BANK FRAUD DETECTION WITH IBM SPSS MODELER

BY

AYÇA NUR VANLI

TABLE OF CONTENTS

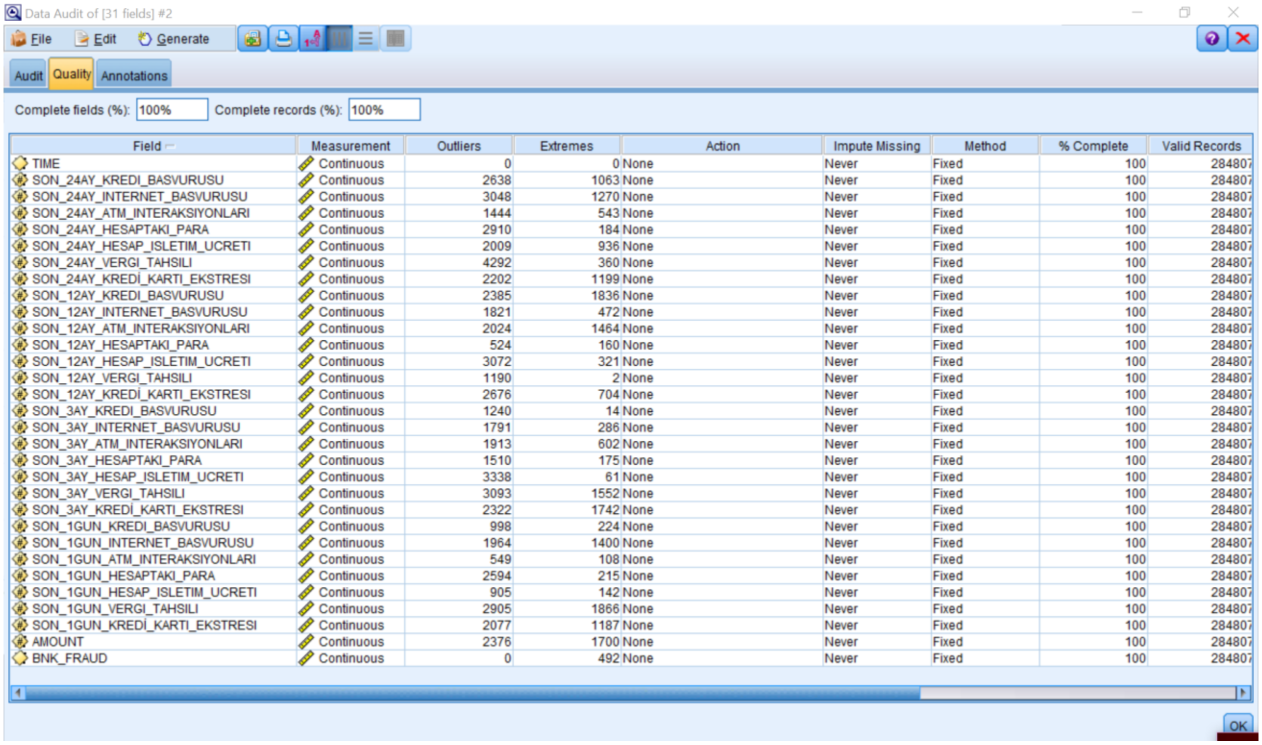
1. Data Preparation

Data sets can have missing, irrelevant or noisy data. We must understand these obstacles that in our way to have a successful model. Without prepared data set, the model’s success rate is very low and inaccurate. In order to make the data set prepped and ready to use for the model, we must clean and transform the data sets.

According to what we need and different perspectives, the steps that we are taking in this phase of the data mining can differentiate. If some steps are not fitting the data set’s nature we can skip or have a different approach to it.

1.1. Data Cleaning

We insert our credit card data that in excel (creditcard.csv) form. After that we look inside of our data by Data Audit. We can see our parameters.

