Università Degli Studi di Milano

Master Degree in Computer Science

BPI Challenge 2019

Giovanni Arvati

Business Information Systems

Prof. Paolo Ceravolo

A.Y. 2023/2024

Contents

1 Case Study Description

The case study considered for this project is the BPI challenge 2019. For that year challenge, the work focused on data from a multinational company operating from The Netherlands in the area of coatings and painting. The goal of the challenge was to investigate the purchase order handling process of the company.

The code relative to the work explained in the following section is shown in this Notebook.

1.1 Dataset

The dataset provided by the company is a selection of some of its 60 subsidiaries and consists of a collection of purchase orders. Each purchase order contains one or more line items and for each line item there are roughly four types of flows possible in the data:

1. **3-way matching, invoice after receipt**
   For these items, there is always an invoice receipt after a goods receipt.

2. **3-way matching, invoice before receipt**
   For these items, invoices can be entered before the goods are receipt.

3. **2-way matching, no receipt**
   For these items, there is no separate goods receipt message required.

4. **Consignment**
   For these items, there are no invoices as this is handled fully in a separate process

The dataset refers to orders submitted in 2018 and the log is anonymized, but some semantics are left, like:

- The **resources** are split between *batch users*, that refers to automated processes executed by the system, and *normal users* , that refers to human actors in the process.
- **Company**, **vendor**, **system** and **document names** and **IDs** are again fully anonymized in a consistent way throughout the log.

The event log is fully IEEE-XES compliant and contains a total of 251,734 cases from 76,349 purchase documents. In these cases, there are 1,595,923 events relating to 42 activities performed by 627 users (607 human users and 20 batch users).
In the event log the following attributes are recorded:

- **concept:name**: A combination of the anonymized purchase document id and the anonymized item id,
- **Purchasing Document**: The anonymized purchasing document ID,
- **Item**: The anonymized item ID,
- **Item Type**: The type of the item,
- **GR-Based Inv. Verif.**: Flag indicating if GR-based invoicing is required (see above),
- **Goods Receipt**: Flag indicating if 3-way matching is required (see above),
- **Source**: The anonymized source system of this item,
- **Doc. Category name**: The name of the category of the purchasing document,
- **Company**: The anonymized subsidiary of the company from where the purchase originated,

- **Spend classification text**: A text explaining the class of purchase item,
- **Spend area text**: A text explaining the area for the purchase item,
- **Sub spend area text**: Another text explaining the area for the purchase item,
- **Vendor**: The anonymized vendor to which the purchase document was sent,
- **Name**: The anonymized name of the vendor,
- **Document Type**: The document type,
- **Item Category**: The category as explained above.

2 Organizational Goals

As already mentioned above, the goal of the challenge was to investigate the purchase order handling process of the company. More in detail, the company's interest can be summarised in three main questions to answer:

The first one, is to try to **find a collection of process models which together properly describe the process in this data**. The four categories explained previously will suggest that at least four models are needed to describe the whole process in the data, but the complexity of the latter goes further than just the division in four categories.

The second question is relative to the invoicing process. the company is interested in knowing which is the **throughput of the invoicing process**, i.e. the time between goods receipt, invoice receipt and payment (clear invoice).

The third, and last, part is about what it has been called *anomalies*, that is all of that behaviour that are considered unexpected. In particular, the focus of this part is about the Purchase Documents that have the most activities related to them, and to the users that in the log have produced a lot of rework.

3 Knowledge Uplift Trail

In an attempt to answer the questions raised by the challenge, described in the previous section, the work has been divided mainly into four parts. This four parts are: **preprocessing**, **process discovery**, **invoicing throughput analysis** and **anomalies analysis**. A schematic and direct representation of the work performed is shown in Table 1.

The preprocessing part was done on the original dataset as it was made available. For all the following parts the work was instead done on the preprocessed dataset.

