not incorporating time factors into the model, so these 2 columns will be deleted. After preprocessing the data, the final raw dataset is saved on the Google Drive directory using the to_csv() function to proceed with splitting the dataset.

Data Processing. Dataset for customer clustering:

The groupby() function: This function groups some attribute columns in the table together and uses in conjunction with common calculation functions to statistically analyze for an analysis problem. In this study, three columns client_id, category_treatment, and health_facility are grouped together.

The count() function is used to count the number of non-empty values in a data column, specifically here, the disease code and compensation status columns. The mean() function is used to find the mean value of numeric data columns.

The max() function: From the total compensation and guarantee times data, the study selects for a customer, which service they use and at which healthcare facility the most. The purpose is to initially grasp the characteristic habits of each customer.

The groupby() function to know the total compensation amount, hospital fee guarantee, and the total number of records submitted by each customer. Then, the study calculates ACC based on the formula:

$$Average\ Value = \frac{Total\ Claims\ and\ Direct\ Billing\ Amount}{Number\ of\ Claims\ and\ Direct\ Billing\ Records} \quad (16)$$

The purpose of calculating the total number of medical visits and ACC for customers is because these will be the two main factors for clustering customers with K-Means. The total number of medical visits is related to the severity of illnesses and is necessary to determine compensation or guarantee rights. Additionally, the ACC factor represents the efficiency and quality of claims and direct billing services. If the ACC is too high compared to the market average, it may indicate expensive or ineffective compensation claims. After calculating these indices, the study combines the factors of total medical visits, ACC, the type of service used, and the healthcare facility most visited by customers to generate the final dataset. For the dataset used to predict customer churn, the raw dataset from the data preprocessing step, containing the history of medical visits, the type of service used, and the healthcare facility most visited, is combined with the clustering results of customers and their ACC. The purpose of predicting from this dataset is to consider whether customers in a specific ACC segment continue their participation after using the service. Customers are expected to be segmented into four customer tiers based on the value they bring to the insurance company:

Diamond tier customers

Gold tier customers

Silver tier customers

Bronze tier customers

In terms of value brought by customers, from lowest to highest, the tiers are as follows: Bronze < Silver < Gold < Diamond. The customer clustering data is updated daily to best reflect the customer tiers. Additionally, the clustering data is stored in the customer profile data.

4 Results and Discussions

The result of data collection and preprocessing is a dataset containing 1,082,337 customers, among which 54,356 customers continue to use the claims and direct billing services.

4.1 Clustering Results of Customers with K-Means

In customer clustering, K-Means is widely used in practice and is one of the fundamental algorithms for solving customer segmentation problems in many businesses. In this study, to cluster customers using K-Means, this research utilizes two fields of data: the total number of compensation requests and the average value of a compensation and hospital fee guarantee record.

Table 2. Input data for customer segmentation

No.	Attribute	Explanation
1	client_id	Client ID
2	total_treatment	Total number of medical treatment visits
3	client_acc	Average value of a claims or direct billing records

With K-Means, this research chose the Elbow method as the approach to determine the value of K. The result of the Elbow method suggests selecting K = 4, indicating that there will be 4 clusters of customers assigned labels.

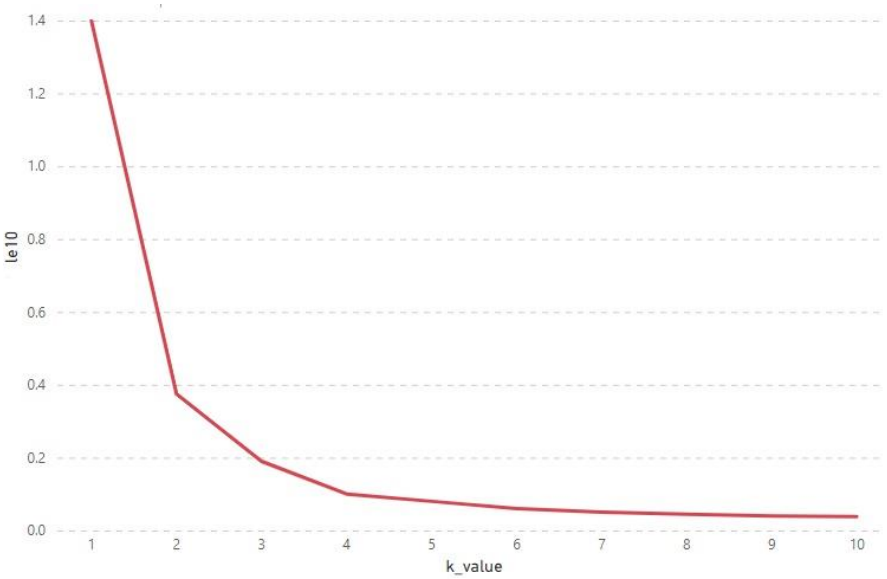

Fig. 2. Visualization of cluster values using the Elbow method

With a cluster number of K = 4, customers are allocated based on the average value of claims and direct billing records, while we do not observe significant differences from the number of compensation claims submitted by customers to the business.

