**Predicting Medical Insurance Premiums**

**Author: Lewis So**

**Introduction**

The accurate prediction of medical insurance premiums is a fundamental challenge within the insurance industry, directly impacting pricing fairness, financial risk management, and competitive positioning. This project, undertaken as a core component of the Introduction to Business Analytics course, aims to develop a sophisticated predictive model for PremiumPrice using a comprehensive dataset of 986 individual records. These records encompass a wide array of medical and demographic attributes, including Age, BMI, Diabetes, AnyTransplants, AnyChronicDiseases, HistoryOfCancerInFamily, and NumberOfMajorSurgeries, providing a robust foundation for in-depth analysis. The central objective is to construct a model that elucidates a minimum of 60% of the variance in PremiumPrice, with validation conducted on an independent test set to ensure reliability. This initiative is driven by the need to support data-informed pricing strategies, a critical aspect of modern insurance operations.

The dataset, potentially derived from a course-provided resource or a simulated insurance database, presents a diverse set of predictors that mirror individual health risk profiles. Initial explorations hinted at non-linear dynamics, particularly between Age and PremiumPrice, alongside evident skewness in the response variable, necessitating the application of advanced statistical methodologies. The business context is highly pertinent, as accurate premium forecasting can optimize revenue streams, enhance customer satisfaction through equitable pricing, and ensure adherence to regulatory standards. This report offers an exhaustive examination of the methodology employed, presents a detailed exposition of the model’s results with extensive validation, engages in a thorough discussion of the findings and their implications, and concludes with actionable recommendations for practical application and potential future enhancements.

**Methodology**

**Data Preparation**

The dataset, initially comprising 986 rows and 13 columns, underwent a rigorous preparation process to guarantee its readiness for modeling. A preliminary inspection was conducted to assess data integrity, employing Python’s pandas library to execute commands such as df.isnull().sum() to detect missing values. The analysis confirmed the absence of significant gaps, though minor inconsistencies—such as potential encoding errors or extreme outliers—were identified and addressed. For continuous variables like Age and BMI, where anomalies were suspected, a median-based imputation strategy was adopted to preserve the dataset’s statistical properties, ensuring that the distribution remained representative. This approach was chosen over mean imputation to avoid skewing the data due to potential outliers, a decision validated by subsequent distribution checks.

Feature selection emerged as a cornerstone of the preparation phase, guided by insights from exploratory data analysis. Correlation analysis, facilitated by pandas’ .corr() function, revealed a strong linear relationship between Weight and BMI (correlation coefficient r = 0.820679), indicating redundancy. Similarly, Height exhibited a negligible correlation with PremiumPrice (r = 0.026910), suggesting limited predictive power. Consequently, KnownAllergies, Height, and Weight were systematically excluded to mitigate multicollinearity and enhance model interpretability. The refined feature set—comprising Age, BMI, Diabetes, BloodPressureProblems, AnyTransplants, AnyChronicDiseases, HistoryOfCancerInFamily, and NumberOfMajorSurgeries—was selected as the basis for prediction, with PremiumPrice designated as the target variable. To further validate this selection, initial variance inflation factor (VIF) calculations were performed on preliminary models, confirming that the retained predictors maintained independence, with VIF values generally below 5 prior to centering adjustments.

**Exploratory Data Analysis**

The exploratory data analysis phase served as a critical preliminary step, providing a deep understanding of the dataset’s underlying structure and inter-variable relationships. Utilizing matplotlib, a series of scatter plots were generated to visualize the relationships between predictors and PremiumPrice. A notable non-linear trend was observed between Age and PremiumPrice, characterized by an initial increase in premiums that tapered off in older age groups, suggesting the necessity of incorporating a quadratic term to capture this curvature accurately. Correlation analysis, conducted with pandas’ .corr() function, further substantiated the relevance of BMI as a predictor, showing a moderate positive correlation with PremiumPrice, while reinforcing the decision to exclude Height and Weight due to their overlap with BMI.

Detailed histograms of PremiumPrice were constructed to assess its distribution, revealing pronounced right-skewness, a common trait in monetary data where a few high premiums skew the distribution. This finding underscored the need for a logarithmic transformation to stabilize variance and improve model fit. Binary variables such as AnyTransplants were analyzed for prevalence, showing a low incidence rate of 5–6%, which hinted at their potential significance as rare but impactful risk factors. These comprehensive insights shaped the modeling strategy, prioritizing non-linear terms, data transformations, and a focus on key predictors to maximize predictive power. Additional exploratory efforts included box plots to identify potential outliers and pairwise scatter plots to explore interactions, though these were deferred to the modeling phase for formal handling.