	Step	**Input**	**Acquired Knowledge**		**Output**
			Analytics/Models	**Type**	
Preprocessing	1	BPI Challenge 2019 Dataset	Timestamp	Prescr	Dataset filtered by date
	2	Step 1	Zero duration cases	Prescr	Dataset filtered by zero duration cases
	3	Step 2	Incomplete and non-compliant cases	Prescr	Dataset with only complete cases

Process Discovery	4	Step 3	Splitting dataset by Category, Doc. Type and Item Type	Prescr	8 datasets
	5	Step 4	Filtering by infrequent variants	Prescr	Frequent variants datasets
	6	Step 5	As-Is models	Descr	8 bpmn models
	7	Step 6	Fitness	Descr	Models fitness for each categories
Inv. Throughput	8	Step 3	Checking conformance of invoicing loops	Prescr	Dataset filtered on invoicing data
	9	Step 8	Time taken by the invoicing process	Descr	Invoicing throughput by categories
Anomalies Analysis	10	Step 3	Activity *Remove Payment Block*	Descr	Evidence of unexpected bahaviour
	11	Step 3	Activities per Purchase Document	Descr	Document with higher number of activities
	11	Step 3	Rework activities	Descr	Users with most rework

Table 1: Knowledge Uplift Trail Table

4 Project Results

In this section all the work that has been done will be explained in detail and the results will be shown using tables and graphs.

4.1 Preprocessing

The **preprocessing** was a key part of the work, because the dataset was very large and varied and there was the necessity to reduce it to a dataset without undesired records or cases.

The first thing done was **filtering data based on the timestamp of events**. In particular, as mentioned in the reference page of the challenge, the dataset refers to data collected during 2018, so all events with a timestamp preceding the first of January 2018 were removed. In addiction, the start date of the challenge was 28 January 2019, so all events with a equal or later date have also been removed.

Then, all the cases with **zero duration** were checked. Thanks to the analysis of the ending activities of the zero duration cases, it was possible to keep in the dataset those in which the ending activity suggested that there is no error. All the cases that ended up with the activity *Delete Purchase Order Item* were retained, while the others were removed.

The last step in the preprocessing process was to filter out all the **incomplete and non-compliant cases**. For the incomplete cases the focus was on the end activities. In particular, all the cases belonging to the first three category, i.e. *3-way matching, invoice after receipt*, *3-way matching, invoice before receipt* and *2-way matching, no receipt*, are involved in the invoicing process. This means that all this cases need to end with the event *Clear Invoice*. For *Consignment* cases, in which the invoicing process is not requested, the only end activity accepted became *Record Good Receipt*.

The non-compliance analysis focused on the invoicing process, because it is probably the most important part of the whole process and because the invoicing process is the main focus of the third part of this work, so it was necessary to not have compliance problem with it.
As already mentioned above, for the *Consignment* cases there are no invoices, so these cases were ignored during this analysis. Then, there were identified

some rules to ensure compliance. If any case did not follow one or more of the rules, that case would be eliminated.
For the cases of the *3-way match* categories the rules identified were the following:

- `n° Record Goods Receipt > 0`
- `n° Record Invoice Receipt > 0`
- `n° Record Goods Receipt = n° Record Invoice Receipt`
- `n° Record Invoice Receipt = n° Clear Invoice`

In that way it was possible to ensure that every case had at least one *Record Goods Receipt*, one *Record Invoice Receipt* and one *Clear Invoice*, and for the cases with more than one receipt, that the number of *Record Goods Receipt*, *Record Invoice Receipt* and *Clear Invoice* matched.
For the *2-way match* cases, due to the absence of *Goods Receipts*, the rules were slightly different:

- `n° Record Invoice Receipt > 0`
- `n° Record Invoice Receipt = n° Clear Invoice`

After all the preprocessing step it was possible to obtain the filtered dataset that would be the working dataset for all the subsequent steps. A numerical comparison between the original dataset and the filtered one in shown in Figure 1 (Note that in the figure it is used the *incomplete* notation to identify all the cases removed from the orginal dataset). From the original dataset containing 248204 cases we arrived to the filtered one with a total of 183131 cases.

