Smart Medical Insurance Cost Prediction using Machine Learning

Dr. R. Vijayarajeswari
B.Tech, M.E., Ph.D.,
Professor
Department of Computer Science and Engineering
Velalar College of Engineering and Technology
Thindal, Erode - 638012

Keerthana S.
Department of Computer Science and Engineering
Velalar College of Engineering and Technology
Thindal, Erode - 638012

Gunabharathi M. K.
Department of Computer Science and Engineering
Velalar College of Engineering and Technology
Thindal, Erode - 638012

Indhumathi S.
Department of Computer Science and Engineering
Velalar College of Engineering and Technology
Thindal, Erode

Abstract—Predicting who will get health insurance is a big problem in healthcare because it can help find people who might not have insurance and might benefit from specific outreach and education. In this project, we investigate health insurance prediction using Multi-Layer Preceptron and Linear Regression, two well-known machine learning models. Individual's age, gender, income, occupation, and medical history were pre-processed and divided into training and testing sets. An MLP is a type of Artificial Neural Network that is made up of multiple layer of interconnected nodes, or "neurons". After that, we used a variety of metrics like accuracy, precision, recall, and F1-score to evaluate the model's performance on the test data after they had been trained on the training data. MLP outperformed Linear Regression in terms of accuracy and F1-score, as demonstrated by our findings, when it comes to predicting health insurance status. MLP, on the other hand, required more computational resources and required more time to train than Linear Regression, particularly for larger datasets or models that were more complex.

Keywords—Multi-Layer Preceptron, Linear Regression, Artificial Neural Network, neurons

I. INTRODUCTION

Heath insurance is an essential component of healthcare access, because it protects the high costs of medical care. However, a significant portion of the population is still uninsured or underinsured, despite the significance of health insurance. In 2018, approximately 27.5 million people in the United States lacked health insurance, according to the National Center for Health Statistics. Insurance prediction is a significant issue in healthcare because it can assist in identifying individuals who are at risk of not having health insurance and may benefit from specialized outreach and education. Based on a variety of input characteristics, such as age, gender, income, occupation, and medical history, machine learning models can be used to predict health insurance status.

Considering the rising cost of medical care, each person needs health insurance. Emotional and financial trauma can be severe in medical emergencies. As a result, in unpredictability, a health insurance policy can assist in mitigating financial risks. The current insurance system, on the other hand, is expensive because thousands of people pay the premiums, but few people pay the premiums, but few people file claims.

Additionally, the claim settlement procedure is exhaustingly lengthy. In this article, we focus on creating a machine learning for the health. In this article, we investigate the potential advantages of MLP and Linear Regression for insurers and policyholders in health insurance prediction. We talk about the factors that affect health insurance premiums and claim likelihood. Insurers can improve individual's access to healthcare and promote the overall health of populations by utilizing machine learning algorithms to provide more individualized and cost-effective health insurance services.

A. Machine Learning

A subfield of artificial intelligence known as machine learning focuses on the development of computer systems that are capable of learning from data and making predictions or decisions based on that data. Without being explicitly programmed to do so, computers with machine learning algorithms are able to identify patterns in data and use those patterns to make predictions or take actions. There are a few kinds of AI calculations, including managed learning, unaided learning, and support learning. In supervised learning, a model is trained to make predictions on new, unlabeled data for which the correct answers are known. The process of training a model on unlabeled data in order to identify patterns and group data points that are unsupervised learning. A model is trained to learn through trial and error in the reinforcement learning by receiving feedback in the form of rewards or punishments.

II. LITERATURE REVIEW

A. Relational Random Forests based on Random Relational Rules

Propositional learning has demonstrated that Grant Anderson R random Forests perform exceptionally well. FORF is a relational data version of Random Forests. In order to generate Random Forests over relational data, we examine the drawbacks of FORF and propose an alternative method called R4F. R4F utilizes haphazardly produced social standards as completely independent Boolean test insisde every hun in a tree and in this manner can be seen as a case of dynamic propostionalitt. The way R4F is put into practice makes it possible for all of the ensemble's trees to grow simultaneously or in parallel in a shared, but still single-threaded, efficient manner. Both FORF and the combination of static propostionalization and standard Random Forests are favourably compared in the Random Forests [Breiman, 2001] are one of the best-performing off-the-shelf methods, competing with both support vector machines and boosted decision trees, Additionalyy, various strategies for tree initialization and splitting of nodes, as well as the ensemble size, diversity, and computational complexity of R4F that result form these actions, are the subjects of investigation. Random Forests an densemble methods in general have not received a lot of attention in relational learning.

