Report

Medical Insurance Cost Prediction

Elnaz Khaveh

Abstract. This paper is a report for the project "Insurance Cost Prediction" which is done to predict the insurance cost, from the "Insurance" dataset. The prediction was done by training Linear Models (LM) and Random Forest (RF) to check which one performs better. The whole project was done in R. To compare these 2 models, Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and R Squared were used. The results indicate that Random Forest outperforms the other model.

1. Introduction

Living in a world full of unexpected disasters, makes people and all their properties exposed to a lot of risks and dangers, such as illnesses, accidents, natural disasters, theft etc. One of the largest impacts of these happenings is the financial loss. Insurance is a policy that decreases the expenses of these catastrophes in different areas. For example, health insurance, car insurance, etc. Premium is the amount that should be paid by people to the insurance companies on a regular basis, and it differs from person to person based on some of their characteristics which affect the probability of occurring the risks, because if all people pay the same amount of premium, the insurance company would lose the low-risk policy holders and only the high-risk ones will be kept for them. One factor in predicting these premiums is looking at the costs of the insured person, for example how much charges is billed by the insurance company for the policy holder, to predict the future costs and calculate the corresponding premium.

The remainder of the report is structured as what follows:

In Section 2, Materials and Methods, the dataset and the methodologies used for this purpose are explained, as well as some explanatory data analysis and data visualizations.

Section 3 discusses the results gained from the implemented methods. Finally, section 4, concludes the main findings of these work.

2. Materials and Methods

2.1 Preliminary definitions

Linear Model (LM): A Linear Model is a model that specifies the linear relationship between a dependent variable and several independent variables:

$$y = a_0 + a_1 x_1 + a_2 x_2 + \ldots + a_n x_n$$

Where y is the dependent variable, $\{x_i\}$ are independent variables, and $\{a_i\}$ are parameters of the model. It is used in machine learning prediction tasks which can be used for predicting the unseen future data points.

Random Forest: To know the definition of Random Forest, first we should know what decision trees are. Decision Tree is a supervised learning algorithm that can be used for both classification and

regression tasks. It is a tree-like structure in which each internal node denotes a test on a variable, each branch shows an outcome for a test, and each leaf shows a class that the input belongs to or a continuous output. Random Forests are a committee of decision trees.

K-Fold Cross Validation: A K-Fold Cross Validation is a resampling technique to assess machine learning models. The parameter K refers to the number of groups the data is split into. In this procedure, the dataset is shuffled randomly, then 1/K of the groups is taken as test set and the rest as train set, the model is fitted on the train set and evaluated on the test set, it continues until all the K groups are taken as test set once.

2.2 Dataset

The dataset used for this project was the "Insurance" dataset from Kaggle, but it is originally for the book "Machine Learning with R" by Brett Lantz. It consists of 1338 rows and 7 columns. The predictors are as the following:

- Age: Age of primary beneficiary
- Sex: Insurance contractor gender
- BMI: Body Mass Index of the policy holders
- Children: Number of children covered by the health insurance
- Smoker: Smoking status
- Region: The beneficiary's residential area in US

The only response variable is "charges" which is the individual medical cost billed by the health insurance and our goal is to predict this.

The following plots show some information about the correlations of the independent variables with the dependent one, also the independent variables with each other.

