R. Calabrese: Sample selection bias in peer-to-peer lending market

Credit scoring models for peer-to-peer (P2P) lending platforms are usually estimated only on the sample of accepted applicants. This may lead to biased estimates of the risk drivers. In this paper, we compare a model based only on accepted P2P applicants with one built on the sample selection approach applied to all applicants. To correct for the sample selection bias, we propose a new flexible regression model suitable for binary imbalanced samples. To relax the usual assumption in scoring models of symmetric link function, the quantile function of the generalised extreme value distribution is considered as the link function. We also use different copula functions to model the dependence structure between the selection and outcome equations. We implement the proposed model in the R package BivGEV. The application of this proposal to a comprehensive dataset provided by Lending Club shows that parameter estimates obtained only on accepted P2P applicants are biased. The predictive accuracy of our proposal is higher than those obtained using univariate approaches or a sample selection probit model, commonly employed by P2P lending platforms. Our proposal also provides more conservative estimates of the Value at Risk and the Expected Shortfall.