First Part of the Assignment for BPI 2021

Introduction

The first part of the assignment guides you through the analysis of two datasets using the techniques and tools provided in the course. It tests the understanding of the material in the lectures 1 through 6. It is necessary to follow the assignment in the given order since the results of specific questions might depend on answers to previous questions. You have to use a student ID of one of the group members as seed to sample the data in the assignments. We do that to make sure that no two groups get exactly the same results. You need to specify the student ID you used as seed so that we can check your solution.

Note: The submission of the assignment part 1 should be done using two files:

- A PDF file (at most 10 pages) including your answer to the questions. Please name the file *BPI_assignment_part_1_[student_id_1]_[student_id_2]_[student_id_3].pdf*
- A ZIP file including all the other requested files (e.g., workflow file from RapidMiner), datasets, and intermediate results. If needed, you can include more (not requested) files and pictures to clarify your answers. Please name the file *BPI_assignment_part_1_[student_id_1]_[student_id_2]_[student_id_3].zip*

Datasets

The *FirstAssignmentBPI-DataSet1.csv* dataset includes data about applicants for working at a company which is active in the field of Big Data and Data Science. In the following, you can see the description of the different features in this dataset:

- enrollee_id: Unique ID of the candidate
- city: City code
- city_development_index: Development index of city (scaled)
- gender: Gender of the candidate
- relevent_experience: Relevant experience of candidate
- enrolled_university: Type of University course enrolled, if any
- education_level: Education level of candidate
- major_discipline: Education major discipline of candidate
- experience: Candidate's total work experience in years
- company_size: Number of employees in the current employer's company
- company_type: Type of current employer
- last_new_job: Timespan in years between finishing the previous job and starting the current job
- training_hours: Training hours completed
- another_job: 0 – Not looking for a job change, 1 – Looking for a job change

The second dataset (*FirstAssignmentBPI-DataSet2.xes*) is a real-life event log containing events for loan applications.

Question 1 – Preprocessing the data (4 points)

This task should be done using RapidMiner

Please carry out the following preprocessing steps on *FirstAssignmentBPI-DataSet1.csv*: **(4 points)**

1. In column "experience", replace value **>20** with **23** and **<1** with **0**. Change the attribute type to numerical.
2. In column "last_new_job", replace value **>4** with **5** and **never** with **7**. Change the attribute type to numerical.
3. Save this dataset with the name *dataset1.csv* and upload it with your assignment.
4. Show a snapshot of the first 10 rows of the dataset.

Question 2 - Clustering (26 points)

This task should be done using RapidMiner. When asked, please provide a screenshot of the main process design (design tab) that you used in RapidMiner.

We want to analyze the differences between the instances (candidates) in dataset1. For this question, just consider the variables "last_new_job", "another_job", "experience", "city_ development _index", and "training_hours".

Use the student ID of one of the students in the group as the seed and sample 95 percent of the instances. Write down the student ID you used in your submission.

1. Normalize "last_new_job", "experience", "city_development_index" and training_hours" using Z-transformation. Using the K-Means algorithm, cluster instances in the sampled dataset into 3, 5, and 7 clusters. We name these three clusterings as A, B, and C respectively. Show the main process design (design tab) that you used in RapidMiner and upload the designed process with your assignment (the name of the process should be "processQ21"). The setting of the uploaded version should be the same as the setting that you have used for the assignment. Answer the following questions. **(9 points)**
   a. What are the centroids (using the default parameter settings) of each clustering (A, B, and C)?
   b. For each clustering (A, B, and C), which cluster is the most coherent one?
   c. Which one of A, B, and C is a better clustering? Justify your answers?
2. Consider clustering B (i.e. five clusters) of the previous task. Describe the differences between the instances (candidates) in the smallest and the biggest cluster (in terms of the number of instances in the cluster). **(4 points)**
3. Consider clustering B (i.e. five clusters) again. Show the means of the clusters (centroids) using a line plot. Is/Are there any feature(s) that can be considered redundant while determining the clusters? Explain why. **(3 points)**
4. Again, sample 95 percent of the instances in dataset1 using the same student ID and then divide it into two subsets:
   a. Subset1: those candidates that are looking for a job change
   b. Subset2: those candidates that are not looking for a job change

   Consider "last_new_job", "experience", "city_development_index" and "training_hours". Cluster subset1 and subset2 in five clusters using these features. Show the main process design (design tab) that you used in RapidMiner. Upload the

designed process with your assignment (the name of the process should be "processQ24"). The setting of the uploaded version should be the same as the setting that you have used for the assignment. **(4 points)**

5. Use the centroids or the line plots of the generated clusters and explain the similarities and dissimilarities between the clusters on each subset in terms of values of centroid attributes. **(6 points)**

Question 3 - Association Rule (15 points)

This task should be done using RapidMiner. When asked, please provide a screenshot of the main process design (design tab) that you used in RapidMiner.

The goal is to find the association rules in dataset1.

