German Credit

Background

The second-oldest profession in the world, money lending has existed since the invention of money. However, the systematic assessment of credit risk is a very recent development, as lending was formerly largely dependent on reputation and incredibly sketchy data. The third President of the United States, Thomas Jefferson, was continually in debt and untrustworthy with his loan payments, but people kept giving him money anyhow. The Retail Credit Company was established to disseminate credit information only around the turn of the 20th century. Equifax, one of the top three credit scoring companies, is now that business (the other two are Transunion and Experion).

These days, the use of local and individual human judgment is mostly irrelevant in the credit reporting process. Numerous customer and transactional details are used by credit agencies and other major financial firms that issue credit to consumers to make predictions about the likelihood of defaults and other unfavorable events.

Data

This study focuses on a historical transition to predictive modeling at an early stage when records were classified as having good or bad credit by humans. The German Credit dataset contains 1000 records with 30 variables, each representing a previous credit applicant. In 700 cases, each applicant received a "good credit" or "poor credit" rating (300 cases). The values of these variables for the first four records are displayed in ***Table 1***. ***Table 2*** provides explanations for each variable. Additionally, based on the values of the 30 predictor variables, new credit applicants can be categorized as either good or bad credit risks.
According to assessments, the following are the effects of misclassification: The expenses of a false positive—saying an application is a good credit risk when you shouldn't—outweigh the advantages of a real positive—saying an applicant is a good credit risk—by a factor of five. ***Table 3*** provides an overview. The average net profit per loan as stated in ***Table 4*** was used to create the opportunity cost table. We use these tables to evaluate how well the various models perform since decision-makers are accustomed to thinking

Table 1: First four records from German Credit dataset

OBS#	CHK_ACCT	DURATION	HISTORY	NEW_CAR	USED_CAR	FURNITURE	RADIO/TV	EDUCATION	RETRAINING	AMOUNT	SAV_ACCT	EMPLOYMENT	INSTALL_RATE	MALE_DIV	MALE_SINGLE	MALE_MAR_WID	CO-APPLICANT	GUARANTOR
1	0	6	4	0	0	0	1	0	0	1169	4	4	4	0	1	0	0	0
2	1	48	2	0	0	0	1	0	0	5951	0	2	2	0	0	0	0	0
3	3	12	4	0	0	0	0	1	0	2096	0	3	2	0	1	0	0	0
4	0	42	2	0	0	1	0	0	0	7882	0	3	2	0	1	0	0	1

PRESENT_RESIDENT	REAL_ESTATE	PROP_UNKN_NONE	AGE	OTHER_INSTALL	RENT	OWN_RES	NUM_CREDITS	JOB	NUM_DEPENDENTS	TELEPHONE	FOREIGN	RESPONSE
4	1	0	67	0	0	1	2	2	1	1	0	1
2	1	0	22	0	0	1	1	2	1	0	0	0
3	1	0	49	0	0	1	1	1	2	0	0	1
4	0	0	45	0	0	0	1	2	2	0	0	1

(Data adapted from German Credit)

