Introduction to Supervised Learning

Logistic Regression

Logistic Regression and classification model evaluation– Topics

Logistic Regression:

- Introduction to Logistic Regression.
- Odds ratio and Logit function.
- Math behind Logistic Regression.
- Pros and Cons of Logistic Regression.

Model Evaluation:

- Introduction to different metrics used for model evaluation in classification problems.
- Confusion matrix and related metrics like F1 Score.
- Gini Coefficient and ROC and relation between them.

Introduction to Logistic Regression

- In statistics, the logistic model is a statistical model that is usually taken to apply to a binary dependent variable.

- In regression analysis, logistic regression or logit regression is estimating the parameters of a logistic model.

- In Logistic Regression, the dependent variable is binary rather than continuous and it can also be applied to ordered categories (ordinal data).

The term “Odds”

- Popular in horse races, sports, gambling, epidemiology etc.

- Instead of talking about the *probability* of winning or contacting a disease, people talk about the *odds* of winning or contacting a disease

- How are these two different?

Odds vs Probability

- What is probability of A– P(A)?
- $Odds\ ratio = \frac{Probability\ of\ event\ occuring}{Probability\ of\ event\ not\ occuring}$
- Odds ratio = $\frac{P}{1-P}$
- Probability= $\frac{Odds\ ratio}{1+Odds\ ratio}$

Logit function in Logistic Regression

In logistic regression, the dependent variable is a logit, which is the natural log of the odds, that is,

$$Odds = \frac{P}{1-P}$$

$$\log(odds) = logit(P) = \ln\left(\frac{P}{1-P}\right)$$

Math behind Logistic Regression

So a logit is a log of odds and odds are a function of P, where P is the probability. It is a type of function that creates a map of probability values from (0 , 1) to (-∞ , +∞)

In logistic regression, we find

$$logit(P) = a + bX$$

$$\ln\left(\frac{P}{1-P}\right) = a + bX$$

$$\frac{P}{1-P} = e^{a+bX}$$

$$P = \frac{e^{a+bX}}{1 + e^{a+bX}}$$

Math behind Logistic Regression

- Predict likelihood or probability
- Predicted value - >0 and <1
- Use of sigmoid function to achieve this
- $Probability\ (P) = \frac{e^z}{1+e^z}\ \ Where,\ z = \beta_0 + \beta_1 x$
- $Odds\ Ratio = \frac{P}{1-P}$
- Substituting for P, $Odds\ Ratio = \frac{P}{1-P} = e^z = e^{(\beta_0+\beta_1 x)}$

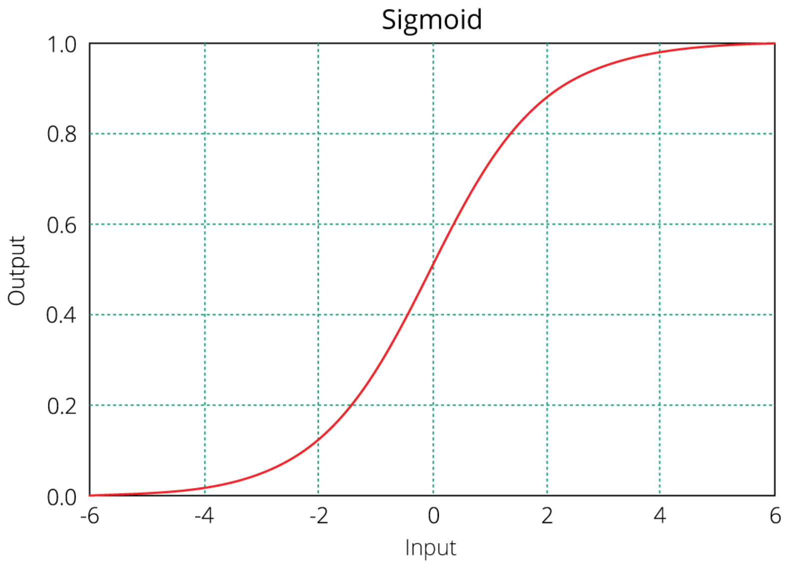
Equation of logistic regression

log - odds or **odds ratio** or **logit** function and is the link function for Logistic Regression

$$\ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x$$

Regression intercept & coefficient

This link function follows a sigmoid function which limits its range of probabilities between 0 and 1.



Pros and Cons of logistic regression

Pros:

- It is a model that gives probabilities.
- It can be easily scaled to multiple classes.
- It is very quick to train and very fast at classifying unknown records.

Cons:

- The classifies constructs liner boundaries.
- Assumes that the variables are independent.
- Interpretation of coefficients is difficult.

Confusion Matrix

Well, it is a performance measurement for machine learning classification problem where output can be two or more classes. It is a table with 4 different combinations of predicted and actual values.

		Actual Values	
		Positive (1)	Negative (0)
Predicted Values	Positive (1)	TP	FP
	Negative (0)	FN	TN

Confusion Matrix Contd

		Actual	
		1	0
Predicted	1	TP	FP
	0	FN	TN

Employees who will actually not attrite but predicted as will attrite

*Employees **who will actually attrite** but predicted as **will not attrite***

Confusion Matrix Contd

		Actual	
		1	0
Predicted	1	TP	FP
	0	FN	TN

$$Specificity = \frac{TN}{TN + FP}$$

$$Precision = \frac{TP}{TP + FP}$$

$$Recall\ or\ Sensitivity = \frac{TP}{TP + FN}$$

Why accuracy is not a good model performance measure?

