Health Care Cost Prediction

Final Project Report for Data 1030, Fall 2021 at Brown University

Yifei Song

December 7, 2021

1 Introduction

We spend part of our money on health care every year, so I think it is necessary to collect some data to predict the annual expenses according to each person's own situation, and then purchase appropriate medical insurance or save enough money to deal with the possible expenses according to the conclusion.

This project attempts to create a regression model from the health condition of an individual to predict his or her annual health care cost. The dataset used for this project came from the Medical Cost Personal Datasets in Kaggle[Cho], which includes age, sex, bmi, children, smoker, region and charges. Through 6 features of 1338 individuals, I want to explore which factors are closely related to our target variable 'charges', and whether we can predict some person's medical insurance expenditure once we obtain these factors. This will greatly help the insurance company price medical insurance. Furthermore, individuals will have the benefit of understanding the medical expenses he will pay as well.

Several authors have uploaded their study to kaggle that use this dataset to establish various models to predict the health care cost. Shubham Shrimant used Decision Tree Regression to do the prediction based on all 6 features. The model he made performed well for accuracy, with a r2 score of 0.69.[Shr]. Praviin Jaiswal used the sequential model in Keras multiple times with different hyper-parameters and obtained a final lose and mae as 1264.5 which is also pretty good.[Jai]. Different from these examples, I want to explore a different method to build different models to achieve this goal, in order to get a higher accuracy score.

Here's the github link for my project: github.

2 Exploratory Data Analysis

The following section contains several of the images I created during the exploratory data analysis:

The figure 1 is a catplot that shows the number of smokers among patients of different genders. From the figure, we can know that the total number of smokers is about 260, and the total number of non-smokers is about 1050, so the total number of smokers is much smaller than that of non-smokers, which is about 25% in percentage. In addition, for different genders, there are about 40 more male smokers than female smokers, which means that male smokers are about 36% more than female smokers, and male non-smokers are about 10% less than female non-smokers.

The figure 2 is the heatmap that shows the correlation between all features. Except for the target value 'charge', the correlation between each feature is not very large. There are some relationships between bmi and age, which may be due to the different physical conditions of young people and old people. Adolescents are in the developmental stage, and most of them will have a healthy bmi value, while the physical fitness of the elderly is declining, and the bmi data may not be very healthy. In addition, there are some correlations between bmi and region, which may be due to the different dietary customs in different regions that lead to different physical conditions of local people. Since these correlations are not very large, the correlation between these features will be ignored in the following studies in this article. The correlation between features and target value 'charge' will be considered emphatically. As we can see from the heatmap, the influence of gender on charges seem to be negligible. Instead, bmi, age and smoking are closely related to charges, especially smoking.

