Journal of Computer Information Systems

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/ucis20

Predicting Cross-Selling Health Insurance Products Using Machine-Learning Techniques

Khulekani Mavundla, Surendra Thakur, Emmanuel Adetiba & Abdultaofeek Abayomi

To cite this article: Khulekani Mavundla, Surendra Thakur, Emmanuel Adetiba & Abdultaofeek Abayomi (05 Sep 2024): Predicting Cross-Selling Health Insurance Products Using Machine-Learning Techniques, Journal of Computer Information Systems, DOI: 10.1080/08874417.2024.2395913

To link to this article: https://doi.org/10.1080/08874417.2024.2395913

Published online: 05 Sep 2024.

JOURNAL OF COMPUTER INFORMATION SYSTEMS
https://doi.org/10.1080/08874417.2024.2395913

Taylor & Francis
Taylor & Francis Group

Predicting Cross-Selling Health Insurance Products Using Machine-Learning Techniques

Khulekani Mavundla[a], Surendra Thakur[a], Emmanuel Adetiba[a,b], and Abdultaofeek Abayomi[c]

[a]Durban University of Technology, Durban, South Africa; [b]Covenant University, Ota, Nigeria; [c]HRA, Walter Sisulu University, East London, South Africa

ABSTRACT

This study delves into the utilization of Machine Learning (ML) techniques for predicting health insurance cross-selling behavior in South African consumers. The main goal is to create a robust ML model that assists health insurance companies in pinpointing potential customers with higher probabilities of purchasing additional health insurance products. Employing quantitative methodology, the study extracted consumer data and applied various ML algorithms such as random forest, K-nearest neighbors, XGBoost classifier, and logistic regression using Python. Tailored feature engineering techniques were employed to enhance predictive accuracy. Analyzing 1,000,000 customer records with 16 features, Random Forest emerged as the top-performing model, achieving an accuracy score of 0.99 and F1 score of 1.00. The study reveals that customers aged 25–70, with prior insurance and longer service history, are more inclined to purchase additional health insurance products. These findings provide actionable insights for refining marketing strategies, boosting customer acquisition, and increasing revenue.

KEYWORDS

Health insurance; cross-selling; customer churn; machine learning algorithms; prediction; model training

Introduction

The intersection of data analytics and the insurance industry has paved way for innovative approaches to customer relationship management and revenue generation. One particularly intriguing avenue is the exploration of cross-selling opportunities within the health insurance industry.[1] Health insurance cross-selling is the practice of insurance companies offering health insurance products to existing consumers (policyholders) in addition to the other products that they are currently being covered, which is a critical component of the broader insurance sector that plays a pivotal role in the business growth and profitability of insurance companies and in safeguarding individuals and families against the financial burdens associated with medical expenses. With the evolving landscape of healthcare and insurance, providers are increasingly recognizing the importance of enhancing customer engagement and expanding their product offerings beyond traditional coverage.[2]

Healthcare is a basic human right and important to the society and the economy. It consists of all sectors (including insurance), providing services that promote the safety and well-being of the society. The health insurance industry is incredibly significant by ensuring affordable and accessible healthcare for people. It offers financial protection, promotes preventive care, enables risk pooling, and expands access to healthcare services, thus enhancing health outcomes and overall well-being. However, the healthcare sector faces various challenges such as pandemic, health disparity, infectious disease, etc. and it is continuously impacted by technology, thus necessitating data scientists to continuously monitor and respond toward a resilient system.

By harnessing the power of data analytics, insurers can gain valuable insights into customer behavior, preferences, and risk profiles.[3] The utilization of health insurance customer datasets in this context represents a paradigm shift in how insurance companies approach their business strategies. This research specifically focuses on understanding the potential for cross-selling within the health insurance domain by examining existing customer data to identify customers who are more likely to purchase additional insurance products to help insurers optimize their marketing strategies and increase business revenue. In South African insurance companies where the insurance market is highly competitive, accurate prediction of health insurance cross-selling becomes imperative for insurers to stay competitively ahead.[4]

CONTACT Khulekani Mavundla 20907985@dut4life.ac.za Department of Information Technology, Faculty of Accounting and Informatics, Durban University of Technology, PO Box 1334, Durban 4000, South Africa

Historically, cross-selling has been dependent on manual procedures and subjective decision-making.[5] Nevertheless, the advent of machine learning and predictive analytics provides insurers with the chance to employ data-driven methods for more efficient identification of potential cross-selling opportunities.[6] Applying machine learning algorithms enables insurance companies to analyze large volumes of health insurance customer datasets and uncover patterns and insights that may not be covered through traditional methods.[7]

The need for this current study is triggered by the dynamic business environment currently driven by the Fourth Industrial Revolution (4IR), which compel organizations to embrace emerging technologies like big data, AI, ML, IoT, edge, and cloud computing in order to stay competitive and drive innovation.[8] These advancements promise operational efficiencies while redefining customer expectations and market dynamics. This necessitates investigating how these technologies particular machine learning can be applied in the insurance sector, including the health insurance category, as it faces unique challenges in adapting to these technological shifts, particularly in cross-selling.[9] Unlike other financial institutions such as banks where transactional engagement is straightforward, insurance companies often encounter customer reluctance when promoting additional health insurance products, thus hindering efforts to expand market share and improve profitability.[10] Devising means to overcome this using ML approach is considered germane and innovative.

The aim of this research is to develop a health insurance cross-selling predictive model using machine learning algorithms to identify South African consumers who are more likely to purchase additional health insurance products. Thus, the following research questions are targeted in order to achieve the research aim. These include: (i) How can machine learning approach be utilized for building a predictive model for health insurance cross-selling? (ii) What marketing insights can be developed by insurance companies from the results of the health insurance cross-selling predictive model?

Through the examination of diverse consumer characteristics, including demographics, purchasing history, and socio-economic factors, such a predictive model seek to discern patterns and indicators that play a role in predicting health insurance cross-selling opportunities.[11] The ultimate goal is to assist health insurance companies in planning targeted marketing campaigns and personalized offers to maximize health insurance cross-selling success and profitability. This study builds upon existing research in the field of cross-selling prediction. While previous studies have explored cross-selling in various domains such as banking, there is a limited body of research specifically focused on health insurance cross-selling prediction of South African consumers using machine learning models.[12] The contributions of this current study address the research gap of a limited body of research specifically focused on health insurance cross-selling prediction of South African consumers using machine learning models; provide valuable marketing insights from health insurance cross-selling prediction; utilizing four varied algorithms and; overcoming the data barrier problem including dataset size, quality, accuracy, and access by utilizing a potent data source with 1 million records, such that the results obtained are credible and tailored toward generalization.

The significance of this study lies in its potential to explore an approach to cross-selling health insurance products through data analytics and machine learning while addressing two research questions. It examines how machine learning approach can be applied to build a predictive model for healthcare insurance cross-selling and, discover the demographic and behavioral factors that have the greatest influence on health insurance customers' likelihood to engage in cross-selling purchases using the specified dataset. It also states that marketing insights can be developed from the best results obtained thus strengthening the marketing strategies of health insurance companies in South African consumers.

By accurately predicting health insurance cross-selling, insurers can identify their potential customers more efficiently, improve customer acquisition and retention rates, and increase business revenue. Additionally, the findings of this study can help insurers to better understand the features influencing cross-selling among South African health insurance consumers toward developing suitable products and services for their target markets.

Literature review

AI and ML applications in the insurance sector highlight significant strides in enhancing operational efficiencies and customer engagement.[1,3] Despite advancements in predictive analytics in insurance and an evolving AI and ML with traditional business strategies to refine customer segmentation and personalized marketing,[6,7] few studies have addressed the unique market dynamics and consumer behaviors prevalent in South African consumers,[4] thus necessitating focused research.[5] Understanding and addressing these dynamics could provide insurers with actionable

insights to improve cross-selling effectiveness and customer satisfaction.

The application of machine learning techniques to predict cross-selling of health insurance products is hereby reviewed to include the concept of cross-selling prediction, analysis of consumer behavior in the health insurance sector, and exploring the utilization of machine learning techniques within the health insurance category.

Cross-selling prediction

Cross-selling prediction refers to the process of using data analysis and machine learning algorithms to forecast or estimate the likelihood of a customer purchasing additional products or services from a company.[13] The cross-selling prediction model aims to identify existing health insurance consumers who are more likely to purchase additional insurance products. By leveraging historical health insurance consumer data, such as demographic information, past purchasing behavior, and interactions with the company, machine learning models can be trained to recognize patterns and make predictions about potential cross-selling opportunities.[14] These predictions can significantly help insurance companies design their marketing strategies and offerings to target existing consumers who are more inclined to purchase additional health insurance products.

The cross-selling prediction process typically involves several steps, including data collection and preprocessing, feature engineering, model training, and evaluation.[15,16] Machine learning techniques such as classification algorithms, regression models, and ensemble methods are commonly employed to build predictive models that can identify potential cross-selling opportunities accurately.[10] By accurately predicting cross-selling opportunities, insurance companies can improve customer satisfaction, increase revenue, and improve overall business performance.[17] Additionally, customers may benefit from personalized recommendations and offers that align with their specific insurance needs and preferences.

South African health insurance consumer behavior analysis

Healthcare is a fundamental human right, essential to the society and the economy, with healthcare insurance instrumental in providing medical financial protection toward ensuring accessible and affordable healthcare for the populace for better health outcomes and improved productivity.

