Loan Dataset Analysis

Possible Solution

11/07/2023

Disclaimer

The file contains a possible solution for the exam. It has been considered a good work by professor, but still there can be some errors.

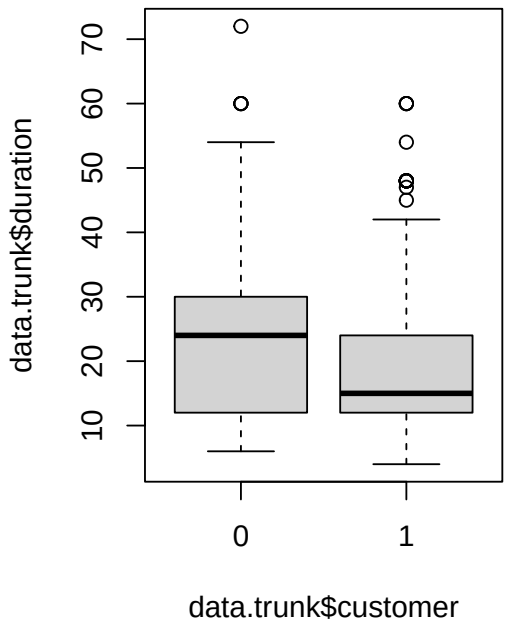
First Part

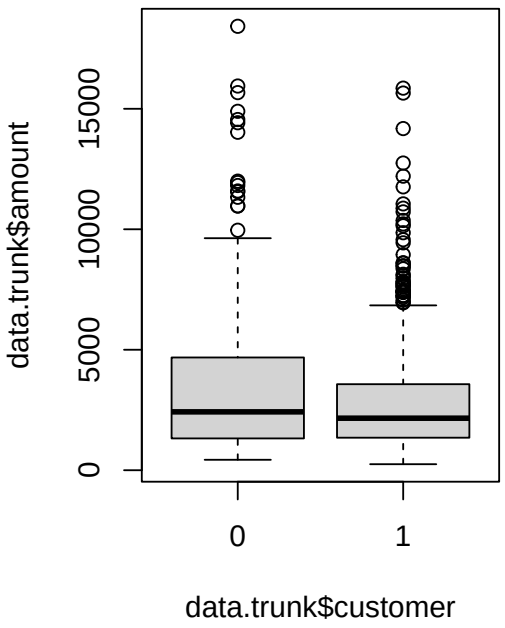
First of all, we load the dataset and we control the correctness of the type of all variables. Then we trunk the dataset.

The response variable is a categorical one (*customer*). Since it has only two levels, we will consider a logistic regression model.

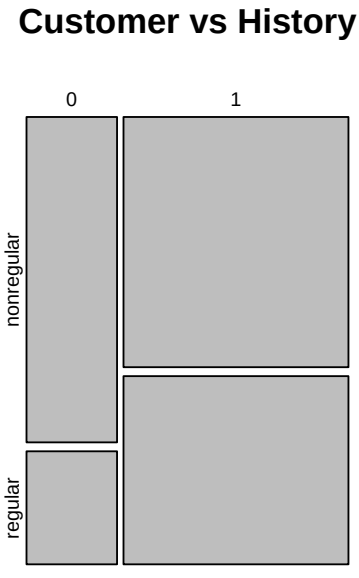
Let's look into the relationships among the response variable and the covariates.

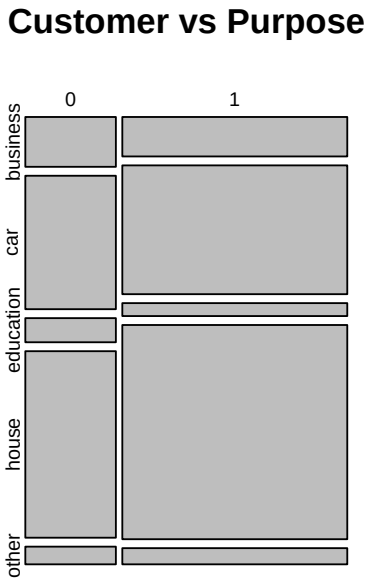
The relationship among *customer* and *duration* seems to be significant, as boxes have different dimensions. In addition medians are not at the same level and tails have different lengths. The same applies to the relationship between *customer* and *amount*, but it's not really clear.





The relationship among *customer* and *history* seems to be significant, as boxes have different dimensions and they are not overlapping each other. The same applies to the relationship between *customer* and *purpose*.

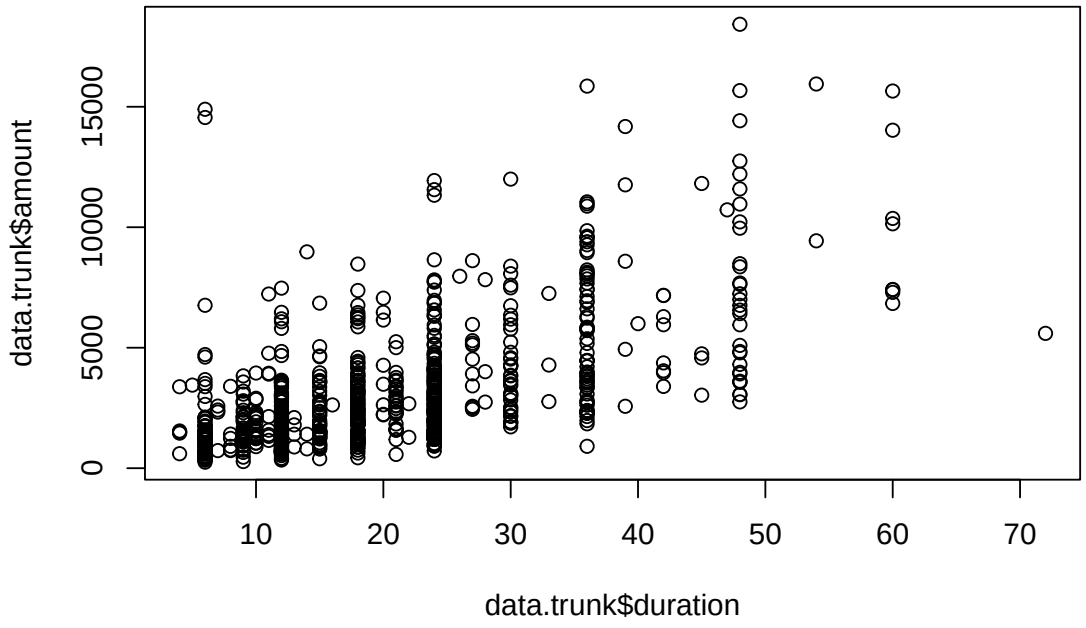




Now let's look at the interactions among covariates.

