Outline

ANONYMOUS (for peer-evaluation)

Present your findings as a clear and structured argument: connect the bullet points with appropriate signal words into a coherent line of thought. Your argument should be built around at least two pieces of evidence from your analysis. The outline for the final report consists of three parts:

Introduction. In no particular order:

- *Briefly introduce the process.*

The process is related to the loan application and offer process between a bank and its applications. The process owner is the bank, and the main external stakeholder is the loan applicant. Depending on several characteristics of the loan and its applicant, the bank rejects or makes an offer of varying amounts to the applicant. If the applicant rejects the offer, the bank can follow up with other offers. The bank's objective maximize revenue and minimize costs. This can be approached two-fold: increase number of accepted offers, as well as reduce non-value added time.

- *Inform the reader about prediction task you identified and its feasibility. This is essentially the summary of your main analysis results.*

Following up from the process introduction, we identified a prediction task to predict which offers will be accepted or rejected by the applicant. Potentially, the company can reduce efforts on applicant that might likely reject the offer, or conversely increase efforts on such applicants to increase the likelihood that they will accept the offer, and also increase customer service efforts on those that are likely to accept. Moreover, by modeling this as a regression task or decision tree, the company can better understand the relationships between customer characteristics and those that are likely to accept an offer. Lastly, this task can be extended with even more informative prediction tasks if time allows, for instance predicting the offer amount that might make an applicant more likely to accept the offer, based on the application characteristics.

- *Motivate why you recommend this prediction task in terms of business value to the stakeholder.*

Please view previous section.

Main part. The main part is a coherent line of thought that explains the suitability and value of the proposed prediction model to the process owner. Cover the following aspects of your process-mining and prediction analysis:

A. Selecting a relevant prediction task:

- *Describe the prediction task that you will solve with your prediction model.*

Predict if an applicant will accept or reject an offer.

- *Describe why you chose this prediction task. Explain, both, why it is relevant (wrt. the business) and feasibly (wrt. the data):*
  - *i. Motivate the value of the prediction task to the process owner. Assume an ideal situation, i.e., the model does what you expect it to do with sufficient performance. What would be the business gain of implementing this model? Explain how the model deals with pertinent aspects of the process.*

Business: can focus on clients that are likely to increase an offer and reduce those that will reject, or conversely improve efforts on those that might reject to increase probability that they will accept.

Data: the model uses features that are correlated to the dependent feature (offer rejected/accepted) to achieve this task

- Since an applicant can receive multiple offers, and this may impact if a future offer will be accepted or not, we will (for now) construct a new feature that embeds the offer history by counting the number of rejected offers (there are probably better ways to incorporate this information which we need to investigate in the future).
- This way, for applicants who receive multiple offers, we will predict for each offer if this will be accepted or rejected.

ii. *Argue about the suitability and feasibility of the prediction task in light of the process and the data. Your argument is data-based. Its justification follows from your understanding of the data and the pieces of evidence you collected. Describe each piece of evidence (process model/visualization/table) you use in your argument precisely and with a fitting figure/table. Explain why these pieces of evidence are a relevant and valid basis for your choice of the proposed prediction task. Did you ensure that a layperson can understand the piece of evidence?*

Correlation between independent variables and dependent variables:

- Independent variables:
  - monthly cost
  - credit score
  - offer amount
  - number of terms
- Dependent variables:
  - offer selected or rejected

B. Method for solving the prediction task:

- *Provide a description of the envisioned prediction model relevant to a stakeholder, e.g., what features will be focused on, will a specific subset be selected, and other relevant choices. Provide a justification for these choices wrt. the data (by referring to the pieces of evidence) and/or wrt. the business.*

The first prediction model implemented is a Random Forest Classifier. The decision to proceed with said method has been made to account for: the variability of the available features, the intrinsic complex nature of human behavior, and the reliability of the method. The dataset presents several variables that seem, at first sight, reliable indicators for the selection of a proposed offer (e.g. Credit score, Monthly cost, etc.). Unfortunately, upon examination, some of the intuitive predictors fail to demonstrate reliability. On the other hand, a distinguishing feature of Random Forest is the deliberate inclusion of randomness throughout the construction of individual trees. This is accomplished using bootstrapped samples during tree construction and the selection of a random subset of features for each split. This deliberate randomness serves to decorrelate the trees within the ensemble, enhancing the model's resistance to noise and generalizability that is part of our dataset.

The data exploration and model implementation were preceded by some measures of data pre-processing to clean the dataset from irrelevant data and better handle the relevant information. For this task we used Disco, for the offered tool of data filtering.