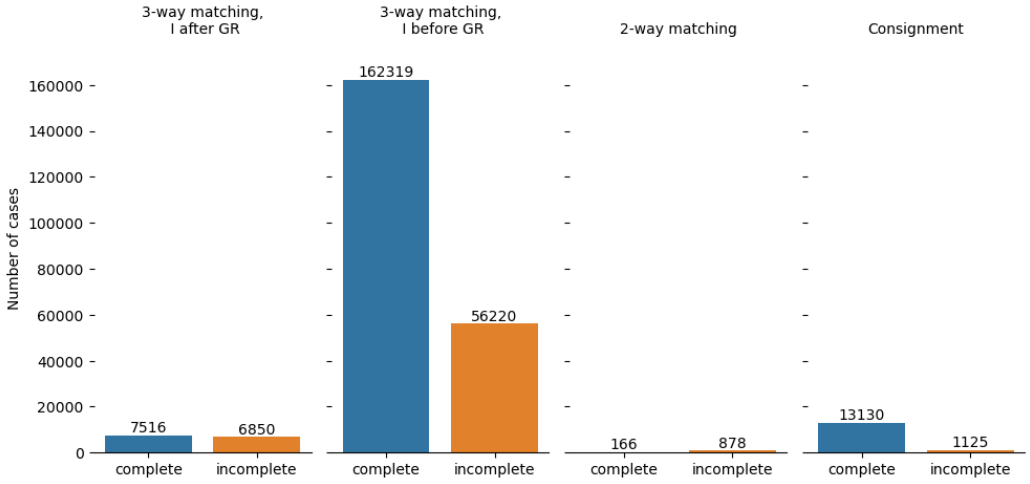

Figure 1: Complete and Incomplete cases divided by Item Category

4.2 Process Discovery

The first request of the challenge was to find a collection (at least 4) of process models that which properly describe the whole process in the data. The idea was to try to find some clusters for which the cases inside the same cluster could follow the same process model. To do this, the first thing was to group data based on some attributes that seemed to be useful for this purpose, i.e. *Item Category*, *Document Type* and *Item Type*. The result of the grouping is shown in Figure 2 with the number of cases for each group.

Looking at Figure 2 it was possible to learn some information from the data in the log file. For example, that the Item Type *Limit* is linked only with *2-way match* category. Another important thing is that the category identified by the tuple (*3-way match, invoice before GR*, *Standard PO*, *Standard*) is by far the one with the most cases, around 900000 with the second highest having only around 35000.

case:Item Category	case:Document Type	case:Item Type	
2-way match	Framework order	Limit	1016
3-way match, invoice after GR	EC Purchase order	Service	4518
		Standard	1163
	Framework order	Service	4563
	Standard PO	Service	4815
		Standard	35381
		Subcontracting	32
		Third-party	1313
3-way match, invoice before GR	EC Purchase order	Standard	8582
	Standard PO	Standard	912185
		Subcontracting	10562
		Third-party	21833
Consignment	Standard PO	Consignment	32574

Figure 2: Table representing the number of cases for each category obtained after grouping by *Item Category*, *Document Type* and *Item Type*

4.2.1 SRM activities

After grouping the cases as above and observing the event log, it was possible to see that there were a number of activities named by the prefix SRM that had a distinctive behaviour. The acronym *SRM* usually stands for *Supplier Relationship Management*, so all the activities with this prefix refers to activities not performed by a user. Furthermore, it could be seen that activities named by the prefix SRM only appeared in cases where the Document Type was *EC Purchase Order*. Due to the evidence just described, it seemed reasonable to **isolate cases that concerned SRM** from those that did not.

