GLM:

Delete all observations with loan\_status 'Current' since we are looking at predicting whether a client will eventually be charged off.

Delete all observations with incomplete data (924 out of 82537, so it is okay).

Notice we will assign 0 to 'Charged Off' status and 1 to all other statuses.

From now on, we will consider only binomial GLM.

1. Build intercept-only GLM as null model

2. Build full model:

2.1. build numerical-only model

Use scatter-plot and correlation plot to seek possible numerical-numerical interaction terms (model l2)

2.2. simplify numerical-only model

Delete insignificant terms (model l3) and check ANOVA(l3,l2) which has p-value 0.244

2.3. include numerical-categorical interactions

For each categorical predictor, build GLM with numerical terms and numerical to that specific predictor interaction terms (model l4 to l13). Each GLM is compared to numerical-only GLM (model l3) with drop-in deviance test to see whether this categorical predictor's interaction should be included. If so, check summary to see whether level-combination is likely to be useful and compare releveled GLM to un-releveled GLM (p-value should be >0.05).

It turns out that only 'grade' and 'emp\_length' interactions should be included into full model.

3. forward selection from null to full models based on AIC with 20 steps to get final1 model.

4. check summary of final1 to remove insignificant terms to get final3 model

5. determine classifier to separate predicted values for different observed values (in our case, any logit(odds)>0.42 is classified as 1 (non charged off) and vice versa)

prediction accuracy 80.39%

all plots and data can be retrieved from R code.

Good luck