Predicting Health Expenses Using Linear Regression

Rishi Raj Makkar
Raj Kumar Goel Institute of Technology

Mrityunjay
Raj Kumar Goel Institute of Technology

Mayank Pandey
Raj Kumar Goel Institute of Technology

Birendra Kumar Saraswat
Computer Science & Engineering
Raj Kumar Goel Institute of Technology

***Abstract*—A branch of artificial intelligence called machine learning employs statistical and computational methods to let computers "learn" from data instead of having to be explicitly programmed. It entails utilizing a model that has been trained on a data set to make predictions or judgements without being explicitly instructed on how to do so. The idea is for the model to be able to recognize patterns in the data and extrapolate those patterns to brand-new, unexplored data. Various applications, such as computer vision, natural language processing, and recommendation systems, use machine learning extensively.**

Keywords—Linear Regression, regression, style, styling, insert

I. INTRODUCTION (*HEADING 1*)

Healthcare costs continue to be a major concern for individuals, families and governments around the world. Predictive modelling using machine learning provides valuable insight into future healthcare costs and helps healthcare organizations plan their budgets more effectively. The goal of this research is to apply machine learning techniques to predict healthcare costs based on relevant factors such as demographic information, medical history and lifestyle choices.

This study collects data from a variety of sources, including Kaggle records and insurance claims. The data are pre-processed to handle missing values and outliers and split into training and test datasets. The machine learning algorithms used in this study are trained on training data and evaluated on test data to determine their accuracy in predicting medical costs. The results of this study will help healthcare organizations make more informed decisions about budgeting and resource allocation by understanding how machine learning can be used to predict healthcare costs.

A branch of artificial intelligence called machine learning employs statistical and computational methods to let computers "learn" from data instead of having to be explicitly programmed. It entails utilising a model that has been trained on a data set to make predictions or judgements without being explicitly instructed on how to do so. The idea

is for the model to be able to recognise patterns in the data and extrapolate those patterns to brand-new, unexplored data. Various applications, such as computer vision, natural language processing, and recommendation systems, use machine learning extensively.

Global healthcare expenditures are rising, making it more crucial than ever to develop efficient strategies to control them. Machine learning-based predictive modelling can offer useful insights into upcoming healthcare costs and aid healthcare companies in budget planning.

Medical health expenses prediction is a major problem. Linear regression , supervised learning , un-supervised learning, fuzzy logic, reinforcement learning are some of the important machine learning approaches used to predict health expenses.

Predicting medical bills is a significant issue. Some of the key technologies utilised in this include reinforcement learning, fuzzy logic, supervised learning, and linear regression.

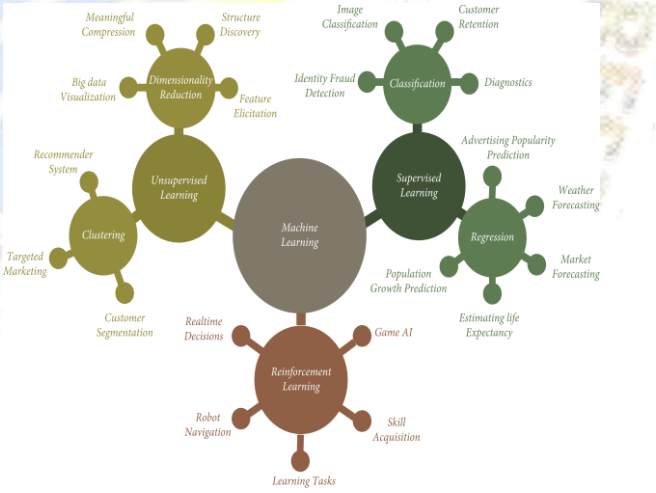

Fig. 1 Types of Machine Learning

Linear regression is a statistical technique that is widely used to predict continuous variables. It is a linear approach to modeling the relationship between a dependent variable and one or more independent variables. In this study, we will use linear regression to predict medical expenses.