4.2.2 Invoice before Receipt

The dataset presented the events already divided into four categories. Of these, the two categories named with *3-way match* differed only in the order

in which invoice and receipt could appear. The cases belonging to the category *3-way match, invoice after GR* always had *Invoice Receipt* after *Goods Receipt*, in contrast, the cases belonging to the category *3-way match, invoice before GR* could present I*nvoice Receipt* before *Goods Receipt*, but this is not a rule but rather simply a possibility.
Knowing that, the cases in which the *Invoice Receipt* appeared before a *Goods Receipt* were counted, resulting in this happening only the 8.9% of the total cases of *3-way match, invoice before GR*. Being aware of these numbers, the decision was made to **merge the two 3-way match categories**, as only in a small percentage did the behaviours differ.

4.2.3 Models' categories

As a result of the evidence explained in 4.2.1 and 4.2.2 it was possible to identify 8 different categories that properly describe the whole process in the data. The 8 categories are the following:

1. **3-way match - SRM - Service**
2. **3-way match - SRM - Standard**
3. **3-way match - NO SRM - Service**
4. **3-way match - NO SRM - Standard**
5. **3-way match - NO SRM - Subcontracting**
6. **3-way match - NO SRM - 3rd Party**
7. **2-way match**
8. **Consignment**

It is important to note that in the above list, the notation SRM and NO SRM has been used to specify `Document Type` = `EC Purchase Document` and `Document Type` ≠ `EC Purchase Document`, respectively.

After splitting the dataset according to these 8 categories, each dataset is filtered by removing the infrequent behaviour. In particular, for each dataset, only variants appearing in at least 2% of the cases were retained. The number of cases for each category before and after filtering is shown in the Table 2.

cat	# cases	# filtered cases	ratio
1	213	111	52.1%
2	787	666	84.6%
3	831	753	90.6%
4	162306	128962	79.5%
5	1487	829	55.7%
6	4211	3797	90.2%
7	166	128	77.1%
8	13130	11905	90.6%

Table 2: Number of cases before and after the infrequent behaviour filtering

Having the datasets filtered by infrequent behaviour it was then possible to generate the process models. To visualize these models, BPMN was chosen as it represents the standard for business process modelling. To generate the models it has been used the *Inductive Miner algorithm* with a noise threshold of 0.4. The models are shown from Figure 3 to Figure 10.

