Health Insurance Lead Prediction

Sanket Sonu StudentId: x19206071
Sachin Muttappanavar StudentId: x20144253
Saranya Varshni Roshan Karthikha StudentId: x20154801
MSc in Data Analytics
National College of Ireland Dublin, IRELAND
URL: www.ncirl.ie

***Abstract*—Nowadays, health is very uncertain and unpredictable because of the volatile world. Nature is one of the main reasons behind this and uncertain accidents, which can occur anytime without knocking on the door, can cost our life or can do serious injury. Health insurance is a trend that is very necessary for everyone, however, health insurance is expensive because it takes care of many factors. The research is using a FinMan dataset, which is a financial service company that provides loans, insurance, investment funds, etc. The aim of this research is to find whether a previous or new customer, who is taking other services, will continue with the health insurance? Companies use demographics and other information to predict the insurance premium value. This project is using Logistic Regression to Predict the Response of the customers after they visited the website to check the insurance premium value.**

***Index Terms*—Health Insurance, Logistic Regression, Python, Business Analysis, Financial Service**

I. Introduction

Health Insurance is one of the dominant parts of anyone's life. Nature and the world are very unpredictable. The human body is very sensitive to nature and lifestyle. The lifestyle that most of the population is living can lead to many health issues, the common one is diabetes which can lead to heart attack and many diseases like cancer is uncertain. There are many health issues that are very severe and hence cost a huge amount of money to cure. Moreover, health care bills and medical expenses are very expensive nowadays, for example, the Covid-19 pandemic has shown the real price of health care expenses. In the worst-case scenario, India had 425,000 cases per day and many of the patients needed ventilators and normal oxygen and ventilators were costing them a huge amount of money. There were millions of patients who were able to get good treatment because of expensive medical facilities in a hard time.

Data Source: Health Insurance dataset from Kaggle. The data has been sourced from Kaggle. This dataset has 14 columns and 50,882 observations. This research paper uses a dataset of FinMan company. FinMan is a financial service provider, which provides financial services such as loans, insurances, investment funds, and much more. There are many customers, who are already using other financial services of FinMan. Many customers are using loans, vehicle insurance, life insurance, etc. The company is trying to cross-sell health insurance to their existing customers, who are already using other financial services, also FinMan is targeting new customers, who are visiting their website to check the health insurance premium plans. First, customers must visit any customer service agent or FinMan's website to enter the details such as demographic (age, city, region, etc.), other holding policies information, and based on those variables website will predict the premium policy to an individual. If the user fill-up the form, the website will assign a positive 'Response' for those users.

There may be cases when a person is already using other financial services of the FinMan, which will allow them to get some benefits while purchasing health insurance. If a new user visit website to check the premium price for health insurance, FinMan will also take care of other factors and will try to provide the best premium to the user and will give an offer to them, so that they can purchase other services as well.

This research paper will aim in predicting the 'Response' of the user based on the demographic and existing policies they are using. We are using Logistic regression to predict the 'Response'. Logistic Regression is a simple classification method and computationally less expensive when compared to decision tree, ensemble, boosting and bagging, and deep learning models.

II. Related Works

In [18] mining hidden patterns of disease with the medical data using different ML algorithms. The traditional data analysis considers only the doctor's experience while the new proposed system helps the doctor to predict disease precisely. This helps the insurance

provider and the patient to choose a more appropriate policy with the greatest benefits. The proposed system K-Nearest Neighbour and Decision Tree. These classification techniques provide 100% accuracy with the training dataset with the decision tree. For the testing dataset C4.5 showcases, 90.43% and K-Nearest Neighbour provides 76.96% of accuracy. Hence it can be observed that the decision tree in this case of prediction provides better accuracy. So considering a decision tree would be a good choice from the list of applicable techniques.

