Counterfeit detection

Billy Tran

04/01/2020

Counterfeit detection

My IT consulting company offers a new mission to the Ministry of the Interior, as part of the fight against organized crime, the central office for the suppression of counterfeiting money. My mission is to create a counterfeit detection algorithm.

Part 1 - Introduction

This section describes the dataset and summarizes the goal of the project and key steps that were performed.

1.1 Installing and loading packages

We are going to use the following library:

1.2 Loading data

Here is a sample of my data.

```
##   is_genuine diagonal height_left height_right margin_low margin_up length
## 1       True   171.81      104.86       104.95       4.52      2.89 112.83
## 2       True   171.67      103.74       103.70       4.01      2.87 113.29
## 3       True   171.83      103.76       103.76       4.40      2.88 113.84
## 4       True   171.80      103.78       103.65       3.73      3.12 113.63
## 5       True   172.05      103.70       103.75       5.04      2.27 113.55
## 6       True   172.57      104.65       104.44       4.54      2.99 113.16
```

1.3 Data description

```
## 70 are fake bills.
```

```
## 100 are real bills.
```

There is 170 bills in total. Each bill is described by a state (real : TRUE & fake : FALSE) and 6 geometric characteristics lengths : diagonal, height_left, height_right, margin_low, margin_up, length

My aim is to create a counterfeit detection algorithm according to this dataset.

Part 2 - Analysis

This section that explains the process and techniques used, such as data cleaning, data exploration and visualization, any insights gained, and my modeling approach.

2.1 Data cleaning

We work with real data: we must take into account the fact that some data may be missing, outliers or atypical. We must clean our dataset before our analyzes.

Data Structure

```
## 'data.frame':    170 obs. of  7 variables:
##  $ is_genuine  : Factor w/ 2 levels "False","True": 2 2 2 2 2 2 2 2 2 2 ...
##  $ diagonal    : num  172 172 172 172 172 ...
##  $ height_left : num  105 104 104 104 104 ...
##  $ height_right: num  105 104 104 104 104 ...
##  $ margin_low  : num  4.52 4.01 4.4 3.73 5.04 4.54 3.97 3.54 4.06 4.63 ...
##  $ margin_up   : num  2.89 2.87 2.88 3.12 2.27 2.99 2.9 3.19 3.33 3.02 ...
##  $ length      : num  113 113 114 114 114 ...
```

The categorical variable is_genuine is of class 'factor'. All the other columns represent the dimensions of the ticket: they are in digital format. There are 170 observations for 7 variables.

Missing data

Here are the number of missing data per variable.

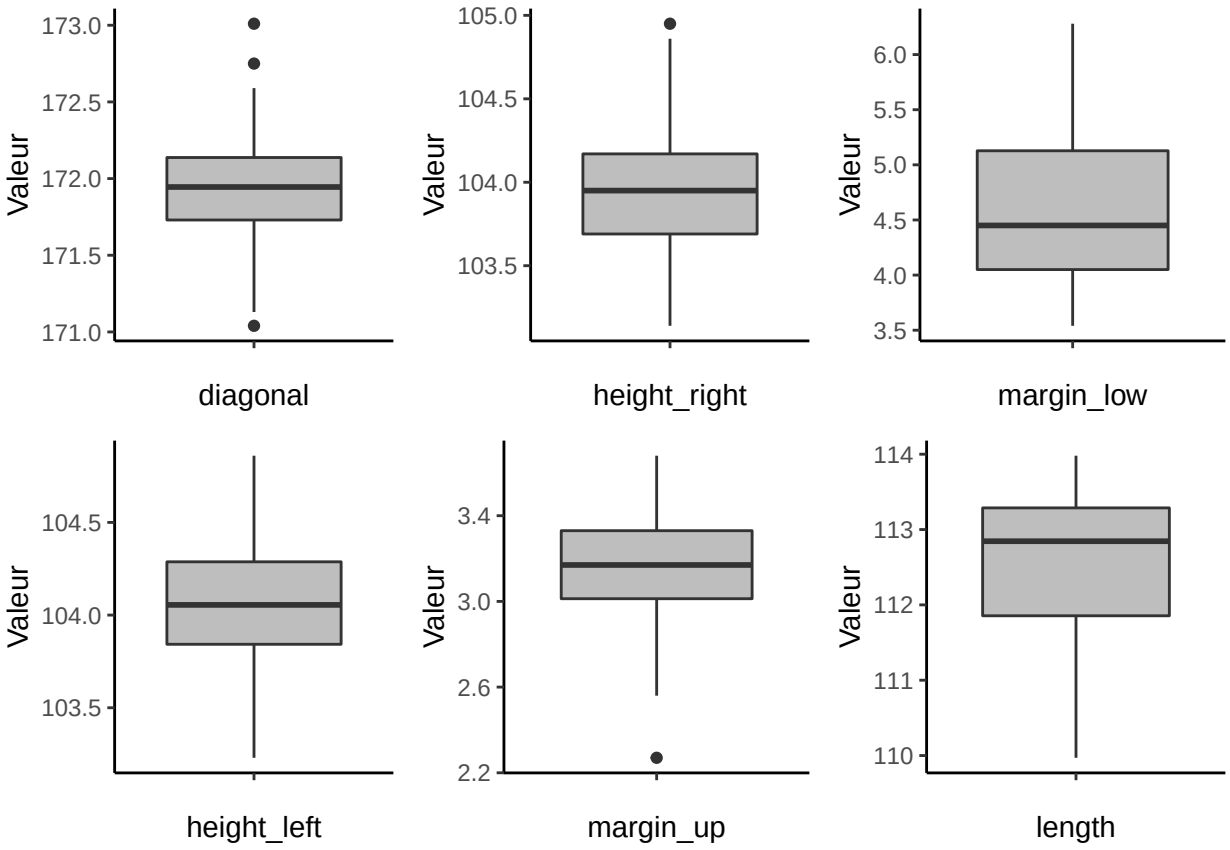
```
##   is_genuine     diagonal  height_left height_right   margin_low    margin_up
##            0            0            0            0            0            0
##       length
##            0
```

There is no missing data.

