RUTENDO DARANGWA, ISOM835: COMPREHENSIVE ANALYSIS OF THE HEALTH INSURANCE PREMIUMS DATASET

1. EXECUTIVE SUMMARY

1.1. Brief Overview

Mispriced health premiums can lead to financial losses for insurers or unaffordable coverage for individuals. This project applies predictive analytics to a medical insurance dataset to understand which factors most strongly drive premium prices and to build models that can predict premiums for new customers.

1.2. Dataset Description

The project will use a dataset of 1,003 individuals with 10 predictor variables and a target variable, Premium Price. The original dataset contained slightly fewer than 1,000 rows. A few additional records, consistent with the original feature definitions, were added to reach the 1,003-row requirement for the project. The raw features were directly used for model training after splitting the data. The predictors describe demographic and health-related characteristics, including age, chronic disease status, transplant history, number of major surgeries, allergies, and family cancer history.

1.3. Approach Summary

After exploratory data analysis (EDA) and model development, three models were trained and compared on a held-out test set. The model used are Linear Regression, Decision Tree Regressor and Random Forest Regressor. The Random Forest Regressor achieved the best performance and these are the results, **MAE = 1,367, RMSE = 3,257 and R^2 = 0.784.**

1.4. Key Findings and Recommendations

The model explains about 78% of the variation in premium prices and typically predicts premiums within approximately $1,400 of the true value. Across EDA, correlation analysis, and Random Forest feature importance, age consistently emerged as the dominant driver of premium prices, followed by weight and transplant history. Individuals who are older, have higher body weight, or have a history of organ transplant are associated with significantly higher premiums. Chronic diseases, number of major surgeries, and history of cancer in the family also contribute positively to premiums but with smaller influence. The results suggest that premiums are largely driven by age and indicators of severe health risk. An insurer could use a Random-Forest–like model as a decision-support tool for pricing and risk assessment, while carefully considering fairness and regulatory constraints. Opportunities for targeted wellness or risk-management programs aimed at older, higher-risk customers to help control long-term costs can be implored.

2. INTRODUCTION AND BUSINESS CONTEXT

2.1. Business Problem and Objectives

Health insurance providers operate in an environment of rising medical costs, aging populations, and increasing prevalence of chronic diseases. To remain financially sustainable, insurers must accurately estimate the expected medical expenses of their customers and translate those into appropriate premium prices. At the same time, premiums must remain fair and affordable to avoid excluding vulnerable populations from coverage.

Predictive analytics and machine learning provide powerful tools for understanding how different risk factors contribute to medical costs and for building models that can estimate premiums more accurately. This project simulates a professional analytics engagement in which an insurer wants to use historical data on individuals' demographic and medical characteristics to better understand and predict premium prices.

2.2. Problem Definition

The goal of this project is to use a real-world dataset to:

2.2.1. Understand how demographic and health factors are associated with medical insurance premiums, and

2.2.2. Develop regression models that can predict premium prices for new individuals.

2.3. Research Questions

The analysis is framed by two main questions:

1. **RQ1 (Factor Influence**) *How do demographic and health factors (age, chronic diseases, transplant history, number of surgeries, etc.) influence individual medical insurance premium prices?*
2. **RQ2 (Predictive Modeling**) *Can we build accurate regression models to predict premium prices, and which features emerge as the most important predictors in these models?*

2.4. Contribution

This project delivers the following;

2.4.1. An end-to-end predictive analytics workflow: data understanding, EDA, preprocessing, modeling, evaluation, interpretation, and ethical reflection.

2.4.2. A comparison of three regression models: Linear Regression, Decision Tree Regressor, and Random Forest Regressor.

2.4.3. Interpretation of feature effects using both linear coefficients and Random Forest feature importances, translated into business insights relevant to health insurance pricing.

3. DATA DESCRIPTION AND EXPLORATORY DATA ANALYSIS (EDA)

3.1. Dataset Overview

The dataset used is a Medical Insurance Premium Prediction dataset (Medicalpremium.csv) consisting of 1,003 observations and 11 variables. The dataset was found to be clean with no missing values and all features were already in a suitable numerical format (int64). Therefore, no explicit feature engineering steps were performed. This makes the dataset convenient for modeling without extensive cleaning. This clean structure allowed the analysis to focus on modeling and interpretation rather than data cleaning.

