GERMAN_LOGISTIC_MODEL

The functional form of the Logistic Regression Model is given by

$$P(Y=1) = \ln [e^z/(1+ e^z)]$$

Where $Z = \beta_0+\beta_1X_1+\beta_2X_2+\beta_3X_3+\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\beta_{11}X_{11}+\beta_{12}X_{12}+\beta_{13}X_{13}$

Where

Y = 1 : Credit Classification as Bad

Y = 0 : Credit Classification as Good

X_1 = CHK_ACCT X_2 = Duration X_3 = Credit History

X_4 = Credit Amount X_5 = Balance in Savings A/C X_6 = Employment

X_7 = Install_rate X_8 = Marital status X_9 = Present Resident

X_{10} = Age X_{11} = Other installment X_{12} = Num_Credits

X_{13} =Job

Model 1 Summary

$P(Y=1) = \ln [e^z/(1+ e^z)]$

Where

Z = -0.0754+0.0104*Duration+0.0002*CreditAmount+0.3399*Install_rate+……..0.1159*Job_ skilled-0.3676*Job_Unemployed-0.0018*Job_Unskilled

The model summary suggest that as per walds test, only 7 features are statistically significant at a significance value of $\alpha=0.05$. The significant variables are

Credit Amount

Install_rate

Age

CHK_ACCT_no-account

CHK_ACCT_over-200DM

Credit History_ critical

Credit History_ delay

Model 2 Summary

The new Model with significant variables are

$P(Y=1) = \ln [e^z/(1+ e^z)]$

Z = -0.4385+0.0002*CreditAmount+0.2987*Install_rate-0.0267*Age-1.7346*CHK_ACCT_no-account-

1.1121*CHK_ACCT_over-200DM-0.6778*Credit History_ critical-0.4645*Credit History_ delay

Some observations from the Model output are

- The probability of being bad customer increases as credit amount and install rate increases.
- The probability of being bad customer decreases as Age decreases.

Confusion matrix

	Bad credit	Good credit
Bad credit	24 (True Negative)	37 (False Positive)
Good credit	30 (False Negative)	149 (True Positive)

In the confusion matric the columns represent the predicted class while the rows represent the actual class.

Accuracies

Sensitivity = TP/(TP+FN) = 149/(149+30) = 0.832

Specificity = TN/(TN+FP) = 24/(24+37) = 0.393

Precision = TP/(TP+FP) = 149/(149+37) = 0.801

ROC Curve Summary

The Receiver Operating Curve can be used to understand the overall performance of Logistic Regression Model. It is a plot between sensitivity(True Positive Rate) on the vertical axis and 1-Specificity (false positive rate) on the horizontal axis.

The diagonal line in the figure represents the case of not using a model. The sensitivity and Specificity are likely to change when the cut-off probability is changed. The line above the diagonal line in the figure captures how sensitivity and 1-Specificity changes when the cut-off probability is changed. Model with higher AUC is preferred. The AUC for the model obtained is 0.72 validates the model.