Title: Exploratory Medical Insurance Prediction Practice

Lormel Bationo, Esteban Valladares

Abstract:

Medical insurance is an essential tool that facilitates access to quality healthcare and preventive services. Our aim in this study is to create different models using both supervised and unsupervised learning methods to create models that will help us understand the different factors that lead to high medical insurance cost and also the mechanisms, insurance companies use to come up with premium prices for different individuals. We used data gathered by an insurance company from 1000 customers. The dataset also contains different information such as age, medical conditions, history of family cancers, chronic diseases. We used best subset selection to identify relevant variables and used a linear regression model to fit the model and obtain predicted vs actual values to get an estimation of medical insurance premium.

1. Background and Research Objective:

Having medical insurance is associated with lower death rates, better health outcome and improved productivity. However, due to its costs, many people are still unable to afford insurance premiums. In the United States about 45% of the population report poor perceived value regarding insurance cost as their household usually pay too much for the quality of care they received. It is important to perform regression analysis on insurance premiums so that we may understand the factors that influence a high or low premium price. We wish to answer the question of how we can determine which factors influence the price the most and how we can use them to predict an individual's price based on health indicators.

2. Methods:

We used a dataset from an insurance company with 1000 customers and different variables. We started our analysis by creating a model to estimate insurance premium price. We used a mix of techniques under supervised learning to obtain comprehensive results and explore the underlying patterns in the dataset.

Under supervised learning, we started with best subset selection to identify the different variables that were significant for our model. After that we used the results from the best subset to fit the model using a linear regression. We obtained a summary and estimated insurance premium price with a plot of predicted vs actual premium price. We also created a regression tree, selecting some variables seen from the best subset selection to act as a visual indication of what factors influence a high or low premium price.

2.1 Best subset selection

Best subset selection is an alternative to the least squares method, which aims to perform variables selection. We fit a separate least squares regression for each possible combination of the p predictors. That is, we fit all p models that contain exactly one predictor, all $\frac{p(p-1)}{2}$ models. We then look at all the resulting models, with the goal of identifying the one that is the "best". In this case, we define "best" as the models with the highest r^2 value and lowest cp and BIC value. Using the programming language R, we can automatically perform this process and out the number of variables which give us the best models under each definition. From there, we use a helper function to find the model that is best given the pre-specified number of variables selected. in our case we use seven variables since that amount gave us the lowest cp criterion.

2.2 Linear Regression Model

Linear regression is a common statistical tool used to model the relationship between variables in a dataset. In the context of our project, we want to estimate and predict insurance premium price based on the explanatory variables. After fitting the model, we proceed to the model evaluation by estimating the residuals, MSE (means squared error) and RMSE (Root mean squared error).

2.3 Regression Tree

Tree-based methods involve stratifying or segmenting the predictor space into several simple regions. A regression tree will predict the mean response value of a response variable that is quantitative. Our response variable is the probability of having a high premium insurance price, and the chosen predictors are the same ones found in best subset selection. Our initial regression tree had 13 nodes (see Figure 3); we then pruned it so that there were only 5 nodes. We chose 5 because it results in the lowest cross validation rate, as shown in Figure C.

3. Results:

From our results of doing best subset selection (see Figure 6), we chose to select the model that resulted in the lowest cp criterion. The included predictors in this model were Age, Diabetes, Any Transplants, Chronic Diseases, Weight, History of Cancer in Family, and Number of Major Surgeries. We continued to consider these variables in our future methods.

The linear model results figure(x) highlighted the variables that were significant in estimating insurance premium. While variables like high age, blood pressure, having any transplant or chronic disease is associated with an increase in premium price, variables like diabetes heights and the number of major surgeries were associated with a decrease in insurance premiums. Our model was statistically significant, and the adjusted R indicates that roughly 64% of the variation in the model are explained by our model.

Lastly, creation of a regression tree proved to show insight on the predictive and interpretation of the predictor classes. For example, in Figure 3, we can see that the regression tree has many nodes and is difficult to interpret. This initial tree has age as the top split assigning observations(clients) that are greater than 45.5 years old to the right branch, and those less than to the left branch. Continuing this will result in the mean response value of the client paying a "high" premium, high being considered above the average premium price of ₹24,337. After pruning the regression tree by comparing test error rates, we are left with the tree seen in figure 5. This model gives a readable and interpretable regression tree where age, history of cancer and any chronic diseases are the predictor classes, again, showing the probability of the client having a high premium. Since the calculated MSE was very high, a value of 25,221, we cannot assume high accuracy with these predictions.