Table 2: Variables for the German Credit Dataset

Variable number	Variable name	Description	Variable type	Code description
1	OBS#	Observation numbers	Categorical	Sequence number in dataset
2	CHK_ACCT	Checking account	Categorical	0: <0 DM
		status		
				1: 0–200 DM
				2 : >200 DM
				3: No checking account
3	DURATION	Duration of	Numerical	
		credit in months		
4	HISTORY	Credit history	Categorical	0: No credits taken
				1: All credits at this bank paid back duly
				2: Existing credits paid back duly until now
				3: Delay in paying off in the past
				4: Critical account
5	NEW_CAR	Purpose of	Binary	Car (new), 0: no, 1: yes
		credit		
6	USED_CAR	Purpose of	Binary	Car (used), 0: no, 1: yes
		credit		
7	FURNITURE	Purpose of	Binary	Furniture/equipment,
		credit		0: no, 1: yes
8	RADIO/TV	Purpose of	Binary	Radio/television,
		credit		0: no, 1: yes
9	EDUCATION	Purpose of	Binary	Education, 0: no, 1: yes
		credit		
10	RETRAINING	Purpose of	Binary	Retraining, 0: no, 1: yes
		credit		
11	AMOUNT	Credit amount	Numerical	
12	SAV_ACCT	Average balance	Categorical	0: <100 DM
		in savings		1 : 101–500 DM
		account		2 : 501–1000 DM
				3 : >1000 DM
				4 : Unknown/no savings account
13	EMPLOYMENT	Present	Categorical	0 : Unemployed
		employment		1: <1 year
		since		2: 1–3 years
				3: 4–6 years
				4: ≥7 years
14	INSTALL_RATE	Installment	Numerical	
		rate as		
		% of disposable		
		income		
15	MALE_DIV	Applicant is male	Binary	0: no, 1: yes
		and divorced		
16	MALE_SINGLE	Applicant is male	Binary	0: No, 1: Yes
		and single		
17	MALE_MAR_WID	Applicant is male and married or a widower	Binary	0: No, 1: Yes
18	CO-APPLICANT	Application has a coapplicant	Binary	0: No, 1: Yes
19	GUARANTOR	Applicant has a guarantor	Binary	0: No, 1: Yes
20	PRESENT_RESIDENT	Present resident	Categorical	0: ≤1 year
		since (years)		1: 1–2 years
				2: 2–3 years
				3: ≥3 years
21	REAL_ESTATE	Applicant owns	Binary	0: No, 1: Yes
		real estate		
22	PROP_UNKN_NONE	Applicant owns no property (or unknown)	Binary	0: No, 1: Yes
23	AGE	Age in years	Numerical	
24	OTHER_INSTALL	Applicant has other installment plan credit	Binary	0: No, 1: Yes
25	RENT	Applicant rents	Binary	0: No, 1: Yes
26	OWN_RES	Applicant owns residence	Binary	0: No, 1: Yes
27	NUM_CREDITS	Number of	Numerical	
		existing credits		
		at this bank		
28	JOB	Nature of job	Categorical	0 : Unemployed/unskilled— non-resident
				1 : Unskilled— resident
				2 : Skilled employee/official
				3 : Management/self-employed/highly qualified employee/officer
29	NUM_DEPENDENTS	Number of people for whom liable to provide maintenance	Numerical	
30	TELEPHONE	Applicant has phone in his or her name	Binary	0: No, 1: Yes
31	FOREIGN	Foreign worker	Binary	0: No, 1: Yes
32	RESPONSE	Credit rating is good	Binary	0: No, 1: Yes

The original dataset had a variety of category variables, some of which were converted into a collection of binary variables and others of which were retained in their original format to be handled as numerical variables. Information taken from German Credit.

Table 3: Opportunity Cost Table (Deutsche Marks)

	Predicted (decision)	
Actual	**Good (accept)**	**Bad (reject)**
Good	0	100
Bad	500	0

Table 4: Average Net Profit (Deutsche Marks)

	Predicted (decision)	
Actual	**Good (accept)**	**Bad (reject)**
Good	100	0
Bad	−500	0

Assignment

1. Review the predictor variables (descriptive statistics) and guess what their role in a credit decision might be. Are there any surprises in the data?

Predictor variables

Category	Credit Purpose	Credit Information	Financial Status of Applicant	Information of Applicant
Predictors	New_CAR	DURATION	CHK_ACCT	EMPLOYMENT
	USED_CAR	AMOUNT	HISTORY	MALE_DIV
	FURNITURE	INSTALL_RATE	SAV_ACCT	MALE_SINGLE
	RADIO/TV	CO_APPLICANT	OTHER_INSTALL	PRESENT_RESDENT
	EDUCATION	GUARANTOR	NUM_CREDITS	REAL_ESTATE
	RETRAINING	RESPONSE		PROP_UNKN_NONE
				AGE
				RENT
				OWN_RES
				NUM DEPENDENTS
				JOB
				NUM_DEPENDENTS
				TELEPHONE
				FOREIGN

Surprising data

Variable Name	Reason / Comments
AMOUNT	This variable indicates how much debt an individual already has.
RENT / OWN_RES	The complement of these two is one another. It would have the same financial outcome whether they own and rent or own and don't rent. Only one will be used.
NEW_CAR / USED_CAR	As a rule, the same variable is used, therefore these variables look odd. The owner of a used automobile also owns a brand-new vehicle. There should only be one employed for analysis, as a result. I only intend to utilize NEW_CAR.
DURATION	This reveals the length of time the creditor has been paying on time or not at all.
CHK_ACCT / SAV_ACCT	The person may be able to repay more money if they have a larger income.

Others

All other factors are normal and unremarkable.

2. Partition the dataset into 60% training and 40% validation (set the seed to 12345). Develop at least two classification models of your choice. Describe the two models that you chose, with sufficient details (method, parameters variables, etc.) so that it can be replicated.

