INTI JOURNAL | Vol.2023:65
eISSN:2600-7320

Forecasting Member Churn in Medical Insurance through Machine Learning Analysis

Chee Wen Jet[1], Goh Ching Pang[1*]

[1] Tunku Abdul Rahman University of Management and Technology, Kuala Lumpur, Malaysia

[*]**Email**: gohcp@tarc.edu.my

Abstract

The insurance industry faces an escalating challenge with increasing customer churn, spurred by global advancements in technology. The ease with which customers can compare policies, explore new offers, and switch providers online has intensified industry competition. This phenomenon has led to substantial revenue loss for many companies, as acquiring new customers often incurs higher costs than retaining existing ones. Recognizing the paramount importance of client retention, this research addresses the issue by proposing a Churn Prediction System tailored for the medical insurance sector. The system leverages machine learning models to forecast whether an existing customer is likely to churn, crucial for proactive retention strategies. To determine the most effective algorithm for this task, four models—Logistic Regression, Random Forest Decision Tree, Support Vector Machine, and Artificial Neural Network—are tested. The Random Forest Classifier emerges as the optimal performer which achieves accuracy of 90%.

Keywords

Logistic Regression, Random Forest Decision Tree, Support Vector Machine, Artificial Neural Network, Churn Analysis

Introduction

Health constitutes the cornerstone of a fulfilling life, yet Malaysians face challenges in maintaining optimal well-being. Alarming statistics from Malaysia's health minister reveal that one in two Malaysians grapples with obesity and overweight issues, contributing to persistently high rates of heart disease and obesity (BERNAMA, 2021). The silver lining emerges as an increasing number of Malaysians recognize the significance of medical insurance, spurred by the profound impact of the Covid-19 pandemic on health (Raynaud et al., 2021). Notably, a study by Balqis-Ali (2021) indicates a rise in medical insurance adoption, with 43.4% of Malaysians now covered—a noteworthy increase from previous years.

Medical insurance, elucidated by Abdul Rahman (2010), serves as a financial safeguard, covering medical and surgical costs for policyholders who pay monthly premiums to insurance providers. This protective measure ensures financial resilience during emergencies, alleviating concerns about escalating medical expenses. Simultaneously, insurance companies leverage

Submission: 14 November 2023; **Acceptance:** 28 November 2023

premiums to fuel investments, as outlined by Sheth and Pawar (2021), forming a symbiotic relationship with policyholders. However, the harmony in this mutualistic relationship faces disruption due to customer churn—an issue outlined by Begum et al (2023), where customers cease using a company's products or services. Customer churn is a critical concern, given that acquiring new customers costs significantly more than retaining existing ones, as demonstrated by Claudine and Tobin (2023). Moreover, a mere 5% increase in customer retention can increase a company's profits by at least 25%. To address the challenge of customer churn, understanding its underlying causes is paramount. Various factors, such as dissatisfaction with pricing, mismatched product offerings, or competition, contribute to churn (Siddika et al., 2021). Traditionally, manual churn analysis methods were employed, but their complexity and inefficiency have rendered them obsolete. In response, researchers have turned to machine learning models, recognizing their suitability for predicting customer churn. However, there are limited studies on using various machine learning techniques to study the churn in medical insurance.

Methodology

This research makes use of the dataset obtained from kaggle.com. The dataset consists of 4 separate excel files where each of them represent different fields of information and are related to each other which are Client_data.csv, Payment_history.csv, Policy_data.csv and train.csv. Each of them is built up from a different number of rows and variables. The definition of variables in the dataset are shown in Figure 1(a) to (d). Client_data.csv (Figure 1(a)) contains some personal information on the principal member. Payment_history.csv (Figure 1(b)) contains partial payment history up to the end of 2018, tied to Policy ID. Policy_data.csv (Figure 1(c)) describes the policies themselves. There may be multiple rows for each Policy ID since policies can cover more than one person. train.csv contains a list of all the policies. Policies that lapsed in 2017, 2018 or 2019 are identified with a 1 in the 'Lapse' column, and the year is provided. All other columns have a '?'