4. Discussion:

The models we used for our study unraveled a lot of the mechanisms behind insurance premiums calculations and helped us discover some of the things that positively or negatively impact the cost of insurance premiums. We made discoveries that were unexpected for example with our linear model, we found out that having diabetes increased the chances of having a lower insurance premium rate compared to not having the diseases. This could probably be explained by the government subsidizing diabetic treatments in the US for example. However, the study also helped us discover that the premium cost was impacted at 64% by people age, weight, medical history and depending on whether they had certain condition or had a history of cancer in the family. The best subset selection helped in identifying the significant variables and creating a new model. However, the new model did not perform better than the model with all the variables. The assumptions made when determining a high premium price may be reasonable to some but for a lower income family, could be a price way over their budget. In the future, we would like to create different brackets of wealth, so that there is a wider range of predicted prices the clients could expect. The use of both supervised and unsupervised learning methods offers a good balance.

However, the nature of our data made it harder for the linear models to capture anything significant. The error rates were very high, showing that the prediction accuracy was not optimal. I think that to make the study more complete we could for the same dataset variables with values assigned instead of dummy variables and including more variables that could including race and gender, and the geographical location.

5. Appendix:

```
Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)               5480.610   2095.294   2.616  0.00904 **
Age                        329.367      9.839  33.474  < 2e-16 ***
Diabetes                  -429.120    251.419  -1.707  0.08818 .
BloodPressureProblems      180.504    252.421   0.715  0.47472
AnyTransplants            7894.201    521.963  15.124  < 2e-16 ***
AnyChronicDiseases        2654.886    313.990   8.455  < 2e-16 ***
Height                      -5.822     11.919  -0.488  0.62535
Weight                      69.675      8.428   8.267 4.45e-16 ***
KnownAllergies             300.882    295.796   1.017  0.30931
HistoryOfCancerInFamily   2311.829    385.373   5.999 2.80e-09 ***
NumberOfMajorSurgeries    -654.186    186.103  -3.515  0.00046 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3753 on 975 degrees of freedom
Multiple R-squared:  0.6429,    Adjusted R-squared:  0.6392
F-statistic: 175.5 on 10 and 975 DF,  p-value: < 2.2e-16
```