B. An Ensemble Random Forest Algorithm for Insurance Big Data Analysis

Weiwei Lin Directly applying big data techniques to realistic business data typically deviations from the business objectives for a variety of reasons, including the unbalanced distribution of business data, the absence of user features, and numerous other factors. Classification algorithms like Logistic Regression, SVM, and other similar techniques make it challenging to model the insurance industry data. In this paper, we use an ensemble learning algorithm and a heuristic bootstrap sampling method to mine large amounts of insurance company data. We then propose an ensemble random forest that makes use of Spark's parallel computing capability and memory-cache mechanism. To use the proposed algorithm, we gathered insurance industry data from China Life Insurance Company and analyzed potential clients. The algorithm's performance is evaluated with F-Measure and G-mean. The experiment shows that the ensemble random forest algorithm performs better than SVM and other classification algorithms in the imbalanced data, both in terms of performance and accuracy. This means that it can be used to improve product marketing accuracy over the conventional artificial approach. The third industrial revolution, which is represented by information technology, entered a new era with the advent of big data.

C. Pricing insurance contracts and determining optimal capital of insurers

Hong Mao: In this paper, we discuss how to figure out the best capital level for an insurer and extend Kliger and Levikson's method for pricing insurance contracts by taking into account the effect that a company's level of capital has on the price of insurance contracts. Under the condition of optimal capital level, out analysis reveals that the optimal number of insured, maximum value of expected profit increase, premium (price), and probability of insolvency decrease. In addition, we apply the previous strategy to multi-line cases and figure out the best prices, policy numbers, and capital allocation strategy for multi-line insurance contracts. We assume that the insurer's goal is maximize its net expected profit, which is calculated as the difference

between the insurer's expected loss from insolvency and the expected net revenue from insurance contract sales. The fact that we add the capital cost factor to the expected net revenue and the capital factor to the probability of insolvency distinguished our model from their model.

D. Towards definition of the risk premium function

Nikola Krear: The ability of the market actor to accurately forecast the price of electricity is essential for successful trading in electricity markets. The residual, which contains a risk premium, is one of the price drivers that is used in the fundamental electricity price models. Other price drivers also provide market information. In the past, researchers investigating risk premium ignored intraday information, making it difficult to accurately determine risk premium, and instead focused primarily on daily spot price levels. A new KGB method for modelling risk premium is presented in this paper. It is based on an "ex-ante" approach and is focused on a yearly product. The novel KGB Model and its linearized formulation, the KGB Linear Model, are used in this strategy, which makes it possible to capture the impact of renewable energy sources on risk premiums. An insight into the impact of RES generation on the development of risk premiums was obtained through the utilization of the four primary drivers of the KGB Linear Model. Using historical data from the German electricity market, the method was tested. The overall influence of PV production share on risk premium is greater than that of wind production share, as both increase risk premium dur to their variability and uncertainty, as shown by the results for the periods from 2010 to 2014. The electricity spot price's volatility and mean reversion are two of the most important characteristics.

E. Real Estate price prediction with Regression and Classification

Hujia Yuet .al., has proposed in this system Housing prices are an important reflection of the economy, and housing price ranges are of great interest for both buyers and sellers. In this project, house prices will be predicted given explanatory variables that cover many aspects of residential houses. As continuous house prices, they will be predicted with various regression techniques include Lasso, Ridge, SVM, and Random Forest regression; as individual price ranges, they will be predicted with classification methods including Naïve Bayes, logistic regression, SVM classification, and Random Forest classification. We will also perform PCA to improve the prediction accuracy. The goal of this project is to create a regression model and a classification model that are able to accurately estimate the price of the house given the features. Thus, large amounts of features enable us to explore various techniques to predict.

