PROCESS ORIENTED DATA SCIENCE

Master in Data Science

Final Project: BPI Challenge 2016: Analyzing customer service event logs for the Dutch Employee Insurance Agency

Group 5 (class 11)

December 20th, 2022

Sebastian Paglia……..**sebastian.paglia@estudiantat.upc.edu**
Agustina Martinez…….**agustina.martinez@estudiantat.upc.edu**
Rasmus Siljander……..**rasmus.siljander@estudiantat.upc.edu**
Mattéo Boursault……..**matteo.boursault@estudiantat.upc.edu**
Ibrahim Braham……..**ibrahim.braham@estudiantat.upc.edu**

INDEX

1. Introduction 3

2. Raw Data 3

3. Preprocessing 4
   3.1. Customer-identified events 4
   3.2. Not Logged-in 4

4. Analysis 5
   4.1. Not-logged in events 5
   4.2. Logged-in events 9
      4.2.1. Event type patterns 9
      4.2.2. Website Activity 12
      4.2.3. User Demographics **13**
      4.2.4. Common topics **13**
      4.2.5. Activities performed by duration 13**4**
      4.2.6. Transition from website to more expensive channels **14**

5. Conclusion 15

6. Sources 17**7**

7. Appendices 17**7**
      A. Preprocessed logged-in event data columns 17**7**

1. Introduction

This project aims to analyze the process log data given in the BPI Challenge of 2016, which inspects event data of the Dutch Employee Insurance Agency (UWV). It is a Dutch government organization responsible for providing services for the Dutch labor market, these services include employee insurance, data solution, and other labor services.
This dataset is part of the Business Processing Intelligence Challenge 2016.

Events were collected into Click Logged In for the customers that are logged in to the website, Click Not Logged In, Werkmap Questions asked by customers, Messages sent by customers and complaint data filed by customers. These were used to build an understanding of customer activity and movement. The following questions are posed and will be answered:

1. Is it possible to obtain patterns of usage of the website for not-logged users?
2. Are there patterns of the activities in the webpage of the logged users?
3. Do exist patterns in the complaints, messages and questions registered through the website?

With this analysis, we aim to contribute to finding optimal solutions capable of meeting customer needs as effectively as possible. We will use information from the interaction of UWV customers through its different channels and the use of the website, in general, using the registration data of arranged events. This analysis focuses on establishing patterns of activity, obtaining user statistics, requirements and perceptions of the procedure as a basis for comparison with the established procedure once it becomes available.

2. Raw Data

This report used all of the five available datasets to answer the aforementioned research questions. All data was available for download as .csv from the challenge website. The datasets were the following:

Dataset	Rows	Variables	Description
Click-data (not logged in)	1.557.352	17	Website activity data for customers not logged in as UWV website users. Contains Session, IP IDs, page name, URL, action information.
Click-data (logged in)	223.051	20	Website activity data for customers logged in as UWV website users. Contains Contains Session, IP IDs, demographic information, page name, URL, action information.
Questions	123.403	17	Werkmap questions by customers
Messages	66.058	8	Werkmap messages sent by the customers
Complaints	289	18	Werkmap complaints sent by the customers

One of the challenges differentiating the datasets is a customer key. While all Werkmap related contact events and the logged in click-data could be identified to a particular customer (and hence had a customer ID), the click-data for the not-logged-in customers only had a session ID to differentiate user events from each other. Hence, these two data profiles need to be analyzed separately, resulting in a "not logged-in" vs. "logged in" data split in analysis. Closer details about how this was executed can be found in the next section.

3. Preprocessing

This section outlines the processing techniques used to build usable event files for analysis.

3.1. Customer-identified events

As mentioned in the previous section, logged in clicks data was aggregated with Werkmap events in order to form a larger, customer ID identifiable dataset. This means merging the Click data with Werkmap questions, messages, and customer complaints. The aim of such a merge is to be able to explore various different types of usage patterns in relation to each other.

The merge was done using the software R. The different datasets were eventually concatenated by adding all together, but before that, it needed to be ensured that all the column information was aligned.

First, this meant initializing a unique customer event-ID, which was done by concatenating an event's CustomerID and contact date (e.g. "*1503890 2015-09-01*"). Concatenating allows for the establishment of a unique event ID for the case where there is more than one event row for a unique customer. The drawback of this is that with this data we are not able to handle situations where more than one contact event happened for the same customer within a particular day.

The second thing is choosing the appropriate columns for each individual dataframes. Columns that were not found in the data were initialized as a N/A-filled vector, and the rest were selected with minor cleanup processing. A closer breakdown of what columns were used for analysis can be found in Appendix A. Finally, event types were given a categorical label ("*click/message/question/message*") based on where the data was collected from.

