Business Process Intelligence Challenge 2018: Analysis and Prediction of Undesired Outcomes.

J.H.H. Brils[1,2], N.A.F. van den Elsen[1,3], J. de Priester[1,4], and T.A. Slooff[1,5]

[1] Eindhoven University of Technology, 5612AZ Eindhoven, Netherlands
[2] j.h.h.brils@student.tue.nl
[3] n.a.f.v.d.elsen@student.tue.nl
[4] j.d.priester@student.tue.nl
[5] t.a.slooff@student.tue.nl

Abstract. In the rapidly emerging discipline of process mining, the Business Process Intelligence Challenge offers an enticing way of applying novel techniques onto real-world process data. In this paper, we analyze a process of handling payment applications of German farmers to receive funding from the European Agricultural Guarantee Fund [4]. We focus particularly on the aspect of so-called 'undesired outcomes', and analyze what characterizes such cases. We outline the differences between such cases, with a focus on throughput time and departments. Lastly, we process the data such that they may be used by machine learning algorithms to predict undesired outcomes as early as possible. With these processed data, we create models that have a high accuracy in predicting cases with undesired outcomes before an initial payment decision is made.

Keywords: Business Process Intelligence · BPIC 2018 · ProM · Disco.

1 Introduction

Process mining is an emerging discipline that has experienced major developments over the last decade [5]. This year, it is the focus of the yearly Business Process Intelligence Workshop. Two of the aims of the workshop are to discuss the current state of the research discipline and to share practical experiences [3]. One means of doing so is the Business Process Intelligence Challenge (BPIC).

The BPIC 2018 considers the yearly process of handling payment applications of German farmers from the European Agricultural Guarantee Fund [4]. The German company **data experts** provided the data [7]. It provided an application log and a collection of document logs. Although the eldest data comes from May 2014, the data actually describes the process over three years. In this report, we define these years as cases that started in either 2015, 2016 or 2017.

The BPIC consists of four business questions. However, in the student category of the challenge it is suggested to focus on a specific aspect. Therefore, in this report, the focus lies on the business question related to **undesired outcomes**. The business question is defined as follows [4]:

A usual case is opened around May of the respective year and should be closed by the end of the year. By "closed", we refer to the timely payment of granted subsidies. There are, however, several cases each year where this could not be achieved. We would like to detect such cases as early as possible. Ideally, this should happen before a decision is made for this case.

Additionally, we take into account another business question. This is a question which can be combined easily with the question related to undesired outcomes. The question deals with **differences across departments** and is defined as follows:

How can one characterize the differences between departments and is there indeed a relation?

Our project approach consists of three parts: an overall analysis of the process, an analysis of cases with undesired outcomes and finally predicting cases with undesired outcomes. However, before we actually start predicting, data has to be prepared and that is where this report becomes unique. We introduce 'context switches' in order to generate valuable input data. A context switch is a change to either another document type or to another sub process. The concept is illustrated in Figure 1. A further elaboration on context switches and their added value to prediction can be found in Chapter 3.
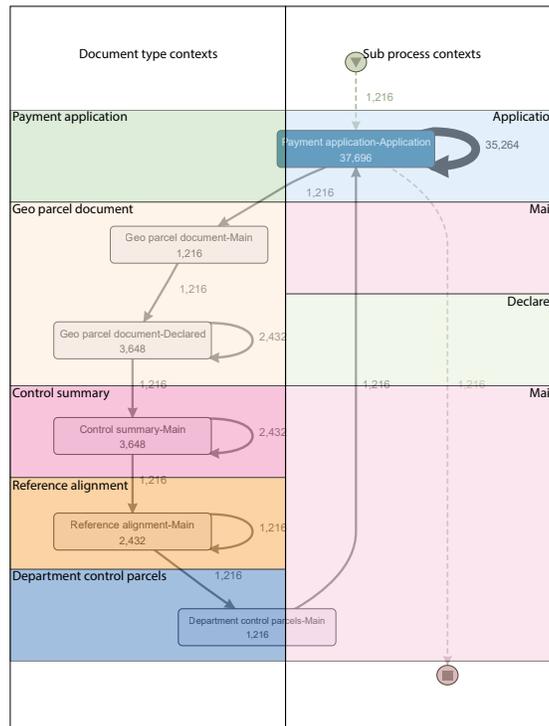

Fig. 1: Context switches

This report has the following structure. Chapter 2 contains the first overall analysis of the process. Chapter 3 is focused on the undesired outcomes. In this chapter, firstly, an analysis is performed. The second part of Chapter 3 contains two prediction models: a gradient boosted trees classifier to predict late cases and a gradient boosted trees classifier to predict reopened cases. Chapter 4 contains our conclusions.