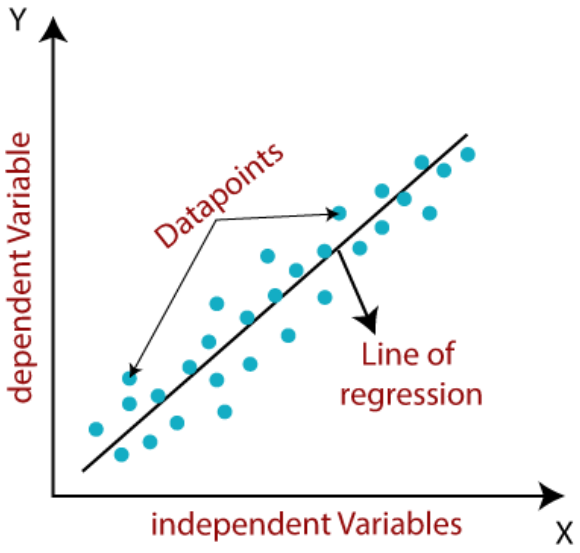

Fig. 2 Linear Regression

II. RELATED WORK

In year 2019, Rama Devi[3] talked about benefits and dynamics factors that impact the implementation of AI and ML like claim processing, price optimization,

Personalized Marketing, Customer segmentation, Fraud Prevention, Risk Management.

In 2020, Sam Goundar[2] talked about Artificial Neural Network. Using data from 1999 to 2017, a neural network model was trained.

After analyzing the implementation's outcomes, it is clear that forecasting medical claims data is feasible and can yield reliable outcomes when employing artificial neural network models.

In 2021,Ch.Anwar ul Hussan[1] used different methods to forecast the cost of health insurance.

The KAGGLE repository was used to obtain the medical insurance dataset, which was then used to train and test the machine learning algorithms.

This dataset underwent preprocessing, feature engineering, data splitting, regression, and evaluation processes before being subjected to regression analysis.

Stochastic Gradient Boosting (SGB) had a high accuracy of 86% and an RMSE of 0.340, according to the final results.

In 2020, Belisario Panay[4] discussed different approaches to predict health care cost by using Weighted Evidential Regression.

1.Classifying patients into cost buckets is one strategy for enhancing performance.

2.Use of a nearness to death feature could be another strategy because it has been seen that costs increase significantly when it is used.

3.Deep learning techniques can also be used to try to tackle the problem of predicting the cost of medical treatment,

but this goal may only be possible with the availability of a larger dataset.

In 2021, Aman Kharwal[10] talked about using Python to anticipate health insurance premiums using machine learning. He takes dataset from Kaggle for the purpose of predicting health insurance premiums. After experimenting with various machine learning algorithms, he discovered that the random forest method performs the task the best.

In 2019, Eline M. van den Broek-Altenburg and Adam J. Atherly[11] suggested that other, non-financial elements, such as consumer attitudes, may be significant when deciding on a health insurance plan. It may seem strange to mention "fear" in regard to selecting a health insurance plan, but the fundamental economic theory of health insurance holds that risk aversion is a major driver of insurance purchases. In some ways, the word "fear" in common usage simply means "risk-averse". In another sense, however, this study clarifies the nature of risk aversion and contends that customers lack trust in their decisions and exhibit fear of unfavorable health outcomes and unexpected costs.

III. PROPOSED WORK

Our objective is to find a way to estimate the value in the “charges” column using the values in the other columns. If we can do so for the historical data, then we should be able to estimate charges for new customers too, simply by asking for information like their age, sex, BMI (Body Mass Index), number of children, smoking habits and region.

Our model predicts the right expense, which is based off of the existing data which gives the customer ensured of transparency and is not biased because all the work is done by the algorithms and the decision of predicting any data is completely unbiased as there is no human intervention. This model helps people to encourage to be insured and that too at a price that everyone can agree. Our model also makes sure that the health insurance company does not take unfair advantage of their customers and are also not underpaid. This insures credibility and most importantly trust of the system.

Creating a machine learning model which predicts medical expenses of the individual will provide accurate and precise solution to the problem. We use the data of existing customers and put it into machine learning model which uses linear regression to calculate the prediction of medical expenses of the customer. In this machine learning model, we try to find the relationship between various variables. We also draw some visual representation between different factors using histogram, bar plot, scatterplot, heatmap and py plot so that it is easier to understand the relationship between them for example a simple histogram representing a visual of distribution of BMI(Body Mass Index) reveals that there is no meaningful relation between higher or lower BMI with the frequency of people taking a medical insurance but on the other hand there is a strong relation between the people who smoke and their annual medical charges. With a simple histogram it is very easy to conclude that people who smoke certainly are prone to shelling out higher annual medical charges as compared to the people who don’t smoke. It is also seen that smokers also don’t necessarily take medical insurance at all which can be due to various reasons like lack of health awareness. Unnecessary data is removed from our

dataset which is then put into our machine learning model to predict medical expenses of the customer.

