**Name: Yeap Jiun Aing - Assignment 3**

**1)**

Other.installment.plans and Years.with.present.employer

have value not as per defined in code book.

The undefined values are marked as NA. I decided to exclude the two records having such issues.

sapply(gcr\_raw, function(x) sum(is.na(x)))

sapply(gcr\_raw, function(x) length(unique(x)))

# Found Other.installment.plans and Years.with.present.employer have values no defined in the code book

gcr$purpose.of.loan[gcr$purpose.of.loan>10|gcr$purpose.of.loan<0]

gcr$Other.installment.plans[gcr$Other.installment.plans>3|gcr$Other.installment.plans<1] <-NA

gcr$Years.with.present.employer[gcr$Years.with.present.employer>5|gcr$Years.with.present.employer<1] <-NA

library(Amelia)

missmap(gcr, main ="Missing values vs observed")

nrow(gcr)

gcr <-gcr[complete.cases(gcr),] # remove rows having NA value

nrow(gcr)

2)

Personal status and sex  
1: Male + Divorced / Separated  
2: Female + Divorced / Separated / Married 3: Male + Single  
4: Male + Married / Widowed  
5: Female + Single

The above column has been found to contain two type of info, which entails further splitting to create two new fields which are **sex** and **isSingle** .

gcr$sex <- 0 # default as male  
gcr$sex[gcr$Personal == 2|gcr$Personal == 5] <- 1 ## 1 as female  
gcr$IsSingle <-0 # Is not single  
gcr$IsSingle[gcr$Personal == 3|gcr$Personal == 5] <-1 ## is single

Credit Offer is the value which the model to built needs to predict. It has value of 1 (approve) and 2 (reject). Transform the value of 2 to 0 for fitting to logistic regression formula which predict in the range [0..1]

gcr$Credit.offered[gcr$Credit.offered == 2] <- 0 # To factor 0 (original raw value=2) as reject him and 1 (original raw value) as approve



3)

Try 4, 5, 6, 7 clusters plotting.

6 clusters probably has best distinct grouping in 2-D plot.

set.seed(20)  
grp <- c(1:21)  
gcrCluster <- kmeans(gcr[, grp], 4, nstart = 50)  
autoplot(gcrCluster,gcr[, grp])





**4)**

A **model** was first built with all the possible dependable (predictive) variables.

Then reviewing the model using chisq test across each predictor in sequence to determine the relative importance of each predictor.

null hypothesis is that the predictors are independent -- they have no contribution to the prediction

anova(model, test="Chisq")

it is found that Status.of.account (\*\*\*), Loan.Duration(\*\*\*), Credit.history(\*\*\*), purpose.of.loan(\*\*\*), Savings.account(\*\*), Other.installment.plans(\*) are significance, we rejected null hypothesis that they are independent (has no contribution to the model prediction).

6 other models with different predictors were randomly picked and tried - model1, model2, .. model6

It is found that **model6** has the lowest AIC among them. Hence it is chosen.

Df Deviance Resid. Df Resid. Dev Pr(>Chi)

NULL 802 986.24

Status.of.account 3 112.688 799 873.56 < 2.2e-16 \*\*\*

Loan.Duration 1 21.611 798 851.95 3.339e-06 \*\*\*

Credit.history 4 25.394 794 826.55 4.191e-05 \*\*\*

purpose.of.loan 9 28.903 785 797.65 0.0006729 \*\*\*

Savings.account 4 17.384 781 780.26 0.0016273 \*\*

Other.installment.plans 2 7.549 779 772.71 0.0229461 \*

other.debtors.guarantors 2 8.829 777 763.89 0.0121012 \*

Age 1 5.713 776 758.17 0.0168397 \*

Foreign.Worker 1 5.124 775 753.05 0.0236004 \*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’

5)

myTrain<- train[,c(1,2,3,6,11,13,19)]

head(train[,c(1,2,3,6,11,13,19)])

round(cor(myTrain),2)

Status.of.account Loan.Duration Credit.history Savings.account Assets Other.installment.plans Foreign.Worker

Status.of.account 1.00 -0.05 0.17 0.23 0.01 0.03 -0.06

Loan.Duration -0.05 1.00 -0.06 0.04 0.32 0.00 -0.12

Credit.history 0.17 -0.06 1.00 0.03 -0.05 0.12 0.04

Savings.account 0.23 0.04 0.03 1.00 0.01 0.03 -0.01

Assets 0.01 0.32 -0.05 0.01 1.00 -0.06 -0.14

Other.installment.plans 0.03 0.00 0.12 0.03 -0.06 1.00 0.01

Foreign.Worker -0.06 -0.12 0.04 -0.01 -0.14 0.01 1.00

There isn’t any strong correlations amongst the predictors suggesting multicollinearity in the predictors.

