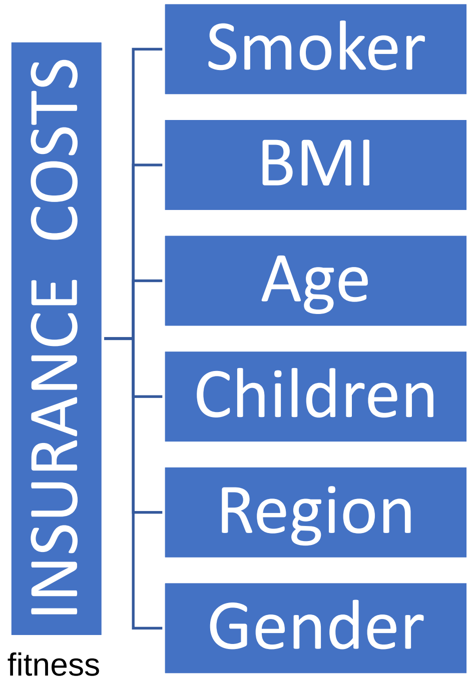
Boosted NNs and Medical Costs

Introduction



"Relative to the size of the economy, healthcare costs [in the U.S.] have increased over the past few decades, from 5 percent of GDP in 1960 to 18 percent in 2021", according to a recent article by the Peter G. Peterson Foundation[1]. After unprecedented government spending on healthcare related to the COVID pandemic, the question has to be asked: "How much is too much to spend on healthcare? Insurance companies have worked hard to combat costs by enticing their consumers to focus on prevention of high-costing medical care by partnering with health monitors like Apple watches, paying part of gym memberships and giving discounts if companies generate their own fitness programs as a way of improving the health of their consumers. For our data set, we are looking at five variables of health factors: Smoker (Yes/No), Body Mass Index (BMI), Age, Number of children covered by health insurance, and Gender (Sex) in addition to a Residential Area (Region) in order to predict the price of Individual medical costs or Charges.

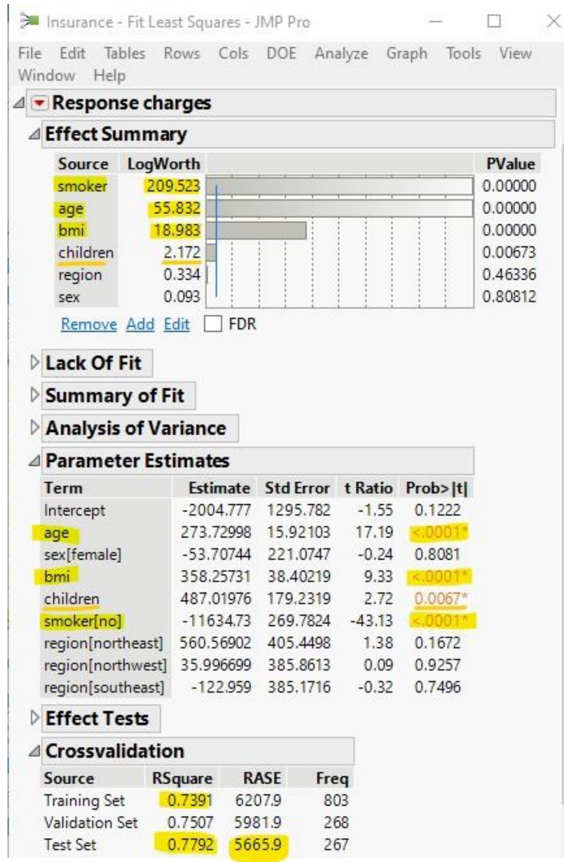
Our given data set has six variables and 1.338 undated observations or rows. There are three continuous variables of **BMI** [*Body mass index, providing an understanding of body, weights that are relatively high or low relative to height, objective index of body weight (kg / m ^ 2) using the ratio of height to weight, ideally 18.5 to 24.9*], **Age** [*insurance contractor age, years*], and number of **Children** [*number of children covered by health insurance / Number of dependents*]. There are two binary variables in the data set: **Smoker** [smoking, [yes, no]] and Gender or **Sex** [*insurance contractor gender, [female, male]*]. The data set also has 1 nominal or categorical descriptor variable with **Region** [*the beneficiary's residential area in the US, [northeast, southeast, southwest, northwest]*]. For the analysis of determining which of the variables predict the **Charges**, we will use all the given variables.

