User manual

Made by : ROBLIN Caroline, TELLIER Kevin, YE Maxime

08 novembre 2019

The purpose of our demonstrator is to compare the performance of the SVM machine learning method with other machine learning methods (Logistic Regression, Decision Tree, Random Forest, Gradient Boosting, Xgboost).

Note : the loading of the application depends on the network connection.
Depending on the network speed, the application may take further time to load.

Explanations of the methodology

Data presentation

Comparison of SVM with another model

Inputs

Model to compare with svm

- Logistic Regression
- Decision Tree
- Random Forest
- Gradient Boosting
- XGBoost

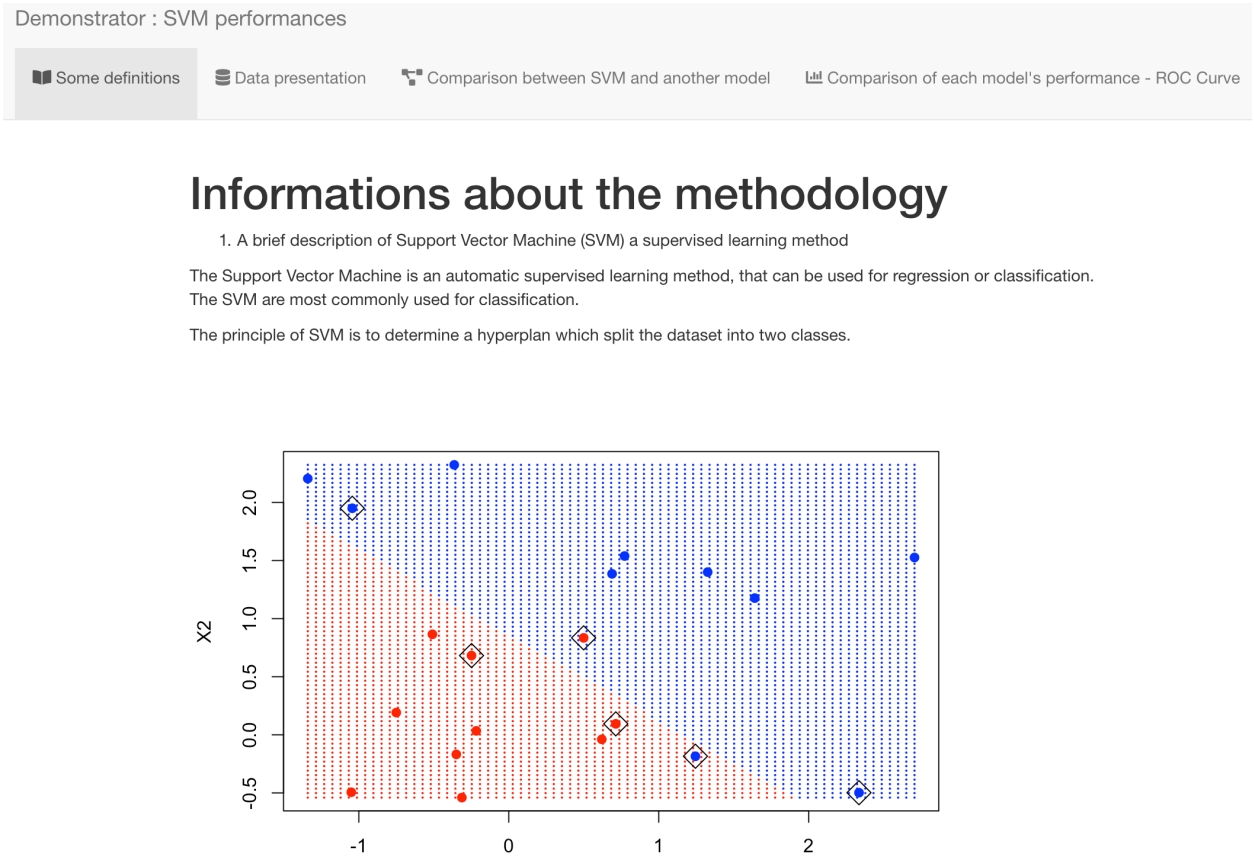

Figure 1: 1st panel : Brief description of SVM and methodology

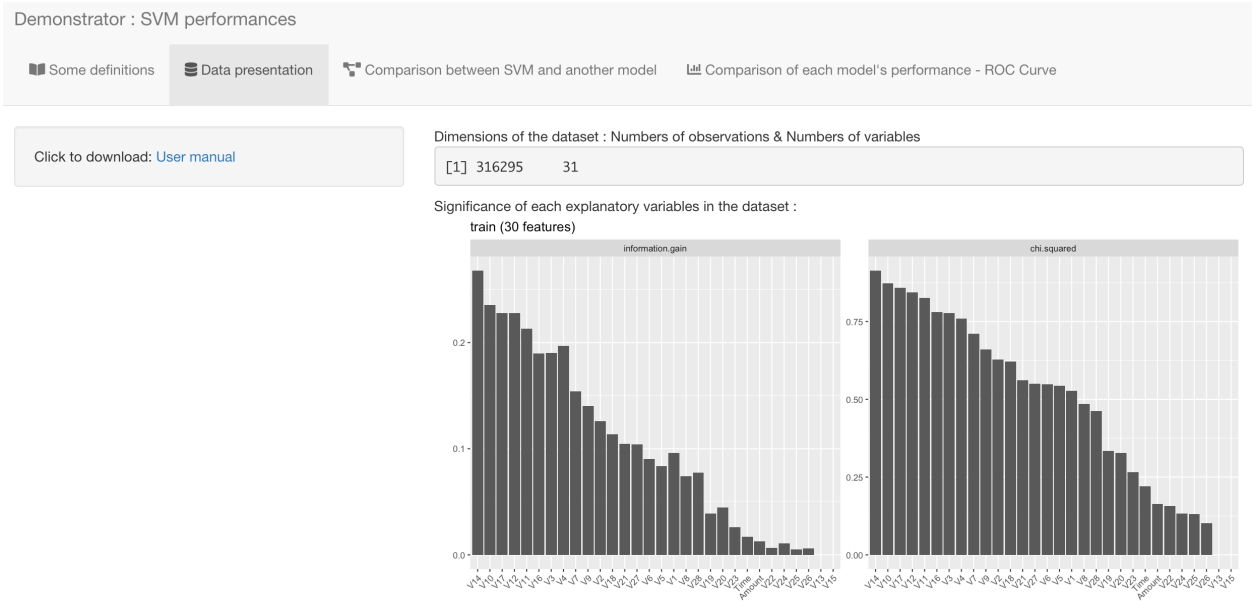

Figure 2: 2nd panel : Data presentation

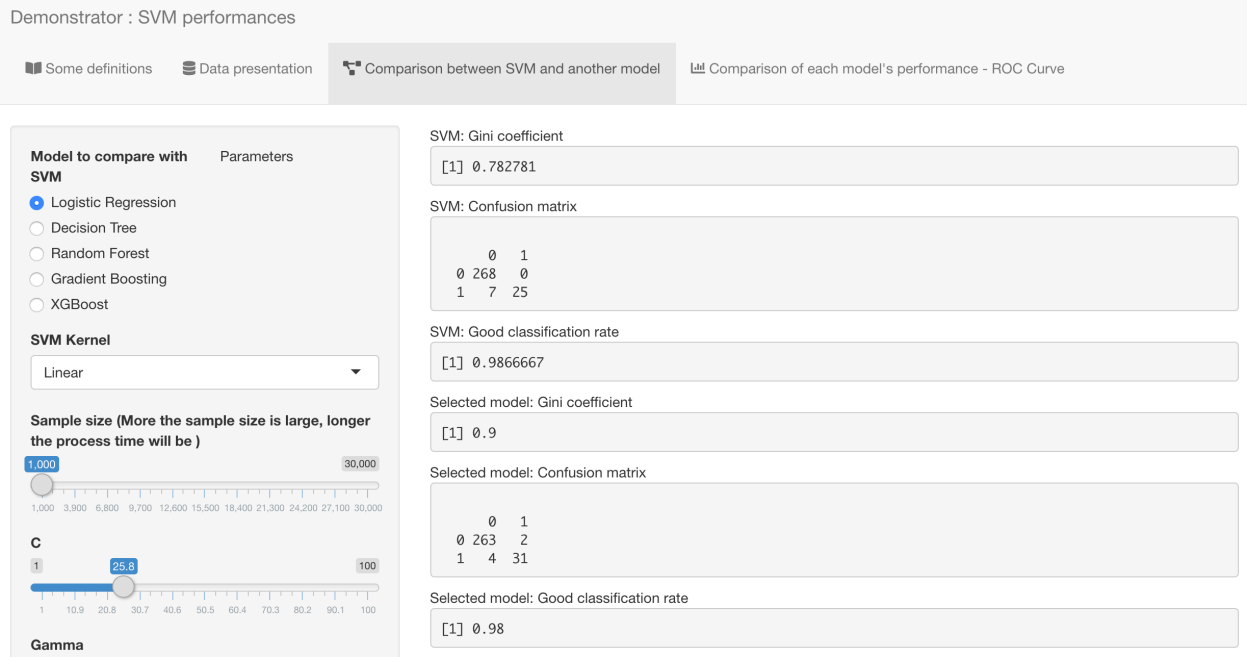

Figure 3: 3rd panel : Comparing SVM with another Machine Learning model