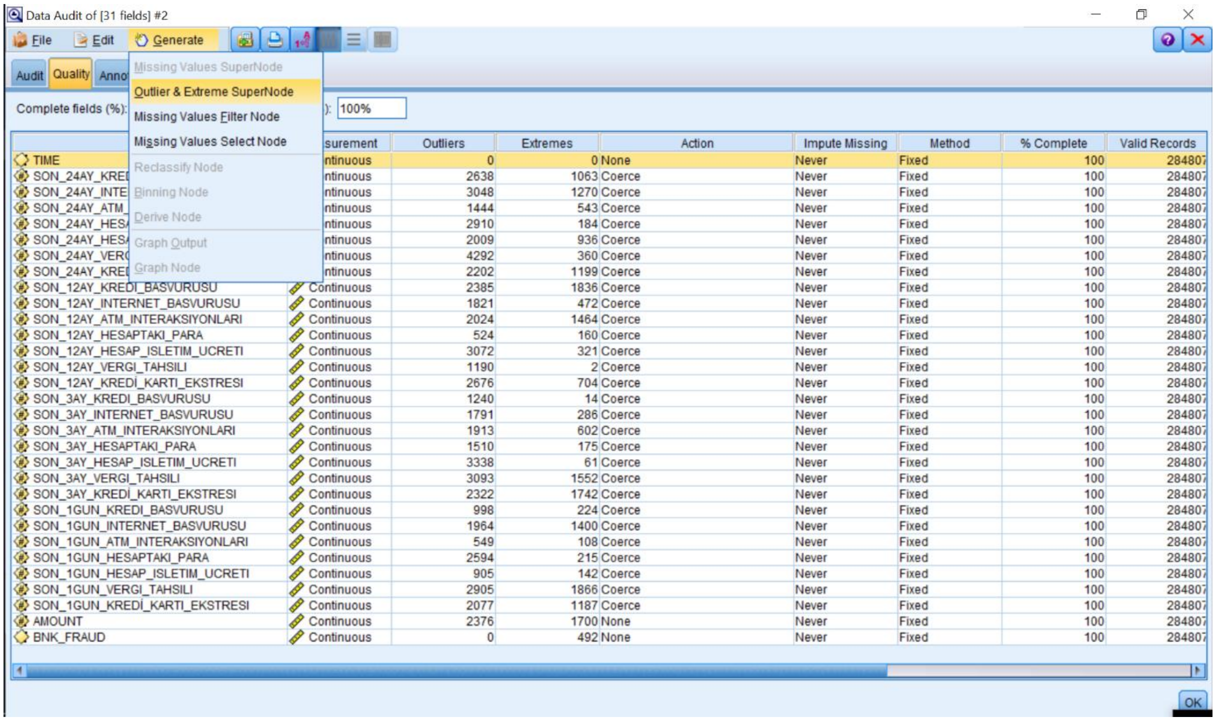


Complete fields (%): 100% Complete records (%): 100%

Field	Measurement	Outliers	Extremes	Action	Impute Missing	Method	% Complete	Valid Records
TIME	Continuous	0	0	None	Never	Fixed	100	284807
SON_24AY_KREDI_BASVURUSU	Continuous	2638	1063	None	Never	Fixed	100	284807
SON_24AY_INTERNET_BASVURUSU	Continuous	3048	1270	None	Never	Fixed	100	284807
SON_24AY_ATM_INTERAKSIYONLARI	Continuous	1444	543	None	Never	Fixed	100	284807
SON_24AY_HESAPTAKI_PARA	Continuous	2910	184	None	Never	Fixed	100	284807
SON_24AY_HESAP_ISLETIM_UCRETI	Continuous	2009	936	None	Never	Fixed	100	284807
SON_24AY_VERGI_TAHSILI	Continuous	4292	360	None	Never	Fixed	100	284807
SON_24AY_KREDİ_KARTI_EKSTRESI	Continuous	2202	1199	None	Never	Fixed	100	284807
SON_12AY_KREDI_BASVURUSU	Continuous	2385	1836	None	Never	Fixed	100	284807
SON_12AY_INTERNET_BASVURUSU	Continuous	1821	472	None	Never	Fixed	100	284807
SON_12AY_ATM_INTERAKSIYONLARI	Continuous	2024	1464	None	Never	Fixed	100	284807
SON_12AY_HESAPTAKI_PARA	Continuous	524	160	None	Never	Fixed	100	284807
SON_12AY_HESAP_ISLETIM_UCRETI	Continuous	3072	321	None	Never	Fixed	100	284807
SON_12AY_VERGI_TAHSILI	Continuous	1190	2	None	Never	Fixed	100	284807
SON_12AY_KREDİ_KARTI_EKSTRESI	Continuous	2676	704	None	Never	Fixed	100	284807
SON_3AY_KREDI_BASVURUSU	Continuous	1240	14	None	Never	Fixed	100	284807
SON_3AY_INTERNET_BASVURUSU	Continuous	1791	286	None	Never	Fixed	100	284807
SON_3AY_ATM_INTERAKSIYONLARI	Continuous	1913	602	None	Never	Fixed	100	284807
SON_3AY_HESAPTAKI_PARA	Continuous	1510	175	None	Never	Fixed	100	284807
SON_3AY_HESAP_ISLETIM_UCRETI	Continuous	3338	61	None	Never	Fixed	100	284807
SON_3AY_VERGI_TAHSILI	Continuous	3093	1552	None	Never	Fixed	100	284807
SON_3AY_KREDİ_KARTI_EKSTRESI	Continuous	2322	1742	None	Never	Fixed	100	284807
SON_1GUN_KREDI_BASVURUSU	Continuous	998	224	None	Never	Fixed	100	284807
SON_1GUN_INTERNET_BASVURUSU	Continuous	1964	1400	None	Never	Fixed	100	284807
SON_1GUN_ATM_INTERAKSIYONLARI	Continuous	549	108	None	Never	Fixed	100	284807
SON_1GUN_HESAPTAKI_PARA	Continuous	2594	215	None	Never	Fixed	100	284807
SON_1GUN_HESAP_ISLETIM_UCRETI	Continuous	905	142	None	Never	Fixed	100	284807
SON_1GUN_VERGI_TAHSILI	Continuous	2905	1866	None	Never	Fixed	100	284807
SON_1GUN_KREDİ_KARTI_EKSTRESI	Continuous	2077	1187	None	Never	Fixed	100	284807
AMOUNT	Continuous	2376	1700	None	Never	Fixed	100	284807
BNK_FRAUD	Continuous	0	492	None	Never	Fixed	100	284807

After looking at our data, we must look at the “Extremes” column. We can see that there are a lot of extremes in our data so we must clean that. In order to clean the extremes, we must do some changes in the “Action” column. We must change the “None” to “Coerce” for each parameter that does not have 0 in the Extremes column.

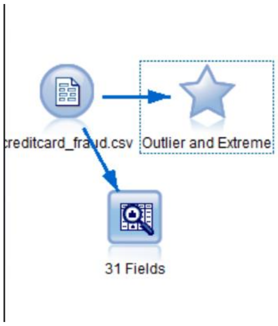
After changing the Action column, we must click the Generate button from the toolbar. Then clicking the Outlier & Extreme Supernode.



Complete fields (%):): 100%

	surement	Outliers	Extremes	Action	Impute Missing	Method	% Complete	Valid Records
TIME	ntinuous	0	0	None	Never	Fixed	100	284807
SON_24AY_KREI	ntinuous	2638	1063	Coerce	Never	Fixed	100	284807
SON_24AY_INTE	ntinuous	3048	1270	Coerce	Never	Fixed	100	284807
SON_24AY_ATM_	ntinuous	1444	543	Coerce	Never	Fixed	100	284807
SON_24AY_HES/	ntinuous	2910	184	Coerce	Never	Fixed	100	284807
SON_24AY_HES/	ntinuous	2009	936	Coerce	Never	Fixed	100	284807
SON_24AY_VER(	ntinuous	4292	360	Coerce	Never	Fixed	100	284807
SON_24AY_KREI	ntinuous	2202	1199	Coerce	Never	Fixed	100	284807
SON_12AY_KREDI_BASVURUSU	Continuous	2385	1836	Coerce	Never	Fixed	100	284807
SON_12AY_INTERNET_BASVURUSU	Continuous	1821	472	Coerce	Never	Fixed	100	284807
SON_12AY_ATM_INTERAKSIYONLARI	Continuous	2024	1464	Coerce	Never	Fixed	100	284807
SON_12AY_HESAPTAKI_PARA	Continuous	524	160	Coerce	Never	Fixed	100	284807
SON_12AY_HESAP_ISLETIM_UCRETI	Continuous	3072	321	Coerce	Never	Fixed	100	284807
SON_12AY_VERGI_TAHSILI	Continuous	1190	2	Coerce	Never	Fixed	100	284807
SON_12AY_KREDİ_KARTI_EKSTRESI	Continuous	2676	704	Coerce	Never	Fixed	100	284807
SON_3AY_KREDI_BASVURUSU	Continuous	1240	14	Coerce	Never	Fixed	100	284807
SON_3AY_INTERNET_BASVURUSU	Continuous	1791	286	Coerce	Never	Fixed	100	284807
SON_3AY_ATM_INTERAKSIYONLARI	Continuous	1913	602	Coerce	Never	Fixed	100	284807
SON_3AY_HESAPTAKI_PARA	Continuous	1510	175	Coerce	Never	Fixed	100	284807
SON_3AY_HESAP_ISLETIM_UCRETI	Continuous	3338	61	Coerce	Never	Fixed	100	284807
SON_3AY_VERGI_TAHSILI	Continuous	3093	1552	Coerce	Never	Fixed	100	284807
SON_3AY_KREDİ_KARTI_EKSTRESI	Continuous	2322	1742	Coerce	Never	Fixed	100	284807
SON_1GUN_KREDI_BASVURUSU	Continuous	998	224	Coerce	Never	Fixed	100	284807
SON_1GUN_INTERNET_BASVURUSU	Continuous	1964	1400	Coerce	Never	Fixed	100	284807
SON_1GUN_ATM_INTERAKSIYONLARI	Continuous	549	108	Coerce	Never	Fixed	100	284807
SON_1GUN_HESAPTAKI_PARA	Continuous	2594	215	Coerce	Never	Fixed	100	284807
SON_1GUN_HESAP_ISLETIM_UCRETI	Continuous	905	142	Coerce	Never	Fixed	100	284807
SON_1GUN_VERGI_TAHSILI	Continuous	2905	1866	Coerce	Never	Fixed	100	284807
SON_1GUN_KREDİ_KARTI_EKSTRESI	Continuous	2077	1187	Coerce	Never	Fixed	100	284807
AMOUNT	Continuous	2376	1700	None	Never	Fixed	100	284807
BNK_FRAUD	Continuous	0	492	None	Never	Fixed	100	284807