Variable	Definition
Policy_ID	Policy ID for the main member
NPH_LASTNAME	Last name of principle member
NPH_SEX	Gender of principle member
NPH_BIRTHDATE	Birth year of principle member
NAD_ADDRESS1	Main address of principle member
NAD_ADDRESS2	Second address of principle member

(a)

Variable	Definition
Policy_ID	Policy ID for the main member
AMOUNTPAID	Amount paid for policy
DATEPAID	Date of payment made
PREMIUMDATE	The next date on which the premium should be paid by to keep the policy alive

(b)

Variable	Definition
Policy_ID	Policy ID for the main member
NPI_EFFECTDATE	Date started
PPR_PRODCD	Specifies the product code of product subscribed by the client
NPR_PREMIUM	Extra premium
NPH_LASTNAME	Last name of member
CLF_LIFECD	Differentiates relationship to principal. 1 principal, 2 spouse, 3 child, 4 parent, 5 parent, 6 extended
NSP_SUBPROPOSAL	Unique identifier of a life on a policy on a specific policy we can have the principal as 111, child1 as 222, child2 as 333
NPR_SUMASSURE	Amount that is paid to the nominee of the plan in the unfortunate event of the policyholder's demise.
NLO_TYPE	Identifies the premium types charged on a policy
NLO_AMOUNT	Amount if there's an extra charge
AAG_AGCODE	Agent code of the policy agent
PCL_LOCATCODE	Branch code
OCCUPATION	Job
CATEGORY	Area of work

Variable	Definition
Policy_ID	Policy ID for the main member
Lapse	Is the policy churned. (1 - churned, ? - not churned)
Lapse Year	The year where the policy churned. (? - not churned)
Customer Info Exist	Other information about clients (1- other information about client, 0- client has no other information)
Has Payment History	Client has payment history (1- client has payment history, 0- client has no payment history)
Policy Info Exist	Policy information exist (1- other information about client policies, 0- client has no other information on his/her policies)

(c) (d)

Figure 1. Dataset

Prior to machine learning model input, the dataset undergoes crucial pre-processing steps. Data from four Excel files are merged to create a comprehensive dataset, with payment_history.csv uniquely processed to extract due counts. Incomplete or improperly format data is handled by either removing rows or strategically filling missing attributes. Addressing imbalanced rows, the under-sampling technique is employed to reduce the non-churning group, enhancing model performance. The dataset is then split into a 7:3 ratio of training and test sets for model learning and validation, aligning with standard community practices. To construct an effective prediction system, it is imperative to employ a high-performance machine learning model for accurate predictions. Accordingly, four distinct algorithms will be used to train the dataset to determine the optimal model for our specific use case. The selected algorithms, along with their respective parameters, are outlined in Table 1.

Table 1. Algorithms and parameters

Algorithm	Parameters
Artificial neural network *from Tensorflow Keras*	hidden layers = 2 layers (units = 8, 6 ; dropout rate = 0.2, 0.1), output unit = 1, batch_size = 128, epochs = 50
Random Forest Classifier *from Sklearn.ensemble*	Criterion = entropy, min_samples_leaf = 3, min_samples_split = 5, n_estimators = 300, random_state = 50
Logistic Regression *from Sklearn.linear_model*	Solver = liblinear, class_weight = balanced, random_state = 50
Support Vector Machine *from sklearn.svm.svc*	Kernel = rbf, class_weight = balanced

Results and Discussion

Table 2. Summary of performance of all models

Model	Positive Class F1-score	Negative Class F1-score	Accuracy
Artificial Neural Network	47%	0.87	81%
Random Forest Classifier	70%	0.94	90%
Logistic Regression Model	47%	0.82	73%
Support Vector Machine	38%	0.90	83%

The results obtained reveal a notable trend across all models, indicating higher F1-scores for the negative class. This discrepancy can be attributed to the dataset's inherent imbalance,

lacking sufficient positive cases for robust model learning. As described in Table 2, the Random Forest Classifier emerges as the top performer, boasting a remarkable 90% accuracy and a positive class F1-score of 70%. Given the research focus on predicting customer churn, the positive class F1-score assumes paramount significance, ensuring the effective identification of potential churn instances. This strategic approach empowers users to proactively address customer retention, minimizing the risk of customer loss. In contrast, the artificial neural network, logistic regression, and support vector machine exhibit lower positive class F1-scores, ranging from 38% to 47%, alongside accuracy values spanning 73% to 83%.

