Calculating Healthcare Premiums Using Statistical Computing

By Noah Glosson

Introduction

Insurance premiums remain a crucial component of actuarial science in the medical industry. Insurance premiums have skyrocketed substantially in recent years, as nearly every household in the country needs medical insurance. Many health insurance companies rely on accurate models and algorithms to provide accessible coverage to consumers while maintaining financial stability. Predicting insurance premiums for each consumer is challenging, as companies must assess a range of factors, including demographics, health conditions, medical history, family size, lifestyle choices, and economic status. With accurate predictive models, insurers can predict risk, optimize pricing strategies, and provide accessible coverage across their consumer base.

In this project, I utilize a dataset from a recently concluded Kaggle competition, specifically the Regression with an Insurance Dataset. This data was created from a training model on another dataset and has very similar features, although it is not identical. Additionally, the data is divided into training and testing datasets (Kaggle). The objective of this analysis is to explore which variables impact the price of insurance premiums on the training set, then use the data to generalize the model before testing its efficacy on the testing set. We plan to utilize linear regression to start and operate under the assumptions of normal distribution. If those assumptions are violated, then non-parametric models such as K Nearest Neighbors (KNN) regression would be necessary in predicting insurance premiums compared to simple linear regression. By utilizing different methods, I plan to examine which factors contribute the most to premium prices and how to create algorithms that can predict insurance premiums for each individual.

Data Overview and Exploratory Data Analysis (EDA)

The initial training set contained 1,200,000 entries, with 800,000 in the testing set. The training set contained ID, Age, Gender, Annual Income, Marital Status, Number of Dependents, Education Level, Occupation, Health Score, Location, Policy Type, Previous Claims, Vehicle Age, Credit Score, Insurance Duration, Policy Start Date, Customer Feedback, Smoking Status, Exercise Frequency, Property Type, and Premium Amount. Premium Amount was the response variable while the rest were predictors, excluding ID. When looking at the testing set, it lacked the response variable of Premium. Amount. Given the nature of the Kaggle Competition this dataset originated from, we couldn’t use the testing data. Instead, the training set will be split up into a training and a testing set. After using na.omit(), we were left with 384,004 rows of data to work with. Following a 70-30 split, the training set contained 268,802 entries while the testing set had 115,202. Then, all the categorical predictors were converted to factors for computational purposes.

We then proceeded to plot histograms of various predictors to see the ratio of the various categorical predictors that consist of marital status, location, smoking status, and gender. When looking at the categorical predictors such as gender, location, and marital status, we notice a fairly uniform distribution across the board. There were no major differences among any of the categories within each predictor. In fact, smoking status only had 22 more smokers in the data compared to non-smokers.