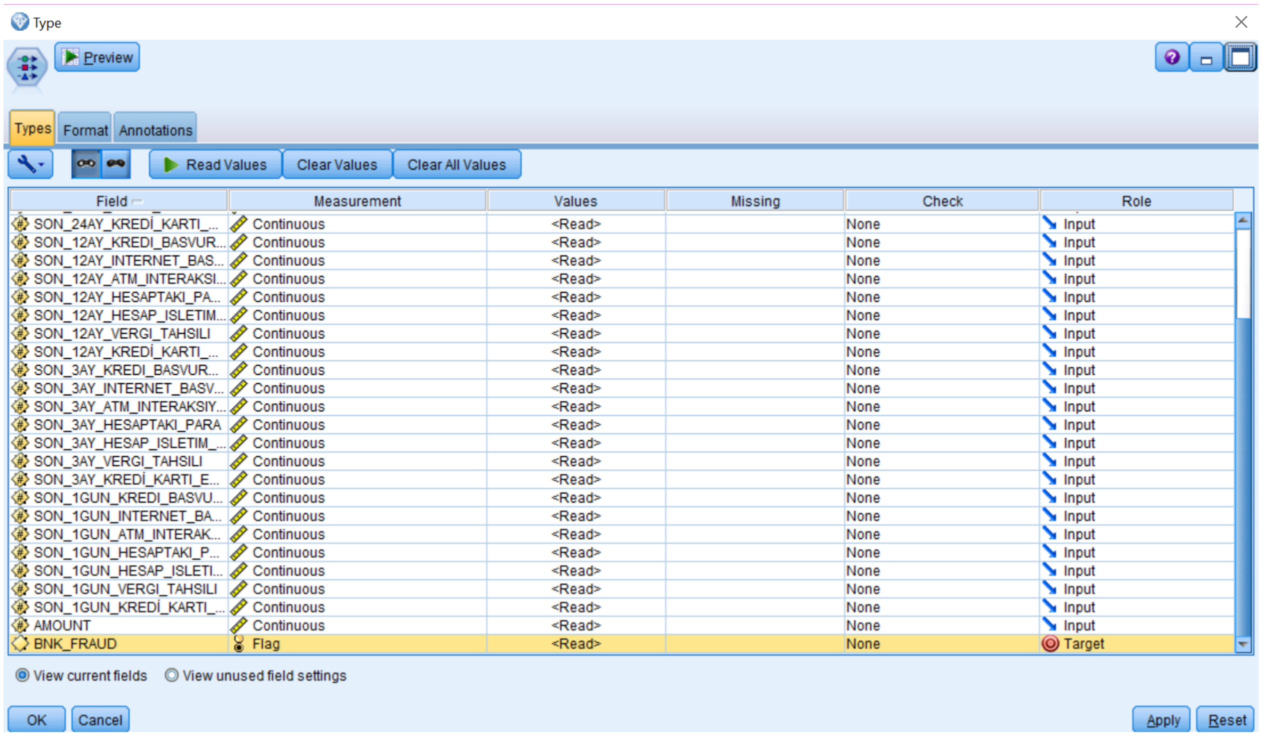
After clicking we can see that there is a star that represent the updated data set.



1.2. Feature Selection

After this step, we must introduce the data to the system such as giving what is parameters are input or target. In our data set, all the parameters that other than BNK_FRAUD is considered as input and BNK_FRAUD parameter considered as a target.

Also, we must consider the parameters measurements. If in that parameter our data can be floating point then its measurement is continuous. But if we look at the BNK_FRAUD data set. It is [0,1]. Therefore, it can only have two values. With that information, we must set its measurement as flag.



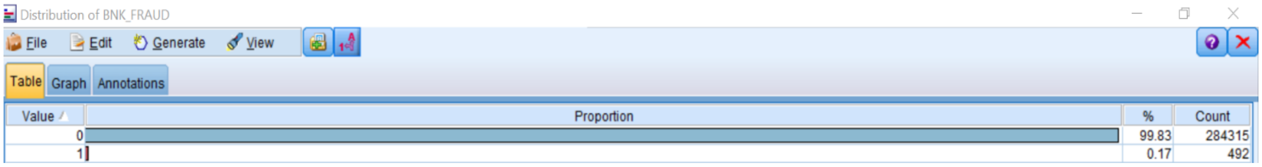
Field	Measurement	Values	Missing	Check	Role
SON_24AY_KREDİ_KARTI_...	Continuous	<Read>		None	Input
SON_12AY_KREDI_BASVUR...	Continuous	<Read>		None	Input
SON_12AY_INTERNET_BAS...	Continuous	<Read>		None	Input
SON_12AY_ATM_INTERAKSI...	Continuous	<Read>		None	Input
SON_12AY_HESAPTAKI_PA...	Continuous	<Read>		None	Input
SON_12AY_HESAP_ISLETIM...	Continuous	<Read>		None	Input
SON_12AY_VERGI_TAHSILI	Continuous	<Read>		None	Input
SON_12AY_KREDİ_KARTI_...	Continuous	<Read>		None	Input
SON_3AY_KREDI_BASVUR...	Continuous	<Read>		None	Input
SON_3AY_INTERNET_BASV...	Continuous	<Read>		None	Input
SON_3AY_ATM_INTERAKSIY...	Continuous	<Read>		None	Input
SON_3AY_HESAPTAKI_PARA	Continuous	<Read>		None	Input
SON_3AY_HESAP_ISLETIM_...	Continuous	<Read>		None	Input
SON_3AY_VERGI_TAHSILI	Continuous	<Read>		None	Input
SON_3AY_KREDİ_KARTI_E...	Continuous	<Read>		None	Input
SON_1GUN_KREDI_BASVU...	Continuous	<Read>		None	Input
SON_1GUN_INTERNET_BA...	Continuous	<Read>		None	Input
SON_1GUN_ATM_INTERAK...	Continuous	<Read>		None	Input
SON_1GUN_HESAPTAKI_P...	Continuous	<Read>		None	Input
SON_1GUN_HESAP_ISLETI...	Continuous	<Read>		None	Input
SON_1GUN_VERGI_TAHSILI	Continuous	<Read>		None	Input
SON_1GUN_KREDİ_KARTI_...	Continuous	<Read>		None	Input
AMOUNT	Continuous	<Read>		None	Input
BNK_FRAUD	Flag	<Read>		None	Target

