Deloitte.

**Risk Advisory**
August 2019

# Point of View:
## Using Random Forest for credit risk models

### Machine learning and Credit Risk: a suitable marriage?

Since the financial crisis, regulators have put an important focus on risk management supervision and expect banks to have transparent, auditable risk measurement frameworks dependent on portfolio characteristics for regulatory, financial or business decision-making purposes. Quantitative modelling techniques are used to get better insights from data, reduce cost and increase overall profitability.

In this disruptive era of Big Data and Artificial Intelligence, banks are considering the adoption of evolving technological capabilities whilst arbitrating between heightened regulatory demands and business objectives.

Banks have been focusing on revisiting risk management frameworks with advanced analytics to develop end-to-end approaches that combine data management, innovative technological solutions and analytics platforms to foster a global, 360-degree client view.

The use of Machine Learning methods faces skepticism for credit risk model development – notably for regulatory purposes, because of the lack of transparency and the percieved "black box" effect of these techniques.
Nevertheless, Machine Learning systems can help model developers to reduce model risk and improve general model predictive power.

A powerful benefit of these techniques is that they may allow model developers to significantly reduce the time spent on data management and data pre-processing steps before the development of the actual model.

As a first step toward credit risk modelling with Machine Learning algorithms, Machine Learning techniques can be explored to reduce the time spent on data management, or to get genuine insights on data at hand.

Model aggregation methods are good at selecting important variables and handling data quality issues.

Innovation in credit risk modelling is a complex challenge in an environment of evolving regulatory expectations.

## Model aggregation methods, what are they?

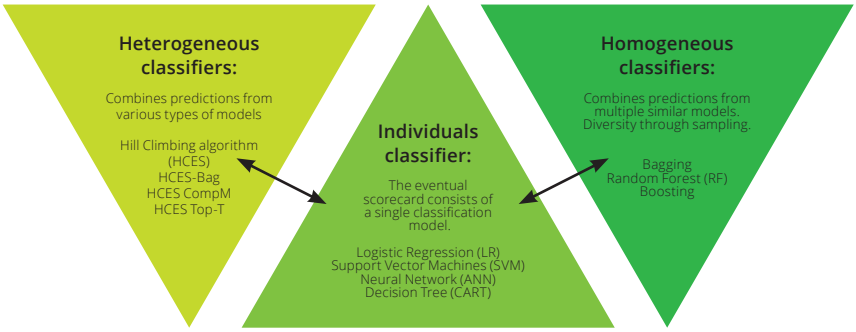



Figure 1: A panorama of identified algorithms to train models

**Model aggregation methods** or **ensemble learning** are algorithms that work on a "divide and conquer" approach in order to improve prediction performance. The principle is to combine several statistical models and then classify new inputs by taking a weighted vote of the predictions made by the models to improve overall accuracy.

These ensemble learning models combine learning algorithms such as classification and regression trees (CART), neural network (ANN), support vector machine (SVM) and many others.

Ensemble learning algorithms can be divided into three types: **individual**, **homogeneous** and **heterogeneous** methods.

The homogeneous methods are built on combining the same model several times while the heterogeneous methods combine different learning algorithms.

Homogeneous methods are built on the aggregation of the same learning algorithm. Repeating these combinations focuses on increasing robustness to variance and reducing the overall sensibility to model parameters and noise.

Two strategies can be used to train Machine Learning aggregation methods. On the one hand, **sequential ensemble** methods use learning models that are generated sequentially. The purpose is to exploit the dependence among the base models.
On the other hand, **parallel ensemble** methods use randomly generated learning models. The model gains diversity through sampling.

The **Random Forest** model is a particular case of a "bagging" (bootstrap aggregating) classifier applied to classification and regression trees (CART) with an additional randomized process regarding the set of explanatory variables.

To illustrate the differences between models, bagging (bootstrap aggregating) and boosting are the two most popular homogeneous models. The former is a parallel ensemble while the latter is a sequential ensemble.

