Data Mining project

**IS 665 Spring 2018**

****

Sahana Umashankar

**Contents:**

1. Abstract………………………………………………………………………………3
2. Introduction………………………………………………………………………..3
3. Classification Analytics………………………………………………………..4
4. Data Set Description……………………………………………………………4
5. Problem Formulation………………………………………………………….6
6. Classification Models………………………………………………………….6-7
   1. Naïve Bayes Algorithm…………………………………………………………..6
   2. Neural Network Algorithm…………………………………………………….7
7. Experimental Results………………………………………………………….8-14
   1. Naïve Bayes Algorithm…………………………………………………………..8-12
      1. Implementing
      2. Simple Distributor
      3. Confusion matrix with accuracy, precision & recall
      4. F-measure & Lift
   2. Neural Network Algorithm…………………………………………………….12-14
      1. Implementing
      2. Simple Distributor
      3. Confusion matrix with accuracy, precision & recall
      4. F-measure & Lift
8. Performance Vector …………………………………………………………..15
9. Conclusion…………………………………………………………………………..15-16

**ABSTRACT :**

Over the past decade with improved trust for financial banking, the credit card system has boomed widely creating easy access to this privilege by major portion of the population. However as a banking institute it is equally necessary to take responsibility in monitoring customers who are entitled to such privilege. This paper discusses about the two different systematic and comprehensive Data Mining techniques i,e. Naive Bayes and Neural Networks to create a model which can predict whether a customer would default on his payment or not. The data set used- **“Default of Credit Card Clients Dataset”** has been obtained from kaggle.com and contains information on default payments, demographic factors, credit data, history of payment, and bill statements of credit card clients in Taiwan from April 2005 to September 2005.

**INTRODUCTION:**

In today’s increasingly electronic society and with the rapid advances of electronic commerce on the internet, the use of credit cards for purchases has become convenient and necessary. Credit card transactions have become de facto standard for internet and web based e-commerce. However, the growing number of credit card users alters the banking institutes to take responsibility in monitor customers who are entitled to such services. Though Banks have been using conventional predictive systems to check on their customer’s financial records, Data mining has proven to be the new found systematic approach to deal with such huge data, at the same time being accurate and authentic. In this paper we discuss about the most widely used Data mining techniques such Naive Bayes and Neural Networks to analyse and create a model for **“Default of Credit Card Clients Data set”** which can predict whether a customer would default on his payment or not. The data set was obtained from kaggle.com and contains valuable information of 25 variables such as default payments, demographic factors, credit data, history of payment, bill statements of credit card clients in Taiwan from April 2005 to September 2005. Thus obtained results from study of data set through Naive Bayes and Neural Networks are compared to determine which model produces best and accurate results to the given problem.

**CLASSIFICATION ANALYSIS:**

Classification is a data mining function that assigns items in a collection to target categories or classes. The goal of classification is to accurately predict the target class for each case in the data. Classifications are discrete and do not imply order. Continuous, floating-point values would indicate a numerical, rather than a categorical, target.

To meet the challenges of Big data, which is about volume,velocity,variety and veracity of data , a range of automatic methods for extracting useful information has been developed, among them classification.

In the model build (training) process, a classification algorithm finds relationships between the values of the predictors and the values of the target. Different classification algorithms use different techniques for finding relationships. These relationships are summarized in a model, which can then be applied to a different data set in which the class assignments are unknown.Classification has many applications in customer segmentation, business modeling, marketing, credit analysis, and biomedical and drug response modeling.

There are various classification models available today which include neural networks, k-nearest neighbor, decision tree, Naive bayes etc. For our project we are considering Naive bayes and neural networks classification models.

**DATASET DESCRIPTION:**

We have chosen **“Default of Credit Card Clients Dataset”** . This dataset contains information on default payments, demographic factors, credit data, history of payment, and bill statements of credit card clients in Taiwan from April 2005 to September 2005. In this project we are trying to predict the target variable which is default payment, which means whether a customer would default on his payment or not.

**Default Credit Card:** Happens when clients fail to adhere to the credit card agreement, by not paying the monthly bill.

**Main Goal:** Development of a system capable of detecting clients that will not be able to pay next month.

It has 25 features:

* ID: ID of each client
* LIMIT\_BAL: Amount of given credit in NT dollars (includes individual and family/supplementary credit
* SEX: Gender (1=male, 2=female)
* EDUCATION: (1=graduate school, 2=university, 3=high school, 4=others, 5=unknown, 6=unknown)
* MARRIAGE: Marital status (1=married, 2=single, 3=others)
* AGE: Age in years
* PAY\_0: Repayment status in September, 2005 (-1=pay duly, 1=payment delay for one month, 2=payment delay for two months, ... 8=payment delay for eight months, 9=payment delay for nine months and above)
* PAY\_2: Repayment status in August, 2005 (scale same as above)
* PAY\_3: Repayment status in July, 2005 (scale same as above)
* PAY\_4: Repayment status in June, 2005 (scale same as above)
* PAY\_5: Repayment status in May, 2005 (scale same as above)
* PAY\_6: Repayment status in April, 2005 (scale same as above)
* BILL\_AMT1: Amount of bill statement in September, 2005 (NT dollar)
* BILL\_AMT2: Amount of bill statement in August, 2005 (NT dollar)
* BILL\_AMT3: Amount of bill statement in July, 2005 (NT dollar)
* BILL\_AMT4: Amount of bill statement in June, 2005 (NT dollar)
* BILL\_AMT5: Amount of bill statement in May, 2005 (NT dollar)
* BILL\_AMT6: Amount of bill statement in April, 2005 (NT dollar)
* PAY\_AMT1: Amount of previous payment in September, 2005 (NT dollar)
* PAY\_AMT2: Amount of previous payment in August, 2005 (NT dollar)
* PAY\_AMT3: Amount of previous payment in July, 2005 (NT dollar)
* PAY\_AMT4: Amount of previous payment in June, 2005 (NT dollar)
* PAY\_AMT5: Amount of previous payment in May, 2005 (NT dollar)
* PAY\_AMT6: Amount of previous payment in April, 2005 (NT dollar)
* **Important:** default.payment.next.month: Default payment (1=yes, 0=no)

**Data Set:**

****

**PROBLEM FORMULATION:**

The goal of this project is to analyse data set and create a predictive model based on supervised learning. The given data set can create a model which is helpful to explain whether a **customer would default on his payment or not.** In this data set, we are choosing **“default.payment.next.month”** as a target variable and rest all variables are independent variables.

To achieve this goal, we have used two classification algorithms on the dataset which are:

* Naive Bayes
* Neural Networks

We compare the results to determine which model produces best and accurate results to the given problem.