3.2. Not Logged-in

Regarding the not logged-in file, it was necessary to apply some preprocessing before the analysis. We decided in this case to use python instead of R, just to include different tools into our project. First of all, there were many events without an activity associated, which we decided not to take into account (using the abstract version). Next, we needed to decide which features were useful for our analysis and

which ones were not, so those with all of the entries or above 99% of missing values were deleted ('REF_URL_category', 'page_action_detail', 'tip', 'xps_info', 'page_action_detail_EN', 'tip_EN').

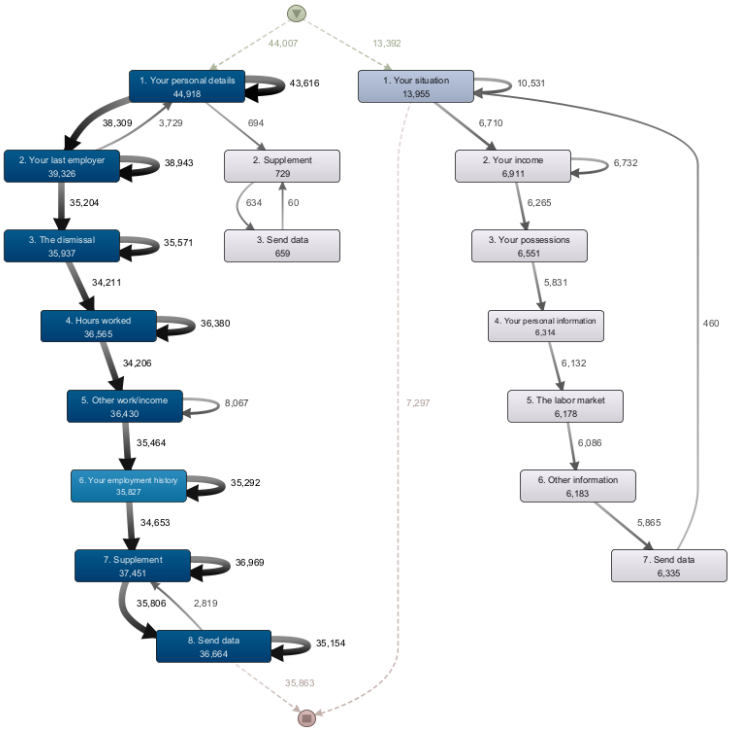
Other features were also removed, because of different reasons: 'VHOST' and 'page_load_error' have only one category so for our analysis, we decided not to keep them; 'URL_FILE' has the same information as 'PAGE_NAME' and we will not include duplicate information. Similarly, 'service_detail' has the same information as 'Activity' so it was also removed.

4. Analysis

4.1. Not-logged in events

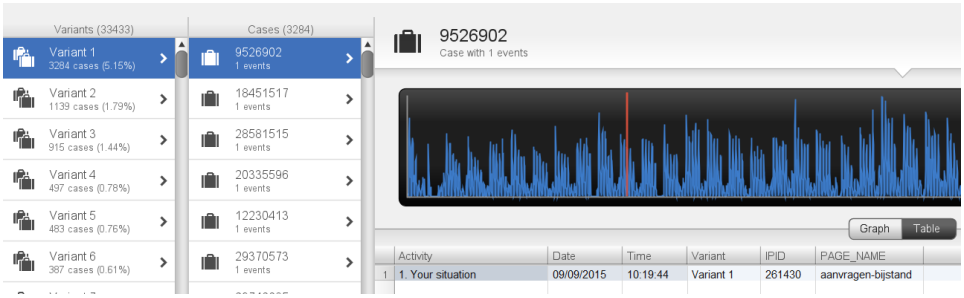
From the point of view of the Not-Logged in dataset (unknown users), we started from an initial dataset using the Abstract version, downloaded from the webpage. We preprocessed the data as mentioned in the previous chapter, removing unnecessary columns (redundant data, repetitive information or too many missing values). Finally, we loaded the preprocessed file into Disco, taking into account that Apromore has a restriction on the amount of events. Once the dataset was loaded, the first step was to check whether we could simplify the data even more, using the tools provided in Disco, so we checked its performance and realized that 99% of the cases lasted less than an hour (compared to the longest 3:06 hs), therefore we filtered those cases that exceeded the threshold.

We reached a dataset with 63.751 cases with 1.523.850 events in total, with a median and mean duration of 11,4 mins and 13,5 mins respectively and a total of 33.433 variants. As we have too many variants, there is no point in analyzing the map setting paths at 100%, so as shown in the map, we decided to set the activities at 100% (17 different activities) and paths at 15%, which will allow us to get a better understanding of the behaviors at first sight.



