Lauren Cardieri
DataMining
Project 1: Healthcare Charge Predictor
August 13, 2025

For this project, I decided to make a model that predicts the charge of someone's health care insurance. The model is based on the following features: age, BMI, number of children, sex, if they smoke, and region. This specific business problem is important because it determines the cost of treatment someone will have to pay based on certain aspects of their life. This could be a determining factor between life and death, as proper healthcare treatment can be too expensive for people with lower incomes or financially struggling. This model will help individuals determine if they are getting charged correctly based on recent data, and allow them to dig into certain aspects of their lives to see what impacts their rate. For example, an individual can see how much their insurance charge may differ if they quit smoking or if they work on their body fat.

The dataset used for my model can be found on Kaggle at the following link, https://www.kaggle.com/datasets/willianoliveiragibin/healthcare-insurance, hosted by Willian Oliveira. For preprocessing, I went through the following procedures to clean my data. Firstly, I checked to see if there were any missing values in my dataset, and found that there were none. Then, I transformed the gender column into 0s and 1s using the get_dummies function. I attempted to do the same for the region column, but found that it had a negative affect on the performance of my model. Instead, I decided to find the mean charge of each unique region, then substitute it in as a value the model can read, which worked better for my model. After this, my data was ready to be processed.

Next, it was time to determine which model would be best for my data. I started with the base model of Linear Regression, which did not do entirely well with my data. Performing slightly better on the testing data, it scored an R^2 score of .7411 on training and .7810 on testing. In an attempt to find a better model, I decided to test Random Forest Regressor on my data, which ended up performing better than the linear regression. It scored an R^2 score of .9743 on testing and .8633 on training. While this shows the slighting hint of overfitting, this is still in the reasonable range to assume that the model is good. I also tried XGBoost, however it performed worse than random forests. Thus, this was the model I chose for my data.

The key insights I received from my model is that the three features that truly affect insurance charge are age, BMI, and smoking, with the order being smoking first, then age, then BMI. Therefore, since age is unchangeable, the model recommends that someone can focus on lowering their body fat and to quit smoking to lower their insurance charges.

For this model, there were a few limitations. Firstly, I wish there was a column that showed what insurance people have that covers certain costs, as this could also explain lower charges. I wish the location column was more specific, possibly done by state instead of region. Lastly, I wish there was a column that included if people had prior health issues or a family

history of health problems, as this definitely has an effect on healthcare charges. For future improvements, I hope to find a better data set that includes these aspects.