The Application of Decision Tree Regression to Optimize Business Processes

Mike Sishi and Arnesh Telukdarie
Post Graduate School of Engineering Management
University of Johannesburg, South Africa
mike.sishi@gmail.com, arnesht@uj.ac.za

Abstract

Many organizations use business processes as a tool to realize and sustain competitive advantage in the market. A business process is a structured collection of activities with comprehensible sequence and dependency to yield a required outcome. The optimization of these processes is of paramount importance because optimized processes yield adaptability, accurate information, enhanced efficiency, accountability through performance monitoring, and improved quality. Relying on business people such as executives and management to identify areas of improvement in the business processes is potentially subjective. This research commences on the assumption that business processes are fully constituted for a business and on this premise seeks an alternate, none subjective, optimization technique. A Decision Tree (DT) is a tool that supports decision making by means of a tree-structured modeling approach to map possible outcomes of a chain of interconnected choices. When applied in statistical regression modeling, a DT model employs supervised learning techniques to model decisions in a tree structure with possible results, input costs, and usefulness. In a DT model, aspects of an element are monitored and the model is trained to predict the future. DT can be applied to improve business processes by identifying activities or elements with significant impact when enhanced. This paper demonstrates business process optimization via DT regression modeling by the use of Python programming.

Keywords
Decision Tree, Process Optimization, Python, Standard Deviation, Optimization.

1. Introduction

Organizations strive to achieve and maintain competitive market advantage by establishing business processes. These processes need to be revisited, assessed, and optimized periodically. Therefore, the process models employed need to be of high quality in order to enhance the administration of business processes (Jiménez-Ramírez et al. 2015). Traditional business process optimization procedures include modeling current processes, identifying gaps in the current processes, and creating and modeling optimized processes. The manual enhancement and modeling of business processes become challenging when input is uncertain, events are correlated, and resources are to be allocated. Moreover, the manual procedure for enhancing business processes is subject to individual experience, preferences, and discretions which might be inadequate, unsuitable, or inefficient for the business (Parody et al. 2016).

Defining, managing, and optimizing business processes is critical for the business to achieve and maintain competitive advantage. This research commences by studying existing literature pertinent to business processes and the significance of optimizing these processes.

Decision Trees (DT's) can be used for both classification and regression; however, this paper demonstrates DT regression techniques to predict future outcome (Gokgoz and Subasi 2015).

The study continues by investigating Artificial Intelligence (AI) techniques suitable for optimizing business processes. DT is one of the Data Mining (DM) tools for categorizing and predicting class variables. DT is known for its ability to utilize historical and current data to predict future outcome (Yan et al. 2016). DT is devised by recursive partitioning of data into multiple splits and fitting the prediction model with each split (De Oña et al. 2014). The splits are represented graphically as a decision tree. The DT regression method is selected due to its ability to examine the past and current conditions and predict the future outcome graphically. Having knowledge of the future outcome could enable business to fine-tune processes to yield better results.

A DT regression model is created and implemented on the dataset obtained from an electronics manufacturing company in South Africa. This dataset encompasses a number of items manufactured between the years 2013 to 2016. Subsequently, a DT regression model is devised and applied using Python programming language to examine the dataset and predict productivity for each month of the year 2017. A detailed procedure to implement DT regression via Python is presented comprehensively. Furthermore, the reverse engineering of DT regression is demonstrated to enhance Overall Equipment Effectiveness (OEE).

1.1 Business Process Definition

A business process (BP) is defined as a collection of activities executed in tandem within a company to yield a specific purpose (Jiménez-Ramírez et al. 2015). These undertakings can come in the form of manual actions or software components. Parody et al. (2016) simply define BP as a string of actions devised to yield a product or service. Insufficiently designed processes could increase inefficiency and consequently yield futility. Therefore, it is pivotal for an organization to have clearly defined process design techniques. Studies (Parody et al. 2016, Cimino and Vaglini 2014) emphasize the importance of analyzing business processes prior to commencing with production and during production time. Failure to analyze business processes could result in long cycle times, backlogs, underperforming activities, inefficient resource utilization and poor service levels, disgruntled customers, and diminished company reputation (Lohrmann and Reichert 2016).

BP's are documented using Business Process Model (BPM), a graphical description of how PB instances should be executed (Lohrmann and Reichert 2016). Figure 1 depicts a business process model for handling incoming employment applications documented based on Business Process Model and Notation (BPMN) Object Oriented Management Group standard (OMG). The example of employment application process shows that a job offer will be made only if these conditions are satisfied: (1) application documents are accepted, (2) interview has been conducted with positive results, (3) the two parties (i.e., employer and employee) have agreed on certain basic conditions, and (4) management has approved the job offer.

