Manasvi Ponaka
Workshop-I "Acturial science"
Master of Data Science
Winter semester '22 - '23
University of Luxembourg

Medical Insurance cost prediction

Introduction

In this project, we aim to address the challenging problem of insurance prediction. The insurance industry is subject to numerous uncertainties that must be taken into account at every stage of the process. Despite these difficulties, machine learning algorithms have been shown to have a high degree of accuracy in forecasting insurance costs. These algorithms can be used to predict an individual's risk of making an insurance claim, which in turn can aid insurance companies in better understanding and managing their financial risk. Moreover, they can also be used to predict the cost of an insurance claim and thus help insurance companies set more accurate prices for their products. Furthermore, they can assist underwriting decisions by analyzing data on individuals and businesses and make more accurate decisions on the best coverage or plan, or even improve customer service by providing personalized recommendations. Not only does this benefit the insurance companies but also individuals who can make informed decisions about their medical coverage. In this project, we model medical insurance cost data using various machine learning models and evaluate the performance using standard evaluation metrics such as root mean squared error and R-squared.

Exploratory Data Analysis

The data used in this research project was obtained from Kaggle[1]. It includes various characteristics of the policyholder such as age, sex, body mass index (BMI), number of children, smoking status, and medical charges billed by the insurance company. A sample of the dataset is presented in Table 1.

Table 1: Medical cost insurance dataset

age	sex	bmi	children	smoker	region	charges
19	female	27.9	0	yes	southwest	16885
18	male	33.77	1	no	southeast	1726
28	male	33	3	no	southeast	4449
33	male	22.7	0	no	northwest	21984
32	male	28.88	0	no	northwest	3867
31	female	25.74	0	no	southeast	3757

Table 2: Summary Medical cost insurance dataset (continued below)

age	sex	bmi	children	smoker	region	charges
Min. :18.00	female:662	Min. :15.96	Min. :0.000	no :1064	northeast:324	Min. : 1122
1st Qu.:27.00	male :676	1st Qu.:26.30	1st Qu.:0.000	yes: 274	northwest:325	1st Qu.: 4740
Median :39.00	NA	Median :30.40	Median :1.000	NA	southeast:364	Median : 9382
Mean :39.21	NA	Mean :30.66	Mean :1.095	NA	southwest:325	Mean :13270
3rd Qu.:51.00	NA	3rd Qu.:34.69	3rd Qu.:2.000	NA	NA	3rd Qu.:16640
Max. :64.00	NA	Max. :53.13	Max. :5.000	NA	NA	Max. :63770

The dataset was analyzed through a visual representation of the data, as illustrated in Fig 1. This analysis revealed that there is a balance in the data and no evident outliers were observed. The plots in Fig 2, which depict the relationship between age and medical charges, as well as between body mass index (BMI) and medical charges, illustrate the influence of smoking status on medical charges. When we draw a regression line for smokers and non-smokers on these plots, we can observe a clear pattern in the charges for each

group. This suggests that smoking may be acting as a confounding variable in the relationship between these factors and medical charges. Understanding interactions between features is important information which will aid in the modelling of the data. Another critical observation we can make is about the non-linear relationship between age, BMI with charges. The consequences of our observations are explained elaborately in further sections.