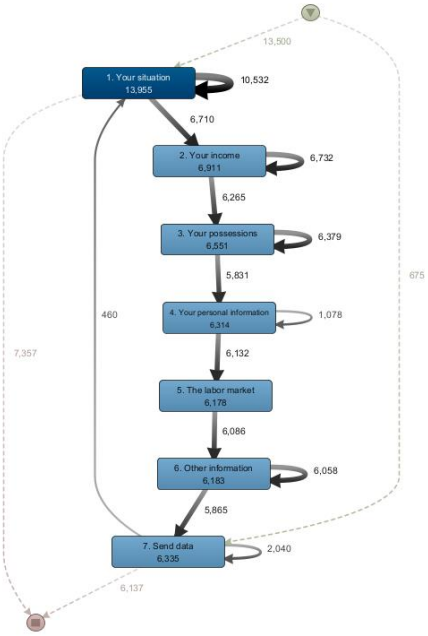
The first insight that we can deduce from the map is that there is a clear common path for most of the cases on the left side of the map, starting in “1. Your personal details”, representing 71% of the cases. A second insight that arises is the rework (“loop”) that all of the activities have in almost all the cases for each activity in the left side of the map, except for “5. Other work/income” where only 22% of the cases go directly through this activity again.

Bearing this information in mind, several filters and explorations of the activities paths were carried out in order to obtain a better understanding of the behavior of the data flow. Analyzing the first variant (3.284 cases, 5,15%) we can see that there is only one activity “1. Your situation”, which probably are all the customers that did not continue the process after the initial activity. However, even though it is the most common variant, as we already saw on the right side of map, 7.279 cases flow through this activity to the case closure, explained by the loop in the activity or also in some cases after sending the data, returning to the first activity.

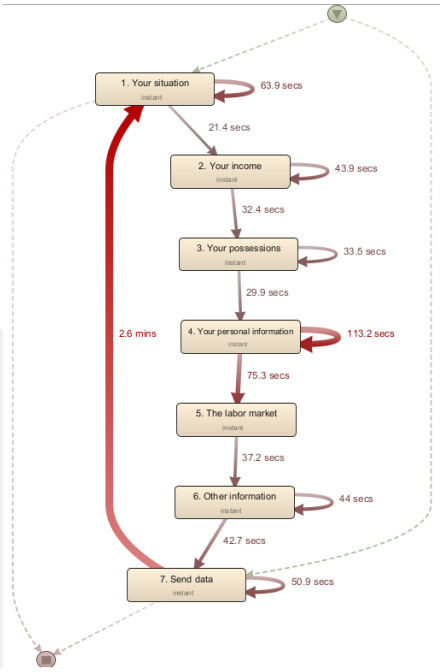


In this dataset, we decided not to apply the removal filter for incomplete cases due to the lack of characteristics within the registry or complementary information that establishes which activities are specifically for case closure. Therefore, we can not consider only “Send data” as the case closure.

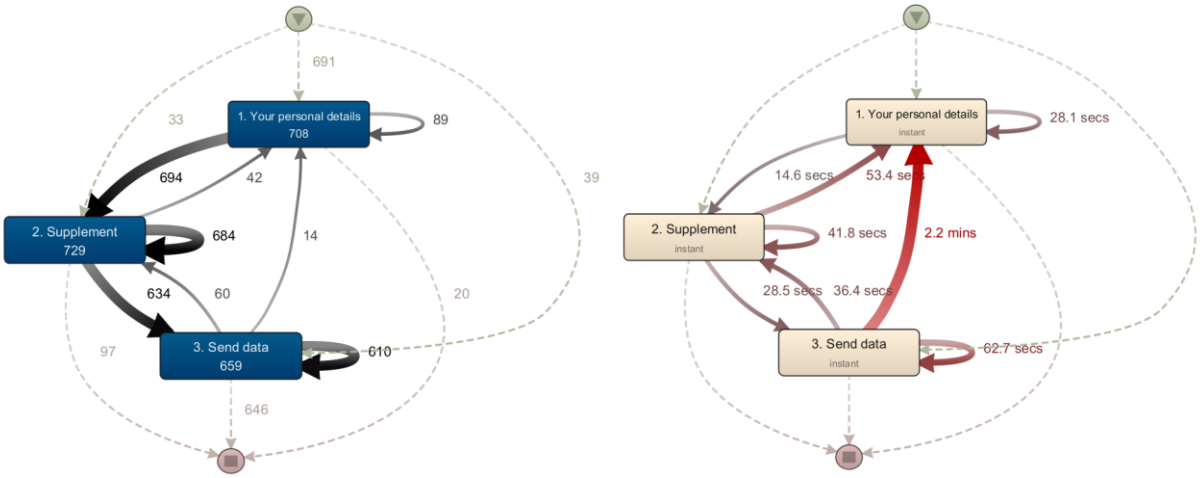
Now we will analyze if there is any difference between the three types of PAGE_NAME attributes, starting from the “aanvragen-bijstand” (“assistance request”) . We can see that it is mainly represented with the right side of the complete process map seen before. As the size of events was reduced by the filter, we allowed to increment up to 30% the paths, always maintaining the activities at 100%. It is important to realize that the median and mean case duration were modified by -5,3 mins (-46%) and -4 mins (-30%) respectively. Also, we can see on the performance map (using mean duration) that the transition between “7. Send data” to “1. Your situation” is the most expensive one (2.6 mins) in terms of time spent.



