RISK PREDICT

A MACHINE LEARNING PIPELINE FOR LOCALIZED RISK PREDICTION IN HEALTH INSURANCE

Madhusha Hansani Welihinda

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

Department of Computing

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

2022

Abstract

In Sri Lanka, one of the main challenges faced by health insurance providers is identifying future risk of policyholders to come up with competitive and affordable premiums. Risk is a factor that differ from person to person and careful identification of risk is a crucial part of underwriting process. With the increase of insurance data and the use of machines learning algorithms the underwriting process can be further improved by faster data processing and identification of risks. This research aims at providing a framework coupled with a machine learning pipeline to predict health insurance risk amount in an accurate manner. A real-world dataset containing five years of claimed data has been used to conduct the analysis. The research was carried out by training seven classification models and four regression models to develop a machine learning pipeline to classify risk and then to predict risk amount. The experimental results showed that the proposed machine learning pipeline has obtained acceptable results and ensemble algorithms works well on health insurance data compared to other machine learning algorithms. In classification part of the study, Random Forest and XGBoost achieved the best accuracies. The regression part of the study reveals that XGBoost perform well in predicting amounts for low-risk policyholders and Random Forest generate better results for normal, high, and bad risk data after applying data augmentation techniques.

Key words: Health insurance risk, Underwriting process, Machine learning, Data augmentation