MEDICAL INSURANCE PREDICTION

By

1.Name: Aishwarya Devisetti (1211921)
2.Name: Sucharitha Swargam (1212077)
3.Name: Praveen Kollipara (1211436)

Adviced By
Prof. Asif Mahmood

SUBMITTED IN PARTIAL FULFILMENT
OF THE REQUIRMENTS FOR THE
DEGREE OF MASTER OF SCIENCE IN
COMPUTER ENGINEERING

THE SCHOOL OF ENGINEERING
UNIVERSITY OF BRIDGEPORT
CONNECTICUT

Abstract

Health insurance is becoming an essential commodity for everyone. Insurance data has increased dramatically in the last decade. The health insurance system explores predictive modeling to boost its business operations and services. Computer algorithms and Machine Learning are used to study and analyze past insurance data and predict new output values based on trends in customer behavior, insurance policies, and data-driven business decisions. So, in this project we would like to develop a Health insurance cost prediction system using ML algorithms. The proposed model incorporates and demonstrates different models of regression.

We are using Random Forest and Linear Regression Algorithm. This project aims to brief the introduction of various computational intelligence techniques and algorithms-based approaches used for the prediction of Medical Insurance using python. This project, we are implementing using Insurance dataset. The main aim of the data is to investigate several variables like age, BMI number, Gender, Region etc., and their impact on the estimated amount was investigated.

1. Introduction

This chapter discusses the aim and project objectives of the research. This chapter also discusses the project objectives in detail along with its research approach and structure of the dissertation.

The goal of this project is to allow a person to get an idea about the necessary amount required according to their own health status. Later they can comply with any health insurance company and their schemes benefits keeping in mind the predicted amount from our project. This can help a person focus more on the health aspect of insurance rather than the futile part.

Health insurance is a necessity nowadays, and almost every individual is linked with a government or private health insurance company. Factors determining the amount of insurance vary from company to company. Also, people in rural areas are unaware of the fact that the government of India provide free health insurance to those below poverty line. It is a very complex method and some rural people either buy some private health insurance or do not invest money in health insurance at all. Apart from this people can be fooled easily about the amount of the insurance and may unnecessarily buy some expensive health insurance.

Our project does not give the exact amount required for any health insurance company but gives enough idea about the amount associated with an individual for his/her own health insurance. Prediction is premature and does not comply with any particular company so it must not be only criteria in selection of a health insurance. Early health insurance amount prediction can help in better contemplation of the amount needed. Where a person can ensure that the amount, he/she is going to opt is justified. Also, it can provide an idea about gaining extra benefits from health insurance. 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, it is to know a person to get a clear idea about the necessary amount required according to their own health status.

Concerning the value of insurance in the lives of individuals, the policy changes overall. It is therefore critical that these tasks are performed with high accuracy. Use different tools to calculate the insurance premium. 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. As indicated by the World Bank, the absolute use on medicinal services as an extent of GDP in 2015 was 3.89extent of GDP is simply use was 65.06in clinical innovation has made it conceivable to fix illnesses that were once viewed as serious. In any case, the expense of their treatment is so high, it is practically incomprehensible for a white-collar class individual to manage the cost of them. As indicated by insights. Rs 5 lakh family floater strategy will cover self, mate and one kid will cost anyplace between Rs 10,000 and Rs 17,000 on a yearly premise though Rs. 5 lakh singular wellbeing plan will cost a multiyear old Rs. 4,000-7000 per year.

2. Statement of The Problem

Healthcare costs continue to rise, making it essential for individuals to make informed decisions when selecting health insurance plans. However, determining the expected insurance cost based on personal health factors can be challenging, leading many to either overpay for coverage they don’t need or choose insufficient plans that leave them financially vulnerable.

This project aims to develop a predictive model that estimates an individual’s potential health insurance costs based on key personal health attributes such as age, BMI, smoking status, and other relevant factors. By leveraging machine learning techniques, this model can provide accurate cost predictions, enabling individuals to better understand their expected expenses and make well-informed choices when selecting health insurance plans.

The ability to predict insurance costs can benefit not only individuals but also insurance providers by helping them assess risk more effectively. Overall, this project seeks to enhance transparency in health insurance pricing, reduce unnecessary financial burdens, and promote fairer, data-driven decision-making in the healthcare sector.

3. Significance of The Study

Access to healthcare is a fundamental need, yet many individuals struggle to navigate the complexities of health insurance pricing and coverage. This challenge is especially pronounced in rural areas, where awareness of government-provided health insurance schemes is often limited. Additionally, private insurance companies use intricate pricing structures that can make it difficult for consumers to understand the true cost of coverage, leading to potential overpayment or underinsurance.