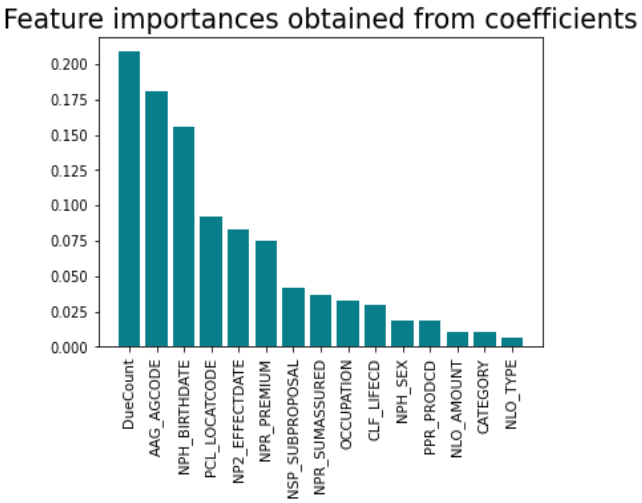

Figure 2. Feature importance from Random Forest Classifier

Upon closer examination of the Random Forest Classifier, Figure 2 illustrates the feature importance, revealing valuable insights. Notably, the 'DueCount' feature emerges as the most influential (0.215), underscoring its pivotal role in the model's outstanding performance. The introduction of this manually crafted feature proves to be a successful augmentation, significantly enhancing the model's predictive capabilities. Additionally, the second highest-ranking feature, 'AAG_AGCODE' (insurance agent code), exhibits a noteworthy correlation with churn cases (0.177). This suggests a potential impact of insurance agents on customer churn, where factors such as agent behavior and professionalism may influence customer decisions to discontinue services. For instance, occurrences of rudeness or impatience displayed by an insurance agent after a contract has been signed can cause customers to perceive a lack of respect, which in turn may increase the likelihood of them discontinuing their relationship with the company.

Conclusion

In conclusion, this research addresses the escalating challenge of customer churn in the medical insurance sector, leveraging machine learning models. Four machine learning models have been used to study the churn in medical insurance. The study reveals a substantial increase in Malaysians embracing medical insurance, spurred by the profound impact of the Covid-19 pandemic. Among the algorithms tested, the Random Forest Classifier emerges as the optimal performer, achieving a notable accuracy of 90% and a crucial positive class F1-score of 70%. The feature analysis

underscores the significance of the 'DueCount' variable and highlights the potential impact of the 'AAG_AGCODE' (insurance agent code) on churn instances. These findings provide valuable insights for proactive customer retention strategies in the medical insurance industry.

Acknowledgement

This study is supported by Tunku Abdul Rahman University of Management and Technology.

References

Abdul Rahman, Z. (2010). Adverse selection and its consequences on medical and health insurance and takaful in Malaysia. *Humanomics: The International Journal of Systems and Ethics, 26*(4), 264–283. https://doi.org/10.1108/08288661011090875

Balqis-Ali, N., Jailani, A.-S., Fun, W. H., & Sararaks, S. (2021). Private health insurance in Malaysia: Who is left behind? *Asia-Pacific Journal of Public Health, 33*(1), 10105395211000912. https://doi.org/10.1177/10105395211000913

Begum, T., Pravallika, E., Rushitha, M., & Kavya, K. (2023). Customer churn analysis. *International Journal for Research in Applied Science and Engineering Technology, 11*(3), 1464–1468. https://doi.org/10.22214/ijraset.2023.55720

BERNAMA. (2021, November 11). Malaysia, unhealthy nation with low health awareness – Khairy. *The Malaysian Reserve.* https://themalaysianreserve.com/2021/11/11/malaysia-unhealthy-nation-with-low-health-awareness-khairy/

Howard-James, C., & Tobin, A. (2023). An analysis of direct and indirect costs in hidradenitis suppurativa. *Skin Health and Disease*, *3*(1), e306. https://doi.org/10.1002/ski2.306

Raynaud, M., Goutaudier, V., Louis, K., Al-Awadhi, S., Dubourg, Q., Truchot, A., Brousse, R., Saleh, N., Giarraputo, A., Debiais, C., Demir, Z., Certain, A., Tacafred, F., Cortes-Garcia, E., Yanes, S., Dagobert, J., Naser, S., Robin, B., Bailly, E., & Loupy, A. (2021). Impact of the COVID-19 pandemic on publication dynamics and non-COVID-19 research production. *BMC Medical Research Methodology, 21,* Article 255. https://doi.org/10.1186/s12874-021-01404-9

Sheth, R., & Pawar, S. (2021). Life and health insurance – Tools to safeguard your investment. In *Pandemic-induced creative disruptions: Issues, challenges, and prospects* (Conference presentation).

Siddika, A., Faruque, A., & Masum, A. K. M. (2021). Comparative analysis of churn predictive models and factor identification in telecom industry. In *2021 24th International Conference on Computer and Information Technology (ICCIT)* (pp. 1–6). IEEE. https://doi.org/10.1109/ICCIT54785.2021.9689881