Prediction of Process Failure Approach Using Process Mining

Diahame Mamadou(✉) and Camara Mamadou Samba

Ecole Supérieur Polytechnique, Université Cheikh Anta Diop de Dakar (UCAD), Dakar, Senegal
{diahame.mamadou,mamadou.camara}@esp.sn

Abstract. Events log are a collection of events that concern a business process. In them, we may find cases where its output is different from what expected. These differences are considered as failure and many publications usually propose prediction model to improve the business model. But existing approach of prediction rarely take into account the loops. The aim of this work is to propose a prediction of business process failure while considering loops as failure. So, in order to introduce the loop, we need first to determine how to implement the loop in existent event log. We propose some machine learning model in order to do the prediction. And then, compare the prediction model in order to get the best one. The prediction model is made by using the event log's dataset of a loan application performed in a financial institution.

Keywords: Process failure · Process mining · Event log · Event loop · Machine learning

1 Introduction

Nowadays with the boom of data, we easily find massive data that can be used. The generation of those data are mostly related to event. These events are collected and organized into event log with which we can determine the process model of a system [1]. A process instance may contain some error that can be consider as failure. Sometimes, during a given case or instance, we can find excessive repetitions (>3) of an or multiple event(s). this repetition can be seen as failure because it can affect the performance of the business process.

Differentiating the terms fault, error and failure is significant in the context of business process failure prediction. In present literature, they have a well-defined semantic by [2] as follow:

- A fault is the adjudged or hypothesized cause of an error.
- An error is the deviation of the system from its desired state.
- Finally, a failure occurs when the system is not able to deliver its output as it is supposed to, leading to an undesirable outcome.

A. D. Mambo et al. (Eds.): InterSol 2022, LNICST 449, pp. 82–95, 2022.
https://doi.org/10.1007/978-3-031-23116-2_6

The aim of this work is to propose a prediction of business process failure while considering loops as failure. So, in order to introduce the loop, we need first to determine how to implement the loop in existent event log. We propose some machine learning model in order to do the prediction. The prediction model is made by using the event log of a loan application performed in a financial institution.

The remainder of this paper is organized as follows: Sect. 2: we present the related work. Section 3: the description of the loan application data. Section 4: describe the preprocessing step. Section 5: define the metric that are going to be used for the comparison. Section 6: describe the learning model technique that are used. Section 7: Comparison of the result.

2 Related Work

2.1 Process Failure

There are publications that address the process failure in different angle. For [3], they propose the PreMiSE (PREdicting failures in Multi-tIer distributed SystEms) which according to them is a novel approach that can accurately predict failures and precisely locate the responsible faults in multi-tier distributed systems. In order to identify the failure, they use a Key Performance Indicators (KPI) for instance CPU utilization for each CPU processor in the system. For [4], the goal of their paper is to examine the exploitation of events in order to find errors and predict potential failures during (distributed) process execution. They use artificial neural network for the prediction model. In [5], they use the local outlier factor (LOF) algorithm which is an unsupervised fault detection algorithm with rule-based monitoring approaches. They used the LOF with the KNNI prediction algorithm for the prediction of abnormal termination of a real-time business process. For [6], they propose a novel method for predicting the next process event and also a novel application for deep learning methods. This application consists of using the Long Short-Term Memory (LSTM) which is a Recurrent Neural Network (RNN) architecture.

2.2 Loop

There are documents that talk about the loop. For [7, 8], a loop causes a task to be executed multiple times during a given case. In [8] and [9], they define two types of loops:

- Basic loops: which according to [10] can be compared to WHILE loops.
- Arbitrary loops: which are like the GOTO statements according to [11].

According to [7, 8], in order to defined a representation of loops inside the event log, one can add the element task_instance. We used a similar representation of the loops where we named it “Iteration”. In [10], they try to address the loop failure by proposing a methodology based on the CRISP-DM (Cross Industry Standard Process for Data Mining). Their methodology consists of a number of Step based on the 6 phases of the CRISP-DM process [12] which are: Business understanding, Data understanding, Data preparation, Modeling, Evaluation and Deployment.

2.3 Critics

For [3–6], they use some predictive model for different objective and did not treat the loop issue. Their propositions show different ways to identify the failure of a business process. In [10] even if they made a proposition of methodology, there wasn't an implementation of it. For the [7–9], they made a good explanation for the loop but since then there are not much publication that take it into account.

