**Assignment 1: Boosted NNs and Medical Costs**

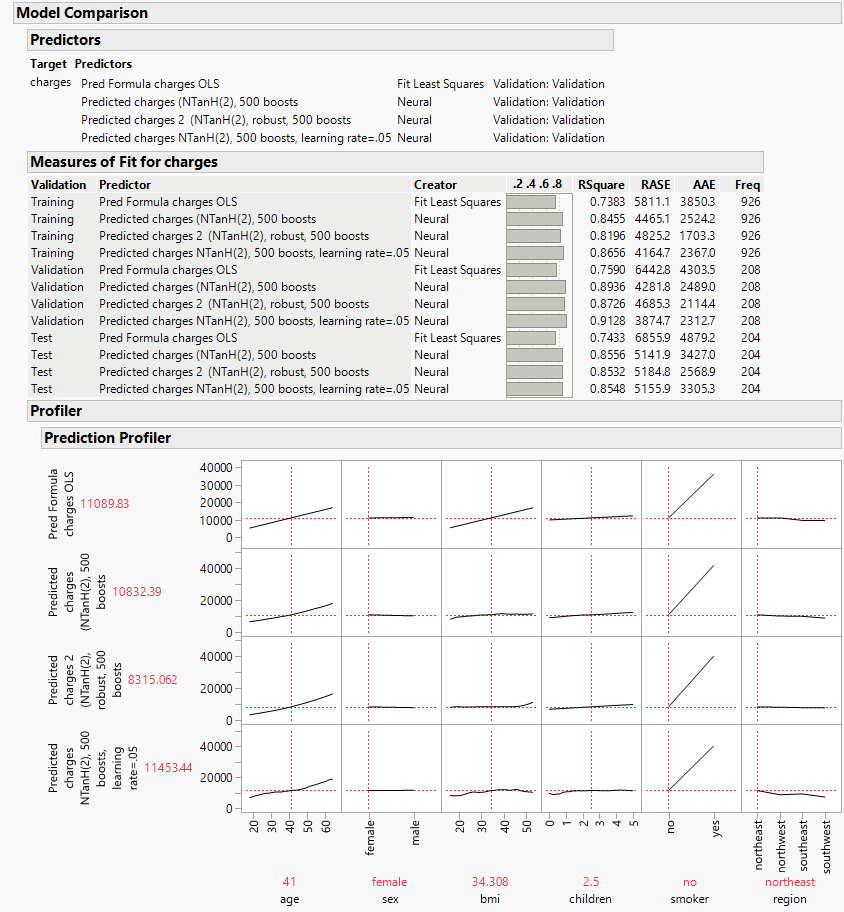
**Introduction**

The problem addressed in this analysis is the prediction of charges based on various factors such as age, BMI, smoker status, and region. Accurately predicting charges has applications in fields like insurance pricing and healthcare, where understanding how individual factors impact cost can support better pricing and resource allocation decisions. This study employs a neural network approach with boosting to develop a model that can make these predictions with high accuracy and generalizability. The dataset used includes individual-level attributes relevant to predicting charges, with the primary variables of interest being age, sex, BMI, number of children, smoker status, and geographic region.

The response variable in this analysis is the Charges variable, representing the cost or charge associated with each individual. The predictor variables consist of continuous (e.g., age, BMI) and categorical (e.g., smoker status, region) variables. The goal is to model the relationships between these predictors and the charges variable, aiming for both accurate predictions and insight into how each factor influences charges. The statistical method employed is a Boosted Neural Network (NTanH(2), 500 boosts, learning rate = 0.05), chosen after comparing several models. The boosted neural network was selected due to its ability to capture complex patterns in the data, providing both high predictive accuracy and the potential for more stable predictions on unseen data. This paper will cover the model comparison process, the interpretation of the results, and an exploration of the key determinants of charges.

