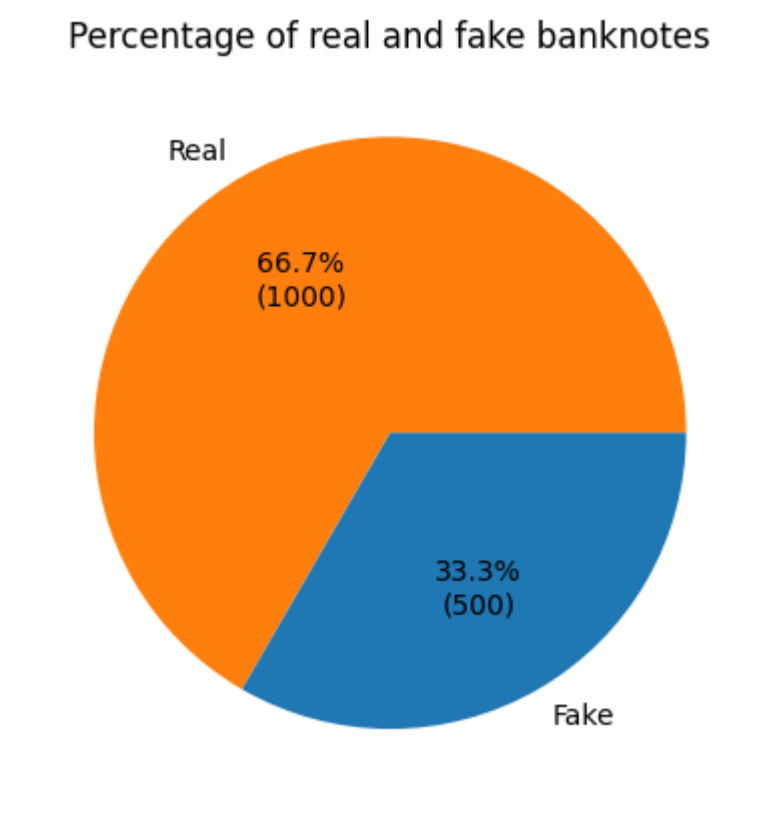
AUTOMATIC DETECTION OF FAKE BANKNOTES

ONCFM

Context

- The National Organization for the Fight against Counterfeiting – A government organisation
- The organisation has noticed differences in the dimensions of real and fake banknotes.
- Project objective - create an algorithm capable of automatically differentiating real and fake banknotes using the dimensions of the banknotes.

ONCFM



The Data

1500 banknotes

- 1000 real, 500 fake

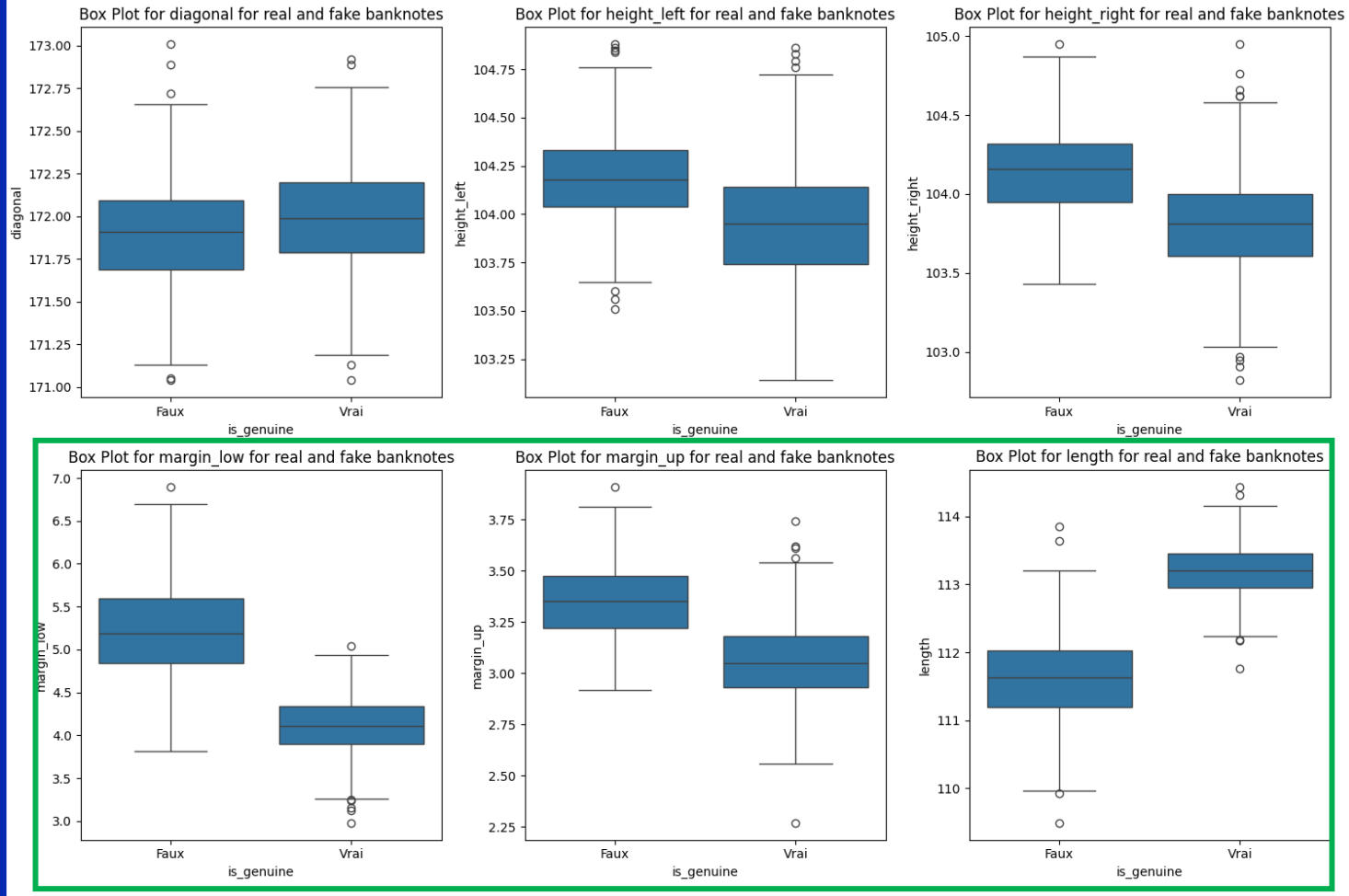
7 columns

- is_genuine: if the banknote is real or fake
- length
- height_left
- height_right
- margin_up: the margin between the top edge of the banknote and the image on the banknote
- margin_low: the margin between the bottom edge of the banknote and the image on the banknote
- diagonal: measurement of banknote when measured diagonally

