Medical Insurance Premium Price Prediction

Erika, Fiona, Jason

Contents

Medical Premium: Data Set Information

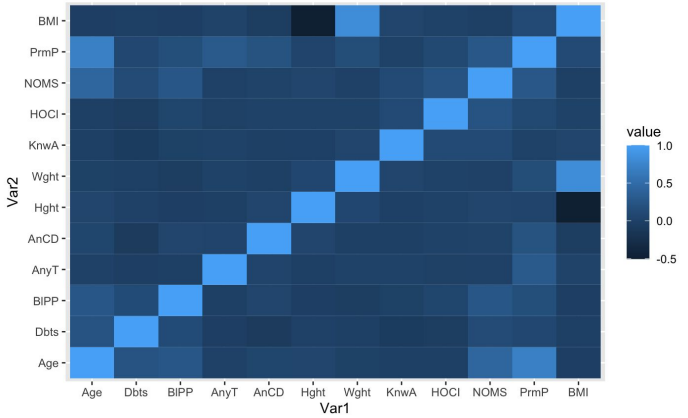
Data set information

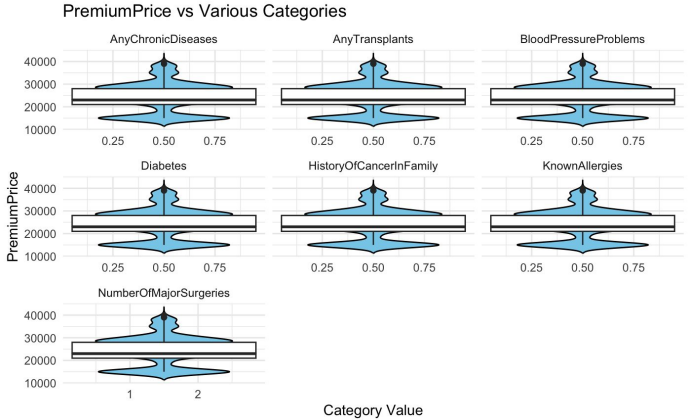
Variables: Age, BMI, Diabetes, Blood Pressure Problems, Organ Transplants, Any Chronic Diseases, Height, Weight, Known Allergies, Family History of Cancer, and Number of Major Surgeries

- As potential predictors for Insurance premiums
- 986 observations

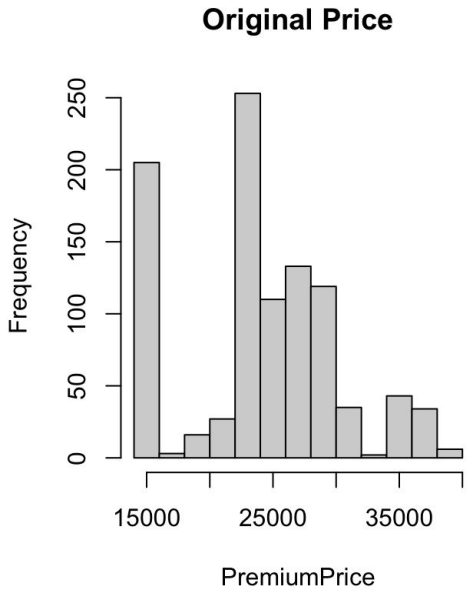
Data Preprocessing

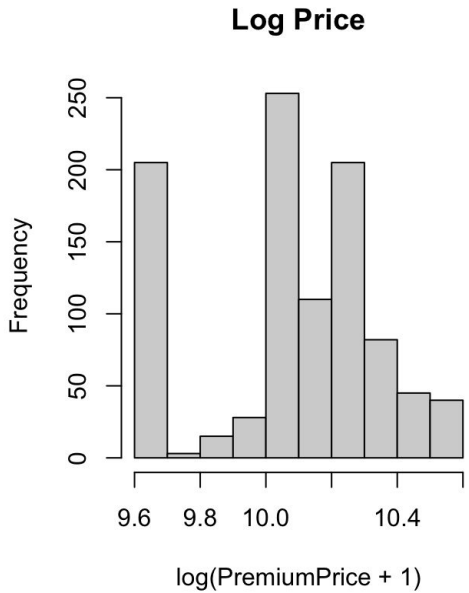
- T-tests of individual features to determine significance.
- Created histograms and boxplots to assess the distribution and potential outliers of each variable.
- Created a BMI variable using the existing height and weight features.
- Conducted ANOVA tests to quantify if differences were significant.





