EPiC Series in Education Science

Volume 6, 2024, Pages 78–94

Proceedings of the NEMISA Digital Skills Summit and Colloquium 2024

Employing Machine Learning Techniques to Analyze Customer Records for Cross-Selling Probability

Khulekani Mavunla[1*] and Surendra Thakur[2]

[1,2] Durban University of Technology, Durban, South Africa

20907985@dut4life.ac.za, thakur@dut.ac.za

Abstract

The study delved into health insurance cross-selling, where additional insurance products are promoted to existing policyholders, suggesting supplementary coverage such as dental or life insurance to those with basic health insurance. The study focused on applying machine learning to predict cross-selling opportunities among South African customers. The aim was to develop a predictive model to aid health insurers in identifying potential cross-selling customers. Utilizing a quantitative research methodology, a comprehensive dataset of health insurance consumer information was analyzed using various machine learning algorithms, including Random Forest, K-Nearest Neighbors, XGBoost classifier, and Logistic Regression in Python. Results revealed Logistic Regression as the top-performing model, achieving an accuracy score of 0.83 and an F1 score of 0.91 when trained on a dataset of 1,000,000 health insurance customers with 17 features comprising health insurance customer information. The analysis uncovered that customers aged 25-70 with prior insurance and longer service history are more likely to purchase additional health insurance products. These insights empower health insurers to enhance revenue through improved customer targeting and retention strategies, thereby providing valuable information for the industry's understanding of effective cross-selling approaches. The methodology comprised quantitative data extraction and machine learning application, thus contributing to advancements in cross-selling strategy comprehension.

1 Introduction

Health insurance cross-selling, wherein insurance companies offer existing policyholders additional insurance products, is a pivotal factor in business growth and profitability of these firms (Sekeroglu,

* Corresponding author

H. Twinomurinzi, N.T. Msweli, S. Gumbo, T. Mawela, E. Mtsweni, P. Mkhize and E. Mnkandla (eds.), NEMISA DigitalSkills 2024 (EPiC Series in Education Science, vol. 6), pp. 78–94

2022). This research aimed to establish a predictive model for health insurance cross-selling, utilizing machine learning algorithms to identify South African consumers inclined to acquire supplementary health insurance products. By analyzing diverse consumer attributes, including demographics, purchasing history, and socio-economic factors, the model sought to discern patterns and indicators crucial for predicting health insurance cross-selling. The objective was to assist health insurance companies in developing targeted marketing campaigns and tailored offers to optimize the success of cross-selling among existing customers (Sekeroglu, 2022).

Significant in enhancing the marketing strategies of health insurance companies targeting South African consumers, precise predictions of health insurance cross-selling enable insurers to efficiently identify potential customers, improve customer acquisition and retention rates, and ultimately increase business revenue (Ozdemir & Bayrakli, 2022). In the financial services domain, cross-selling has been extensively studied in sectors such as banking and retail, recognized as a potent strategy for boosting revenue and fortifying customer relationships (Dutta & Bhattacharya, 2019). Notably, a research gap remains, particularly in the realm of health insurance within the South African market. The scarcity of research specifically addressing machine learning applications for cross-selling in the South African health insurance sector is evident (Ndoro et al., 2020). Most existing studies focus on global markets, neglecting the unique dynamics and challenges that characterize the South African landscape. A thorough review of the literature has revealed a paucity of research exploring the intricacies of customer records within the context of the South African health insurance market (Pillay & Van der Merwe, 2021).

The South African population exhibits unique patterns of behavior when it comes to health insurance purchasing decisions. Cultural, economic, and social factors influence customer choices in ways that differ from other regions (Katrodia, 2021; Wrottesley et al., 2021). Machine learning models trained on global datasets might not effectively capture these nuances, thus emphasizing the need for targeted research within the South African market to enhance the accuracy of predictive analytics. Additionally, a comprehensive analysis of data availability and quality issues specific to the South African health insurance market highlights gaps in the health insurance industry's current datasets (Day & Zondi, 2019).

In the absence of comprehensive studies, there is a lack of established best practices for implementing machine learning techniques in the South African health insurance sector. Furthermore, the lack of standardized data collection methods, disparities in data formats, and incomplete datasets hinder the development and application of accurate machine learning models (Fletcher et al., 2021; Javaid et al., 2022). Collaborative efforts between industry stakeholders, researchers, and regulatory bodies are essential for establishing guidelines and frameworks that promote the ethical and effective use of machine learning for cross-selling within this market (Shaikh et al., 2022). The study addressed this gap by utilizing machine learning techniques to analyze historical customer data and forecast cross-selling potential, providing valuable insights for refining cross-selling strategies within the South African health insurance sector and ensuring the responsible use of machine learning in the context of customer record analysis (Qadadeh & Abdallah, 2018).

This research offered a distinctive opportunity to explore an untapped research area and contribute to the existing body of knowledge. Through addressing this research gap, the study aimed not only to fill this void but also to provide profound insights with the ability to enhance and guide more effective cross-selling practices within the health insurance industry. This paper commences with a review of relevant literature on the prediction of health insurance cross-selling, followed by a delineation of the research methodology, which includes data collection, preprocessing, exploratory data analysis, and feature selection. The paper proceeds with the construction and assessment of four machine learning models designed to predict cross-selling probability. The subsequent sections outline the experimental results. The paper concludes by summarizing key findings and proposing potential avenues for future research.

2 Literature Review

This study extends the current body of research in the realms of cross-selling prediction and the application of machine learning in the insurance industry. By pinpointing customers with a higher likelihood of acquiring additional insurance products, insurers can refine their marketing strategies, leading to increased business revenue. In South African insurance companies, which operate in a highly competitive market, precise prediction of health insurance cross-selling becomes crucial for insurers to maintain a competitive edge within the industry (Ozdemir & Bayrakli, 2022).

Historically, cross-selling has been dependent on manual processes and subjective decision-making. However, the advent of machine learning and predictive analytics offers insurers the chance to employ data-driven approaches for more efficient identification of potential cross-selling opportunities (Kumar et al., 2019). The application of machine learning algorithms allows insurance companies to analyze extensive datasets of health insurance customer information to reveal patterns and insights that may elude traditional methods (Sidorowicz et al., 2022).