Events	191,727
Cases	14,979
Activities	7
Median case duration	6.1 mins
Mean case duration	9.5 mins
Start	07/01/2015 08:48:23
End	02/29/2016 00:59:55

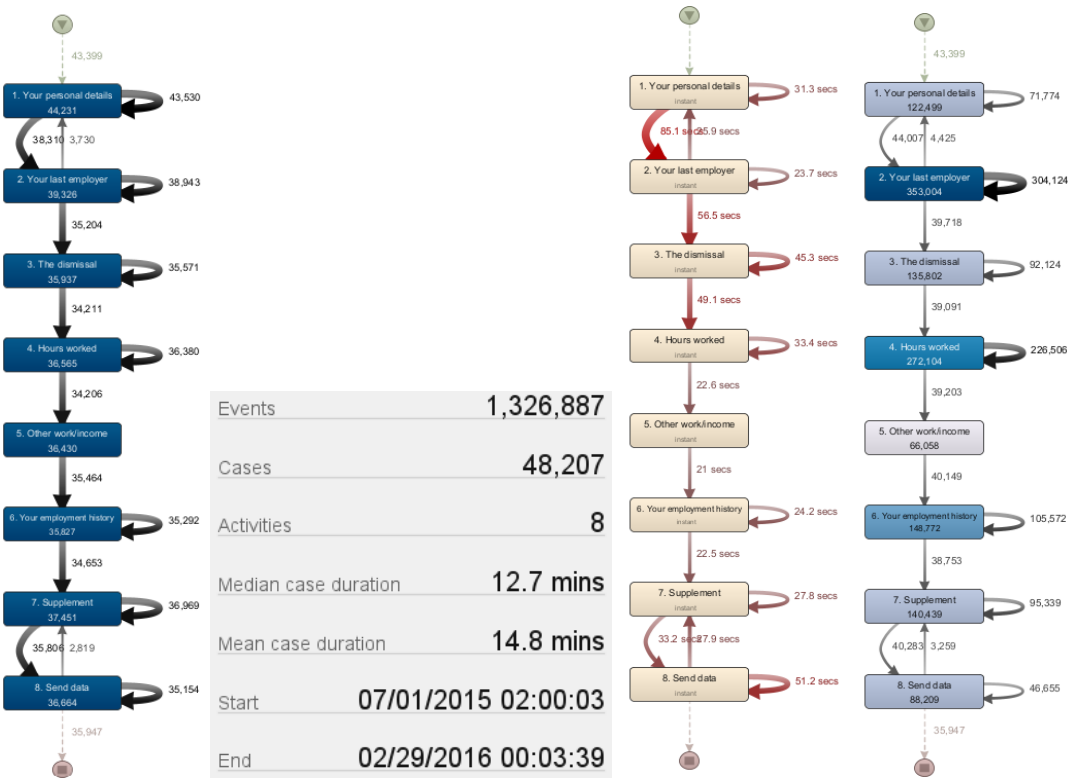


Secondly, we will analyze the PAGE_NAME “aanvragen-tw” (“employment request”). We can see that even appling the 100% on the paths, the graph is smaller, taking into account that it has only 3 activities that correspond to the middle side of the initial map. In this case, the median and mean are -8,8 mins (-77%) and -9,7 mins (-72%) with respect to the complete data. Similarly to the previous page name, the transition between Send data and Personal details / Yours situation is the most expensive one.



Events	5,236
Cases	763
Activities	3
Median case duration	2.6 mins
Mean case duration	3.8 mins
Start	07/01/2015 13:16:15
End	02/28/2016 09:41:59

Finally, analyzing PAGE_NAME “aanvragen-ww” (“unemployment request”) we can realize that it corresponds to the left side of the original map. Its median and mean is +1,3 mins (+11%) and +1,3 mins (+10%) with respect to the dataset without this filter. It is comprehensible that in this page category, the median and mean throughput time are closer to the complete version because most of the cases are within this category, and also taking into account that the rest of the categories throughput time are fewer, this one should increase the values. This higher value of case duration time may be also explained as it has more activities to go through or also because there is a “rework” meaning that one customer goes through the same activities more than once. That is why we will analyze in this case not only the case frequency but also the absolute frequency, showing that the most incurred activity is “2. Your last employer” where we can also see a notorious loop in this activity.