Figure 1. The effect of sex, smoking, and age on charges

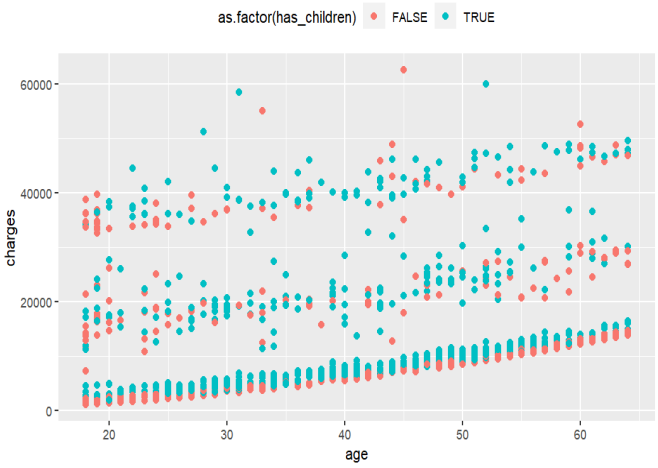

Figure 3. The effect of having children and age on charges

Figure 2. The effect of region and age on charges

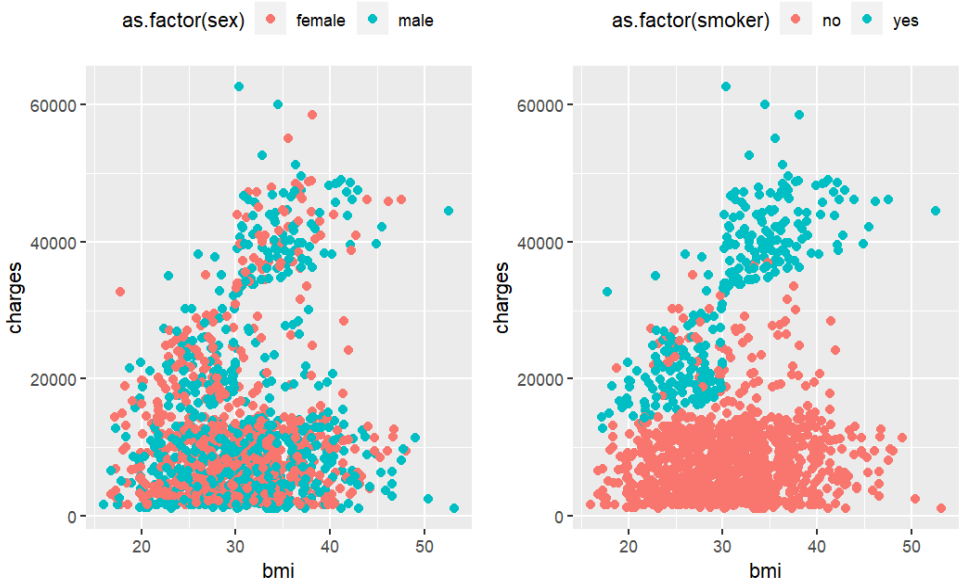

Figure 4. The effect of sex, smoking, and bmi on charges

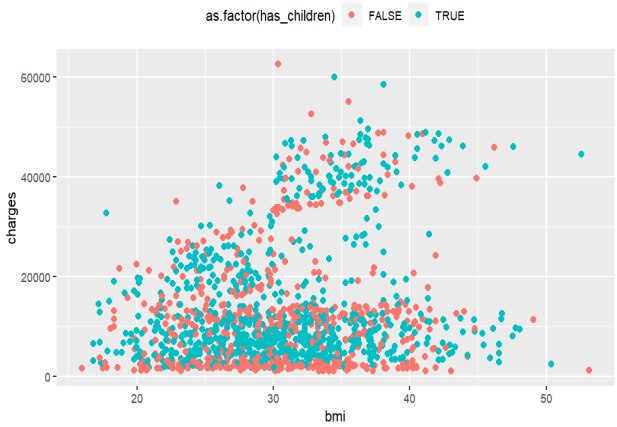

Figure 5. The effect of having children and bmi on charges

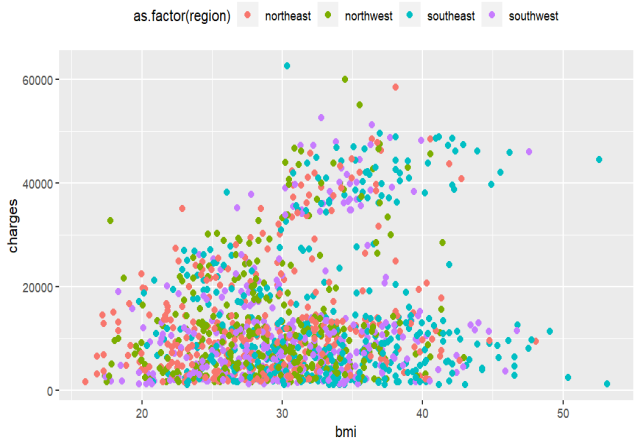

Figure 6. The effect of region ang bmi on charges

Figure 7. Correlation heatmap

In figure 1, the correlation between 2 discrete variables and one continuous variable and the target variable is shown. It is evident that smoking has a positive correlation with the charges, meaning that smokers have higher charges than non-smokers. Also, the older the people, the more their costs, but there is no correlation between the age and smoking, or age and sex. In figures 2 and 3, it is obvious that there is no significant correlation between having children and region with charges. The same can be concluded from the other plots for the other continuous predictor, bmi, the only difference is that bmi does not affect the amount of charge unlike age. In figure 6, the correlation heatmap indicates the severity of correlations between the attributes.

2.3 Pre-Processing

