International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Insurance Premium Prediction and Forecasting Using Machine Learning

Prof. H. R. Agashe[1], *Payal S. Bhangre*[2], *Aditya R. Karle*[3], *Krishna S. Kharde*[4], *Abhishek S. Niphade*[5]

[1,2,3,4,5]**Savitribai Phule Pune University**

ABSTRACT-

People are exposed to different risk forms and the risk levels can vary. These dangers contain the risk of death, health and property loss or assets. But, risks cannot usually be avoided. Insurance is, therefore, a policy that decreases or removes loss costs incurred by various risks. Insurance provider they use different tools to calculate insurance premium. ML is beneficial here. These ML models can be learned by themselves, the model is trained on insurance data from the past. The requisite factors to measure the payments can then be defined as the model input, then the model can correctly anticipate insurance policy costs. The regression is the best choice available to fulfill our needs. We use Multiple Linear Regression in this analysis since there are many independent variables used to calculate dependent (Target) variable. To make this prediction more visualize Forecasting can be useful in this case. The key reason for this study is to include a new way of estimating insurance costs.

Keywords: Machine Learning, Linear Regression, Streamlit, Jupyter Notebook, Tkinter.

I. INTRODUCTION

We are on a planet full of threats and uncertainty. People, households, companies, properties, and property are exposed to different risk forms and the risk levels can vary. these dangers contain the risk of death, health, and property loss or assets. Life and wellbeing are the greatest parts of people's lives. But, risks cannot usually be avoided, so the world of finance has developed numerous products to shield individuals and organizations from these risks by using financial capital to reimburse them. Insurance is, therefore, a policy that decreases or removes loss costs incurred by various risks. Concerning the value of insurance in the lives of individuals, it becomes important for the companies of insurance to be sufficiently precise to measure or quantify the amount covered by this policy and the insurance charges which must be paid for it.

Various variables estimate these charges. Each factor of these is important. If any factor is omitted when the amounts are computed, the policy changes overall. It is therefore critical that these tasks are performed with high accuracy. As human mistakes are could occur, insurers use people with experience in this area. they also use different tools to calculate the insurance premium. ML is beneficial here. ml may generalize the effort or method to formulate the policy. These ml models can be learned by themselves. The model is trained on insurance data from the past. The requisite factors to measure the payments can then be defined as the model inputs, then the model can correctly anticipate insurance policy costs. This decreases human effort and resources and improves the company's profitability. Thus the accuracies can be improved with ml. our objective is to forecast insurance charges in this article. The value of insurance fees is based on different variables. as a result, insurance fees are continuous values. The regression is the best choice available to fulfill our needs. we use multiple linear regression in this analysis since there are many independent variables used to calculate the dependent(target) variable. For this study, the dataset for cost of health insurance is used. Preprocessing of the dataset is done first. Then we trained regression models with training data and finally evaluated these models based on testing data. The key reason for this study is to include new way of estimating insurance cost.

II. LITERATURE SURVEY

1) 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 SymposiumProceedings, vol. 2017, pp. 1312, 2017.An important informatics tool for controlling healthcare costs is accurately predicting the likely future healthcare costs of individuals. To address this important need, we conducted a-systematic literature review and identified five methods or predicting healthcare costs. To enable a direct comparison of these different approaches, we empirically evaluated the predictive performance of each reported approach, as well as other state-of- the-art.

2) Philipp Drewe-Boss, Dirk Enders, JochenWalker 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.Accurate prediction of healthcare costs is important for optimally managing health costs. However, methods leveragingthe medical richness from data such as health insurance claims or electronic health records are missing.

Here, we developed a deep neural network to predict future cost from health insurance claims records. We applied the deep network and a ridge regression model to a sample of 1.4 million German insurance to predict total one-year health care costs.