3 Data Description

The dataset used is from the BPIC 2017 [13]. This dataset is from a financial institution. The event log provided contains all loan applications filed in 2016 and their subsequent handling up to February 2nd 2017. The characteristic of the dataset is as follow:

- Number of variables: 19. Among them, we have 11 categorical variables, 6 numerical and 2 Booleans
- Number of observations (events): 1,202,267: These observations represent the event of the application process
- Number of instances (cases): 31,509 cases it is represented by the variable 'case:concept:name'

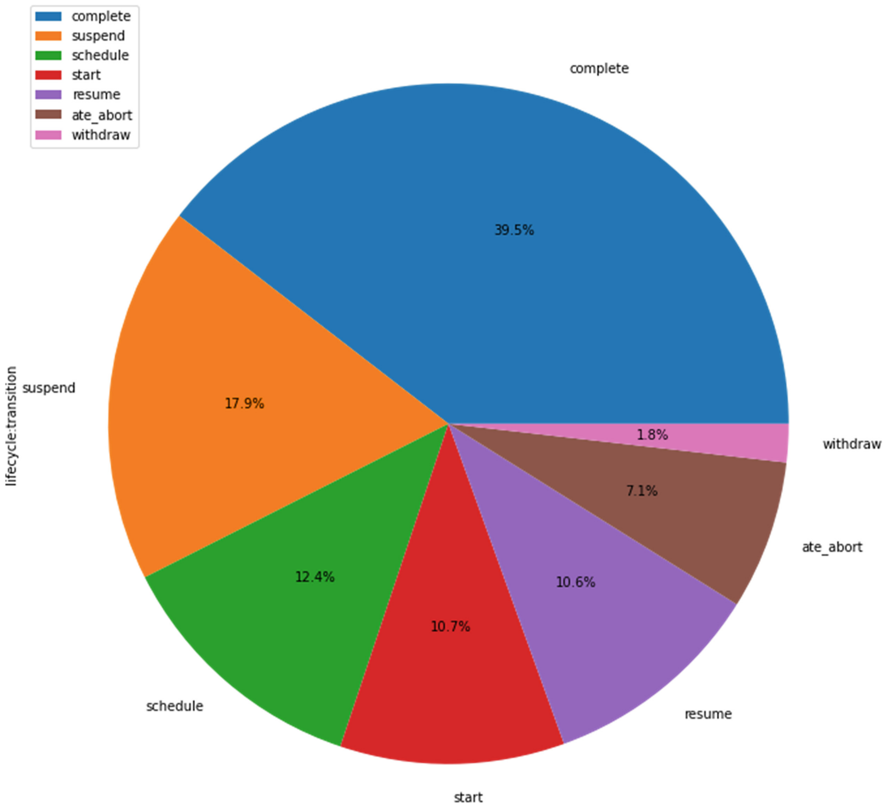

Fig. 1. Distribution of the lifecycle:transition value

The variable that contains the event is named “concept:name” which can take 26 different values. These values will be referred in the document as event’s types. We can see a disparity between the number of instances and the number of observations. The variable that concerns the failure of a business process is the lifecycle:transition. In fact, it can take a total of 7 value which are represented in the Fig. 1.

The Fig. 1 represent the status of the execution of an event. In fact, based on the figure, we can see that, after the execution of an event during a trace, the status of it can take different values like: ‘complete’, ‘resume’, ‘suspend’, ‘ate_abort’. Among these values, the value that can represent a business process failure is “ate_abort”.

