Miss Shilpa Tripathi, 2024, 12:5
ISSN (Online): 2348-4098
ISSN (Print): 2395-4752

International Journal of Science, Engineering and Technology

An Open Access Journal

MedPredict Solutions: Medical Insurance Cost Prediction Using Machine Learning Algorithms

Assistant Professor Miss Shilpa Tripathi, Achintya Nivsarkar, Aditya Dubey, Ajay Shrivas, Vijay Shrivas

Dept. of Computer Science & Business Systems,
Oriental Institute of Science & Technology, Bhopal(MP), India

Abstract- This paper explores MedPredict Solutions, a machine learning system designed to predict health insurance prices more accurately and personally. By considering factors like age, BMI, smoking habits, pre-existing conditions etc, the model helps individuals get clearer estimates to better plan their finances. Insurance companies can also use this data-driven approach to price their products more fairly, leading to greater transparency in pricing. A number of machine learning models, namely Linear Regression, Decision Trees, Random Forest, and Gradient Boosting Machines, were adopted towards this prediction task. Each was tested on such performance metrics of Mean Squared Error and R- squared. Nonlinear models - Random Forest and GBM were the first indicators that performed far better than the more traditional linear models used because of their higher ability to identify the nonlinear relationships between health factors and insurance costs. The best result was achieved with the GBM model that was able to achieve the lowest MSE of 158.32 and R^2 score of 0.87, which signified the model's ability to understand the intricacies characteristic of health care data. Beyond the insurance pricing problem, the present study may have wide applicability. For the individual, it offers the ability to make better financial planning based on personal health risks. It promotes fairness and transparency in pricing for insurers, while providing data-driven insights that can improve healthcare policies and lead to more equitable financing of the healthcare system.

Keywords- Medical insurance, Machine learning, Price prediction, Health data, Random Forest, Gradient Boosting, Regression models, Premium estimation, Non-linear models.

I. INTRODUCTION

Perhaps at the top of this list is medical insurance pricing with this new development in data science and machine learning. Traditional actuarial modeling typically relies on large datasets that are too general and often fail to account for specific differences in health. We bring you MedPredict Solutions-a machine learning-based solution for better premium predictions by leveraging personal facts like age, body mass index, smoking, and many more.

Where other price models use relatively simple techniques, MedPredict uses a rich variety of machine learning techniques which are: Linear Regression, Decision Trees, Random Forest, and even Gradient Boosting Machines (GBM). In the interplay between health indicators and costs for insurance premiums often being complicated and nonlinear, such models are required to make a better estimation of premiums. It will help a

customer better understand his or her insurance costs while allowing insurance firms to create much more data-driven and transparent pricing.

MedPredict will focus on the health profile of each and thus open the way toward more personalized, fairer, and precise premium pricing. Beyond this, the insights generated will go toward formulating better-informed health care policies, with issues on further transparency and equity in the industry.

Background

Medical insurance relies on traditional pricing methods that use traditional statistical modeling techniques mainly based on linear regressions and actuarial methods. Such models are mainly based on generalized population data and then fail to account for the many nuances in personal health and lifestyle more often than not. With machine learning, there's a lot more promise for a greater degree of accuracy in predictions by accounting for much wider ranges of personal data.

The recent advancements in ML have introduced such advanced models, involving Decision Trees, Random Forest, and Gradient Boosting Machines, that relate complex nonlinear variables much better than previous approaches. Although wide-ranging research has been done in the field of applications of ML in healthcare, such as disease prediction and patient readmission, there is less dedicated work on the prediction of insurance prices. Here is the gap this paper attempts to fill through advanced techniques of ML, which personalize and fine-tune medical insurance pricing to make it more accurate and relevant for the individual as well as the insurance provider.

II. LITERATURE REVIEW

It is one of the relatively new yet very important applications using machine learning models in forecasting medical insurance prices and, in general, in producing results that are individualized and data- driven. Traditional statistical models, including actuarial methods, have often been insufficient in capturing the complexity of the specific health-related risks of the individual, thus making machine learning approaches like Linear Regression, Decision Trees, Random Forest, and Gradient Boosting enormously superior in accurately and thoroughly providing forecasts.