Consider the interaction between 2 numerical covariates.

The graph suggests a positive linear relationship between *duration* and *amount*.



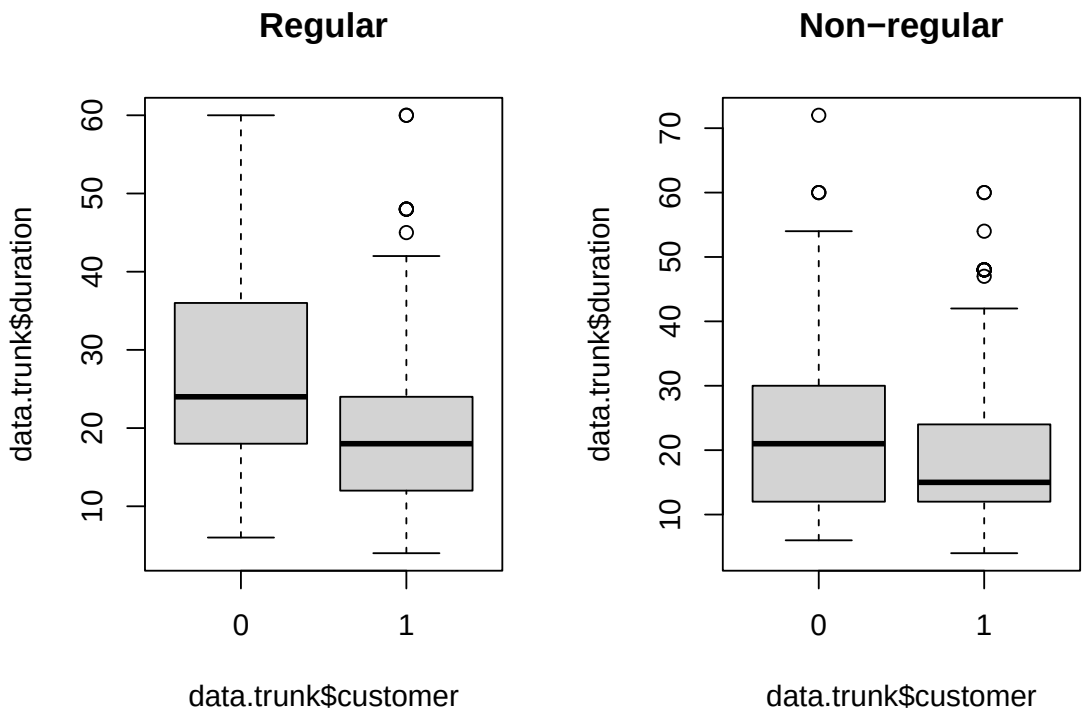
The intuition is correct, as show in the following correlation table.

```
##           duration    amount
## duration 1.0000000 0.6207223
## amount   0.6207223 1.0000000
```

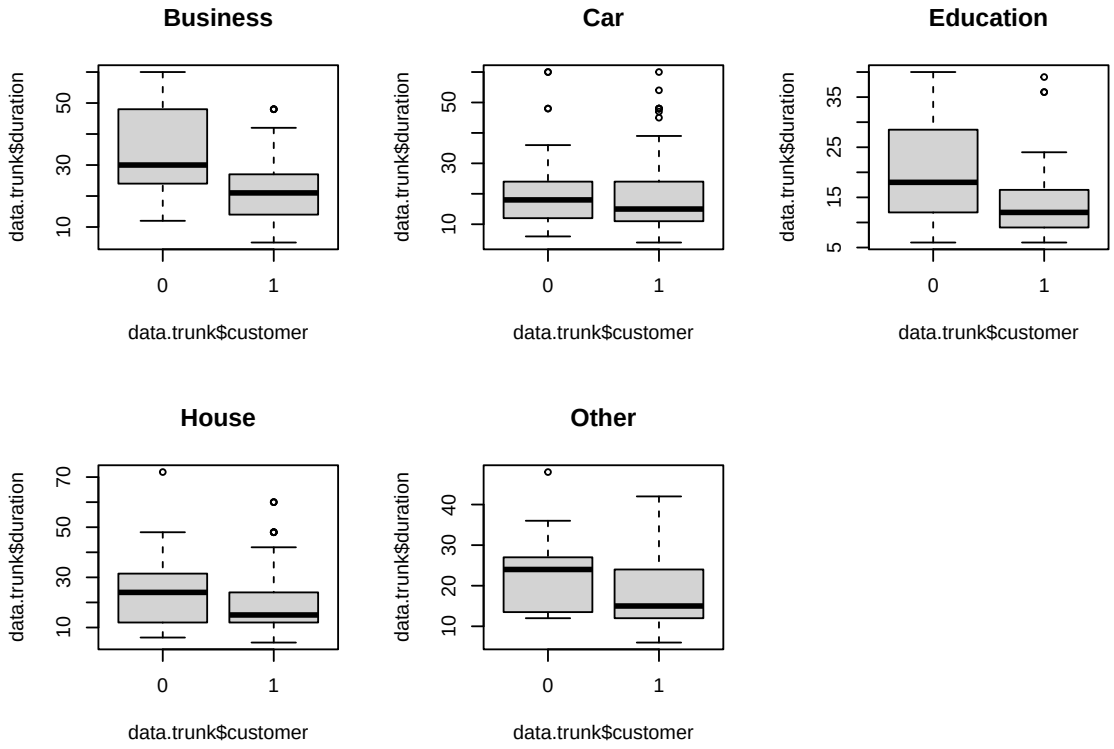
Focus now on the interaction between 1 numerical and 1 qualitative covariate.