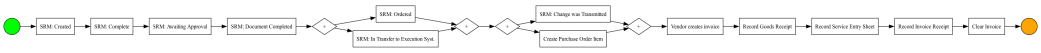

Figure 3: Category 1 BPMN model

Figure 4: Category 2 BPMN model

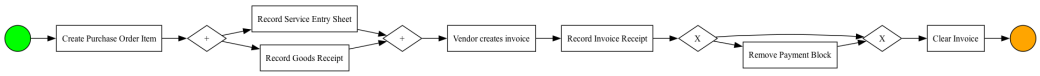

Figure 5: Category 3 BPMN model

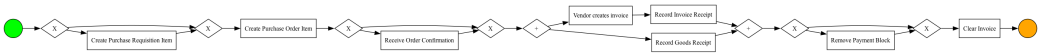

Figure 6: Category 4 BPMN model

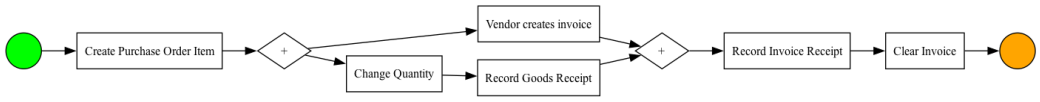

Figure 7: Category 5 BPMN model

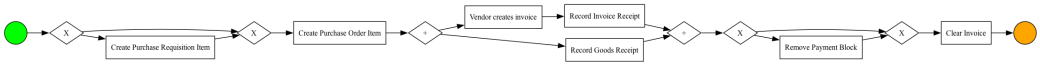

Figure 8: Category 6 BPMN model

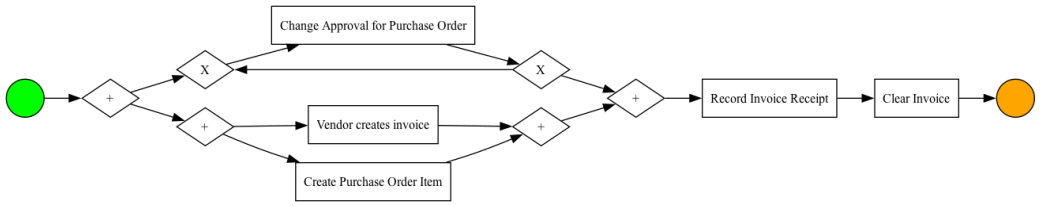

Figure 9: Category 7 BPMN model

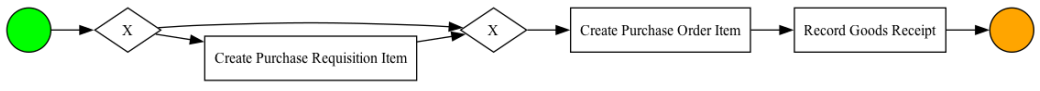

Figure 10: Category 8 BPMN model

4.2.4 Fitness

To understand whether the generated models are a good representation of the entire process, the fitness level for each category was calculated. The models were generated using the dataset filtered by excluding infrequent behaviours, but, obviously, the fitness levels were computed by testing the entire dataset against the models. The results of these analysis are reported in Table 3 where are shown the number of cases that fit completely the models and the number of cases that have a level of fitness greater than 0.8.

cat	tot. cases	80% fitness		100% fitness	
		# cases	ratio	# cases	ratio
1	213	199	93.4%	93	43.7%
2	787	748	95.0%	284	36.1%
3	831	723	87.0%	695	83.6%
4	162306	162302	100.0%	157334	96.9%
5	1487	1417	95.3%	980	65.9%
6	4211	4209	99.9%	3667	87.1%
7	166	164	98.8%	148	89.2%
8	13130	11850	90.3%	11850	90.3%

Table 3: Number and Percentage of cases for value of fitness 1 and 0.8

Considering fitness greater than 0.8, the results are very good because almost for all categories the percentage is above 90%, and for the largest category (cat. 4) this percentage is approximately 100%. When considering true fitness, i.e. fitness equal to 1, the percentages will obviously lower but the results are still good. This because the 3 largest categories still have high percentages, with the largest one that has 96.9% of cases that completely fit the model. The lower percentages belong to cat. 1 and 2 which contain relatively few cases and these two categories are also the two related to

SRM management, which presents very diversified cases, therefore difficult to summarise in two simple models.

The results in Table 3 tell us that the models showed in Figures from 3 to 10 are a good representation of the process identified by the 8 categories explained in the previous section.

4.3 Invoicing Process Throughput

As explicitly requested in the challenge outline, one of the factors studied was the invoicing process and in particular its throughput. The invoicing process concern only three activities: *Record Goods Receipt*, *Record Invoice Receipt* and *Clear Invoice*, but the ways these appear in the event log are not always the same. For this reason, the first step done was to reduce the original event log to a version with only the 3 activities mentioned above.

4.3.1 Invoice iterations

To be able to effectively calculate the throughput of the invoicing process the event log had to be slightly modified. Specifically, a case could present multiple invoice and goods receipts with corresponding *Clear Invoice* events. This mean that there was a need to have a system to associate each *Record Goods Receipt* with its corresponding *Record Invoice Receipt* and then with the corresponding *Clear Invoice*, because, in a case with multiple invoices, there is no guarantee that a second invoice is registered after the first invoice is cleared.
In this regard, it was added to the event log a column called *invoice iteration* which labels with the same integer number the events that belongs to the same invoice. An example of this invoice iteration column is shown in Figure 11.

