Technical Manual

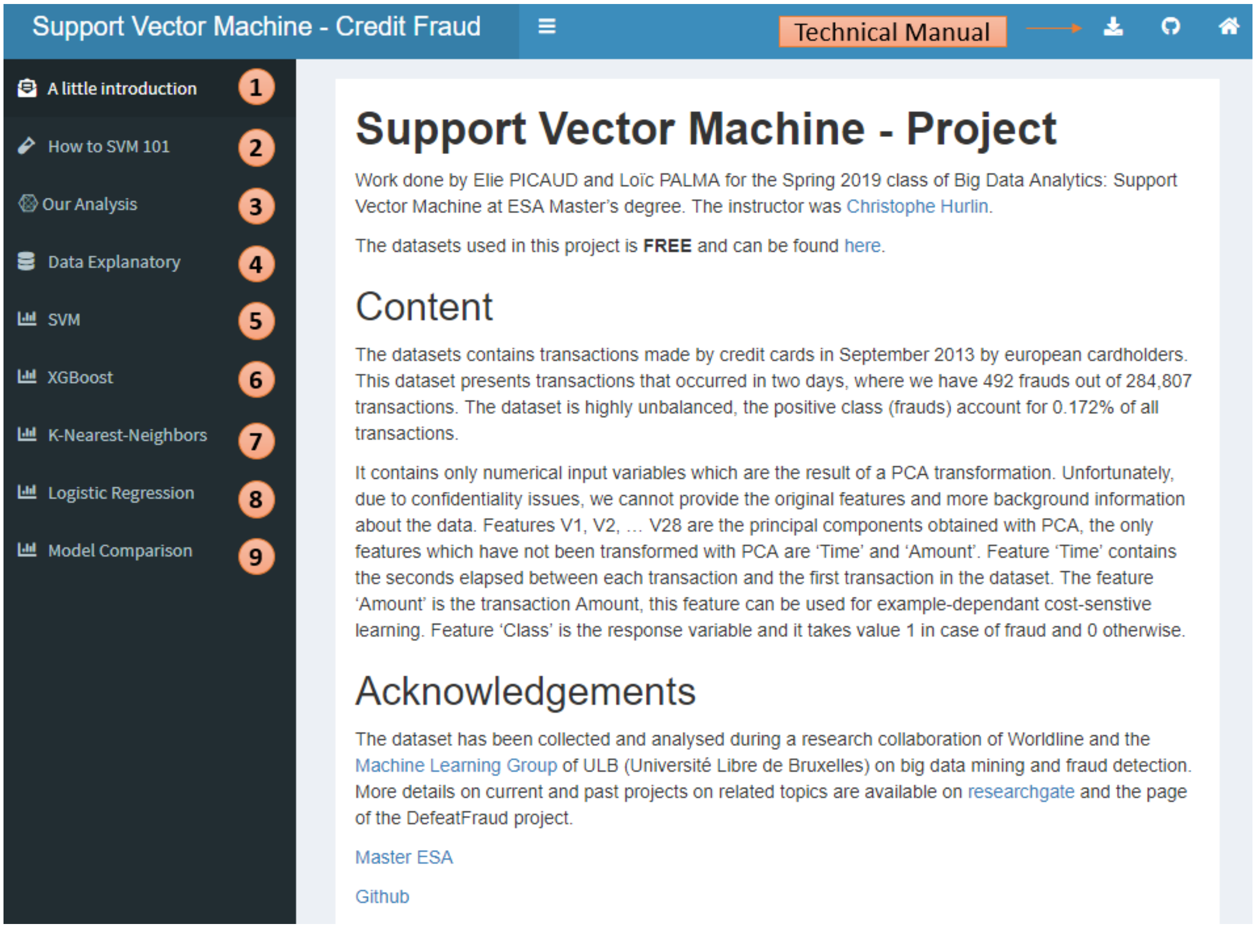
Technical Documentation

Welcome to the Technical Manual of the Support Vector Machine Dashboard.
If you are there, that is most likely because you have clicked the download button in the top-right hand corner of the dashboard's header.

The objective of this application is to propose to the customer a site allowing to understand the implementation of the SVM and to visualize online the performances of this solution by having the possibility to choose certain hyper-parameters or methods of selecting these hyper-parameters.

If you want to get in touch with R Shiny or you just want to know more about it, please have a look at the following link.
If you want to know more about Support Vector Machine without going further into the maths behind it, we strongly recommend An Idiot's guide to Support Vector Machines (SVMs) from R. Berwick[1].
If you want a more in-depth introduction to Support Vector Machine, consider reading Support Vector Machine by C. Hurlin[2].