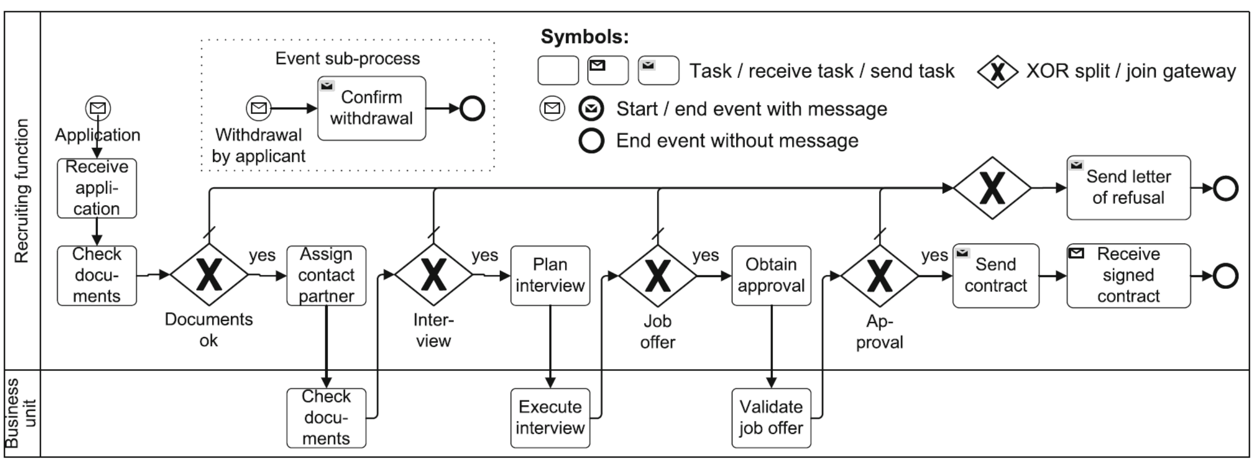

Figure 1. Business process model (Lohrmann and Reichert 2016)

1.2 Business Process Optimization

To achieve augmented management of BP's, high quality BP models should be maintained. These models could be software products for capturing correlations between business units (Weidlich et al. 2014). BP's are influenced by co-operating human activities, resources, business rules, and constrictions which are uncertain, inaccurate, variable, and dynamic in nature. The modeling of BP's is usually performed manually, which is very laborious (Clempner 2014). This becomes a challenge when input is uncertain, activities are related, and resource allocation is taken into account (Fang et al. 2018). Moreover, constructing effective business models becomes intricate when optimization is required to achieve flexibility and robustness, resulting in diminished models, inaccuracies, and absence of flexibility. BP's could be improved by devising configurable process models that accommodate the aforementioned requirements of flexibility and robustness (Weidlich et al. 2014, Fang et al. 2018). The optimization of processes should not focus only on reducing execution time or resources employed in a PB but improving the output of each case as well (Hadded et al. 2018). BP optimization enables organizations to become more competitive over their rivals in the market. Therefore, organizations embark on process optimization in order to, amongst other reasons, decrease the cost of cycle times or acquire the best product on the market to boost their competitive advantage (Montarnal et al. 2018).

2. Decision Tree

A Decision Tree (DT) is one of the tools used to underpin decision making in various types of industries. A DT uses a tree-structured modeling approach to traverse through possible results of a string of related choices (Yan et al. 2016). DT's are also used as Data Mining (DM) tools for classifying and predicting class variables. DT's construct classification or regression models using a tree structure. A classification tree and regression tree are used for discrete and continuous target variables respectively (De Oña et al. 2014, Yan et al. 2016). A dataset is partitioned into smaller subgroups that gradually develop into an associated tree with decision nodes and leaf nodes. A decision node encompasses two or more subdivisions that each represents values for the element tested. On the other hand, a leaf node denotes a decision on the arithmetic end. The node located at the top of the tree, considered to be the best predictor, is called the root node (Zhao et al. 2014). DT's can accommodate both statistical and categorical data (Candanedo and Feldheim 2016). DT's are extolled over other modeling techniques for the main advantage of being able to be presented via visual bifurcating images that clearly depict effective *If-Then* rules (Yan et al. 2016). Each leaf on a decision tree represents a decision rule that encompasses metadata. DT's depict the value and impact of independent variables on the model (Candanedo and Feldheim 2016). An example of a decision tree structure is depicted in Figure 2.

