Exploratory Data Analysis

Zhiquan Cui

2023-04-04

Contents

1 Load Required Libraries

```
library(dplyr)
library(insight)
library(knitr)
library(kableExtra)
library(ggplot2)
library(tidyverse)
library(corrplot)
library(patchwork)
library(rcompanion)
library(gridExtra)
```

2 Load Data & Inspect Variables

```
# Read the data
data <- read.csv("Credit.csv")
# Check the number of observations and number of variables
n <- nrow(data)
m <- ncol(data)
n
```

```
## [1] 1000
```

```
m
```

```
## [1] 21
```

```
# Check the data
kable(head(data[, 1:8]), format = "latex", align=rep("c", 8), booktabs=TRUE)
```

status	duration	credit_history	purpose	amount	savings	employment_duration	installment_rate
1	18	4	2	1049	1	2	4
1	9	4	0	2799	1	3	2
2	12	2	9	841	2	4	2
1	12	4	0	2122	1	3	3
1	12	4	0	2171	1	3	4
1	10	4	0	2241	1	2	1

```
kable(head(data[, 9:14]), format = "latex", align=rep("c", 6), booktabs=TRUE)
```

personal_status_sex	other_debtors	present_residence	property	age	other_installment_plans
2	1	4	2	21	3
3	1	2	1	36	3
2	1	4	1	23	3
3	1	2	1	39	3
3	1	4	2	38	1
3	1	3	1	48	3

```
kable(head(data[, 15:21]), format = "latex", align=rep("c", 7), booktabs=TRUE)
```

housing	number_credits	job	people_liable	telephone	foreign_worker	credit_risk
1	1	3	2	1	2	1
1	2	3	1	1	2	1
1	1	2	2	1	2	1
1	2	2	1	1	1	1
2	2	2	2	1	1	1
1	2	2	1	1	1	1

```
# Check invalid or missing values
anyNA(data)
```

```
## [1] FALSE
```

```
# Check the data type of each column
sapply(data, class)
```

```
##                  status                duration          credit_history
##               "integer"               "integer"               "integer"
##                 purpose                  amount                 savings
##               "integer"               "integer"               "integer"
##     employment_duration        installment_rate     personal_status_sex
##               "integer"               "integer"               "integer"
##           other_debtors       present_residence                property
##               "integer"               "integer"               "integer"
##                     age other_installment_plans                 housing
##               "integer"               "integer"               "integer"
##          number_credits                     job           people_liable
##               "integer"               "integer"               "integer"
##               telephone          foreign_worker             credit_risk
##               "integer"               "integer"               "integer"
```

As we can see from the above outputs, there is no NaN values so the data is clean. And all of the columns are of type integer. Some of them are quantitative variable while some of them are qualitative variables. Here is a summary of the variables:

- status: status of the debtor's checking account with the bank (categorical)
- duration: credit duration in months (quantitative)
- credit_history: history of compliance with previous or concurrent credit contracts (categorical)
- purpose: purpose for which the credit is needed (categorical)
- amount: credit amount in DM (quantitative; result of monotonic transformation; actual data and type of transformation unknown)
- savings: debtor's savings (categorical)
- employment_duration: duration of debtor's employment with current employer (ordinal; discretized quantitative)
- installment_rate: credit installments as a percentage of debtor's disposable income (ordinal; discretized quantitative)
- personal_status_sex: combined information on sex and marital status (categorical)
- other_debtors: is there another debtor or a guarantor for the credit? (categorial)

- present_residence: length of time (in years) the debtor lives in the present residence (ordinal; discretized quantitative)
- property: the debtor's most valuable property (ordinal)
- age: age in years (quantitative)
- other_installment_plans: installment plans from providers other than the credit-giving bank (categorical)
- housing: type of housing the debtor lives in (categorical)
- number_credits: number of credits including the current one the debtor has (or had) at the bank (ordinal; discretized quantitative)
- job: quality of debtor's job (ordinal)
- people_liable: number of persons who financially depend on the debtor (binary; discretized quantitative)
- telephone: is there a telephone landline registered on the debtor's name? (binary)
- foreign_ worker: is the debtor a foreign worker? (binary)
- credit_risk: has the credit contract been complied with (good) or not (bad)? (binary)

We can see that the **quantitative variables** include duration, amount and age, while **qualitative variables** include status, credit_history, purpose, savings, employment_duration, installment_rate, personal_status_sex, other_debtors, present_residence, property, other_installment_plans, housing, number_credits, job, people_liable, telephone, foreign_worker and credit_risk.

