2023 International Conference on Artificial Intelligence and its Applications
https://doi.org/10.59200/ICARTI.2023.031

Analysing Health Insurance Customer Dataset to Determine Cross-Selling Potential

Khulekani Mavundla and Surendra Thakur
Department of Information Technology
Durban University of Technology
Durban, South Africa
20907985@dut4life.ac.za, thakur@dut.ac.za

***Abstract*— Health insurance cross-selling refers to the practice of offering additional or complementary insurance products to existing policyholders. Insurance providers leverage cross-selling, offering customers additional policies like dental or life insurance when they already have a basic health insurance plan. This study is conducted to focus on the application of machine learning techniques to predict health insurance cross-selling opportunities among South African customers. The aim of this study is to develop a cross-selling predictive machine learning model that can assist health insurance companies to identify potential customers for cross-selling probabilities. To achieve this goal, a quantitative research methodology is adopted, focusing on extracting a comprehensive dataset of health insurance consumer information and employing various machine learning algorithms using the Python programming language, including Random Forest, K-Nearest Neighbours, XGBoost classifier, and Logistic Regression algorithms to build the cross-selling predictive machine learning model. The experimental results obtained demonstrate the accuracy scores of four different machine learning algorithms trained using 1,000,000 customer dataset with 17 features, logistic regression is considered as the top-performing model. It achieved an accuracy score of 0.83 and an F1 score of 0.91. The analysis indicates that customers aged 30-60, with prior insurance, and longer service history are more likely to buy additional health insurance products. The findings of this research can help health insurers boost revenue by improving their customer targeting and retention strategies.**

Keywords*— *Cross-Selling, Machine Learning algorithms, Health Insurance, Prediction, Model training

I. INTRODUCTION

Health insurance cross-selling is the practice of insurance companies offering additional insurance products to existing customers (policyholders), which plays a crucial role in the business growth and profitability of insurance companies [1]. 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.

By analysing various consumer attributes, such as demographics, purchasing history, and socio-economic factors, the model aims to identify patterns and indicators that contribute to health insurance cross-selling prediction. The goal is to assist health insurance companies in planning targeted marketing campaigns and personalized offers to maximize health insurance cross-selling success to their existing customers [1].

The importance of this research resides in its ability to strengthen the marketing strategies of health insurance companies for 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 [2].

Additionally, the findings of this study are a valuable resource for health insurance providers aiming to optimize their cross-selling predictions in the South African market. By exploring the factors influencing cross-selling in the health insurance sector, insurers can gain a deep understanding of customer behaviour patterns. With this knowledge, they can make data-driven decisions to improve marketing strategies, adapt product offerings, and create customized services that meet the specific needs and preferences of their target customers. This proactive approach not only boosts customer satisfaction but also cultivates stronger and enduring client relationships, ultimately contributing to the growth and sustainability of the health insurance industry.

In the realm of financial services, cross-selling has been explored in various sectors, such as banking and retail, as an effective strategy to increase revenue and enhance customer relationships. However, a research gap persists, particularly in the context of health insurance, especially within the South African market. This gap highlights the distinct challenges and opportunities within South Africa's health insurance industry. This study aims to bridge the gap by leveraging machine learning models to analyse historical customer data and predict cross-selling potential to provide valuable insights for optimizing cross-selling strategies within the South African health insurance sector. [3].

This research represents a unique opportunity to delve into an unexplored research area and enrich the existing body of knowledge. By bridging this research gap, the study aims to not only fill this void but also to deliver profound insights that can improve the way for more effective cross-selling practices in the health insurance industry.

The paper proceeds by first providing a review of related literature in health insurance cross-selling prediction. It then outlines the research methodology employed which includes data collection, data preprocessing, exploratory data analysis, and feature selection. The paper develops and evaluates four machine learning models to predict cross-selling potential. It then outlines the experimental results and concludes the paper by summarizing and suggesting future research direction.

II. LITERATURE REVIEW

This study builds upon existing research in the field of cross-selling prediction and machine learning applications in the insurance industry. By identifying customers who are more likely to purchase additional insurance products, insurers can 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 ahead in the industry [2].

Traditionally, cross-selling has relied on manual processes and subjective decision-making. However, with the arrival of machine learning and predictive analytics, insurers now have the opportunity to leverage data-driven approaches to identify potential cross-selling opportunities more effectively [4]. Applying machine learning algorithms enable the insurance company to analyse large volumes of health insurance customer datasets and uncover patterns and insights that may not be covered through traditional methods [5].

A. Predictive Modelling Techniques

Several studies have investigated the effectiveness of different predictive modelling techniques to identify potential cross-selling opportunities within the health insurance domain. A gradient-boosting algorithm is utilized to predict the likelihood of customers purchasing additional coverage based on their historical health claims and demographic information [6]. Their results demonstrated that ensemble methods such as gradient boosting outperform traditional logistic regression models in terms of predictive accuracy.

By exploring the application of machine learning techniques utilizing gradient bosting model [21]. Their studies demonstrated the effectiveness of this approach in identifying potential customers for additional insurance products based on historical data and customer attributes. Similarly, [22] employed random forests and neural networks to develop predictive models for cross-selling health insurance policies. Their comparative analysis highlighted the importance of feature engineering and model selection in achieving accurate predictions.

B. Imbalanced Data Handling

Imbalanced data is a common challenge in cross-selling prediction, where the number of customers who do not purchase additional coverage often outweighs those who do [23]. This issue addressed by applying various techniques, including oversampling the minority class and using different evaluation metrics to account for the class imbalance [5]. Their findings highlighted the importance of handling imbalanced data to prevent biased predictions.

C. Customer Segmentation

Segmenting customers based on their characteristics and behaviors can enhance the effectiveness of cross-selling prediction models [1]. Employed unsupervised clustering techniques to group customers with similar profiles can significantly improve the accuracy of cross-selling prediction models by identifying distinct customer segments, enabling more targeted marketing strategies and product recommendations[5]. They then built separate predictive models for each cluster, leading to more accurate predictions by capturing the distinct purchasing behaviors within different customer segments [24].

This study endeavours to fill critical research gaps in the realm of health insurance cross-selling within the South African context. While cross-selling has been studied in various industries, there is gap specifically tailored to the South African health insurance sector. By employing machine learning models, this research seeks to pioneer the application of advanced predictive analytics techniques for predicting cross-selling opportunities.

