RWTH Aachen University

Business Process Intelligence

BPI 2021 ASSIGNMENT - PART 1 REPORT

Karthika Vijay (417307)
Manjari Chaudhri (416114)
Shilpa Chatterjee (404788)

June 4, 2021

Contents

1 Preprocessing the data

1. The column **experience** has been replaced values >20 with 23 and <1 with 0 and the attribute type has been changed to numerical. Please refer to Figure 1

2. The column **lastnewjob** has been replaced values >4 with 5 and never with 7 and the attribute type has been changed to numerical

enrollee_id	city	city_development_index	gender	relevent_experience	enrolled_university	education_level	major_discipline	experience	company_	company_type	last_new_job	training_h	another_job
29725	city_40	0.776	Male	No relevent experien	no_enrollment	Graduate	STEM	15	50-99	Pvt Ltd	5	47	0
666	city_162	0.767	Male	Has relevent experien	no_enrollment	Masters	STEM	23	50-99	Funded St	4	8	0
402	city_46	0.762	Male	Has relevent experien	no_enrollment	Graduate	STEM	13	<10	Pvt Ltd	5	18	1
27107	city_103	0.92	Male	Has relevent experien	no_enrollment	Graduate	STEM	7	50-99	Pvt Ltd	1	46	1
23853	city_103	0.92	Male	Has relevent experien	no_enrollment	Graduate	STEM	5	5000-9999	Pvt Ltd	1	108	0
25619	city_61	0.913	Male	Has relevent experien	no_enrollment	Graduate	STEM	21	1000-4999	Pvt Ltd	3	23	0
6588	city_114	0.926	Male	Has relevent experien	no_enrollment	Graduate	STEM	16	Oct-49	Pvt Ltd	5	18	0
31972	city_159	0.843	Male	Has relevent experien	no_enrollment	Masters	STEM	11	100-500	Pvt Ltd	1	68	0
19061	city_114	0.926	Male	Has relevent experien	no_enrollment	Masters	STEM	11	100-500	Pvt Ltd	2	50	0

Figure 1: Snapshot of the first 10 rows of the dataset

2 Clustering

Student ID used as seed in submission: 416114. Please refer to Figure 2 for a snapshot of the designed process of the uploaded setting.

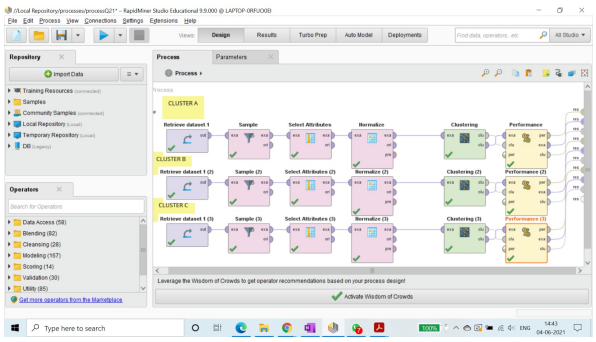

Figure 2: Snapshot of the designed process of the uploaded setting

1. Centroids of clustering

   (a) The centroids of cluster A,B,C are:

      i. The centroid of Cluster A is shown in Figure 8

Attribute	cluster_0	cluster_1	cluster_2
city_development_index	0.506	-1.672	0.262
experience	0.222	-0.695	0.049
last_new_job	0.074	-0.210	-0.018
training_hours	-0.343	-0.167	2.016
another_job	0.084	0.450	0.106

Figure 3: Centroid of Cluster A

      ii. The centroid of Cluster B is shown in Figure 4

Attribute	cluster_0	cluster_1	cluster_2	cluster_3	cluster_4
city_development_...	0.483	-1.785	0.061	0.172	0.490
experience	1.373	-0.741	-0.230	-0.010	-0.511
last_new_job	0.484	-0.477	1.430	-0.166	-0.680
training_hours	-0.226	-0.144	-0.233	2.396	-0.291
another_job	0.072	0.495	0.160	0.122	0.083