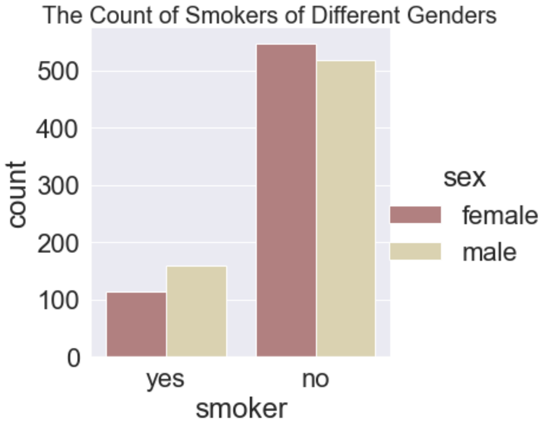

Figure 1: The count of smokers of different genders

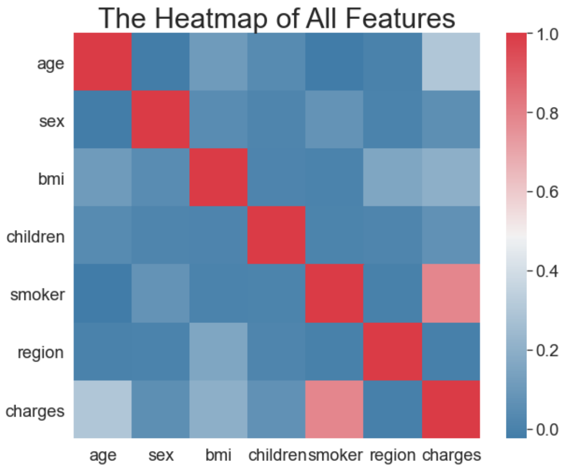

Figure 2: The Heatmap of All Features

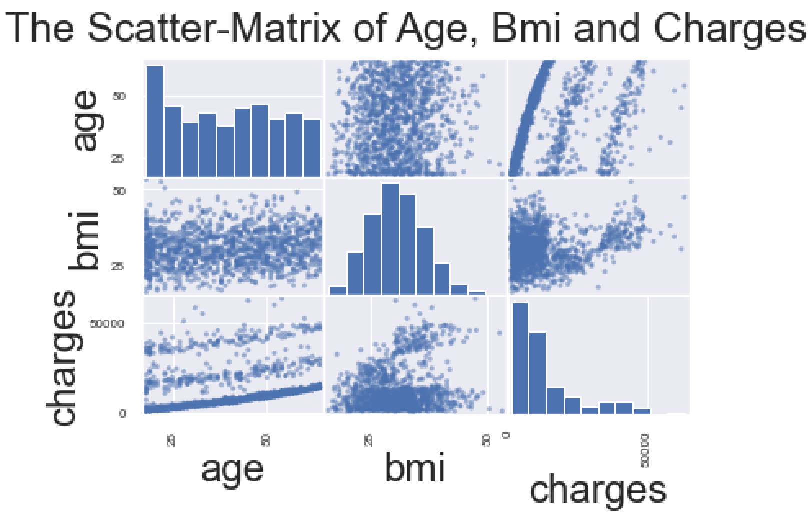

Figure 3: This scatter matrix is built from 2 features: age, bmi and our target variable charges. From these 9 plots, we can see clearly the relations between these variables.

Figure 3 is the scatter-matrix of age, bmi and charge, which contains 9 different scatter plots between these three factors. some of them are just random, but some of them are really interesting. For example, we can see from the graph of bmi that the bmi value generally follows a Gaussian distribution. From the charge chart, it can be seen that the overall average charge of patients is more than 10,000, but some extreme values exceed 50,000. In addition, it can be seen from the graphs of age and charge that the distribution has three linear curves, which is very interesting.

3 Methods

3.1 Data Splitting and Preprocessing

In the data preprocessing part, given each observation represents the health condition of a single individual, the data is assumed to be independent and identically distributed with no group structure or time-series property. Thus, we can just use the train-test-split to split the dataset into training data and testing data. Our dataset is not very big and larger training data would increase the accuracy of the model, so I chose to split the dataset into 80% for training and 20% for testing. In previous steps, the ordinary data sex, smoker and region data have been transferred using LabelEncoder in order to make it clear to analyse during the EDA part. While in the preprocessing part, we need to do more encoding on these features for better analyse. Since the age, bmi and children features are reasonably bounded, which are suitable for the MinMaxEncoder, and sex, smoker and region are categorical features, which are suitable for ordinalEncoder, we can use the pipeline to fit and transfer the 6 features.