Outliers and unusual data

Here is a summary of the data :

```
##  is_genuine     diagonal      height_left     height_right     margin_low
##  False: 70   Min.   :171.0   Min.   :103.2   Min.   :103.1   Min.   :3.540
##  True :100   1st Qu.:171.7   1st Qu.:103.8   1st Qu.:103.7   1st Qu.:4.050
##              Median :171.9   Median :104.1   Median :104.0   Median :4.450
##              Mean   :171.9   Mean   :104.1   Mean   :103.9   Mean   :4.612
##              3rd Qu.:172.1   3rd Qu.:104.3   3rd Qu.:104.2   3rd Qu.:5.128
##              Max.   :173.0   Max.   :104.9   Max.   :105.0   Max.   :6.280
##    margin_up         length
##  Min.   :2.270   Min.   :110.0
##  1st Qu.:3.013   1st Qu.:111.9
##  Median :3.170   Median :112.8
##  Mean   :3.170   Mean   :112.6
##  3rd Qu.:3.330   3rd Qu.:113.3
##  Max.   :3.680   Max.   :114.0
```



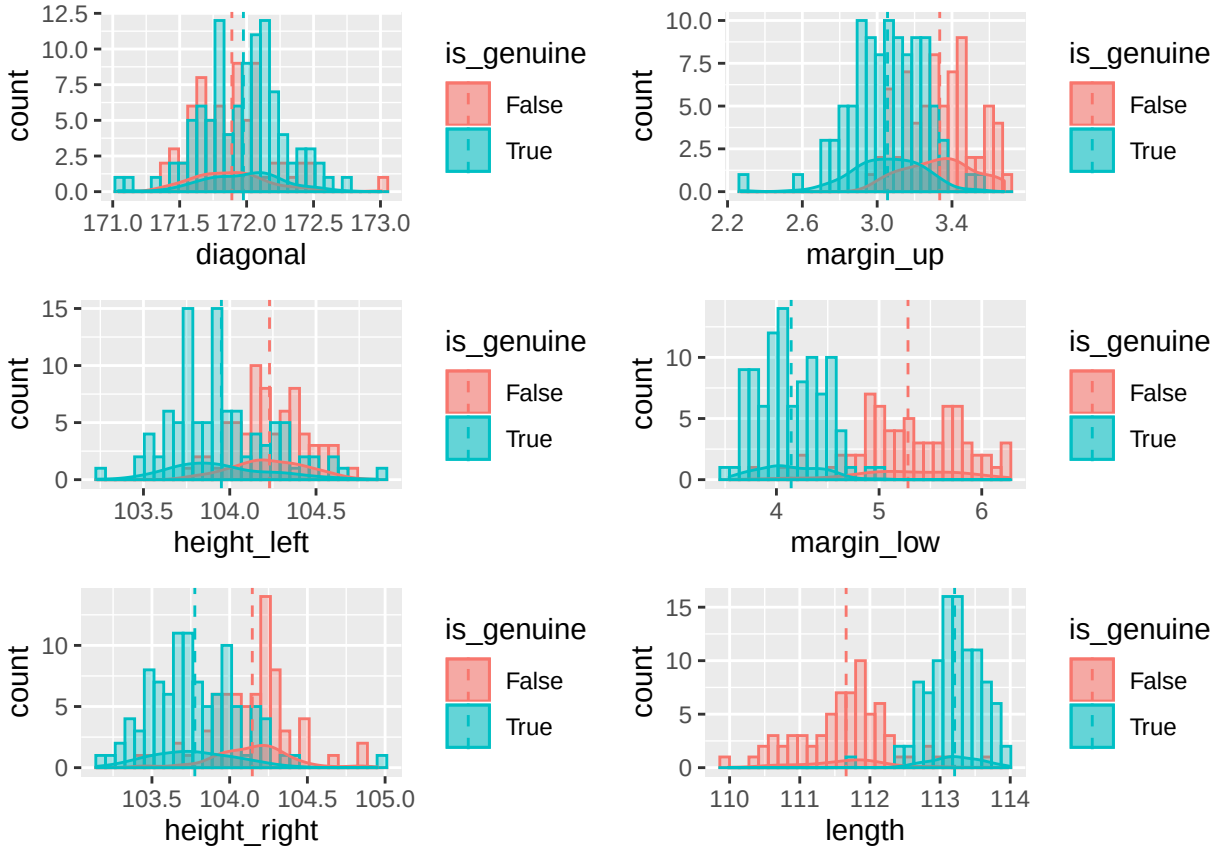
Paying attention to the min and max values, we deduce that there is no outlier or atypical value. The fact that there are some (weak) outliers is possible in the context of the mission. In addition to that, we note that values are very large compared to others: it will have to standardize the data so that the big values take no precedence over small.

Categorical variable

```
## [1] "False" "True"
```

There are two categories of tickets: real and fake.

2.2 Data visualization



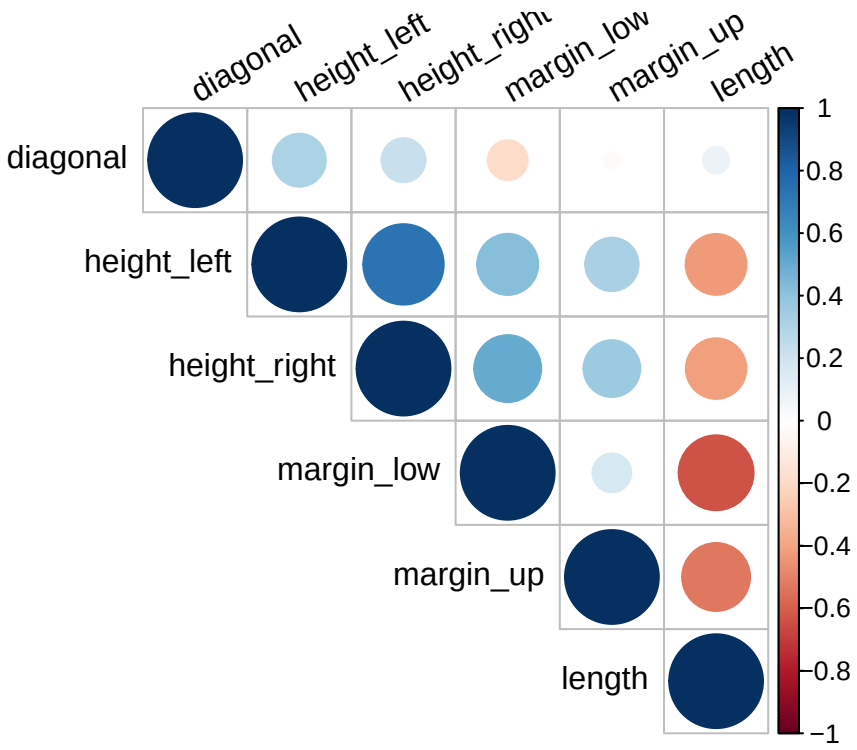
- Observations
  - In both cases: the diagonal, height_left, height_right, margin_up values are approximately distributed in the same range or the False interval is included in the True range.
  - For margin_low: the interval of the real bills is included in [3.5 - 5] while that of the counterfeit bills is included in [4 - 6.5].
  - For length: the interval of the real notes is included in [112 - 114] while that of the counterfeit bills is included in [110 - 114].
- It would seem, then, that what distinguishes the true from the counterfeit bills are :
  - The top margin of the bills
  - The lower margin of the bills
  - The length of the bills

2.3 Data exploration

2.3.1 Correlation Matrix

To observe the possible correlations between varibales, the correlation matrix between the variables is represented.