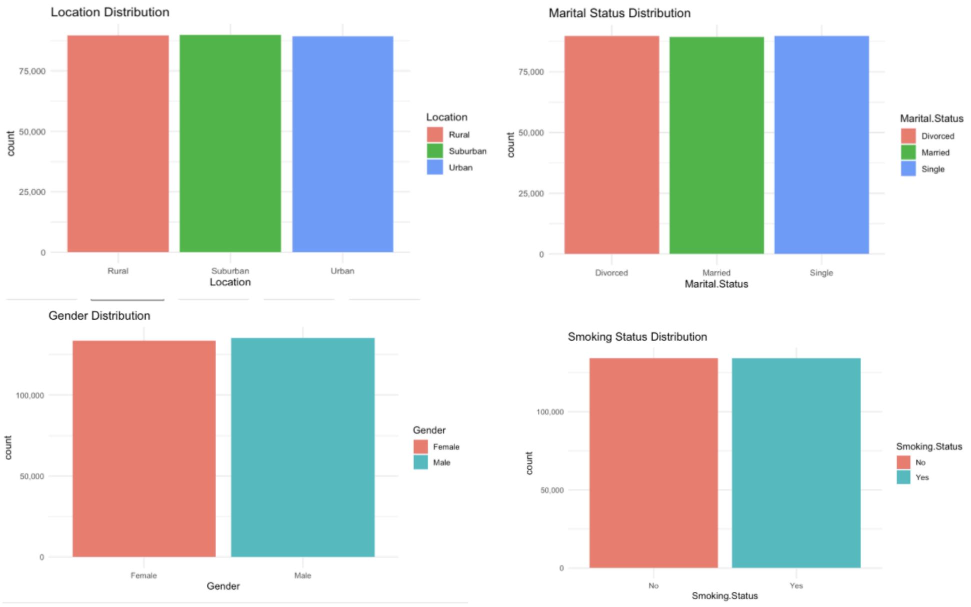

Figure 1: Distribution of Location, Marital Status, Gender, and Smoking Status

When looking at the numerical predictors such as age, credit score, and income, the numbers tell a different story. Age and Credit Score don't follow any symmetrical pattern, however, annual income is right-skewed, with most of the people earning less than 50,000 dollars. In addition, the response variable, Premium Amount, also follows a similar pattern of right skewness (Figure 2).

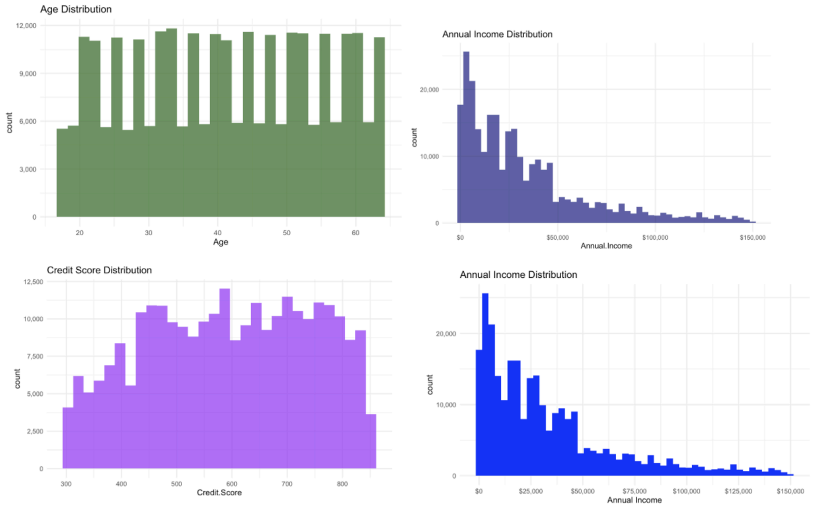

Figure 2: Distribution of Premium Prices and Annual Income

When looking at the summary statistics of the numerical predictors, including mean and median, it confirms that Premium Prices and Annual Income are right-skewed. For Age and Credit Score, it appears to be more uniform overall. When looking at the summary statistics, we can see that the median age in this data set is 41, and everyone is between the ages of 18 and 64, below retirement age. About 75 percent of the data includes individuals who make less than $45,000 a year, much lower than the national average. With regard to credit score, we see a median and mean just under 600, which indicates poor credit on average. In addition, we see some outliers like annual salary being $2 and only having to pay $20 for a premium.

	Min	1st Quarter	Median	Mean	3rd Quarter	Max
Age	18	30	41	41.13	53	64
Annual Income	$2	$8062	$24140	$33042	$44792	$149996
Credit Score	300	470	596	594.2	772	849
Premium Amount	20	520	883	1114	1522	4997

Figure 3: Summary Statistics for Numerical Predictors

After performing a correlation heat map on the numerical predictors, there was no significant correlation between any of the predictors. The highest correlation between predictors was -0.21 between credit score and annual income, giving us confidence that multicollinearity won’t be an issue.