2.1 Customer Satisfaction to Enhance Probability of Cross-Selling

The success of cross-selling is profoundly influenced by customer satisfaction, as content individuals are inclined not only to avail themselves of existing services but also demonstrate a heightened willingness to explore and acquire additional insurance products. In a comprehensive study conducted by Eckert et al. (2021), the evaluation of customer satisfaction levels was undertaken through sentiment analysis on customer reviews. This analytical approach involved the application of advanced machine learning techniques to discern patterns within sentiments, thereby establishing crucial connections between customer satisfaction and the positive outcomes of cross-selling endeavors.

The utilization of machine learning techniques, with a specific emphasis on the gradient boosting model (Vandrangi, 2022), proved to be highly effective in discerning potential customers who may be amenable to acquiring additional insurance products. This effectiveness was demonstrated through the analysis of historical data and customer attributes. Furthermore, a parallel study conducted by Vandrangi (2022) delved into the construction of predictive models for cross-selling health insurance policies, employing random forests and neural networks. Notably, Vandrangi's research underscored the significance of meticulous feature engineering and judicious model selection in ensuring precise and reliable predictions in the context of cross-selling endeavors.

2.2 Predictive Modeling for Cross-Selling in Health Insurance

Several studies have explored the efficacy of predictive modeling methodologies in terms of discerning potential cross-selling prospects within the realm of health insurance. One noteworthy example involves a comprehensive study that employed a sophisticated gradient-boosting algorithm. This algorithm was utilized to prognosticate the probability of customers opting for supplementary coverage, leveraging valuable insights gleaned from their extensive historical health claims data and pertinent demographic information (Weerasinghe & Wijegunasekara, 2016). The findings underscored the heightened predictive precision offered by ensemble methods, exemplified by gradient boosting when compared to conventional logistic regression models.

Furthermore, in the dynamic landscape of health insurance, the exploration of predictive modeling techniques has become increasingly paramount. Beyond the study of Weerasinghe and Wijegunasekara (2016), various researchers have embarked on endeavors to uncover novel approaches for identifying cross-selling opportunities (Ozdemir & Bayrakli, 2022). One notable avenue involves the integration of artificial intelligence and machine learning methodologies, which not only enhances the accuracy of predictions but also contributes to a deeper understanding of customer behavior patterns.

2.3 Customer Segmentation for Health Insurance Cross-Selling

Implementing customer segmentation in the context of health insurance cross-selling can significantly enhance the precision of prediction models and boost the overall effectiveness of marketing strategies (Khan & Aziz, 2023). By leveraging unsupervised clustering techniques, characteristics and behaviors are utilized to identify distinct customer segments, thereby enabling a more nuanced understanding of diverse preferences and needs (Sekeroglu, 2022; Sidorowicz et al., 2022). This strategic segmentation lays the foundation for targeted marketing efforts and personalized recommendations to ensure that cross-selling initiatives are tailored to the specific requirements of each segment (Altun & Yucekaya, 2021). With separate predictive models dedicated to individual clusters, the approach not only acknowledges the heterogeneity within the customer base but also allows for a more accurate prediction of purchasing behaviors, ultimately optimizing the cross-selling process.

Additionally, the adoption of customer segmentation not only refines predictive models but also empowers health insurance providers to align their cross-selling efforts with the evolving demands of different market segments. Through a data-driven understanding of customer characteristics, preferences, and behaviors, insurers can tailor their communication strategies and product offerings to cater to the unique needs of each segment. This targeted approach both enhances customer satisfaction and contributes to increased cross-selling success rates. In an era where personalization is paramount, leveraging customer segmentation is a strategic imperative for health insurance providers seeking to navigate the complexities of cross-selling while staying ahead in a competitive market.

2.4 Addressing Imbalanced Data in Cross-Selling Prediction

Addressing imbalanced data is a pervasive challenge in the realm of cross-selling prediction, particularly in the context of health insurance, where a substantial majority of customers may opt not to purchase additional coverage (Hanafy & Ming, 2021). The inherent disparity in the number of customers across the two classes, with non-purchasers significantly outnumbering those who choose additional coverage, necessitates specialized techniques for effective model training and evaluation. To take on this issue, a variety of approaches are commonly employed, including oversampling the minority class to balance the dataset and adapting evaluation metrics to account for the class imbalance (Sidorowicz et al., 2022). These strategic interventions aim to rectify the skewed distribution of data and foster a more equitable representation of both purchasing and non-purchasing behaviors.

The significance of addressing imbalanced data becomes even more pronounced in cross-selling prediction, where biased models may lead to suboptimal marketing strategies and inaccurate predictions (Werb & Schmidberger, 2021). The findings of Werb and Schmidberger emphasize the critical role of these techniques in mitigating the risks associated with imbalanced data to ensure that predictive models are robust and reliable in capturing the complexities of customer behaviors in the health insurance cross-selling landscape. As the industry increasingly relies on data-driven insights, addressing imbalanced data emerges as a pivotal step toward fostering more accurate and unbiased predictions, ultimately enhancing the efficacy of cross-selling endeavors in the competitive health insurance market (Ghavidel & Pazon, 2023).

This study aspired to fill existing research gaps and offer valuable insights with practical implications to make a significant contribution to current knowledge. The primary focus centered around enhancing cross-selling practices within the South African health insurance sector through the implementation of machine learning algorithms. The overarching objective was to streamline and optimize the efforts of insurance companies, ultimately leading to increased business revenue and heightened customer satisfaction by identifying potential customers more effectively. By addressing specific challenges within the South African context, this research endeavored to provide tailored solutions that can be applied readily by insurance providers, thereby aligning theoretical advancements with practical applications in the dynamic landscape of the health insurance industry. The study's

unique emphasis on the South African market ensured a contextually relevant approach, acknowledging the nuances of the local insurance landscape and contributing insights that can foster improvements in cross-selling strategies.

3 Research Methodology

Employing a quantitative research methodology, this study adopted a deductive approach to illuminate patterns within the realm of human existence. The methodology involved dissecting the social landscape into measurable components, referred to as variables that can be expressed and quantified numerically.