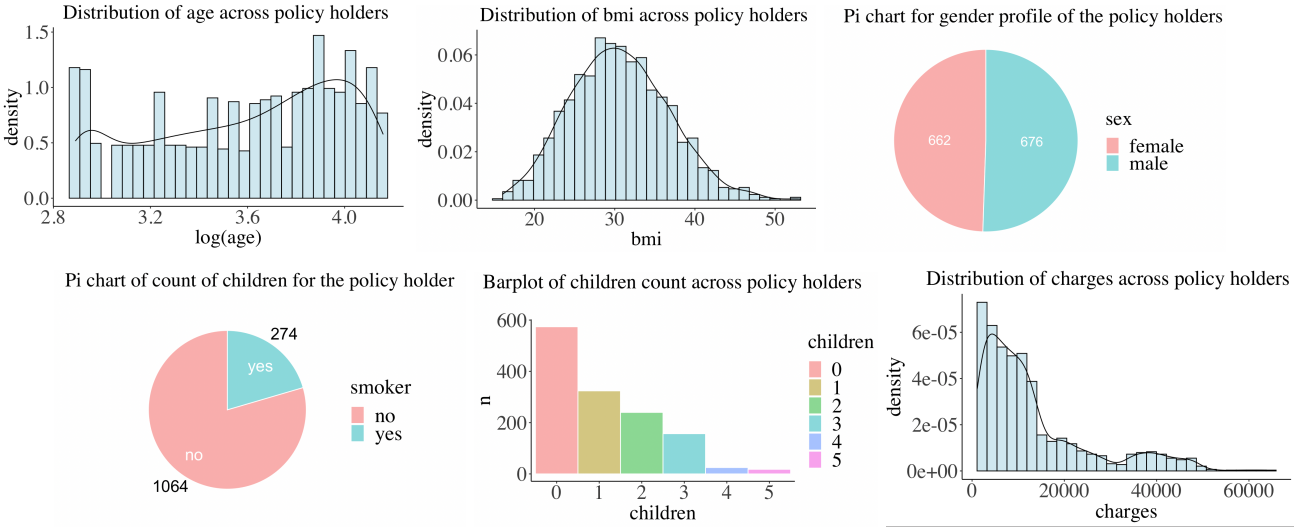

Fig 1: Visual analysis of features to check for imbalances

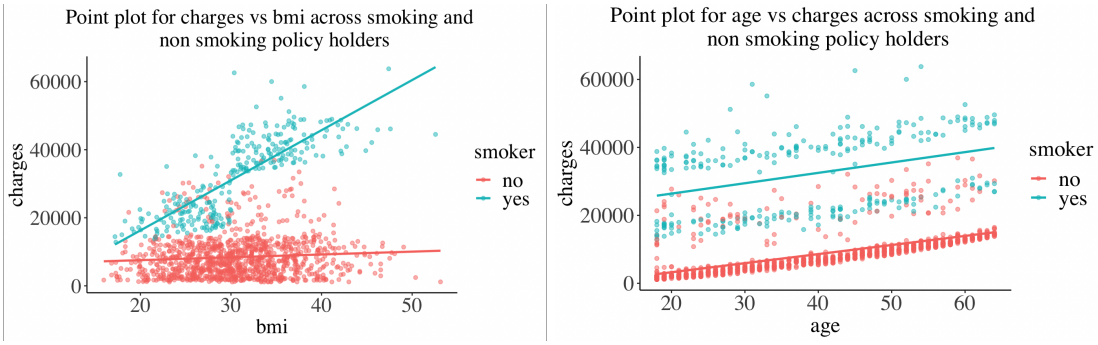

Fig 2: Interactions with explanatory variable smoking in scatter plot age ~ charges and BMI ~ charges

Methodology

We start with a simple linear regression model for our dataset $(Y_i, X_{i1}, \ldots, X_{ip})$, $i = 1, \ldots, n$. Y_i is a function of X_i and β, with representing an additive error term that may stand in for un-modelled determinants of Y_i or random statistical noise:

$$Y_i = f(X_i, \beta) + e_i$$

For linear modelling ,

$$Y_i = \beta_{i0} + \beta_{i1}X_{i1} + ..., \beta_{ip}X_{ip} + e_i$$

In a simple linear regression model (LM) , the relationship between the dependent variable (charges) and independent variables, $f(X_i, \beta)$, is assumed to be linear and must satisfy certain assumptions, including normality of errors and homoscedasticity (constant variance of errors). However, as depicted in Figure 1, the distribution of the dependent variable does not conform to a normal distribution. To address this issue, the logarithm of the charges was used as the dependent variable in the model. Despite this transformation, as shown in Figure 3, the normality assumption for residuals was not met, as confirmed by a Wilk-Shapiro test[1] (p-value < 0.05). Additionally, the presence of heteroscedasticity (unequal variances) in the residuals was indicated by a significant p-value (<0.05) in an OLS test[2] for heteroscedasticity. These violations of assumptions can be attributed to several factors such as outliers in the data, non-linearity between the dependent and independent variables, or the presence of an omitted variable that is correlated with the independent variables. Further investigation, such as checking for non-linearity or omitted variable, is needed to fully understand the underlying causes of these violations and to improve the model's accuracy.