Figure 3 − Correlation matrix



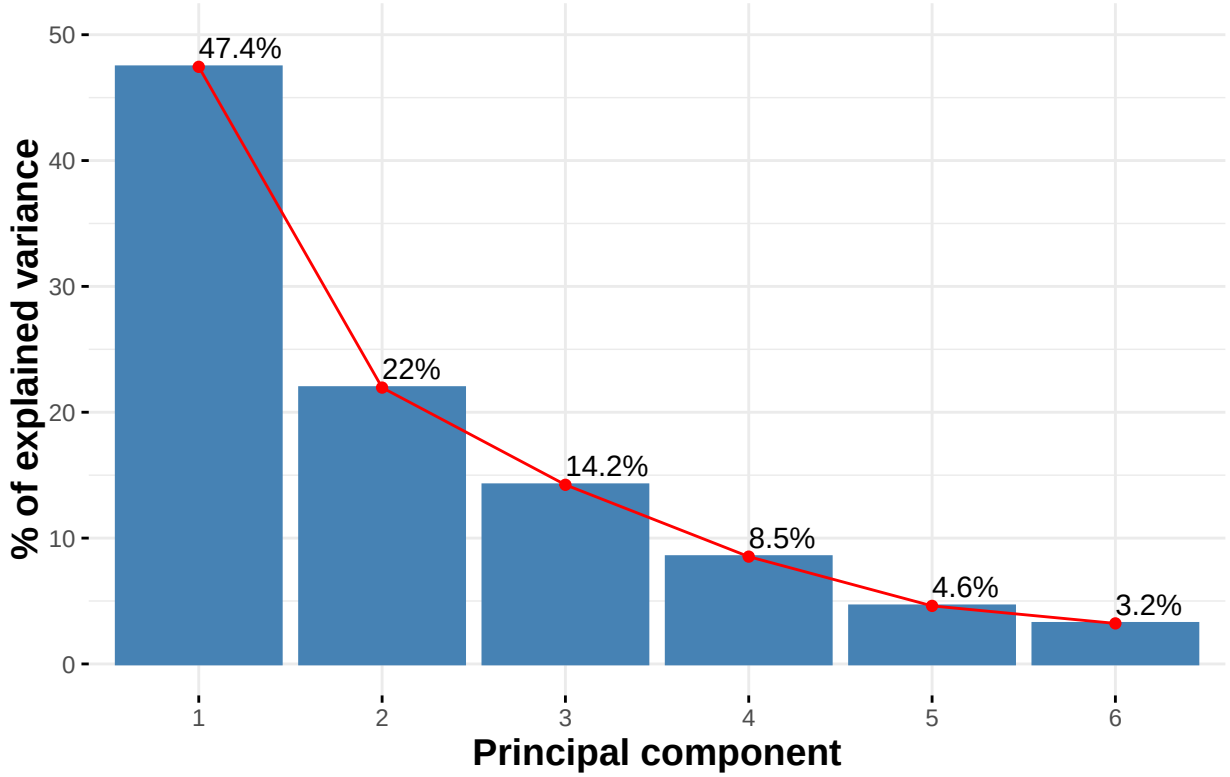
In the first part, we describe our data, our bills and the correlations which could exist between the geometrical characteristics of the ticket. We have noticed from the histograms that what distinguishes the most the counterfeits from the real bills are the lower margin, the upper margin, and the height.

2.3.2 Principal Component Analysis

- Active individuals (in blue, lines 1: 170): Individuals that are used in the principal component analysis.
- Active variables (in pink, columns 2: 6): variables used for PCA.
- Additional qualitative variables (in green, column 1): the coordinates of these variables will be predicted: column 1 characterizes whether the ticket is a real one or not. This is a categorical variable. It can be used to color individuals in groups.

Scree Plot Eigenvalues : Percentage of variances explained by each principal axis