Following the health factors as shown in the graphic above, it would be reasonable to expect that Gender or Sex would not heavily weigh on healthcare charges as a women's increased reproductive costs would be balanced out by a man's increased cardiac costs as they age. It would also be reasonable that BMI would affect the healthcare costs a lot as it is known as a

[1] https://www.pgpf.org/blog/2023/01/why-are-americans-paying-more-for-healthcare

comorbidity of costly conditions like Cardiovascular disease and diabetes. Likewise, whether a person smokes would increase healthcare costs as the inflammation that it costs increases negative health symptoms such as lung condition and cardiovascular health.

Analysis and Model Comparison



Insurance - Fit Least Squares - JMP Pro

File Edit Tables Rows Cols DOE Analyze Graph Tools View Window Help

Response charges

Effect Summary

Source	LogWorth	PValue
smoker	209.523	0.00000
age	55.832	0.00000
bmi	18.983	0.00000
children	2.172	0.00673
region	0.334	0.46336
sex	0.093	0.80812

Remove Add Edit FDR

Lack Of Fit

Summary of Fit

Analysis of Variance

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	-2004.777	1295.782	-1.55	0.1222
age	273.72998	15.92103	17.19	<.0001*
sex[female]	-53.70744	221.0747	-0.24	0.8081
bmi	358.25731	38.40219	9.33	<.0001*
children	487.01976	179.2319	2.72	0.0067*
smoker[no]	-11634.73	269.7824	-43.13	<.0001*
region[northeast]	560.56902	405.4498	1.38	0.1672
region[northwest]	35.996699	385.8613	0.09	0.9257
region[southeast]	-122.959	385.1716	-0.32	0.7496

Effect Tests

Crossvalidation

Source	RSquare	RASE	Freq
Training Set	0.7391	6207.9	803
Validation Set	0.7507	5981.9	268
Test Set	0.7792	5665.9	267

The analysis will start with the Ordinary or Least Squares Model, which is the benchmark, the oldest of the strategies, to compare with. Then four additional models will be used, all of them will be a neural network model which is a simplified model based on the way neurons operate in layers within the human nervous system. According to IBM[2], there are typically three layers in a neural network model: 1) an input layer, 2) one or more hidden layers and 3) an output layer. Neural Networks and all machine learning models tend to overfit (where you get a great result in training data but will fare poorly in testing or new data).

To overcome overfitting, similar to the penalizing methods with Lasso and Elastic Net, there are built in penalties with neural networks which make estimation models perform more accurately. In SAS's JMP program, these penalties include "Squared", "Absolute", "Weighted Decay" or "None". Neural Networks also require a Validation method which splits data into 3 subsets of data: Training, Validation, and Testing so that the Estimation with penalties can take place at the same time as determining the best model to use. Additionally, there is a Random Seed of 123 in the models so that each time the model is run, the data is split in the same ways and is reproducible. The four neural network models will vary with the number of models to average together, however, the model fitting option for number or Tours is set to 20 so that the process is repeated 20 times to ensure a better prediction.