Outlier Detection: Since there are no missing values in the dataset, so there is no need to fill them, and we go directly to detect the outliers. Outlier detection is done to ensure that the trained model is not skewed and can generalize well, so it is an important step in pre-processing. To do that the function "rosnerTest" from "EnvStats" library is used. It is a method which is used when we have a large sample size ($n > 20$) and it has 2 advantages over other outlier detection techniques: One is that it detects several outliers at once, and the other is that it avoids the masking problem (when an outlier which is close to another outlier remains undetected). As a result, there is no outlier in bmi, but there is one in charges so it can be removed since it is the only one.

Normalization: This step helps the models to predict more accurate since the range of the data points is kept in a particular scale. To normalize the data in this work, log-transformation is used on the target variable because the range of the costs were so wide. Log-Transformation can make a more skewed distribution, less skewed. In other words, it will make the distribution of the data to be more like the normal distribution and it enables us to use statistical properties for normal distribution. Figure 8 illustrates the distribution fitted to the variable charges before transformation and after it.

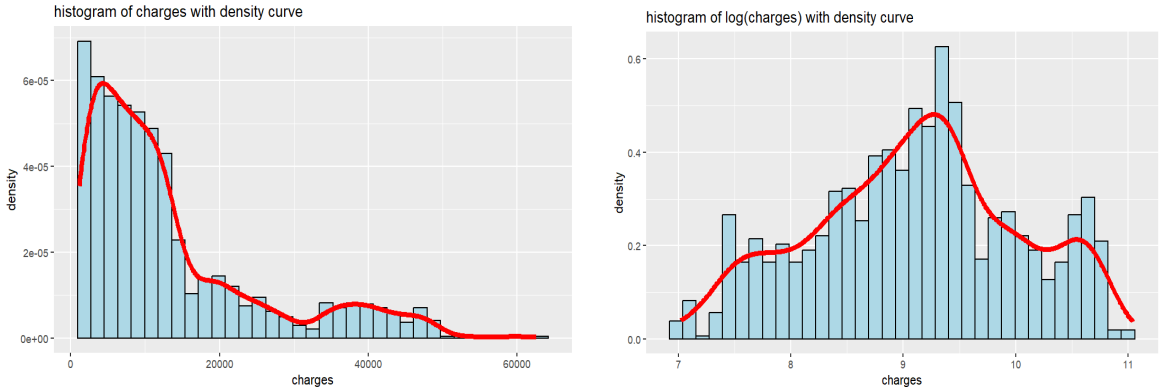

Figure 8. Distribution of charges before transformation and after transformation

2.4 Models

Firstly, the dataset is split to train and test sets, with 80% for the train and 20% for the test set. A 5-Fold Cross Validation is applied for both models.