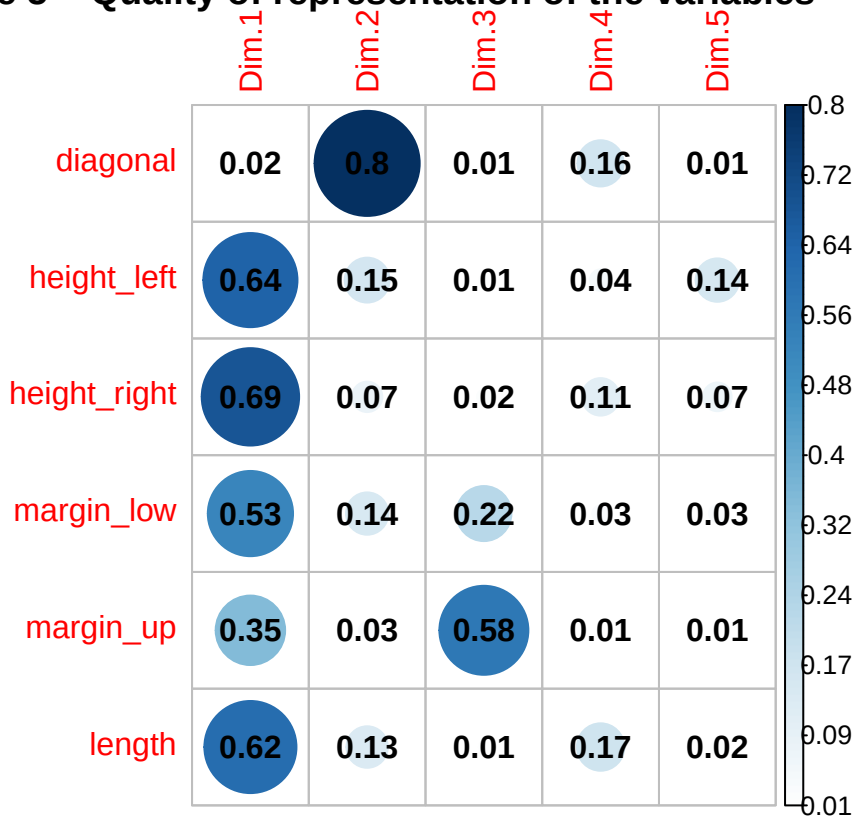
Figure 4 – Scree Plot Eigenvalues

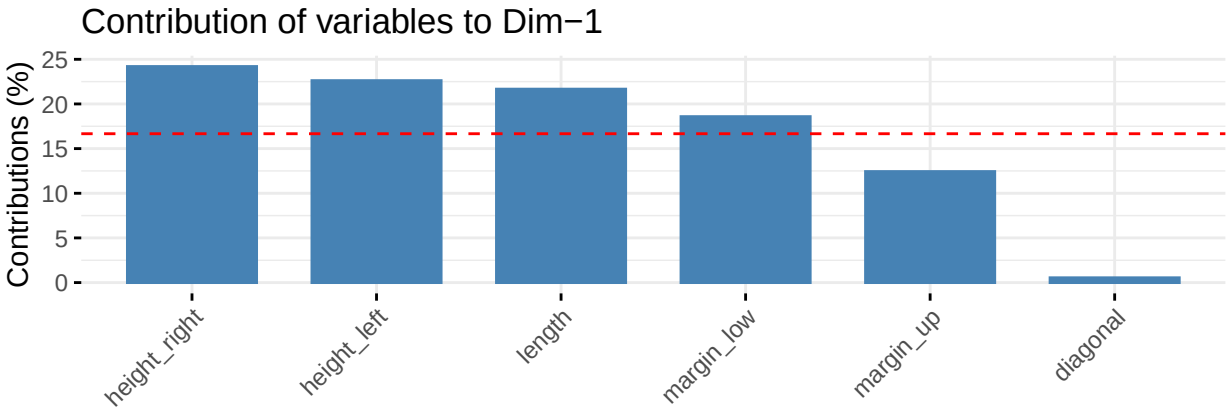


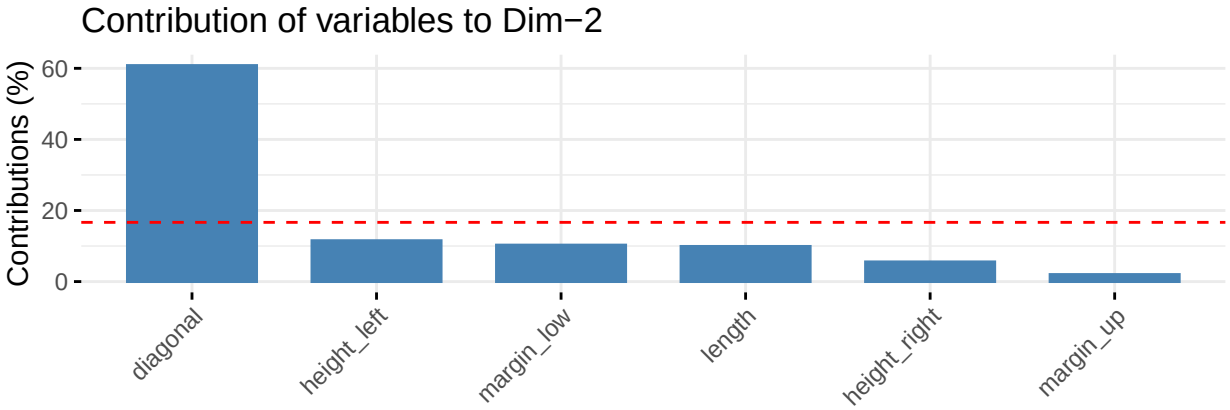
Elbow Method : from the 2nd factorial axis, we observe a setback (or bend) in the decay followed by a more regular decay. We will interpret the first two axes that explain 69.9% of the variance

Quality of representation of the variables on the factorial axes

Figure 5 − Quality of representation of the variables



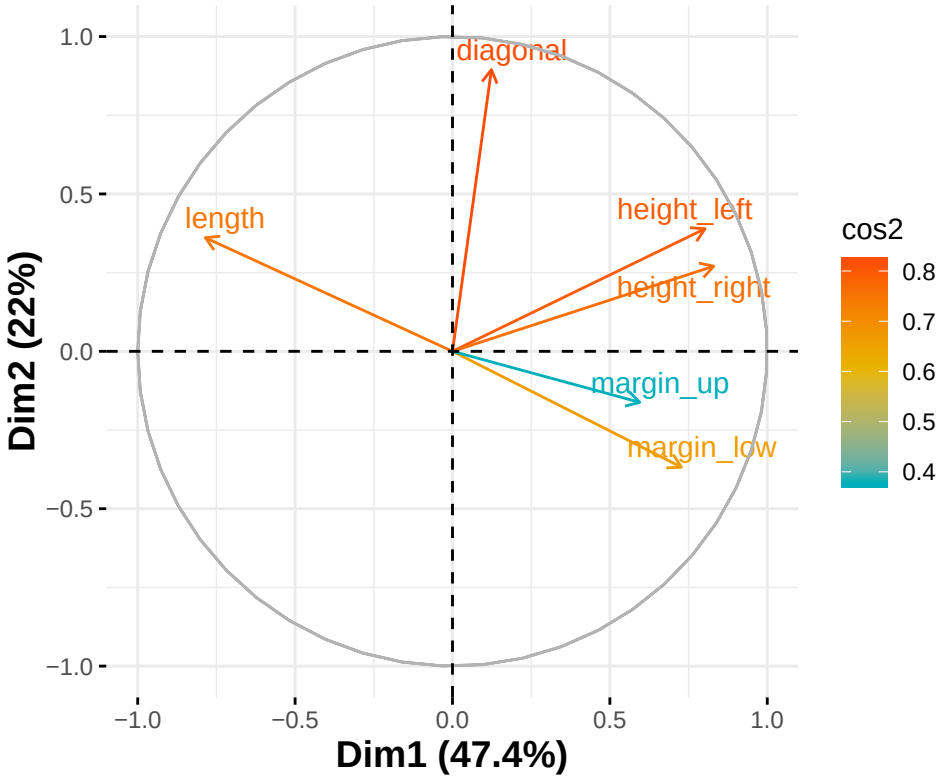




- height_left, height_right, margin_low and length contribute the most and are perfectly represented by the first main component (minus for margin_up)
- Diagonal contributes the most and is very well represented by the second main component.

Correlation Circle

Figure 7 − Correlation Circle



The correlation graph of the variables shows the relationships between all the variables and the factorial axes. It can be interpreted as follows:

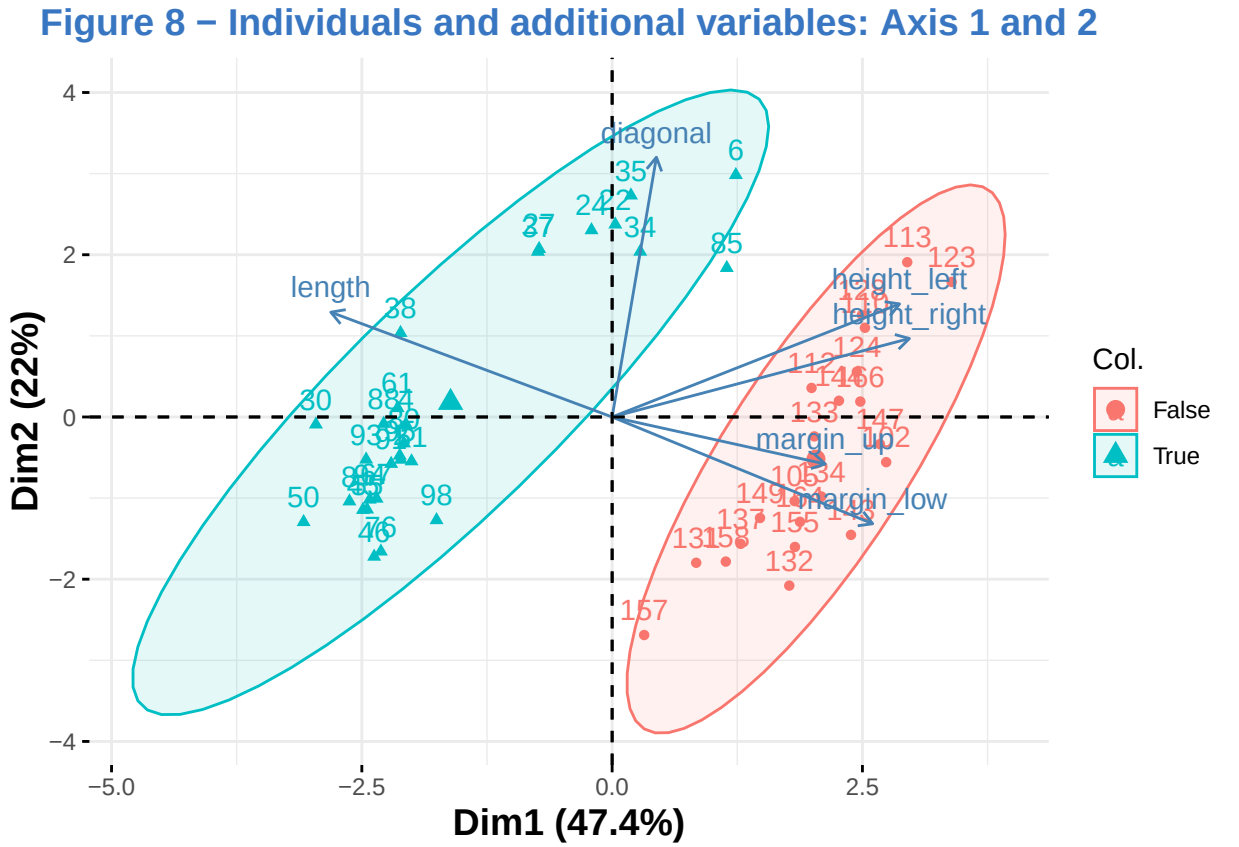
- The positively correlated variables are grouped:
  - height_left and height_right
  - margin_low and margin_up
- The negatively correlated variables are positioned on the opposite sides of the graph origin (opposite quadrants):
  - length compared to others
- The distance between the variables and the origin measures the quality of representation of the variables. Variables that are far from the origin are well represented by the ACP (red / orange color). All the variables (slightly less for margin_up: blue color) are well explained by the first two main components (Dim.1 & Dim.2) because they are positioned close to the correlation circle.
- Axis 1 represents the vertical and horizontal dimensions of the bill : + a bill is on the right, + its vertical dimensions are large and - its horizontal dimensions are high.
- Axis 2 represents the diagonal of the note. : + a bill is up, + its diagonal is high

We find the observations on the first factorial axis: the correlated variables (close to the first factorial axis) are grouped in the same direction:

- on the right in the factorial plane and in blue: height_left, height_right, margin_low, margin_up

- on the left in the factorial plane and in red: length

Representation of individuals



The projection of the 50 individuals who contribute the most to the formation of the first factorial plane and having a quality of representation cos2 greater than 0.8 separates very well the two categories of bills. We note that the reality of a bills is directed only on one axis: the one led by lenght and margin_low.

- The "true" criterion of a bill is characterized by a good value of length.
- The "false" criterion of a bill is characterized by a high value for margin_low.

In this part, by using a principal component analysis, we were able to confirm our initial hypothesis. The variables that contribute the most to the categorization of the bills are the length of the bill, its lower margin and its top margin.

2.4 My modeling approach

Here is my modeling approach of my next part. In the next Two models will be evaluated by cross-validation: **decision tree** and **logistic regression**.

- Cross validation is performed with 5 folds, 5 parts stratified with respect to the class of the note.
- At each iteration, 4 parts will be used to train the dataset. And a part will be used to calculate a performance indicator: the RSME mean squared error. This is a total distance between the actual classes and the values taken by the model made on the training game.
- At the end of the 5 iterations of cross-validation, the average RSME is calculated from the 5 RSMEs.

- The best model is the one that minimizes the average RSME. The best model will then be trained on the entire dataset. Modeling derived from cross validations will not be used in the final predictive model since they are only trained on 80% of the dataset.

Part 3 - Modeling results and model performance

This section presents the modeling results and discusses the model performance.

3.1 Preparing cross-validation

3.2 Modeling : Decision tree

```
## CART
##
## 170 samples
##   6 predictor
##   2 classes: 'False', 'True'
##
## No pre-processing
## Resampling: Cross-Validated (5 fold)
## Summary of sample sizes: 136, 136, 136, 136, 136
## Resampling results across tuning parameters:
##
##   cp         Accuracy   Kappa
##   0.0000000  0.9470588  0.8905634
##   0.2214286  0.9411765  0.8775193
##   0.4428571  0.9411765  0.8775193
##   0.6642857  0.9411765  0.8775193
##   0.8857143  0.7117647  0.3143424
##
## Accuracy was used to select the optimal model using the largest value.
## The final value used for the model was cp = 0.
```

