Novel Data Mining Approach for Prognostication and Customer Identification at Small Medium Enterprises

W.M.H.G.T.C.K. Weerakoon*, and R.M.K.T. Rathnayaka

Department of Physical Sciences and Technology, Faculty of Applied Sciences,
Sabaragamuwa University of Sri Lanka, Belihuloya, Sri Lanka
**thariweera@gmail.com*

Almost every industry, organizations use Information Technology for their core business activities. Similarly, SMEs tend to use technologies for their businesses and the usage has increased data generation. Even though the amount of data generated low compared to the Large-Scale Enterprises, yet SMEs' data reserves can be converted into meaningful information with the aid of data mining and machine learning concepts. When using data mining and machine learning approaches for datasets generated at SMEs need special cautions, due to constraints such as limited computational capacity at SMEs and as well as comparatively less data capacity at data repository than convectional Big data. This study aims to formulate a data mining model for sales prediction and as well as focus on initial Customer Relationship Management stage, customer identification within the context of SMEs, which is effective and efficient under the constraints identified. The data sets comprise of demographic features; Gender, Age, Residency information, Marital Status and Occupation. Preprocessed repository data used to initiate descriptive statistical analysis, variable correlation analysis and feature engineering. The processed variables used within number of machine learning algorithms; Linear Regression (LR), Ridge Regression (RR), Decision Tree Regression (DTR), Random Forest Regression (RFR), Multiple Layer Perceptron Regression (MLPR) models to find most efficient algorithm for sales prognostication. Furthermore, the customer identification had initiated with Principle Component Analysis (PCA) along with classification techniques. The finalized model for predictive analytics at SMEs determined through the lowest Root Mean Square Error (RMSE) and then a validation process carried out to assess the performance. For the predictive analytics in sales, DTR was suitable due to lowest RMSE of 2689 and through PCA and classification 3 customer bases were identified.

Keywords: *SME, Prognostication, Customer identification, Regression analysis, Principle component analysis*