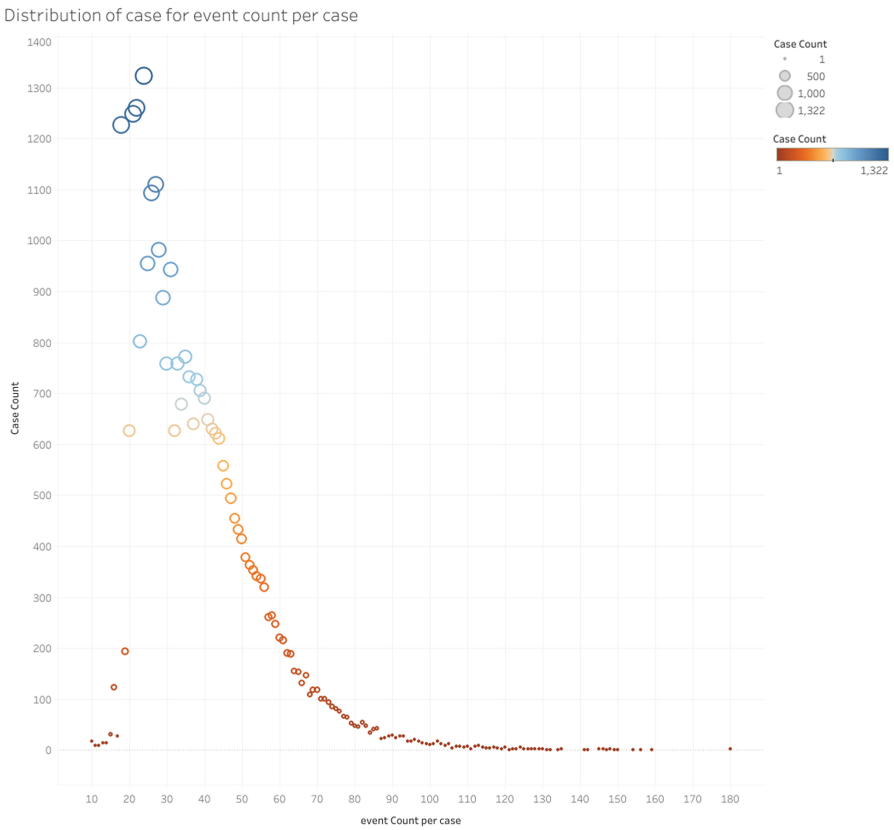

Fig. 2. Distribution of cases for the event count per case

In the Fig. 2 we have the distribution of the case that have the same number of events. The figure permits us to see the relation between the number of events for a case and how much cases have the same number. The Abscissa represent the number of events per cases. We can see that the smallest number of events for a case is 10 and the maximum number of events is 180. The ordinate represents the count of the cases that have the same number of events. We can say that the count is between 1 and 1322. The most common number of events is around 24 events which has a total of 1322 related cases.

We can also see a trend as the number of events increase, there is less likely cases that have the same number of events.

Therefore, considering the loop as failure (>3) can have a valuable effect in the business process. In order to show it in a more detail way, we have the Fig. 3 which show the ratio of repetition of a case. The ratio is obtained by doing the relation between the number of event's types in a case and the number of events that are in that same case.