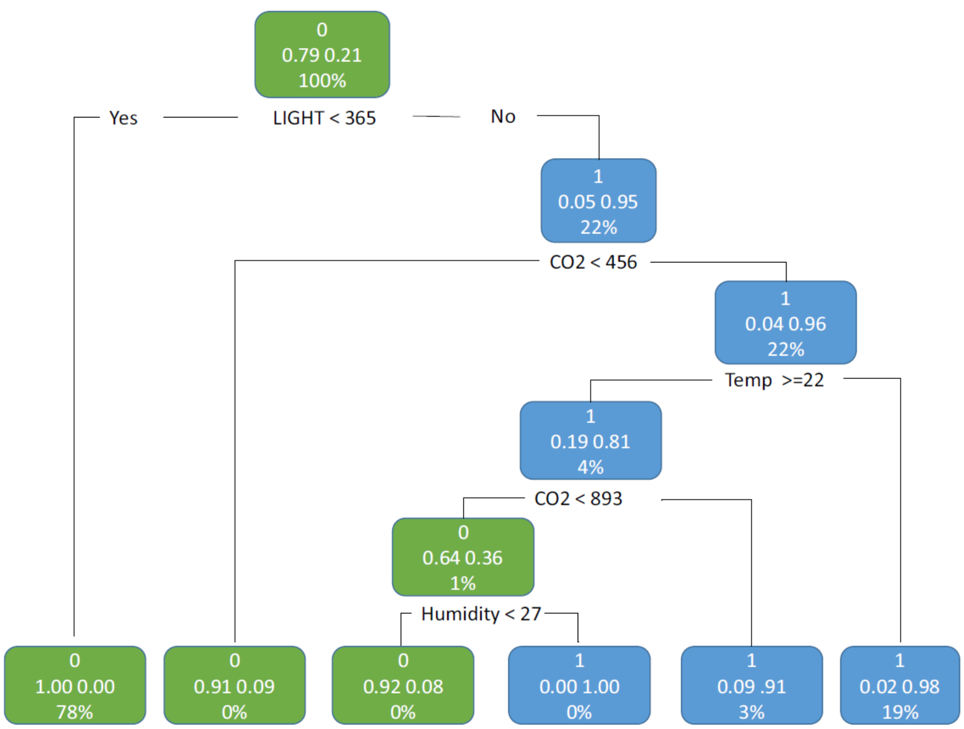

Figure 2. Decision Tree structure (Candanedo and Feldheim 2016)

2.1 Decision Tree Algorithms

Iterative Dichotomiser (ID3), invented by Ross Quinlan, is a technique used to create DT's. ID3 builds decision tress by applying Standard Deviation Reduction (SDR) (Yan et al. 2016) discussed in the succeeding subsections.

2.1.1 Standard Deviation (SD)

Standard Deviation (SD) is the square root of the variance (V), V is the average (Avg) of the squared deviations from the mean, and coefficient of variation (CV) measures the dispersal of a probability distribution. The mathematical expressions of these elements are illustrated as follows (Gokgoz and Subasi 2015, Yan et al. 2016).

Mean (Avg) $= \frac{\sum Y_i}{N}$

Variance (V) $= [\sum_{i=1}^{N} (Y_i - Y)^2] / N$

SD $= \sqrt{variance}$

Coefficient of Variation (CV) $= \frac{\text{SD}}{\text{Avg}}$ x 100%

Where mean represents the average, Y is the distribution mean, Y_i denotes each value of the distribution, $(Y_i - Y)$ is the deviation from the mean, and N is the total count of values in the distribution. The sum of squares is defined as deviations from the mean that are each squared and added together. SD is for building tree nodes, CV is a determinant for when to stop branching, and Avg forms the value of the leaf nodes. Consequently, a DT is built by dividing data, from top to bottom, into homogenous elements. SD is used to determine the homogeneity of the group of elements. The SD becomes zero (0) if the elements in the group are totally homogeneous.

2.1.2 Standard Deviation Reduction (SDR)

Standard Deviation Reduction (SDR) process is illustrated in Figure 3 as adapted from Seera and Lim (2014) and discussed in detail thereafter.

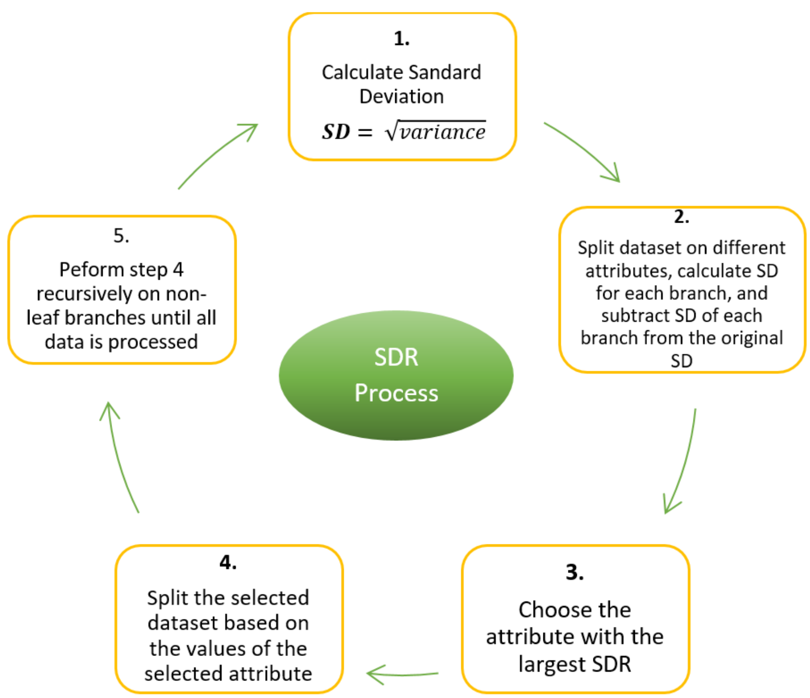

Figure 3. Standard Deviation Reduction process adapted from Seera and Lim (2014)

