Model to predict the Probability of Default

Jaime Prades

mar 2023 - Chile

Introduction section: Project's goal and database used.

This project aims to develop a code that calculates the likelihood of a financial entity's customer defaulting on their credit card debt. The data for this project was sourced from a small sample from a financial entity in Chile where I am employed. The paper will showcase the various methods and models that were tested and the final selection of the most suitable one for the specific case study.

Exploratory analysis: understanding the data.

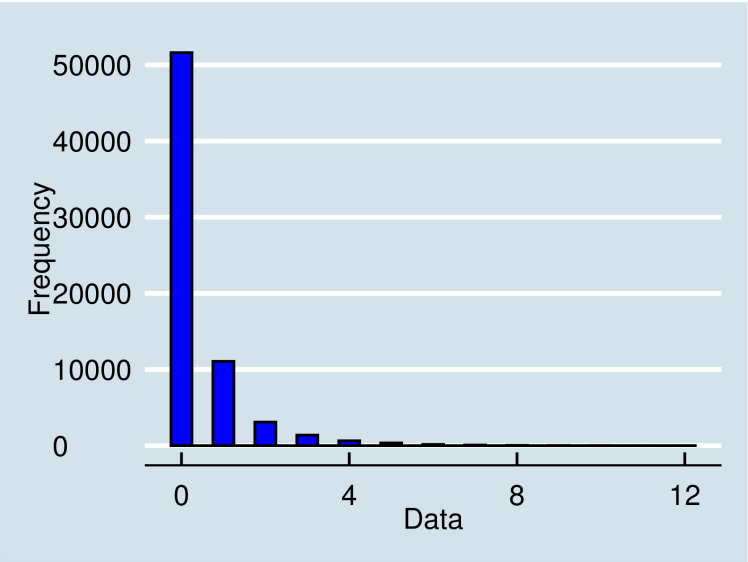
The data used in this project was sourced from a Chilean financial entity and consists of 12 columns, including 11 independent variables (X) and one dependent variable (Y). The dependent variable, represented as a binary flag, indicates the individual's credit card debt payment status (0 = Debt Paid, 1 = Debt Not Paid, Default). The data consists of information from 68471 individuals, with independent variable information collected in January 2021 and the Default/No Default flag reflecting the payment behavior throughout the entire year of 2021. If an individual has a delayed payment for more than 90 days, the flag is set to Default (1), which is considered an absorbing state and cannot be changed once set.

Variables Analysis.

Univariate Analysis.

Let's begin by analyzing the independent variables for better understanding. In this section, I listed the variables and analyzed each one separately. I also applied some techniques to capture better the estimation power of the data. To examine the distribution of these variables, histograms were utilized.

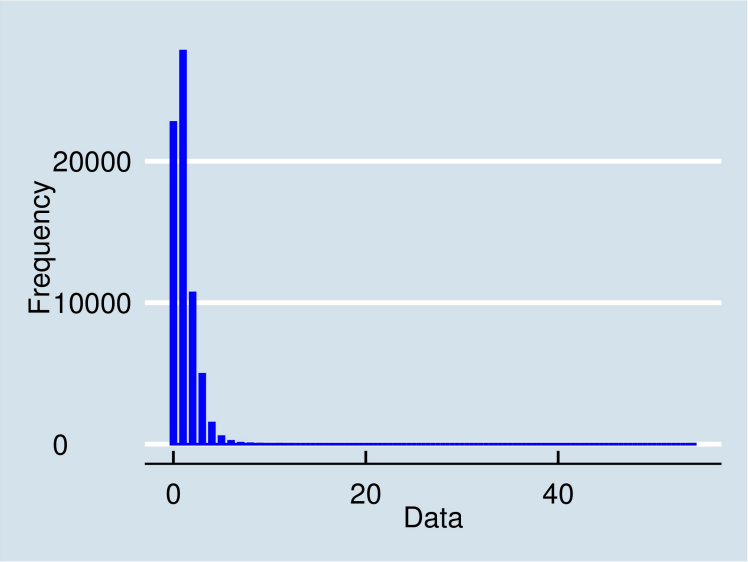
The first variable, "Commercial References," represents the number of times the person being analyzed has failed to pay bills such as electricity and cell phone bills (this information can be purchased in Chile). The histogram for this variable is shown below:



The majority of the observations are distributed around the "0", as expected.