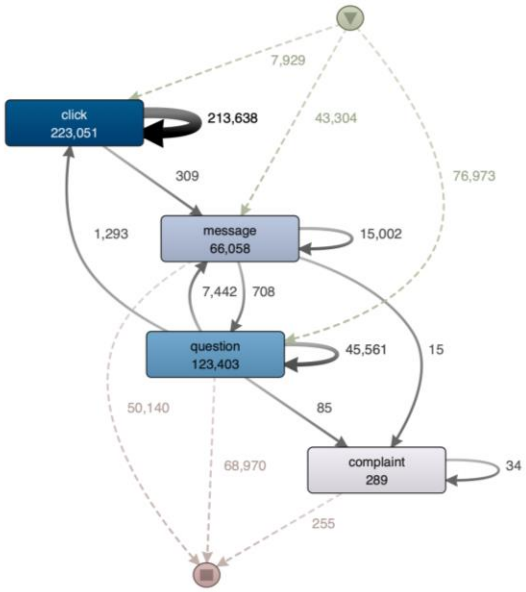


4.2. Logged-in events

4.2.1. Event type patterns

The primary visualization of the logged-in data can be seen in the process map (75% of the paths visualized). The most notable observation about the click data is that the most common event following a website click was another click. This suggests that users are more likely to return to the website (or continue using the site on a different webpage), indicating that there might be underlying usage patterns within a particular event type. Such observations are analyzed in more detail in Section 4.2.2.

Of the 412.801 events outlined in the data, we can see that 54% (223.051/412.801) of those were a click. Further, it is worth noting that no traces exist where a customer has moved directly to file a complaint after a click event. This suggests that the company website is working as intended and that no complaints arise from website use itself. Instead, complaints seem to have risen from questions and message actions, indicating that they result from an already increased need for more expensive means of help.

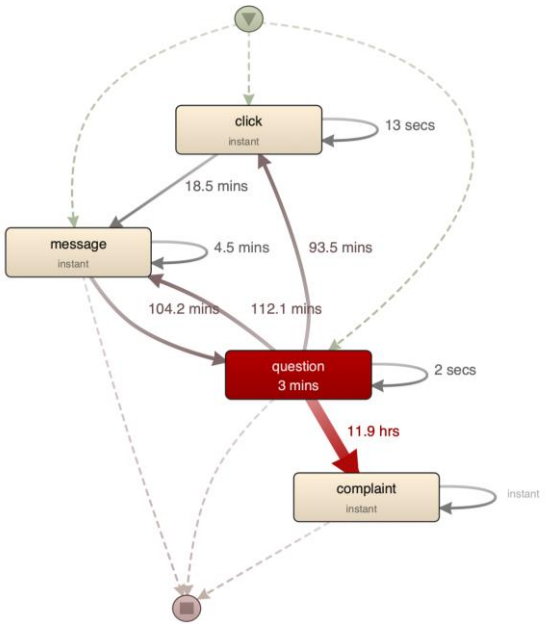


The process map above also shows evidence for the Werkmap questions working both efficiently (as intended) and inefficiently. First, we notice that for a significant number of events (68.970), the process ends with a question. That is, for a number of customers there is no need to return to a message or a website activity after posting a question. However, we see that there are also 45.561 question events repeated. This might suggest that customers are not satisfied with a single question answer, but need to return to the questions platform for receiving answers. This is poor use of resources, since it is previously known in the challenge description that the Werkmap platforms are expensive and time-consuming to maintain. It might therefore be useful for UWV to focus on developing their platform in a way that repeating question events are minimized, especially considering that *questions* is the most common starting event for all cases (76.973/128.360).

There is, however, an exception to this observation, we do not know exactly the topics of the questions, and it might be possible that repeating events are simply due to customers having more than one question from different areas. This would register as separate events and would hence not be evidence of poor question management, but instead an artifact of dimension reduction in aggregating multiple question topics to one category. Such analysis also falls outside of the scope of this report.

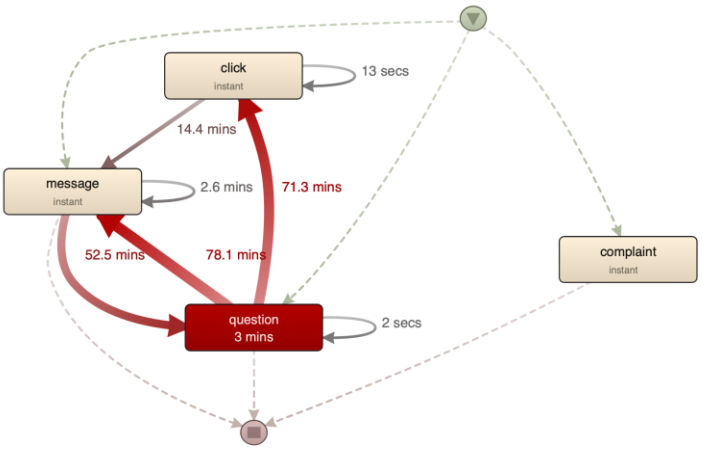
Overall, no other extremely relevant finding can be presented from this model. We already know that website activity does not seem to lead to complaints or messages (in large numbers). In addition, we can notice that there are some paths that lead from questions to messages and back. This might be expected, however, since customers with questions are more likely to contact UWV via message as well. We will next move on to analyzing whether website activity reveals any additional insights.