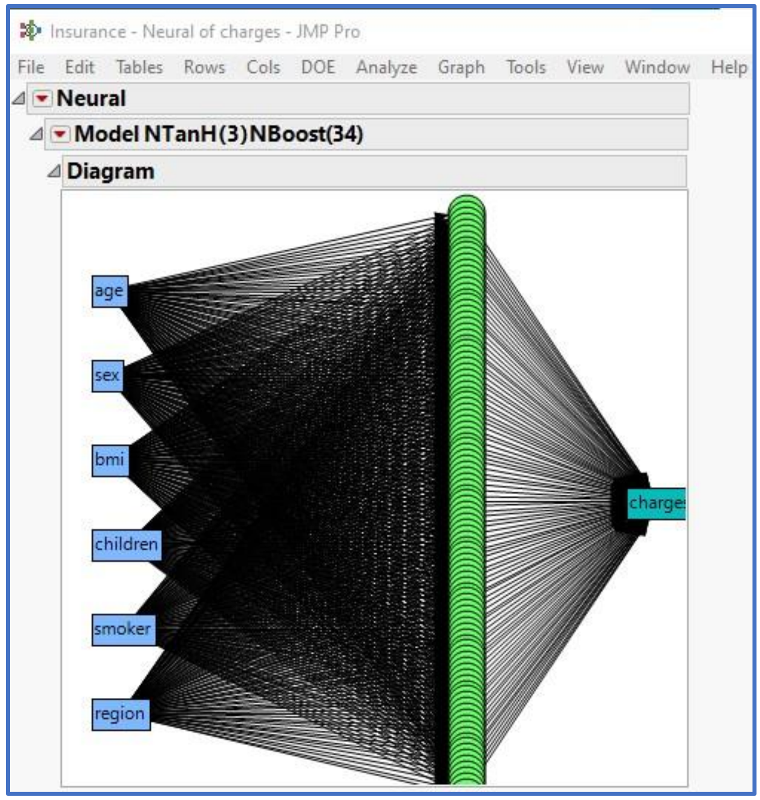
The difference between the four neural network models will be in the type of penalty applied and the number of nodes of each activation type that decides whether a neuron will be activated or

---

[2] https://www.ibm.com/docs/en/spss-modeler/18.0.0?topic=networks-neural-model

not. Both neural network models will use the **Squared Penalty method** as well as an **Absolute Penalty method** with a 10% weight to help prevent over-fitting. The first set of models will be calculated using the default Neural Network settings which include a first layer with **three nodes** in the TanH activation function (*the hyperbolic tangent function which is supposed to account for linear and non-linear functions*) but no other layers or activation types. The second set of models will be which include a first layer with only **one node** in the TanH activation function but no other layers or activation types.

For the data set, a Validation column will be created to allow cross-validation to occur within the modeling training phase at the same time estimation of the parameters of variables occurs. The training phase will contain about 802 observations or 60%. The Validation data subset contains 20% of the data which works to determine when to stop when it has found that the model is no longer improving, has about 268 observations or rows. Lastly, the Testing data subset is also 20% of the data and is used to determine how good the final model is once the Training phase is complete.



<u>The Best-Chosen Model</u>

In comparing the five Estimating models [**1**- the Ordinary or Least Squares (OLS) Method Model, **2**-the Default Neural Network Model (NTanH 3 nodes) with Squared penalty, **3**- the Default Neural Network Model (NTanH 3 nodes) with Absolute penalty, **4**- the Neural Network Model (NTanH 1 node) with Squared penalty, **5**- the Default Neural Network Model (NTanH 1 node) with Absolute penalty], an interesting pattern began to appear. While the OLS method needed no changes, almost all the neural models that were initialized to have 40 models to train and estimate with had the 1st computation come back with the average of the 40 models being the best model. This mean that the actual best model might be higher than 40. So, the model dialogue was run again with an additional 10 models until the best model came back les than the initial number of models. In other words, if the model dialogue said that there should be no more than 50 models, an acceptable "best model" could be 47 models. So, the **1st** Neural Network (3 nodes) model with Squared Penalty had the average of **47** models, the **2nd** Neural Network (3 nodes) model with Absolute Penalty had the average of **34** models, the **3rd** Neural Network (1 node) model with Squared Penalty had the average of **84** models, and the **4th** Neural Network (1 node) model with Absolute Penalty had the average of **61** models.