Fix the numerical covariate *duration*.

The interaction between *duration* and *history* could be significant, as boxes are not completely overlapping, there are different tails and also medians are different. Generally speaking, it is possible to recognize different behaviors with respect to different levels.

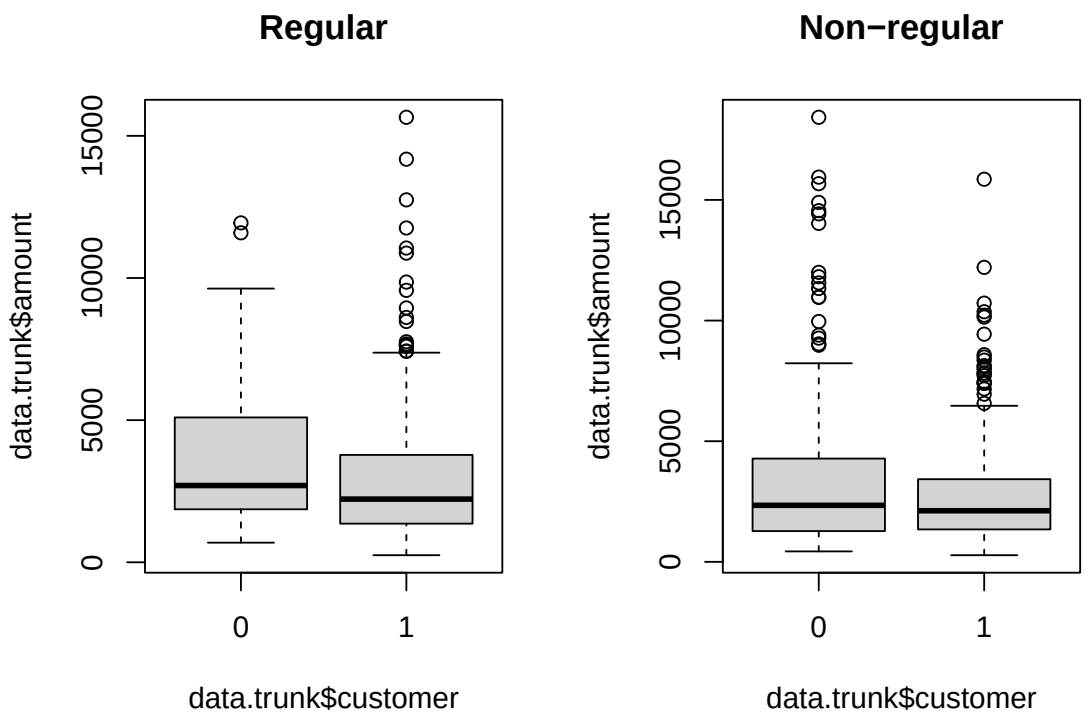


As we can see from the graphs interaction between *duration* and *purpose* could be significant, as boxes are not completely overlapping, there are different tails and also medians are different. So there are different behaviors with respect to different levels.

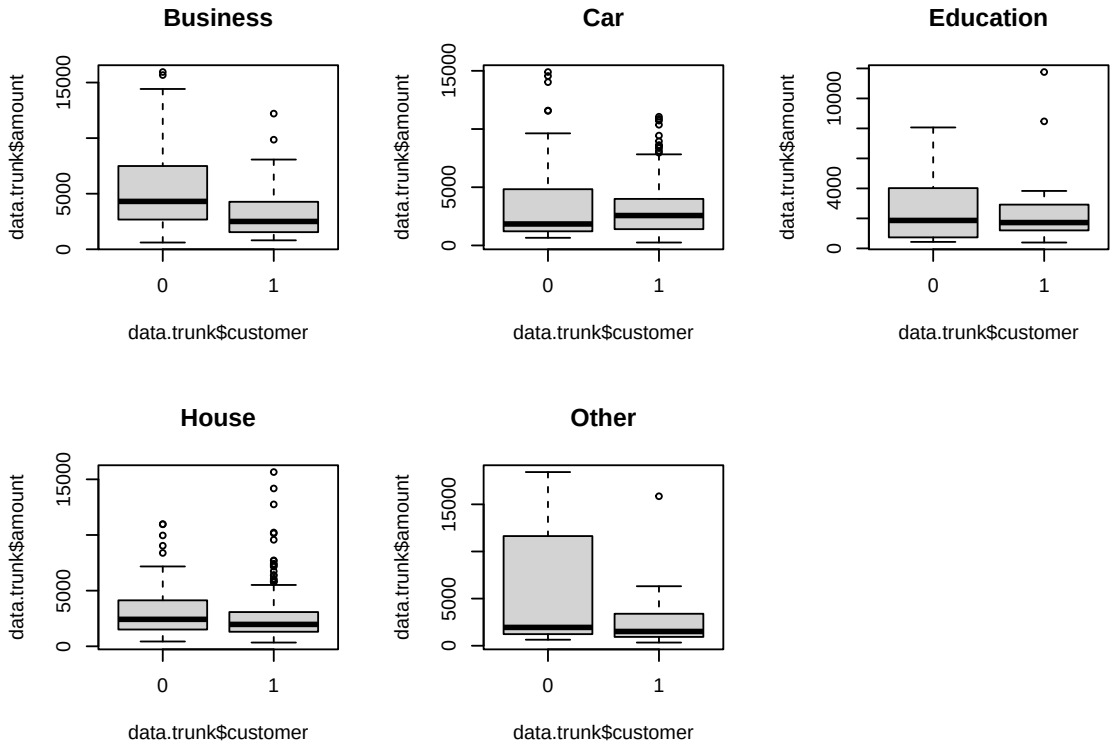


Fix the numerical covariate *amount*.

The interaction between *amount* and *history* could be significant, as boxes are not completely overlapping, there are different tails and also medians are different. Generally speaking, it is possible to recognize different behaviors with respect to different levels.

