**Medical Insurance Premium Prediction**

1. Himanshu Prajapati – 002774274 – [hprajapati1@student.gsu.edu](mailto:hprajapati1@student.gsu.edu)
2. Salini Mahapatra – 002693890 – [smahapatra4@student.gsu.edu](mailto:smahapatra4@student.gsu.edu)
3. Riya Singh - 002774332 - [rsingh28@student.gsu.edu](mailto:rsingh28@student.gsu.edu)
4. **Problem Statement:** Medical Insurance firms face the challenge of striking a balance between providing affordable premiums to policyholders while ensuring their own financial viability. This intricate task involves navigating a complex interplay of diverse factors. To remain competitive and profitable, insurance companies must adeptly consider these variables.

The primary goal of this undertaking is to construct robust predictive models that harness historical data and advanced analytical techniques. The precision of these predictions enhances risk assessment, pricing precision, customer contentment, and the financial solidity of insurance companies. This, in turn, safeguards equitable and sustainable healthcare coverage for policyholders.

Addressing this challenge yields mutual advantages for both insurers and the insured. It promotes a more streamlined and equitable healthcare insurance ecosystem, fostering efficiency and fairness.

1. **Data Sources:** <https://www.kaggle.com/datasets/tejashvi14/medical-insurance-premium-prediction>
2. **Data Preparation:** The data for this project was found to be well prepared and required minimal additional processing or cleaning. It was obtained in a structured format (CSV file) with the following attributes:

1. Age

2. Diabetes

3. BloodPressureProblems

4. AnyTransplants

5. AnyChronicDiseases

6. Height

7. Weight

8. KnownAllergies

9. HistoryOfCancerInFamily

10. NumberOfMajorSurgeries

11. PremiumPrice

No missing values, duplicates, or outliers were identified in the dataset. Additionally, the data was consistent with the project's requirements, and no data transformation or feature engineering was deemed necessary.

The dataset was ready for direct use in the model building, training, and evaluation.

1. **Data Exploration:** The exploratory data analysis (EDA) conducted using Tableau & seaborn library of Python provides valuable insights into the relationship between age groups, weight class, chronic disease history, blood pressure problems, and diabetes, and their impact on average premium prices.
2. **Correlation matrix using Seaborn library of Python :**

We analyzed correlations among variables using Python's Seaborn library to uncover data relationships

A screenshot of a graph

Description automatically generated

**Visualization using Tableau (**[**Tableau dashboard link**](https://public.tableau.com/app/profile/riya.singh5778/viz/MedicalInsurancePremium-DAProject/Dashboard1)**) :**

**Age vs. Average Premium:** The dashboard reveals that as age increases, the average premium also increases. This suggests that older individuals tend to pay higher premiums for insurance. When this is seen in context of body weight, we observe that within the age group of 18-30, the analysis indicates that higher weight classes are associated with higher premiums, highlighting the significance of weight for younger adults. In the 31-55 age range, individuals with mid-range weight classes tend to have higher average premiums, suggesting a nuanced relationship within this bracket. For those aged 50 and above, the data again shows that higher weight classes result in higher average premium prices. These findings underscore the complex interplay between age and weight class, providing insurers with valuable information for more accurate premium rate determinations.

**Chronic Disease History vs. Premium:** The tree chart demonstrates that individuals with a history of chronic diseases in their family tend to pay higher premiums compared to those without such a history. Furthermore, within each age group, older individuals consistently pay higher premiums, indicating that age is a significant factor in premium pricing.

**Blood Pressure Problems vs. Premium:** The bubble chart divides age groups into two classes: those with blood pressure problems and those without. Surprisingly, the data suggests that, for similar age groups, the average premium paid is relatively similar, regardless of whether individuals have hypertension or not. This finding challenges the assumption that blood pressure problems significantly impact premium pricing.

**Diabetes vs. Premium:** In the last bar graph, the analysis differentiates average premiums for each age group based on diabetes status. Interestingly, within each age group, individuals without diabetes tend to pay either the same or slightly higher premiums than those with diabetes. This peculiar finding raises questions about the factors influencing premium pricing for diabetic and non-diabetic populations.

In conclusion, this EDA on Tableau sheds light on the relationships between various factors and average premium prices. Age, weight class, chronic disease history, and diabetes status all play roles in determining premium pricing, but the interactions between these factors are complex and require further investigation to fully understand their impact.