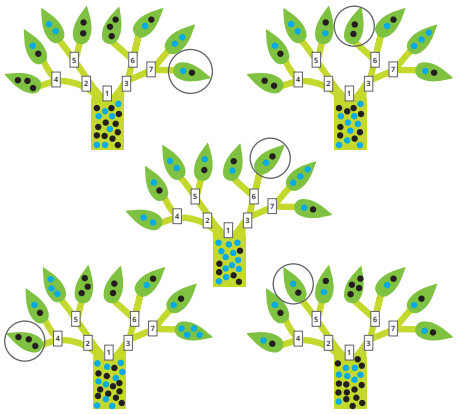
## Random Forest: how does it work?

The Random Forest algorithm is based on the construction of a myriad of different decision trees, creating a "forest" that is then aggregated. The diversity of these trees comes from two aspects of the manner of constructing the forest.

Firstly, each tree is built on a random sample of the observations, according to the bagging method.

Secondly, a random set of features is chosen to split nodes for each tree of the forest (feature sampling).

Finally, in order to use the model for prediction, the trees are aggregated. This is accomplished by averaging the results when the outcome is numerical, and by conducting a plurality vote when predicting a class variable.



The Random Forest algorithm is based on the construction of a myriad of different decision trees, creating a forest.

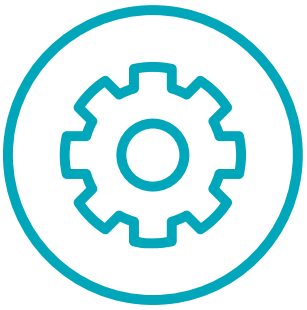
## Random Forest: which properties?

Random Forest was tested on different datasets (different populations with different characteristics) and the general properties observed for this type of aggregation method are indicated below:

**Random Forest characteristics**

| | |
|---|---|
| **Pros** | • Limits overfitting<br>• High accuracy (pruning is not used)<br>• Easy choice of relevant variables<br>• Stable against the data (good handling of missing values)<br>• Can handle a high dimensional set of variable |
| **Cons** | • Low interpretability<br>• Parameter choice (number of trees, proportion of the observation in each sample, depth, etc.)<br>• Computation time |

Table 1: Pros and Cons of the Random Forest algorithm applied to credit risk



**The following insights are examples of Random Forest algorithm usage that prove its utility in the field of credit risk. Random Forest can be used to save time on data management steps and to identify the most important characteristics in a dataset.**

## Insight 1 Random Forest is capable of identifying important features

As with many machine learning algorithms, the Random Forest model can be seen as a black box: it is not possible to define the process completely from input to output. However, there are ways to look inside the algorithm. An extremely powerful example is the ability to compute a feature's importance.

To measure the importance of the variable *X*, the values of *X* are randomly permuted, to mimic the absence of the variable from the model. The difference in the accuracy of the prediction before and after permuting the variable *X*, i.e. with and without the help of the variable *X*, is used as a measure of its importance.

To illustrate an application of this technique, the most powerful variables for the prediction from a dataset were selected with two different methods: Random Forest and a benchmark model based on p-value and correlation analysis. The probability of default is then assessed with the logistic regression using the two selection techniques.

As shown in Figure 2 below, when looking at the two models the logistic regression calibrated on the set of variables selected by Random Forest performs better than the logistic regression calibrated on the set of variables selected by the benchmark selection process.

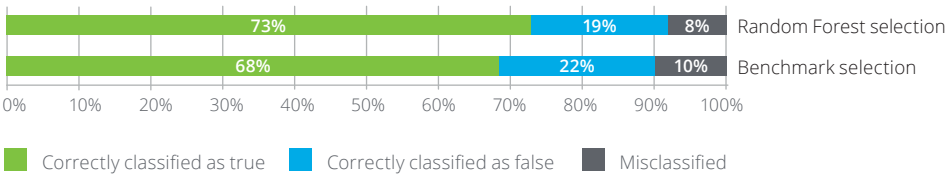



Figure 2: Confusion matrix for the logistic regression run on a selection of two different variables

**According to this test, feature selection with Random Forest is an accurate measure that allows a quick and relevant selection of features for credit risk models.**

## Insight 2 Random Forest can help to reduce the time spent on data management

Another important Random Forest feature is that the algorithm handles missing data in the dataset. In this section, we will test the predictive power and the behavior of the Random Forest algorithm on different credit risk datasets with different level of data transformations.

