Daniel Sanchez

BAN 525

Dr. Cetin Ciner

Module 6: Assignment 1 -- Boosted NNs and Medical Costs

June 21st, 2021

## Introduction

Imagine walking into a store, knowing that you need to purchase something there. You are not sure what that item is, or when you might need it. But nevertheless, you do or will eventually need what this store is selling. You walk down an aisle, only to find nondescript boxes of different shapes and sizes lining the shelves. You notice that none of these boxes has its price listed. You approach a store clerk and ask how much one these mystery items costs. Instead of answering plainly, the clerk responds with a series of questions: What is your age? Have you been here before? Where are you from? What is your level of physical activity? Do you smoke? This line of questioning makes you suddenly realize that the store you've entered is a health insurance marketplace.

The health insurance industry can be baffling, with its complicated systems and jargon, quotas and limitations. The fixed cost of health insurance premiums can be particularly mystifying. Data analysis may help us better understand and target the determinants of overall medical costs.

The data set in this analysis contains observations of seven variables from 1,338 individuals in a healthcare setting. The variables and their definitions are listed below.

- **Age**: insurance contractor age, years
- **Sex**: insurance contractor gender, [female, male]
- **BMI**: Body mass index, providing an understanding of body, weights that are relatively high or low relative to height, objective index of body weight (kg / m ^ 2) using the ratio of height to weight, ideally 18.5 to 24.9
- **Children**: number of children covered by health insurance / Number of dependents
- **Smoker**: smoking, [yes, no]
- **Region**: the beneficiary's residential area in the US, [northeast, southeast, southwest, northwest]
- **Charges**: Individual medical costs billed by health insurance, in dollars

Charges will be selected as the response variable from this data set. Of the remaining features, Age and Sex are immutable. Meanwhile, BMI, Children, Smoker, and Region stand out as lifestyle choices that could potentially be modified. If these latter features prove to be predictive of medical costs, they could be targeted by campaigns that promote health and financial wellness.

The analysis in this report has been performed in JMP Pro 16, using an array of different methods. These include Ordinary Least Squares (OLS) and three versions of Boosted Neural Nets, with variations in their model specifications. The configurations of the boosted neural net model candidates are as follows:

1.) Hyperbolic Tangent Activation Function, 3 Nodes, 1 Layer, 40 Models
2.) Hyperbolic Tangent Activation Function, 1 Node, 1 Layer, 40 Models
3.) Hyperbolic Tangent Activation Function, 3 Nodes, 1 Layer, 40 Models, Absolute Penalty

## Analysis and Model Comparison

*Ordinary Least Squares (OLS):* This method will simply be used as a benchmark for the analysis of the continuous variable, Charges. OLS fits a model by minimizing the squared differences between observed and predicted values. Since all of the model effects are included in the final model, overfitting can become an issue.

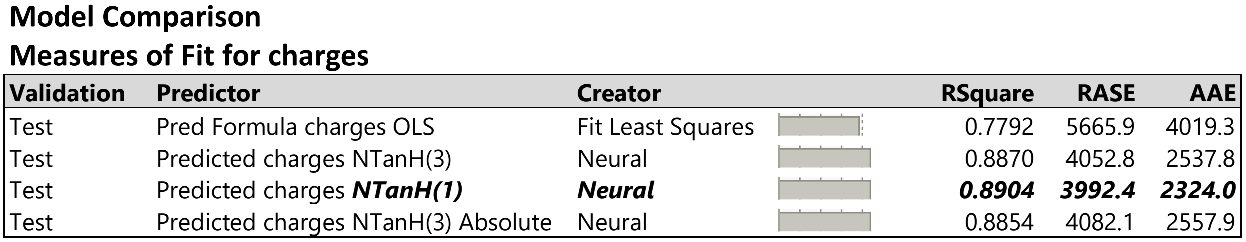
*Boosted Neural Net (Hyperbolic Tangent Function, 3 Nodes, 1 Layer, 40 Models):* This method is helpful for modeling very complex non-linear relationships. Designed to simulate how the human brain processes information, this method sends inputs through nodes (3 in this case), where a transformation function is applied. The zero-centered, S-shaped hyperbolic tangent(TanH) function allows for a non-linear relationship with outliers and approximates a linear regression for mid-range values.