2   Analysis of the process

This section contains a first analysis of the process. This analysis will provide us essential insights for a successful approach of the business questions. The analysis includes a generation and analysis of a process map and some descriptive statistics.

For this analysis, the application log is used [7]. This event log contains events for over 43,000 cases over a period of three years. In this log, each event describes the state of a specific document. Therefore, also this process analysis is centred around documents.

2.1   Process map

The analysis is performed using the process mining tool Disco [1]. A process map is created, as shown in Figure 2. In order to create the process map, the event log was first transformed to a comma separated value file format using Disco [1]. This way, we could centre the process around documents as mentioned before. Additionally, we could now simplify the process as we only focus on documents and not on activities or subprocesses. The process map shows all document types as activities and the arrows represent paths. To reduce the complexity and enhance understanding of the map only the most frequent paths are shown. The numbers represent the amount of cases that performed an activity or followed a path.

When analyzing Figure 2, one easily sees that there are multiple loops in the process. The most important observation regarding those loops is that most of the time the entire process starts and ends with the activity Payment application. Another observation is the repetition of activities. This is due to the fact that the process map is based on document types only and not on activities within a document type.

An analysis of the activities shows that there are differences in the amount of cases that have performed the activity. In the process map, one can see this difference by the difference between the colors of the activities. The darker the activity, the larger the amount of cases that have performed the activity. This analysis gives three key activities of the process: Payment application, Control summary, and Reference alignment. However, there is no case that only performed these three activities. One reason for that is that some changes have occurred in the process over time. These changes were already given by the challenge. From 2016, the activity Geo parcel document replaced the activity Parcel
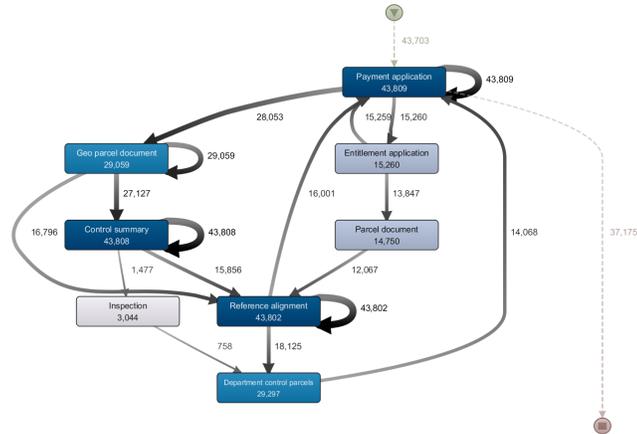
Fig. 2: Process map

document. In the process map, this can easily be seen as approximately only one third of the cases performed the activity Parcel document and approximately two third of the cases performed the activity Geo parcel document. Actually, there is no case that does not include either the activity Parcel document or Geo parcel document. After this change, the activity Entitlement application was only performed by a few percent of the cases. Later, from 2017, the Geo parcel document also replaced the Department control parcels, which therefore also occurred in only two third of the cases. Therefore, over the years the process is simplified to four main activities. The activity Inspection remained part of the process throughout the years, although only part of the cases perform the activity.

2.2 Descriptive statistics of the process

Now that we have created a visualization of the process and we thus understand what the process looks like, it is time to have a look at the numbers. This section therefore contains some descriptive statistics of the process.

The analysis is performed on the data from the event log. We have looked at all data, data per department and data per year. We manually created a spreadsheet using the statistics from Disco [1]. In the rest of this section, some tables with data derived from that spreadsheet are discussed.

Firstly, the case duration is discussed. The case duration is measured as the difference between the first event of a case and the last event of the case When aggregating over a department or over a year, one can look to either the mean or the median of all cases. Therefore, Table 1 and Table 2 respectively show the mean and median of the case duration over years and over departments. Both

tables show a decline in the case duration over the years. Especially between 2015 and 2016, progress has been made. Another similarity regarding the tables is that there are not many differences between the departments in the two outer right columns. The last similarity is that both tables show that in 2015, the cases handled in department d4 appeared to have a significantly longer duration than cases handled in the other departments. On the contrary, there are also some differences between Table 1 and Table 2. One difference is that Table 1 contains larger values than Table 2. Another difference is that the drop in case duration takes two years in Table 1 and only one year in Table 2. Both differences can be explained by the fact that the mean takes extreme values into account to a larger extend than the median.