Model selection

Utilizing the data mining methods of classification trees, logistic regression and KNN, in R. We also divide the data into training and validation divisions and create classification models.

Data exploration

After talking about the data and information other banks use to determine which customers get credits, we'll compare it to the data provided in the GC dataset and carefully examine and investigate the various variables. In addition to examining the correlations between the variables, we also want to make sure that the dataset is free of major outliers, inaccurate or misleading data, and confusing or significant outliers. All of these factors should enable us to identify a good creditor (grant loan) or a bad creditor in the GC dataset (reject application). Checking account status and average savings account balance, two categorical variables in the dataset, may cause the results to be skewed. The highest categories, 3 and 4, were given to not having a checking account and not having a saving account or having one that is unknown. In order to resolve this issue, we changed the hierarchical order of both sets of data by designating category "0" for neither no checking account nor no savings account, respectively, and modifying the other categories as necessary.

We separate the datasets into training set and validation set by 4:6

```
set.seed(12345)
train.index <- sample(row.names(GC.df), 0.6*dim(GC.df)[1])
valid.index <- setdiff(row.names(GC.df), train.index)
train.df <- GC.df[train.index, ]
valid.df <- GC.df[valid.index, ]
```

Data Selection

Therefore we decide create 2 datasets. All variables are included in the first dataset. We chose the entire dataset because we wanted an extreme point to gauge how well our additional data reduction performed. Furthermore, as was already said, the majority of the data in GC's dataset is crucial for other banks operating in the same sector.

For the second model, we decide to take out the following variables:

```
GC.df = subset(GC.df, select = -c(1,6,8,10,20,26,29,30,31) )
```

#6 We delete USED_CAR since we think it is duplicated with NEW_CAR, even though a new car may be more expensive than a used car on average.

#8 We delete RADIO/TV variable since we think it is repetitive compare to FURNITURE, since radio/TV can be part of the furniture.

#10 We delete RETRAINING and contain the EDUCATION. Since it is all part of the education even retraining means the age of applicants can be bigger in general. But the purpose are similar.

#26 We delete OWN_RES since it is overlaping with REAL-ESTATE, because they all in a way indicate that whether applicants own property.

31.We delete foreign worker since its low observation which may cause the bias in our model.

#20/#29/#30 We delete the Present resident/Telephone/People for whom liable to provide maintenance due to their low correlation to responsive variable.

Data Normalization

We normalize both datasets due to the high variance created by numerical variables.

```
norm.values <- preProcess(train.df[, c(2,7,10,18)], method=c("center", "scale"))

train.norm.df[, c(2,7,10,18)] <- predict(norm.values, train.df[, c(2,7,10,18)])

valid.norm.df[, c(2,7,10,18)] <- predict(norm.values, valid.df[, c(2,7,10,18)])

GC.norm.df[, c(2,7,10,18)] <- predict(norm.values, GC.df[, c(2,7,10,18)])
```

Describe Classification tree and Logistic model with details

This model has 1000 observations and 32 variables like CHK_ACCT, DURATION, HISTORY, NEW_CAR ⋯, and we can use the function > describe(GC.df) to know some details about these parameter variables, as shown below:

```
> describe(GC.df)
GC.df

 32  Variables      1000  Observations
---------------------------------------------------------------------------------------------------------------
OBS.
       n  missing distinct     Info     Mean      Gmd      .05      .10      .25      .50      .75      .90
    1000        0     1000        1    500.5    333.7    50.95   100.90   250.75   500.50   750.25   900.10
     .95
  950.05

lowest :    1    2    3    4    5, highest:  996  997  998  999 1000
---------------------------------------------------------------------------------------------------------------
CHK_ACCT
       n  missing distinct
    1000        0        4

Value          0     1     2     3
Frequency    274   269    63   394
Proportion 0.274 0.269 0.063 0.394
---------------------------------------------------------------------------------------------------------------
DURATION
       n  missing distinct     Info     Mean      Gmd      .05      .10      .25      .50      .75      .90
    1000        0       33    0.985     20.9    12.98        6        9       12       18       24       36
     .95
      48

lowest :  4  5  6  7  8, highest: 47 48 54 60 72
---------------------------------------------------------------------------------------------------------------
HISTORY
       n  missing distinct
    1000        0        5

lowest : 0 1 2 3 4, highest: 0 1 2 3 4

Value          0     1     2     3     4
Frequency     40    49   530    88   293
Proportion 0.040 0.049 0.530 0.088 0.293
---------------------------------------------------------------------------------------------------------------
NEW_CAR
       n  missing distinct
    1000        0        2

Value          0     1
Frequency    766   234
Proportion 0.766 0.234
---------------------------------------------------------------------------------------------------------------
```

We don't show all the information for each variable in this report, you can get all the details in the code using the method above. Through analysis, removing some irrelevant variables like OBS#, we will select 23 variables with important influence among the 32 variables as predictors.

