Christopher Dillard 10/8/2023

Boosted Neural Networks and Insurance Charges

In this assignment, I conducted an in-depth analysis of a regression problem to assess the predictive accuracy of boosted neural networks in estimating insurance charges. I used an Ordinary Least Squares (OLS) model as a baseline for comparison. The dataset comprises individual medical records from health insurance companies, with six predictor variables, being: Age, Sex, BMI, Children, Smoker, and Region. These variables encompass both discrete and continuous data, providing for a complex analysis of the many relationships between the predictors and response variable, Charge.

This aspect of the data helps to showcase the robustness of Boosted Neural Networks, in how they mimic human learning in all of its implied complexities. That is, Neural Networks identify patterns iteratively across random samples in the training sets. The Boosted concept refers to the ensembling strategy that serves to continuously learn from poor predictions (residuals) in previous iterations by applying higher weights on these weakest residual. This process is continued until the model ceases to improve its predictive accuracy on the validation split, after which the model is then run on the new, unbiased data in the testing split. As such, this assignment is much more computationally intensive than previous ones and lends itself to a richer discussion of the how Neural Networks can handle the imperfections present in data collected in an uncontrolled, “real-world” setting.

My three Boosted Neural Network models have the following characteristics:

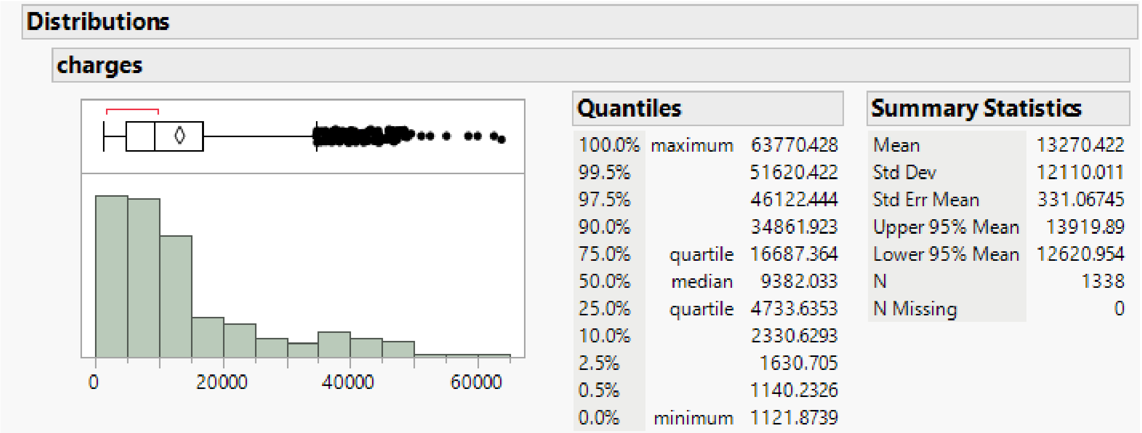
1. NTanH(2), 500 Boosts (Boosts are entered as “Number of Models” in JMP) – with a default 0.1 learning rate.
2. Same as the first model, but with the “Robust Fit” option checked.
3. Like the first one (not robust) but with a 0.05 learning rate.

I used a random seed split of 123 to ensure that the result were as replicable as possible. While setting up the Boosted Neural Networks, I also set the number of tours to be 20. This number refers to how many random starting points in the training set will be chosen by JMP. Since the inherent randomness in Neural Network can lead their results to be slightly different every time, choosing 20 tours should yield a similar result since JMP will keep the starting point that produced the best model. The lower learning rate significantly slows the models performance, yet also can have the benefit of avoiding overfitting by not overshooting the optimal weights on the residuals. So, models with smaller learning rates are more computationally intensive, but could be necessary to avoid overfitting it the potential is suspecting in the dataset like where the distribution of several variables are very skewed, which is the case for this dataset, which I will discuss later in this paper). This dataset does not have too many predictor variables, however, and has a relatively good ratio of rows to columns, which does lower the risk of overfitting. So, by including models with different learning rates in this assignment, I can address these issues accordingly and determine where a lower learning rate could produce a better model at predicting the charges in the testing split.