**CLASSIFICATION MODELS:**

**(1) Naive Bayes Algorithm**

Naive Bayes is a classification technique based on Bayes’ Theorem with an assumption of independence among predictors. In simple terms, a Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature.

Naive Bayes model is easy to build and particularly useful for very large data sets. Along with simplicity, Naive Bayes is known to outperform even highly sophisticated classification methods.

Bayes theorem provides a way of calculating posterior probability P(A|B) from P(A), P(B) and P(B|A). Look at the equation below:

**P(A|B) = (P(B|A) \* P(A)) / P(B)**

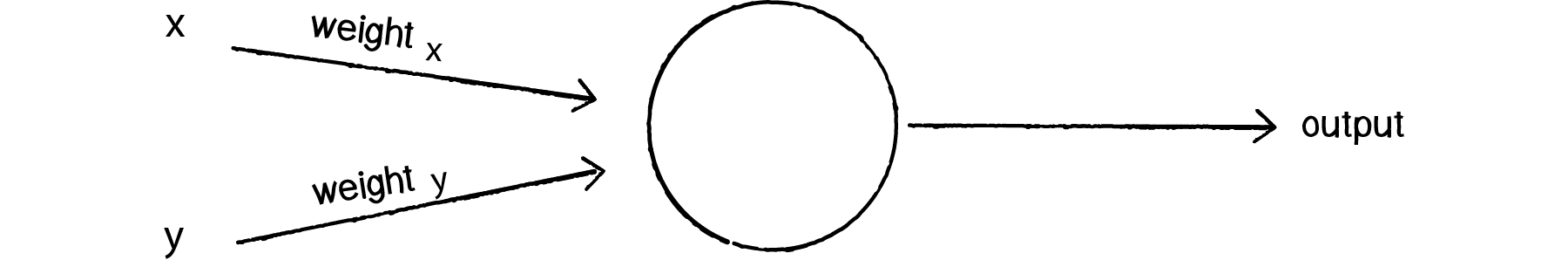
* P(A|B) is the posterior probability of target A given predictor B.
* P(A) is the prior probability of A.
* P(B|A) is the likelihood which is the probability of predictor given A.
* P(B) is the prior probability of predictor B.

**Why we choose Naive Bay:**

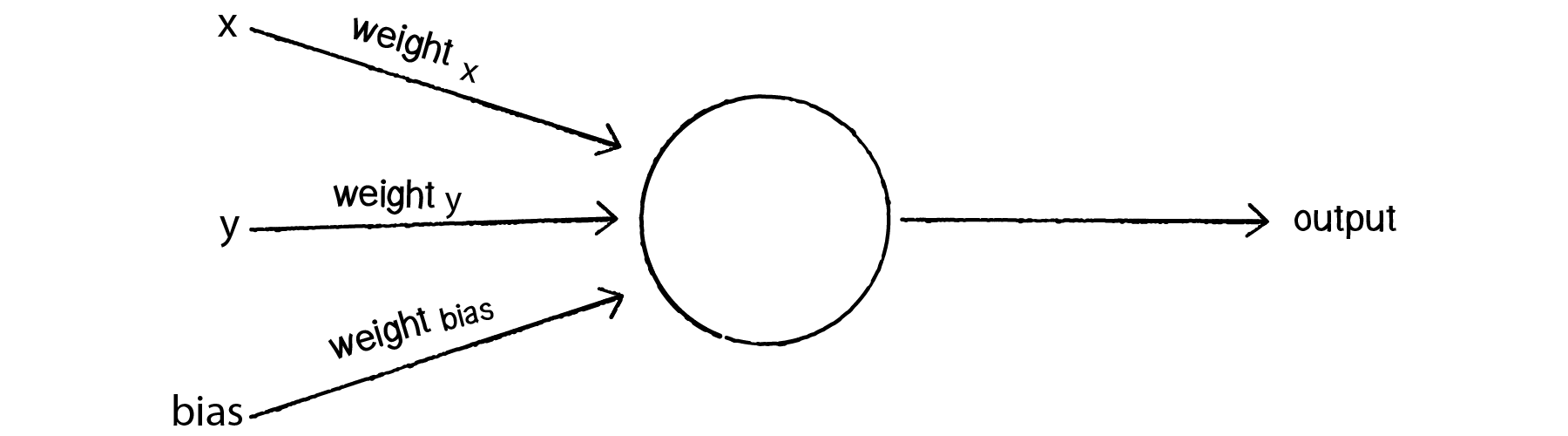
* It is simple, easy to implement and fast.
* It returns not only the prediction but also the degree of certainty, which can be very useful.
* It handles continuous and discrete data.
* It can be used for both binary and multi-class classification problems
* It’s not sensitive to irrelevant features.

**(2) Neural Network Algorithm**

Neural Network algorithm is an implementation of the popular and adaptable neural network architecture for machine learning. The algorithm works by testing each possible state of the input attribute against each possible state of the predictable attribute, and calculating probabilities for each combination based on the training data. You can use these probabilities for both classification or regression tasks, to predict an outcome based on some input attributes. A neural network can also be used for association analysis.



We can see how there are two inputs (x and y), a weight for each input (weightx and weighty), as well as a processing neuron that generates the output.

****

0 \* weight for x = 0

0 \* weight for y = 0

1 \* weight for bias = weight for bias

**Why we chose Neural Network:**

* A neural network model supports regression, association, and classification analysis, Therefore, the meaning of each prediction might be different.
* We can also query the model itself, to review the correlations that were found and retrieve related statistics.
* ANN is nonlinear model that is easy to use and understand compared to statistical methods.
* ANN with Back propagation (BP) learning algorithm is widely used in solving various classification and forecasting problems.