**Model Building**

The modeling process was meticulously designed to tackle the dataset’s inherent complexities through a multi-step approach. The data was partitioned into training (70%, 690 observations) and testing (30%, 296 observations) sets using the train\_test\_split function from scikit-learn, with a random\_state=42 to ensure reproducibility and maintain a representative sample across both subsets. This split was validated by comparing means and proportions of key variables (e.g., Diabetes at 58% vs. 58%, AnyTransplants at 94.6% vs. 93.9%), confirming balanced representation and minimizing selection bias.

A logarithmic transformation was applied to PremiumPrice to address the observed right-skewness, a technique that not only normalizes the distribution but also mitigates potential heteroskedasticity in the residuals of initial ordinary least squares (OLS) models. The initial OLS regression identified a core set of significant predictors, but diagnostic checks revealed multicollinearity between Age and its quadratic term (Age²), with variance inflation factors exceeding acceptable thresholds (e.g., >10). To resolve this, Age was centered by subtracting its mean value across the dataset (Age\_centered), and the squared term was computed as Age2\_centered. This centering technique, a widely recognized method in regression analysis, effectively reduced multicollinearity, with post-centering VIF values dropping to 1.01–1.03, enhancing the model’s numerical stability and reliability of coefficient estimates.

Outlier management was a pivotal aspect of the modeling process, with 30 observations identified based on residuals exceeding 2 standard deviations from the mean. These outliers, potentially representing extreme premium cases or data entry errors, were capped to their predicted values to minimize their disproportionate influence on the model’s fit. This decision was informed by a detailed residual analysis, which highlighted heavy tails in the distribution, a common issue in insurance data where high-cost claims can skew results. To further tackle heteroskedasticity, confirmed by the Breusch-Pagan test in prior iterations (p = 0.0000), weighted least squares (WLS) was implemented. Weights were derived from the inverse of the predicted values from an initial OLS model, a strategy designed to adjust for varying error variances across the range of PremiumPrice. This approach, rooted in advanced regression techniques, aimed to enhance model robustness by giving less weight to observations with larger predicted values, where variance was observed to increase. The final WLS model underwent rigorous validation on the test set, with predictions transformed back to the original scale using exponentiation to facilitate practical interpretation and assessment.

**Validation**

The validation phase was a critical step to ensure the model’s real-world applicability, involving the conversion of log-predicted PremiumPrice values back to their original monetary scale through exponentiation. This transformation was essential to align the model’s outputs with the currency-based expectations of insurance stakeholders, providing a direct measure of predictive performance in financial terms. The process included generating sample predictions and comparing them against actual values from the test set to evaluate individual accuracy, offering a qualitative perspective on the model’s effectiveness. Additionally, the Mean Squared Error (MSE) was computed on the test set to offer a quantitative assessment of the model’s predictive power across all observations. This dual-validation approach, combining sample-level insights with aggregate metrics, ensured a comprehensive evaluation, aligning with best practices for model assessment in business analytics and providing a bridge between statistical modeling and practical application.

**Results**

**Model Performance**

The weighted least squares model exhibited a robust fit, achieving an R-squared value of 0.696, signifying that 69.6% of the variance in log(PremiumPrice) is accounted for by the selected predictors. The adjusted R-squared, calculated at 0.693, adjusts for the inclusion of six predictors in the model, providing a more conservative estimate of explanatory power that accounts for degrees of freedom. This adjustment confirms the model’s efficiency and its ability to generalize beyond the training data. The F-statistic, recorded at 260.8 with an exceedingly small p-value of 6.06e-173, offers compelling evidence of the model’s overall statistical significance, decisively rejecting the null hypothesis that all coefficients are zero. These performance metrics collectively affirm that the model surpasses the project’s initial target of explaining at least 60% of the variance, establishing a solid foundation for predictive accuracy and practical utility.

