Zackery Bradley

Ban 525

20220227

**Model 6 Assignment 1**

On behalf of this week’s assignment, we have been tasked with constructing predictive models to examine the determinants of medical costs billed by health insurance companies. Our dataset consists of seven variables. These variables include charges as our response variable followed by age, sex, BMI, children, smoker, and region. Regarding our pre-analysis hypothesis, it is our opinion that whether or not the person is a smoker will have the most significant impact on health insurance costs. This hypothesis is based on strong scientific evidence that smoking greatly reduces one’s life expectancy. Studies suggests that one cigarette shortens a person’s life by 11 minutes and for chronic smokers can shorten a person’s life by up to 10 years! Throughout this analysis, we will be utilizing four different modeling methods. These methods include the OLS method, along with Boosted NN models with consists of NTanH(3), NTanH(1), and NTanH(3) with absolute penalty. OLS is typically used as a method for estimating the parameters that are unknown in the model. This method helps us to find the relationships between independent and dependent variables. NN models are considered very flexible algorithms that are able to model extraordinarily complex relations between model inputs in outputs. Each NN consists of an unput layer, in addition to one or more hidden layers followed by an output layer. The hidden layers are made up of one or more nodes, in which a linear combination of the input variables is transformed. Boosted NN models follow this same methodology and is considered a relatively recent and widely popular method in machine learning. Boosting models are built in sequence, learning from its error within each step of the process. Boosting begins by estimating a model and obtaining residuals in the case of a regression and misclassification in the class of classification. From there, observations with the largest errors (where the model performed poorly) are given additional weight. The goal of this is that with each additional step the model fit is improved by correcting the subpar fit of the previous steps. Next, the estimates from these models are averaged to produce a prediction, or classification for each observation.

Analysis and Comparison

Regarding our predominant methods as previously stated, OLS can be described as the “workhorse” method and is a standard operating tool in statistics analysis. On the contrary, OLS tends to produce estimates with a great deal of variance regarding big data. In all, OLS typically generates poor forecasts. Our second method, NN also comes with its own list of advantages and shortcomings. As previously mentioned, NN typically has good predictive ability and is capable of capturing complex relationships. However, NN has no variable selection mechanism, so utilizing JMP’s variable importance is vital when identifying the most important variables in the data. Typically, NN is referred to as a “black box” method, meaning that findings are often difficult to interpret. Procedurally, we will be employing a cross-validation analysis. Cross-validation techniques are built on the concept of leaving out a part of the data out of the estimation process as a buffer. As the predictions of the model stop improving as the data holdout process occurs, model growth then stops, and estimates are obtained. This creates a ‘split’ in the data. These splits are then broken down into three parts in the terms of our model. Those parts include training, validation, and testing splits respectfully. Training data is the portion that is used to estimate our model. Validation data is not used in estimate directly, but instead works in the background to determine the optimal point in which the model stops. Lastly, the test data is never actually used in the model estimation. Test data is used to represent “new observations” and assists in providing and unbiased analysis of the predictive ability of the model. With the validation method only being utilized indirectly, we choose our model that displays the best performance on the test data. After conducting our analysis, we can see that one model proved to be more statistically proficient in relation to the others. Admittedly, my expectation was that NTanH(3) with absolute penalty would be the superior model. However, NTanH(1) proved to be the dominant performer, as it recorded an RSquare of 0.8904, with a RASE and AAE or 3992.4 and 2324 respectfully. This performance was followed by our NTanH(3) method, which calculated an Rsquare of 0.8870, followed by an RASE of 4052.8 and AAE of 2537.8. Third, was our NTanH(3) with absolute penalty method, with an RSquare of 0.8854 in addition to a RASE of 4082.1 and of AAE 2557.9. Lastly, our baseline model, OLS proved to be the least efficient model with an RSquare of 0.7792 a RASE of 4052.8 and an AAE of 4019.3.

**Measures of Fit for charges**

| **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(3) Squared | Neural |  | 0.8491 | 4721.7 | 2958.6 | 803 |
| Training | Predicted charges NTanH(1) Squared | Neural |  | 0.8433 | 4810.9 | 2774.0 | 803 |
| Training | Predicted charges NTanH(3) Boosted | Neural |  | 0.8475 | 4746.1 | 2968.8 | 803 |
| Validation | Pred Formula charges OLS | Fit Least Squares |  | 0.7507 | 5981.9 | 4097.9 | 268 |
| Validation | Predicted charges NTanH(3) Squared | Neural |  | 0.8516 | 4614.8 | 2693.8 | 268 |
| Validation | Predicted charges NTanH(1) Squared | Neural |  | 0.8451 | 4714.5 | 2474.8 | 268 |
| Validation | Predicted charges NTanH(3) Boosted | Neural |  | 0.8533 | 4588.0 | 2681.3 | 268 |
| Test | Pred Formula charges OLS | Fit Least Squares |  | 0.7792 | 5665.9 | 4019.3 | 267 |
| Test | Predicted charges NTanH(3) Squared | Neural |  | 0.8870 | 4052.8 | 2537.8 | 267 |
| Test | Predicted charges NTanH(1) Squared | Neural |  | 0.8904 | 3992.4 | 2324.0 | 267 |
| Test | Predicted charges NTanH(3) Boosted | Neural |  | 0.8854 | 4082.1 | 2557.9 | 267 |

Interpretation

After examining model performance and discovering NTanH(1) as our superior method, we performed additional measures to determine the impact of each variable in our dataset. We can see from the visual, our hypothesis was proved to be factual, as smoking was deemed to be the most significant factor regarding insurance costs by a substantial margin, having a total effect of 0.835. BMI proved to be our second most important, with a total effect of 0.23. Lastly, our third most important variable was age, as it carried a total effect of 0.048. By seeing these numbers, it becomes clear why health insurance companies always ask questions such as if you are a smoker, along with body measurement and age questions prior to insuring you.

| **Column** | **Main Effect** | **Total Effect** |  |
| --- | --- | --- | --- |
| smoker | 0.72 | 0.835 |  |
| bmi | 0.117 | 0.23 |  |
| age | 0.027 | 0.048 |  |
| children | 0.001 | 0.003 |  |
| region | 0.001 | 0.002 |  |
| sex | 1e-4 | 3e-4 |  |

Additionally, we were tasked with predicting the insurance charges for a 45-year-old non-smoker male, who resided in the Southeast region and had two children and a BMI of 38. We determined that this individual would be billed total of $10,799.35 in charges.

Graphical user interface, chart

Description automatically generated

In continuation, we conducted further analysis by breaking down just how costly smoking is in regard to insurance prices. We took the same male in the experiment above, and simply studied what the impact would be if he was a smoker. As we can see, the effects were devastating, as just being a smoker increased his billing charges by over 400% to $43,403.10.

Chart

Description automatically generated

Lastly, we examined all the variables within a marginal model plot to paint a bigger picture on how medical costs are billed throughout the given data.

Chart

Description automatically generated

In closing of our analysis, we can say with certain confidence NTanH(1) performed adequately in the discussion of this particular dataset. This can be seen on our overall analysis of the other performance methods as NTanH(1) was proven to be superior statistically speaking. Given how enormous the need for medical care is in the world, it is vital that companies have access to the best modeling analysis methods available in order to stay relevant. By seeing the significance of each variable associated with the pricing of medical costs billed by health insurance providers, companies are able to make better economic decisions on each individual they potentially insure, as well as having strong statistics as to what to charge those individuals. Having mathematical models such as this one saves the companies millions in lost revenue by effectively mitigating risk potential.