Predicting the Probability of Critical Illness Claims

Kelan Laban
Department of Computing and Information Technology
The University of the West Indies
St. Augustine, Trinidad
Email: kelan.laban@my.uwi.edu

Patrick Hosein
Department of Electrical and Computer Engineering
The University of the West Indies
St. Augustine, Trinidad
Email: patrick.hosein@uwi.edu

***Abstract*—The World Health Organisation reported that 74% of global deaths in 2019 were attributed to Noncommunicable Diseases (NCDs). Preventive care best improves NCD outcomes, while insurance provides financial protection in the Critical Illness (CI) coverage. The two may be combined through a data science approach. Thus, a method was proposed for identifying clients at risk of NCDs using their probability of a CI claim. Five Machine Learning algorithms were tested: Logistic Regression with Elastic Net (LREN), Support Vector Machine (SVM), Extreme Gradient Boosting (XGBoost), Random Forests (RF), and Artificial Neural Network (ANN). Experiments tested the predictive power of health claim data from health insurance coverages and prediction time windows of 0, 90, 180 and 360 days. Health claim data provided better predictors of CI claims, improving log loss skill scores from 0.074 - 0.098 to scores of 0.13 - 0.39. Increasing prediction time windows offered more effective interventions, but at a performance cost. Insurance providers can use the approach to assist customers with preventive NCD Management.**

***Index Terms*—Critical Illness Insurance, Noncommunicable Disease, Data Science, Claim Probability, Customer Ranking**

I. Introduction

Sources such as Our World in Data [1] and the World Health Organization [2] reported that 74% of global deaths in 2019 were attributed to Noncommunicable Diseases (NCDs). Similarly, in Trinidad and Tobago (TTO), NCDs accounted for 79.3% of deaths in 2000 and 82.7% in 2019, showing that the burden of NCDs has not improved in decades. NCDs are characterised by being noninfectious and having prolonged courses with difficult resolutions [3]. Injuries with prolonged recuperation may also qualify [3]. Reports suggest that NCDs are difficult to control due to complex and compounding risk factors such as smoking, air pollution, and indicators of poor health [1]. Identifying crucial risk factors is of particular interest.

Many insurance companies offer financial protection against NCDs through plans often referred to as Critical Illness (CI). These plans pay a lump sum upon a CI diagnosis for medical treatment. What qualifies as a CI varies, but the commonly insured diagnoses are mostly NCDs or their complications, many of which are controllable when detected early. Early detection, screening and medical intervention are well understood to improve NCD outcomes [4], but the current insurance model is quite reactive. The insurance industry can adopt a role in NCD management by covering the costs of preventive care needed by high-risk clients as a feature of CI insurance. This is significantly cheaper than the payout on CI diagnosis and can retain clients as premium-paying customers while providing investment to the wider NCD management in healthcare.

Early detection is key in NCD management, but a prior level of action can be achieved through a data science approach. This "predictive stage" can frequently evaluate health status without direct actions such as doctor visits. For example, a person may be considered at high risk for an NCD based on medical histories and recent drug purchases. Then, they may be prompted to enter the detection stage by receiving a formal checkup, where they may be diagnosed. The importance of different inputs in the predictive stage varies and can be measured, known as the feature importance. This can reveal underlying information, such as which risk factors should be monitored.

The goal of this research was twofold. A model for predicting the probability of a CI claim is developed. This probability is then used to rank clients. Considering CIs and NCDs equal, a client with a high probability of CI is at a high probability of NCD diagnosis. As such, the model is designed as a predictive model for NCDs through the use of its insights. The paper is structured as follows. Firstly, a review of related work in this area shall be presented. Then, the specific details of the dataset, processes and predictive model are given. Finally, experimental evaluations and observations are discussed.