3 Univariate Data Analysis & Visualization

3.1 Histogram of Quantitative Variables

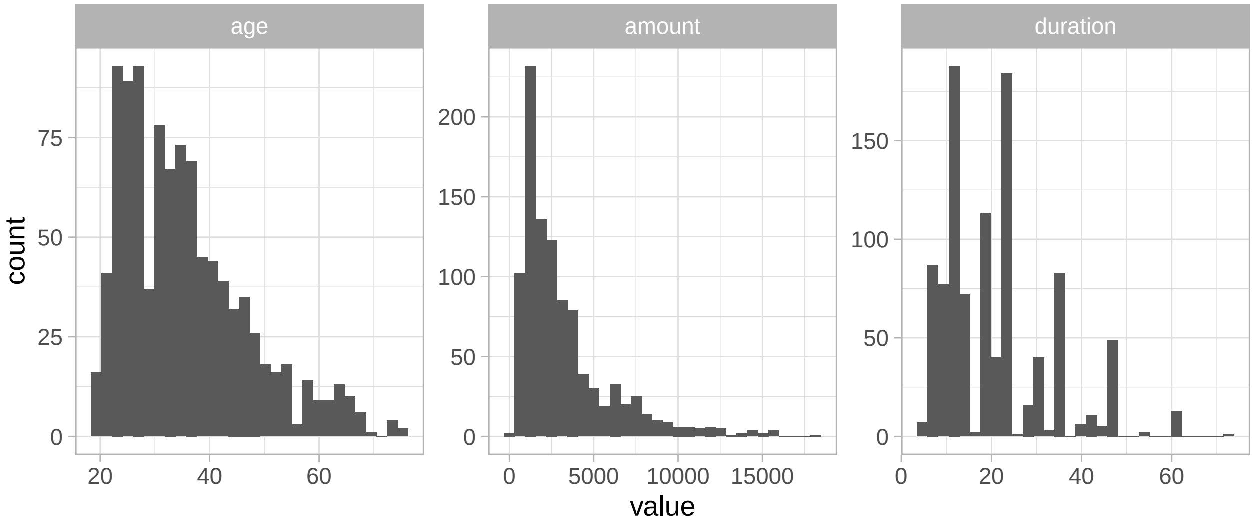
First we will perform univariate analysis on each of the variables and look at their distribution. Here is the summary statistics:

```
quant_vars <- c("duration", "amount", "age")
qual_vars <- c("status", "credit_history", "purpose", "savings", "employment_duration",
               "installment_rate", "personal_status_sex", "other_debtors", "present_residence",
               "property", "other_installment_plans", "housing", "number_credits", "job",
               "people_liable", "telephone", "foreign_worker", "credit_risk")
summary(data[, quant_vars])
```

```
##     duration        amount           age
##  Min.   : 4.0   Min.   :  250   Min.   :19.00
##  1st Qu.:12.0   1st Qu.: 1366   1st Qu.:27.00
##  Median :18.0   Median : 2320   Median :33.00
##  Mean   :20.9   Mean   : 3271   Mean   :35.54
##  3rd Qu.:24.0   3rd Qu.: 3972   3rd Qu.:42.00
##  Max.   :72.0   Max.   :18424   Max.   :75.00
```

Next, let us check the histograms of the quantitative variables:

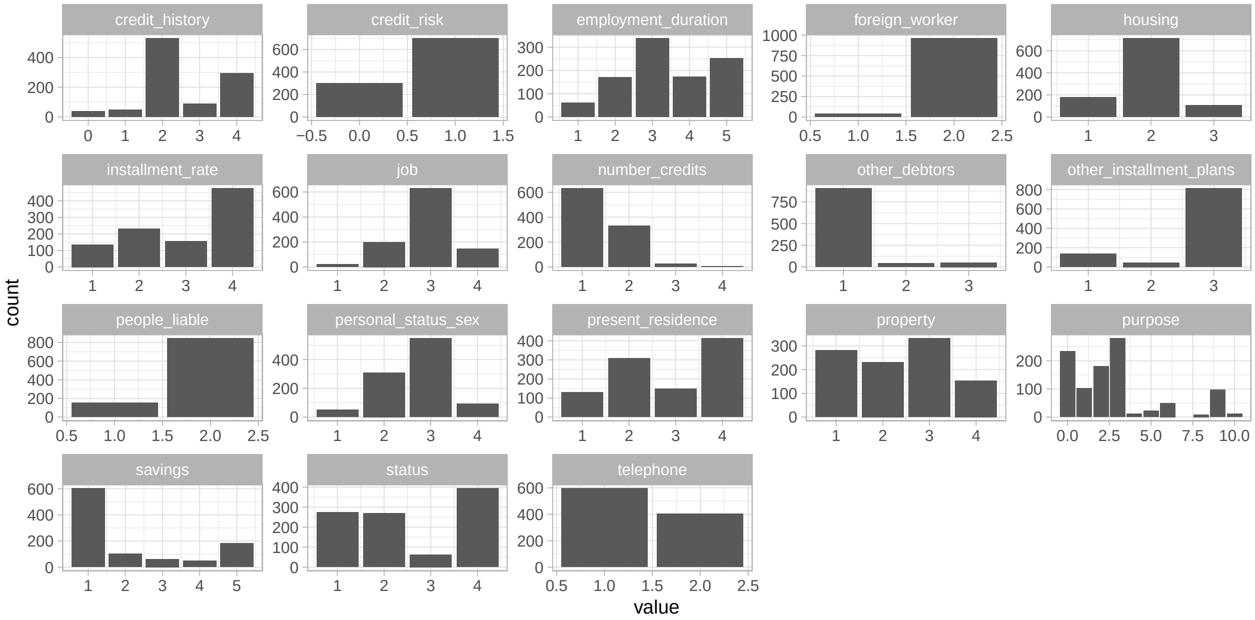
```
data[, quant_vars] %>%
  gather() %>%
  ggplot(aes(value)) +
    facet_wrap(~ key, scales = "free") +
    geom_histogram() +
    theme_light()
```



3.2 Barplot of Qualitative Variables

Then, let us check the barplots of qualitative variables:

```
data[, qual_vars] %>%
  gather() %>%
  ggplot(aes(value)) +
    facet_wrap(~ key, scales = "free") +
    geom_bar() +
    theme_light()
```

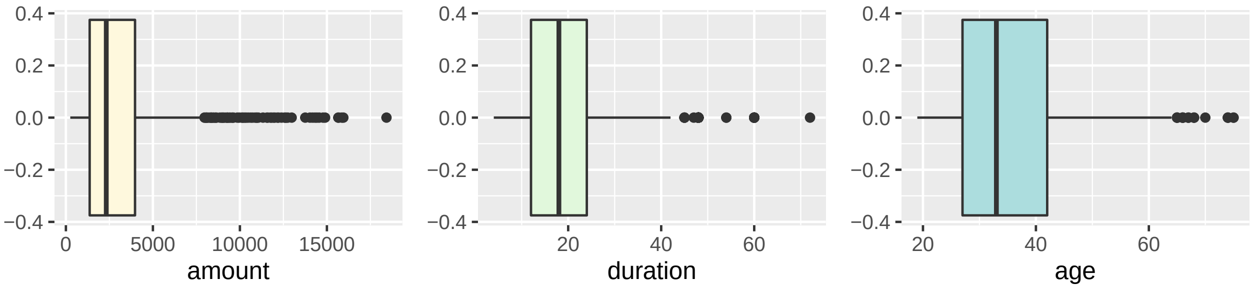


As we can see, the response variable credit risk is a binary variable while we have more than 2 predictors. This indicates that it is a good idea to use Multiple Logistic Regression as our model.

3.3 Boxplot of Quantitative Variables

After checking the histograms and barplots, we will check the boxplots of the quantitative variables. Here we will not check barplots for qualitative variables because it only makes sense to examine the median, first and third quartiles and maximum value for quantitative variables.

```r
g1 <- ggplot(data, aes(x = amount)) + geom_boxplot(fill="#FEF8DD")
g2 <- ggplot(data, aes(x = duration)) + geom_boxplot(fill="#E1F8DC")
g3 <- ggplot(data, aes(x = age)) + geom_boxplot(fill="#ACDDDE")
g1 + g2 + g3
```



From the above box plots, we can see that there are a few outliers for the variable amount. If we look at the histogram of variable amount, we can see that it is a right skewed distribution with a long right tail, which results in these outliers.