3.2. Variable Definitions

The target variable

Premium Price – the insurance premium amount charged to each individual.

The predictor variables

i. Age: Age of the individual

ii. Diabetes: 0 = no diabetes, 1 = has diabetes

iii. Blood Pressure Problems: 0 = no BP problems, 1 = has problems

iv. Any Transplants: 0 = no transplants, 1 = has had a transplant

v. Any Chronic Diseases: 0 = none, 1 = at least one chronic disease

vi. Height: Height (same unit across individuals)

vii. Weight: Weight (same unit across individuals)

viii. Known Allergies: 0 = no known allergies, 1 = has allergies

ix. History of Cancer in Family: 0 = no history, 1 = family history present

x. Number of Major Surgeries: Number of major surgeries undergone

This is a supervised learning regression problem, given the above features, predict the numeric outcome Premium Price.

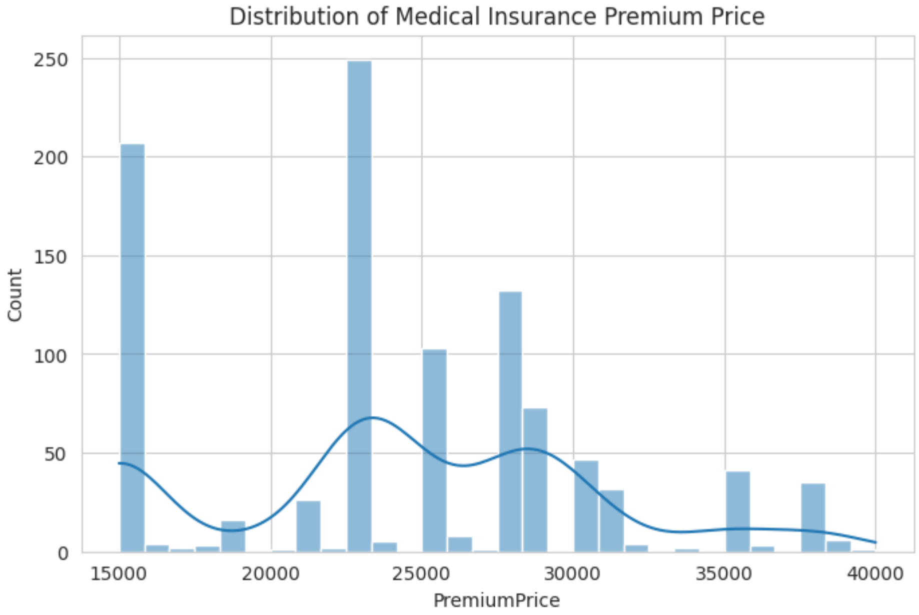
3.3. Univariate Analysis

Several histograms and density plots were created to understand variable distributions.

3.3.1. Premium Price

Distribution appears right-skewed with a concentration of mid-range premiums and fewer very high premiums. This suggests that while most individuals pay moderate premiums, a subset of high-risk individuals incur significantly higher charges.

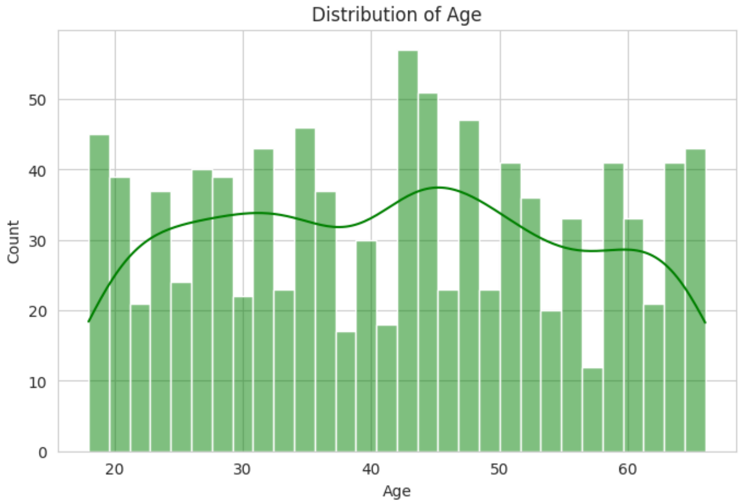
Figure 1: Distribution of Medical Insurance Premium Price



3.3.2. Age

The sample covers a broad range of ages, from younger adults to older individuals. Middle-aged and older adults are prominent, which is consistent with greater demand for medical insurance in these age groups.

