Predicting and classifying insurance charge

Dr Mohammad Zavid Parvez, Dr Md Golam Rakibul Alam,
Fariha Nawaz(15301121)
Department of Computer Science & Engineering
BRAC University
66 Mohakhali, Dhaka, Bangladesh.
E-mail : zavidparvez@gmail.com, golam.rakibul.alam@bracu.ac.bd ,
fariha.nawaz@g.bracu.ac.bd

***Abstract*— One of the most common tasks performed by data scientists are prediction and machine learning. In this paper we are going to predict the insurance cost by using Logistic Regression Analysis, Multiple Linear Regression Analysis and binary Tree. We are using these predictions so that we can find the cost by analysis and it will help the customers to be prepared for the insurance cost the have to give.**

Keywords—Insurance prediction; linear regression; logictic regression; decision tree

I. INTRODUCTION

A. Motivation

Multiple linear regression is one of the most common form of linear regression. As predictive analysis linear regression is the assumption between the dependent variable and independent variable [1]. In our findings we are basically using multivariable linear regression. We are selecting the parameters age, sex, bmi, children, smoker and region and according to that we wanted to predict the insurance charge. So insurance charge is the dependent variable and the other parameters are the independent variable. We use this multiple linear regression mainly for three reasons. First, is might be used to identify strength of the dependency of independent and dependent variable. Second, it can be used to forecast effects or impacts of charges. Third, multiple linear regression analysis predicts trends and future values. **Logistic regression** is a statistical method for analyzing a dataset in which there are one or more independent variables that determine an outcome. The outcome is measured with a dichotomous variable (in which there are only two possible outcomes) [2]. We used logistic regression for this because unlike actual regression, logistic regression does not try to predict the value of a numeric variable given a set of inputs. Instead, the output is a probability that the given input point belongs to a certain class. A **decision tree** is a graphical representation of possible solutions to a decision based on certain conditions. It's called a decision tree because it starts with a single box (or root), which then branches off into a number of solutions, just like a tree [3]. Binary decision tree are those trees which has only two branches. It has a start node and an end node. It decide it should go to that road or not be checking the given parameters.

B. Findings

After running three types of prediction model we found that for the same data the results can be different as the models are calculating the data in different way. Linear regression can work with the original data but the logistic regression and decision tree can only work with the binary data so for the real life scenario ,we think it's best to work with the linear regression as it gives us a specific number, not just zero and one. However, we found that decision tree is the most accurate and then the logistic regression and lastly the linear regression as we didn't change any data here. This is a fully supervised learning as all the data are given and linear regression, logistic regression and the decision tree will learn from the given data. We covering the investing strategy analysis area as we are predicting the insurance charge.

II. EXPERIMENTAL SETUP

A. Description

We used three types of models in our project. For this three cases we used pandas, numpy, Scikit-learn and matplotlib library. We use the age, sex, BMI, smoker ,children and region for predicting the insurance cost.

For the linear regression, first we read the dataset using pandas. We convert the dataset into DataFrame. We divided the features into two part one is X and the other is Y. Y is the insurance charge in our case, and the rest of them are in X for calculating the feature. Then we split the whole dataset into training and testing. Then we fit the dataset in Linear Regression Classifier. We found the coefficient, intercept and then calculate the accuracy of this model. Then we converted the DataFrame into numpy for plotting and plot the original value and predictive value.

For the logistic regression we read the data similarly with the pandas and convert them in DataFrame. Then we calculate the median of the features as we need to convert the data into binary. The data which is greater than the median will be counted as high and the rest will be counted as low. Then again we split the data into X and Y then then split the dataset into train and test. We fit the dataset into Logistic Regression Classifier. We create the confusion matrix from the converted data and then found the true positive, true negative, false positive, false negative. And then we found the accuracy, sensitivity, specificity, positive predicted value, negative predicted value and false negative rate. Finally we plotted the roc curve for the problem

For the binary decision tree we did the similar thing just we took the mean to convert the data in the case of median.

