**BANK FRAUD DETECTION WITH IBM SPSS MODELER**

**BY**

**AYÇA NUR VANLI**

# TABLE OF CONTENTS

1. Data Preparation

**1.1. Data Cleaning**

**1.2. Feature Selection**

**1.3. Data Balancing**

**1.4. Anomaly Detection**

**1.5. Anormaly Omitting**

**1.6. Parameter Filtering**

**1.6. Data Partition**

2. Neural Network

**2.1. Inside of Neural Network**

**2.2. Confusion Matrices**

**2.3. Top 10 Frauds**

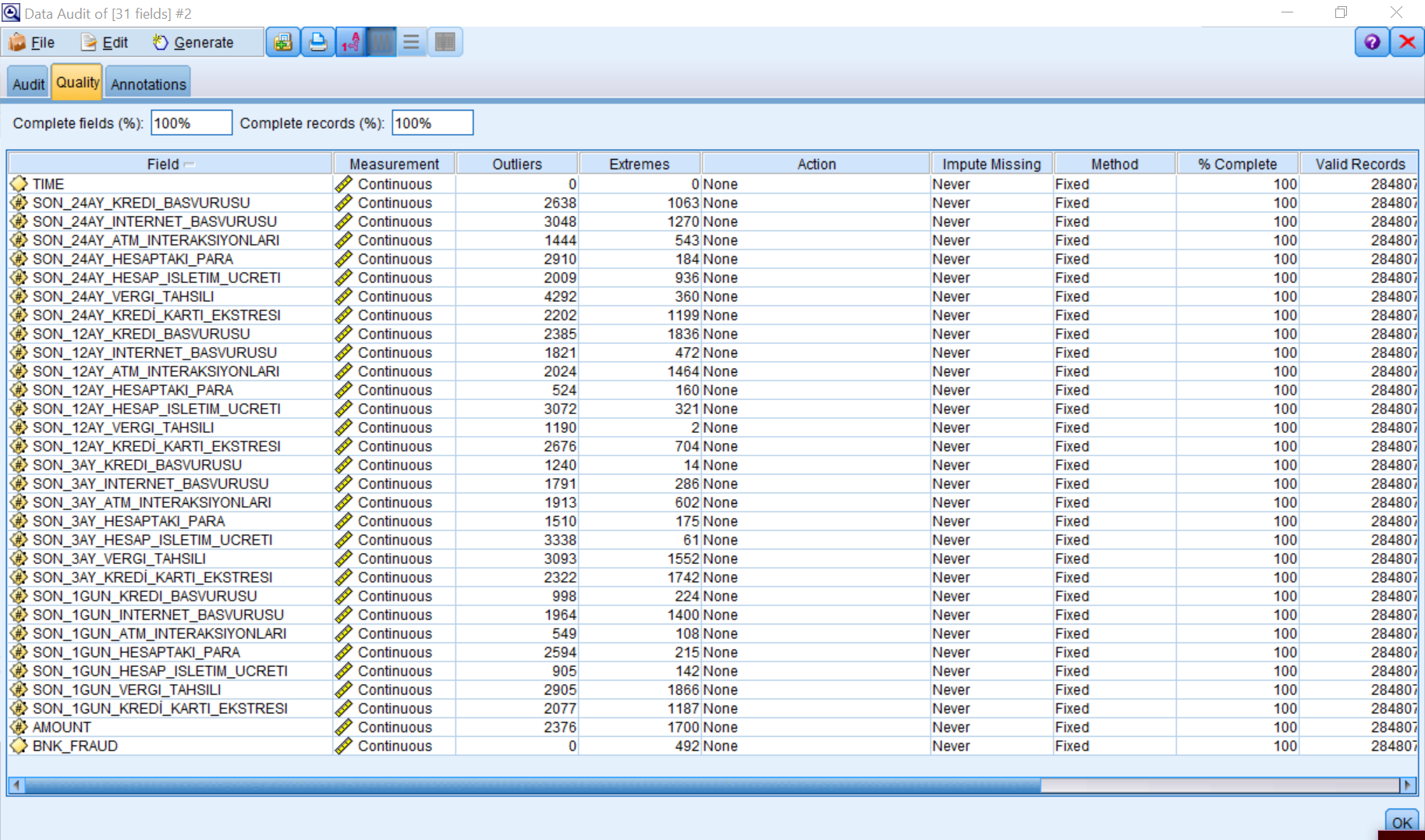
**2.4. Evaluation of Model**

**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.

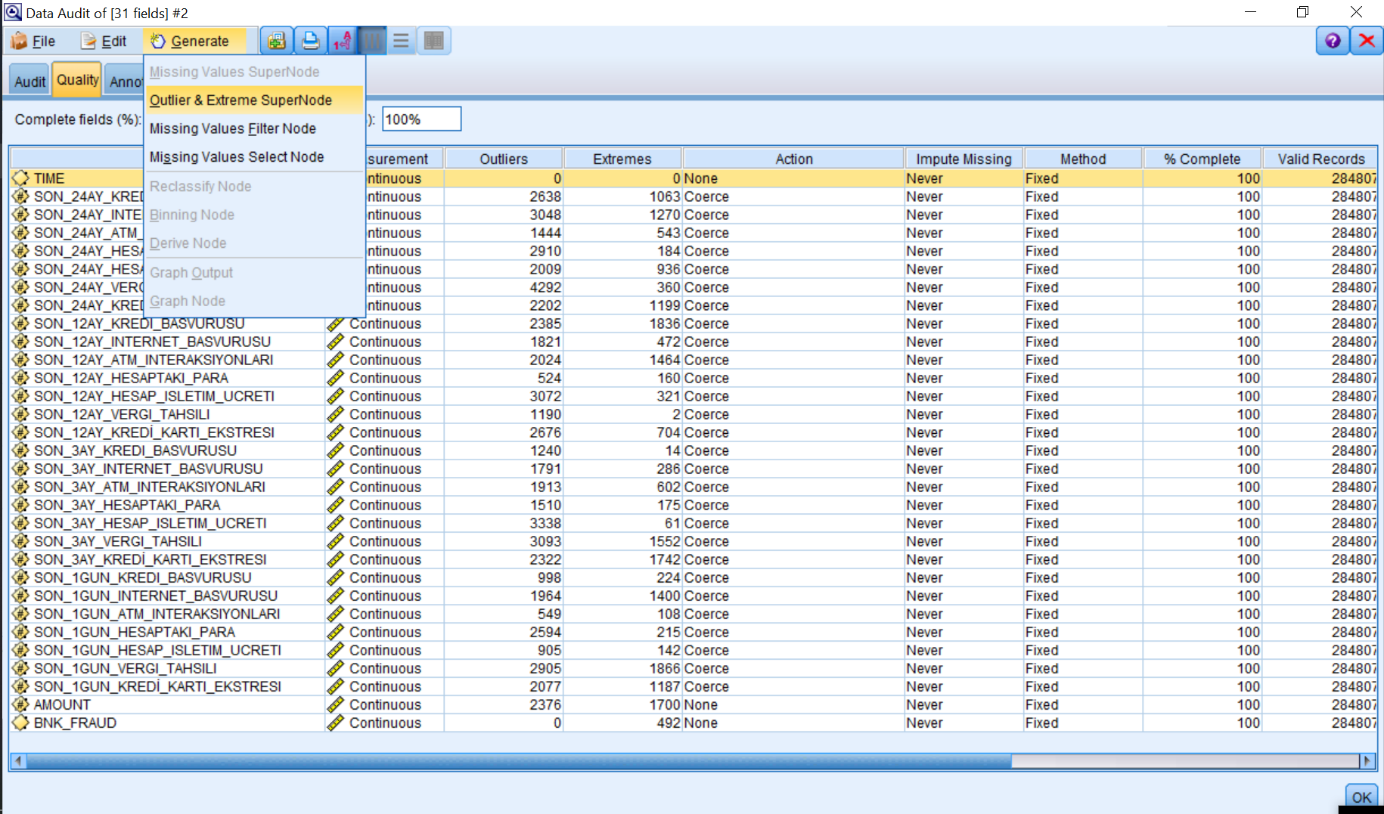
According to what we need and different perspectives, the steps that we are taking in this phase of the data mining can differentiate. If some steps are not fitting the data set’s nature we can skip or have a different approach to it.

