Biometrical Journal **52** (2010) 61, zzz–zzz / DOI: 10.1002/bimj.200100000

Detecting possible persons of interest in a physical activity program using step entries: including a web based application for outlier detection and decision making

S.S.M.Silva[*,1], **Denny Meyer** [1], and **Madawa Jayawardana** [1,2,3]

[1] Department of Statistics, Data Science and Epidemiology, Swinburne University of Technology, Hawthorn 3122 Victoria, Australia

[2] Peter MacCallum Cancer Centre, Melbourne 3000 Victoria, Australia

[3] Sir Peter MacCallum Department of Oncology, The University of Melbourne, Parkville 3010 Victoria, Australia

Received zzz, revised zzz, accepted zzz

According to recent statistics from the World Health Organisation, 23% of people aged 18 years and over are not sufficiently physically active. Strangely this is at a time when, due to the improvement in sensor technology, physical activity programs which track physical activity have become popular. However, some participants who enrol in these programs cheat, by manipulating the data they enter. This can be discouraging for other participants, also invalidating the overall accuracy of program outcomes. Therefore, detecting these participants and discarding their manipulated entries is important in order to maintain the quality of the program. Currently most of these physical activity programs use manual processes to detect and reject fraudulent step entries by reviewing the participant's demographic profiles along with their longitudinal step count performance data. In this study a process, including two parallel models for detecting person of interest characteristics and abnormal step count entries, is developed. The first model uses the penalised logistic regression with SMOTE sub-sampling to address the imbalance in the proportion of genuine and persons of interest. Having a highly imbalanced distribution between genuine and person of interest profiles makes this task more challenging. The second model uses a variety of outlier detection methods to detect and reject abnormal step entries based on previously entered data. This process will be more efficient and productive compared to the current manual system and will support better decision making in the future. The proposed system can be applied for other fraud detection applications, after suitable adjustments.

Key words: decision making; fraud; imbalanced; outlier; physical activity;

1 Introduction

Virgin Pulse Global Challenge (VPGC) is a scientifically designed program, which consists of four modules, namely physical activity, nutrition, balance and sleep, together aiming to enhance the physical activity and psychological well-being of all participants. This program runs for 100 consecutive days with thousands of participants enrolled from all over the world.

During the 100 day program participants collect and enter/sync step counts using devices such as pedometers or accelerometers or other wearable fitness tracking devices. These step entries can be edited to include other activities such as yoga, weightlifting etc., which typically would not contribute steps measurable by wearable devices. This can lead to some abnormal entries which might result in outliers. In addition, there may be some participants who cheat during the program. These anomalous data points may

[*]Corresponding author: e-mail: sssilva@swin.edu.au, Phone: +061-451-361-991, Fax: -

adversely affect the accuracy of program outcomes and erode confidence in the program's leaderboard. According to the program experts, these outliers discourage other participants and teams from participating in the program. Currently, a semi-automated process is carried out to detect these anomalous data points and to send the participants a notification asking them to check and verify their entry. In this process, any step entry which is higher than an arbitrary value (e.g:-30,000 or 25,000) will be flagged. The justifications returned by the participants are reviewed manually by the customer service representatives in order to accept or reject these specific entries, also taking into account the past performance of these participants. This process is time consuming and potentially prone to subjectivity. This study develops a combined methodology for detecting persons of interest and outlier step entries, using participant profiles as well as their longitudinal performance data for step counts. The main objective of developing a 'person of interest' detector is to identify the probability of being a 'person of interest', so as to reduce the subjectivity in decision making and support the process. The objective of the outlier detector application is to enhance the efficiency and productivity of the current process in outlier detection and decision making. It is also expected that this system will decrease the workload of customer service representatives and improve the accuracy of the program outcomes data.

The statistical definition for outliers should ideally depend on the underlying distribution of the variable in question (Last and Kandel, 2001). According to Pincus (1995), a general definition for an outlier is an observation (or subset of observations) which appears to be inconsistent with the remainder of the data. This definition is most applicable for the current context. The step entries which are inconsistent with a participant's previous behaviour are defined as outliers or anomalies. If there are several unjustified step entries which are identified as outliers, then that participant's profile should perhaps be labelled as a 'suspected person of interest'.

In most cases the cause for outliers is unknown to the data analyst or user. Data entry errors associated with malfunctioning technology may also cause outliers. Alternatively, outliers can represent exceptional performance and correct information (Mendenhall *et al.*, 1993; Last and Kandel, 2001). This may be due to a rare event that has taken place such as a marathon. These two types of outliers occur in the current program. For this reason human intervention needs to play a role in the outlier detection process in order to ensure that these outliers are correctly handled, taking the justifications of participants into account.

The majority of the anomaly detection methods can be categorised as classification based, nearest neighbour based, clustering based or statistical test based techniques (Chandola *et al.*, 2009). Classification based outlier detection techniques work in two phases. The first phase is the learning phase for the classifier using the labelled data, and the second phase is the testing phase which classifies unseen observations as in-range or anomalous, using the trained classifier (Chandola *et al.*, 2009). Nearest neighbour-based anomaly detection methods can be categorised into two groups (Chandola *et al.*, 2009); using the distance of a data point to its k^{th} nearest neighbour as the anomaly score (Eskin *et al.*, 2002; Zhang and Wang, 2006) or computing a relative density for each data point to find its anomaly score (Breunig *et al.*, 2000; Ismo *et al.*, 2004). Clustering based anomaly detection techniques work on the assumption that in-range data points belong to a cluster in the data whereas anomalous data points do not belong to any cluster (Chandola *et al.*, 2009). Among these techniques the DBSCAN (Ester *et al.*, 1996), ROCK (Guha *et al.*, 2000) and SNN (Ertöz *et al.*, 2004) clustering techniques are important. Statistical outlier detection techniques usually fit a model to the given data and then test whether a new unseen data point supports the model using a statistical inference test (Chandola *et al.*, 2009). Gaussian model based techniques assume data generated from a Gaussian distribution, where the parameters are estimated using Maximum Likelihood Estimation(MLE). A threshold is defined to identify an anomaly depending on the distance of the point from the estimated mean (Chandola *et al.*, 2009). All the above mentioned methods focus on point anomalies. Contextual anomalies are anomalous only in a specific context but not otherwise (Chandola *et al.*, 2009; Song *et al.*, 2007). For example, time is the contextual attribute in time series data, longitude and latitude are the contextual attributes of spatial data. The anomalies found in these cases also need to be handled with care. In this study the time context is very important so that the outliers are detected for

any participant by comparing step count entries for that specific participant with his/her the previous step count entries rather than the step count entries of other participants.