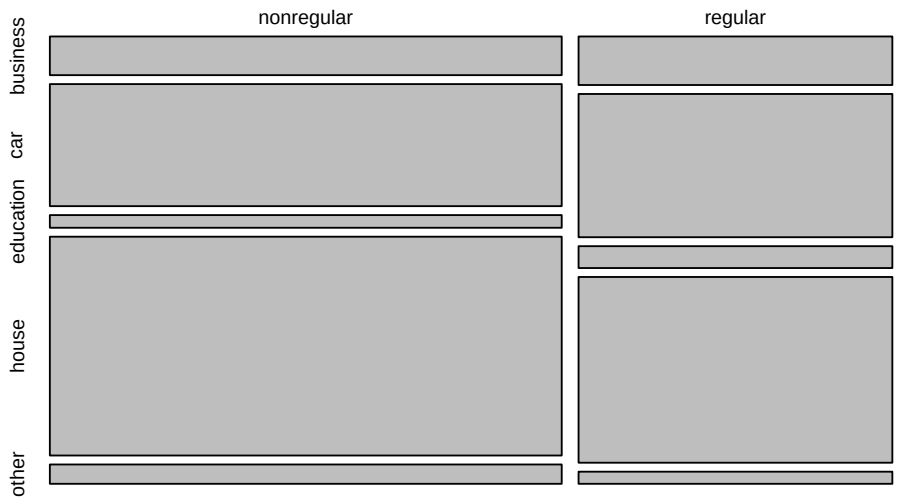


As we can see from the graphs interaction between *amount* and *purpose* could be significant, as boxes are not completely overlapping, there are different tails and some medians are different. So there are different behaviors with respect to different levels.



Focus now on the interaction between 2 qualitative covariates.

The interaction among *history* and *purpose* could be significant, as boxes have different dimensions and they are not overlapping each other. So the following mosaicplot are showing the presence of a kind of interaction.



Let's estimate the logistic regression model. We will include all the covariates and their interactions first.

```
##
## Call:
## glm(formula = customer ~ duration * amount + duration * history +
##     duration * purpose + amount * history + amount * purpose +
##     history * purpose, family = binomial, data = data.trunk)
##
## Coefficients:
##                                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)                    2.670e+00  7.136e-01   3.742 0.000182 ***
## duration                      -5.986e-02  2.778e-02  -2.155 0.031153 *
## amount                        -2.314e-04  1.343e-04  -1.724 0.084786 .
## historyregular                 9.496e-01  6.712e-01   1.415 0.157084
## purposecar                    -1.503e+00  7.206e-01  -2.086 0.036942 *
## purposeeducation              -1.759e+00  9.977e-01  -1.763 0.077817 .
## purposehouse                  -7.176e-01  7.031e-01  -1.021 0.307379
## purposeother                  -1.103e+00  1.101e+00  -1.002 0.316193
## duration:amount                3.340e-06  2.160e-06   1.546 0.122023
## duration:historyregular       -2.999e-02  1.971e-02  -1.522 0.128127
## duration:purposecar            2.582e-02  2.990e-02   0.864 0.387807
## duration:purposeeducation      1.044e-03  5.270e-02   0.020 0.984190
## duration:purposehouse          5.273e-03  2.914e-02   0.181 0.856419
## duration:purposeother          4.285e-02  5.401e-02   0.793 0.427524
## amount:historyregular          7.332e-05  8.448e-05   0.868 0.385437
## amount:purposecar              1.588e-04  1.220e-04   1.302 0.193035
## amount:purposeeducation        2.460e-04  2.274e-04   1.082 0.279338
## amount:purposehouse            1.186e-04  1.251e-04   0.948 0.342986
## amount:purposeother           -2.195e-05  1.493e-04  -0.147 0.883143
```

```
## historyregular:purposecar        3.565e-01  5.900e-01   0.604 0.545672
## historyregular:purposeeducation -3.079e-01  9.422e-01  -0.327 0.743838
## historyregular:purposehouse      5.286e-01  5.750e-01   0.919 0.357883
## historyregular:purposeother      2.793e-01  1.066e+00   0.262 0.793277
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1069.40  on 891  degrees of freedom
## Residual deviance:  980.87  on 869  degrees of freedom
## AIC: 1026.9
##
## Number of Fisher Scoring iterations: 4
```

We can perfom variable selection over the model, obtaining the following output.

```
##
## Call:
## glm(formula = customer ~ duration + amount + history + purpose +
##     duration:amount + amount:purpose, family = binomial, data = data.trunk)
##
## Coefficients:
##                           Estimate Std. Error z value Pr(>|z|)
## (Intercept)              2.496e+00  5.271e-01   4.735 2.19e-06 ***
## duration                -5.770e-02  1.218e-02  -4.737 2.17e-06 ***
## amount                  -2.745e-04  1.130e-04  -2.430   0.0151 *
## historyregular           8.778e-01  1.725e-01   5.090 3.58e-07 ***
## purposecar              -9.721e-01  4.866e-01  -1.998   0.0458 *
## purposeeducation        -1.717e+00  6.754e-01  -2.542   0.0110 *
## purposehouse            -4.290e-01  4.734e-01  -0.906   0.3649
## purposeother            -3.083e-01  6.780e-01  -0.455   0.6493
## duration:amount          4.144e-06  2.075e-06   1.997   0.0458 *
## amount:purposecar        2.233e-04  9.466e-05   2.359   0.0183 *
## amount:purposeeducation  2.056e-04  1.576e-04   1.304   0.1921
## amount:purposehouse      1.337e-04  9.513e-05   1.406   0.1599
## amount:purposeother      5.606e-05  1.154e-04   0.486   0.6271
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1069.4  on 891  degrees of freedom
## Residual deviance:  986.6  on 879  degrees of freedom
## AIC: 1012.6
##
## Number of Fisher Scoring iterations: 4
```

After performing model selection, the resulting model is ***customer ~ duration + amount + history + purpose + duration:amount + amount:purpose***. We removed all the interactions, apart from *duration:amount* and *amount:purpose*. *purpose* and *amount:purpose* seem not to be significant for some levels, but since there one of their level significant, they cannot be removed.