This study aims to bridge this gap by developing a predictive model that provides individuals with an independent, data-driven estimation of their potential health insurance costs. By utilizing personal health data such as age, BMI, smoking status, and medical history, the model can generate accurate cost predictions, helping individuals make more informed decisions when selecting an insurance plan.

The significance of this study extends beyond just individual consumers. Insurance providers can use predictive models to better assess risk and tailor their offerings, potentially leading to fairer and more transparent pricing structures. Additionally, policymakers can leverage such models to identify gaps in insurance coverage and design targeted awareness campaigns for underinsured populations.

By empowering individuals with knowledge about their estimated insurance costs, this project promotes financial planning, reduces the likelihood of unexpected medical expenses, and encourages a more transparent and accessible health insurance marketplace.

4. Purpose of The Study

The primary objective of this study is to develop a predictive model that estimates an individual's potential health insurance costs based on their current health status and potential future risks. By analyzing key health factors such as age, BMI, smoking status, and medical history, the model can provide a data-driven approach to determining insurance premiums.

This predictive system serves multiple purposes. For individuals, it acts as a decision-making tool that helps them understand how various health factors influence insurance costs, allowing them to choose the most suitable insurance plan. This is particularly important for those who may be unaware of the factors that impact their premiums or who struggle with complex insurance pricing structures.

For insurance providers, the model can enhance risk assessment, leading to more accurate pricing strategies that balance affordability for consumers with sustainability for insurers. Additionally, policymakers can use such a model to identify trends in insurance accessibility and affordability, helping them design better public healthcare policies and awareness programs.

Ultimately, the study aims to make health insurance more transparent, accessible, and tailored to individual needs, reducing financial uncertainty and ensuring better healthcare planning for the future.

5. Research Hypothesis

Author Muhammad Arief et al. [1], used XGBoost machine learning model to find the insurance claim frequency and severity which is most important one to find claims faster with accurate result. Sahar F. Sabbeh [1-3], compared all the machine learning models with their performance result to yield customer retention. Among the other models, the derived result showed that Random Forest and Ada boost algorithm perform better and attained 96% accuracy. Cunningham et al. [4-5], presented a paper to find the classification using K nearest neighbor. Kayri et al. [6-8], compared the results of three algorithms such as Random Forest classifier, Multiple Linear regression, and Artificial Neural Network for atmospheric data.

6. Definitions of The Keywords

Machine Learning: A subset of AI that enables systems to learn from data.

Regression Models: Statistical models predicting continuous outcomes.

Neural Networks: Algorithms inspired by the human brain to process complex data patterns.

7. Research Limitations

While this study provides insights into the effectiveness of machine learning for insurance cost prediction, it has several limitations:

- **Data Availability:** The dataset used may not cover all possible patient demographics and conditions, limiting the generalizability of results.
- **Feature Bias:** Some features may have a stronger influence on predictions than others, leading to potential biases.
- **Model Interpretability:** Advanced models like deep learning may provide high accuracy but lack interpretability, making it difficult to understand their decision-making process.
- **Regulatory Constraints:** Real-world insurance pricing is influenced by regulations, which are not fully accounted for in this study.
- **Computational Limitations:** The performance of machine learning models can vary depending on hardware capabilities, and extensive computations may not be feasible in all settings.

2. Existing System

At present the model for health insurance prediction is about how a person can get a price estimation. But there are many disadvantages for existing model. Model cannot predict the perfect estimation for insurance.

2.1 Medical Insurance Industry

The medical insurance industry is a critical component of global healthcare systems, ensuring financial security for individuals against unforeseen medical expenses. The industry has evolved significantly, integrating data analytics and AI to improve risk assessment and pricing models.

2.1.1 Factors Affecting Insurance Premiums

Insurance premiums are determined based on several factors:

1. **Demographics**: Age, gender, and location impact premium rates.
2. **Medical History**: Pre-existing conditions and past claims history influence costs.
3. **Lifestyle Choices**: Smoking, alcohol consumption, and obesity contribute to higher premiums.
4. **Family Medical History**: Genetic predispositions to diseases can affect risk assessment.
5. **Policy Terms**: Coverage limits, deductibles, and additional benefits alter pricing structures.

2.2 Role of Data Analytics in Insurance

2.2.1 Traditional Statistical Methods

Historically, insurance companies have relied on linear regression models and actuarial tables to estimate policy costs. While effective to some extent, these methods struggle to capture complex relationships between multiple factors.

2.2.2 Machine Learning in Insurance Prediction

Machine learning offers a robust approach to insurance prediction by leveraging large datasets and complex algorithms. Some widely used machine learning techniques include:

- **Linear Regression**: Establishes relationships between dependent and independent variables.
- **Decision Trees**: Splits data based on attribute conditions to make predictions.
- **Random Forest**: Uses multiple decision trees for improved accuracy.
- **Neural Networks**: Simulates human brain function to detect intricate data patterns.