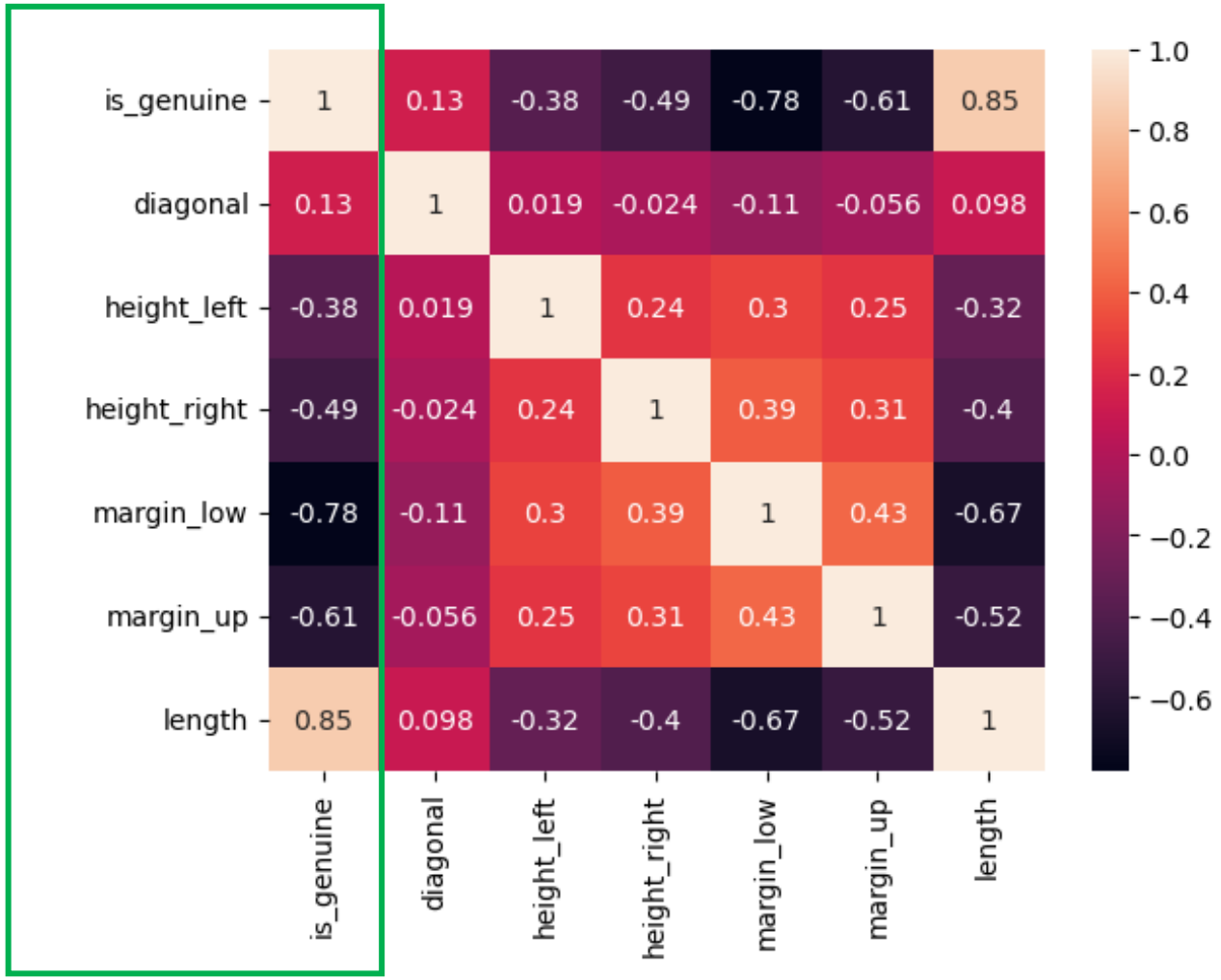
ONCFM

Descriptive Analysis

- There are quite big differences between real and fake banknotes in terms of the “length”, “margin_up” and “margin_low” dimensions.
- However, the “diagonal”, “height_left”, and “height_right” dimensions are quite similar.



ONCFM



ONCFM

Descriptive Analysis

Most relevant variables for “is_genuine”:

- length
- margin_low
- margin_up

Least important variable for “is_genuine”:

- diagonal

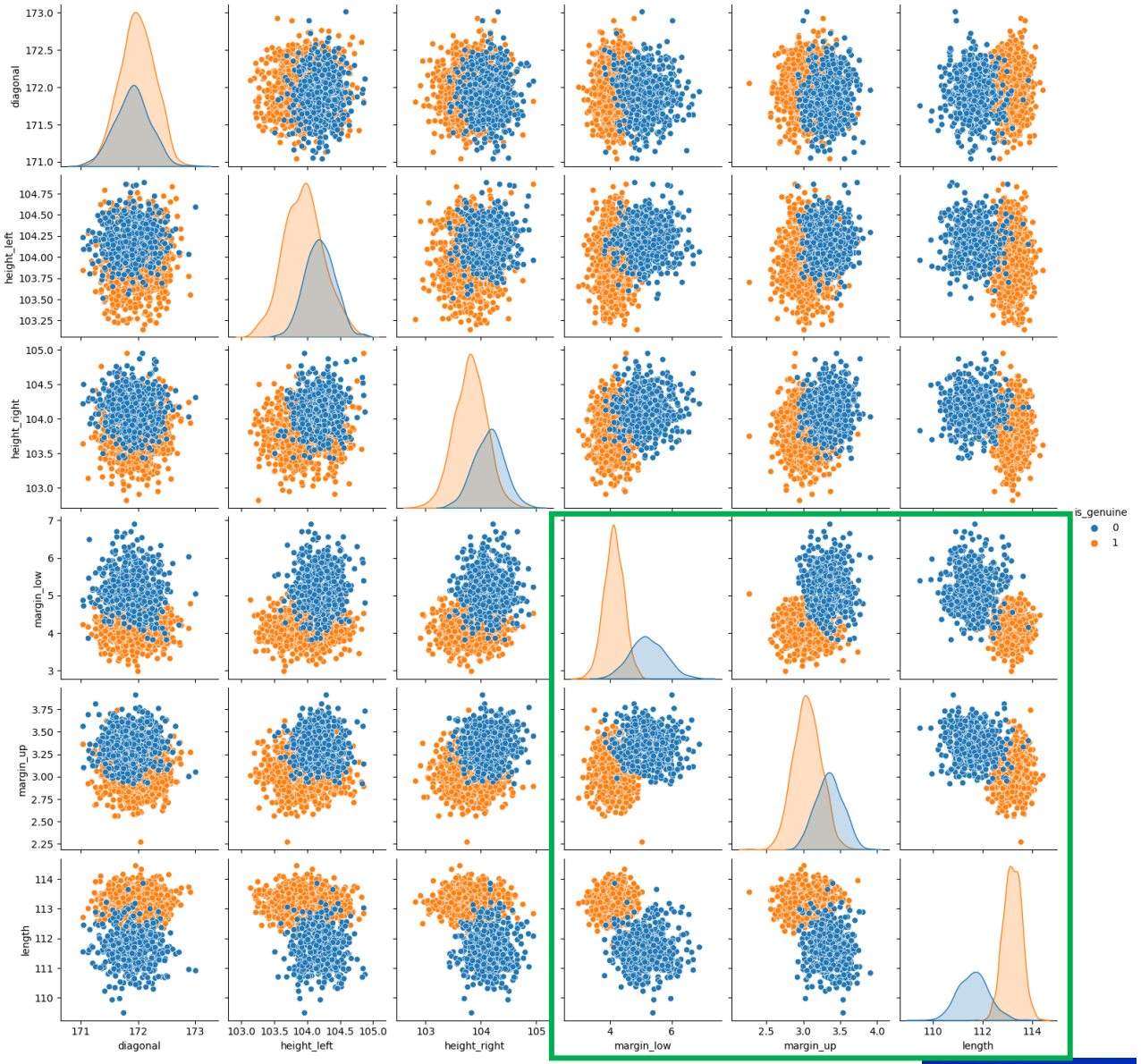
Descriptive Analysis

Kernel density plots

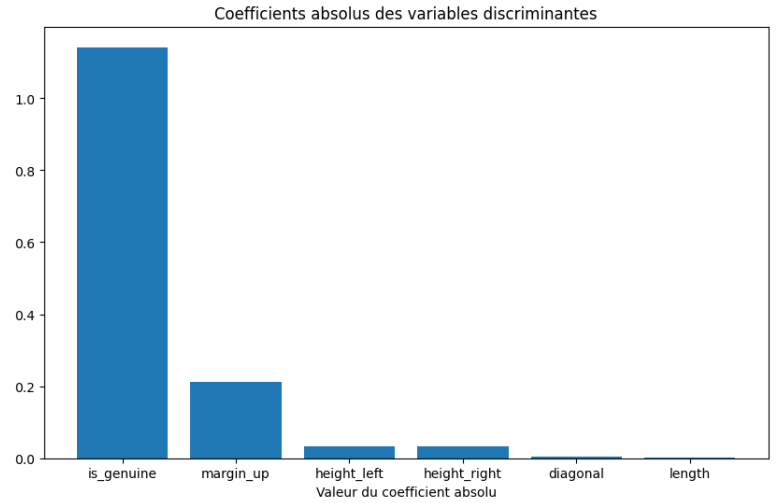
- The real and the fake banknotes are clearly differentiated by the “length” and “margin_low” variables.
- This is not true for "diagonal", "height_left", and "height_right".

Scatterplots

- Scatterplots that contain “length,” “margin_low,” and “margin_up” have fairly well-separated clusters.
- Useful for determining which banknotes are real and which are fake.



ONCFM



OLS Regression Results

Dep. Variable:	margin_low	R-squared:	0.617
Model:	OLS	Adj. R-squared:	0.616
Method:	Least Squares	F-statistic:	1174.
Date:	Thu, 08 Feb 2024	Prob (F-statistic):	1.24e-304
Time:	17:48:57	Log-Likelihood:	-774.73
No. Observations:	1463	AIC:	1555.
Df Residuals:	1460	BIC:	1571.
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	5.9263	0.198	30.003	0.000	5.539	6.314
margin_up	-0.2119	0.059	-3.612	0.000	-0.327	-0.097
is_genuine	-1.1632	0.029	-40.477	0.000	-1.220	-1.107

Omnibus:	22.365	Durbin-Watson:	2.041
Prob(Omnibus):	0.000	Jarque-Bera (JB):	39.106
Skew:	0.057	Prob(JB):	3.22e-09
Kurtosis:	3.793	Cond. No.	65.0

Linear Regression

Bar plot

- “is_genuine” has a lot of influence on the “margin_low” variable.
- “margin_up” also has a slight influence.

Model

- The best model uses only “margin_up and “is_genuine”
- R2 = 0.617
- Statistically significant (1.24e-403<0.05)
- Coefficients – “margin_up” = -0.2119, is_genuine = -1.1632

ONCFM

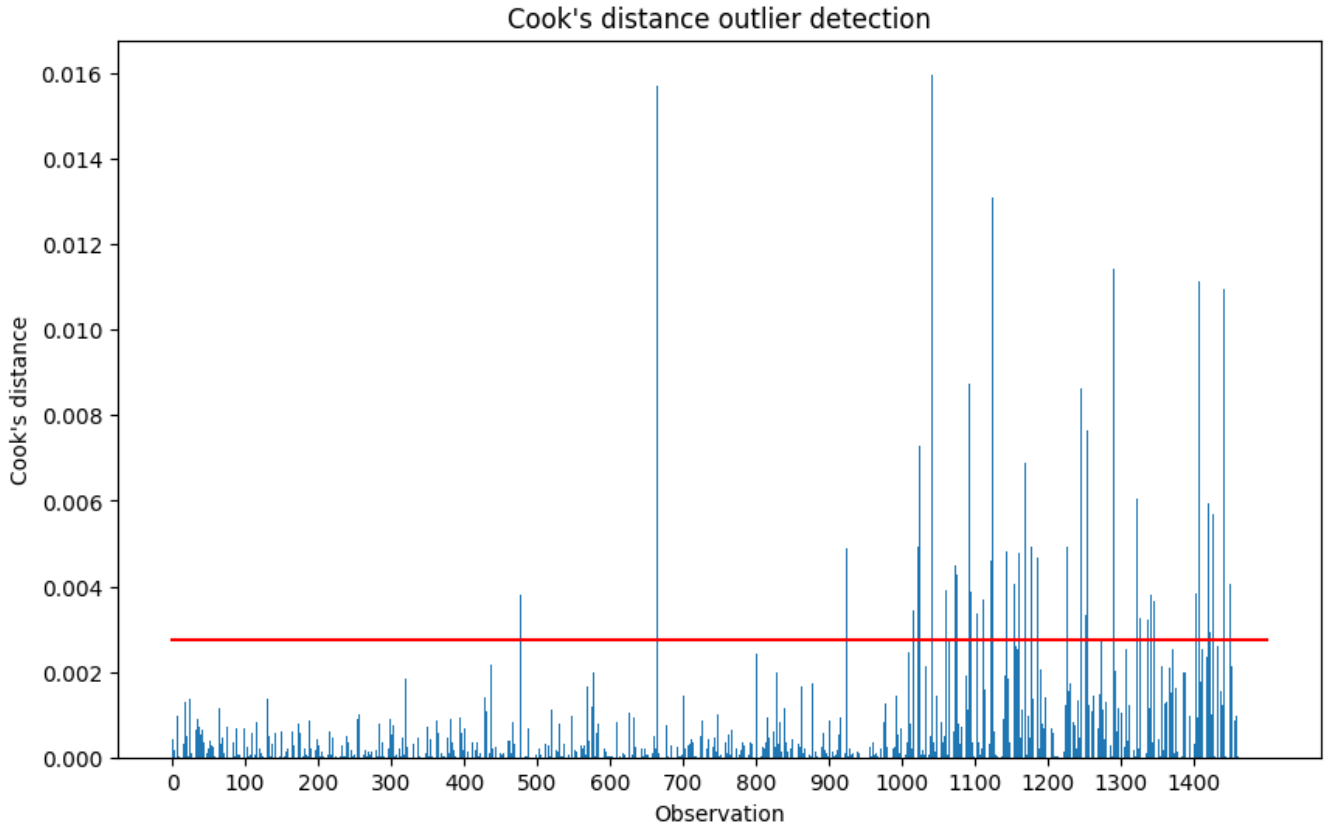
Linear Regression

Cook's Distance Outliers

- 79 fake and 14 real banknotes classified as outliers

Process

- Outliers for fake banknotes are not very surprising
- But the relative differences between the outliers of real banknotes and other real banknotes are more surprising.
- After looking at the differences, I removed the 14 real banknotes that were outliers.