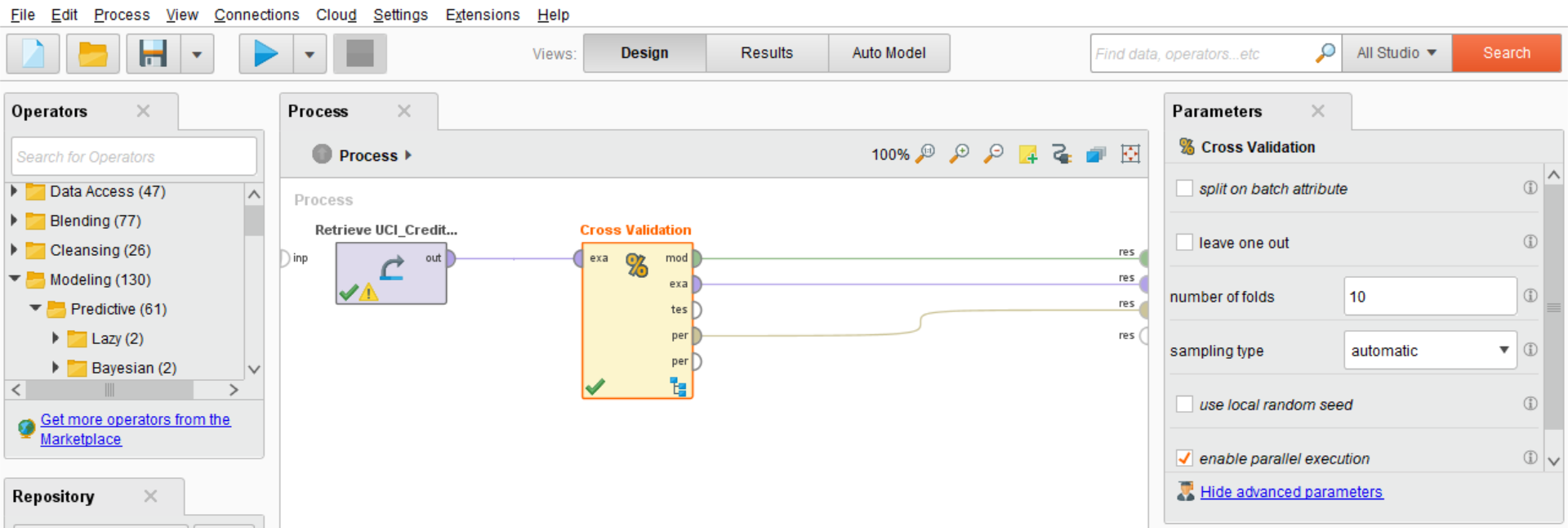
**EXPERIMENTAL RESULTS:**

Before using the dataset we formatted one of the columns **“Default.payment”** - Changed its data type to **“binomial”** and changed its role to “label” since it is the outcome variable.

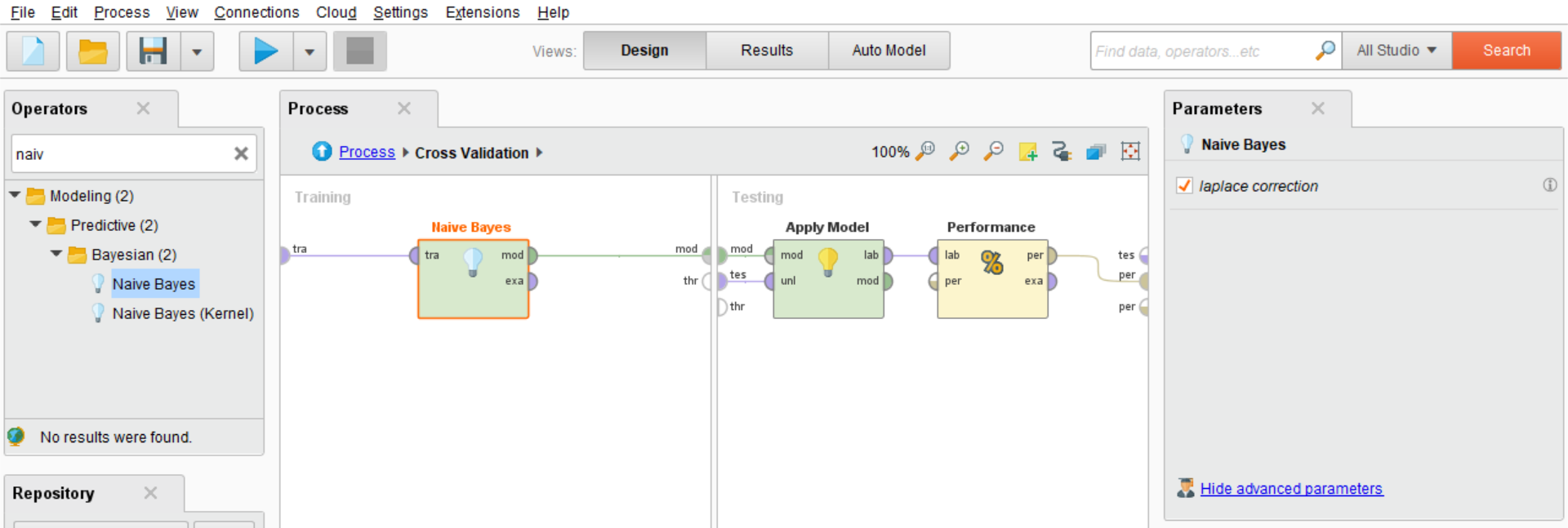
All the other variables were by default “integers” and no changes were necessary.

**Naïve Bayes**

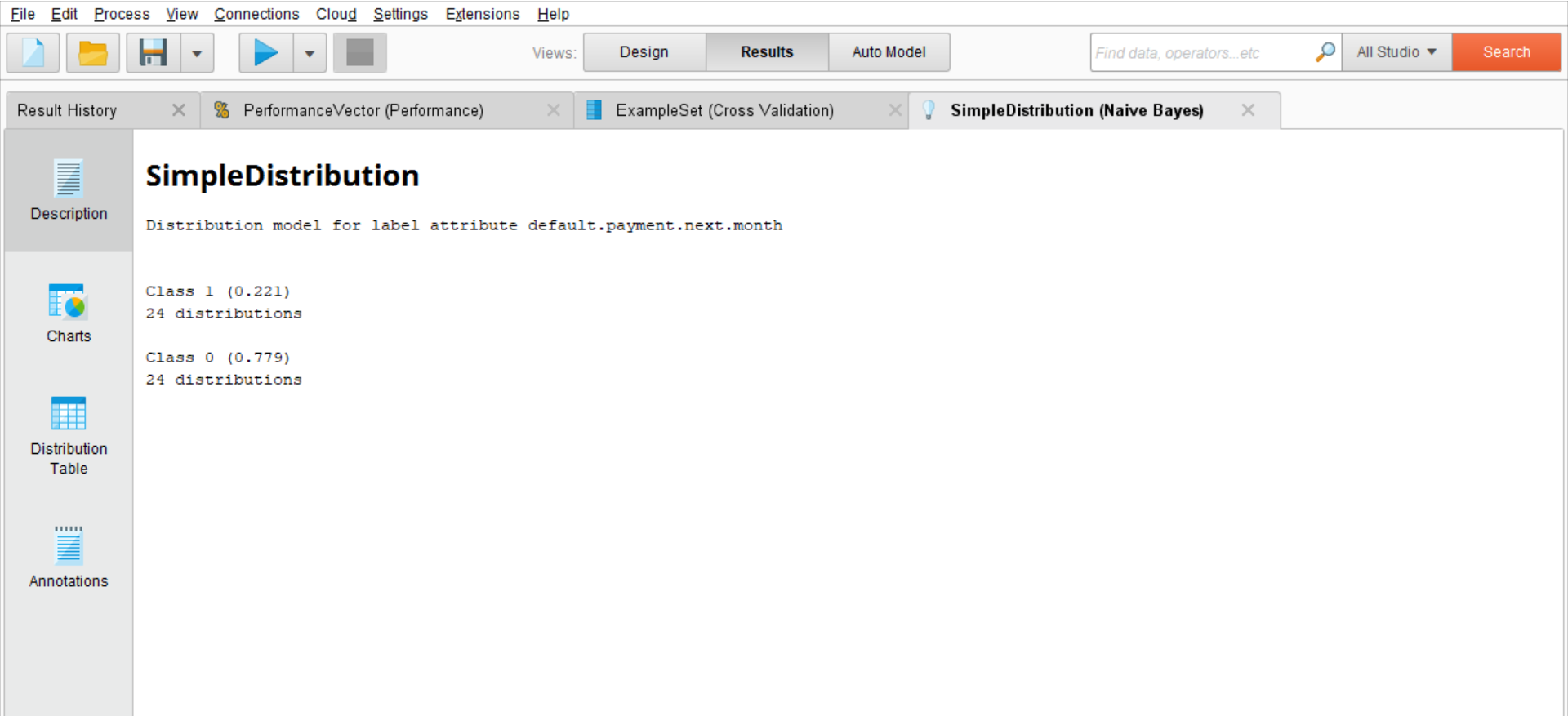
**Selecting Cross Validation operator after loading the dataset**

