EFFECTIVE PRE-AUCTION SAMPLE SELECTION THROUGH A MACHINE LEARNING MODEL FOR SRI LANKA TEA BOARD

IIT No. 1613570

A Dissertation By

Wijesinghe Mudiyanselage Indika Hiran Wijesinghe

Supervised by
Dr. Sameera Viswakula

Submitted in partial fulfilment of the requirements for the

M.Sc Big Data Analytics

Robert Gordon University
August 2018

Abstract

A Scientific method of selecting all possible fraudulent tea samples from the weekly pre-auction; where tea samples will be bided by the tea exporting companies to be purchased for exporting purposes is critical to mitigate frauds in tea exports from Sri Lanka.

There was no such system built to predict fraudulent samples tea samples in tea industry, but there are some research being conducted to predict anomalies in other domains. Therefore, building a proper mechanism to predict fraudulent samples, received by the tea tasting unit of the Sri Lanka Tea Board is a challenge for the government's regulatory/testing body. The focus of this research is to implement a pre-auction sample selection system with Machine Learning approach to develop a self-learning sampling system to pick the most likely fraudulent samples for physical testing.

This report aims at producing optimized Automated Sample Selection Model with Machine Learning algorithm. The training dataset obtain from main tea board MS SQL database with a customized query. The core algorithm used for the system is Random Forest (RF) algorithm and it optimized with values obtained in Receiver Operating Characteristic (ROC) curve and Precision and Recall curve. The model implements with well-known web based micro service-based architecture to keep efficiency and scalability of the product. The Flask was use as RESTful API that offers web services with Python. In addition, Jinja2, JavaScript, ML based open source libraries and technologies were used to finalize the product.

Results obtained through the system were impressive as the system offered higher accuracy and could be used to predict frauds in other domains with modifications. Furthermore, researcher introduced modification for the current architecture and IOT based intelligent fraud detection mechanism with Google Machine Learning Platform (GMLP) as Future enhancement.

Keywords- Supervised Learning, Classification; Machine Learning; Random Forest; ROC curve; precession score, recall score, frauds prediction, Tea Industry