Credit risk model for forecasting loan default

O.G.V.V Bandara(SC/2019/10723),E.R.D.M Kumari(SC/2019/10836),E.A.T Maduwanthi(SC/2019/10751)

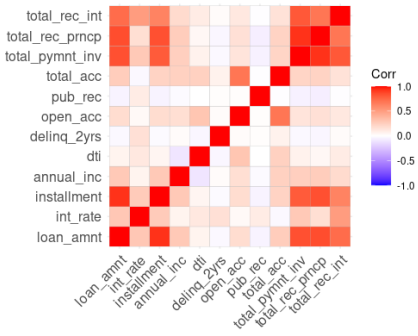
I. Introduction

Credit risk arises when a corporate or individual borrower fails to meet their debt obligations. Loan repayment deafault can result in various factors that effect the borrower. Also loan default can result in loss and conflict for the lender.The objective was to design a model for predicting loan default using actual data. A fitted model after validation has the ability to make prediction for new data.

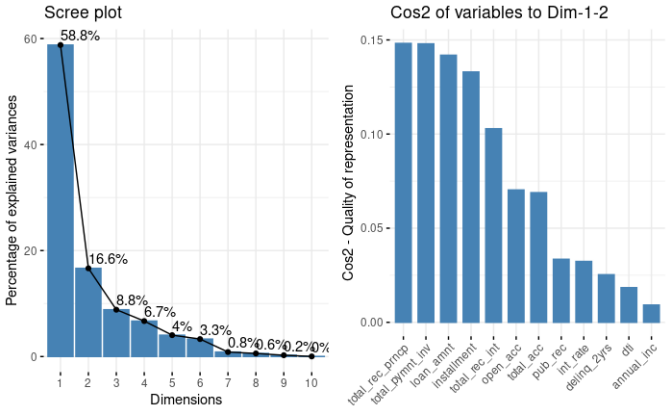
II. Methedology

A. Feature selection

Features selection was performed using principle component analysis.Covariance matrix for numerical features represent the intensity of correlation between features.



principle components were generated corresponding to the number of features present.Each component explains a percentage of the total variance in the data set. The scree plot depicts the contribution percentages to the principle components.The first 5 components show a major contribution. Analyzing the 'sqaure cosine', the position of leveling shows the end of the major contributors to the principle components hence contributes to the selection of the 5 major contributors.



B. Model fitting

The Generalized linear model was utilized for model fitting. Model fitted for three linkage functions namely "logit","probit" and "cloglog". GLM's have non normal errors. GLM is a generalization of ordinary linear regression.

III. Conclusion

holdout validation for model selection

A. "logit" link model

R2 score	RMSE	MAE
0.5398092	0.7942828	0.4489687

B. "probit" link model

R2 score	RMSE	MAE
0.5549441	0.7899474	0.4460073

C. "cloglog" link model

R2 score	RMSE	MAE
0.4703521	0.814688	0.516473

comparison of float values,
$R22 > R21 > R23$
$RMSE3 > RMSE1 > RMSE2$
$MAE2 > MAE1 > MAE3$
model 1 with "logit" link is the most suitable model

10 fold cross validation for model selection

D. "logit" link model

R2 score	RMSE	MAE
0.901279	0.1122364	0.0286609

E. "probit" link model

R2 score	RMSE	MAE
0.8828325	0.1228718	0.03587401

F. "cloglog" link model

R2 score	RMSE	MAE
0.8704998	0.128783	0.0170227

$R21 > R22 > R23$
$RMSE3 > RMSE2 > RMSE1$
$MAE2 > MAE1 > MAE3$
Through 10 fold cross validation it is possible to determine that model 1 is the best model with "logit" link.

model selection through akike information criteria (AIC)

aic_1	aic_2	aic_3
4913.425	5700.689	47085.01

$aic_1 < aic_2 < aic_3$
The minimum value is for model 1, Therefore model 1 with "logit" link is the most suitable.

overall conclusion is that the model 1 with link "logit" is the most suitable