Figure 4: Centroid of Cluster B

iii. The centroid of Cluster c is shown in Figure 5

Attribute	cluster_0	cluster_1	cluster_2	cluster_3	cluster_4	cluster_5	cluster_6
city_develop...	-1.662	0.179	0.512	0.473	-1.761	0.500	0.492
experience	-0.267	-0.073	-0.246	1.023	-0.795	1.377	-0.719
last_new_job	1.508	-0.219	1.263	-0.600	-0.678	1.241	-0.676
training_hours	-0.110	2.474	-0.277	-0.303	-0.149	-0.104	-0.275
another_job	0.416	0.121	0.112	0.073	0.484	0.067	0.087

Figure 5: Centroid of Cluster C

(b) The cluster which is the most coherent one is the cluster where the distance between the cluster instances and the centroid of the cluster is the least i.e the the cluster consists of the most homogeneous set of examples.

i. Clustering A: cluster 0 is the most coherent one.

PerformanceVector

```
PerformanceVector:
Avg. within centroid distance: 2.572
Avg. within centroid distance_cluster_0: 2.410
Avg. within centroid distance_cluster_1: 2.477
Avg. within centroid distance_cluster_2: 3.553
Davies Bouldin: 1.325
```

Figure 6: cluster 0 is the most coherent one

ii. Clustering B: cluster 4 is the most coherent one

PerformanceVector

```
PerformanceVector:
Avg. within centroid distance: 1.632
Avg. within centroid distance_cluster_0: 1.619
Avg. within centroid distance_cluster_1: 1.755
Avg. within centroid distance_cluster_2: 2.041
Avg. within centroid distance_cluster_3: 3.301
Avg. within centroid distance_cluster_4: 0.964
Davies Bouldin: 1.120
```

Figure 7: Clustering B: cluster 4 is the most coherent one

iii. Clustering C: cluster 6 is the most coherent one.

PerformanceVector

```
PerformanceVector:
Avg. within centroid distance: 1.248
Avg. within centroid distance_cluster_0: 2.043
Avg. within centroid distance_cluster_1: 3.194
Avg. within centroid distance_cluster_2: 1.109
Avg. within centroid distance_cluster_3: 0.974
Avg. within centroid distance_cluster_4: 1.431
Avg. within centroid distance_cluster_5: 0.925
Avg. within centroid distance_cluster_6: 0.809
Davies Bouldin: 1.054
```

Figure 8: Clustering C: cluster 6 is the most coherent one

(c) Clustering C seems to be the best clustering with a low davies bouldin index 1.054 (the ratio of the within cluster scatter, to the between cluster separation) and low average within clustering distance 1.248 compared to other clusterings.

2. The smallest cluster is cluster 3: 762 items and the largest cluster is cluster 4: 2908 items. The average within centroid distance for cluster 3 is 3.301. This implies that the difference between instances in cluster 3 is relatively high despite being the smallest cluster.The average within centroid distance for cluster 4 is 0.964 indicating small difference within the instances in the biggest cluster.

3. The feature **another_job** can be considered redundant when determining the clusters as the values for the centroids for this feature is almost the same in all 5 clusters.

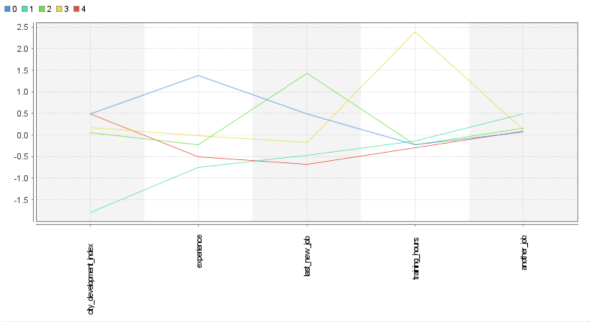

Figure 9: Means of clusters (centroids) using a line plot

4. Before applying the clustering algorithm, we normalise the required attributes - **last_new_job**, **experience**, **city_development_index** and **training_hours**. Process design submitted as processQ24.