Table 1: Mean duration

Years	2015	2016	2017
Department 4e (weeks)	60	42	36
Department 6b (weeks)	63	44	36
Department d4 (weeks)	74	43	36
Department e7 (weeks)	67	42	36

Table 2: Median duration

Years	2015	2016	2017
Department 4e (weeks)	44	36	36
Department 6b (weeks)	44	36	35
Department d4 (weeks)	68	36	36
Department e7 (weeks)	50	36	36

Secondly, the amount of activities is discussed. In the event log, an activity is defined as a combination of a document type, a sub process and an activity. Therefore, there are a lot more activities than in the process map, which only focused on the document type. Table 3 shows that over the years, the amount of activities has declined. This is similar to what was discussed in the analysis of the process map. The amount of document types is reduced and logically, the amount of activities reduced as well. Table 3 does not show major differences between departments.

Thirdly, the success ratio is discussed. Table 4 shows the percentages of cases that ended successfully over the years and over departments. Something remarkable is that every year, the success ratio of department e7 is lower than the overall

Table 3: Activities

Years	2015	2016	2017
All departments	147	145	102
Department 4e	138	129	88
Department 6b	138	127	89
Department d4	139	127	81
Department e7	139	133	88

success ratio. All other departments have higher success ratios than the overall success ratios.

Table 4: Acceptance ratio

Years	2015	2016	2017
All departments (%)	99.39	99.31	97.46
Department 4e (%)	99.49	99.43	97.74
Department 6b (%)	99.65	99.32	97.7
Department d4 (%)	99.45	99.4	97.85
Department e7 (%)	99.06	99.12	96.46

3   Undesired outcomes

data experts, the company that provided the data for this year's BPI Challenge, has formulated several business questions they would like participants to focus on. The first of these questions has to do with undesired outcomes. Such undesired outcomes can be divided into two categories: late payments, and reopened payments. These events result in untimely closing of a case.

Therefore, we have decided we would look into such cases, and see what we can do to detect such cases before they occur. To do this, we first analyze the available data, to see if we can find patterns, or other notable characteristics of cases with undesired outcomes. We also analyze if different departments affect cases with undesired outcomes differently, which is part of another of *data experts*' business questions. Then, we try to see if we can predict such cases, based on data that was gathered before a first decision is made for a case. We accumulate relevant features for predictions, and then we apply well-known machine learning algorithms to make these predictions.

3.1   Analysis of the data

In order to identify differences, it may be valuable to keep the undesired outcomes separated as they may have different causes. In this subsection we will look at the given data, and mine it for valuable information. This will be done primarily through the tools Disco [1] and ProM [9,8]. We will look at information that can be gathered from the data, but also try to find some differences between departments.

Undesired outcome 1: *The payment is late. A payment can be considered timely, if there has been a begin payment activity by the end of the year that was not eventually followed by abort payment.*

Undesired outcome 2: *The case needs to be reopened, either by the department (sub process Change) or due to a legal objection by the applicant (sub process Objection). This may result in additional payments or reimbursements (payment_actual$\{x\}$ > 0, where $x \geq 1$ refers to the xth payment after the initial one).*

Late Payments There is no easy way of filtering the cases to find cases with late payments. So in order to look at the differences, we added a case attribute ourselves with the use of Python. With the use of the ProM forum, we identified 3 situations in which a case is considered late:

1. There is no 'begin payment' event in the case.
2. The last 'begin payment' event is followed by an 'abort payment' event.
3. The last 'begin payment' event occurs in a later year than what the case was started in.

Please note that in all cases there is no 'finish payment' event. After this case attribute was added to the log, we used this log to filter on this case attribute and look at the statistics.

The natural result of having a late payment would be to have a longer case. So the first thing to look at is case duration. The median case duration is 22 months, and the mean case duration is 20.6 months. Looking at all cases; these have a median duration of 38.1 Weeks and a mean duration of 47.9 Weeks. The median and mean values for late payments are somewhat weird, as normally the mean duration is larger than the median. This is because it suffers more from outliers than the median, and especially outliers to the right (longer cases) have a large effect as these can be further away. So cases with late payments are quite a lot longer, with their median difference being more than a year.

When comparing the process maps of the cases with late payments and the general process map, the bottleneck is clearly visible with the payment application document. This is not surprising of course. There are some events which take a lot of time, which is definitely expected. However, a little bit less expected is the amount of cases that go through the change process. The change sub process occurs in 16.47% of the events in cases with late payments, while it only

occurs in 2.15% of the general cases. So relatively more cases with late payments change departments, compared to the general cases. And this takes part in the longer case durations, as this adds a lot of extra time to the case.

Table 5: General statistics cases

Attribute	small farmer	young farmer	selected risk
All cases (%)	3.98	8.55	8.2
Late payment cases (%)	1.54	10.55	5.02

In Table 5 the general statistics of the cases are combined. There are some small differences, with the most interesting being a small difference in the cases that are selected by risk assessment. It seems that cases with late payments were originally less likely to be selected as a risk. In the case attribute 'young farmer' there is a small increase in young farmers relative to the general cases. And lastly, contrary to what may be expected, the percentage of small farmers is smaller for cases with late payments. These statistics were selected because they could be interesting, but the differences here are quite small, so no conclusions should be drawn from them.