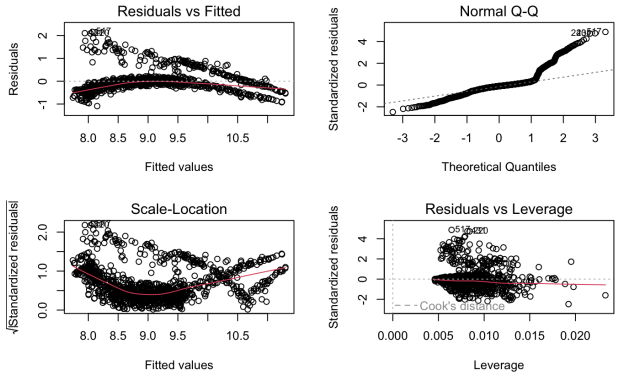

Fig 3: Residual analysis plots for a simple Linear Regression

In order to account for possible violations of linearity in the relationship between the dependent variable (charges) and the independent variables, we have developed a Non-linear model (NLM1) that includes interaction terms and non-linear transformations of the independent variables. The presence of omitted variable can be ruled out as all the independent variables in the dataset are used to model charges. This is supported by the observation in Figure 2, which illustrates a clear interaction between smoking status, age, and BMI on charges. To further examine this non-linear relationship, we have included interaction terms between smoking status and age and BMI in the model. Additionally, we have applied a logarithmic transformation to the age and BMI independent variables. The resulting model is detailed shown below.

$$\begin{aligned} Y_{i,charges} = \beta_{i,0} + \beta_{i,age}X_{i,age} + \beta_{i,bmi}X_{i,bmi} + \beta_{i,smoker}X_{i,smoker} + \beta_{i,children}X_{i,children} \\ +\beta_{i,region}X_{i,region} + \beta_{i,age*smoker}X_{i,age} * X_{i,smoker} + \beta_{i,bmi*smoker}X_{i,bmi} * X_{i,smoker} + \\ \beta_{i,log(age)}log(X_{i,age}) + \beta_{i,log(bmi)}log(X_{i,bmi}) + e_i \end{aligned}$$

[1] Wilk–shapiro test – H_0: the errors(residuals) are normal H_1: the errors are not normal.
[2] OLS test – H_0: the variance of errors is homogeneous H_1: the variance of error is not homogeneous.

After fitting the model with interaction parameters, we address the presence of outliers in the data by using the Mahalanobis distance method, which is a multivariate method that calculates the distance of an observation from the mean of the data, taking into account the covariance of the variables. The outliers are removed and the model(NLM2) is fit. The results of this model's performance will be shown in the next section .

Performance & Evaluation:

To evaluate the performance of our non-linear model in predicting medical insurance costs, we have chosen to compare it to several other popular non-linear modeling techniques. Specifically, we have chosen to compare our model to Random Forest (RF), Classification and Regression Trees (CART), k-Nearest Neighbors (KNN), and Support Vector Machines (SVM).

Random Fores (RF) is an ensemble learning method that utilizes multiple decision trees to improve the accuracy and stability of the model. It works by creating multiple decision trees using different subsets of the data and features. The final prediction is made by averaging the predictions of all the trees (for regression) or by taking a majority vote (for classification). It is known to handle complex data well.

Classification and Regression Trees (CART) is a decision tree-based model that is commonly used for both classification and regression tasks. It recursively splits the data into subsets based on the values of the independent variables, and the final prediction is made based on the majority class or the mean of the dependent variable in the final leaf node of the tree. It's a popular method as it can handle non-linear relationships and interactions between variables and it's easy to interpret.

k-Nearest Neighbors (KNN) is a non-parametric method that is used for both classification and regression. It works by finding the k-nearest observations to a new observation based on a distance metric and the final prediction is made based on the majority class or the mean of the dependent variable of the k-nearest observations. It's a common method as it can handle large datasets, it's easy to interpret and it's not affected by outliers.

