Chapter 13
Boosted Decision Trees for Credit Scoring

Luca Di Persio
University of Verona, Italy

Alberto Borelli
University of Verona, Italy

ABSTRACT

The chapter developed a tree-based method for credit scoring. It is useful because it helps lenders decide whether to grant or reject credit to their applicants. In particular, it proposes a credit scoring model based on boosted decision trees which is a technique consisting of an ensemble of several decision trees to form a single classifier. The analysis used three different publicly available datasets, and then the prediction accuracy of boosted decision trees is compared with the one of support vector machines method.

MACHINE LEARNING AND FINANCIAL RISK MANAGEMENT

During last decades we have witnessed to an impetuous growth of Machine Learning (ML) applications in almost all areas related to everything could have a relation with data, no matter about their proper organisation, source, deep, sampling frequency, even independently from the fact that they have been artificially generated or taken from real life scenario. From this point of view, one of the areas that have been received most attention has certainly been the one related to the modern theory of financial markets, particularly within the financial risk management (FRM) sector. The latter is mostly justified by the huge amount of organised data, as in the case of financial time series, that have been made available, often real time, by several both private and open access sources. Nevertheless, such a reason, being necessary, is far from being sufficient, since, as to efficiently deal with thousands of giga of heterogeneous data aiming at providing, e.g., accurate forecasting methods for risky quantities, we also need effective computational solutions. Toward this direction, classical approximation/calibration schemes, as in the case of Monte Carlo-based approaches, see, e.g., (Sandmann, G., Koopma, S.J., 1998), spectral and/or asymptotic solutions, see, e.g., (Albeverio, Cordoni, Di Persio, Pellegrini, 2019, Bonollo, Di Persio,

DOI: 10.4018/978-1-7998-8609-9.ch013

Pellegrini, 2015) and references therein, simply fail to produce results in reasonable amount of times, no matter about the use of personal computers or in-cloud solutions. From a theoretical point of view, an alternative has been given by schemes based on nested neural networks, ensemble of ML-methods, and possibly hybridized alternatives, e.g. mixing strong results belonging to the realm of S(P)Des with calibration procedures coming from the ML world. Nevertheless, such approaches started to be computationally walkable just during recent years, namely when proper algorithms have been developed, exploiting modern programming languages and shared programming platforms, together with a massive increasing of computational strength essentially realized by on-line calculus platforms, nowadays also reliable and secure, as in the case AWS, Azure, Kamatera, etc. As a result, aforementioned sophisticated, but classical, statistical-based models, have been progressively substituted by a massive use of ML-related solutions within everyday financial institutions' activities, spanning within different economics areas, such, e.g., management, banking, insurance, etc., where ML/NNs solutions are higher and higher implemented to derive financial predictions, risk modelling scenarios, optimal portfolio constructions, insurance strategies, etc.

The just depicted scenario traduces in a set of both theoretical and applied methodologies, applications, analytical solutions, numerical schemes, practical determinations, data collection and relative evaluations, which form a world simply too wide to be collapsed in just one chapter, or even in a series of books. Rather, we preferred to focus our attention on a specific class of NN-based approach to a specific financial tasks which is known as the credit scoring problem. As to better introduce both the methods used and the financial application considered, let us provide a brief overview about the financial risk management world. First of all, let us recall that financial risk sources are classically grouped into the following classes: adverse financial market movements, loan defaults, unexpected insurance claims, fraudulent activities, customers loss. According with such risk clusters, it is then possible to classify risk into different categories, i.e.: market risk, insurance or demographic risk, operating risk and credit risk .

Market risk refers to the uncertainties in the value of the company's underlying assets, liabilities, or income due to exposure to a highly dynamic financial market.

Insurance and Demographic risk is more specific to the insurance industry. Indeed, it refers to the variance in insurance claim experience due to unpredictable events (e.g. catastrophes, car accidents, etc.) as well as uncertainties involved with the demographic profile of its policyholders (e.g. mortality).

Operational risk refers to the risk of loss due to the unpredictability of business operation or loss of performance due to faulty or fraudulent business practices.

Finally, *Credit risk* refers to the uncertainty involving creditors' ability to perform their contractual obligation (loan defaults or bankruptcy). This is applicable for both retail lenders (lenders who provide loans to individuals or retail customers) and corporate lenders (lenders who provide loans to businesses).

For what concerns market risk, one of the principal tasks one has to consider when treating both its possible sources and related solution, is linked to the volatility measurement and associated forecasting. One of the possible models that can be effectively implemented to address such tasks is the Generalized Autoregressive Conditional Heteroskedasticity (GARCH) one, considering its output as one of the inputs for a Multi-Layered Perceptron model. Indeed, such a solution has shown to be able to provide accurate volatility prediction (Kristjanpoller, Fadic, & Minutolo, 2014), improving performances obtained implementing traditional methods like Exponentially Weighted Moving Average (EWMA), Autoregressive Conditional Heteroskedasticity (ARCH) or the classical GARCH. Experiments suggest also that Long Short-Term Memory (LSTM), which is a variant of Recurrent Neural Networks (RNN), can outperform the traditional GARCH model, while providing more robust forecasts (Liu, 2019).