Figure 2: Distribution of Age



3.4. Other numeric variables

Number of Major Surgeries is mostly low values (0–2), with relatively few individuals having many past surgeries. Binary variables such as Any Chronic Diseases, Any Transplants show typical imbalanced distributions. Most individuals do not have these severe conditions, while a minority do. These distributions highlight a realistic risk mix, predominantly lower-risk individuals with a smaller group of high-risk, high-cost cases.

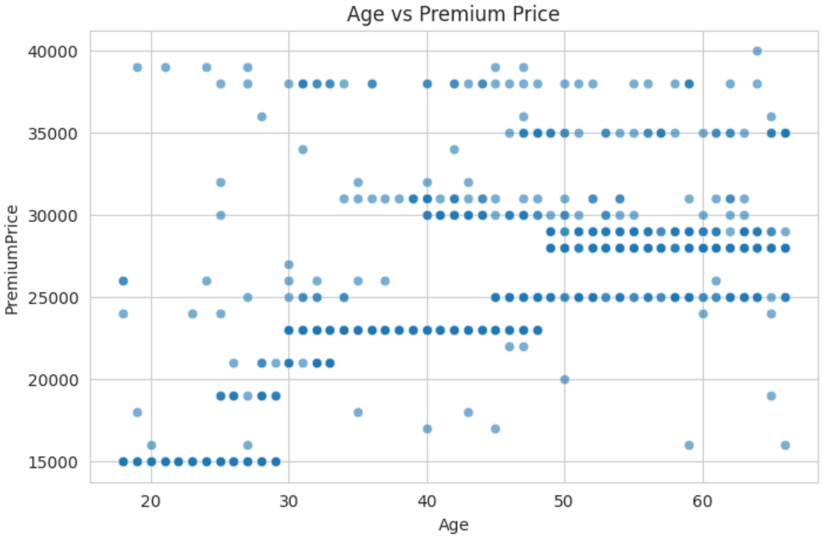
3.5. Bivariate Analysis with Premium Price

To understand how each factor relates to premium prices on the first research question, several scatterplots and boxplots were analyzed.

3.5.1. Age vs Premium Price (scatterplot)

Clear upward trend: as age increases, premium price tends to rise. Older individuals are clustered at higher premium levels, indicating that age is a key cost driver.

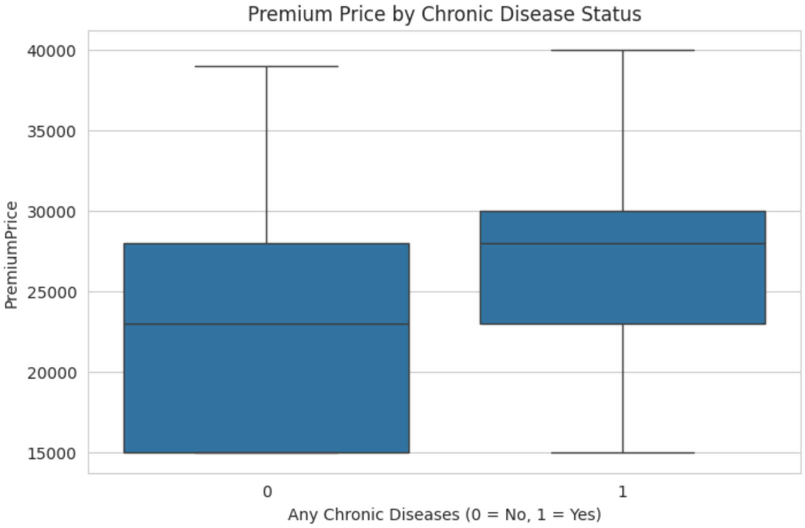
Figure 3: Age vs Premium Price



3.5.2. Chronic Diseases vs Premium Price (boxplot)

Individuals with Any Chronic Diseases = 1 have noticeably higher median and upper-quartile premium prices compared to those without chronic diseases.