This study has applied seven different outlier detection techniques including statistical, nearest neighbour and contextual anomaly detection methods to detect anomalous step entries along with 'person of interest' identification. The study is organised into seven parts as follows. In section 2, the proposed system to detect fraudulent participant profiles ('person of interest') and anomalous step entries is described in broad terms. Section 3 describes the data availability for both outlier detector and person of interest detector. Section 4 includes the methods used in the study to detect the persons of interest in the program. In section 5, methods used for outlier detection have been described along with the application system. Section 6 summarises the results obtained from the proposed system. Finally, section 7 and 8 provide concluding remarks for the study.

2 Proposed system to detect fraudulent profiles and step entries

This study will be focused on the physical activity module of a workplace health and exercise program in which the participants track their day to day physical activity using step counts. Fraudulent participant profiles can be defined as step count trajectories for participants who provide inflated step entries on a regular basis. These participants may be using fraudulent step entries with the intention of moving to the top of the leaderboard. Therefore these participants are named as 'persons of interest' throughout this study. The participants who provide genuine step entries and have genuine intentions for participating in the program are called 'genuine'.

The current system of flagging step entries uses an arbitrary value in order to identify fraudulent step entries. This cut-off is not a personalised value although the physical activity of participants is likely to vary depending on various psychosocial, demographic, weather and climatic factors. Moreover, accepting and rejecting flagged step entries based on past performance and reasons provided by participants tends to be subjective. Furthermore, once a flagged step entry is rejected or accepted, these findings are not being taken into consideration in order to identify the genuineness of participants. It is expected that the proposed framework will overcome most of these issues.

The following combined framework shows the proposed system to detect 'person of interest' as well as their individual fraudulent step entries.

The framework in figure 1 can be subdivided mainly in to two main systems namely 'person of interest detector' and 'the abnormal activities detector' (TAAD). Differentiation between 'person of interest' profiles and 'genuine' profiles will be conducted using the initial T1 survey responses and the past longitudinal performance data of the participants using the 'person of interest detector'. The T1 survey is an initial online survey completed by all participants, which provides demographic data, physical data (e.g:- height and weight), current behaviour (physical activity, nutrition intake, sleep, stress) and future goals. Some of these data are also used to segment all participants in relation to their intentions when joining the program. In parallel to this process, once a participant has been flagged as a 'person of interest' due to his/her fraudulent initial responses and performance data, the system will send this class probability to the TAAD, providing prior information for anomaly detection.

The TAAD detects the outliers and anomalies that exist in the step counts of participants. When a step entry has been detected as an outlier/anomaly, the participants are asked to provide a valid reason for that specific step entry. The provided reasons are reviewed by the customer service representatives, taking the 'person of interest' probability provided by the 'person of interest detector' into consideration.

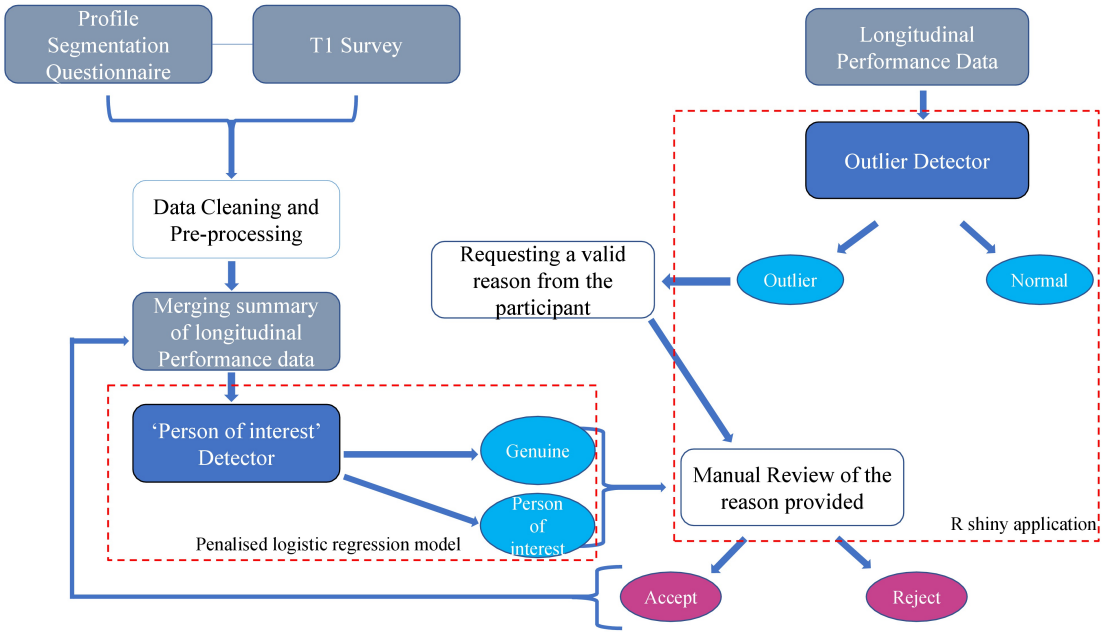

Figure 1 Proposed system to detect fraudulent profiles and step entries.

3 Data availability

As mentioned in section 2, the proposed framework consists of two main sub systems for 'person of interest' detection and 'outlier detection'. Therefore, two different data sets were used for these two sub systems.

3.1 Person of interest Detector

Currently Virgin Pulse Global Challenge does not have an automatic system for identifying the 'person of interest' profiles. Therefore, classification of profiles in to 'person of interest' and 'genuine' has not been included in their database. In order to overcome this issue, it was decided to consult with the VPGC experts to identify a clear cut person of interest profile. From this qualitative process it was decided to classify a profile as a 'person of interest' if more than 2 percent of the step entries were rejected by the current reviewing process.