Figure 1 linear regression model output

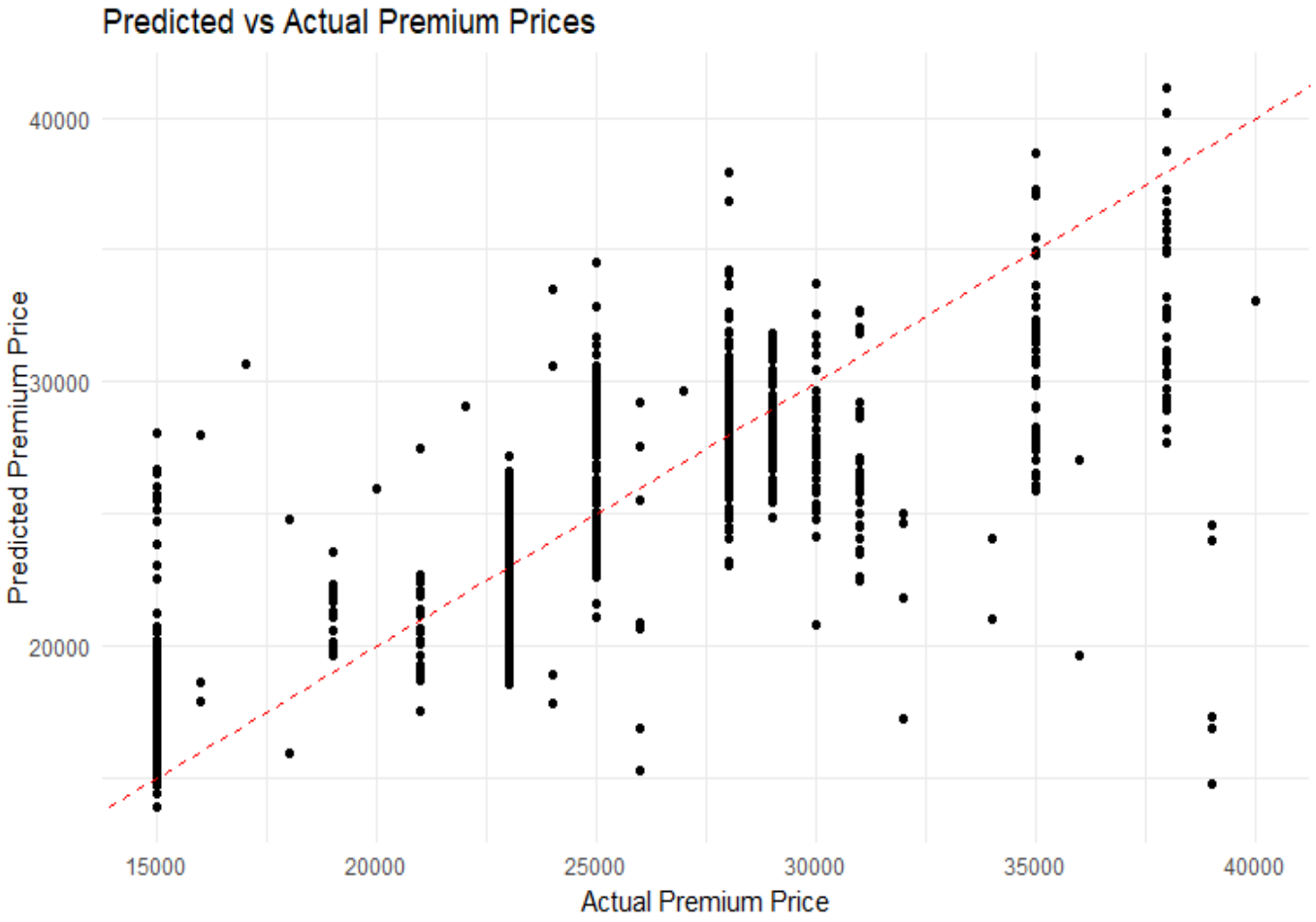

Figure 2 Plot of predicted vs actual Premium price

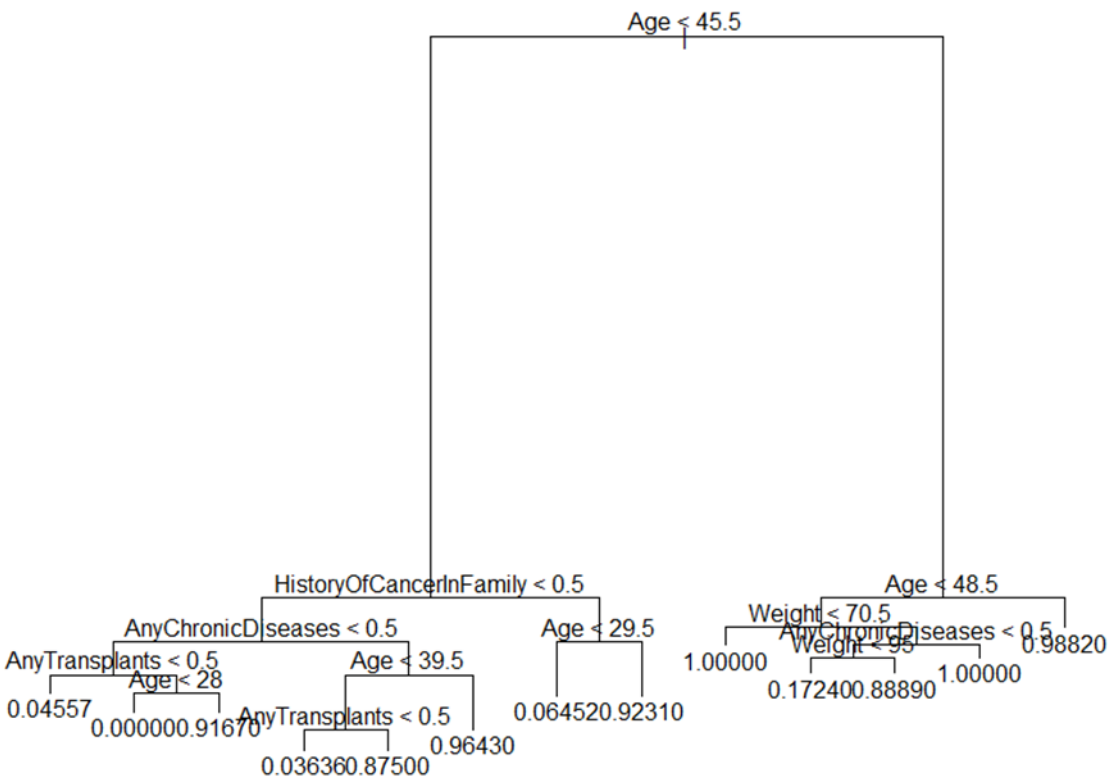

Figure 3 Regression Tree Unpruned

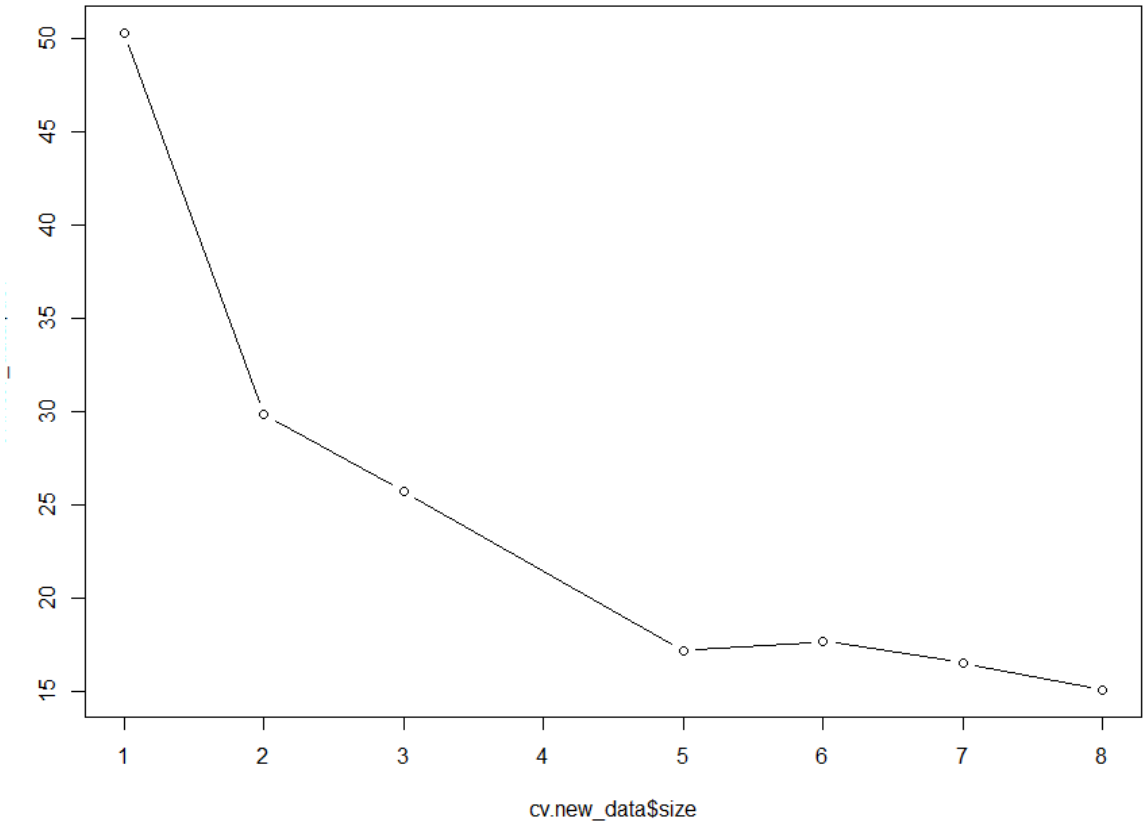

Figure 4 Terminal Node Vs Cross-Validation error rate

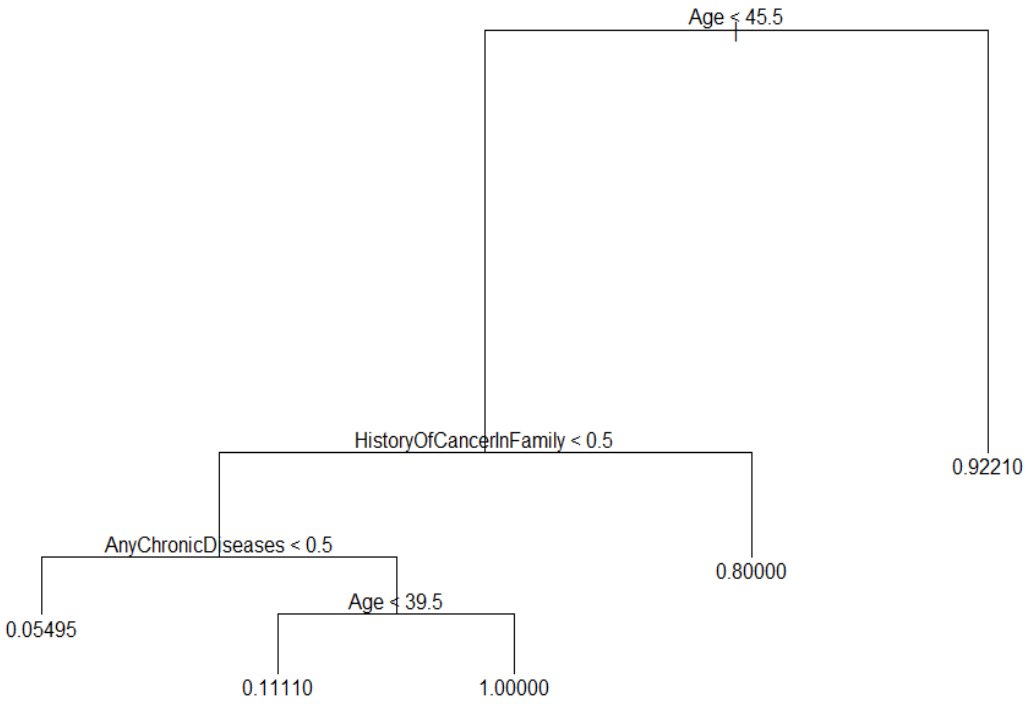

Figure 5 Regression Tree Pruned

```
Subset selection object
Call: regsubsets.formula(PremiumPrice ~ ., og_dataset, nvmax = 40)
10 variables  (and intercept)
                        Forced in Forced out
Age                         FALSE      FALSE
Diabetes                    FALSE      FALSE
BloodPressureProblems       FALSE      FALSE
AnyTransplants              FALSE      FALSE
AnyChronicDiseases          FALSE      FALSE
Height                      FALSE      FALSE
Weight                      FALSE      FALSE
KnownAllergies              FALSE      FALSE