II. Literature Review

NCD prediction has generally been performed with data sourced from national health systems [5]–[7], but studies [8]–[10] have also used insurance claim data. However, these studies have provided little benefit to the insurance sector. In modelling, variables such as sociodemographic information, medical conditions, medications, procedures, costs, and health screening results were typically used. Standard preprocessing and feature engineering techniques were seen. Statistical tests and multicollinearity analysis were repeated for feature selection. Imbalanced data was addressed through class weighting in the model objective [9] and data undersampling [10]. Models tested included ensembles of Random Forests (RF) and XGBoost algorithms [8], XGBoost, Neural Network (NN), Logistic Regression (LR), RF, and CatBoost [9], [10]. Grid search and Cross-Validation (CV) optimised metrics such as out-of-bag errors and ROCAUC [8], [9], while Platt calibration was used to tune probability estimates [10]. Sensitivity

analyses involved varying sensitivity and specificity, training set sizes, and years of test data [10]. Finally, risk factors were assessed via CoxPH ratios [8], XGBoost feature importance and statistical tests.

Despite the popularity of CI insurance, Machine Learning (ML) approaches were seldom seen in its management. Incidence rates and annual changes for common NCDs were modelled with joinpoint trend analysis validated by the Monte Carlo permutation method [11]. The CoxPH, cause-specific, and subdistribution hazard models were tested [12], using hazard ratios to establish links between demographic variables and CI events. A continuous multistate Markov Chain was presented [13] with healthy, diagnosed and dead states. The transition intensities and probabilities between states were then used to determine an appropriate premium for the insurance.

Claim probabilities [14], rates [15], and counts [16] can be predicted for any insurance type. The studies here followed standard preprocessing and feature engineering procedures. A novel "Kappa algorithm" was proposed [15] with feature selection via model error minimisation and CV optimising the MAE. A Support Vector Machine (SVM) model tuned with Particle Swarm Optimisation was tested in [14], while [16] tested LR and XGBoost models. Feature importance was assessed with the Kappa algorithm [15], while [16] used LR coefficients and the Gini impurity metric.

Studies also evaluated high-cost clients via ranking [17], [18]. Undersampling and ROSE methods [17] were observed for imbalanced data. LR, RF, and XGBoost models were tested [17], as well as MLR, Boosting Machine (GBM), and NN models [18]. XGBoost information gain and the RF mean decrease accuracy scores [17], and MLR coefficients and information gain in the GBM [18] were used to assess feature importance. The prediction window was varied between 1-12 months, with increased performance with time [18]. Both studies showed that the top percentiles of patients were the most important to manage and used rank values to calculate performance metrics.

III. Methodology

In Trinidad and Tobago, implementing large-scale ML applications in NCD management is difficult due to poor national health data [19]. In the literature, insurance data has been a successful alternative, but NCD models typically used health claim data over CI. Studies focused on the burdened health sector and have not considered integrating support from resourceful industries such as insurance. As such, this section presents a predictive model which ranks NCD risk in insurance clients, novelly using both CI and health claim data from a Trinidadian insurance company. These predictions may be used for early interventions with mutual benefit for the insurance and healthcare industries.

A. Dataset Description

A private insurance company provided the data, with approval granted through a Non-Disclosure Agreement (NDA) and access given to a database schema. Personal information was removed to protect customer privacy. Data from three areas of business were provided: CI coverage, client, and health claims. Health claims were used as an active medical history of a client. The dataset was prepared using SQL and was loaded into the development environment using Python libraries.

B. Data Preprocessing

Data partitioning was the first step to avoid "data leakage" from test sets into training sets. Partitions were based on date fields as businesses commonly desire periodic analyses (yearly, quarterly). Test sets contained CI coverages ending after 2023 and 2024. All previous data were used for training and validation. Validation sets were created through ScikitLearn's (SKL) TimeSeriesSplit CV [20]. Multiple coverages with the same contractual dates for the same client were aggregated. Duplicate clients were identified using a provided CRM table and similarly aggregated. Duplicate health claims were not considered since procedures could be legitimately repeated. The "Sum Insured", "Income", and "Years Employed" fields were adjusted by Winsorisation (replacement with the closest inlier) using information such as labour laws in TTO to determine the inlier ranges. All sparse variables in the base data were excluded. Min-max standardisation was implemented via the SKL MinMaxScaler class [20] to ensure that features varied on comparable scales.