As mentioned in Figure 1, the person of interest detector model used initial T1 survey responses of the participants to predict whether a participant is a person of interest. The initial T1 survey contained 23 Likert scale questions, covering the following five core fields:

i Health and physical activity,

ii Sleep,

iii Nutrition,

iv Stress, happiness and productivity

v Psychological wellbeing

The Likert scale questions under each core field were then averaged so as to extract a mean value for each core field.

In addition, the T1 initial survey contained the following five questions regarding alcohol consumption, sleep and physical activity that were used to predict if a participant matched the person of interest profile:

i Whether the participant is a smoker or not

ii Number of alcohol drinks the participant typically have

iii Number of days per week the participant had alcohol

iv Number of hours sleep per night

v Number of days per week the participant has undertaken 30 minutes of moderate intensity physical activity

In addition, the profile segmentation questionnaire was used to predict whether a participant is a person of interest. This questionnaire contained questions related to the demographic status of the participants (e.g., age, sex, height, weight etc.) and it asked the participants about their motivation to join VPGC and expectation of enrolling to VPGC. Finally, the model used three summary statistics for longitudinal performance data, namely, personal best step count (PersonalBest), average step count for valid entries (StepAverage) and average step count per entry (StepAveragePerEntry), for which invalid step counts were also considered.

3.2 Outlier Detection

In order to test the suitability of different outlier methods for the current program, 560 participants who had completed the program with at least 100 step entries were randomly selected. This data set contained 55,998 random step entries along with the other variables such as;

i Event day (1 to 100)

ii Created date of the step entry

iii The decision of the reviewing process (whether it is a regular or approved or rejected step entry)

4 Methodology for detecting persons of interest

In this section, methodology considered for detecting persons of interest will be described. To detect persons of interest in the program it was decided to test two different models namely penalised logistic regression and random forest. These two methods have different strengths in classification problems as follows.

1. Penalised logistic regression

   The `R package glmnet` (Friedman *et al.*, 2010a; Simon *et al.*, 2011) includes an efficient procedure for fitting the entire lasso (least absolute shrinkage and selection operator) and an elastic net regularization path for linear regression, logistic and multinomial regression models. The elastic net regularization includes lasso regression and ridge regression where each case obtained when $\alpha = 1$ and $\alpha = 0$ respectively. In this study a generalized linear model is fitted with a penalised maximum likelihood assuming a binary response (y). This is done by minimising the following objective function in terms of the β parameters across a grid of α and λ values (Jayawardana, 2016) using inputs x:

$$\left(-\frac{1}{N} \left[\sum_{j=1}^{N} y_j (x_j^T \beta - \log(1 + \exp(x_j^T \beta) \right] + \lambda \left[\frac{(1-\alpha)}{2} ||\beta||_2^2 + \alpha ||\beta||_1 \right] \right)$$

According to Friedman *et al.* (2010b) this elastic net regularization is very effective when p $\gg$ N, where p denotes the number of variables and N denotes the sample size and when there are many correlated predictor variables. λ is the tuning parameter which controls the overall strength of the parameter penalties.

Ridge regression, shrinks the coefficients of the correlated predictors closer allowing them to borrow strength from each other (Friedman *et al.*, 2010b). However, the lasso is different from ridge regression, in that it tends to pick only one of the correlated predictors, ignoring the rest. The elastic net penalty may therefore force many coefficients to be close to zero and a small subset of coefficients to be non-zero and large (Friedman *et al.*, 2010b).

2. Random Forest

A random forest is a classifier which consists of a collection of tree structured classifiers $\{h(x, \theta_k), k = 1, ...\}$ where the $\{\theta_k\}$ are independent and identically distributed random vectors and each tree casts a unit vote for the most popular class at input x (Breiman, 2001). This method chooses from a random selection of variables to split the data at each node for each classifier with a random selection of data for each tree. This yields a more robust model with respect to noise.

4.1 Data cleaning

To differentiate between 'person of interest' profiles and 'genuine' profiles, initial T1 survey responses and a three summary statistics for step count performance data were used. From the initial survey responses, participants' current state of physical and psychological well-being were extracted along with demographics. Reasons for joining the program, provided in the T1 survey, were used to produce the profile segmentation of participants. In addition, the personal best step count (PersonalBest), average step count for valid entries (StepAverage) and average step count per entry (StepAveragePerEntry), where invalid step counts are also considered, were used.

The data set had to be cleaned since there were observations with missing values. Originally the dataset included 94,776 observations. The person of interest to genuine participant ratio in the original data set was highly imbalanced. There were 1,691 (1.78%) persons of interest while 93,085 (98.22%) participants were categorised as genuine participants. It was decided to use only the complete observations for model building. Once the data set has been cleaned by extracting the complete observations, persons of interest were reduced to 687 (1.45%) and genuine participants were reduced to 46,720 (98.55%). This highly imbalanced data set provides challenges when building a classification model.

As in section 3.1, once the complete observations were extracted, Likert scale question responses under each core field were averaged, so as to take an index representing the respective core field for each participant.

4.2 Model building procedure for person of interest detector

The cleaned data set with a ratio of 1.45:98.55, for the persons of interest to genuine participants ratio, was split 70:30 for training and testing purposes respectively, while retaining the same class imbalance that existed in the original data set in both these data sets. In order to overcome the highly imbalanced nature of the data set in model training, the Synthetic Minority Over-sampling Technique (SMOTE) (Chawla *et al.*, 2002) was used. The SMOTE technique creates synthetic minority class examples in order to oversample the minority class rather than oversampling with replacement of the existing minority class observations (Chawla et al., 2002). The synthetic examples for the minority class are generated as follows. The algorithm first takes a distance between a feature vector (of the sample in consideration) and its nearest neighbour. Then this difference is multiplied by a random number between 0 and 1 where the random distance derived by the multiplication is used to select a random point along the line segment between these

two specific observations (Chawla et al., 2002). This process will force the decision region of the minority class to become more general (Chawla et al., 2002).

