INFORMATICS INSTITUTE OF TECHNOLOGY

In Collaboration with

UNIVERSITY OF WESTMINSTER

Churn Prediction System for Pre-paid Subscribers in Sri Lanka

Kanishka Udugoda

w1800735 / 20200010

Supervisor: Dr. Alroy Mascrenghe

Date: 16th May 2022

Department: Computer Science

Submitted in partial fulfilment of the requirements for the MSc in Advance Software Engineering degree at the University of Westminster.

ABSTRACT

Since the customer is the biggest asset to a company considering most of the business scenarios, it is very important to ensure the customer's retainability for business continuity. Specifically, in the telecommunication business, subscriber churn will most likely increase further and it will be highly affected for all the telecommunication service providers in Sri Lanka due to the upcoming number portability feature which will be introduced by the Telecommunication Regulatory Commission of Sri Lanka (TRCSL). Therefore, it is very important for the telecommunication companies to monitor the churning tendency of subscribers closely in order to maintain a sufficient subscriber base to make profits.

The purpose of this study is to research, design, and develop an accurate churn prediction model for pre-paid subscribers. Here, the pre-paid subscriber churn is more difficult to identify than the post-paid subscriber churn due to the fact that pre-paid subscribers are not bound to any contract with the service provider.

Using machine learning technology, the predictive models are designed and developed with supervised learning principals. For the implementation attempt, a dataset with 15000 records of pre-paid subscribers including 21 features was obtained from one of the leading telecommunication service providers and was used in this research project. Stacking ensemble technique was mainly selected to build and analyze since that method had not been used in recent research work in the telco domain. Apart from the mentioned technique, ensemble techniques such as bagging Classification and Regression Trees (CART) and Random Forest algorithms were used alongside with Logistic regression, K-nearest neighbor (KNN) and Naïve bayes (NB) algorithms for the predictive modeling and analysis. Predictive model which was built using Random Forest algorithm showed the best results among the other models with accuracies of **90.11%** and **89.1%** with caret and sklearn packages respectively.

Key Words: Churn, Subscriber, Pre-paid, Machine learning, Ensemble, Classification.