In [15] several machine learning algorithms are used to predict the presence of cancer at the initial stage. Cancers detected at the initial stages are 100% curable. Input dataset similar to what sonography would be required is used in this project. Factors like lump size, patient's age, BMI, glucose, leptin data points are considered. ML algorithms like logistic regression produced 83.33%, KNN with 82.12% and a Decision tree of 90% accuracy could be observed. Three different types of cancers like in the bladder, breast cancer, and cervical were checked upon. This system not just identifies the cancer presence but also suggests a suitable insurance policy for the patients. It is quite evident that logistic regression and KNN provide similar accuracy whereas the Decision tree outperforms all.

This [10] paper deals with the prediction of health insurance premium quotes and typically calculates the monthly fee that is to be paid by the policyholder. Creating such policies needs to consider factors like lifestyle, income, family medical history, own medical history, age, gender, etc. A proposed combination of K-mean and Elbow method was deployed to accurately predict and group the people in an optimal way to cluster based on some input parameters.

In [14] challenges in the health care claim process are highlighted. A lot of misuse in the Medicare that is the medical insurance system could be observed. Hence, the author proposed a new machine learning model to detect anomalous behavior with the data points in the medical insurance claims made. This model was built to identify when the doctors act disgracefully with the insurance claim to indicate the fraud activities. This model uses a multinomial Naïve Bayes algorithm for identifying such strange data values. Evaluation parameters like recall, precision, and F-score were used. The results depict that successful prediction was done with F-score over 0.9.

This [11] paper deals with, identifying the misrepresentation of manipulated results to take more benefits of healthcare insurance. To eschew such unfair benefits, a statistical approach to analyse the dataset on filtering the fraudulent activity is considered. Techniques like random forest regression, decision tree with X-Gradient Boost is used to determine the false insurance claims. The results depict that Random Forest exhibits 81% of accuracy whereas XGB of random sampling was successful with 86% of accuracy. These results provides more evidences on the applicable techniques that can be chosen for the healthcare lead prediction.

	ID	Region_Code	Upper_Age	Lower_Age	Holding_Policy_Type	Reco_Policy_Cat	Reco_Policy_Premium	Response
count	50882.000000	50882.000000	50882.000000	50882.000000	30631.000000	50882.000000	50882.000000	50882.000000
mean	25441.500000	1732.788707	44.856275	42.738866	2.439228	15.115188	14183.950069	0.239947
std	14688.512535	1424.081652	17.310271	17.319375	1.025923	6.340663	6590.074873	0.427055
min	1.000000	1.000000	18.000000	16.000000	1.000000	1.000000	2280.000000	0.000000
25%	12721.250000	523.000000	28.000000	27.000000	1.000000	12.000000	9248.000000	0.000000
50%	25441.500000	1391.000000	44.000000	40.000000	3.000000	17.000000	13178.000000	0.000000
75%	38161.750000	2667.000000	59.000000	57.000000	3.000000	20.000000	18096.000000	0.000000
max	50882.000000	6194.000000	75.000000	75.000000	4.000000	22.000000	43350.400000	1.000000

Fig. 1. Statistical Description about the dataset

III. METHODOLOGY

For this research work, we are following Knowledge Discovery in Database.

A. *Data Selection:*

For this project, we are using the Kaggle dataset. Data is already available in the structured format. Kaggle: https://www.kaggle.com/imsparsh/jobathon-analytics-vidhya

B. *Exploratory Data Analysis:*

Data set consists of the 50,882 number of rows and 14 columns. Statistical information about the data is studied and same is showed in the (Figure 1).

Data Cleaning and Data Pre-processing: In this step, we are cleaning data by removing Null values. Null values are replaced with high-frequency values. 'ID' column is dropped as it won't convey any information in finding the target variable. We also checked for the correlation among the input variables. There is no high correlation among the input variables (Figure 2). To examine outliers in the input variable we have plotted box plots (Figure 3). Outliers are handled by replacing them with the average value. (Figure 4) demonstrates the pair plot of numerical variables. 'Reco Policy Premium' and 'Upper Age', 'Reco Policy Premium' and 'Lower Age' shows the linear relationship. After observing 'Reco Policy' it can be concluded that premium policy is around 11,000 $.