Analyzing consumer behavior in the South African health insurance market involves studying the patterns, preferences, and decision-making processes of individuals regarding purchasing and utilizing health insurance products.[18] Understanding consumer behavior is crucial for insurance companies to effectively market their offerings, tailor products to meet customers' needs, and improve customer satisfaction.[19] By conducting a comprehensive analysis of some factors including demographics, purchasing behavior, product preferences, utilization patterns, affordability, and accessibility, insurance companies can gain a deeper understanding of the South African consumers' behaviors and preferences. Thus, enabling them to develop targeted marketing strategies, improve product offerings, and enhance customer satisfaction in the health insurance market.[20]

Customer churn in health insurance

When conducting a consumer behavior analysis in the South African health insurance market, it is important to include an examination of customer churn, which refers to the rate at which customers discontinue their health insurance policies.[21,22] Analyzing and understanding this provides valuable insights into the factors that contribute to customer attrition and allows insurance companies to implement strategies to improve customer retention.[21,23]

Customer churn prediction is an ever-growing field of study globally including South Africa.[24] Hence, marketing and actively managing customer churn is now more critical and costly due to high competition and the aftermath of COVID-19 pandemic which has influenced many customers to churn or switch between companies.[25,26] Conducting study on customer churn prediction can support insurance companies in segregating their customers. Companies will be able to reprioritize their business strategies and customer service support to target customers appropriately by giving some incentives to retain them and focusing on cross-selling additional products to increase business revenue.[26]

Machine learning and artificial intelligence

Machine learning is a sub-domain of artificial intelligence (AI) that involves the use of algorithms and statistical models that enable computer systems to learn from data, make predictions or decisions, and improve their performance on a specific task based on the dataset utilized.[6,27] It involves training computer systems to learn from data, rather than relying on

explicit instructions or rules. In health insurance cross-selling prediction, it is used to develop models that analyze health insurance data on consumer demographics, past purchase behaviors, and other factors to predict cross-selling of additional health insurance products to policyholders. Machine learning can also be used in language translation, sentiment analysis, voice recognition, chatbots, etc., using the natural language processing (NLP) technique[28] as well as in fraud detection task by analyzing transaction data and detecting suspected fraudulent activities.[29]

The three types of machine learning[30] are hereby described.

Supervised learning model

In supervised learning, the model is trained on a labeled dataset, where each input is associated with a corresponding output.[31] The goal is to learn a mapping between the input and output so that the model will make accurate predictions on new (unseen) data. An example is medical diagnosis where extracted patient dataset is labeled with patients' corresponding diagnoses to train and build a supervised learning model to predict the diagnosis of new patients based on their symptoms and medical history.

Unsupervised learning model

In unsupervised learning, the model is trained on an unlabeled dataset, where there is no corresponding output for each input. The goal is to learn patterns or structures in the data, through techniques such as clustering or dimensionality reduction.[32] In clustering technique, similar customer data points are grouped together based on their characteristics. For instance, using a customer dataset with purchase history to identify other groups of customers with similar purchasing patterns using the K-means algorithm.[32]

Reinforcement learning model

In reinforcement learning, the model learns to make decisions based on feedback from the environment. The goal is to maximize a reward signal, such as a score or a profit, by taking actions that lead to positive outcomes.[33] Reinforcement learning applications include robotics and automation, gaming and entertainment, finance and trading, and in autonomous vehicles whereby the model is trained to make decisions in self-driving cars to detect when to change lane, accelerate or brake, and navigate intersections.[34]

Application of machine learning in health insurance cross-selling predictions

Machine learning is being used to develop a predictive cross-selling model by analyzing existing health insurance consumer data to determine the likelihood of consumers purchasing additional health insurance products.[34] It is used in consumer segmentation based on characteristics, such as age, gender, income, location, and previous purchase behavior to identify consumers who are most likely to be interested in cross-selling.[34,35]

This study utilizes a supervised learning approach to analyze existing health insurance consumer data for cross-selling predictions.[36] The health insurance historical customer data will be labeled with information on whether the existing customer will be interested to purchase an additional insurance product or not, thus providing a target variable for the algorithm to predict. By analyzing historical data with machine learning algorithms, insurance companies can identify patterns in customer behavior and make data-driven decisions on how to target cross-selling efforts for increasing business revenue and improving customer retention.[37]

The use of supervised learning in health insurance cross-selling prediction involves mapping between a set of input variables and an output variable, using health insurance data to train a model to accurately predict and provide an indication of whether a customer will likely buy an additional insurance product or not. In order to achieve this, the health insurance customer dataset will be divided into a training set and a test (validation) set.[38] The training set is used to train the model, while the test set is used to evaluate the performance of the model.

In supervised learning, a model predicts a continuous value, such as a house price, based on its features like location, size, number of bedrooms, and bathrooms. For instance, a random forest model can be trained on relevant dataset to predict new (unseen) house' prices. It will output both the predicted house value and the true sale prices from the training data.[33]

Machine learning life cycle for cross-selling prediction

This current study systematically applies the comprehensive machine learning life cycle,[33] which consists of eight distinct stages as shown in Figure 1, to a health insurance dataset for the purpose of cross-selling prediction. This approach ensures a thorough and structured analysis of the data to enhance the

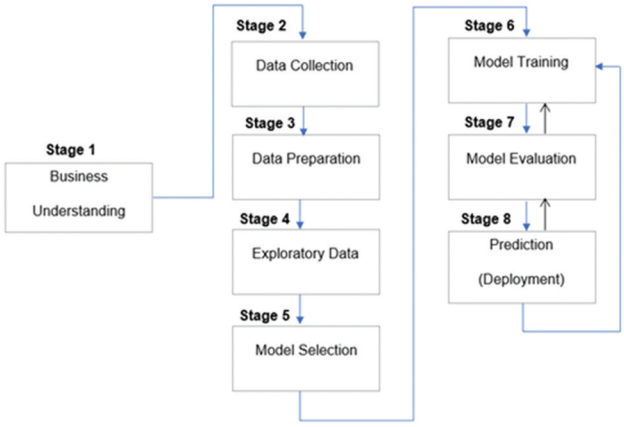

Figure 1. Eight stages of machine learning lifecycle[33].

accuracy and effectiveness of the prediction model. These steps are briefly discussed in subsequent relevant sections

Dataset and methodology

The health insurance dataset utilized in this study is a compilation of 1,000,000 customer records obtained from a large insurance company database in South Africa. We utilized SQL queries to extract health insurance customers' records from various tables within the expansive database.

Quantitative technique

Quantitative research utilizes deductive logic aimed at identifying patterns for instance in human lives by separating the social world into empirical components called variables which can be represented numerically.[39] For this study, the health insurance dataset was extracted from a large insurance company database in South Africa, focusing on customer behavior which can be quantified and patterned to extract meanings and insights.

Thus, this research used quantitative technique to focus on the collection of raw data and interpretation of numeric data where each health insurance dataset has a numerical value associated with it and its quantified information can be used for calculating statistical analysis so that decisions can be made after identifying if the existing health insurance customer will be interested in purchasing additional health insurance products.[40]

Health insurance dataset acquisition

Data acquisition is the second stage of the machine learning life cycle.[33] A health insurance customer dataset typically contains a collection of individual customers' records. The structure of the extracted dataset is as shown in Table 1.

Data collection in the context of the health insurance company in this study involves the compilation of related information concerning policyholders and insured members to construct a comprehensive dataset. The data primarily originates from a large insurance company database in South Africa, encompassing details relating to insurance policies, members, claims, and other relevant variables. The method employed for data collection entails using an SQL query to extract health insurance customer records from diverse database tables. A total of 1,000,000 records were harvested and compiled into a single table in the staging database. To protect privacy and comply with regulations including the Protection of Personal Information Act (POPIA) 2013, the dataset underwent anonymization, thus ensuring the security and confidentiality of customer information. This comprehensive dataset with 1,000,000

Table 1. Health insurance cross-selling dataset variables, data types, and definitions.

Variable	Definition	Data type
Gender	Gender of the consumer 0 = Other, 1 = Male, 2 = Female	Integer
Age	Age of the consumer	Integer
Region Code	Unique code for the region of the customer	Float
Race Code	Different race (Unknown, White, Black, Indian, Coloured) 1 = Unknown, 2 = White, 3 = African, 4 = Coloured, 5 = Asian	Integer
Previously Insured	1: Customer already has Health Insurance, 0: Customer does not have Health Insurance	Integer
Initial Sum Assured	An original fixed amount that will be paid to the nominee	Float
Current Sum Assured	A current fixed amount that will be paid to the nominee	Float
Monthly Income	Current consumer monthly income	Float
Monthly Premium	Amount consumer needs to pay every month for Health Insurance	Float
Annual Premium	Amount consumer needs to pay as a premium in the year for Health Insurance	Float
Vintage	Number of days, the consumer has been associated with the company	Integer
Insurer Type	1 = Consumer uses Internal Health Insurance product, 0 = Customer uses external Health Insurance product	Integer
Policy Status Type	1 = Active, 0 = Inactive	
Product Type Type	1 = Consumer has comprehensive health insurance cover, 2 = consumer has accident only cover, 3 = consumer has standard cover	Integer
Insurance Condition	1 = Health Insurance condition is Compulsory, 0 = health insurance condition is optional	Integer
Response	1: Consumer is perceived to be interested, 0: Customer is perceived not to be interested	Integer

Table 2. A sample of the raw health insurance dataset.