-

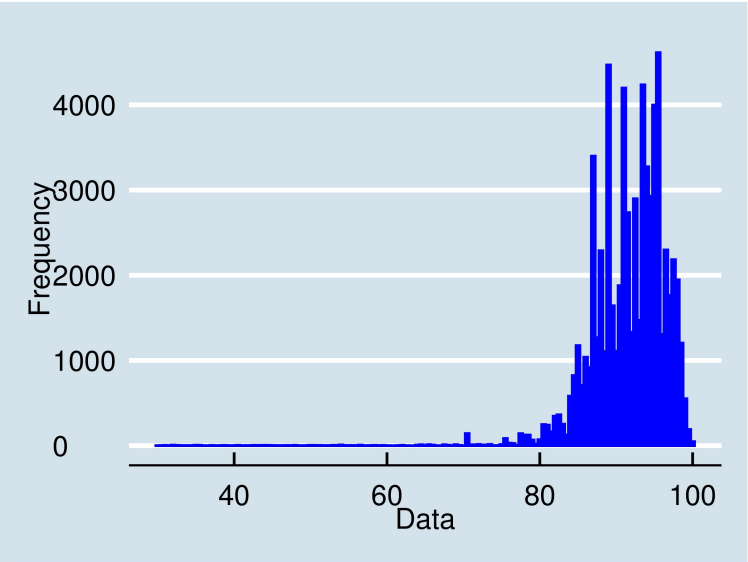
The second variable is "Quantity of credit cards the person owns". This is its histogram:



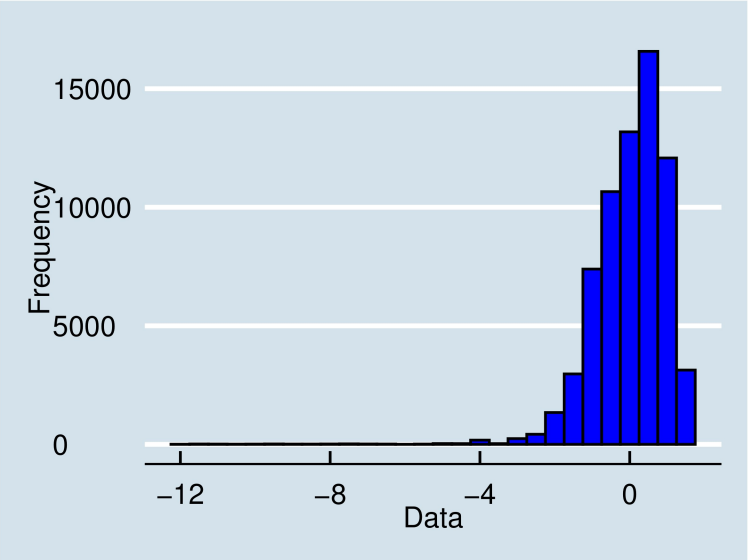
As expected, the distribution of the variable is also around "0".

-

The third variable is a very interesting one, it is "Basic Needs". It shows the percentage of basic needs that are covered in the region the person lives (it can be a huge indicator for the good or bad behavior in payments). This is its histogram:



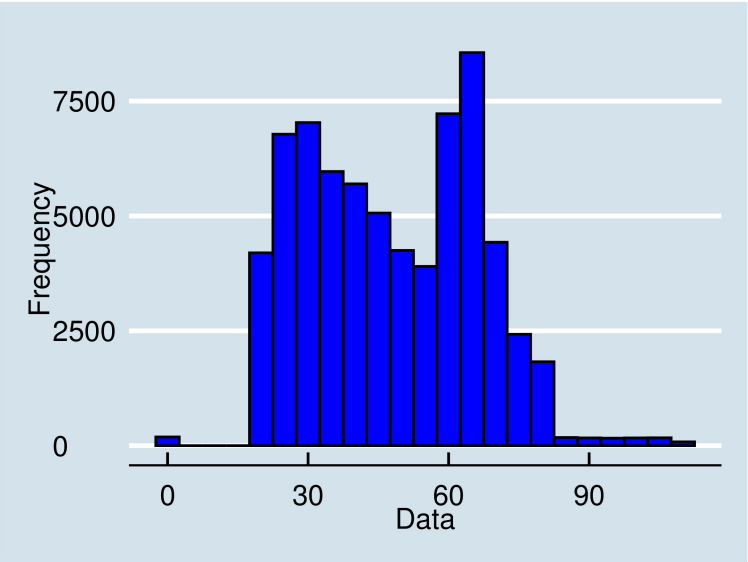
It is distributed around 91.6027082 (its mean). The histogram displays a low variability, with a limited number of observations below 60 points. To normalize the variable, I carried out mean subtraction and division by the standard deviation. The resulting histogram is displayed below:



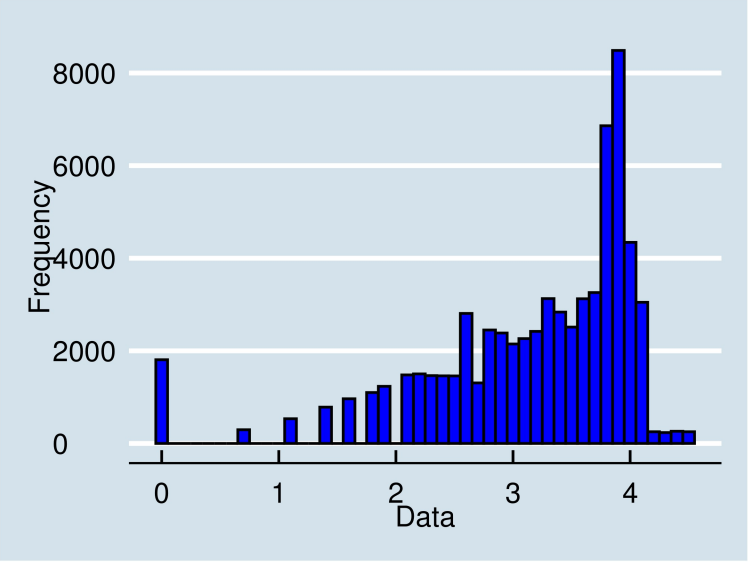
Now it is distributed around “0”.

-

The fourth variable is “Age”. This is its histogram:

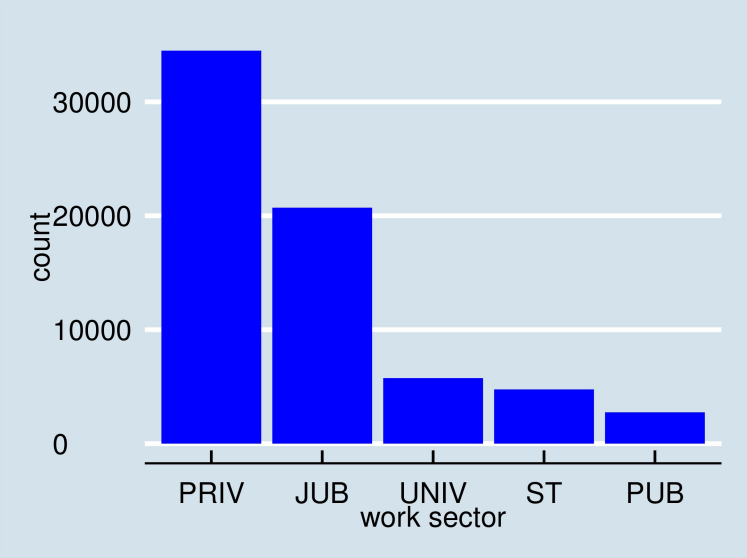


Given that the analysis focuses on individuals who are of legal age, the data was truncated at 18 years old. There are only a limited number of observations for individuals over 80 years old, so the data was also truncated at 80. The histogram shows that the data has a high variability and does not appear to follow a normal distribution. To enhance the predictive power of this variable, the natural logarithm (LN) function was applied. The resulting histogram is displayed below:



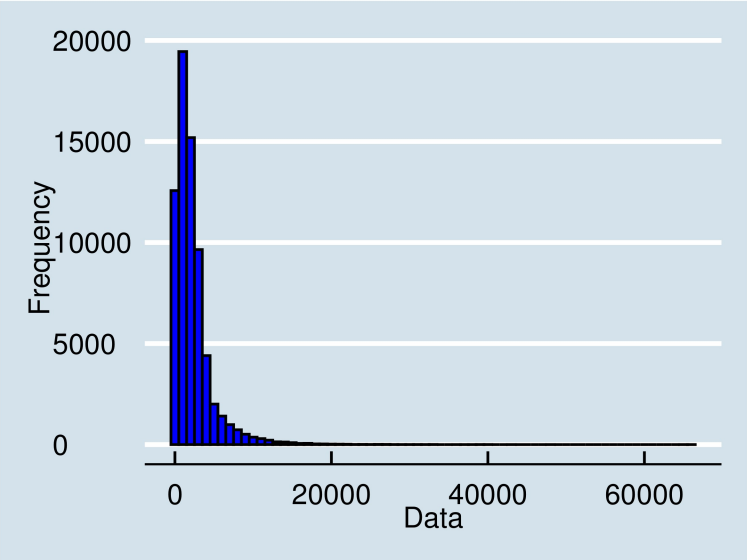
-

The fifth variable is the individual's sector of employment, represented as PRIV (Private sector), PUB (Public sector), JUB (Retired), UNIV (University teacher), and ST (Worker located in Santiago, Chile). The plot for this variable is shown below:

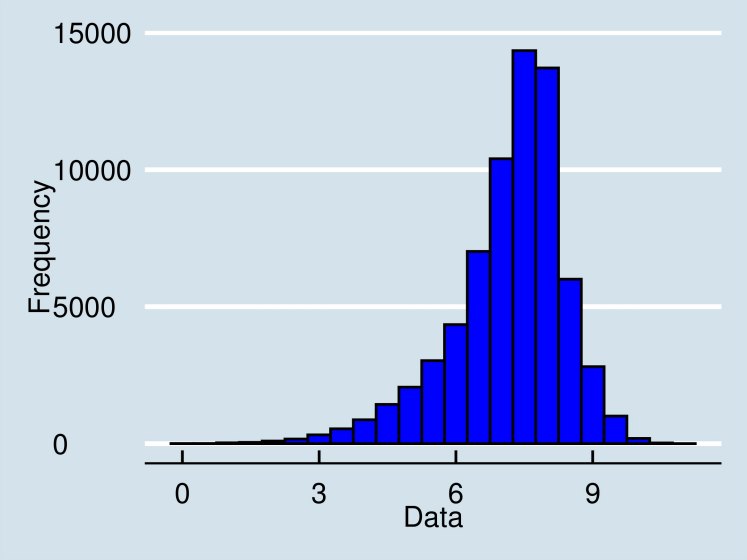


-

The sixth variable is the amount of money spent in Debit Card. This is its histogram:

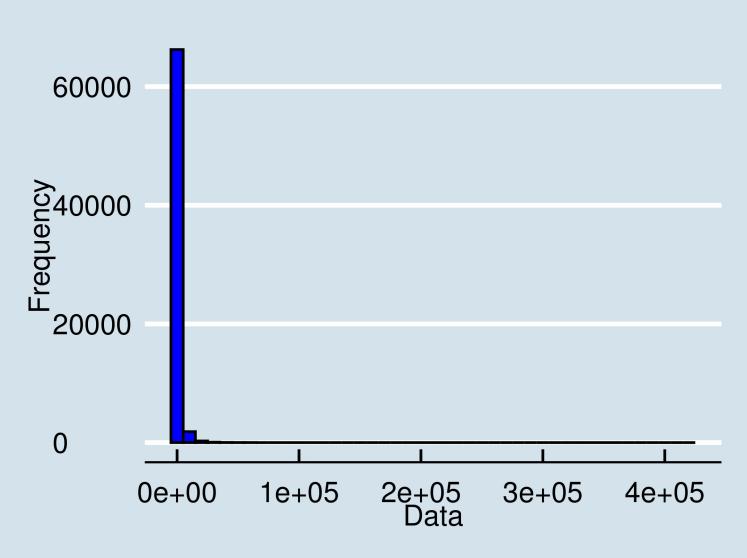


Since this is a variable in money units, it could be appropriated to apply LN function here, it should make the distribution of the variable more normal. This is the new histogram:

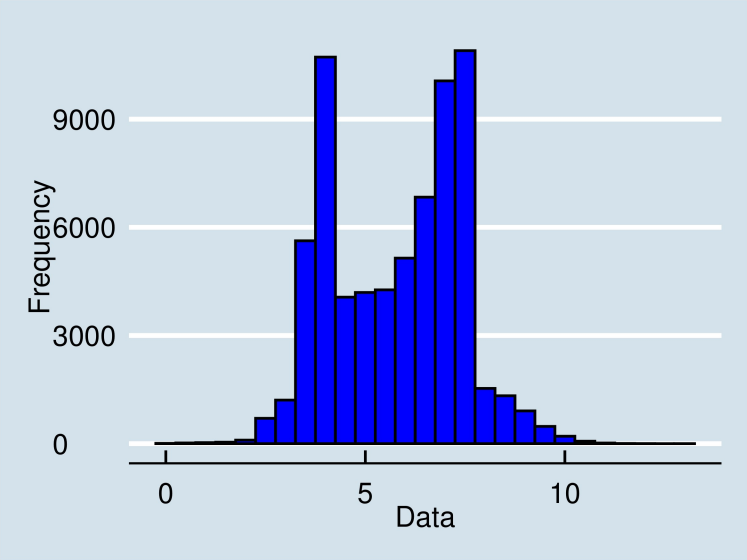


-

The seventh variable is Monthly Expenses. This is its histogram:

