Alec Engel
June 19, 2021

Introduction

The question at hand here is to determine what if any patient details play a role in predicting someone's health insurance premium charges. We want to run several cross-validation models to see what variables can be strongly correlated with our response variable of "charges". These cross-validation methods will allow us to split our dataset into training, validation and test portions. The split allows for better predictions of future outcomes with reduction in potential overfitting and removing unnecessary model noise amongst highly correlated independent variables. We will next take a look at our variables using a benchmark model of ordinary least squares followed by three varying neural network boosted models.

Predictor variables consist of patient history data and their current medical records that may cause increase or decrease of insurance premiums. The most important variables in this dataset are the response variable of charges; along with predictor variables of smoker or non-smoker, age, BMI, and how many children (dependents) they have. As we will elaborate later, whether a patient is a smoker or not will have highly unique information as it relates to the correlation of insurance premium charges.

In order to narrow down our predictor variables, we ran several statistical models and judged our outcomes. Statistical methods include cross-validation tactics of ordinary least squares (OLS) as well as a default and two more customized boosted neural network models (NN). The latter method of boosted neural networks produces extremely powerful and telling models that create a solid narrative when dealing with nonlinear data. Neural networks use hidden layer nodes to help model very complex variable relations between a model's inputs and outputs. Transformation is then applied and used to provide stronger model predictions. With neural networks, it is also important that we include a validation portion of our dataset as one drawback with neural networks is that they can overfit if validation is not present. Each neural network performed here is also being boosted which allows a model to be built sequentially by collecting and correcting residual errors. By placing heavier weight on these errors as each iteration is run, we get a more accurate prediction and avoid making the same model mistakes twice.

Analysis and Model Comparison

For the insurance premium charge response variable, we first used OLS which involves the inclusion of all variables. While using all variables, there are often noisy and unnecessary data variables that are accounted for. Our boosted neural network models performed much better while taking several different transformation approaches. As a default, we first used one layer of transformation with TangentH(3) and TangentH(1) while including 40 model runs and squared penalties. For comparison, we then used one layer of transformation with TangentH(3) while including 40 models and an absolute penalty. TangentH transformation function uses a similar "S" shape as a logarithmic function while Linear is similar to a linear regression model and Gaussian transformation function is similar to a normal (bell shaped) distribution. Taking the absolute value of our penalties for each model run in this dataset seems to be more powerful than squaring our errors. To get the clearest picture and better predictive abilities, we take the most stock in our absolute penalty version of boosted neural network that includes one layer and three nodes for TangentH transformation functions. Although more complex, this advanced model is our best bet for accurately predicting the charge of a patient's insurance premiums.

We have created a cross-validation column which splits our data into training (60%), validation (20%) and testing (20%) sets to allow our model to gain predictive attributes and score our R squared values and error rates for how we believe future data will look. As previously mentioned, this is an important step and validation must be included in neural network models to tell the model when to stop running and prevent overfitting. As you can see with this dataset, all models ran through 40 iteration boosts meaning no overfitting occurred and we probably could have even added additional models for a more enhanced view. We then created predictor formulas with each of the above four methods described. Each model was added as a dataset column so that we could easily perform model comparisons.

Model comparison results can be seen below. First with the training sets, we see all four models with relatively high performance. All four are above 0.77 R squared with somewhat low route average standard error rates. These however are our training sub sets of data and we want to focus more towards our validation and testing metrics. Moving to validation and test notations, we see consistency

with our training data but the winner with a slight edge in R squared value is our more complex model of boosted neural networks using TanH3 and absolute penalties. This model explains over 85% on our unbiased test data (R squared of 0.8547) and produces the lowest RASE. Although I will dig into each model to investigate strongest profile variables, I want to choose the strongest neural network boosted model here as my predominant model.