Ten-fold cross validation was used for model evaluation and repeated five times. The ten-fold cross validation resampling technique was done for the 70% training split in order to choose the best model. The best cross validated model was then tested using the 30% split of test data set. Moreover, the study compared the performance on various person of interest to genuine participant ratios for the SMOTE algorithm. The study tested the penalized logistic regression model from the `glmnet R package` as well as a random forest from the `randomForest R package` (Liaw and Wiener, 2002). Both these models were trained using the `caret package` (Kuhn, 2008) in `R`. The three custom oversampling and under sampling percentages used for the SMOTE algorithm in the training stage are as follows.

i SMOTE with a person of interest : genuine ratio of 1:1

ii SMOTE with a person of interest : genuine ratio of 2:3

iii SMOTE with a person of interest : genuine ratio of 3:2

For the penalized logistic regression model fitted using the `caret package`, the predictors were standardized using z scores inside the 'train' function. A random grid search of the penalty parameters was used to find the optimum model using cross validation. It was found that the optimum model was obtained when alpha equals one, which means that only the lasso component of the elastic net was required (Friedman *et al.*, 2010b). The selected lambda values varied depending on the SMOTE subsamples within the model training process. The other classification model that was tested was the random forest model with a manual and an automatic grid search for initial parameter estimates. For the manual grid search the number of input variables considered for splitting at each node (mtry) covered the range 3 to 15 using the *tuneGrid* argument, whereas in the automatic grid search the number of input variables considered for splitting at each node was randomly selected from 20 random values using the *tuneLength* argument.

4.3 Person of interest detector model evaluation

Goodness of fit for both these methods was evaluated using a 2×2 confusion matrix. Moreover, different model evaluation measures such as accuracy, sensitivity, specificity (Altman and Bland, 1994) and area under the curve(AUC) were derived for model comparison.

Accuracy:
Accuracy measures the proportion of correctly classified observations out of the total number of observations (Metz, 1978), when estimated probabilities of above 0.5 are used to classify person of interest.

Area under the curve (AUC):
The Receiver Operating Characteristics (ROC) graph is a technique for visualising classifier based performance (Fawcett, 2006). ROC graphs are drawn in a two dimensional plane with $sensitivity$ plotted on the y axis and $(1 - specificity)$ plotted on x axis. This graph denotes the trade off between the benefits (sensitivity) and costs $(1 - specificity)$ in response to all possible positive probability cut-points (Fawcett, 2006).

Area Under the ROC Curve (AUC) is a single value which can be used to compare ROC curves derived from different prediction models. It is equivalent to the probability that the classifier ranks a randomly chosen positive observation higher than a randomly chosen negative observation (Fawcett, 2006).

5 Methods for detecting outliers

5.1 The Abnormal Activities Detector (TAAD)

As explained above, identification of outlier step entries is a huge challenge faced by the program administrators of physical activity programs. Customer service attendants of these physical activity programs check and, if necessary, they remove these data points manually. However, this process is very subjective and highly time consuming. To alleviate this problem, we have developed an automated tool known as 'the abnormal activities detector' (TAAD) using the `R Shiny` (Chang *et al.*, 2018) environment to detect multiple outliers in step entries, while allowing the customer representatives to investigate the detected outliers more effectively and efficiently. This application includes two main parts namely a 'descriptive analysis window' and a 'panel of statistical methods'. Figure 2 illustrates the interface of the TAAD application.

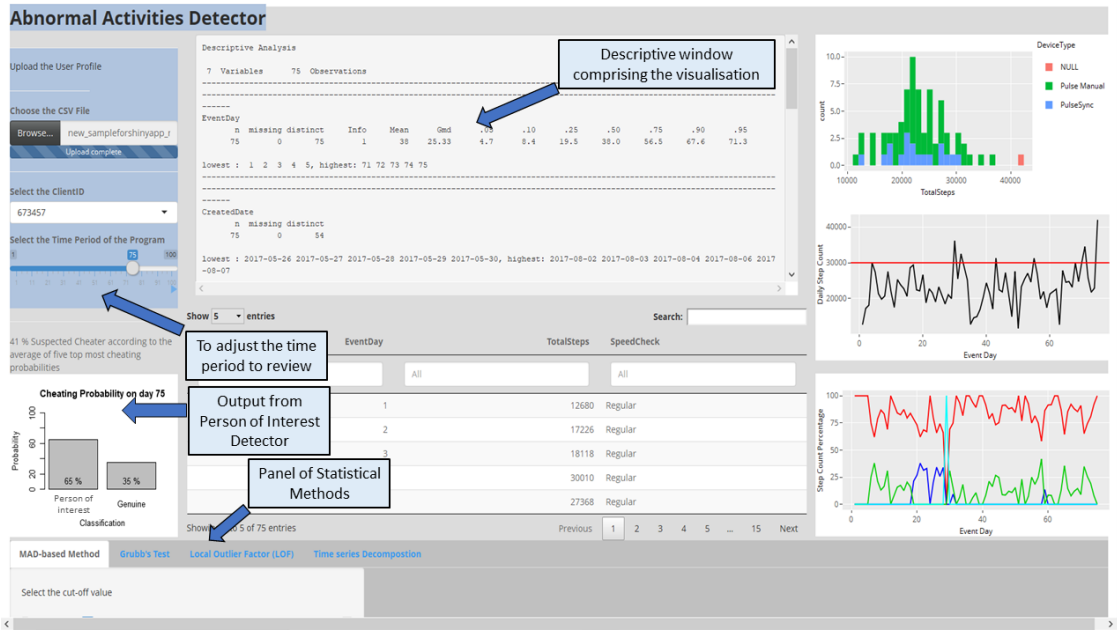

Figure 2 The abnormal activities detector interface.

The 'Descriptive analysis window' includes plots to visualise the step count distributions and trajectories of participants, along with the descriptive statistics of each and every main physical activity (e.g: swimming, cycling etc), up to the current time point. This will provide the customer service attendant with comprehensive background information about the participant's physical activity. This window also indicates whether this profile belongs to a suspected 'person of interest' as identified by the previous 'person of interest detector' classifier.

