PREDICTION OF INTEREST FOR MOTOR INSURANCE PRODUCTS BASED ON MACHINE LEARNING APPROACHES USING INFORMATION ON EXISTING HEALTH INSURANCE CUSTOMERS TO SUPPORT CROSS SELLING OPPORTUNITIES

Wasana Erangi Perera

A dissertation submitted in partial fulfilment of the requirement for Master of Science degree in Big Data Analytics

Department of Computing

Informatics Institute of Technology, Sri Lanka in collaboration with

Robert Gordon University, Aberdeen, Scotland

2022

Abstract

Insurance sector is the most important domain currently available in all over the world which defined as a mechanism of transferring financial risk. Even though, preventing risk is unpredictable, insurance can financially safeguard people from the risk. So, it is important to keep people engaged in insurance sector by having more insurance products rather than having one insurance product. Through this study, a timely need solution will be developed to help insurance industry for engaging people more on their products by considering cross-selling concept. Since most of the people have health insurance products, the study area will be focused on selling motor insurance products for the health insurance policyholders. A system will be developed to predict policyholders' interest status on motor insurance products. Several predictive algorithms will be carried out to predict the interest status on motor insurance products. Logistic Regression, Decision Tree, Random Forest, XG Boost, Adaboosting and Light Gradient Boosting will be evaluated to carry out the best fitted model for prediction. SMOTE techniques will be used to handle the class imbalance in response class. As the model evaluation criteria, accuracy, precision, recall and F1-score has been conducted through the advanced analysis process. Furthermore, validation of the best fitted model will be conducted using K-fold cross validation approach. Based on the results of the model outcomes, XG Boost model with SMOTE technique is selected as the best fit model for cross-selling prediction model with 80% recall and 76% accuracy.

Keywords: logistic regression, random forest, extreme gradient boosting, Adaboosting, light gradient boosting, accuracy, precision, recall, f1-score, SMOTE, K-fold cross validation