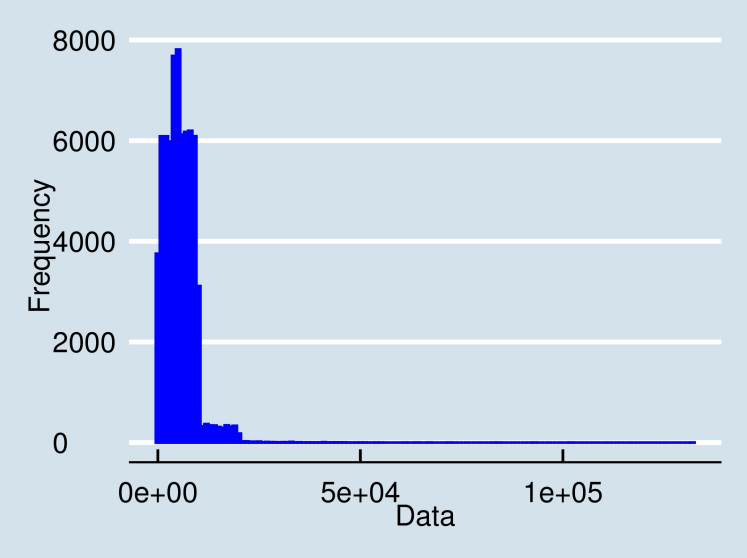


Same as DebitCard, I applied LN function. This is the new histogram:

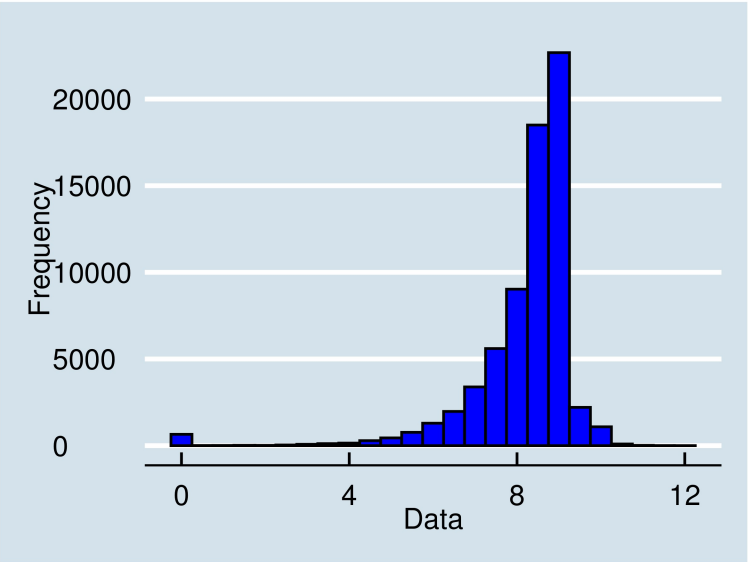


-

The eighth variable is Wage, this is a very important variable to estimate Default. This is its histogram:

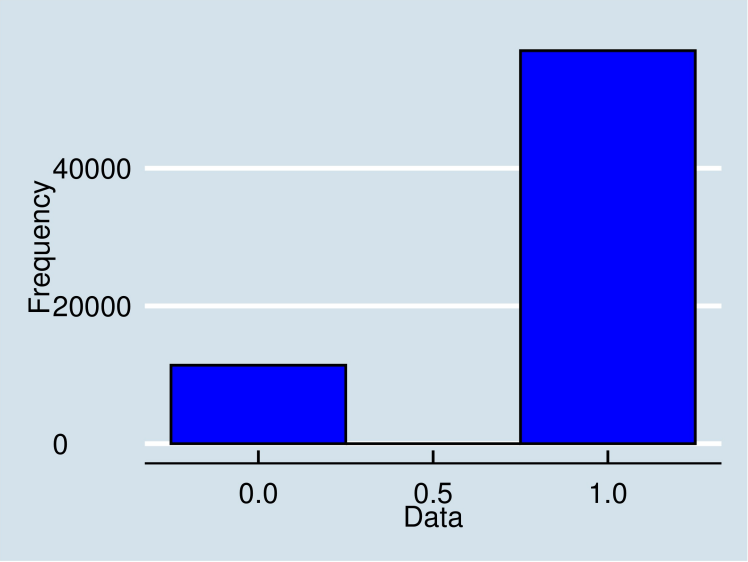


With the same logic as before, I applied the LN function. This is the new histogram:



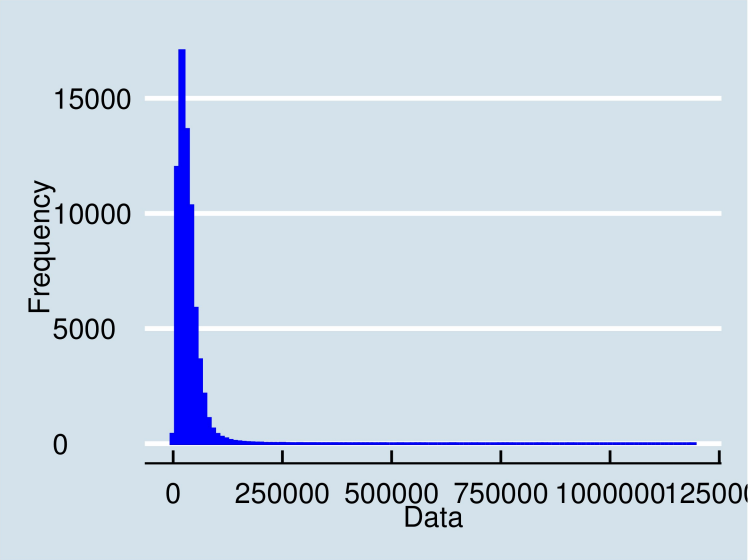
-

The ninth variable is a flag that shows if the person has used a passive product (like debit card) 12 months in a row. (It could be a good indicator of how much use the person gives to the bank's products). This is its histogram:

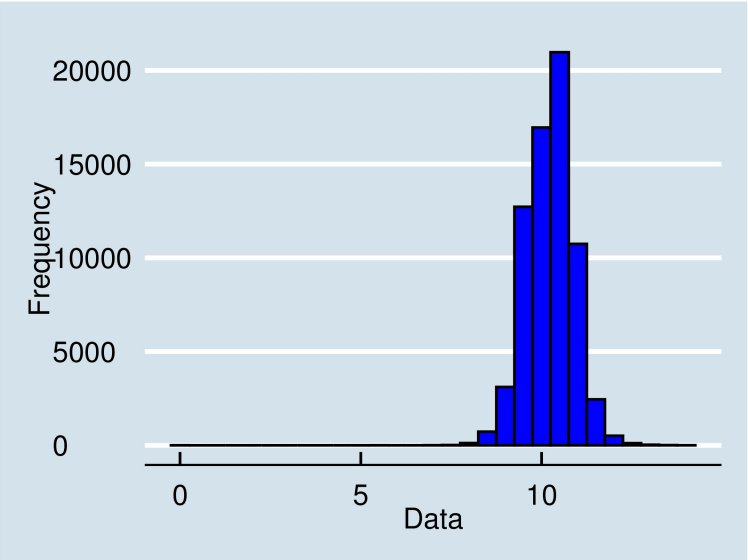


-

The tenth variable is the limit the person has in his/her credit card in the financial sector in Chile. This is its histogram:

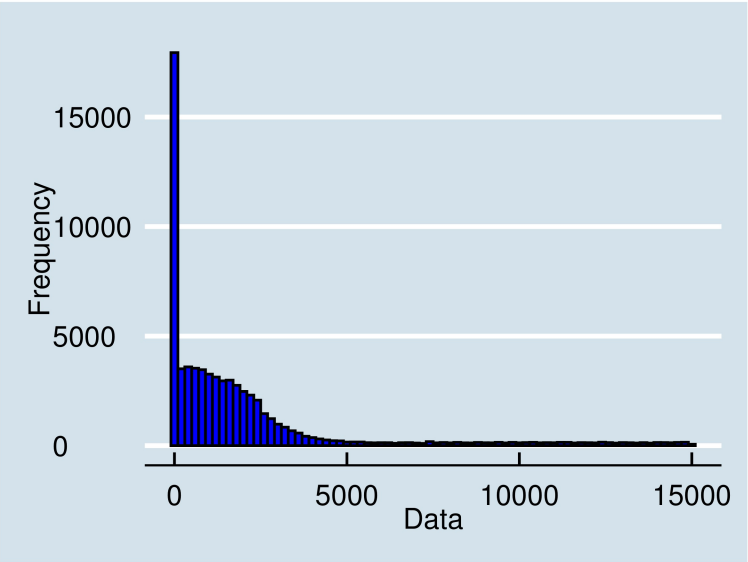


I applied the LN function. This is the new histogram:

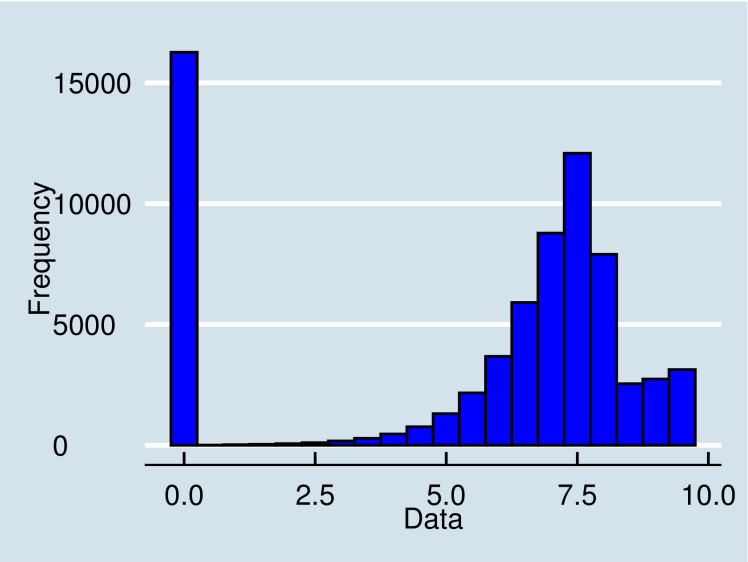


-

The last variable is Interest, and it shows the amount of interest (it could be in fix rent) the person won the last year. This is its histogram:



With the same logic as before, I applied LN function. This is the new histogram:



We have seen the distribution of all the variables. It was also shown how, on some particular cases, a transformation was applied, this was done trying to capture better the prediction power of the variables.