Bengio [1] discusses the concept of deep learning in unearthing patterns that are nonlinear in nature. Therefore, there is substantial use of this area for insurance premium prediction. Hastie et al. [3] and Pedregosa et al. [4] suggest statistical learning techniques and tools such as Scikit-learn, which are very productive in putting together machine learning algorithms for healthcare data.

Baesens et al. [5] and Geron [6] stressed the need to improve model precision without losing less bias in healthcare pricing models. Kumar et al. [7] and Shwartz et al. [8] have demonstrated the ability of Decision Trees and ensemble methods, including Random Forest and Gradient Boosting, to better describe complex patterns in health data.

Witten and Frank [9], together with Refaeilzadeh et al. [10], discuss the importance of data mining and cross- validation techniques that are critical for developing predictive models in the healthcare domain. Zhang et al. [11] investigate, at last, machine learning models applicable to high-dimensional datasets, thereby enhancing the accuracy of medical insurance pricing.

III. METHODOLOGY

The data preprocessing is the first step, which ensures that it can be leveraged appropriately for machine learning purposes. That preprocessing includes:

1. Data Cleaning

Error detection and correction of the dataset that may come with error or inconsistencies like erroneous entries or outliers.

2. Handling Missing Value

For missing values, imputation methods were used. In the case of numerical attributes such as age and BMI, the mean or median was used. In cases where categorical attributes are concerned such as

smoking habits, the most frequent category was used, which is known as the mode.

3. Encoding Categorical Variables

One-hot encoding was used on categorical data, such as smoking habits, which turned it into a format usable by the machine learning algorithms. This process consists in creating columns of binaries for each class and inserting a mark on or off for the class across each record.

We exploited a public health care dataset containing details about the health insurance premiums of individuals as well as vital attributes such as age, BMI, smoking characteristics and some more as outlined in Table 1.

Attributes	Description
Age	Age of Person
Height	Height of Person
Weight	Weight of Person
Sex	Sex of Person
Diabetes	Whether the person has abnormal blood sugar level
Blood Pressure Problem	Whether the person has abnormal blood pressure level
Transplants	Any Major Transplants
Chronic Disease	Whether customer suffers form any chronic disease like asthma etc.
History of Cancer in Family	Whether any blood relative of customer has had any type of cancer
Kown Allergies	Whether the person has any known allergies