Table 6: Distribution of departments

Department	e7	4e	6b	d4
All cases (%)	29.31	29.25	25.8	15.64
Late payment cases (%)	30.12	25.57	22.92	21.39

Table 6 shows the distribution of departments over the cases. There is not a lot of difference between the distribution. Department 4e and 6b occur a little less in cases with a late payment, and department d4 a little more. Department e7 is very similar in both scenarios.

So the most interesting finding in these cases with late payments, is that they change departments significantly more often than general cases. This most probably involves some more detailed problems that we have no insight into, but further research into this could be useful for the company.

Re-opened cases This section looks at the cases which have been re-opened. This means they contain either the sub process 'Change' or the sub process 'Objection' (or both.) This does not necessarily mean that these cases lead to additional payments. Cases that need further payments or reimbursements are

discussed later in this section. The cases discussed here thus concern all re-opened cases, 11% of all cases had to be re-opened.

Cases that had to be re-opened had a median duration of 21.4 months, and interestingly the mean duration is 20.6 months. This is remarkably similar to the statistics of case durations for cases with late payments. Figure 3 shows a distribution plot of the case duration for re-opened cases. This figure was taken from Disco. In Disco itself this graph is interactive so you can investigate further, but when exporting the graph you can not add axes. From the figure it can be seen that the distribution of case duration is quite varied, there are quite some outliers to the left (short cases) as well as to the right (long cases). Comparing these numbers to the numbers of all cases combined, which have a median of 38.1 weeks and a mean of 47.9 weeks it shows that the cases which have to be re-opened cost more time. This is logical since there are extra processes involved with the re-opening of a case.
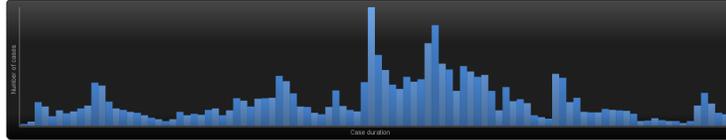

Fig. 3: Case duration for the re-opened cases

Table 7 shows the distribution of the departments over the cases. This shows the percentage of the cases handled by that department. Department e7 is the largest department, and handles a surprisingly similar percentage of re-opened cases compared to the normal percentage. Departments 4e and 6b both have a drop of around 5% in the cases they handle. That leaves department d4, which has an increase of around 10%. This indicates that department d4 may get more difficult cases, which need to be re-opened.

Table 7: Distribution of departments

Department	e7	4e	6b	d4
All cases (%)	29.31	29.25	25.8	15.64
Re-opened cases (%)	30.02	25.14	19.84	24.99

It may be helpful to look at the sub processes for the re-opened cases in department d4, and compare this to the rest of the re-opened cases. In department d4, the relative frequency of 'Change' and 'Objection' is 12.66% and 2.81% respectively. Compare this to the rest of the re-opened cases, where we find a

relative frequency of 12.84% for 'Change', and 4.76% for 'Objection'. So department d4 does not seem to have a problem with having more cases changed or having more legal objections. As a matter of fact, they have relatively fewer legal objections. So it seems that department d4 simply has bad luck with getting more tough cases, or they get more cases transferred to them.

We have found some statistics indicating that these cases cause problems, in that they cost quite a bit more time. However, in the given characteristics of undesired outcome 2, it is also noted that some of these cases may need additional payments or reimbursements. These cases are more valuable to investigate, since they cost even more. We look at these in more detail later in this section, where we look at difference between the general cases, the re-opened cases, and these re-opened cases that need additional payments or reimbursements.

Cases with additional payments or reimbursements For this section the analysis is concerning all cases that have some sort of additional payments or reimbursements. This means *payment_actual1* was not 0 and not empty. 5% of cases have some sort of additional payment or reimbursement. For the rest of this section, these cases may sometimes be referred to as 'problematic cases'.
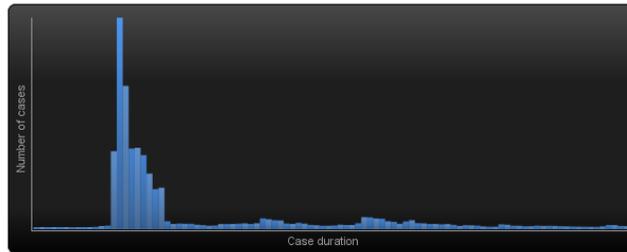
(a) Case durations of all data
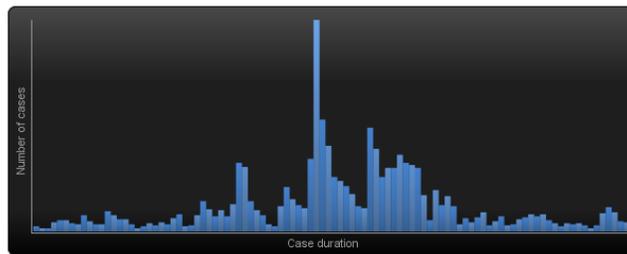
(b) Case durations of problematic cases