****

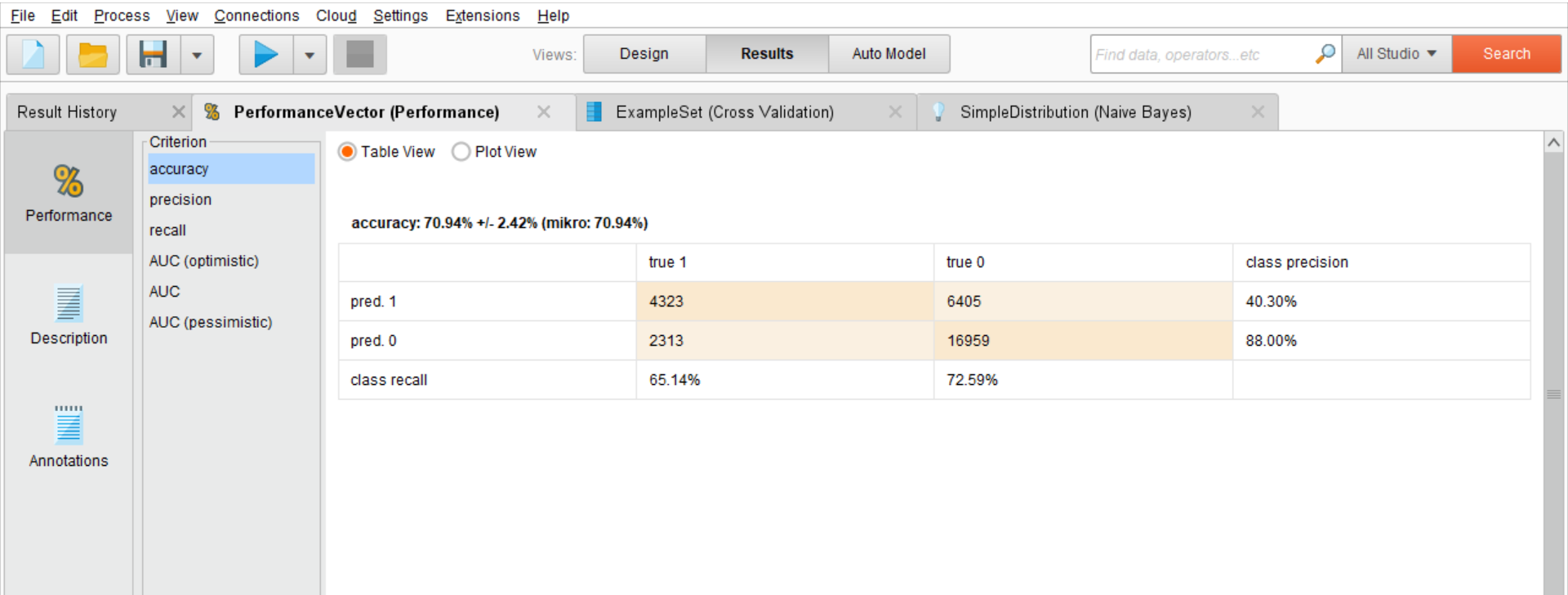
**Implementing Naive Bayes classifier**

****

**Simple Distribution**

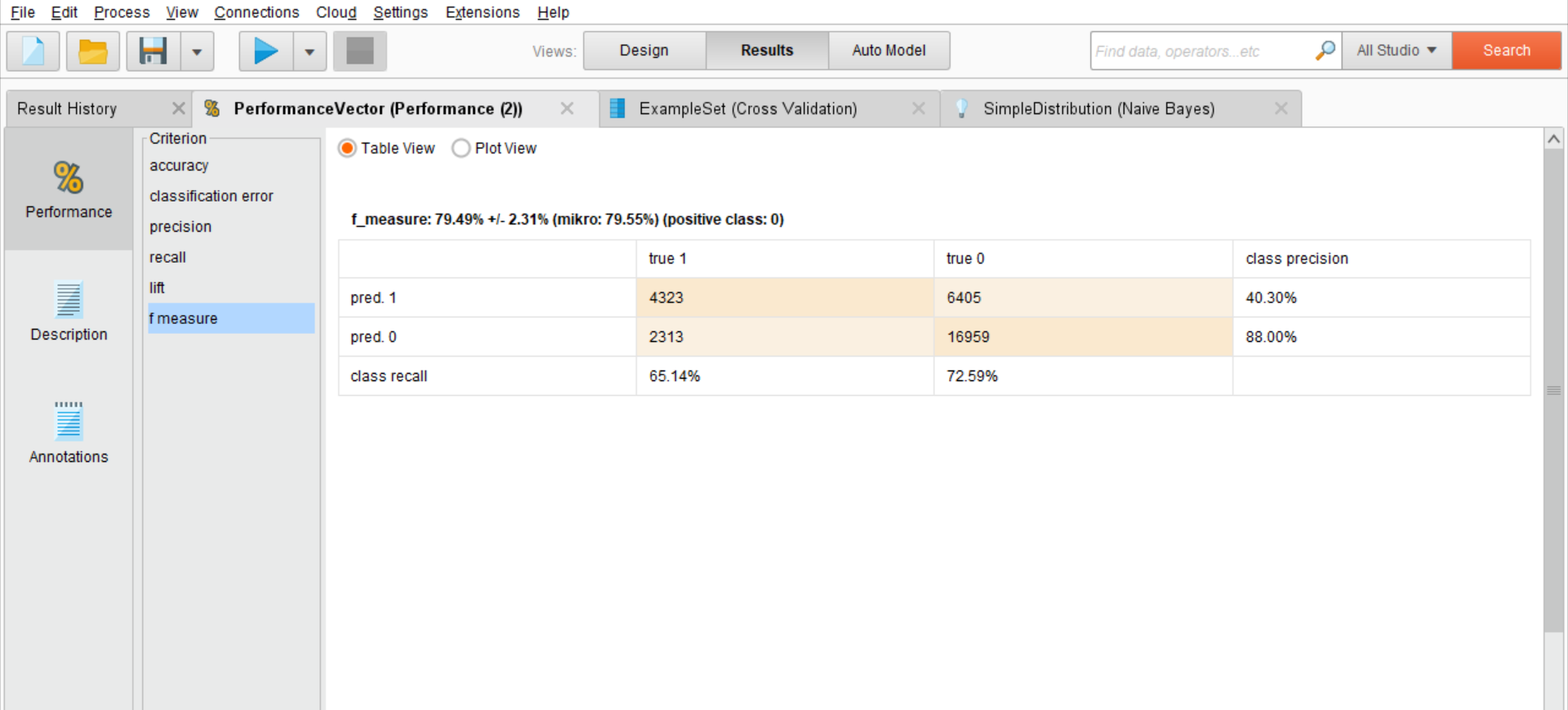