Bivariate Analysis

Next, I will conduct a bivariate analysis to compare the independent variables with each other and with the dependent variable. To evaluate the individual explanatory power of the independent variables on the dependent variable, I will use two performance metrics: the Kolmogorov-Smirnov (KS) statistic and the Area Under the Receiver Operating Characteristic (AUC-ROC) curve.

The KS statistic calculates the distance between the empirical distribution function of the sample and the cumulative distribution function of the reference distribution. A higher KS score indicates a better model. A score of 50 or higher is commonly considered a good score for a model.

The ROC curve plots the true positive rate against the false positive rate at various thresholds. The AUC

measures the performance of a classification model across various thresholds. A higher AUC score indicates a better model. A score of 80 or higher is commonly considered a good score for a model.

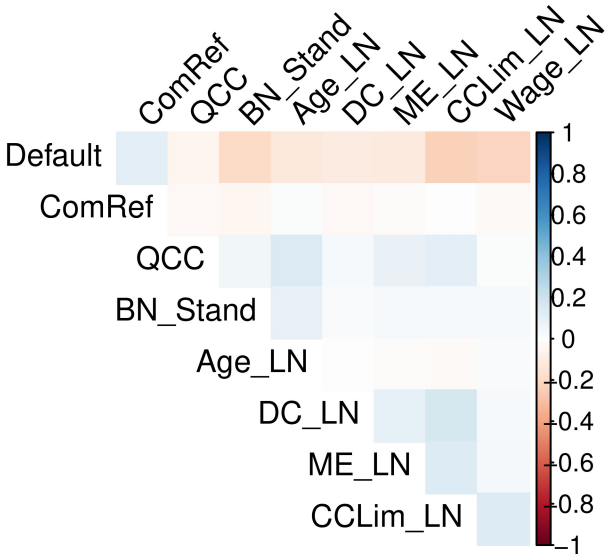
I will now create a univariate logistic regression model for each independent variable and the dependent variable. The results of the KS and AUC-ROC scores for each model are presented below:

```
##       varaible    KS
## 1   KS_Wage_LN 22.05
## 2    KS_ComRef 24.51
## 3       KS_QCC 15.36
## 4     KS_DC_LN 37.44
## 5    KS_Age_LN 31.48
## 6  KS_BN_Stand 22.91
## 7     KS_ME_LN 44.41
## 8      KS_Work 33.64
## 9      KS_I_LN  0.80
## 10   KS_M12_PP  0.64
## 11 KS_CCLim_LN 42.65
```

```
##        variable AUC.ROC
## 1   ROC_Wage_LN   47.30
## 2    ROC_ComRef   62.72
## 3       ROC_QCC   59.09
## 4     ROC_DC_LN   68.19
## 5    ROC_Age_LN   69.57
## 6  ROC_BN_Stand   66.30
## 7     ROC_ME_LN   67.85
## 8      ROC_Work   67.94
## 9      ROC_I_LN   49.90
## 10   ROC_M12_PP   50.32
## 11 ROC_CCLim_LN   77.30
```

It is evident that the variables “M12_PP” and “I” have a nearly 0 KS and nearly 50 ROC score. This indicates that they do not have any significant impact on the dependent variable, so they can be safely removed from the database.

Let’s proceed to create a correlation plot to analyze the correlations between the variables. To do this, I will exclude the “Work” variable as it is a categorical variable.



There is not a strong correlation between independent variables.

Multivariate models

Let's now start with the different models that will be tested. I have tested 5 different models:

1_ Linear Regression with the intercept only

2_ Multivariate Linear Regression

3_ Logistic Regression with the intercept only

4_ Multivariate Logistic Regression

5_ Decision Trees

First of all, it is necessary to transform the factor variable into dummies and to divide the database between train and test

```
# First, I have to convert to dummies the factor variable
factor_variable <- "Work"
DataBase_wDummies <- as.data.frame(get_dummies.(DataBase))
DataBase_wDummies <- DataBase_wDummies[,-c(which(colnames(DataBase_wDummies) == "Work"))]


# Then, I set a seed
seed = 1234567
set.seed(seed)


# I divided the database between train and test.
```

```
Muestra = 0.9
Coord = sample(nrow(DataBase_wDummies),nrow(DataBase_wDummies)*Muestra)
Test = DataBase_wDummies[-Coord,]
Train = DataBase_wDummies[Coord,]
```

1ST MODEL: LINEAR MODEL - Intercept Only

```
Model_Linear_null <- glm(Default ~ 1, data=Train)
```

Let's see the results:

```
KS_lm_null
```

```
## [1] 0
```

```
ROC_lm_null
```

```
## [1] 50
```

This model does not explain the default. A KS = 0 and a ROC = 50 are similar to an aleatory decision between default and no default.