*Boosted Neural Net (Hyperbolic Tangent Function, 1 Node, 1 Layer, 40 Models):* This iteration of the boosted neural net model will apply a hyperbolic tangent function through just a single node.

<u>*Boosted Neural Net (Hyperbolic Tangent Function, 3 Nodes, 1 Layer, 40 Models, Absolute Penalty):*</u> This final iteration of the boosted neural net models is nearly identical to the first version. However, instead of applying a squared error penalty, it will use an absolute error penalty.

Since the health insurance data set in this analysis is cross-sectional (as opposed to being time-series), the cross-validation method for this analysis involves randomness. The data set has been split into three parts. The *Make Validation Column* facility in JMP Pro 16 was selected, applying a random seed of 123. Implementing a 60-20-20 split of the 1338 total observations, 803 rows were used to train the models, 268 rows were tied to validation, and 267 rows were set aside for testing.

The Charges variable was selected in JMP Pro 16 as the Y variable. All of the other variables in the data set were included as candidate predictors. The Validation Column was implemented, and the analysis was conducted using all four methods. The random seed of 123 was applied to the boosted neural net models, running 40 models through 20 tours. As predictions from each model were generated, they were saved and stored in new columns in the data set. Only the profile formulas were saved, so that the hidden layer values remained hidden. The figure below shows the comparison of how the four models performed with the test set.



**Model Comparison**
**Measures of Fit for charges**

| Validation | Predictor | Creator | | RSquare | RASE | AAE |
|---|---|---|---|---|---|---|
| Test | Pred Formula charges OLS | Fit Least Squares | | 0.7792 | 5665.9 | 4019.3 |
| Test | Predicted charges NTanH(3) | Neural | | 0.8870 | 4052.8 | 2537.8 |
| Test | Predicted charges ***NTanH(1)*** | ***Neural*** | | ***0.8904*** | ***3992.4*** | ***2324.0*** |
| Test | Predicted charges NTanH(3) Absolute | Neural | | 0.8854 | 4082.1 | 2557.9 |

These models were evaluated in terms of RSquare, RASE, and AAE. The Boosted Neural Net model with one layer and one node, applying a hyperbolic tangent function outperformed the other models. Its RSquare value of 0.8904 was a significant improvement over the OLS benchmark model, and marginally better than the other Boosted Neural Net configurations.

The single-node model also had the lowest RASE and AAE values (3992.4 and 2324.0, respectively). This model will be more closely examined in the rest of this analysis.

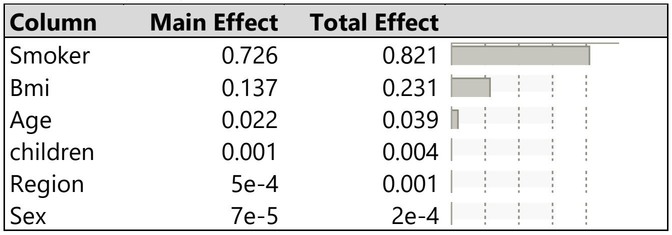
## Interpretation

The single-node boosted neural net model achieved optimal results after creating all 40 models. Variable importance analysis reveals that the binary variable Smoker is the most predictive of Charges. It is startling to find that the total effect of this one variable is 82.1%. The next important variable is BMI, which has only a 23.1% total effect. Surprisingly, Age comes in a distant third place, with a total effect of 3.9%. Children, Region, and Sex do not appear to be predictive of Charges.