Gender	Age	Region Code	Race Code	Previously Insured	Initial Sum Assured	Current Sum Assured	Monthly Income	Monthly Premium	Annual Premium	Vintage	Resp
int64	int64	int64	int64	int64	float64	float64	float64	float64	float64	int64	int64
2	62	4093	2	1	195332.4	138986.52	0	367.23	4244.143	8	0
2	54	2162	3	1	575058.65	563762.120	23994	0	0	10	0
1	44	2302	2	1	542714.65	729075.35	51878.45	992.17	11357.52	3	0
1	41	3610	3	1	2600187.1	2544079.21	0	479.25	5485.024	9	0
2	60	4037	5	1	252602.74	136684.35	0	430.22	5011.394	4	0
1	53	1813	2	1	1457869.95	1000000	0	1271.11	14712.9	1	0
1	43	163	2	1	1109078.81	759778.59	24500	707.04	8189.997	10	0
1	45	182	3	1	445814.88	445814.88	0	557.3	6439.693	15	0
2	54	7490	4	1	361963.09	290403.15	0	376.14	4326.319	8	0
2	56	2190	3	1	301754.35	252398.73	18840.25	618.49	7106.655	9	0

records and 16 feature variables provides valuable insights into diverse customers and their respective health insurance policies.

Thereafter, the extracted dataset was imported using the Python Pandas library in Jupiter Lab so that it can be trained using the random forest, k-nearest neighbors, XGBoost classifier, and logistic regression algorithms considered in this study. To further ensure data accuracy, completeness, integrity, and validity before proceeding to process the data, we verified the extracted dataset by comparing it with the source of truth and found correctness. The extracted dataset was viewed to ensure that the fields and features were imported successfully. A sample of the top 10 records is shown in Table 2.

Table 3 indicates the dimensions or structure of the dataset, representing the number of rows (instances) and columns (feature variables) it contains. In the Table, the validation dataset consists of 15 variables due to the independent variable that is changed or controlled in order to assess the effects it will have on the dependent variable. The independent variable takes all 16 columns and removes the last column "Response" which is a target variable that indicates 1 for consumers who perceive to be interested and 0 for consumers who perceive not to be interested in health insurance cross-selling.

Data pre-processing

Data pre-processing and preparation which is the third stage in the machine learning life cycle[33] is a crucial stage in the data analysis and machine learning pipeline, involving the cleaning, transformation, and preparation of the raw data for subsequent analysis. Python code was employed to clean the extracted data and eliminate duplicates, correct inaccuracies, normalize the data, and address data entry errors for an enhanced feature engineering. Techniques such as imputation and removal of rows/columns with missing values were applied to handle missing data. The pre-processed dataset represents a curated subset, carefully refined to enhance its suitability for machine learning task in order to ensure optimal model training. It serves as a refined and standardized input for machine learning algorithms, facilitating more accurate and effective model development.

Following data pre-processing, some specific health insurance dataset records were eliminated, resulting in a modified dataset shape with 713,538 instances and 16 feature variables each as indicated in Table 4. The pre-processed dataset was viewed before splitting into training and testing subsets to assess the model's performance accurately during the training and testing tasks. A sample of the top 10 records is shown in Table 5.

Exploratory data analysis (EDA)

In the exploratory data analysis process, as the fourth stage of the machine learning life cycle,[33] a component of data pre-processing was executed to underscore the significance of delving into and comprehending the extracted dataset. The EDA in this study encompassed evaluating aspects such as the dataset size and variable data types. Additionally, descriptive analysis, distribution assessments, correlation examinations, and the creation of data visualizations were conducted. EDA proves to be pivotal in data analysis task by providing a means to explore and summarize the fundamental characteristics of the dataset. This process aids in gaining insights,

Table 3. Dimensions of the health insurance training and validation dataset.

	Training Dataset Shape	Validation Dataset Shape
Number of Rows	1 000 000	1 000 000
Number of Columns	16	15

Table 4. Dimension of the pre-processed health insurance dataset.

Dataset Shape after Cleaning and Preprocessing	
Number of Rows	713538
Number of Columns	16

Table 5. A sample of the pre-processed health insurance dataset.

Gender	Age	Region Code	Race Code	Previously Insured	Sum Assured	Monthly Income	Monthly Premium	Annual Premium	Vintage	Resp
1	44	2302	2	1	729 075.35	51 878.45	992.17	11 357.52	3	0
1	43	163	2	1	759 778.59	24 500.00	707.04	8 190.00	10	0
2	56	2190	3	1	252 398.73	18 840.25	618.49	7 106.65	9	0
2	59	9786	3	1	285 339.33	15 748.00	471.27	5 465.18	2	0
1	31	157	4	1	823 986.79	29 750.00	526.57	6 055.17	12	0
1	46	1818	3	1	521 995.54	36 904.16	755.98	8 589.81	20	0
2	52	7100	3	1	188 678.85	7 820.55	243.68	2 790.08	20	0
1	40	5201	3	1	759 578.71	40 789.75	767.42	8 761.35	19	0
2	46	4066	3	1	653 353.80	37 303.87	700.46	7 933.77	2	0
2	52	1618	3	1	672 240.25	29 313.59	2 347.55	26 980.11	18	0

understanding underlying patterns, identifying trends, and uncovering potential issues within the dataset.

Variables correlation in the dataset

Correlation refers to the statistical connection between variables, quantifying both the strength and direction of their linear association. Its computation involves comparing the variations in the data points of two variables, and the resulting coefficient ranges from −1 to 1. A coefficient of 1 signifies a perfect positive correlation, while −1 indicates a perfect negative correlation, implying that as one variable increases, the other decreases proportionally. A correlation coefficient of 0 denotes no linear correlation between the variables, suggesting that changes in one variable do not predict changes in the other. This scale provides a clear and concise way to interpret the strength and direction of relationships of variables within the health insurance dataset.

The Pearson correlation coefficient between two variables, X and Y, is calculated thus:

$$r = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2 \sum_{i=1}^{n}(Y_i - \bar{Y})^2}} \quad (1)$$

where Xi and Yi are the individual data points; $\bar{X}$ and $\bar{Y}$ are the means of variables X and Y, respectively; and n is the number of data points.

As presented in Figure 2, the correlation analysis of the feature variables for health insurance cross-selling prediction for some selected features which are age, monthly premium, and gender include:

Age (−0.054): The negative correlation coefficient indicates that as the age of the consumer increases, the likelihood of purchasing cross-selling insurance product decreases. This suggests that younger individuals may be more inclined to buy additional insurance products compared to older individuals.

Monthly Premium (+0.0026): The positive correlation coefficient suggests that as the monthly premium paid by the customer increases, the likelihood of purchasing cross-selling insurance product also increases. This implies that consumers who are already paying premiums for their current policy may be more open to purchasing and willing to pay for additional products.

Gender (+0.0013): The positive correlation coefficient suggests that gender has a slight influence on the likelihood of purchasing the cross-selling insurance product.

Visualization of the health insurance data patterns

In this study, data visualization employs graphical elements such as charts and graphs to represent information and help present the features influencing the target variable in the health insurance dataset, thus facilitating data exploration, pattern identification, data characteristic analysis, outlier detection, and effective communication of findings.

Figure 3 depicts the vintage response, indicating the duration of customer association with the insurance company in terms of years. The vintage represents the number of years consumers use company services from 1 to 30 years versus the number of consumers in the count.

Figure 4 however showcases the age versus frequency of likelihood curve. After analyzing the health insurance consumer data based on consumers who are previously insured, analyses indicated that consumers aged 25 to 70 exhibit a higher likelihood of expressing interest in purchasing additional health insurance products compared to customers younger than 25 or older than 70 years old.

Figure 5 illustrates gender counts, categorized as other (0), male (1), and female (2), depicted in both a bar graph and a pie chart with percentage distribution. This representation unveils a dataset bias toward male records though the margin is not significant.