Log Transformation





Original Premium Price skewed to the right with a larger concentration to the lower end of the premium prices

Log transformation reduces skewness of the premium price and makes more symmetric, and therefore easier to analyze the relationships

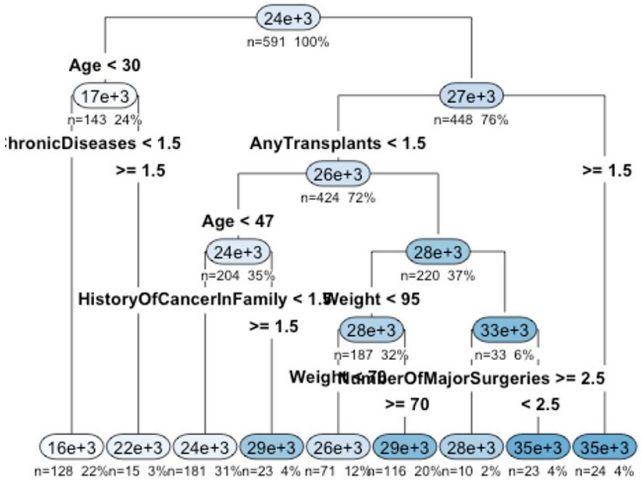
Regression Analysis

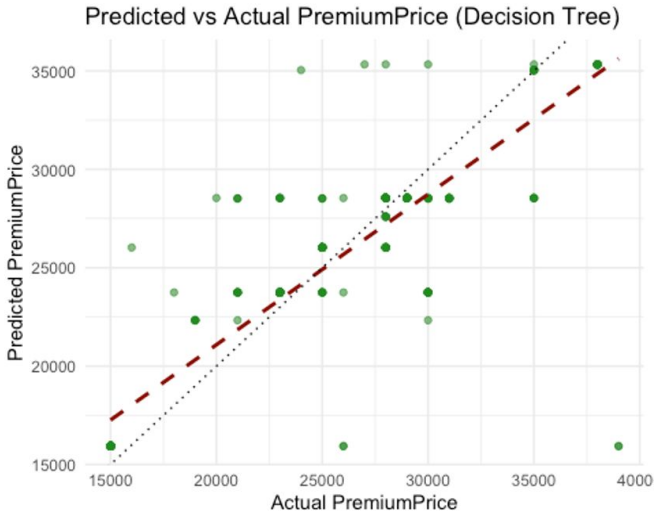
- To determine which features should be used in our models we used different types of automated selection.
- Both stepwise and forward selection determined the same variables to be significant which we then used in our models.
- The variables selected include age, history of transplants, presence chronic diseases, weight, number of major surgeries, and history of cancer.
- The lasso and ridge regression model both boasted and R-squared value of approximately 0.658.
- This indicates that these models can explain an estimated 65.8 percent of variation in medical premium prices.

```
#ridge regression model

ridge_model <- glmnet(X,y,alpha=0, lambda=min_mse)
coef(ridge_model)

## 7 x 1 sparse Matrix of class "dgCMatrix"
##                                       s0
## (Intercept)                9.220703029
## Age                        0.014923564
## AnyTransplants             0.261354117
## Weight                     0.002486957
## AnyChronicDiseases         0.123484357
## HistoryOfCancerInFamily    0.088574406
## NumberOfMajorSurgeries    -0.026603954
```

```
#final_model

final_model <- glmnet(X,y,alpha=1,lambda=min_mse)
coef(final_model)

## 7 x 1 sparse Matrix of class "dgCMatrix"
##                                       s0
## (Intercept)                9.224470905
## Age                        0.014888649
## AnyTransplants             0.259594630
## Weight                     0.002454203
## AnyChronicDiseases         0.122412203
## HistoryOfCancerInFamily    0.086528084
## NumberOfMajorSurgeries    -0.025491272
```

Decision Tree Outputs



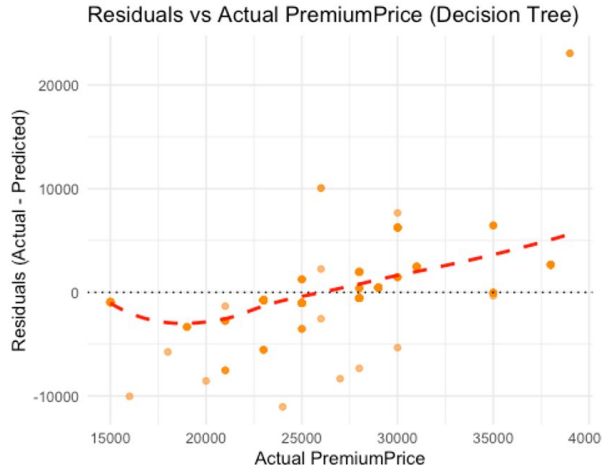


Predicted vs. Actual plot: Shows a general trend following the diagonal, but with some flattening. This indicates the tree predicts close to the average price for many cases (limited flexibility).

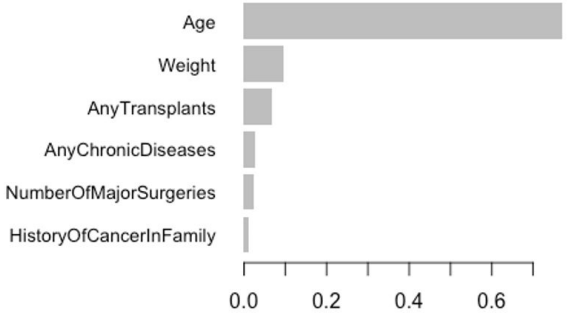
Residual Plot: Shows systematic patterns (underprediction of high prices, overprediction of low prices). The loess curve is curved, indicating bias.

Interpretation Strengths: Easy to follow and understandable for the average business owner or person investigating their premiums.