D. Customer Satisfaction

Customer satisfaction is a critical factor in cross-selling success, as satisfied customers are more likely to consider additional insurance products. In the study conducted by [25] sentiment analysis was applied to customer reviews and feedback to gauge customer satisfaction levels. Machine learning techniques were then used to identify patterns in customer sentiments and link them to cross-selling outcomes.

The field of cross-selling prediction in health insurance has seen significant advancements through the application of various machine learning techniques. Researchers have addressed challenges such as imbalanced data, customer segmentation, and customer satisfaction to enhance the accuracy and effectiveness of cross-selling strategies, ultimately contributing to the growth and profitability of health insurance companies.

This study is intended to contribute significantly to the existing knowledge base by not only addressing these research gaps but also by offering valuable insights and practical applications that can enhance cross-selling practices in the South African health insurance sector seeking to optimize their cross-selling efforts using machine learning algorithms.

III. Methodology

This study adopted quantitative research because it employs a deductive approach to uncover patterns in human existence by separating the social realm into measurable elements known as variables that can be quantified numerically [7].

A. Quantitative Methodology

The quantitative research method was applied to focus on investigating the answers to the questions starting with how many, how much, and to what extent [7]. For this study, health insurance data was extracted from an open-source database, focusing on customer behaviour that can be quantified and patterned to interpret their meanings to create data insights.

Quantitative data is the value of data in the form of counts or numerals where each dataset has unique numerical data [8]. This data was statistically analysed to establish the conclusive results of health insurance cross-selling prediction. Raw quantitative data was collected and interpreted using statistical analysis to determine if the existing health insurance customer will be interested in purchasing additional health insurance products.

B. Data Collection and Processing

For this research, data collection involved gathering relevant information about policyholders and insured members to create a comprehensive dataset. The researcher was guided by the Machine Learning Lifecycle to build a Health Insurance Cross-Selling Prediction model.

1) Machine Learning (ML) Lifecycle (8 stages)

The ML lifecycle encompasses various stages that provide structure to handle data and build a sustainable machine learning predictive model. The importance of noting that the specific techniques and algorithms used within each stage may vary depending on the nature of the health insurance dataset, the available variables, and the objectives of the ML project, and the ML lifecycle is iterative, meaning that the stages are not always strictly sequential [9]. It often involves iterating through stages multiple times, refining the model, and improving its performance based on feedback and new insights gained during the process [9].

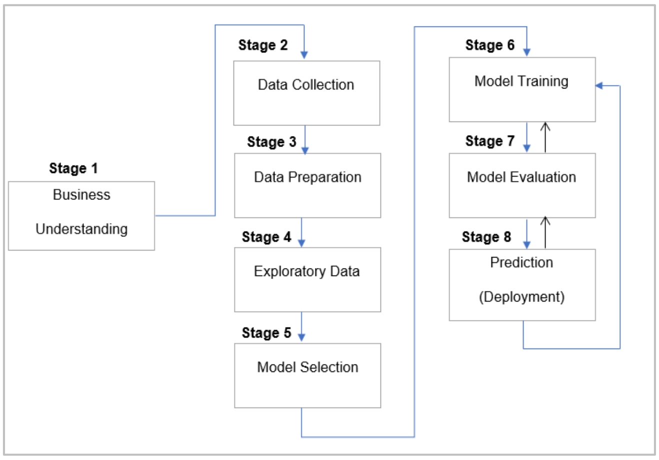


Fig. 1. Graphically Depicts the ML Lifecycle (Source: Researcher)

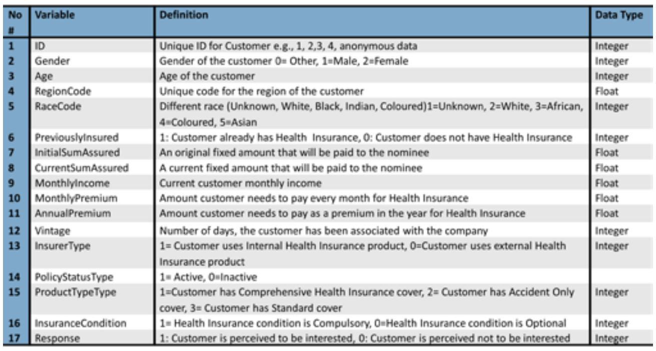
2) Health Insurance Data Source

The health insurance data collected from the large insurance company database contains information related to insurance policies, members, claims, and other relevant variables.

3) Understanding Health Insurance Data Frame (Dataset)

Table 1 illustrates the features that represent the individual customers, descriptions, and data type.

TABLE I. HEATH INSURANCE DATASET



No #	Variable	Definition	Data Type
1	ID	Unique ID for Customer e.g., 1, 2,3, 4, anonymous data	Integer
2	Gender	Gender of the customer 0= Other, 1=Male, 2=Female	Integer
3	Age	Age of the customer	Integer
4	RegionCode	Unique code for the region of the customer	Float
5	RaceCode	Different race (Unknown, White, Black, Indian, Coloured)1=Unknown, 2=White, 3=African, 4=Coloured, 5=Asian	Integer
6	PreviouslyInsured	1: Customer already has Health Insurance, 0: Customer does not have Health Insurance	Integer
7	InitialSumAssured	An original fixed amount that will be paid to the nominee	Float
8	CurrentSumAssured	A current fixed amount that will be paid to the nominee	Float
9	MonthlyIncome	Current customer monthly income	Float
10	MonthlyPremium	Amount customer needs to pay every month for Health Insurance	Float
11	AnnualPremium	Amount customer needs to pay as a premium in the year for Health Insurance	Float
12	Vintage	Number of days, the customer has been associated with the company	Integer
13	InsurerType	1= Customer uses Internal Health Insurance product, 0=Customer uses external Health Insurance product	Integer
14	PolicyStatusType	1= Active, 0=Inactive	
15	ProductTypeType	1=Customer has Comprehensive Health Insurance cover, 2= Customer has Accident Only cover, 3= Customer has Standard cover	Integer
16	InsuranceCondition	1= Health Insurance condition is Compulsory, 0=Health Insurance condition is Optional	Integer
17	Response	1: Customer is perceived to be interested, 0: Customer is perceived not to be interested	Integer

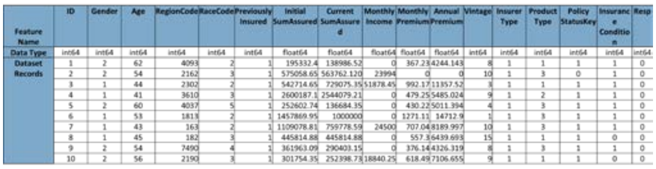
4) Data Collection Method

The dataset was collected from the large database by writing an SQL query to select health insurance customer records from various database tables into one table created in the staging database and selecting the top 1,000,000 records. The extracted dataset contains 1,000,000 records, with 17 features, reflecting diverse individuals and their respective insurance policy information. The dataset was anonymized to comply with privacy regulations (POPIA) and ensure the security and confidentiality of customer information.