III. METHODS

A. Multi-Layer Preceptron (MLP)

The term "MLP" refers to an artificial neural network that is utilized in machine learning. An input layer, one or more hidden layers, and an output layer make up this feed forward neural network's multiple layers. The MLP introduces non-linearity to the model through an activation function, allowing it to learn complex data patterns. In MLPs, the sigmoid, the hyperbolic tangent (tanh), and the rectified linear unit (ReLU) are the activation functions that are utilized the most frequently. A supervised learning method is used to train the MLP, in which the weights of the connections between neurons are changed during training to reduce the difference between the predicted and actual output. This cycle is known as back propagation. The classification of images, speech recognition natural language processing, and financial forecasting are just a few of the many uses for MLPs. However, they can be time-consuming computationally and necessitate a significant amount of training data for optimal performance.

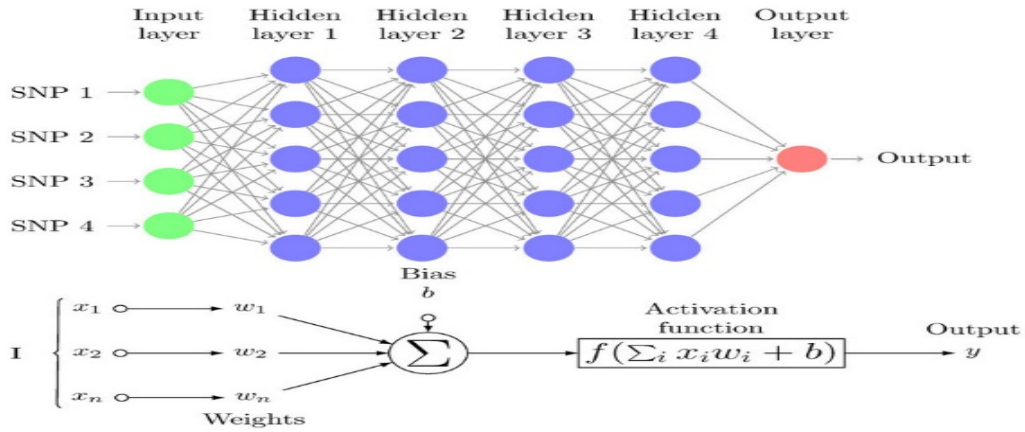

Fig. 1. Working of Multi-Layered Preceptron in the layered manner

1) Working:

a) Forward propagation Algorithm: In order to proceed we need to improve the notation we have been using. That for, for each layer $1 \geq l \geq L$, the activators and outputs are calculated. Where l and L are the number of layers and number of neurons respectively.

b) Calculation of Error Function: It is used to measure performance locality associated with the results produced by the neurons in output layer and the expected result.

c) Activation function: The activation functions like sigmoid function, hyperbolic tangent function, ReLU functions are calculated.

d) Backward Propagation:

- In output layer: Calculate error in output layer and update all weight between hidden and output layer then update bias value in output layer.
- In input layer: calculate error in hidden layer, update all weight between hidden and output Layer and update bias value in output layer.

B. Linear Regression

Linear Regression is the supervised machine learning model in which the model finds the best fit linear line between the independent and dependent variables. It predicts the health insurance cost based on the relationship between the two variables. In our dataset, age, sex, BMI, children, smoker, region are the independent variables. The charges are the variables which depends on all other variables of the dataset. It provides the final cost prediction of the insurance model with great accuracy.