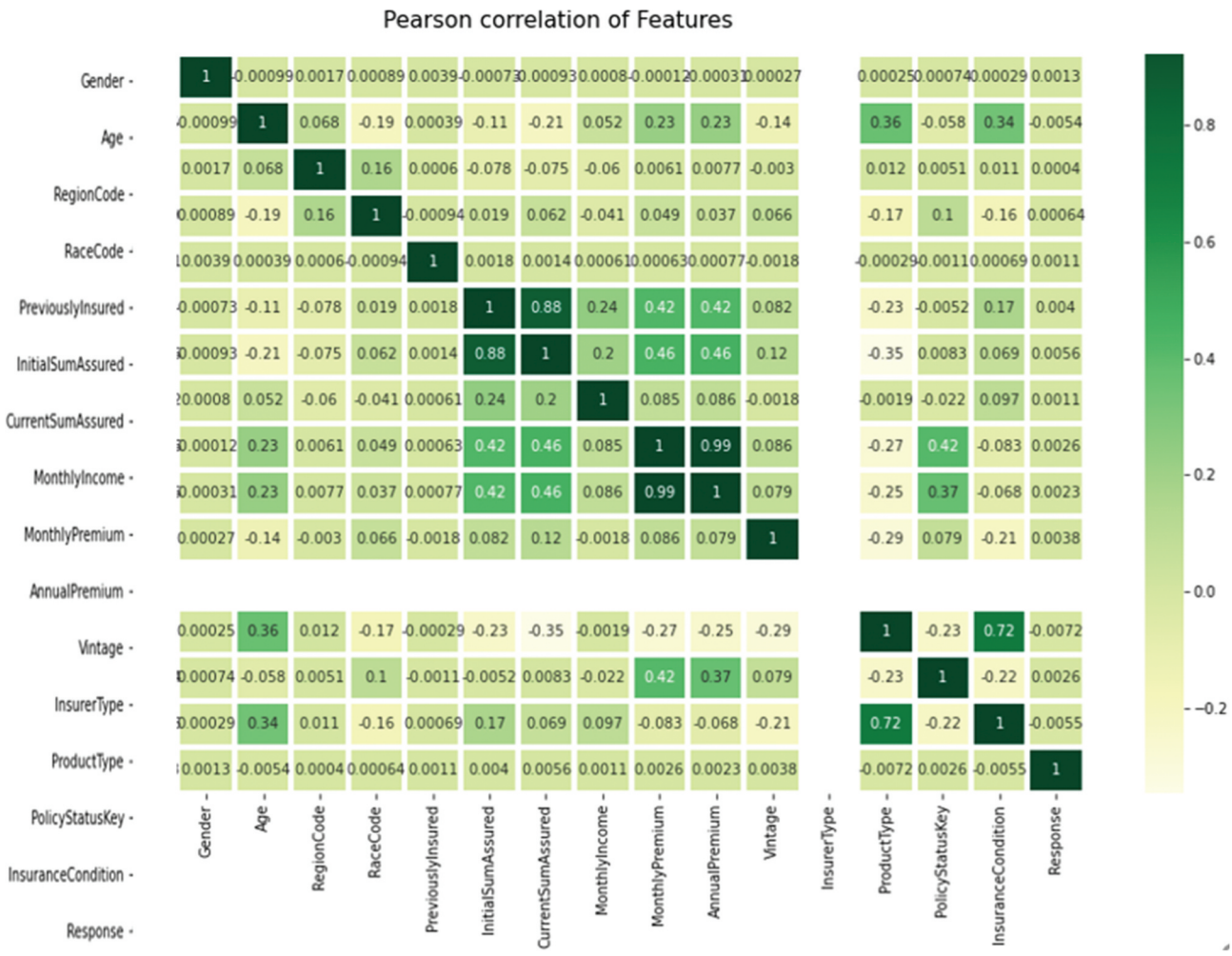

Figure 2. Pearson correlation of the feature variables.

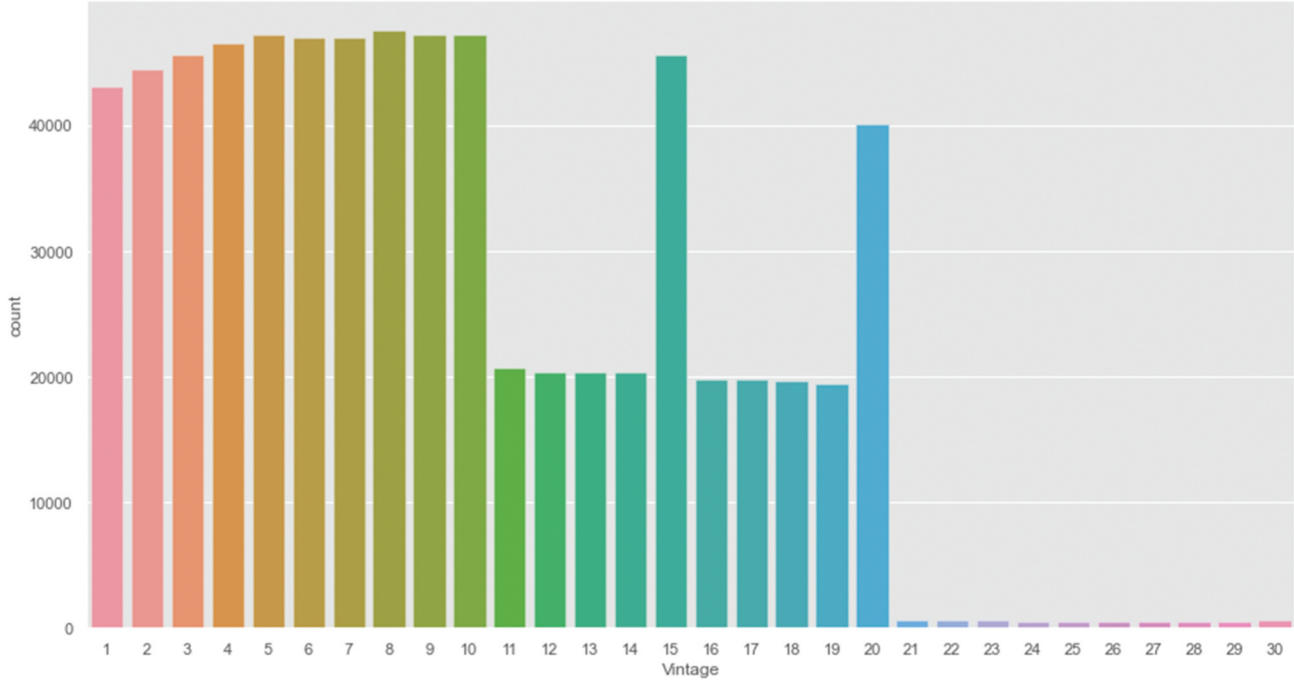

Figure 3. Customer long service response (vintage).

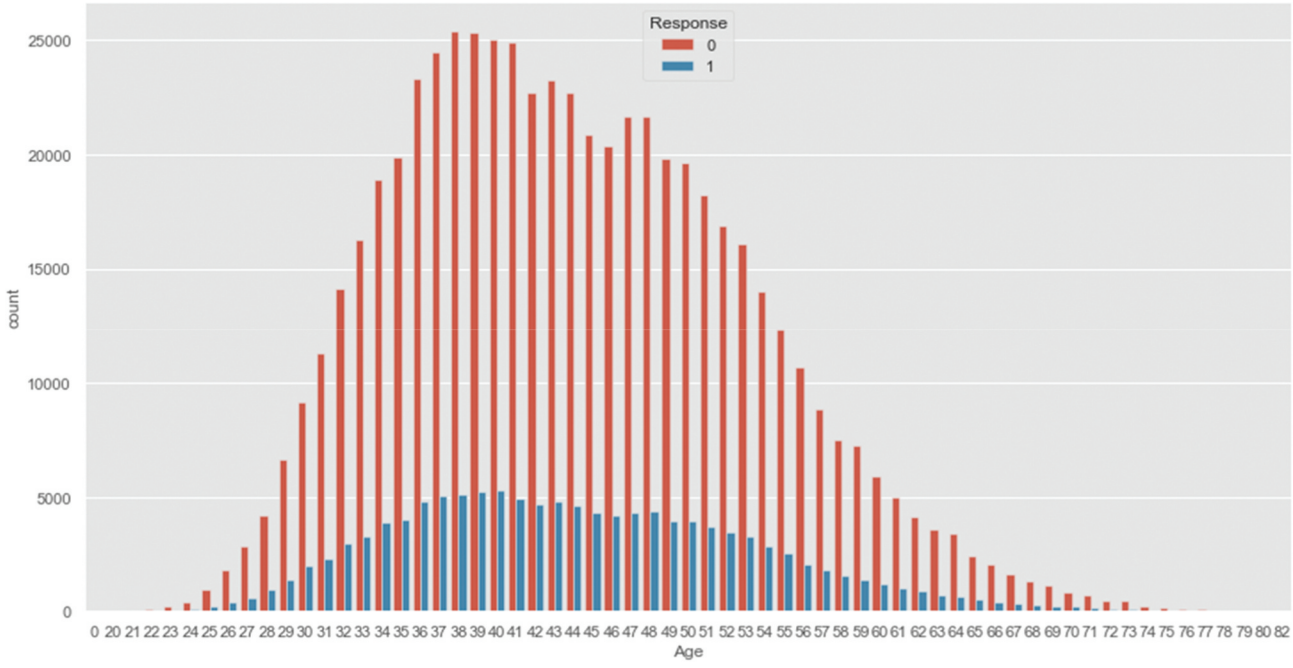

Figure 4. Plot of age and frequency of consumer cross-selling purchase likelihood.