Classification trees in R

Using GC.df <- GC.df[, -c(1,6,8,10,20,26,29,30,31)], we remove some low-impact variables, and then, we can in the function summary(GC.df) , to view the specific information of the 23 variables after screening.

```
> summary(GC.df)
 CHK_ACCT    DURATION     HISTORY NEW_CAR FURNITURE EDUCATION     AMOUNT      SAV_ACCT EMPLOYMENT
 0:274    Min.   : 4.0   0: 40   0:766   0:819     0:950     Min.   :  250   0:603    0: 62
 1:269    1st Qu.:12.0   1: 49   1:234   1:181     1: 50     1st Qu.: 1366   1:103    1:172
 2: 63    Median :18.0   2:530                               Median : 2320   2: 63    2:339
 3:394    Mean   :20.9   3: 88                               Mean   : 3271   3: 48    3:174
          3rd Qu.:24.0   4:293                               3rd Qu.: 3972   4:183    4:253
          Max.   :72.0                                       Max.   :18424
  INSTALL_RATE   MALE_DIV MALE_SINGLE MALE_MAR_or_WID CO.APPLICANT GUARANTOR REAL_ESTATE
 Min.   :1.000   0:950    0:452       0:908           0:959        0:948     0:718
 1st Qu.:2.000   1: 50    1:548       1: 92           1: 41        1: 52     1:282
 Median :3.000
 Mean   :2.973
 3rd Qu.:4.000
 Max.   :4.000
 PROP_UNKN_NONE      AGE         OTHER_INSTALL RENT     NUM_CREDITS    JOB      RESPONSE
 0:846          Min.   :19.00   0:814         0:821   Min.   :1.000   0: 22   0:300
 1:154          1st Qu.:27.00   1:186         1:179   1st Qu.:1.000   1:200   1:700
                Median :33.00                         Median :1.000   2:630
                Mean   :35.55                         Mean   :1.407   3:148
                3rd Qu.:42.00                         3rd Qu.:2.000
                Max.   :75.00                         Max.   :4.000
```