Data Pre-processing

First, we loaded the BPI_Challenge_2017.xes file in Disco. The dataset contains many frequent traces, but 87% of the variants have only 1 case, for this reason we focused our attention on the variants that had at least 10 cases (selection). Then we proceeded by excluding all the events (projection) of “lifecycle:transition” that didn’t result as “complete” to avoid redundancy and have a clearer view of the process. Then we proceeded by exporting the filtered dataset as a CSV file to continue the analysis in python.

The features

An exploratory analysis has been conducted using Disco and both R and the PM4PY library for Python. We focused our analysis on: 'Monthly Cost', 'Number of Terms', 'Credit Score'; but also explored other potential features like: 'Offered Amount', 'Requested Amount', 'First Withdrawal Amount'. The least are trivially correlated among themselves. As for the other features:

- The number of Terms refers to the number of payback terms that the client agrees upon accepting the loan offer. This number is mostly associated with the Offered Amount and, because of the obvious link between the two, also with the Requested Amount.
- The monthly cost represents the amount to be paid monthly by the client to repay the loan if the offer is selected. Given that the monthly payment is based on the Offered Amount, it was expected to find great variability in the data. The average cost, considering the whole dataset is close to 270€, but if we consider the interquartile range (IQR) the mean value is actually 232€. This feature is also related to the Offered and Requested Amount.
- Credit score is a peculiar feature. By means of further analysis it results in an incredibly reliable predictor for the task at hand, but upon inspection there are some irregularities. The credit score should represent the reliability of the client, the higher the score, the lower the risk in giving a loan to him. Unfortunately, 100% of the cases that have a credit score different from zero, also have selected an offer. Usual data manipulation, like replacing the zero’s with median or mean values would simply shift the problem. The most probable explanation for this is that the credit score was recorded in the dataset only after the case ended. If we use this variable as a feature, we would surely have 100% accuracy due to extreme overfitting, thus we exclude this feature from the model.

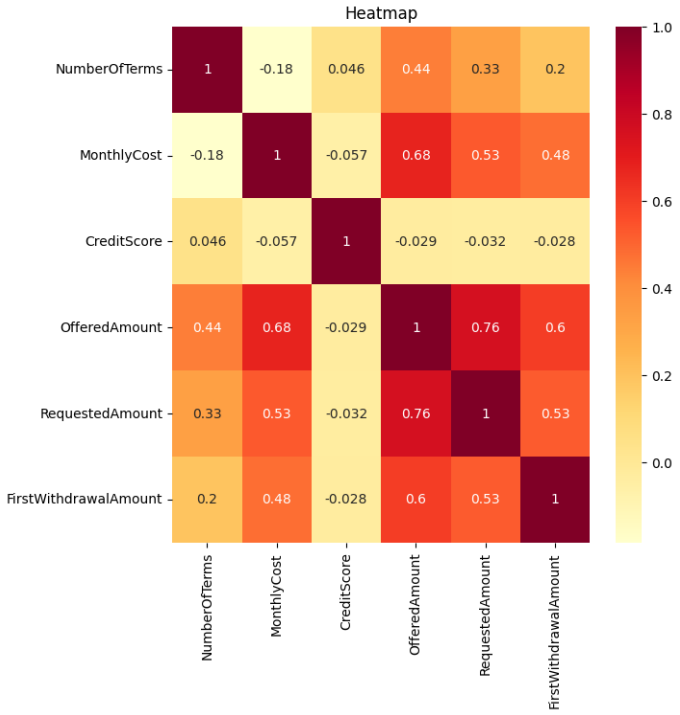

Figure 1 - Correlation Heatmap

Random Forest

Having selected our features (Requested Amount, Offered Amount, First Withdrawal Amount, Monthly Cost, Number of Terms), we proceeded by implementing the Random Forest Classifier of the python library “scikit-learn”. Instead of splitting the dataset we opted for a 5-fold cross validation of the prediction model. The results are quite modest, but expected as this first prediction model is very simple and does not even include trace encodings (that will be implemented in a future model). The resulting mean accuracy was 60%.

C. Preliminary results:

- *Report a work-in-progress result of the prediction model and explain how this should be interpreted. It does not need to be a good result yet, as long as you have some tangible result to show the process owner.*
    - We implemented a basic Random Forest Classifier model as described in section B. While these results are still in early stages and the model is not fully satisfactory, they offer insights into our progress so far.
    - During the data preprocessing (as described in section B), we noticed that the credit score has many missing values. By replacing it with the mean value, we were able to correct the issue, however the potential bias induced remains to be investigated.
    - The naive Random Forest Classifier model was evaluated using cross validation with 5 splits, which all resulted in a test prediction accuracy of 95% or higher
    - The only preprocessing step besides those described in section B was to binarize the dependent variable (‘Selected’)
- *Briefly discuss the results in relation to the current initial implementation. What still needs to be implemented and do you expect this will improve your results?*

How?