Fig. 4: Case durations

It becomes clear from Figure 4 that cases with additional reimbursements do not only increase costs with regards to additional payments, but they also

require much more time. These figures were also taken from Disco, so the axes do not have a scale. To give a clearer picture of the difference; the biggest peak in Figure 4a is for 248 days and 9 hours, and the biggest peak in Figure 4b is 1 year and 250 days. This already indicates a big difference.

The mean for all cases is 47.9 weeks, and the median (which is less sensitive to outliers) is 38.1 weeks. For cases with additional payments or reimbursements the mean is 21.3 months, and the median is 21.1 months. This further establishes that these problematic cases take much more time. However it does not differ largely from all re-opened cases (including re-opened cases without additional payments). The mean does increase, which indicates that more of the longer outliers are cases that need additional payments. Lastly, the amount of events that occur in cases with additional payments are typically much larger than with all cases included. This makes sense, as there are obviously more sub processes involved when (for instance) handling legal objections.

So this establishes these cases as problematic. But what causes this? We will predict these problematic cases in Section 3.2. Here we find interesting differences between these problematic cases and the rest. The percentage of cases that were rejected was not very different with the general percentage at 0.69% and the problematic cases at 0.77%. The amount of small farmers is smaller with these problematic cases: 0.72% compared to 3.98% when looking at all cases. The amount of young farmers increases slightly in the problematic cases: 10.46% compared to 8.55% in general. Table 8 shows an overview.

Table 8: General statistics cases

Attribute	small farmer	young farmer	rejected
All cases (%)	3.98	8.55	0.69
Problematic cases (%)	0.72	10.46	0.77

Lastly, the distribution of cases over the departments changes between the general cases and this particular problematic case. Table 9 shows this distribution, where the % means the percentage of cases being handled by that department. For department e7 and 6b, the percentages are remarkably similar between the general cases and the problematic ones. However, departments 4e and d4 do change.

It also helps to look at the difference in distribution between the problematic cases, which are cases that need additional payments or reimbursements, and the re-opened cases. Re-opened cases also include cases that do not need additional payments, that is: the problematic cases are a subset of the re-opened cases. Interestingly, while department 6b does not change in distribution compared between problematic cases and all cases, there is a difference between the problematic cases and the re-opened cases. This means they handle a fairly larger share of the problematic cases, compared to the amount of re-opened cases they

get. On the other hand, department d4 gets a larger amount of re-opened cases, but has to handle less problematic cases. Although this is still relatively larger than the general cases, as they handle a lot more re-opened cases.

Table 9: Distribution of departments

Department	e7	4e	6b	d4
All cases (%)	29.31	29.25	25.8	15.64
Problematic cases (%)	30.2	23.53	25.87	20.4
Re-opened cases (%)	30.02	25.14	19.84	24.99

When looking at the process maps, it seems that there is a clear bottleneck with the problematic cases in that it takes a long time to insert a document after 'Payment application-Application-finish payment'. However the process map of the problematic cases seems to be a bit more standardized. With this we mean there are less different variations in going from begin-point to end-point. When looking at the general cases this is a lot more diverse and less well-defined. But this is logical as simply the amount of cases plays a role in this. When taking a look at the statistics of the variants, it appears that the general cases have a clear cut most-occurring variant, but suffers from a lot of outliers. On the other hand the problematic cases do not have one clearly more-occurring variant. This is illustrated in Figure 5. Note that again it is not possible to set a scale on these images, so this is just to illustrate the difference roughly. The peak in Figure 5a shows that it is one small section of variants (variants 1 - 289) which addresses the largest amount of cases. On the other hand, with the problematic cases there are only 11 variants which have more than 1 case following that specific variant. All other variants have only 1 case following that variant, i.e. almost every case is unique in what process they exactly follow.

Concluding from this, an opportunity for improvement would be to further standardize the cases where additional payments are necessary. This may help in reducing the time these cases take. Furthermore, small farmers appear to cause less problematic cases, whereas young farmers constitute a slightly higher percentage of problematic cases. Perhaps young farmers and large farmers need to be better informed of regulations and procedures, or they need to be handled different internally. It might also help to look more closely into departments, as there seems to be a large discrepancy for some departments in their rate of problematic and re-opened cases.