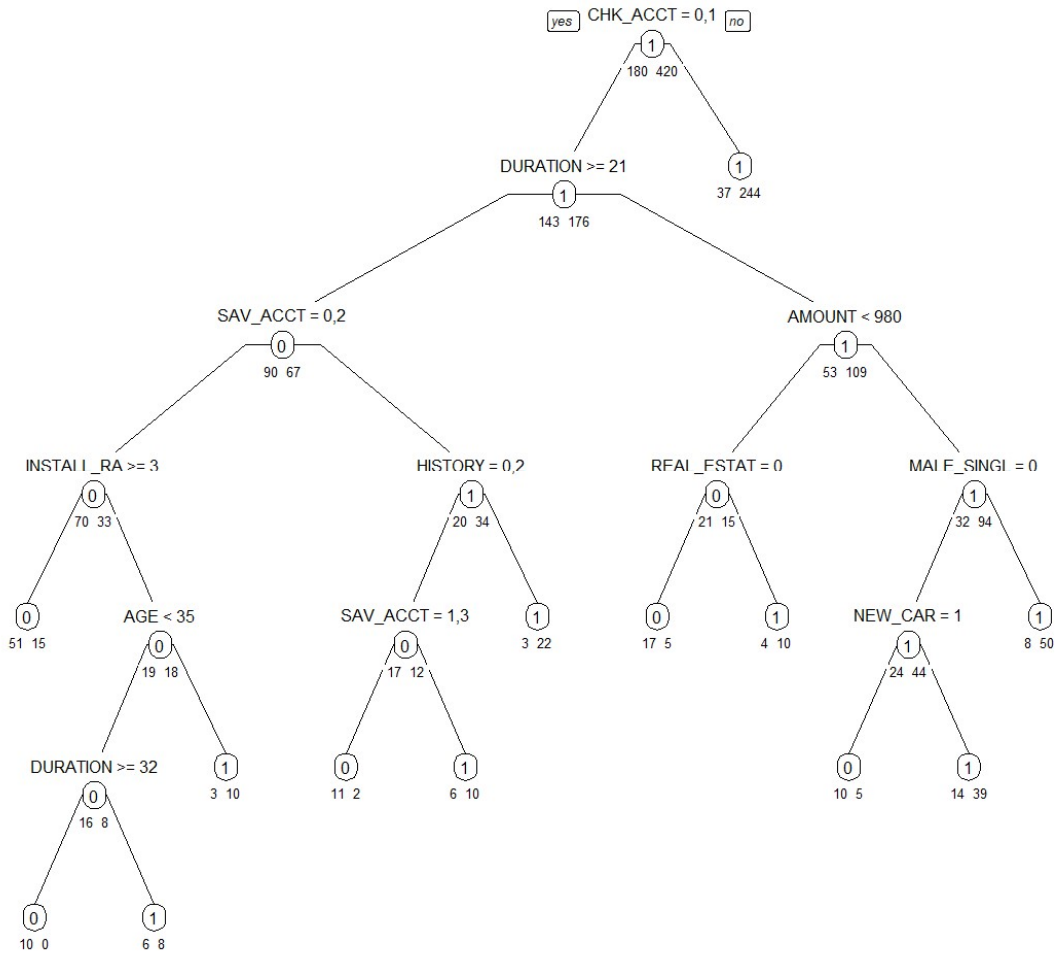
> class.tree

```
> class.tree
n= 600

node), split, n, loss, yval, (yprob)
      * denotes terminal node

 1) root 600 180 1 (0.3000000 0.7000000)
   2) CHK_ACCT=0,1 319 143 1 (0.4482759 0.5517241)
     4) DURATION>=20.5 157  67 0 (0.5732484 0.4267516)
       8) SAV_ACCT=0,2 103  33 0 (0.6796117 0.3203883)
        16) INSTALL_RATE>=2.5 66  15 0 (0.7727273 0.2272727) *
        17) INSTALL_RATE< 2.5 37  18 0 (0.5135135 0.4864865)
          34) AGE< 34.5 24   8 0 (0.6666667 0.3333333)
            68) DURATION>=31.5 10   0 0 (1.0000000 0.0000000) *
            69) DURATION< 31.5 14   6 1 (0.4285714 0.5714286) *
          35) AGE>=34.5 13   3 1 (0.2307692 0.7692308) *
       9) SAV_ACCT=1,3,4 54  20 1 (0.3703704 0.6296296)
        18) HISTORY=0,2 29  12 0 (0.5862069 0.4137931)
          36) SAV_ACCT=1,3 13   2 0 (0.8461538 0.1538462) *
          37) SAV_ACCT=4 16   6 1 (0.3750000 0.6250000) *
        19) HISTORY=1,3,4 25   3 1 (0.1200000 0.8800000) *
     5) DURATION< 20.5 162  53 1 (0.3271605 0.6728395)
      10) AMOUNT< 979.5 36  15 0 (0.5833333 0.4166667)
        20) REAL_ESTATE=0 22   5 0 (0.7727273 0.2272727) *
        21) REAL_ESTATE=1 14   4 1 (0.2857143 0.7142857) *
      11) AMOUNT>=979.5 126  32 1 (0.2539683 0.7460317)
        22) MALE_SINGLE=0 68  24 1 (0.3529412 0.6470588)
          44) NEW_CAR=1 15   5 0 (0.6666667 0.3333333) *
          45) NEW_CAR=0 53  14 1 (0.2641509 0.7358491) *
        23) MALE_SINGLE=1 58   8 1 (0.1379310 0.8620690) *
   3) CHK_ACCT=2,3 281  37 1 (0.1316726 0.8683274) *
```

> prp(class.tree, type = 1, extra = 1, under = TRUE, split.font = 1, varlen = -10)

```
#splitting the data into 60% training and 40% validation data
set.seed(111) #in order to randomize
index <- createDataPartition(GC.df$RESPONSE, p=0.60, list = FALSE)
training <- GC.df[index,]
validation <- GC.df[-index ,]

#####Classification Tree#####
library(rpart)
library(rpart.plot)

class.tree <- rpart(RESPONSE ~ ., data = training, control = rpart.control(maxdepth = 30), method = "class")
class.tree
prp(class.tree, type = 1, extra = 1, under = TRUE, split.font = 1, varlen = -10)
```

Classification Trees



Logistical Regression in R

The goal of the logistic regression approach is to forecast the classification's result based on predictor factors. We do a regression analysis on our data sample to determine the correlation between the applicant's categorization and the explanatory factors, in this example, the responses to our questionnaire. The dependent variable is transformed into a probability score that indicates the likelihood that a loan will be approved for an applicant with a certain combination of qualities based on prior judgments.