3) MC Politi, E Shacham, AR Barker, N George, N Mir, S Philpott et a1., "A Comparison Between Subjective and Objective Methods of Predicting Health Care Expenses to Support Consumers"'. Number of electronic tools help consumers select health insurance plans based on their estimated health care utilization. However, the best way to personalize these tools is unknown. The purpose of this study was to compare two common methods of personalizing health insurance plan displays: 1) quantitative healthcare utilization predictions using nationally representative Medical Expenditure Panel Survey (MEPS) data and 2) subjective-health status predictions.

4) C. A. Powers, C. M. Meyer, M. C. Roebuck healthcare costs using pharmacy claims data: A comparison of alternative econometric cost modeling techniques", Med. Care, vol. 43, pp. 1065-1072, 2005.We sought to evaluate several statistical modeling approaches in predicting prospective total annual health costs (medical plus pharmacy) of health Dimensions (PHD), a pharmacy claims-based risk index. Methods. We undertook a 2-year (baseline year/followup year) longitudinalanalysis of integrated medical

5) C. H. Dove, I. Duncan and A. Robb, "A prediction model for targeting low-cost high-risk members of managed care organizations", Amer. J. Managed Care, vol. 9, pp. 381-389,2003.To describe the development and validation of a predictive model desi ed M likely to incur high costs. Study Design: Split-sample multivariate regression analysis. Patients and Methods: We studied enrollees in a350 000- member HMO with ñ 1 claim in 1998 and 1999. The prediction model uses a combination of clinical and behavioral variables and 1998 and 1999 claims data.

6) Svetlana Sokolov Mladenovic, Milos Milovancevic and IgorMladenovic, "Identification of the important variables for prediction of individual medical in Society, vol. 62, pp. 101307, August 2020. The cost of health care insurance is one of the most important factors in the health care development. To establish a better health care system, there is a need to estimate the cost of health insurance. The prediction of the cost is one possibility to improve health care development. There is a need for more advanced methods other than traditional regression approaches, because the prediction of the health insurance costs.

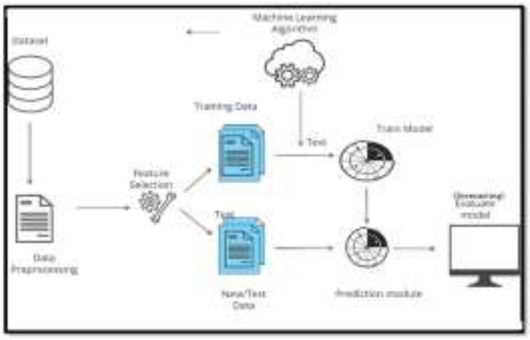
7) Roman Tkachenko, Ivan Izonin, Natalia Kryvinska and Valentyna Chopyak, Piecewise-linear Approach for Medical Insurance Costs Prediction using SGTM Neural-Like Structure Conference: Proceedings of the 1st International Workshop on Informatics & Data-Driven Medicine (IDDM 2018). The article proposes a new insurance medical cost prediction method. It is based on the piecewise-linear approach using the SGTM neural-like structure. Piecewise-linear approach provides high processing efficiency for large amounts of data, and the SGTM neural-like structure provides high accuracy and highspeed training procedure. The simulation of the proposed method using real data on health insurance costs and two SGTM neural- like structure cascades was performed.

8) Muhammad Arief Fauzanl and Hendri Murfi,"he Accuracy of XGBoost for Insurance Claim Prediction Int. J. Advance Soft Compu", Appl, vol. 10, no. 2, July 2018,ISSN 2074-8523.The increasing trend of claim frequency and claim severity for auto insurance result in need of methods to quickly of them is machine learning that treats the problem as supervised learning. The volume of the historical claim data is usually large. Moreover, there are many missing values for many features of the data. Therefore, we need machine learning models that can handle both data characteristics.

9) A. Ravishankar Rao, Daniel Clarke, "A comparison of models to predict medical procedure costs from open public healthcare data",2018 International Joint Conference on Neural Networks (IJCNN), pp.1-8, 2018.In our earlier work, we presented BOAT, a big-data open-source analytics toolkit and framework, and applied it to analyze trends and outliers in public healthcare data. In this paper, we extend different medical procedures that patients Specifically, we analyze de identified patient data from New York StateSPARCS (statewide planning and research cooperative system), consisting ofmore than 2 million records.

