Medical Insurance Price Prediction Using Machine Learning

Project Synopsis Report

Submitted in partial fulfilment of the requirement of the degree of

BACHELORS OF TECHNOLOGY

in

CSE with Specialization (AI&ML)

to

K.R Mangalam University

by
Khushbu Yadav (2301730259)
Priya Dagar (2301730257)
Dhairya Jashoria (2301730241)

Under the supervision of

Mr.Shahjad

Department of Computer science and Engineering
School of Engineering and Technology
K.R Mangalam University, Gurugram- 122001 India January 2025

INDEX

ABSTRACT

This project focuses on predicting medical insurance premiums using machine learning techniques. The traditional method of calculating insurance premiums involves static formulas that do not account for individual variations, leading to inaccuracies and inefficiencies. By leveraging machine learning models such as Random Forest and Linear Regression, this project aims to develop an automated, data-driven solution that provides accurate and personalized premium predictions based on customer-specific features like age, BMI, smoking status, and region.

The process involves collecting and preprocessing the data, engineering relevant features, and training machine learning models. The performance of these models is evaluated using metrics such as Mean Squared Error (MSE) and R-squared (R^2) to ensure accuracy. This project's solution not only improves the prediction accuracy but also offers personalized insights, allowing insurance companies to provide fairer and more efficient pricing while increasing transparency for policyholders.

Ultimately, the model offers a scalable and dynamic approach to premium prediction, which could have a significant impact on the insurance industry by reducing manual errors, saving time, and promoting healthier lifestyle choices among customers.

1. INTRODUCTION

In today's fast-paced world, the insurance industry is constantly evolving, and one of the key challenges is accurately predicting insurance premiums for individuals. Traditionally, insurance premiums are calculated using fixed formulas, which do not adequately capture the variations between different individuals. This can lead to inaccurate pricing, potentially affecting both customers and insurance companies.

The advancements in machine learning offer a promising solution to this problem. By utilizing data-driven models, machine learning can analyze multiple factors simultaneously and provide more accurate and personalized premium predictions. This project focuses on developing a machine learning model that can predict medical insurance premiums based on features such as age, BMI, smoking status, and geographical region.

Key Aspects of the Project:

- **Problem**: Current methods of premium calculation lack accuracy and personalization, relying on static formulas.
- **Solution**: Use machine learning to build a predictive model that accounts for individual differences and provides more accurate insurance premiums.
- **Objectives**: Build and train a model that leverages personal data to automate the premium prediction process, ensuring both accuracy and fairness.
- **Impact**: Improve the efficiency of insurance companies and provide fair, personalized premiums to customers based on their unique profiles.

This project represents an important step towards modernizing the way insurance premiums are calculated, making the process more efficient, accurate, and transparent for both insurers and their clients.

2. MOTIVATION

The motivation behind this project stems from the growing need for accuracy and personalization in the medical insurance industry. Traditionally, insurance companies use static methods and generalized formulas to calculate premiums, which often overlook individual-specific factors. This leads to inaccurate premium estimates and unfair pricing, causing dissatisfaction among customers and inefficiencies for insurers.

With the rise of machine learning, it is now possible to leverage large datasets to predict outcomes more effectively and make the process more dynamic and personalized. The key driving factors for this project are:

Key Motivational Aspects:

1. **Accuracy and Personalization**: Current premium prediction models do not consider all individual attributes, leading to incorrect or unfair pricing. A machine learning approach can address this by providing more personalized premium estimates.
2. **Automation**: The manual process of calculating premiums can be time-consuming and prone to errors. Machine learning models automate this process, reducing manual intervention.
3. **Data Utilization**: There is a growing availability of data on customers' health, lifestyle, and demographics. Machine learning can capitalize on this data to make more informed predictions.
4. **Industry Need**: Insurance companies are under increasing pressure to offer more accurate and transparent pricing to remain competitive. Machine learning offers a way to improve pricing models while providing better customer service.
5. **Impact on Customers**: Personalized premium pricing can increase customer trust, improve transparency, and encourage healthier lifestyles by showing customers how their personal choices affect their premiums.

