Boosted NNs and Medical Costs

BAN 525

Khayrayyah Haamid-Day

Insurance companies utilize a series of their own calculations and statistics to determine how much to charge their customers, this is especially true when it comes to premiums. Certain factors may make a person’s medical cost and premiums far higher than another’s. For example, in the United States, there are health care laws in place to limit what factors health insurance companies are allowed to consider. These five factors are age, location, tobacco use, individual vs family enrollment, and plan category. Factors that are not allowed to be considered are gender, current health, and previous medical history. This analysis will construct a series of predictive models to better understand the potential determinants of medical costs billed by health insurance companies. In turn, this can also eventually be used to determine health insurance premiums.

This dataset contains 7 variables that are known to have an impact on medical costs as well as premiums. The response variable will be “Charges” while the predictor variables will be “Age”, “Sex”, “BMI”, “Children”, “Smoker”, and “Region”. Even though gender is technically not allowed to be considered by health insurance companies, it is still interesting to consider a relationship that may exist. This analysis will construct four statistical methods to investigate the best predictive model. OLS will be used as a baseline model and three different types of Boosted NN will also be used. The first method, OLS, works well as a benchmark model that provide easy interpretation and explanation. On the other hand, this method can be sensitive to outliers and create overfit models. The next three methods used will employ various versions of boosted Neural Networks. NN on their own are highly efficient models that employ continuous learning in order to produce well performing models. The boosted versions of NN adds another layer to its learning, with each step allowing the model to learn from previous errors and improve on the model. Additional weight is added to observations with the largest errors by use of a scaling factor. This process is repeated until the model can no longer be improved. Despite this, NN are known to be black boxes and, as such, they can be very difficult to interpret and understand. The first boosted NN will use one layer, three nodes in NTanH, and a maximum of 40 models. The second boosted NN will also use one layer, only 1 node in NTanH, and 40 models. The final boosted NN will use one layer, 3 nodes in NTanH, 40 models, and absolute penalty. The cross-validation method will use a validation column that splits the data, 60/20/20, into a training, validation, and test set. Using cross-validation in this analysis will allow the test data to be an unbiased look at each methods performance. Both the validation column and the boosted NN methods will use random seed 123.

Text, table

Description automatically generated

Text, table

Description automatically generated

Text, table

Description automatically generated

Text, table

Description automatically generated

At first glance, it looks like every method produced well performing models. There is not much variation between the training, validation, and test set. Though there is a fairly large difference in RASE for OLS compared to the other three methods. Another point to mention is the RSquare value difference between the training and validation set. For example, the training set is lower than the validation for the OLS method. While the other three NN methods have very little difference, essentially the same, between the RSquare values of the training and validation set. The model comparison will provide more insight into which model was the best but so far, it appears that the OLS model is not that good in comparison to the boosted NN models.

Table

Description automatically generated

The model comparison gives clear look at how well each model performed in the test set, as the best model will be chosen using the test set. First it is obvious that OLS is the weakest model, it has the lowest RSquare and the highest errors. The first boosted NN method has a high RSquare and lower errors while the second boosted NN method has a lower RSquare and higher errors. It appears that there is a difference in performance when building models with only one node versus three nodes, that is the model sees a decrease in performance. The final boosted NN method has a slightly higher RSquare than the first boosted NN, but it has the lowest errors out of all the methods. This model was built similarly to the first boosted NN model except that absolute penalty was used. The penalty is already in place to reduce the likelihood of an overfit NN model. Absolute penalty is useful for larger amounts of variables especially when a few of these variables contribute the most to the predictive performance of the model.

Graphical user interface, table

Description automatically generated

Chart, histogram

Description automatically generated

The variable importance profile indicates that the variable smoker has the most importance to charges, followed by BMI and age. This is unsurprising, when considering the long-term health impacts of smoking especially when someone has been smoking for a long period of time. A persons BMI also has a high impact on health, a high BMI can lead to a variety of diseases and long-term illnesses. The older a person is, the more illness they may become susceptible to, and an older age can lead to longer recovery times. The marginal model plots highlight these relationships further. Not smoking leads to far less charges, a rising BMI increasing charges as does age. The other three variables do not appear to have much of an impact on increase or decrease of charges.

Graphical user interface

Description automatically generated

A true test of this predictive model is looking at its ability to accurately predict a potential patient’s medical cost. Suppose there is a 45-year-old non-smoker male with a BMI of 38 who has two children and is from the southeast. There are factors that would lower his costs, like not being a smoker, and other that may raise it some, like being 45-year-old. This model would predict that his medical cost would be $10,358.99. Of course, there are other factors that could raise this price, such as what he is being treated for, cost of prescriptions, follow up care, etc. In this case, this predicted medical cost could indicate a potential minimum that this patient would be expected to pay. A health insurance company could also use a model like this to determine this potential client insurance premium. On average, 45 – 54-year-old men living in the United States could be paying up to $551 monthly for their health insurance though this average does not account for other health and family factors. The average overnight hospital stays in the US hovers around $11,700 and this may vary based on the type of insurance patients have, if they have any.

In conclusion, this final model shows similar charges to real world averages which further proves its predictive performance. Perhaps these charges would be drastically different if compared to countries that offer free, or far cheaper, health insurance and medical costs.