Transformation, aggregation, categorisation, and encoding procedures were utilised. Date format variables were transformed into integers as the difference in days or years between them. Valid health claims were aggregated to create health indicator variables like claim counts, procedure and diagnosis counts, and total health charges. Claim procedure and diagnosis codes, representing thousands of medical events, were grouped. Procedures were categorised into: "DENTAL", "DRUGS", "HOSPITAL", "SERVICE", "TEST", "VACCINATION", "VISION", "VISIT", "OTHER". Diagnoses were grouped into: "CANCER", "CHRONIC", "CARDIAC", "RESPIRATORY", "DENTAL", "VISION", "INFECTION", "COMMON", "INJURY", "PAIN", "ORGAN", "SEXUAL", "EXAM", "OTHER". Lastly, all categorical variables were one-hot encoded.

In ML algorithms, regularisation simplifies the objective function to be optimised. This reduces the number of features and provides robustness to multicollinearity. Explicit regularisation, such as L1/LASSO and L2/Ridge methods, and implicit regularisation, such as early stopping and randomisation (tree methods), were implemented [20]. When dealing with imbalanced data, probability estimation is more appropriate than classification or assigning class labels based on probability thresholds [21]. Thus, resampling methods like SMOTE are unnecessary in preprocessing. The course of action should depend on the probability of the outcome. For example, the severity of medical intervention should be based on the probability of CI. It is also beneficial to build models using "proper scoring rules" instead of threshold-based metrics like

precision, recall, and F1. As such, this study estimates CI claim probability and uses the Log Loss metric in evaluation.

C. Predictive Modelling

Let $p(t)$ denote the PDF of the probability that the customer makes a CI claim at time t. Let c denote the age of the customer given premature coverage closure (death, lapse or surrender if this happens) and let $\tau = \min\{c, 80\}$ (80 is the maximum age for this insurance). The probability of a claim is then given by:

$$L = \int_0^{\tau} p(t)dt \tag{1}$$

It is assumed that each sample contains client features and coverage information. A variable is created where, for sample i, the value is $x_i = 1$ if a claim was made and $x_i = 0$ otherwise. To illustrate the approach, N clients are assumed with all samples having the same features but different x_i values. The average experimental probability that a client makes a claim is given by:

$$\theta = \frac{1}{N}\sum_{i=1}^{N} x_i \tag{2}$$

Hence, we make the approximation:

$$L \approx \theta \tag{3}$$

The basic illustration is expanded by adding several features, and a classification algorithm predicts θ. Customers may then be prioritised by ranking with this measure. Classification algorithms considered include Logistic Regression with Elastic Net (LREN), Support Vector Machine (SVM), Extreme Gradient Boosting (XGBoost), Random Forest (RF), and Artificial Neural Network (ANN). Models were optimised through K-fold cross-validation, randomised grid search for hyperparameter tuning, and probability calibration using the Sigmoid/Platt Logistic Function.

D. Model Evaluation

For the classification, the Log Loss Score (LLS) [20], Log Loss Skill Score (LLSS) [22], and Area Under the Curve (AUC) [20] metrics were used. For the ranking, the top K listed elements were observed. Then, the ratio of the number of relevant elements (claimants) captured to K, the Precision-at-K (PAK) [23], and the ratio of the number of relevant items to the total number of relevant items, the Recall-at-K (RAK) [23], were assessed. Model coefficient values were used to demonstrate the advantages and disadvantages of feature importance analysis. Additionally, sensitivity analysis included the use of health claim data and various prediction windows. The "prediction intervals" evaluated how early CI claims could be predicted. For example, the 360-day interval represents predicting a year in advance, granting more time for preventative action.

IV. Models

The target variable of the probability prediction, "CLAIM" was observed in only 1.93% of coverages, with most claimants aged 38 to 60. Claimants also had longer tenures on their CI products and appeared to open policies at later ages. Claim rates were similar for both sexes. Smoking was unexpectedly less frequent among CI claimants, while marriage was more frequent. Income and years of employment were slightly higher for claimants. CI claimants made more health claims on average and had greater health expenses.

