A Generic Model for End State Prediction of Business Processes Towards Target Compliance

Naveed Khan[1](✉), Zulfiqar Ali[1], Aftab Ali[1], Sally McClean[1], Darryl Charles[1],
Paul Taylor[2], and Detlef Nauck[2]

[1] School of Computing, Ulster University, Newtownabbey, UK
{n.khan,z.ali,a.ali,si.mcclean,dk.charles}@ulster.ac.uk
[2] Applied Research, BT, Ipswich, UK
{paul.n.taylor,detlef.nauck}@bt.com

Abstract. The prime concern for a business organization is to supply quality services to the customers without any delay or interruption so to establish a good reputation among the customer's and competitors. On-time delivery of a customers order not only builds trust in the business organization but is also cost effective. Therefore, there is a need is to monitor complex business processes though automated systems which should be capable during execution to predict delay in processes so as to provide a better customer experience. This online problem has led us to develop an automated solution using machine learning algorithms so as to predict possible delay in business processes. The core characteristic of the proposed system is the extraction of generic process event log, graphical and sequence features, using the log generated by the process as it executes up to a given point in time where a prediction need to be made (referred to here as cut-off time); in an executing process this would generally be current time. These generic features are then used with Support Vector Machines, Logistic Regression, Naive Bayes and Decision trees to predict the data into on-time or delayed processes. The experimental results are presented based on real business processes evaluated using various metric performance measures such as accuracy, precision, sensitivity, specificity, F-measure and AUC for prediction as to whether the order will complete on-time when it has already been executing for a given period.

Keywords: Business processes · Automated system · Process prediction · End state prediction

1 Introduction

Over the last few decades, there has been a significant ongoing interest in research into Business Process Management (BPM) with the aim of predicting future process states [2]. Here a process is a series of tasks or steps, terminated by

M. Bramer and M. Petridis (Eds.): SGAI-AI 2019, LNAI 11927, pp. 325–335, 2019.
https://doi.org/10.1007/978-3-030-34885-4_25

an event and taken in order to achieve a particular end. Such prediction can help to gain operational excellence, and boost productivity, customer satisfaction and cost effectiveness [1]. The monitoring of a complex and dynamic business process is essential for analysis and identification when process instances do not perform as required. The timely prediction of such behaviours from online data can facilitate intervention and avert an undesired state of a process from occurring [1].

Moreover, the existence of such inefficiencies in business processes can greatly affect performance, ultimately increasing cost and having a negative impact on customer satisfaction [7]. Therefore, predictive process monitoring can utilize data generated during process execution so as to continuously monitor process performance [8]. Continuous monitoring of a business process can facilitate pre-emptive actions to attain the desired process outcome.

In this paper, machine learning techniques have been investigated for online process analysis to extract useful and discriminating information from raw data. Such information can be used to discover patterns that characterize an outcome as very likely and subsequently perform online prediction on incomplete process instances when this pattern has been observed. Therefore, the focus is to develop a system that will extract generic features from new processes for early prediction of a timely outcome, i.e. situations where the order can be delivered on-time, as opposed to delayed or cancelled. These techniques will help to uncover deeper insight into patterns which are difficult to execute manually or through visualization. Such analysis will enable domain experts to address these inefficiencies and help to streamline the process. The novelty here resides in the development of online strategies for predicting outcomes based on heterogeneous process data where we train and test by mapping the (online) processes onto the percentage of time until target has been reached. The generic feature selection approach, which is based on a portfolio of process, event log, graphical and sequence features, is also novel.

The remainder of this paper is organised as follows: in Sect. 2 the end state prediction framework is discussed. In Sect. 3, Feature extraction and dataset description is discussed with results and discussions presented in Sect. 4. Finally, Conclusions and Future Work are presented in Sect. 5.

2 A Framework of End State Prediction of Business Process Data

In many enterprises, early predictions of business processes are very helpful and can make the business more cost effective. Although analysis of such processes is quite complex and challenging, the capability of perceiving the likely conclusion of an ongoing process in advance would help business managers to react in time and help to avoid any delays or undesirable situations. In this paper, we consider an example of timely and early prediction of BT consumer processes, where the data contains the information for landline telephone and/or broadband orders. The aim is to develop a system that will extract generic features from processes for early online prediction that either the order will be delayed/cancelled or delivered on-time. Moreover, it is very difficult for the human mind to classify a multidimensional feature vector [6] and hence automated pattern recognition becomes essential and provides help in analyzing and understanding of complex data. Also, if the extracted features are generic in nature, then the system developed using such features can also be used for a new problem without making significant changes in our approach.