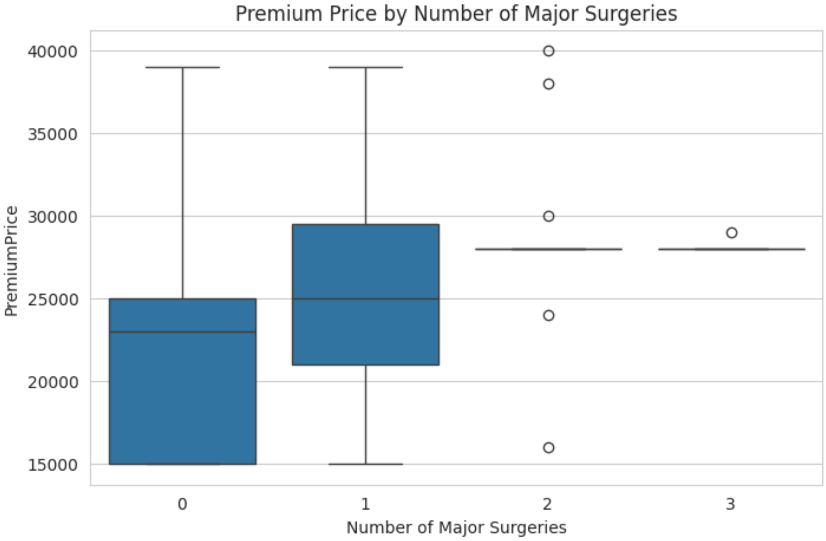
Figure 4: Premium Price vs Chronic Disease Status



3.5.3. Number of Major Surgeries vs Premium Price (boxplot)

Premiums increase as the number of major surgeries rise, although the relationship is less smooth than that of age due to fewer observations at higher surgery counts.

Figure 5: Premium Price vs Number of Major Surgeries

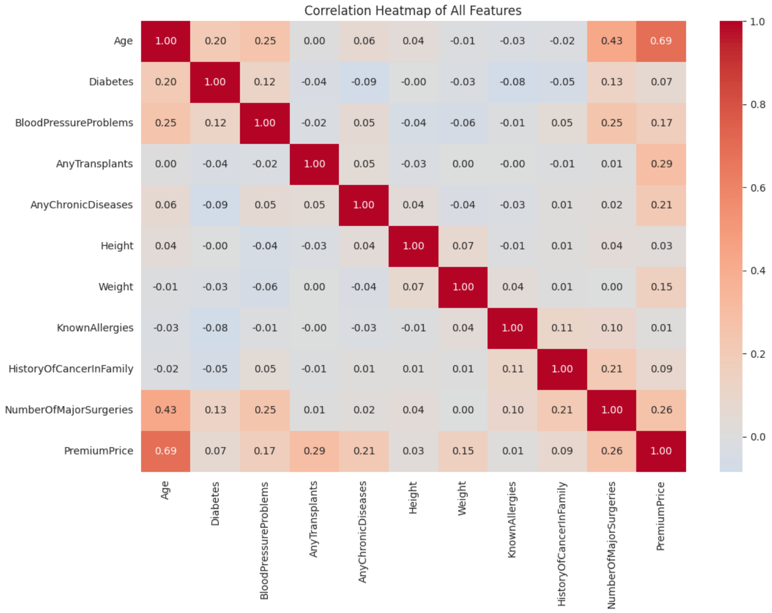


3.5.4. Blood Pressure Problems and Family Cancer History

Both conditions show somewhat higher premiums in the affected group, but with more overlapping ranges than transplants or chronic diseases. Overall, the bivariate plots suggest that age, transplants, chronic diseases, and major surgeries are associated with higher premiums.

3.6. Correlation Analysis

Figure 6: A correlation heatmap of all variables



i. Age and Premium Price: strongest positive correlation (0.69).

ii. Any Transplants and Premium Price: moderate positive correlation (0.29).

iii. Number of Major Surgeries and Premium Price: moderate positive correlation (0.26).

iv. Mild correlations also exist between age and some medical history variables (for example age and number of surgeries).

These correlations confirm EDA observations that older age and major health events are strongly linked with higher premium prices. However, correlation remains pairwise and does not account for the joint effects of all variables, which is tackled through regression modeling.