**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.

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.



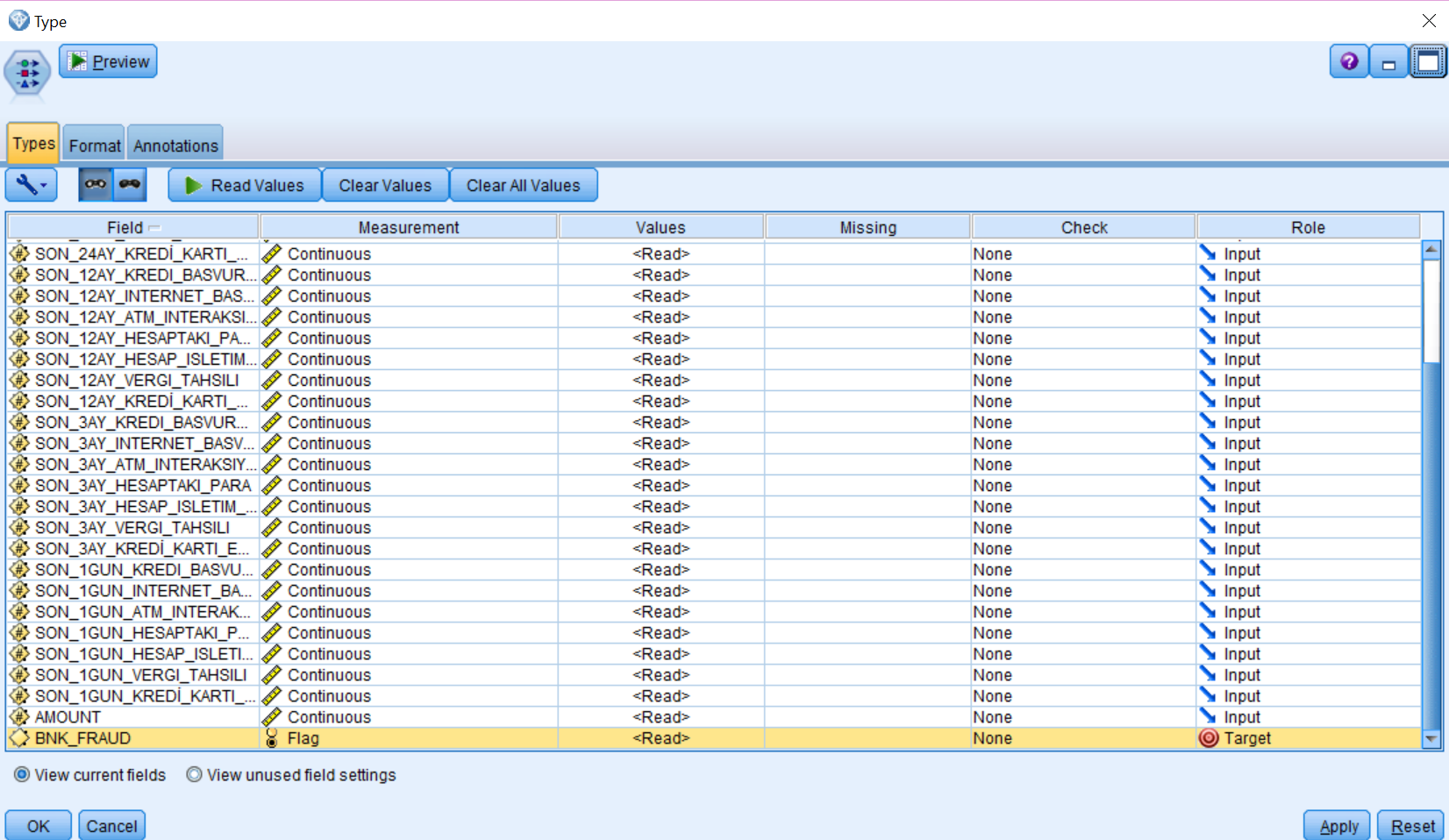
After clicking we can see that there is a star that represent the updated data set.



**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.

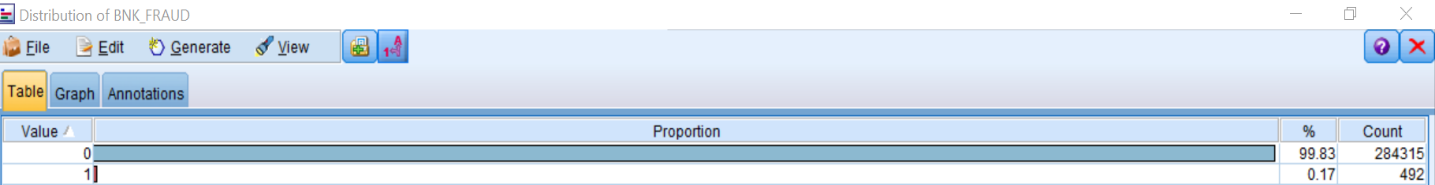
Also, we must consider the parameters measurements. If in that parameter our data can be floating point then its measurement is continuous. But if we look at the BNK\_FRAUD data set. It is [0,1]. Therefore, it can only have two values. With that information, we must set its measurement as flag.