3.2 ML Model Selection

After the splitting and preprocessing part, eight different machine learning models were trained and compared: a linear regression model without any regularization, a Lasso regression model, a Ridge regression model, an ElasticNet regression model, a random forest regression model, a support vector machine regressior, a K-nearest neighbors regressor, and an XGBoost regressor. Except for the linear regression model without regularization, all models were hyperparameter tuned using a k-fold grid search method to find the optimal parameter combination for each model. This process was repeated on 10 different random states for 10 different splits. Below are the parameters tuned and values tried for each model:

Model	Parameters
Lasso	**alpha**: np.logspace(-7,0,29)
Ridge	**alpha**: np.logspace(-10,0,51)
ElasticNet	**alpha**: np.logspace(-10,0,51)
RF	**max_depth**: 1, 2, 3, 4, 5; **n_estimators**: 20,25,30,35,40
SVR	**C**: 0.1, 1, 10, 100, 1000; **gamma**: 1,0.1,0.01,0.001
KNN	**n_neighbors**: 1, 10, 30, 50
XGB	**learning_rate**: 0.03,0.05,0.07; **max_depth**: 2,3,4

After tuning, each grid search's best model parameters were extracted and used for comparison on accuracy score. Since The RMSE scores of these models are too large for comparison, so we choose r2 scores, which captures the fraction of response variance captured by the regression and tend to give better picture of quality of regression model.

4 Results

4.1 Model Evaluation

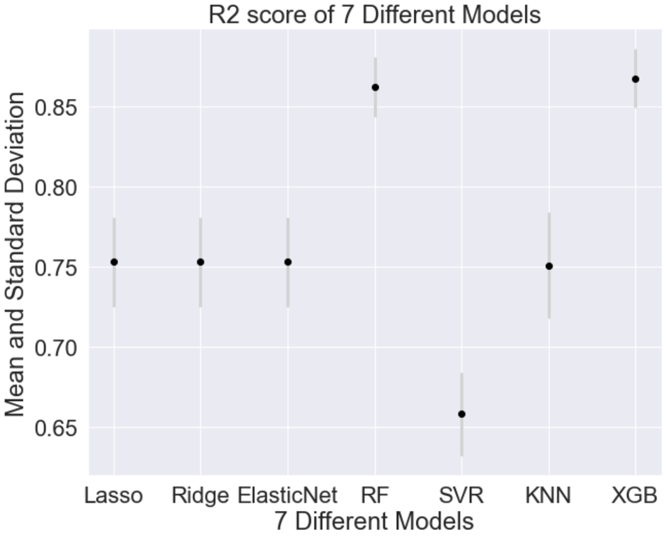

Figure 4: The mean and standard deviation of the r2 scores for 7 different models.

As we can see from the figure 4, which is the average accuracy scores for each model across the ten random states, the r2 scores of the XGBoost regression model and the RandomForests regression model are much higher than the other 4 models. Since the score of the XGBoost regression model is slightly higher than that of the RandomForests regression model, we can conclude that the XGBoost regression model had the highest test set performance and was chosen as the model of choice.

After choosing the best model and hyperparameter choice, the model was retrained on new splits over 50 new random states. For each split, 80% of the data was allocated to training and 20% was allocated to testing. For each random state, the accuracy of our best model and baseline accuracy score were recorded.

Over the 100 random states, the baseline models returned an average accuracy score of 0.7474 with a standard deviation of 0.0335. In comparison, the trained models returned an average accuracy score of 0.8591 with a standard deviation of 0.0276. The trained models achieved an accuracy that is $\frac{10}{3}$ standard deviations above baseline. Similarly, the baseline model's accuracy was 4 standard deviations below the average of the trained models.

4.2 Model Interpretation

Global Feature importance for the model was calculated by 3 ways: a permutation test over 10 shuffles for each random state, SHAP values and mean absolute value of the SHAP values. Below are the results of these evaluations:

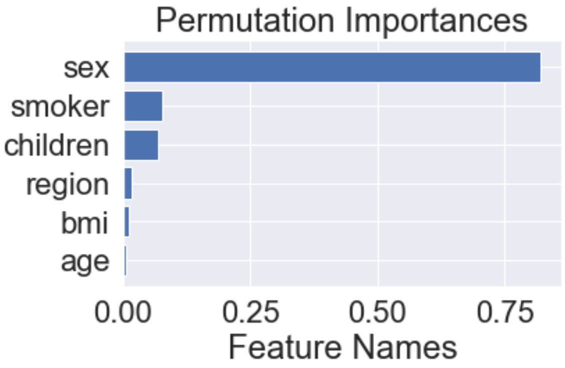

Figure 5: Permutation Importance

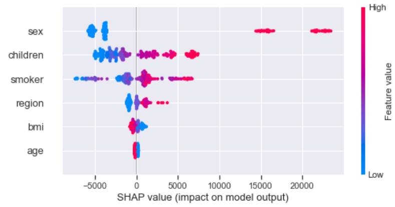

Figure 6: SHAP values

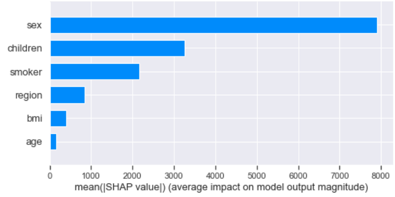

Figure 7: mean absolute value of the SHAP values

From the result, the most important feature is sex, which is much higher than the other 5 features and the age, bmi and region feature have little effect on the model output. This result means that the health care cost prediction result will be highly affected by the sex of the patients. The number of children the patients have and the patients smoked or not would also influenced the model output. This is unexpected in a sense, because in the previous coefficient heatmap, we can see that whether patients smoke or not is the most relevant to the charges, and the feature 'age' and 'bmi' is also highly correlated with the health care cost.

Two specific sample was selected to check the local feature importance by calculating the SHAP values in figure 8.

As we can see from the 2 specific patients, the baseline - the average predicted charge - is about 13160. The first patient has a high predicted charge of 21605.74. In this example, the patient is

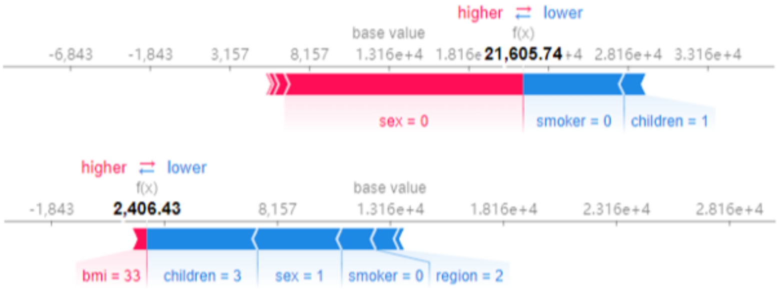

Figure 8: SHAP values to explain the predicted charges of two patients.

female mostly increase her predicted charge, and charge decreasing effects such as non-smoker and less children cannot fully offset by the increasing effect sex. The second patient has a low predicted charge of 2406.43. Despite his bmi somewhat increase the charge a little, his gender, number of children, his living region and that he didn't smoke highly decrease his predicted charge.

5 Outlook

Although the accuracy of the XGB regression model used on this data set is very high, there are still many technical problems that can be improved. For example, you can consider tuning more parameters in the hyper parameter tuning stage. Since the XGB regression model has a lot of adjustable parameters, which may further improve the accuracy of the model. Furthermore, more regression models can be tested in gridsearchCV to see if there are better models. In addition, it is unreasonable that the most important feature is 'sex', which may be because the data set selected for model construction is not large enough and the features of the data set are not enough. It might be helpful to add more observations and features, such as the regularity of work and rest and the diet of patients, etc., which might increase the interpretability of the model and make the practical application value of the model higher.

References

[Cho] Miri Choi. Medical cost personal datasets. https://www.kaggle.com/mirichoi0218/insurance.

[Jai] Praviin Jaiswal. Regression with neural network. https://www.kaggle.com/praviinjaiswal/regression-with-neural-network.

[Shr] Shubham Shrimant. Decision tree regression. https://www.kaggle.com/shubhamshrimant/decision-tree-regression.