5. For Subset 1: The centroid of the clusters vary significantly in terms of centroid attributes. Cluster 0,1, 2, 4 have a similar centroid values but centroid for cluster 3 has a negative value for **city_development_index**. All the cluster centroids differ for attributes: **experience** and **last_new_job**. For **training_ hours**: cluster 0, 1, 2, 3 have similar centroids. It can also be noticed that cluster 0 and cluster 2 have similar centroids for almost all attributes except **experience**. For Subset 2: Cluster 0, 1,

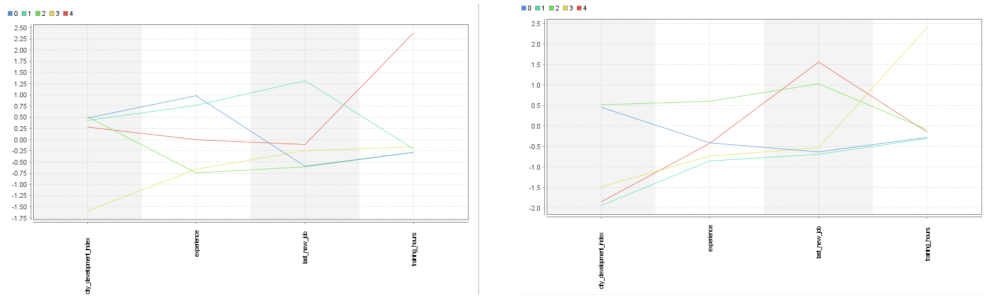

Figure 10: Subset 1 and Subset 2

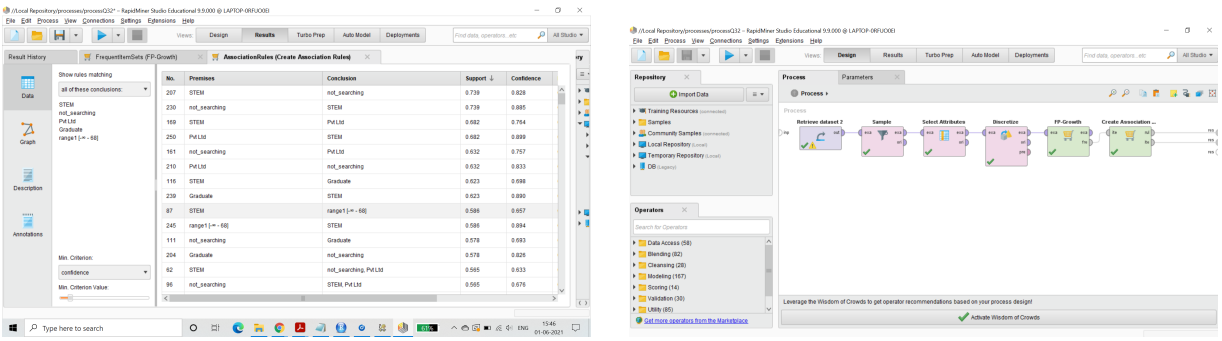

Figure 11: Rules with their support and confidence and Design Tab of processQ32

2, 4 have identical centroids for attribute **training_hours**, cluster 0,1,3 for attribute **last_new_job**. All the cluster centroids differ for attribute **experience** except cluster 0 and 4. Cluster 1, 3, 4 follow similar patterns within but differ from cluster 0, 2 which also have approximately the same centroids for **city_development_index**.

3 Association Rule

1. The values with 1 has been replaced with **searching** and 0 with **not_searching** in column **another_job** and attribute type has been changes to categorical.
2. processQ32 has been uploaded.
3. Three interesting rules that have support values larger than 0.5 are as follows:
   (a) **Graduate** → **not_searching** : **0.577** Most of those with education level as GRADUATE are not looking for jobs.
   (b) **Graduate** → **STEM** : **0.623** Most of those with education level as GRADUATE have STEM as their major discipline
   (c) **STEM** → **not_searching** : **0.739** Most of those with STEM as their major discipline are not searching for jobs.