3. Proposed System

A wide range of statistical techniques are used in building scoring models. Traditionally linear techniques are used such as Linear Regression. The regression analysis is a predictive method that explores the relationship between a dependent (target) and the independent variable (predictor). In regression, learning algorithms map the input data to continuous output like weight, cost, etc. This technology is used to fore- casting, estimate model time series, and find the causal effect relationship among the variables. In this analysis. If we need to analyze the relationship between insurance cost (target variable) and six independent variables based on (age, BMI, child number, individual living area, or sex and whether the customer is a smoking person) on the basis of a regression. Analysis of regression also helps one to compare the results of measured variables at various scales, such as independent variable and de- pendent variable effects. These advantages allow market researchers, data analysts, and data scientists to remove and determine the best range of variables for predictive model.

3.1 Feasibility Study

As the proposed project is present. This can consider the project as a feasible one and get successful in the market. Though the project is a lengthy and time taking one, it can be highly useful in the society.

3.1.1 Model Selection & Implementation

To predict insurance costs, various machine learning models were evaluated based on accuracy, efficiency, and interpretability. The selected models include:

- **Linear Regression**: Baseline model for comparison.
- **Random Forest**: Handles non-linear relationships and reduces overfitting.
- **XGBoost & LightGBM**: Gradient boosting models that iteratively improve predictions.
- **Artificial Neural Networks (ANNs)**: Captures complex patterns but requires higher computational power.

Model Implementation

1. **Data Preprocessing**: Handling missing values, encoding categorical variables, and feature scaling.
2. **Data Splitting**: 80% training, 20% testing.
3. **Model Training**: Each model is trained and optimized using hyperparameter tuning.
4. **Performance Evaluation**: Models are assessed using:
   - **Mean Squared Error (MSE)**
   - **Mean Absolute Error (MAE)**
   - **R-Squared (R^2) Score**
5. **Prediction & Optimization**: Best-performing models (XGBoost & ANN) were fine-tuned for better generalization.

3.1.3 Model Training & Evaluation

Model Training

The dataset was preprocessed (handling missing values, encoding categorical features, and normalizing numerical data) before splitting into training and testing sets. Models including Linear Regression, Random Forest, XGBoost, LightGBM, and Neural Networks (ANNs) were trained to predict insurance costs. Hyperparameter tuning was applied for optimization.

Model Evaluation

Models were assessed using Mean Squared Error (MSE), Mean Absolute Error (MAE), and R-Squared (R^2) scores. Results showed that ensemble models (XGBoost, LightGBM) and ANNs outperformed traditional regression methods, with ANNs achieving the highest accuracy. Overfitting was mitigated through regularization and feature selection.

3.1.4 Deployment

The trained model is deployed using a RESTful API (Flask/FastAPI) and integrated into a user-friendly web interface. It is hosted on cloud platforms like AWS or Google Cloud for scalability. Continuous monitoring ensures accuracy, with periodic retraining for improved predictions.

4. Tools and Technologies Used

4.1 Programming Languages

Python: Primary language for data processing, model development, and evaluation due to its extensive ML libraries and ease of use.

SQL: Used for querying and managing structured insurance datasets.

4.2 Data Processing & Machine Learning Libraries

NumPy & Pandas: For data manipulation, cleaning, and preprocessing.

Matplotlib & Seaborn: For data visualization and exploratory analysis.

Scikit-Learn: Provides essential ML algorithms like Linear Regression, Random Forest, and model evaluation metrics.

XGBoost & LightGBM: Advanced boosting algorithms for better predictive accuracy.

TensorFlow & Keras: Used for building and training Artificial Neural Networks (ANNs).

4.3 Model Deployment & Cloud Technologies

Flask & FastAPI: Used to create APIs for model deployment.

Docker: Ensures model portability and consistent environment setup.

AWS & Google Cloud: Cloud platforms for hosting models, using services like AWS Lambda and GCP AI Platform for scalable deployment.

Streamlit: Provides a simple UI for end-users to interact with the prediction model.

5. Conclusion

5.1 Summary of Findings

The project successfully developed a machine learning model for predicting medical insurance costs. **XGBoost and Neural Networks** performed best, with key factors like age, BMI, and smoking status influencing predictions. Proper **data preprocessing** and **feature selection** significantly improved accuracy.

5.2 Future Enhancements

- **Adding more health indicators** for better prediction.
- **Using advanced deep learning models** like LSTMs.
- **Deploying a real-time API** for insurance companies.
- **Automating hyperparameter tuning** for optimization.

5.3 Final Recommendations

- **Insurance companies** can leverage this model for personalized pricing.
- **Further research** can explore time-series forecasting.
- **Cloud deployment** will enhance scalability and accessibility.

By implementing these enhancements, the predictive model can offer more reliable, efficient, and user-friendly medical insurance cost estimation.