10) Liwen Cui, Xiaolei Xie, Zuojun Shen, Rui Lu, Haibo Wang, "Prediction of the healthcare resource utilization using multi-output regression models", IISE Transactions on Healthcare Systems Engineering, vol.8, no.4, pp.291, 2018.With the rapidly increasing healthcare cost and the scarcity of inpatient resources, it is of paramount importance to accurately predict the healthcare resource utilization. Previous research mainly focuses on predicting the healthcare cost using single-output models. However, the intensity of the healthcare resource utilization is reflected by multiple measures.

III. Architecture Diagram :



IV. INPUT DATA USED:

The following article discusses a dataset that can be accessed on the Kaggle website for the purpose of training and testing. This dataset is saved in a CSV file and is well organized. It is available at the specified link for those interested in using it.

No. of columns = 7

1338 rows total. Total number = 9366

In order to accurately predict the cost of health insurance, it is necessary to clean the dataset before applying regression algorithms. The data shows that age and smoking status have the most significant impact on the amount of insurance, with smoking having the greatest effect. However, factors such as family medical history, BMI, marital status, and geography also play a role. Children's property was found to have little impact on the prediction, so it was removed from the input for the regression model to improve efficiency and accuracy.. The data shows that age and smoking status have the most significant impact on the amount of insurance, with smoking having the greatest effect. This is a preliminary estimate and does not adhere to any company. These algorithms are designed to make classifications or predictions using statistical techniques, which can uncover key insights in data mining processes. The outcomes from these insights can be seen in the given figure 1 key growth indicators in businesses and applications, if used correctly. The data shows that age and smoking status have the most significant impact on the amount of insurance, with smoking having the greatest effect. They will be able to make a more informed decision. Additionally, it may suggest.

V. CONCEPT USED:

Machine Learning:

Machine Learning is a subset of computer science and AI that involves using data and algorithms to replicate the way that humans learn. These algorithms are designed to make classifications or predictions using statistical techniques, which can uncover key insights in data mining processes. The outcomes from these insights can have a significant impact on key growth indicators in businesses and applications, if used correctly. (S. Ramakrishnan, 2016) The data shows that age and smoking status have the most significant impact on the amount of insurance, with smoking having the greatest effect. However, factors such as family medical history, BMI, marital status, and geography also play a role. The data shows that age and smoking status have the most significant impact on the amount of insurance, with smoking having the greatest effect. However, factors such as family medical history, BMI, marital status, and geography also play a role.

Linear Regression Algorithm:

Linear regression is a statistical modelling technique used to understand the relationship between a dependent variable and one or more independent variables. It assumes a linear relationship between the independent variables (also called predictors or features) and the dependent variable (also known as the target or outcome variable). Linear regression aims to find the best-fit line that represents this relationship and allows us to make predictions based on the input variables.

Here's a general outline of the key concepts and steps involved in linear regression:

1. Dataset: Gather or obtain a dataset relevant to your project. The dataset should contain observations of the dependent variable and corresponding values of the independent variables.
2. Data pre-processing: Perform necessary data pre-processing steps such as handling missing values, handling outliers, and transforming variables if required.
3. Split the dataset: Split the dataset into two subsets: the training set and the test set. The training set will be used to build the linear regression model, while the test set will be used to evaluate its performance.
4. Model training: Fit a linear regression model to the training data. The model will estimate the coefficients (slopes) for each independent variable and the intercept (bias term).
5. Model evaluation: Evaluate the performance of the trained model using appropriate evaluation metrics such as mean squared error (MSE), root mean squared error (RMSE), or R-squared. These metrics measure the accuracy of the model's predictions compared to the actual values.
6. Interpretation of results: Analyse the coefficients of the independent variables to understand their impact on the dependent variable. A positive coefficient indicates a positive relationship, while a negative coefficient indicates a negative relationship.
7. Prediction: Use the trained model to make predictions on new, unseen data or test data. Apply the model to your specific project scenario to obtain predictions or insights.