The accuracy of the model, based on the deviance is the following.

```
## [1] 0.006492173
```

The accuracy of the model seems to be very low.

By using *anova*, to compare the first fitted model (all covariates + interactions) and the last one, it looks like sticking on the simpler one is the right choice.

```
## Analysis of Deviance Table
##
## Model 1: customer ~ duration + amount + history + purpose + duration:amount +
##     amount:purpose
## Model 2: customer ~ duration * amount + duration * history + duration *
##     purpose + amount * history + amount * purpose + history *
##     purpose
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1       879     986.60
## 2       869     980.87 10   5.7283   0.8375
```

We can try to add polynomials to *duration* and *amount*.

```
##
## Call:
## glm(formula = customer ~ duration + amount + history + purpose +
##     I(duration^2) + duration:amount + amount:purpose, family = binomial,
##     data = data.trunk)
##
## Coefficients:
##                           Estimate Std. Error z value Pr(>|z|)
## (Intercept)              2.559e+00  5.469e-01   4.679 2.88e-06 ***
## duration                -6.742e-02  2.551e-02  -2.642  0.00823 **
## amount                  -2.546e-04  1.225e-04  -2.078  0.03771 *
## historyregular           8.813e-01  1.727e-01   5.104 3.32e-07 ***
## purposecar              -9.571e-01  4.876e-01  -1.963  0.04968 *
## purposeeducation        -1.719e+00  6.755e-01  -2.545  0.01093 *
## purposehouse            -4.198e-01  4.736e-01  -0.886  0.37539
## purposeother            -3.095e-01  6.781e-01  -0.456  0.64806
## I(duration^2)            2.331e-04  5.359e-04   0.435  0.66360
## duration:amount          3.464e-06  2.611e-06   1.326  0.18470
## amount:purposecar        2.188e-04  9.537e-05   2.294  0.02178 *
## amount:purposeeducation  2.062e-04  1.577e-04   1.307  0.19108
## amount:purposehouse      1.314e-04  9.543e-05   1.377  0.16855
## amount:purposeother      5.855e-05  1.161e-04   0.504  0.61395
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1069.40  on 891  degrees of freedom
## Residual deviance:  986.41  on 878  degrees of freedom
## AIC: 1014.4
##
## Number of Fisher Scoring iterations: 4
```

Polynomials for *duration* seem not to be significant.

```
##
## Call:
## glm(formula = customer ~ duration + amount + history + purpose +
##     I(amount^2) + I(amount^3) + duration:amount + amount:purpose,
```

```
##     family = binomial, data = data.trunk)
##
## Coefficients:
##                            Estimate Std. Error z value Pr(>|z|)
## (Intercept)              2.033e+00  5.417e-01   3.752 0.000175 ***
## duration                -9.786e-02  1.642e-02  -5.959 2.54e-09 ***
## amount                   3.810e-04  2.113e-04   1.803 0.071377 .
## historyregular           8.735e-01  1.746e-01   5.002 5.67e-07 ***
## purposecar              -9.090e-01  4.848e-01  -1.875 0.060801 .
## purposeeducation        -1.522e+00  6.776e-01  -2.246 0.024736 *
## purposehouse            -2.565e-01  4.708e-01  -0.545 0.585906
## purposeother            -1.482e-01  6.874e-01  -0.216 0.829276
## I(amount^2)             -1.144e-07  3.404e-08  -3.360 0.000779 ***
## I(amount^3)              3.659e-12  1.477e-12   2.477 0.013245 *
## duration:amount          1.143e-05  3.090e-06   3.700 0.000216 ***
## amount:purposecar        2.324e-04  9.784e-05   2.376 0.017508 *
## amount:purposeeducation  1.923e-04  1.585e-04   1.213 0.224958
## amount:purposehouse      8.889e-05  9.778e-05   0.909 0.363289
## amount:purposeother      9.645e-05  1.332e-04   0.724 0.468969
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1069.40  on 891  degrees of freedom
## Residual deviance:  967.25  on 877  degrees of freedom
## AIC: 997.25
##
## Number of Fisher Scoring iterations: 4
```

Polynomials for *amount* seem to be significant, until the third degree.

We can use either the model with the polynomials for *amount* or the initial resulting model. We prefer to use the model without the polynomials as it is simpler to handle.

The predictions of the values gives the following error table.

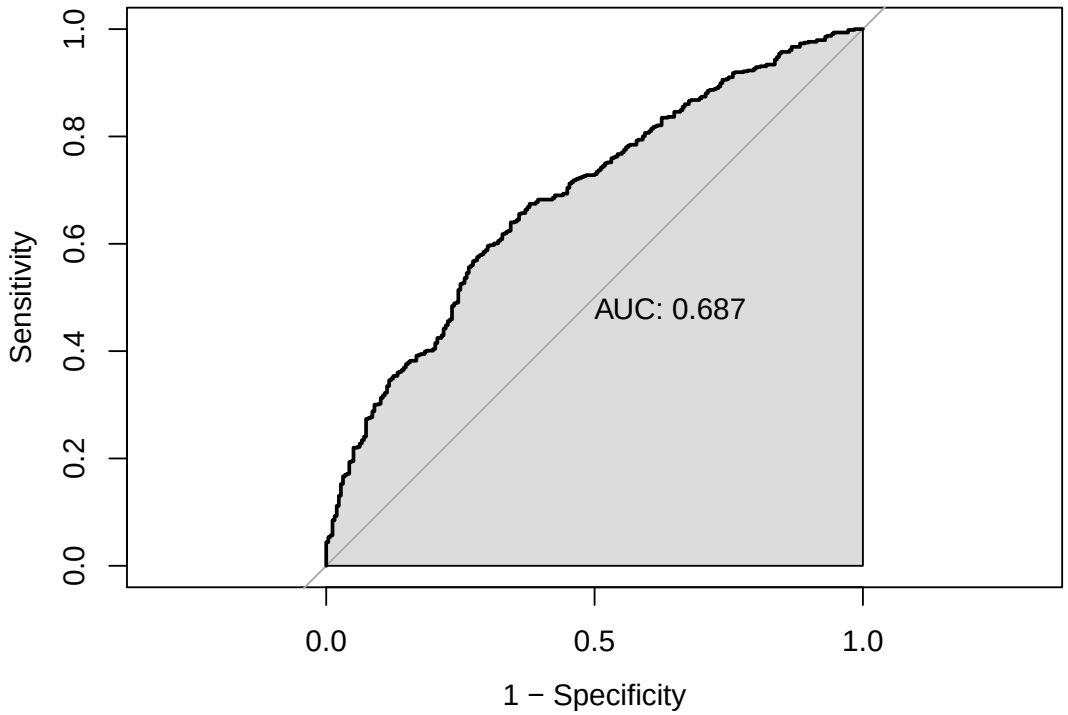
```
##            cliente
## predictions   0   1 Sum
##         0    34  23  57
##         1   222 613 835
##         Sum 256 636 892
```