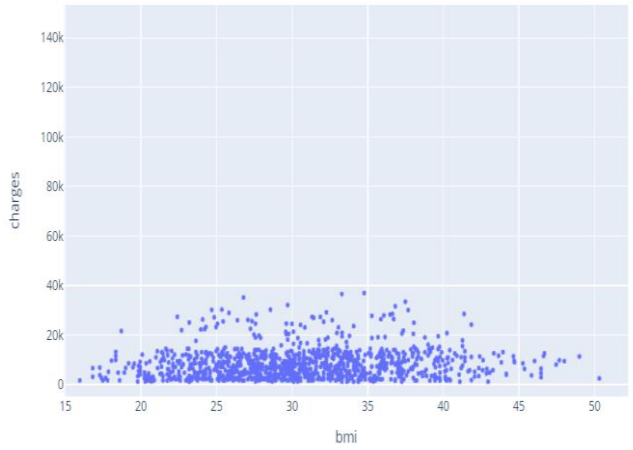

Fig.3 BMI vs Charges of health expenses.

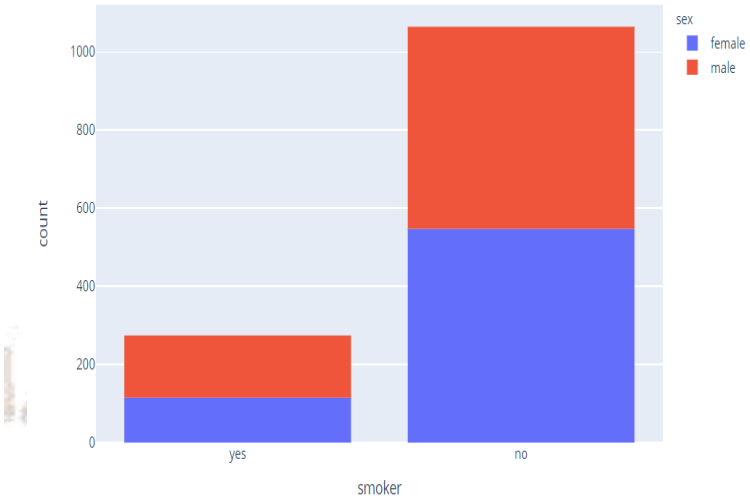

Fig.5 Smoker vs Count

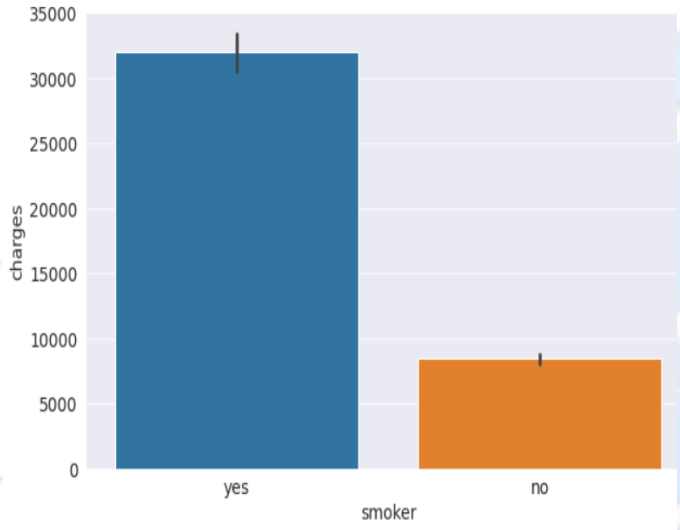

Fig.4 Smoker vs Annual medical charges.