● View current fields ○ View unused field settings

1.3. Data Balancing

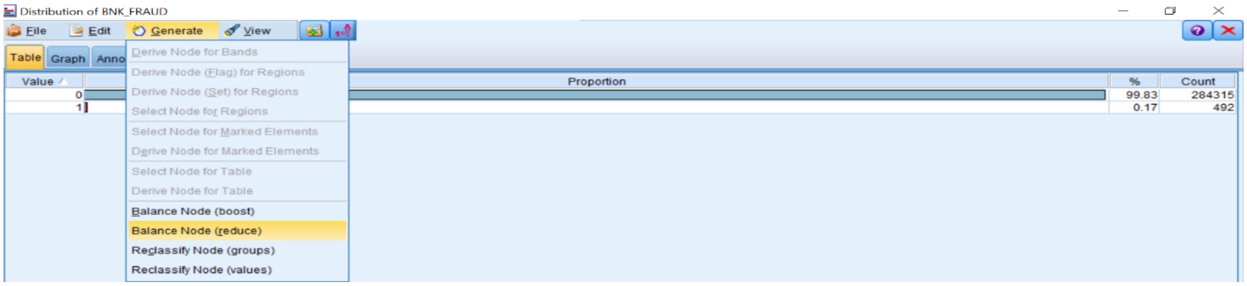
After cleaning and introducing data to the system, the next step is balancing the any unbalance data.

When we look at the BNK_FRAUD parameter, we see that we have very much of an unbalanced data set. If we leave the set like that the model wouldn't even consider the very small set number. So, we have to balance the data set to give each of that an equal consideration chance.



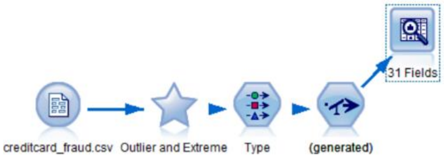
Value	Proportion	%	Count
0		99.83	284315
1		0.17	492

In order to balance the unbalanced data, we must click the Generate and Balance Node(reduce).

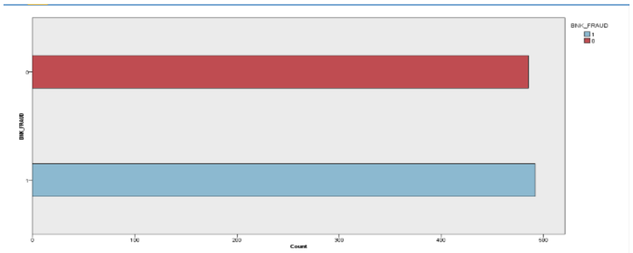


Value	Proportion	%	Count
0		99.83	284315
1		0.17	492

Then we can see the (generated) symbol as the output of our balanced data.



If we Data Audit the (generated), we can see that BNK_FRAUD is now balanced.



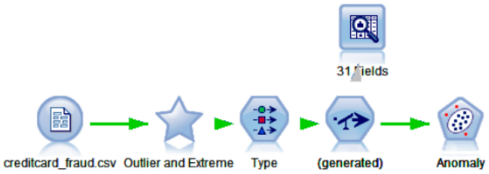
1.4. Anomaly Detection

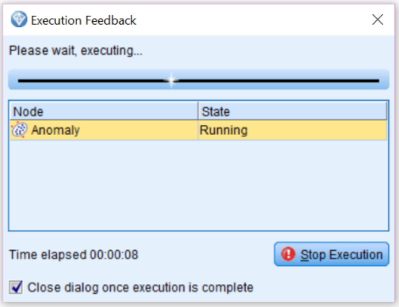
In this step we are looking for data that are abnormal. Anomaly processor automatically scan the dataset to find abnormal or extraordinary one. We call this process anomaly detection.

Anomaly detection uses a two-step clustering algorithm. With this algorithm, it observes the ones that does not fit the nature or characteristic of the data.

Addition to the clustering algorithm, it does;

- Gives a score that shows which cluster it is belong to
- The anomaly index that assigned to each observation shows abnormality scale





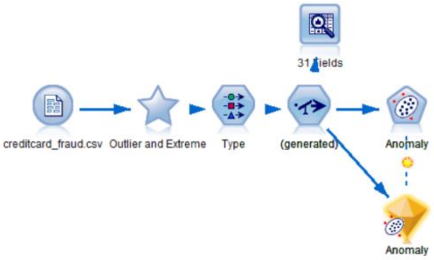
Please wait, executing...

Node	State
Anomaly	Running

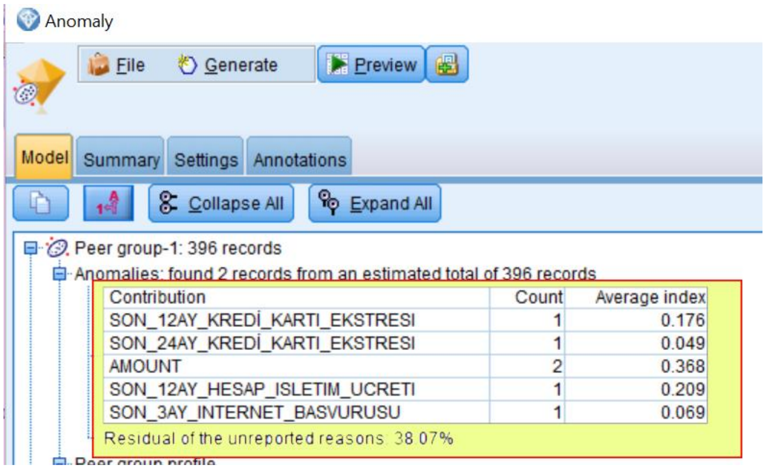
Time elapsed 00:00:08

☑ Close dialog once execution is complete

After it executes there will be a diamond as a result.