Here you can find the training error rate.

```
## [1] 0.2746637
```

The error rate is quite ok (~27%), which means that the model is a quite good fit for the reduced dataset and it is quite good on prediction on the same dataset.

Plotting the *ROC* curve. The results not so accurate.



Based on the fitted logistic regression model, *customer* is associated with:

- *duration*: more high is the duration of the loan, there is less probability to have repaid the loan;
- *amount*: more high is the amount of the loan, there is less probability to obtain the loan back from the customer;
- *history*: if the customer haven't repaid previous loans, there is less possibility the customer will repay the loan than the one that have repaid his loans in the past;
- *purpose*: if the customer take a loan for business, there is more possibility the customer will repay the loan than the other purposes. For each level if the estimate is lower implies a less possibility to have the loan repaid by the customer.

There are also the interaction between covariates *duration:amount* and *purpose:amount*.

Second part

Considering all the variables of the dataset. We will use all the covariates and the interaction *duration:amount* and *amount:purpose*, that we saw being relevant on the first part.

GLM

```
##
## Call:
## glm(formula = customer ~ . + duration:amount + amount:purpose,
##     family = "binomial", data = data.full)
##
## Coefficients:
##                            Estimate Std. Error z value Pr(>|z|)
## (Intercept)              2.050e+00  1.012e+00   2.027  0.04269 *
## duration                -6.014e-02  1.305e-02  -4.609 4.04e-06 ***
## historyregular           9.693e-01  2.133e-01   4.544 5.52e-06 ***
## purposecar              -1.143e+00  5.069e-01  -2.256  0.02409 *
## purposeeducation        -1.939e+00  7.066e-01  -2.745  0.00606 **
## purposehouse            -4.670e-01  4.907e-01  -0.952  0.34118
## purposeother            -3.116e-01  7.142e-01  -0.436  0.66268
## amount                  -2.938e-04  1.188e-04  -2.472  0.01343 *
## bankaccountyes          -1.023e+00  2.466e-01  -4.148 3.35e-05 ***
## genderM                  1.942e-01  1.815e-01   1.070  0.28470
## jobyes                   7.037e-01  3.493e-01   2.015  0.04395 *
## otherloans              -2.769e-01  1.759e-01  -1.574  0.11541
## propertyother            7.197e-01  3.432e-01   2.097  0.03597 *
## propertyrealestate       9.282e-01  3.723e-01   2.493  0.01266 *
## age                      2.126e-02  8.540e-03   2.489  0.01280 *
## houserent                3.835e-01  2.035e-01   1.885  0.05944 .
## unemployedyes            2.499e-01  2.041e-01   1.224  0.22078
## maintenancemore than 1   1.360e-01  2.560e-01   0.531  0.59531
## foreignyes              -9.932e-01  5.864e-01  -1.694  0.09033 .
## duration:amount          4.501e-06  2.162e-06   2.082  0.03738 *
## purposecar:amount        2.453e-04  9.816e-05   2.499  0.01246 *
## purposeeducation:amount  2.630e-04  1.638e-04   1.606  0.10832
## purposehouse:amount      1.423e-04  9.996e-05   1.424  0.15450
## purposeother:amount      6.402e-05  1.205e-04   0.531  0.59534
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1069.40  on 891  degrees of freedom
## Residual deviance:  929.76  on 868  degrees of freedom
## AIC: 977.76
##
## Number of Fisher Scoring iterations: 5
```

Which gives the following accuracy, based on the deviance.

```
## [1] 0.07153815
```

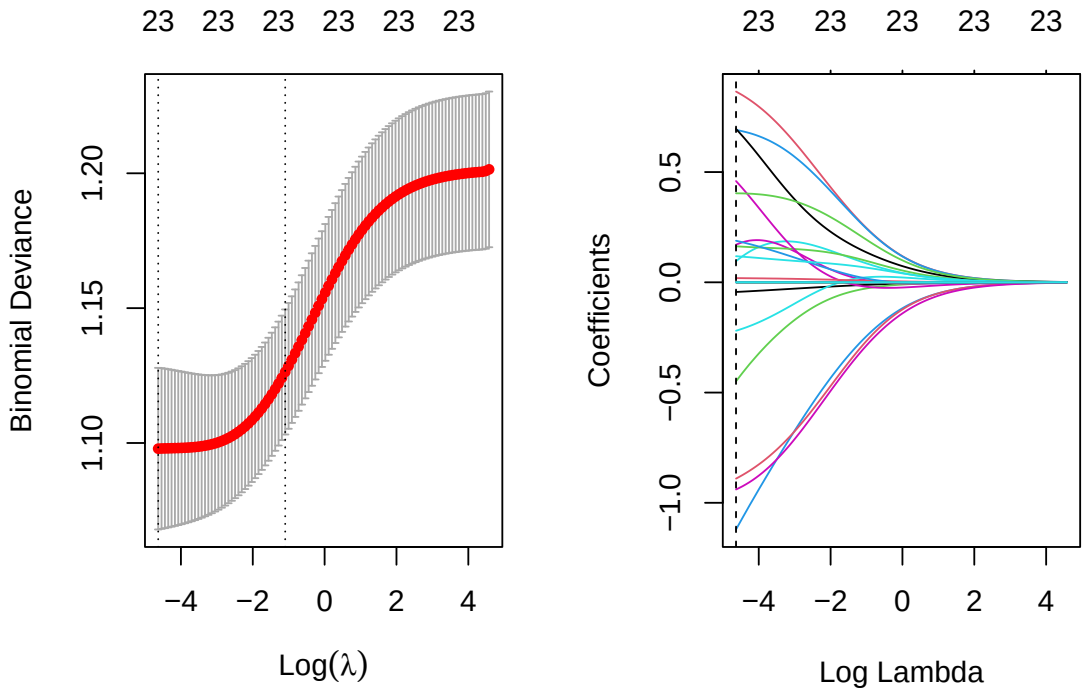
Ridge

Let's try to use *Ridge* to find the best model for the dataset provided.