2ND MODEL: MULTIVARIATE LINEAR MODEL

```
Model_Linear <- lm(Default ~ 1 +  . ,data = Train)
```

Let's see the results:

```
KS_lm
```

```
## [1] 52.95
```

```
ROC_lm
```

```
## [1] 83.44
```

This model is much better than the one with only the intercept. Let's try to improve it with a logistic.

3RD MODEL: LOGISTIC REGRESSION - Intercept Only

```
Model_Logistic_null <- glm(Default ~ 1, data=Train, family = "binomial")
```

Let's see the results:

```
KS_glmL_null
```

```
## [1] 0
```

```
ROC_glmL_null
```

```
## [1] 50
```

The same as the first one, a model with only an intercept does not explain the default.

4TH MODEL: MULTIVARIATE LOGISTIC MODEL

```
Model_Logistic <- glm(Default ~ 1 +  . ,data = Train , family = "binomial")
```

Let's see the results:

```
KS_glmL_log
```

```
## [1] 61.13
```

```
ROC_glmL_log
```

```
## [1] 88.57
```

Here we can see how the logistic model improves the linear one. This is mainly because the dependent variable is binary and the logistic model applies much better in these cases.

An extra analysis could be to take a look at the p-values of the variables:

```
pvalue <- as.data.frame(summary(Model_Logistic)$coefficients)
pvalue <- pvalue[order(pvalue$`Pr(>|z|)`),]
pvalue
```

```
##                Estimate Std. Error    z value      Pr(>|z|)
## (Intercept) 19.22265070 0.41710930  46.085405  0.000000e+00
## CCLim_LN    -1.75154746 0.04093005 -42.793676  0.000000e+00
## BN_Stand    -0.50843254 0.01881561 -27.021849 8.184535e-161
## ME_LN       -0.33665268 0.01572742 -21.405459 1.188436e-101
## ComRef       0.28880871 0.01724073  16.751538  5.518149e-63
## Wage_LN     -0.18553275 0.01144273 -16.214023  4.014127e-59
## DC_LN       -0.24978881 0.01750692 -14.268007  3.463380e-46
## Age_LN      -0.25717356 0.02513033 -10.233591  1.402208e-24
## Work_PRIV    0.97499729 0.11242902   8.672114  4.241716e-18
## Work_JUB    -0.91756570 0.13582940  -6.755281  1.425591e-11
## QCC         -0.06604082 0.02435981  -2.711057  6.706914e-03
## Work_PUB     0.42648340 0.17051053   2.501214  1.237682e-02
## Work_ST      0.23841144 0.15348719   1.553299  1.203518e-01
```

Here we can see that all the p-values are quite small, so all the variables are significant to the model.

5TH MODEL: DECISION TREES

```
# Creation of the model
tree<-rpart(Default~., Train)
```

Let's now see the performance of this model:

```
KS_tree
```

```
## [1] 60.5
```

```
ROC_tree
```

```
## [1] 81.48
```

The KS and the ROC acceptable in this model. It can discriminate quite well the defaults.

Resume and finals results:

Let's analyze all models results at once:

```
##                        model    KS   ROC
## 1   Linear - only Intercept  0.00 50.00
## 2       Multivariate Linear 52.95 83.44
## 3 Logistic - only Intercept  0.00 50.00
## 4     Multivariate Logistic 61.13 88.57
## 5             Decision Tree 60.50 81.48
```

It is important to note that a model with a KS score of 60 or above and an AUC-ROC score of 80 or above can be considered a good model.

The model with the best KS is:

```
## [1] "Multivariate Logistic"
```

The model with the best AUC-ROC is:

```
## [1] "Multivariate Logistic"
```

Conclusion

In conclusion, the analysis suggests that traditional linear regression models are not suitable for predicting the probability of default in this particular project. However, alternative models such as logistic regression and decision trees demonstrate potential for accurately estimating default probabilities. Among these models, the Multivariate Logistic Model performed best in terms of accuracy and demonstrated the highest potential for delivering reliable predictions.

It should be emphasized that selecting the ultimate algorithm is just as crucial as utilizing high-quality information. Achieving KS $>$ 60 and ROC $>$ 80 is attainable with the combined strength of well-prepared data and appropriate algorithms.