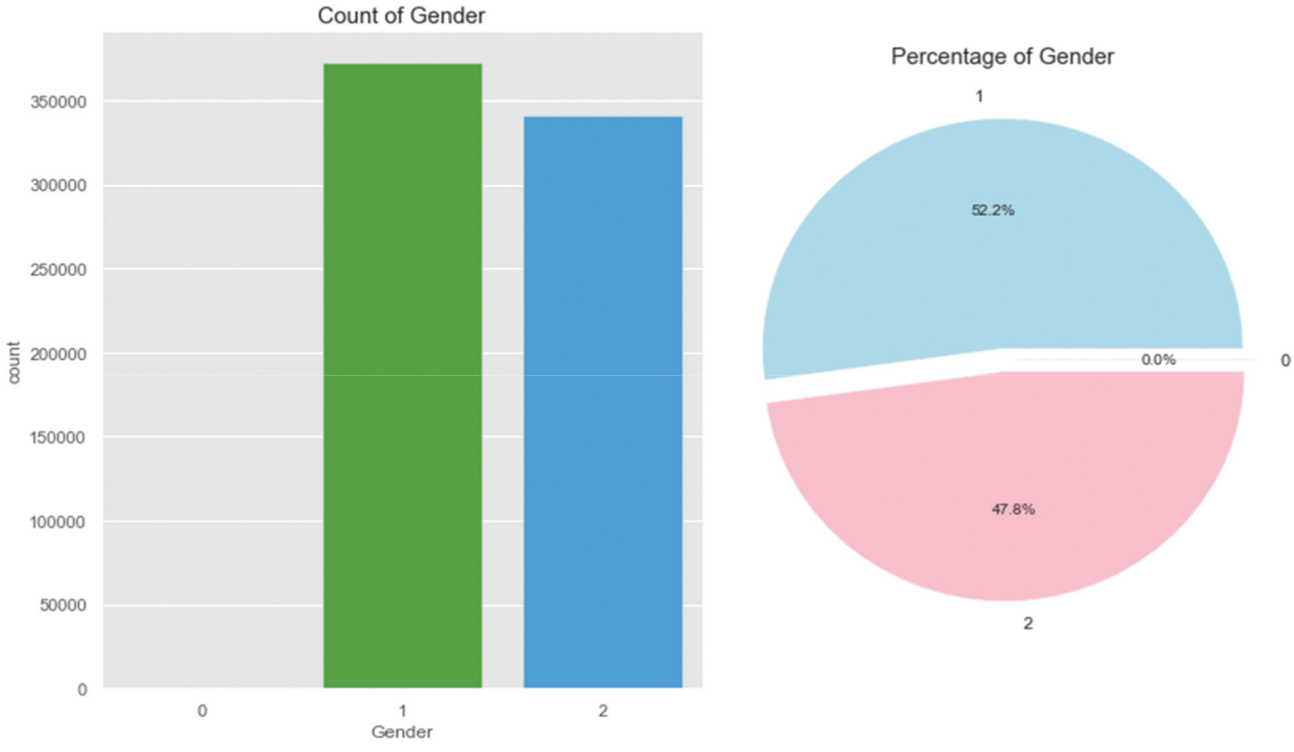

Figure 5. Gender distribution of the health insurance dataset.

Employed machine learning algorithms for model training and evaluation

The next step in the machine learning life cycle process as proposed by[33] and applied in this current study is model selection. In our health insurance cross-selling prediction task, the random forest, k-nearest neighbor, logistic regression, and XGBoost classifier machine-learning algorithms were selected, trained, and their performances compared in order to determine and select the best and most accurate algorithm among

them for the purpose of predicting cross-selling opportunities in health insurance consumers/policyholders in South Africa. These classifiers were chosen because of their popularity in the machine learning domain after considering the type and size of the health insurance consumer dataset, the algorithms' performance potentials including speed and accuracy, the prediction problem at hand, as well as the complexities of the various classifiers.

Random forest algorithm

Random forest algorithm was applied in health insurance cross-selling prediction using the insurance dataset on an ensemble of decision trees.[40] This algorithm works by randomly selecting a subset of the features (variables) from the insurance dataset for each decision tree and splitting the dataset based on the selected features. The result is a collection of decision trees that make independent predictions, which are combined to form the final prediction of whether a customer will accept a cross-selling offer or not. The random forest has been successfully used to predict credit card application approval based on customer's existing dataset using features such as income, age, credit score, and employment status.[41]

K-nearest neighbors algorithm

The K-NN model involve training data set by assigning each datapoint to the class of its k-nearest neighbors.[16] The value of k was selected through cross-validation to predict the possibility of cross-selling additional insurance products to an existing customer. For example, suppose there is a dataset of customer purchase history in a retail store, along with their corresponding feature variables, such as gender, age, and purchase history, K-NN has been used to predict whether consumers will be interested in making a purchase based on these features by analyzing their historical dataset.[16,42]

XGBoost classifier

XGBoost is another powerful machine learning algorithm, that we employed for health insurance cross-selling prediction using a collective of decision trees. This model was built iteratively, with each new tree attempting to correct the mistakes of the previous tree. The hyperparameters of the model, including the learning rate, maximum depth of the tree, and the number of trees, were tuned using common machine learning techniques. XGBoost has been used for automotive insurance claim prediction problem using insurance consumer's historical data.[43] It has also been used for building machine learning models as recommendation systems by training insurance datasets to identify customer purchasing patterns and recommend them for insurance cross-selling of additional products.[44]

Logistic regression algorithm

Logistic regression technique was applied in this study for health insurance cross-selling prediction using the insurance customer dataset cross-selling outcome and the dependent variable including age, gender, previous insured, monthly income, and vintage to predict the health insurance customer likelihood of accepting the cross-selling offer,[45] and consequently making a purchase. The logistic regression technique calculates the probability of an existing customer accepting the purchase offer and purchasing a health insurance product, by using a sigmoid function where the result is either 0 or 1, with 0 indicating the customer is not interested and 1 means interested.[45,46]

After evaluating the performance of the models, the best and most accurate model was selected and applied to the health insurance cross-selling prediction. The most used prediction model is logistic regression[47] and was also considered in this study. The logistic regression model was applied to train the health insurance customer historical data that includes features such as consumer demographics, previous purchase history, and policy information that may impact the consumer's decision to purchase an additional health insurance product.

Model training and evaluation

Training a machine learning model is the sixth stage of the machine learning life cycle as suggested by[33] and adopted in this current study. It is a procedure that entails guiding algorithms to recognize patterns and make predictions using input data.[48] Training each model involving the random forest, k-nearest neighbor, logistic regression, and XGBoost algorithms was done by feeding as input, customers' health insurance historical data consisting of 713, 538 records/instances each with 16 feature variable size which includes gender, age, region code, race code, previously insured, initial sum insured, current sum insured, monthly income, monthly premium, annual premium, vintage, insurance type, policy status key, product type, insurance condition, and response as target variable that will indicate the expected output as 1 for consumers who perceive

and 0 not perceive to purchase health insurance additional products.

The health insurance cross-selling dataset for training was split into 80% training set and 20% testing set. Subsequently, for the model evaluation stage, as suggested by Hong et al.. (2020), some evaluation metrics were employed to gauge the effectiveness of the machine learning model by, assessing its capacity to apply knowledge to new/unseen data. The choice of evaluation metrics is contingent upon the task at hand and the characteristics of the problem being tackled. To enable us to assess the performance and quantify the effectiveness of the machine learning models developed in this current study, we utilized the accuracy precision, recall, and F1 score evaluation metrics.

For our health insurance cross-selling prediction task, these evaluation metrics were chosen because they provide a comprehensive evaluation of the model's performance in identifying customers who are more likely to purchase the cross-selling insurance product. Accuracy is a measure of the total correctly predicted samples out of the entire samples in the dataset, Precision measures the accuracy of positive predictions, Recall assesses the ability to capture all positive instances, and F1-Score combines both precision and recall into a single metric, thereby providing a balanced assessment of the model's performance.

The metrics are computed as follows:

$$Precision = \frac{TruePositives}{TruePositives + FalsePositives} \quad (2)$$

$$Recall = \frac{True\ Positives}{True\ Positives + False\ Negatives} \quad (3)$$

$$F1Score = \frac{2\ x\ Precision\ x\ Recall}{Precision + Recall} \quad (4)$$

The enumerated step-by-step machine learning life cycle for health insurance cross-selling prediction which include data collection, data preparation, exploratory data analysis, model selection, model training, model evaluation, and prediction, as well as other reviewed literature presented in this study, answers the study's research question 1 of how can machine learning approach be utilized for building a predictive model for health insurance cross-selling.

Results

The following results were obtained for each ML algorithm after the model was trained, and its performance was evaluated.

The Random Forest model exhibited an outstanding performance as it achieved an accuracy score of 0.9988 as shown in Table 6. Notably, the model showcased higher precision and recall for the first class (0) in comparison to the second class (1). Specifically, it achieved an F1-Score of 1.00 for class 0 and both the macro-average and weighted average F1-Scores were calculated to be 1.00. These results highlight the model's exceptional capability and accuracy across both classified classes.

However, the testing of the Random Forest model resulted in an accuracy score of 0.7961, as shown in Table 7 using 20% of the consumer dataset, which amounts to 142,708 records. It showed higher precision and recall for the first class (0) compared to the second class (1), with an F1-Score of 0.89 for class 0. However, the macro-average F1-Score is relatively low at 0.48, and

Table 6. Random forest classifier training results.

Random Forest Accuracy Score: 0.99877				
Classification report:	Precision	Recall	F1-score	Support
0	1.00	1.00	1.00	474601
1	1.00	1.00	0.99	96229
Accuracy			1.00	570830
macro avg	1.00	1.00	1.00	570830
weighted avg	1.00	1.00	1.00	570830

Table 7. Random forest classifier testing results.

Random Forest Accuracy Score: 0.79605				
Classification report:	Precision	Recall	F1-score	Support
0	0.83	0.95	0.89	118453
1	0.18	0.06	0.08	24255
Accuracy			0.80	142708
macro avg.	0.50	0.50	0.48	142708
weighted avg.	0.72	0.80	0.75	142708

the weighted average F1-Score is 0.75, indicating varying performance across both classes.