3 Dataset Description

In our experiments, a BT consumer dataset is used for analysis and evaluation. The complete dataset consists of 505,632 instances; however, in current experiments, we have only extracted instances of consumers, who have ordered landline telephone and/or broadband. These instances are used with the labels Y and N, where Y represents on-time delivered orders and N represents delayed/cancelled orders. Initially, a total of 15,523 on-time delivered processes and 1,585 processes delayed processes were extracted. Pre-processing and feature extraction are the most crucial steps in prediction [3] and are used to extract useful information for prediction into timely or delayed process instances.

3.1 Pre-processing

In the pre-processing step, we have processed the raw data of consumers, which are new orders for land line telephone and/or broadband. From the total of 15,523 on-time delivered processes, only 725 processes were used in our experiments to extract useful information since for the remainder of the processes either the tasks are recorded as zero duration or Target date and time (T_{dt}) were missing. Here Target date represents a target by which the process should, be completed. Similarity, for delayed processes, a total of 5830 processes were extracted, out of which 1,585 processes were used in our experiments because for the remaining processes either the task durations are 0 or Target date and time (T_{dt}) were missing, as before. The extracted features from successful on-time processes and unsuccessful delayed processes are then used to predict on-time and delayed/cancelled orders.

In our framework as shown in Fig. 1, initially the features are extracted from a process by taking cut-off time 25%, 50%, 75%, 85%, 95% and 100% from the starting date and time (S_{dt}) of the process until Target date and time (T_{dt}). Here (T_{dt}) is an initial target for the process completion date and time. The cut-off time means the percentage of time that point in target time which we regard as current so that we can make predictions using data only relating to the history up to that point in time and then compare them with the actual compliance, or otherwise. Cut-off time here is calculated as a percentage of the target.

Hence, if a process completes according to the estimated Commitment date and time it will be considered to be an on-time delivery, otherwise it will be considered to be a delayed process. For instance, consider, an example process initiated at a particular instance of date and time (e.g. Start date and time - 11/05/2018 10:35:00) with an initial estimate to complete this process by Commitment date and time (30/07/2018 23:59:59). In order to predict the process at different cut-off time ratios, we calculated the cut-off time using Eq. 1.

$$Cut - off\ \ Time = (T_{dt} - S_{dt}) * Th\% \tag{1}$$

The T_{dt} is the Target date and time, S_{dt} is the process starting date and time and $Th\%$ is the 25%, 50%, 75%, 85%, 95% and 100% is the process cut-off time. As shown in Fig. 1, the cut-off time was explored for different ratios of 25% (dotted lines), 50% (dashed lines) and 75% (dashed line) and 100% (dotted line) of the time difference between start date and time and the corresponding target date and time.

3.2 Feature Extraction

The feature extraction step is first used offline to compute the likely most discriminating features from the data since success of prediction is strongly dependent on the extraction of highly relevant features form the raw data [4]. Moreover, if the extracted features are generic in nature, then the system developed using such features can also be used in a new domain without applying significant changes. Here, the process of feature extraction is elaborated by considering the randomly selected order as an example process from the raw data for ease of understanding. For such processes a number of generic features are typically available, such as process, event log, graphical and sequence features. Thus, for example, process features are known at the start of the process and do not change with time while event log features are revealed as the process executes the different tasks which it comprises. Graphical and sequence features, on the other hand, relate to the currently available event tree, which will be incomplete if the process is still ongoing. Here graphical features are a measure of how complex the event tree is e.g. how many nodes are repeated or how wide the tree is, while sequence features relate to the order of events within the process tree and whether there are common patterns or motifs which might indicate success or failure, in this case a timely outcome, or otherwise.

Moreover, the node features, on the other hand are all types of graphical feature which quantify the complexity of the log graph traversed to date. For example, the repetition of tasks can be a crucial indicator for the prediction of orders, as the repetition of some tasks may cause a delay in process completion.