1. **Prediction Models:** We used various machine learning models to tackle the problem at hand. We used 70% of the data for training these models and the remaining 30% of data for model validation. Our approach involved following 5 prediction models to predict the medical insurance premium:

* Multiple Linear Regression
* Boosted Regression Trees
* Random Forest Regression
* Artificial Neural Networks
* Genetic Algorithm

**A. Multiple Linear Regression:** Multiple Linear Regression was applied as a baseline model to establish a simple linear relationship between the target variable and the predictors. We conducted Multiple Linear Regression using Ordinary Least Squares with L2 regularization (weight: 0.001) and an intercept term included.

Tools used: Microsoft Azure ML

**B. Boosted Regression Trees:** Boosted Regression Trees were employed to capture non-linear relationships and interactions within the data. We employed an ensemble of 100 Boosted Regression Trees with a maximum of 20 leaves per tree, a minimum of 10 samples per leaf, and a learning rate of 0.2.

Tools used: Microsoft Azure ML

**C. Random Forest Regression:** Random Forest Regression was utilized to further enhance predictive accuracy. We implemented a model with 8 decision trees using Bagging resampling, each with a maximum depth of 32 and a minimum of 1 sample per leaf node.

Tools used: Microsoft Azure ML

**D. Artificial Neural Network (ANN):** We employed a fully connected Artificial Neural Network (ANN) with 100 hidden nodes. The ANN was trained with a learning rate of 0.1 over 100 iterations.

Tools used: Microsoft Azure ML

1. **Evaluation:** We critically evaluated all models' performance, analyzing key metrics R-squared and Root Mean Squared Error, to gauge their effectiveness in addressing the problem at hand.
2. **Multiple Linear Regression :** This model achieved a moderate R-squared value of 0.699, explaining a substantial portion of the variance in the data. The Root Mean Squared Error of 3563.216 showed reasonable model fit.

A screenshot of a computer

Description automatically generated

1. **Boosted Regression Tree:** Boosted Regression Tree Model achieved a strong coefficient of determination (R-squared) of 0.787, indicating its ability to explain a significant portion of the variance in the data. The Root Mean Squared Error (RMSE) of 2996.173 indicated good overall model fit.

A screenshot of a computer

Description automatically generated

1. **Random Forest Regression:** The Random Forest model exhibited strong predictive performance with an R-squared of 0.823, highlighting its ability to capture complex relationships within the data. A low MAE of 1212.838 and Relative Absolute Error of 0.235 reflected its accuracy in prediction. The Relative Squared Error of 0.177 and RMSE of 2728.146 indicated excellent overall fit.

A screenshot of a computer

Description automatically generated

1. **Artificial Neural Network (ANN):** The ANN model showed limited predictive capability with a negative R-squared value (-1.7152e-4), implying that it performed no better than a horizontal line fit to the data. The MAE of 5158.784 and RMSE of 6491.348 indicated substantial prediction errors, while Relative Absolute Error and Relative Squared Error were both exceptionally high at 1.0002.

A screenshot of a computer

Description automatically generated

In conclusion, the Boosted Regression Tree and Random Forest models outperformed others, demonstrating strong predictive capabilities and lower error metrics. Conversely, the ANN model exhibited poor performance, while the Multiple Linear Regression model displayed moderate predictive capability. Further optimization may enhance model performance in future iterations.

1. **Recommendations:** Based on our extensive analysis of your health insurance dataset, we have uncovered critical insights that can greatly inform the business strategy. The utilization of Random Forest Regression as the primary prediction method has delivered remarkable results in estimating insurance premium costs. Furthermore, we have observed that age exhibits the highest correlation with premium price, highlighting its pivotal role in determining insurance costs.

Two notable discoveries are the significant impact of prior major surgeries and organ transplants on premium rates. This underscores the need to enhance the pricing models to more accurately account for these variables. Surprisingly, individuals without diabetes are found to carry higher premium prices than their diabetic counterparts, signaling a potential necessity to reassess pricing structures for this demographic. Interestingly, hypertension appears to exert minimal influence on premium pricing, suggesting an opportunity to streamline risk assessment for individuals with this condition.

Considering these findings, we propose a dual-pronged strategy: firstly, further enhance the pricing model to encompass age, prior surgeries, and transplant history more comprehensively. Secondly, evaluate and potentially revise pricing for non-diabetic individuals. These actions will not only enhance the precision of premium calculation but also enhance the competitiveness of the medical insurance offerings, benefiting both the organization and the policyholders.