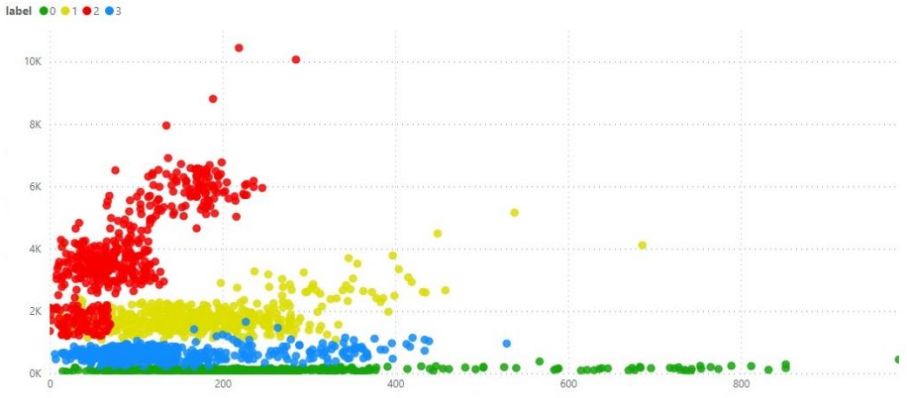

Fig. 3. Visualization of the separated data clusters

With this model, clustering results for 1,082,883 customers show a clear distribution among the clusters. The majority of customers (96%) are in the group with ACC ranging from 0 to 122.02 dollars (cluster 0). For more details, refer to Table 3.

Table 3. Summary of customer cluster analysis results

Cluster number	Number of customers	Customer proportion	Average value of largest records	Average value of smallest records
0	1,039,471	0.96	122.02	0
1	9,284	0.008	1,149.97	528.22
2	2,541	0.002	2,574.04	1,150.47
3	31,587	0.029	527.98	121.24

The cluster number K=4 is chosen because the diferrence between K=4 and K=5 compared to K=2 and K=3 is the lowest; between K=3 and K=4 is also minimal. Additional, after K=4, the elbow curve begins to flatten more visibly to the naked eye. Additionally, the four clusters correspond well to the research's initial segmentation hypothesis with four customer segments: Diamond, Gold, Silver, and Bronze. With the clustered results shown in Fig. 3, this research assigns the cluster numbers with the corresponding segment labels as follows:

Table 4. Labeling for customer clusters

Cluster number	Cluster color	Cluster label	Number of customers
0	Green	Bronze	1,039,471
1	Yellow	Gold	9,284
2	Red	Diamond	2,541
3	Blue	Silver	31,587

Considering the customer hierarchy from low to high as Bronze < Silver < Gold < Diamond, we observe that over 90% of the company's customers belong to the Bronze tier with high service usage frequency and low Average Claims Cost (ACC). Therefore, the insurance company has ample opportunities to develop this customer

segment by gradually elevating their standards. In addition to running online advertising campaigns, the business can create website content or design special health insurance packages with attractive benefits and higher coverage limits to attract and retain a better customer base. For the remaining 3 customer tiers, although their ACC is high, their frequency of using medical services is not substantial. The company can deploy customized strategies for each Silver, Gold, and Diamond tier. The key is to focus on delivering high value, fostering interactive environments, and building long-term relationships to enhance customer satisfaction and experience in the realm of claims reimbursement and medical fee guarantees in health insurance.

4.2 Prediction Results Using Algorithms and Model Evaluation

Model Prediction Results. This study selects 4 classification models: Logistic Regression, Random Forest, XGBoost, and k-Nearest Neighbour, to classify customers into "continue using" or "churn" categories and predict whether each customer who has used the service will continue to participate or not. The confusion matrix results for the 4 classification models are presented in the table.

Table 5. Representation of prediction results using a confusion matrix

Model	TP	TN	FP	FN
Logistic Regression	1,307,590	10,387	31,035	4,764
Random Forest	1,313,575	4,402	5,763	30,036
XGBoost	1,312,606	5,371	15,239	20,560
k-Nearest Neighbour	1,310,489	8,034	7,844	27,409

Model Evaluation Metrics. The values of evaluation metrics Accuracy, Precision, Recall, F1-score, and Support are shown in Table 6.

From the comparison results among the machine learning prediction models, it can be seen that the models yield similar Accuracy values, Combined with Precision, Recall, and F1-score metrics, this research suggests using the Random Forest model as the main prediction algorithm for predicting customer churn, Quick prediction enables the

business to devise prompt strategies to retain customers immediately and encourage them to make renewal decisions faster.