**Comparison of Model Iterations**

The development of the final model involved four iterations, each building on the previous to address identified issues and improve performance. A comparative analysis of these iterations provides insight into the evolution of the modeling process:

* **Initial Model**: The baseline OLS model achieved an R-squared of 0.696 but included insignificant predictors (Diabetes p = 0.378, BloodPressureProblems p = 0.626, NumberOfMajorSurgeries p = 0.936), with a high condition number (2.44e+04) indicating numerical instability due to multicollinearity between Age and Age². The Mean Squared Error (MSE) on the training set was 0.02, but no test set validation was performed, limiting its generalizability.
* **Iteration 1**: This iteration introduced centering of Age to mitigate multicollinearity, but a coding error included both Age and Age\_centered, resulting in an infinite VIF and a condition number of 7.41e+17. The R-squared remained 0.696, with all predictors significant except due to the error, and the training MSE stayed at 0.02. Test set validation was added (MSE 0.02), but the model was unstable.
* **Iteration 2**: Correcting the error, this iteration used only Age\_centered and Age2\_centered, maintaining an R-squared of 0.696 and a condition number of 1.34e+03. All predictors were significant, with VIF values reduced to 1.00–1.03, and the training/test MSE remained 0.02. The identification of 30 outliers marked a step toward robustness, though heteroskedasticity was not yet addressed.
* **Iteration 3**: Outlier capping was implemented, reducing training MSE to 0.01 while retaining an R-squared of 0.696 and a condition number of 1.34e+03. The Breusch-Pagan test confirmed heteroskedasticity (p = 0.0000), prompting the next iteration. Test MSE stayed at 0.02, indicating consistent generalizability.
* **Iteration 4 (Current)**: The final WLS model, with inverse predicted value weights, maintained an R-squared of 0.696, a condition number of 1.35e+03, and low VIF (1.00–1.03). Training MSE improved to 0.01 with capped outliers, and test MSE remained 0.02. However, the Breusch-Pagan test (p = 0.0000) indicated persistent heteroskedasticity, and non-normal residuals (Skew 1.468, Kurtosis 11.686) persisted, though mitigated by outlier handling.

This progression highlights a systematic refinement, addressing multicollinearity, outliers, and heteroskedasticity, though the latter remains a limitation. The current model balances predictive power with practical applicability, informed by iterative improvements.

**Coefficients**

The model’s predictors and their estimated coefficients are detailed in the following table, providing a comprehensive view of their influence on PremiumPrice:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Predictor | Coefficient | Std Err | t-statistic | p-value | Interpretation |
| const | 9.9618 | 0.028 | 357.861 | 0.000 | Establishes the baseline log(premium) when all predictors are zero |
| BMI | 0.0051 | 0.001 | 5.349 | 0.000 | Indicates a 0.51% increase in premium for each unit increase in BMI |
| AnyTransplants | 0.2188 | 0.025 | 8.718 | 0.000 | Suggests a 24.5% premium increase for individuals with a transplant |
| AnyChronicDiseases | 0.0914 | 0.015 | 6.169 | 0.000 | Implies a 9.6% premium increase with the presence of chronic diseases |
| HistoryOfCancerInFamily | 0.0532 | 0.018 | 2.939 | 0.003 | Indicates a 5.5% premium increase with a family history of cancer |
| Age\_centered | 0.0143 | 0.000 | 35.430 | 0.000 | Reflects a 1.44% premium increase per centered age year |
| Age2\_centered | -0.0004 | 3.12e-05 | -12.799 | 0.000 | Captures a negative quadratic effect, reducing the premium growth rate with age |

These coefficients are derived from a log-linear model, where each value approximates the percentage change in PremiumPrice for a one-unit change in the predictor, holding other variables constant. For example, the coefficient for AnyTransplants (0.2188) translates to a 24.5% increase, calculated as 100×(e0.2188−1) 100 \times (e^{0.2188} - 1) 100×(e0.2188−1), highlighting the substantial impact of transplant status on insurance costs. The quadratic term Age2\_centered introduces a diminishing rate of increase, suggesting that the risk premium associated with age peaks and then levels off, a pattern consistent with actuarial risk profiles that account for health deterioration followed by mortality effects.

**Validation Results**

The model’s predictive performance was rigorously validated by reverting log-predicted PremiumPrice values to their original monetary scale through exponentiation, a critical step to align the model’s outputs with the practical expectations of insurance stakeholders. This transformation process was applied to the test set, comprising 296 observations, to assess real-world applicability. Sample predictions were generated and juxtaposed against actual values to evaluate individual accuracy, offering a qualitative perspective on the model’s effectiveness. The following comparisons illustrate this validation:

* **Sample 1**: Predicted = 28,723.97, Actual = 31,000.00 (Absolute Error: 2,276.03, Percentage Error: ~7.3%)
* **Sample 2**: Predicted = 26,256.16, Actual = 31,000.00 (Absolute Error: 4,743.84, Percentage Error: ~15.3%)
* **Sample 3**: Predicted = 28,067.12, Actual = 29,000.00 (Absolute Error: 932.88, Percentage Error: ~3.2%)
* **Sample 4**: Predicted = 19,588.54, Actual = 15,000.00 (Absolute Error: 4,588.54, Percentage Error: ~30.6%)
* **Sample 5**: Predicted = 26,489.49, Actual = 23,000.00 (Absolute Error: 3,489.49, Percentage Error: ~15.2%)