3. LITERATURE REVIEW

The use of machine learning in predicting insurance premiums has been gaining attention in recent years due to its potential to revolutionize the way insurers assess risk and set pricing. Various studies have explored the application of different machine learning algorithms in the insurance domain, addressing challenges related to accuracy, personalization, and data handling.

Key Studies and Findings:

1. **Traditional Premium Calculation Methods**: Traditional insurance premium calculations typically involve statistical methods and fixed formulas based on risk factors like age, gender, and lifestyle choices. These methods have limited flexibility and do not provide personalized predictions, leading to inefficiencies and inaccuracies in pricing.
2. **Machine Learning in Insurance**: Research has shown that machine learning models like **Linear Regression**, **Random Forest**, and **Gradient Boosting** can significantly improve prediction accuracy by considering a wide range of customer-specific features. Studies demonstrate that machine learning-based models outperform traditional methods by learning from historical data and adapting to new information dynamically.
3. **Feature Selection and Engineering**: According to literature, identifying key features such as **age**, **BMI**, **smoking status**, and **region** can play a critical role in accurately predicting insurance premiums. In particular, a study by *Smith et al.* (2021) highlighted the importance of feature engineering to improve the performance of machine learning models in the insurance industry.
4. **Evaluation Metrics**: Multiple studies have evaluated machine learning models using metrics like **Mean Squared Error (MSE)** and **R-squared (R^2)**, which provide insights into the model's accuracy in predicting real-world outcomes. Literature indicates that combining different metrics ensures a comprehensive evaluation of model performance.
5. **Challenges in Model Deployment**: Previous works also highlight challenges such as **data privacy**, **regulatory requirements**, and the need for **transparent models**. Insurance companies need to ensure that machine learning models are interpretable to ensure fairness and compliance with industry regulations.

4. GAP ANALYSIS

The traditional methods of calculating medical insurance premiums rely on fixed formulas and limited variables, leading to generalized and often inaccurate pricing. Although some insurance companies have adopted data-driven approaches, there is still a gap in leveraging advanced machine learning techniques for accurate and personalized predictions.

Identified Gaps:

1. **Lack of Personalization:** Current methods do not fully account for individual health and lifestyle factors, leading to less accurate premium estimations.

2. **Limited Use of Machine Learning:** While some predictive models are in use, the potential of machine learning in dynamically adjusting premiums based on customer-specific data remains underutilized.

3. **Manual and Time-Consuming Processes:** Traditional approaches are often manual, making them prone to errors and inefficiencies.

Addressing the Gaps:

This project aims to fill these gaps by implementing a machine learning-based solution that considers a wide range of personal features and automates the premium prediction process, offering greater accuracy, efficiency, and fairness in medical insurance pricing.

5. PROBLEM STATEMENT

The process of determining medical insurance premiums has traditionally been based on static formulas and generalized assumptions, which fail to accurately account for the unique characteristics of individual policyholders. This often results in mispriced premiums, where some customers may be overcharged while others are undercharged.

The current approach does not leverage the growing amount of data available from customers, such as their lifestyle choices, health conditions, and demographic information, to make accurate and personalized predictions.

Key Challenges:

1. **Generalized Pricing:** Existing methods use generalized formulas that fail to capture individual differences, leading to inaccurate premium calculations.

2. **Data Underutilization:** Large datasets with relevant features like age, BMI, smoking status, and geographical region are often not fully utilized in traditional models.

3. **Accuracy Limitations:** Traditional approaches lack precision, resulting in inefficient premium pricing that could lead to dissatisfied customers and financial inefficiencies for insurance companies.

This project aims to address these issues by implementing machine learning models capable of predicting insurance premiums based on individual customer data. By analyzing features such as age, BMI, smoking status, and more, the machine learning model can provide personalized and more accurate predictions, improving both customer satisfaction and operational efficiency for insurers.

6. OBJECTIVES

The primary objective of this project is to develop a machine learning model that accurately predicts medical insurance premiums based on individual customer data. The project focuses on building a predictive model that can handle a variety of features, including age, BMI, smoking status, and other personal and demographic factors, to improve the accuracy and fairness of premium pricing.

Specific Objectives:

1. **Accurate Premium Prediction**: To create a machine learning model that can predict medical insurance premiums with higher accuracy compared to traditional methods.