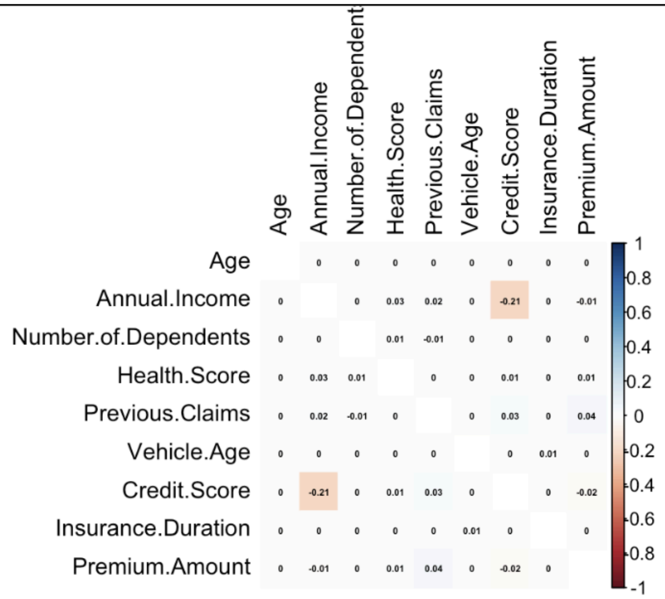

Figure 4: Correlation Heatmap between Predictors

Methodology and Results

Since the response variable is numeric, the first step was to get a preliminary examination of the data with all predictor variables using linear regression models. We found age, being male, annual income, health score, previous claims, credit score, and poor customer feedback were statistically significant predictors in determining premium prices. However, the R^2 was only 0.003, meaning that the model could explain just 0.3 percent of the variance in premium prices. In addition, the Residual Standard Error was $867.90, an incredibly high error when predicting important values such as healthcare premiums. The Akaike Information Criterion (AIC) was also 4,400,336, extraordinarily high for most models. To identify the best overall model, we employed forward stepwise linear regression to determine the optimal linear regression model.

Using the statistical significance of each predictor, we found the best model to have significant predictors of Previous Claims, Credit Score, the Male Gender, an interaction between Annual Income and Previous Claims, an interaction between Previous Claims and Credit Score, an interaction between Credit Score and annual income, a previous claim between annual income and health score, an interaction between credit score and health score, and male gender with poor customer feedback interaction variable. The R^2 value is just 0.009364, with a 0.9% variability of Premium Prices explained by the model. Ultimately, a linear model lacks predictive capabilities on its own. In Figure 5, we have the entire equation below.

$$Premium.Amount = 1316.0 - 138.4 \cdot Previous.Claims - 2.891 \cdot Credit.Score$$

$$+ 0.000106 \cdot Annual.Income - 0.1000 \cdot Health.Score - 8.358 \cdot Age + 19.85 \cdot Gender_{Male}$$

$$- 22.47 \cdot Customer.Feedback_{Good} + 21.71 \cdot Customer.Feedback_{Poor}$$

$$+ 0.1886 \cdot (Previous.Claims \times Annual.Income) + 2.029 \cdot (Previous.Claims \times Credit.Score)$$

$$- 5.987 \times 10^{-6} \cdot (Credit.Score \times Annual.Income) + 2.892 \times 10^{-5} \cdot (Annual.Income \times Health.Score$$

$$- 4.320 \cdot (Previous.Claims \times Health.Score) + 2.447 \cdot (Credit.Score \times Health.Score)$$

$$+ 1.237 \times 10^{-4} \cdot (Credit.Score \times Age) - 12.36 \cdot (GenderMale \times Customer.FeedbackGood)$$

$$- 25.09 \cdot (GenderMale \times Customer.FeedbackPoor) + 0.03145 \cdot (Credit.Score \times Customer.FeedbackGood)$$

$$- 1.535 \cdot (Credit.Score \times Customer.FeedbackPoor) - 5.000 \times 10^{-6} \cdot (Annual.Income \times Age)$$

$$+ 5.009 \cdot (Customer.FeedbackGood \times Customer.FeedbackGood) +$$

$$8.171 \cdot (Previous.Claims \times Customer.FeedbackPoor) + \varepsilon$$

Figure 5: Equation of Best Model Through Stepwise Selection and Statistics

Upon examining the model diagnostics (Figure 4), we can further conclude that the assumptions of normality are violated, and linear regression is insufficient for effectively predicting insurance premiums. With the residual plot, there is no consistent variance throughout the plot, indicating heteroskedasticity. About the QQ Plot, we see the tail end of the distribution deviates substantially, violating the assumption of normally distributed residuals. The Scale Location plot further indicates heteroskedasticity with increasing spread. With cook's distance, there are no extreme outliers that influence the graph, but there were some high leverage points. Overall, linear regression was insufficient.