3.2   Processing the data and predicting undesired outcomes

To be able to predict whether a case has an outcome that is not desired, we first do some preparations of the data. The provided data cannot be used with common machine learning algorithms for several reasons. Firstly, the format of

(a) Case variants for all data

(b) Case variants for problematic cases

Fig. 5: Case variants

the data is an event log. Such event logs do not have a well-defined tabular structure. To resolve this, we use Disco's export feature to generate a comma-separated values (csv) file for the data. However, this generated file can still not be used for prediction, since each case now has a row for every event in that case. Prediction does not aggregate these rows, which means that one could only be able to predict undesired outcomes based on single events.

Ideally, we would want one row per case, with some well-defined features for these cases. These features would then summarize the events for this case. Fortunately, we already have some per-case data available, namely the so-called 'trace attributes'. These provide information regarding the application, and potential penalties or inspection selections. However, by merely using these data, we omit the traces completely, which is not something we would wish for.

Thus, we run some scripts on the data to retrieve further statistics from the traces. Note that we only generate these statistics from the events that occur before the first occurrence of 'Payment application-Application-decide'. This allows us to try to predict undesired outcomes before a decision for a case is made.

The first statistic we generate is simply counting how many times an event in a certain sub process or with a certain document type has occurred. The reasoning behind this is that perhaps more activities within a sub process or on a document type could indicate doubt, or uncommon behavior.

A further statistic that may provide additional value is the time in days between the first event in a case and the 'Payment application-Application-decide' event. This gives us an indication of the duration of the overall process so far. Say that an application was delayed several times, and therefore takes a relatively long time before a decision is made, this could indicate that payments may be expected to be late as well.

We can also look at the number of distinct documents that have been handled in a case so far. By counting the number of unique values for the 'docid' attribute in a case, we can easily spot outliers or nominal cases.

Beyond counting occurrences of sub processes or document types, we can look into how much time was spent working on a specific sub process or document type. To do this, we define 'context switches'. A context switch can occur either by switching to a different sub process, or by switching to a different document type. This is illustrated in Figure 1, where different colors indicate different contexts. Note that in this figure, the left and right column correspond to document type and sub process contexts respectively, and similar colors between the columns have no additional meaning. Now, we count the time that is spent between the first event and last event between two context switches. This gives us a time that was spent on a specific sub process or document type. Note that if a sub process or document type occurs in multiple contexts, we calculate the sum of the time spent. Furthermore, we calculate and store the number of document type switches, the number of sub process switches, and the mean, minimum, and maximum time spent in document type and sub process contexts.

Now, we have generated the following attributes from all data until the first decision ('Payment application-Application-decide'), alongside the given trace-specific attributes that were already included in the data:

– Number of events in each sub process
– Number of events with each document type
– Average, minimum, and maximum amount of time spent in a sub process context
– Average, minimum, and maximum amount of time spent in a document type context
– For each sub process: the time spent in such sub process contexts
– For each document type: the time spent in such document type contexts
– The number of days until a first decision is made

Predicting undesired case #1: late payments In order to predict whether or not a case is going to result in a late payment, machine learning is used. Machine learning allows for automatic generation of decisions, based on provided data. Alternatively, we could make hypotheses manually and verify these with the available data, but machine learning allows for a greater degree of automation [6]. For machine learning it is essential to clean your data set prior to classification. This means that the attributes in the data set have to be checked

on relevance to the prediction class. The attributes that are removed from the data set can be seen in Table 10.

Table 10: Attributes removed from the data set with motivation

Removed attribute	Motivation
Case_id	Adds no information; all values are unique
Application	Adds no information; all values are unique
Basic payment	All values true; no unique values
Greening	All values true; no unique values
Program_id	No unique values
Rejected	Removed since they do not apply for this problem
Year	Prediction based on years is not applicable for future predictions
Penalty_amount	Almost all data is missing

Furthermore, cases that have started in 2017 are removed from the data set. This is due to the fact that it is unknown whether or not these cases were labelled as late. The final preprocessing step is creating a data set with a 50-50 division for late and on-time cases. This is done to remove bias from the original data set, which contains a 95-5 division between on-time and late cases. The sampling for creating this 50-50 division data set is random. If we do not create this 50-50 division, most algorithms that we used predict too many cases to not have an undesired outcome, as this would result in a higher overall accuracy.