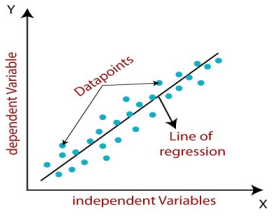

Fig. 2. Graphical representation of Linear regression

IV. EXPERIMENTAL SETUP

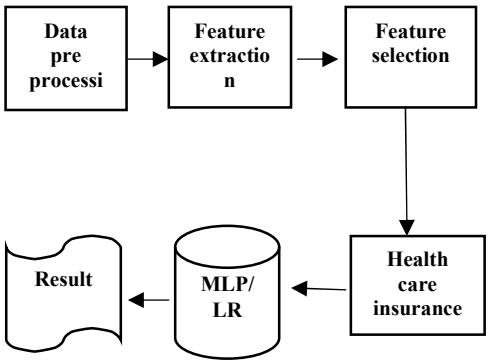

Fig. 3. System Architecture

A. Load input data

The health insurance dataset must be loaded into the environment as the first step. Age, gender, medical history, lifestyle factors, and insurance premiums ought to be included in the dataset. Reading the health insurance dataset into the environment is necessary for loading the input data. The health insurance dataset can be loaded from a .csv file or a database. Input variables like age, gender, occupation and pre-existing medical conditions, among others, ought to be included in the dataset and an output variable that tells you if the individual has health insurances. The dataset must be examined for errors, outliers, and missing values. Using statistical techniques, we can either remove the rows with missing values or infer them. For this purpose, we can employ approaches like standard scaling or Min-Max scaling.

B. Data Preprocessing

It involves a number of tasks, including handling missing values, encoding categorical variables, scaling numerical features, and removing duplicates. Repetitive observations can lead to model bias, so it's important to get rid of duplicates. For model training and inference to be error-free, missing value handling is essential. To resolve missing values, imputation methods like mean, median, and mode imputation can be used. The ranges, units, and scales of the dataset's input variables may vary. In order for machine learning algorithms to interpret the dataset's categorical variables, they must be encoded as numerical values. One-hot encoding and label encoding are two techniques that can be used to accomplish this. Numerical features must also be scaled to ensure that they all contribute equally to the model and are the same size. Standardization and normalization are two methods that can be used to accomplish this. The dataset can be made suitable for machine learning modelling by carrying out these pre-processing steps, which will result in predictions that are more accurate and dependable.

C. Feature Extraction

The process of transforming the input data into a set of new features that can help the model's predictive performance is known as feature extraction. Rather than involving age as a constant variable, we can bunch people into various age groups (e.g., 20-30, 31-40, and so forth) to record the variable risk profiles of various age groups. For each pre-existing medical condition (such as diabetes, hypertension, etc.,), we are able to generate binary features, to record hot these conditions affect your health insurance status. We are able to develop a new feature that will allow us to record the number of family members who are covered by the health Insurance policy.

D. Prediction of Health Insurance

Machine learning methods like Linear regression and Multi-Layer Preceptron health insurance prediction are common. After loading and pre-processing the dataset, relevant features are extracted, the dataset is divided into training and evaluated. Multiple layers of neurons make up the MLP model, which is capable of modeling intricate connections between input and output variables. To get the best performance, it needs to adjust hyper parameters like the number of layers, the number of neurons in each layer, and the learning rate. Based on the input variables, the model can be used to predict whether an individual has health insurance after it has been trained. Based on the input variables, logistic regression calculates the probability that an individual has health insurance. It has fewer hyper parameters to tune and is a simpler model than MLP.

V. Result Analysis

For our proposed problem statement, we talked about some traditional regression models in this paper , moving on to other techniques like MLP or LINEAR REGRESSION that will be addressed in subsequent work. On top of model evaluation, a number of optimization methods, like the Genetic Algorithm and the Gradient Descent Algorithm, can be used. Because some of the features may be omitted when predicting the charges, we can also apply some features selection techniques to our dataset before training our model to achieve a high accuracy value. In addition, a well-balanced dataset with a greater number of observations is required for a model to perform well. This will reduce the model's variability in the future in the event that we obtain more data than the model can handle.