We will use the following seed to find the best lambda.

```
set.seed(222)
```

The best lambda can be found in the following graph. There is also the graph of the coefficients with respect to the different value of lambda.



The values of the coefficients are the following. Obviously, no model selection has been made since we are using *Ridge* regression.

```
## 24 x 2 sparse Matrix of class "dgCMatrix"
##                                          s0
## (Intercept)              2.050368e+00  1.285808e+00
## duration                -6.014356e-02 -4.374903e-02
## historyregular           9.692948e-01  8.649904e-01
## purposecar              -1.143434e+00 -4.515714e-01
## purposeeducation        -1.939386e+00 -1.123617e+00
## purposehouse            -4.670469e-01  9.351292e-02
## purposeother            -3.115618e-01  1.682345e-01
## amount                  -2.937796e-04 -9.187952e-05
## bankaccountyes          -1.023066e+00 -8.903067e-01
## genderM                  1.942014e-01  1.632931e-01
## jobyes                   7.036798e-01  6.911569e-01
## otherloans              -2.768564e-01 -2.196364e-01
## propertyother            7.196962e-01  4.585126e-01
## propertyrealestate       9.282174e-01  6.949449e-01
## age                      2.125921e-02  1.904925e-02
```

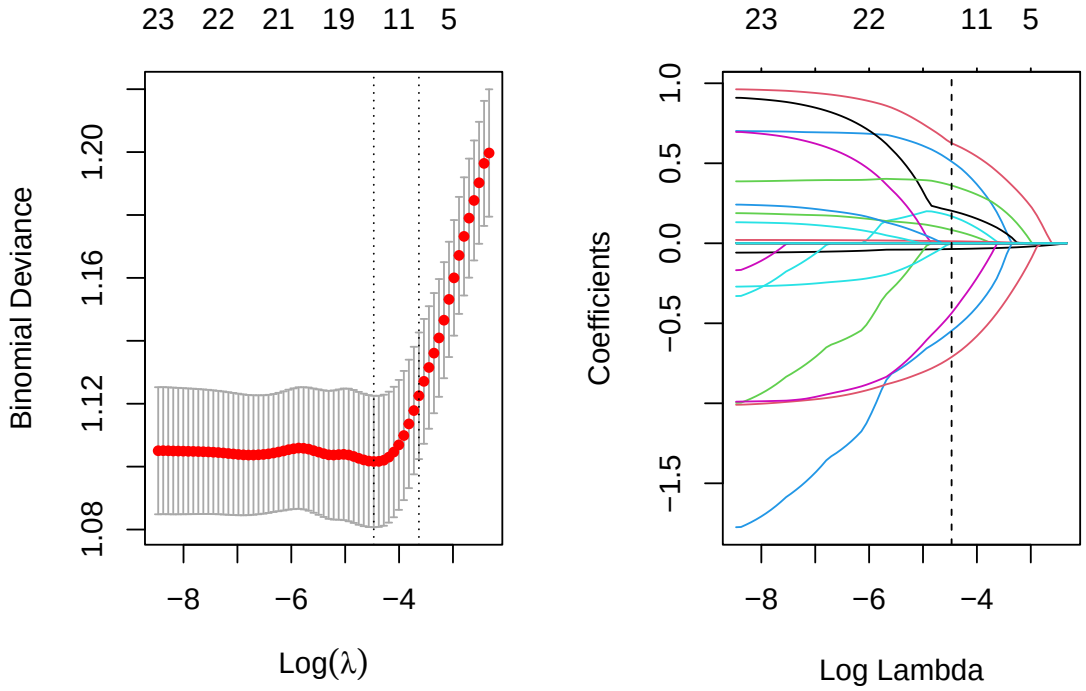
```
## houserent                 3.835152e-01  4.035186e-01
## unemployedyes             2.498802e-01  1.884512e-01
## maintenancemore than 1    1.359786e-01  1.183794e-01
## foreignyes               -9.931678e-01 -9.394948e-01
## duration:amount           4.500838e-06  1.339195e-06
## purposecar:amount         2.453050e-04  1.025344e-04
## purposeeducation:amount   2.630442e-04  9.695156e-05
## purposehouse:amount       1.423293e-04  2.851721e-05
## purposeother:amount       6.402089e-05 -3.373093e-05
```

We can see that *Ridge* has penalized a lot of coefficients and it has given more importance to others.

Lasso

Let's try to use *Lasso* and check if there is model selection. We will compare the results with the ones from *Ridge* later.

The best lambda can be found in the following graph. There is also the graph of the coefficients with respect to the different value of lambda.