C. *Data Transformation:*

To prepare data in a machine-understandable way, we are converting categorical variables into numerical by

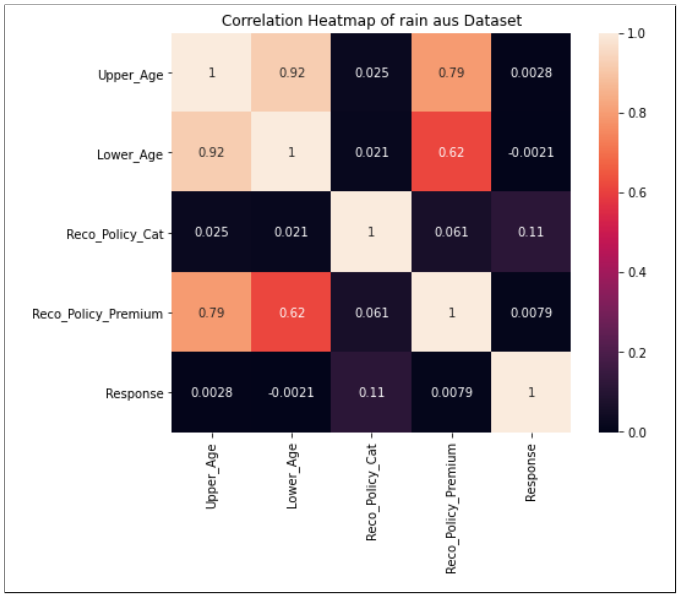

Fig. 2. Correlation among the input variables

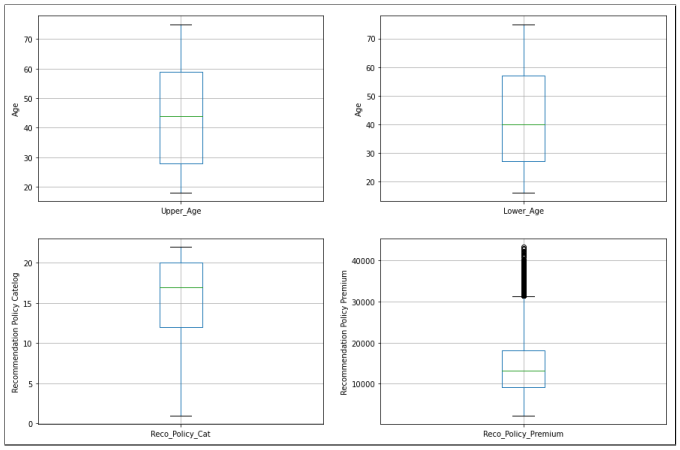

Fig. 3. Box plot of input variables

applying a one-hot encoding technique. Also, among numerical input variables, there is a variation in the magnitude range. To bring all the input variable's values into one level we have transferred values by applying the normalization technique. Dataset is divided into training and test data with test size being 20% of data.

D. Model Building:

We have used the Logistic Regression model to solve the classification problem. sklearn library being used for this work. We have used 'saga' as a solver which supports the Elastic-net regularization. Model is trained over 40705 instances of the input data and the same is tested over 10177 instances of data.