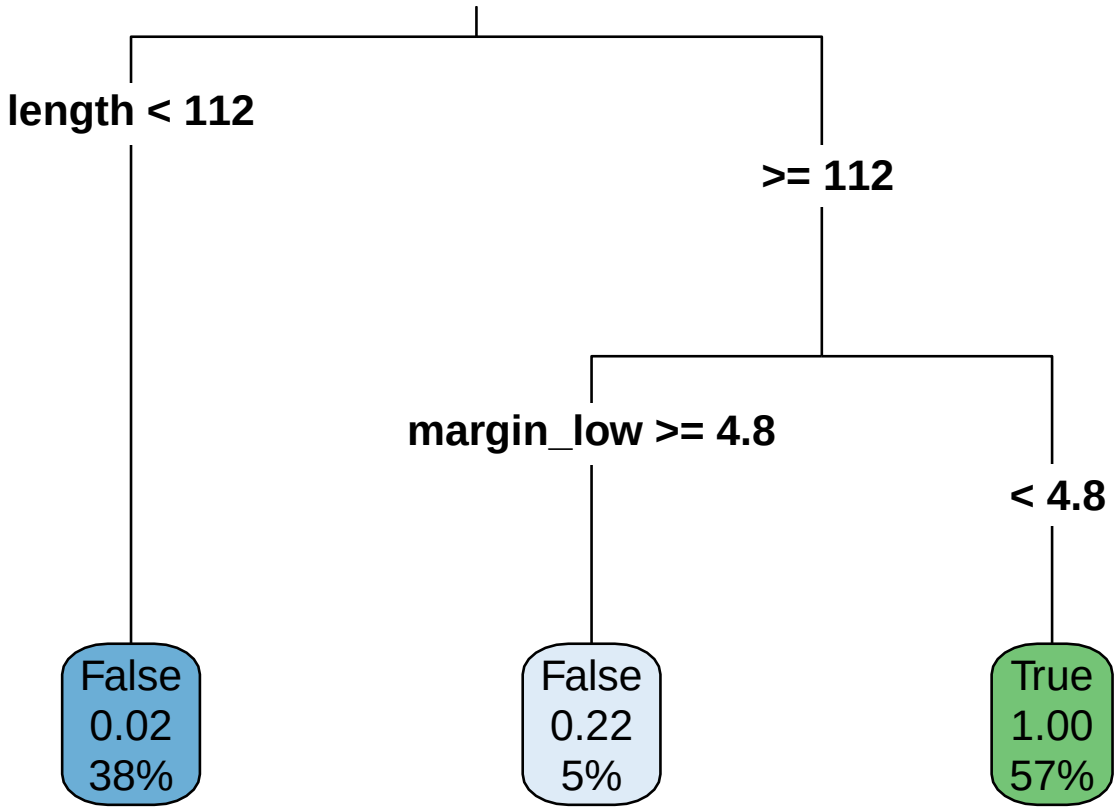
Let be : alpha = accuracy et RMSE = root mean squared error. In a case of a binary problem, these scores can be calculated according to true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN). The total size of the dataset is : Omega=TP+TN+FP+FN. Otherwise :

- alpha=(TP+TN)/Omega
- RMSE=sqrt((FP+FN)/Omega

So : alpha+RMSEˆ2=1

```
## With the decision tree, the average RSME for cross-validation is 0.19 .
```

In the following figure, the tree is plotted according to the model driven on 80% of the dataset. This is an illustrative example for understanding.



Reading the decision tree According to the decision tree, a bill is true if:

- first, if its length is higher than or equal to 112 mm
- then, if its lower margin is lower than 4.8 mm.

This decision tree confirms our observations made in Parts 1 and 2 where we said that the most distinguishing features of counterfeits were their lower margin and their length.

