Medical Insurance Cost prediction using Machine Learning: an Individual project for the Workshop-I "Actuarial Science" course

Amir H. Talebi

January 14, 2024

1 Introduction

Insurance provides financial protection/coverage against potential risks, thanks to actuarial sciences which provide the analytical foundation for managing risks and losses. The effectiveness of insurance policy terms in an insurance sector can be improved by the utilization of machine learning methods. In this project, we have used medical data to forecast the individual medical costs billed by health insurance for various categories of people. We have used linear regression, XGBoost regression, Random Forest regression models and finally a feed forward neural network. These models are trained using the provided dataset. We have also found the RMSE value for each one of them and have reported the values.

Attribute	Description
Age	Age of primary beneficiary
Sex	Insurance contractor gender, female, male
BMI	Body mass index
Children	Number of children covered by health insurance / Number of dependents
Smoker	Smoking
Region	The beneficiary's residential area in the USA
Charges	Individual medical costs billed by health insurance (out-of-pocket costs)

Table 1: Seven attributes.

2 Problem statement

Digital health and health insurance problems have doubled globally during the last few years. In this trend, the health insurance has faced with two obstacles in industrialized nations: (1) Growing health care costs and (2) Increase in the amonut of people with no coverage! Solitary health insurance plays an important role in the healthcare system, especially for people with rare diseases. For this class of people, the medical and preventive insurance can help to reduce the treatment expenses. The world in which we are living has been filled with unknown events and there is always potential for disease, death and loss of assets. Gladly, the financial industry has developed some methods to protect and shield businesses from them.

In the world of healthcare, where a pivotal component is medical insurance, advancements in machine learning algorithms have been important for predictive analysis. With the help of machine learning algorithms, we can make predictions, but it's important to note that the accuracy of these predictions may vary. Also note that, a substantial amount of money used for medical expenses is

contributed directly by the patients themselves, rather than being covered entirely by external sources such as insurance. This study goes in the topic of claim prediction employing statistical methods, ML techniques and deep learning.

3 Descriptive Analysis of the Dataset

Dataset has been already published and is accesible online [1]. For the modeling part, we have split the dataset into two parts: test and train (20 and 80 percent, respectively) and it has 7 attributes as Table 1 shows. The goal is to use this dataset in a model and forecast medical insurance costs (per year) with the help of train data. Test data can help for the assessment of the model. Based on the calculations for the summary of numerical features (age, MBI, number of children and charges), there is a huge difference between the third quartile and the Max value in "charges" column. Maybe a patient had a rare disease.