Implementation

> logistic.GC

> summary(logistic.GC)

This way can help us know the relationship between German Credit and other factors.

```
> logistic.GC <- glm(RESPONSE ~ ., data = train.df, family = "binomial")
> logistic.GC

Call:  glm(formula = RESPONSE ~ ., family = "binomial", data = train.df)

Coefficients:
     (Intercept)          CHK_ACCT          DURATION           HISTORY          NEW_CAR          USED_CAR
       -0.6767092         0.5818602        -0.0183430         0.5250817        -0.2271161         1.2708336
        FURNITURE          RADIO.TV         EDUCATION        RETRAINING            AMOUNT          SAV_ACCT
         0.1109019         0.3841585        -0.2449803        -0.4187910        -0.0001532         0.2357953
       EMPLOYMENT      INSTALL_RATE      CO.APPLICANT         GUARANTOR  PRESENT_RESIDENT       REAL_ESTATE
         0.2278111        -0.3327431        -0.3475391         1.6676609        -0.0227440         0.2495687
              AGE     OTHER_INSTALL           OWN_RES       NUM_CREDITS               JOB    NUM_DEPENDENTS
         0.0129467        -0.6458734         0.6699873        -0.2176836        -0.1151231        -0.0612055
        TELEPHONE            FOREIGN
         0.5265383         1.3963030

Degrees of Freedom: 599 Total (i.e. Null);  574 Residual
Null Deviance:      741.3
Residual Deviance: 525.2        AIC: 577.2
```

