In this part of the project, we built a regression model to predict annual insurance premiums using the dataset premium\_data.csv. The main goal was to simulate the insurance pricing process, where demographic, financial, and medical risk factors are combined to determine fair premiums. The dataset included numerical variables such as Age, BMI, Credit Score, and Sum Insured, as well as categorical factors like Smoking Status, Gender, Region, Education Level, Age Group, Income Level, Credit Category, Pre-existing Conditions, Family Medical History, and High Risk. To prepare the data, we removed missing values, recoded blank entries in Income Level as NA and dropped them, and converted all categorical variables into factors so that the regression could produce consistent and interpretable results. Continuous predictors were kept in their original form.

The theoretical premium structure was based on the regression formulation



\[

\text{Premium}\_i \approx \alpha + \beta\_1 \cdot \text{Age}\_i + \beta\_2 \cdot \text{BMI}\_i + \beta\_3 \cdot \text{SumInsured}\_i + \beta\_4 \cdot \text{CreditScore}\_i + \sum\_{k} \gamma\_k \cdot X\_{ik},

\]where the categorical indicators$X\_{ik}$ represent smoking, income, education, or medical risk attributes. Exploratory analysis revealed patterns consistent with actuarial reasoning: smokers and individuals with pre-existing conditions or family medical history consistently faced higher premiums, while income level influenced policy sizes, with elite groups tending to purchase larger sums insured and thereby incurring higher premiums. Among continuous variables, Premium was almost linearly related to Sum.Insured, confirming proportionality between coverage and cost, while Age and BMI also showed positive associations and Credit.Score was negatively related, implying that lower creditworthiness corresponds to higher premiums.

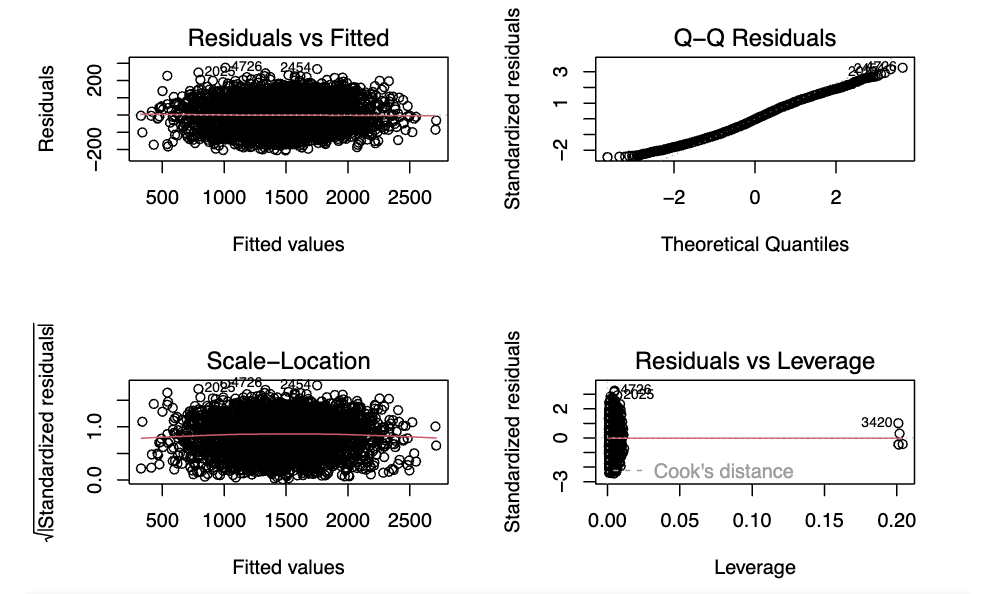
日程表

AI 生成的内容可能不正确。图表, 条形图, 瀑布图

AI 生成的内容可能不正确。

Based on these findings, two models were developed: a linear regression model and a random forest. The linear regression provided interpretable coefficient estimates, while the random forest captured nonlinear interactions and ranked variable importance, highlighting Sum.Insured, Credit.Score, and Smoking.Status as dominant predictors. Model validation was conducted via a 70-30 train-test split. The linear regression achieved RMSE ≈ 85 and MAE ≈ 71 on the test set, while the random forest was less accurate with RMSE ≈ 109 and MAE ≈ 88. Diagnostic checks confirmed that residuals of the linear regression were approximately normal and homoscedastic, with no multicollinearity issues (all VIF values below 5). Interestingly, the High\_Risk variable was statistically insignificant, suggesting that its impact overlapped with smoking and medical history.

图表, 散点图

AI 生成的内容可能不正确。

图示

AI 生成的内容可能不正确。图表, 散点图

AI 生成的内容可能不正确。

From a business perspective, the regression model is highly interpretable and matches real-world expectations. Age, BMI, and Sum Insured all raise premiums; Credit Score lowers them; smokers and medically at-risk individuals face extra charges; and income level reflects affordability and policy size choices. These results show that the model captures the main risk factors used in insurance pricing. However, there are some limitations: the imbalance between smokers and non-smokers may bias results, and while linear regression fit the data well, additional variables and nonlinear methods could improve accuracy in larger datasets. Stress testing with extreme profiles would also help check robustness.Overall, we recommend using the linear regression model as the main pricing tool because it balances accuracy with transparency, making it suitable for both actuarial practice and regulatory reporting. The random forest can be used as a secondary benchmark for nonlinear effects. As more customer data becomes available, the company should continue refining the model, expanding the feature set, and exploring advanced machine learning methods to further improve predictive accuracy and business value.