5) Jupiter Lab and Python Programming

The extracted dataset was imported using the Python Pandas library in Jupiter Lab to train a model. The extracted dataset was viewed by selecting the top 10 records to ensure that the fields and features were imported successfully. Table 2 illustrates the dataset before the cleaning process commenced.

TABLE II. TRAINING DATASET BEOFRE CLEANING PROCESS

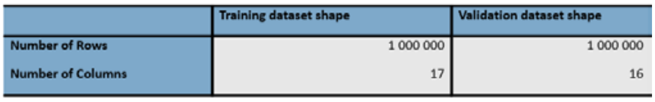


Feature Name	ID	Gender	Age	RegionCode	RaceCode	Previously Insured	Initial SumAssured	Current SumAssured	Monthly Income	Monthly Premium	Annual Premium	Vintage	Insurer Type	Product Type	Policy StatusKey	Insurance Condition	Resp
Data Type	int64	int64	int64	int64	int64	int64	float64	float64	float64	float64	float64	int64	int64	int64	int64	int64	int64
Dataset Records	1	2	62	4093	2	1	195332.4	138986.52	0	367.23	4244.143	8	1	1	1	1	0
	2	2	54	2162	3	1	575058.65	563762.120	23994	0	0	10	1	3	0	1	0
	3	1	44	2302	2	1	542714.65	729075.35	51878.45	992.17	11357.52	3	1	1	1	1	0
	4	1	41	3610	3	1	2600187.1	2544079.21	0	479.25	5485.024	9	1	2	1	1	0
	5	2	60	4037	5	1	252602.74	136684.35	0	430.22	5011.394	4	1	3	1	1	0
	6	1	53	1813	2	1	1457869.95	1000000	0	1271.11	14712.9	1	1	3	1	1	0
	7	1	43	163	2	1	1109078.81	759778.59	24500	707.04	8189.997	10	1	3	1	1	0
	8	1	45	182	3	1	445814.88	445814.88	0	557.3	6439.693	15	1	1	1	0	0
	9	2	54	7490	4	1	361963.09	290403.15	0	376.14	4326.319	8	1	3	1	1	0
	10	2	56	2190	3	1	301754.35	252398.73	18840.25	618.49	7106.655	9	1	1	1	0	0

The dataset shape refers to the dimensions or structure of the dataset, specifically the number of rows (instances) and columns (features) it contains.

Table 3 shows this study's training and validation dataset shape.

TABLE III. TRAINING AND VALIDATION DATASET SHAPE



	Training dataset shape	Validation dataset shape
Number of Rows	1 000 000	1 000 000
Number of Columns	17	16

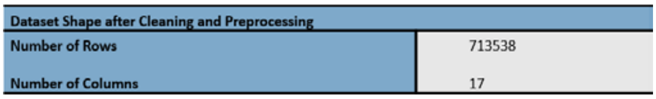
6) Health Insurance Data Preprocessing

Data preprocessing is a critical step in the data analysis and machine learning pipeline that involves cleaning, transforming, and preparing raw data to make it suitable for further analysis or modelling [10]. In the process, the Python code was used to remove duplicates, correct inaccuracies, and address data entry errors.

All missing values were handled by applying the techniques used including imputation (replacing missing values with estimated ones) or removing rows/columns with missing data. The extracted dataset was split into training and testing subsets to evaluate the model's performance effectively.

After performing the data preprocessing step, the selected health insurance dataset records were eliminated and dropped, meaning that the dataset shape has changed and is illustrated in Table 4.

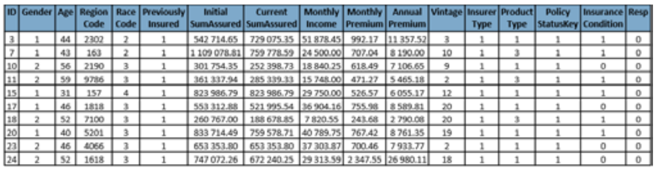
TABLE IV. TRAINING DATASET SHAPE



Dataset Shape after Cleaning and Preprocessing	
Number of Rows	713538
Number of Columns	17

For the training dataset, the top 10 records were selected after the data cleaning and pre-processing process. The following table 3 illustrates that the data was now cleaned and ready for model training.

TABLE V. TRAINING DATASET AFTER CLEANING PROCESS



ID	Gender	Age	Region Code	Race Code	Previously Insured	Initial SumAssured	Current SumAssured	Monthly Income	Monthly Premium	Annual Premium	Vintage	Insurer Type	Product Type	Policy StatusKey	Insurance Condition	Resp
3	1	44	2302	2	1	542 714.65	729 075.35	51 878.45	992.17	11 357.52	3	1	1	1	1	0
7	1	43	163	2	1	1 109 078.81	759 778.59	24 500.00	707.04	8 190.00	10	1	3	1	1	0
10	2	56	2190	3	1	301 754.35	252 398.73	18 840.25	618.49	7 106.65	9	1	1	1	0	0
11	2	59	9786	3	1	361 337.94	285 339.33	15 748.00	471.27	5 465.18	2	1	3	1	1	0
15	1	31	157	4	1	823 986.79	823 986.79	29 750.00	526.57	6 055.17	12	1	1	1	1	0
17	1	46	1818	3	1	553 312.88	521 995.54	36 904.16	755.98	8 589.81	20	1	1	1	0	0
18	2	52	7100	3	1	260 767.00	188 678.85	7 820.55	243.68	2 790.08	20	1	3	1	1	0
20	1	40	5201	3	1	833 714.49	759 578.71	40 789.75	767.42	8 761.35	19	1	1	1	1	0
23	2	46	4066	3	1	653 353.80	653 353.80	37 303.87	700.46	7 933.77	2	1	1	1	0	0
24	2	52	1618	3	1	747 072.26	672 240.25	29 313.59	2 347.55	26 980.11	18	1	1	1	0	0