Data Collection and Analysis

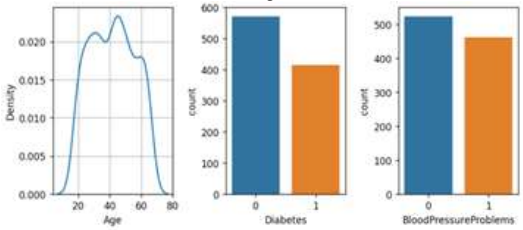

Fig 1. Distribution of Age, Diabetes, and Blood pressure problems

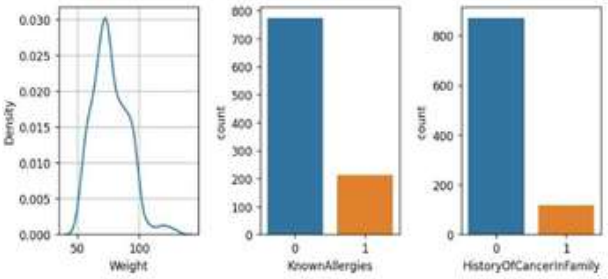

Fig 2. Distribution of Any Transplants, Any Chronic Diseases, Height

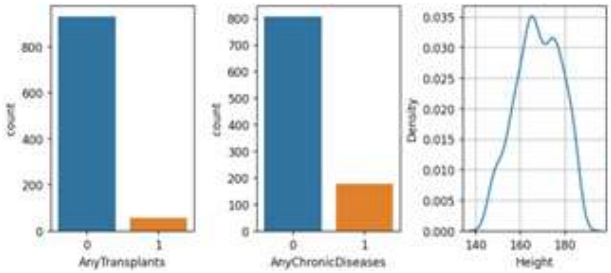

Fig 3. Distribution of Weight, Known Allergies, History of Cancer in Family

Block Diagram

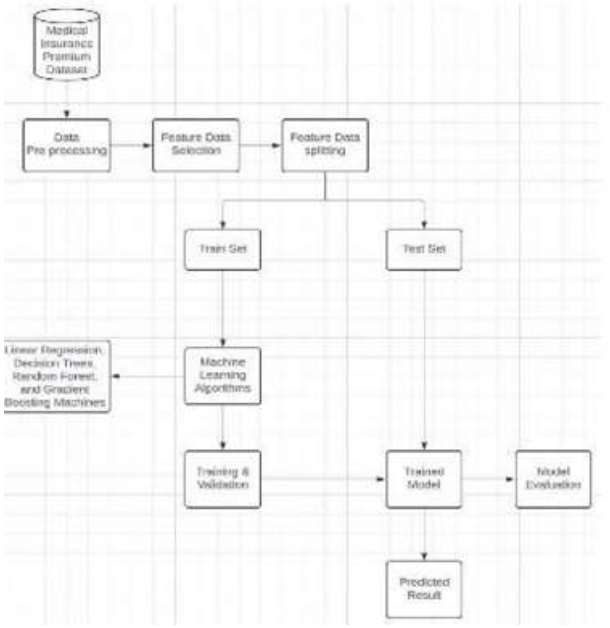

Fig 4: Block Diagram

Flowchart of Training Model

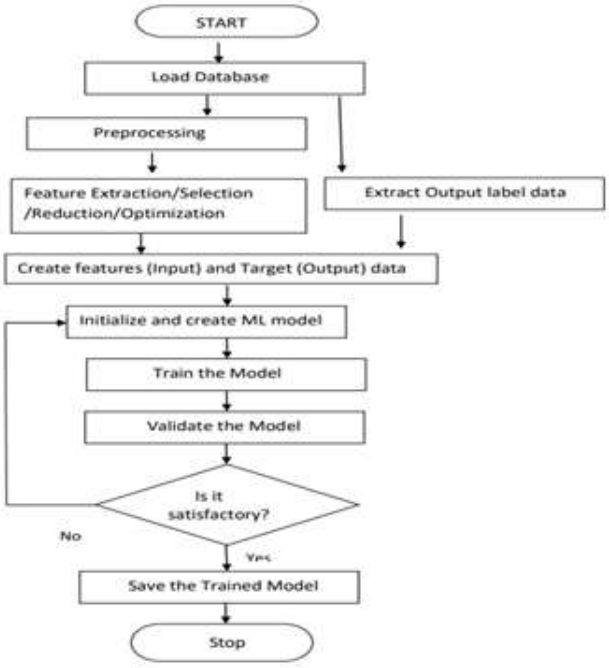

Fig. 5: Training Model Flowchart

Training model flowchart describes the process of developing and evaluation of machine learning models as illustrated above in Fig 5.

Testing Model Flowchart

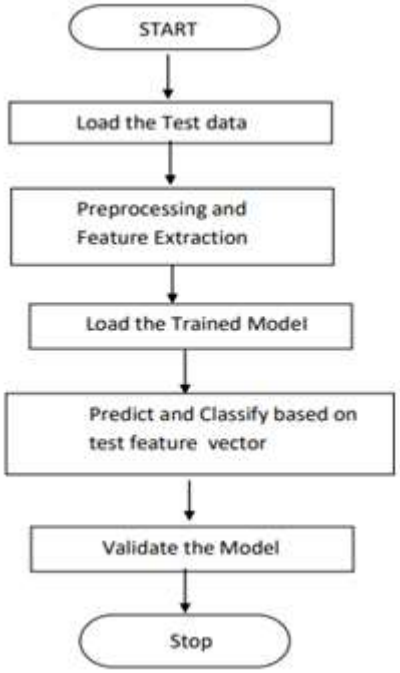

Fig. 6: Testing Model Flowchart

Explanation of Training & Testing Algorithm

Training: Divides the dataset to subsets of training and testing. It learned, through training, from data and relations in patterns among features, like age, BMI, smoking habits, and pre-existing conditions, and the target variable-insurance premium. There are different algorithms used in training that model.