The Robust Fit is another measure to avoid overfitting by reducing the influence of outliers in the response variable. The NTanH(2) specification for each model refers to how many nodes there are for each activation type in the Hidden Layer Structure. In this case, all models had 2 nodes on the TanH activation function in only the First layer option, leaving all other boxes blank for this section. The number of nodes essentially defines the capacity of the model to accurately learn complex patterns in the training set. Adding too many nodes can however lead to overfitting, in which case retrofitting the model, or “pruning” the model, would be useful to optimize its performance if I suspect this to be the case when comparing each model’s RSquare values for the Testing split, compared to the Training and Validation splits.

For the sake of some preliminary data exploration, which could shape my interpretation of the predictive model, I explored the distribution of the response variable, Charges. Charges is the individual medical costs billed by health insurance.

Here is the output of the Quantiles table for the Charges column for all 1338 rows of data:


Distributions

charges

Quantiles

100.0%	maximum	63770.428
99.5%		51620.422
97.5%		46122.444
90.0%		34861.923
75.0%	quartile	16687.364
50.0%	median	9382.033
25.0%	quartile	4733.6353
10.0%		2330.6293
2.5%		1630.705
0.5%		1140.2326
0.0%	minimum	1121.8739

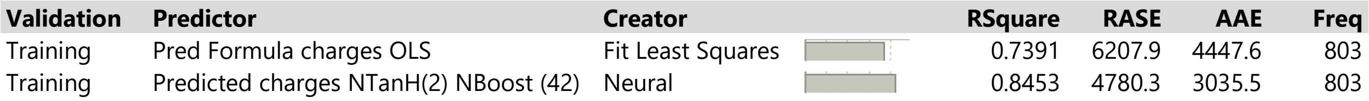
Summary Statistics

Mean	13270.422
Std Dev	12110.011
Std Err Mean	331.06745
Upper 95% Mean	13919.89
Lower 95% Mean	12620.954
N	1338
N Missing	0

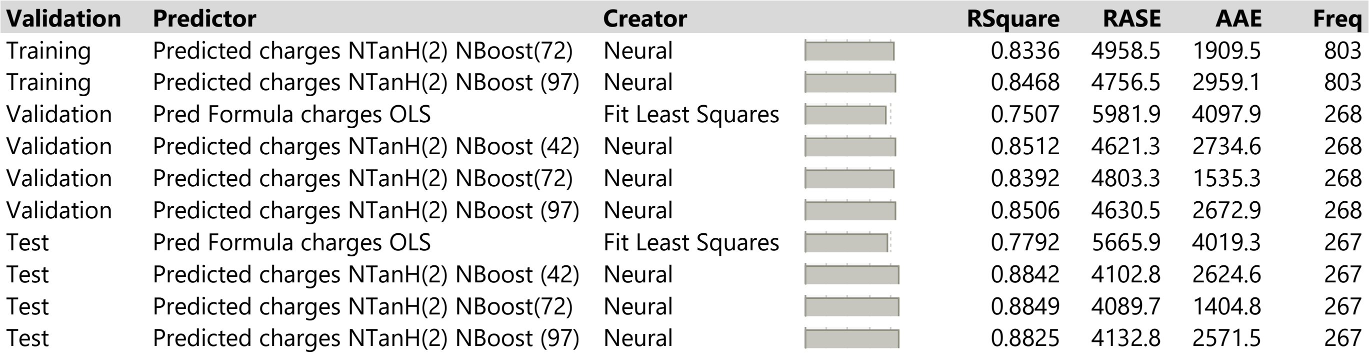
The distribution of the charges is highly skewed to the right, with many outliers identified as dots in the boxplot. My goal in this assignment is to identify which predictor variables have the greatest impact on charges to determine why so many people in the dataset were paying far more than the median cost of $9,382.03.

Prior to constructing the models, I needed to ensure that I had a correctly formulated Cross-Validation Column with a 60-20-20 split for the Training, Validation, and Testing data. Neural Networks, unlike previous methods explored throughout this course, require a validation split on the data as a means to control for the comparatively higher potential for overfitting.

So, after creating the validation column (with a random seed of 123), I then ran the models and saved their predictions into their own respective columns. To test the accuracy of each model and determine which one performed the best, I used the “Model Comparison” tool, under Predictive Modeling option in the Analyze drop down box. The four columns containing the predictions were entered into the “Y, Predictors” section and grouped by the “Validation” column. Upon running this Model Comparison, JMP assesses how close the models were to predicting the actual charges for each individuals, where the variance of the prediction from the actual values is listed in the RSquare column. The RSquares on the “Test” split in the Validation column is the most important when identifying the most accurate model, where the best model would have the highest RSquare.



