Business Process Automation: Automating the Analysis of Anomaly Data

Tristan Nolan [1], Enda Fallon [1], Paul Connolly [2], Kieran Flanagan [2]

[1] *Software Research Institute,* Athlone Institute of Technology, Athlone, Ireland.
[2] *The NPD Group, Inc,* Athlone, Ireland.

Email: A00242605@student.ait.ie; efallon@ait.ie; paul.connolly@npd.com; kieran.flanagan@npd.com

***Abstract* - This research proposes a method which evaluates a real-world data analytics business process to identify key performance variables for the manual business process. Once complete, we incorporate these finding to an unsupervised machine learning algorithm to allow for tuning of the outputs. Experiments show that using this approach can reduce the overall number of undesired outputs, giving an overall higher effectiveness for the ML system in a real-world application. The proposed method offers consistency while providing an organization the option of focusing resources on high value activities.**

Keywords - Automation, Business Process, Machine Learning, Data Analysis

I. INTRODUCTION

Machine learning has become more popular and is playing a pivotal role in many sectors such as healthcare [1] Karatekin et al, financial [2] Hasan et al, manufacturing [3] Wang et al and the software industry [4] Meinke et al. Currently business processes/set of tasks can have a high amount of manual input. Manual processes are often slower and struggle to scale as demand can rise and fall. Manual processes also are very prone to inconsistencies in data and vary greatly in detail due to knowledge available.

Moving to an automated business process is highly beneficial for organizations, automated systems in place of their manual counterpart would allow an organization to scale on demand and lead to higher efficiency. An automated system offers consistency and helps focus resources to high value activities due to the automation of the work.

The use of ML processes would allow organizations to migrate from activities such as mass-based inspection which require a large amount of resources. Manual inspection cannot be dependent on large amounts of historical data as there can be a large resource requirement. An ML process would reduce the resources required due to it being fully automated which would aid measuring past and current data without harsh limits that a manual process may face.

KPIs (Key performance Indicators) from the current process are identified and used in the creation of the automated system.

II. LITERATURE SURVEY AND LIMITATIONS OF CURRENT METHODS

In recent years work on automating business processes has increased dramatically due to the demand on resources that manual processes require. Traditionally business processes are mapped out and executed manually for a specific task. Manual processes are often resource intensive and may not achieve all the desired output depending on unusual use cases. There are many ways of changing a manual process to a fully-fledged automatic process. This review will investigate how people are currently doing this and highlight the limitations of the current method.

A. Supervised Learning

Supervised learning is a subcategory of machine learning [5] Pahwa et al. It uses a training set of data to reach a desired outcome. Classification and regression are a part of supervised learning. Classification can be commonly seen in image classification; this is where for example after being trained in images that are labeled correctly and provide identifiable characteristics i.e., face shape, size, eyes, it would tell you the resulting image is from a human face [6] et al. Regression is used to predict continuous values, common models for regression are linear and polynomial, typical use of regression is trend analysis.

B. Unsupervised Learning

Unsupervised learning is another subcategory of machine learning. Clustering is a part of unsupervised learning and it involves splitting data and putting them into groups with other data points which are like each other. Common use of clustering is anomaly detection as seen in previous research [7] Flanagan et al.

C. BPA/RPA Software (Business Process Automation/ Robtotic Process Automation)

The use of business process automation/robotic process automation software is one possible solution to go from a manual to automated system. The RPA/BPA software industry has seen a major rise in popularity in the last five years.

The industry is projected to reach two billion dollars in revenue by the end of 2021 while also projected to grow at double digit rates through to 2024 and beyond. RPA software can be used for a wide variety of tasks, most of which requires no coding background. RPA software is designed with highly repetitive work in mind like data entry [8] et al, As seen in prior research [9] Ortiz et al, [10] Raissa et al. they were able to use RPA software to turn their manual procedure of calculating the Sharpe ratio (measure of risk-adjusted return) for stocks in an index fund to an automated process. Using RPA software, they were able to reduce the time to calculate this significantly, manually this would take a few minutes, but with the use of RPA they were able to get that down to two minutes, in those two minutes they are scraping the internet for data, completing the calculation from the scraped data, and then sending the results of the calculations in an email.