The reduction in SD, following the data split on an attribute (feature or input variable), is the basis of standard deviation reduction (SDR). The main aim in the DT construction is to determine the attribute with the largest reduction in SD; i.e., the mostly identical splits (Zhao et al. 2014). The SDR process commences with the calculation of SD for the target variable in a dataset. The dataset in subsequently partitioned into different attributes and an SD is calculated for each of the attributes. Subsequently, the SD for each attributed is deducted from the SD of the target variable of the entire dataset, resulting in an SDR per attribute. The decision (root) node is derived by choosing the attribute encompassing the greatest SDR. Successively, the dataset is split based on the values of the chosen root node. The aforementioned SDR process is applied repeatedly to create more nodes, exclusive non-leaf branches, until all data is processed.

3. Results

3.1 Optimizing Production with Decision Tree Regression

Suitable methods and tools are selected by the authors of this paper to demonstrate the optimization of business processes via a DT regression model. The main aim of using DT regression model in this research is to use historical production data to predict future productivity. Historical dataset is acquired from a manufacturer of electronic devices in South Africa. Raw data encompasses the number of items manufactured between the year 2013 and 2016. Figure 4 depicts a sample of the aforementioned raw data. Raw data is transmuted, with Structured Query Language (SQL), to produce production unit count for each month of year 2013 to 2016 and saved into a comma separated values (CSV) file. Python code is applied on the transformed dataset to create a DT regression model. The process commences by importing Python packages and loading the CSV file containing the dataset. Data is arranged for creating a range of values from the minimum to the maximum value of $\boldsymbol{x}$ with a deference of 0.01 between two consecutive values. Data is further reshaped to inform the model that the dataset is made of one input variable $\boldsymbol{x}$ (time in months) and one target variable $\boldsymbol{y}$ (estimated productivity). The dataset is divided into 80% of training data and 20% of testing data. Subsequently, the model is tested with test data (20%) and the results are presented in the scatter plot (Figure 5). The red dots on the scatter plot represent actual productivity while the blue line represents the predicted productivity. The results on the scatter plot reveal very good prediction accuracy by the model.

To create the predictive DT regression model, Python built-in functions are applied to process production dataset and yield mean, V, SD, and CV based on the formulas discussed in Section 1.4.1 of this paper. The SDR methodology detailed in Section 1.4.2 is followed to obtain the decision tree depicted in Figure 6. For each node of the decision tree, the value of input variable $\boldsymbol{x}$ is represented by Time (In Months), value of V is represented by mse, value of N is represented by samples, and the value of target variable $\boldsymbol{y}$ is represented by value. The actual values of the elements in the root node (top of the tree) is illustrated in Table 1. The resulting DT reveals that the model is trained to use time (in months) as the determinant of productivity. This means that each month of the year is very crucial in terms of productivity. Having the knowledge of which months have a potential of yielding low and high productivities, respectively, would help a manufacturing organization plan better. The rectangles below the root node are branches that the model uses to traverse through each month of the year to predict productivity. The rectangles with zero mse and one sample are leaf nodes that represent the end of splitting for a particular branch. As part of continuous improvement, the model can be retrained with new data if an organization notices significant changes in actual productivity over time.