An SKL Pipeline [20] was implemented with Winsorisation, MinMaxScaler and the regularised ML algorithms as classifiers producing probability scores via "predict_proba" or "decision_function". To implement the client ranking, N probability estimates were obtained from the best classifier, depending on the dataset's size. The N predictions were averaged as the claim probability, and clients were ranked in descending order. Since calibration and grid search require separate CVs, nested CV was implemented to maximise data usage at the cost of execution speed, as seen in Fig. 1. 5-Fold CV was implemented through SKL's "TimeSeriesSplit" class in both layers. This gave a good balance between performance and execution speed. The Log Loss was optimised in the outer grid search using the probabilities from the inner calibration.

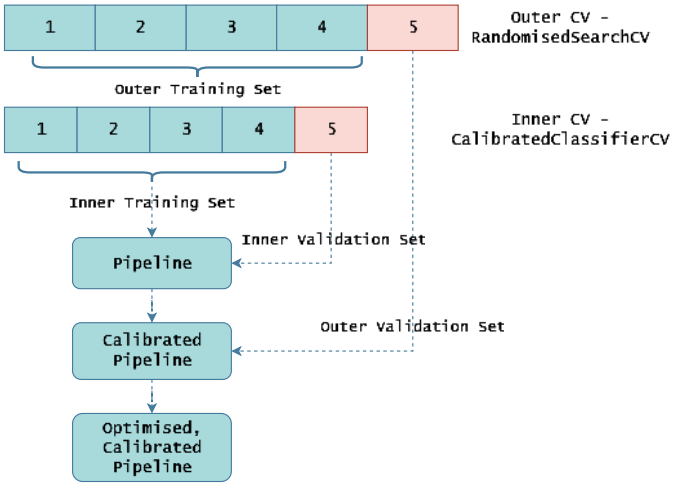

Fig. 1. Nested CV Implementation

There were two base test sets: coverages ending after 2023, with a training set up to "2022-12-31", and coverages ending after 2024, with a training set up to "2023-12-31". The experiments were repeated with and without health claim data and for time intervals: 0, 90, 180, and 360 days. Experiments were also repeated for numerical stability. To evaluate the "prediction interval", time-based features like "tenure", and "age" were reduced by 0, 90, 180 and 360 days accordingly. Health claims were excluded if made within the time interval. Using 360 as an example, all claims made in the last 360 days were excluded. This simulates early prediction by limiting the accessible data.

V. Results

The code and experiments of the proposed approach can be found at https://github.com/kelanRoss/Predicting-CI-Claims-RTSI-2025.git.

A. Without Health Claims (Demographic-Only)

With 110402 rows and 15 features, the demographic-only dataset was allotted 5 runs per test. LLSS was low with a best mean score of 0.098, but still above a dummy classifier. AUC was reasonable, with a best mean score of 0.79. These scores suggested that demographic variables were weak predictors of CI claims. The LREN model showed stable but average performance, while the SVM and ANN displayed high variance. The RF and XGBoost models gave the highest scores and performed stably. The RF slightly outperformed the XGBoost in most intervals. These algorithms showed steadily declining performance with increasing 0-360 intervals. The models' AUC scores followed identical patterns. Table I displays the model achieving the best average LLSS and ROCAUC scores for each experiment. This was most frequently the RF for this dataset.

TABLE I
Best Model Scores Without Health Claims

Year	Prediction Window	Clf	Mean LLSS	Mean AUC
2023	0 day	RF	0.098	0.79
2023	90 day	RF	0.093	0.79
2023	180 day	XGB	0.085	0.78
2023	360 day	RF	0.081	0.78
2024	0 day	RF	0.088	0.78
2024	90 day	RF	0.085	0.77
2024	180 day	RF	0.078	0.77
2024	360 day	RF	0.074	0.77

B. With Health Claims

With 1758 rows and 46 features, experiments were repeated 20 times for the health claim dataset. LLSS was 2 to 4 times higher than the demographic-only data, with a best mean score of 0.39. AUC was high, with a maximum mean of 0.91. These scores suggested that health claim variables were stronger predictors of CI claims. Scores were higher in the 2024 test, but both years showed similar trends in algorithmic performance. Here, the LREN outperformed all other models across all intervals. The LREN and RF performed stably, while the SVM and XGBoost showed high variance. The models showed declining performance as the prediction window increased from 0 to 360 days. Again, the trend of the AUC was similar. Table II displays the model achieving the best average LLSS and ROCAUC for the 20 runs. This was consistently the LREN for this dataset.

