Logistic Regression

Subham Roy
Department of Computer Science and Engineering
University at Buffalo, Buffalo, NY 14260
subhamro@buffalo.edu

1 Overview

The linear regression models assumes that the response variable is quantitative. But in many situations, the response variable is Qualitative. For example, Fish Species is qualitative. Often qualitative variables are referred to as categorical variables. In this report, we work on predicting qualitative responses, a process that is known as Classification.

2 About Logistic Regression

This type of statistical model is often used for classification. Logistic regression estimates the probability of an event occurring, such as did someone vote or did not, based on a given dataset of independent variables. Since the outcome is a probability, the dependent variable is ranges between 0 and 1. In logistic regression, a logit transformation is applied on the odds— i.e., (the probability of success/ the probability of failure). This is also commonly known as the log odds and this logistic function is represented by the following formulas:

$$\text{Logit}(\pi) = 1/(1+ e^{(-\pi)})$$

$$\ln(\pi\ /(1-\ \pi)) = \beta_0 + \beta_1 * X_1 + \ldots + \beta_n * X_n$$

In this logistic regression equation, Logit(π) is the dependent variable and X is the independent variable. The β_i parameters, or coefficient, in this model is commonly estimated via maximum likelihood estimation (MLE). This method tests different values of β through multiple iterations to optimize for the best fit of log odds. All of these iterations produce the log likelihood function, and logistic regression seeks to maximize this function to find the best parameter estimate. Once the optimal coefficients are found, the conditional probabilities for each observation can be evaluated, logged, and summed together to yield a predicted probability.
Here, we particularly work on Binary logistic regression. In this approach, the response variable is divided into two parts (e.g., 0 or 1). Some popular examples of its use include predicting if an e-mail is spam or not spam and the example, we will analyze is whether Customer will churn or not. Within logistic regression, this is one of the most commonly used approach, and more generally, it is one of the most common classifiers for binary classification.

3 Experiment

The Dataset used in this experiment was taken from Kaggle <https://www.kaggle.com/datasets/gangliu/customerchurnrate>

Initially, we start with loading the Dataset and checking the dimensions and the number of variables within the Dataset. In this case the dataset has 28 variables and a dimension of 200 rows and 28 columns.

Next step is we perform data preprocessing. We perform the following steps before we can be certain that we can apply the regression model. This is almost similar for all the techniques we apply. First, we check for missing data, then check for the spread of the data that means where the Mean, Median of the data, are there any outliers, we take care of that. Now that we have cleaned the dataset, we check for categorical data, and if needed we format that to binary using One Hot Encoding or simply sequentially replacing it with factors. This helps the regression model evaluate all the parameters without any bias.

In this case study, the dataset had no particular Null values or outliers. One thing we noticed; our particular dataset had variables of various different scales. So, we implemented the scaled feature over the dataset, first on X_train set and then scaled X_test dataset. One might ask why it is needed to scale separately, rather than scale the whole dataset. This is because in reality we do not have the Test set. If we scale the whole set, we skew the outcome which we do not want. Just so that we can simulate the reality, we scale the training and test set separately. Now we check how each predictor are correlated to the response variable. This helps us identify important variables and ignore the ones that might be redundant, because we already know more data always doesn't equate to good data and can lead to very misleading estimates of the association. Below is the Correlation matrix over the dataset –

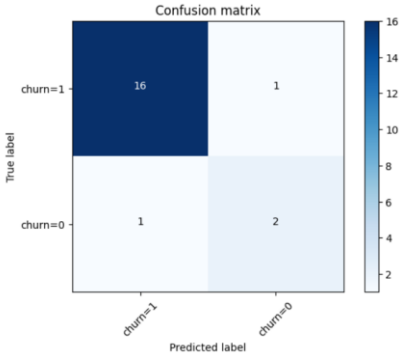
	tenure	age	address	income	ed	employ	equip	churn
tenure	1.000000	0.431802	0.456328	0.109383	-0.070503	0.445755	-0.117102	-0.376860
age	0.431802	1.000000	0.746566	0.211275	-0.071509	0.622553	-0.071357	-0.287697
address	0.456328	0.746566	1.000000	0.132807	-0.145550	0.520926	-0.148977	-0.260659
income	0.109383	0.211275	0.132807	1.000000	0.141241	0.345161	-0.010741	-0.090790
ed	-0.070503	-0.071509	-0.145550	0.141241	1.000000	-0.213886	0.488041	0.216112
employ	0.445755	0.622553	0.520926	0.345161	-0.213886	1.000000	-0.174470	-0.337969
equip	-0.117102	-0.071357	-0.148977	-0.010741	0.488041	-0.174470	1.000000	0.275284
churn	-0.376860	-0.287697	-0.260659	-0.090790	0.216112	-0.337969	0.275284	1.000000

We see particularly these attributes are highly correlated to churn while other remaining attributes are very low compared to that. Here we can ignore those other attributes and select only the significant predictors.

Now in the next step, by using x-test dataset we predict y-test.

4 Analysis

We find out a lot of information about our prediction from this model. Firstly, the model accuracy for all the predictions is above 90%, which is excellent for any model. The below plot depicts the Confusion matrix for the prediction. Confusion Matrix is generally used to validate the Type I and Type II errors occurred during the predictions. Here also, we see excellent results, as we see that both False Positives(Type I) and False Negatives(Type II) occurred only once, and rest was accurately predicted–



Overall, it has an RMSE of 0.3162277 which is very low error for any model and is great considering it is Logistic Regression model.

References

[1] Gareth James [at], Daniela Witten [aut], Trevor Hastie [aut, cre], Rob Tibshirani [aut], Balasubramanian Narasimhan [ctb], *Introduction to Statistical Learning, Second Edition*.

[2] Tom M. Mitchell, *Machine Learning: A multistrategy approach, 1997 Edition*.