3.4 Sample Odds of Binary Variables

For binary variables people_liable, telephone, foreign_worker and credit_risk, we can calculate and interpret the sample odds:

```r
binary_var <- c("Statistics", "people_liable", "telephone", "foreign_worker", "credit_risk")
odds <- c("Sample Odds")
for (var in binary_var[2:5]) {
  if (var == "credit_risk") {y <- sum(data[, var] == 1)}
  else {y <- sum(data[, var] == 2)}
  n <- length(data[, var])
  odds <- append(odds, round(y / (n - y), 2))
}
kable(data.frame(t(odds)), col.names = binary_var, format = "latex") %>%
  kable_styling(position = "center", latex_options = "hold_position") %>% row_spec(0, bold = TRUE)
```

Statistics	**people_liable**	**telephone**	**foreign_worker**	**credit_risk**
Sample Odds	5.45	0.68	26.03	2.33

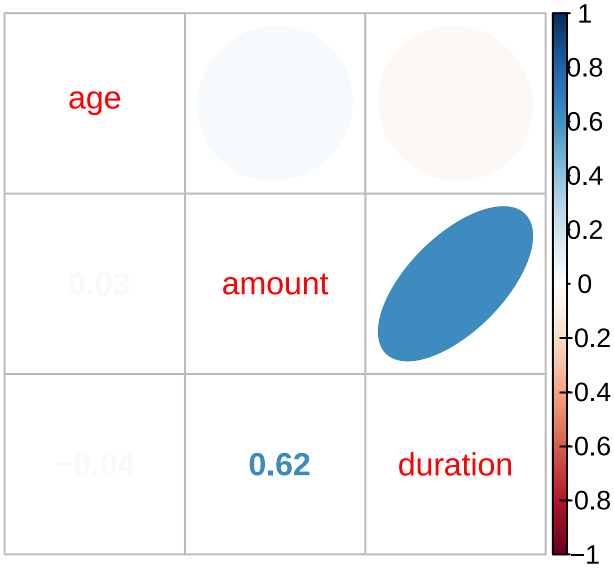
Based on our sample, the estimated probability of a person to have good credit is 2.33 times as likely as having a bad credit. Similarly, the estimated probability of a person to have a telephone landline registered on his/her name is 0.68 times as likely as not having such a telephone landline.

4 Multivariate Data Analysis & Visualization

4.1 Quantitative Variable

First, let us look at the correlation plots of the quantitative variables.

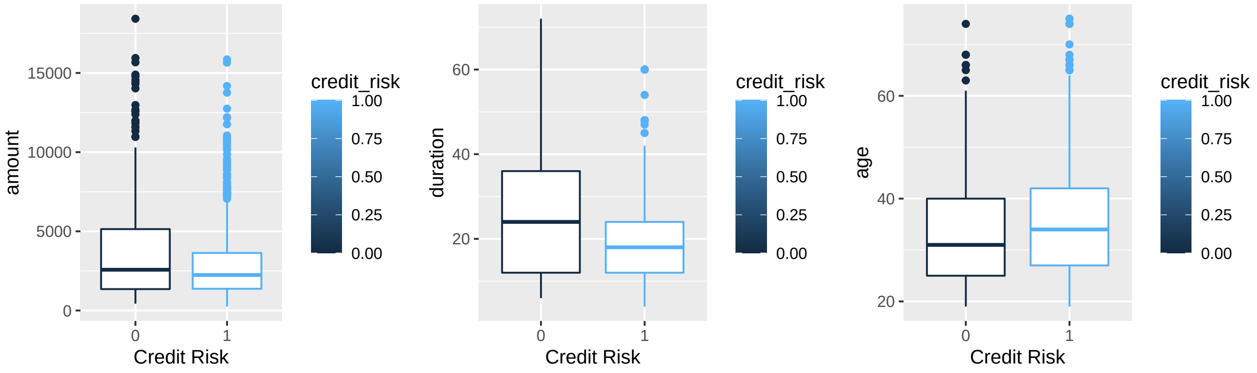
```
corrplot.mixed(cor(data[quant_vars]), lower='number', upper='ellipse', order='AOE')
```