A complete process with all associated tasks is shown in Fig. 1. Different measures are computed from the raw data to form the feature vector. The general form of seven-dimensional feature vector F used for prediction is given in Eq. 2.

$$F = [aT \quad uT \quad rT \quad nRT \quad perT \quad t_{Time} \quad r_{Time}] \tag{2}$$

where

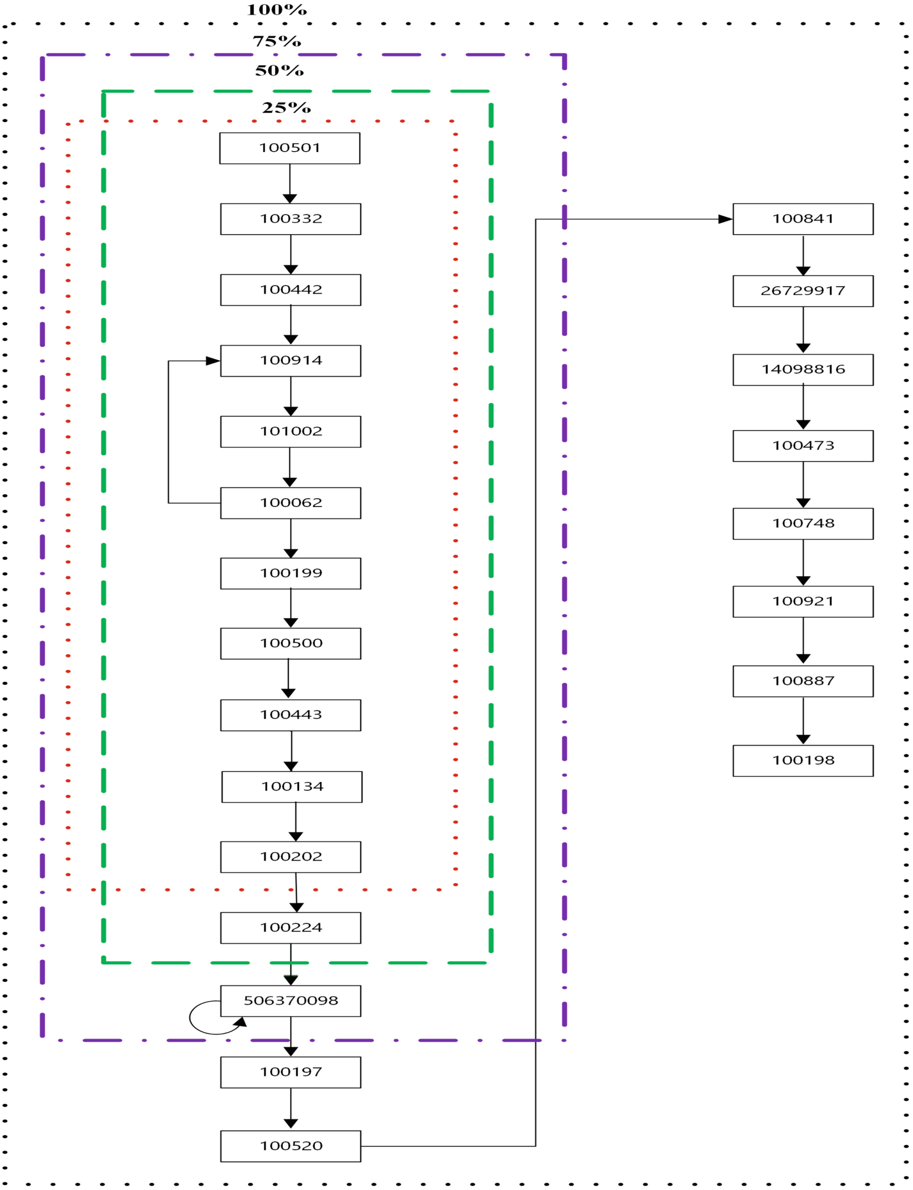

Fig. 1. Task execution of a process for different cut-off percentages.

1. The number of all traversed nodes including repetition are Appeared Tasks. This feature is denoted by aT.
2. All nodes having a degree exactly equal to 1, in fact, these are Unique Tasks which appear only once and are denoted by uT.