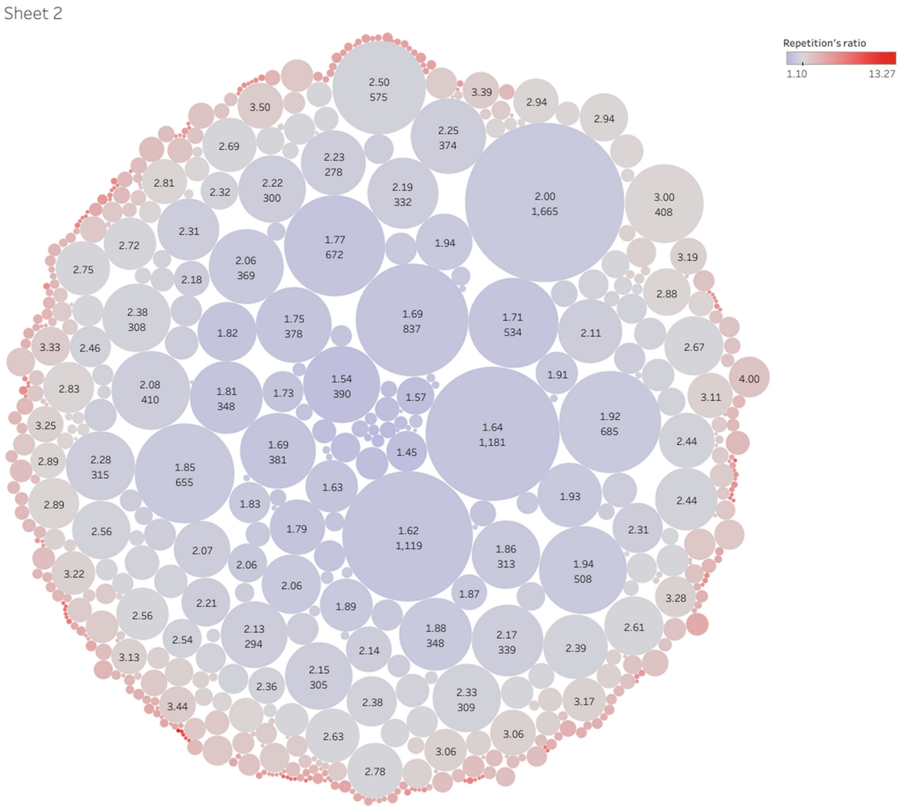

Fig. 3. Repetition's ratio of the cases

We can say based on the figure that the common ratio of the case is 2 that means that for a case, the number of the events is two times superior to the number of event's types. Therefore, for that ratio, there are 1665 cases that have the same ratio. The Fig. 3 shows that in the dataset that we use, there are some case where predicting loops can have a great impact in the quality of the business process model.

There are some documents that made an advance analysis of the dataset [14–16].

4 Preprocessing

Based on the analysis of the data, we find that there no variable that materialize the loop inside an instance. In order to materialize the loop inside the dataset, we added a new variable named "iteration" which count each event during the execution of a process. We also decide to delete the variable that has high number of unique values. Those variables are:

- EventID
- time: timestamp

Then, we have to also find the variable that have high correlation with our prediction variables ("lifecycle:transition", "iteration"). The first one represents the final state of an event. The ladder counts each event during the execution of a process. In order to do so, we first need to determine the correlation between the variable. That correlation is show in the Fig. 4 through a heatmap chart.

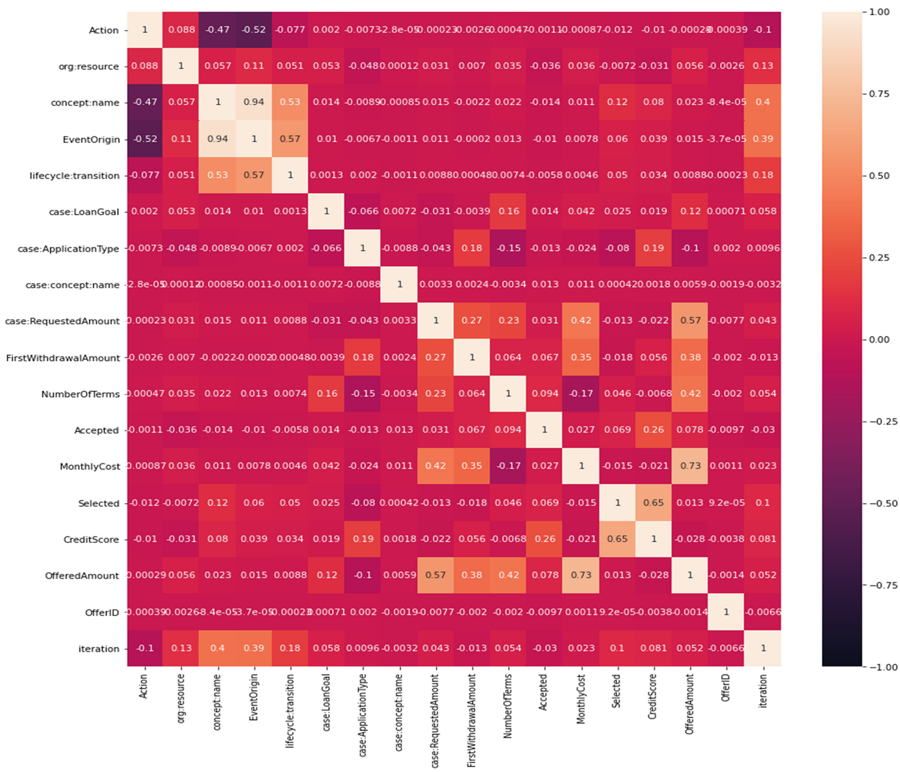

Fig. 4. Heatmap of the preprocess dataset

According to the heatmap, we can determine that the variable that have high correlation to our prediction variable are:

- concept:name
- EventOrigin

- Org:resource
- Selected
- CreditScore
- NumberOfTerms
- OfferedAmount
- Case:RequestedAmount
- Case:LoanGoal

This selection is motivated by all variable that has a correlation superior to 0.01 to the prediction variable which are "lifecycle:transition", "iteration". The specificity of those variables is described in the Table 1.