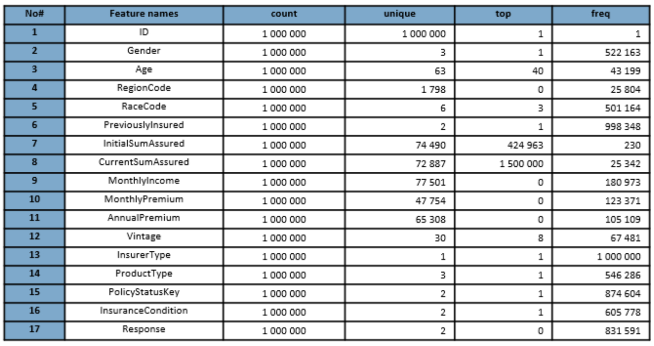
C. Exploratory Data Analysis (EDA)

The EDA process, which forms part of the data pre-processing process, was followed [11] to emphasize the importance of data exploration in thoroughly understanding and analysing the extracted dataset. During EDA, the researcher explored the health insurance dataset extracted from the Large Insurance Company database. This involved assessing factors like the dataset's size, column count, and variable data types, as well as conducting descriptive analyses, examining distributions, and correlations, and creating data visualizations. EDA is a crucial step in the data analysis process that involves exploring and summarizing the main characteristics of a dataset to gain insights and better understand its underlying patterns, trends, and potential issues.

1) Summary Statistics

Summary statistics refer to computing basic statistical measures, such as count, unique, top, and frequent values for each variable (Table 6).

TABLE VI. SUMMARY STATISTICS



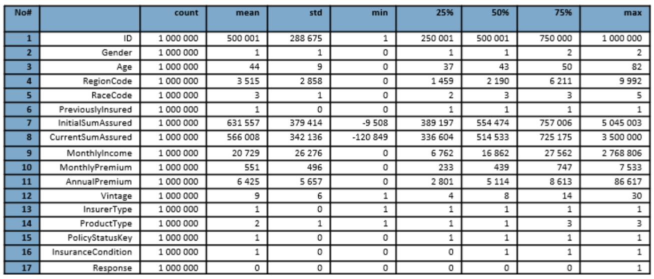
No#	Feature names	count	unique	top	freq
1	ID	1 000 000	1 000 000	1	1
2	Gender	1 000 000	3	1	522 163
3	Age	1 000 000	63	40	43 199
4	RegionCode	1 000 000	1 798	0	25 804
5	RaceCode	1 000 000	6	3	501 164
6	PreviouslyInsured	1 000 000	2	1	998 348
7	InitialSumAssured	1 000 000	74 490	424 963	230
8	CurrentSumAssured	1 000 000	72 887	1 500 000	25 342
9	MonthlyIncome	1 000 000	77 501	0	180 973
10	MonthlyPremium	1 000 000	47 754	0	123 371
11	AnnualPremium	1 000 000	65 308	0	105 109
12	Vintage	1 000 000	30	8	67 481
13	InsurerType	1 000 000	1	1	1 000 000
14	ProductType	1 000 000	3	1	546 286
15	PolicyStatusKey	1 000 000	2	1	874 604
16	InsuranceCondition	1 000 000	2	1	605 778
17	Response	1 000 000	2	0	831 591

2) Descriptive Statistics and Distribution

Descriptive statistics and distribution are important components of Exploratory Data Analysis (EDA) that help in understanding the main characteristics of a dataset and how its values are distributed across various variables.

Table 7 shows the descriptive statistics and distribution of the health insurance records selected.

TABLE VII. DESCRITIVE STATISTICS AND DISTRIBUTION



No#		count	mean	std	min	25%	50%	75%	max
1	ID	1 000 000	500 001	288 675	1	250 001	500 001	750 000	1 000 000
2	Gender	1 000 000	1	1	0	1	1	2	2
3	Age	1 000 000	44	9	0	37	43	50	82
4	RegionCode	1 000 000	3 515	2 858	0	1 459	2 190	6 211	9 992
5	RaceCode	1 000 000	3	1	0	2	3	3	5
6	PreviouslyInsured	1 000 000	1	0	0	1	1	1	1
7	InitialSumAssured	1 000 000	631 557	379 414	-9 508	389 197	554 474	757 006	5 045 003
8	CurrentSumAssured	1 000 000	566 008	342 136	-120 849	336 604	514 533	725 175	3 500 000
9	MonthlyIncome	1 000 000	20 729	26 276	0	6 762	16 862	27 562	2 768 806
10	MonthlyPremium	1 000 000	551	496	0	233	439	747	7 533
11	AnnualPremium	1 000 000	6 425	5 657	0	2 801	5 114	8 613	86 617
12	Vintage	1 000 000	9	6	1	4	8	14	30
13	InsurerType	1 000 000	1	0	1	1	1	1	1
14	ProductType	1 000 000	2	1	1	1	1	3	3
15	PolicyStatusKey	1 000 000	1	0	0	1	1	1	1
16	InsuranceCondition	1 000 000	1	0	0	0	1	1	1
17	Response	1 000 000	0	0	0	0	0	0	1

3) Correlation

The correlation refers to the statistical relationship between two or more variables [12]. It measures the strength and direction of the linear association between variables, indicating how changes in one variable are related to changes in another variable. In a correlation process, statistical measurements are made to describe how much two variables are linearly related to one another and to determine whether two variables are correlated when both move in the same direction.

The following figure shows the Pearson correlation of 17 features on how data were correlated in this study.

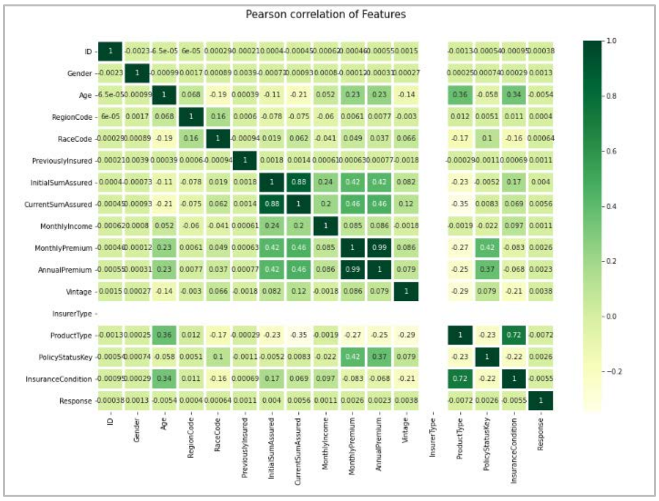


Fig. 2. Pearson correlation of 17 features (Source: Researcher)

4) Data Visualization

Data visualization is the graphical representation of data and information using visual elements such as charts, graphs, maps, and infographics [13]. Data is presented in graphically using features that affect the target variable. In this study, the data has been presented through different visual elements to ensure that the extracted health insurance dataset is distributed correctly. This process helped the researcher to explore the data, uncover patterns, identify outliers, and communicate the study's findings effectively.