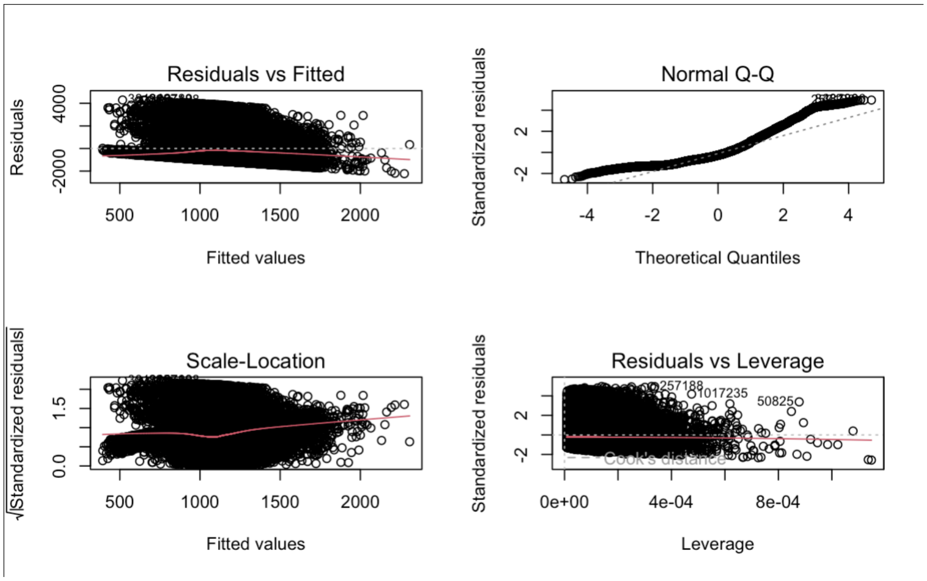

Figure 6: Linear Regression Model Diagnostics

Given what's happening with the linear regression model, we decided to try the Box-Cox transformation. The Box-Cox transformation is used to stabilize the residuals and make them more normally distributed to meet linear regression assumptions. Despite these efforts and a much lower residual standard error, the model still suffered from a low R^2 of 0.0176, indicating the transformation failed to make a significant improvement. As a result, we were forced to move to non-parametric methods to capture the complex relationships of premium pricing and various health-related factors.

With regards to non-parametric models, we approached the issue with K Nearest Neighbors (KNN) Regression. Unlike linear regression, KNN doesn't assume a linear relationship between the predictors and response variables (Yao, Z. and Ruzzo). Instead, it makes predictions based on the average premium amounts of the k most similar observations in the training data. For all KNN implementations in this analysis, we consistently used k = 10 as the number of neighbors, balancing local sensitivity with overall generalization. This approach allowed us to look at flexible models and non-linear relationships within the data without relying on parametric assumptions.

To simplify our analysis and reduce dimensionality, we looked at numerical predictors only. As shown in Figure 7 below, the KNN model produced a tight clustering and a linear relationship. This indicates strong performance in approximating premium amounts for a wide range of values and supports the flexibility of KNN in capturing complex data patterns. In addition, our Root Mean Squared Error (RMSE) was just \$142.92 and our R^2 result turned out to be 0.9728, meaning the model captured over 97% of the variability captured by the model.