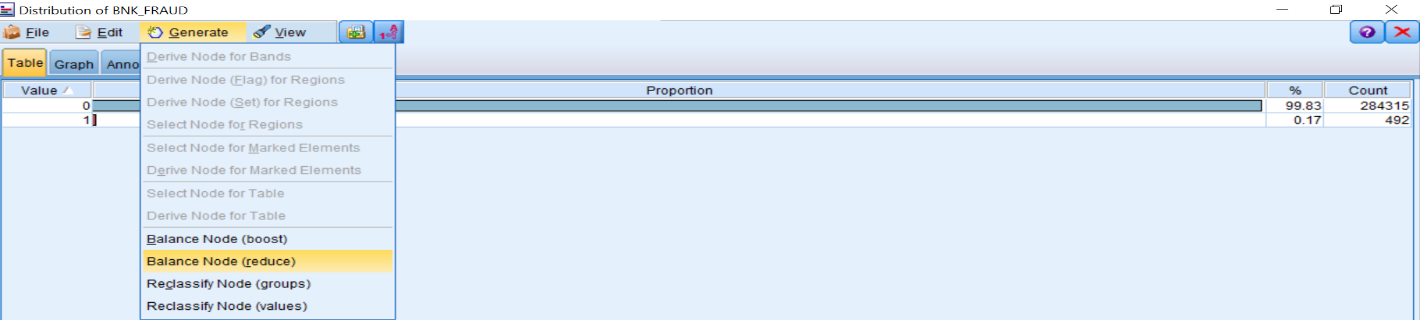
**1.3. Data Balancing**

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

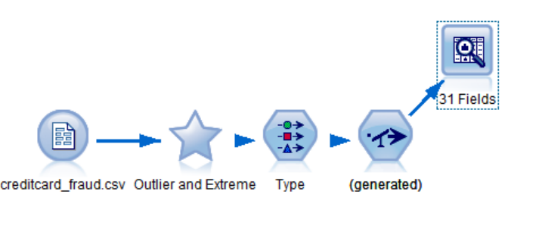
When we look at the BNK\_FRAUD parameter, we see that we have very much of an unbalanced data set. If we leave the set like that the model wouldn’t even consider the very small set number. So, we have to balance the data set to give each of that an equal consideration chance.



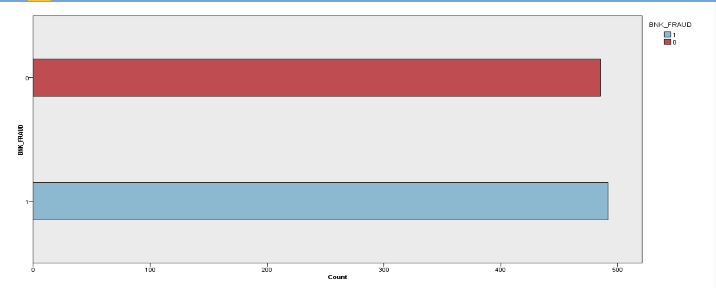
In order to balance the unbalanced data, we must click the Generate and Balance Node(reduce).



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.



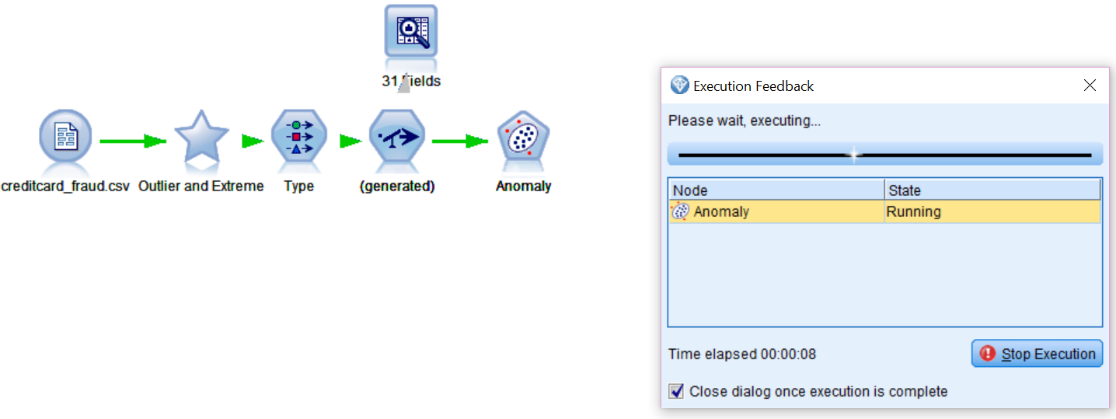
**1.4. Anomaly Detection**

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.

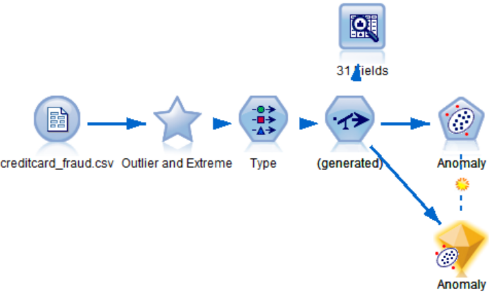
Anomaly detection uses a two-step clustering algorithm. With this algorithm, it observes the ones that does not fit the nature or characteristic of the data.

Addition to the clustering algorithm, it does;

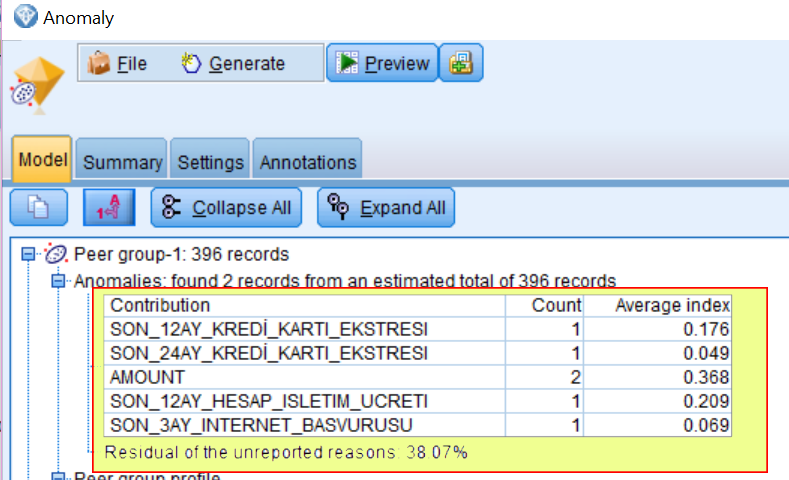
* Gives a score that shows which cluster it is belong to
* The anomaly index that assigned to each observation shows abnormality scale