93 outliers

ONCFM

Linear Regression

Before

OLS Regression Results

Dep. Variable:	margin_low	R-squared:	0.617
Model:	OLS	Adj. R-squared:	0.616
Method:	Least Squares	F-statistic:	1174.
Date:	Thu, 08 Feb 2024	Prob (F-statistic):	1.24e-304
Time:	17:48:57	Log-Likelihood:	-774.73
No. Observations:	1463	AIC:	1555.
Df Residuals:	1460	BIC:	1571.
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	5.9263	0.198	30.003	0.000	5.539	6.314
margin_up	-0.2119	0.059	-3.612	0.000	-0.327	-0.097
is_genuine	-1.1632	0.029	-40.477	0.000	-1.220	-1.107

Omnibus:	22.365	Durbin-Watson:	2.041
Prob(Omnibus):	0.000	Jarque-Bera (JB):	39.106
Skew:	0.057	Prob(JB):	3.22e-09
Kurtosis:	3.793	Cond. No.	65.0

After

OLS Regression Results

Dep. Variable:	margin_low	R-squared:	0.627
Model:	OLS	Adj. R-squared:	0.627
Method:	Least Squares	F-statistic:	1249.
Date:	Sat, 10 Feb 2024	Prob (F-statistic):	1.00e-318
Time:	16:42:32	Log-Likelihood:	-749.04
No. Observations:	1486	AIC:	1504.
Df Residuals:	1483	BIC:	1520.
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	6.0124	0.194	30.969	0.000	5.632	6.393
margin_up	-0.2377	0.058	-4.119	0.000	-0.351	-0.124
is_genuine	-1.1671	0.028	-41.835	0.000	-1.222	-1.112

Omnibus:	30.919	Durbin-Watson:	2.046
Prob(Omnibus):	0.000	Jarque-Bera (JB):	59.806
Skew:	0.087	Prob(JB):	1.03e-13
Kurtosis:	3.967	Cond. No.	66.0

ONCFM

Linear Regression

Distribution of residuals

- The statistic of the model is good.
- But the p value < 0.05 calls into question the normality of the residuals.
- The residuals are not very different from a symmetric distribution and the sample has more than 30 individuals
- So, the results obtained by the model are not absurd

Conclusion

- I will use this model to impute the missing values (project requirements explicitly required me to).

No problem with colinearity = Valid

VIF for the coefficients = [1.6202, 1.6202]

(Less than 10)

Homoscedasticity = Not valid

Breusch Pagan p-value: 1.9624e-39

Distribution of residuals = Not valid

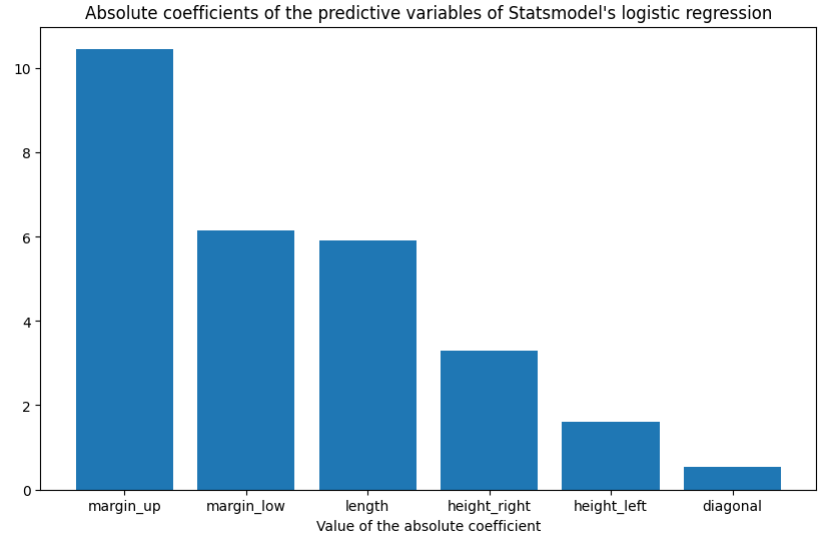
Statistic = 0.9928

p-value = 1.7644e-06

ONCFM

CREATION OF THE ALGORITHMS

ONCFM




```
is_genuine ~ margin_up + height_right + length + margin_low + 1
Optimization terminated successfully.
         Current function value: 0.026782
         Iterations 13
is the final model !

                           Logit Regression Results
==============================================================================
Dep. Variable:              is_genuine   No. Observations:                 1500
Model:                          Logit   Df Residuals:                     1495
Method:                           MLE   Df Model:                            4
Date:                Thu, 08 Feb 2024   Pseudo R-squ.:                 0.9579
Time:                        17:49:00   Log-Likelihood:                -40.173
converged:                       True   LL-Null:                       -954.77
Covariance Type:            nonrobust   LLR p-value:                    0.000
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept   -282.4740    139.419     -2.026      0.043    -555.731      -9.217
margin_up    -10.4098      2.197     -4.738      0.000     -14.716      -6.103
height_right  -3.3512      1.123     -2.984      0.003      -5.553      -1.150
length         6.1592      0.889      6.931      0.000       4.418       7.901
margin_low    -6.3058      0.963     -6.550      0.000      -8.193      -4.419
==============================================================================
```

ONCFM

Statsmodels: Logistic Regression

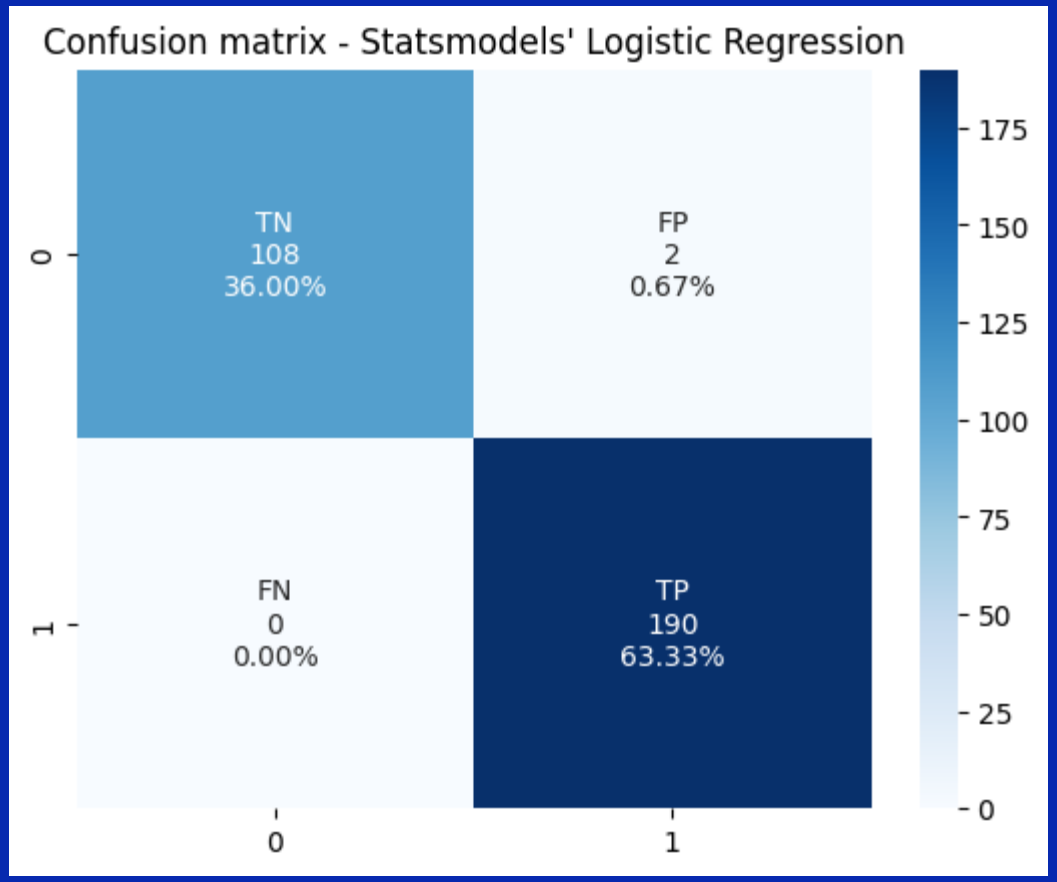
Bar plot

- "margin_up", "margin_low", and "length" have a relatively strong influence on "is_genuine".