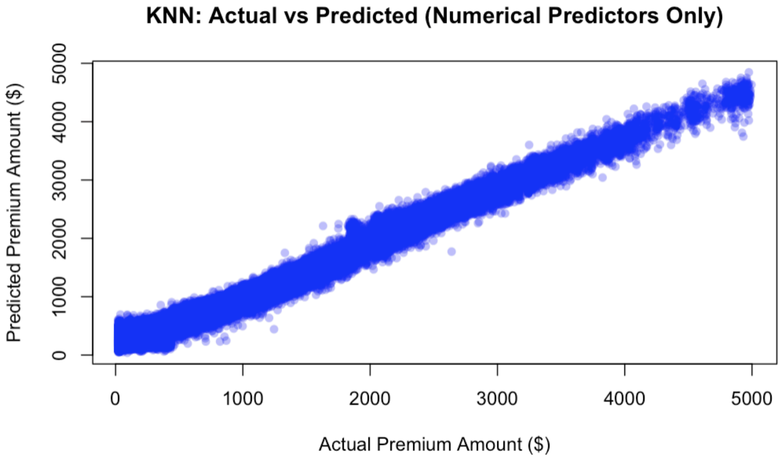

Figure 7: KNN Regression with Only Numerical Predictors

In our second KNN model, we expanded our input to include all predictors by encoding the categorical variables and scaling the design matrix. This approach will capture more complex interactions beyond numerical values. Despite these enhancements, the model performance crumbled, posting an RMSE of $904.33 and an R^2 of -0.0872, meaning our model performed worse than simply predicting the mean of the response for all observations. Figure 8 confirms this as the plot was very compressed across the y-axis and it failed to capture any variance in actual premium amounts.

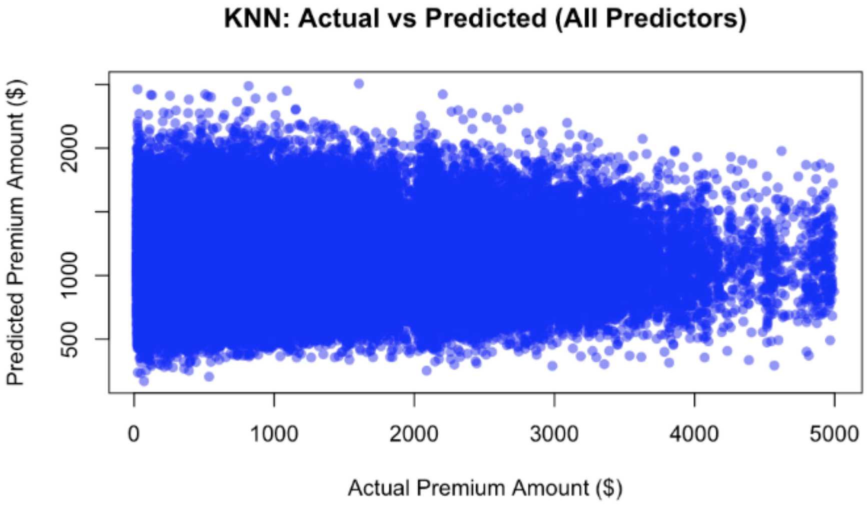

Figure 9: Predictions versus Actual Values using KNN

Since the response variable was right-skewed, we used a logarithmic transformation to minimize the skew and variance. The transformation compressed the data to reduce the influence of outliers and minimize variance. After training the model with k = 10 neighbors, we transformed the predictions back to the original scale. The final KNN model yielded a R^2 value of 0.641 and a residual standard error of $519.56. While these are far from perfect, it is a significant improvement over linear regression and box-cox transformations.