The correlation between status.of.account , credit.history and saving.account are found to be relatively stronger.

Model7 -9 were built by removing one of the predictors (status.of.account , credit.history and saving.account.) to see it would improve the results. It is found that the model 6 which maintains status.of.account , credit.history and saving.account as predictors still better.

6)

Model6

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) -1.218542 0.804399 -1.515 0.129811

Status.of.account2 0.376295 0.233404 1.612 0.106918

Status.of.account3 0.899182 0.384465 2.339 0.019347 \*

Status.of.account4 1.707840 0.247974 6.887 5.69e-12 \*\*\*

Loan.Duration -0.037636 0.007963 -4.727 2.28e-06 \*\*\*

Credit.history1 0.334064 0.563372 0.593 0.553199

Credit.history2 0.813563 0.446570 1.822 0.068484 .

Credit.history3 1.076800 0.515680 2.088 0.036787 \*

Credit.history4 1.619878 0.473379 3.422 0.000622 \*\*\*

purpose.of.loan1 1.589202 0.396143 4.012 6.03e-05 \*\*\*

purpose.of.loan2 0.690109 0.271039 2.546 0.010891 \*

purpose.of.loan3 1.031915 0.262311 3.934 8.36e-05 \*\*\*

purpose.of.loan4 0.420132 0.760852 0.552 0.580821

purpose.of.loan5 0.219568 0.605441 0.363 0.716861

purpose.of.loan6 -0.038187 0.434294 -0.088 0.929932

purpose.of.loan8 1.553743 1.194155 1.301 0.193216

purpose.of.loan9 0.999249 0.359174 2.782 0.005401 \*\*

purpose.of.loan10 1.524437 0.776945 1.962 0.049752 \*

Savings.account2 0.208664 0.307530 0.679 0.497444

Savings.account3 0.415124 0.449826 0.923 0.356083

Savings.account4 1.155162 0.561039 2.059 0.039497 \*

Savings.account5 1.098032 0.284837 3.855 0.000116 \*\*\*

Other.installment.plans2 0.570172 0.459641 1.240 0.214801

Other.installment.plans3 0.849942 0.262061 3.243 0.001182 \*\*

other.debtors.guarantors2 -0.193400 0.431183 -0.449 0.653769

other.debtors.guarantors3 1.131623 0.436217 2.594 0.009482 \*\*

Age 0.021247 0.008781 2.420 0.015530 \*

Foreign.Worker1 -1.255314 0.607873 -2.065 0.038914 \*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Status of account, loan duration, credit history, purpose of loan, saving account, other instalment plan and guarantor forms the predictors of the logistic regression model.

It has intercept of -1.218542.

As some of the fields are categorial type of value, when built in the regression model, we could see some of the categorial values have statistical significance (which have \*\*\*), for example:

4: No checking account

4: Critical account / other credits existing (not at this bank)

7)

Accuracy : 0.735

95% CI : (0.6681, 0.7948)

No Information Rate : 0.715

P-Value [Acc > NIR] : 0.2945

Kappa : 0.3097

Mcnemar's Test P-Value : 0.1696

Sensitivity : 0.4386

Specificity : 0.8531

Pos Pred Value : 0.5435

Neg Pred Value : 0.7922

Prevalence : 0.2850

Detection Rate : 0.1250

Detection Prevalence : 0.2300

Balanced Accuracy : 0.6459

The model has an accuracy of 0.735. It has sensitivity of 0.4386. That means the True Positive hit rate isn’t too high but with a good a specificity rate 0.8531 which translate to (1- specificity) = false positive rate = **0.1469.**

But to Note that it is worse to class a customer as good when they are bad, than it is to class a customer as bad when they are good. The model must be geared towards reducing false positive error.

8)

AUC = 0.7733407 for model 6, which is the best AUC values compare to others models tried.

As a rule of thumb, a model with good predictive ability should have an AUC closer to 1 (1 is ideal) than to 0.5.