**Report for Question g: Predicting Premium Amount**

**1. Introduction**

In this part of the project, we aim to construct a regression model that predicts the annual premium amount for customers in the given insurance dataset (*premium\_data.csv*). The purpose of this exercise is to mimic the process of insurance pricing, where customer demographics, health indicators, and risk factors are taken into account to determine the appropriate premium. To achieve this, we carried out several steps: cleaning and preparing the data, conducting exploratory data analysis (EDA), engineering relevant features, developing and comparing multiple models, validating model performance, and finally diagnosing potential model violations. By systematically following these steps, we ensure that the final model is not only accurate in prediction but also interpretable and practically applicable in an insurance business context.

**2. Data Preparation and Feature Engineering**

The raw dataset contained both numerical and categorical variables. To begin, we removed observations with missing values and specifically addressed the issue of blank entries in the variable *Income\_Level*. These blank values were first converted to missing values and then removed from the dataset to avoid misleading baseline comparisons. Following this, all categorical variables were properly converted to factor variables, which included *Smoking.Status, Gender, Region, Educational.Level, Age\_Groups, Income\_Level, Credit\_Category, Pre.existing.Conditions, Family.Medical.History,* and *High\_Risk*. Numerical variables such as *Age, BMI, Credit.Score,* and *Sum.Insured* were retained in their original form. Through this preparation, the dataset was cleaned and structured in a way that supported reliable regression analysis and machine learning modeling. This step is crucial because insurance premium prediction depends heavily on correctly distinguishing categorical effects such as smoking behavior or income group from continuous variables such as credit score or age.

**3. Exploratory Data Analysis**

Exploratory analysis was conducted to understand the distribution of premiums across different customer groups and to evaluate the correlation structure among continuous variables. Boxplots were generated to illustrate how premium amounts vary across categorical predictors. For example, smokers were shown to have significantly higher premiums compared to non-smokers, while income level also displayed systematic differences, with elite groups tending to purchase larger policies and therefore paying higher premiums. Similarly, customers with pre-existing conditions or family medical history of illness were charged higher premiums on average. In parallel, correlation plots of the continuous predictors revealed that *Premium Amount* is strongly related to *Sum.Insured*, as expected, since the insured face amount is a direct driver of premium levels. Age and BMI also showed positive associations with premiums, while Credit.Score was negatively correlated, indicating that customers with lower credit scores tend to pay higher premiums. These findings from EDA guided the inclusion of these variables in the regression model and confirmed their practical relevance.

**4. Model Development and Selection**

Based on the exploratory findings, we developed two predictive models: a linear regression model and a random forest model. The linear regression model included all key predictors: age, BMI, credit score, sum insured, smoking status, pre-existing conditions, family medical history, income level, and the high-risk indicator. This model allowed us to quantify the effect of each variable on the premium in an interpretable way. In contrast, the random forest model was constructed as a non-parametric alternative that can capture complex, nonlinear interactions between predictors. Variable importance plots from the random forest highlighted *Credit.Score, Smoking.Status,* and *Sum.Insured* as the most influential variables. However, the random forest model tended to smooth out the strong linear relationship between *Sum.Insured* and *Premium Amount*, which limited its predictive accuracy compared to the regression model. Overall, the two models provided complementary perspectives: the linear model offered clear interpretation of coefficients, while the random forest confirmed the ranking of important predictors.

**5. Model Validation and Diagnostics**

To validate the models, the dataset was split into training and testing subsets with a 70-30 split. Predictions were generated on the test dataset and compared to actual premium values. For the linear regression model, the root mean squared error (RMSE) was approximately 85 and the mean absolute error (MAE) around 71, indicating strong predictive accuracy. For the random forest model, the RMSE was higher at approximately 109 and the MAE was around 88, suggesting weaker performance. Diagnostic checks were also conducted on the linear regression model. Residual plots and Q-Q plots confirmed that residuals were approximately normally distributed and homoscedastic. The Variance Inflation Factor (VIF) test showed no significant multicollinearity among the predictors, with all values well below critical thresholds. While the high-risk variable was not statistically significant in the regression model, its effect may already have been captured indirectly through other correlated variables such as smoking status or medical history. These diagnostics reassured us that the linear regression model was appropriate and stable for this dataset.

**6. Interpretation and Business Insights**

The regression results provide meaningful business insights into premium pricing. Age, BMI, and sum insured all exert a strong positive influence on premium amounts, reflecting the higher mortality and claim risk associated with older, less healthy individuals, as well as the proportionality between coverage size and premium. Credit score was negatively related to premiums, consistent with the idea that individuals with lower creditworthiness are viewed as higher risk. Smoking status was highly significant, with smokers facing considerably higher premiums, while non-smokers received a discount relative to the baseline. Customers with pre-existing conditions or family medical history were also charged higher premiums, reflecting the elevated health risks. Interestingly, the high-risk flag itself was not significant, which implies that its explanatory power may overlap with other health indicators. Income level was also important, with low and medium income groups associated with systematically lower premiums, which may reflect differences in policy choices and affordability. These findings confirm that the model not only predicts premiums effectively but also aligns with actuarial intuition.

**7. Limitations and Recommendations**

Although the linear regression model performed well, there are some limitations. First, the dataset is limited in scope and may not represent the full population of insurance customers. Important variables such as lifestyle habits beyond smoking, geographic location, or occupation are absent and could further improve the predictive power. Second, the imbalance between smoker and non-smoker groups means that the model could underpredict risk for minority groups. Third, while the linear model captured relationships well, nonlinear effects or interactions might become important in larger, more complex datasets. Future work could involve testing regularized models such as Lasso or Ridge regression to reduce potential overfitting, as well as expanding the feature set to include additional demographic and health factors. Stress testing the model with simulated customer profiles would also help assess robustness in extreme scenarios.

From a practical business perspective, we recommend adopting the linear regression model as the primary pricing tool because of its superior accuracy and interpretability. It provides clear guidance on how premiums respond to changes in customer characteristics, which is essential for both actuarial decision-making and transparent communication with regulators. The random forest model, while less accurate here, could still be used as a supplementary tool to cross-check nonlinear interactions. Ultimately, the company should continue refining its pricing models with additional data and explore more advanced machine learning techniques once sufficient real-world policyholder data is accumulated.