Figure number 3 shows a set of screenshots of the 'Panel of statistical methods', which allows the customer service attendants to analyse the anomalies/outliers that exist in each participant's profile using four main statistical methods, each possessing different strengths, as described below.

i A Median Absolute Deviation Method

ii Grubb's test

iii Local Outlier Factor

iv Timeseries Decomposition

Figure 3 Screen shots of all the tabs in statistical panel for detecting outliers.

5.2 Median absolute deviation (MAD) method

A common practice in outlier detection is to use an interval spanning over the mean plus or minus three standard deviations. This method has high sensitivity to outliers (Leys *et al.*, 2013). Median Absolute Deviation is a more robust method. According to Huber (2011) Median Absolute Deviation can be defined as follows where M_N is the median of the original series of x_i's and M_i is the median of the absolute values of x_i's deviations from the original median.

$$\text{MAD} = bM_i(|x_i - M_N|)$$

where x_i represents i^{th} of N original observations. Under the normality assumption b equals a constant of 1.4826, disregarding the effect of outliers (Rousseeuw and Croux, 1993). More generally the value of b equals $1/Q(0.75)$ where $Q(0.75)$ is the 75th percentile of the distribution (Leys *et al.*, 2013). The decision rule for detecting outliers is given by;

$$\left|\frac{x_i - M_N}{\text{MAD}}\right| > 3$$

where the threshold value (3) can be adjusted depending on the researcher's stringency criteria. The median is a robust measure of central tendency in the presence of outliers. Under this method the step counts of the specific participant will be arranged in ascending order, so that the time component of the step count entries will be ignored when detecting outliers. This window of the 'Shiny application' has enabled the users to select a cut-off value depending on their requirements. In the current physical activity program context, a value of three is recommended as the threshold to define the decision rule.

5.3 Grubb's test

Grubb's test which is also known as extreme studentized deviate test (Grubbs, 1969), is a procedure for determining whether the highest observation, the lowest observation, the highest and the lowest observations, the two highest observations, the two lowest observations or more of the observations are possible outliers in a sample. In Grubb's test the hypotheses will depend on the above testing objectives. For example, for a two sided test the hypothesis might be.

H_0 : There are no outliers in the dataset

H_1 :There is exactly one outlier in the data set

In this case the test statistics is defined as

$$G = \frac{max|X_i - \overline{X}|}{s}$$

Where $\overline{X}$ is the sample mean and s denotes the standard deviation. For a two sided test, the null hypothesis will be rejected for a sample size N when

$$G > \frac{(N-1)}{\sqrt{N}}\sqrt{\frac{(t_{\alpha/(2N),N-2})^2}{N-2+(t_{\alpha/(2N),N-2})^2}}$$

when $t_{\alpha/(2N),N-2}$ denotes the critical value of the t distribution with $(N-2)$ degrees of freedom and a significance level $\alpha/(2N)$ (Grubbs, 1969).

Grubb's test is capable of detecting the most outlying observations in a univariate series. Therefore, this window can be used to detect the most outlying step entries one by one. In this application this test has been coded so that the most outlying entry detected will be flagged and removed before the next cycle for finding the most outlying entry begins. However the flagged outliers are not removed from the visualisation plot, allowing the customer service attendants to view the entire step entry history for each participant. The application enables the user to define a user specific significance level to detect the outliers, with a value of $\alpha = 5\%$ recommended for this application.

5.4 Local outlier factor (LOF)

The LOF method provides a local outlier factor, which estimates the likelihood of an outlier for each observation in the data set (Breunig *et al.*, 2000). This factor depends on how isolated a specific observation is with respect to its neighbourhood. In other words this method evaluates each observation's uniqueness depending on the distance from the k nearest neighbours (Breunig *et al.*, 2000). This method is capable of detecting outliers regardless of the distribution of the data.

This method is a density based outlier detection method which relies on a nearest neighbour search. In this application the LOF method will compute a LOF value for each and every step entry of a user's profile, indicating the degree to which each point is an outlier. In the current context this score is computed using daily fluctuating variables from each participant's profile, namely 'total step count', 'whether that specific entry has been edited before entering', 'created date' etc. To enhance the user-friendliness of this application, these scores have been visualised using a plot. The application allows the user to select a cut off value for the LOF scores in order to identify outliers. In the application, the user can also define the number of nearest observations (k) used in finding the local neighborhood for the LOF score calculations. According to Breunig *et al.* (2000) the standard deviation of the LOF only tends to stabilize when the local neighborhood size is at least 10 for Gaussian distributions. Therefore, for the current study k=10 is recommended as the minimum value when defining the local neighbourhood size.

5.5 Seasonal Decomposition Method

Time series decomposition is a technique used to decompose a time series into latent subseries such as a trend component, cyclic component, seasonal component and irregular component. This is a popular technique in time series analysis. This technique can be used to detect the outliers in a time series by using the residual series once the trend and seasonality have been removed. In the current study this has been tested using two decomposition methods, a Seasonal Trend decomposition using Loess smoothing (STL) (Cleveland *et al.*, 1990) and a twitter method which decomposes the trend component using a piecewise median approach (Vallis *et al.*, 2014).

In the STL method, seasonality is initially extracted after smoothing the original series using LOESS (Cleveland, 1979). Then the estimated seasonal component is subtracted from the original series and the remaining series is smoothed by LOESS in order to extract the trend component. This procedure is repeated until convergence occurs.

In the 'twitter' method, seasonality extraction is similar to the STL method, whereas the piecewise median will be used with non overlapping windows to extract the long term trend in the original time series (Vallis *et al.*, 2014).

Once the decomposition is completed and the residual series computed, anomaly detection is carried out using the inter quartile range (IQR) or generalized extreme studentized deviate test (GESD) (Rosner, 1975, 1983).

In the Shiny application the user can select either the Twitter or STL seasonal decomposition methods. Once the decomposition is complete the user can select either one of the IQR and GSED 'anomalize' methods to detect the anomalies in the residual series of the step counts.

In our context the physical activity of the participants is assumed to have a weekly seasonal pattern. This window will provide the users of the application with a clear idea about the seasonality and trend for the physical activity patterns of all participants. This information will enhance the decision making of the program administrators.

These four main methods have different strengths as described below.