The research chose a quantitative approach to explore questions such as "the number of high-earning customers," "the proportion of previously insured females and males," and "the extent of insurance coverage among different age groups" (Rahman, 2016). To execute this approach, health insurance data was meticulously extracted from an open-source database, honing in on quantifiable customer behaviors. The primary objective was to conduct a comprehensive analysis of these behaviors, discern patterns, and interpret their underlying meanings to generate valuable insights. By adopting a quantitative lens, the study aimed to provide a structured and numerical understanding of the intricacies within the health insurance dataset, allowing for rigorous statistical analysis and the extraction of meaningful patterns that contribute to a nuanced comprehension of customer behaviors in the context of health insurance cross-selling. This methodological choice ensured a systematic exploration of the quantitative dimensions of customer interactions, facilitating a more precise and data-driven examination of the study's focal area.

In the pursuit of health insurance cross-selling prediction, this study exclusively employed quantitative data, represented in numerical counts. The primary focus was on systematically collecting and rigorously analyzing this numerical information to discern the interest levels of existing customers in acquiring additional health insurance products (Goertzen, 2017). By leveraging statistical analysis techniques, the study aimed to unveil patterns and trends within the quantitative data, thereby providing a data-driven foundation for predicting cross-selling behaviors among the target customer base. This approach facilitated a precise examination of numerical indicators and enabled the identification of key factors influencing customers' inclination toward purchasing supplementary health insurance products. The emphasis on quantitative data underscored the study's commitment to a robust and empirically grounded exploration of customer behaviors in the health insurance domain, with the ultimate goal of enhancing predictive accuracy in the context of cross-selling initiatives.

The study collected important information about policyholders and insured members to create a comprehensive health insurance customer dataset. The researcher endeavored to build a health insurance cross-celling prediction model and followed the step-by-step approach in the machine learning lifecycle. This method provided a structured guide during the model-building process, which made it easier to explore complex patterns in the collected data. By using this systematic methodology, the study improved its analytical depth and played a role in developing an advanced and dependable predictive model specifically designed for cross-selling health insurance products.

3.1 Phases in the Machine Learning (ML) Lifecycle

The Machine Learning (ML) lifecycle (Figure 1) is designed to organize data and construct resilient predictive models. It encompasses several stages where techniques and algorithms may vary depending on factors such as health insurance dataset characteristics and project objectives (Ritz et al., 2022). It is worth emphasizing that the lifecycle is inherently iterative, permitting multiple passes through its stages. This iterative nature allows for refining the model and enhancing its performance based on feedback and insights acquired throughout the process (Ritz et al., 2022).

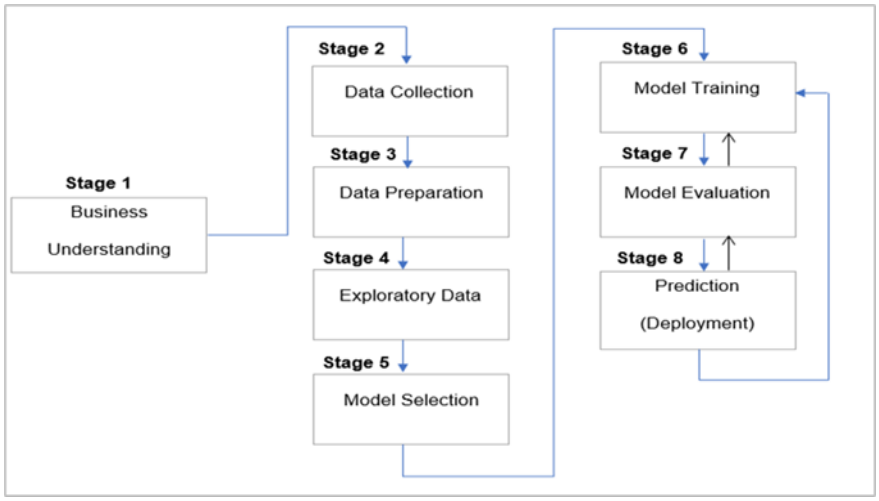

Figure 1: Visualization of the Machine Learning Lifecycle (Source: Researcher)

3.2 Health Insurance Customer Dataset Source

The data extracted from the extensive database of the insurance company in South Africa encompassed a wealth of information related to various aspects, including insurance policies, members, claims, and other essential variables. This comprehensive dataset offered a detailed overview of the diverse facets within the insurance domain, which provided a nuanced understanding of the intricacies associated with policies, membership details, and claims data, among other relevant factors. The richness of this dataset ensured the availability of a broad spectrum of information for analysis, which, in turn, contributed to a comprehensive exploration of the insurance landscape and facilitated a thorough examination of the relationships and patterns embedded within the collected data.

3.3 Analyzing the Health Insurance Dataset

Table 1 presents the features representing individual customers along with their descriptions and data types.

No#	Variable	Definition	Data Type
1	ID	Unique ID for consumer, e.g., 1, 2,3, 4, anonymous data	Integer
2	Gender(Gen)	Gender of the consumer 0=Other, 1=Male, 2=Female	Integer
3	Age	Age of the consumer	Integer
4	RegionCode	Unique code for the region of the customer	Float
5	RaceCode	Different races (Unknown, White, Black, Indian, Coloured), where 1=Unknown, 2=White, 3=African, 4=Coloured, 5=Asian	Integer
6	PreviouslyInsured (Prev Insured)	1: Customer already has health insurance, 0: Customer does not have health insurance	Integer
7	InitialSumAssured (Sum Assured)	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 consumer monthly income	Float

No#	Variable	Definition	Data Type
10	MonthlyPremium	Amount consumer needs to pay every month for health insurance	Float
11	AnnualPremium	Amount consumer needs to pay as a premium in the year for health insurance	Float
12	Vintage(Vin)	Number of days the consumer has been associated with the company	Integer
13	InsurerType	1=Consumer uses internal health insurance product, 0=Customer uses external health insurance product	Integer
14	PolicyStatusType	1=Active, 0=Inactive	
15	ProductTypeType	1=Consumer has comprehensive health insurance cover, 2= consumer has accident-only cover, 3=consumer has standard cover	Integer
16	InsuranceCondition	1=Health insurance condition is compulsory, 0=health insurance condition is optional	Integer
17	Response	1=Consumer is perceived to be interested, 0=Customer is perceived not to be interested	Integer

Table 1: Health insurance data features

3.4 Approach to Data Collection

A meticulous health insurance dataset comprising records of health insurance customers was acquired through a thoughtfully constructed SQL query. The dataset extraction process involved careful selection from various tables and was refined to include only the top 1,000,000 entries, ensuring a robust representation of the population under study. This extensive dataset encapsulated 17 features encompassing both individual- and insurance-specific details. Furthermore, to align with privacy regulations for the Protection of Personal Information (POPIA) Act, the dataset underwent a thorough anonymization process. This meticulous step was implemented to safeguard the security and confidentiality of customer information, attesting to the study's commitment to ethical data handling practices and compliance with privacy norms.