Model Comparison

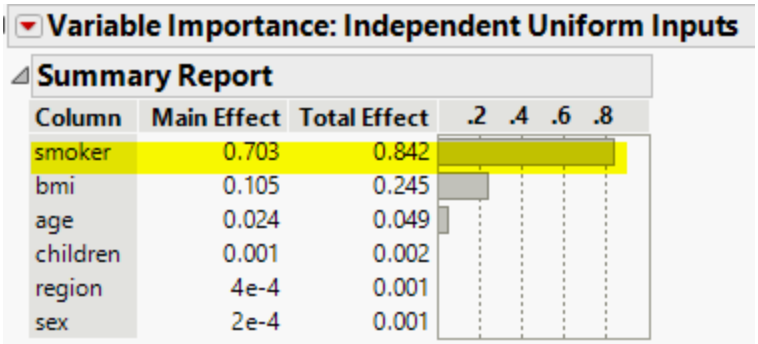
Predictors

Measures of Fit for charges

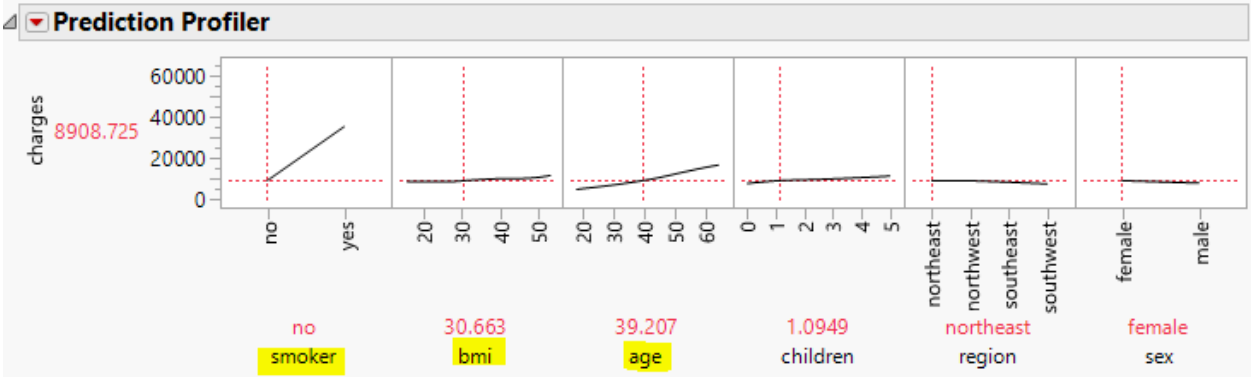
Validation	Predictor	Creator	.2 .4 .6 .8	RSquare	RASE	AAE	Freq
Training	Pred Formula charges OLS	Fit Least Squares		0.7710	5941.0	4093.0	803
Training	Predicted charges Boosted N TanH3 Squared	Neural		0.8830	4247.6	2402.0	803
Training	Predicted charges Boosted NN TahH1 Squared	Neural		0.8667	4533.5	2724.5	803
Training	Predicted charges Boosted NN TanH3 Absolute	Neural		0.8816	4272.2	2415.7	803
Validation	Pred Formula charges OLS	Fit Least Squares		0.6807	6581.1	4399.1	268
Validation	Predicted charges Boosted N TanH3 Squared	Neural		0.8114	5058.5	2758.1	268
Validation	Predicted charges Boosted NN TahH1 Squared	Neural		0.7971	5245.9	3023.2	268
Validation	Predicted charges Boosted NN TanH3 Absolute	Neural		0.8131	5035.1	2727.3	268
Test	Pred Formula charges OLS	Fit Least Squares		0.7473	5821.7	4114.2	267
Test	Predicted charges Boosted N TanH3 Squared	Neural		0.8531	4437.9	2604.0	267
Test	Predicted charges Boosted NN TahH1 Squared	Neural		0.8437	4578.6	2837.4	267
Test	Predicted charges Boosted NN TanH3 Absolute	Neural		0.8547	4413.9	2605.0	267

Interpretation

Per below, we can see that three prediction profiler variables were leaned on to maximize our overall model performance on test data. These top three parameter estimates in our boosted neural network model are smoker, BMI and age. We can see that a patient's sex, region and number of dependents don't play much if any role in determining premium charges. However, there is obviously heavy weight tied between insurance premium charges and whether a patient smokes or not. Per our model, a smoker will have a total effect (including non-linear variable interactions) of over 84% explanatory power. So, when determining a patient's insurance premium charge, the first thing that needs to be deciphered is whether they are a smoker or not. This makes a lot of sense since it's a well-known fact that smoking is not healthy and if you choose to smoke then you will pay higher premiums with insurance companies since they know you are at higher risk of major medical procedures.

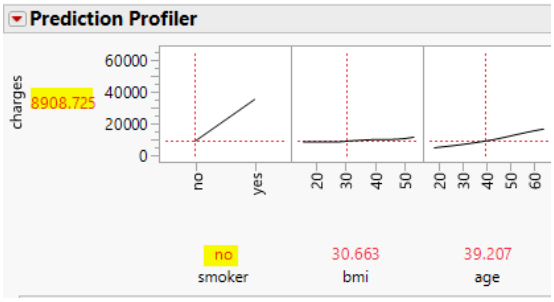


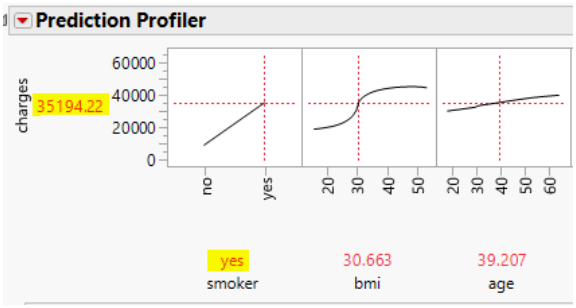
Smoker and BMI will be our most important parameter estimates as they show both high estimate numbers as well as low standard errors. For smoker, we see that as the classification of someone goes from non-smoker to smoker, the charges simultaneously rise in nearly a linear 45-degree angle. There looks to be no leveling off here either. As stated previously, if you are a smoker, your insurance premium charges will be much higher than if you didn't smoke. A similar notion can be used with a person's BMI in relationship to their charges. Someone with a high body mass index ratio will have a higher premium charge than someone with a lower BMI ratio. Both of these positive relationships with charges help us to accurately predict someone's insurance premium charges. These factors all make complete sense along with age because someone is more prone to need medical assistance and increased risk the older and more unhealthy they get.



When predicting with our profilers and performing some manipulation, we can back up our theory of important variables being smoker, BMI and age. I have highlighted our most valuable variable of smoker and amended from no to yes. As you can see, charges nearly quadruple from $8,908.73 to $35,194.22 just with the fact that someone smokes. This again

backs up our notion that smoker is our strongest predictor variable as its relationship explains over 84% of variations in a health insurance premium charge.





Finally, we will create a new case and use our preferred boosted neural network model with absolute penalties to predict insurance charges. We are using a 45-year-old non-smoking male with two kids who has a BMI of 38 and is from the southeast. Using our model to predict, this man's medical costs will be $9,613.25.

age	sex	bmi	children	smoker	region	charges	Validation	Pred Formula charges OLS	Predicted charges Boosted N TanH3 Squared	Predicted charges Boosted NN TahH1 Squared	Predicted charges Boosted NN TanH3 Absolute
45	male	38	2	no	southeast	•	•	12445.645806	9740.5860905	9954.478718	9613.2454258