Larger Datasets Lead to More Inaccurate Credit Scoring

Mimi Chong*
Western University Canada, London, Canada mchong9@uwo.ca

Matt Davison
Western University Canada, London, Canada mdavison@uwo.ca

Credit scoring is an automated, objective and consistent tool which helps lenders to provide quick loan decisions. It can replace some of the more mechanical work done by experienced loan officers whose decisions are intuitive but potentially subject to bias. Existing credit scoring models are built using as many historical data as possible from past borrowers with known repayment performance. Analysts believe that using a larger dataset to build credit scoring models will always increase model accuracy. However, previous findings show that increasing the amount of data used to build models may result in a more complex model with no significant increase in accuracy. We will show that if some borrowers respond untruthfully to some questions, using higher dimensional data may even reduce the model's predictive power compared against using a dataset with lower dimensions. Using more data to build the model will increase the associated accumulated error and results in an overestimated model with low accuracy. The proposed issue will be studied using simulated data and discriminant analysis based on the credit scoring context. Knowing the optimal amount of data that is required to build credit scoring models can both improve accuracy and reduce operational cost and processing time. This research can help lending financial institutions to maximize profit in their loan activities.

Keywords: classification, discriminant analysis, lying, normal distribution