Nonetheless, the K-nearest neighbors (KNN) classifier as shown in Table 8 achieved an accuracy score of 0.8601 using 80% of the dataset, which is 570,830 records for training. It displayed higher precision and recall for class (0) than for class (1), resulting in an F1-Score of 0.92 for class 0. The macro-average F1-Score, indicating a balanced evaluation across both classes, was observed at 0.67, while the weighted average F1-Score stood at 0.84. With an accuracy score of 0.86, the model demonstrates its strength to make accurate predictions, considering both true positives and true negatives.

The testing of the K-nearest neighbors (KNN) classifier posted an accuracy score of 0.7802, using 20% of the dataset (142708 records) as displayed in Table 9. In the KNN results, it was evident that the classifier displayed higher precision and recall for class (0) compared to class (1), resulting in an F1-Score of 0.87 for class 0. The macro-average F1-Score, reflecting a balanced evaluation across both classes, was observed at 0.49, while the weighted average F1-Score stood at 0.74.

In the next experiment, the training of the XGBoost model yielded an accuracy score of 0.8316, utilizing 80% of the dataset (570,830 records) for training, as indicated in Table 10 thus showcasing its overall performance. Notably, the XGBoost model exhibited a precision of 0.83 and a recall of 1.00 for class (0), underscoring its capability to accurately identify instances of this class. The accuracy score of 0.8316 underscores the model's proficiency in making precise predictions, considering both true positives and true negatives.

As indicated in Table 11, the testing of the XGBoost model achieved a commendable accuracy score of 0.8300, utilizing a testing dataset comprising 142,708 records, thus indicating its strong performance. The model's precision is 0.83 and recall is 1.00, respectively,

Table 8. K-nearest neighbors (KNN) classifier training results.

K-Neighbors Classifier Accuracy Score: 0.86007				
Classification report:	Precision	Recall	F1-score	Support
0	0.87	0.97	0.92	474601
1	0.69	0.31	0.43	96229
Accuracy			0.86	570830
macro avg.	0.78	0.64	0.67	570830
weighted avg.	0.84	0.86	0.84	570830

Table 9. K-nearest neighbors (KNN) classifier testing results.

K-Neighbors Classifier Accuracy Score: 0.78016				
Classification report:	Precision	Recall	F1-score	Support
0	0.83	0.92	0.87	118453
1	0.18	0.08	0.11	24255
Accuracy			0.78	142708
macro avg.	0.50	0.50	0.49	142708
weighted avg.	0.72	0.78	0.74	142708

Table 10. XGBoost classifier training results.

XGBoost Accuracy Score: 0.83157				
Classification report:	Precision	Recall	F1-score	Support
0	0.83	1.00	0.91	474601
1	0.88	0.00	0.00	96229
Accuracy			0.83	570830
macro avg.	0.85	0.50	0.46	570830
weighted avg.	0.84	0.83	0.76	570830

Table 11. XGBoost classifier testing results.

XGBoost Accuracy Score: 0.82995				
Classification report:	Precision	Recall	F1-score	Support
0	0.83	1.00	0.91	118453
1	0.23	0.00	0.00	24255
Accuracy			0.83	142708
macro avg.	0.53	0.50	0.45	142708
weighted avg.	0.73	0.83	0.75	142708

Table 12. Logistic regression classifier training results.

Logistic Regression Accuracy Score: 0.83142				
Classification report:	Precision	Recall	F1-score	Support
0	0.83	1.00	0.91	474601
1	0.00	0.00	0.00	96229
Accuracy			0.83	570830
macro avg.	0.42	0.50	0.45	570830
weighted avg.	0.69	0.83	0.75	570830

for class (0), highlighting its capability to precisely identify instances of this class.

The training of the Logistic Regression model, however, displayed an accuracy score of 0.8314 as indicated in Table 12, leveraging 80% of the dataset (570830 records) for training, it exhibits a commendable precision of 0.83 and a recall of 1.00 for class (0), indicating its proficiency in accurately identifying instances of this class. With an accuracy of 0.8314, the model demonstrates its capability in making precise predictions, encompassing both true positives and true negatives.

As shown in Table 13, the testing of the logistic regression model revealed a significant accuracy score of 0.8300, employing a testing dataset of 142,708 records, hence highlighting its strong overall performance. The logistic regression model displays commendable precision of 0.8300 and recall of 1.00 for class (0), indicating its capability to accurately identify instances of this class. With an accuracy of 0.8300, the model demonstrates its adeptness in making precise predictions, covering both true positives and true negatives.

The results obtained unequivocally demonstrate that all the machine learning algorithms performed very well by achieving relatively high accuracy levels. However, among the algorithms considered, the Random Forest demonstrated superior overall performance, boasting an accuracy of 0.99 and an F1 score of 1.00. Consequently, the Random Forest model is selected as the optimal choice for building a health insurance cross-selling prediction model with the historical health insurance customer dataset that was utilized in this study. The exceptional accuracy and F1 score of the Random Forest model signify its robustness and effectiveness in accurately predicting potential customers for health insurance cross-selling. This model's high performance makes it a reliable tool for guiding marketing strategies and maximizing cross-selling opportunities within the insurance industry.

Discussion

Upon analyzing the results obtained for the four models as shown in Tables 6–13, it becomes evident that the random forest model stands out with 6 min and 35.6 s computation time, boasting the highest F1-Score of 1.00 and an accuracy score of 0.99. This underscores its exceptional equilibrium between precision and recall, indicating proficiency in accurately identifying positive instances while minimizing both false positives and false negatives. Conversely, the KNN demonstrated an F1-Score of 0.92 while XGBoost classifier, and Logistic Regression posted an F1-Score of 0.91, showcasing comparable performances in accurately classifying positive instances and maintaining a balance between precision and recall.

The difference in computation times among the training algorithms for health insurance cross-selling predictions as indicated in Table 14 suggests variations in their computational complexity and efficiency. Random Forest with 6 m 34.6s and Logistic Regression with 4 m 5.2s demonstrated relatively shorter computation times compared to KNN that took 30 m 12.4s, thus implying faster processing and model training. XGBoost classifier fell between these extremes in terms of computation time of 5 m 5.5s. However, while computation time provides insights into model's efficiency and speed, it cannot alone determine the model's overall performance. Other performance metrics like accuracy, precision, and recall should also be considered for a comprehensive assessment of each model's effectiveness in making accurate predictions for health insurance cross-selling.

Table 13. Logistic Regression testing results.

Logistic Regression Accuracy Score: 0.83003				
Classification report:	Precision	Recall	F1-score	Support
0	0.83	1.00	0.91	118453
1	0.00	0.00	0.00	24255
Accuracy			0.83	142708
macro avg.	0.42	0.50	0.45	142708
weighted avg.	0.69	0.83	0.75	142708

Table 14. Summary of models' results.

ML Algorithms model	Accuracy score		F1-Score	Computation (min & sec)
	Training	Testing		
Random Forest	0.9987	0.7960	1.00	6m 34.6s
K-nearest neighbors	0.8600	0.7801	0.92	30m 12.4s
XGBoost	0.8315	0.8299	0.91	5m 5.5s
Logistic Regression	0.8314	0.8300	0.91	4m 5.2s

A summary of the results obtained after training and testing each model is illustrated in Table 14.

However, the visual representation in Figure 6 provides a comparative analysis between consumers with a history of prior insurance coverage, who exhibit a higher likelihood of responding to the health insurance supplementary products, and consumers without previous insurance. This visual comparison sheds light on the distinctive response patterns between these two customer segments, contributing valuable insights into the factors influencing the propensity to engage with additional health insurance products based on their insurance history.

The results indicating 83.1% of consumers responded negatively to cross-selling and 16.9% responded positively were obtained by analyzing the dataset containing consumers' policy information and their previous responses to health insurance cross-selling offers. The dataset sample size was used to calculate the percentage of consumers interested and those not interested in the supplementary product. Additionally, a comparative analysis was conducted between consumers with and without previous insurance coverage to determine their respective likelihood of responding to the cross-selling offer.

The examination of the health insurance dataset unveiled a notable trend: individuals aged 25 to 70 display a heightened probability of acquiring supplementary health insurance products. This underscores the predictive significance of age in understanding and forecasting customer behavior. The practical implications of this for corporate marketing are that they provide insights into informed actionable strategies to be implemented for cross-selling insurance products. By focusing on personalized approaches and leveraging the identified characteristics, preferences, and behaviors of the target demographic, insurance companies can, for