The following figure illustrates gender count, which includes *Other, Male, and Female* in a bar graph and pie chart depicted in percentage. The Gender attribute has been defined numerically as follows: 0 = Other, 1 = Male, and 2 = Female.

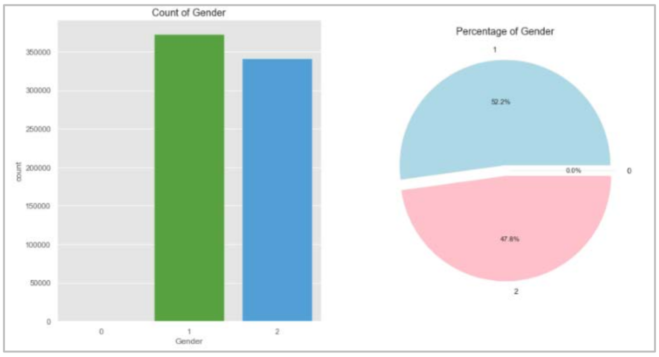


Fig. 3. Gender Counts (Source: Researcher)

The above bar graph shows that the dataset contained more records where the Gender Count is male than female.

The following figure presents the age versus response curve graph, illustrating that customers between the ages of 25 and 60 are more likely to be interested in purchasing additional health insurance products compared to younger than 30-year-old customers.

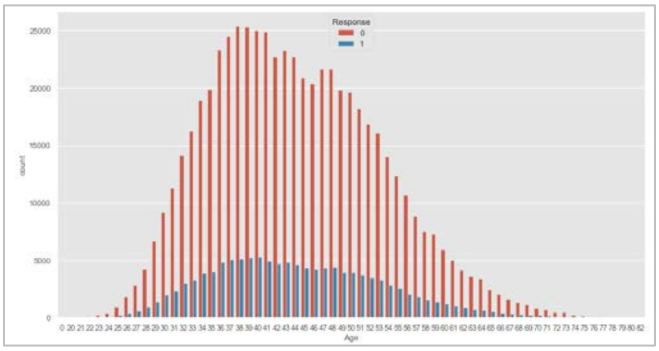


Fig. 4. Age vs. Response (Source: Researcher)

The following figure shows the vintage response, which is the number of days that customers have been associated with the company.

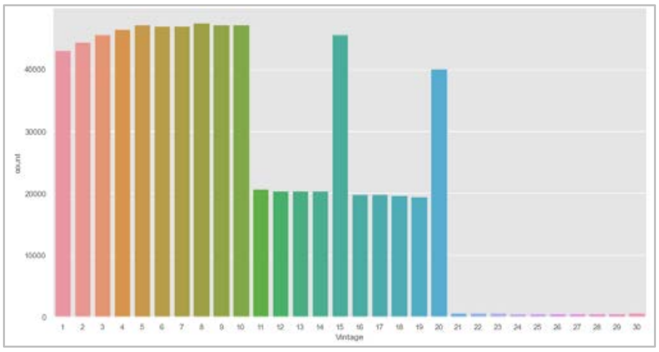


Fig. 5. Vintage Response (Source: Researcher)

The following figure shows the percentage of previously insured versus not insured customers. The analysed data shows that previously insured customers are more likely to respond to additional health insurance products compared to customers who have not been previously insured.

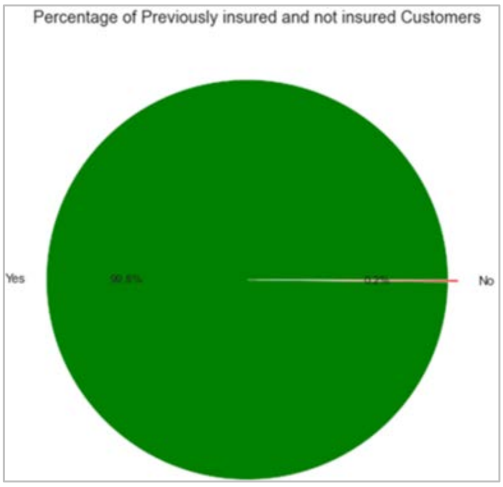


Fig. 6. Percentage of Previouslely Insured vs. Not (Source: Researcher)

D. Feature Engineering

Feature engineering is the process of creating new or modifying existing features (variables) from the raw data to improve the performance of machine learning models and data analysis tasks [14]. For feature engineering in this study, the Correlation Analysis method was used. Correlation analysis is a statistical technique used to quantify the degree of association or relationship between two or more variables [14]. It helps to understand how changes in one variable relate to changes in another. For feature selection, the following steps have been applied using Python code:

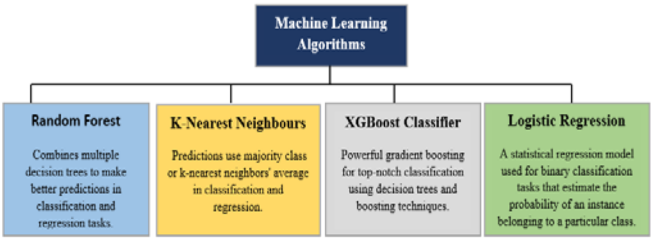
- Calculate the correlation matrix between all features and the target variable.
- Filter the correlation values related to the target variable.
- Sort the features based on their correlation with the target variable (in descending order).
- View the features and their correlation values.

E. Model Selection

Model selection is the process of choosing the most appropriate machine learning algorithm or model for a specific task or dataset [15].

Model selection involves evaluating different models, comparing their performance, and selecting the one that best fits the data and yields the most accurate predictions.

The following shows the machine learning algorithms selected, trained, and evaluated to build a cross-selling predictive model.


Machine Learning Algorithms

Random Forest

Combines multiple decision trees to make better predictions in classification and regression tasks.

K-Nearest Neighbours

Predictions use majority class or k-nearest neighbors' average in classification and regression.

XGBoost Classifier

Powerful gradient boosting for top-notch classification using decision trees and boosting techniques.