Table 1. Description of the selected variable

Variable name	Type	Number of unique value
concept:name	Categorical	26
EventOrigin	Categorical	3
Org:resource	Categorical	149
Selected	Boolean	2
CreditScore	Numerical	520
NumberOfTerms	Numerical	147
OfferedAmount	Numerical	663
Case:RequestedAmount	Numerical	701
Case:LoanGoal	Categorical	14

Based on our selected variables, we can see that there are some of them that are categorical. For those categorical variables, we need to encode them in order to be used inside the learning model. To do that we use the OneHotEncoder explained in Fig. 5.

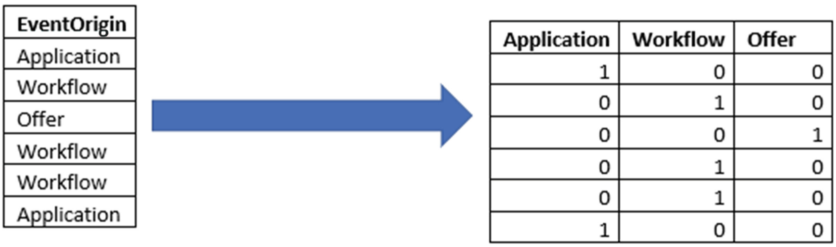

Fig. 5. How OneHotEncoder work

As for the prediction model, our target is when the "lifecycle: transition" is equal to "ate_abort" and when the "iteration" is superior or equal to 3.

Then we split the data into training data and test data with a 0.8 ratio. Which mean that the training data will have 80% of the total data and the test data will have 20%. We didn't do the shuffle because that won't represent the execution of a process.

5 Metric

In order to compare the predictions models, we need a uniform metric to compare them to. There is a common metric that usually permit to determine the others metrics. That metric is known as confusion matrix and is show in the Table 2. The matrix is composed of 4 values. According to [4], the terms ''positive" and "negative" refer to the classifier's prediction, and the terms "true" and ''false" refer to whether that prediction corresponds to the external judgment (correspond to the observation). The Table 2 represent the confusion matrix.

Table 2. Confusion matrix

		Prediction value	
		Positive	Negative
Observation	True	True Positive (TP)	False Negative (FN)
	False	False Positive (FP)	True Negative (TN)

The metrics used are the accuracy, the precision, the recall, the F1 score [17] and the ROC graph [18].

6 Prediction Model

With the popularization of learning technique, there are a lot more that are discover. For our prediction model we decide to limit our self to 4 prediction model. The prediction that we use are the Random Forest Classifier (RF), the Decision Tree Classifier (DT), the Logic Regressor (LR) and the Multi-Layer Perceptron Classifier (MLP). Among these models, we can say that the first 3 are machine learning algorithm and the last one use neural network for the prediction. To facilitate the learning, we decide to use pipeline. In the pipeline, we decide to use the column transformer parameter with the OneHotEncoder. This encoder concerns the following columns: 'org:resource', 'lifecycle:transition', 'EventOrigin', 'concept:name', 'case:LoanGoal'.

Since we split the date into train part and test part, we use a 80% ratio for the split for the train dataset. Therefore, the training dataset is a total of 961,813 events.

7 Result

In order to get the result, we need to test the predictive model with the test dataset. This test dataset consists of 240,454 events. The result we obtained are displayed in this section. So, in this section, we will first display the confusion matrix of each predictive model then, we will show a comparative ROC graph.

As described earlier, based on the confusion graph, we can determine others metrics like accuracy, precision etc. So, after we show the confusion graph, we will also display the other metric into a table named classification report. The classification report shows the precision, recall and f1-score metrics through a table. The percentage obtain in the confusion matrix is based on the number of events for the test dataset which is equal to 240,454.

7.1 Random Forest Classifier (RF)

The Fig. 6 represent the confusion matrix of the random forest classifier.

