Logistic Regression in Python

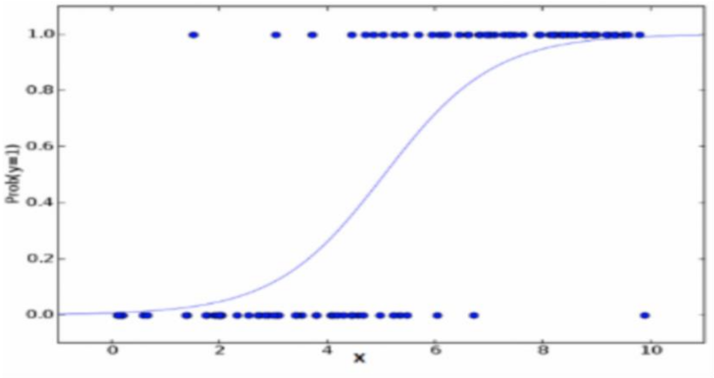
1. Introduction

1.1. Objective

Logistic Regression is a "Supervised machine learning" algorithm that can be used to model the probability of a certain class or event. It is used when the data is linearly separable, and the outcome is binary or dichotomous in nature. That means Logistic regression is usually used for Binary classification problems. Binary Classification refers to predicting the output variable that is discrete in two classes. A few examples of Binary classification are Yes/No, Pass/Fail, Win/Lose, Cancerous/Non-cancerous, etc.

The sigmoid function is a mathematical function used to map the predicted values to probabilities. It maps any real value into another value within a range of 0 and 1. The value of the logistic regression must be between 0 and 1, which cannot go beyond this limit, so it forms a curve like the "S" form. The S-form curve is called the Sigmoid function or the logistic function.

- logit(p) = b0 + b1X1 + b2X2 + —— + bk Xk
- where logit(p) = ln(p / (1-p))
- p: the probability of the dependent variable equalling a "success" or "event".



1.2. Data

There are 1,100 observations and 7 columns.

- **AGE:** Age of the applicant (between 3 and 78 years)
- **HOME:** Home owner or Home renter (indicated by H and R)
- **INCOME:** Income of the applicant in USD (between 20,000 and 70,000)
- **GENDER:** Male or Female (indicated by M and F)
- **HOUSEHOLD_N:** Number of dependents (count of household)
- **CREDIT_LINES_N:** Number of credit lines (between 0 and 6)
- **DEAFULTED:** Dependent variable (1 indicates defaulted and 0 indicated non-defaulted)

1.3. Dependent variable

59% of the observations did not default and **41%** of the observations defaulted.

2. Feature Engineering

2.1. Categorical Features

For each category the log-odds is calculated. An example is provided below:

- For home-owners (H) the probability of default is 31.8% and the log-odds is -0.761
- For home-renters (R) the probability of default is 59.7% and the log-odds is 0.394
- The categories are replaced by the log-odds. Hence the categorical variable is converted to numerical variable

There are 2 categorical variables –home and gender

Home

Row Labels	Count of Rows	Average of Defaulted	Log Odds
H	735	31.8%	-0.761
R	365	59.7%	0.394

Gender

Row Labels	Count of Rows	Average of Defaulted	Log Odds
F	647	40.5%	-0.385
M	453	41.9%	-0.325

2.2. Numerical Features

The numerical variable is first binned. Then for each bin the log-odds is calculated. An example is provided below:

- For income less than or equal to USD 20,000 the probability of default is 73.2% and the log-odds is 1.006
- For income more than USD 20,000 and less than or equal to USD 40,000 the probability of default is 55.2% and the log-odds is 0.208
- For income more than USD 40,000 and less than or equal to USD 50,000 the probability of default is 26.8% and the log-odds is -1.003
- For income more than USD 50,000 the probability of default is 17.4% and the log-odds is -1.559
- The bins are replaced by the log-odds.

There are 4 numerical variables – income, age, number of credit lines and number of dependents (count of household)

Income

Row Labels	Count of Rows	Average of Defaulted	Log Odds
a. 2	127	73.2%	1.006
b. 3-4	424	55.2%	0.208
c. 5	313	26.8%	-1.003
d. 6-7	236	17.4%	-1.559

Age

Row Labels	Count of Rows	Average of Defaulted	Log Odds
a. le 30	103	60.2%	0.414
b. 31-45	269	42.8%	-0.292
c. 46-60	421	31.8%	-0.762
d. ge 61	307	45.9%	-0.163

Number of Credit Lines

Row Labels	Count of Rows	Average of Defaulted	Log Odds
0	192	26.0%	-1.044
1	216	34.3%	-0.652
2+	692	47.4%	-0.104

Number of Dependents

Row Labels	Count of Rows	Average of Defaulted	Log Odds
a . Even	487	34.5%	-0.641
b. Odd	613	46.3%	-0.147

3. Model

3.1. Model Assumptions

Assumptions of Logistic Regression

- Logistic regression requires the appropriate structure of the outcome variable. Binary logistic regression requires the dependent variable to be binary.
- Logistic regression requires there to be little or no multicollinearity among the independent variables.
- Logistic regression requires the observations to be independent of each other.
- Logistic regression assumes linearity of independent variables and log odds.

The final model has 4 independent variables. All the 4 variables have p-value less than 0.05. This indicates that all the variables are significant.

	coef	std err	z	P>|z|
const	1.024	0.158	6.496	0.000
home_bin	0.463	0.140	3.312	0.001
income_bin	1.031	0.093	11.097	0.000
household_bin	0.622	0.295	2.109	0.035
creditlines_bin	1.771	0.215	8.221	0.000

3.2. Model Summary

- **AUC-ROC:** Area under the curve of Receiver Operating Characteristic examines how well the model can distinguish between the positives and negatives. It is the plot between True Positive Rate (Sensitivity) and False Positive Rate (1-Specificity).

  AUC-ROC: 0.780

- **Gini Coefficient:** Gini coefficient gives a summary of the CAP curve. It tells us about the proximity of out model to the perfect model and how far it is from a random model. It is measured by calculating the area between the CAP and the diagonal as the proportion of are in the ideal rating procedure. Relation between Gini Coefficient and AUROC: 2*AUROC-1

  Gini: 0.560

- **Confusion Matrix:** It is a table that is often used to describe the performance of a classification model. The attributes of a confusion matrix can be divided into 4 categories – True Positive (TP) , True Negative (TN) , False Positive (FP) , False Negative (FN).

	Predicted = 0	Predicted = 1
Actual = 0	**501**	**147**
Actual = 1	**150**	**302**

- **Accuracy:** This statistic gives the percentage how often is the classifier correct. Accuracy = (TP+FP)/ (TP+FP+TN+FN)

  Accuracy: 0.730

- **F1 Score:** F1 score is the weighted average of precision and recall. This metric is more useful than accuracy, especially when the model has uneven class distribution. Precision statistic gives when predicted 1, how often it is correct. Recall statistic gives out of actual true, how often it is predicted True. F1 Score = (2 * P * R) / (P + R)

  Precision: 0.673
  Recall: 0.668
  F1 Score: 0.670