

Abstract

This thesis investigates the effect of binning numerical variables on the performance of credit risk models. The differences are evaluated utilizing five publicly available data sets, six evaluation metrics, and a rigorous statistical test. The results suggest that the binning transformation has a positive and significant effect on the performance of logistic regression, feedforward artificial neural network, and the Naïve Bayes classifier. The most affected aspect of model performance appears to be its ability to differentiate between eligible and ineligible customers. The obtained evidence is particularly pronounced for moderately-sized data sets. In addition, the findings are robust to the inclusion of missing values, the elimination of outliers, and the exclusion of categorical features. No significant positive effect of the binning transformation was found for the decision tree algorithm and the Random Forest model.