- # Total customers – 100
  # of customers who churn - 8
- Churn rate – 8%
- Model predicts that no one churns
- Accuracy – 92%
- Misses out critical customers who churn

		Actual	
		1	0
Predicted	1	TP	FP
	0	FN=8	TN=92

F1 Score

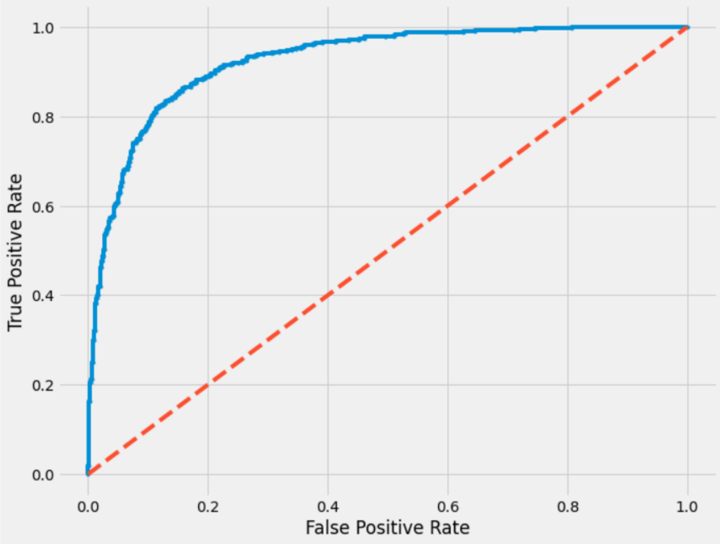
A single metric is not sufficient for the evaluation of classification models. We have seen that we need to use recall and precision together along with accuracy to evaluate our model.
Let us consider another metric that puts together the recall and precision metrics. We call it F1 Score which is the harmonic mean of the two metrics – Precision and Recall.

F1 Score = 2(precision*recall)/precision + recall

The F1 score can also be used to evaluate the model.

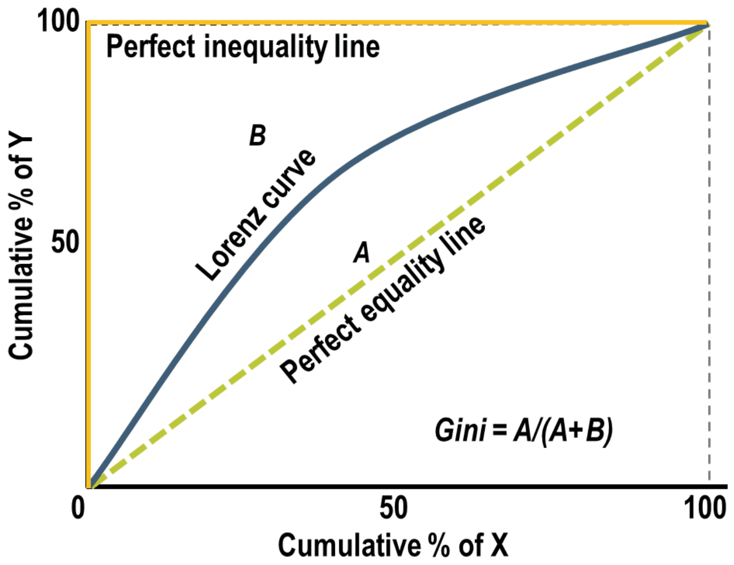
ROC AUC

- Roc is a curve which allows us to compare models.
- It is plot between TPR(true positive rates) and FPR(false positive ratio).
- The area under the ROC curve (AUC) is a measure of the how good a model is.
- The value for AUC lies between 0 and 1. Larger AUC implies better model.



Gini Coefficient

- It is also used to measure the goodness of a fit.
- It is the ratio of areas in a roc curve and is scaled version of the AUC.
- GI = 2*AUC -1
- The larger the Gini coefficient the better the model is.



Hands on exercise on Logistic Regression

Problem Overview

The dataset used here is Credit risk data.

Credit risk is nothing but the default in payment of any loan by the borrower. In Banking sector this is an important factor to be considered before approving the loan of an applicant. Dream Housing Finance company deals in all home loans. They have presence across all urban, semi urban and rural areas. Customer first apply for home loan after that company validates the customer eligibility for loan.

Company wants to automate the loan eligibility process (real time) based on customer detail provided while filling online application form. These details are Gender, Marital Status, Education, Number of Dependents, Income, Loan Amount, Credit History and others. To automate this process, they have given a problem to identify the customers segments, those are eligible for loan amount so that they can specifically target these customers. Here they have provided a partial data set.

Hands on Contd.

Dataset features:

- Variable Description
- Loan_ID Unique Loan ID
- Gender Male/ Female
- Married Applicant married (Y/N)
- Dependents Number of dependents
- Education Applicant Education (Graduate/ Under Graduate)
- Self_Employed Self employed (Y/N)
- ApplicantIncome Applicant income
- CoapplicantIncome Coapplicant income
- LoanAmount Loan amount in thousands
- Loan_Amount_Term Term of loan in months
- Credit_History credit history meets guidelines
- Property_Area Urban/ Semi Urban/ Rural
- Loan_Status Loan approved (Y/N)

Hands on Contd.

Steps to follow:

- Importing Libraries.
- Check for the features, use the tools at your dispense to gather insights of the features.
- Identify the opportunities to do the required pre processing of your data.
- Get rid of the inconsistency, remove missing values.
- Take 70:30 training and test set.
- Fit the model on train data.
- Check for the fit on the test data.
- Choose the best metric and give business insights.

greatlearning

Power Ahead

Happy Learning !