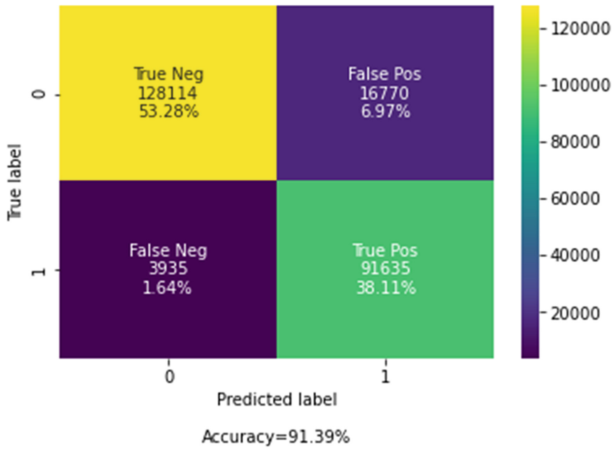

Fig. 6. Confusion matrix of the Random Forest Classifier

In the figure above, we have the confusion matrix for the Random Forest Classifier. In it, we can see that the True Negative is equal to 53.36% where the True Positive is 38.07%. As for the accuracy, it is equal to 91.39%. The Table 3 show the classification report of the RF.

Table 3. Classification report of RF

		Precision	Recall	f1-score	Support
Observation	0	0.88	0.97	0.93	132049
	1	0.96	0.85	0.90	108405
Average		0.92	0.91	0.91	240454

7.2 Decision Tree Classifier (DT)

The Fig. 7 represent the confusion matrix of the decision tree classifier.

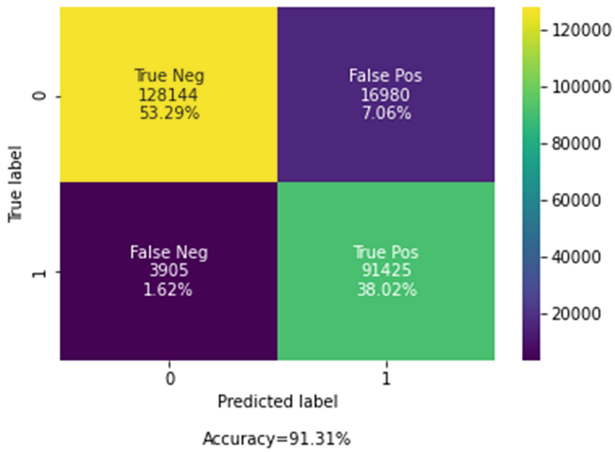

Fig. 7. Confusion matrix of the Decision Tree Classifier

In the figure above, we have the confusion matrix for the Decision Tree Classifier. In it, we can see that the True Negative is equal to 53.29% where the True Positive is 38.02%. As for the accuracy, it is equal to 91.31%. The Table 4 show the classification report of the DT.

Table 4. Classification report of DT

		Precision	Recall	f1-score	Support
Observation	0	0.88	0.97	0.92	132049
	1	0.96	0.84	0.90	108405
Average		0.92	0.91	0.91	240454

7.3 Logistic Regression (LR)

The Fig. 8 represent the confusion matrix of the logistic regression.

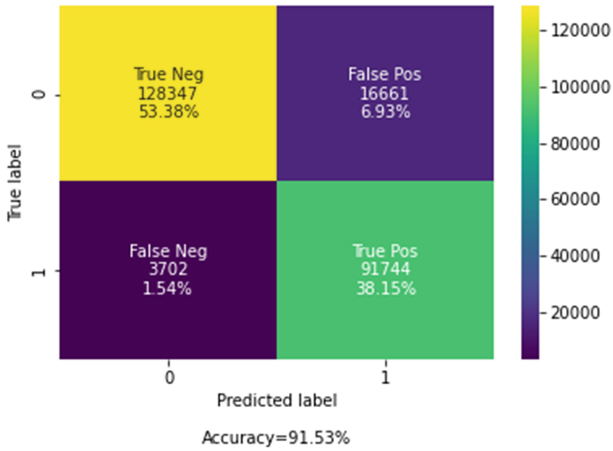

Fig. 8. Confusion matrix of the Logistic Regression

In the figure above, we have the confusion matrix for the Logistic Regression. In it, we can see that the True Negative is equal to 53.38% where the True Positive is 38.15%. As for the accuracy, it is equal to 91.53%. The Table 5 show the classification report of the LR.