HistoryOfCancerInFamily     FALSE      FALSE
NumberOfMajorSurgeries      FALSE      FALSE
1 subsets of each size up to 10
Selection Algorithm: exhaustive
          Age Diabetes BloodPressureProblems AnyTransplants AnyChronicDiseases Height Weight
1  ( 1 )  "*" " "      " "                   " "            " "               " "    " "
2  ( 1 )  "*" " "      " "                   "*"            " "               " "    " "
3  ( 1 )  "*" " "      " "                   "*"            "*"               " "    " "
4  ( 1 )  "*" " "      " "                   "*"            "*"               " "    "*"
5  ( 1 )  "*" " "      " "                   "*"            "*"               " "    "*"
6  ( 1 )  "*" " "      " "                   "*"            "*"               " "    "*"
7  ( 1 )  "*" "*"      " "                   "*"            "*"               " "    "*"
8  ( 1 )  "*" "*"      " "                   "*"            "*"               " "    "*"
9  ( 1 )  "*" "*"      "*"                   "*"            "*"               " "    "*"
10  ( 1 ) "*" "*"      "*"                   "*"            "*"               "*"    "*"
          KnownAllergies HistoryOfCancerInFamily NumberOfMajorSurgeries
1  ( 1 )  " "            " "                     " "
2  ( 1 )  " "            " "                     " "
3  ( 1 )  " "            " "                     " "
4  ( 1 )  " "            " "                     " "
5  ( 1 )  " "            "*"                     " "
6  ( 1 )  " "            "*"                     "*"
7  ( 1 )  " "            "*"                     "*"
8  ( 1 )  "*"            "*"                     "*"
9  ( 1 )  "*"            "*"                     "*"
10  ( 1 ) "*"            "*"                     "*"
```

Figure 6 Best Subset Selection R Output