3.7. Key EDA Insights

i. Premium prices rise substantially with age, making it the dominant risk factor visually and statistically.

ii. Severe medical conditions—transplants, chronic diseases, and major surgeries—are associated with higher premiums.

iii. The dataset is clean (no missing values, all numeric), which simplifies preprocessing.

iv. There is potential for multicollinearity, as age and some medical history variables are correlated, motivating the use of models beyond simple linear regression.

4. METHODOLOGY

4.1. Data Preprocessing

Preprocessing steps were relatively simple due to the dataset's cleanliness:

4.1.1. Missing values

df.isnull().sum() confirmed zero missing values across all variables. No imputation or row removal was needed.

4.1.2. Data types

All variables were stored as integers. Predictors were already numeric, so no encoding or transformation was required.

4.1.3. Feature and target definition

Features (x): all columns except Premium Price, **Target (y)**: Premium Price.

4.1.4. Data splitting

An **80/20 train–test split** was used. X_train, X_test, y_train, y_test, and used test_size=0.2, random_state=42 for reproducibility.

4.2. Feature Engineering decisions

No explicit feature engineering such as BMI was performed. The focus was on understanding how the original, interpretable variables relate to premium prices.

4.3. Modeling Approaches

Three regression models were implemented using scikit-learn:

4.3.1. Linear Regression

Serves as a simple baseline model assuming a linear relationship between each predictor and premium price. The advantage is that coefficients are easy to interpret and explain to non-technical stakeholders. However, it cannot capture nonlinear interactions or complex patterns.

4.3.2. Decision Tree Regressor

A tree-based model that recursively splits the data into subsets based on feature values was used. It captures nonlinear relationships and interactions among predictors. It was configured with max_depth=4 to limit complexity and reduce overfitting.

4.3.3. Random Forest Regressor

An ensemble of many decision trees trained on bootstrap samples with random feature subsets was done. More robust and accurate than a single tree and it reduces variance through averaging. The configuration was n_estimators=100 (number of trees), max_depth=10 and random_state=42.

4.4. Evaluation Strategy

To fairly compare models, all were trained on the same training set, and evaluated on the same held-out test set. The following metrics were used;

i. **Mean Absolute Error (MAE):** Average absolute difference between predicted and actual premiums, interpretable directly in currency units.

ii. **Root Mean Squared Error (RMSE):** Square root of the average squared error; penalizes large errors more heavily than MAE.

iii. **R^2 (Coefficient of Determination):** Proportion of variance in Premium Price explained by the model (1.0 = perfect fit; 0 = no explanatory power).

This evaluation approach avoids data leakage and provides an honest assessment of how each model might perform on new insurance applicants.

5. RESULTS & MODEL COMPARISON

5.1. Model Performance

The three models' performances on the test set are summarized below;

Table 1: Model Performance Results

Model	MAE	RMSE	R²
Linear Regression	2,923.46	4,289.27	0.625
Decision Tree	2,070.48	3,625.07	0.732
Random Forest	1,367.29	3,257.23	0.784

5.1.1. Linear Regression

i. Explains about 62.5% of the variance in premium prices.

ii. Average error of about $2,923 per person.

iii. As a simple baseline, it demonstrates that linear relationships capture much but not all of the signal.

5.1.2. Decision Tree Regressor

i. Improves R² to approximately 0.73, with noticeable reductions in MAE and RMSE versus Linear Regression.

ii. Captures nonlinearities, such as thresholds where premiums increase sharply for higher-risk individuals.

5.1.3. Random Forest Regressor

i. This delivered the best performance with R² = 0.784 (explains ~78.4% of variance) and MAE = 1,367, RMSE = 3,257

ii. It Indicates that ensemble tree methods are particularly effective for this dataset.

In overall, both tree-based methods outperform linear regression, and the Random Forest achieves the strongest predictive performance.