C. Ranking

Three strategies for the K value in the PAK and RAK metrics were used: $K = 10$, $K = 20$ and K equal to the CI claim rate in the training set times the size of the test set.

TABLE II
Best Model Scores With Health Claims

Year	Prediction Window	Clf	Mean LLSS	Mean AUC
2023	0 day	LREN	0.29	0.88
2023	90 day	LREN	0.22	0.86
2023	180 day	LREN	0.13	0.84
2023	360 day	LREN	0.14	0.84
2024	0 day	LREN	0.39	0.91
2024	90 day	LREN	0.29	0.86
2024	180 day	LREN	0.17	0.87
2024	360 day	LREN	0.2	0.88

1) Without Health Claims: Fig. 2 and Fig. 3 show the ranking results for this dataset. The RF model was used, and tests were repeated 5 times. Performance fluctuated with low to moderate PAK and RAK scores. Mean PAK scores ranged from 0.06 to 0.28 in the 2023 test and from 0.1 to 0.38 for 2024. Mean RAK scores ranged from 0.002 to 0.084 for 2023 and 0.017 to 0.09 for 2024. The CI rate gave the top 237 clients for 2023 and 104 clients for 2024. As expected, the PAK decreased with K while the RAK increased. Both metrics fluctuated as the interval increased, but were more stable at larger K values.

Fig. 2. Ranking Analysis for the 2023 Test Without Health Claims

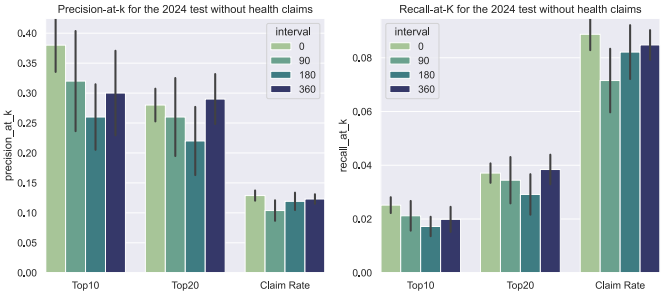

Fig. 3. Ranking Analysis for the 2024 Test Without Health Claims

2) With Health Claims: Fig. 4 and Fig. 5 display the PAK and RAK scores with health claims. The LREN model was used with 20 experimental runs. Here, the CI rate gave the top 20-24 clients in 2023 and 10 in 2024. Scores were significantly higher than those without health claims, especially for RAK, but similar trends were observed. Average PAK scores ranged from 0.2 to 0.66 for 2023 and 0.35 to 0.6 for 2024. Average RAK scores were 0.11 to 0.49 for 2023 and 0.32 to 0.8 for 2024. Again, PAK decreased with K, while RAK increased. Both scores appeared stable at the 0 and 90-day intervals but

decreased sharply at 180 days. Overall, including health claims resulted in more acceptable PAK and RAK scores.