3.3 Modeling : Logistic regression

```
##   is_genuine diagonal height_left height_right margin_low margin_up length
## 1          1   171.81      104.86       104.95       4.52      2.89 112.83
## 2          1   171.67      103.74       103.70       4.01      2.87 113.29
## 3          1   171.83      103.76       103.76       4.40      2.88 113.84
## 4          1   171.80      103.78       103.65       3.73      3.12 113.63
## 5          1   172.05      103.70       103.75       5.04      2.27 113.55
## 6          1   172.57      104.65       104.44       4.54      2.99 113.16
```

```
## Warning in train.default(trainX, trainY, method = "glmStepAIC", direction =
## "both", : You are trying to do regression and your outcome only has two possible
## values Are you trying to do classification? If so, use a 2 level factor as your
## outcome column.
```

```
## Start:  AIC=-84.78
## .outcome ~ diagonal + height_left + height_right + margin_low +
```

```
##     margin_up + length
##
##                Df Deviance     AIC
## - diagonal      1   3.8093 -86.278
## - height_left   1   3.8142 -86.102
## <none>              3.7953 -84.779
## - height_right  1   3.8840 -83.636
## - length        1   5.0423 -48.139
## - margin_up     1   5.5232 -35.751
## - margin_low    1   7.0210  -3.119
##
## Step:  AIC=-86.28
## .outcome ~ height_left + height_right + margin_low + margin_up +
##     length
##
##                Df Deviance     AIC
## - height_left   1   3.8485 -86.884
## <none>              3.8093 -86.278
## - height_right  1   3.8922 -85.349
## + diagonal      1   3.7953 -84.779
## - length        1   5.1005 -48.579
## - margin_up     1   5.5871 -36.186
## - margin_low    1   7.3645   1.377
##
## Step:  AIC=-86.88
## .outcome ~ height_right + margin_low + margin_up + length
##
##                Df Deviance     AIC
## - height_right  1   3.8936 -87.299
## <none>              3.8485 -86.884
## + height_left   1   3.8093 -86.278
## + diagonal      1   3.8142 -86.102
## - length        1   5.1013 -50.558
## - margin_up     1   5.6068 -37.708
## - margin_low    1   7.4465   0.884
##
## Step:  AIC=-87.3
## .outcome ~ margin_low + margin_up + length
##
##                Df Deviance     AIC
## <none>              3.8936 -87.299
## + height_right  1   3.8485 -86.884
## + diagonal      1   3.8842 -85.631
## + height_left   1   3.8922 -85.349
## - length        1   5.1452 -51.392
## - margin_up     1   6.2011 -26.006
## - margin_low    1   8.4267  15.701
## Start:  AIC=-74.26
## .outcome ~ diagonal + height_left + height_right + margin_low +
##     margin_up + length
##
##                Df Deviance     AIC
## - diagonal      1   4.1077 -76.020
## - height_left   1   4.1082 -76.004
```

```
## - height_right  1   4.1499 -74.630
## <none>              4.1005 -74.260
## - length        1   5.0672 -47.470
## - margin_up     1   6.3604 -16.557
## - margin_low    1   7.9350  13.524
##
## Step:  AIC=-76.02
## .outcome ~ height_left + height_right + margin_low + margin_up +
##     length
##
##                Df Deviance     AIC
## - height_left   1   4.1131 -77.842
## - height_right  1   4.1677 -76.047
## <none>              4.1077 -76.020
## + diagonal      1   4.1005 -74.260
## - length        1   5.1110 -48.300
## - margin_up     1   6.4023 -17.663
## - margin_low    1   8.4708  20.410
##
## Step:  AIC=-77.84
## .outcome ~ height_right + margin_low + margin_up + length
##
##                Df Deviance     AIC
## <none>              4.1131 -77.842
## - height_right  1   4.1843 -77.508
## + height_left   1   4.1077 -76.020
## + diagonal      1   4.1082 -76.004
## - length        1   5.1233 -49.974
## - margin_up     1   6.4179 -19.333
## - margin_low    1   8.4764  18.502
## Start:  AIC=-92.98
## .outcome ~ diagonal + height_left + height_right + margin_low +
##     margin_up + length
##
##                Df Deviance     AIC
## - diagonal      1   3.5774 -94.819
## - height_right  1   3.5939 -94.193
## <none>              3.5731 -92.982
## - height_left   1   3.6573 -91.817
## - length        1   4.7025 -57.628
## - margin_up     1   5.6079 -33.681
## - margin_low    1   7.6945   9.339
##
## Step:  AIC=-94.82
## .outcome ~ height_left + height_right + margin_low + margin_up +
##     length
##
##                Df Deviance     AIC
## - height_right  1   3.6016 -95.902
## <none>              3.5774 -94.819
## - height_left   1   3.6578 -93.797
## + diagonal      1   3.5731 -92.982
## - length        1   4.7359 -58.666
## - margin_up     1   5.6940 -33.609
```

```
## - margin_low    1   8.2691  17.134
##
## Step:  AIC=-95.9
## .outcome ~ height_left + margin_low + margin_up + length
##
##                Df Deviance     AIC
## <none>              3.6016 -95.902
## - height_left   1   3.6590 -95.752
## + height_right  1   3.5774 -94.819
## + diagonal      1   3.5939 -94.193
## - length        1   4.7387 -60.585
## - margin_up     1   6.0319 -27.770
## - margin_low    1   9.1675  29.160
## Start:  AIC=-91.9
## .outcome ~ diagonal + height_left + height_right + margin_low +
##     margin_up + length
##
##                Df Deviance     AIC
## - diagonal      1   3.6090 -93.624
## - height_left   1   3.6187 -93.257
## <none>              3.6018 -91.895
## - height_right  1   3.7055 -90.036
## - margin_up     1   5.3506 -40.070
## - length        1   5.3867 -39.154
## - margin_low    1   6.8289  -6.891
##
## Step:  AIC=-93.62
## .outcome ~ height_left + height_right + margin_low + margin_up +
##     length
##
##                Df Deviance     AIC
## - height_left   1   3.6350 -94.647
## <none>              3.6090 -93.624
## - height_right  1   3.7066 -91.996
## + diagonal      1   3.6018 -91.895
## - length        1   5.3878 -41.127
## - margin_up     1   5.4234 -40.231
## - margin_low    1   7.3233   0.614
##
## Step:  AIC=-94.65
## .outcome ~ height_right + margin_low + margin_up + length
##
##                Df Deviance     AIC
## <none>              3.6350 -94.647
## - height_right  1   3.7087 -93.916
## + height_left   1   3.6090 -93.624
## + diagonal      1   3.6187 -93.257
## - length        1   5.3885 -43.110
## - margin_up     1   5.4470 -41.640
## - margin_low    1   7.3558  -0.783
## Start:  AIC=-87.62
## .outcome ~ diagonal + height_left + height_right + margin_low +
##     margin_up + length
##
```

```
##                Df Deviance     AIC
## - height_left   1   3.7409 -88.740
## - diagonal      1   3.7519 -88.343
## <none>              3.7168 -87.619
## - height_right  1   3.7808 -87.297
## - length        1   4.8919 -52.260
## - margin_up     1   5.9394 -25.871
## - margin_low    1   8.1432  17.048
##
## Step:  AIC=-88.74
## .outcome ~ diagonal + height_right + margin_low + margin_up +
##     length
##
##                Df Deviance     AIC
## - height_right  1   3.7811 -89.286
## <none>              3.7409 -88.740
## - diagonal      1   3.8032 -88.494
## + height_left   1   3.7168 -87.619
## - length        1   4.8945 -54.188
## - margin_up     1   5.9446 -27.752
## - margin_low    1   8.1445  15.068
##
## Step:  AIC=-89.29
## .outcome ~ diagonal + margin_low + margin_up + length
##
##                Df Deviance     AIC
## - diagonal      1   3.8154 -90.059
## <none>              3.7811 -89.286
## + height_right  1   3.7409 -88.740
## + height_left   1   3.7808 -87.297
## - length        1   4.9169 -55.565
## - margin_up     1   6.3733 -20.282
## - margin_low    1  10.1955  43.615
##
## Step:  AIC=-90.06
## .outcome ~ margin_low + margin_up + length
##
##                Df Deviance     AIC
## <none>              3.8154 -90.059
## + diagonal      1   3.7811 -89.286
## + height_right  1   3.8032 -88.494
## + height_left   1   3.8046 -88.445
## - length        1   4.9388 -56.961
## - margin_up     1   6.4216 -21.255
## - margin_low    1  10.5737  46.569
## Start:  AIC=-109.49
## .outcome ~ diagonal + height_left + height_right + margin_low +
##     margin_up + length
##
##                Df Deviance      AIC
## - diagonal      1   4.7614 -111.354
## - height_left   1   4.7935 -110.211
## <none>              4.7576 -109.489
## - height_right  1   4.8406 -108.551
```

```
## - length        1   6.3665  -61.969
## - margin_up     1   7.2413  -40.081
## - margin_low    1   9.4446    5.079
##
## Step:  AIC=-111.35
## .outcome ~ height_left + height_right + margin_low + margin_up +
##     length
##
##                Df Deviance      AIC
## - height_left   1   4.8087 -111.674
## <none>              4.7614 -111.354
## - height_right  1   4.8410 -110.537
## + diagonal      1   4.7576 -109.489
## - length        1   6.3666  -63.965
## - margin_up     1   7.3457  -39.647
## - margin_low    1  10.1586   15.468
##
## Step:  AIC=-111.67
## .outcome ~ height_right + margin_low + margin_up + length
##
##                Df Deviance      AIC
## - height_right  1   4.8433 -112.454
## <none>              4.8087 -111.674
## + height_left   1   4.7614 -111.354
## + diagonal      1   4.7935 -110.211
## - length        1   6.3681  -65.925
## - margin_up     1   7.3930  -40.555
## - margin_low    1  10.2248   14.573
##
## Step:  AIC=-112.45
## .outcome ~ margin_low + margin_up + length
##
##                Df Deviance      AIC
## <none>              4.8433 -112.454
## + height_right  1   4.8087 -111.674
## + height_left   1   4.8410 -110.537
## + diagonal      1   4.8417 -110.510
## - length        1   6.3826  -67.538
## - margin_up     1   7.9008  -31.263
## - margin_low    1  11.8369   37.461

## Generalized Linear Model with Stepwise Feature Selection
##
## 170 samples
##   6 predictor
##
## No pre-processing
## Resampling: Cross-Validated (5 fold)
## Summary of sample sizes: 136, 136, 136, 136, 136
## Resampling results:
##
##   RMSE       Rsquared   MAE
##   0.1755559  0.8788229  0.1379961
```

```
##        RMSE  Rsquared       MAE Resample
## 1 0.1679119 0.8846697 0.1457480    Fold1
## 2 0.1477160 0.9231594 0.1205645    Fold2
## 3 0.1960706 0.8529226 0.1475754    Fold3
## 4 0.1902345 0.8547055 0.1422176    Fold4
## 5 0.1758463 0.8786571 0.1338749    Fold5
```

How does the algorithm work ? For each iteration, an ascending and descending stepwise method is applied. We start from a model that expresses the class of the note according to all the geometric variables. Then we add or remove variables until we get the linear combination of variables that minimizes the AIC (Akaike Information Criterion). This is a measure of the quality of the statistical model that depends on:

- the complexity of the model: k: the number of parameters used in the model
- the quality of the fit: L: the maximum of the model's likelihood function

With logistic regression, the mean RSME for cross-validation is 0.18. Since the mean RSME for cross-validation of logistic regression is lower than that of the decision tree, the logistic regression model is considered to be more efficient. It will be used on the entire dataset for the prediction.