In order to compare all these models, I first compared the all Neural Network (3 nodes) models together, then I compared all the Neural Network (1 node) models, then finally, I compared the best of those two models with the OLS model. Both the selected Neural Network models that had **Absolute Penalties** with a 10% weight, outperformed the ones with Squared penalties. As the chart to the right reflects, the results were mixed. The Ordinary or Least Squares Model had the lowest R^2 value, as well as the lowest the Root Average^2 Error (RASE) and the Mean/Average Absolute Error (RMSE) when compared to both the Neural Network models. The two Neural Network models were comparable, as the neural network NTanH (1 node) with 61 models had a lower Average Absolute Error (AAE), but the neural network **NTanH (3 nodes) with 34 Models** stood out with a higher R^2 value and a lower RASE value.

Insurance - Model Comparison - JMP Pro

File Edit Tables Rows Cols DOE Analyze Graph Tools View Window Help

Model Comparison

Predictors

Measures of Fit for charges

Validation	Predictor	Creator	.2 .4 .6 .8	RSquare	RASE	AAE	Freq
Training	NTanH(1) 61Mod 10% Abs 20x Predicted charges	Neural		0.8428	4819.3	2887.7	803
Training	NTanH(3) 34Mod 10% Abs 20x Predicted charges	Neural		0.8496	4713.3	2975.2	803
Training	OLS Predicted charges	Fit Least Squares		0.7391	6207.9	4447.6	803
Validation	NTanH(1) 61Mod 10% Abs 20x Predicted charges	Neural		0.8463	4695.9	2578.9	268
Validation	NTanH(3) 34Mod 10% Abs 20x Predicted charges	Neural		0.8533	4587.8	2702.0	268
Validation	OLS Predicted charges	Fit Least Squares		0.7507	5981.9	4097.9	268
Test	NTanH(1) 61Mod 10% Abs 20x Predicted charges	Neural		0.8862	4067.8	2393.6	267
Test	NTanH(3) 34Mod 10% Abs 20x Predicted charges	Neural		0.8894	4009.0	2521.6	267
Test	OLS Predicted charges	Fit Least Squares		0.7792	5665.9	4019.3	267

Insurance - Neural of charges - JMP Pro

File Edit Tables Rows Cols DOE Analyze Graph Tools View Window Help

Neural

Validation Column: Validation

Model Launch

Model NTanH(3)NBoost(40)

Model NTanH(3)NBoost(47)

Model NTanH(3)NBoost(34)

Training — **charges**

Measures	Value
RSquare	0.8496159
RASE	4713.253
Mean Abs Dev	2975.2179
-LogLikelihood	7931.2889
SSE	1.784e+10
Sum Freq	803

Validation

Test — **charges**

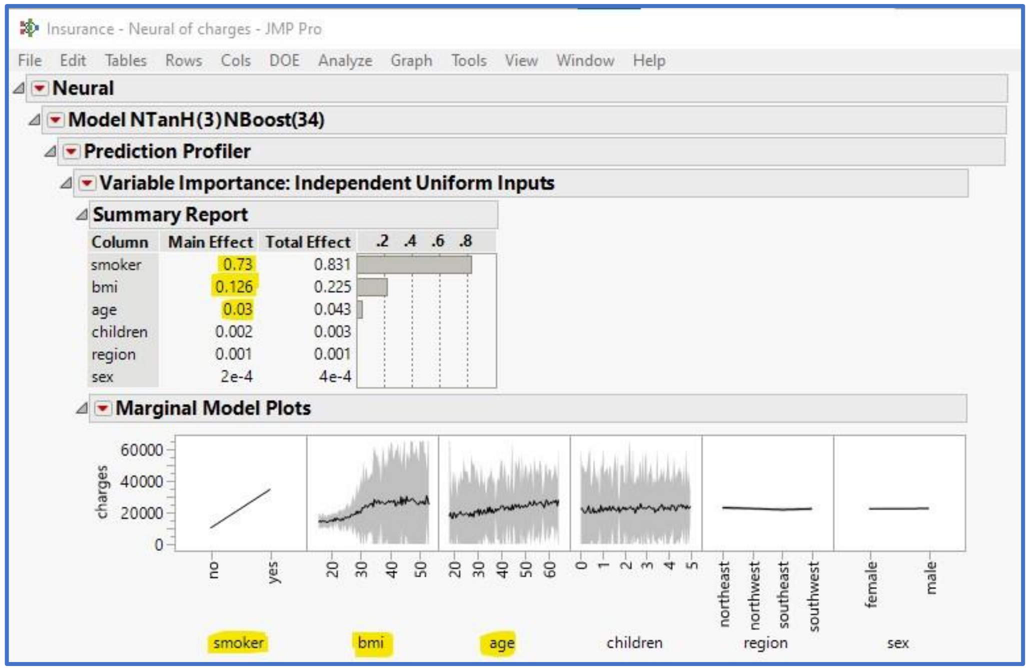
Measures	Value
RSquare	0.8894395
RASE	4008.996
Mean Abs Dev	2521.5857
-LogLikelihood	2593.9677
SSE	4.2912e+9
Sum Freq	267