5.2. Interpretation of Linear Regression Coefficients

The linear regression model's coefficients (sorted) were;

i. **Large positive coefficients**

Any Transplants = +7,016, Any Chronic Diseases = +2,579 and History of Cancer in Family = +2,140

ii. **Moderate positive coefficients**

Known Allergies = +326, Age = +325, Weight = +74 and Blood Pressure Problems = +38

iii. **Negative coefficients**

Height = −4, Diabetes = −454 and Number of Major Surgeries = −653

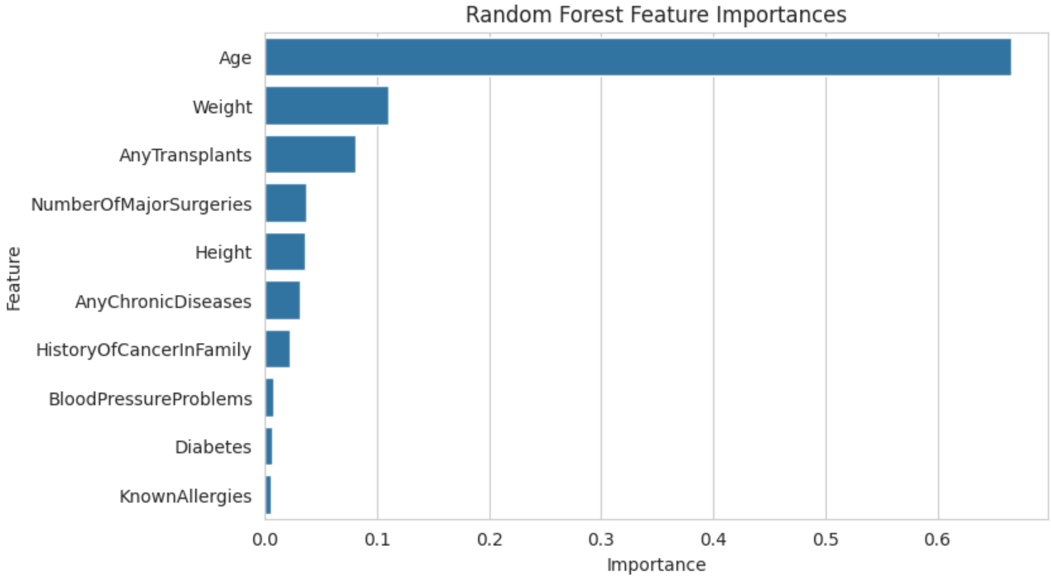
5.2.1. Interpreting these coefficients

Holding all else constant, having a transplant is associated with an increase of roughly 7,000 in Premium Price, while having chronic disease or a family history of cancer increase premiums by around 2,200 – 2,600. Age has a positive coefficient, each additional year is associated with 325 higher premium, so a 20-year difference could respond to 6500 extra. Some negative coefficients, for Diabetes and Number of Major Surgeries are somewhat counterintuitive. This may be due to multicollinearity, where related variables (such as age, surgeries, and other conditions) share explanatory power, leading some coefficients to become negative when controlled together. This highlights a limitation of linear regression in this context, while interpretable, its coefficient signs can be influenced by overlapping information among predictors.

5.3. Random Forest Feature Importances

5.3.1. The Random Forest model provides feature importances, measuring each feature’s contribution to reducing prediction error across all trees.

Figure 7: **Random Forest Feature Importances**



i. Age – **0.666**
ii. Weight – **0.110**
iii. Any Transplants – **0.080**
iv. Number of Major Surgeries – 0.036
v. Height – 0.035
vi. Any Chronic Diseases – 0.031
vii. History of Cancer in Family – 0.022
viii. Blood Pressure Problems – 0.007
ix. Diabetes – 0.006
x. Known Allergies – 0.006

Age is by far the most important predictor, accounting for about two-thirds of the model’s explanatory power. Weight and Any Transplants are also substantial contributors. Other factors—chronic diseases, surgeries, cancer history—still matter but are less critical than age in the Random Forest model. Unlike the linear coefficients, Random Forest importances align more closely with the correlation analysis, reinforcing the conclusion that age is the dominant driver of premium prices.

5.4. Final Model Selection

Given its superior performance higher R^2 and lower MAE/RMSE, Random Forest was selected as the final model for this project. This is because it captures nonlinear and interaction effects. It is more robust than a single decision tree. The model also provides interpretable feature importances that are consistent with domain expectations.

5.5. Summary with Respect to Research Questions

RQ1 (Factor Influence): The analysis shows that age, weight, and severe medical history (transplants, chronic diseases, surgeries) are positively associated with higher premium prices. Age is consistently the dominant factor across correlation, linear regression, and Random Forest feature importance.

RQ2 (Predictive Modeling): Among the three regression models tested, the Random Forest Regressor achieves the best predictive performance ($R^2 = 0.784$), indicating that machine learning can explain a substantial share of the variation in premium prices using the available features.