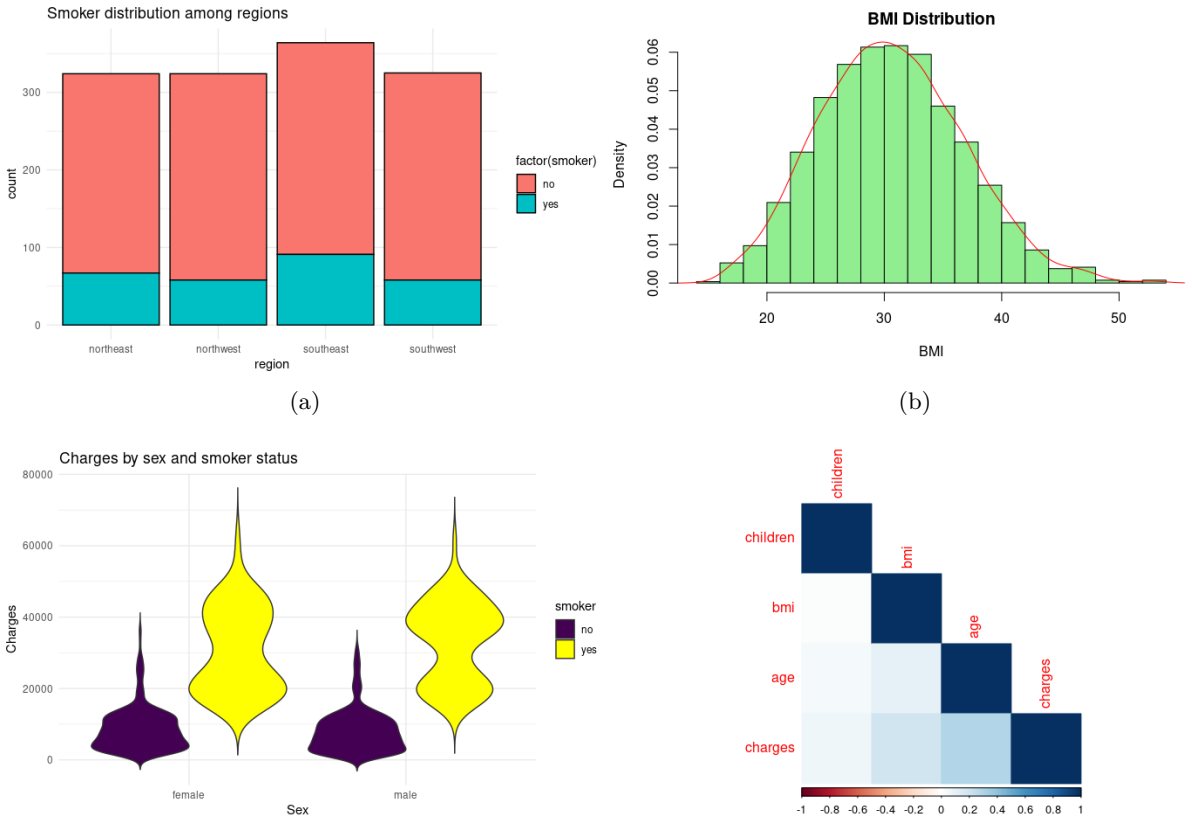


Figure 1: Different types of plots which are helpful in this study.

The original dataset has 1338 rows and 7 columns. The "charges" variable with `datatype = float` is the target variable which needs to be predicted based on the medical records of the patients. The age is between 18-64 years old and the majority of them are male. Few people have more than three children and the majority of them have a MBI between 26-34 kg/m^2. Four main regions are considered: northeast, northwest, southeast, and southwest. The largest concentratoin of smokers is in southeast.

Fig. 1 shows some selected data visualizations of the features. Only the frequncy distributoin of BMI value is presented, Fig. 1 (b). Note that a whole range of plots can be found in the source code file or the `html` file, other information in this report are based on the calculations we have made. It is also possible to get some intution about the statistical metrics. It is shown that the individual medical

[1] data are from `https://www.kaggle.com/datasets/mirichoi0218/insurance/data`.

costs billed by health insurance (i.e. charges) is skewed. In this column we have more outliers than BMI column. This shows that the insurance companies need to have reserve funds for the people with higher values of costs in the hospital or doctors.

4 Data Pre-processing and Cleaning

In the original dataset, three features are numerical and three are categorical. We have used label encoding to give those categories qualities numerical lables and make the machine learning part possible. There is no null value, but, there is one repeated row that we removed. After that, the number of unique rows becomes 1337. We didn't apply feature scaling for XGBoost and Random Forest regressions since these ensemble methods, relying on decision trees, are robust to the scale of features. However, when we specifically scale our data for the linear regression model, `lm`, the amount of residual standard error is lower than the case with no scaling [2].

In Fig. 1 (a), we can see that most of the customers are from the southeast region, however, differences between number of customers among all regions are not that high. We also infer that the southeast region has the highest mean "charges" while mean "charges" for the rest are not much of a difference from each other. Additionally, there is almost equal distribution of female and male on the dataset. Mean for the "charges" of male gender is higher than for female by around 1400. Given that the southeast region has the most number of customers, this region also has the most number of males and females. For the rest, number of males and females are almost the same. The mean "charges" of the males are higher for all regions except the northwest region while southeast region has the highest mean "charges".

There are more non-smokers than smokers and mean "charges" for smokers is significantly higher than those of non-smokers. Also, based on the calculations, being a smoker will incur higher "charges" than being a non-smoker. Southeast region has the most number of smokers which may be why the mean "charges" for southeast is the highest. There are more smokers in males than in females which may be why mean "charges" for males are slightly higher than in females. As we can see based on the violin plots, Fig. 1 (c), being a smoker, either female or male, is highly correlated to having higher insurance "charges". Fig. 1 (d) shows the correlation matrix between numerical values, we can observe that age has the highest correlation with "charges", although not that high. Additionally, numeric values are not highly correlated with each other.

5 Modelling, Description and Prediction

With the help of available models, the goal is to study and forecast the insurance costs based on a variety of factors like age, sex, number of children, BMI [3], location of living and smoking situation.

Four models are used in this project to calculate the amount of charges. We have calculated the RMSE for each model and have compared them. One of the important issues in this dataset that impacts on the value of RMSE, is the presence of outliers in BMI and charges columns (numerical features). To do modeling, we need to use label encoding for the categorical features. We have further proceed and detected the outliers based on their z-score (which measures how many standard deviations a data point is far from the mean). As we can see in the programming part, the percentage of outliers in boht BMI and charges columns is less than 0.6 percent. We cannot define a theory or distribution for them with this few amount. One way is to use Trimming, otherwise the amount of RMSE would be larger. Note that, the given outlier list for both BMI and charges is for the data that are not in $+3\sigma$, both BMI and charges are positive quantities. We could also use Winsorizing which involves replacing the outliers with the nearest non-outlier values.