Within RPA/BPA software there are subcategories, as seen in another research [11] Gupta et al. it can be split into four different categories, Assisted, Unassisted, Autonomous and Cognitive RPA. Assisted RPA will automate a task on a user's machine, while it reduces time it still requires human input at certain points i.e., Authentication Unassisted RPA is where tasks are deployed on several machines at once, it differs from assisted as it does not need a user to log on to start the process and unlike humans' robots can work 24/7. Autonomous RPA automates rules-based and mundane tasks. Cognitive RPA uses algorithms i.e., NLP (Natural Language Processing) which can be used in the process of making a chat bot.

D. Intracranial Aneurysm Detection using Image Classification

Intracranial aneurysms are a weakening of a vein or artery which can cause them to balloon and If left untreated and if it were to rupture could result in a fatal incident [12] Lauric et al. It is critical for doctors to be able to diagnose aneurysms before they have ruptured. Currently the process for diagnosing an aneurysm involves using an MRI, CT Scanner or DSA imaging machine [13] Nejati et al. Not all patients can use the above machines due to many circumstances, for example patients with hearing aids, implanted pacemakers, or Intracranial aneurysm clips (restricts blood flow from reaching an aneurysm) cannot use an MRI machine as it would interfere with the strong magnet within the machine. The proposed method in that paper uses 2d x-ray images from a DSA machine (Digital Subtraction Angiography machine which shows an image of blood vessels). The method at the time of its publication has a 96% accuracy of detecting aneurysms, while not 100% it can be used in conjunction with a doctor's expertise to help them diagnose a patient with aneurysms [14] Rahmany et al. A process like this would help doctors diagnose patients faster which is crucial for aneurysms.

E. Use of AI and NLP in the Automation of Chat Bots (Artificial Intelligence/Natural Language Processing)

Chat support, seen in many ecommerce businesses used to require whole teams to operate the needs of consumers who need help with a certain query. Nowadays with the use of AI (Artificial Intelligence) and NLP fewer people are needed, and human support should only be the last resort if the bot cannot solve the customer's query. Natural Language Processing is a type of Artificial Intelligence with the goal of helping machines understand how people speak and write [15] Zong et al. The automation of chat bots first began with the use of a rules-based approach, while this did work at the time, the process of implementing a system such as this is very tedious, and you must consider every case of how a customer could interact with the bot. The main issue with rules-based chat bots is they can handle simple or complex tasks but only tasks that the creator of the system has defined. If you vary from the pre-defined queries the bot would not respond. The other issue that NLP and AI chat bots solve that rules-based bots suffer from is that rules-based bots do not learn from any of the interactions they have with customers. With the use of Neural networks, the NLP bots can not only identify what words the user is typing but crucially they understand how the person is saying it. When first training an NLP chat bot, they should be monitored closely but once up and running and provided with enough training data they can give quicker results than humans and offer consistent responses to the customer. Crucially for businesses less people needed to operate support results in sizable cost savings as support teams can be scaled down.

F. Current Method: Filter Followed by Manual Process

The current method follows a manual process where users are manually inspecting records using criteria they have come up with through experience. The records are gathered from anomaly detection as seen in a prior research [16] Flanagan et al.The method in the research offers an algorithm which can detect anomalies fast and reliably. It works by using a clustering algorithm that can take in a feed of items and produces a list of anomalies based on regression analysis of the variance of items. Anomalies are something that deviates from what is expected.

The current method deployed involves using a filter on anomalies to find samples that need to be further investigated. This process is arbitrary in nature. As seen in Equation (1) below is the formula used in the current method. The formula contains three features. Each feature was selected from the dataset and classified as highly impactful by the manual process SMEs (Subject Matter Experts).

The features have a rules-based value assigned to them, all three features must exceed the value from the formula e.g. if feature 1 is greater than 10 and feature 2 is greater than 20 and feature 3 is greater than 100 it will be detected as an anomaly. If all three features do not exceed their set values the resulting output will not be detected as an anomaly.

$$Current\ Met\square od = (Feature1 > 10)\ \&\ (Feature2 > 20)\ \&\ (Feature3 > 100) \qquad (1)$$

The other problem with this filter is the manual process that comes with it, once an anomaly is found within the filter it must be investigated, the manual process involved in this is different on a per user basis due to knowledge available, some users perform differently from others so we have no consistency that an automated process could give. The filter was born out of necessity due to the ever-increasing amount of data and the need for investigating them without a filter would be too costly on resources and would not be feasible due to those constraints.

In short there are numerous amounts of ways to implement an automated business process, from the different sectors (financial, healthcare, software, and IT support) it was quite clear why it was needed to move from manual to automated processes. In this paper we propose the use of a method that in the future ML processes will help automate for the use of automating data analysis, we will compare the current process against the proposed concept and measure the effectiveness of the proposed implementation and what drawbacks need to be addressed before it can be implemented.