It is expected that the above set of statistical methods will look differently at each participant's step count profile. The MAD method, Grubb's test and four seasonal decomposition methods use a univariate approach while the LOF method is a multivariate approach, making the LOF method more powerful. The MAD and Grubb's test consider only the distance between the point of interest and the median or mean step counts, respectively, when identifying outliers. However, in the seasonal decomposition methods the

trend and seasonality component of the participant's step counts are also considered. Since participant step counts tend to exhibit weekly seasonality, it can be said that seasonal decomposition methods are also very important particularly in the context of physical activity programs. In the application there are two different seasonal decomposition methods (STL and twitter) and two different methods for detecting anomalies (GESD and IQR). Therefore, the user can select either of the combination (STL with GESD or STL with IQR or twitter with GESD or twitter with IQR) to detect outliers. In comparison to the above set of methods, the LOF method considers the distance for a particular observation from its k-nearest neighbours, where the density of the neighbourhood is defined by a set of features, including total step counts of participants, step entry created date, event day of the program and whether the step entry is being edited before entering. Therefore, the LOF method considers each observation in a multidimensional plane in order to detect whether that particular step count is an outlier or not. It is clear that these methods will view each participant step count differently when deciding if any step count entry is an outlier, thereby reducing the subjectivity of the process. Finally the TAAD provides a confirmatory indication for the application user about each participant step count being an outlier, based on the results obtained from the methods selected.

6 Results

This section describes the results obtained for person of interest detector model and outlier detection methods respectively.

6.1 Person of interest identification

As mentioned in the methods section 4, a penalised logistic regression and a random forest were trained using different person of interest:genuine ratios. Table 2 provides the summarised results for model performance.

Table 1 Goodness of fit measures for test data

	Random forest (automatic grid search)			Random forest (manual grid search)			Penalized logistic regression-Lasso		
Person of interest:Genuine Ratio / Measure	3:2	1:1	2:3	3:2	1:1	2:3	3:2	1:1	2:3
On training data									
Hyperparameters Selected	mtry = 22	mtry = 15	mtry = 25	mtry = 14	mtry = 14	mtry = 13	α=1 λ=6.17 $\times 10^{-5}$	α=1 λ=1.12$\times 10^{-4}$	α=1 λ=5.11$\times 10^{-5}$
ROC	0.9163306	0.9250101	0.9326848	0.9151263	0.9237783	0.9304311	0.9917162	0.9922007	0.9919274
On testing data									
Accuracy	0.8112	0.8631	0.8986	0.8044	0.8616	0.8961	0.9727	0.9763	0.9786
Sensitivity	0.87117	0.8589	0.82822	0.8589	0.81595	0.77301	0.96319	0.95706	0.93865
Specificity	0.81026	0.8632	0.89964	0.80362	0.86228	0.89798	0.97288	0.97657	0.97915
AUC	0.8407	0.8611	0.8639	0.8313	0.8391	0.8355	0.968	0.9668	0.9589
95% Confidence Interval	(0.8147-0.8668)	(0.8341-0.888)	(0.8348-0.8931)	(0.8042-0.8583)	(0.8091-0.8691)	(0.8031-0.8679)	(0.9535-0.9826)	(0.9511-0.9825)	(0.9404-0.9774)

At the training stage, the optimal model for each technique was selected from 10 fold cross validation repeated 5 times, using the 'AUC' metric to evaluate the fit. Then the other performance measures including AUC were used to compare the models on testing data. It is clear that according to the AUC, the best model fit was obtained with the penalised logistic regression-Lasso, when the SMOTE subsample contains 60% persons of interest. The penalised logistic regression lasso is therefore chosen to flag the suspected persons of interest.

6.2 Combined system for person of interest / outlier detection

In the current system, all the step entries which exceed a pre-defined cut-off value (e.g. 30,000 steps) require a valid explanation from the participant. This is clearly not a personalised approach as mentioned in section 2. According to the random sample of complete cases (participants who complete the program with at least 100 valid step entries), outlying entries (e.g. $\geq$ 30,000 steps) account for nearly 4.4% of all step entries. However, from this sample of detected anomalies, only 2.8% of step entries, were rejected by the program administrators. As mentioned above, this review process is very subjective and labour intensive. Therefore, an automatic system is proposed in which the entries which are anomalies will be filtered using one of the best methods out of the seven statistical methods compared in table 3. Moreover, these seven statistical methods were also included in TAAD to reduce the subjectivity in decision making for outlier detection.

Table 2 Outlier detection percentage.

Method	Entries required to be reviewed (out of 55,998)	Higher than the cut-off		
		$\geq$20,000	$\geq$30,000	$\geq$40,000
Step counts ($\geq$20,000)	8,950 (16%)			
Step counts ($\geq$30,000)	2,468 (4%)			
Step counts ($\geq$40,000)	712 (1%)			
	Higher than the median			
Median absolute deviation method	1,864 (3.3%)	1,317 (2.4%)	588 (1.1%)	261 (0.5%)
Grubb's test	740 (1.3%)	520 (0.9%)	263 (0.5%)	131 (0.2%)
Local outlier factor method (cut off value set at 2)	1 698 (3.0%)	1 028 (1.8%)	435 (0.8%)	181 (0.3%)
Time series decomposition (Twitter method plus GESD)	2,234 (4.0%)	1 398 (2.5%)	598 (1.1%)	242 (0.4%)
Time series decomposition (STL Method plus IQR)	914 (1.6%)	667 (1.2%)	330 (0.6%)	162 (0.3%)
Time series decomposition (Twitter method plus IQR)	713 (1.3%)	554 (1.0%)	283 (0.5%)	134 (0.2%)
Time series decomposition (STL Method plus GESD)	2,404 (4.3%)	1,491 (2.7%)	619 (1.1%)	244 (0.4%)

It is clear from the first three rows of table 3 that when the cut-off value for the step entries is increased, the system required fewer step entries to be reviewed. However, setting the same cut-off value for everyone is not appropriate when the physical activity of participants vary depending on lifestyle and other factors. Moreover, this approach creates much workload for the customer service attendants, with valuable labour hours used for reviewing many 'genuine step entries'. Furthermore, some participants would be reluctant to record explanations for each step entry which is higher than the cut-off value (e.g. 30,000 steps). This can also affect the decision making process when the subjectivity of the reviewer affects the acceptance or rejection of step entries. Therefore it is recommended that more personalised approaches be used for finding the anomalous entries.