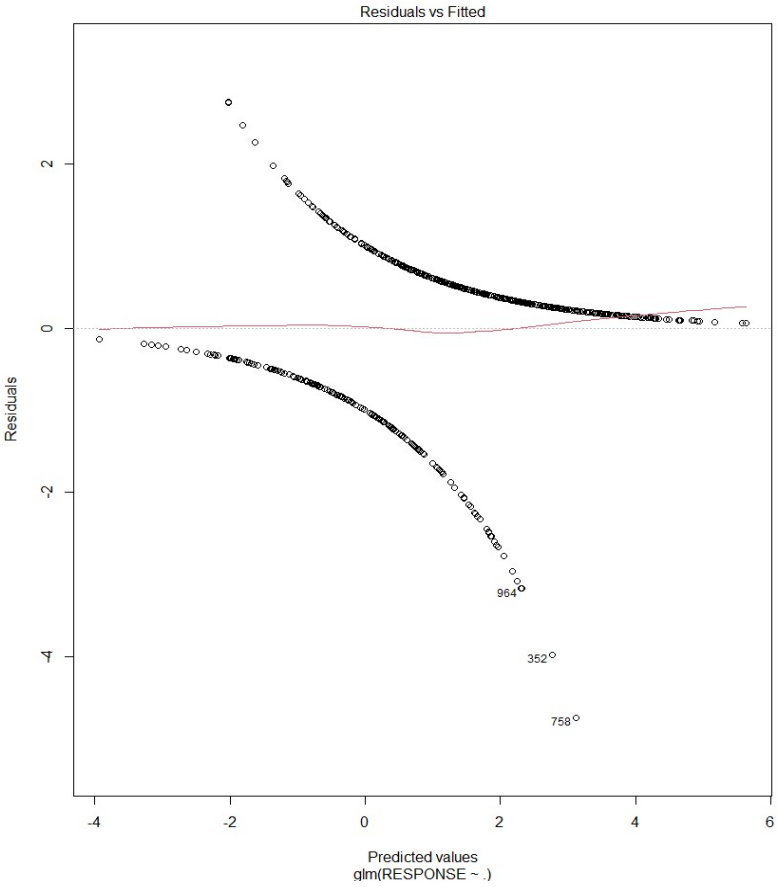
Some details to explain

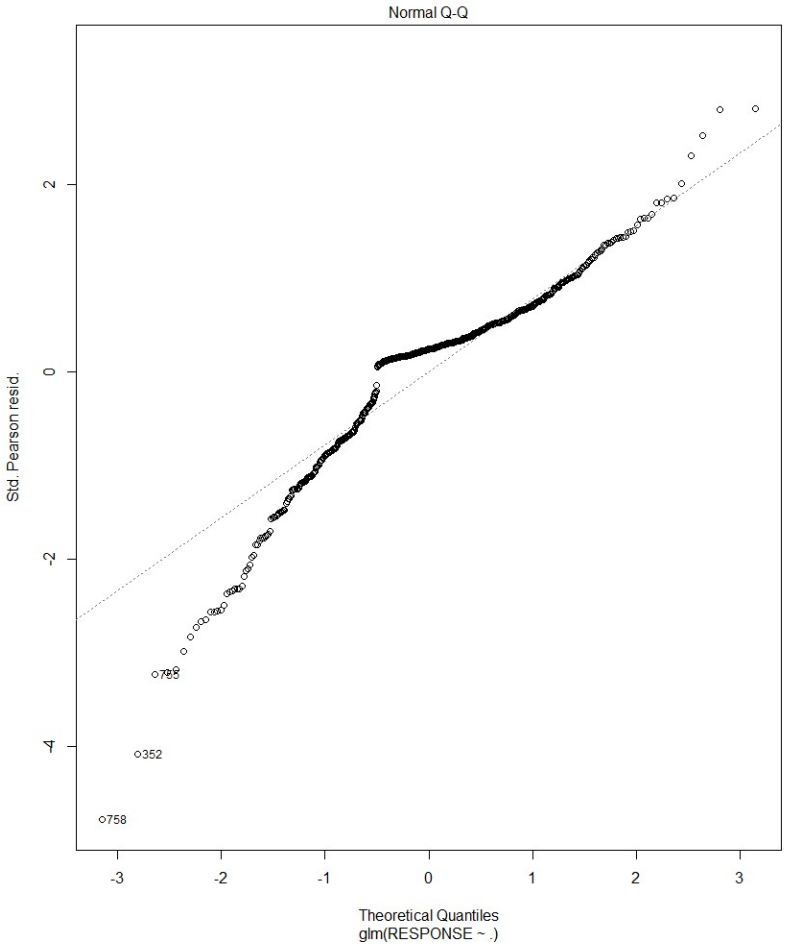
Each of our models received the logistic regression treatment. The error rate and net profitability of our sample model are determined by one setscrew for our statistical approach. The cutoff value is this.

The cutoff for the likelihood of success is determined by the value, therefore admitted candidates. The likelihood of approving a candidate increases with the chosen cutoff value.

As a result, a lower cutoff value might raise the possibility of a type two mistake, which is the acceptance of a candidate without the necessary loan attributes. The best cutoff value in our report depends on the particular data sample.

```
> summary(logistic.GC)

Call:
glm(formula = RESPONSE ~ ., family = "binomial", data = train.df)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.5137  -0.6848   0.3403   0.6902   2.0747

Coefficients:
                   Estimate  Std. Error z value      Pr(>|z|)
(Intercept)      -0.67670923  0.93717799  -0.722       0.47025
CHK_ACCT          0.58186023  0.09702108   5.997 0.00000000201 ***
DURATION         -0.01834296  0.01145019  -1.602       0.10916
HISTORY           0.52508171  0.12311644   4.265 0.00001999742 ***
NEW_CAR          -0.22711608  0.50883077  -0.446       0.65535
USED_CAR          1.27083363  0.65182013   1.950       0.05122 .
FURNITURE         0.11090189  0.53169796   0.209       0.83478
RADIO.TV          0.38415847  0.51401615   0.747       0.45484
EDUCATION        -0.24498028  0.67407096  -0.363       0.71628
RETRAINING       -0.41879104  0.56120710  -0.746       0.45553
AMOUNT           -0.00015324  0.00005347  -2.866       0.00416 **
SAV_ACCT          0.23579526  0.07833421   3.010       0.00261 **
EMPLOYMENT        0.22781110  0.09834471   2.316       0.02053 *
INSTALL_RATE     -0.33274314  0.11537481  -2.884       0.00393 **
CO.APPLICANT     -0.34753908  0.52492821  -0.662       0.50793
GUARANTOR         1.66766086  0.66818092   2.496       0.01257 *
PRESENT_RESIDENT -0.02274398  0.11298383  -0.201       0.84046
REAL_ESTATE       0.24956873  0.28474598   0.876       0.38078
AGE               0.01294667  0.01049207   1.234       0.21722
OTHER_INSTALL    -0.64587339  0.26872167  -2.404       0.01624 *
OWN_RES           0.66998728  0.25799382   2.597       0.00941 **
NUM_CREDITS      -0.21768363  0.21314509  -1.021       0.30712
JOB              -0.11512307  0.18889708  -0.609       0.54223
NUM_DEPENDENTS   -0.06120552  0.32292124  -0.190       0.84967
TELEPHONE         0.52653834  0.26749596   1.968       0.04902 *
FOREIGN           1.39630304  0.79453891   1.757       0.07885 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 741.31  on 599  degrees of freedom
Residual deviance: 525.24  on 574  degrees of freedom
AIC: 577.24

Number of Fisher Scoring iterations: 5
```


