EpiMetrics, Inc.

Big Data in Universal Health

Case 3: Development of Outlier-Based Fraud Detection Model for PhilHealth Insurance Claims

EXECUTIVE SUMMARY

Abstract

The project applied big data methods in three topics highly relevant to the DOH – social determinants of health (SDH) versus universal health care (UHC) inequalities, geographical inequalities, and outlier detection. The third case was for PhilHealth outlier detection for insurance claims together with Thinking Machines.

Since the application of big data in the Philippine setting is an emerging field, this project was a proof of concept for the DOH and PhilHealth to explore the field for important tools in the monitoring and evaluation of public health. With big data, analysis of diverse and voluminous data becomes possible, and the DOH can gain insight into the factors that significantly impact and influence the health of the nation. However, mechanisms and policies must be put in place to ensure that data is available, accessible, and of good quality in order to properly apply big data analytics. Case 3 aimed to demonstrate how big data and computer science can both be utilized in detecting fraud in health insurance claims.

This case employed the use of Data Science, particularly Machine Learning (ML), which utilizes computer programming to detect patterns and predict outcomes. The following data sets were included: 1) claims data from 2014-2016 on all pneumonia and cataract cases using all case rates, 2) providers database, with all variables for healthcare institutions and healthcare professionals, 3) members and dependents databases, and 4) Medical Post Audit Module or MPAM data. Analysis was conducted according to claims groups namely, Cataract, Moderate-Risk Pneumonia and High-Risk Pneumonia using Microsoft Azure. Five models were included into the analysis which allowed for an unsupervised machine learning that focused on outlier detection. The five models were: Isolation Forest, On-Class Support Vector Machine (SVM) PCA- Based Anomaly Detection, and Local Outlier Factor (LOF). Outlier status derived from each model is then entered a majority ensemble, such that if 3 out of 5 models classifies a claim as an outlier, it is considered an outlier overall.

By undergoing a majority ensemble of 5 unsupervised machine learning models, outliers were detected. Under moderate risk pneumonia, 2,127 (0.095%) outliers out of 2,228,800 claims were identified. Under high risk pneumonia, 1,335 (0.78%) outliers out of 171,517 claims were identified. Under cataract requiring CPSA, 695 (0.198%) outliers out of 351,500 were identified. It is important to note that the outliers identified in this case are not necessarily fraudulent in nature. Due to the lack of training data, it was not possible to undergo a supervised type of machine learning. As such, it is these outliers should undergo an audit, and based on the audit's results, we can then use those fraud cases to jumpstart a fraud classifier.

In case 3, machine learning models can be utilized to detect outliers related to probable fraud among the health sector as these can speed up the auditing process with the system.

Executive Summary

A. Introduction

Government agencies, specifically, the Department of Health (DOH) and the Philippine Health Insurance Corporation (PhilHealth) have been amassing data about the nation's health, universal health care, and on the social determinants of health (SDH). To assist the Department of Health (DOH) and the Philippine Health Insurance Corporation (PhilHealth) in the achievement of national and global health goals for sustainable development, this project used big data tools to maximize the generation of actionable information from the amassed data. Moreover, since the application of big data in the Philippine setting is an emerging field, this project was a proof of concept for the DOH and PhilHealth to explore the field for important tools in the monitoring and evaluation of public health.

B. Methods

This case employed the use of Data Science, particularly Machine Learning (ML), which utilizes computer programming to detect patterns and predict outcomes. For this case, two types of ML were intended: supervised ML, which is commonly used to detect outliers by classifying data into two or more categories, and unsupervised ML, which utilizes specialized mathematical models to detect the existence of clusters, patterns, and outliers. The following data sets were included: 1) claims data from 2014-2016 on all pneumonia and cataract cases using all case rates, 2) providers database, with all variables for healthcare institutions and healthcare professionals, 3) members and dependents databases, and 4) Medical Post Audit Module or MPAM data.

Analysis was conducted according to claims groups namely, Cataract, Moderate Risk Pneumonia and High Risk Pneumonia. Data analysis was conducted using a cloud-based service, namely Microsoft Azure. Five models were included into the analysis which allowed for an unsupervised machine learning, which focused on outlier detection. The five models were: Isolation Forest, One-Class Support Vector Machine (SVM), PCA- Based Anomaly Detection, Local Outlier Factor (LOF), and Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN). Outlier status derived from each model is then entered a majority ensemble, such that if 3 out of 5 models classifies a claim as an outlier, it is considered an outlier overall.

C. Results and Discussion

By undergoing a majority ensemble of 5 unsupervised machine learning models, outliers were detected. Under moderate risk pneumonia, 2,127 (0.095%) outliers out of 2,228,800 claims were identified. Under high risk pneumonia, 1,335 (0.78%) outliers out of 171,517 claims were identified. Under cataract requiring CPSA, 695 (0.198%) outliers out of 351,500 were identified. It is important to note that the outliers identified in this case are not necessarily fraudulent in nature. Due to the lack of training data, it was not possible to undergo a supervised type of machine learning. As such, it is these outliers should undergo an audit, and based on the audit's results, we can then use those fraud cases to jumpstart a fraud classifier. Additionally, there was no viable training data available for supervised machine learning due to the limited access of PhilHealth to the diseases. Therefore, only an Outlier Detection Model was made.

