MSc Project Report

PREDICTION OF CUSTOMER CHURN STATUS AND POLICY BENEFITS FOR PASSENGER CAR SEGMENT OF MOTOR INSURANCE BASED ON MACHINE LEARNING APPROACHES USING SRI LANKAN MOTOR INSURANCE DATA

Erangi Wijerathne

2023

A report submitted as part of the requirements for the degree of MSc Big Data Analytics at

Robert Gordon University, Aberdeen, Scotland

Abstract

General (Non-life) insurance has based on losses or damage related to particular financial event. Motor insurance is a subset of general insurance and it is a kind of insurance that acquired vehicles and in return offers a payment in case of an accident to insured vehicle within their terms and conditions. The insurance sector has become a one of the most competitive industry in Sri Lanka due to the recent political and economic movements. It is important to keep existing customers since the process of acquiring a new customer is more expensive compared to keeping an existing customer. Thus, it is significant to identify potential churn customer in advanced and make strategies to keep them in current company.

This study will be focused on developing a classification model to predict customer churn status using different machine learning algorithms including Logistic regression, Random forest, Decision tree, XGBoost, Ada Boost and, Light Gradient Boost. The SMOTE technique was used to overcome class imbalanced problem in the used dataset. The model evaluation metrics such as Accuracy, Precision, Recall and, F1-score were used to evaluate fitted models. XGBoost model was selected as the best fitted model based on Accuracy and F1-score.

Churn customer profile analysis is performed using K-Means in order to develop strategies to keep churn customers. The Principal Component Analysis (PCA) was performed on churn customer to reduce nine predictors to two components since it's difficult to visualize nine predictors in a successful way. Then, K-Means clustering technique was performed on two component dataset to group churn customer based on their characteristics. The model evaluation was unable to perform on K-Mean cluster model due to unavailability of ground truth. Finally, special policy benefits were created based on characteristics of obtained cluster to offer to customers.

Keywords: Logistic regression, Random forest, XG Boost, Ada Boost, Light Gradient Boosting, Accuracy, Precision, Recall, F1-score, SMOTE, PCA, K-Means, K-Fold Cross Validation