6. BUSINESS INSIGHTS AND RECOMMENDATIONS

6.1. Key Insights

It is critical to note that these key insights are derived from the combined EDA and modeling results

6.1.1. **Age is the primary cost driver.** Premium prices increase sharply with age, and age is the single most important predictor in the Random Forest model.

6.1.2. **Severe medical history significantly increases premiums.** Individuals with past transplants, chronic diseases, or multiple major surgeries tend to face higher premiums, reflecting their elevated expected costs.

6.1.3. **Body weight is an important secondary factor.** Weight emerges as the second-most important variable in the Random Forest model, suggesting that higher body weight—often a proxy for future health risk—is associated with higher premiums.

6.2. **Other factors have smaller but positive contributions.** Family history of cancer, allergies, and blood pressure issues all contribute to higher premiums, though their influence is smaller compared with age and severe events.

7. RECOMMENDATIONS FOR AN INSURANCE COMPANY

7.1. Use the model as decision support, not automatic pricing

If this modeling approach were used in a real insurance context, it would be advisable to use the model as decision support, not automatic pricing. A Random-Forest–based model can provide a strong, data-driven estimate of appropriate premium levels. However, final pricing decisions should remain under human and policy control, incorporating regulatory constraints and fairness considerations.

7.2. Develop risk-based wellness programs

In addition to that, the insurance company would be advised to develop risk-based wellness programs. Since age, weight, and serious medical conditions drive premiums upward, insurers could offer wellness or chronic-disease management programs targeted at older and higher-weight individuals. The Insurance companies can also provide incentives for example, reduced premium increases for participation in preventive health initiatives.

7.3. Segment customers based on risk factors

The model's insights support risk segmentation, such as Lower-risk segment for young individuals with no chronic conditions; moderate-risk for middle-aged individuals or those with moderate health issues as well as high-risk for older individuals with transplants or multiple major surgeries.

7.4. Monitor model performance over time

It may be also advisable for insurance companies that the model should be retrained and evaluated regularly using new data to maintain accuracy. This is because medical costs and population health trends change.

8. ETHICS AND RESPONSIBLE AI

8.1. Potential Biases and Fairness Concerns

Even though the dataset does not explicitly include protected characteristics such as race or income, the model raises important ethical questions:

8.1.1. Age discrimination

Age is the strongest predictor of premium price. Heavy reliance on age could be considered discriminatory against older individuals, even if it reflects actual cost differences.

8.1.2. Health-status discrimination

Using variables like transplants and chronic diseases to set premiums may penalize those who already face significant health burdens, many of which may be outside their control.

8.1.3. Proxy bias

Variables like weight or chronic diseases may indirectly correlate with socioeconomic status or lifestyle factors, potentially introducing hidden bias.

8.2. Privacy and Data Protection

It is of great importance to note that real-world deployment of such a model would require strong data privacy controls for sensitive health information. In this context, it is critical to maintain compliance with health data regulations such as HIPAA in the U.S. Also, clear consent mechanisms for customers whose data is being used for modeling should be sought.

8.3. Fairness Mitigation Strategies

In order to address fairness insurers could introduce caps or subsidies for vulnerable groups such as older adults to prevent unaffordable prices. In addition to that, regular fairness audits could be performed by checking whether error rates or premium recommendations differ systematically across age groups or health conditions in problematic ways. Moreover, policymakers could define acceptable ranges for how much premiums can vary based on particular health factors.

8.4. Transparency and Explainability

In undertaking such researches, transparency is critical for customer trust. Linear Regression coefficients and Random Forest feature importances give interpretable explanations for why premiums differ between people. Stakeholders should receive plain-language explanations, such as; "Your premium is higher primarily because of your age and history of serious medical conditions." By combining predictive accuracy with clear explanations and fairness checks, insurers can use AI responsibly in pricing decisions.

9. CONCLUSION AND FUTURE WORK

9.1. Summary of Findings

This project applied a complete predictive analytics workflow to a medical insurance premium dataset of 1,003 individuals. The main findings are that EDA and correlation analysis showed that age and serious medical history (transplants, chronic diseases, surgeries) are strongly associated with higher premiums.