****

**Confusion Matrix with accuracy, precision and recall %**

****

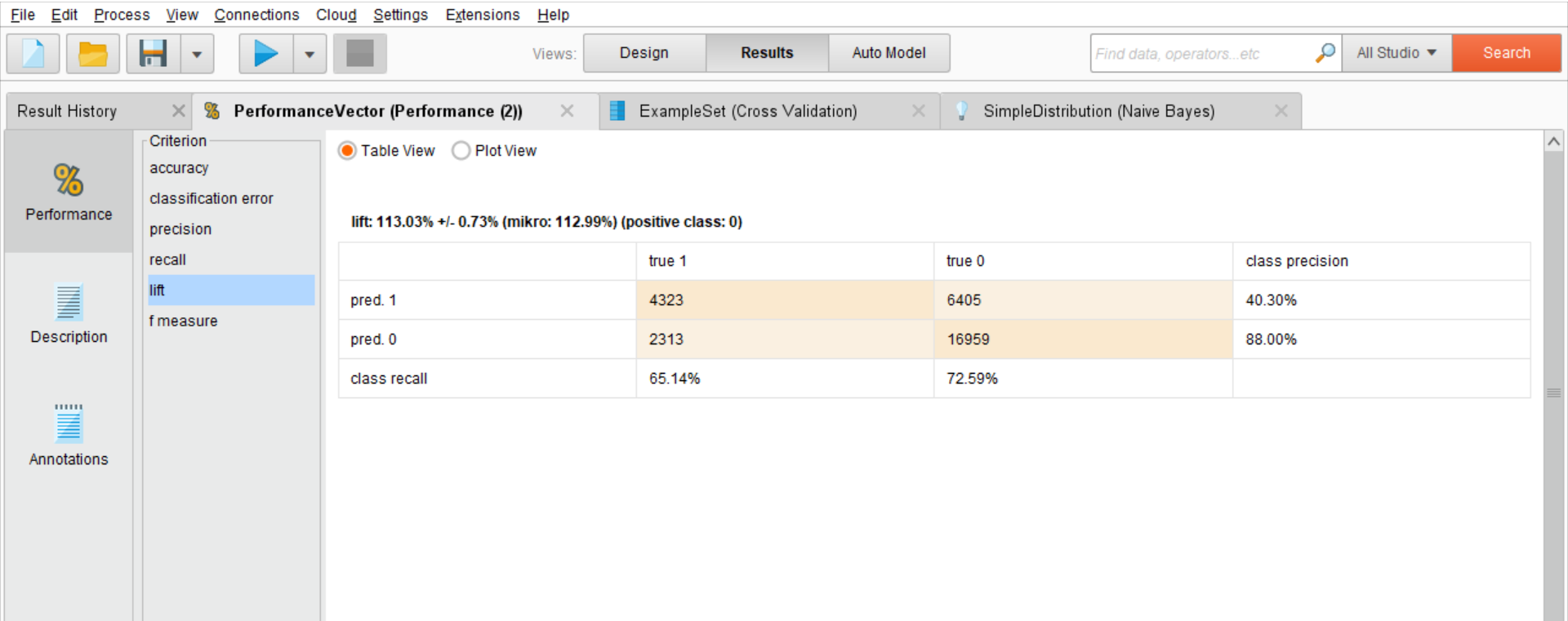
Accuracy = 70.94%, Precision = 88%, Recall = 72.59%