Activity	Timestamp	Iteration
Record Goods Receipt	10:00:00	1
Record Invoice Receipt	10:54:00	1
Record Invoice Receipt	11:05:00	2
Clear Invoice	11:46:00	1
Record Goods Receipt	11:52:00	2
Record Goods Receipt	12:13:00	3
Record Invoice Receipt	14:34:00	3
Clear Invoice	15:06:00	2
Clear Invoice	17:34:00	3

Figure 11: Invoice iterations example

The example in Figure 11 refers to a 3-way match case, so, each *invoice iteration* is composed by one *Record Goods Receipt*, one *Record Invoice Receipt* and one *Clear Invoice*. For each case, the first of each of the above-mentioned activities will be marked with the number 1, the second to appear will be marked with the number 2 and so on.
It is important to notice that even in the 2-way cases the invoice iterations are added but in those cases there will only be *Record Invoice Receipt* and *Clear Invoice* activities to match.

4.3.2 Throughput Results

Having the dataset formatted as shown in the section above, it was possible to compute the throughput of the invoicing process for all the dataset. The computation was split in three categories:

- **3-way match, Invoice after GR**

- **Invoice-Receipt time** - Time between the *Record Goods Receipt* and the *Record Invoice Receipt* of the same iteration
- **Clearing time** - Time between the *Record Invoice Receipt* and the *Clear Invoice* of the same iteration

- **3-way match, Invoice before GR**
  - **Invoice-Receipt time** - Time between the *Record Goods Receipt* and the *Record Invoice Receipt*, or vice versa, of the same iteration
  - **Clearing time** - Time between the *Record Invoice Receipt* and the *Clear Invoice* of the same iteration

- **2-way match**
  - **Clearing time** - Time between the *Record Invoice Receipt* and the *Clear Invoice* of the same iteration

For the cases with multiple invoices it has been taken the average of the *Invoice-Receipt times* and *Clearing times* computed for each iteration.

The results obtained grouped for *Item Category*, *Item type* and *Document Type* are shown in Figure. 12.