4 Process Mining

1. File imported as sampled_log_Dataset2.csv. Student ID used for sampling: 416114

2. (a) There are 8 activities in the log

   (b) Most frequent activity - Validation ; Least frequent activity "Personal Loan Collection"

3. (a) Fraud suspicion check takes on average the most time i.e 3 days and 1 hour.

   (b) Performance

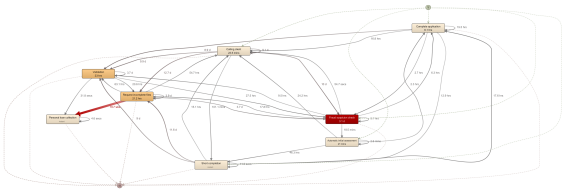

Figure 12: Question 4_3_B_Performance

4. (a) **Short completion** did not appear before 30-06-2016. We filtered out the traces which happened before 30-06-2016 and noticed that one activity was missing from the list of activities.

   (b) Model of all cases containing **Short completion**

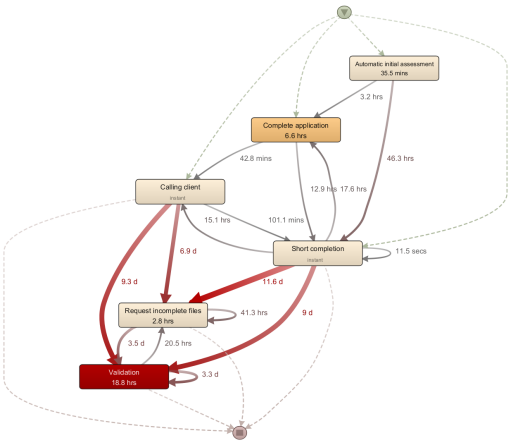

Figure 13: Model of all cases containing **Short completion**

   (c) 33 variants include this activity.

5. (a) User-138 takes on average the longest time i.e. 2 days and 22 hours.

   (b) Process model for resource :User-138

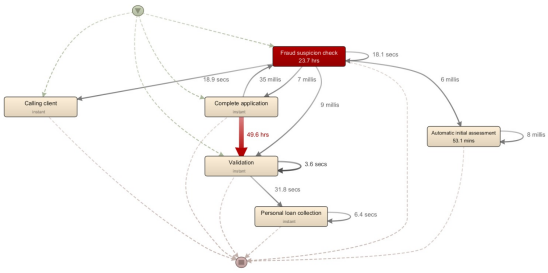

Figure 14: Process model for resource :User-138

(c) This resource is primarily performing Fraud suspicion check. Total time spent by User-138 on it is 22.3 months.

6. (a) Variant 1 is most frequent with 7535 cases.

   (b) Process model of Variant 1 on frequency and performance.

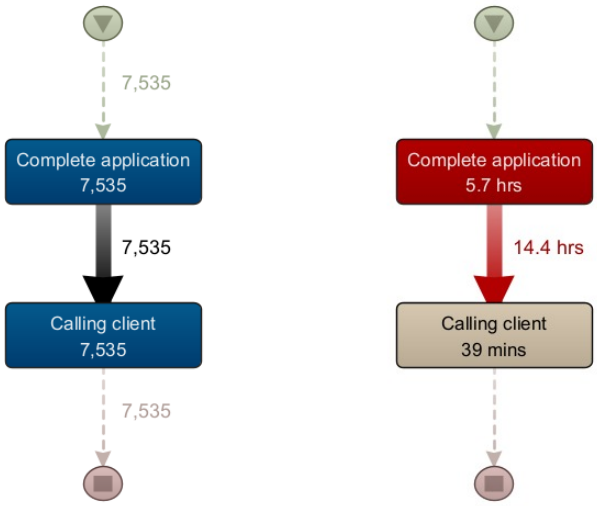

Figure 15: Process model of Variant 1 with frequency and performance

   (c) The longest case of this variant takes 29 days and 5 hours taken by Application_715212713.

