ABSTRACT

mgr Grzegorz Migut

This dissertation is concerned with the multi-aspect assessment of the quality of classification data mining models. Its aim is to identify determinants affecting the quality of models and to determine the relationships between them. The substantive area of the above investigations was limited to customer loyalty models. In addition to the main objective of the study, secondary goals include: assessing the impact of selected data transformation techniques on the quality of the built classification models; identifying the optimal path for building an econometric model, using logistic regression as an example; assessing the effectiveness of classification tree models built using alternative partitioning paths; comparing the performance of the econometric model with machine learning models; assessing the impact of hybridization, segmentation and aggregation on the quality of the built models; comparison of interpretable models with advanced "black-box" methods.

The research procedure concerned the performance of simulation studies assessing the impact of determinants affecting the quality of customer migration models and determining the relationships between them. Based on the available dataset, a number of models were built in accordance with the CRISP-DM methodology. The following factors were taken into account during the simulations: ***T**ransformation* - method of preparing predictors, discretization, standardization, etc.; ***I**nteraction* - the fact of adding derived variables to the dataset; ***V**ariables* - method of selecting variables for the model; ***H**yperparameters* - methods of optimizing hyperparameters; ***E**nsembles* - additional learning strategies: segmentation, hybridization, model aggregation. Aspects of TIVHE were taken into account in a way that took into account the specific properties of the analytical methods used.

The dissertation consists of five chapters. The first chapter introduces customer retention as an area of marketing modelling. The second chapter deals with the data preparation for the construction of customer retention models. The third chapter presents issues related to the construction of an optimal classification model. Chapter four discusses issues relating to the validation and application of customer retention models. Among the aspects presented in the theoretical chapters, the part related to the measures of the model predictive power, which includes additional research on assessing the sensitivity of such measures to changes in the proportions of classes of the dependent variable, may deserve additional attention.

Chapter five presents the results of research work related to the identification of the optimal selection path of the customer migration model. Based on the research conducted, it can be concluded that the factors influencing the quality of customer retention models affect them differently across different methods. The aspect of variable transformation turned out to be significant only in the case of logistic regression and had a large impact on the quality of the model. The addition of derived variables was the only factor that had a positive impact on the results of all analyzed methods. A similar effect was observed for hyperparameter optimization. Segmentation of the data set turned out to be a factor that only improved the quality of "white-box" models. Its effect was particularly visible when combined with the addition of derived variables. For "black-box" models, a negative impact of segmentation on the predictive power was found. Model aggregation was found to improve the quality of all models except logistic regression.