Support Vector Machines (SVM) is a powerful technique that can be used for both classification and regression. it works by finding the hyperplane that best separates the data into different classes or predicts the value of the dependent variable. It's a popular method as it can handle high-dimensional datasets and non-linear relationships and its effective in finding complex boundary decision.Here we choose the RBF kernel to linear kernel because it can model non-linear decision boundaries and adapt to the density of the data, while the linear kernel can only model linear decision boundaries.

All of these models have their own strengths and weaknesses, and the choice of which model to use depends on the specific problem and the characteristics of the data. By comparing our model to these other non-linear models,, we can get a better sense of its performance and how it compares to other popular approaches.

Table 3: Performance of models on train data

model	parameters	rmse	r2	mae
KNN	k = 5	11193	0.18	7793
LM		8288	0.64	4191
CART	cp = 0.0605	5478	0.79	3910
NLM1		5249	0.8	2747
SVM	sigma=0.065 & C=1	4895	0.84	2629
RF	mtry=5	4599	0.85	2564
NLM2		4314	0.88	1996

Table 4: Performance of models on test data

rmse	r2	mae
10457	-2.09	7473
8315	0.63	4378
6314	0.61	4557
5572	0.77	2853
5020	0.77	2600
4985	0.79	2824
4057	0.87	1865

Table 3 & Table 4 : Tables reporting the performance of models and are arranged wrt rmse. params column indicates the hyperparameters of the respective mdoels

Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), R-squared (R^2) are performance metrics used to evaluate the performance the models. RMSE measures the difference between predicted and actual values, with lower values indicating a better fit of the model. MAE measures the average absolute difference between predicted and actual values, with lower values indicating a better fit of the model. R-squared measures the proportion of the variance in the dependent variable that is predictable from the independent variable(s), with high values indicating a good fit of the model.

$$RMSE = \sqrt{\frac{1}{n}\sum(\hat{Y}_i - Y_i)^2}$$

$$MAE = \frac{1}{n}\sum|\hat{Y}_i - Y_i|$$

$$R^2 = 1 - \frac{\sum(\hat{Y}_i - Y_i)^2}{\sum(\hat{Y}_i - \bar{Y}_i)^2}$$

Table 3 & Table 4 compares compared the performance of various machine learning models for predicting medical insurance costs. The models were evaluated using a 10-fold cross-validtion strategy. The results presented in Table 3 & Table 4 show that the non-linear model without outliers(NLM2) performed the best, with an R-squared score of 0.88 on the training data and 0.87 on the test data. This indicates that 88% of the variation in the dependent variable (charges-> medical insurance costs) can be explained by the independent variables included in the model. It is important to note that R-squared can be artificially inflated when a large number of independent variables are included in the model. However, this is not the case in our study, as the model includes only a limited number of independent variables. Additionally, the lower values of rmse

and mae for the non-linear model with outliers further support its superior performance. While the random forest and SVM models performed comparably, they did not perform as well as the non-linear model without outliers.

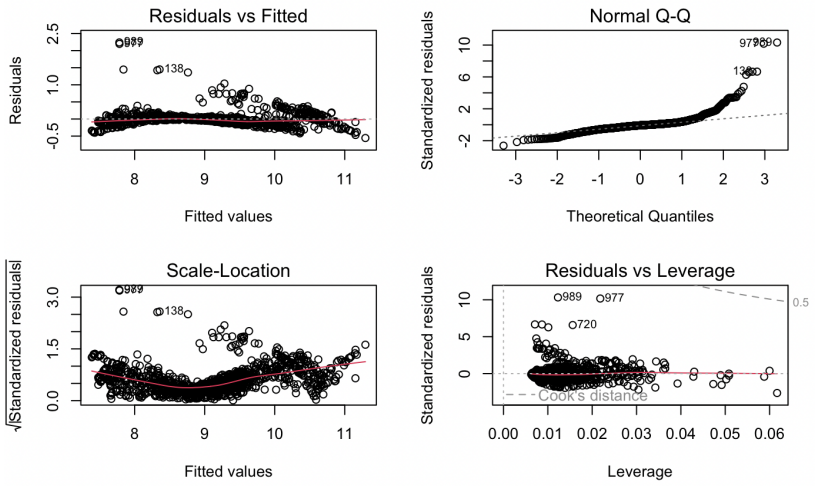

Fig 5: Residual analysis plot for non-linear model with out outliers(NLM2)