Model

- The best model uses "margin_up", "margin_low", "length" and "height_right".
- Pseudo R2 = 0.9579
- Statistically significant (0.00<0.05)
- Coefficients – "margin_up" = -10.41, "height_right" = -3.35, "length" = 6.16, "margin_low" = -6.31

Statsmodels – Logistic Regression



Accuracy: 99.33%

Precision: 98.96%

Recall: 100%

ROC-AUC: 0.9909

Statsmodels: Logistic Regression

Increase the threshold = Reduce the number of false positives

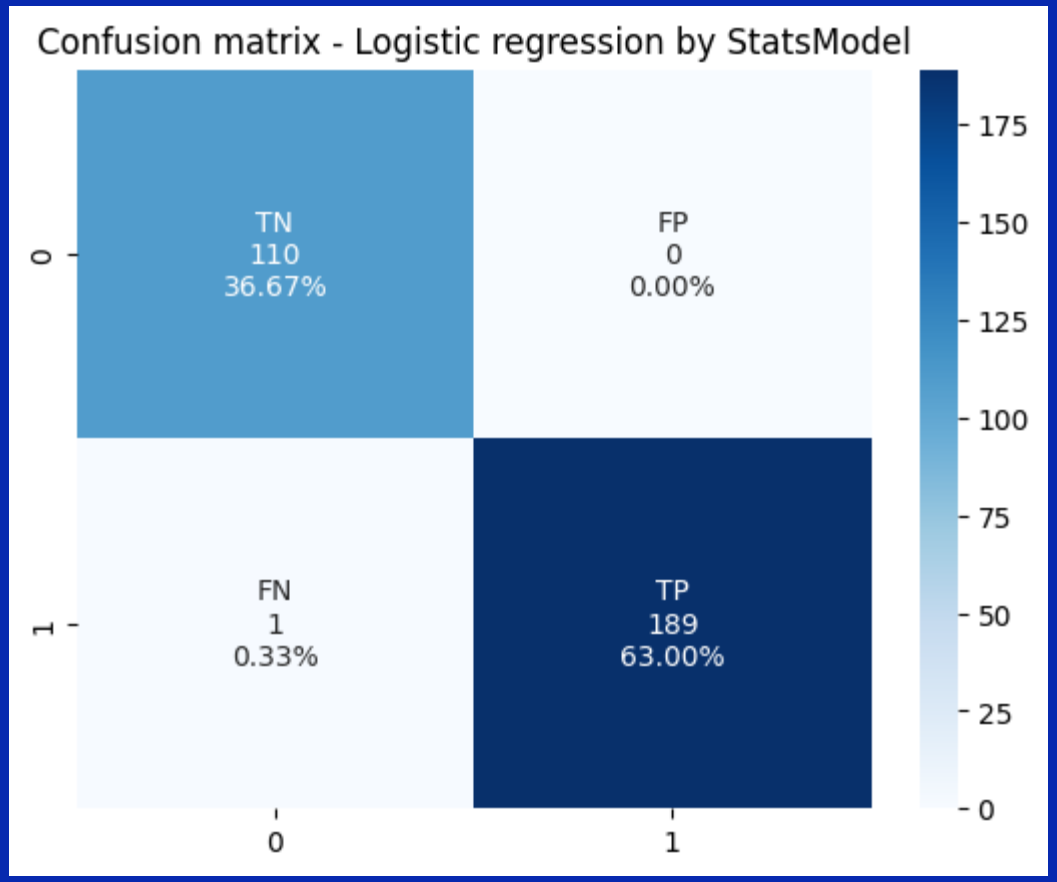
Reduce the threshold = Reduce the number of false negatives

Threshold of 0.7

- No false positives
- BUT 1 false negative

```
Seuil 0.3 - Nombre de True Negative = 106
Seuil 0.3 - Nombre de True Positive = 190
Seuil 0.3 - Nombre de False Negative = 0
Seuil 0.3 - Nombre de False Positive = 4
-------------------------------------------------
Seuil 0.4 - Nombre de True Negative = 106
Seuil 0.4 - Nombre de True Positive = 190
Seuil 0.4 - Nombre de False Negative = 0
Seuil 0.4 - Nombre de False Positive = 4
-------------------------------------------------
Seuil 0.5 - Nombre de True Negative = 108
Seuil 0.5 - Nombre de True Positive = 190
Seuil 0.5 - Nombre de False Negative = 0
Seuil 0.5 - Nombre de False Positive = 2
-------------------------------------------------
Seuil 0.6 - Nombre de True Negative = 108
Seuil 0.6 - Nombre de True Positive = 189
Seuil 0.6 - Nombre de False Negative = 1
Seuil 0.6 - Nombre de False Positive = 2
-------------------------------------------------
Seuil 0.7 - Nombre de True Negative = 110
Seuil 0.7 - Nombre de True Positive = 189
Seuil 0.7 - Nombre de False Negative = 1
Seuil 0.7 - Nombre de False Positive = 0
-------------------------------------------------
Seuil 0.8 - Nombre de True Negative = 110
Seuil 0.8 - Nombre de True Positive = 187
Seuil 0.8 - Nombre de False Negative = 3
Seuil 0.8 - Nombre de False Positive = 0
-------------------------------------------------
```

Statsmodels – Logistic Regression

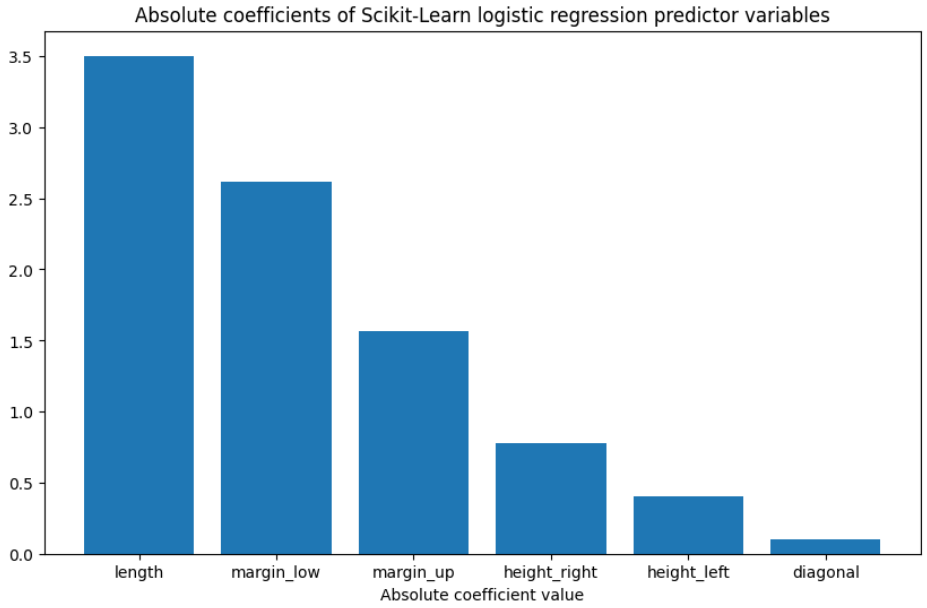


Accuracy: 99.67%

Precision: 100%

Recall: 99.47%

ROC-AUC: 0.9974



	Accuracy	Precision	Recall	ROC-AUC score
Length, margin_low, margin_up, height_right, height_left, diagonal	0.9933	0.9896	1	0.9909
Length, margin_low, margin_up, height_right, height_left	0.9933	0.9896	1	0.9909
Length, margin_low, margin_up, height_right	0.99	0.9845	1	0.9864
Length, margin_low, margin_up	0.9933	0.9896	1	0.9909
Length, margin_low	0.9867	0.9794	1	0.9818
Length	0.9533	0.94	1	0.9402