Weaknesses: Underfits complex relationships. Not optimal for explaining variations in detail.



XGBoost Outputs

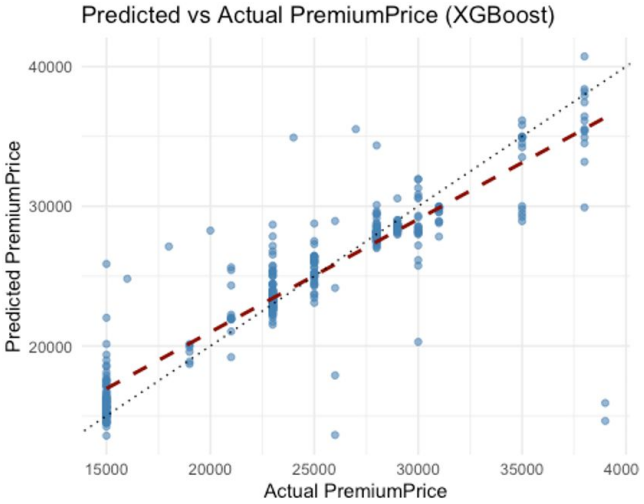


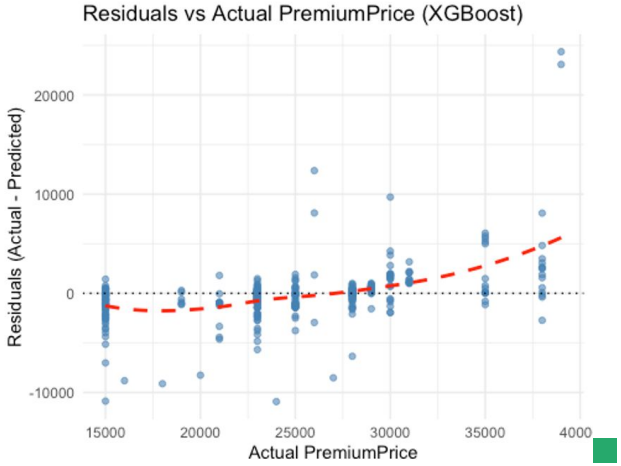
Predicted vs. Actual plot: Shows the predicted and actuals are much closer fit as shown in the line of best fit. Therefore, more accuracy across all premium price levels.

Residual Plot: Shows systematic patterns (underprediction of high prices, overprediction of low prices). The loess curve is curved, indicating bias. However, it is less bias than the decision tree.

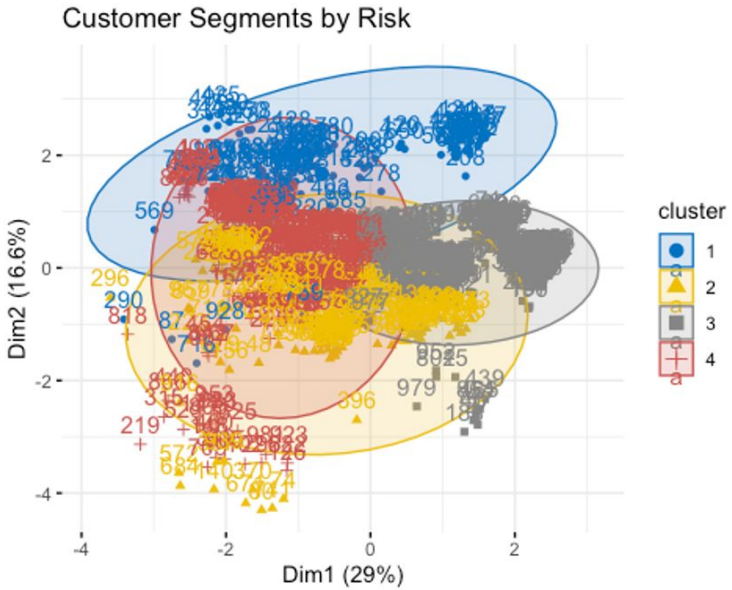
Interpretation Strengths: Easy to follow and understandable for the average business owner or person investigating their premiums.

Weaknesses: The regression model is less interpretable, therefore, the feature importance model is used to show which feature is best to predict the premium price.





K-Means Clustering



Cluster 1: Low Risk | 116 observations | Likely younger, healthier, fewer surgeries

Cluster 2: Moderate Risk | 156 observations | Mild health conditions or aging effects

Cluster 3: High Risk | 421 observations | Largest group — possibly middle-aged with issues

Cluster 4: Very High Risk | 293 observations | Older or multiple chronic conditions

The Elbow Method determined k = 4 clusters

Conclusion

Results: We find a significant relationship between Age, Any Chronic Diseases, Weight, Family History of Cancer, and Number of Major Surgeries to be significant predictors of insurance premiums

Recommendations:

- Multi-tiered pricing strategies for medical premiums.
- Create insurance discount incentives for more health stable clients who participate in health or other wellness programs.
- Continue calculating relationship to ensure future predictions are data-driven and updated.
- Refine the predictive performance tools via feature interaction modeling, more ensemble models (e.g. Random Forest), and interpretability tools (e.g. SHAP).