We are creating an automated system to estimate the annual medical expenditure of new customers for an insurance company using information such as their age, sex, BMI, children, smoking habits and region of residence, finding the appropriate correlation between various parameters with the help of supervised machine learning. The model used here is linear regression, the independent and dependent variables causal links can be inferred using regression analysis. Regressions by themselves, it should be noted, only illuminate connections between a dependent variable and a group of independent variables in a given dataset. Researchers must carefully explain why existing correlations have predictive value in a new context or why a link between two variables has a causal meaning before using regressions for prediction or to infer causal relationships, respectively. When attempting to estimate causal linkages using observational data, the latter is particularly crucial. which is used to predict medical expense using the information of the customers provided. For the application of regression Al analysis, we use scikit learn library in which the method of Support Vector Classification can be extended to solve regression problems. This method is called Support Vector Regression. The model produced by support vector classification (as described above) depends only on a subset of the training data, because the cost function for building the model does not care about training points that lie beyond the margin. Analogously, the model produced by Support Vector Regression depends only on a subset of the training data, because the cost function ignores samples whose prediction is close to their target. A support vector machine constructs a hyper-plane or set of hyper-planes in a high or infinite dimensional space, which can be used for classification, regression or other tasks. Intuitively, a good separation is achieved by the hyper-plane that has the largest distance to the nearest training data points of any class (so-called functional margin), since in general the larger the margin the lower the generalization error of the classifier. The figure below shows the decision function for a linearly separable problem, with three samples on the margin boundaries, called “support vectors “Our model predicts the right expense, which is based off of the existing data which gives the customer ensured of transparency and is not biased because all the work is done by the algorithms. Our model helps the people to live a healthy and safe lives.

IV. CONCLUSION

There were difference ways and technologies and methods used in research papers like according to Health Insurance Claim Prediction Using Artificial Neural Networks[1]: In the insurance business, there are importantly two things which are required to be considered for analysing losses, one of them being frequency of loss and the other one is severeness of loss. Previous research investigated the use of artificial neural networks (ANNs) to develop models as aids to the insurance underwriter when finding acceptability and price on insurance policies. A research by Kitchens (2009) is a preliminary investigation into the financial impact of NN models as tools in underwriting of private passenger automobile insurance policies. In the past, research by Mahmoud et al. (2013) and Majhi (2018) on recurrent neural networks (RNNs) have also demonstrated that it is really an improved forecasting model for time series. To demonstrate this, NARX model (nonlinear autoregressive network having exogenous inputs), is a recurrent dynamic network was compared and tested against feed forward artificial neural network.

A study done in 2018, for developing an algorithm for stroke prediction. A National Health Insurance Database Study in Korea[2]: The second leading cause of worldwide death is stroke and continues to remain an important health burden both for the individuals, government and for the national healthcare systems and infrastructure. Risk factors which can potentially cause stroke include hypertension, cardiac disease, diabetes, and dysregulation of glucose metabolism, atrial fibrillation, and other common lifestyle factors. The final aim of this study was to develop a model equation for developing a pre-diagnosis algorithm for stroke with the potentially modifiable risk factors. Logistic regression for model derivation is the method which was used in this study, together with data from the database of the Korea National Health Insurance Service (NHIS). A total of 500,000 enrolees' were reviewed from the NHIS records. For the regression analysis, data of 367 stroke patients were selected. The control group consisted of 500 patients followed up for 2 consecutive years having no history of stroke.

We also looked at different models and after studying Modelling risk using generalized linear models[3]: Traditionally, linear regression has been the choice for predicting medical risk of the human beings. This paper presents a different approach to modelling the second part of two-part models utilizing extensions of the (GLM) Generalized Linear Model. Maximum likelihood is the primary method used for estimation. This method is discussed in alignment with generalizations quasi-likelihood and extended quasi-likelihood. An example using medical expense data from Washington State employees is used to illustrate the methods. The model also has demographic variables and Ambulatory Care Group variable to get the information of prior health status of individuals.

In Measuring overfitting in nonlinear models: A new method and an application to health expenditures[4]: We start by restricting the nonlinear models analysed to the members of the generalized linear model (GLM) family (Nelder and Wedderburn, 1972), which notably include the linear model for untransformed continuous variables, Poisson regression for counts, logistic and probit regression for binary variables, and parametric proportional hazard models for durations. In the GLM family, the distribution of the observed outcomes, is assumed to be a member of the exponential family where the expectation is related to the linear predictor. In addition, the variance is supposed to be a function of its expectation.