3. The number of nodes having a degree greater than 1. In other words, the number of all Repeated Tasks. This feature is represented by rT.
4. The sum of degrees of all nodes which are traversed more than once. This feature represents the number of occurrences of all repeated tasks and are denoted by nRT.
5. The actual percentage of cut-off time till the Target date and time (T_{dt}) for each process and represented by $perT$.
6. The total time taken by a process to execute tasks from start to end of a process. This feature is denoted by (t_{Time}).
7. The total time taken by repeated tasks in a process. This feature is denoted by (r_{Time}).

It is critical to know the capability of each attribute in the feature vector F to discriminate between two types of orders and understand how it can help in accurate early prediction of the process completion.

Hence, the Fisher discriminant ratio (FDR) is used in this study to calculate the discriminating power of the features. FDR has been previously used successfully for this task e.g. [2]. Fisher discriminant ratio (FDR) is calculated using Eq. 3 and implemented for each attribute of the feature vector F.

$$FDR_f = \frac{(\mu_Y - \mu_N)^2}{\sigma_Y^2 + \sigma_N^2} \tag{3}$$

where μ and σ represents mean and variance of data, respectively, Y and N stands for the on-time and delayed orders, respectively. The variable f signifies the number of attributes and $f = 1, 2, 3, ..., n$. The higher the ratio, the more successful the discriminating feature is in terms of predicting whether the target is successfully achieved or not. The last two features (t_{Time}) and (r_{Time}) were less discriminant based on FDR analysis and have been eliminated. The discriminant features used here along with their corresponding Fisher ratio are represented in Fig. 2.

A generic form of a 5-dimensional feature vector F used for prediction is given as:

$$Feature\,Vector = [aT \quad uT \quad rT \quad nRT \quad perT]$$

where the feature vector for an example process is given by:

$$F = [13 \quad 9 \quad 2 \quad 4 \quad 45.35\%]$$

These features were used to predict the process compliance with the target at different cut-off time ratios to predict whether the process will complete on-time or will be delayed. For simplicity, we diagrammatically represented an example of 50% cut-off time for a process as shown in Fig. 3, where 15 tasks appeared before the cut-off time, of which 9 were unique tasks and 3 tasks were repeated 6 times and 50% is the actual percentage of cut-off time until the Target date and time (T_{dt}).