These samples demonstrate a range of predictive accuracy, with percentage errors varying from a minimal 3.2% to a substantial 30.6%. The largest discrepancy, observed in Sample 4, may be attributed to an outlier or a case where non-linear effects or unmodeled interactions are not fully accounted for, such as extreme low premiums. The Mean Squared Error on the testing set, computed on the original scale, is 12,531,667.25, corresponding to a root mean squared error (RMSE) of approximately 3,539.67. This RMSE, representing 12–23% of the typical premium range (15,000–40,000), indicates that the average prediction error is significant in monetary terms, reflecting the challenges of exponentiating log-transformed predictions. This high MSE underscores the impact of log transformation bias, a well-documented issue in statistical modeling where the reversal to the original scale amplifies residual variances, a factor that could be explored further with bias correction techniques.

**Diagnostics**

* **Multicollinearity**: The Variance Inflation Factors (VIF) for the predictors range from 1.00 to 1.03, with the exception of the constant term at 24.60, which is a typical outcome when using centered variables and does not indicate a problem among the predictors themselves. This low VIF range confirms that multicollinearity has been effectively mitigated through the centering of Age, ensuring that the model’s estimates are not unduly influenced by intercorrelations among the independent variables.
* **Condition Number**: Recorded at 1.35e+03, the condition number remains elevated, suggesting potential numerical instability. This value is likely influenced by the inclusion of the quadratic term Age2\_centered or the inherent scale of the dataset, though it has remained stable across iterations, indicating consistency rather than a new issue.
* **Heteroskedasticity**: The Breusch-Pagan test provided an LM Statistic of 48.80 with a p-value of 0.0000, offering conclusive evidence of significant heteroskedasticity. This statistical outcome aligns with the visual observation of a fan-shaped pattern in the residual plot from prior iterations, where the spread of residuals increases with the magnitude of predicted values, indicating that the variance of errors is not constant across the range of PremiumPrice.
* **Residuals**: The Omnibus test statistic of 266.153, with a p-value of 0.000, alongside the Jarque-Bera test statistic of 2416.695 and a p-value of 0.00, highlights significant deviations from normality in the residuals. The skewness measure of 1.468 and kurtosis of 11.686 further suggest the presence of heavy tails, likely due to remaining outliers or unmodeled non-linear effects, even after the capping of 30 outliers based on a 2 standard deviation threshold.

**Discussion**

**Interpretation**

The developed model provides a wealth of actionable insights for the insurance industry, offering a detailed lens through which to view the determinants of medical insurance premiums. The most striking finding is the substantial coefficient for AnyTransplants at 0.2188, which translates to an approximate 24.5% increase in PremiumPrice for individuals with organ transplants. This significant uplift reflects the elevated medical risk and associated healthcare costs tied to transplant patients, a critical consideration for risk assessment and premium setting. The Age\_centered coefficient of 0.0143, coupled with the negative quadratic term Age2\_centered at -0.0004, delineates a nuanced age-related effect. Premiums increase with age at a rate of about 1.44% per centered year, but this growth diminishes as age advances, suggesting a risk profile that peaks in later years before stabilizing, possibly due to mortality effects or reduced healthcare utilization among the eldest.

The BMI coefficient of 0.0051 indicates a modest but statistically significant 0.51% premium increase per unit of BMI, reflecting a gradual impact of body mass on health-related costs, consistent with its role as a general health indicator. Similarly, AnyChronicDiseases with a coefficient of 0.0914 suggests a 9.6% premium hike, underscoring the financial burden imposed by ongoing medical conditions. The weaker but still significant effect of HistoryOfCancerInFamily at 0.0532, equating to a 5.5% increase, points to a genetic predisposition that influences pricing, albeit to a lesser extent. These findings collectively align with actuarial principles, where risk factors are weighted based on their likelihood and cost impact, providing a data-driven basis for premium adjustments.