TABLE I. COMPARTSON OF VARIOUS ALGORITHMS

Algorithm	Accuracy
CNN	62.86
ANN	75.86
K-SVM	75.82
MLP AND LR	80.97
DNN	75.86

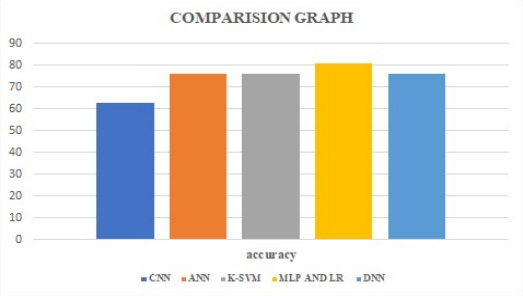

Fig. 4. Comparison Graph for various algorithms

VI. CONCLUSION

In conclusion, there are significant advantages for both insurers and policyholders when machine learning algorithms like Linear regression and MLP are used in health insurance prediction. Insurers can better comprehend the risks associated with providing coverage and adjust their services accordingly by analyzing historical data and predicting health insurance premiums and claim likelihood. Improved access to healthcare, improved health outcomes, and more cost-effective insurance services for policyholders are all possible outcomes of this. By automating numerous aspects of the insurance process and identifying fraudulent claims, the use of machine learning algorithms can also help insurers save money and increase efficiency. The potential for machine learning algorithms to transform the health insurance industry is enormous due to the availability of large amounts of data and the continuing development of technology. However, it is essential to keep in mind that the application of machine learning algorithms to the predictions of health insurance has some drawbacks. It is necessary to carefully consider and address privacy concerns as well as potential biases in the data that is utilized to train the algorithms.

ACKNOWLEDGMENTS

We would acknowledge our guide for development of the Smart Medical Insurance cost prediction model with full support and guidance. We would like to thank the user who contribute their data to study.

REFERENCES

[1] Azzone M., Barucci E., Giuffra Moncayo G., Marazzina D. *A Machine Learning Model for Lapse Prediction in Life Insurance Contracts*. Expert Syst. Appl. 2022;191:116261. doi: 10.1016/j.eswa.2021.116261.

[2] B. Milovic and M. Milovic, "*Prediction and decision making in health care using data mining,*" Kuwait Chapter of the Arabian Journal of Business and Management Review, vol. 1, no. 12, 2021.

[3] Ch. Anwar ul Hassan, Jawaid Iqbal, Saddam Hussain, Hussain Al Salman, Mogeeb A. A. Mosleh and Syed Sajid Ullah, "A Computational Intelligence Approach for Predicting Medical Insurance Cost", *Hindawi*, 2021.

[4] Ejiyi C.J., Qin Z., Salako A.A., Happy M.N., Nneji G.U., Ukwuoma C.C., Chikwendu I.A., Gen J. Comparative *Analysis of Building Insurance Prediction Using Some Machine Learning Algorithms*. Int. J. Interact. Multimed. Artif. Intell. 2022;**7**:75–85. doi: 10.9781/ijimai.2022.02.005.

[5] Fauzan M.A., Murfi H. *The Accuracy of XGBoost for Insurance Claim Prediction.* [(accessed on 9 May 2022)];Int. J. Adv. Soft Comput. Appl. 2018 **10**:159–171.

[6] Gupta, S., & Tripathi, P. (2016, February). *A leading trend of data analytics with health insurance in India.* In 2016 International Conference on Innovation and Challenges in CS (ICICCS-INBUSH) (pp. 64-69). IEEE

[7] J.N.V.R. Swarup Kumar et al., "Evolution of Television Shows Popularity based on T witter Data using Sentiment Analysis Techniques", *International Journal of Engineering and Advanced Technology (IJEAT)*, vol. 8, no. 6S2, August 2019, ISSN 2249-8958.

[8] Karthik Venkat, Tarika Gautam, Mohit Yadav and Mukhtiar Singh, "An XGBoost Ensemble Model for Residential Load Forecasting", *Proceedings of International Conference on Intelligent Computing Information and Control Systems*, pp. 321-334, 2021.

[9] K. Shaukat, F. Iqbal, T. M. Alam et al., "The impact of artificial intelligence and robotics on the future employment opportunities," Trends in Computer Science and Information Technology, vol. 5, no. 1, pp. 50–54, 2020.

[10] Kumar Sharma D., Sharma A. Prediction of Health Insurance Emergency Using Multiple Linear Regression Technique. Eur. J. Mol. Clin. Med. 2020;7:98–105.

[11] Nagarajan C., Neelakrishnan G., Akila P., Fathima U., Sneha S. "Performance Analysis and Implementation of 89C51 Controller Based Solar Tracking System with Boost Converter" Journal of VLSI Design Tools & Technology. 2022; 12(2): 34–41p

[12] M. A. Morid, K. Kawamoto, T. Ault, J. Dorius and S. Abdelrahman, "Supervised Learning Methods for Predicting Healthcare Costs: Systematic Literature Review and Empirical Evaluation", *AMIA Annual Symposium Proceedings*, vol. 2017, pp. 1312, 2017.