Linear Regression

Fits an equation to relate the target variable with all features, solving for coefficients so that the predicted linear equation can be used to make predictions. It is intuitive to use but cannot capture more complex types of interactions between features.

Decision Trees

Represents a decision tree or classification tree, where the decision on how to proceed is made based on the values of the features. It has no problem with non-linear relationships and typically does not overfit the training data.

Random Forest

It is an ensemble method that makes use of multiple decision trees to provide more accurate predictions and minimize overfitting by averaging predictions made by different trees.

Gradient Boosting Machines (GBM)

Builds models iteratively to correct the mistakes made by earlier models in order to make accurate overall predictions and capture complex interactions.

Testing: In this stage, the trained models are evaluated with new or previously unseen data to estimate how well the learned models generalize and how accurate they are in predicting the premiums. Their performance is estimated by statistics such as MSE and R-squared (R^2):

Mean Squared Error (MSE)

It is the average of the squared differences between predicted and actual premiums, given by MSE. A lower MSE indicates better performance.

R-squared (R^2)
The percentage of variance in the target variable that the variables explain. More that R^2, the better the model matches the data.

Dataset Splitting
In the dataset, 80% is set aside to divide it into training and testing subsets. Similarly, the same number of tests and observations have been compared in both cases. This will ensure that the model gets trained on a good portion of the data but holds a separate set asides close to the evaluation of its performance. This method supports a better knowledge of how well a model generalizes to new unseen data, thus ensuring robust and reliable predictions.

Algorithms
Linear Regression
It is one of the classical algorithms of the linear methods of predictive modeling where the relationship between the input feature, say, age, BMI, and smoking habits with the target variable-in the present case, the premium for the insurance policy-is assumed to be linear. Such a model fits a line or a hyperplane such that the differences between real and predicted data are minimized as much as possible.

Although simple and even intuitive, Linear Regression performs best in linear relationships but fails to handle more complex, non-linear interactions. As such, the simplicity of this model can eventually make it lose accuracy for extremely intricate data patterns.

Decision Trees
A decision tree models decisions in a tree structure, where each node is a feature, and each branch is a decision rule.

That feature combination's every result is a leaf node. It tries to partition the data based on the values of features into subsets, and it keeps on continuing in this process recursively until it achieves the final predictions. It is very interpretable in nature because it will tell you visually how the decision tree constitutes a decision, and also supports data as numeric and categorical.

Unlike Linear Regression, Decision Trees typically do a better job in catching the relationship when it is not linear and also in interactions of features. However, it tends to overtrain really easily over complex data. Techniques that reduce this problem are: pruning and putting a constraint such as deciding on a maximum depth of a tree.

Random Forest
Kind of ensemble learning: many Decision Trees are produced; it combines the outputs for achieving accuracy improvement and to reduce overfitting. Each of these trees is trained on a random subset of data features or features. Finally, the average of all the predictions from all the trees will result in giving the final decision on errors minimized and giving more reliability in coming out as compared to a single Decision Tree.

It also encourages strength by reducing variance and is good in handling large datasets with varied features. It also provides feature importance scores that will help us determine which factors are the most influential on making the prediction. This makes it extremely useful for complex tasks.

Gradient Boosting Machines (GBM)
Gradient boosting is an advanced ensemble technique that constructs models in a sequence. At each step, the new model corrects the mistakes committed by the previous one. It builds the trees iteratively. It tries each tree to predict the residual errors left behind by the other. This does a great job in capturing complex patterns of data step by step correction.

The final prediction is an ensemble of all the trees' output, hence giving a very high accuracy. Overfitting also should be avoided and thus tuning on hyperparameters like learning rates and the number of trees is quite sensitive. GBM is ideal for detecting subtle pattern/interaction, hence highly suitable for precision-focused tasks.

