Since the financial crisis, regulators put a great focus on risk management supervision and expect banks to have transparent, auditable risk measurement frameworks depending on portfolios 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.

**Point of View:**

Using Random Forest for credit risk models

**Machine learning and Credit Risk : a suitable marriage ?**
Banks focus on revisiting risk management frameworks with advanced analytics to converge towards end-to-end approaches combining data management, innovative technological solutions and analytics platforms fostering a 360 degrees client global view.

The usage of Machine Learning techniques faces skepticism for credit risk model development, notably for regulatory purposes because of the lack of transparency and the known “black box” effect of these techniques. Nevertheless, Machine Learning techniques can help model developers to reduce model risk and improve general model predictive power.

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

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

Indeed, 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?

Model aggregation methods or ensemble learning are algorithms working on a diver and conquer approach in order to improve performance of the predictions. The principle is to combine several statistical models and then classify new individuals by taking a (weighted) vote of their predictions 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 the aggregation of the same learning algorithm. Repeating these combinations aim at increasing robustness to variance and reducing the overall sensibility to model parameters and noise.

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

The random forest model is a particular case of bagging classifier applied on classification and regression trees (CART) with in addition another randomize process regarding the set of explanatory variables.

To illustrate these differences between models, bagging (bootstrap aggregating) and boosting are the two most known homogeneous models. The first one 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 composing a forest that are then aggregated. The diversity of these trees comes from two aspects of the construction of the forest.

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

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

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

The random forest algorithm is based on the construction of a myriad of different decision trees composing 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

<table>
<thead>
<tr>
<th>Pros</th>
<th>Cons</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Limits overfitting</td>
<td>• Low interpretability</td>
</tr>
<tr>
<td>• Method with high accuracy (pruning is not used)</td>
<td>• Parameter choice (number of trees, proportion of the observation in each sample, depth ...)</td>
</tr>
<tr>
<td>• Easy choice of relevant variables</td>
<td>• Computation time</td>
</tr>
<tr>
<td>• Stable against the data (good handling of missing values)</td>
<td></td>
</tr>
<tr>
<td>• Can handle a high dimensional set of variable</td>
<td></td>
</tr>
</tbody>
</table>

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.
**Point of View: Using Random Forest for credit risk models**

**Insight 1 Random forest is capable of identifying important features**

As many of machine learning algorithms, the random forest can be seen as a black box. Namely, it is not possible to define completely the process from the input to the output. However, there are ways to peek into the algorithm. Among them, a very powerful one is the ability to compute features 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 prediction accuracy 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.

As an example of application, the most powerful variables for the prediction from a dataset were selected with two different techniques: the 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 showed in the figure below, when looking at the two models, the logistic regression calibrated on the set of variables selected by the random forest performs better than the logistic regression calibrated on the set of variables selected by the benchmark selection process.

![Confusion matrix for the logistic regression run on two different variables selection](image)

**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 are going to 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 graph, knowing that each of these datasets were tested before and after the data imputation steps:

![Description of the datasets used for the study](image)
The Random Forest was calibrated by tuning the following hyper-parameters:

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

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

- Early Stopping, consisting in stopping the iterations early when validation indicators get away from training ones.
- Time reduction, which consists in reducing the execution time of the algorithm.
- Cross-validation option, which tests the model on more validation samples.

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

![Data with imputation vs Data without imputation](image)

**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).

In the same time, 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 well raw inputs with no prior transformations or pre-processing steps.
Random Forest a good choice for Machine Learning applications in credit risk

As one can notice, Machine Learning techniques, especially the Random Forest algorithm, manages well raw datasets without data preprocessing, notably the imputation of missing values. The benefit of these techniques is that it may allow model developers to reduce significantly the time spent on data management and data pre-processing steps prior to the model development step. It can also reduce bias in modelling though limiting the data transformations made prior to the model development.

Sure, one may argue the stability of this technique when there are structural changes over time in a banking portfolio. Banks should then consider increasing monitoring and backtesting frequency to keep model behavior on track.

This analysis is part of the global Smart Credit Risk Modelling work within Deloitte France, an approach allowing banks to improve models’ performance 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 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 is coordinating the development of both Model Risk Management and Smart Credit Risk Modelling offers.

Anca Maria Alvirescu
Senior Consultant, Risk Advisory
Anca Maria Alvirescu works as senior consultant and data scientist at Deloitte France. Her work consists in 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 and she specializes in treasury services. Besides working on financial regulations and financial instruments, she is involved in developing machine learning applications for quantification purposes.

Koen Dessens
Partner, Risk Advisory
Kdessens@deloitte.nl
Koen is partner at Deloitte Netherlands, in the Financial Risk Management team. He is responsible for the Modelling competency.

Dirard Mikdad
Director, Risk Advisory
dmikdad@deloitte.nl
Dirard is director within the Financial Risk team in the Netherlands, with a strong focus on managing regulatory change programs in the credit and market risk area.

And a special thanks to our interns Alexis Gerbeaux and Kevin Vuong.

Article written in collaboration with:

NL Contacts

Deloitte refers to one or more of Deloitte Touche Tohmatsu Limited (“DTTL”), its global network of member firms and their related entities. DTTL (also referred to as “Deloitte Global”) and each of its member firms are legally separate and independent entities. DTTL does not provide services to clients. Please see www.deloitte.com/about to learn more.

This communication contains general information only, and none of Deloitte Touche Tohmatsu Limited, its member firms or their related entities (collectively, the “Deloitte network”) is, by means of this communication, rendering professional advice or services. Before making any decision or taking any action that may affect your finances or your business, you should consult a qualified professional adviser. No entity in the Deloitte network shall be responsible for any loss whatsoever sustained by any person who relies on this communication.

© 2018. For information, contact Deloitte Touche Tohmatsu Limited.