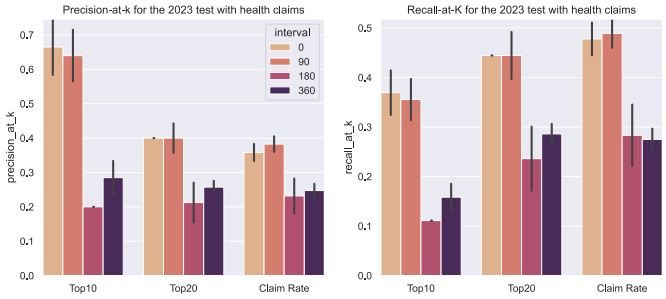

Fig. 4. Ranking Analysis for the 2023 Test With Health Claims

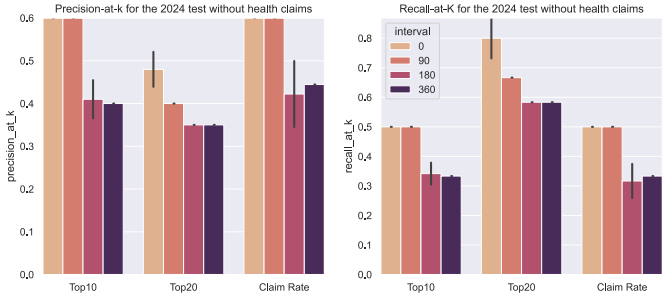

Fig. 5. Ranking Analysis for the 2024 Test With Health Claims

D. Comparison of Prediction Windows/Intervals

Fig. 6 and Fig. 7 show that classifier performance dropped as the prediction window increased from 0 to 360 days. This was expected, as predictors of a client's CI claim likely develop over time. Still, with reasonable performance across intervals, some clients can be detected early. This grants the insurance company and the client time for preventive care. There were greater drops in performance when health claims were included, likely because claims accumulated over time and were stronger predictors of CI claims. Similar observations were made with the ranking performance across intervals. PAK and RAK decreased with increasing window length and more sharply with health claims. However, ranking performance usually remained stable between the 0 and 90-day intervals, even though classifier performance decreased. This suggests that at-risk clients can be identified with 90 days' notice.

E. Feature Importance

The best performance was observed for the 2024 test using health claims for the 0-day interval and the LREN model. This experiment was repeated 50 times, and the feature coefficients of the LREN were averaged and plotted. The top 10 are shown in Fig. 8. Final feature sets and coefficients were observed to vary greatly between individual runs. Many features were set to zero, proving that regularisation was a reliable method for feature selection. Correlated feature pairs were handled correctly, as one variable in each pair was set to zero.

The analysis identified features like the number of hospital procedure claims and having a marital status of single as

Fig. 6. Comparison of Datasets and Models for the 2023 Experiments

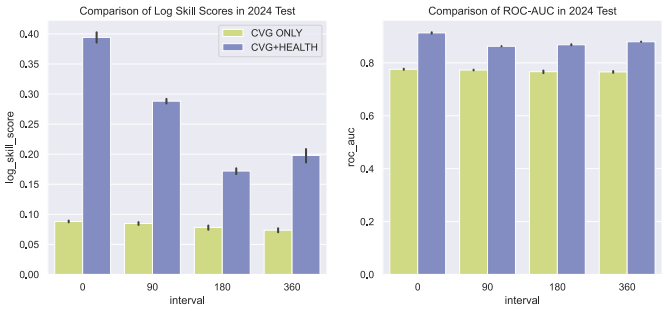

Fig. 7. Comparison of Datasets and Models for the 2024 Experiments

strong positive predictors of CI claims, while income, number of vision procedures, and female sex were strong negative predictors. While logic and medical evidence support some of these findings, non-smoking was also highlighted as a strong positive predictor. This illustrates the challenges of feature importance analysis. Clients accumulating hospital procedures on their health coverage can be closely monitored, but it would not be wise to advise clients to smoke to reduce their CI risk. Nonetheless, it was strange that non-smoking was highlighted, and this finding should signal the insurer to launch an investigation into whether smoking status is being filed correctly, or if clients were dishonest on their applications.

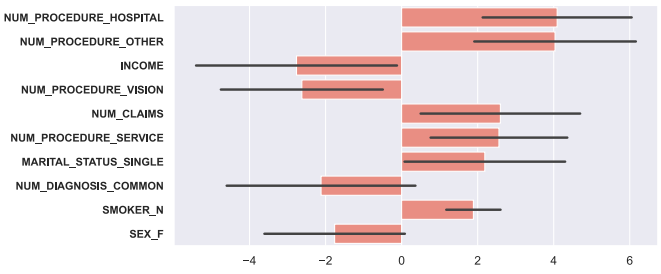

Fig. 8. Feature Importance Analysis

F. Limitations

This study developed an innovative method for monitoring NCDs using two types of insurance products: the unique critical illness coverage, and health claims as proxies for medical histories. This study also opted for probability prediction over classification labels, using the Log Loss instead

of common threshold metrics like F1, recall, precision, etc. This combination of choices made it difficult to find directly comparable studies, but we add to the works supporting such design decisions.