From the above correlation plot, we can see that the correlation coefficient between amount and duration is as high as 0.62, which indicates a strong positive correlation between the two variables. This also makes sense intuitively because the longer credit duration one has in months, he/she will have a higher chance to build up his/her credit and obtain a higher credit amount. Similarly, if one has a high credit amount, then he/she is more likely to have a long credit duration. In order to avoid multicollinearity, we will consider droping one of amount and duration in our model. However, before making a decision, we shall examine the side by side box plots.

```
g1 <- ggplot(data, aes(x=as.factor(credit_risk), y=amount, color=credit_risk)) +
      geom_boxplot() + xlab("Credit Risk")
g2 <- ggplot(data, aes(x=as.factor(credit_risk), y=duration, color=credit_risk)) +
      geom_boxplot() + xlab("Credit Risk")
g3 <- ggplot(data, aes(x=as.factor(credit_risk), y=age, color=credit_risk)) +
      geom_boxplot() + xlab("Credit Risk")

grid.arrange(g1, g2, g3, nrow=1)
```



From the above side by side box plots, we can see that for variables duration and age, there are significant differences on the box plots between two levels of credit risks. This indicates a significant association between credit risk and these two variables. However, we don't see a significant difference between two credit risk levels for variable amount.

Therefore, we will drop the variable amount.

4.2 Qualitative Variables

After examining the quantitative variables, we will now look at the qualitative variables. Since they are not continuous and numeric data, we should not use the same methodology as above. Instead, we will use Pearson's Chi-sq Test of Indepence and Cramer's V designed for qualitative variables to examine the data.

```
Pearson_chisq_test <- data.frame(matrix(0, ncol = length(qual_vars),
                                 nrow = length(qual_vars)), row.names = qual_vars)
colnames(Pearson_chisq_test) <- qual_vars
for (var in qual_vars) {
  for (var_2 in qual_vars) {
    test <- chisq.test(table(data[, var], data[, var_2]), simulate.p.value = TRUE)
    Pearson_chisq_test[var, var_2] <- test$p.value
  }
}
kable(Pearson_chisq_test[, 1:6], format = "latex", booktabs=TRUE) %>%
  kable_styling(font_size = 6, latex_options = "hold_position")
```

	status	credit__history	purpose	savings	employment__duration	installment__rate
status	0.0004998	0.0004998	0.0004998	0.0004998	0.0109945	0.4302849
credit__history	0.0004998	0.0004998	0.0004998	0.2053973	0.0019990	0.6956522
purpose	0.0009995	0.0004998	0.0004998	0.0554723	0.0109945	0.0019990
savings	0.0004998	0.1899050	0.0399800	0.0004998	0.0274863	0.3263368
employment__duration	0.0074963	0.0004998	0.0079960	0.0274863	0.0004998	0.0029985
installment__rate	0.4457771	0.6826587	0.0009995	0.3248376	0.0004998	0.0004998
personal__status__sex	0.1409295	0.0119940	0.0004998	0.4577711	0.0004998	0.0009995
other__debtors	0.0009995	0.0584708	0.0004998	0.0224888	0.0794603	0.9755122
present__residence	0.0004998	0.1089455	0.0079960	0.1364318	0.0004998	0.4257871
property	0.0459770	0.0839580	0.0004998	0.1019490	0.0004998	0.4607696
other__installment__plans	0.2598701	0.0004998	0.0064968	0.9995002	0.2883558	0.6136932
housing	0.0044978	0.0134933	0.0004998	0.8235882	0.0004998	0.1109445
number__credits	0.0429785	0.0004998	0.2783608	0.1144428	0.0004998	0.4302849
job	0.0449775	0.3893053	0.0004998	0.3453273	0.0004998	0.0624688
people__liable	0.1134433	0.0499750	0.0009995	0.8865567	0.0459770	0.1029485
telephone	0.0844578	0.2818591	0.0004998	0.0689655	0.0004998	0.4542729
foreign__worker	0.1149425	0.3988006	0.0029985	0.9160420	0.4607696	0.0029985
credit__risk	0.0004998	0.0004998	0.0014993	0.0004998	0.0009995	0.1429285

```
kable(Pearson_chisq_test[, 7:12], format = "latex", booktabs=TRUE) %>%
  kable_styling(font_size = 6, latex_options = "hold_position")
```