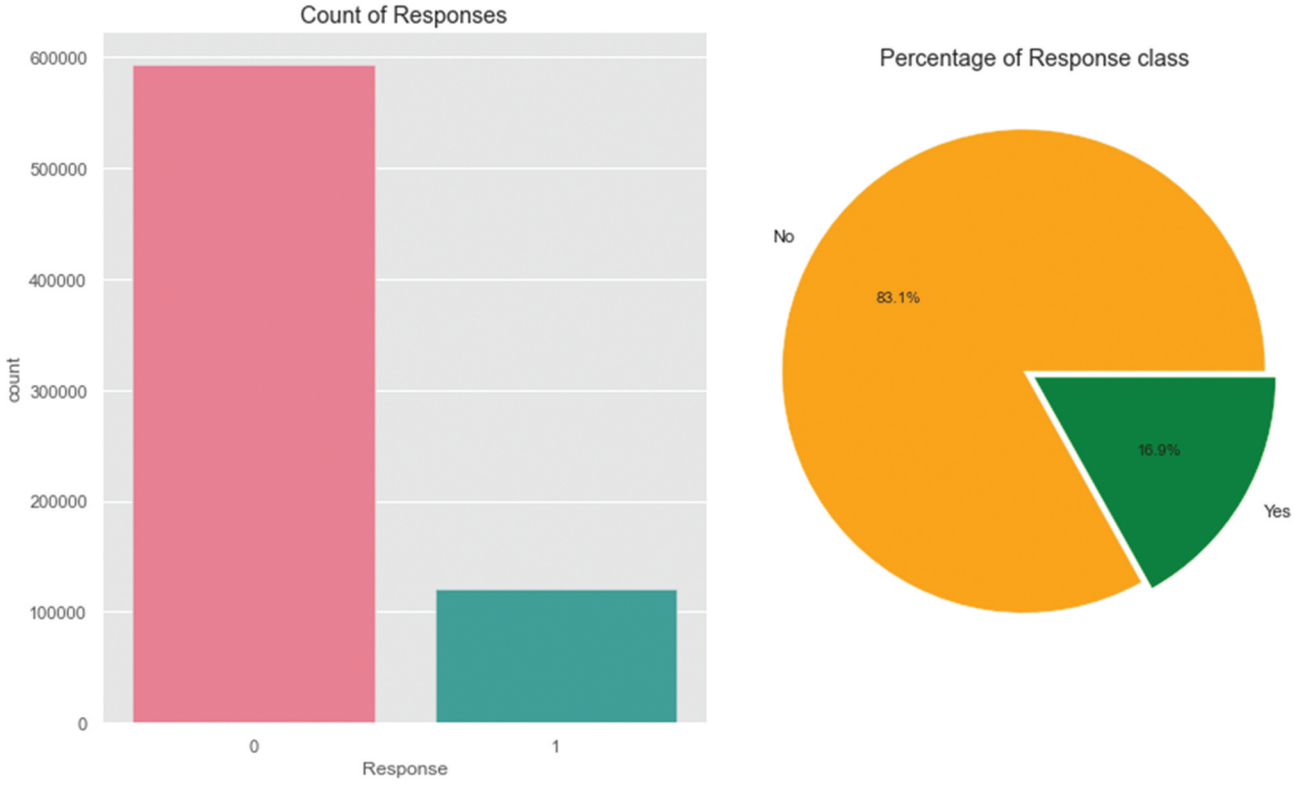

Figure 6. Result of consumers interested vs. not interested in insurance cross-selling.

instance, implement targeted e-mail campaigns and personalized messaging that highlight the specific benefits of additional coverage tailored to their needs. This provides answers to this study's research question 2 of what marketing insights can be developed by insurance companies from the results of the health insurance cross-selling predictive model. In addition, it is recommended that insurance companies should provide insurance education, develop incentives backed loyalty programs and referrals for the identified group to encourage them to purchase more products and refer others while also using the appropriate message communication channels and appeals.

Furthermore, the duration of a customer's association with the company referred to as "Vintage," surfaced as a pivotal factor. Consumers with a history of previous insurance coverage demonstrated an inclination toward additional health insurance products. This underscores the importance of leveraging historical data in developing an effective predictive model for customer preferences and behavior in the insurance domain. Additionally, the study identified features significantly influencing the prediction model, including *Gender, Age, Previously Insured status, Monthly Income, Monthly Premium*, and *Annual Premium.*

These features were identified during the training phase, as the model learns the relationships between these features and the response as the target variable. This learning process involves finding coefficients or weights for each feature that optimize the model's ability to make accurate predictions. Larger coefficients and statistical significance indicate a stronger influence on predictions.

The correlation coefficients indicate the strength and direction of the linear relationship between each feature and the response target variable. A positive coefficient suggests a positive correlation (as the feature variable increases, the likelihood of purchasing the cross-selling product also increases), while a negative coefficient suggests a negative correlation (as the feature variable increases, the likelihood of purchasing the cross-selling product decreases). Other factors such as feature importance analysis, model performance metrics (e.g., accuracy, precision, and recall), and domain knowledge can help in assessing the importance of these features in making predictions.

In this study, the insights from analyzing the health insurance customer records can be applied to enhance marketing strategies and increase revenue through targeted health insurance cross-selling. This analysis will advance our understanding of applied machine learning techniques in the health insurance industry and provide a framework for optimizing these methods in real-world scenarios. Considering the 1million dataset we utilized in this study, our understanding of applied machine learning in healthcare (and health insurance) is also forwarded in respect of breaking the data barrier challenges such as data access, data quality, accuracy, and dataset size (quantity). Algorithms are only as good as the data source and size as they determine performance and generalization. Small datasets used to train and validate AI can cause unavoidable problems, thus the creation of a health insurance data system that will feed prospective data into AI algorithm is suggested.

Conclusion

The current research underscores the significant potential inherent in the application of machine learning algorithms to meticulously analyze health insurance datasets, and discerning nuanced cross-selling probabilities. The abundance of extensive customer data within the health insurance sector presents a distinctive and advantageous opportunity, such that machine learning models are exceptionally well-suited to harness and capitalize on. Given the vast reservoirs of customer data within the health insurance domain, machine learning models can proficiently sift through this information to unveil patterns, behaviors, and preferences. Consequently, these models play a pivotal role in identifying potential customers who may express interest in acquiring additional insurance products.

Through the adept utilization of predictive machine learning models for cross-selling in health insurance, these algorithms precisely forecast the likelihood of a customer's inclination toward purchasing supplementary products. This capability empowers insurance companies to tailor their cross-selling endeavors, offering personalized recommendations that align with individual customer needs and preferences. As the industry undergoes continual evolution, the adoption of machine learning algorithms emerges as increasingly vital, promising to foster innovation, enhance operational efficiency, and elevate overall customer satisfaction.

The current study's limitations include non-consideration for external factors such as changes in healthcare regulations or economic conditions. These factors can significantly influence consumer behavior, and purchasing decisions regarding insurance products but might not have been accounted for in the analysis. As a result, the predictive models may not accurately capture real-world scenarios, leading to potentially flawed predictions and ineffective cross-selling strategies.

Essentially, incorporating machine learning into the comprehensive analysis of health insurance datasets for

the purpose of predicting cross-selling probabilities is strategically imperative. This integration ensures that the industry remains not only competitive but also adaptive, and adept at meeting the dynamic and evolving needs of its customers. Ultimately, this optimization of business processes leads to a substantial increase in revenue. Other practical implications for research may include refining strategies for boosting customer retention through personalized insurance offerings, driven by advanced ML algorithms integrated with customer relationship management systems, should be prioritized, while ML ethical considerations relating to customer data usage should be evaluated to ensure confidentiality and foster trust. Future research endeavors could consider further analysis, incorporating, and refining the insights gained from health insurance cross-selling predictions.

Disclosure statement

No potential conflict of interest was reported by the author(s).

References

1. Son W, Kang S. Determinants of insurance products cross-selling performance: focusing on career experience. J Serv Res Stud. 2019;9(3):39–60. doi:10.18807/jsrs.2019.9.3.039.
2. Agarwal R, Dugas M, Gao G, Kannan PK. Emerging technologies and analytics for a new era of value-centered marketing in healthcare. J Acad Mark Sci. 2020;48:9–23. doi:10.1007/s11747-019-00692-4.
3. Beaudon G, Soulier E. Customer experience analytics in insurance: trajectory, service interaction, and contextual data. In: Rocha A, Ferras C Paredes M. editors. Information technology and systems. Advances in intelligent systems and computing. Vol. 918. Cham: Springer; 2019. p. 203–214. doi:10.1007/978-3-030-11890-7_19.
4. De Jager JH. A strategic framework for relationship development in a South African short-term insurance intermediary company [Doctoral dissertation]. South Africa (SA): North-West University; 2021.
5. Gupta S, Justy T, Kamboj S, Kumar A, Kristoffersen E. Big data and firm marketing performance: findings from knowledge-based view. Technol Forecast Soc Change. 2021;171:120986. doi:10.1016/j.techfore.2021.120986.
6. Zarifis A, Holland CP, Milne A. Evaluating the impact of AI on insurance: the four emerging AI- and data-driven business models. Emerald Open Res. 2023;1(1). doi:10.1108/EOR-01-2023-0001.
7. Rawat S, Rawat A, Kumar D, Sabitha AS. Application of machine learning and data visualization techniques for decision support in the insurance sector. Int J Inf Manag Data Insights. 2021;1(2):100012. doi:10.1016/j.jjimei.2021.100012.
8. Molloy L, Ronnie L. Sustaining the life insurance industry in the fourth industrial revolution. South Afr Actuarial J. 2020;20(1):81–107.
9. Zeier Roschmann A, Erny M, Wagner J. On the (future) role of on-demand insurance: market landscape, business model, and customer perception. Geneva Pap Risk Insur-Issues Pract. 2022;47:603–642. doi:10.1057/s41288-022-00265-7.
10. Ozdemir YE, Bayraklı S. A case study on building a cross-selling model through machine learning in the insurance industry. Avr Bilim Teknol Derg. 2022;35:364–372. doi:10.31590/ejosat.895069.
11. Khar MA, Irfan M. The impact of socioeconomic factors on consumer buying behavior: a case of mobile phone market of Pakistan. Indian J Econ Bus. 2021;20(4):1183–1193.
12. Ndoro H, Johnston K, Seymour LF. Artificial intelligence uses, benefits and challenges: a study in the Western Cape of South Africa financial services industry. Proceedings of SACAIR 2020; 2020. Cape Town, South Africa. p. 58.
13. Dutta S, Bhattacharya S. Cross selling of investment products and services: a case study of leading financial services organisation. Int J Bus Forecast Mark Intell. 2019;5(2):241–248. doi:10.1504/IJBFMI.2019.101608.
14. Boustani N, Emrouznejad A, Gholami R, Despic O, Ioannou A. Improving the predictive accuracy of the cross-selling of consumer loans using deep learning networks. Ann Oper Res. 2024;339:613–630. doi:10.1007/s10479-023-05209-5.
15. Laxmi KR, Srivastava S, Madhuravani K, Pallavi S, Dewangan O. Modified cross-sell model for telecom service providers using data mining techniques. In: Raja R, Nagwanshi K, Kumar S Laxmi K. editors. Data mining and machine learning applications. 2022. p. 195–207. doi:10.1002/9781119792529.ch8.
16. Lico T, Peres P, Li Y. A novel approach for cross-selling insurance products using positive unlabelled learning. Proceedings of the 2022 International Joint Conference on Neural Networks (IJCNN); 2022 Jul 18–23; Padua, Italy: IEEE, 2022. p. 1–8. doi:10.1109/IJCNN55064.2022.9892762.
17. Mavundla K, Thakur S. Analysing health insurance customer dataset to determine cross-selling potential. Proceedings of the International Conference on Artificial Intelligence and its Applications; 2023 Nov 9–10; Mahebourg, Mauritius: Preskil Island Resort, 2023. p. 219–226. doi:10.59200/ICARTI.2023.031.
18. Nayak B, Bhattacharyya SS, Krishnamoorthy B. Application of digital technologies in health insurance for social good of bottom of pyramid customers in India. Int J Sociol Soc Policy. 2019;39(9/10):752–772. doi:10.1108/IJSSP-05-2019-0095.
19. Jeanningros H, McFall L. The value of sharing: branding and behaviour in a life and health insurance company. Big Data Soc. 2020;7(2). doi:10.1177/2053951720950350.
20. Mensah K, Amenuvor FE. The influence of marketing communications strategy on consumer purchasing behaviour in the financial services industry in an emerging economy. J Financ Serv Mark. 2022;27:190–205. doi:10.1057/s41264-021-00121-0.