Logistic Regression

A statistical regression model used for binary classification tasks that estimate the probability of an instance belonging to a particular class.

Fig. 7. Machine Learning Algorithms (Source: Researcher)

F. Model Training and Evaluation Model Training

Machine learning model training refers to the process of teaching machine learning algorithms to recognize patterns and make predictions based on input data [16]. Model evaluation metrics are used to assess the performance of an ML model and measure how well it generalizes to unseen data, and the choice of evaluation metrics depends on the specific task and the nature of the problem being solved [17].

Below are some commonly used evaluation metrics for different types of ML models which have been outlined and provided examples for each metric.

1) Confusion Matrix

A confusion matrix is a table used to evaluate the performance of a classification model. It is particularly useful for binary classification problems where there are only two classes (Predicted Positive and Predicted Negative).

TABLE VIII. CONFUSION MATRIX TABLE

Actual Negative	True Negative(TN)	False Positive (FP)
Actual Positive	False Negative (FN)	True Positive (TP)

2) Model Evaluation Metrics

Model evaluation metrics are used to assess the machine learning model's performance and measure. The following equations were used [17][18]:

$$Precision = \frac{True\ Posivitives}{True\ Posivitives + False\ Positives}$$

$$Recall = \frac{True\ Positives}{True\ Positives + False\ Negatives}$$

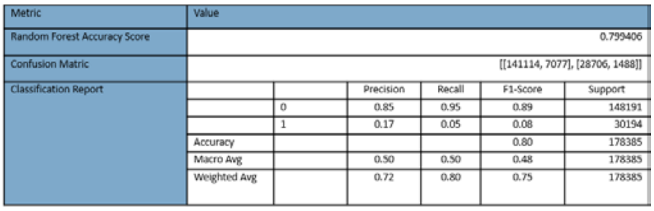
$$F1\ Score = \frac{2\ x\ Precision\ x\ Recall}{Precision + Recall}$$

IV. EXPERIMENTS, RESULTS, AND DISCUSSION

The researcher conducted an experiment to predict cross-selling opportunities for health insurance products using machine learning techniques. The goal was to develop a predictive model that could identify customers who are more likely to purchase additional coverage based on their historical behaviours, demographics, and engagement with the insurance provider.

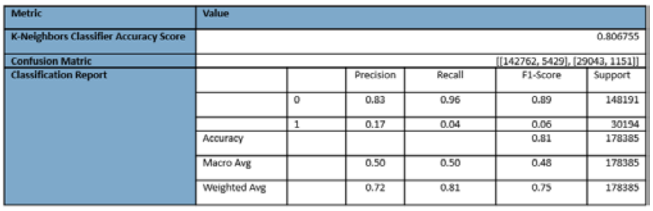
The results obtained from each machine learning model are as follows:

TABLE IX. RANDOM FOREST



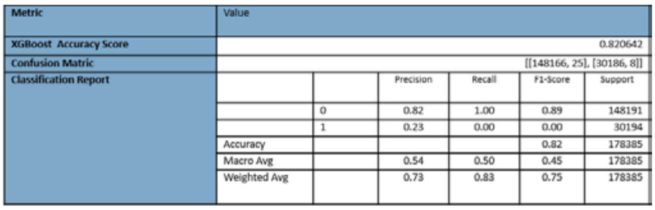
Metric	Value					
Random Forest Accuracy Score	0.799406					
Confusion Matric	[[141114, 7077], [28706, 1488]]					
Classification Report			Precision	Recall	F1-Score	Support
		0	0.85	0.95	0.89	148191
		1	0.17	0.05	0.08	30194
	Accuracy				0.80	178385
	Macro Avg		0.50	0.50	0.48	178385
	Weighted Avg		0.72	0.80	0.75	178385

TABLE X. K-NEAREST NEIGHBOUR (KNN)



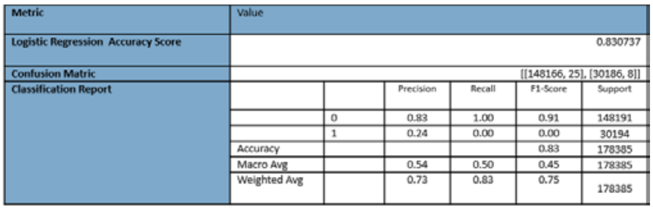
Metric	Value					
K-Neighbors Classifier Accuracy Score	0.806755					
Confusion Matric	[[142762, 5429], [29043, 1151]]					
Classification Report			Precision	Recall	F1-Score	Support
		0	0.83	0.96	0.89	148191
		1	0.17	0.04	0.06	30194
	Accuracy				0.81	178385
	Macro Avg		0.50	0.50	0.48	178385
	Weighted Avg		0.72	0.81	0.75	178385

TABLE XI. XG BOOST CLASSIFIER



Metric	Value					
XGBoost Accuracy Score	0.820642					
Confusion Matric	[[148166, 25], [30186, 8]]					
Classification Report			Precision	Recall	F1-Score	Support
		0	0.82	1.00	0.89	148191
		1	0.23	0.00	0.00	30194
	Accuracy				0.82	178385
	Macro Avg		0.54	0.50	0.45	178385
	Weighted Avg		0.73	0.83	0.75	178385

TABLE XII. LOGISTIC REGRESSION



Metric	Value					
Logistic Regression Accuracy Score	0.830737					
Confusion Matric	[[148166, 25], [30186, 8]]					
Classification Report			Precision	Recall	F1-Score	Support
		0	0.83	1.00	0.91	148191
		1	0.24	0.00	0.00	30194
	Accuracy				0.83	178385
	Macro Avg		0.54	0.50	0.45	178385
	Weighted Avg		0.73	0.83	0.75	178385

Among the algorithms, Logistic Regression showed the best overall performance, achieving an accuracy of 0.83 and an F1 score of 0.91. Thus, the logistic regression model is chosen to be the best for building a Health Insurance Cross-Selling Prediction model.

The following graphically depicts the comparison of the machine learning algorithms.