Revealing the cost of Type II diabetes in Europe[5]: n this study a design was used which was based on the prevalence and done in a bottom-up manner, which resulted in great optimization of the collection of data at the national level while making sure to maintain maximum international comparability. Considerable amount of effort was made to ensure consistency in terms of data specification, data collection tools and methods, sampling design, doing the analysis and reporting of results. Results reported are for individual countries and in aggregation for the total study of the population. The total medical costs of Type II diabetes in the eight European countries which were studied was estimated at 29 billion Euros a year (1999 values). The estimated average yearly cost per patient was 2834 Euros a year, hospitalizations accounted for the maximum proportion which was at 55%. This resulted in total of 15.9 billion Euros for the given eight countries. In the 6-month evaluation period, approximately 13% of the Type II diabetic patients were hospitalized, with an annual projection of average of 23 days in hospital. In contrast, drug costs for managing Type II diabetes were relatively low, with antidiabetic drugs and insulin accounting for only 7% of the total healthcare costs for Type II diabetes.

V. RESULT

Our model uses linear regression for finding medical expenses of the new customers. Data of existing customers is used to train the model and find relations in between different variables like age distribution of customers availing medical insurance, distribution of BMI (Body Mass Index), annual medical charges of individuals based on their smoking habits, Age of the customers availing health insurance. Upon training our model, we found that our test loss was approximately Rs 4684 which means our model yields an accuracy of approximately 80% based on the dataset provided.

REFERENCES

[1] Ch.Anwar ul Hassan,[1] awaid Iqbal, Saddam Hussain, Hussain AlSalman, Mogeeb A. A. Mosleh,[4]and Syed Sajid Ullah, "A Computational Intelligence Approach for Predicting Medical Insurance Cost" *Hindawi Mathematical Problems in Engineering Volume 2021, Article ID 1162553.*

[2] Sam Goundar, Suneet Prakash, Pranil Sadal, Akashdeep Bhardwaj, "Health Insurance Claim Prediction Using Artificial Neural Networks" *International Journal of System Dynamics Applications Volume 9 • Issue 3 • July-September 2020.*

[3] Rama Devi Burri, Ram Burri, Ramesh Reddy Bojja, Srinivasa Rao Buruga "Insurance Claim Analysis Using Machine Learning Algorithms" *International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-8, Issue- 6S4, April 2019.*

[4] Belisario Panay, Nelson Baloian , José A. Pino , Sergio Peñafiel , Horacio Sanson and Nicolas Bersano "Feature Selection for Health Care Costs Prediction Using Weighted Evidential Regression" *Sensors 2020, 20(16), 4392.*

[5] Viktor von Wyl "Proximity to death and health care expenditure increase revisited: A 15-year panel analysis of elderly persons." *Health Econ Rev* 9, 9 (2019).

[6] K. Bhatia, S. S. Gill, N. Kamboj, M. Kumar and R. K. Bhatia, "Health Insurance Cost Prediction using Machine Learning," 2022 3rd International Conference for Emerging Technology (INCET), Belgaum, India, 2022, pp. 1-5, doi: 10.1109/INCET54531.2022.9824201.

[7] Kaushik K, Bhardwaj A, Dwivedi AD, Singh R. Machine Learning-Based Regression Framework to Predict Health Insurance Premiums. Int J Environ Res Public Health. 2022 Jun 28;19(13):7898. doi: 10.3390/ijerph19137898. PMID: 35805557; PMCID: PMC9265373.

[8] E. M. van den Broek-Altenburg and A. J. Atherly, "Using Social Media to Identify Consumers' Sentiments towards Attributes of Health Insurance during Enrollment Season," *Applied Sciences*, vol. 9, no. 10, p. 2035, May 2019, doi: 10.3390/app9102035.

[9] Nidhi Bhardwaj , Rishabh Anand "Health Insurance Amount Prediction" *International Journal of Engineering Research & Technology (IJERT) ISSN: 2278-0181 Vol. 9 Issue 05, May-2020.*

[10] *Aman Kharwal "Health Insurance Premium Prediction with Machine Learning"issue 26,10,20201,*

[11] *Eline M.van den Broek-Altenburg, Adam J.Atherly "Using Scoial Media to Identify Consumers' Sentiments towards Attributes of Health Insurance during Enrollment Season", Center for Health Services Research , The Larner College of Medicine, University of Vermont, Burlington, VT 05405, USA.*