Validation	Predictor	Creator	RSquare	RASE	AAE	Freq
Training	Pred Formula charges OLS	Fit Least Squares	0.7391	6207.9	4447.6	803
Training	Predicted charges NTanH(2) NBoost (42)	Neural	0.8453	4780.3	3035.5	803

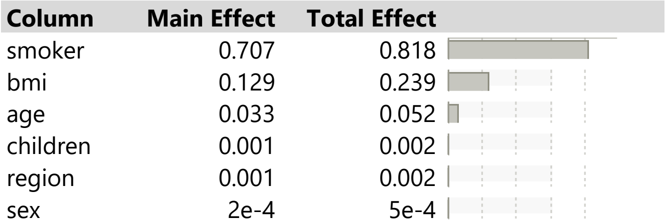


Validation	Predictor	Creator	RSquare	RASE	AAE	Freq
Training	Predicted charges NTanH(2) NBoost(72)	Neural	0.8336	4958.5	1909.5	803
Training	Predicted charges NTanH(2) NBoost (97)	Neural	0.8468	4756.5	2959.1	803
Validation	Pred Formula charges OLS	Fit Least Squares	0.7507	5981.9	4097.9	268
Validation	Predicted charges NTanH(2) NBoost (42)	Neural	0.8512	4621.3	2734.6	268
Validation	Predicted charges NTanH(2) NBoost(72)	Neural	0.8392	4803.3	1535.3	268
Validation	Predicted charges NTanH(2) NBoost (97)	Neural	0.8506	4630.5	2672.9	268
Test	Pred Formula charges OLS	Fit Least Squares	0.7792	5665.9	4019.3	267
Test	Predicted charges NTanH(2) NBoost (42)	Neural	0.8842	4102.8	2624.6	267
Test	Predicted charges NTanH(2) NBoost(72)	Neural	0.8849	4089.7	1404.8	267
Test	Predicted charges NTanH(2) NBoost (97)	Neural	0.8825	4132.8	2571.5	267

The second Neural Network, which stopped on the 72nd boost, performed the best on the test split. As expected, this model also had the lowest Root Average Square Error (RASE) and Average Absolute Error (AAE) values. This model has a learning rate of 0.1 and was my Robust model.

Now that I have identified the best performing model, I have assessed each variable's importance by their Independent Uniform Inputs in the Model's Prediction Profiler. This information is helpful to answer my previous point about what could be causing so many outliers in the distribution of the Charge's.