With the result we can see the abnormal values.



Peer group-1: 396 records

Anomalies: found 2 records from an estimated total of 396 records

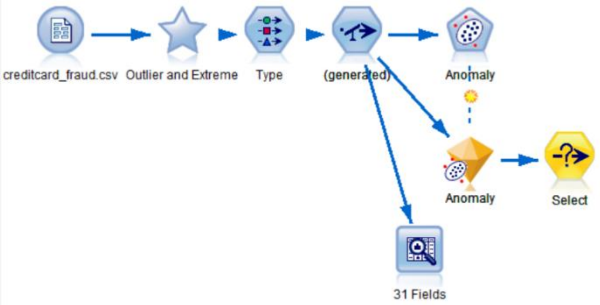
Contribution	Count	Average index
SON_12AY_KREDİ_KARTI_EKSTRESI	1	0.176
SON_24AY_KREDİ_KARTI_EKSTRESI	1	0.049
AMOUNT	2	0.368
SON_12AY_HESAP_ISLETIM_UCRETI	1	0.209
SON_3AY_INTERNET_BASVURUSU	1	0.069

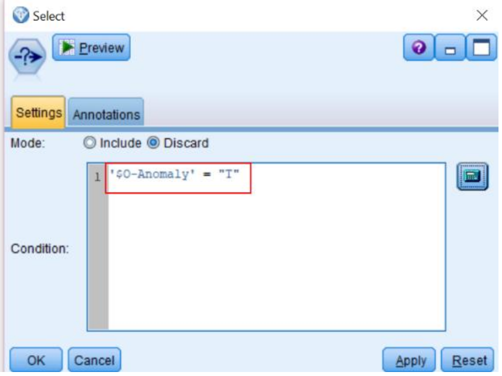
Residual of the unreported reasons: 38.07%

Peer group profile

1.5. Anomaly Omitting

These are the ones that we want to omit. So, we select these ones by select operation. We are selecting by coding '$0-Anomaly' = "T" and discard it. Therefore, in the result we must omit the ones that anomaly detection detected as anomaly.





Mode: ○ Include ● Discard

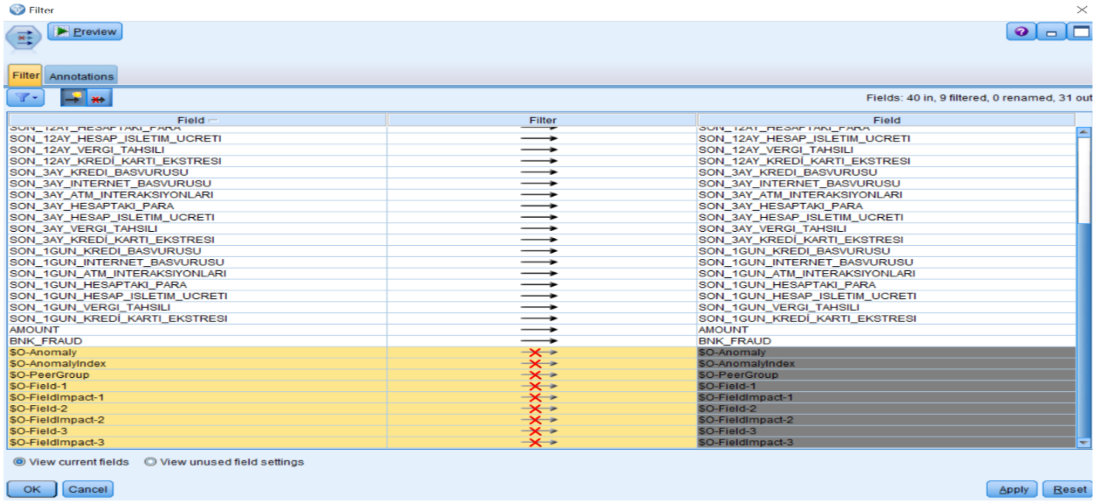
Condition:

```
'$0-Anomaly' = "T"
```

1.6. Filtering Parameters

Before giving the data set to the model, we must make sure that we don't give any unnecessary parameters to the model. Since the other parameters that we don't care can affect the result if we don’t discard it. So, we must filter the parameters according to the needs of the problem that we are facing.

After anomaly detection there is some parameters that added to the data set. We don’t want it go in the model so we must filter those. We used filter operation in this case.



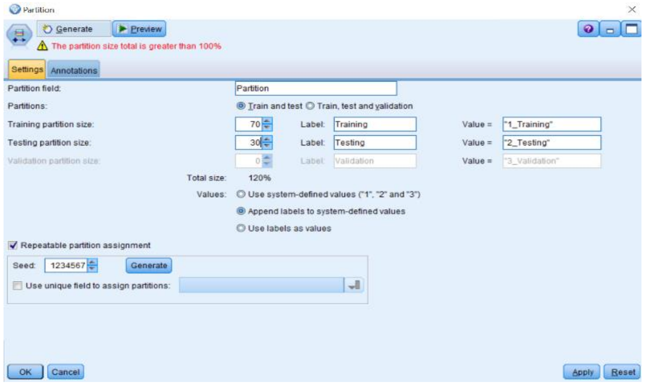
Fields: 40 in, 9 filtered, 0 renamed, 31 out