III. PROPOSED METHOD

In this section we will present a methodology on adapting KPI’s from a manual business process for use in an automated system.

A. Proposed Method: Filter that can be Automated in the Future

The proposed method uses an activation function as a filter. When creating the activation function, KPI’s were identified from the manual process through interaction with SMEs. Once they were identified the activation function could be made. The activation function ranks anomalies based on the change of the four KPIs from one period to another. The bigger the variation in the KPIs results in a higher activation function output. The output is ranked in descending order. Below in table I you can see the output from the activation function showing how the output is ordered.

TABLE I. ACTIVATION FUNCTION DESCENDING OUTPUT

Date	Output from Activation Function
	Descending order
06/07/20	587350
06/07/20	538197
06/07/20	499068
06/07/20	478018

Equation (2) below is the formula for how the activation function is calculated, it works by calculating the absolute of KPI one through three, then this is multiplied by KPI4 and divided by the attribute which is a feature within the dataset.

$$Activation\ Function = (|KPI1| + |KPI2| + |KPI3|) \times (KPI4 \div (Attribute\ 1)) \qquad (2)$$

This proposed method needs to be broken down into multiple steps, first we must take in all the anomalies from one time period to another, we then work out the average of the anomalies, the standard deviation of the anomalies and finally the number of anomalies that have occurred that time period. With this information we can solve the confidence intervals of 65%, 95% and 99% and with these confidence intervals we can use the average value on the corresponding confidence value to get our upper bound value. With the upper bound value, we can now create our own filter and use it for comparison against the current method. The anomalies that appeared were then exported and analyzed to find edge cases that the current method would miss out on. The next stage is the implementation of an ML process which would be automated. Once automated the next step would be to provide context to the anomalies using 3rd party data. The results at first may not be what we want when compared with the current process but moving to the new process is vital for introducing consistency to anomaly analysis.

Fig 1 below shows a model of how the current and proposed methods work. The current method takes in the incoming anomaly data, it then goes through its manual process which includes what items have been detected by the filter and then they are investigated. Once investigated the data is then prepared to be processed and then outputted.

The proposed method takes in the incoming data and goes through the automated process, the manual process is fed back into the automated process via the activation function. The calculations on the activation function are completed and this gives us the confidence values. The confidence values are used on the activation function as a filter. Only anomalies that are above the threshold will show up. The output of this filter is then exported. The exported outputs of the current and proposed methods are compared and anomalies that do not appear in the current method are then outputted. These outputted anomalies are then analyzed to see if they perform better than the manual process.

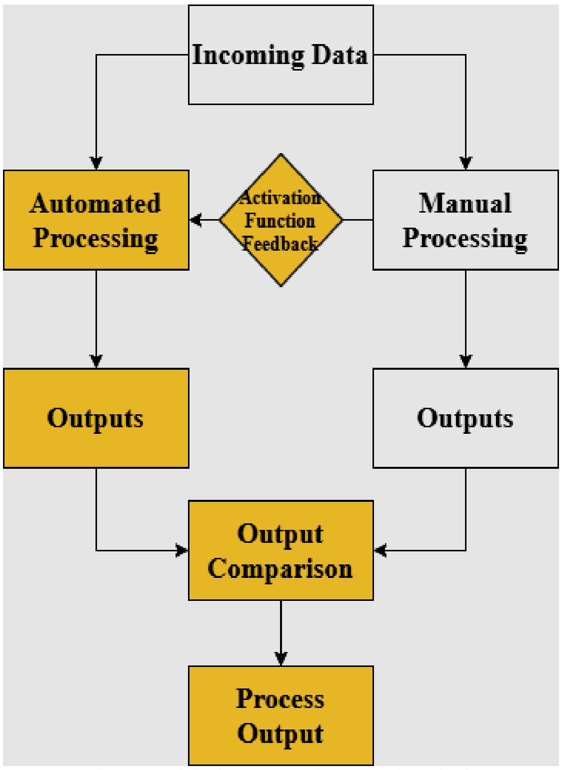

Fig. 1. Model of current and proposed methods

IV. RESULTS AND DISCUSSION

The proposed method was applied to real-world anomaly data obtained from The NPD Group, Inc. The results from solving the confidence values were split into the three different confidence intervals 65%, 95% and 99%. To work out the three intervals the confidence interval equation was used on the exported data as seen below in Equation (3).

$$CI = X \pm Zs\sqrt{n} \qquad (3)$$

X is the mean, Z is the confidence value or can also be known as an alpha value, this value can be calculated by subtracting one from the confidence level required e.g., 1 — 0.65 would result in an alpha value of 0.35, this is repeated for the three confidence intervals of 65%, 95% and 99%. s is for standard deviation, and this is put over $\sqrt{n}$ which is the square root of the sample size. Table II below shows the result of using the formula. For this table a mean of 1234, alpha value of 0.01, standard deviation of 1860 and a sample size of 7508 were used to calculate the 99% value and the upper bound value.