3.5 Utilizing Jupyter Lab and Python Programming

The dataset was imported into the Jupyter Lab environment by utilizing the Python Pandas library for the purpose of training the cross-selling predictive model. To verify the successful import, a review of the top ten records within the dataset was conducted. The initial state of the dataset, prior to the commencement of the cleaning process, is presented in Table 2 below.

	Training Dataset Shape	Validation Dataset Shape
Number of Rows	1 000 000	1 000 000
Number of Columns	17	16

Table 2: Training dataset prior to the cleaning process

3.6 Preprocessing the Health Insurance Dataset

The data preprocessing phase, integral in both data analysis and the machine learning pipeline, was executed using Python code to eradicate duplicates and rectify inaccuracies and data entry errors (Misra & Yadav, 2019).

Addressing missing values involved employing techniques such as imputation, which replaces missing values with estimated ones, and removing rows and/or columns containing missing data. Following these steps, the extracted dataset was partitioned into training and testing subsets to comprehensively evaluate the model's performance.

Post data preprocessing, specific records from the health insurance dataset were dropped, leading to a notable alteration in the dataset's structure, as depicted in Table 3 below.

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

Table 3: Training dataset shape

Following the completion of the data cleaning and pre-processing procedures, the top ten records and 12 features were chosen from the training dataset. The presentation in Table 4 below attests to the successful cleaning of the data, signaling its readiness for subsequent model training.

ID	Gen	Age	Region Code	Race Code	Prev Insured	Sum Assured	Monthly Income	Monthly Premium	Annual Premium	Vin	Resp.
3	1	44	2302	2	1	729 075.35	51 878.45	992.17	11 357.52	3	0
7	1	43	163	2	1	759 778.59	24 500.00	707.04	8 190.00	10	0
10	2	56	2190	3	1	252 398.73	18 840.25	618.49	7 106.65	9	0
11	2	59	9786	3	1	285 339.33	15 748.00	471.27	5 465.18	2	0
15	1	31	157	4	1	823 986.79	29 750.00	526.57	6 055.17	12	0
17	1	46	1818	3	1	521 995.54	36 904.16	755.98	8 589.81	20	0
18	2	52	7100	3	1	188 678.85	7 820.55	243.68	2 790.08	20	0
20	1	40	5201	3	1	759 578.71	40 789.75	767.42	8 761.35	19	0
23	2	46	4066	3	1	653 353.80	37 303.87	700.46	7 933.77	2	0
24	2	52	1618	3	1	672 240.25	29 313.59	2 347.55	26 980.11	18	0

Table 4: Training dataset prior cleaning process

4 Analyzing Dataset Attributes

The exploration of data through the process of exploratory data analysis (EDA), a crucial component of the data pre-processing stage, was implemented (Boodhun & Jayabalan, 2018). This step was taken to emphasize the importance of thoroughly examining the data for a comprehensive understanding and analysis of the extracted dataset. For the EDA, the researcher conducted an analysis of the health insurance dataset, considering aspects such as size, columns, and variables. This involved employing descriptive and correlation assessments, complemented by data visualizations, to gain deeper insights into the characteristics and relationships present within the dataset.

4.1 Examining Correlations Among Variables in the Dataset

Correlation refers to the statistical relationship between variables, quantifying both the strength and direction of their linear association. Statistical measurements are employed to assess the extent of this relationship, especially when variables exhibit movement in the same direction (Mikalef et al., 2020). This analytical approach enables a quantitative understanding of how changes in one variable correspond to changes in another, thus providing valuable insights into their interconnectedness and facilitating a nuanced interpretation of their relationship within a dataset.

A visual representation in the form of a figure showcases the Pearson correlation among the 17 features analyzed in this study (see Figure 2). This graphical depiction offers an accessible and intuitive way to interpret the relationships and patterns present within the dataset, highlighting the degree and direction of correlation between each pair of features. Such visual aids contribute to a clearer understanding of the interconnections among the variables, enhancing the overall comprehension of the dataset's intricate dynamics.

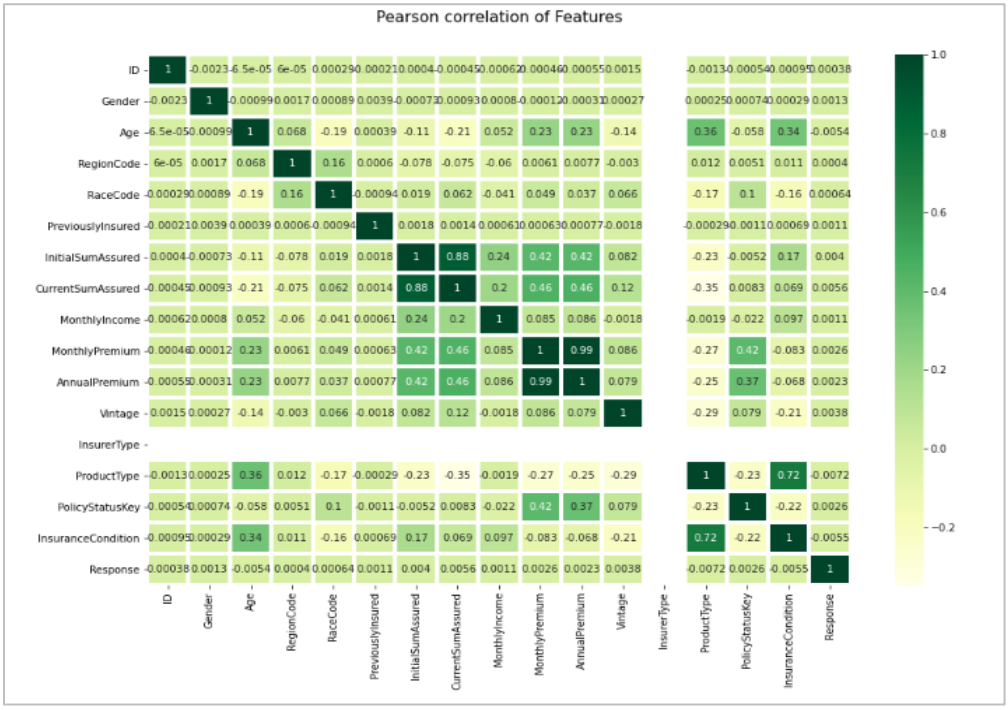

Figure 2: Pearson correlation analysis of 17 features (Source: Researcher)