Table 5. Classification report of LR

		Precision	Recall	f1-score	Support
Observation	0	0.89	0.97	0.93	132049
	1	0.96	0.85	0.90	108405
Average		0.92	0.91	0.91	240454

7.4 Multi-Layer Perceptron Classifier (MLP)

The Fig. 9 represent the confusion matrix of the Multi-Layer Perceptron classifier.

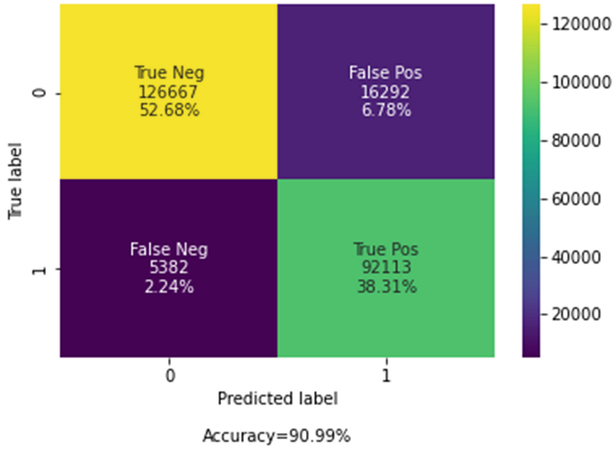

Fig. 9. Confusion matrix of the Multi-Layer Perceptron Classifier

In the figure above, we have the confusion matrix for the Multi-Layer Perceptron Classifier. In it, we can see that the True Negative is equal to 52.68% where the True Positive is 38.31%. As for the accuracy, it is equal to 90.99%. The Table 6 show the classification report of the MLP.

Table 6. Classification report of MLP

		Precision	Recall	f1-score	Support
Observation	0	0.89	0.96	0.92	132049
	1	0.94	0.85	0.90	108405
Average		0.92	0.91	0.91	240454

7.5 ROC Graph

The roc graph does a comparison of the different model use early.

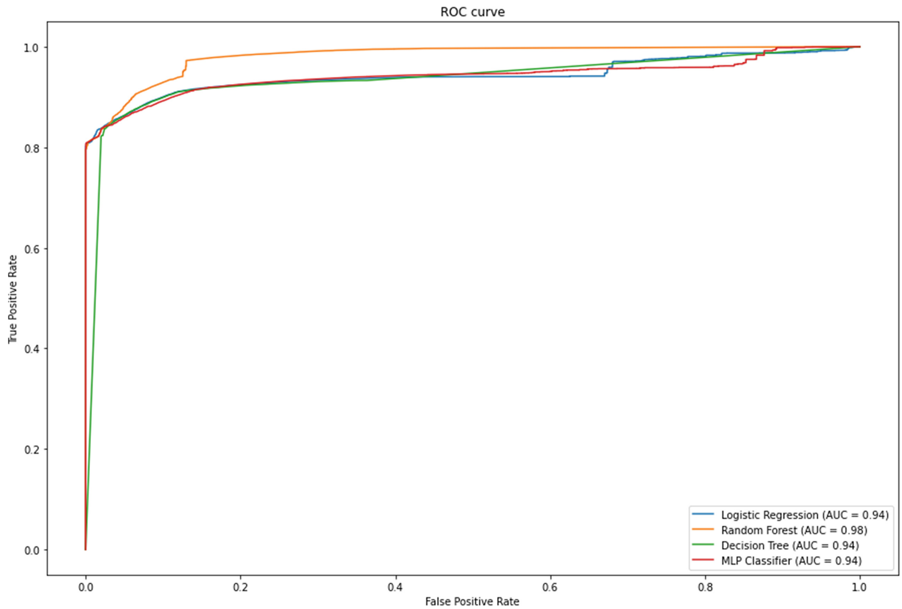

Fig. 10. ROC curve which compares the LR, RF, DT, MLP

The Fig. 10 represent the ROC graph which compares the different prediction model named earlier. In it, we can see the evolution of the False Positive Rate compare to the True Positive Rate during the training. According to [18], the Area Under ROC Curve (AUC) of a classifier is equivalent to the probability that the classifier will rank a randomly chosen positive instance higher than a randomly chosen negative instance. So, in our ROC curve, we can see that the RF score an AUC of 0.98 while the other prediction model has 0.94. However, it cannot determine clearly the final accuracy of the prediction model.