TABLE II. RESULTS FROM ACTIVATION FUNCTION CALCULATION

Date	Activation Function Output		
	99% Value	*99% Interval*	*Upper Bound*
05/03/20	55.29	1234 ± 55.29	1289.29

TABLE III. 99% ACTIVATION FUNCTION CONFIDENCE VALUES FROM A THREE WEEK PERIOD

Date	Confidence Value Results		
	99% Value	*99% Interval*	*Upper Bound*
27/04/20	86544	39819 ± 86544	126364
04/05/20	122286	88577 ± 122286	210863
11/05/20	12742	10949 ± 12742	23692

TABLE IV. COMAPRISON OF NUMBER OF ANOMALIES USING CURRENT AND PROPOSED FILTERS

Date	Filters			
	Current Filter	*65% Filter*	*95% Filter*	*99% Filter*
27/04/20	2	19	14	13
04/05/20	4	15	12	10
11/05/20	2	7	5	5

Table III shows the results from the calculations performed on the activation function in a three-week period. The upper bound value is being used as a filter on the activation function so that only anomalies above the upper bound value would show up. Table IV shows the number of anomalies and the comparison of the current methods filter, and the proposed methods filters over a three-week period. The current filter shows a lower number of anomalies compared to the proposed methods, but the current filter misses out on edge cases that still need to be investigated.

Comparing Anomaly Filters

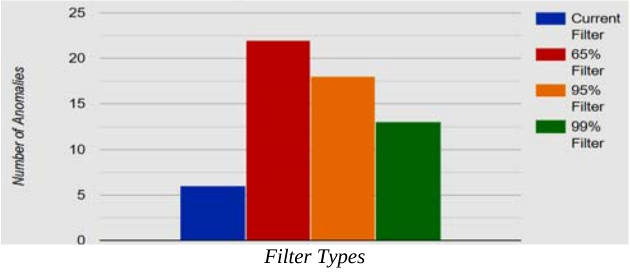

Fig 2. Anomaly filter comparison 1.

Comparing Anomaly Filters

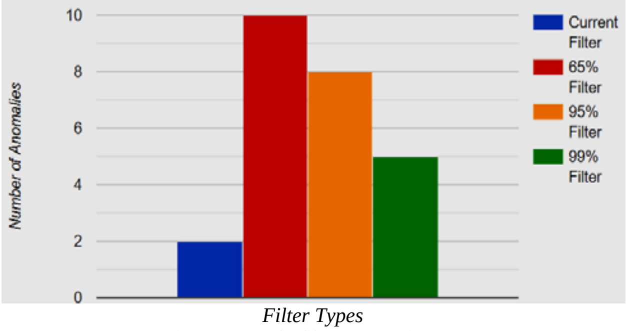

Fig 3. Anomaly filter comparison 2.

The two graphs above are from two different experimental runs, which show anomalies that were detected on real-world data from the NPD group. The number of anomalies is on the y axis and on the x axis from left to right are the following filters, Current Filter (Blue), 65% filter on the Activation Function (Red), 95% filter on the AF (Orange) and a 99% filter on the AF (green). The current filter (blue) offers the lowest number of anomalies, which is to be expected due to it taking a rules-based approach. Having a low number of anomalies is crucial as it is not feasible to investigate over 15 anomalies due to time constraints. As expected, using the confidence values on the activation function as a filter resulted in an increase of anomalies appearing with 99% confidence offering the lowest of the confidence values which again was expected. While the number of anomalies is high for one to two people to investigate fully, we are seeing more anomalies that are on the edge that the current filter cannot detect due to it being rules-based. Future work must be done on the current implementation of the activation function to solve what needs to be changed to bring the number of anomalies down so that the number of people required for analysis can be minimized. Removing the excess anomalies is the next step that needs to happen before the implementation of the ML process can take place.