Table 6. Evaluation results with metrics Accuracy, Precision, Recall, F1-score and Support

Model	Accuracy	Precision	Recall	F1-score	Support
Logistic Regression	0.9653	0.9754	0.9862	0.9733	1,317,977
Random Forest	0.9822	0.9983	0.9974	0.9960	1,317,977
XGBoost	0.9768	0.9851	0.9901	0.9883	1,317,977
k-Nearest Neighbour	0.9887	0.9801	0.9235	0.8019	1,318,523

For the first group - classified as "Yes", customers who are likely to continue using the service:

Customer Care: Ensure the company provides excellent customer service, promptly addressing customer inquiries, concerns, or complaints, This helps build trust and increase customer satisfaction,

Enhance Customer Experience: Upgrade processes and interfaces to make the customer experience convenient and seamless, This may include providing online services, mobile apps, or communication through online channels,

Product Diversification: Offer a range of diverse insurance products and packages to meet customers' needs and preferences, providing flexible and suitable options helps customers feel satisfied and gives them more choices,

For the second group - classified as "No", customers likely to churn:

Improve Service Quality: Ensure that the claims and reimbursement processes are carried out quickly, fairly, and efficiently, Address customer complaints promptly and professionally to build trust and enhance customer confidence,

Provide Value-Added Services: Offer additional services and related utilities to increase value for customers, For example, providing disease prevention programs, free health consultations, promotional programs, or special discounts for loyal customers,

Communication and Customer Outreach: Regularly communicate and engage with customers after they have used the service to monitor their satisfaction and address any issues that arise, Send notifications and updates about products, programs, or other relevant information to maintain engagement and motivate customers to stay.

5 Conclusion

The main contribution of this paper is the proposal of multiple calculation steps to achieve the best prediction results with the current dataset of clients using claims and direct billing services in the health insurance sector, Using machine learning will help optimize health insurance policies, By analyzing claim history, information about diseases, and treatment costs, the machine learning model can suggest policy adjustments such as liability limits, premium adjustments, or offering customized insurance packages suitable for each client, thereby significantly reducing personnel costs and optimizing business operations, In the long run, machine learning will be an effective tool for widespread application in business service analysis, However, the health insurance sector is very diverse and complex, including information about diseases, medical history, medical costs, medications, and many other factors, Collecting, analyzing, and understanding these factors requires specialized knowledge and high-level data analysis skills, Given these limitations, the future direction of the research is to continue proposing additional Deep Learning models to detect and classify fraudulent behaviors in the health insurance sector, such as dredging medical records, registering incorrect information, or misuse, Additionally, the research could also be used for analyzing and evaluating risk factors, thereby proposing effective risk control and management measures.

References

1. Bach, M, P,, Pivar, J,, & Jaković, B, (2021), Churn Management in Telecommunications: Hybrid Approach Using Cluster Analysis and Decision Trees, Journal of Risk and Financial Management, 14(11), 544, doi: 10,3390/jrfm14110544
2. Bardicchia, M, (2020), Digital CRM-Strategies and Emerging Trends: Building Customer Relationship in the Digital Era, United Kingdom: Independently published

3. Buckinx, W,, & Poel, V, D, P, (2005), Customer base analysis: Partial defection of behaviourally loyal clients in a non-contractual FMCG retail setting, European Journal of Operational Research, 164, 252-268, doi:10,1016/j,ejor,2003,12,010
4. Cover, T, & Hart, P, (1967) Nearest Neighbor Pattern Classification, IEEE Transactions on Information Theory, 13, 21-27, doi: 10,1109/TIT,1967,1053964
5. Dick, A, S,, & Basu, K, (1994), Customer Loyalty: Toward an Integrated Conceptual Framework, Journal of the Academy of Marketing Science, 22, 99 – 113
6. Tamaddoni, A,, Sepehri, M,, M,, Teimourpour, B,, & Choobdar, S,, (2010), Modeling Customer Churn in a Non-Contractual Setting: The Case of Telecommunications Service Providers, Journal of Strategic Marketing, 18(7), 587 - 598, doi: 10,1080/0965254X,2010,529158
7. Ji, H,, Ni, F,, & Liu, J, (2021), Prediction of telecom customer churn based on XGB-BFS feature selection algorithm, Frontiers in Computing and Intelligent Systems, 17(2), 458 - 475, doi: 10,3390/jtaer17020024
8. Khadka, N, (2019), General machine learning practices using Python, Boston, MA: Pearson Education
9. Mozer, M, C, (2000), Predicting Subscriber Dissatisfaction and Improving Retention in the Wireless Telecommunications Industry, Piscataway, New Jersey, United States: IEEE
10. Neslin, S, A,, Gupta, S,, Kamakura, W, A,, & Lu, J, (2006), Defection detection: Measuring and understanding the predictive accuracy of customer churn models, Journal of Marketing Research, 43(2), 204 - 211, doi: 10,1509/jmkr,43,2,204
11. Oliveira, V, L, M, (2012), Analytical Customer Relationship Management in Retailing Supported by Data Mining Techniques, Ph, D, Thesis, Universidade do Porto, Porto, Portugal
12. Renjith, S, (2015), An integrated framework to recommend personalized retention actions to control B2C E-commerce customer churn, International Journal of Engineering Trends and Technology, 27(3), 152 - 157, doi: 10,14445/22315381/IJETT-V27P227