Assignment 2: Logistic Regression

Charushree Madhugiri Bahubali - *MSBA '23*

1 Introduction and objectives

The assignment involves analysing an Insurance policy company's data. The analysis involves building a Logic Regression model for figuring out which individuals are likely to purchase the policy. The model is built to predict purchase (1) vs. non-purchase (0). This analysis is crucial in understanding the customers behaviour towards purchasing the policy. In addition, the analysis can help send sales agents to individuals the models predict as being likely to be interested in the product. The main objective of the project is to build Logistic Regression model, determine the confusion table, and calculate fit of the model with area under the Receiver Operating Characteristic (ROC) curve.

2 Dataset

The Caravan dataset contains 85 predictors that measure demographic characteristics for 5,822 individuals and "Purchase," which indicates whether or not a given individual purchases a caravan insurance policy. The column named "Purchase" is the binary dependent variable with *Yes* and *NO* values. Further analysis reveals that the data set is imbalanced with 94% of *NO*s and only 6% of *YES* as shown in Figure 1.

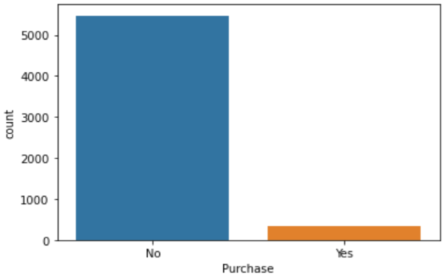

Figure 1: Class imbalance in the dataset

3 Logistic Regression

Logistic regression is the method of estimating the parameters of a logistic model, that models the probability of an event taking place by having the log-odds for the event be a linear combination of one or more independent variables. Binary Logistic Regression is a special case of logistic regression where the dependent variable has only two categories as in our current example.

Metric	Training data	Test data
Sensitivity	0.01	0.03
Specificity	0.99	0.99
Precision	0.33	0.67
Accuracy	0.94	0.93

Table 1: Metrics for training and the test data.

4 Results and Analysis

In this section, I briefly discuss the implementation and the metrics used to evaluate the models followed by analysis of the results.

4.1 Implementation

I use the popular *sklearn*[1] library which provides a rich set of methods like *roc_curve()*, etc., that helps in the easy analysis of the model. Additionally, the *predict_proba()* method allows analysis of the effect of different thresholds in the binary classification.

4.2 Metrics

4.2.1 Confusion matrix

Confusion matrix is a table that is used to define the performance of a classification algorithm. Mainly, it visualizes and summarizes the performance of a classification algorithm.

4.2.2 Sensitivity, Specificity, Precision, and Accuracy

Sensitivity is the ratio of correctly labeled positive data points to the total positive data points in the dataset. *Specificity* is the ratio of correctly labeled negative data points to the total negative data points in the dataset. *Precision* is the ratio of correctly labeled positive data points to the total data points labeled positive by the model. *Accuracy* is the ratio of number of correctly labeled data points to the total number of data points.

4.2.3 Receiver Operating Characteristic (ROC) curve

ROC curve is a graph that illustrates the diagnostic ability of a binary classifier system as its discrimination threshold is varied. Area Under the Curve(AUC) is a determining factor of model fit.

Actual/Predicted	0	1
0	4532	8
1	278	4

(a) Training data.

Actual/Predicted	0	1
0	933	1
1	64	2

(b) Test data.

Table 2: Confusion matrix.

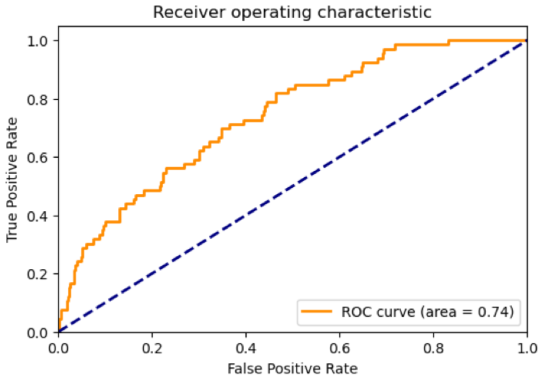

Figure 2: ROC curve for the logistic regression model.

4.3 Analysis

I split the dataset into train and test, with the test dataset consisting of 1000 data points. After training the logistic regression model with the training data, I obtained the predictions for both train and test dataset using a *classification threshold of 0.5*. The metrics corresponding to the training and test data are given in Table 1. The confusion matrix for these two data are also given in Table 2. Figure 2 shows the ROC curve for the model along with the AUC score of 0.74. It shows that the model performs decently when the right classification threshold is chosen but the default classification threshold of 0.5 performs poorly, especially, for the positive class as the dataset is highly imbalanced as shown in Figure 1. The metrics in Table 1 also support this argument. The negative effect of data imbalance can be addressed by various approaches [2].

References

[1] "sklearn Logistic Regression," https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html, online; accessed 2^{nd} Oct, 2022.

[2] "7 Techniques to Handle Imbalanced Data," https://www.kdnuggets.com/2017/06/7-techniques-handle-imbalanced-data.html, online; accessed 2^{nd} October, 2022.