Interpretable models of loss given default

João Bastos (ISEG – CEMAPRE)

ABSTRACT

Model interpretability may be defined as the degree to which a human can understand the cause of its outputs. In statistical modeling, there is a trade-off between interpretability and prediction accuracy: interpretable models are usually less accurate, whereas complex models have greater out-of-sample precision so long as we control overfitting. Credit risk management is an area where regulators expect banks to have transparent and auditable risk models, which would preclude the use of black-box models. Furthermore, unknown biases in the risk models may lead to unfair lending decisions. In this study, we show that banks do not have to sacrifice prediction accuracy at the cost of model transparency. In particular, we show that the predictions given by black-box models for credit losses given default can be interpreted in terms of their inputs. Therefore, banks can comply with stringent regulatory requirements while pursuing a competitive advantage.

Joint work with Sara M. Matos