**Analysis and Model Comparison**

The Boosted Neural Network model was chosen due to its advantages in handling complex, nonlinear relationships between predictors and response variables. Unlike traditional linear models, which assume a linear relationship between predictors and the response, neural networks can capture more intricate patterns in data, accommodating interactions and nonlinearities that may be crucial in predicting charges. Boosting, an ensemble technique, further enhances the model's accuracy by sequentially training multiple models, where each new model seeks to correct the errors made by the previous ones. This approach reduces bias and variance, helping the neural network generalize better to new data. However, neural networks with boosting are computationally intensive and require careful tuning of parameters such as the number of boosts, learning rate, and network structure.



The cross-validation procedure used in this analysis involved dividing the data into training (0.6), validation (0.2), and test sets (0.2). The training set was used to fit the model, the validation set to tune model parameters, and the test set to evaluate the final model's performance on unseen data. This split ensures that the model's performance is not inflated by overfitting on the training data and that it generalizes well to new instances. Cross-validation was especially important in this context because neural networks, given their complexity, can overfit if not properly regularized or validated.

The analysis involved setting up the neural network with two hidden nodes in a single layer, using the Tanh activation function. The model was then boosted with 500 iterations at a learning rate of 0.05, with the number of tours for each model set at twenty (20), which allows gradual improvement in prediction accuracy over each iteration without overshooting. Various models were compared, including a baseline ordinary least squares model, a neural network without boosting, and several boosted neural networks with different configurations of learning rates and robust fitting. After examining the performance metrics—R-squared, RASE (Root Average Square Error), and AAE (Average Absolute Error)—the model with NTanH(2), 500 boosts, and a learning rate of 0.05 was chosen as the best model.

The chosen model demonstrated a high R-squared value on the training set (0.8656) and an even higher R-squared on the validation set (0.9128), indicating strong predictive power and effective generalization. On the test set, the R-squared was 0.8548, showing that the model performed well even on unseen data. Additionally, the model's RASE and AAE values were comparatively low, suggesting that its predictions were not only accurate in terms of variance explained but also close to actual values on an absolute scale. In comparison, the other models showed lower R-squared values and/or higher error metrics, making them less optimal for this task. These results support the selection of the NTanH(2), 500 boosts, learning rate 0.05 model as the best option for accurate and reliable charge predictions.

**Interpretation**

Although neural networks lack traditional parameter estimates, the Prediction Profiler offers insights into how each predictor influences the response variable. According to the profiler, age has a positive impact on charges; as age increases, predicted charges rise, though the relationship is relatively moderate compared to other predictors. BMI, a measure of body mass, also positively impacts charges. Higher BMI values lead to increased predicted charges, which aligns with common knowledge in healthcare, where higher BMI often correlates with increased health risks and associated costs. Smoker status, one of the categorical predictors, has a particularly strong effect on charges. The profiler shows a marked increase in predicted charges for smokers compared to non-smokers, indicating that smoking is one of the most influential factors in predicting charges. This aligns with healthcare data trends, as smokers are generally at higher risk for numerous health issues, which translate into higher charges.

To further quantify the influence of each predictor on charges, a variable importance analysis was conducted based on the profiler’s output and the model's configuration. The table below ranks the top predictors according to their importance in determining charges, offering a clearer picture of the variables that drive the predictions.

| **Variable** | **Importance Ranking** | **Impact on Charges** |
| --- | --- | --- |
| Smoker Status | 1 | Large increase in charges for smokers |
| BMI | 2 | Higher BMI associated with higher charges |
| Age | 3 | Charges increase with age |
| Region | 4 | Slight regional variations in charges |
| Number of Children | 5 | Minor impact on charges |
| Sex | 6 | Minimal influence |

The table above highlights the primary determinants of charges, with smoker status and BMI identified as the most influential. These two variables have a significant impact on predicted charges, as seen in the profiler output and model results. Smoker status, in particular, causes a substantial increase in predicted costs for individuals who smoke, reflecting the increased healthcare costs associated with smoking. BMI follows closely, as it correlates with various health conditions that increase medical expenses. Age also plays a critical role, with costs rising moderately as age increases, which aligns with the increasing likelihood of age-related health issues. Region and the number of children provide minimal adjustments to charges, while sex has a negligible effect in this model.