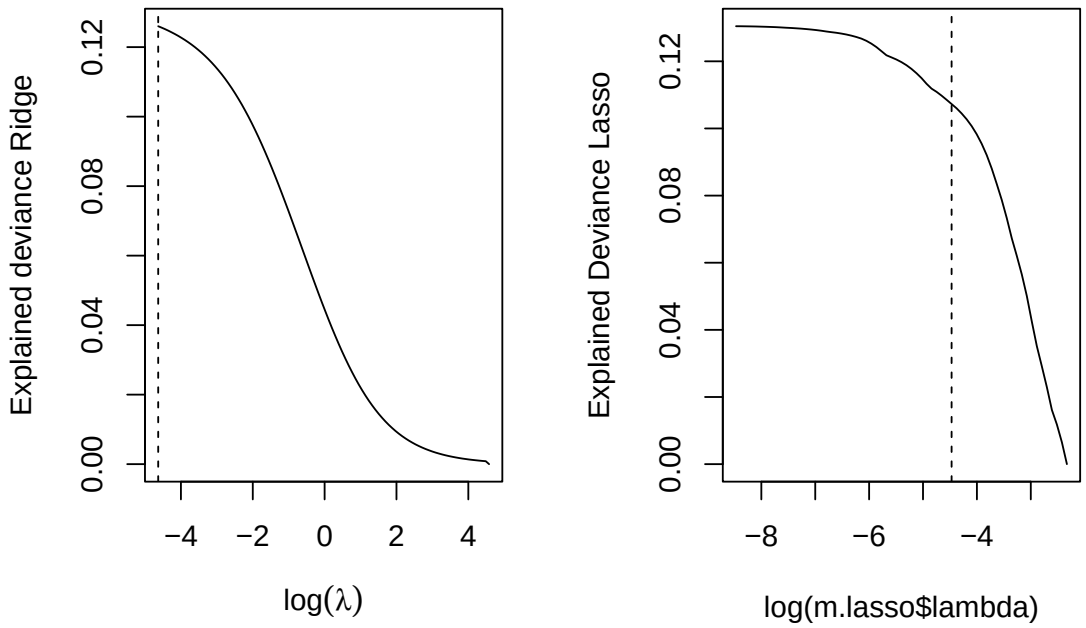
Lasso has performed model selection, in fact many coefficients are equal to 0.

```
## 24 x 3 sparse Matrix of class "dgCMatrix"
##                                          s0            s0
## (Intercept)               2.050368e+00  1.285808e+00  1.087360e+00
## duration                 -6.014356e-02 -4.374903e-02 -3.592764e-02
## historyregular            9.692948e-01  8.649904e-01  6.254748e-01
## purposecar               -1.143434e+00 -4.515714e-01  .
## purposeeducation         -1.939386e+00 -1.123617e+00 -5.468219e-01
## purposehouse             -4.670469e-01  9.351292e-02  1.689664e-01
```

```
## purposeother            -3.115618e-01  1.682345e-01  .
## amount                  -2.937796e-04 -9.187952e-05  .
## bankaccountyes          -1.023066e+00 -8.903067e-01 -7.108295e-01
## genderM                  1.942014e-01  1.632931e-01  8.307986e-02
## jobyes                   7.036798e-01  6.911569e-01  5.108662e-01
## otherloans              -2.768564e-01 -2.196364e-01  .
## propertyother            7.196962e-01  4.585126e-01  .
## propertyrealestate       9.282174e-01  6.949449e-01  2.030759e-01
## age                      2.125921e-02  1.904925e-02  1.291556e-02
## houserent                3.835152e-01  4.035186e-01  3.614590e-01
## unemployedyes            2.498802e-01  1.884512e-01  .
## maintenancemore than 1   1.359786e-01  1.183794e-01  .
## foreignyes              -9.931678e-01 -9.394948e-01 -4.366298e-01
## duration:amount          4.500838e-06  1.339195e-06  .
## purposecar:amount        2.453050e-04  1.025344e-04  .
## purposeeducation:amount  2.630442e-04  9.695156e-05  .
## purposehouse:amount      1.423293e-04  2.851721e-05  .
## purposeother:amount      6.402089e-05 -3.373093e-05 -7.520124e-06
```

From these coefficients, it's possible to understand which is the best model obtained with *Lasso*.

The graphs of the explained deviance of *Lasso* and *Ridge* are the following.



The maximum explained deviances, obtained with the minimum lambda (from *Ridge* and *Lasso*), are the following.

```
## [1] 0.1260124
```

```
## [1] 0.1072775
```

Ridge performs better than *Lasso*, as the explained deviance shows. Therefore *Ridge* seems to be preferable

respect to this parameter, but the explained deviance is pretty low.

The *MSE*s for the two models are the following (1^{st} *Ridge*, 2^{nd} *Lasso*).

```
## [1] 1.097891
```

```
## [1] 1.10163
```

In conclusion, *Ridge* seems to be preferable also in terms of *MSE* against the model fitted by *Lasso*.

Based on the model fitted with *Ridge*, *customer* is associated with:

- *duration*: more high is the duration of the loan, there is less probability to have repaid the loan;
- *history*: if the customer hasn't repaid previous loans, there is less probability the customer will repay the loan than the one that have repaid his loans in the past;
- *purpose*: if the customer takes a loan for business, there is more probability the customer will repay the loan than the other purposes. For each level if the estimate is lower than 0 implies a less probability to have the loan repaid respect to the loan for business, else implies a more probability to have the loan repaid by the customer respect to the loan for business;
- *amount*: more high is the amount of the loan, there is less probability to obtain the loan back from the customer;
- *bankaccount*: if the customer hasn't a bank account, there is more probability the customer will repay the loan than the one that has a bank account;
- *gender*: if the gender is female, there is less probability the customer will repay the loan than a male;
- *job*: if the customer hasn't any job, there is more probability the customer will repay the loan than the one that has a job;
- *otherloans*: more high is the number of the loan, there is less probability to obtain the loan back from the customer;
- *property*: if the customer hasn't any property, there is less probability the customer will repay the loan than the other levels. For the other levels more high is the estimate, there is more probability to repay the loan by the customer;
- *age*: more high is the age, there is more probability to obtain the loan back from the customer;
- *house*: if the customer owns an house, there is less probability the customer will repay the loan than the one that has rented a house;
- *unemployed*: if the customer is employed, there is less probability the customer will repay the loan than the one that is unemployed;
- *maintenance*: if the customer are able to provide maintenance for 1, there is less probability the customer will repay the loan than the one that is able to provide maintenance for more than 1;
- *unemployed*: if the customer is not foreign, there is more probability the customer will repay the loan than the one that is foreign.

There are also the interaction between covariates *duration:amount* and *purpose:amount*.