LM: Before training the model, some feature selection should be done to have better performance. Forward stepwise criterion, which is used in this project, is a method that initiates the linear model with no variable and then adds the variables until all of them are added, then at each step AIC (Akaike Information Criterion) is calculated to have a comparison between the models. AIC is a statistical measure that helps in choosing the best model. This variable selection criteria shows that the model with all the variables is the best one since it has a lower AIC; In other words, all the attributes are significant. So, the LM is used for the target variable versus all the independent variables.

Random Forest: The Random Forest is trained for the same variables, charges versus all the other variables, to be able to compare the results of models with similar conditions.

3. Results

The performance metrics applied in this work, to evaluate the regression models, are RMSE (Root Mean Squared Error), MAE (Mean Absolute Error), and R Squared.

RMSE, shows the square root of the average squared difference between the predicted values and the actual values. While, MAE, indicates the mean absolute difference between these 2 values. Lower values for these two errors show a better model performance. However, R Squared shows how much variation of the dependent variable can be explained by the independent variables. So higher values are preferred. Table below illustrates these results:

	LM	**Random Forest**
RMSE	7537.859	4302.2004134
MAE	3845.617	2091.2961397
R Squared	0.5743797	0.8635256

As it is evidenced from the table, random forest advances the other model in performance.

In figure 9, plots show how distant are the true values and predicted values in LM and Random Forest Y-axis indicates the actual value and X-axis indicates the predicted value. (Note that the red line in both is showing x=y)

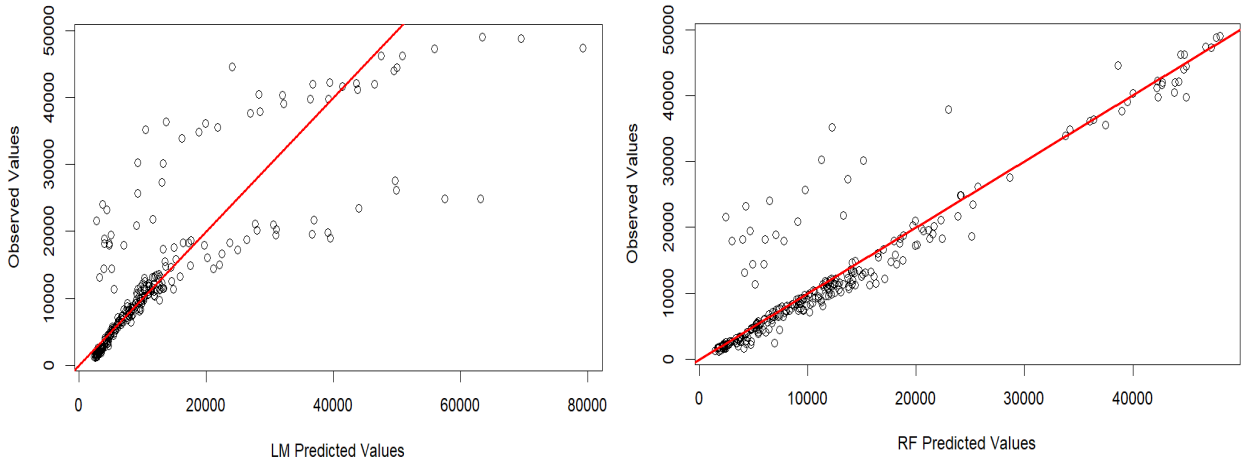

Figure 9. LM and RF predicted vs. actual values.

4. Conclusion and future work

Smoking has a significant correlation with the medical costs of a policy holder, age also is correlated with the costs. As discussed in the results, Random Forest outperformed LM in predicting the medical costs of the insurance. Some polynomial feature selections could have improved the results of both models. Also trying GLM and GLMNET and compare them with these two models can be done for the future projects.