Other predictors such as sex, region, and the number of children have comparatively smaller impacts on charges. For instance, the effect of the region is modest, with slight variations in predicted charges across different geographic areas. This suggests that regional differences might not be a major determinant of charges, at least in this dataset. Similarly, the number of children shows a slight increase in charges with more children, likely due to family-based pricing adjustments, but the impact is minimal compared to smoker status and BMI. These insights allow us to see the model's predictive logic and align with our understanding of risk factors associated with healthcare costs.

An important aspect of interpreting neural network models involves investigating the significance of each predictor in the context of the model’s profilers. By examining the profiler output, we can observe that smoker status and BMI are among the most critical variables driving predictions. This observation is further supported by the profiler tables, where smoker status and high BMI values cause substantial increases in charges, emphasizing these variables' importance in the prediction model. The Prediction Profiler thus provides a practical way to understand the variable relationships without direct parameter estimates, demonstrating the boosted neural network’s effectiveness in highlighting key factors.

**Conclusion**

In applying the model to make a specific forecast, suppose we want to predict charges for an individual with specific characteristics: a 40-year-old female with a BMI of 30, who is a smoker with no children and lives in the southeast region. Inputting these values into the profiler, we observe that the predicted charges are significantly higher due to the combination of smoking and a relatively high BMI. The model’s output aligns with expectations, reinforcing its reliability and practicality for making individualized predictions based on a combination of factors.

The Boosted Neural Network model (NTanH(2), 500 boosts, learning rate 0.05) was selected as the optimal model for predicting charges based on a combination of performance metrics, including high R-squared, low RASE, and low AAE. This model proved capable of capturing complex relationships between predictor variables and charges, demonstrating both accuracy and stability in prediction. Key predictors such as smoker status and BMI were identified as primary drivers of predicted charges, aligning with known risk factors in healthcare and supporting the model’s interpretability through the Prediction Profiler.

A screenshot of a graph

Description automatically generated

Using the chosen model, **NTanH(2) with 500 boosts and a learning rate of 0.05**, the Prediction Profiler provides an estimated medical cost of approximately **$9,511.79** for a 45-year-old non-smoker male with a BMI of 38, from the Southeast, who has 2 children. This model was selected as the best because of its high predictive accuracy and stability, evidenced by high R-squared and low error rates. In this prediction, the model’s output reflects how each variable impacts charges, with age, BMI, and smoking status identified as significant determinants based on previous analysis.

In particular, the model captures the moderate influence of both age and BMI on costs, as healthcare expenses generally increase with age and higher BMI, which often correlates with elevated health risks. However, since this individual is a non-smoker, the predicted cost is considerably lower than it would be if smoking were present; smoker status is a powerful predictor of increased costs, and the model effectively accounts for this impact through the boosted network structure. Additionally, the Southeast region and family size contribute slight adjustments to the overall prediction, aligning with typical healthcare cost patterns observed in regional and family-based pricing. The model’s accuracy and ability to capture these nuanced influences reinforce its suitability for predicting medical charges.

The analysis highlights the potential of boosted neural networks in predictive modeling for cost estimation, particularly in fields requiring a nuanced understanding of multifaceted relationships. By selecting and optimizing this model, we achieved an effective balance between variance explained and error stability, providing a model that generalizes well to new data. Future studies could explore the application of this model in larger datasets or test alternative boosting techniques to further improve prediction accuracy and adaptability. Overall, this boosted neural network model offers a robust tool for estimating charges, with implications for decision-making in insurance pricing, healthcare resource allocation, and risk assessment.