8 Conclusion

The prediction of business process failure is always useful for improving the efficiency of a business process. Therefore, taking into account the loop can be benefiting for a process model. Because it can reduce the number of repetitive events into an efficient one. In our process, with the applied data, we conclude that the logistic regression algorithm fit the best for the prediction with an accuracy of 91.53%. There may be better algorithm for the prediction. The hole process of prediction and analysis has been made by using the IBM cloud service. A comparison to other cloud service can be made to determine the one with best performance. Even if in our proposition we didn't use the timestamp, it may affect the final state of a process case or it can be the cause of loops.

References

1. Van der Aalst, W.: Process mining software. Process Min., 325–352 (2016). https://doi.org/10.1007/978-3-662-49851-4_11
2. Avizienis, A., Laprie, J.-C., Randell, B., Landwehr, C.: Basic concepts and taxonomy of dependable and secure computing. IEEE Trans. Dependable Secur. Comput. **1**, 11–33 (2004). https://doi.org/10.1109/TDSC.2004.2
3. Mariani, L., Pezzè, M., Riganelli, O., Xin, R.: Predicting failures in multi-tier distributed systems (2019).
4. Borkowski, M., Fdhila, W., Nardelli, M., Rinderle-Ma, S., Schulte, S.: Event-based failure prediction in distributed business processes. Inf. Syst. (2017) https://doi.org/10.1016/j.is.2017.12.005
5. Kang, B., Kim, D., Kang, S.H.: Real-time business process monitoring method for prediction of abnormal termination using KNNI-based LOF prediction. Expert. Syst. Appl. **39**(5), 6061–6068 (2012) ISSN 0957-4174. https://doi.org/10.1016/j.eswa.2011.12.007
6. Evermann, J., Rehse, J.R., Fettke, P.: A deep learning approach for predicting process behaviour at runtime, pp. 327–338 (2017). https://doi.org/10.1007/978-3-319-58457-7_24
7. van der Aalst, W.M.P., Weijters, A.: Process mining: a research agenda. Comput. Ind. **53**(3), 231–244 (2004)
8. van der Aalst, W., Dongen, J., Herbst, B.F.V., Maruster, G.W., Schimm, L.: Workflow mining: a survey of issues and approaches. Data Knowl. Eng. **47**(2), 237–267 (2003)
9. van der Aalst, W.M.P.: Process Mining: Discovery. Springer, Conformance and Enhancement of Business Processes (2011)
10. Camara, M.S., Fall, I., Mendy, G., Diaw, S.: IEEE 2015 19th International Conference on System Theory, Control and Computing (ICSTCC)—Cheile Gradistei, Romania (2015.10.14–2015.10.16). 2015 19th International Conference on System Theory, Control and Computing (ICSTCC)—Activity failure prediction based on process mining, pp. 854–859 (2015). https://doi.org/10.1109/ICSTCC.2015.7321401
11. van der Aalst, W.M.P., Hofstede, A.H.M.T., Kiepuszewski, B., Barros, A.P.: Workflow patterns. Distrib. Parallel Databases **14**(1), 5–51 (2003)
12. Chapman, P., Clinton, J.: Crisp-dm 1.0. SPSS Inc, Tech. Rep. (2000)
13. https://doi.org/10.4121/uuid:5f3067df-f10b-45da-b98b-86ae4c7a310b
14. Povalyaeva, E., Khamitov, I., Fomenko, A.: From the Moscow Higher School of Economics, with their submission entitled: Density Analysis of the Interaction with Client
15. Blevi, L., Robbrecht, J., Delporte, L.: From KPMG Technology Advisory, Belgium, with their submission entitled: Process mining on the loan application process of a Dutch Financial Institute
16. Rodrigues, A et al.: From Pontifícia Universidade, with their submission entitled: STAIRWAY TO VALUE: Mining the loan application process
17. Metrics and scoring: quantifying the quality of predictions—Scikit-learn 0.24.2 documentation 2021/09/10
18. ROC Graphs: Notes and Practical Considerations for Researchers Tom Fawcett March 16, 2004.