Residuals vs Fitted
Residuals
Predicted values
glm(RESPONSE ~ .)
964
352
758


Normal Q-Q
Std. Pearson resid.
Theoretical Quantiles
glm(RESPONSE ~ .)
352
758

KNN Model

When using KNN method to classify the customers, we decided to use two datasets we created before in order to get different and various insights from forming two models. And comparing the statistics between two models will help us to decide the final KNN model for prediction.

We decide to test K value from 1 to 44 for both sets and find the optimal accuracy value. And we found out the accuracy bestly achieve around 0.77. And we choose the smallest K to improve the performance.

```
accuracy.df <- data.frame(k = seq(1, 44, 1), accuracy = rep(0, 44))
for(i in 1:44) {

  knn.pred <- knn(train.norm.df[,1:22], valid.norm.df[,1:22],
                  cl = train.norm.df[,23], k = i)
  accuracy.df[i, 2] <- confusionMatrix(knn.pred, as.factor(valid.norm.df[,
23]))$overall[1]
}
accuracy.df
plot(accuracy.df)
```

BEST K

By generating the dataframe to track accuracy rate. We made final decesion

4	4	0.7450
5	5	0.7600
6	6	0.7550
7	7	0.7350
8	8	0.7550
9	9	0.7425
10	10	0.7400
11	11	0.7600
12	12	0.7600
13	13	0.7625
14	14	0.7600
15	15	0.7675
16	16	0.7650
17	17	0.7675
18	18	0.7600
19	19	0.7625
20	20	0.7575
21	21	0.7700
22	22	0.7650
23	23	0.7700
24	24	0.7575

Model with First dataset

k	accuracy
1	0.7025
2	0.6950
3	0.7275
4	0.7125
5	0.7300
6	0.7275
7	0.7375
8	0.7375
9	0.7425
10	0.7450
11	0.7575
12	0.7500
13	0.7475
14	0.7625
15	0.7600
16	0.7650
17	0.7550
18	0.7550
19	0.7675
20	0.7675
21	0.7625
22	0.7625
23	0.7600
24	0.7700

Model with Second dataset

We found out for the first dataset which include all variables, best K is 21

knn.pred.whole <- knn(train.whole.norm.df[,1:30], valid.whole.norm.df[,1:30], cl=train.whole.norm.df[,31], k=21)

We use K= 24 for the second model using second datasets with variables selection.
knn.pred1 <- knn(train.norm.df[,1:22], valid.norm.df[,1:22], cl=train.norm.df[,23], k=24)

Model Selection

```
> table(actual=valid.whole.norm.df$RESPONSE,predicted=knn.pred.whole )
      predicted
actual   0   1
     0  51  66
     1  26 257
> table(actual=valid.norm.df$RESPONSE,predicted=knn.pred1)
      predicted
actual   0   1
     0  49  68
     1  28 255
```

By comparing two models we build. We found the first model has better performance.

3. Choose one model from each technique and report the confusion matrix and the cost/gain matrix for the validation data. Which technique has the highest net profit?

KNN Model

Confusion Matrix

	predicted	
actual	0	1
0	51	66
1	26	257

Profit/gain Matrix

	Predicted(decision)	
Actual	Good(accept)	Bad(reject)
Good	100*257 = 25700	0
Bad	-500*66 = -33000	0

PROFIT:-7300

Logistical Regression

Confusion Matrix

	actual	
predicted	0	1
0	52	43
1	63	242

Profit/gain Matrix

	Predicted(decision)	
Actual	Good(accept)	Bad(reject)
Good	100*242 = 24200	0
Bad	-500*63 = -31500	0

PROFIT:-7300

Classification trees

Confusion Matrix

	actual	
predicted	0	1
0	57	35
1	63	245

Profit/gain Matrix

	Predicted(decision)	
Actual	Good(accept)	Bad(reject)
Good	100*245 = 24500	0
Bad	-500*63 = -31500	0

PROFIT:-7000

In conclusion, the Classification tree model has the highest profit which is -7000. Though they are all negative