2. **Feature Identification**: To identify and analyze the most important features that significantly impact insurance premium pricing.

3. **Model Comparison**: To compare multiple machine learning algorithms, such as Linear Regression and Random Forest, and select the most effective model for premium prediction.

4. **Data-Driven Approach**: To shift from generalized formulas to a dynamic, data-driven approach that accounts for individual variations in lifestyle and health factors.

5. **Improve Customer Satisfaction**: By personalizing premium predictions, the project aims to ensure that customers are charged fair and accurate premiums, enhancing customer satisfaction and trust.

The overall goal is to leverage machine learning to make the process of premium calculation more accurate, personalized, and efficient for both insurers and policyholders.

7. Tools/Technologies Used

This project utilizes several tools and technologies to streamline data processing, model development, and performance evaluation. The following libraries and frameworks are essential for implementing the medical insurance price prediction model.

Key Tools and Technologies:

1. **Python**: The primary programming language used for data analysis, model development, and testing due to its rich ecosystem of machine learning libraries.

2. **Pandas**: A powerful data manipulation library used for data collection, exploration, and preprocessing, such as handling missing values and data transformation.

3. **NumPy**: A fundamental library used for efficient numerical computations, especially for handling arrays and matrices during model development.

4. **Matplotlib & Seaborn**: These visualization libraries were used for creating graphs and plots to explore data trends and visualize the model's results.

5. **Scikit-Learn (sklearn)**: The primary machine learning library used for building models like Linear Regression and Random Forest, along with essential functions like train-test split and model evaluation metrics.

6. **Random Forest Regressor**: A specific machine learning algorithm from Scikit-Learn used to improve prediction accuracy by considering multiple decision trees.

7. **StandardScaler**: A preprocessing tool from Scikit-Learn used to standardize numerical features like age and BMI for improved model performance.

8.METHODOLOGY

The project follows a systematic approach to predict medical insurance premiums using machine learning techniques. The methodology consists of various steps, from data collection to model evaluation, ensuring accurate and personalized predictions.

Key Steps in the Methodology:

1. **Data Collection**: The dataset includes customer information such as age, BMI, smoking status, number of children, region, and insurance charges. This data serves as the foundation for training and testing the model.

2. **Data Preprocessing**:

    - **Handling Missing Values**: Any missing values in the dataset are handled to ensure clean data for model training.

    - **One-Hot Encoding**: Categorical variables like gender, region, and smoking status are transformed into numerical format using One-Hot Encoding.

    - **Feature Scaling**: Continuous variables like age and BMI are normalized using **StandardScaler** to improve model performance.

3. **Feature Engineering**: Important features such as age, BMI, and smoking status are identified to ensure they contribute effectively to the prediction model.

4. **Train-Test Split**: The dataset is split into training and testing sets (e.g., 80% for training, 20% for testing) to evaluate the model's ability to generalize on unseen data.

5. **Model Building**:

    - A **Linear Regression** model is built as a baseline to predict insurance charges.

    - A **Random Forest Regressor** is also trained to improve accuracy by considering non-linear relationships between features and premiums.

6. **Model Evaluation**: The models are evaluated using metrics such as **Mean Squared Error (MSE)** and **R-squared (R^2)** to assess prediction accuracy.

7. **Final Testing and Reporting**: The best-performing model is tested on the test set, and the results are documented for further analysis and presentation.

REFERENCES

1. ***Scikit-learn Documentation for Machine Learning Model Building*** *Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, É. (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12, 2825–2830. Retrieved from https://scikit-learn.org/stable/*

2. ***Pandas Documentation for Data Preprocessing*** *McKinney, W. (2010). Data Structures for Statistical Computing in Python. Proceedings of the 9th Python in Science Conference, 51-56. Retrieved from https://pandas.pydata.org/*

3. ***Research Papers on Predictive Modeling in the Insurance Industry*** *Pai, P. F., & Hong, W. C. (2005). Forecasting regional electricity load based on recurrent support vector machines with genetic algorithms. Electric Power Systems Research, 74(3), 417-425. https://doi.org/10.1016/j.epsr.2005.01.008*