In this dataset, I examined 1338 datapoints of individual medical costs billed by health insurance and how that variable of interest correlates with 6 variables. I evaluated four models.

Model Comparison of 4 Models:Table

Description automatically generated

The best performing model on the test data is the boosted Neural Network TanH(1) with squared errors. This model has the best RSquared value and the lowest RASE and AAE.

This is how the model was formedChart

Description automatically generated but it’s nearly impossible to discern the relationships here.

Table

Description automatically generatedIn this model, we learn that the most important variable to predict individual medical costs is whether the person smokes or not. Also related as a predictor variable is BMI, but smoker vs non-smoker is the highest correlation, explaining 79% of the variability in medical costs billed by health insurance. These are very logical findings.

Chart, histogram

Description automatically generated

Common sense also says that medical costs paid out by health insurance companies rise for smokers and fall for non-smokers. Also, these same costs go up with rising BMI. This is confirmed in the profilers.

In the case of a 45 year old non-smoker male with a BMI of 38 from the southeast, who has 2 children. The predicted medical costs for him = $10,078.

As an additional note, this is not a uniformly distributed dataset. Here is the distribution of the response variable, charges (units of x axis is dollars).

Chart, histogram

Description automatically generatedThere are many outliers high in the 4th quartile. There are many datapoints in the first three bins, first two quartiles.

|  |  |
| --- | --- |
| Mean | 13270.422 |
| Std Dev | 12110.011 |
| Std Err Mean | 331.06745 |
| Upper 95% Mean | 13919.89 |
| Lower 95% Mean | 12620.954 |
| N | 1338 |
| Skewness | 1.5158797 |
| Kurtosis | 1.6062987 |
|  |  |

Normal kurtosis is 3, but we see by the 1.6 number that there is weight in the lower tail. Symmetrical dataset has a skewness of 0, here we see a 1.5 positive skew. The data is “longer” so to speak (we noted outliers previously) on the right side of the peak. It is asymmetrical. It’s just all worth noting. Of course, anyone who has battled with an insurance company for medical cost reimbursement is not surprised that the payouts are not normally distributed and that there is a lot of weight in the lower tail. This distribution shows their effectiveness of keeping the payouts skewed to the low side.