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

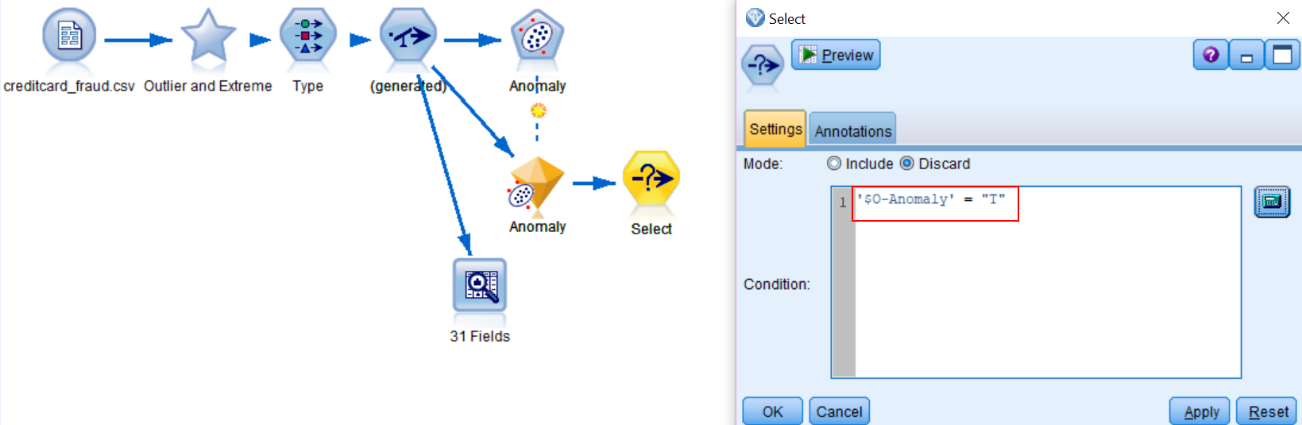


With the result we can see the abnormal values.



**1.5. Anomaly Omitting**

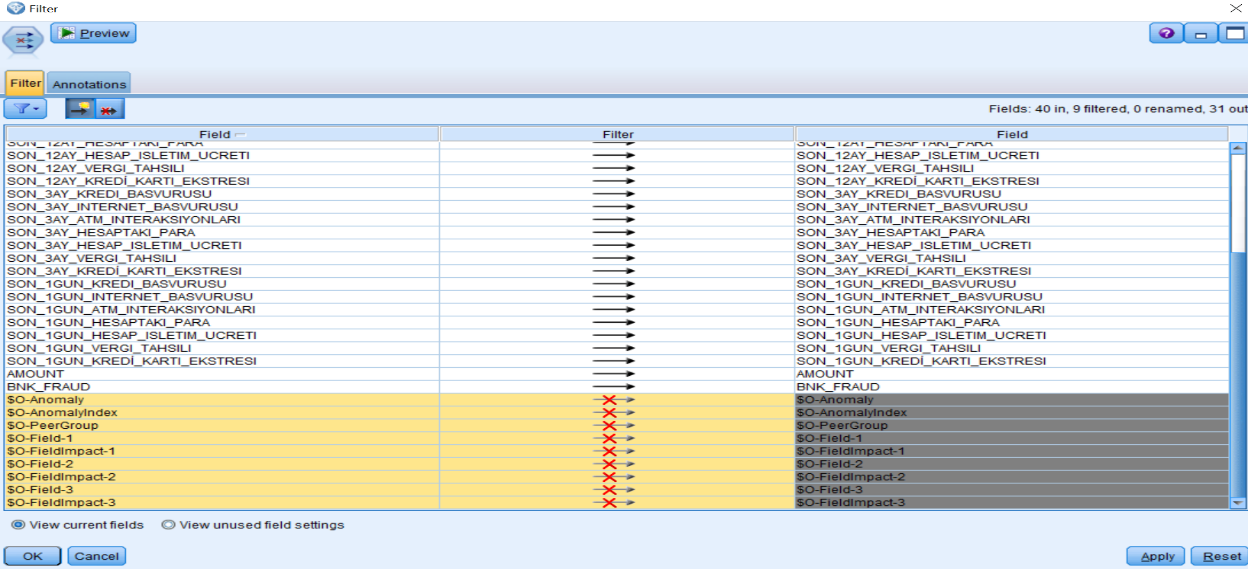
These are the ones that we want to omit. So, we select these ones by select operation. We are selecting by coding ‘$0-Anomaly’ = “T” and discard it. Therefore, in the result we must omit the ones that anomaly detection detected as anomaly.



**1.6. Filtering Parameters**

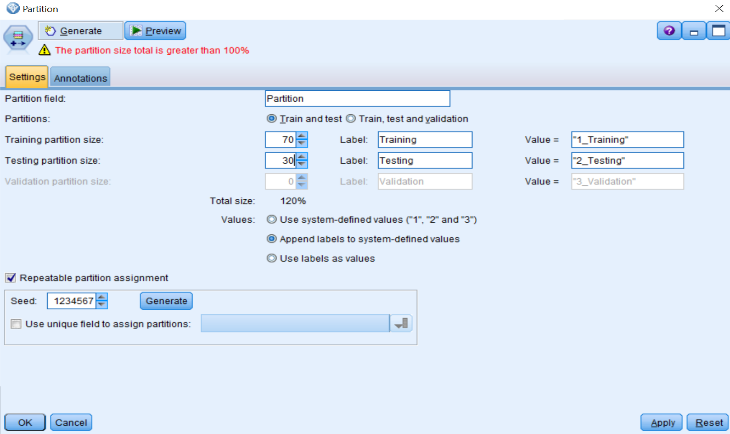
Before giving the data set to the model, we must make sure that we don’t give any unnecessary parameters to the model. Since the other parameters that we don’t care can affect the result if we don’t discard it. So, we must filter the parameters according to the needs of the problem that we are facing.

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.



**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.



**2. Neural Network**

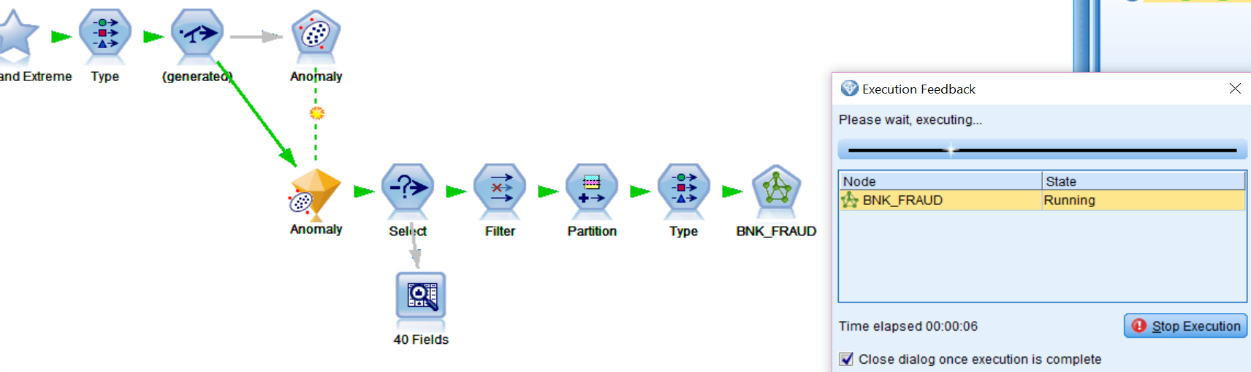
Neural networks are inspired by human brain when comes to decision making.