SciKit-Learn: Logistic Regression

Bar plot

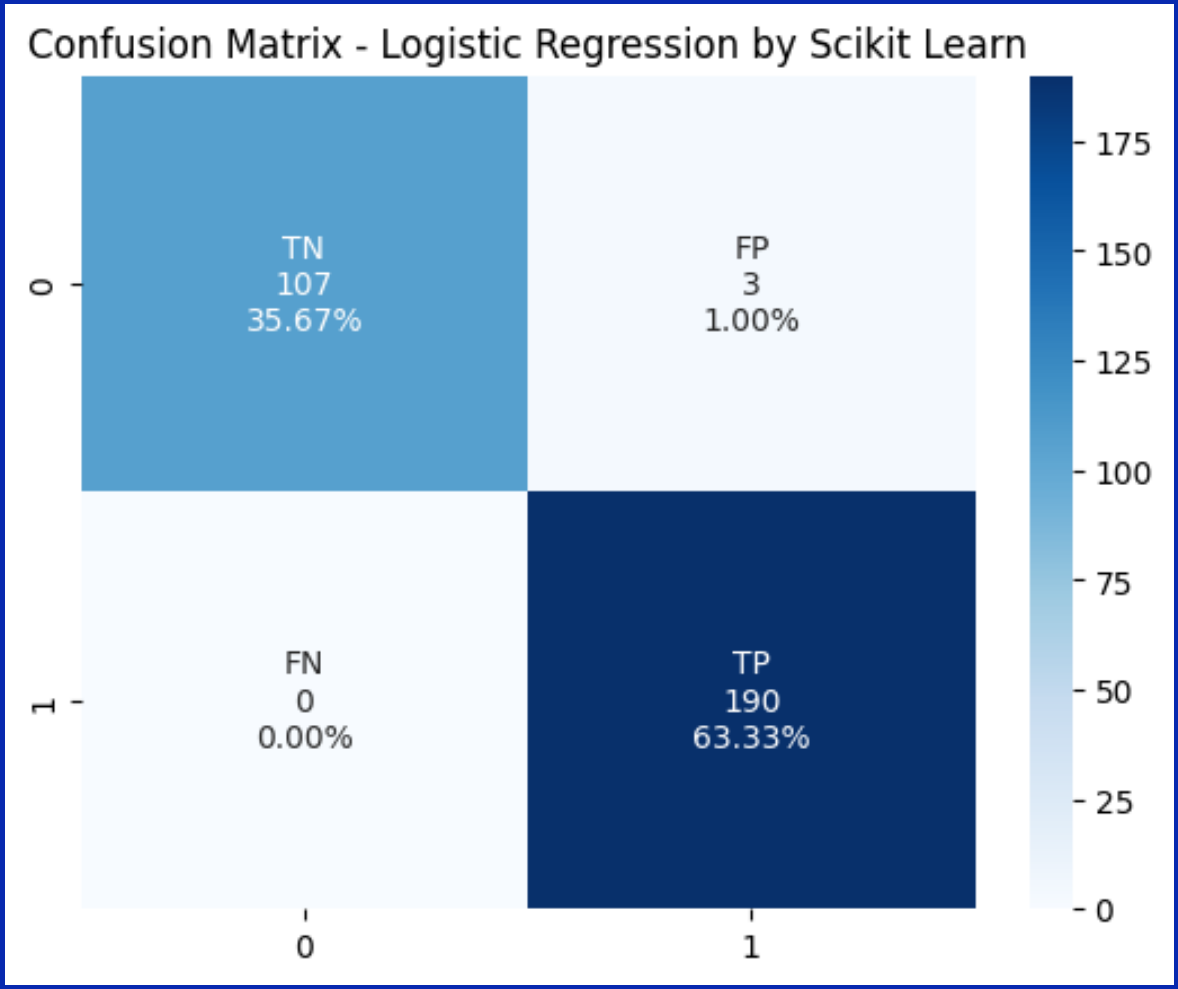
- "length", "margin_low" and "margin_up" have the strongest influence on "is_genuine".

Model

- The best model uses "length", "margin_low", and "margin_up", and "height_right".
- Using fewer variables simplifies the model and reduces the risk of overfitting.

ONCFM

SciKit-Learn – Logistic Regression



Accuracy: 99%

Precision: 98,45%

Recall: 100%

ROC-AUC: 0,9864

Scikit-Learn: Logistic Regression

Increase the threshold = Reduce the number of false positives

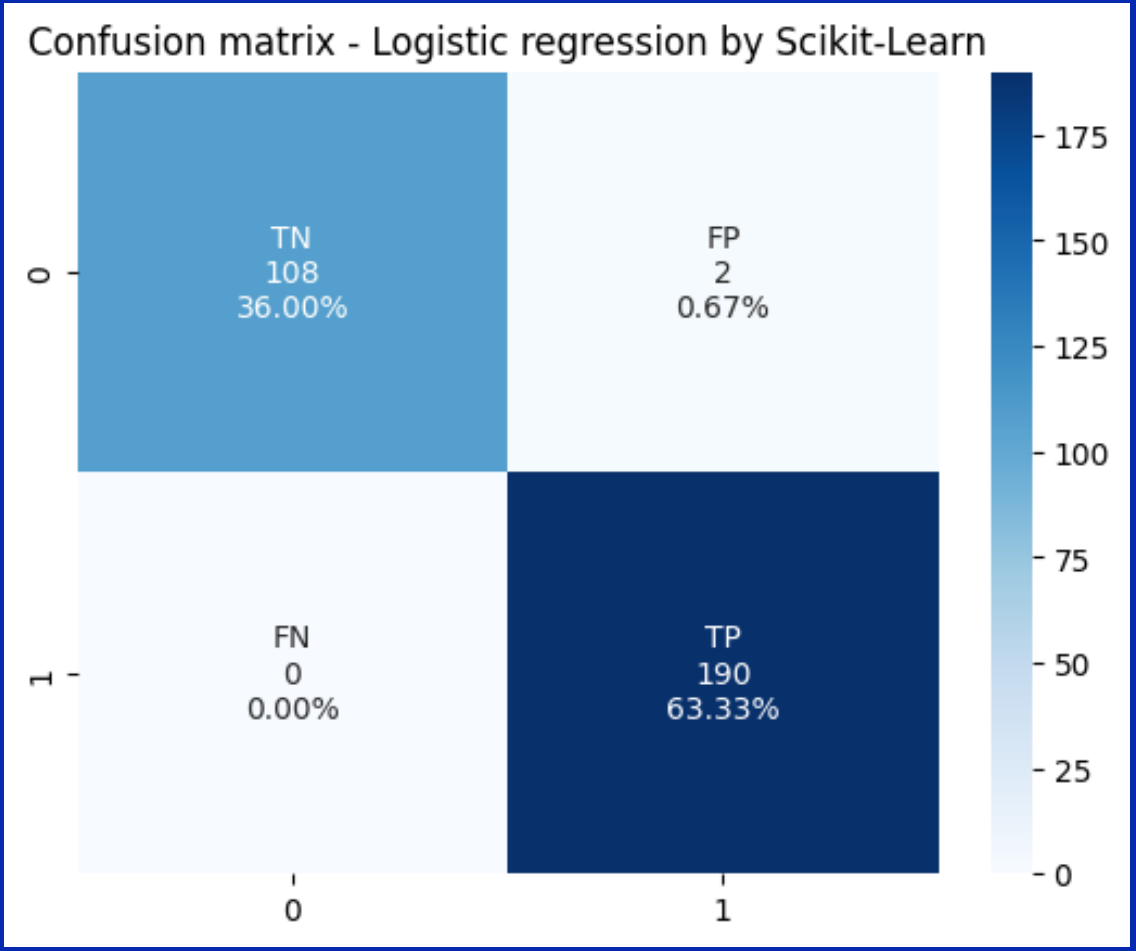
Reduce the threshold = Reduce the number of false négatives