Field	Filter	Field
SON_12AY_HESAPTAKI_PARA	→	SON_12AY_HESAPTAKI_PARA
SON_12AY_HESAP_ISLETIM_UCRETI	→	SON_12AY_HESAP_ISLETIM_UCRETI
SON_12AY_VERGI_TAHSILI	→	SON_12AY_VERGI_TAHSILI
SON_12AY_KREDİ_KARTI_EKSTRESI	→	SON_12AY_KREDİ_KARTI_EKSTRESI
SON_3AY_KREDI_BASVURUSU	→	SON_3AY_KREDI_BASVURUSU
SON_3AY_INTERNET_BASVURUSU	→	SON_3AY_INTERNET_BASVURUSU
SON_3AY_ATM_INTERAKSIYONLARI	→	SON_3AY_ATM_INTERAKSIYONLARI
SON_3AY_HESAPTAKI_PARA	→	SON_3AY_HESAPTAKI_PARA
SON_3AY_HESAP_ISLETIM_UCRETI	→	SON_3AY_HESAP_ISLETIM_UCRETI
SON_3AY_VERGI_TAHSILI	→	SON_3AY_VERGI_TAHSILI
SON_3AY_KREDİ_KARTI_EKSTRESI	→	SON_3AY_KREDİ_KARTI_EKSTRESI
SON_1GUN_KREDI_BASVURUSU	→	SON_1GUN_KREDI_BASVURUSU
SON_1GUN_INTERNET_BASVURUSU	→	SON_1GUN_INTERNET_BASVURUSU
SON_1GUN_ATM_INTERAKSIYONLARI	→	SON_1GUN_ATM_INTERAKSIYONLARI
SON_1GUN_HESAPTAKI_PARA	→	SON_1GUN_HESAPTAKI_PARA
SON_1GUN_HESAP_ISLETIM_UCRETI	→	SON_1GUN_HESAP_ISLETIM_UCRETI
SON_1GUN_VERGI_TAHSILI	→	SON_1GUN_VERGI_TAHSILI
SON_1GUN_KREDİ_KARTI_EKSTRESI	→	SON_1GUN_KREDİ_KARTI_EKSTRESI
AMOUNT	→	AMOUNT
BNK_FRAUD	→	BNK_FRAUD
$O-Anomaly	✗→	$O-Anomaly
$O-AnomalyIndex	✗→	$O-AnomalyIndex
$O-PeerGroup	✗→	$O-PeerGroup
$O-Field-1	✗→	$O-Field-1
$O-FieldImpact-1	✗→	$O-FieldImpact-1
$O-Field-2	✗→	$O-Field-2
$O-FieldImpact-2	✗→	$O-FieldImpact-2
$O-Field-3	✗→	$O-Field-3
$O-FieldImpact-3	✗→	$O-FieldImpact-3

● View current fields ○ View unused field settings

1.7. Data Partition

We will be giving the models to how much of a percentage that it must participate the dataset. We use the partition operation. We decided that for train it is %70, and for test %30.



⚠ The partition size total is greater than 100%

Partition field: Partition

Partitions: ● Train and test ○ Train, test and validation

Training partition size: 70 Label: Training Value = "1_Training"

Testing partition size: 30 Label: Testing Value = "2_Testing"

Validation partition size: 0 Label: Validation Value = "3_Validation"

Total size: 120%

Values: ○ Use system-defined values ("1", "2" and "3")

● Append labels to system-defined values

○ Use labels as values

☑ Repeatable partition assignment

Seed: 1234567

☐ Use unique field to assign partitions:

2. Neural Network

Neural networks are inspired by human brain when comes to decision making.

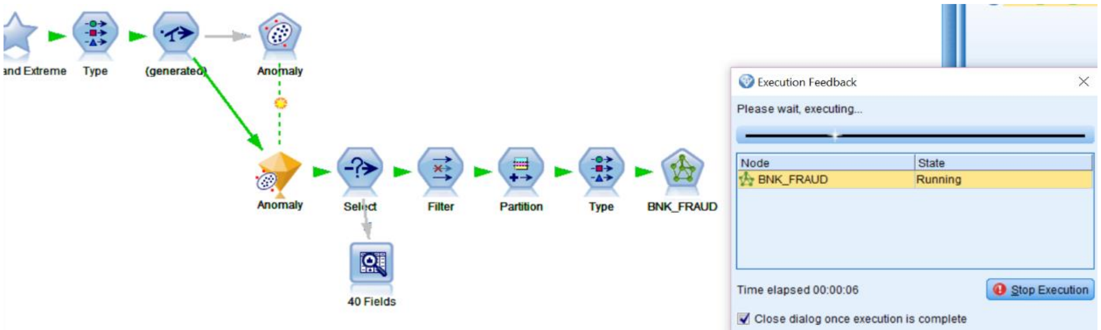
A neural network consists of a 3 main layer. Which they are

- Input layer

- Hidden layer
- Output layer

In IBM Spss Modeler, it only allows two hidden layers.

Basically, neural network is basis of a neurons that allows us a recognize the pattern. It helps us to clustering and classification.



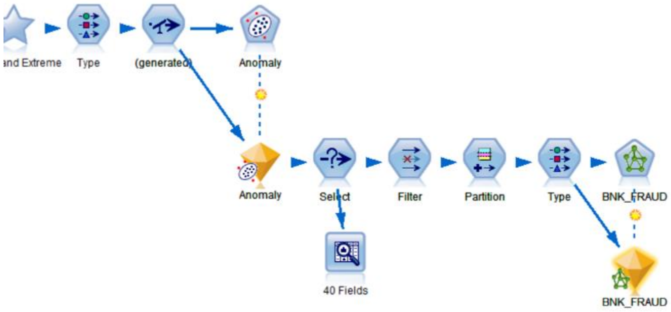
Please wait, executing...

Node	State
BNK_FRAUD	Running

Time elapsed 00:00:06

☑ Close dialog once execution is complete

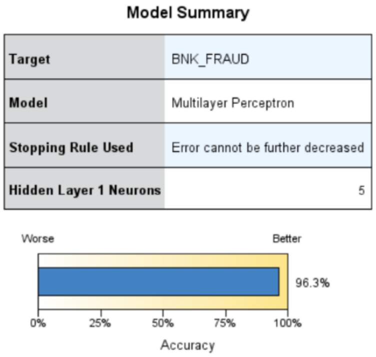
After its execution as a result we can see a diamond.



If we go inside of a neural network by clicking the diamond, we can see the structure of it.