21. Dulhare UN, Ghori I. An efficient hybrid clustering to predict the risk of customer churn. Proceedings of the 2018 2nd International Conference on Inventive Systems and Control (ICISC); 2018 Jan 19–20; Coimbatore, India: IEEE, 2018. p. 673–677. doi:10.1109/ICISC.2018.8398883.
22. Olaniyi BY, Fernandez Del Rio A, Perianez A, Bellhouse L. User engagement in mobile health applications. Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining; 2022 Aug 14–18; (WA), (DC), USA, 2022. p. 4704–4712. doi:10.1145/3534678.354268.
23. Awit NT, Marticio RM. Customer churn prediction using predictive analytics: basis for the formulation of customer retention strategy in the context of web-based collaboration platform. Proceedings of the International Conference on Industrial Engineering and Operations Management; 2023 Mar 7–9; Manila, Philippines. 2023.
24. Lappeman J, Franco M, Warner V, Sierra-Rubia L. What social media sentiment tells us about why customers churn. J Consum Mark. 2022;39(5):385–403. doi:10.1108/JCM-12-2019-3540.
25. Bettiol M, Capestro M, Di Maria E, Micelli S. Reacting to the COVID-19 pandemic through digital connectivity with customers: the Italian experience. Ital J Mark. 202:305–330. doi:10.1007/s43039-021-00031-y.
26. Nkolele R, Wang H. Explainable machine learning: a manuscript on the customer churn in the telecommunications industry. Proceedings of the 2021 Ethics and Explainability for Responsible Data Science (EE-RDS); 2021 Oct 27–28; Johannesburg, South Africa: IEEE, 2021. p. 1–7. doi:10.1109/EE-RDS53766.2021.9708561.
27. Mukhamediev RI, Popova Y, Kuchin Y, Zaitseva E, Kalimoldayev A, Symagulov A, Levashenko V, Abdoldina F, Gopejenko V, Yakunin K, et al. Review of artificial intelligence and machine learning technologies: classification, restrictions, opportunities and challenges. Mathematics. 2022;10(15):2552. doi:10.3390/math10152552.
28. Kaddari Z, Mellah Y, Berrich J, Belkasmi MG, Bouchentouf T. Natural language processing: challenges and future directions. Proceedings of the International Conference on Artificial Intelligence & Industrial Applications; 2022 Apr 25–28; Montreal, QC, Canada; Cham: Springer International Publishing, 2020. p. 236–246. doi:10.1109/SysCon53536.2022.9773855.
29. Yee OS, Sagadevan S, Malim NHAH. Credit card fraud detection using machine learning as data mining technique. J Telecommun Electron Comput Eng. 2018;10(1–4):23–27.
30. Morales EF, Escalante HJ. A brief introduction to supervised, unsupervised, and reinforcement learning. In: Cingolani M. editor. Biosignal processing and classification using computational learning and intelligence. Academic Press; 2022. p. 111–129. doi:10.1016/B978-0-12-820125-1.00017-8.
31. Sen PC, Hajra M, Ghosh M. Supervised classification algorithms in machine learning: a survey and review. In: Mandal J Bhattacharya D. editors. Emerging technology in modelling and graphics. Advances in intelligent systems and computing. Singapore: Springer; 2020. p. 37–157. doi:10.1007/978-981-13-7403-6_11.
32. Dike HU, Zhou Y, Deveerasetty KK, Wu Q. Unsupervised learning based on artificial neural network: a review. Proceedings of the 2018 IEEE International Conference on Cyborg and Bionic Systems (CBS); 2018 Oct 25–27; Shenzhen, China: IEEE, 2018. p. 322–327. doi:10.1109/CBS.2018.8612259.
33. Hong T, Wang Z, Luo X, Zhang W. State-of-the-art on research and applications of machine learning in the building life cycle. Energy Build. 2020;212:109831. doi:10.1016/j.enbuild.2020.109831.
34. Kumar V, Garg ML. Predictive analytics: a review of trends and techniques. Int J Comput Appl. 2018;182(1):31–37.
35. Alt MA, Saplacan Z, Benedek B, Nagy BZ. Digital touchpoints and multichannel segmentation approach in the life insurance industry. Int J Retail Distrib Manag. 2021;49(5):652–677. doi:10.1108/IJRDM-02-2020-0040.
36. Amani FA, Fadlalla AM. Data mining applications in accounting: a review of the literature and organizing framework. Int J Acc Inf Syst. 2017;24:32–58. doi:10.1016/j.accinf.2016.12.004.
37. Riikkinen M, Saarijarvi H, Sarlin P, Lahteenmäki I. Using artificial intelligence to create value in insurance. Int J Bank Mark. 2018;36(6):1145–1168. doi:10.1108/IJBM-01-2017-0015.
38. Chowdhury S, Mayilvahanan P, Govindaraj R. Optimal feature extraction and classification-oriented medical insurance prediction model: machine learning integrated with the internet of things. Int J Comput Appl. 2020;44(3):278–290. doi:10.1080/1206212X.2020.1733307.
39. Bonache J, Festing M. Research paradigms in international human resource management: an epistemological systematisation of the field. Ger J Hum Resour Manag. 2020;34(2):99–123. doi:10.1177/23970022209097.
40. Speiser JL, Miller ME, Tooze J, Ip E. A comparison of random forest variable selection methods for classification prediction modeling. Expert Syst Appl. 2019;134:93–101. doi:10.1016/j.eswa.2019.05.028.
41. Subasi A, Cankurt S. Prediction of default payment of credit card clients using data mining techniques. Proceedings of 2019 International Engineering Conference; 2019 Jun 23–25; Erbil, Iraq: IEEE, 2019. p. 115–120. doi:10.1109/IEC47844.2019.8950597.
42. Lico L, Enesi I, Çiço B. Analyzing performance of clustering algorithms on a real retail dataset. Proceedings of the 2021 International Conference on Information Technologies (InfoTech); 2021 Sep 16–17; Varna, Bulgaria: IEEE, 2021. p. 1–6. doi:10.1109/InfoTech52438.2021.9548359.
43. Hanafy M, Ming R. Machine learning approaches for auto insurance big data. Risks. 2021;9(2):42. doi:10.3390/risks9020042.
44. Rusdah DA, Murfi H. Xgboost in handling missing values for life insurance risk prediction. SN Appl Sci. 2020;2:1336. doi:10.1007/s42452-020-3128-y.

45. Shahbazi Z, Byun YC. Product recommendation based on content-based filtering using XGBoost classifier. Int J Adv Sci Technol. 2019;29(4):6979–6988.
46. Li X, Li Z. A hybrid prediction model for e-commerce customer churn based on logistic regression and extreme gradient boosting algorithm. Ingenierie des Systemes Inf. 2019;24(5):525. doi:10.18280/isi.240510.
47. Helal S, Li J, Liu L, Ebrahimie E, Dawson S, Murray DJ, Long Q. Predicting academic performance by considering student heterogeneity. Knowl Based Syst. 2018;161:134–146. doi:10.1016/j.knosys.2018.07.042.
48. Sarker IH. Machine learning: algorithms, real-world applications and research directions. SN Comput Sci. 2021;2:160. doi:10.1007/s42979-021-00592-x.