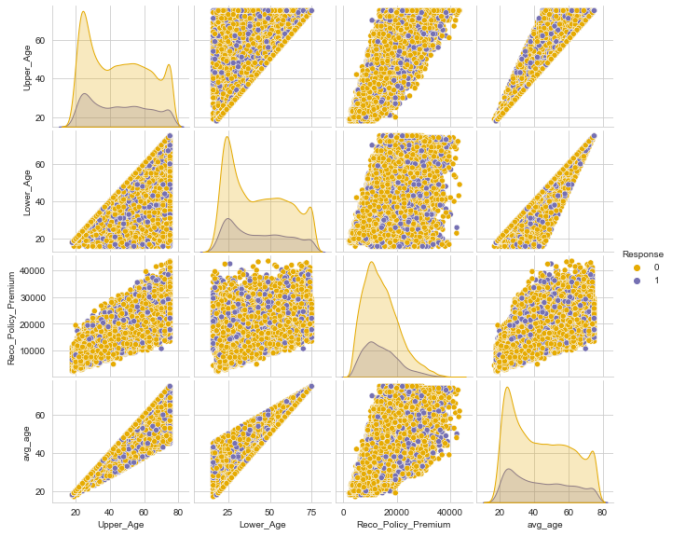

Fig. 4. Pair Plot

E. Evaluation and Interpretation:

To check how the model behaves with a different set of data as a train and test data, we tested model consistency over different sets of train and test split. To diagnose the model we ran it through 10 folds. Model is providing the same accuracy over different folds of the data as train and test split. Model is giving accuracy around 76% with a standard deviation of 0.0001 which is too small. That depicts model accuracy is consistent over different folds of data as a train and test. Then, to evaluate the model confusion metrics is constructed. Metrics like recall, precision, true positive rate are considered in evaluating the model performance in classification. The earlier model was producing 100% recall and 76 % precision. As the recall value is high, we reduced it by varying thresholds. It would be good to alert stakeholders even if there is a small probability that customer will purchase their premium. We tried varying the threshold to different values, among all 0.4 thresholds yielded good results. Confusion metrics report for the dame is shown in the 5. We can see in the confusion matrix report(5), recall value is reduced to 94%. Model is predicting the target results with an accuracy of 73% which is acceptable.

IV. Business Analysis

After a series of exploratory analysis using logistic regression, it is quite evident that demographic features influence the health insurance policy purchase. For instance, unmarried couples, house owners, customers between ages 20- 30 are more inclined towards purchasing the premium. Followed by the customers

	precision	recall	f1-score	support
0	0.76	0.94	0.84	7765
1	0.21	0.06	0.09	2412
accuracy			0.73	10177
macro avg	0.49	0.50	0.46	10177
weighted avg	0.63	0.73	0.66	10177

Fig. 5. Correlation among the input variables

around the senior age group (70 above) also show interest in policy. It can be said that people of both extreme ages are willing to buy health insurance policies suitable to their needs. The more accurate will be the premium policy the more the probability of customers purchasing the insurance policy.

V. Conclusion

Conclusively, nowadays nature and health conditions are very unpredictable. Healthcare costs and medical expenses are expensive. So, people tend to purchase health insurance to manage their medical expenses during emergency conditions. To predict the cross sell of health insurance based on other financial service purchases of Finman company, Logistic regression was implemented to analyse the response of customers based on the survey. The model was recorded with 73% of accuracy rate and favourable recall precision metric values.

References

[1] *Jödicke, A.M., Zellweger, U., Tomka, I.T. "Prediction of health care expenditure increase: how does pharmacotherapy contribute?" BMC Health Serv Res 19, 953 (2019)*

[2] *Hanafy, Mohamed. "Predict Health Insurance Cost by using Machine Learning and DNN Regression Models". International Journal of Innovative Technology and Exploring Engineering (2021). Volume-10. 137. 10.35940/ijitee.C8364.0110321*

[3] *Bertsimas, Dimitris & Bjarnadóttir, Margrét & Kane, Michael & Kryder, J. & Pandey, Rudra & Vempala, Santosh & Wang, Grant. "Algorithmic Prediction of Health Care Costs and Discovery of Medical Knowledge". Operations Research. (2007). 56. 10.1287/opre.1080.0619*

[4] *Yang, C., Delcher, C., Shenkman. "Machine learning approaches for predicting high cost high need patient expenditures in health care". BioMed Eng OnLine 17, 131 (2018)*

[5] *Sales AE, Liu C-F, Sloan KL. "Predicting costs of care using a pharmacy-based measure risk adjustment in a veteran population". Med Care. 2003; 41(6): 753–60*

[6] *Morid MA, Kawamoto K, Ault T. "Supervised learning methods for predicting healthcare costs: systematic literature review and empirical evaluation". AMIA Annu Symp Proc. 2017:1312–21*

[7] *Nidhi Bhardwaj, Rishabh Anand. "Health Insurance Amount Prediction" INTERNATIONAL JOURNAL OF ENGINEERING RESEARCH & TECHNOLOGY (IJERT) Volume 09, Issue 05 (May 2020)*