4.2 Illustrating Data Patterns Through Visualization

Data visualization utilizes graphical elements, including charts and graphs, to effectively represent information (Wang et al., 2020). In the context of this study, visualization served as a valuable tool for presenting features that influence the target variable within the health insurance dataset. The use of visualizations aids in the exploration of data, identification of patterns, detection of outliers, and the clear and impactful communication of findings. By visually representing complex relationships and trends, data visualization enhances the interpretability of the dataset, making it more accessible and facilitating meaningful insights for analysis and decision-making.

Figure 3 shows the vintage response, representing the duration of customer association with the company in days.

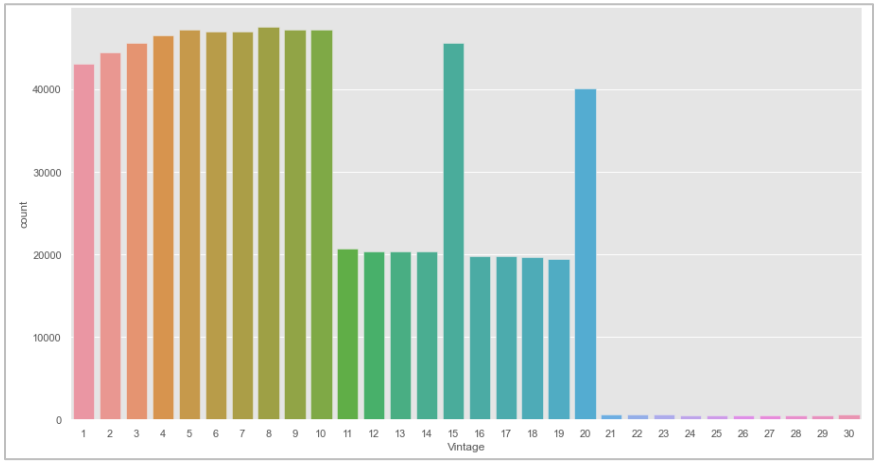

Figure 3: Customer long service response (Source: Researcher)

The graph presented in Figure 4 displays the age versus response curve, highlighting that individuals aged 25-60 are more likely to express interest in purchasing additional health insurance products compared to those younger than 30 years old.

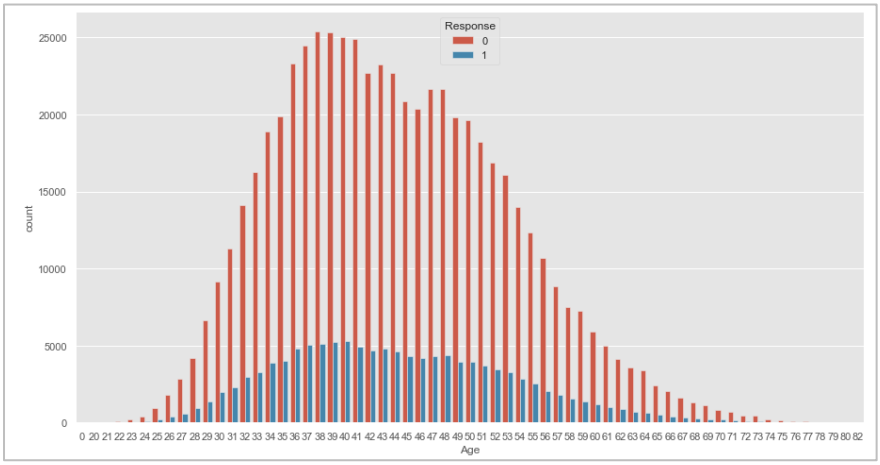

Figure 4: Age vs. response relationship visualization (Source: Researcher)

The chart below (Figure 5) exhibits gender counts categorized as *Other*, *Male*, and *Female*, depicted through both a bar graph and a pie chart with percentage distribution. Gender is numerically represented as 0 for *Other*, 1 for *Male*, and 2 for *Female*, exposing dataset bias towards male records.

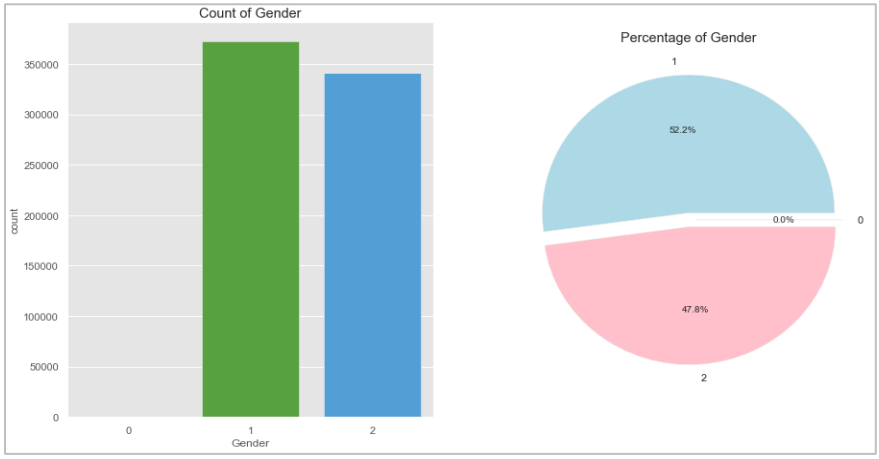

Figure 5: Gender distribution visualization (Source: Researcher)

4.3 Feature Enhancement Through Engineering

Feature engineering is the process of creating new features or modifying existing ones from raw data to improve the performance of machine learning models and data analysis tasks (Sekeroglu, 2022). In this study, the correlation analysis method was utilized for feature engineering. Correlation analysis, a statistical technique, quantifies the degree of association or relationship between two or more variables (Sekeroglu, 2022). It facilitates an understanding of how changes in one variable correspond to changes in another, providing valuable insights for enhancing the effectiveness of the features used in the analysis.