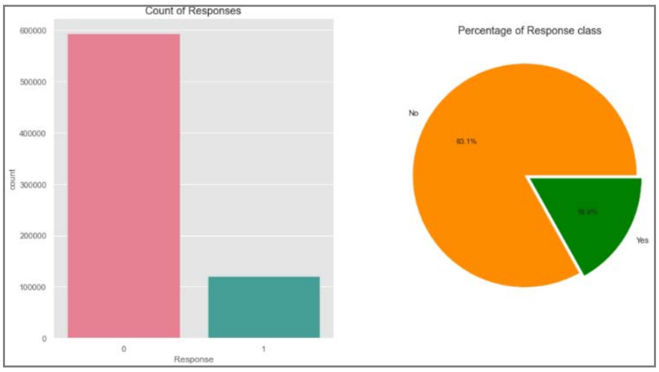


Fig. 8. Machine Learning Algorithms Comparison(Source: Researcher)

TABLE XIII. MACHINE LEARNING ALGORITHMS PERFORMANCE

No	Machine Learning Algorithms		
	Model Name	***Accuracy Score***	***F1-Score***
1	Random Forest	0.799406	0.89
2	K-Nearest Neighbours	0.806755	0.89
3	XGBoost classifier	0.820642	0.89
4	Logistic Regression	0.830737	0.91

Interpretation of the F1-Scores for the above four models in table 14, Among the four models evaluated, the Logistic Regression model stands out with the highest F1-Score of 0.91, indicating its superior balance between precision and recall. This implies that it excels in correctly identifying positive instances while keeping false positives and false negatives to a minimum. In contrast, the Random Forest, K-Nearest Neighbours (KNN), and XGBoost Classifier all exhibit an F1-Score of 0.89, demonstrating similar performance in terms of accurately classifying positive instances and maintaining a balance between precision and recall. While these models are effective, the Logistic Regression model appears to offer a slightly better overall classification performance.

A. Equations

For the Logistic Regression, the following equations were used [19]: Using Logistic Regression with two variables where one is the dependent variable (Y) and the other is the independent variable (X). Using Equation 1 to describe the relationship between X and Y where Y is binary.

$$h\theta(X) = \frac{1}{1+e^{-(\theta_0+\theta_1 X_1+\theta_2 X_2+\ldots+\theta_n X_n)}} \quad (1)$$

Logistic equation $h\theta(X)$ predicts the probability of a binary outcome (0 and 1) being based on the input variables $X_1, X_2, \ldots X_n$ and associated coefficients $\theta_0, \theta_1, \theta_{2,\ldots}\theta_n$. This equation uses the logistic (sigmoid) function to model the relationship between the predictor variables and the probability of the binary outcome being 1.

Equation 2, called the sigmoid function known as a logistic function, is a mathematical function that maps any real-valued number to a value between 0 and 1. It is an essential component of logistic regression and other ML algorithms used for binary classification.

$$\sigma(z) = \frac{1}{1+e^{-z}} \quad (2)$$

The sigmoid function is used to transform the log odds (z) into a cross-selling probability value between 0 and 1.

Equation 3 is the linear combination of predictor variables, used as an input to the logistic function.

$$f(x) = \beta_0+\beta_1 X_1 + +\beta_2 X_2 + \cdots +\beta_r X_r \quad (3)$$

β_0 is the intercept term.

$\beta_1, \beta_2, ..., \beta_r$ are the coefficients of the predictor variables $x_1, x_2, ..., x_r$.

The four distinct machine learning models—Random Forest, K-Nearest Neighbours, XGBoost Classifier, and Logistic Regression—were trained and evaluated to assess their performance and select the most suitable model for building a predictive ML model [20]. The results of the health insurance cross-selling prediction utilizing machine learning revealed that after a comprehensive evaluation of four distinct machine learning models, logistic regression emerged as the top-performing model. It achieved an impressive accuracy rate of 0.83 and an F1 score of 0.91, solidifying its position as the optimal choice for predicting health insurance cross-selling.

The analysis of the health insurance dataset delved into the factors influencing customer behaviors. It was evident that individuals aged between 30 and 60 are more inclined to purchase additional health insurance products, indicating the significance of age as a predictive variable. The duration of a customer's association with the company, known as "Vintage", emerged as a pivotal factor, with higher Vintage values correlating with a high probability of purchasing additional health insurance. The results revealed that the customers with a history of prior insurance coverage demonstrated that they might be interested in purchasing additional health insurance products, highlighting the value of historical data in the prediction model.

The study also emphasized the key features that significantly impact the prediction model, including *Gender, Age, Previously Insured status, Monthly Income, Monthly Premium, and Annual Premium.*

V. CONCLUSION

In conclusion, research highlights the remarkable potential of machine learning algorithms to analyse health insurance datasets to determine cross-selling probabilities. The availability of vast customer data in this industry presents a unique opportunity, one that machine learning models are exceptionally well-suited to exploit. With massive amounts of customer data available in the health insurance sector, machine learning models analyse the data to identify patterns, behaviours, and preferences, thereby help identifying potential customers who may be interested in purchasing additional insurance products. By leveraging predictive machine learning models for cross-selling in health insurance, these algorithms accurately predict the likelihood of a customer's interest in purchasing additional products, enabling insurance companies to personalize their cross-selling efforts.

As the industry evolves, the adoption of machine learning algorithms becomes increasingly vital, promising to drive innovation, efficiency, and customer satisfaction. In essence, the utilization of machine learning in analysing health insurance datasets for cross-selling probabilities is not just a choice but a strategic imperative that ensures the industry remains competitive, adaptive, and capable of meeting the evolving needs of its customer, thereby optimizing and increasing business revenue. Future research could focus on incorporating the health insurance cross-selling prediction findings.

REFERENCES

[1] A. G. Sekeroglu, "Impacts of Feature Selection Techniques in Machine Learning Algorithms for Cross Selling: A Comprehensive Study for Insurance Industry," [Online]. Available: https://www.researchgate.net/profile/Ali-Galip-Sekeroglu/publication/353072980_Impacts-of-Feature-Selection-Techniques-in-Machine-Learning-Algorithms-for-Cross-Selling-A-Comprehensive-Study-for-Insurance-Industry/links/60e6cb851c28af345851e1c7/Impacts-of-Feature-Selection-Techniques-in-Machine-Learning-Algorithms-for-Cross-Selling-A-Comprehensive-Study-for-Insurance-Industry.pdf. [Accessed: August 15, 2023].

[2] Y.E. Ozdemir, S. Bayrakli, "A Case Study on Building a Cross-Selling Model through Machine Learning in the Insurance Industry", vol. 35, pp.364-372. May 2022.

[3] W. Qadadeh, S. Abdallah, "Customer Segmentation in the Insurance Company (TIC) Dataset," INNS Conference on Big Data and Deep Learning, vol. 144, University of Edinburgh, Edinburgh, UK, 2018, pp. 277-290.