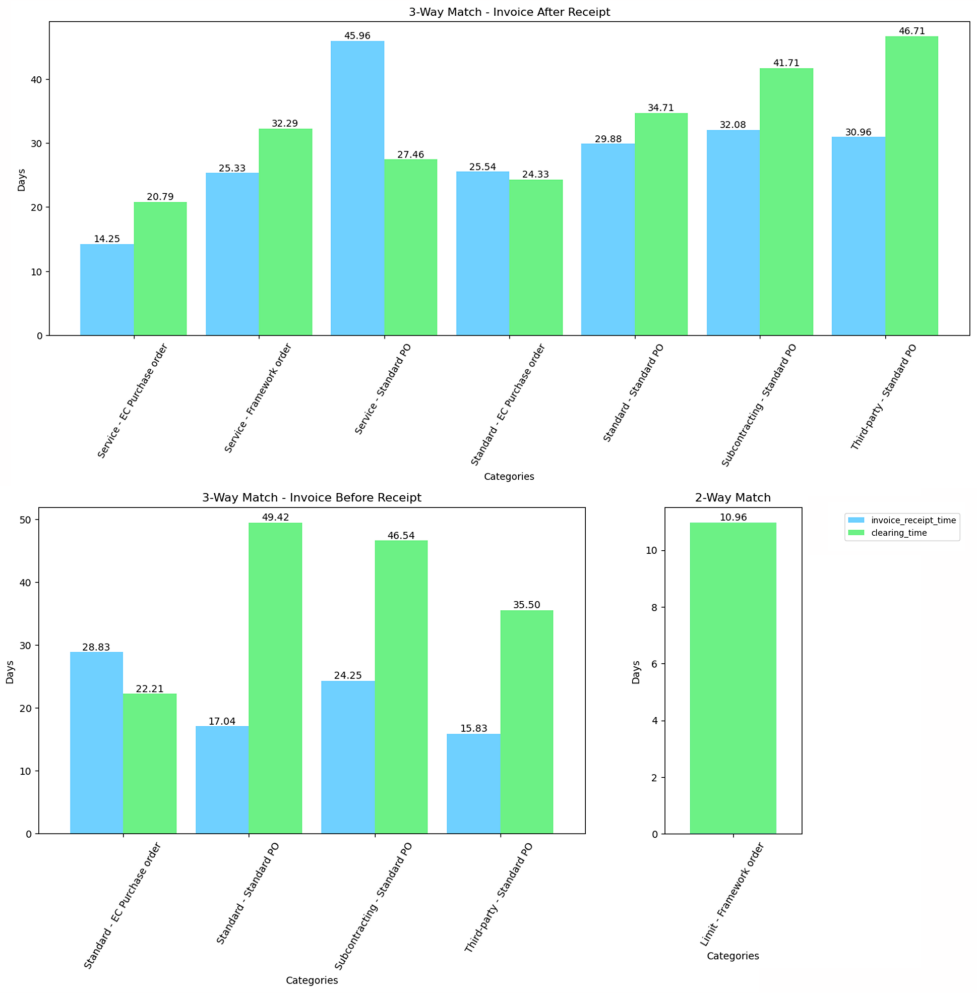

Figure 12: Invoicing throughput results group by Item Category, Item Type and Document Type

4.4 Anomalies Analysis

In this last section, the work focused on analysing and trying to understand the behaviours behind certain anomalies noted in the previous parts. It is

important to specify that, due to the limited information available about the data, abnormal behaviour highlighted in this section may not necessarily be undesirable behaviour. Hence, the term anomalies refers to non-usual behaviour that cannot be judged as conforming or non-conforming.

4.4.1 Remove Payment Block

The first point to be noted is the massive presence of the activity *Remove Payment Block*. This is visible also in the process models in section 4.2.3, where data were filtered by keeping only frequent behaviour before generating the models. Despite this, the activity *Remove Payment Block* appears in Figure 5, 6 and 8.

Having so many activities of this kind in itself might not be a problem, but the name *Remove Payment Block* suggests that the payment block needs to be set before being removed. In fact, among all the different activities of the dataset there is also an activity called *Set Payment Block*.
For this reason, it seems logical to expect each activity *Remove Payment Block* to be preceded by an activity *Set Payment Block*. However this almost never happen.

The activity *Remove Payment Block* occurs a total of 44468 times throughout the all dataset, but the activity *Set Payment Block* only 44 times. Moreover, of this 44 times in which the payment block is set, in only 10 of them the block is then removed. Therefore, a difference this big is very difficult to explain.

4.4.2 Largest Purchase Documents

The next analysis, following the challenge outline that asked for which Purchase Documents stand out from the log, was looking at the number of activities per Purchase Document with special emphasis on the documents that

has the highest number of activities.

First of all, it was computed the average number of activities per Purchase Document throughout the all dataset, resulting in a mean of 18.49 activities and a median of 6.0 activities. Subsequently, for each document was counted how many activity has and then the entire dataset was grouped by the number of activities in order to have for each number of activities the number of documents that has that certain amount of activities. The 10 number of activities that have the highest counter are shown in table 4 and in figure 13 is also visible a histogram showing the results on all the dataset. Notice that for visualisation purposes the y axis is cut at 5000 and each bar of the histogram represent a range of 20 values, so the first bar, the largest one, represent number from 1 to 20.

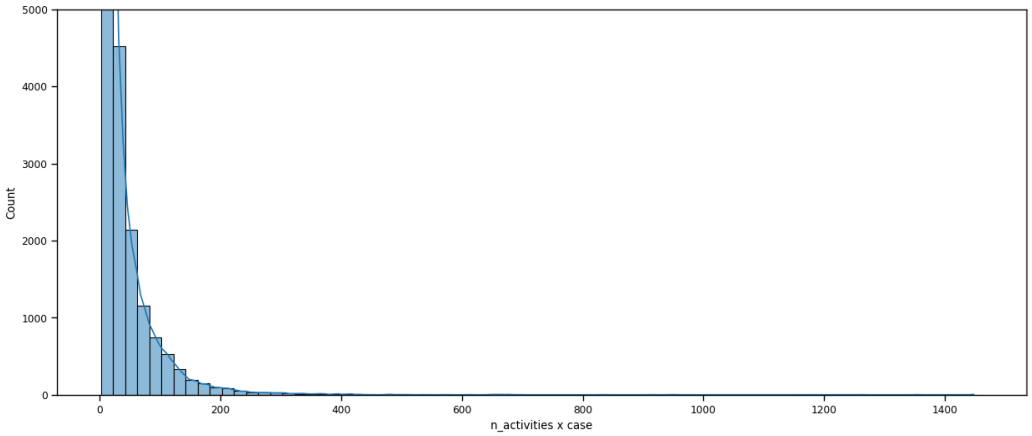

Figure 13: Counts for number of cases with a certain amount of activities per case