In the process of feature selection, Python code has been utilized to implement the following steps:

- Compute the correlation matrix between all features and the target variable.
- Apply a filter to extract the correlation values associated with the target variable.
- Arrange the features in descending order based on their correlation with the target variable.
- Display the features along with their corresponding correlation values.

4.4 Model Selection for Predictive Analysis

Choosing a model involves the selection of the most suitable machine learning algorithm or model for a specific task or dataset (Anitha & Patil, 2022). The process of model selection includes evaluating various models, comparing their performance, and selecting the one that best fits the data, ultimately yielding the most accurate predictions.

The subsequent details highlight machine learning algorithms meticulously chosen, trained, and evaluated for the construction of a predictive model geared toward health insurance cross-selling. This comprehensive selection encompasses a diverse set of algorithms, including Random Forest, K-Nearest Neighbors (KNN), XGBoost classifier, and Logistic Regression. Each algorithm underwent rigorous training and assessment to discern its efficacy in predicting cross-selling outcomes, ensuring a thorough exploration of diverse modeling approaches.

4.5 Training and Evaluating Models for the Cross-Selling Predictive Model

Training a machine learning model involves instructing machine learning algorithms to recognize patterns and make predictions based on input data (Severino & Peng, 2021) Subsequently, model

evaluation metrics are utilized to analyze the effectiveness of a machine learning model, determining its capacity to extrapolate to unfamiliar data. The choice of evaluation metrics hinges on the specific task and the nature of the problem at hand (Vujovic, 2021).

Presented below are commonly employed evaluation metrics tailored to different types of ML models, each elucidated with examples for clarity.

- Confusion matrix for evaluating cross-selling prediction model: A confusion matrix serves as a valuable table for assessing the performance of a classification model. Its usefulness is particularly prominent in the context of binary classification problems, where the focus is on two distinct classes: *Predicted Positive* and *Predicted Negative*. This matrix allows for a detailed examination of the model's predictions, breaking down the outcomes into true positives, true negatives, false positives, and false negatives. Such granularity aids in a comprehensive understanding of the model's strengths and weaknesses, providing insights into its ability to correctly classify instances and identify areas for improvement in the prediction process.
- Evaluation metrics for health insurance prediction model: The assessment of a machine learning model's performance and the quantification of its effectiveness hinge significantly on the use of model evaluation metrics. In this critical process, a systematic approach is adopted by employing the equations set out in section 5 below, as delineated in the literature (Vujovic, 2021; Grandini et al., 2020). These equations serve as valuable tools for both evaluating the model's accuracy and providing a nuanced understanding of its predictive capabilities across diverse metrics.

5 Experiments, Results Analysis, and Discussion

The study aimed to predict health insurance cross-selling opportunities using machine learning, incorporating historical behaviors, demographics, and interactions. The purpose was to create a robust predictive model that understands customer behaviors and characteristics, allowing for targeted identification of cross-selling opportunities in the health insurance domain.

The results of the predictive analysis on cross-selling probability in the insurance sector revealed significant implications for health insurers seeking to identify potential cross-selling customers. The accuracy scores for the models evaluated (see section 4.4) range from 0.799 to 0.831, with Logistic Regression achieving the highest accuracy score of 0.831 and a F1 score of 0.9, which offers a precise means of assessing the likelihood of customers engaging in cross-selling activities. This clarity allows insurers to tailor their marketing efforts, focusing on individuals with a higher propensity for accepting additional health insurance products or services.

The predictive analysis offers health insurers an effective way to identify potential cross-selling customers. High-accuracy models, such as Logistic Regression, allow for the precise targeting of individuals likely to respond positively to cross-selling. This precision enables optimized marketing resource allocation, thereby focusing efforts on customer segments inclined to purchase additional products. Leveraging segmentation strategies based on cross-selling probabilities enable insurers to tailor marketing approaches to specific customer groups. Additionally, the predictive analysis facilitates personalized product recommendations, thus ensuring that additional offerings resonate with individual customer needs.

The discussion of this study revolved around the actionable insights derived from the analysis. Logistic Regression emerged as a standout performer, confirming its effectiveness in predicting cross-selling behavior. This model's success can be harnessed by insurers to optimize their cross-selling strategies. By incorporating these predictive findings, insurers can develop targeted marketing campaigns by segmenting customers based on their predicted probabilities. This segmentation enables

personalized product recommendations, thus ensuring that cross-selling offers align with the specific needs and preferences of diverse customer segments.

Figure 6 compares the performance of various machine learning algorithms utilized in the study.

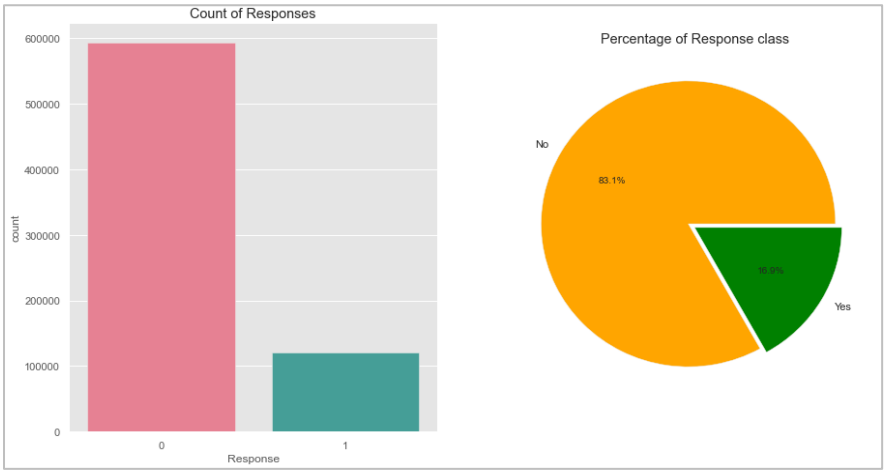

Figure 6: Comparison of machine learning algorithms (Source: Researcher)

In practical terms, health insurers can utilize these results to streamline their cross-selling initiatives. The predictive models not only enhance precision targeting but also contribute to resource optimization by directing marketing efforts toward those more likely to respond positively. This ensures efficient use of resources and maximizes the impact of cross-selling campaigns.