[4] N. Kumar, J.D. Srivastava, H. Bisht."Artificial Intelligence in Insurance Sector," [Online]. Available: https://www.researchgate.net/profile/Naman-Kumar-3/publication/337305024_Artificial_Intelligence-in-Insurance-Sector/links/5dd00e33a6fdcc7e138761cc/Artificial-Intelligence-in-Insurance-Sector.pdf. [Accessed: August 18, 2023].

[5] T. Sidorowicz, P. Peres, Y. Li, A, "Novel Approach for Cross-Selling Insurance Products Using Positive Unlabelled Learning," 2022 International Joint Conference on Neural Networks (IJCNN), pp. 1-8.IEEE Xplore.

[6] K.P.M.L.P. Weerasinghe, and M.C. Wijegunasekara, "A comparative study of data mining algorithms in the prediction of auto insurance claims," European International Journal of Science and Technology, vol. 5, pp.47-54.

[7] M.S. Rahman, "The Advantages and Disadvantages of Using Qualitative and Quantitative Approaches and Methods in Language "Testing and Assessment" Research," Journal of Education and Learning, vol.6, University of Plymouth, pp.102-102. October 2016.

[8] M.J. Goertzen, "Introduction to Quantitative Research and Data", Library Technology Reports, vol. 53, pp.12-18. May 2017.

[9] F. Ritz, T. Phan, A. Sedlmeier, P. Altmann, J. Wieghardt, R. Schmid, H. Sauer, C. Klein, C. Linnhoff-Popien, and T. Gabor, "Dependencies within Machine Learning via a Formal Process Model," In International Symposium on Leveraging Applications of Formal Methods, pp.249-265. Cham: Springer Nature Switzerland. October 2022.

[10] P. Misra, and A.Yadav, "Impact of Preprocessing Methods on Healthcare Predictions," [Oline]. Available: https://www.researchgate.net/profile/Puneet-Misra-3/publication/332436103_Impact-of-Preprocessing-Methods-onHealthcarePredictions/links/5d37ac2192851cd04680da45/Impact-of-Preprocessing-Methods-on-Healthcare-Predictions.pdf. [Accessed: August 19, 2023].

[11] N. Boodhun, and M. Jayabalan. "Risk prediction in life insurance industry using supervised learning algorithms," Complex & Intelligent Systems, vol. 4, pp.145-154.

[12] P. Mikalef, J. Krogstie, I.O. Pappas, P. Pavlou, "Exploring the relationship between big data analytics capability and competitive performance," The mediating roles of dynamic and operational capabilities. Information & Management, vol.57.pp-103169. March 2020.

[13] Q. Wang, Z. Chen, Y. Wang, H. Qu, "Applying Machine Learning Advances to Data Visualization: A Survey on ML4VIS," [Online]. Available: https://www.researchgate.net/profile/Yong-Wang-149/publication/346555391_Applying-Machine-Learning-Advances-to-Data-Visualization-A-Survey-on-ML4VIS/links/603cd29e92851c4ed5a5590d/Applying-Machine-Learning-Advances-to-Data-Visualization-A-Survey-on-ML4VIS.pdf. [Accessed: August 19, 2023].

[14] A. G. Sekeroglu, "Impacts of Feature Selection Techniques in Machine Learning Algorithms for Cross Selling: A Comprehensive Study for Insurance Industry," [Online]. Available: https://www.researchgate.net/profile/Ali-Galip-Sekeroglu/publication/353072980_Impacts-of-Feature-Selection-Techniques-in-Machine-Learning-Algorithms-for-Cross-Selling-A-Comprehensive-Study-for-Insurance-Industry/links/60e6cb851c28af345851e1c7/Impacts-of-Feature-Selection-Techniques-in-Machine-Learning-Algorithms-for-Cross-Selling-A-Comprehensive-Study-for-Insurance-Industry.pdf. [Accessed: October 19, 2023].

[15] P. Anitha, M.M. Patil (2019). “RFM Model for Customer Purchase Behaviour Using K-Means Algorithm,” Journal of King Saud University-Computer and Information Sciences, vol. 34, pp.1785-1792. May 2022.

[16] M.K. Severino, and Y. Peng, “Machine learning algorithms for fraud prediction in property insurance: Empirical evidence using real-world microdata,”. Machine Learning with Applications, vol. 5, p.100074.

[17] Z. Vujovic, “Classification model evaluation metrics,” International Journal of Advanced Computer Science and Applications, vol.12, pp.599-606.

[18] M. Grandini, E. Bagli, G. Visani, “Metrics for multi-class classification” an overview. arXiv preprint arXiv: 2008.05756. August 2020.

[19] E.Y. Boateng, and D.A. Abaye, “A review of the logistic regression model with emphasis on medical research,” Journal of data analysis and information processing, vol.7, pp.190-207.

[20] M.T. Akter, M. Begum, and R. Mustafa, “Bengali sentiment analysis of e-commerce product reviews using k-nearest neighbors,” In 2021 International conference on information and communication technology for sustainable development (ICICT4SD), pp. 40-44. IEEE, 2021.

[21] N. Dheib, H. Ghazzai, H. Besbes, Y. Massoud, “Extreme Gradient Boosting Machine Learning Algorithm For Safe Auto Insurance Operations,” 2019 IEEE International Conference on Vehicular Electronics and Safety (ICVES), pp. 04-06. IEEE, 2019.

[22] S. Vandrangi, “Predicting the Insurance Claim by each user using Machine Learning Algorithms,” Journal of Emerging Strategies in New Economics, vol.1, pp.1-11. October 2022.

[23] M. Hanafy, R. Ming, “Using Machine Learning Models to Compare Vrious Resampling Methods in Predicting Insurance Fraud,” Journal of Theoretical and Applied Information Technology, vol.99, pp.12-23. June 2021.

[24] Y. Altun, A. Yucekaya, “A Probabilistic Approach to Maximize Cross-Selling Revenues of Financial Products,” American Scientific Research Journal for Engineering, Technology, and Sciences (ASRJETS), vol.79, pp.1-14. Industrial Engineering Department, Kadir Has University, Istanbul, Turkey. 2021.

[25] C. Eckert, C. Neunsinger, K. Osterrieder, “Managing customer satisfaction: digital applications for insurance companies,” The Geneva Papers on Risk and Insurance - Issues and Practice, vol.47, pp.569-602. February 2022.