2.1. Inside of Neural Network

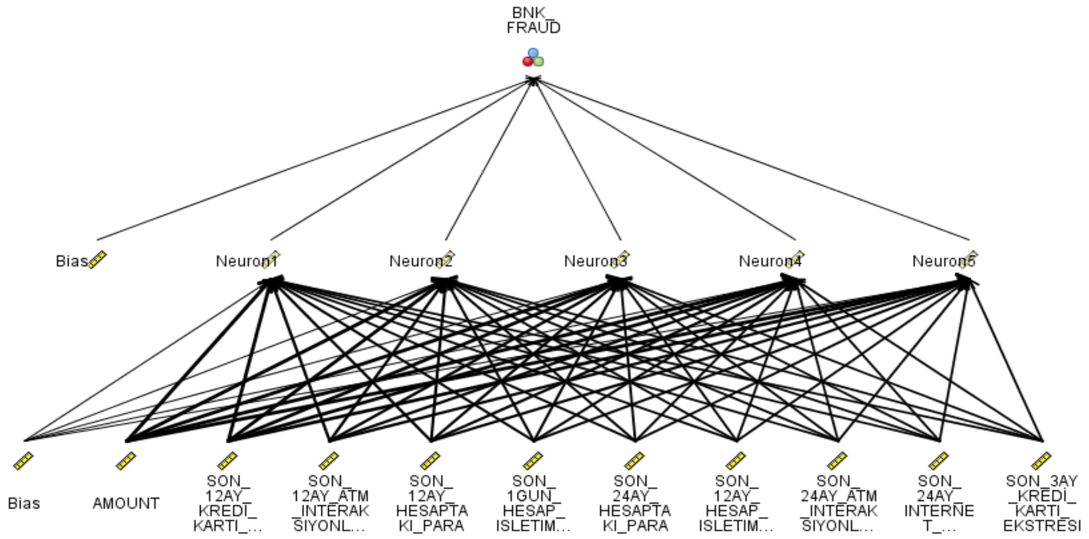
Inside the model we can see the general performance rate of our accuracy.



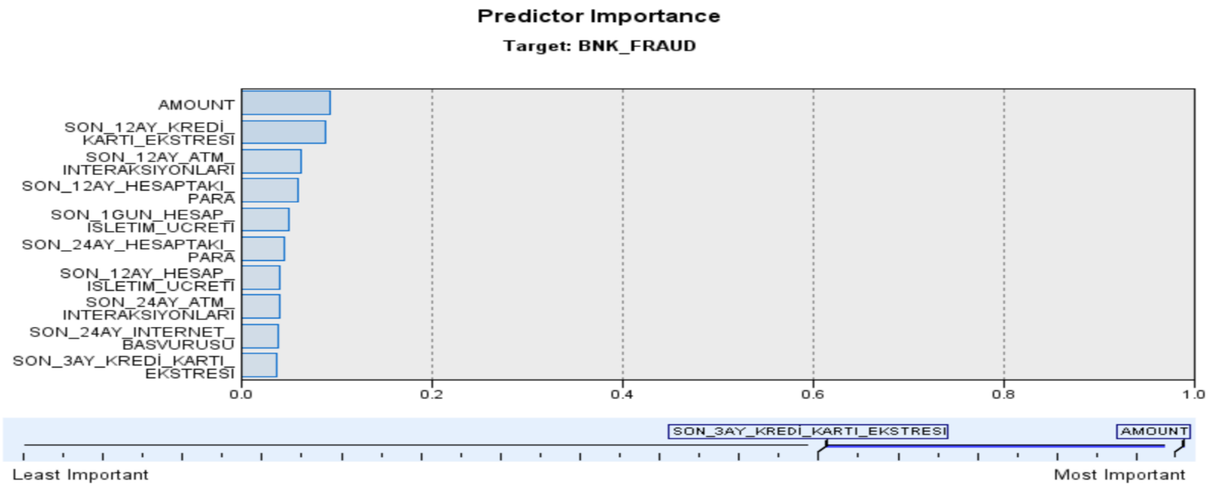
Model Summary

Target	BNK_FRAUD
Model	Multilayer Perceptron
Stopping Rule Used	Error cannot be further decreased
Hidden Layer 1 Neurons	5

We can see five neuron and bias, also the input parameters.



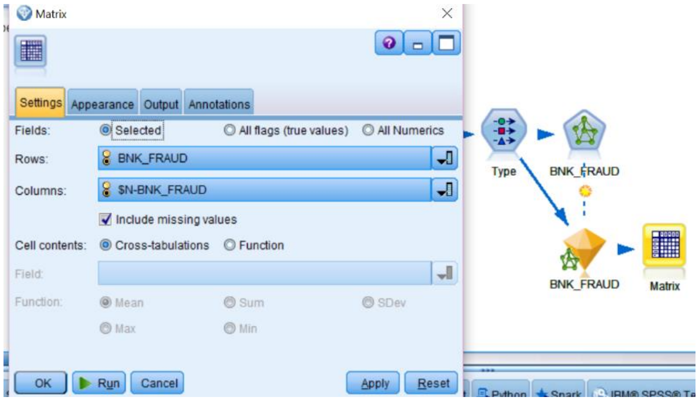
Also, we can see the importance degree of them.



2.2. Confusion Matrices

If we insert a matrix and configure it, we can see the neural networks confusion matrix.

Confusion matrix shows us the performance of our model and how accurately does is predict.



Fields: ● Selected ○ All flags (true values) ○ All Numerics

Rows: BNK_FRAUD

Columns: $N-BNK_FRAUD

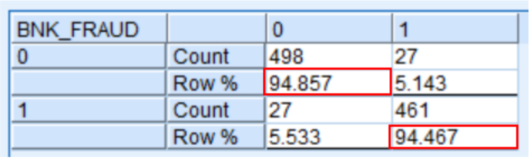
☑ Include missing values

Cell contents: ● Cross-tabulations ○ Function

Field:

Function: ● Mean ○ Sum ○ SDev ○ Max ○ Min

The confusion matrix can be seen in that table:

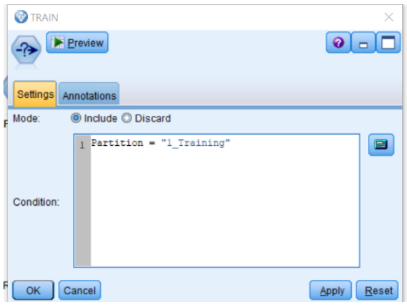


BNK_FRAUD		0	1
0	Count	498	27
	Row %	94.857	5.143
1	Count	27	461
	Row %	5.533	94.467

So,

- True Positive = %94.857
- False Positive = %5.143
- True Negative = %94.467
- False Negative = %5.533

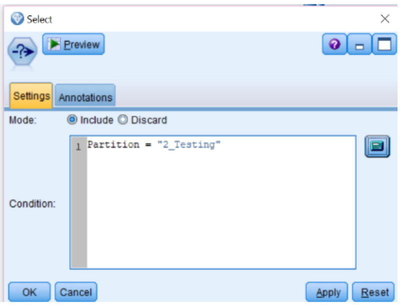
To be sure that model doesn't overfitting, we are going to try with two different subsets of our dataset. Overfitting occurs when a model learns specifically just for that data set. So, in that case we are providing new two data sets with select operation.