The study aimed to offer a comprehensive overview, dissecting the intricacies of each model's performance and providing valuable insights into their respective strengths and areas for potential refinement.

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

Table 5: Performance of machine learning algorithms

Upon analyzing the F1-Scores (Table 5) of the four models, a clear pattern emerged, highlighting the Logistic Regression model as the standout performer with the highest F1 score of 0.91. The high F1 scores across models emphasize balanced precision and recall, affirming the reliability of the predictive analysis in identifying potential cross-selling customers.

5.1 Equations Employed for Health Insurance Cross-Selling

In the case of Logistic Regression, the following equations (Boateng & Abaye, 2019) were applied: Logistic Regression was employed with two variables, where one served as the dependent variable (Y), and the other functioned as the independent variable (X). Equation 1 was utilized to characterize the relationship between X and Y, with Y being binary.

$$h\theta(X) = \frac{1}{1+e^{-(\theta_0+\theta_1X_1+\theta_2X_2+\ldots+\theta_nX_n)}} \qquad (\textit{equation } 1)$$

In the context of Logistic Regression, the application of the following equations (Boateng & Abaye, 2019) was integral to the analytical process. Logistic Regression was implemented with two variables, where one assumed the role of the dependent variable (Y), and the other operated as the independent variable (X). Equation 1 played a pivotal role in delineating the relationship between X and Y, specifically tailored for scenarios where Y is binary. This equation provided a nuanced characterization of the intricate dynamics between the independent and dependent variables, contributing to a comprehensive understanding of the logistic regression model's predictive capabilities.

Equation 2, the sigmoid function, is crucial in logistic regression and binary classification algorithms. Also known as the logistic function, it maps real-valued numbers to a range between 0 and 1. This function is significant for compressing diverse input values into a bounded output range, making it pivotal in modeling binary outcomes and defining nuanced decision boundaries in classification algorithms.

Equation 3 in logistic regression combines predictor variables, influencing log odds and predicting probabilities. This mathematical formulation enables the model to systematically integrate multiple predictors, discern patterns, and accurately predict health insurance cross-selling opportunities.

$$f(x) = \beta_0 + \beta_1 X_1 + + \beta_2 X_2 + \cdots + \beta_r X_r \qquad (equation\ 2)$$

β_0 in the logistic regression equation denotes the intercept term, representing the baseline or constant component. This term is essential for determining the starting point of the logistic regression curve and establishes the baseline probability or log odds.

β_1, β_2, ..., β_r are the coefficients associated with the predictor variables x_1, x_2, ..., x_r in the logistic regression equation. These coefficients determine the magnitude and direction of the impact that each predictor variable has on the log odds or probability of the event being predicted. The interplay of these coefficients with the predictor variables contributes to the overall predictive power of the logistic regression model.

The study aimed to determine the best machine learning model for predicting health insurance cross-selling opportunities. After evaluating the Random Forest, K-Nearest Neighbors, XGBoost Classifier, and Logistic Regression models, Logistic Regression emerged as the most suitable model, with an accuracy rate of 0.83 and an impressive F1 score of 0.91. The analysis of the health insurance dataset revealed that individuals aged 25-70 and those with a longer history of insurance coverage ("Vintage") were more likely to purchase supplementary health insurance products, which emphasizes the importance of age and historical data in the predictive model.

Moreover, the study discerned pivotal features that wield significant influence on the predictive model. Notable among these are *Gender, Age, Previously Insured status, Monthly Income, Monthly Premium,* and *Annual Premium*. The identification of these critical features enhances the model's interpretability and offers valuable insights into the key determinants shaping the predictions for health insurance cross-selling opportunities.

6 Conclusion

In a comprehensive overview, the current research underscores the immense potential inherent in the application of machine learning algorithms for the analysis of health insurance datasets, particularly in the realm of cross-selling probabilities. This cross-selling predictive approach capitalizes on the vast reservoir of customer data within the industry, allowing for the discernment of intricate patterns, behaviors, and preferences. The utilization of predictive machine learning models emerges as a highly effective strategy, proficiently gauging the likelihood of customer interest in supplementary insurance products. This serves as a testament to the prowess of these models in the dynamic landscape of the health insurance sector.

Machine learning algorithms in health insurance enhance cross-selling efficiency and contribute to personalized healthcare. These MLAs decode customer patterns, thus allowing tailored insurance offerings and fostering a responsive approach. This predictive model is set to revolutionize customer engagement, shifting towards a more adaptive and anticipatory model that meets evolving consumer expectations.

Looking ahead, prospective research could delve into the integration of insights derived from health insurance cross-selling predictions. Such explorations hold the promise of further enriching the industry's comprehension and utilization of machine learning, thereby refining customer-centric strategies and optimizing overall business outcomes. The potential implications of these findings of cross-selling predictions extend beyond the immediate scope of current research, paving the way for continuous advancements in leveraging machine learning within the intricate domain of health insurance.

References

Altun, Y., & Yucekaya, A. (2021). A probabilistic approach to maximize cross-selling revenues of financial products. *American Scientific Research Journal for Engineering, Technology, and Sciences (ASRJETS)*, 79, 1–14. Industrial Engineering Department, Kadir Has University, Istanbul, Turkey.

Anitha, P., & Patil, M. M. (2022). RFM Model for customer purchase behavior using K-Means algorithm. *Journal of King Saud University-Computer and Information Sciences*, 34, 1785–1792.

Boateng, E. Y., & Abaye, D. A. (2019). A review of the logistic regression model with emphasis on medical research. *Journal of Data Analysis and Information Processing*, 7, 190–207.

Day, C., & Zondi, T. (2019). Measuring national health insurance: Towards universal health coverage in South Africa. *South African Health Review*, 2019(1), 55–68.

Dutta, S., & Bhattacharya, S. (2019). Cross selling of investment products and services: A case study of leading financial services organization. *International Journal of Business Forecasting and Marketing Intelligence*, 5(2), 241–248.

Eckert, C., Neunsinger, C., & Osterrieder, K. (2021). Managing customer satisfaction: Digital applications for insurance companies. *The Geneva Papers on Risk and Insurance – Issues and Practice*, 47, 569–602.

Fletcher, R. R., Nakeshimana, A., & Olubeko, O. (2021). Addressing fairness, bias, and appropriate use of artificial intelligence and machine learning in global health. *Frontiers in Artificial Intelligence*, 3, 561802.