D. Conclusions and Recommendations

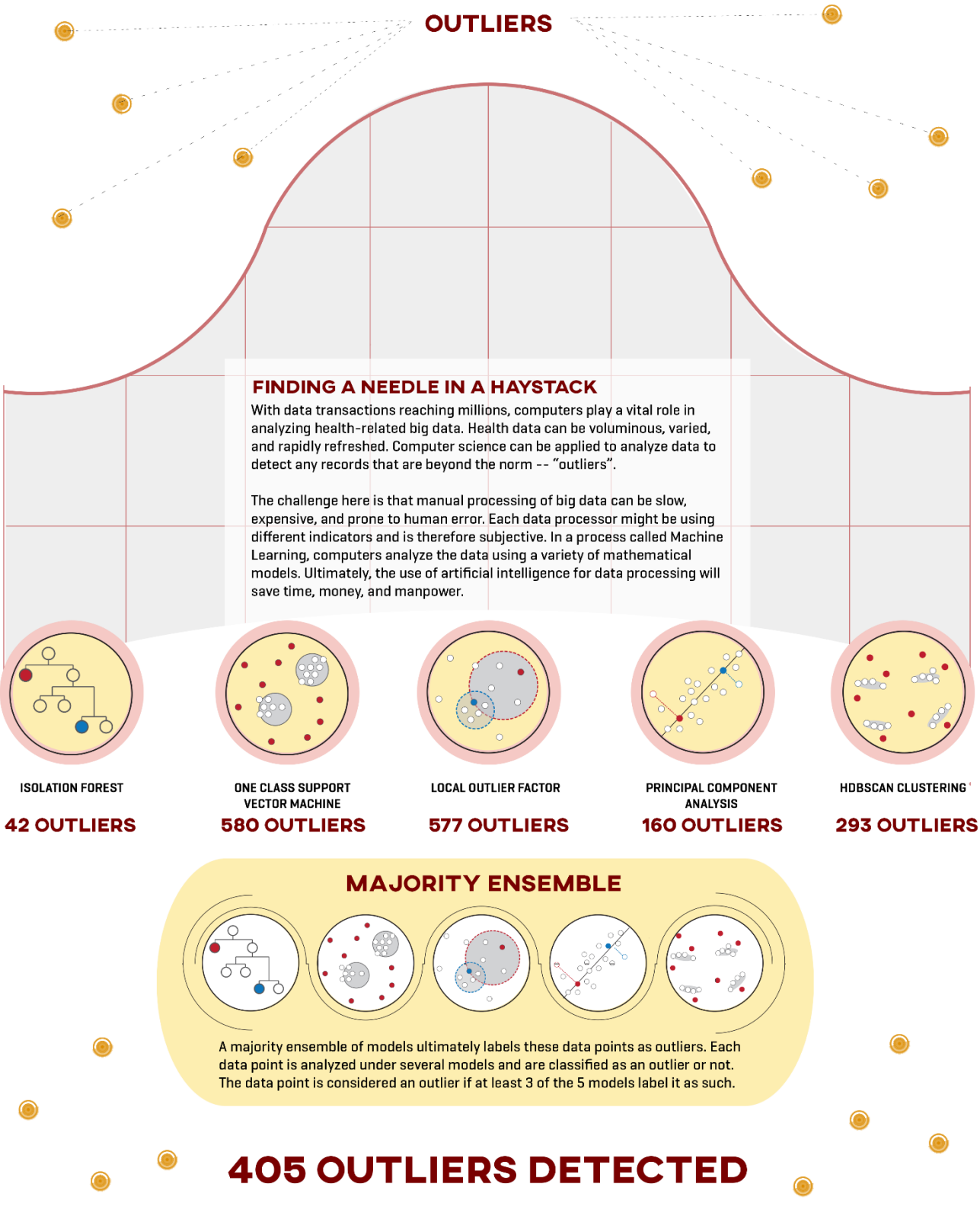
The study proved the use of big data to detect outliers related to probable fraud as a strategy for short-listing probably fraud and speeding up the audit process.

As technology advances, the Philippines must take advantage of its available healthcare related data and improve on it for better utilization and application. It is important that the country, particularly the Department of Health, invests in data governance. Mechanisms and policies must be put in place to ensure that data is available, accessible and of good quality in order to properly apply Big Data analytics.

.

OUTLIER DETECTION USING MACHINE LEARNING

EpiMetrics
HEALTH EQUITY THROUGH RESEARCH



The capacity to detect outliers in large volumes of data has significant applications in Public Health. In this case, outliers are data points that are categorically "different" from the rest of the data. These outliers can be unique data on compliance to company policy, quality of care, performance tracking, or can be unique in other applications. Nonetheless, outliers are indications for audit, investigation, or research.

EpiMetrics

Principal Investigator

John Q. Wong, MD, MSc

Co-Investigator

Carlo Emmanuel L. Yao MD, MBA

Research Associates

Christian Alis, PhD
Regina Ong-Drillon
Stephanie Sy
Raymond Sarmiento, MD