A number of models are compared on their accuracy and ROC curve, as can be seen in Figures 6 and 7 respectively. This process is done in Rapidminer [2].
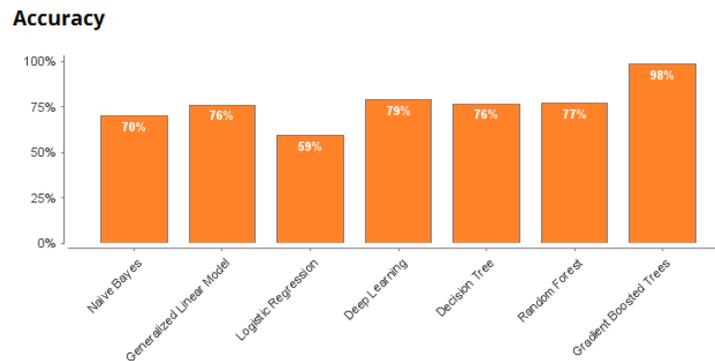

Fig. 6: Accuracy of different models on 50-50 division training set when predicting late payments

The accuracy of a model is calculated by dividing the number of correctly predicted instances over the total number of instances. The ROC curve is created
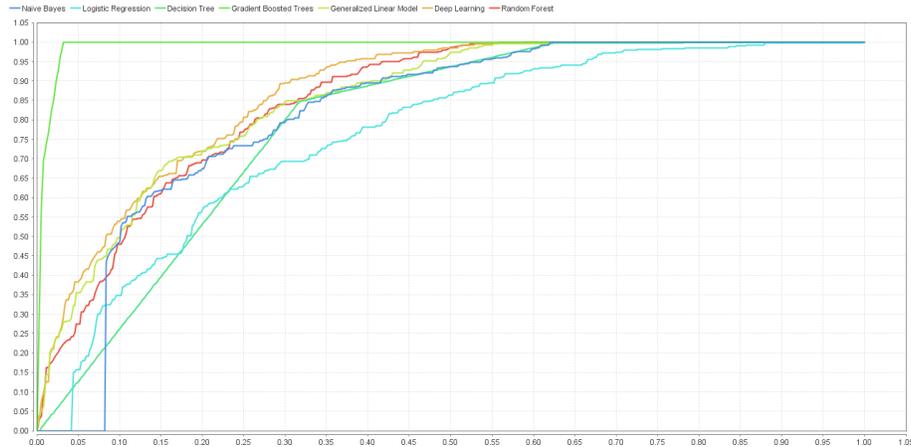

Fig. 7: ROC curves of different models on 50-50 division training set when predicting late payments

by plotting the true positive rate (TPR) over the false positive rate (FPR) at different thresholds. The best performing classifier is the *gradient boosted trees*, according to Figures 6 and 7 with parameters:

– Number of trees = 20
– Maximal depth = 2

These parameters were determined by first varying the depth of the trees to get the highest accuracy, and then varying the number of trees until the accuracy would no longer increase.

Going back to the full size date set, the *gradient boosted trees* classifier is trained on the 50-50 division data set and then applied to the whole data set. This results in an accuracy of 95.52%.

Predicting undesired case #2: reopened cases A similar approach as for predicting late payments is used for predicting reopened cases. The same attributes listed in Table 10 are removed from the data set. Again the distribution between reopened and regular cases is not even. The division is 13% reopened 87% regular. As stated before it is important to train a classifier on a 50-50 division training set. Multiple classifiers are compared based on their accuracy and ROC curve on the 50-50 division training set. Results can be seen in Figures 8 and 9.
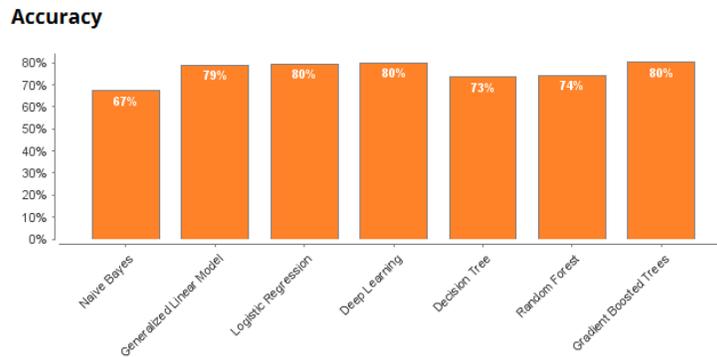
Accuracy