As it is visible, the majority of documents have a number of activities between 1 and 50, and higher is the number of activities lower will be the number of documents with that amount of activities.
Analyzing the opposite side, the 20 documents with most activities have

n activities	n documents
6	12167
5	10710
7	4243
10	2949
2	2660
12	1940
3	1632
15	1341
8	1047
11	886

Table 4: 10 highest counts for number of cases with a certain amount of activities per case

more than 500 activities with the largest one having 1448 activities. The only pattern noticeable in these large document cases is that all of them belong to the *3-way match, Invoice before GR* Item Category and *Standard PO* Document Type, which is unfortunately the largest category, so it is hard to understand if this is a particular behaviour or if this is a results due to statistics.

4.4.3 Rework Activities

The last aspect studied for this challenge was the *rework activities*, with special emphasis on which customer produced a lot of rework.

The first thing done was to identify all of activities that are considered *rework activities*. These types of activities are all those that have one of the following words in their name: *change*, *delete* or *cancel*. Using this rule, the

activities identified as *rework activities* are reported in Table 5 together with the number of occurrences for each activity.

activity	occurrences
Change Quantity	13344
Change Price	8614
Change Approval for Purchase Order	2617
Change Delivery Indicator	1294
SRM: Change was Transmitted	1000
Change Storage Location	258
Cancel Invoice Receipt	245
Delete Purchase Order Item	225
Cancel Goods Receipt	183
Cancel Subsequent Invoice	122
SRM: Deleted	73
Change Currency	22
Change Final Invoice Indicator	6
Change payment term	5
Change Rejection Indicator	2

Table 5: Number of occurrences per rework activity

Then, the goal was to identify the users that produced a lot of rework. To do this, the dataset was grouped by users and for each user the number of activities and rework activities have been counted. From this it was possible to compute the percentage of activities that were *rework activities*. In Table 6 are shown the users with at least 50 rework activities and a rework percentage greater than 50%, ordered by rework percentage.

User	activities	rework activities	rework percentage
user_037	1190	1190	100.0%
user_039	1037	1037	100.0%
user_081	276	276	100.0%
user_317	260	260	100.0%
user_603	256	256	100.0%
user_069	246	246	100.0%
user_115	214	214	100.0%
user_307	161	161	100.0%
user_358	53	53	100.0%
user_071	1886	1876	99.5%
user_146	189	186	98.4%
user_169	122	103	84.4%
user_090	61	36	59.0%
user_243	1382	800	57.9%
user_084	4478	2276	50.8%

Table 6: User with at least 50 rework activities and rework percentage greater than 0.5

5 Conclusions

In conclusion, the project focused on three areas of interest for the company that proposed the challenge. These three areas were:

- **Process Discovery**

  It was possible to identify up to 8 models that correctly describe the entire process. As the dataset is very large and varied, a possible future development of this work could be to specialise the model search to a single area of interest, in order to have more circumscribed and precise

results.

- **Invoice Throughput**

  For the invoicing process, it was possible to calculate the time this process takes, even for the most complex cases where several invoices and receipts were present. The results were divided by Item Category, Item Type and Document Type in order to make easier to understand if and where there could be problem in this process.

- **Anomalies Analysis**

  In this part of the work it was possible to discover and highlight some unexpected behaviour, such as users with a lot of reworks or documents with an enormous amount of activities. Unfortunately, there are no enough information from the company to understand whether these behaviours are indeed improper behaviours and were therefore simply reported.