**F\_Measure:**

****

F\_measure = 79.49%

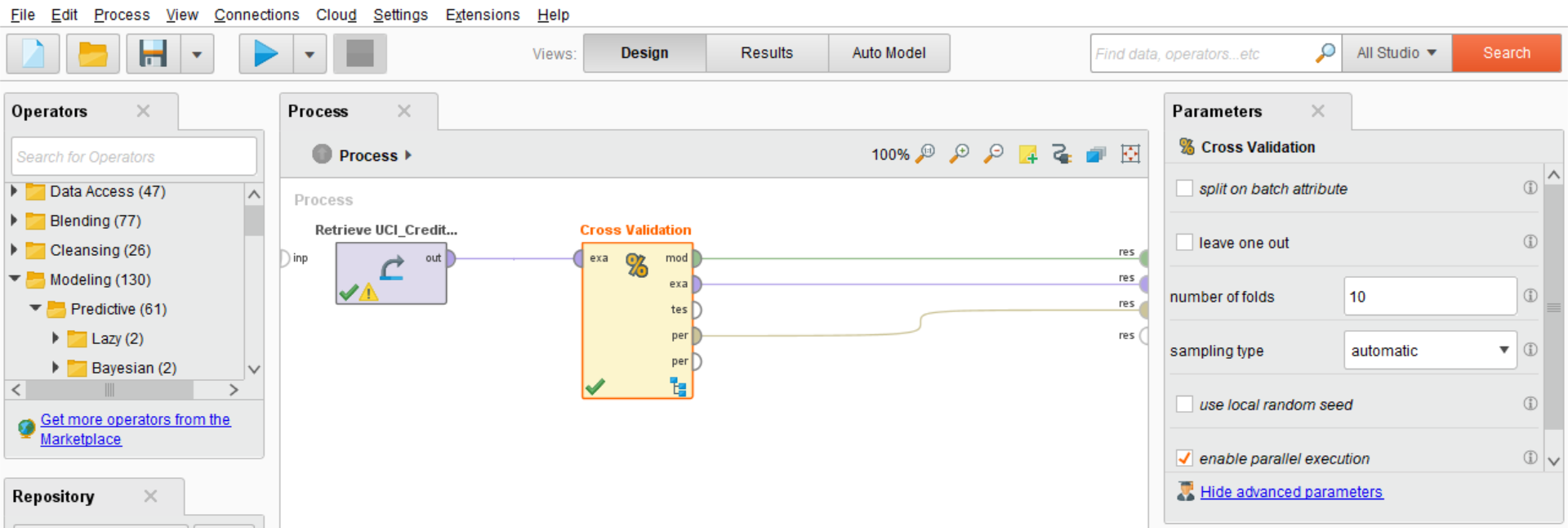
**Lift**

****

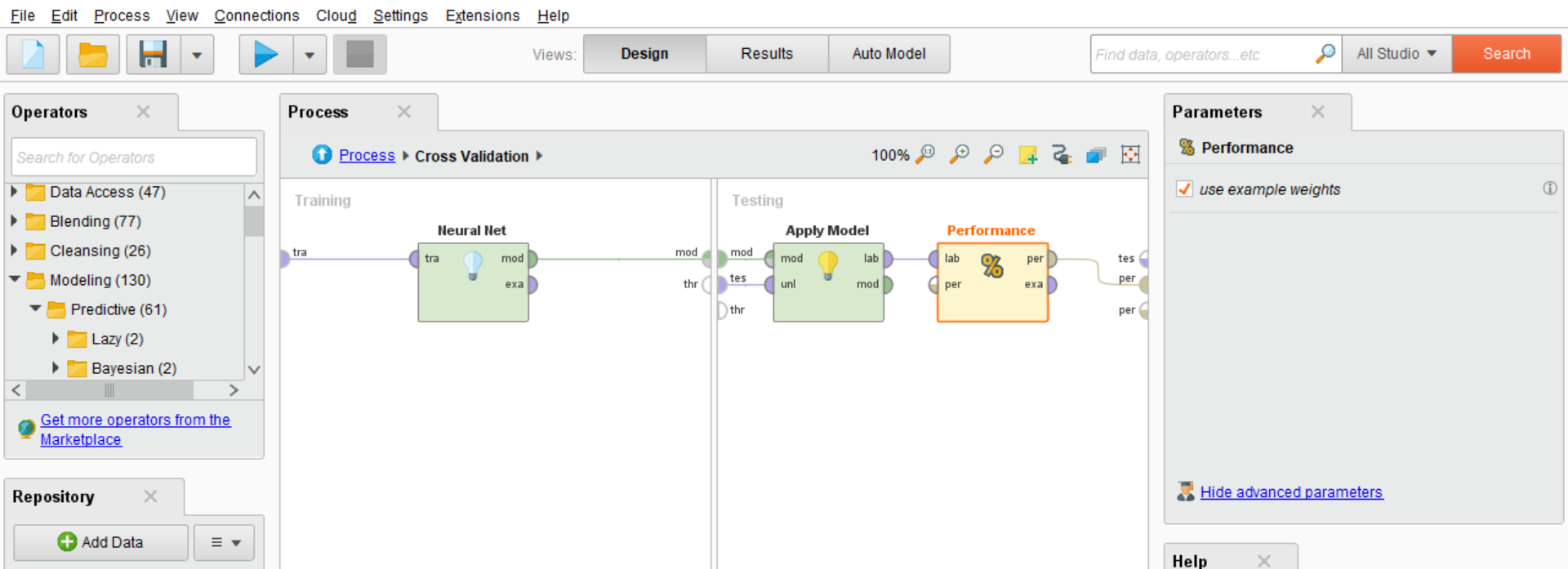
Lift = 113.03%

**Neural Network**

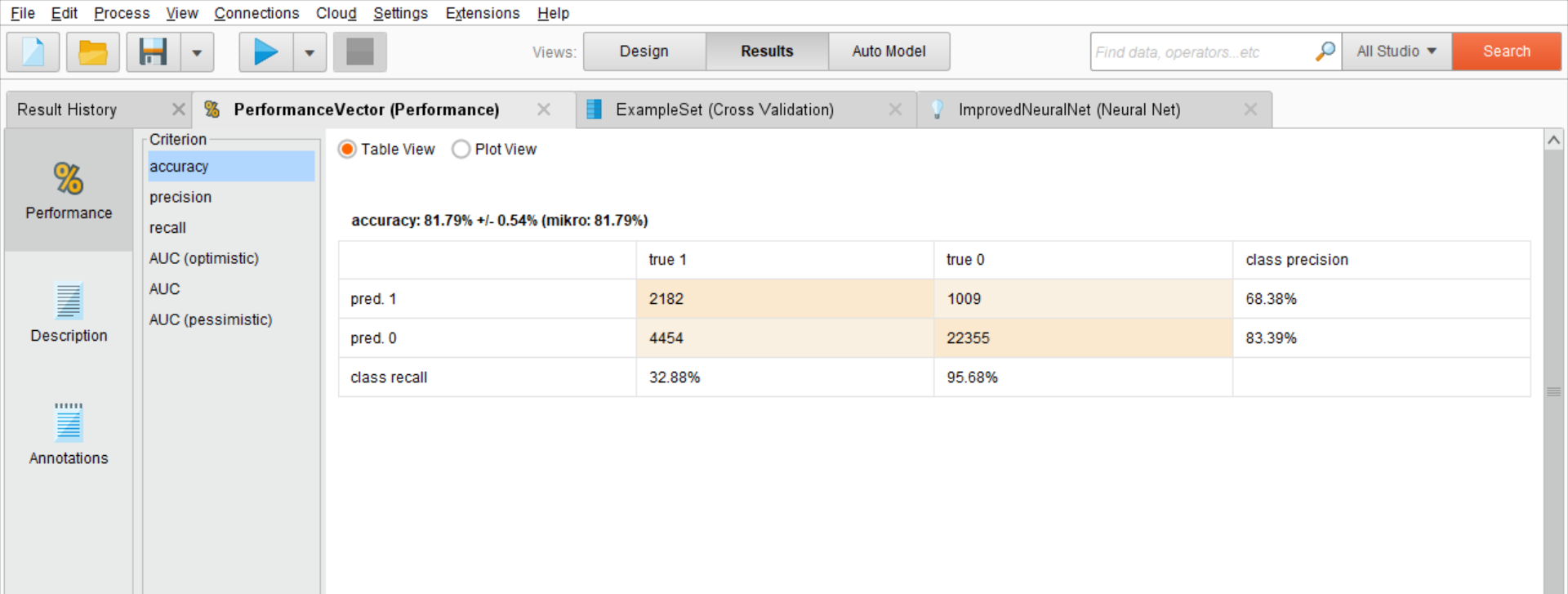
**Selecting Cross Validation operator after loading the dataset**