Performance and Case times

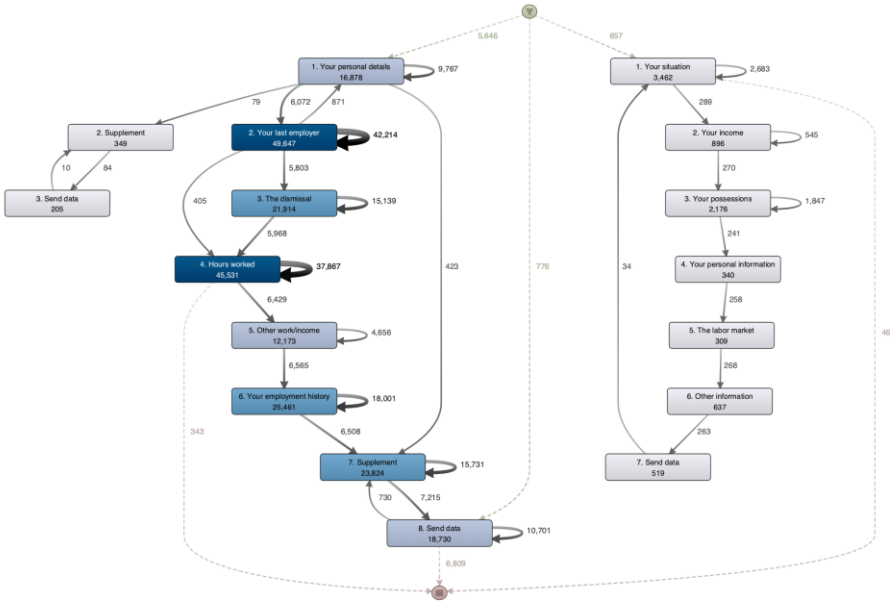


Median case times (unfiltered data) are visualized in the process map above. We can see that there is a significant bottleneck in the transition between "*question*" and "*complaint*", indicating that it might be challenging for customers to transition from a question stage to complaints. Similarly, there is an increased transition time from the questions to other events as well (93.5min for *question* -> *click* and 112.1min for *question* -> *message*). This shows that UWV might struggle with efficiently responding to customer questions and this might lead to expenses in customer service, a factor that UWV could focus on the process of becoming more efficient.

However, it is worth noting that the process map above was generated with unfiltered case data. Since it was found that case duration was highly skewed and that affected case time modeling. By filtering out the top 5% longest cases (>2h 36min), we receive the case duration map shown below (70% paths). The first thing we notice is that the large *question* -> *complaint* arch is removed, suggesting that this artifact might have actually just been due to outliers in the case data. What is left is the patterns of high transition times to and from Werkamp questions. This supports the previously mentioned argument that *Questions* might be a time-consuming event type that UWV could focus on to make their processes more efficient.



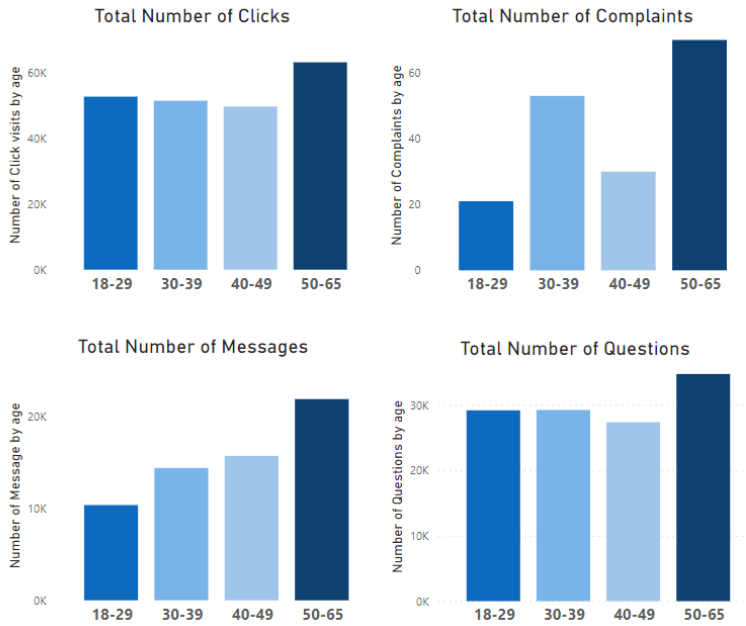
4.2.2. Website Activity



The clicks visualization shows an expected result, something very similar to what was discovered with the not-logged in data. Events seem to follow a logical, form-like usage pattern, where people fill in relevant Werkmap information. In other words, the usage pattern resembles that of filling in a traditional application form, with different sections for filling in relevant information and then sending it. All paths have a "*send data*" as their final event. What makes the process more efficient is the lack of back-loops amongst different parts of the process. Apart from some exceptions (*Supplement <-> Send Data*), there are no paths where a customer has gone backwards in their website activity (that is, assuming that the numbers in the activities imply chronological order), where they have gone back to a site they have already visited during a particular session. From that perspective, there is no need to streamline the website event process. Overall, the results look similar to what was presented in Section 4.1

4.2.3. User Demographics

To see in detail the behavior of the data, it was found that 99% of the cases lasted about eight hours, so the cases that exceeded the threshold were filtered out. In addition, from the customer's point of view, we have observed that according to the age range, those aged between 50 and 65 tend to be those who use the tools the most, both in raw figures and on average. In particular, the average use of the tools by age tends to differ more in the case of the web application (clicks), while the others (messages, questions, complaints) are more similar in all age groups.