id_production_event	pcb_num	prod_id	id_production_order	id_status	timestamp	stb_num	id_smartcard	carton_num	pallet_num
1060921	1181718524	27	240	7	2015-03-10 11:06:41.083	471900000001	NULL	I00000000011	UME0910101
1060922	1181718531	27	240	20	2015-02-27 11:07:12.230	471900000002	NULL	I00000000011	UME0910101
1060923	1181718548	25	240	19	2014-11-09 18:00:23.943	NULL	NULL	NULL	NULL
1060924	1181718555	25	240	19	2014-11-09 18:00:23.943	NULL	NULL	NULL	NULL
1060925	1181718562	25	240	19	2014-11-09 18:00:23.943	NULL	NULL	NULL	NULL
1060926	1181718579	25	240	19	2014-11-09 18:00:23.943	NULL	NULL	NULL	NULL
1060927	1181718586	25	240	19	2014-11-09 18:00:23.943	NULL	NULL	NULL	NULL
1060928	1181718593	25	240	19	2014-11-09 18:00:23.943	NULL	NULL	NULL	NULL
1060929	1181718609	25	240	19	2014-11-09 18:00:23.943	V6044870035	NULL	NULL	NULL
1060930	1181718616	25	240	19	2014-11-09 18:00:23.943	V6044870027	NULL	NULL	NULL
1060931	1184656199	25	241	24	2014-11-09 18:00:23.943	V6044913605	NULL	EW0000002...	UME0990032
1060932	1184656205	25	241	24	2014-11-09 18:00:23.943	V6044917520	NULL	EW0000001...	UME0990017
1060933	1184656212	25	241	24	2014-11-09 18:00:23.943	V6044910339	NULL	EW0000002...	UME0990030
1060934	1184656229	25	241	24	2014-11-09 18:00:23.943	V6044910223	NULL	EW0000001...	UME0990015
1060935	1184656236	25	241	24	2014-11-09 18:00:23.943	V6044910959	NULL	EW0000001...	UME0990016
1060936	1184656243	25	241	24	2014-11-09 18:00:23.943	V6044909586	NULL	EW0000001...	UME0990015
1060937	1184656250	25	241	24	2014-11-09 18:00:23.943	V6044909870	NULL	EW0000001...	UME0990015
1060938	1184656267	25	241	24	2014-11-09 18:00:23.943	V6044913796	NULL	EW0000002...	UME0990032
1060939	1184656274	25	241	24	2014-11-09 18:00:23.943	V6044910509	NULL	EW0000001...	UME0990016
1060940	1184656281	25	241	24	2014-11-09 18:00:23.943	V6044913583	NULL	EW0000002...	UME0990033
1060941	1184656298	25	241	24	2014-11-09 18:00:23.943	V6044910983	NULL	EW0000001...	UME0990016
1060942	1184656304	25	241	24	2014-11-09 18:00:23.943	V6044910100	NULL	EW0000001...	UME0990015
1060943	1184656311	25	241	24	2014-11-09 18:00:23.943	V6044912528	NULL	EW0000002...	UME0990033
1060944	1184656328	25	241	24	2014-11-09 18:00:23.943	V6044911599	NULL	EW0000002...	UME0990033
1060945	1184656335	25	241	24	2014-11-09 18:00:23.943	V6044913834	NULL	EW0000002...	UME0990032
1060946	1184656342	25	241	24	2014-11-09 18:00:23.943	V6044910908	NULL	EW0000001...	UME0990016
1060947	1184656359	25	241	24	2014-11-09 18:00:23.943	V6044926758	NULL	EW0000002...	UME0990032
1060948	1184656366	25	241	24	2014-11-09 18:00:23.943	V6044910002	NULL	EW0000001...	UME0990016
1060949	1184656373	25	241	24	2014-11-09 18:00:23.943	V6044910371	NULL	EW0000001...	UME0990016
1060950	1184656380	25	241	24	2014-11-09 18:00:23.943	V6044913788	NULL	EW0000002...	UME0990033
1060951	1184656397	25	241	24	2014-11-09 18:00:23.943	V6044913842	NULL	EW0000002...	UME0990033
1060952	1184656403	25	241	24	2014-11-09 18:00:23.943	V6044911190	NULL	EW0000002...	UME0990033
1060953	1184656410	25	241	24	2014-11-09 18:00:23.943	V6044909748	NULL	EW0000001...	UME0990015
1060954	1184656427	25	241	24	2014-11-09 18:00:23.943	V6044911777	NULL	EW0000002...	UME0990033