****

**Implementing Neural Network**

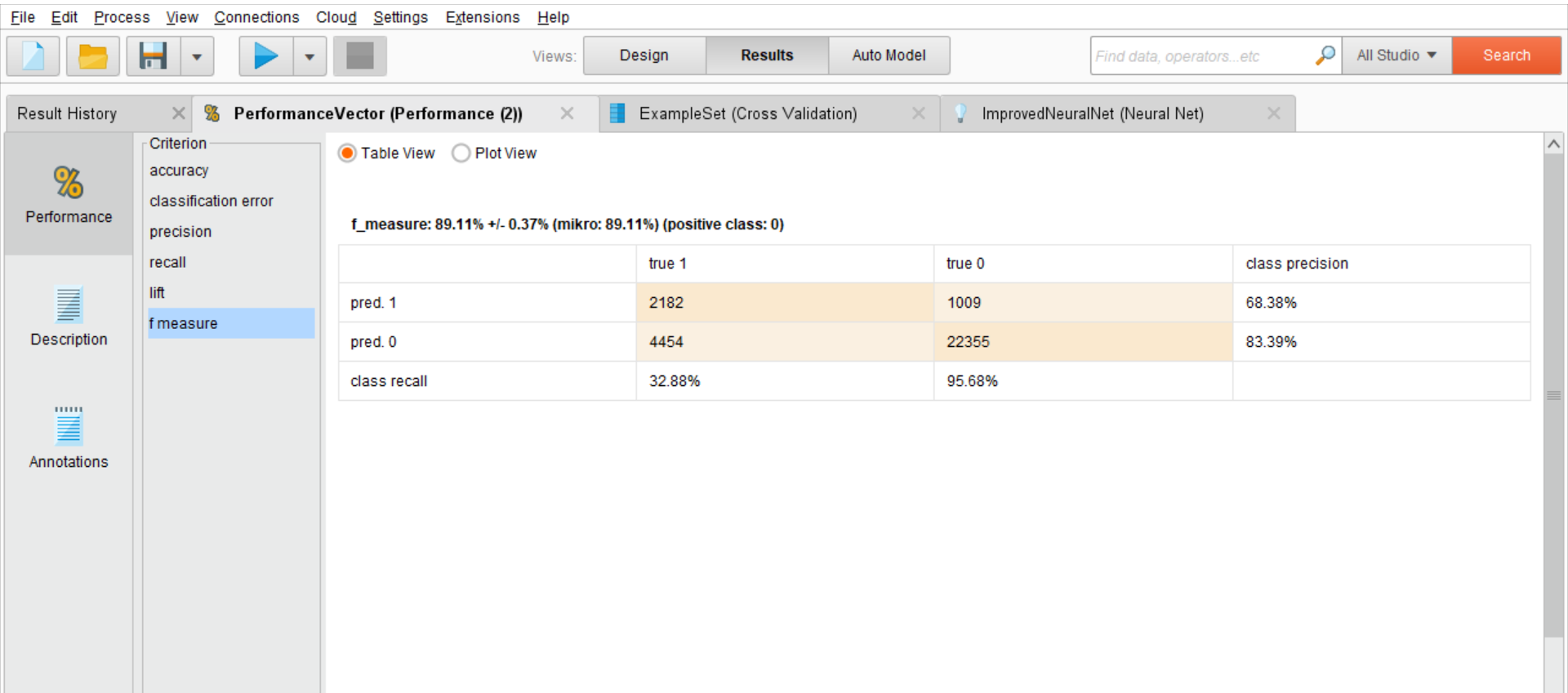
****

**Confusion Matrix with accuracy, precision and recall%**

****

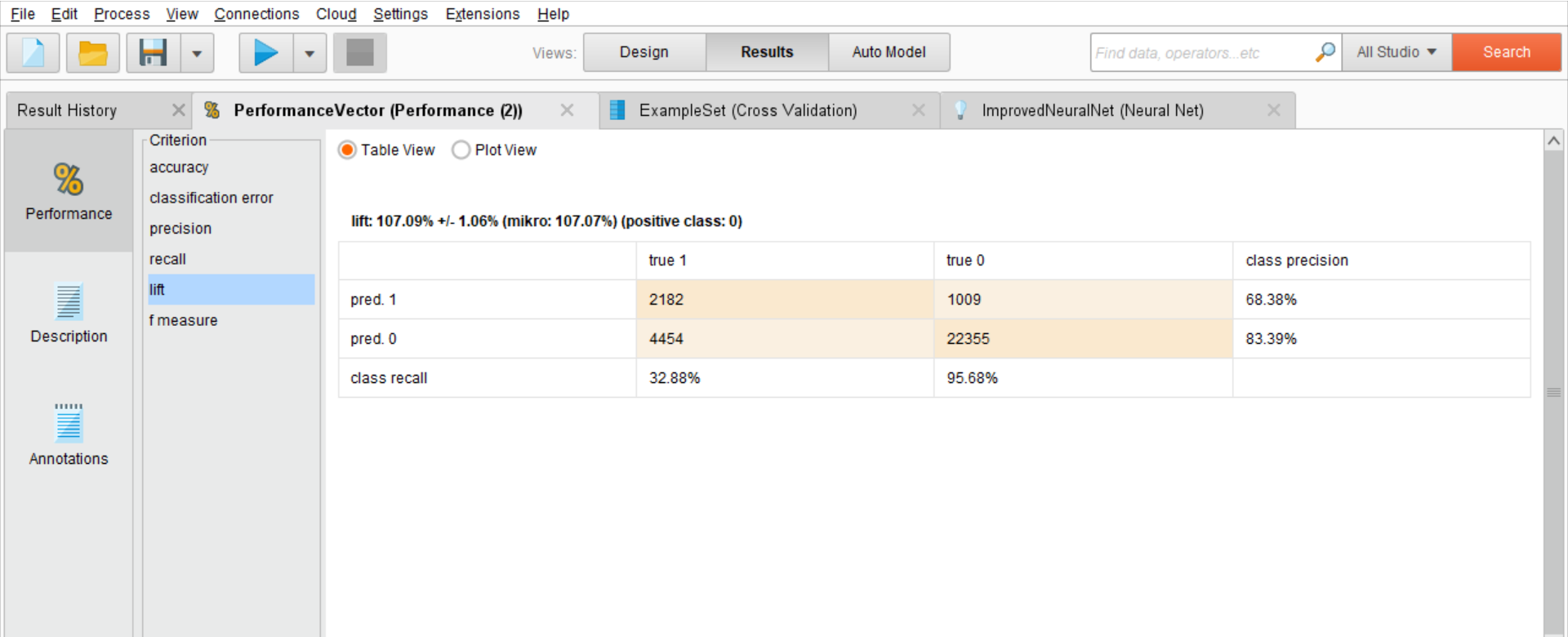
Accuracy = 81.79%, Precision = 83.39%, Recall = 95.68%