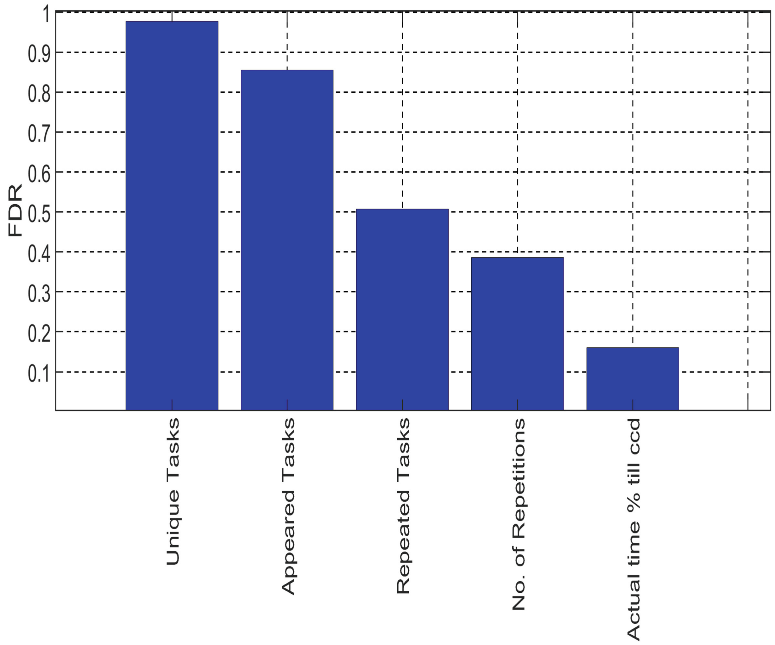

Fig. 2. Discriminant features using fisher ratio

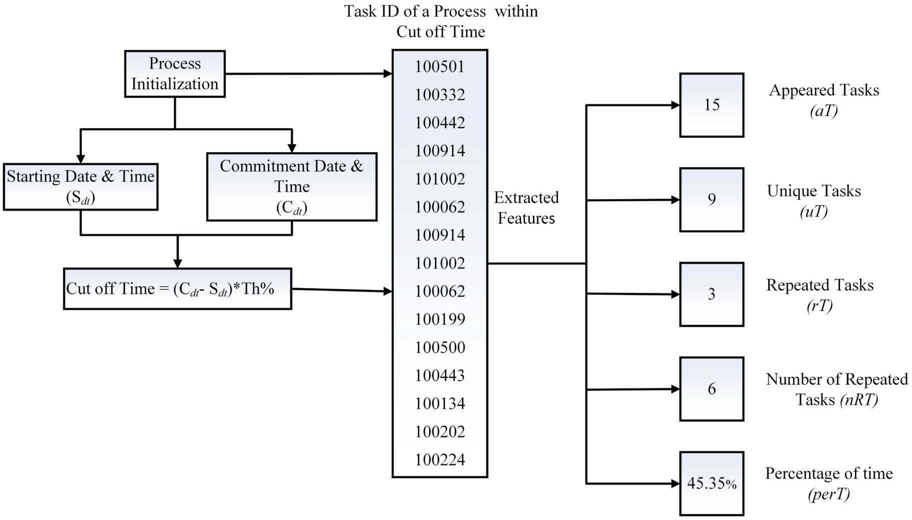
Task ID of a Process within Cut off Time
Process Initialization
Starting Date & Time (S_{dt})
Commitment Date & Time (C_{dt})
Cut off Time = (C_{dt}- S_{dt})*Th%
100501
100332
100442
100914
101002
100062
100914
101002
100062
100199
100500
100443
100134
100202
100224
Extracted Features
15
Appeared Tasks (aT)
9
Unique Tasks (uT)
3
Repeated Tasks (rT)
6
Number of Repeated Tasks (nRT)
45.35%
Percentage of time (perT)

Fig. 3. Feature Extraction Framework of an Incomplete Process at 50% cut-off time with $Th = 0.5$

4 Results and Discussion

We have implemented various machine learning algorithm using the computed multi-dimensional feature vector for prediction of process compliance with target. Our experiments have been performed using 10-fold cross validation to evaluate the predictive model and partition the original sample into a training set and a testing set to evaluate the model.

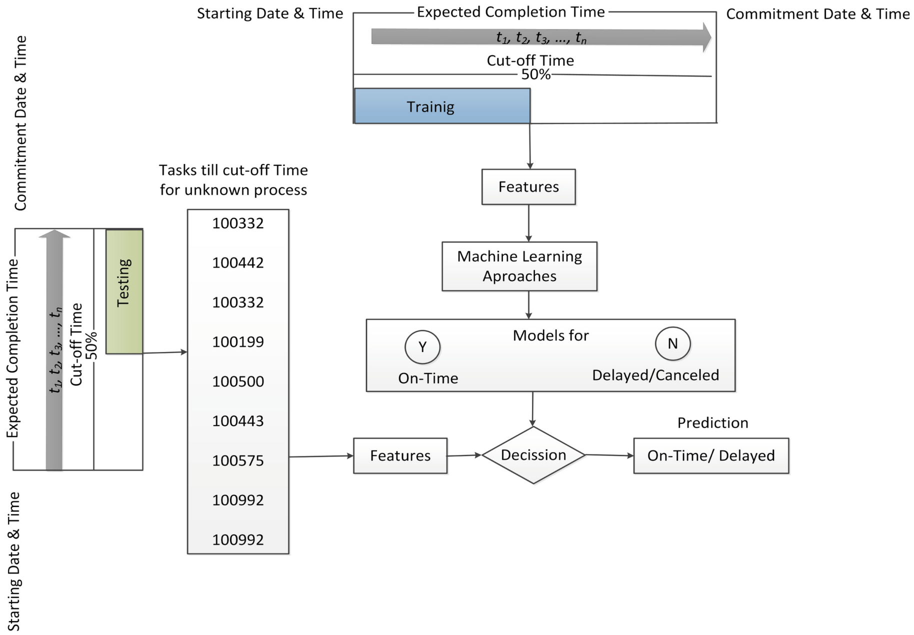

Fig. 4. Block diagram for process prediction

The advantage of this approach is that process values are used for both training and testing, and each observation is used for testing exactly once. During the testing phase, an unknown process is fed to the system, the relevant features are extracted, and the system records if the process finishes before the cut-off time ratio (i.e. 25%, 50%, 75%, 85%, 95%, and 100%) it will be considered as an on-time delivery otherwise we predict it as a delayed process. In our approach, we consider the processes start time (time since initialization of that process) with different cut-off time ratio (i.e. 25%, 50%, 75%, 85%, 95% and 100%) until the initial estimation of process completion (Commitment date and time) as shown in Fig. 4. The features are extracted from the raw data and these features were used to train the model. Different machine learning algorithms such as Support Vector Machines (SVM) [9], Logistic Regression [5], Naive Bayes [11] and J48 [10], are used to predict the outcome of the process.