- for the Decision Tree :
  Minsplit : represents the minimum number of observations in a node for a split to take place (35)
  Minbucket : says the minimum number of observations I should keep in terminal nodes (10)
  Cp : it's the complexity parameter (0.167)
- for the Random Forest :
  Number of trees (108)
  Node Size (11)
  Mtry 11
- for the Gradient Boosting :
  N trees (414)
  interaction depth (7)
  Min obs in node : refers to the minimum number of observations in a tree node (17)
  shrinkage : it's the regulation parameter which dictates how fast / slow the algorithm should move (0.268).
- for the XGBoost :
  Nround (481)
  Max depth (16)
  Lambda (0.563)
  Eta (0.183)
  Sub sample (0.328)
  Min child weight (1.83)
  Cold sample by tree (0.41)

SVM Kernel

- Linear
- Polynomial
- Radial Basis
- Sigmoid

Kernel

C

Sample size (the larger the size chosen, the longer the processing time will be)

Note that the sample size chosen has been split into two samples, with 70% of the data for the train dataset, and 30% for the test dataset. The performances displayed are based on the test sample.

Outputs

- SVM : Gini coefficient
- SVM : confusion matrix
- SVM : good classification rate
- Selected model : Gini coefficient
- Selected model : Confusion matrix
- Selected model : good classification rate
- ROC Curve comparison between the SVM and the selected model

Comparison of each model's performance - ROC Curve

Inputs

Sample size (the larger the size chosen,the longer the processing time will be)

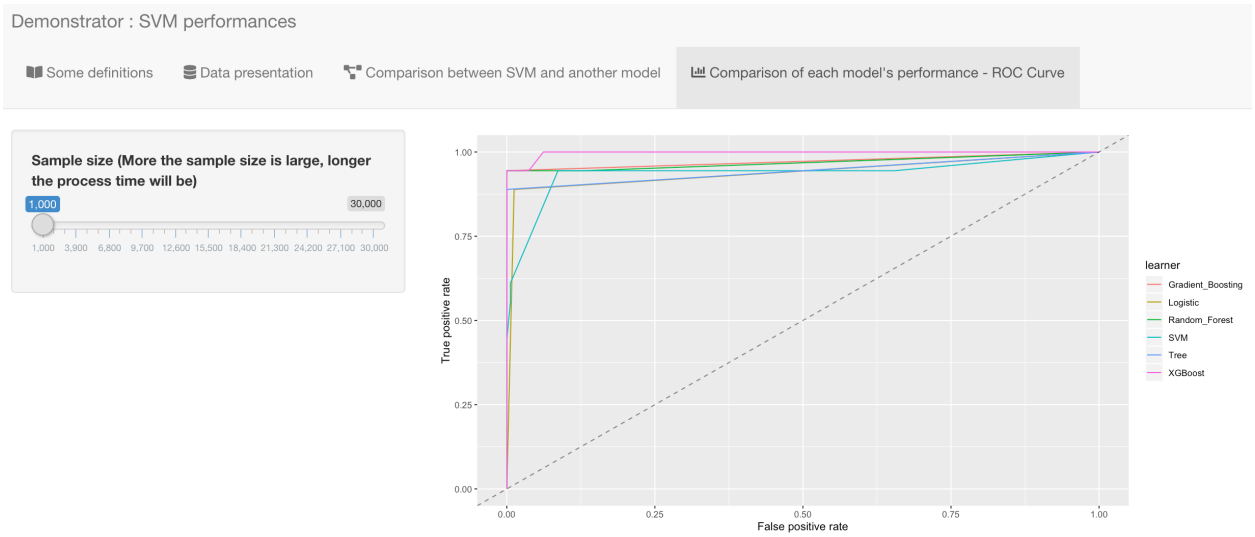

Figure 4: 4th panel : Comparing model performance

Outputs

You can see that it isn't easy to choose the best model regardless of the sample size because of the crossing. It's better to refer to the Gini index or the good classification rate we have in the previous tab.