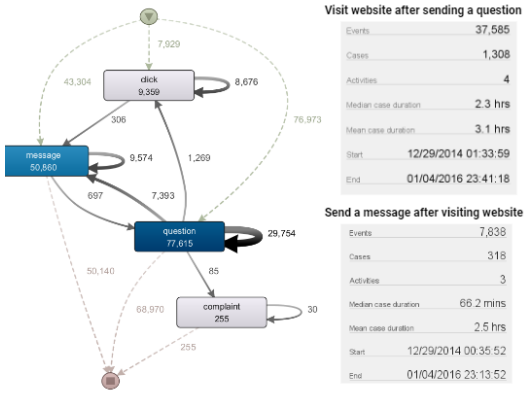
4.2.4. Common topics

Additionally, in order to improve the quality of the customer experience, it is interesting to analyze which are the most common topics of questions and complaints, as these are very expensive communication channels that should ideally be reduced. According to the dataset, the frequently asked questions are related to income declaration (22%) and unemployment benefits (13,3%) , whereas the most common complaints come from inconsistent information(9.5%) and the treatment received (8.9%).

Question	Subtheme	Count
When is transferred my unemployment benefits?	Payment	13078
General: When should I send the form Revenue Problem?	Income from declaration	6027
What is the status of my application WW?	Status	4885
I want to report a change	Report changes	3543
General: Where can I find the form Income Problem?	Income fron declaration	3358

Complaints	Subtheme	Count
Information: incorrect/inconsistent	Information-communication to the customer	17
No respect - Not taken seriously	Treatment (attitude-behavior)	16
Payment over a certain period is missing	Income from declaration	15
Income from ww unreachable	Availability- Accesibility	14
Information: no-insufficient	Information-communication to the customer	11

4.2.5. Activities performed by duration



The process mining results also indicate that around 60% of traces include a question and about 40% of them include a message.

As far as duration is concerned, some insights can be extracted:

1. The media time for visiting the website after asking a question is around 93 minutes, while the complete case lasts 2.3 hours.
2. The median time for sending a message after visiting the website is around 18 minutes, while the complete case lasts 66 minutes.

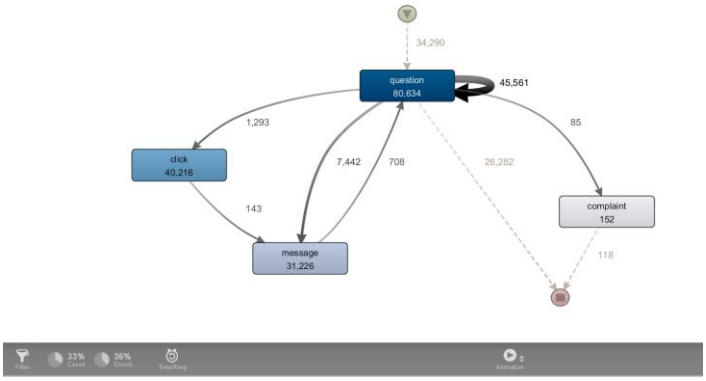
From an operational perspective, this is an undesirable situation because it means that the web application alone is not enough for customers to solve their problems, forcing them to use more expensive channels of communication and thus overloading the contact center.

4.2.6. Transitions to from website to more expensive channels

It is important to have an idea about the users that ask questions through the call center or send messages through the werkmap platform or fill complaints when they get problems because actually these acts are expensive for the company and the goal of this study is to minimize all the costs. So we should know all the information about these users and understand the customer behavior when performing this transition to support the decision making process.

To perform this analysis, the process map was filtered by using “Follower” filter and to make “complaint” , “message” and “question as reference event values and “complaint”, “message”, “question” and “click” as follower event values and we selected “eventually followed” because we only search for the existing of theses events. The goal of this filter was to drop the transactions that contain only “click” events.

The process map generated :



This map shows only 33% of the cases and only 36% of the events and it is important, because it means that the rest of the customers only use the website without any claim and so normally they didn't get problems.

The process maps show that questions(call-center) and messages are the most expensive used channels by the customers. It also shows that after questions there are only 9,2% of the users who send messages after asking a question and only 0,1% of the users do complain after asking a question, which means that the majority got their problems solved after only asking questions. Besides, 50% of users re-ask questions.