Threshold of 0.6 or 0.7

- No false negatives
- 2 false positives

```
Seuil 0.3 - Nombre de True Negative = 102
Seuil 0.3 - Nombre de True Positive = 190
Seuil 0.3 - Nombre de False Negative = 0
Seuil 0.3 - Nombre de False Positive = 8
-------------------------------------------------
Seuil 0.4 - Nombre de True Negative = 105
Seuil 0.4 - Nombre de True Positive = 190
Seuil 0.4 - Nombre de False Negative = 0
Seuil 0.4 - Nombre de False Positive = 5
-------------------------------------------------
Seuil 0.5 - Nombre de True Negative = 107
Seuil 0.5 - Nombre de True Positive = 190
Seuil 0.5 - Nombre de False Negative = 0
Seuil 0.5 - Nombre de False Positive = 3
-------------------------------------------------
Seuil 0.6 - Nombre de True Negative = 108
Seuil 0.6 - Nombre de True Positive = 190
Seuil 0.6 - Nombre de False Negative = 0
Seuil 0.6 - Nombre de False Positive = 2
-------------------------------------------------
Seuil 0.7 - Nombre de True Negative = 108
Seuil 0.7 - Nombre de True Positive = 190
Seuil 0.7 - Nombre de False Negative = 0
Seuil 0.7 - Nombre de False Positive = 2
-------------------------------------------------
Seuil 0.8 - Nombre de True Negative = 109
Seuil 0.8 - Nombre de True Positive = 187
Seuil 0.8 - Nombre de False Negative = 3
Seuil 0.8 - Nombre de False Positive = 1
-------------------------------------------------
```