All the other methods mentioned in the table 3 use a personalised approach based on the historical data for each participant to identify anomalous step entries, using appropriate cut-off values. For the current study, only outliers with values above the median are of interest, so only these results are reported in Table 3. It is found that in comparison to the current methods all these methods reduce the number of entries which need to be reviewed by the consumer service attendants. The Grub's test and twitter methods with IQR have reduced the number of entries to be reviewed more than the other methods.

Table 4 compares the results of the current manual system of reviewing with a cut-off value of 30,000 steps, with the proposed automatic outlier detection methods.

It is clear from Table 4 that the twitter method with GESD detects most of the rejected participant step entries, with the STL method with GESD also mirroring these results quite closely. It should be noted that this simulation was run retrospectively only for participants who entered at least 100 step counts, whereas,

Table 3 Decisions for each detected anomalous step entry

Method	Entries required to be reviewed (out of 55,998)	Decision of the reviewing process		
		Regular	Approved	Rejected
Step counts ($\geq$30,000)	2,468 (4%)	2,216 (89.7%)	184 (7.5%)	68 (2.8%)
	Higher than 30,000			
Median absolute deviation method	588 (1.1%)	430 (19.4%)	117 (63.6%)	41 (60.2%)
Grubb's test	263 (0.5%)	167 (7.5%)	68 (37.0%)	28 (41.1%)
Local outlier factor method (cut off value set at 1.6)	612 (1.1%)	464 (20.9%)	108 (58.7%)	40 (58.8%)
Time series decomposition (Twitter method plus GESD)	598 (1.1%)	437 (19.7%)	116 (63.0%)	45 (66.2%)
Time series decomposition (STL Method plus IQR)	330 (0.6%)	220 (42.9%)	79 (10.0%)	31 (45.6%)
Time series decomposition (Twitter method plus IQR)	283 (0.5%)	187 (8.4%)	69 (37.5%)	27 (39.7%)
Time series decomposition (STL Method plus GESD)	619 (1.1%)	462 (20.8%)	113 (61.4%)	44 (64.7%)

in the current process the anomalous step entries are detected immediately when the participant enters his/her step entries. This puts the automatic outlier detection methods at a disadvantage when testing.

Therefore the twitter method with GESD can be embedded in the mobile application where the anomalous step entries will be detected when the participants enter step counts. For the flagged anomalous step entries, participants have to enter a reason which will be directed to the TAAD along with the 'person of interest probability'. Then this will be evaluated utilizing the methods presented in table 3. Finally, the customer service attendants will make the decision to accept or reject these suspected entries based on all the information provided.

All the above mentioned methods have been implemented using the `R shiny` environment. This environment will provide the reviewer with a comprehensive visualisation environment to facilitate a better quality decision. This process will be more personalised for participants than the current system. It will enhance the filtering process for anomalous step entries, taking into account the past performance and behaviour of participants.

This application provides outlier confirmation for the customer service representatives. This will be helpful for the review of step entries which have been flagged based on past performance and the reasons provided. The application visualises the participant step count distribution and trajectories, along with the contribution of other physical activities, providing a platform which is a decision support system for the customer service attendants. The application has allowed the user to enter the user specific parameter values for each method. For example, in the 'time series decomposition' window, the user can select the desired decomposition method along with the seasonal and trend adjustment parameters. Moreover, the user can also select the desired outlier detection method for the decomposed series. Greater flexibility will allow the reviewer to have a closer look at suspected anomalous entry. Furthermore this application is interactive, instantly responsive to the decisions taken by the user.

7 Discussion

Many physical activity programs allow the participants to track their physical activity using wearable devices such as accelerometers, pedometers etc. Step counts are therefore one of the main outcome variables in these programs. However, there are a few participants who exaggerate their step counts and provide incorrect information. These participants can manipulate the leaderboards and this may demotivate other participants and distort the overall program outcomes. So, detecting these participants as well as detecting their fraudulent step entries is important. This study has proposed a process to achieve these objectives, for our industry partner.

The 'person of interest detector' estimates the probability of a fraudulent step entry series based on the initial (T1) survey responses and summarised longitudinal performance data. This can provide an initial classification that can be applied for every participant after the 5^{th} day of the program. The probability of being a fraudulent participant is updating every day after this using continually updated performance data. In other words, the classifier captures the current status of the participant depending on his/her behaviour. The lasso regularized penalised GLM is used for model training and the imbalanced nature of the data was addressed using the SMOTE technique, which allows a similar proportion for both 'person of interest' and 'genuine' profiles at the training stage. Person of interest probabilities are linked to the 'abnormal activities detector', in order to enhance the decision making process for rejecting or accepting anomalous step entries.

The proposed system will save many valuable labour hours by filtering out the most anomalous step entries. Since this process uses participants' intentions and characteristics as well as latest step count entries to identify fraudulent profiles, the system will provide more accurate information. This will lead to reduced subjectivity in the decision making process while boosting its efficiency. Furthermore, the 'abnormal activities detector' provides more flexibility for the users to adjust the parameters of the methods employed, enabling more interactive decision making. This will further enhance the effectiveness of the decision making process compared to the current system.

Currently the new system can be used to support the decision making process to approve or reject abnormal step entries. As future work it is expected that this system can be improved, so that the decision making can be conducted without human interaction. This could be achieved through text mining and sentiment analysis, making the process more efficient and unbiased. However, it is important to ensure that the system does not dehumanise the experience of participants.

This process and framework have been structured mainly for physical activity programs which allow the participants to enter their physical activities during the Global Challenge. However, this method could also be used for other fraud detection activities such as monitoring of financial transactions, which requires close human monitoring and intervention to identify anomalies and then to decide which of these transactions would be approved or rejected.