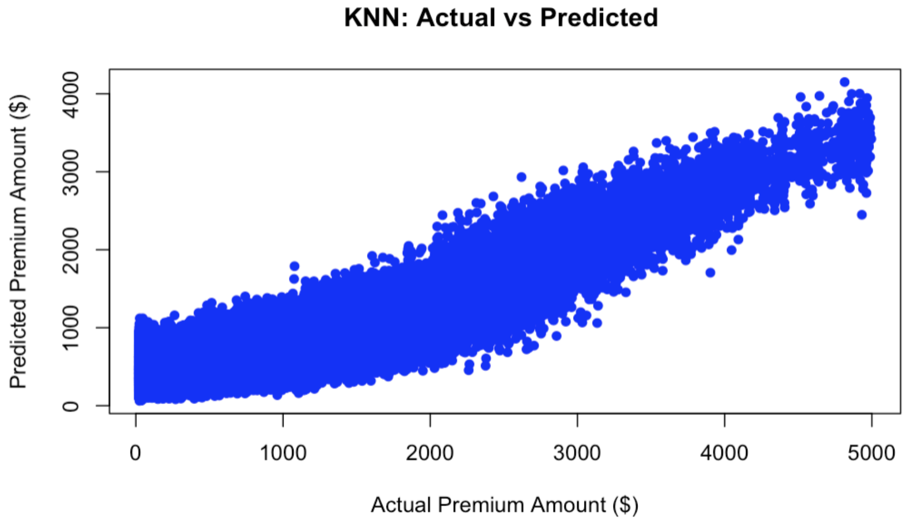

Figure 9: Log Method of Predictions versus Actual Values using KNN

The scaled plot in Figure 9 shows a nonlinear relationship between the predictors and the response variable. As explored in the EDA, this indicates a right-skewed distribution of response values. This is why there appears to be higher variance as premium prices go higher, hence why we see curving.. Since there is so much data for premiums under $2,000, KNN performs much better at understanding the relationships compared to those over $ 3,000. To check the model's performance, we implemented a 5-fold cross-validation procedure, where the model was trained on 80% of the data and tested on the remaining 20%.

Additionally, predictions were transformed back to the original scale for ease of interpretation. Our Root Mean Squared Error was $528.73, and our R-Squared was 0.63. This tells us the model explains 63% of our variation, and there is an average error of almost $530.

Overall, this model highlights the limitations of KNN in prediction, as well as the challenges of understanding a dataset with skewed distributions.

Discussion

Modeling this dataset highlights the challenges of predicting insurance premiums using statistical models. While linear regression is easy to interpret, its inability to capture complex, nonlinear relationships limits its predictive power when dealing with complex data. The abysmal R^2, high standard error, and violation of normality assumptions proved that linear regression would be insufficient in properly predicting insurance premiums. We then attempted to do a Box-Cox transformation, aiming to stabilize variance and normalize residuals. However, even with the transformation, the model still failed to capture the variance in premium prices adequately. This told us the true relationship was driven by nonparametric statistics.

By contrast, the K Nearest Neighbors (KNN) model, particularly with log-transformed premiums, offered a major improvement in predictive performance. The KNN model achieved an R^2 of 0.641 and a substantially lower residual standard error, indicating that it was able to better adapt to the skewness and nonlinear structure in the data. The residual histogram and prediction plots confirmed that KNN handled the bulk of moderate-to-high-premium observations well, although it struggled with extremely high premium values due to the sparsity in that range. Overall, the results underscore the importance of non-parametric algorithms when basic models such as linear regression fail to capture the complexities of real-world data.

Conclusion

This analysis highlights the critical role of predictive analytics in the health insurance industry. While linear regression is a staple of statistical analysis due to its simplicity and interpretability, we found it to be inadequate for modeling this complex, skewed healthcare dataset. Even after transformations, the linear model could only account for 1% of the overall variance. By contrast, the K-Nearest Neighbors Regression (KNN), especially when coupled with logarithmic transformation and feature scaling, demonstrated stronger predictive ability. With a R^2 of 0.641 and a cross-validated R^2 of 0.63, the KNN model was able to capture the meaningful non-linear relationships and drastically reduce prediction error. These results reinforce the value of non-parametric methods in the real world, where assumptions of linearity and normality do not hold. This project illustrated how statistical computing can enhance actuarial decision-making and pricing strategies in the insurance business.

References

1. https://www.kaggle.com/competitions/regression-with-an-insurance-dataset-bt-2-ds/data
2. Yao, Z., Ruzzo, W.L. A Regression-based K nearest neighbor algorithm for gene function prediction from heterogeneous data. BMC Bioinformatics 7 (Suppl 1), S11 (2006). https://doi.org/10.1186/1471-2105-7-S1-S11