The raw dataset contains financial ratios, behavioral and descriptive characteristics. Complementary features with no pertinent information were removed.

The target variable is defined as the event of default on a historical period from 2012 to 2015.

Three capital data management steps were tested:

- Model behavior with regard to missing values
- How Random Forest responds to correlations between variables
- The impact on grouping modalities of categorical variables on model behavior

**The different data management processes are described in the following figure, knowing that each of these datasets were tested before and after the data imputation steps:**

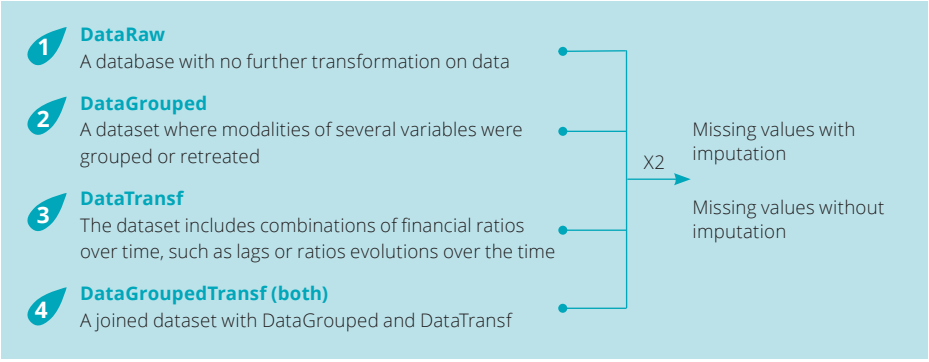



Figure 3: Description of the datasets used for the study

The Random Forest was calibrated by tuning the following hyper-parameters:

- Maximum number of trees
- Number of variables selected per tree
- Maximum depth of each tree
- Minimum of observations for each leaf
- Rate of the observations used for each tree

To avoid the overfitting effects – so that that the model does not generalize to new data and its predictive power is weakened – other preliminary techniques have been considered:

- Early Stopping: stopping the iterations early when validation indicators move away from training ones
- Time reduction: reducing the execution time of the algorithm
- Cross-validation option: testing the model on more validation samples

The results on the cross validation sample for the different sets of data are presented below:

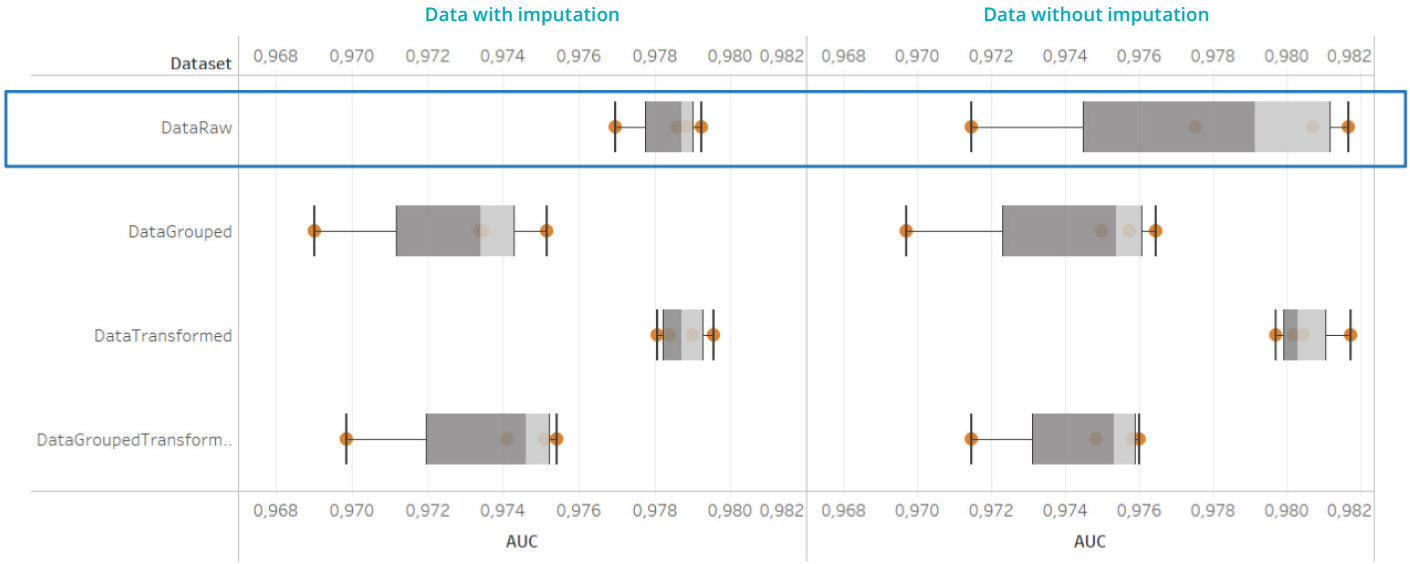



Figure 4: Random Forest performance (AUC) measured on cross validation sample

Globally the algorithm performs equally well on all datasets within a range interval of 96% and 98% in terms of AUC (Area Under the ROC Curve).

When the algorithm was tuned and tested on the datasets without imputing the variables with missing values, we can observe that the range of performance is the same.

The imputation step has almost no impact on model performance and predictive power, because the algorithms trained on the sets of data presented above perform well on both train and cross validation.

This cross validation analysis confirms that Random Forest techniques handle raw inputs well, with no prior transformations or pre-processing steps.

# Random Forest: a good choice for Machine Learning applications in credit risk

Machine Learning techniques, especially the Random Forest algorithm, manage raw datasets effectively without data preprocessing, notably the imputation of missing values.

The benefit of these techniques is that they may allow model developers to significantly reduce the time spent on data management and data pre-processing phases, prior to the model development phase.

They can also reduce bias in modelling through limiting the data transformations made prior to model development.

An area of concern is the stability of these techniques in scenarios where there are structural changes over time in a banking portfolio. In these situations, banks should consider increasing monitoring and backtesting frequency to keep model behavior on track.

**This analysis is part of the global Smart Credit Risk Modelling work ongoing within Deloitte France: an approach that allows banks to improve the performance of models by using innovative techniques.**
***https://www2.deloitte.com/fr/fr/pages/risque-compliance-et-controle-interne/solutions/smart-credit-risk-modelling.html***

**Nadège Grennepois**
**Partner, Risk Advisory**

Nadège Grennepois is a partner at Deloitte France, in charge of Advanced Modelling project for Credit Risk. She has 19 years' experience in the modelling field. She coordinates the development of both Model Risk Management and Smart Credit Risk Modelling offers.

Special thanks to our interns Alexis Gerbeaux and Kevin Vuong.

Article written in collaboration with:

**Anca Maria Alvirescu**
**Senior Consultant, Risk Advisory**

Anca Maria Alvirescu is a senior consultant and data scientist at Deloitte France. Her work consists of applying innovative solutions to credit risk management, notably for credit risk modelling and quantification topics.

**Margaux Bombail**
**Consultant, Risk Advisory**

Margaux is a consultant at Deloitte France where she specializes in treasury services. Besides working on financial regulations and financial instruments, she is involved in developing machine learning applications for quantification purposes.

## Southeast Asia contact

**Frederic Bertholon-Lampiris**
**Executive Director, Risk Advisory**
flampiris@deloitte.com

Frederic Bertholon-Lampiris is a Partner, leading the Financial and Regulatory Risk Advisory practice, advising more than 40 banks in Singapore and across South East Asia. He has over 20 years of multidisciplinary experience in risk management, regulatory compliance and model development for regulatory or business purposes in the banking and asset management industries.