Interpretation

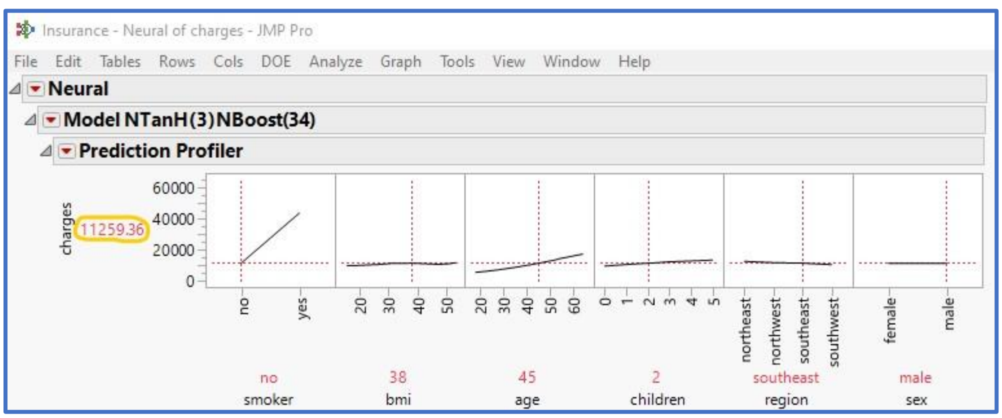
The first thing to look at is whether or not there is an overfitting problem with our Neural Network. With the Training R^2 value being 84.96% and a Testing R^2 value at 88.94%, it is safe to say that there does not appear to be an overfitting issue with this model. So, we can then go ahead and examine the variables.

Boosted NNs and Medical Costs

With the **Neural Network (3 nodes in the TaNH activation function) with 34 Models and Absolute penalties (10% weight)**, the data can then be best analyzed using JMP's Prediction Profiler Tools. The order of the importance of variables is examined using the Variable Importance: Independent Uniform Imports tool which clearly sets being a Smoker as the largest factor of Insurance charges. Since there a small difference between the Main Effect and the Total Effect, there might be some variable interaction but it has a whopping 73% of the Main effect on Insurance charges, all on its own. BMI and age both show some variable interactions as well, but if you look at the Main Effect on Charges, BMI has just 12.6% and Age has just 3%. As expected, the Gender or Age is the least important and just barely registers an effect on Insurance Charges.



The model was then used to formulate a prediction of new case with a 45-year-old non-smoker male with a BMI of 38 from the southeast, who has 2 children. Using JMP's Prediction Profiler, the data was manipulated to reflect the test case and the estimated predicted Insurance Charges was found to be **$11,259.36**. If we only change the Smoker to a "Yes", the Insurance charges jump up to $43,584.38 which is almost 4 times the cost of the non-Smoker!



The prices in this undated data set must be accepted to be a sample of the insurance charges in the snapshot of time from which they were taken from. As stated in the 1st paragraph, U.S. healthcare is costing more every year. Adding into the costs are the Baby Boomers in retirement and Americans poor eating habits which lead to serious comorbidities.