### Model NTanH(1)NBoost(40)

### Variable Importance: Independent Uniform Inputs

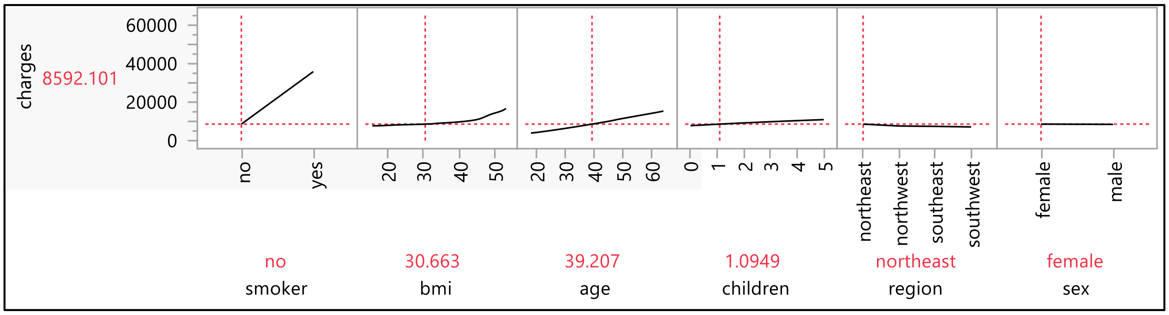
### Summary Report



| Column | Main Effect | Total Effect |
|---|---|---|
| Smoker | 0.726 | 0.821 |
| Bmi | 0.137 | 0.231 |
| Age | 0.022 | 0.039 |
| children | 0.001 | 0.004 |
| Region | 5e-4 | 0.001 |
| Sex | 7e-5 | 2e-4 |

The Prediction Profiler shown below confirms the variable importance relationships of Smoker, BMI, and Age with Charges. These all have a positive correlation with medical costs. Alternating between “No” and “Yes” for the Smoker variable alone results in a more than four-fold increase to Charges.

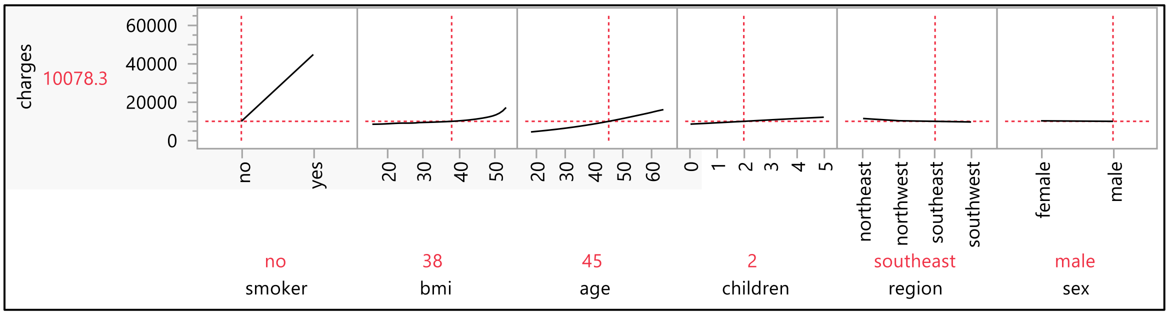
### Prediction Profiler



Adjusting values of each variable often shows how they may influence one another. Toggling the Smoker variable to “Yes,” for example, reveals a more dramatic curve of BMI in relation to Charges. However, adjusting any of the other variables does not have a similar effect on their counterparts. Smoking indeed does appear to have the greatest impact on medical costs and overall health.

Using the single-node Boosted Neural Net model, we are able to predict the medical costs for an individual with these sample traits: 45 years old, non-smoker, male, BMI of 38, from the southeast region, with 2 children. The predicted medical costs for this individual amount to $10,078.30. This is represented by the Prediction Profiler below, applying the appropriately adjusted values.

**Prediction Profiler - Sample Case**


charges
no
yes
northeast
northwest
southeast
southwest
female
male
no
smoker
bmi
age
children
southeast
region
male
sex