Ghavidel, A., & Pazos, P. (2023). Machine learning (ML) techniques to predict breast cancer in imbalanced datasets: A systematic review. *Journal of Cancer Survivorship*, 1–25.

Goertzen, M. J. (2017). Introduction to quantitative research and data. *Library Technology Reports, 53*(5), 12–18.

Grandini, M., Bagli, E., & Visani, G. (2020). Metrics for multi-class classification: An overview. *arXiv preprint arXiv:2008.05756*.

Hanafy, M., & Ming, R. (2021). Using machine learning models to compare various resampling methods in predicting insurance fraud. *Journal of Theoretical and Applied Information Technology, 99*(1), 12–23.

Javaid, M., Haleem, A., Singh, R. P., Suman, R., & Rab, S. (2022). Significance of machine learning in healthcare: Features, pillars and applications. *International Journal of Intelligent Networks, 3*(1), 58–73.

Katrodia, A. (2021). A study of identity consumer purchasing behavior and factors that influence consumer purchase decision: With reference to Durban. *Journal of the Research Society of Pakistan, 58*(3), 60.

Khan, A. R., & Aziz, M. T. (2023). Harnessing Big Data for precision marketing: A deep dive into customer segmentation and predictive analytics in the digital era. *AI, IoT and the Fourth Industrial Revolution Review, 13*(7), 91–102.

Kumar, N., Srivastava, J. D., & Bisht, D. (2022). *Artificial Intelligence in insurance sector*. Retrieved from https://www.researchgate.net/profile/Naman-Kumar-3/publication/337305024_Artificial_Intelligence-in-Insurance-Sector/links/5dd00e33a6fdcc7e138761cc/Artificial-Intelligence-in-Insurance-Sector.pdf

Mikalef, P., Krogstie, J., Pappas, I. O., & Pavlou, P. (2020). Exploring the relationship between big data analytics capability and competitive performance: The mediating roles of dynamic and operational capabilities. *Information & Management, 57*, 103169.

Misra, P., & Yadav, A. (2019). *Impact of preprocessing methods on healthcare predictions*. Retrieved from https://www.researchgate.net/profile/Puneet-Misra-3/publication/332436103_Impact-of-Preprocessing-Methods-on HealthcarePredictions/links/5d37ac2192851cd04680da45/Impact-of-Preprocessing-Methods-on-Healthcare-Pred Sidorowicz ictions.pdf

Boodhun, N. N., & Jayabalan, M. (2018). Risk prediction in life insurance industry using supervised learning algorithms. *Complex & Intelligent Systems, 4*, 145–154.

Ndoro, H., Johnston, K., & Seymour, L. F. (2020). Artificial intelligence uses, benefits and challenges: A study in the western cape of South Africa financial services industry. *SACAIR 2020, 58*.

Ozdemir, Y. E., & Bayrakli, S. (2022). A case study on building a cross-selling model through machine learning in the insurance industry. *Avrupa Bilim ve Teknoloji Dergisi, 35*, 364–372.

Pillay, K., & Van der Merwe, A. (2021, August). Big Data driven decision making guidelines for South African banking institutions. In *2021 International Conference on Artificial Intelligence, Big Data, Computing and Data Communication Systems (icABCD)* (pp. 1–6). IEEE.

Qadadeh, W., & Abdallah, S. (2018). Customers segmentation in the insurance company (TIC) dataset. *Procedia computer science*, *144*, 277–290.

Rahman, M. S. (2016). The advantages and disadvantages of using qualitative and quantitative approaches and methods in language testing and assessment research. *Journal of Education and Learning, 6*, 102.

Ritz, F., Phan, T., Sedlmeier, A., Altmann, P., Wieghardt, J., Schmid, R., & Gabor, T. (2022, October). Capturing 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.

Sekeroglu, A. G. (2022). *Impacts of feature selection techniques in machine learning algorithms for cross selling: A comprehensive study for insurance industry.* Retrieved from 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

Sekeroglu, A. G. (2022). *Impacts of Feature selection techniques in machine learning algorithms for cross selling: A comprehensive study for insurance industry.* Retrieved from 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

Severino, M. K., & Peng, Y. (2021). Machine learning algorithms for fraud prediction in property insurance: Empirical evidence using real-world microdata. *Machine Learning with Applications*, *5*, 100074.

Shaikh, A. A., Lakshmi, K. S., Tongkachok, K., Alanya-Beltran, J., Ramirez-Asis, E., & Perez-Falcon, J. (2022). Empirical analysis in analysing the major factors of machine learning in enhancing the e-business through structural equation modelling (SEM) approach. *International Journal of System Assurance Engineering and Management*, *13*(Suppl 1), 681–689.

Sidorowicz, T., Peres, P., & Li, A. Y. (2022). Novel approach for cross-selling insurance products using positive unlabeled learning. In *2022 International Joint Conference on Neural Networks (IJCNN)* (pp. 1–8). IEEE Xplore.

Vandrangi, S. K. (2022). Predicting the insurance claim by each user using machine learning algorithms. *Journal of Emerging Strategies in New Economics*, *1*(1), 1–11.

Vujović, Z. (2021). Classification model evaluation metrics. *International Journal of Advanced Computer Science and Applications*, *12*(6), 599–606.

Wang, Q., Chen, Z., Wang, Y., & Qu, H. (2020). *Applying machine learning advances to data visualization: A survey on ML4VIS.* Retrieved from 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

Weerasinghe, K. P. M. L. P., & Wijegunasekara, M. C. (2016). A comparative study of data mining algorithms in the prediction of auto insurance claims. *European International Journal of Science and Technology*, *5*(1), 47–54.

Werb, G. A., & Schmidberger, M. (2021). Predictive modeling in marketing. *Die Unternehmung*, *75*(3), 376–396.

Wrottesley, S. V., Bosire, E. N., Mukoma, G., Motlhatlhedi, M., Mabena, G., Barker, M., Hardy-Johnson, P., Fall, C., & Norris, S. A. (2021). Age and gender influence healthy eating and physical activity behaviors in South African adolescents and their caregivers: Transforming Adolescent Lives through Nutrition Initiative (TALENT). *Public Health Nutrition*, *24*(16), 5187–5206.