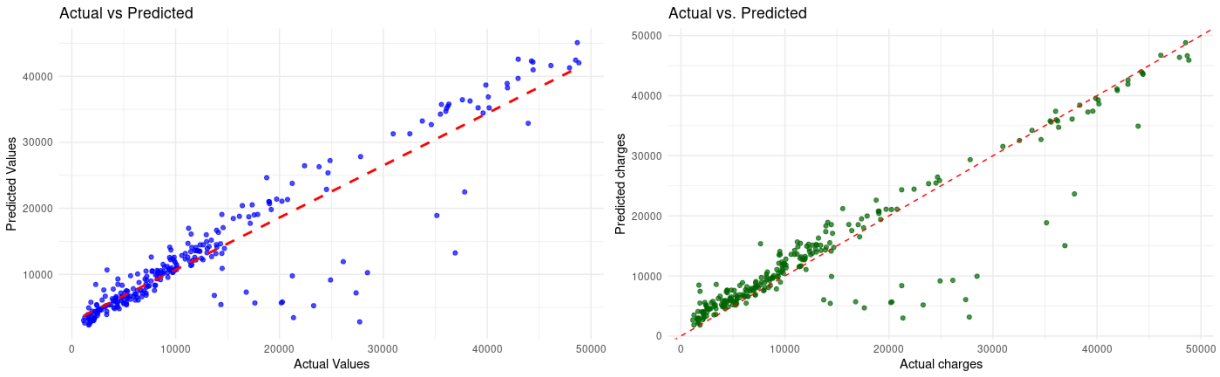
Based on the output for the linear regression model, the F-statistic is equal to 668.2 and p-value is very close to zero, indicating that the model is statistically significant (it is not random effect) and at least one predictor variable has a significant relationship with the response variable. So, we conclude

[2] Scaling here involves centering (subtracting the mean) and standardizing (dividing by the standard deviation) each numeric column.

[3] BMI: a good gauge of risk for diseases that can occur with more body fat, in units of $\frac{kg}{m^2}$.

that not all the coefficients are zero. Also, the coefficients for 'age', 'BMI', 'children', 'smoker', and 'region' are statistically significant (asterisks). Fig. 2 shows the results for the Random forest and XGBoost regression models as two supervised learning algorithms. Random forest regression (similar to XGBoost) uses ensemble learning approach for regression. In this case the predictions from multiple machine learning algorithms are combined to make a more accurate prediction than a single model. At first, the model picks at random k data points and builds a decision tree to these k data points. We set N to be the number of trees and repeat the previous steps. For a new data point, we get the result of prediction from all N trees and average across all the predicted values. For the case of XGBoost, trees are build sequentially, with new tree correcting the errors of the previous ones. Indeed, XGBoost assumes that its base learners are consistently poor at capturing the remaining errors. In this way, when aggregating all predictions, the inaccuracies neutralize each other, and the more accurate predictions gather to produce a final high-quality prediction.

Based on Fig. 2, for our situation and problem, XGBoost regression shows the best regression model. As we can see the trained model can better predict the test data and the data are closer to the dotted red line than Random forest. The ideal case is that all the point to be on top of the base line. The amonut of RMSE for the case of linear regression, Random forest and XGBoost is 5807.57, 4714.73 and 4567.69, respectively. Finally, in order to leverage the power of deep learning models in our computation, we have trained a feed forward neural network with two hidden layers with 8 and 4 neurons. First input layer has 16 neurons with `relu` activation function, while the last layer has just one neuron and the activation function is `linear`. `Adam` optimization algorithm with a loss metric `mse` are considered. The number of `epochs` is set to 100 and then `batch-size` equal to 32. The amount of RMSE for this type of NN model has been obtained to be 16686,39. Note that for the case of NN, it is much better to feed the NN with scaled data.



(a) (b)

Figure 2: Comparing the predicted and actual values for RF and XGBoost regression models. Dashed line represents a perfect match ruler.

6 Summary

In the provided dataset for the medical cost prediction of insured people, males have higher insurance "charges" and being a smoker will incur higher "charges" as well. Having upto 3 children incurs higher "charges" than having 4 or 5 children. And finally, obese people (with higher BMI values) pay higher "charges" than people with lower BMI. Living in different regions does not lead to different results necessarily. We have also looked at forecasting the amount of "charges" with the help of ML regression models. Machine learning methods could significantly reduce efforts in price analysis since they can compute costs quickly while doing so would take a person a long time. We have used linear regression, Random forest, XGBoost and neural network to train models and used them for the forecasting. Based on the values of RMSE and also Fig. 2, up to now XGBoost works better on this dataset.