Figure 4. Dataset sample

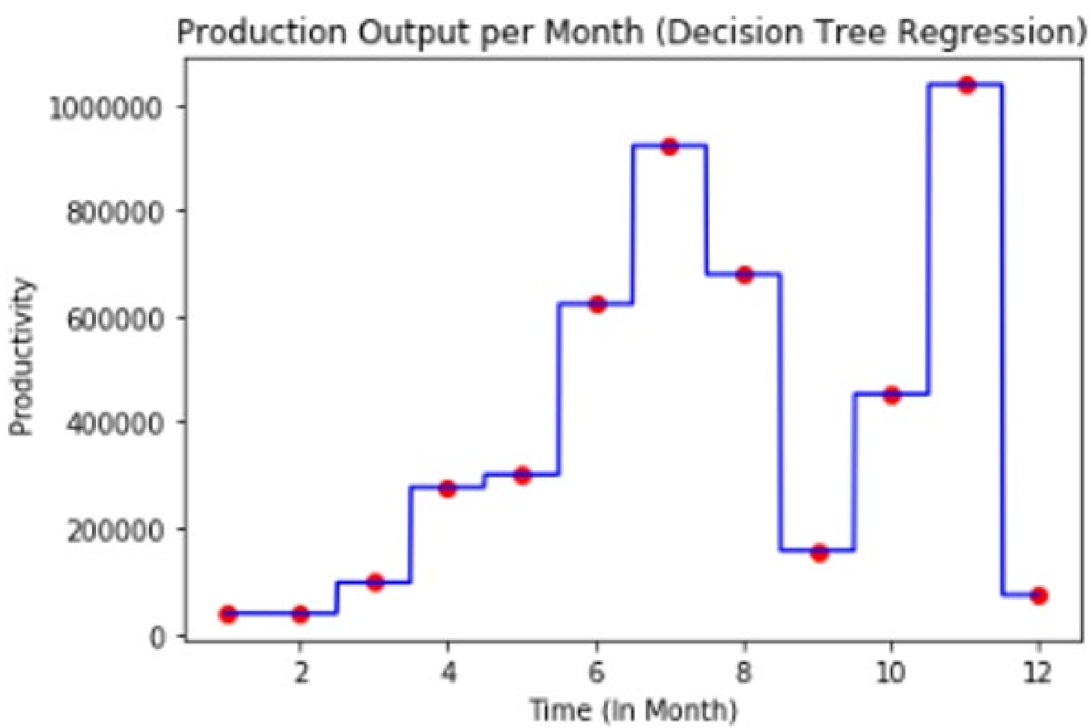

Figure 5. Regression model performance on the scatter plot

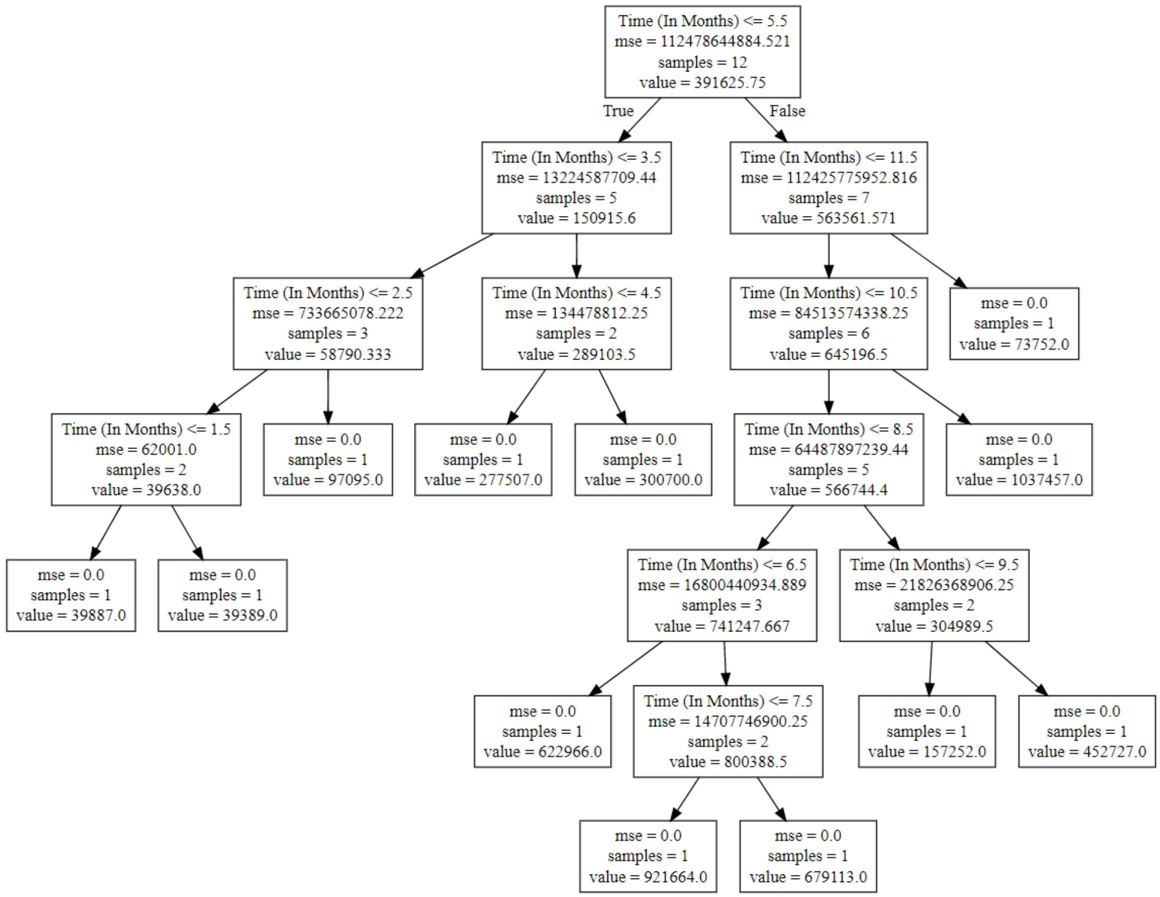

Figure 6. Decision Tree structure for production output

Table 1. Values of decision tree components

Decision Tree Component	**Value**
Mean	391625.75
Variance (mse)	112478644884.521
Standard Deviation	335378.36
Coefficient of Variation	85.64%

3.2 Reversed Engineering of a Decision Tree

The DT regression model created in the preceding section requires historical data in order to predict the future outcome. In contrast, the model discussed in this section introduces a reversed engineering of a DT regression model that allows for the specification of a predefined future outcome. The authors devised a DT regression model for calculating Overall Equipment Effectiveness (OEE). OEE encompasses components illustrated in Table 2. The model allows for the specification of a desired OEE output (e.g., 70%). When a desired OEE output is defined, the model calculates and returns the best possible ways of achieving the desired OEE output. This includes suggesting different values for each OEE component. A change in any OEE constituent's value triggers a change in values of other corresponding constituents. Figure 7 depicts five possible ways, recommended by the model, to achieve OEE of 70%, 50%, and 10%, respectively. The model reconfigures itself every time a constituent's value is changed. The philosophy of this model is intended not only for OEE, but for any business process. This would help businesses select the most efficient methods to execute business processes.