Table 1. Prediction results of different machine learning algorithms for different metric measures

Algorithm	25% cut-off time					
	Accuracy %	Precision %	Sensitivity %	Specificity %	F-measure %	AUC
Logistic Regression	73.66	90.73	76.07	61.85	82.76	0.73
J48	74.61	88.86	77.92	61.93	82.94	0.70
SVM	74.73	89.19	77.78	62.61	83.10	0.73
Naive Bayes	62.75	60.39	81.33	42.85	69.31	0.71
Algorithm	50% cut-off time					
	Accuracy %	Precision %	Sensitivity %	Specificity %	F-measure %	AUC
Logistic Regression	75.10	92.22	77.03	65.33	83.85	0.74
J48	75.07	84.95	80.59	59.36	82.71	0.68
SVM	75.07	90.45	77.70	63.32	83.59	0.74
Naive Bayes	55.72	44.14	85.95	38.67	58.33	0.72
Algorithm	75% cut-off time					
	Accuracy %	Precision %	Sensitivity %	Specificity %	F-measure %	AUC
Logistic Regression	74.60	91.30	76.07	61.85	82.75	0.77
J48	73.32	87.39	77.60	56.79	82.20	0.69
SVM	75.14	90.45	77.82	62.67	83.66	0.76
Naive Bayes	64.16	57.02	87.91	44.14	69.18	0.75
Algorithm	85% cut-off time					
	Accuracy %	Precision %	Sensitivity %	Specificity %	F-measure %	AUC
Logistic Regression	75.41	90	78.35	62.75	83.77	0.78
J48	74.38	87.83	78.43	59.09	82.87	0.72
SVM	75.71	89.45	78.93	62.85	83.86	0.78
Naive Bayes	65.54	58.46	88.90	45.31	70.54	0.76
Algorithm	95% cut-off time					
	Accuracy %	Precision %	Sensitivity %	Specificity %	F-measure %	AUC
Logistic Regression	77	90.18	79.95	65.39	84.76	0.78
J48	75.74	87.74	79.96	61.03	83.67	0.76
SVM	77.21	89.45	80.53	64.97	84.76	0.78
Naive Bayes	66.81	63.96	85.54	45.72	73.19	0.75
Algorithm	100% cut-off time					
	Accuracy %	Precision %	Sensitivity %	Specificity %	F-measure %	AUC
Logistic Regression	77.20	90.45	80	65.80	84.90	0.78
J48	79.16	86.39	84.56	64.96	85.47	0.82
SVM	77.12	89.36	80.51	64.56	84.71	0.77
Naive Bayes	67.53	67.29	83.74	46.06	74.62	0.73

During the testing phase, an unknown process is fed to the system, the relevant features are extracted, if the processes finishes before the cut-off time ratio (i.e. 25%, 50%, 75%, 85%, 95%, and 100%) it will be considered as an on-time delivery, and if not then the decision is made by the predictive model that the process is delayed as shown in Fig. 4.

Different performance measures such as accuracy, precision, sensitivity, specificity, F-measure and AUC are used to evaluate the performance of the chosen machine learning algorithms. In business processes, accuracy is not considered to be the best measure for performance evaluation of prediction algorithms due to the fact that accuracy is sensitive towards class imbalance [7], which is a characteristic of our current problem. Therefore, true positive (the proportion of correctly classified instances) and false positive (proportion of incorrectly classified instances) rates are more important from a cost-benefit perspective and due to their being agnostic towards data skewness. The results of our experiments are evaluated using various performance measures such as accuracy precision, sensitivity, specificity, F-measure and AUC as presented in Table 1. The F- measure is regarded as the best performance metric for imbalanced data and AUC is used to determine which of the used models have predicted the classes best.