V. CONCLUSION AND FUTURE WORK

This paper proposes a method to enable the transition from a manual business process to a fully automatic process. To achieve this, research focused on understanding the current manual process and then generate the proposed method. The method works by calculating confidence values on an “activation function” which is a ranking function of anomalies based on SME feedback. With the confidence values calculated; upper bound values of those calculations are used as a filter. The current and proposed methods have been compared with one another. The results from the proposed method show that the number of anomalies is higher. The higher results provide the edge case anomalies that the current method does not show due to its rules-based nature. This paper used real world data from the NPD Groups dataset. Data of over a year’s period was used to calculate confidence values for every week of the year where anomalies had occurred. To fully test the method future work will investigate whether the “activation function” as it stands needs to be changed to reduce the number of anomalies that appear whilst keeping edge cases that need to be investigated. The next stage is an automatic ML process which would calculate the confidence values and apply them automatically. This process would introduce consistency and provide the option of comparing old data with new data on a larger scale that a manual process cannot handle due to set limitations. In Future once the ML process has been implemented further work will investigate providing context to the data via 3rd party data.

REFERENCES

[1] P. K. S. S. G. C. A. O. Tamar Karatekin, "Interpretable Machine Learning in Healthcare through Generalized Additive Model with Pairwise Interactions (GA2M): Predicting Severe Retinopathy of Prematurity," International Conference on Deep Learning and Machine Learning in Emerging Applications, pp. 1-2, 2019.

[2] O. K. S. A. Afan Hasan, "Predicting Financial Market Big Data: Deep Learning," International Conference on Computer Science and Engineering, p. 1, 2017.

[3] P. M. Jane Wang, "Machine Learning for Quality Prediction in AbrasionResistant Material Manufacturing Process," IEEE Canadian Conference on Electrical and Computer Engineering (CCECE), pp. 1-2, 2016.

[4] A. B. Karl Meinke, "Machine Learning for Software Engineering," ACM/IEEE 40th International Conference on Software Engineering: Companion Proceedings, pp. 1-3, 2018.

[5] N. A. Kunal Pahwa, "Stock Market Analysis using Supervised Machine Learning," International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (Com-IT-Con), pp. 1-2, 2019.

[6] D. R. K. G. R. Senthilkumar, "Performance Improvement in Classification Rate of Appearance Based Statistical Face Recognition Methods using SVM classifier," International Conference on Advanced Computing and Communication Systems (ICACCS), pp. 1-2, 2017.

[7] E. F. A. A. P. C. Kieran Flanagan, "Network Anomaly Detection in Time Series using Distance Based Outlier Detection with Cluster Density Analysis".

[8] P. J. Sutipong Sutipitakwong, "The Effectiveness of RPA in Fine-tuning Tedious Tasks," 2020.

[9] C. J. C. Felipe C. Magrin Ortiz, "RPA in Finance: supporting portfolio management," 15th Iberian Conference on Information Systems and Technologies (CISTI) , pp. 1-4, 2020.

[10] K. Z. S. R. M. A. K. A. Uskenbayeva Raissa, "Applying of RPA in administrative processes of public administration," IEEE 21st Conference on Business Informatics (CBI), pp. 1-2, 2019.

[11] S. R. D. A. D. Saurabh Gupta, "Recent Trends in Automation-A study of RPA Development Tools," International Conference on Recent Developments in Control, Automation and Power Engineering (RDCAPE), pp. 1-3, 2019.

[12] E. L. M. M. I. B. A. M. M. Alexandra Lauric, "Rupture Status Discrimination in Intracranial Aneurysms Using the Centroid–Radii Model," IEEE TRANSACTIONS ON BIOMEDICAL ENGINEERING, vol. 58, no. 10, pp. 1-2, 2011.

[13] R. A. S. S. Mansour Nejati, "A Fast Image Registration Algorithm for Digital Subtraction Angiograph," 17th Iranian Conference of Biomedical Engineering (ICBME), pp. 1-2, 2010.

[14] R. G. N. K. Ines Rahmany, "A Fully Automatic based Deep Learning Approach for Aneurysm Detection in DSA Images," IEEE International Conference on Image Processing, Applications and Systems (IPAS), pp. 1-4, 2018.

[15] C. H. Zhaorong Zong, "On Application of Natural Language Processing in Machine Translation," 3rd International Conference on Mechanical, Control and Computer Engineering, pp. 1-2, 2018.

[16] E. F. A. A. P. C. Kieran Flanagan, "Self-Configuring NetFlow Anomaly Detection using Cluster Density Analysis," International Conference on Advanced Communications Technology - Virtual Conference, pp. 1-5, 2017.