Final Project Write-Up

DSC550 Fall 2022

Karlie Schwartzwald

Bellevue University

In this analysis I will be discussing the creation of a model that predicts the medical insurance premium of an individual based on factors related to the physical health and condition of the individual. Essentially, I have a database of patients and their corresponding health information like their age, height, number of major surgeries, whether they have diabetes, and so on. I plan to use these attributes to make a model that can make a prediction about the cost of the medical premium for an individual. The model I will be using is a random forest regressor, and I will also be doing hyperparameter tuning so I can optimize the random forest regressor for best results.

I can think of a variety of reasons why this could be helpful to an individual or to a company in a business setting. For a business example, consider those who work for the Health Insurance Marketplace in the United States government. This organization provides a comparison tool to help users shop for health insurance. In order to give users an accurate idea of how much they might pay for insurance premiums at a particular insurance company, we can build a model to predict that cost. This allows users to voluntarily enter anonymized health data to predict their cost and accurately compare insurance products.

Suppose you had a company that wanted to be price competative with another health insurance provider. If we had the ability to predict the premium for an individual at the other insurance company, we could then offer them a slightly smaller premium to undercut the competition and win the customers’ business. This would be a valuable tool for insurance salesman to use in the field.

My data was obtained from a database on Kaggle which was curated just for this sort of project that data science students work on. I think that what makes my project unique is my approach with the random forest regressor and hyperparameter tuning. The variables ‘Age’, ‘Height’, ‘Weight’ and 'NumberOfMajorSurgeries' are numeric variables, while 'Diabetes', 'BloodPressureProblems', 'AnyTransplants', 'AnyChronicDiseases', 'KnownAllergies', and 'HistoryOfCancerInFamily' are binary categorical variables. The target feature is ‘PremiumPrice’ which is a numerical variable as well. This data was provided by the individuals themselves and was given voluntarily. The URL can be found here: <https://www.kaggle.com/datasets/tejashvi14/medical-insurance-premium-prediction> .

In summary of Milestone 1, we were hoping to find some features that correlated with ‘PremiumPrice’ so that we could have some hope that modeling the ‘PremiumPrice’ with the given features was possible. Our exploratory data analysis leads us to believe that there are in fact some features correlated with ‘PremiumPrice’, notably ‘Age’ has a strong correlation. See the correlation heat map below (fig 1) for a visualization regarding correlation coefficients in our feature set.

Figure 1. A correlation heat map of the features in this dataset.

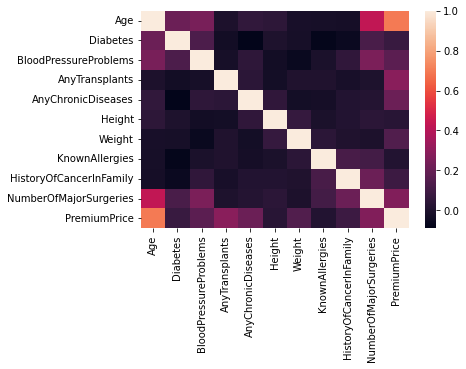
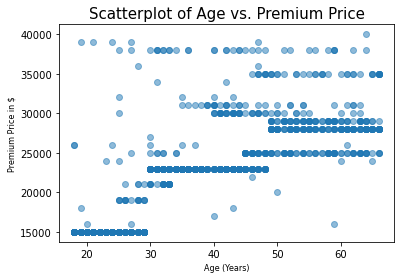


Figure 2. Scatterplot between the Premium Price and the Age of the individual

In figure 2 we can see more closely the relationship between ‘Age’ and ‘PremiumPrice’. The correlation is even obvious from the scatterplot alone. Since this was the only continuous variable that showed a correlation between itself and ‘PremiumPrice’, it was the only scatterplot I chose to include. In Milestone 2, I created the new feature ‘BMI’ from weight and height. I confirmed that there was no missing data, so no need to drop any rows or columns. Because of the variability in the units of each of the features, I decided to use a standard scalar to scale the features.

In Milestone 3 I built the model and evaluated its performance. I started with a grid search that performed hyperparameter tuning on my random forest regressor. I then used those optimized parameters to fit the model to the data. I used Out-of-bag errors to evaluate the performance of the model and got a score of approximately 0.8. Regarding any changes that were made between milestone three and now, there have been none. I received perfect marks on those milestones and no feedback about how to change or improve was suggested.

In conclusion, I was able to build a model that predicted “close” to the correct medical insurance premium on our given testing set approximately 80% of the time. I am still slightly confused about how to translate the out-of-bag error into a regression model understanding. This model is probably not ready to be deployed, given the relatively low accuracy of the model. My recommendations would be to do hyperparameter tuning on a larger variety of hyperparameters, and to also compare the accuracy of this model to the accuracy of a basic linear regression model so that we can have a basic performance threshold. I think that a challenge to still be explored is how to really measure the accuracy of this model, since Out-of-bag errors are typically only used on classification problems. Furthermore, I would like to better understand what is considered a “correct” prediction using a random forest regressor rather than a random forest classifier.