Unfortunately, the provided Customer Relationship Management system was incomplete, making it impossible to link health claims to all CI clients. This resulted in the small 1758 row dataset for the experiments. Still, the results support the hypothesis that health claims predict CI and the experiments can be repeated when more data are available. Models displaying high variance, like the SVM and ANN, perhaps needed more focused tuning and better understanding of hyperparameters. Different model optimisation strategies and objectives, like cost-sensitive learning, can be explored in further work.

VI. Conclusion

This paper presented a data science approach to incorporating NCD management in the administration of CI insurance. This involved estimating the probability of a CI claim and ranking clients. Experiments tested and confirmed the predictive power of health claims, but showed that more work is needed to attain early predictions. Without health claims, LLSS of 0.074 to 0.098 and AUC scores of 0.77 to 0.79 were observed. Performance was significantly improved with health claims, with LLSS of 0.13 to 0.39 and AUC of 0.84 to 0.91.

Insurers can use this model to monitor CI risk in clients and potentially cover preventive care costs, providing benefits for the insurance and healthcare industries and the general public. The top "K" ranked clients, where budgets determine the value of "K", can be focused on as these are the most valuable. Providing care to false-positive clients boosts customer satisfaction, while false-negative clients are still protected by the normal insurance function. Corporate responsibility is required to prevent customer profiling by denying risky clients CI products, or encouraging coverage closure or exchange. In the future, time-to-claim and lifetime value analyses can be performed to provide a complete analysis of CI insurance.

References

[1] S. Dattani, F. Spooner, H. Ritchie, and M. Roser, "Causes of death," *Our World in Data*, 2023. [Online]. Available: https://ourworldindata.org/causes-of-death

[2] World Health Organization. (2024) The top 10 causes of death. Fact sheet by WHO. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death

[3] C. for Disease Control and Prevention, *Overview of NCD's and Related Risk Factors*, Centers for Disease Control and Prevention, 2023. [Online]. Available: https://www.cdc.gov/globalhealth/healthprotection/fetp/training-modules/new-8/overview-ncds-fg-qa-review-091113.pdf

[4] M. Y. Bertram, K. Sweeny, J. A. Lauer, D. Chisholm, P. Sheehan, B. Rasmussen, S. R. Upreti, P. D. Lonim, K. George, and S. Deane, "Investing in non-communicable diseases: an estimation of the return on investment for prevention and treatment services," *The Lancet*, vol. 391, no. 10134, pp. 2071–2078, 05 2018. [Online]. Available: https://www.proquest.com/scholarly-journals/investing-non-communicable-diseases-estimation/docview/2040721945/se-2

[5] K. Davagdorj, J.-W. Bae, V.-H. Pham, N. Theera-Umpon, and K. H. Ryu, "Explainable artificial intelligence based framework for non-communicable diseases prediction," *IEEE Access*, vol. 9, pp. 123 672–123 688, 2021.

[6] K. Davagdorj, V. H. Pham, N. TheeraUmpon, and K. H. Ryu, "Xgboost-based framework for smoking-induced noncommunicable disease prediction," *International Journal of Environmental Research and Public Health*, vol. 17, no. 18, p. 6513, 2020. [Online]. Available: https://www.proquest.com/scholarly-journals/xgboost-based-framework-smoking-induced/docview/2441894015/se-2

[7] L. J. Muhammad, I. Al-Shourbaji, A. A. Haruna, I. A. Mohammed, A. Ahmad, and M. B. Jibrin, "Machine learning predictive models for coronary artery disease," *SN Computer Science*, vol. 2, no. 350, 06 2021.

[8] C. An, J. W. Choi, H. S. Lee, H. Lim, S. J. Ryu, J. H. Chang, and H. C. Oh, "Prediction of the risk of developing hepatocellular carcinoma in health screening examinees: A korean cohort study," *BMC Cancer*, vol. 21, no. 755, 06 2021.