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

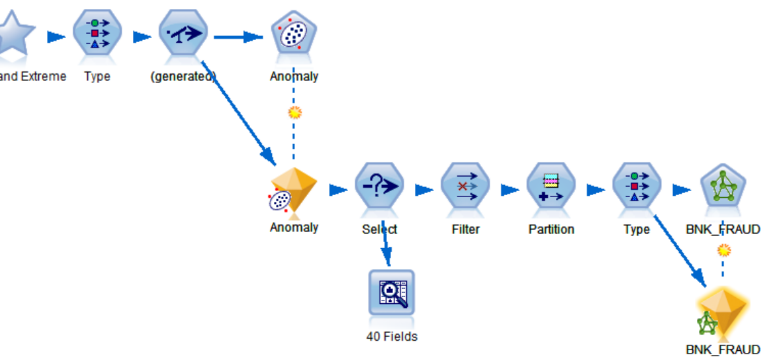
* Input layer
* Hidden layer
* Output layer

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.



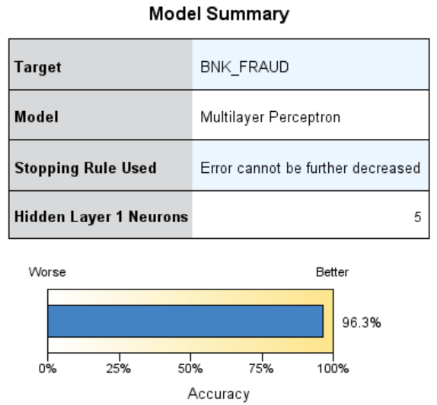
After its execution as a result we can see a diamond.



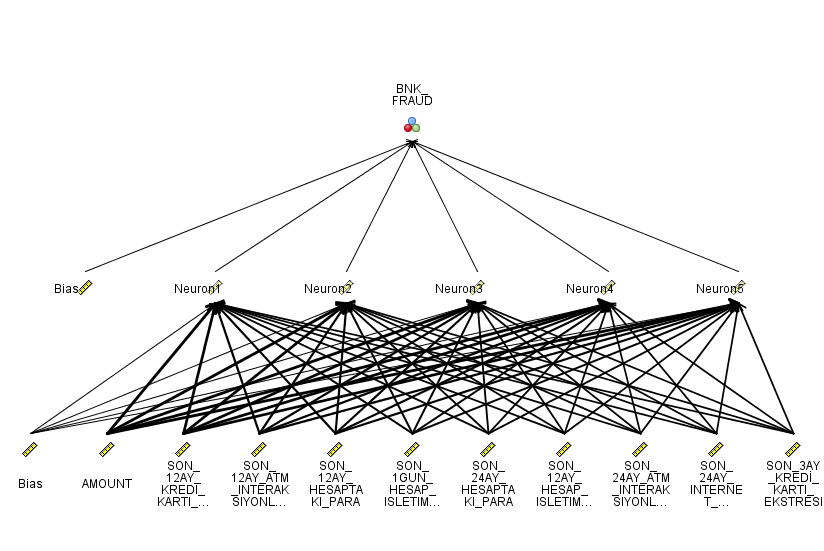
If we go inside of a neural network by clicking the diamond, we can see the structure of it.

**2.1. Inside of Neural Network**

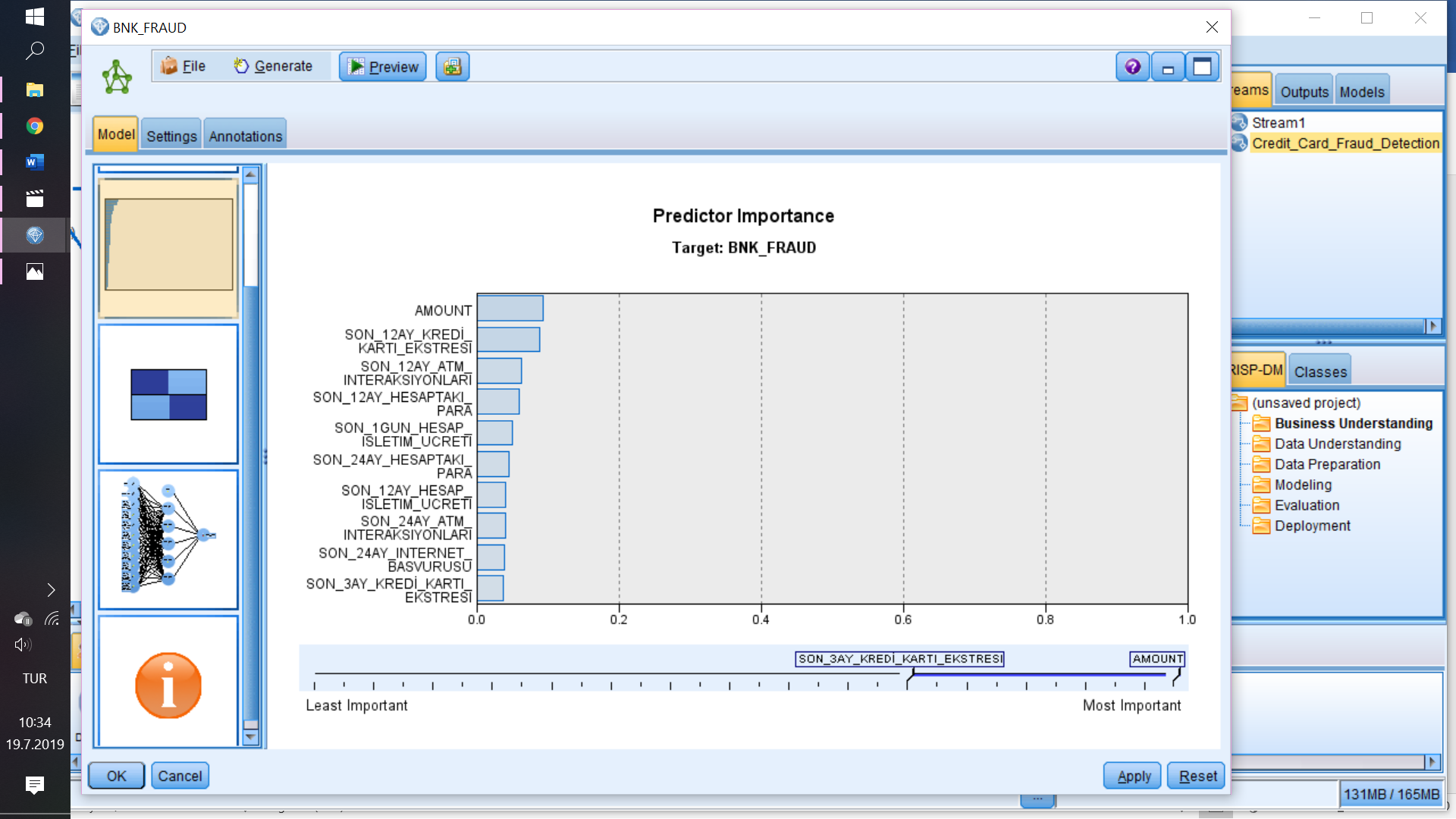
Inside the model we can see the general performance rate of our accuracy.