[13] M.David*,"Auto insurance premium calculation using generaized linear models,"Proc. Econ. Finance,* vol. 20, 00. 147-156 Jan. 2015.

[14] M. U. Ghani, T. M. Alam, and F. H. Jaskani, *"Comparison of classification models for early prediction of breast cancer,"* in Proceedings of the International Conference on Innovative Computing (ICIC), Lahore, Pakistan, November.2021.

[15] N. K. Kumar, G. S. Sindhu, D. K. Prashanthi and A. S. Sulthana, "Analysis and Prediction of Cardio Vascular Disease using Machine Learning Classifiers", *2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS). 2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS)*, March 2020.

[16] Pesantez-Narvaez, J., Guillen, M., & Alcañiz, M. (2019). *Prediction vehicles insurance claims using telematics data—XGBoost versus logistic regression. Risks*, 7(2), 704. Singh, R., Ayyar, M. P., Pavan, T. S., Gosain, S., & Shah, R. R. (2019, September). Automating Vehicle Car Insurance Claims Using Deep Learning Techniques. In 2019 IEEE Fifth International Conference on Multimedia Big Data (BigMM) (pp. 199-207). IEEE.

[17] Philipp Drewe-Boss, Dirk Enders, Jochen Walker and Uwe Ohler, "Deep learning for prediction of population health costs", *BMC Medical Informatics and Decision Making*, vol. 22, no. 1, pp. 1-10, 2022

[18] C. Nagarajan, G.Neelakrishnan, R. Janani, S.Maithili, G. Ramya "Investigation on Fault Analysis for Power Transformers Using Adaptive Differential Relay" Asian Journal of Electrical Science, Vol.11 No.1, pp: 1-8, 2022.

[19] C.Nagarajan and M.Madheswaran - 'Experimental verification and stability state space analysis of CLL-T Series Parallel Resonant Converter' - Journal of ELECTRICAL ENGINEERING, Vol.63 (6), pp.365-372, Dec.2012.

[20] P. Motarwar, A. Duraphe, G. Suganya and M. Premalatha, "Cognitive Approach for Heart Disease Prediction using Machine Learning", *2020 International Conference on Emerging Trends in Information Technology and Engineering (Ic-ETITE). 2020 International Conference on Emerging Trends in Information Technology and Engineering (ic-ET ITE)*, February 2020.

[21] R. SwamySirisati, M. S. Rao and S. Thonukunuri, "Analysis of Hybrid Fusion-Neural Filter Approach to detect Brain Tumor", *2020 Sixth International Conference on Parallel Distributed and Grid Computing (PDGC)*, pp. 460-464, 2020.

[22] R. Tkachenko, H. Kutucu, I. Izonin, A. Doroshenko and Y. Tsymbal, "Non-iterative Neural-like Predictor for Solar Energy in Libya", *Proceedings of the 14th International Conference on ICT in Education Research and Industrial Applications. Kyiv Ukraine May 14-17 2018*, vol. 2105, pp. 35-45, 2018

[23] Nagarajan and M.Madheswaran - 'Experimental Study and steady state stability analysis of CLL-T Series Parallel Resonant Converter with Fuzzy controller using State Space Analysis'- Iranian Journal of Electrical & Electronic Engineering, Vol.8 (3), pp.259-267, September 2012.

[24] Yerpude, P., Gudur, V.: Prediction modeling of crime dataset using data mining. Int. J. Data Min. Knowl. Manage. Process (IJDKP) 7(4) (2017)

[25] C.Nagarajan and M.Madheswaran - 'Performance Analysis of LCL-T Resonant Converter with Fuzzy/PID Using State Space Analysis'- Springer, Electrical Engineering, Vol.93 (3), pp.167-178, September 2011.

[26] C.Nagarajan and M.Madheswaran - 'Stability Analysis of Series Parallel Resonant Converter with Fuzzy Logic Controller Using State Space Techniques'- Taylor & Francis, Electric Power Components and Systems, Vol.39 (8), pp.780-793, May 2011.

.