[9] Z. Segal, D. Kalifa, K. Radinsky, B. Ehrenberg, G. Elad, G. Maor, M. Lewis, M. Tibi, L. Korn, and G. Koren, "Machine learning algorithm for early detection of end-stage renal disease," *BMC Nephrology*, vol. 21, no. 518, 11 2020.

[10] C.-Y. Hung, C.-H. Lin, T.-H. Lan, G.-S. Peng, and C.-C. Lee, "Development of an intelligent decision support system for ischemic stroke risk assessment in a population-based electronic health record database," *PLOS ONE*, vol. 14, no. 3, pp. 1–16, 03 2019. [Online]. Available: https://doi.org/10.1371/journal.pone.0213007

[11] Y. Yang, X. Sun, J. Wang, C. Yang, and L. Zhang, "Incidence rates of four major non-communicable chronic diseases in the chinese adult population from 2007 to 2016: A study based on a national commercial claims database," *Clinical Epidemiology*, vol. 12, pp. 215–222, 2020. [Online]. Available: https://www.proquest.com/scholarly-journals/incidence-rates-four-major-non-communicable/docview/2377791185/se-2

[12] D. Zapletal, "Application of the cox proportional hazards model and competing risks models to critical illness insurance data," *Statistical Analysis and Data Mining: The ASA Data Science Journal*, vol. 14, no. 4, pp. 342–351, 2021.

[13] U. S. Pasaribu, H. Husniah, R. K. N. Sari, and A. R. Yanti, "Pricing critical illness insurance premiums using multiple state continous markov chain model," *Journal of Physics: Conference Series*, vol. 1366, no. 1, p. 012112, 11 2019. [Online]. Available: https://dx.doi.org/10.1088/1742-6596/1366/1/012112

[14] S. Anam, M. R. A. Putra, Z. Fitriah, I. Yanti, N. Hidayat, and D. M. Mahanani, "Health claim insurance prediction using support vector machine with particle swarm optimization," *BAREKENG: Jurnal Ilmu Matematika dan Terapan*, vol. 17, no. 2, pp. 0797–0806, 2023.

[15] P. Hosein, "A data science approach to risk assessment for automobile insurance policies," *International Journal of Data Science and Analytics*, vol. 17, no. 1, p. 127–138, 03 2023.

[16] J. Pesantez-Narvaez and M. Alcañiz, "Predicting motor insurance claims using telematics data—xgboost versus logistic regression," *Risks*, vol. 7, no. 2, p. 70, 2019. [Online]. Available: https://www.proquest.com/scholarly-journals/predicting-motor-insurance-claims-using/docview/2550241103/se-2

[17] Y. Choi, J. An, S. Ryu, and J. Kim, "Development and evaluation of machine learning-based high-cost prediction model using health check-up data by the national health insurance service of korea," *International Journal of Environmental Research and Public Health*, vol. 19, no. 20, p. 13672, 2022. [Online]. Available: https://www.proquest.com/scholarly-journals/development-evaluation-machine-learning-based/docview/2728481305/se-2

[18] C. Yang, C. Delcher, E. Shenkman, and S. Ranka, "Machine learning approaches for predicting high cost high need patient expenditures in health care," *Biomedical engineering online*, vol. 17, pp. 1–20, 2018.

[19] Connect-To-Health, "Strategic plan for strengthening the national health information system, 2012-2016," Feb. 2012. [Online]. Available: http://www.health.gov.tt/downloads/DownloadItem.aspx?id=387

[20] Scikit-learn Developers. (2024) Scikit-learn documentation. Accessed: 2024-11-29. [Online]. Available: https://scikit-learn.org/1.5/index.html

[21] F. Harrell. (2023) Classification: The false dichotomy. Accessed: 2024-11-29. [Online]. Available: https://www.fharrell.com/post/classification/

[22] J. Brownlee. (2020) How to score probability predictions in python. Accessed: 2024-11-29. [Online]. Available: https://machinelearningmastery.com/how-to-score-probability-predictions-in-python/

[23] Evidently AI. (2024) Precision and recall at k: Ranking metrics explained. Blog post by Evidently AI. [Online]. Available: https://www.evidentlyai.com/ranking-metrics/precision-recall-at-k