B. Equations

$$\hat{y} = w_0x_0 + w_1x_1 + \cdots + w_nx_n + b \qquad (1)$$

Equation (1) is used for multiple linear regression where input feature vector: x$= (x_0, x_1, x_2, \dots, x_n)$, $\hat{y}$ is the predicted output and $\mathbf{w} = (w_0, w_1, \dots, w_n)$ is model coefficient and b is the constant bias term.

$$l = \beta_0 + \beta_1 x_1 + \beta_2 x_{2+} \dots + \beta_n x_n \qquad (2)$$

Equation (2) is used for logistic regression where the coefficient is βn and l is the result.

$$Entropy(S) = \sum_{i=1}^{c} -p_i \log_2 p_i \qquad (3)$$

Equation (3) is used for calculating the entropy of the decision tree. Here, c is the total number of classes or attributes and “pi” is number of examples belonging to the ith class.

$$Gain(S, A) = Entropy(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Entropy(S) \quad (4)$$

Equation (4) is used for calculating the gain of the information tree. S refers to the entire set of examples that we have. A is the attribute we want to partition or split. |S| is the number of examples and |Sv| is the number of examples for the current value of attribute A.

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \qquad (5)$$

Equation (5) is used for calculating the accuracy where TP is true positive, TN is true negative, FP is false positive and FN is false negative of the confusion matrix.

$$Sensitivity = \frac{TP}{FP + FN} \qquad (6)$$

Equation (6) is used for calculating the sensitivity from the confusion matrix where TP is true positive, FP is false positive and FN is false negative.

$$Specificity = \frac{TN}{FP + TN} \qquad (7)$$

Equation (7) is used for calculating the specificity where TN is true negative and FP is false positive and FN is false negative of the confusion matrix.

$$Positive\ predicted\ value = \frac{TP}{TP + FP} \qquad (8)$$

Equation (8) is used for calculating the positive predicted value where TP is true positive and FP is false positive of the confusion matrix.

$$Negative\ prdicted\ value = \frac{TN}{TN + FN} \qquad (9)$$

Equation (9) is used for calculating the negative predicted value where TN is true negative and FN is false positive of the confusion matrix.

$$False\ negative\ rate = \frac{FN}{TP + FN} \qquad (10)$$

Equation (10) is used for calculating the negative predicted value where FN is false negative, TP is true positive and FN is false positive of the confusion matrix.

$$False\ positive\ rate = \frac{FP}{FP + TN} \qquad (11)$$

Equation (11) is used for calculating the false positive rate from the confusion matrix.

III. Results and Discussions

A. Results found,

For the linear regression we found, Accuracy=80.53%

For the logistic regression we found, Accuracy=85.07%

For the decision tree we found. Accuracy=92.53%

B. Figures and Tables

TABLE I. COEFFICIENT OF THE FEATURES

FEATURE	COEFFICIENT
AGE	253.26
SEX	-3.38
BMI	330.15
CHILDREN	493.59

SMOKER	23712.73
REGION	-98.74

Table I is showing the coefficients that we found from our linear regression model.

Fig.1 LINEAR REGRESSION OUTPUT

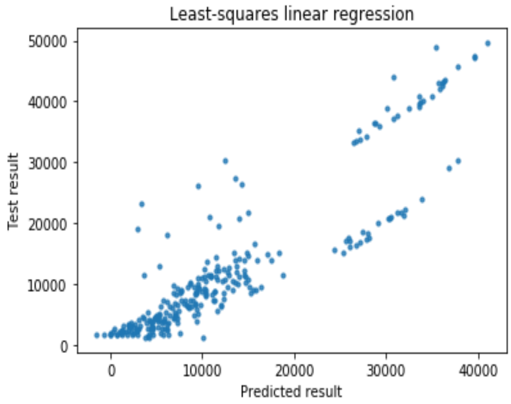

Fig.1 shows the predicted value and the original value of the linear regression.