Prediction:

The model used for predicting the insurance sum for health was based on the relationship between certain features and the label. The accuracy of this prediction was determined by the extent to which the expected value matched the actual amount.

In order to improve the accuracy, the model employed various characteristics, methods, and train-test split sizes. It was found that the amount of data used for training had a significant impact on the accuracy, with a larger train size leading to better results.

The model also employed multiple algorithms in order to forecast the premium amount, and showed how each attribute affected the outcome(Kaggle, Regression data)

VI. RESULTS:

r-squared scores for the different models

Linear Regression(Lr)	0.7833
Support Vector Machine(svm)	-0.0723
Random Forest Regression(rf)	0.8620
Gradient Boosting Regression(gr)	0.8779

The following results can be seen in Prediction:

Linear Regression Algorithm: The accuracy of the Linear Regression Algorithm is 78.33%.

Support Vector Machine:

The accuracy of the Support Vector Machine Algorithm is -0.072%

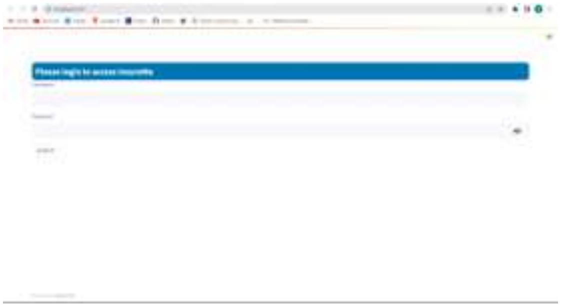
Random Forest Regression:

It was found that the amount of data used for training had a significant impact on the accuracy, with a larger train size leading to better results (Kenward , J.A. , 2019)

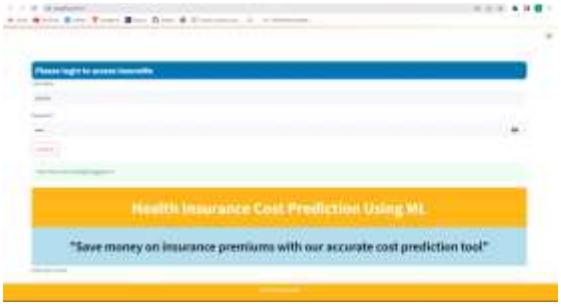
From the **Figure 2** we can see that the best optimum Algorithm for the Amount prediction is Gradient Boosting Algorithm with the highest accuracy .(H. Demirtas, J. Stat Soft.)

OUTPUT:

Screenshot 1:



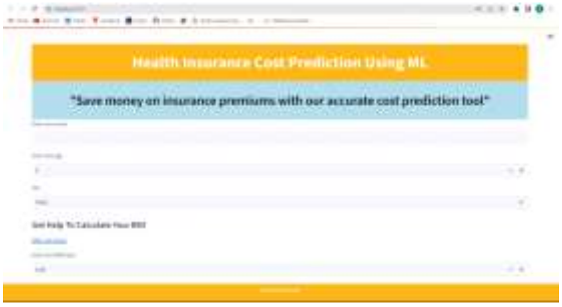
Screenshot 2:



Health Insurance Cost Prediction Using ML

"Save money on insurance premiums with our accurate cost prediction tool"

Screenshot 3:

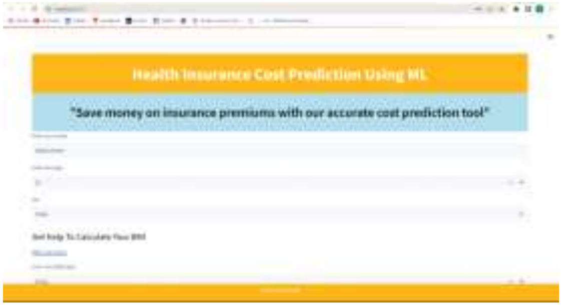


Health Insurance Cost Prediction Using ML

"Save money on insurance premiums with our accurate cost prediction tool"