SciKit-Learn – Logistic Regression

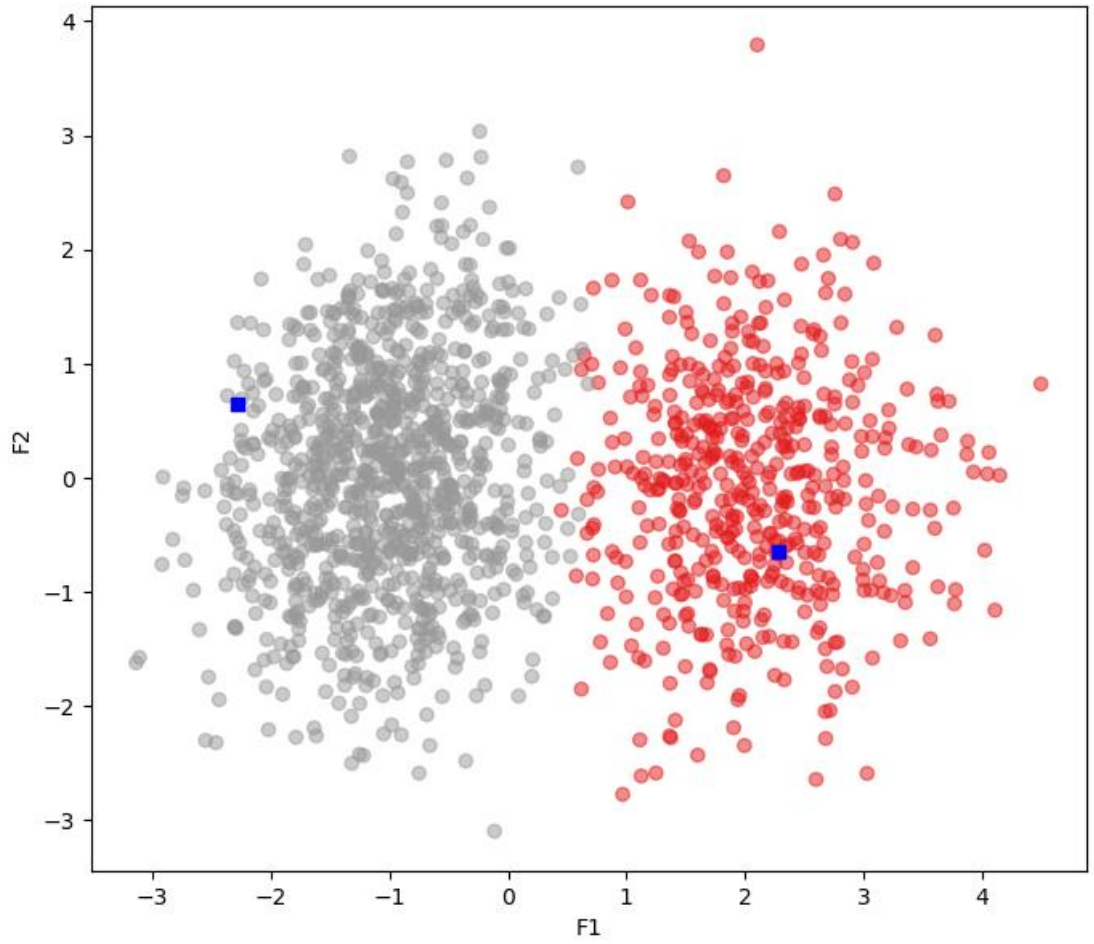


Accuracy: 99,33%

Precision: 98,96%

Recall: 100%

ROC-AUC: 0,9909



K-means

Number of clusters

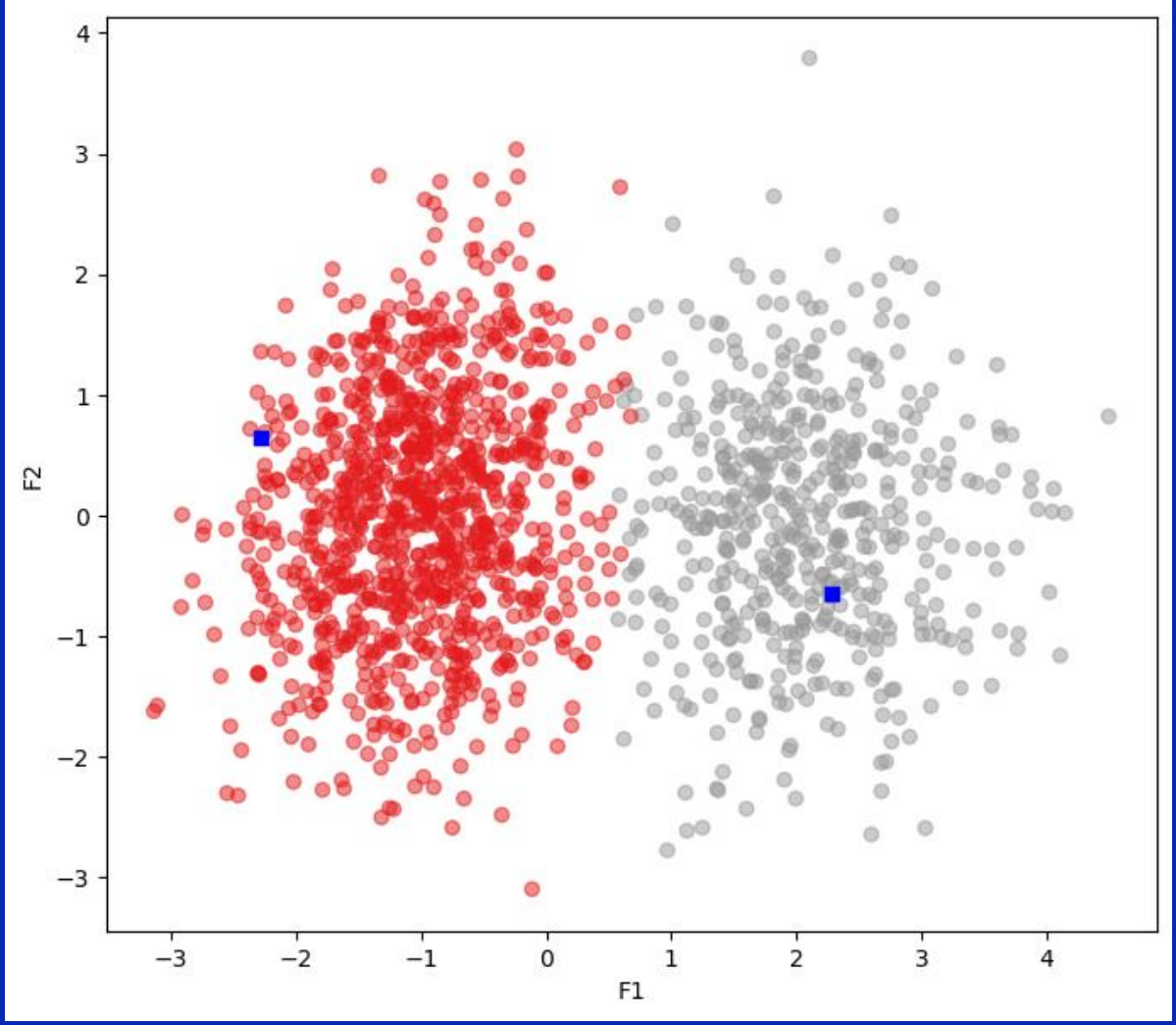
- No need for the elbow method – 2 clusters – the real and the fake banknotes

K-means plot

- 2 well-defined clusters

ONCFM

K-means



Accuracy: 98,67%

Precision: 99,49%

Recall: 98,48%

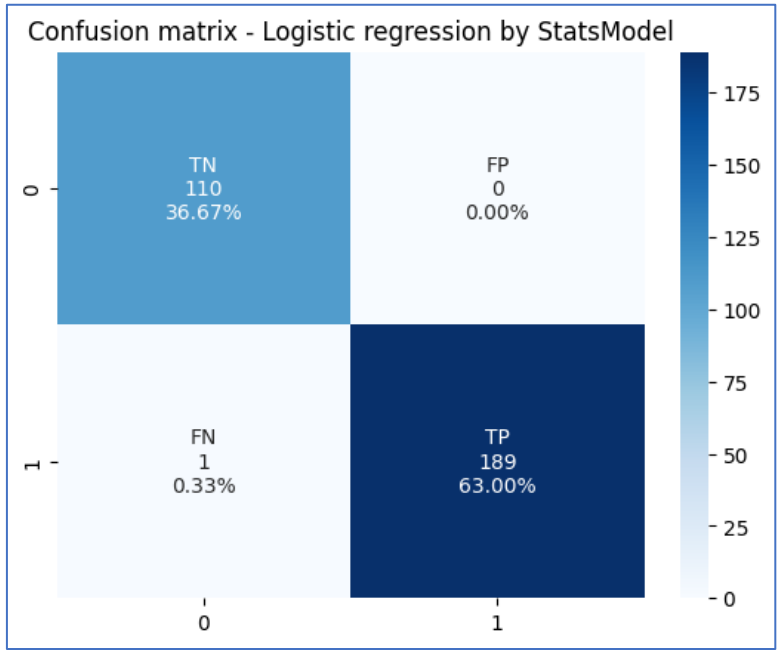
ROC-AUC: 0,9875

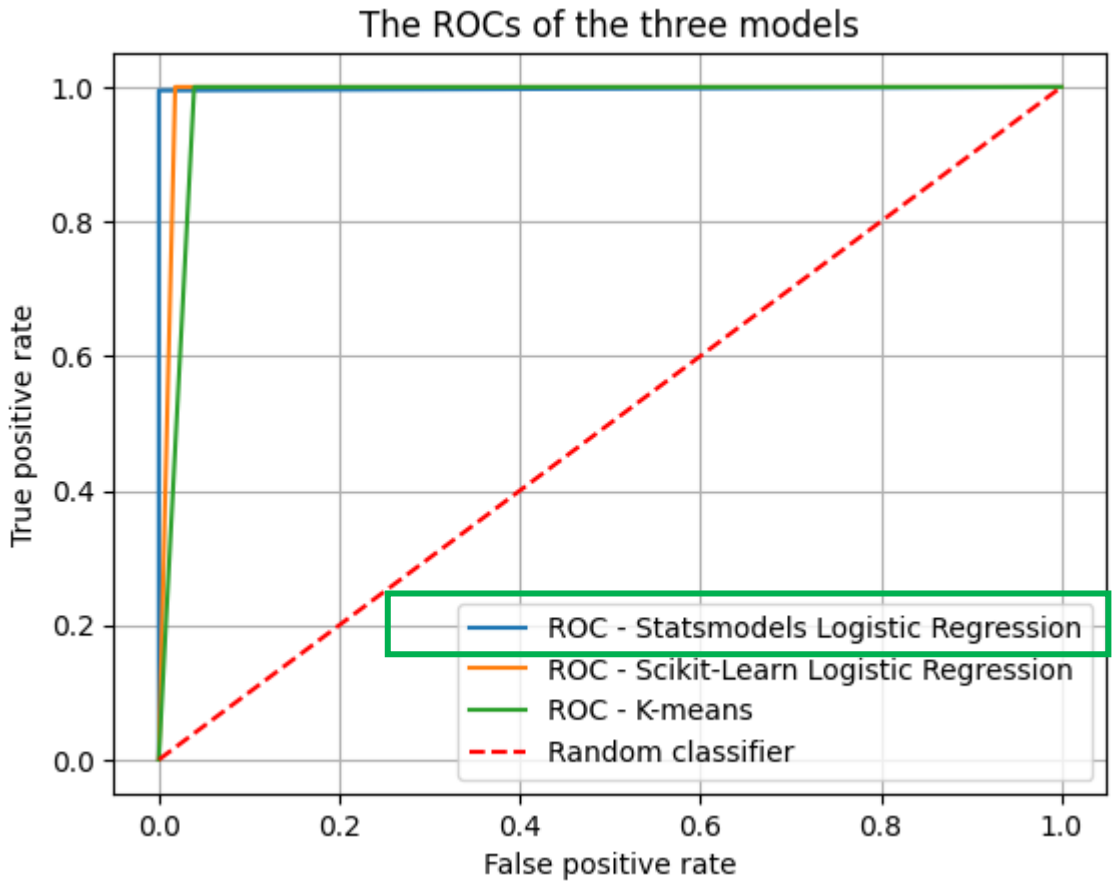
Conclusion

Best algorithm:

Statsmodels logistic regression

- Best accuracy score
- **Best precision score**
- Best ROC-AUC score





	Accuracy	Precision	Recall	ROC-AUC score
Statsmodels Logistic Regression	0.9967	1	0.9947	0.9974
Scikit-Learn Logistic Regression	0.9933	0.9896	1	0.9909
K-means	0.9867	0.9802	1	0.9804