TABLE II. LOGISTIC REGRESSION RESULT

NAME	VALUE
SENSITIVITY	85.07%
SPECIFICITY	94.82%
POSITIVE PREDICTED VALUE	95.16%
NEGATIVE PREDICTED VALUE	76.39%
FALSE NEGATIVE RATE	22.36&
FALSE POSITIVE RATE	51.72%

Table II is showing the result we got from the logistic regression.

Fig.2 ROC CURVE OF LOGISTIC REGRESSION

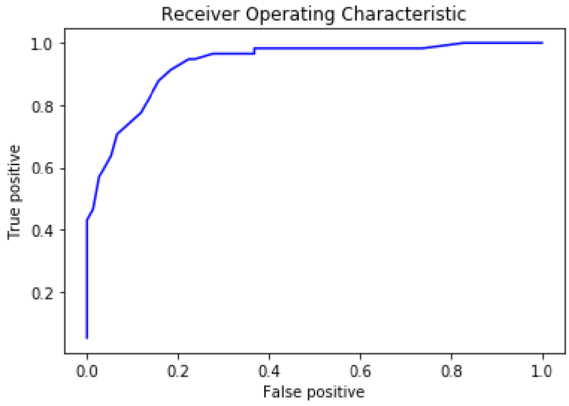

Figure (2) is showing the ROC curve we found by doing the logistic regression.

TABLE III. DECISION TREE RESULT

ORIGINAL	PREDICTED
SENSITIVITY	100.0%
SPECIFICITY	75.61%
POSITIVE PREDICTED VALUE	90.29%
NEGATIVE PREDICTED VALUE	100.0%
FALSE NEGATIVE RATE	0.00%
FALSE POSITIVE RATE	24.39%

Table III is showing the result we found from the decision tree

Fig.3 ROC CURVE OF DECISION TREE

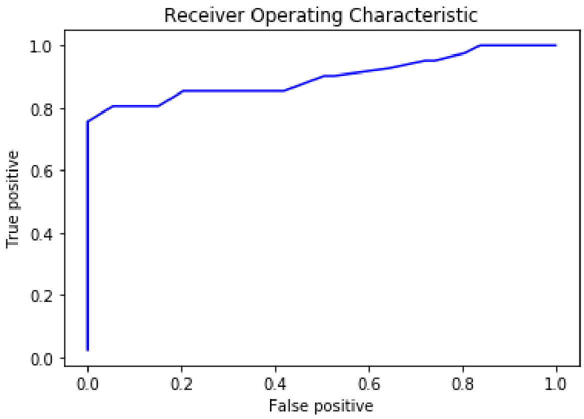

Fig.3 is showing the ROC curve we found from our decision tree.

C. Discussin

From the above result we can say that the result of the decision tree is better than logistic regression as the false negative is zero in binary decision tree.

And not only that, the sensitivity, positive predicted value and the negative predicted value is higher in decision tree, so we can say that decision tree is a much better algorithm in this case.

The data is not that much high in this dataset. If we could find more data, the result would be more accurate.

IV. Conclusion

After finding the value from different models it has become easier to predict weather the insurance cost is high or low and what could be the approximate value of insurance charge. We are hoping that this project will lead the insurance companies to keep up with the competitive market. In future we want to apply different machine learning algorithm in this project to find more accurate value. And we are hoping to publish a paper from this project.

References

[1] "What is Multiple Linear Regression? - Statistics Solutions", *Statistics Solutions*,2018.[Online].Available: https://www.statisticssolutions.com/what-is-multiple-linear-regression/.

[2] F. Schoonjans, "Logistic regression", *MedCalc*, 2018. [Online]. Available: https://www.medcalc.org/manual/logistic_regression.php.

[3] *What Is a Decision Tree? - Examples, Advantages & Role in Management - Video & Lesson Transcript | Study.com*. (2018). *Study.com*. Retrieved 18 November 2018, from https://study.com/academy/lesson/what-is-a-decision-tree-examples-advantages-role-in-management.html