Table 2. OEE constituents

Component	Description	Unit of Measure
OEE	Availability x Performance x Quality	%
Availability	Runtime / Available Time	%
Performance	Actual / Target	%
Quality	Good Count / Total Count	%
Runtime	Available Time – Downtime	hr
Downtime	Available Time – Runtime	hr
Available Time	Available time to produce	hr
Actual	Actual units produced	ea
Target	Targeted production count	ea
Good Count	Total good units produced	ea
Bad count	Total bad units produced	ea

Targeted OEE (%)	Availability (Runtime / Available Time)				Performance (Actual / Target)			Quality (Good Count / Actual)		
	Available Time (hr)	*Downtime (hr)*	*Runtime (hr)*	**Availability (%)**	*Actual (each)*	*Target (each)*	**Performance (%)**	*Good Count (each)*	*Bad Count (each)*	**Quality (%)**
70%	8.00	0.30	7.70	96.26%	7650	9000	85.00%	6501	1148	85.00%
70%	8.00	0.38	7.62	95.26%	7740	9000	86.00%	6579	1161	85.00%
70%	8.00	0.64	7.36	92.00%	8190	9000	91.00%	6339	1351	83.50%
70%	8.00	1.44	6.56	82.00%	8640	9000	96.00%	7646	994	88.50%
70%	8.00	1.04	6.96	87.00%	8010	9000	89.00%	7249	760	90.50%
50%	12.00	2.40	9.60	80.00%	8000	10000	80.00%	6240	1760	78.00%
50%	12.00	2.64	9.36	78.00%	6500	10000	65.00%	6370	130	98.00%
50%	12.00	0.24	11.76	98.00%	5300	10000	53.00%	5096	104	97.00%
50%	12.00	5.28	6.72	56.00%	9300	10000	93.00%	8928	372	96.00%
50%	12.00	0.48	11.52	96.00%	9400	10000	94.00%	5170	4230	55.00%
10%	24.00	19.20	4.80	20.00%	18000	20000	90.00%	9900	8100	55.00%
10%	24.00	18.00	6.00	25.00%	17000	20000	85.00%	8330	8670	49.00%
10%	24.00	15.60	8.40	35.00%	6000	20000	30.00%	5940	60	99.00%
10%	24.00	8.40	15.60	65.00%	10000	20000	50.00%	3000	7000	30.00%
10%	24.00	12.00	12.00	50.00%	9200	20000	46.00%	4140	5060	45.00%

Figure 7. Results of self-configuring OEE model

4. Discussion

The literature discusses business processes, importance and outcomes of enhancing these processes, and process enhancement tools and techniques. The dataset acquired from an electronics manufacturing company is used to create a DT regression model via Python programming language. The model is trained to calculate mean, V, SD, CV, SDR, and consequently create a DT structure. The results of a tested TD model is presented in a scatter plot to reveal, based on the previous data, projected productivity for each month of year 2017.

When populated with time (in months), the model predicts low production output from January to May and increased productivity between June and August. September, October, and December are predicted to be poor performing months. In contrast, November seems to be a productive month. The underlying reasons for poor and better performances can be subject to various circumstances for various industries. Some of the potential reasons are discussed in following subsections.

The reason for poor performance at the beginning of the year (January to May) could be poor performance of the customer order book. This could alert and prompt the company to enhance performance of the sales team during this period.

Other possible reasons for low productivity could be the rising trend of machine breakdown toward the end of the year (September and October). Breakdowns are usually caused by the absence of predictive maintenance plans, amongst other things. This problem could be avoided or lessened by enhancing the company's maintenance plan and consequently implement predictive maintenance techniques, automate alerts for equipment breakdown, arrange resident technicians to respond promptly to breakdowns, and , perform route cause analysis (RCA) when attending to breakdowns.

Possible reasons for low productivity in the month of December could be due to an increased number of holidays or the majority of workers (i.e., staff and shop floor) takes annual leave from the middle of December. Therefore, the company can take measures to increase productivity from the 1st to the 11th month in order to compensate for low production output experienced during the 12th month.

The reversed engineering of a DT regression model allows businesses to specify the desired target and be presented with the most efficient ways to achieve the target. Specifically, the results of the reversed model shows that the same OEE value can be achieved by different values of availability, performance, and quality. This means that an organization might not always need to achieve 100% of availability, performance and quality. Instead, the model can recommend appropriate combinations that will yield a targeted OEE value. Because quality of semi-finished and finished products is pivotal, an organization can opt to keep the quality values constantly high but adjust the values for availability and performance accordingly.

The authors of this paper recommend for the reverse-engineered model to be applied to all types of business processes, in addition to OEE. This would allow businesses to be proactive and choose the best possible and efficient ways to execute business processes. This is different from providing historical data and ask the model to predict the future. Instead, the business specifies the future and the model provides multiple options to get into the future.

5. Conclusion

This paper emphasizes the importance of business process optimization. The right techniques and tools are pivotal to achieve optimized business processes. DT regression model has proven to be efficient in using historical and present data to predict future outcome and provide best possible ways to execute business processes. It is also possible for business to specify the desired future outcome and ask the model to identify aspects of the process to be adjusted accordingly.