5. Conclusion

The dataset for not-logged-in events was preprocessed and filtered to remove cases that lasted more than an hour, resulting in 63,751 cases with 1,523,850 events in total and 33,433 variants. The data was then analyzed to understand common behaviors, and the PAGE_NAME attribute was found to have an impact on the duration and flow of cases. The transition between "Send data" and "Your situation" was the most expensive in terms of time spent for all three PAGE_NAME attributes. The removal filter for incomplete cases could not be applied due to a lack of a characteristic within the registry indicating case closure.

Concerning the event type patterns, the majority of logged-in events in the dataset began with a click and there were no instances of customers moving to file a complaint after a click event. There were also instances of questions being repeated, which may suggest that customers are not satisfied with a single answer or that they have multiple questions in different areas. Besides, website activity did not lead to many complaints or messages, but some paths led from questions to messages and back. Overall, there is limited information that can be gleaned from this model.

Regarding website activity, customers tend to follow a logical usage pattern when using the website, with few exceptions. The case duration data for website clicks is skewed, with most cases lasting a shorter time but some lasting longer due to self-loops in the process. It may be valuable to improve the website design to streamline the user experience and reduce the number of expensive questions. In addition, customers aged 50-65 tend to use the

website the most, with the average use of the website differing more among age groups compared to other channels such as messages, questions, and complaints.

It is important to have an idea on the most common topics of questions and complaints, and in majority they are related to income declaration and unemployment benefits, and inconsistent information and treatment received, respectively. Analyzing these common topics can help improve the quality of the customer experience by reducing the use of expensive communication channels.

As the study aims to minimize costs by understanding customer behavior when transitioning to more expensive channels for support, we looked-up for this point and we generated a process map that used to study the transition to more expensive channels showing that most customers do not have problems and only a small percentage escalate their issues to more expensive channels. Additionally, the majority of customers who ask questions have their issues resolved without needing to escalate to more expensive channels.

Finally, this project shows the potential of business process mining within companies and organizations. Where the most difficult component was cleaning and transforming the data to make it suitable for analysis, which demonstrates the importance of having strong measures to improve data quality in business systems.

6. Sources

BPI Challenge 2016: Processing Intelligence Challenge 2016.

7. Appendices

A. Preprocessed logged-in event data columns

Column name	Description
CustomerID_Date	Unique date-based customer event
AgeCategory	Customer age category. Categories are split into four bins
Gender	Binary “V/M” gender value
Office_U	Benefits Office handling the customer
Office_W	Employment Service Office handling the customer
SessionID	Unique session identifier
Datetime_start	Event start datetime
Datetime_end	Event end datetime
Service_detail	Detail of the provided service
xps_info	Extra information about the event.
VHost	Domain
Page_name	Page/event element name
Topic/Subtheme	Main subject of inquiry, complaint or question
Type	Event type

Contributions

Agustina Martínez

- Preprocessing:
  - Contribution to Pre-processing, identifying the variables/columns to select so that they can be used for an analysis that improves decision-making.
- Disco:
  - Analyze performance in Logged-in events, determine the main patterns that lead the user to generate a complaint or send messages. Additionally patterns regarding demographic characteristics.
- Report:
  - Section 1. Introduction
  - Parts of Section 5. Conclusion
  - Analysis: Section 4.2.3, 4.2.4, 4.2.5

Sebastián Paglia

- Preprocessing:
  - Using python, cleaning of Not logged-in dataset, feature selection and deciding which columns we would use as CaseID, Activities, and attributes.
- Disco:
  - Upload the preprocessed dataset, find insights, analyze performance, case flows and time costs among the activities, differentiating between the different page_name attributes.
- Report:
  - Preprocessing: section 3.2
  - Analysis: section 4.1

Rasmus Siljander

- Preprocessing:
  - Investigation of raw data contents/columns, preprocessing design planning (what columns to choose, what they mean, how can we unify the logged-in data)
  - Fine-tuning (adding some columns/extra column cleanup) of the R preprocessing script created by Mattéo and Ibrahim
- Report:
  - Document outline structure
  - Parts of Section 1 (Introduction)
  - Section 2 (Raw data)
  - Sections 4.2.1, 4.2.2 (Event type patterns, Website activity)

- Disco:
    - Importing the preprocessed logged-in data (ensuring correct activity, timestamp types) and uploading that for group use.
- **Other aspects** (e.g. project management and scheduling) was maintained by other group members.

Ibrahim Braham

- Disco :
    - Section 4.2.5 (transition from website to more expensive channels) using the pre-processed data and a filter
- Report :
    - Section 4.2.5
    - Conclusion of all sections
- Preprocessing :
    - Analyzing the four datasets that must be joined and trying to find the best solution to perform the join
    - R script to do the preprocessing

Mattéo Boursault

- Disco :
    - Section 4.2.6 (transition from website to more expensive channels) using the pre-processed data and a filter
- Report :
    - Section 4.2.6
- Preprocessing :
    - Analyzing the four datasets that must be joined and trying to find the best solution to perform the join
    - Features creation / combination / selection → unification of databases
    - R script to do the preprocessing