Here is an overview of what the dashboard (should) look like:



Support Vector Machine - Project

Work done by Elie PICAUD and Loïc PALMA for the Spring 2019 class of Big Data Analytics: Support Vector Machine at ESA Master's degree. The instructor was Christophe Hurlin.

The datasets used in this project is **FREE** and can be found here.

Content

The datasets contains transactions made by credit cards in September 2013 by european cardholders. This dataset presents transactions that occurred in two days, where we have 492 frauds out of 284,807 transactions. The dataset is highly unbalanced, the positive class (frauds) account for 0.172% of all transactions.

It contains only numerical input variables which are the result of a PCA transformation. Unfortunately, due to confidentiality issues, we cannot provide the original features and more background information about the data. Features V1, V2, ... V28 are the principal components obtained with PCA, the only features which have not been transformed with PCA are 'Time' and 'Amount'. Feature 'Time' contains the seconds elapsed between each transaction and the first transaction in the dataset. The feature 'Amount' is the transaction Amount, this feature can be used for example-dependant cost-senstive learning. Feature 'Class' is the response variable and it takes value 1 in case of fraud and 0 otherwise.

Acknowledgements

The dataset has been collected and analysed during a research collaboration of Worldline and the Machine Learning Group of ULB (Université Libre de Bruxelles) on big data mining and fraud detection. More details on current and past projects on related topics are available on researchgate and the page of the DefeatFraud project.

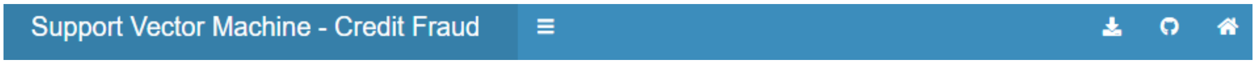
Master ESA

Github

[1] Professor, Computational Linguistics, MIT.
[2] Professor, Economics, University of Orleans.

Basically, there is 9 tabs in our application:

1. **"A little Introduction"** : Here you can find information about the authors as well as where to find the datasets in case you want to try it.
2. **"How to SVM 101"** : Here is a little introduction about Support Vector Machine, explaining briefly what is a SVM, the different types of SVM, the different types of kernel, what is the "kernel trick" as well as the different kernels that exists.
3. **"Our Analysis"** : Here is what we have done on our side. We explained the basics (recall, precision, confusion matrix, ROC and so on). We tried to fit the best models for each algorithms that we implemented, showing the results and briefly conclude with a model comparison. We also discuss the limits of doing Random Under Sampling on our datasets by doing a comparison with algorithms that are being trained on the whole training set.
4. **"Data Explanatory"** : Here you can play around with the data. The first box is meant to show a few observations with all the variables so you can have a look at what the variables looks like. Then on the other boxes, you can have some simple and basics informations about a variable. At the end, you get the correlation matrix between all the variables.
5. **"SVM"** : Here is the first tab where you can fit a model to our training set. For instance, you can fit a SVM model by choosing all the parameters and hyperparameters that are explained in the "How to SVM 101" tab.
6. **"XGBoost"** : Here you can fit a XGBoost model. If you want more details about XGBoost please follow this link. It is worth mentionning that XGBoost is widely used and is a really popular tool among Kaggle competitors and Data Scientists in industry. More details about the parameters/hyperparameters can be found here.
7. **"K-Nearest-Neighbors"** : Here you can fit a KNN model. If you want more details about what KNN is, please have a look at this link.
8. **"Logistic Regression"** : Here you can fit a Logistic Regression. Feel free to choose the variables you want the logit to take consideration of. Please be aware that you basically need to **always** have the target variable (Class) selected. More information about the Logistic Regression here.
9. **"Model Comparison"** : Here you have the summary of all the ROCs and PROCs of the models you have trained. Please be aware that in order this tab to work, you need to have trained all 4 type of models (SVM, XGBoost, KNN and Logistic Regression).



Apart from the download button, you can find a link to our GitHub repository by clicking on the GitHub icon in the top-right hand corner of the dashboard's header.
You will also find a home button. This one will bring you back to the introduction tab.

NOTA BENE : We are using a free version of Shinyapps.io. That is to say the application might be sensible and could **crash**. This comes from the fact that we are loading huge datasets as well as fitting huge memory-demanding models. We are aware of this issue and would like to apologize.