	personal_status_sex	other_debtors	present_residence	property	other_installment_plans	housing
status	0.1504248	0.0029985	0.0019990	0.0424788	0.2833583	0.0024988
credit_history	0.0174913	0.0499750	0.1209395	0.0814593	0.0004998	0.0194903
purpose	0.0004998	0.0014993	0.0039980	0.0004998	0.0039980	0.0004998
savings	0.4692654	0.0219890	0.1314343	0.0804598	0.9990005	0.8235882
employment_duration	0.0004998	0.0879560	0.0004998	0.0004998	0.2718641	0.0004998
installment_rate	0.0004998	0.9785107	0.4232884	0.4652674	0.6211894	0.1144428
personal_status_sex	0.0004998	0.6101949	0.0004998	0.0004998	0.4447776	0.0004998
other_debtors	0.6301849	0.0004998	0.6481759	0.0004998	0.2168916	0.1299350
present_residence	0.0004998	0.6426787	0.0004998	0.0004998	0.7531234	0.0004998
property	0.0004998	0.0004998	0.0004998	0.0004998	0.0034983	0.0004998
other_installment_plans	0.4462769	0.2193903	0.7366317	0.0029985	0.0004998	0.0014993
housing	0.0004998	0.1354323	0.0004998	0.0004998	0.0044978	0.0004998
number_credits	0.0519740	0.8425787	0.0054973	0.1329335	0.0199900	0.0064968
job	0.0439780	0.0479760	0.8830585	0.0004998	0.0329835	0.0004998
people_liable	0.0004998	0.3433283	0.2663668	0.0309845	0.0429785	0.0014993
telephone	0.0464768	0.0504748	0.0194903	0.0004998	0.2753623	0.0009995
foreign_worker	0.0779610	0.0024988	0.3843078	0.0009995	0.8555722	0.0239880
credit_risk	0.0254873	0.0309845	0.8680660	0.0009995	0.0009995	0.0004998

```
kable(Pearson_chisq_test[, 13:17], format = "latex", booktabs=TRUE) %>%
  kable_styling(font_size = 6, latex_options = "hold_position")
```

	number_credits	job	people_liable	telephone	foreign_worker
status	0.0394803	0.0499750	0.1224388	0.0979510	0.1279360
credit_history	0.0004998	0.4047976	0.0514743	0.2628686	0.3998001
purpose	0.2823588	0.0004998	0.0039980	0.0004998	0.0049975
savings	0.1154423	0.3478261	0.8915542	0.0684658	0.8980510
employment_duration	0.0009995	0.0004998	0.0404798	0.0004998	0.4612694
installment_rate	0.4267866	0.0689655	0.1109445	0.4722639	0.0044978
personal_status_sex	0.0509745	0.0349825	0.0004998	0.0434783	0.0814593
other_debtors	0.8565717	0.0504748	0.3138431	0.0594703	0.0004998
present_residence	0.0064968	0.8965517	0.2678661	0.0159920	0.3963018
property	0.1244378	0.0004998	0.0274863	0.0004998	0.0004998
other_installment_plans	0.0224888	0.0329835	0.0509745	0.2893553	0.8650675
housing	0.0089955	0.0004998	0.0024988	0.0024988	0.0304848
number_credits	0.0004998	0.0004998	0.0059970	0.0524738	0.9160420
job	0.0039980	0.0004998	0.0004998	0.0004998	0.0199900
people_liable	0.0054973	0.0014993	0.0004998	0.6596702	0.0264868
telephone	0.0669665	0.0004998	0.6591704	0.0004998	0.0234883
foreign_worker	0.9220390	0.0179910	0.0199900	0.0259870	0.0004998
credit_risk	0.4657671	0.6011994	1.0000000	0.2583708	0.0094953

Based on the above table, we conclude that the following predictors are dependent to most of the predictors with $\alpha = 0.05$ according to Pearson's Chi-sq Test of Independence, and we consider dropping these predictors:

- job
- credit_history
- purpose
- employment_duration
- housing
- people_liable

Also, we can see that the following predictors have very weak association with the response variable:

- installment_rate

- personal_status_sex
- other_debtors
- present_residence
- number_credits
- job
- people_liable
- telephone
- foreign_worker

To summarize, the variables we will use in model building are:

- status
- duration
- savings
- property
- age
- other_installment_plans