

**RARE EVENT PREDICTIVE MODELLING  
APPROACH TO IDENTIFY POTENTIAL  
CUSTOMERS WITH HIGH INTENSITY TO  
PURCHASE FIXED BROADBAND PACKAGES IN  
UWV COMPANY**

**Brandigampolage Chankami Chithrika Perera**

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

**Department of Business  
Informatics Institute of Technology, Sri Lanka  
in collaboration with  
Robert Gordon University, UK**

**2022**

## **Abstract**

Mobile telecommunications, which were an uncommon communication medium few decades ago, are now become vitally important to every individual as well as the economy of a country. The most important revenue part of telecommunication is mobile broadband connection where the possibility of accessing internet via mobile phones. Today accessing internet through mobile data can be seen in everywhere rather than any other medium. The mobile data usage has increased globally and locally during the last few years compared to the previous years. This rising mobile data usage can be expected in the future as well. Due to this rapid increase in mobile data usage, most of the service providers are intended to introduce new marketing strategies to increase sales in fixed broadband connections more as the penetration of fixed broadband connection is less compared to mobile data. This study is based on developing a model to identify customers who have higher propensity to buy fixed broadband packages in one of telecommunication company in Sri Lanka. The classification techniques; Logistic regression, Complimentary log-log link model, Decision tree, Random Forest classifier, Naïve Bayes classifier, Support vector machine (SVM), K-Nearest neighbor (KNN) and Extreme Gradient Boost (XGBoost) are mainly considered in this study with the intention of finding the most appropriate model with high prediction rate. Each classification model is tested in three approaches respectively feeding all variables, with only important features of Boruta algorithms and oversample balanced dataset. Class imbalance problem, a major issue arising very often in practice can be seen in this study sample due to low dispersion of fixed broadband connections among telecommunication customers.

In order to overcome the class imbalance in the training set, ROSE (Random Over-Sampling Ex-amples) technique is considered over synthetic minority oversampling technique due to unusual pattern in generated synthetic data. Reason behind selecting oversampling techniques over undersampling technique is loss of data will be minimized. Several evaluation measures such as sensitivity, specificity, accuracy and area under the ROC curve are used to evaluate the performance of each classifier. Sensitivity which is

most suitable model evaluation measure in this scenario is used to compare and select best classification model in this study.

The obtained results indicate that, for every classifier balancing the training set initially is essential to get a better performance. The XGBoost model after balancing the training set using oversampling technique has given better performance than the performances of other techniques. The model indicates Average revenue as most important feature and Online channeling as least important feature. Sensitivity of the best model is 95.201% which implies there is 95% possibility of correctly predicting positive class (has purchased fixed broadband connection) in this best XGBoost model.

**Key words: - Classification, telecommunication, fixed broadband connections, imbalance dataset.**