Gil Graybill – BAN 525
Assignment 6 – Boosted NNs and Medical Costs
Professor Cetin Ciner
June 21, 2020

Introduction

How much should you pay for health insurance? It's obvious that some individuals require more health care than others, but how do you determine that in advance? And if you could determine that, how do you determine a fair price to charge the individual? Do smokers pay more? Do people who exercise pay less? What if they eat at McDonalds every day? What if you lie to your insurance company about your health habits? The health insurance industry would love to have a dependable model that could predict how much a potential customer will cost them in a year. Too many variables would be unwieldy and delve too much into the details. Too little and they're stuck with a bad model that pays out too much or pays out too little (and upsets customers and regulators. Call in the lawyers!).

The dataset we're using consists of 1338 rows of individuals and properties about them, consisting of the following fields:

- **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, $ (*predicted value)*

Our response variable will be "Charges". We will be running the following types of analysis on this dataset:

Nominal Logistic Regression is a type of logistic regression model calculates the best fit through the data points that minimizes the squared differences between the observed values and the fitted values, and uses the calculated value to further determine the value of the nominal (or categorical) response variable.

A ***Neural Network*** is a set of algorithms that are designed to recognize patterns, simulating the human brain. As these patterns are recognized they are grouped together and virtual connections are formed. As with a neuron in the brain, an input (or firing) determines what other nodes are signaled. Each Neural Network has an input layer, one or many “hidden layers” where these connected are formed and revised, and an output layer. Within the “hidden layer” is where various algorithms are applied to the data across one or more nodes. A Neural Network can be “boosted” when multiple trees are built, with each new one dependent on prior trees that use residual error fitting. Another option that can be used is “Tours”, which tells how many times to create the model, with the best one chosen at the end. While the Tours option will increase processing time, choosing the best of many possible models can be a very good feature.

In our first Neural Network model we will have one layer with three nodes, and the algorithm applied will be the TanH activation function. It will be boosted with 40 models, and 20 tours will be run on it.

In our second Neural Network model we will have one layer with one node, and one algorithm will be applied: TanH. It will be boosted with 40 models, and 20 tours will be run on it.

In our third Neural Network model we will have one layer with three nodes, and one algorithm will be applied: TanH. 20 tours will be run on it.

Analysis and Model Comparison

OLS regression will find any relationships that exist between response variable and the potential dependent variables. One disadvantage for OLS is that if there are many variables that are highly correlated, OLS may model random errors. All variables will be included in the model, for better or worse. This is our “simple” benchmark analysis.

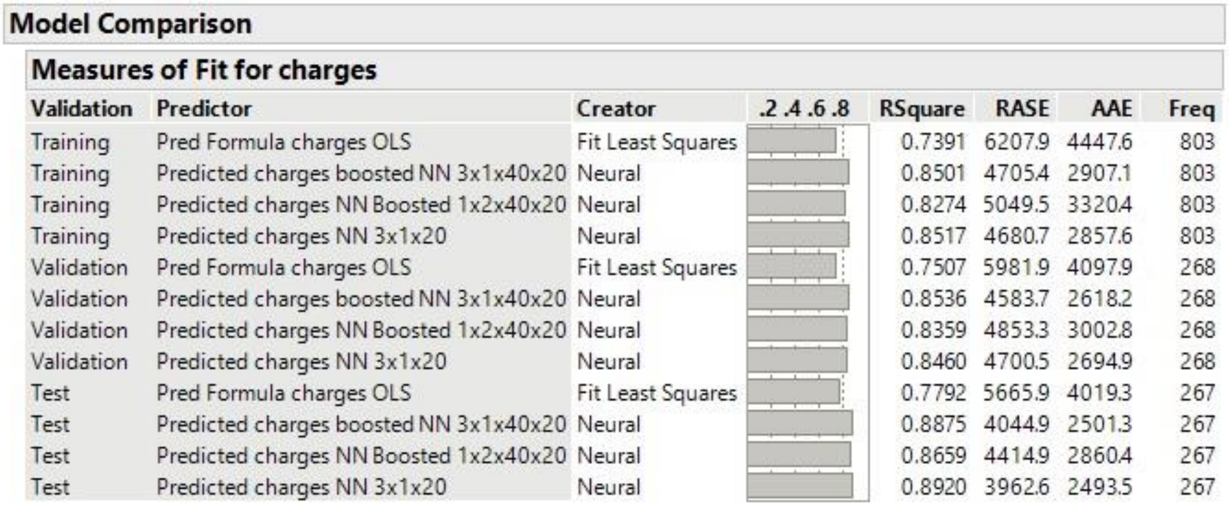
Neural Networks are a very versatile model that can be used on any data. If there is a complex relationship (that is to say, not linear or represented by a polynomial expression), NN will find it. With many different options for the structure of the hidden layer, boosting, and tours, there are countless way to create models. One disadvantage is that a Neural Network is a "black box" with its hidden layer, which makes it hard to troubleshoot or graphically represent the chosen model. Another is increased processing time.