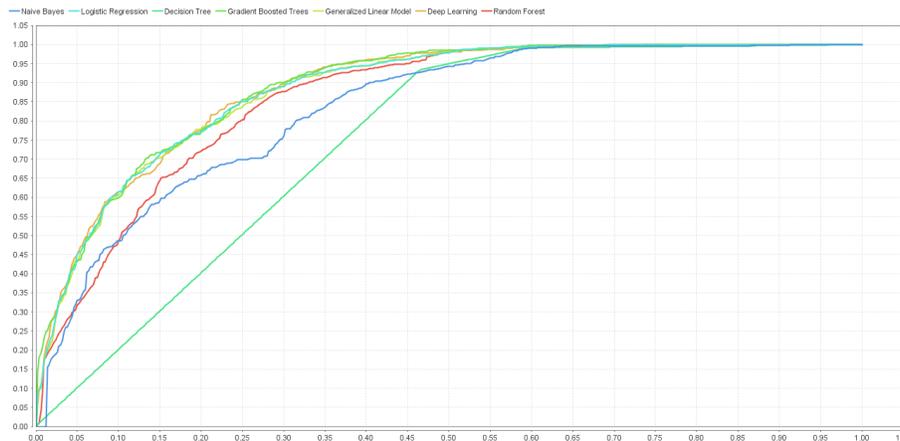Fig. 8: Accuracy of different models on 50-50 division training set when predicting reopened cases

Fig. 9: ROC curves of different models on 50-50 division training set when predicting reopened cases

The best performing classifier is *gradient boosted trees* according to Figures 8 and 9, with parameters:

– Number of trees = 100
– Maximal depth = 7

These parameters were determined by first varying the depth of the trees to get the highest accuracy, and then varying the number of trees until the accuracy would no longer increase.

Going back to the full size date set, the *gradient boosted trees* classifier is trained on the 50-50 division data set and than applied to the whole data set. This results in an accuracy of 82.61%.

The value of predictions We noticed that we achieved a prediction accuracy of 95% on late payments, and 82% on reopened cases. This means that using data that is available before a first decision is made in a case, we can, with high accuracy, predict whether this case will have an undesired outcome. Such predictions can then be used to focus more on cases with expected undesired outcomes, to prevent future problems. Perhaps providing simple tips or guidance on the application process to applying farmers can reduce the number of cases with undesired outcomes.

4   Acknowledgement

5   Conclusion

In this paper, we outlined the process of payment applications to the European Agricultural Guarantee Fund. We analyzed the way the process is executed, and how this varied over the years that the data spans. We clarified how this process differed between the four departments that were specified, by particularly focusing on the duration of the process. We discovered that the process has shortened over the years, and one of the departments has a lower success ratio than the others.

Then, we dove into the question of undesired outcomes. By doing in-depth analyses of cases with undesired outcomes and comparing this with general cases, we made several interesting observations:

- Interestingly, late payments occur less with small farmers
- Departments 4e and 6b have a lower share of late payments
- Whereas department d4 has a notably higher share of late payments
- Department d4 also has a larger share, proportionally, of reopened cases
- Cases with additional payments or reimbursements take quite a bit longer than nominal cases
- Additional payments or reimbursements also occur less with small farmers

We then prepared the data for predictions. Besides using the provided trace-level attributes, we included many additional attributes for prediction. We summarized the events that take place before a first decision is made, and included counts of different sub process or document type occurrences. We also introduce the concept of context switches, to include information regarding changes in sub process or document type. By further including mean, minimum, and maximum times in such contexts, we allow for predictive models to utilize anomalies in the process.

Then, two models have been built for predicting undesired outcomes. The first model uses a tuned gradient boosted trees classifier and can predict whether or not a case will be late with 95.52% accuracy. The second model also consists of a tuned gradient boosted trees classifier and can predict whether or not a case will have to be reopened with 82.61% accuracy.

References

1. Disco, https://fluxicon.com/disco/
2. Rapidminer, https://docs.rapidminer.com/
3. Business process intelligence (May 2018), https://www.win.tue.nl/bpi/doku.php?id=2018:start
4. Business process intelligence challenge (bpic) (May 15 2018), http://www.win.tue.nl/bpi/doku.php?id=2018:challenge
5. van der Aalst, W.: Process Mining: Data Science in Action (2016). https://doi.org/10.1007/978-3-662-49851-4
6. Ertel, W.: Machine Learning and Data Mining, pp. 161–220. Springer London, London (2011). https://doi.org/10.1007/978-0-85729-299-5_8, https://doi.org/10.1007/978-0-85729-299-5_8
7. Van Dongen, B.B., Borchert, F.F.: Bpi challenge 2018 (2018). https://doi.org/10.4121/uuid:3301445f-95e8-4ff0-98a4-901f1f204972, https://data.4tu.nl/repository/uuid:3301445f-95e8-4ff0-98a4-901f1f204972
8. Van Dongen, B.F., de Medeiros, A.K.A., Verbeek, H., Weijters, A., Van Der Aalst, W.M.: The prom framework: A new era in process mining tool support. In: International Conference on Application and Theory of Petri Nets. pp. 444–454. Springer (2005)
9. Verbeek, H., Buijs, J., Van Dongen, B., van der Aalst, W.M.: Prom 6: The process mining toolkit. Proc. of BPM Demonstration Track **615**, 34–39 (2010)