Fig 5 shows the residual analysis plots for the non-linear model without outliers. As we can see, in the residual vs fitted plot, most of the residuals are concentrated around zero and randomly distributed, indicating that the model is correctly capturing the underlying patterns in the data. This is supported by the scale-location plot and the heteroscedasticity test, with a p-value of 0.2706083 > 0.05. The outliers in the Q-Q plot indicate that the model does not completely capture the variation in the dependent variable, indicating that the model is not accounting for some important variable. Fig 6 shows that the predicted values mimic the true distribution of the data. As a future work, we can investigate how adding these missing variable improves the performance of the model.

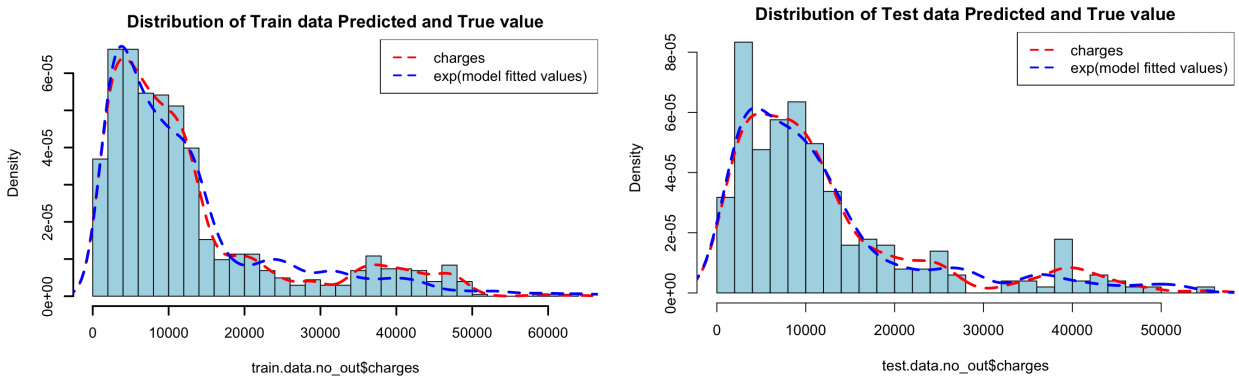

Fig 6: Comparison of Distributions of true values and fitted values for non NLM2.

Conclusion

In conclusion, the goal of this project was to model medical insurance cost data using various machine learning models and evaluate the performance using standard evaluation metrics such as root mean squared error and R-squared. Our analysis revealed that there were non-linear relationships and interactions between features present in the data. The non-linear model without outliers was found to have the highest performance, with an R-squared score of 0.88 on the training data and 0.87 on the test data. Additionally, this model also had the lowest RMSE and MAE values. While Random Forest and SVM algorithms performed comparably, they did not perform as well as the non-linear model without outliers.

The findings of this project highlights the challenges of modeling medical insurance data due to its non-linearity, and how machine learning models can be used to address these challenges. This can aid insurance companies in better understanding and managing their financial risk. Furthermore, these models can be used to predict the cost of an insurance claim and assist underwriting decisions.

Future work can include incorporating more features such as medical history, profession, etc. that would have a high influence on the cost. Additionally, experimenting with more complex non-linear models like neural networks could improve the accuracy of the predictions. Furthermore, experimenting with different feature selection techniques and ensemble methods could also improve the performance of the models.

References :

[1]https://www.kaggle.com/datasets/mirichoi0218/insurance
[2] https://cran.r-project.org/web/packages/caret/index.html