3.4 Prediction

Here are the bills that I am going to classisfy.

```
##   diagonal height_left height_right margin_low margin_up length  id
## 1   171.76      104.01       103.54       5.21      3.30 111.42 A_1
## 2   171.87      104.17       104.13       6.00      3.31 112.09 A_2
## 3   172.00      104.58       104.29       4.99      3.39 111.57 A_3
## 4   172.49      104.55       104.34       4.44      3.03 113.20 A_4
## 5   171.65      103.63       103.56       3.77      3.16 113.33 A_5
```

For the prediction, now that the optimal model has been determined, I carry out a logistic regression on all the data, with the logit parameter so that the probabilities are between 0 and 1. Our program will be able to make a prediction on a bill, that is to say to determine if it is a real or a fake bill. For each bill, the classification algorithm will give the probability that the bill is true. If this probability is greater than or equal to 0.5, the bill will be considered true. Otherwise, it will be considered as a fake.

```
## Warning: glm.fit: algorithm did not converge
```

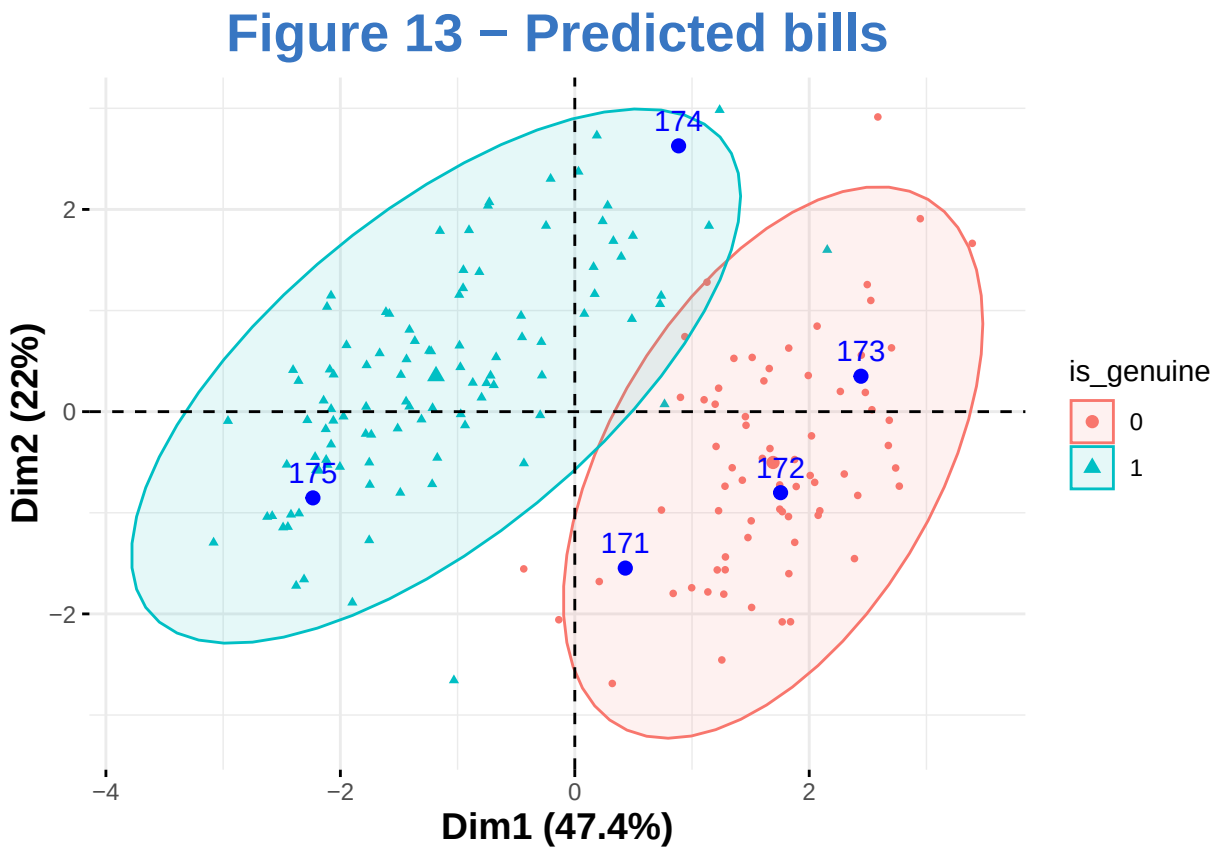
```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
##            1            2            3            4            5
## 2.220446e-16 2.220446e-16 2.220446e-16 1.000000e+00 1.000000e+00
```

```
## [1] "The results above are the probability for each bill that it is true."
```

```
##   diagonal height_left height_right margin_low margin_up length  id is_genuine
## 1   171.76      104.01       103.54       5.21      3.30 111.42 A_1      False
## 2   171.87      104.17       104.13       6.00      3.31 112.09 A_2      False
## 3   172.00      104.58       104.29       4.99      3.39 111.57 A_3      False
## 4   172.49      104.55       104.34       4.44      3.03 113.20 A_4       True
## 5   171.65      103.63       103.56       3.77      3.16 113.33 A_5       True
```

Display on the correlation circle



Part 4 - Conclusion

From a set of bills (characterized by geometrical lengths and a status - real or fake), my aim was to make an algorithm in order to predict if a bill is real or not.

- First, I had to clean all the dataset (missing data, outliers, unusual data...) so that I can see which one of the geometrical characteristics have the most influence on the bill status. It would seem, then, that what distinguishes the true from the counterfeit bills are :
  - The top margin of the bills
  - The lower margin of the bills
  - The length of the bills
- Then by using a principal component analysis, I was able to confirm my initial hypothesis. The variables that contribute the most to the categorization of the bills are the length of the bill, its lower margin and its top margin.
- Finally, I compared 2 models in order to predict a bill status : decision tree and logistic regression. Both models was using at least length and margin low (and margin up for logistic regression). With cross-validation, I showed that the logistic regression performs better (lower RMSE). So I used logistic regression in order to predict the status of new bills.

Once the automatic learning algorithm has been trained on a first set of data, one should evaluate it on a second set of data in order to verify that the model does not over-learn. Finally, the model need to be deployed in production to make predictions, and use the new input data to re-train and improve its model.