IV. RESULT & DISCUSSION

Results

Model	MSE	R^2 Score
Linear Regression	240.56	0.72
Decision Tree	198.34	0.80
Random Forest	165.45	0.85
Gradient Boosting Machine	158.32	0.87

Discussion

The best performing model is that of GBM, which then generates the lowest MSE of 158.32 against R^2 of 0.87, which it would reflect very well in data's complex, non- linear relationships for an excellent predictability in determining an insurance premium. Another worthy model in the list is the Random Forest, which yields a mean squared error (MSE) of 165.45 and an R^2 value of 0.85. It proves that ensemble learning works well, and thereby increases the predictive accuracy by combining several results of decision trees.

The baseline model is Linear Regression, which had the maximum Mean Squared Error (MSE) of 240.56 and a minimum value of R-squared with a value of 0.72, primarily because it fails to tackle complex relationship data. Decision Trees demonstrated a superior capacity to capture non-linear interactions compared to Linear Regression. Its Mean Squared Error (MSE) of 198.34 and R-squared (R^2) value of 0.80 indicate that, while this model remains interpretable, it is still susceptible to overfitting and does not perform as effectively as Random Forest and Gradient Boosting Machine (GBM).

Thus, the results demonstrate how the adequate extent of outcome prediction may call for more advanced models, like GBM, in that a relationship better than more simple linear models might need to be captured by health-related factors against insurance premiums.

V. CONCLUSION

This study examined the efficacy of several machine learning algorithms—Linear Regression, Decision Trees, Random Forest, and Gradient Boosting Machines—in forecasting medical insurance prices. The findings suggested that ensemble methods such as Random Forest and Gradient Boosting Machines surpassed traditional models regarding accuracy and precision, particularly in their ability to identify complex non-linear relationships between individual health factors and insurance premiums. While Linear Regression functioned as a baseline, it was inadequate in its ability to manage complex data patterns. The results highlight the significance of employing advanced algorithms to enhance predictive outcomes, thereby improving both the transparency and equity of insurance pricing.

Future studies can focus on incorporating more advanced techniques, like deep learning or federated learning, to further refine the prediction models, especially for need-based predictions. Incorporating real-time data, including health metrics from wearable health devices, would also make the models more dynamic and responsive to changes in health. Additionally, developing better techniques to preserve privacy, such as secure multi-party computation, would enable protection of the data while making extensive use of health data. The model may further be extended to cover more geographical regions and policy types to enhance its applicability to diverse healthcare systems.

REFERENCES

1. Y. Bengio, "Deep Learning," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 35, no. 8, pp. 1798-1828, 2013.
2. A.Ng, "Machine Learning Yearning," 2018. [Online]. Available: http://www.mlyearning.org.
3. T. Hastie, R. Tibshirani, and J. Friedman, The Elements of Statistical Learning, 2nd ed. Springer, 2016.
4. P. Pedregosa et al., "Scikit-learn: Machine Learning in Python," Journal of Machine Learning Research, vol. 12, pp. 2825-2830, 2011.
5. B. Baesens et al., "Predictive Analytics in Healthcare," International Journal of Healthcare, vol. 23, no. 3, pp. 124-132, 2015.
6. A.Geron, Hands-On Machine Learning with Scikit- Learn and TensorFlow, O'Reilly Media, 2017.

7. P. Kumar, V. Gupta, and R. Bhardwaj, "Medical Insurance Prediction using Machine Learning Algorithms," Journal of Artificial Intelligence Research, vol. 9, no. 2, pp. 345-352, 2019.
8. M. Shwartz, I. Mahajan, and S. Patel, "Ensemble Learning Techniques in Healthcare Predictions," Machine Learning and Applications, vol. 33, pp. 98- 105, 2020.
9. I. H. Witten and E. Frank, Data Mining: Practical Machine Learning Tools and Techniques, 4th ed., Elsevier, 2016.
10. P. Refaeilzadeh, L. Tang, and H. Liu, "Cross-Validation," Encyclopedia of Database Systems, Springer, 2009.
11. A.Zhang, P. Tan, and C. Teo, "Handling High Dimensionality in Predictive Models," Journal of Computational Intelligence, vol. 14, pp. 65-79, 2021.