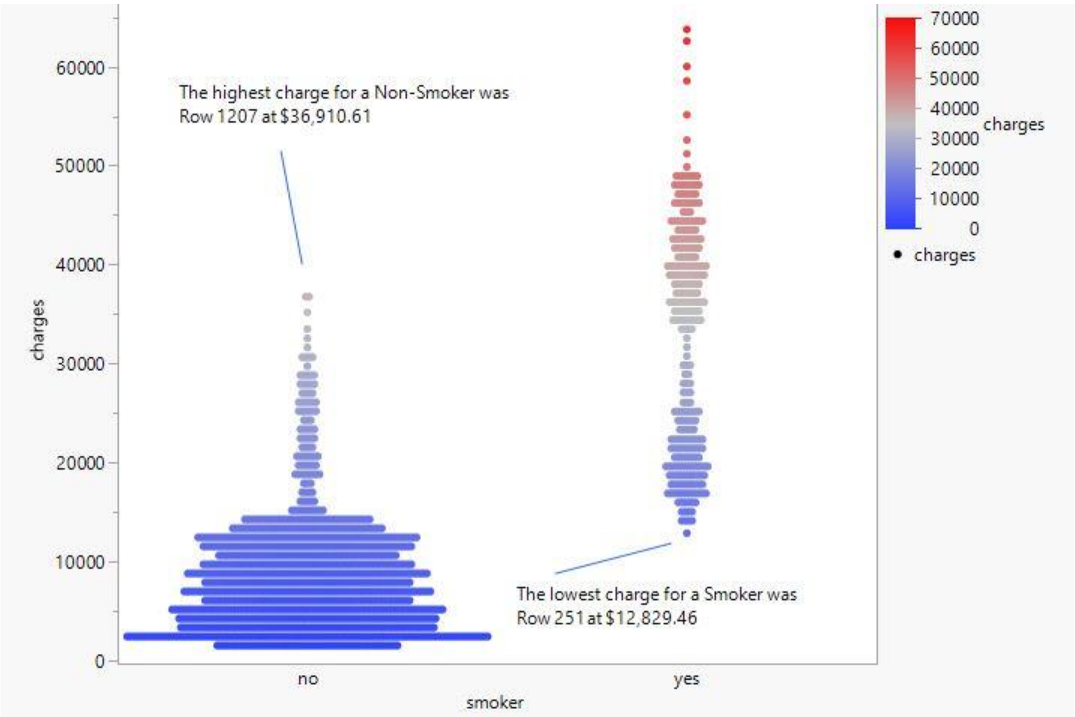
Independent Uniform Inputs



Column	Main Effect	Total Effect
smoker	0.707	0.818
bmi	0.129	0.239
age	0.033	0.052
children	0.001	0.002
region	0.001	0.002
sex	2e-4	5e-4

Whether an individual was a smoker or not was identified as the most impactful variable on the predicted insurance charges. I created the graph below to better visualize how powerful this variable was by evaluating the distribution for of charges for the two Smoking/Non-Smoking gorups in the entire dataset:

Distribution of Charges for Smokers vs. Non-Smokers



The highest charge for a Non-Smoker was Row 1207 at $36,910.61

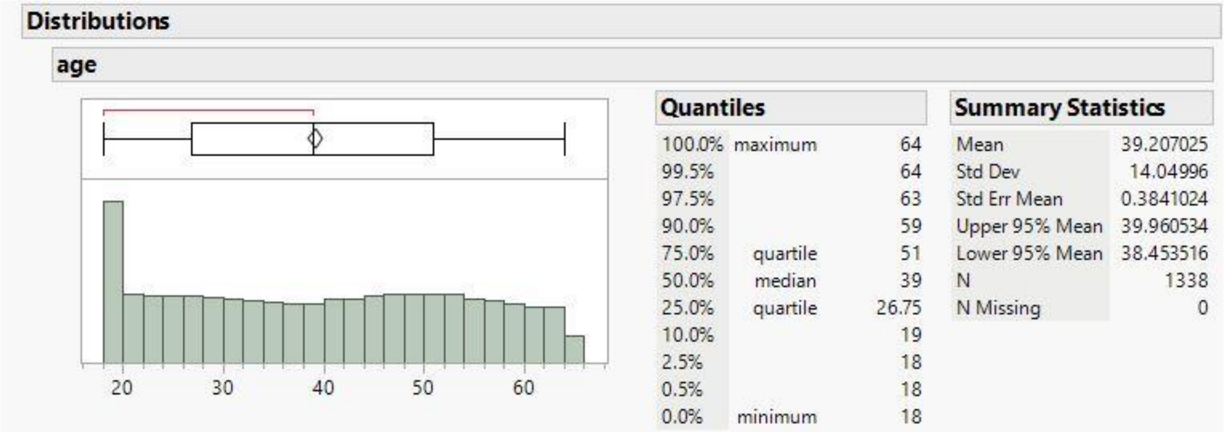
The lowest charge for a Smoker was Row 251 at $12,829.46

The highest charge for a Non-Smoker corresponded to a 59-year-old female with a BMI of 34.8 and 2 children in the southwest region. As for the lowest smoker, the individual was an 18-year-old male with a BMI of 17.29 and 2 children in the northeast region.

Something that struck out as significant in these profiles is how the second and third most impactful variables from the Independent Uniform Inputs table seem to have played a significant role in producing their counterintuitively extreme values for their respective groups. These are of course BMI and Age.

With age intuitively being be an important factor, I would have expect it to have a larger effect than just 5% for health insurance charges. So, while the model was very good at predicting the results on the testing data, one question to be asked is whether this model can be extrapolated to the general population. One particular characteristic in this dataset that may present a challenge when introducing new data from a different source is how skewed the dataset is in terms of reported ages.