8 Conclusion

The study has proposed a new system which is comprised of two sub models to detect fraudulent participant profiles and step entries in a physical activity program. The first model of the system which is the 'person of interest detector' has been trained using penalised logistic regression with lasso regularisation. This technique had performed well compared to random forests. The highly imbalanced nature of the original data set of genuine to person of interest profiles was overcome by the SMOTE technique for model training. The 'person of interest detector' has 97% accuracy along with 97% sensitivity in predicting the person of interest profiles, using summarised longitudinal step count data and the initial (T1) survey responses, including demographics and reasons for joining the program. In the 'abnormal activities detector', when an outlying step entry is detected the mobile application will request a valid reason for that specific entry. These entered reasons will then be reviewed by the customer assistants, along with the computed probability of being a person of interest and previous step count data. This process will make the decision making process more efficient and productive compared to the existing system. This will substantially reduce the workload for the customer service attendants in these types of physical activity program.

Acknowledgements The authors gratefully acknowledge the Virgin Pulse-Global Challenge for providing the data to test the proposed system. Moreover, the authors gratefully acknowledge funding received from the Virgin Pulse-Global Challenge to cover the publication costs of this paper.

Conflict of Interest

The authors have declared no conflict of interest.

Appendix

References

Altman, D. G. and Bland, J. M. (1994). Diagnostic tests. 1: Sensitivity and specificity. *BMJ: British Medical Journal* **308**, 1552.

Breiman, L. (2001). Random Forests. *Machine Learning* **45**, 5–32. URL http://search.proquest.com/docview/757027982/.

Breunig, M. M., Kriegel, H.-P., Ng, R. T., and Sander, J. (2000). LOF: identifying density-based local outliers. In *ACM sigmod record*, volume 29. ACM, pages 93–104.

Chandola, V., Banerjee, A., and Kumar, V. (2009). Anomaly detection: A survey. *ACM computing surveys (CSUR)* **41**, 15.

Chang, W., Cheng, J., Allaire, J., Xie, Y., and McPherson, J. (2018). *shiny: Web Application Framework for R*. URL https://CRAN.R-project.org/package=shiny, r package version 1.2.0.

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research* **16**, 321–357.

Cleveland, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the American statistical association* **74**, 829–836.

Cleveland, R. B., Cleveland, W. S., McRae, J. E., and Terpenning, I. (1990). STL: A Seasonal-Trend Decomposition. *Journal of Official Statistics* **6**, 3–73.

Ertöz, L., Steinbach, M., and Kumar, V. (2004). Finding topics in collections of documents: A shared nearest neighbor approach. In *Clustering and Information Retrieval*. Springer, pages 83–103.

Eskin, E., Arnold, A., Prerau, M., Portnoy, L., and Stolfo, S. (2002). A geometric framework for unsupervised anomaly detection. In *Applications of data mining in computer security*. Springer, pages 77–101.

Ester, M., Kriegel, H.-P., Sander, J., Xu, X., *et al.* (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In *Kdd*, volume 96. pages 226–231.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern recognition letters* **27**, 861–874.

Friedman, J., Hastie, T., and Tibshirani, R. (2010a). Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software* **33**, 1–22. URL http://www.jstatsoft.org/v33/i01/.

Friedman, J., Hastie, T., and Tibshirani, R. (2010b). Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of statistical software* **33**.

Grubbs, F. E. (1969). Procedures for detecting outlying observations in samples. *Technometrics* **11**, 1–21.

Guha, S., Rastogi, R., and Shim, K. (2000). ROCK: A robust clustering algorithm for categorical attributes. *Information systems* **25**, 345–366.

Huber, P. J. (2011). Robust statistics. In *International Encyclopedia of Statistical Science*. Springer, pages 1248–1251.

Ismo, K. *et al.* (2004). Outlier detection using k-nearest neighbour graph. In *null*. IEEE, pages 430–433.

Jayawardana, K. (2016). Prognostic methods for integrating data from complex diseases. Ph.D. thesis, The University of Sydney.

Kuhn, M. (2008). Building Predictive Models in R Using the caret Package. *Journal of Statistical Software, Articles* **28**, 1–26. URL https://www.jstatsoft.org/v028/i05.

Last, M. and Kandel, A. (2001). Automated detection of outliers in real-world data. In *Proceedings of the second international conference on intelligent technologies*. pages 292–301.

Leys, C., Ley, C., Klein, O., Bernard, P., and Licata, L. (2013). Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. *Journal of Experimental Social Psychology* **49**. URL http://search.proquest.com/docview/1347679370/.

Liaw, A. and Wiener, M. (2002). Classification and Regression by randomForest. *R News* **2**, 18–22. URL https://CRAN.R-project.org/doc/Rnews/.

Mendenhall, W., Reinmuth, J. E., Beaver, R. J., and Beaver, B. M. (1993). Statistics for management and economics. Technical report, Duxbury Press Belmont, CA.

Metz, C. E. (1978). Basic principles of ROC analysis. In *Seminars in nuclear medicine*, volume 8. Elsevier, pages 283–298.

Pincus, R. (1995). Barnett, V., and Lewis T.: Outliers in Statistical Data. J. Wiley & Sons 1994, XVII. 582 pp.,£ 49.95. *Biometrical Journal* **37**, 256–256.

Rosner, B. (1975). On the detection of many outliers. *Technometrics* **17**, 221–227.

Rosner, B. (1983). Percentage points for a generalized ESD many-outlier procedure. *Technometrics* **25**, 165–172.

Rousseeuw, P. J. and Croux, C. (1993). Alternatives to the median absolute deviation. *Journal of the American Statistical association* **88**, 1273–1283.

Simon, N., Friedman, J., Hastie, T., and Tibshirani, R. (2011). Regularization Paths for Cox's Proportional Hazards Model via Coordinate Descent. *Journal of Statistical Software* **39**, 1–13. URL `http://www.jstatsoft.org/v39/i05/`.

Song, X., Wu, M., Jermaine, C., and Ranka, S. (2007). Conditional anomaly detection. *IEEE Transactions on Knowledge and Data Engineering* **19**, 631–645.

Vallis, O., Hochenbaum, J., and Kejariwal, A. (2014). A Novel Technique for Long-Term Anomaly Detection in the Cloud. In *HotCloud*.

Zhang, J. and Wang, H. (2006). Detecting outlying subspaces for high-dimensional data: the new task, algorithms, and performance. *Knowledge and information systems* **10**, 333–355.