As presented in Table 1, in the case of a 25% cut-off time, it is observed that the support vector machine (SVM) and logistic regression perform slightly better than the other algorithms in terms of accuracy (74.73% and 73.66%), precision (89.19% and 90.73), sensitivity (77.78%, 76.07%), F-measure (83.10% and 82.76%) and the AUC is the same (0.73%).

Similarly, in the case of a 50% cut-off time, the logistic regression algorithm achieved 75.10% accuracy, 92.22% precision, 77.03% sensitivity, 83.85% F-measure and 0.74% AUC. On the other hand, SVM and J48 are close competitors of logistic regression with the same accuracy 75.07%, precision (90.45% and 84.95%), sensitivity (77.70 and 77.03), F-measure (83.59% and 82.71%) and AUC (0.74% and 0.68%).

As the cut-off time increases (i.e. to 75%, 85% and 95%), SVM and logistic regression perform better than the other algorithms as presented in Table 1. However, in the case of 100% cut-off time, the J48 algorithm performs better than other algorithms and achieved 79.16% accuracy, 86.39% precision, 84.56% sensitivity, 85.47% F-measure and 0.82% AUC. If we rank the algorithms based on the average of all performance parameters, then logistic regression and SVM have a very close contest as is evident from Table 1. AUC is the most important evaluation metric used for checking prediction model performance where a high AUC value represents more accurate prediction made by the model as presented in Table 1.

In case of 25%, 50%, 75%, 85% and 95% cut-off time logistic regression compared to SVM either achieved slightly high or same AUC value of 0.73%, 0.74%, 0.77%, 0.78% and 0.78%. However, in case of 100% cut-off time, J48 achieved AUC value of 0.82%. Also, the result shows that as we increase the cut-off time, prediction is improved, as expected, particularly towards the end of the process.

5 Conclusion

The monitoring of a business processes to avoid delay in the delivery of a customers order is a crucial element for better customer experience and to avoid financial loss for the company. The proposed system tries to provide a solution to this problem, where the real issue is to provide a means of monitoring the progress of an order in real time and assess whether we can infer a likely breach of target compliance and predict delay at an early stage. Finding the optimal time for intervention and mediating against the evolving situation is a trade-off between having confidence in our prediction and intervening in a timely manner. Our current predictions achieve reasonable success where performance is seen to increase as the process evolves towards the target, and more process data is exposed. By using generic features for prediction, we hope in further work to explore different feature selection and classification options as well as diverse problem domains.

Acknowledgement. This research is supported by the BTIIC (BT Ireland Innovation Centre) project, funded by BT and Invest Northern Ireland.

References

1. Aalst, W.: Business process management: a comprehensive survey. ISRN Softw. Eng. **2013**, 1–37 (2013)
2. Al-Nasheri, A., et al.: An investigation of multidimensional voice program parameters in three different databases for voice pathology detection and classification. J. Voice **31**(1), 113–e9 (2017)
3. Alasadi, S.A., Bhaya, W.S.: Review of data preprocessing techniques in data mining. J. Eng. Appl. Sci. **12**(16), 4102–4107 (2017)
4. Aparna, U., Paul, S.: Feature selection and extraction in data mining. In: 2016 Online International Conference on Green Engineering and Technologies (IC-GET), pp. 1–3. IEEE (2016)
5. Menard, S.: Applied Logistic Regression Analysis, vol. 106. Sage, Thousand Oaks (2002)
6. Mesallam, T.A., et al.: Development of the Arabic voice pathology database and its evaluation by using speech features and machine learning algorithms. J. Healthc. Eng. **2017**, 1–13 (2017)
7. Taylor, P.N., Kiss, S.: Rule-mining and clustering in business process analysis. In: Bramer, M., Petridis, M. (eds.) SGAI 2018. LNCS (LNAI), vol. 11311, pp. 237–249. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-04191-5_22
8. Von Rosing, M., Von Scheel, H., Scheer, A.W.: The Complete Business Process Handbook: Body of Knowledge from Process Modeling to BPM, vol. 1. Morgan Kaufmann, Boston (2014)
9. Wang, L.: Support Vector Machines: Theory and Applications, vol. 177. Springer, Heidelberg (2005). https://doi.org/10.1007/b95439
10. Witten, I.H., Frank, E., Hall, M.A., Pal, C.J.: Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann, San Francisco (2016)
11. Zhang, H.: The optimality of naive bayes. Am. Assoc. Artif. Intell. **1**(2), 3 (2004)