As shown below, the dataset is highly skewed to the right:



Distributions

age

Quantiles

100.0%	maximum	64
99.5%		64
97.5%		63
90.0%		59
75.0%	quartile	51
50.0%	median	39
25.0%	quartile	26.75
10.0%		19
2.5%		18
0.5%		18
0.0%	minimum	18

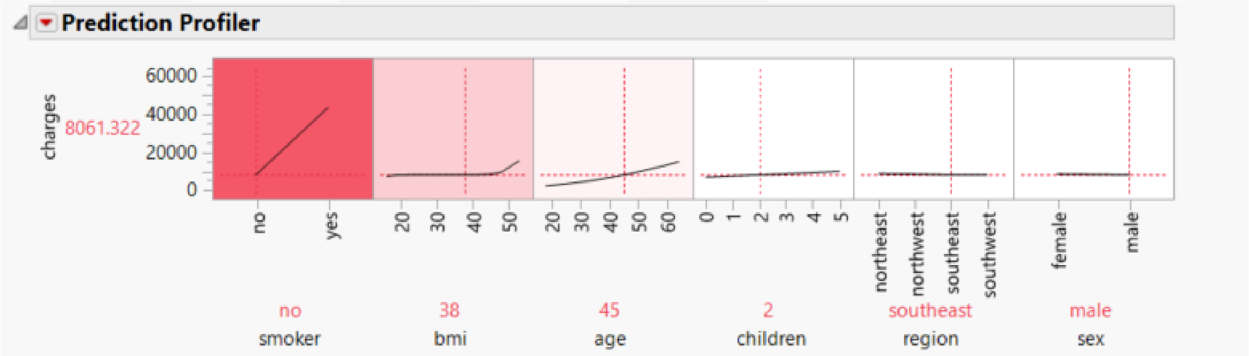
Summary Statistics

Mean	39.207025
Std Dev	14.04996
Std Err Mean	0.3841024
Upper 95% Mean	39.960534
Lower 95% Mean	38.453516
N	1338
N Missing	0

Furthermore, the model could perform poorly on data for adults with ages outside of the range that it was trained on, considering how around 22% of the US population is 65 and older (AECF, 2022).

Nevertheless, I will be using my model to predict the insurance charges for someone with the following characteristics:

Sex: Male, Age: 45, Smoking: No, BMI: 38, Region: Southeast, Number of Children: 2



no smoker

38 bmi

45 age

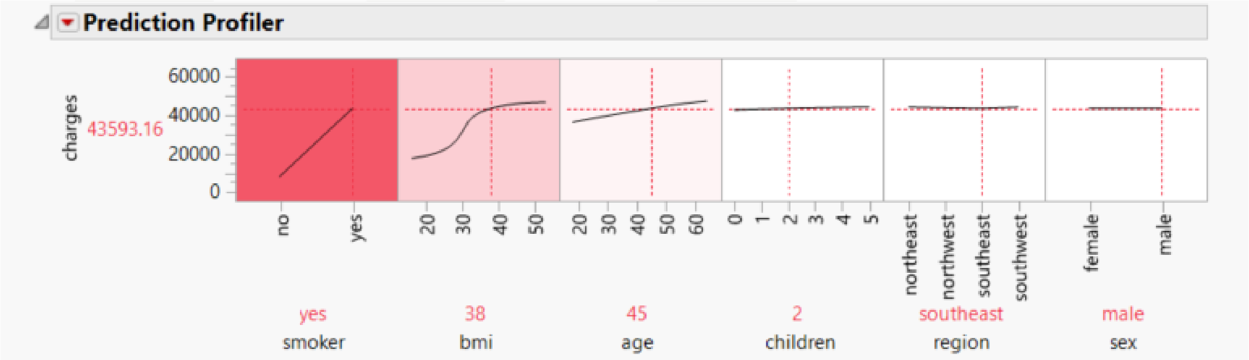
2 children

southeast region

male sex

According to the model, the patient at hand would have a predicted medical insurance cost of $8,061.32, which fell below the dataset's median charge of $9,382.03.

Seeing as though smoking has such a large impact on the model's predictions, the patient can be expected to pay $43,593.16 if they were a smoker:



This alternative, smoker rate for a patient with the same values for all other variables is $35,531.84 higher, which is a 440.77% increase.

References:

“Adult Population by Age Group: Kids Count Data Center.” *Adult Population by Age Group | KIDS COUNT Data Center*, July 2023, datacenter.aecf.org/data/tables/6538-adult-population-by-age-group#detailed/1/any/false/1095,2048,574,1729,37,871,870,573,869,36/117,2801,2802,2803/13515,13516.