Get Help To Calculate Your BMI

Screenshot 4:

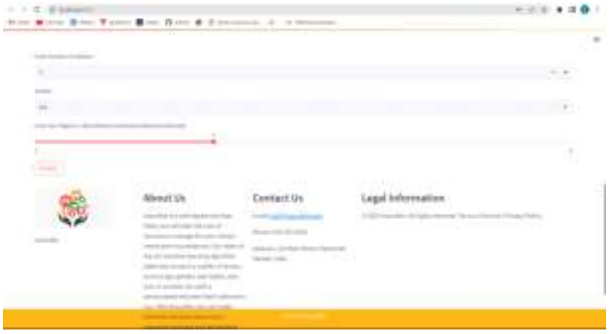


Health Insurance Cost Prediction Using ML

"Save money on insurance premiums with our accurate cost prediction tool"

Get Help To Calculate Your BMI

Screenshot 5:



VII. CONCLUSION :

It uses machine learning regression model to forecast premium of Insurance based on specific attributes. ML models can perform cost calculations in a short time, while a human being would be taking a long time to perform the same task. This will help businesses improve their profitability. The ML models can also manage enormous amounts of data.

VIII. ACKNOWLEDGEMENT

First of all, I am indebted to the god almighty for giving me an opportunity to excel in my efforts to complete this seminar on time. I am extremely grateful to **Dr.G.K.Kharate,** Principal, Matoshri College of Engineering and Research Centre**, Prof.N.L.Bhale** Head of Department, Department of Information Technology**,** for providing all the required resources for the successful completion of my seminar. My heartfelt gratitude to my seminar guide **Prof.H.R.Agashe,** for her value able suggestions and guidance in the preparation of the seminar report. I express my thanks to all my teachers & professors and friends for all the help and co- ordination extended in bringing out this seminar successfully in time. I will be failing in duty if I do not acknowledge with grateful thanks to the authors of the references and other literatures referred to in this seminar. Last but not the least; I am very much thankful to my parents who guided me in every step which I took.

IX. References :

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.

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.

MC Politi, E Shacham, AR Barker, N George, N Mir, S Philpott et al., "A Comparison Between Subjective and Objective Methods of Predicting Health Care Expenses to Support Consumers'".

C. A. Powers, C. M. Meyer, M. C. Roebuck and B. Vaziri, "Predictive modeling of total healthcare costs using pharmacy claim data: A comparison of alternative econometric cost modeling techniques", Med. Care, vol. 43, pp. 1065-1072, 2005.

H. Dove, I. Duncan and A. Robb, "A prediction model for targeting low-cost high-risk members of managed care organizations", Amer. J. Managed Care, vol. 9, pp. 381-389, 2003

Svetlana Sokolov Mladenovic, Milos Milovancevic and Igor Mladenovic, "Identification of the important variables for prediction of individual medical costs billed by health insurance", Technology in Society, vol. 62, pp. 101307, August 2020.

Roman Tkachenko, Ivan Izonin, Natalia Kryvinska and Valentyna Chopyak, Piecewise-linear Approach for Medical Insurance Costs Prediction using SGTM Neural-Like Structure Conference: Proceedings of the 1st International Workshop on Informatics & DataDriven Medicine (IDDM 2018).

Muhammad Arief Fauzanl and Hendri Murfi, "he Accuracy of XGBoost for Insurance Claim Prediction Int. J. Advance Soft Compu", Appl, vol. 10, no. 2, July 2018, ISSN 2074–8523.

Ravishankar Rao, Daniel Clarke, "A comparison of models to predict medical procedure costs from open public healthcare data", 2018 International Joint Conference on Neural Networks (IJCNN), pp.1-8, 2018.

Liwen Cui, Xiaolei Xie, Zuojun Shen, Rui Lu, Haibo Wang, "Prediction of the healthcare resource utilization using multi- output regression models", IISE Transactions on Healthcare Systems Engineering, vol.8, no.4, pp.291, 2018.