Acknowledgements

The authors would like to extend a sincere gratitude to the University of Johannesburg for supporting this research.

References

Candanedo, L.M., and Feldheim, V., Accurate occupancy detection of an office room from light, temperature, humidity and CO2 measurements using statistical learning models, *Energy and Buildings,* Vol. 112, pp. 28-39, 2016.

Chang, D., Chen, B., Li, D., and Han, X., User Experience Oriented Mobile Business Process Optimization, *Studies in Informatics and Control,* Vol. 232, pp. 171-182, 2014, 2014.

Cimino, M., and Vaglini, G., An interval-valued approach to business process simulation based on genetic algorithms and the BPMN, *Information,* Vol. 5, No. 2, pp. 319-356, 2014.

Clempner, J., Verifying soundness of business processes: A decision process Petri nets approach, *Expert Systems with Applications,* Vol. 41, No. 11, pp. 5030-5040, 2014.

De Oña, R., Eboli, L., and Mazzulla, G., Key factors affecting rail service quality in the Northern Italy: a decision tree approach, *Transport,* Vol. 29, No. 1, pp. 75-83, 2014.

Fang, X., Cao, R., Liu, X., and Wang, L., A Method of Mining Hidden Transition of Business Process Based on Region, *IEEE Access,* Vol. 6, pp. 25543-25550, 2018.

Gokgoz, E., and Subasi, A., Comparison of decision tree algorithms for EMG signal classification using DWT, *Biomedical Signal Processing and Control,* Vol. 18, pp. 138-144, 2015.

Hadded, L., Charrada, F.B., and Tata, S., Optimizing Autonomic Resources for the Management of Large Service-Based Business Processes, *IEEE Transactions on Services Computing,* 2018.

Jiménez-Ramírez, A., Weber, B., Barba, I., and Del Valle, C., Generating optimized configurable business process models in scenarios subject to uncertainty, *Information and Software Technology,* Vol. 57, pp. 571-594, 2015.

Lohrmann, M., and Reichert, M., Effective application of process improvement patterns to business processes, *Software & Systems Modeling,* Vol. 15, No. 2, pp. 353-375, 2016.

Montarnal, A., Mu, W., Benaben, F., Lamothe, J., Lauras, M., and Salatge, N., Automated deduction of cross-organizational collaborative business processes, *Information Sciences,* Vol. 453, pp. 30-49, 2018a.

Montarnal, A., Mu, W., Benaben, F., Lamothe, J., Lauras, M., and Salatge, N., Automated deduction of cross-organizational collaborative business processes, *Information Sciences,* Vol. 453, pp. 30-49, 2018b.

Parody, L., Gómez-López, M.T., and Gasca, R.M., Hybrid business process modeling for the optimization of outcome data, *Information and Software Technology,* Vol. 70, pp. 140-154, 2016.

Seera, M., and Lim, C.P., A hybrid intelligent system for medical data classification, *Expert Systems with Applications,* Vol. 41, No. 5, pp. 2239-2249, 2014.

Weidlich, M., Ziekow, H., Gal, A., Mendling, J., and Weske, M., Optimizing event pattern matching using business process models, *IEEE Transactions on Knowledge and Data Engineering,* Vol. 26, No. 11, pp. 2759-2773, 2014.

Yan, R., Ma, Z., Zhao, Y., and Kokogiannakis, G., A decision tree based data-driven diagnostic strategy for air handling units, *Energy and Buildings,* Vol. 133, pp. 37-45, 2016.

Zhao, J., Lasternas, B., Lam, K.P., Yun, R., and Loftness, V., Occupant behavior and schedule modeling for building energy simulation through office appliance power consumption data mining, *Energy and Buildings,* Vol. 82, pp. 341-355, 2014.

Biographies

Mike Sishi is a PhD student and co-supervisor in the Faculty of Engineering and Built Environment (FEBE) at the University of Johannesburg, South Africa. He holds a Master's degree in Engineering Management completed with distinction. His research interests focus on the digitization of the industrial sector, with specific emphasis on the utilization of Industry 4.0 technologies such as Internet of Things (IoT), Industrial Internet of Things (IIoT), Big Data Analytics (BDA), Cyber Physical Systems (CPS), Artificial Intelligence (AI), etcetera. Moreover, he serves as an ICT Professional for an industrial company in South Africa, where he implements Enterprise 4.0 technologies.

Arnesh Telukdarie is an Associate Professor in the Faculty of Engineering and Built Environment (FEBE) at the University of Johannesburg, South Africa. He holds a Doctorate in Chemical Engineering and has published widely in Industry 4.0. He is also a consulting engineer in Industry 4.0, digital business. His academic work focuses on mentoring engineering Master's and PhD students and writing conference and journal papers. He has published in over 30 international journals and has over 100 conference proceeding. His research projects strive to unearth technologies and implementation methodologies that enable the digitizing of the technologically challenged industrial and mining sectors.