Mode: ● Include ○ Discard

Condition:

```
Partition = "1_Training"
```

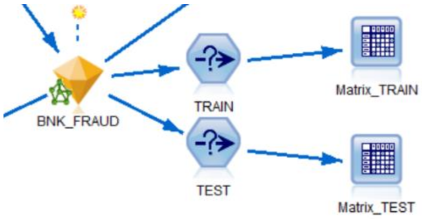


Mode: ● Include ○ Discard

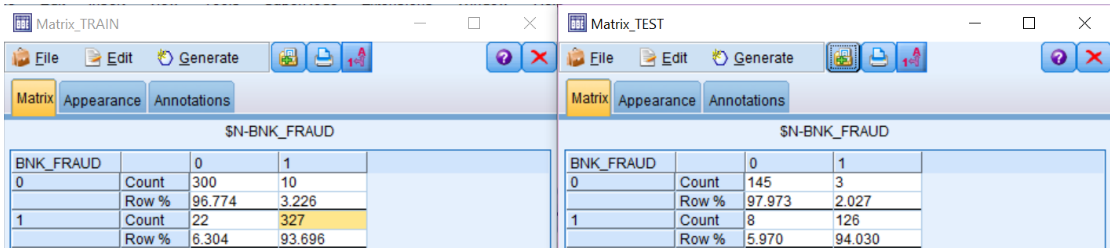
Condition:

```
Partition = "2_Testing"
```

After dividing new two subsets:



If we look at the two new data sets confusion matrices:



Matrix_TRAIN

$N-BNK_FRAUD

BNK_FRAUD		0	1
0	Count	300	10
	Row %	96.774	3.226
1	Count	22	327
	Row %	6.304	93.696

Matrix_TEST

$N-BNK_FRAUD

BNK_FRAUD		0	1
0	Count	145	3
	Row %	97.973	2.027
1	Count	8	126
	Row %	5.970	94.030

We see that the performance of the model stays still.

True Positive of Matrix_TRAIN = %96.774

True Positive of Matrix_TEST = %97.973

96.774 ~ 97.973

True Negative of Matrix_TRAIN = %93.696

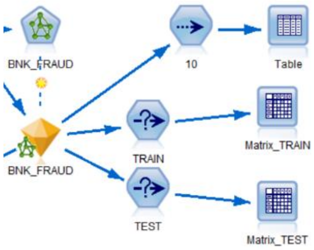
True Negative of Matrix_TEST = %94.030

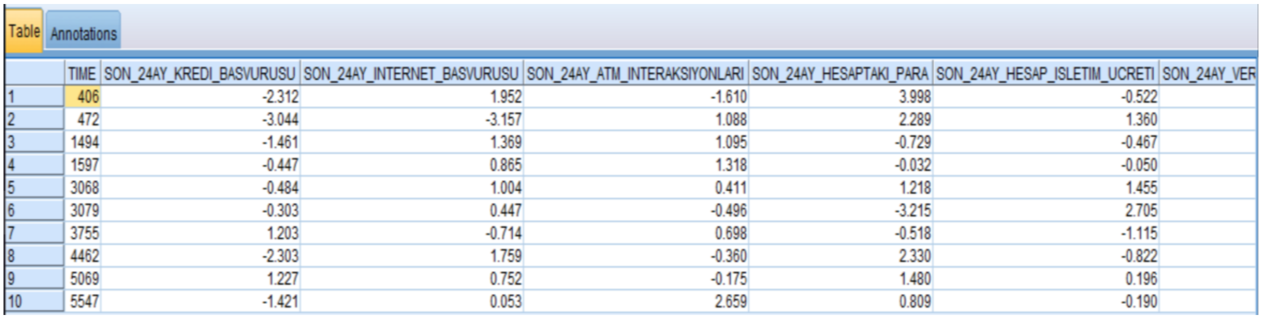
93.696 ~ 94.030

Since these two matrices are consistent, we can say that our model is successful.

2.3. Top 10 Frauds

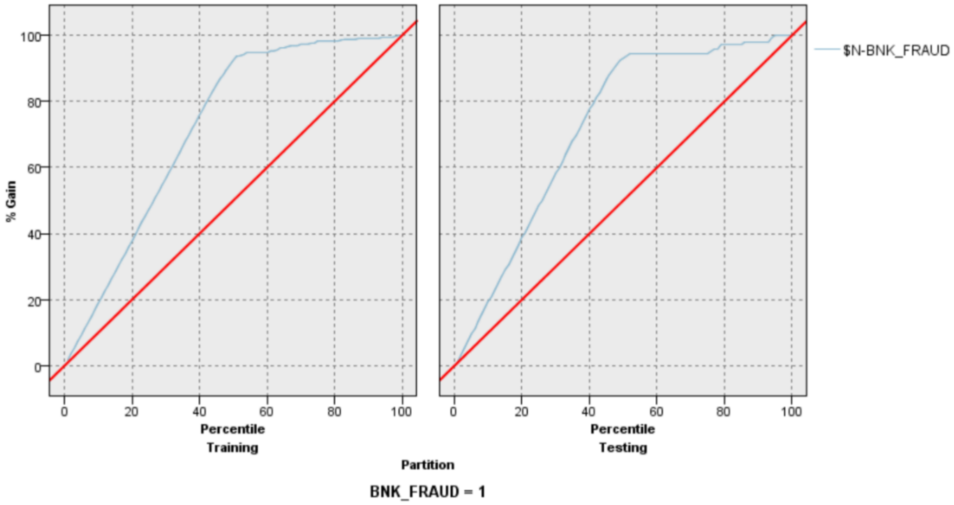
Let’s look at the top 10 frauds that we detect with sample operation that we specialize a first 10. Then we attached a table to look.





	TIME	SON_24AY_KREDI_BASVURUSU	SON_24AY_INTERNET_BASVURUSU	SON_24AY_ATM_INTERAKSIYONLARI	SON_24AY_HESAPTAKI_PARA	SON_24AY_HESAP_ISLETIM_UCRETI	SON_24AY_VER
1	406	-2.312	1.952	-1.610	3.998	-0.522	
2	472	-3.044	-3.157	1.088	2.289	1.360	
3	1494	-1.461	1.369	1.095	-0.729	-0.467	
4	1597	-0.447	0.865	1.318	-0.032	-0.050	
5	3068	-0.484	1.004	0.411	1.218	1.455	
6	3079	-0.303	0.447	-0.496	-3.215	2.705	
7	3755	1.203	-0.714	0.698	-0.518	-1.115	
8	4462	-2.303	1.759	-0.360	2.330	-0.822	
9	5069	1.227	0.752	-0.175	1.480	0.196	
10	5547	-1.421	0.053	2.659	0.809	-0.190	

2.4. Evaluation of Model



In this Gain chart, the blue line represents what would the best performance of the best model behave. The red line represents our models performance. This graphic gives us how near we are compare to perfect behaviour.