The model’s R-squared of 0.696 exceeds the project’s initial target of 60%, demonstrating substantial explanatory power and affirming its utility in capturing the majority of variability in PremiumPrice. The consistent test set MSE of 0.02 on the log scale, coupled with the original-scale RMSE of 3,539.67, provides a practical measure of predictive accuracy. This RMSE, representing 12–23% of the premium range (15,000–40,000), suggests that the model offers reasonable estimates for initial pricing, though the variability in sample prediction errors (3.2% to 30.6%) indicates room for refinement in specific cases, particularly at the lower end of the premium spectrum, such as the 30.6% error observed in Sample 4.

**Limitations**

Despite its robust performance, the model is not without limitations that warrant careful consideration. The Breusch-Pagan test result (p = 0.0000) provides unequivocal evidence of significant heteroskedasticity, where the variance of residuals increases with the predicted values of PremiumPrice. This is visually corroborated by the fan-shaped residual plot observed in earlier iterations, suggesting that the WLS approach with inverse predicted value weights did not fully stabilize the error variance across the range of premiums. The persistence of this issue may stem from unmodeled interactions between predictors—such as between Age and AnyTransplants—or structural characteristics of the data that were not fully addressed by the current weighting strategy, indicating a potential need for more sophisticated variance modeling.

The residuals also exhibit non-normality, as indicated by the Omnibus test (266.153, p = 0.000) and Jarque-Bera test (2416.695, p = 0.00), with a skewness of 1.468 and kurtosis of 11.686. These statistics reveal heavy tails, likely influenced by remaining outliers or unaccounted non-linear effects, despite the capping of 30 observations with residuals exceeding 2 standard deviations. The high original-scale MSE of 12,531,667.25 further highlights a challenge inherent to log-linear models: the bias introduced during exponentiation. This bias amplifies residual variances when reverting to the original monetary scale, a well-documented phenomenon that could be mitigated with a correction factor, such as adding the mean residual to predictions, though this was not implemented in the current iteration. The variability in prediction errors, particularly the 30.6% error in Sample 4, may also suggest that the model struggles with extreme values, potentially due to unmodeled interactions or data-specific anomalies.

**Future Work**

To address these limitations and enhance the model’s utility, several avenues for future work can be pursued. One immediate improvement would be the application of a bias correction to the log-predicted values, such as adding the mean residual—calculated as the average difference between exponentiated training predictions and actual values—before exponentiation. This adjustment could reduce the original-scale MSE and improve prediction accuracy, particularly for cases with large discrepancies like Sample 4. Another promising direction is the exploration of robust regression techniques, such as Huber regression, which are designed to handle heteroskedasticity and outliers more effectively by assigning lower weights to extreme observations, potentially stabilizing the residual variance across the premium range.

Additionally, incorporating interaction terms—such as the product of Age\_centered and AnyTransplants—could capture synergistic effects between predictors, addressing unmodeled non-linearities that may contribute to the observed heteroskedasticity. Further validation through k-fold cross-validation, utilizing 5 or 10 folds, would provide a more robust assessment of the model’s stability across different data partitions, mitigating the risk of overfitting observed in the current test set performance and offering a more comprehensive view of its generalizability. These enhancements would not only refine the model’s predictive precision but also strengthen its statistical inference capabilities, making it more suitable for applications requiring rigorous hypothesis testing or regulatory compliance in the insurance sector, where precise pricing is paramount.

**Conclusion**

This project has successfully engineered a predictive model for medical insurance premiums, attaining a commendable 69.6% explained variance in log(PremiumPrice) through the application of weighted least squares methodology. The model effectively pinpoints Age, AnyTransplants, and AnyChronicDiseases as pivotal predictors, furnishing actionable insights that resonate with the risk-based pricing paradigms of the insurance industry. Validation on the test set, yielding an RMSE of 3,539.67 on the original scale, underscores the model’s practical applicability, offering a reliable foundation for initial premium estimates despite the observed variability in sample predictions (3.2% to 30.6% error). The methodology employed mirrors the analytical rigor expected in business analytics, encompassing data preparation, exploratory analysis, model building, and validation, thereby fulfilling the course’s core objectives.

The model’s ability to leverage data-driven techniques to inform pricing strategies highlights its relevance to real-world applications, providing a valuable tool for insurance providers to optimize revenue and customer satisfaction. The identification of key risk factors, such as the 24.5% premium increase associated with transplants, offers a clear pathway for targeted pricing adjustments. However, the persistent heteroskedasticity and residual non-normality suggest avenues for refinement, which future work can address to enhance both predictive accuracy and statistical robustness. This project lays a solid groundwork for further development, with potential implications for improving financial and operational efficiency in the insurance domain, and serves as a testament to the power of analytical methods in solving complex business problems.