7. (a) The most frequent reason for customers to apply for a loan is **Car**.

   (b) 9297 cases follow this objective. Out of these, 861 apply for limit raise only.

   (c) Process model for all cases that follow loan objective as **Car** with a limit raise. Process model for all users that apply for a Car loan without the added filter of limit raise submitted in zipped file as Figure 16.

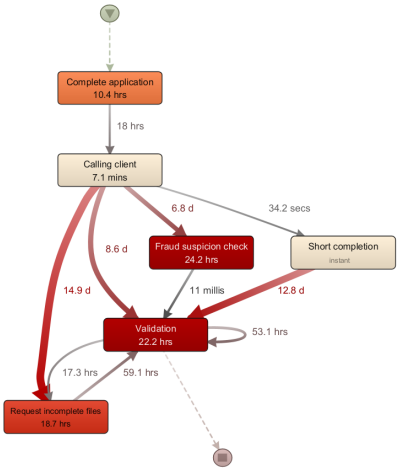

Figure 16: All cases that apply for a Car loan

8. (a) From our analysis we can see that Fraud suspicion check, even though takes the maximum mean duration out of all activities has a very low frequency specially for customers who apply for a car loan. Hence, trying to reduce the throughput time of this event will reduce the average time of the bank but will not affect most customers. A better solution would be to look at activities that take a longer time to complete and are also more frequently done. For example, requesting incomplete files takes on average 21 hours and 7 mins and validation takes 20 hours and 9 mins on average. Both these activities are also frequent in the dataset(customers who apply for a car loan). Furthermore, we can see from the process model that in some events, it is taking a long time to go from one activity to the next. i.e there are long periods of time where no activity is being performed(e.g. Between Short Completion and Request Incomplete files there is a mean time of 13.9 days). To conclude, our two recommendations to the company to reduce the throughput time are -

   i. Reduce the time taken to perform Validation and Requesting Incomplete Files. The customers can perhaps be given a checklist of required documents during the initial Meetings (if this is not already done) to ensure that all documentation is complete without having to repeatedly ask for it.
   ii. Better resource allocation or hiring of new employees to reduce the idle time spent between two activities.

   (b) For this assignment we have worked on the unsampled event log in PROM to convert the CSV to XES and then visualized as dotted chart. In Figure 17 we have time stamp and activity on X and Y-axis respectively with color attribute as Different days of week.This shows that there are certain timespans where no applications are being processed because:

   i. The company does not work on Sundays.
   ii. No applications are being processed after office hours i.e between 16:00(current day) and 8:00(next day).

iii. From Figure 18, the dotted chart has been sorted on "Resource on first event" , we can conclude that more resources have been hired after March 2016.

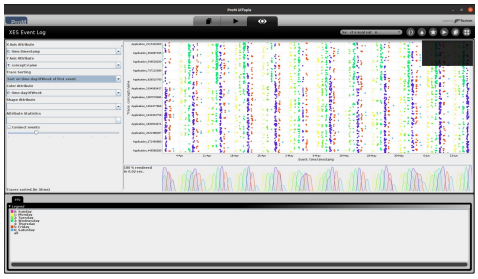

Figure 17: Dotted chart visualization of the event log showing Days of week in different colours.

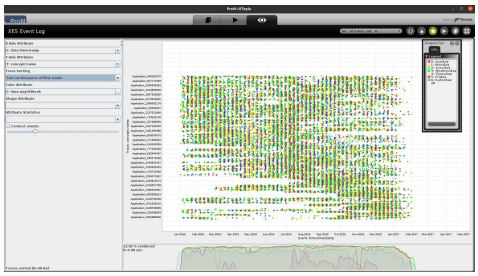

Figure 18: Dotted chart visualization of the event log sorted on "Resources on first event"