In addition to that, three regression models were built and compared the Linear Regression ($R^2 = 0.625$), Decision Tree Regressor ($R^2 = 0.732$) as well as the Random Forest Regressor (R^2, $= 0.784$). The Random Forest model achieved the best predictive performance, with MAE = 1,367 and RMSE = 3,257 on the test set. Feature importance analysis revealed that age is the dominant predictor of premium price, followed by weight and transplant history, with chronic diseases and surgeries contributing additional risk.

Overall, the project demonstrates that machine learning models can effectively use demographic and health information to predict insurance premiums and provide insight into the main cost drivers.

9.2. Limitations

9.2.1. Dataset Size and Representativeness

The findings had several limitations and these should be acknowledged. The dataset size and representativeness are important aspects to know when dealing with such important information with the view to make informed decisions. With 1,003 observations from an unknown source population, results may not generalize to all regions, insurers, or customer types.

9.2.2. Limited feature set

Important factors such as income, lifestyle behaviors such as smoking and exercise, or detailed medical history are not included. Thes factors may have critical impact on the insurance premiums.

9.2.3. Simple model tuning

In addition to that, only basic hyperparameter choices were used such as fixed depth and number of trees. More systematic tuning could further improve performance.

9.2.4. No explicit fairness or cost-sensitivity optimization

The model was evaluated purely on predictive accuracy, not on fairness metrics or business cost considerations.

9.3. Suggestions for Future Improvements

Potential directions for extending this work for improvements include collecting larger and richer datasets covering different geographies and more detailed health and socioeconomic information. Apart from that in future, the researcher may consider applying cross-validation and hyperparameter tuning for instance, Grid Search CV to systematically optimize model performance.

Exploring additional models such as Gradient Boosting Machines (GBM) or XGBoost for potentially better predictive accuracy may be also implored in future. Additionally, integrating fairness metrics and constraints into the modeling process to explicitly measure and mitigate adverse impacts on vulnerable groups may be also considered. It is also to consider segmenting the analysis by subgroups for example, age bands to see whether the model performs equally well across the entire population.

9.4. Lessons Learned

The project helped me connect course concepts such as EDA, model evaluation, overfitting, feature importance, and responsible AI, to a concrete business problem and gave me hands-on experience implementing an end-to-end predictive analytics workflow in Python. The following is an outline of the lessons I learnt from this project.

9.4.1. **Data quality and structure strongly shape the project.** Starting from a clean, fully numeric dataset made it easier to focus on EDA and modeling rather than extensive cleaning or imputation. At the same time, I realized that a simple dataset can limit which questions can be answered, for example, no information about income, lifestyle, or more detailed diagnoses.

9.4.2. **EDA and correlations are necessary but not sufficient.** Initial plots and the correlation heatmap suggested that age and major medical events are strongly related to premium prices. However, the linear regression coefficients did not perfectly match the correlation signs, for example, negative coefficients for diabetes and number of surgeries, which taught me how multicollinearity and model assumptions can affect interpretation.

9.4.3. **Model choice and complexity matter.** Comparing Linear Regression, Decision Tree, and Random Forest showed that more flexible, nonlinear models especially Random Forest can capture patterns that simpler models miss. I learned that a baseline linear model is useful for interpretation, but an ensemble model can provide better predictive accuracy.

9.4.4. **Interpretability vs. performance is a trade-off.** Linear Regression was easier to explain but less accurate, while Random Forest was more accurate but less transparent at the individual decision level. This highlighted the importance of balancing interpretability and performance depending on the business use case.

9.4.5. **Ethical and fairness considerations cannot be an afterthought.** Even though the dataset did not include protected attributes like race or income, thinking through issues such as age discrimination and health-status discrimination made it clear that accurate models can still raise fairness concerns. This reinforced the need to combine technical modeling with ethical and regulatory perspectives when working with sensitive data such as health records.

REFERENCES

1. Medical Insurance Premium Prediction dataset (Medicalpremium.csv), accessed from an open data source.
2. Pedregosa et al. (2011). *Scikit-learn: Machine Learning in Python.* Journal of Machine Learning Research.
3. Wes McKinney (2010). *Data Structures for Statistical Computing in Python.* Proceedings of the 9th Python in Science Conference.
4. Hunter, J. D. (2007). *Matplotlib: A 2D Graphics Environment.* Computing in Science & Engineering.