For cross-validation we employed a 60-20-20 split for training-validation-test split. This gave us 803 rows of training data, 268 rows of validation data, and 267 rows of test data, splitting up the data randomly.

"Charges" was chosen as the response variable and the rest of the variables were tested as dependent variables using JMP Pro 15. The calculated value for the response variable, the calculated value for "Charges" was determined and recorded and the resulting formulas determined using all 4 methods.

To compare results between the models, we calculate the difference between each observed value for the change in charges and the calculated value for the change in charges. These differences can be compared in different ways. "RSquare" is a percentage that tells how much of the variability is explained by the model. "RASE" is "Root Average Standard Error", and because this describes values of error, smaller is better. "AAE" is "Absolute Average Error" is an error term that uses the absolute value of the error to compare the size of the error without regard for direction.

When comparing the models we will be using our cross-validation split. Because the test data is not used to create or refine the models, those are the results we are most interested in. The results are as follows:



Model Comparison

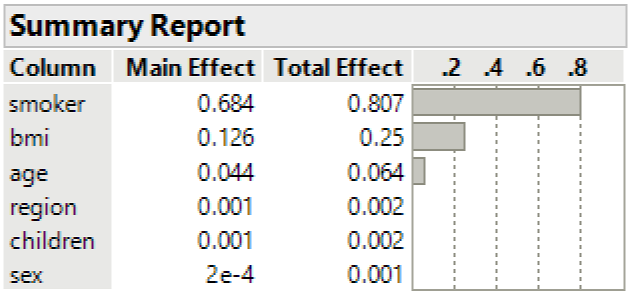
Measures of Fit for charges

Validation	Predictor	Creator	.2 .4 .6 .8	RSquare	RASE	AAE	Freq
Training	Pred Formula charges OLS	Fit Least Squares		0.7391	6207.9	4447.6	803
Training	Predicted charges boosted NN 3x1x40x20	Neural		0.8501	4705.4	2907.1	803
Training	Predicted charges NN Boosted 1x2x40x20	Neural		0.8274	5049.5	3320.4	803
Training	Predicted charges NN 3x1x20	Neural		0.8517	4680.7	2857.6	803
Validation	Pred Formula charges OLS	Fit Least Squares		0.7507	5981.9	4097.9	268
Validation	Predicted charges boosted NN 3x1x40x20	Neural		0.8536	4583.7	2618.2	268
Validation	Predicted charges NN Boosted 1x2x40x20	Neural		0.8359	4853.3	3002.8	268
Validation	Predicted charges NN 3x1x20	Neural		0.8460	4700.5	2694.9	268
Test	Pred Formula charges OLS	Fit Least Squares		0.7792	5665.9	4019.3	267
Test	Predicted charges boosted NN 3x1x40x20	Neural		0.8875	4044.9	2501.3	267
Test	Predicted charges NN Boosted 1x2x40x20	Neural		0.8659	4414.9	2860.4	267
Test	Predicted charges NN 3x1x20	Neural		0.8920	3962.6	2493.5	267

The three Neural Network models all performed better than the OLS, indicating that there are important non-linear relationships that are influencing our results. Another indication of good model creation is that when the test data was used against the models, the results were actually better than the training or validation data. It was actually our "simpler" Neural Network model of running only one layer with three nodes using the TanH algorithm with 20 tours that came out the best, with an R-Squared of 0.892, RASE of 3962, and AAE of 2493 for the test data. These are really outstanding numbers, indicating a very good predictive model.

Interpretation

When running variable importance with these variables against this model, we see the following:



Summary Report

Column	Main Effect	Total Effect	.2 .4 .6 .8
smoker	0.684	0.807	
bmi	0.126	0.25	
age	0.044	0.064	
region	0.001	0.002	
children	0.001	0.002	
sex	2e-4	0.001	

Smoking has a total effect of 81% on the model, followed by BMI with 25% and age with just over 6%. It's not surprising that smoking is a major indicator of health, but the large predictive value shows why everyone in health care benefits from less smoking. A high BMI indicates when an individual is overweight or morbidly obese, and individuals who have this condition are prone to more health problems. Both of these factors can be "multipliers" when it comes to poor health (https://www.sciencedaily.com/releases/2013/09/130911120719.htm). I'm surprised that age does not have a more prominent role, but that just shows what a huge influence smoking and obesity can have on health.

We've been asked to determine the projected health care costs for an individual with the following characteristics:

> 45 year old non-smoker male with a BMI of 38 from the southeast, who has 2 children

From this model, the projected medical costs would be $9,840.