**F\_Measure**

****

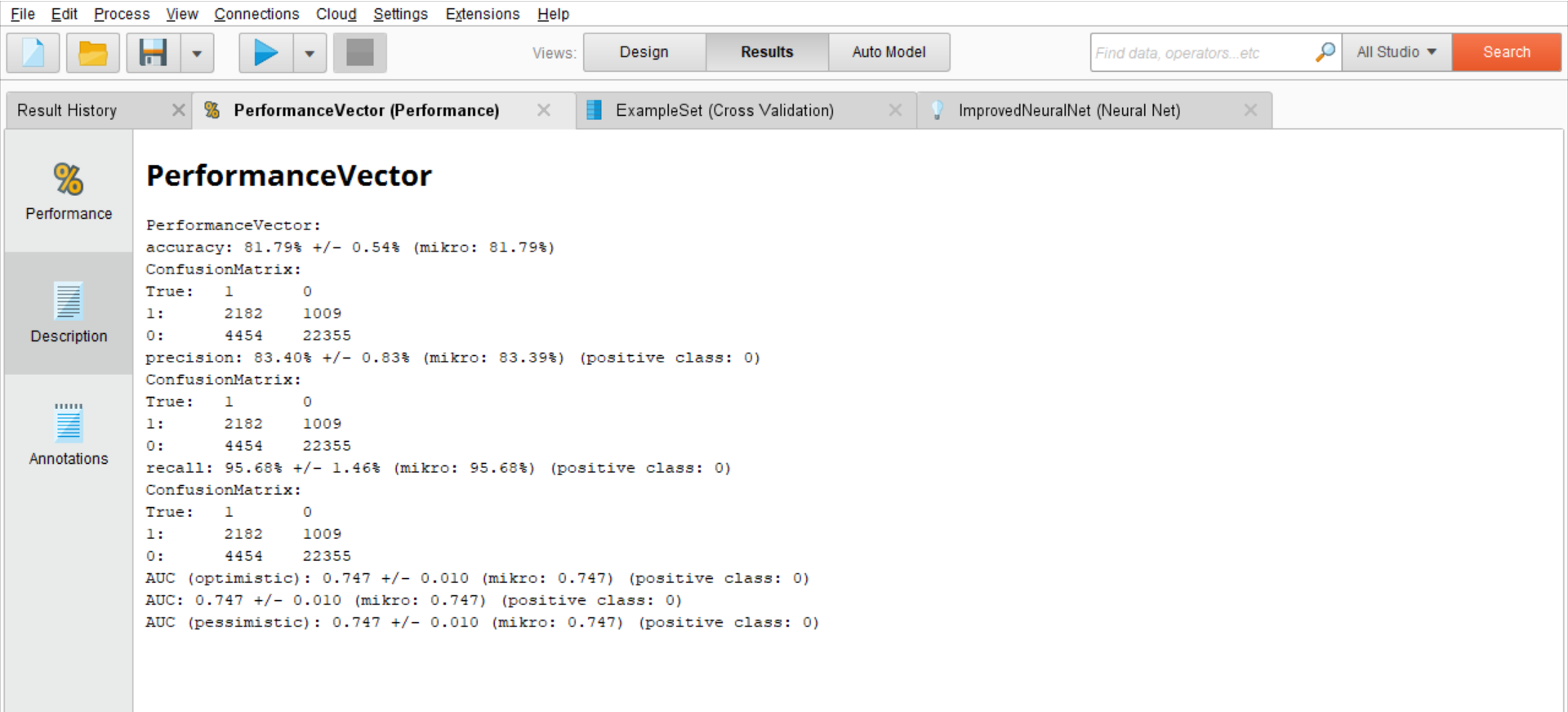
F\_measure = 89.11%

**Lift**

****

Lift = 107.09%

**Performance Vector:**

****

**Conclusion:**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **MODEL** | **ACCURACY** | **PRECISION** | **RECALL** | **F\_MEASURE** | **LIFT** |
| **Neural Net** | 81.79% | 83.39% | 95.68% | 89.11% | 107.09% |
| **Naïve Bayes** | 70.94% | 88.00% | 72.59% | 79.49% | 113.03% |

Application of these algorithms on full scale implementation is cost effective, less error-prone and less time consuming. Application of data mining algorithms in banking sector is gaining momentum. It requires more research and exploration. It has high potential to cater to diversifying needs in the banking sector.Naive Bayes and Neural Networks classification algorithms have classified the dataset to good accuracy thus achieving the objective of this study.

In this project we were tasked to train and test Supervised learning models for Credit Risk prediction, the dataset consisted of all the clients transactional and demographic information(Independent features) and defaulter column(target/dependent feature).

Since we have an Unbalanced dataset, we focused on Precision,Recall and F-measure(to avoid the problem of Accuracy Paradox), Neural nets have high scores for both metrics(above 80%),which makes it the optimal classier for credit risk modeling as it classifies relevant defaulters really well with respect to the data points. Though the Precision and Lift is higher for Naive Bayes, it is irrelevant to predict the data model as they are not much higher and rest measures like Accuracy, Recall and F-measure are not even 80% to make a relevant prediction.

Additionally,to reduce the effect of unbalanced class labels,we have used cross validated training for all models and still have achieved good average scores of all metrics for Neural Nets,this justifies the robustness of our chosen model in predicting Credit risk with respect to a clients information.

**Side Note:**Even though Neural Net may seem like an optimal model for this dataset, in real practice where we have millions of rows,Neural Nets take a lot of time to train, and its due to this very reason along with simplicity of implementation ,we use normal supervised learners like Naive Bayes or Log Regression.