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



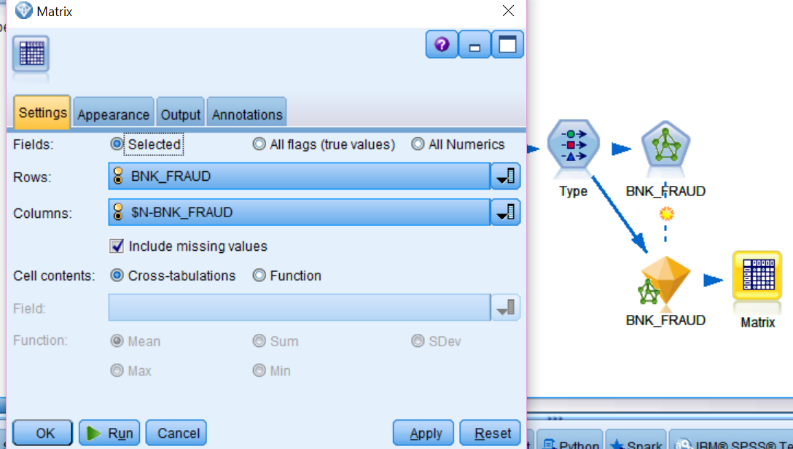
Also, we can see the importance degree of them.



**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.



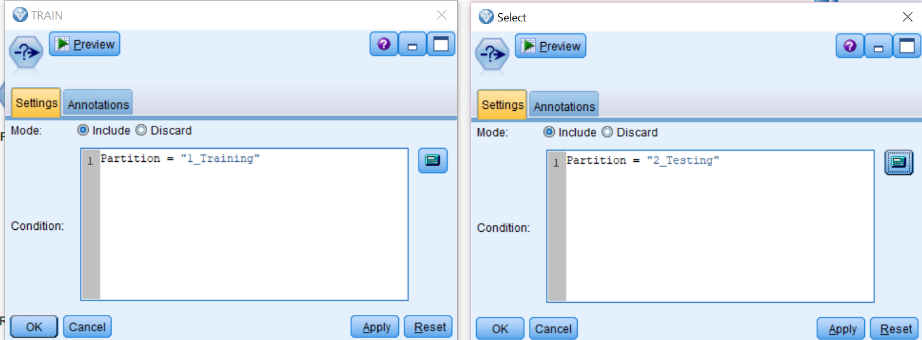
The confusion matrix can be seen in that table:



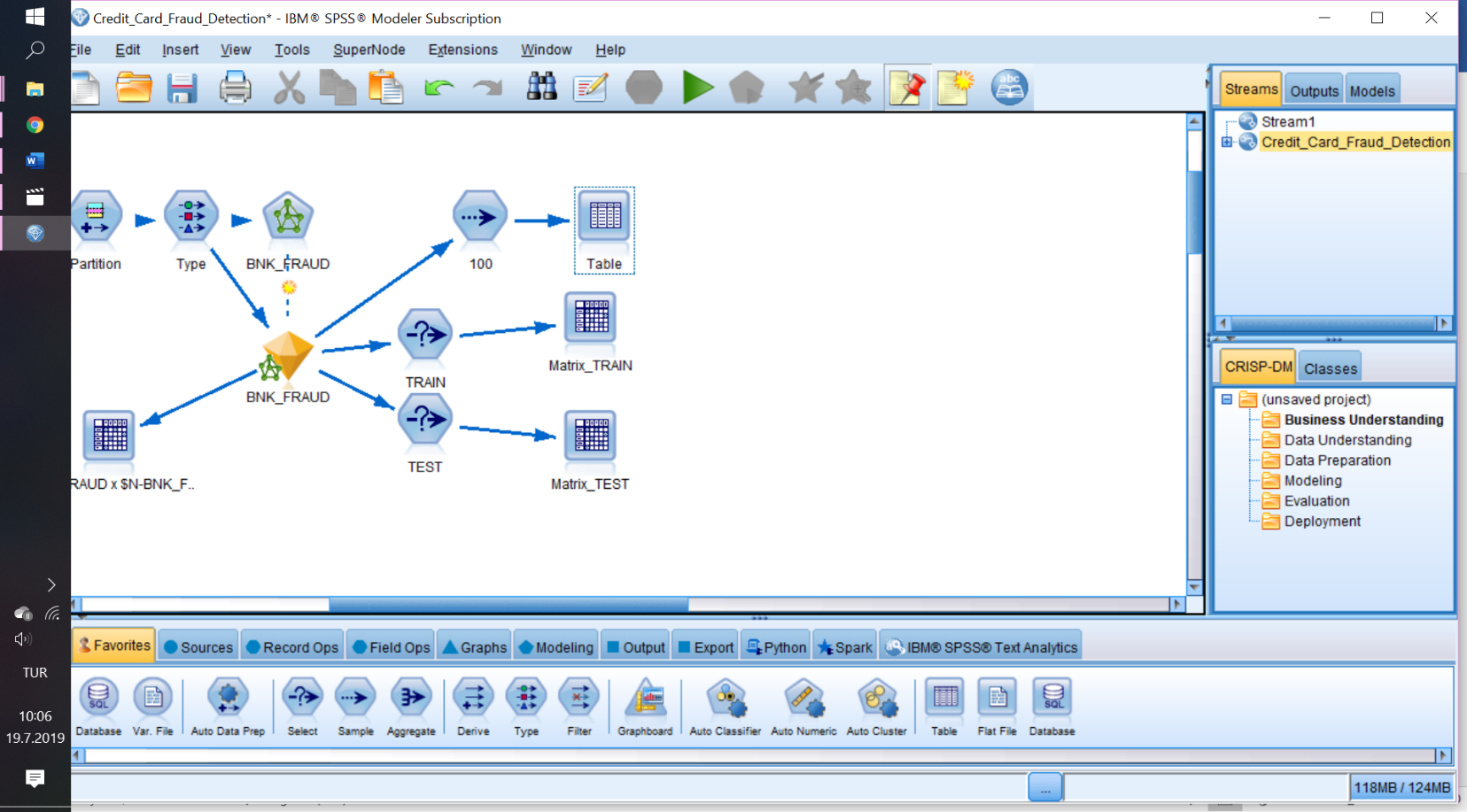
So,

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

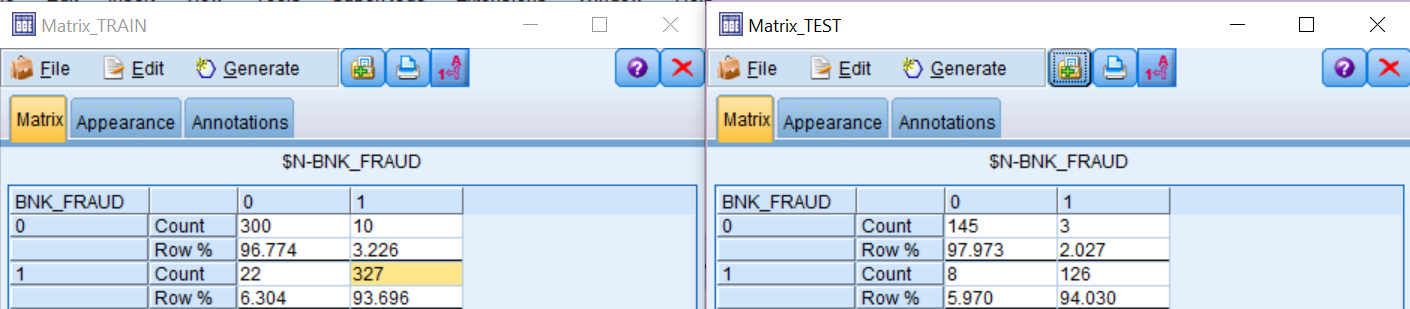
To be sure that model doesn’t overfitting, we are going to try with two different subsets of our dataset. Overfitting occurs when a model learns specifically just for that data set. So, in that case we are providing new two data sets with select operation.



After dividing new two subsets:



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



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

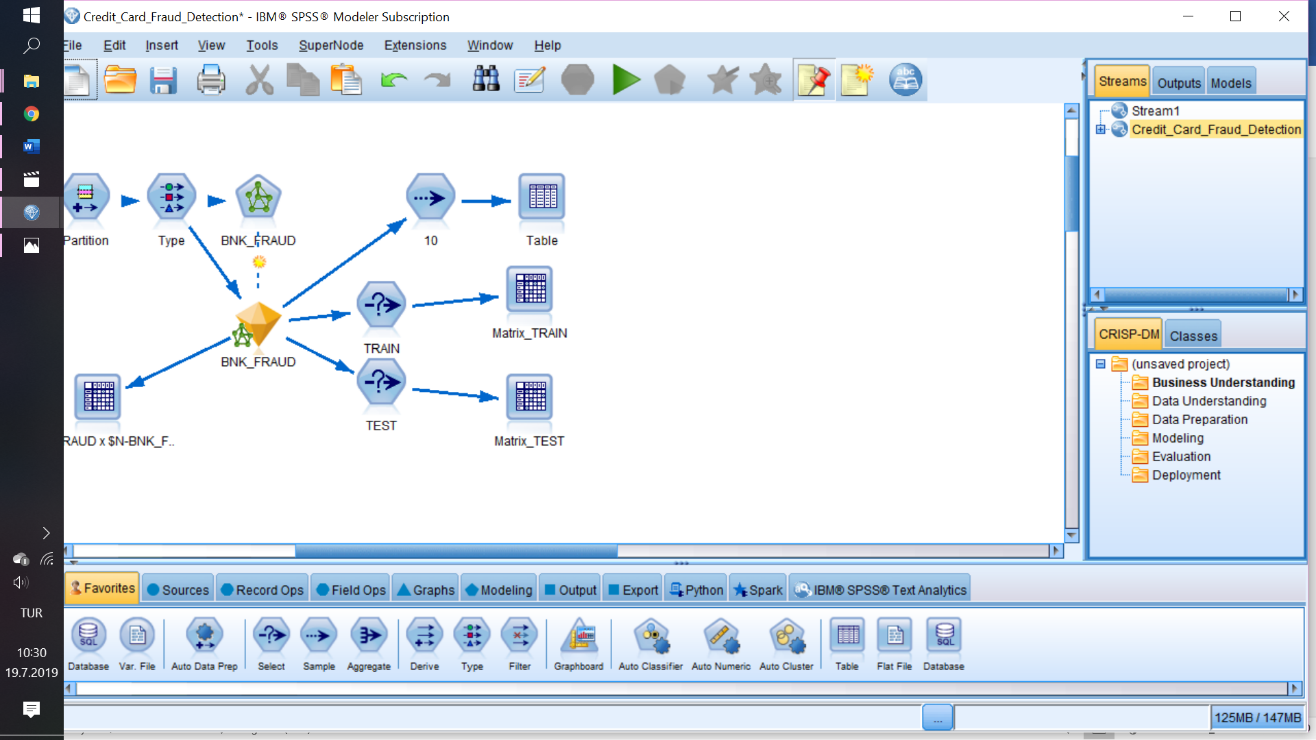
True Negative of Matrix\_TEST = %94.030

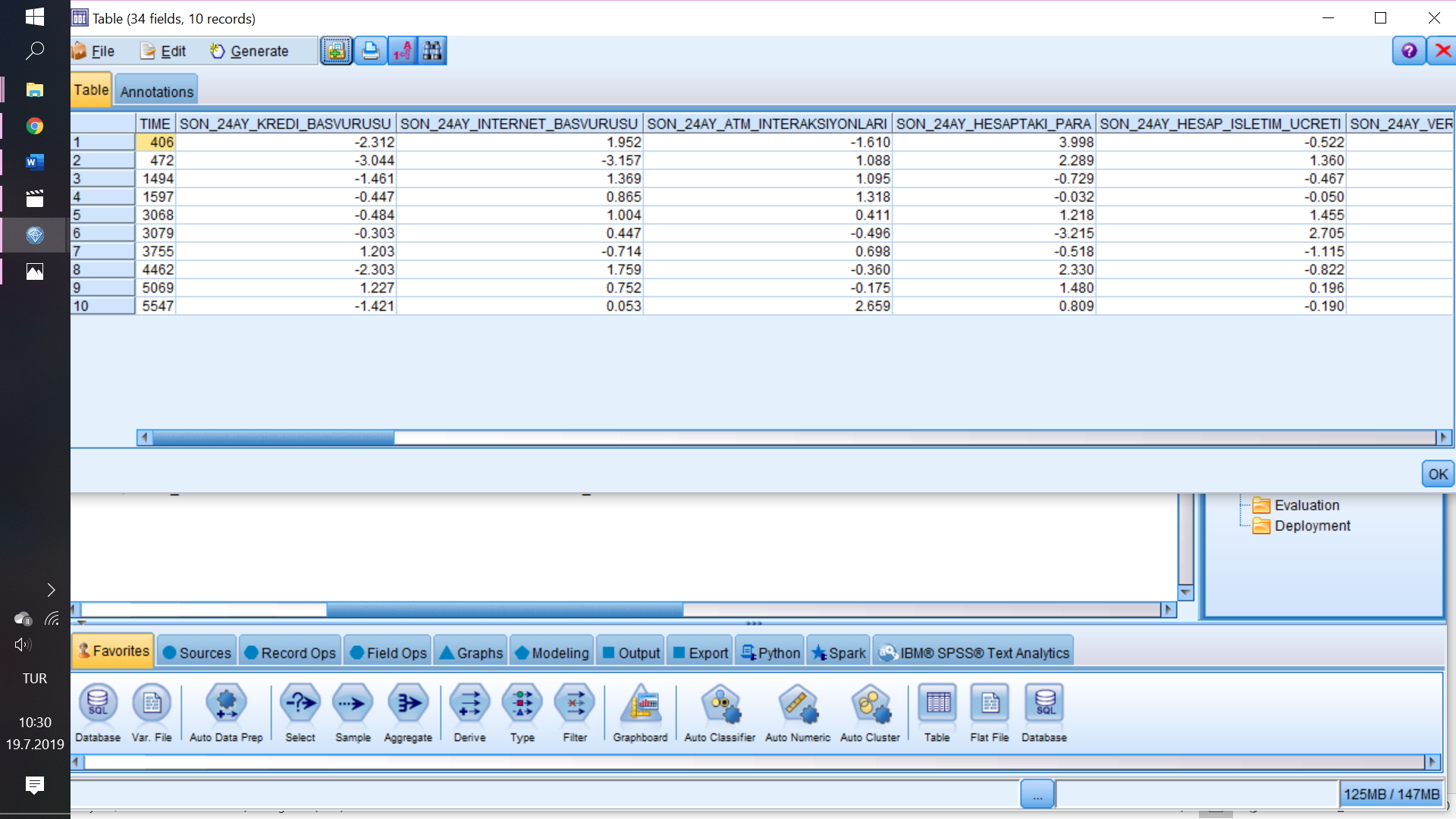
93.696 ~ 94.030

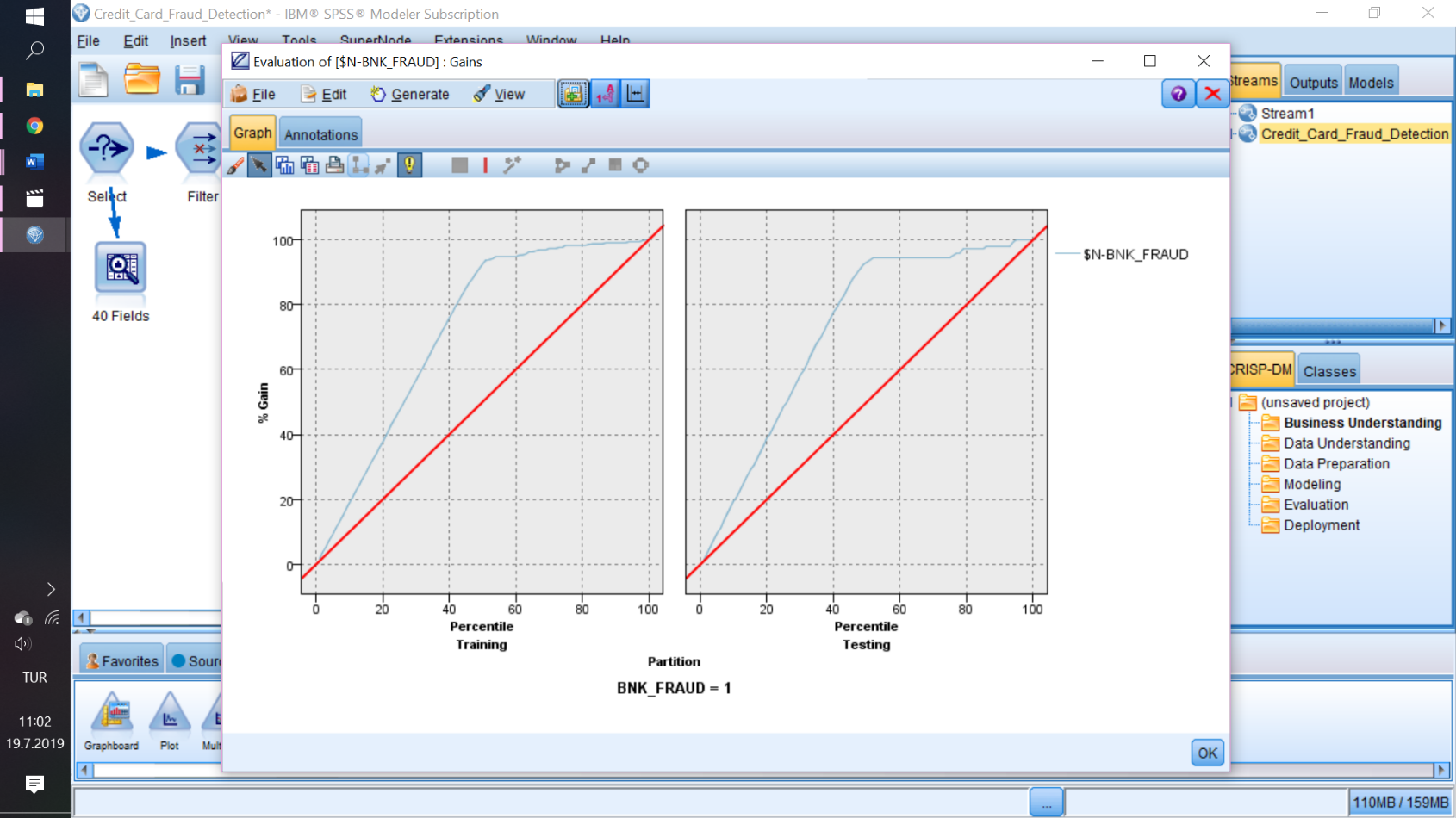
Since these two matrices are consistent, we can say that our model is successful.

**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.



**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.