[8] *Kaur, Tarunpreet, "Factors affecting health insurance premiums: Explorative and predictive analysis" (2018). Creative Components. 72*

[9] *Forrest CB, Lemke KW, Bodycombe DP. "Medication, diagnostic, and cost information as predictors of high-risk patients in need of care management". Am J Manag Care. 2009;15(1):41–8*

[10] *T. Omar, M. Zohdy and J. Rrushi, "Clustering Application for Data-Driven Prediction of Health Insurance Premiums for People of Different Ages," 2021 IEEE International Conference on Consumer Electronics (ICCE), 2021, pp. 1-6, doi: 10.1109/ICCE50685.2021.9427598*

[11] *N. A. Akbar, A. Sunyoto, M. Rudyanto Arief and W. Caesarendra, "Improvement of decision tree classifier accuracy for healthcare insurance fraud prediction by using Extreme Gradient Boosting algorithm," 2020 International Conference on Informatics, Multimedia, Cyber and Information System (ICIMCIS), 2020, pp. 110-114, doi: 10.1109/ICIMCIS51567.2020.9354286*

[12] *Y. Ren, K. Zhang and Y. Shi, "Survival Prediction from Longitudinal Health Insurance Data using Graph Pattern Mining," 2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), 2019, pp. 1104-1108, doi: 10.1109/BIBM47256.2019.8983290*

[13] *Y. Xie, G. Schreier, D. C. W. Chang, S. Neubauer, S. J. Redmond and N. H. Lovell, "Predicting number of hospitalization days based on health insurance claims data using bagged regression trees," 2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, 2014, pp. 2706-2709, doi: 10.1109/EMBC.2014.6944181*

[14] *A. Bauder, T. M. Khoshgoftaar, A. Richter and M. Herland, "Predicting Medical Provider Specialties to Detect Anomalous Insurance Claims," 2016 IEEE 28th International Conference on Tools with Artificial Intelligence (ICTAI), 2016, pp. 784-790, doi: 10.1109/ICTAI.2016.0123*

[15] *S. S. More, V. B. Lobo, R. M. Laban, S. Panchal, M. Patil and G. Pathak, "Cancer Prediction and Insurance Eligibility using Machine Learning Techniques," 2020 5th International Conference on Communication and Electronics Systems (ICCES), 2020, pp. 725-730, doi: 10.1109/ICCES48766.2020.9137936*

[16] *R. A and K. Rohini, "A Survey about Role of Data Mining Techniques and its Applications in Healthcare Sector," 2021 2nd International Conference on Intelligent Engineering and Management (ICIEM), 2021, pp. 277-281, doi: 10.1109/ICIEM51511.2021.9445300*

[17] *E. K. Hashi, M. S. U. Zaman and M. R. Hasan, "An expert clinical decision support system to predict disease using classification techniques," 2017 International Conference on Electrical, Computer and Communication Engineering (ECCE), 2017, pp. 396-400, doi: 10.1109/ECACE.2017.7912937*

[18] *18. J. M. Johnson and T. M. Khoshgoftaar. "Hcpcs2Vec: Healthcare Procedure Embeddings for Medicare Fraud Prediction," 2020 IEEE 6th International Conference on Collaboration and Internet Computing (CIC), 2020, pp. 145-152, doi: 10.1109/CIC50333.2020.00026*

[19] *Chan, Nicholas Khin-Whai and Lee, Angela Siew-Hoong and Zainol, Zuraini. "Predicting Employee Health Risks using Classification Ensemble Model" 2021 Fifth International Conference on Information Retrieval and Knowledge Management (CAMP), 2021, pp. 52-58, doi: 10.1109/CAMP51653.2021.9498106*

[20] *Le Nguyen, Tran and Do, Thi Thu Ha. "Artificial Intelligence in Healthcare: A New Technology Benefit for Both Patients and Doctors" 2019 Portland International Conference on Management of Engineering and Technology (PICMET), 2019, pp. 1-15, doi: 10.23919/PICMET.2019.8893884*