1. In column "another_job", replace the value **1** with **searching** and replace **0** with **not_searching**. Change the attribute type to categorical. (Name this dataset *dataset2.csv* and upload it with your assignment) **(1 point)**
2. Use the same student ID as seed and sample 95 percent of the instances.
   Consider the following variables: "experience", "training_hours", "company_size", "company_type", "education_level", "major_decipline", and "another_job". Discretize "experience" and "training_hours" into 5 bins. Use the *FP-Growth* algorithm to min frequent item-sets with minimum support of 0.7. For association rule mining, use minimum confidence of 0.5. List the rules with their support and confidence in decreasing order w.r.t. their support. Show the main process design (design tab) you used in RapidMiner. Upload the designed process with your assignment (the name of the process should be "processQ32"). The setting of the uploaded version should be the same as the setting that you have used for the assignment. **(5 point)**
3. Interpret and explain at least three interesting rules that have a support value larger than 0.5. For example, an association rule: "Graduate -> not_searching" means that most of those who are graduated do not look for a new job. **(9 points)**

Question 4 - Process Mining (55 points)

This task should be done using Disco and ProM (1 and 8b in ProM, everything else in Disco). Please provide a screenshot of the process model if asked. Also give a brief description of 1-2 sentences on how you came to your answer (e.g., which operations in Disco you used).

For this question, use the dataset *FirstAssignmentBPI-DataSet2.csv*. The event log is obtained from a real-life loan application process. Customers approach this company and request a loan.

1. Preprocessing: Load the event log into ProM, convert it to XES, apply a random trace sampling to retrieve 31400 traces, and use the student ID from the previous tasks as your seed. Export this sampled log as a csv-file. Load the log into Disco. Make sure that the columns are correctly assigned, i.e., case, activity, start and complete timestamp, resource, and other data (loan goal, application type, and requested amount). Write down the student ID used for sampling.
   From now on, use Disco for answering the questions 2-8a.
2. Activities **(3 points)**
   a. How many activities are there in the log? **(1 point)**
   b. Which are the most and the least frequently executed activities? **(2 points)**
3. Performance **(5 points)**
   a. Which is the activity that takes on average the most time? What is the average time? **(2 points)**
   b. Show the process model discovered by Disco with 100% of the activities, 100% of the paths (in further tasks abbreviated with 100%/100%) and annotated with the average duration. **(3 points)**
4. Special behavior **(7 points)**
   a. Which activity did not appear before 30-06-2016? **(2 points)**
   b. Show the 100%/100% model of all cases containing this activity. **(3 points)**
   c. What is the number of variants of the cases that include this activity? **(2 points)**
5. Resources **(8 points)**
   a. Which resource takes on average the longest time? **(2 points)**
   b. Show the 100%/100% process model for only this resource. **(3 points)**
   c. What is this resource primarily doing? Give the total time spent by this resource on that task. **(3 points)**
6. Variants **(5 points)**
   a. Which is the most frequent variant? **(1 point)**
   b. Show the 100%/100% process model of this variant. **(2 points)**
   c. How long does the longest case of this variant take? **(2 points)**
7. Other Data **(7 points)**
   a. What is the most frequent reason for customers to apply for a loan? **(1 point)**

b. How many cases follow this objective? Out of all of these cases, how many do just apply for a limit raise of a previously approved loan? **(4 point)**

c. Show the 100%/25% process model for all cases that follow this loan objective. **(2 point)**

8. Analysis **(20 points)**

a. A large number of customers apply for a car loan. Many of those customers are concerned that their desired car will be gone if the loan is not approved quickly enough. Therefore, the company wants to reduce the throughput time. As a data scientist, you were asked for advice. What would you tell the company? Name two points to improve this situation with respect to the company's current process, e.g., process model, performance analysis, resource analysis, etc... **(10 points)**

b. Use ProM (https://www.promtools.org/) for this task. Use the original (unsampled) event log.

The company from the previous task now also hires a business consultant. The business consultant immediately points out that there are timespans in which no applications are handled and no work is done. As he is confused about why this is the case, he asks for your help. Use the dotted chart in ProM to show patterns in the working times and explain them. Provide your screenshot of the dotted chart. **(10 points)**

Deliverables

The deadline for this part of the assignment is **Friday, 04/06/2021 23:59.** You will need to hand in your submission via **Moodle**. Note, the deadline is strict (i.e., there is no extension possible, and late submissions will not be considered). Do not risk last-minute technical problems due to internet problems or RWTHmoodle failures. Your submission should be two files: A **PDF file**, which presents your results, explanations, and screenshots of the used tools, and a compressed file (.zip/.rar) including all supplementary material that you upload, e.g., the **workflow files** of RapidMiner.

Report requirements:

- Succinct but clear answers to the questions. Avoid being verbose while not answering the question.
- The report file should not exceed **10 pages**.

Grading

Participation in the assignment is one of the prerequisites for taking the written exam. The parts of the assignment and the exam form a whole and it is not possible to retake parts of the course, i.e., the results of the assignment expire after the semester. Furthermore, the assignment can only be redone in the next academic year.

10% of the grading of this part of the assignment is based on the **reporting style** (answers should be well-structured and explanations should be clear).

The grade of this part of the assignment counts 10% towards the final grade (see the study guide for details). Please note that the correctness and completeness of your results, and also the accuracy of your explanation are essential.