- We plan to investigate the effectiveness of random forest models versus regression models
- We plan to investigate more feature analysis, since during our current analysis we did not find significant correlations between predictable variables so far. Potentially, we might look into Principal Component Analysis
- If feasible, we plan to investigate if the model can be extended/modified to predict an offer amount that's likely to be accepted

Conclusion and discussion.

- *Briefly summarize your argument and the evidence you collected.*
  - While our work is still in progress, we conducted a preliminary data and feature analysis, which resulted in an initial Random Forest Classifier which is able to predict whether an applicant will accept or reject an offer with an accuracy of over 95%
  - We did not find strong correlations in our current features analyzed so far, and also encountered many missing values which were imputed by mean value imputation, which may affect the unbiasedness of the results
- *Discuss the suitability of the current prediction model and results to solve the formulated prediction task.*
  - The initial results we have found are promising, but not yet sufficient. The random forest classifier has a high accuracy, but the data imputation method might have introduced rather high bias in the input data. Moreover, we want to analyze more features, as the model may have oversimplified/missed more complex patterns features in the data.
- *Either: conclude with recommendations, or: discuss what has to be done to make the prediction task and model relevant for the stakeholder.*
  - We have thus far identified a model that is able to predict if an offer will be accepted or rejected by a client with high accuracy. While the model needs to be improved, we are convinced that predictive models will yield business value for the company. We intend to further build out our case by improving the model performance, extend the scope of our feature analysis, and possibly investigate if the predicted offer amount (that leads to higher probability of an offer being accepted) can be predcited.

maximum of 6 pages for this outline excluding the appendix

1 Appendix

1.1 Prediction setup

maximum of 1 page

Describe the setup of your implementation and evaluation (see 2AMI20_assignment_phase 3a.pdf – Appendix A) and explain your choices regarding each step. Briefly justify each choice.

Data preparation:

- **Data cleaning.** Disco was used to filter the data, excluding variants that had less than 10 cases (selection), and keeping only events of "lifecycle:transition" set to "complete" (projection). Then the filtered dataset was exported as CSV.
- **Feature selection.** In python we proceeded by analyzing the data with correlation heatmaps and other visualizations. Understanding that we wouldn't manage to implement the feature encoding in time, we decided to try and build a prediction model based entirely on the available numerical data (and not the trace encodings).

Modeling:

- Given that the Random Forest Classifier of scikit-learn had the option of k-fold cross validation we decided to leverage it. We did not split the data into train and test so that the classifier could be trained and evaluated on the full dataset

Evaluation:

- The classification task is evaluated via accuracy, in particular, the mean accuracy of the 5 folds:

```
# print evaluation metrics
print('fit_time:', result['fit_time'])
print('score_time:', result['score_time'])
print('test_score:', result['test_score'])
print('mean accuracy:', sum(result['test_score']) / len(result['test_score']))
```

✓ 1m 5.5s

```
fit_time: [12.74960899 16.19351959 13.38480735 10.50067425 10.85757899]
score_time: [0.34626746 0.32569432 0.31071544 0.33325958 0.37542939]
test_score: [0.59454913 0.61215792 0.60998317 0.61166573 0.59831744]
mean accuracy: 0.6053346784735049
```

1.2 Realization of initial prediction model

maximum of 1 page

Describe how you realized the implementation of your prediction model so that a process mining expert can reproduce your model.

Implementation:

After data pre-processing described above, we splitted the target variable from the feature matrix.

```
# create target variable for prediction
import numpy as np
y = data['Selected']
y = np.array(y).astype(int) # make sure it is an np.array and not a series
x = data.drop(["Selected"], axis=1)
```

Then we instantiated the scikit-learn “RandomForestClassifier” with the specified random state (to reproduce the result set to zero) and used the “cross_validate” function to fit and evaluate the classifier on the default 5-fold cross validation.

```
# Implementation of Random Forest Classifier
from sklearn.ensemble import RandomForestClassifier
clf = RandomForestClassifier(random_state=0)

# fit model with 5-fold Cross Validation
from sklearn.model_selection import cross_validate
result = cross_validate(clf, x, y)  # (defaults) is 5-fold CV
```

1.3 Evidence

1.3.1 Piece of evidence 1: process model (compulsory!)

maximum of ½ page excluding figure/table

Include the visualization/table with a short description of how it was created.

We created multiple visualizations of the process model using ProM, Disco and pm4py. Here is an example of its implementation for the latter using the “pm4py.discover_heuristics_net” function on the pre-processed data.

```
# huristic miner
heu_net = pm4py.discover_heuristics_net(bpi_log, dependency_threshold=0.99)
pm4py.view_heuristics_net(heu_net)
```

From this visualization we can get various insight on the process, for example we can see that there are three ways in which the process can end:

1. A_Pending, if the client selects an offer;
2. O_Refused, if the client refuses all offers;
3. O_Cancelled, if the process owner chooses to end the application.

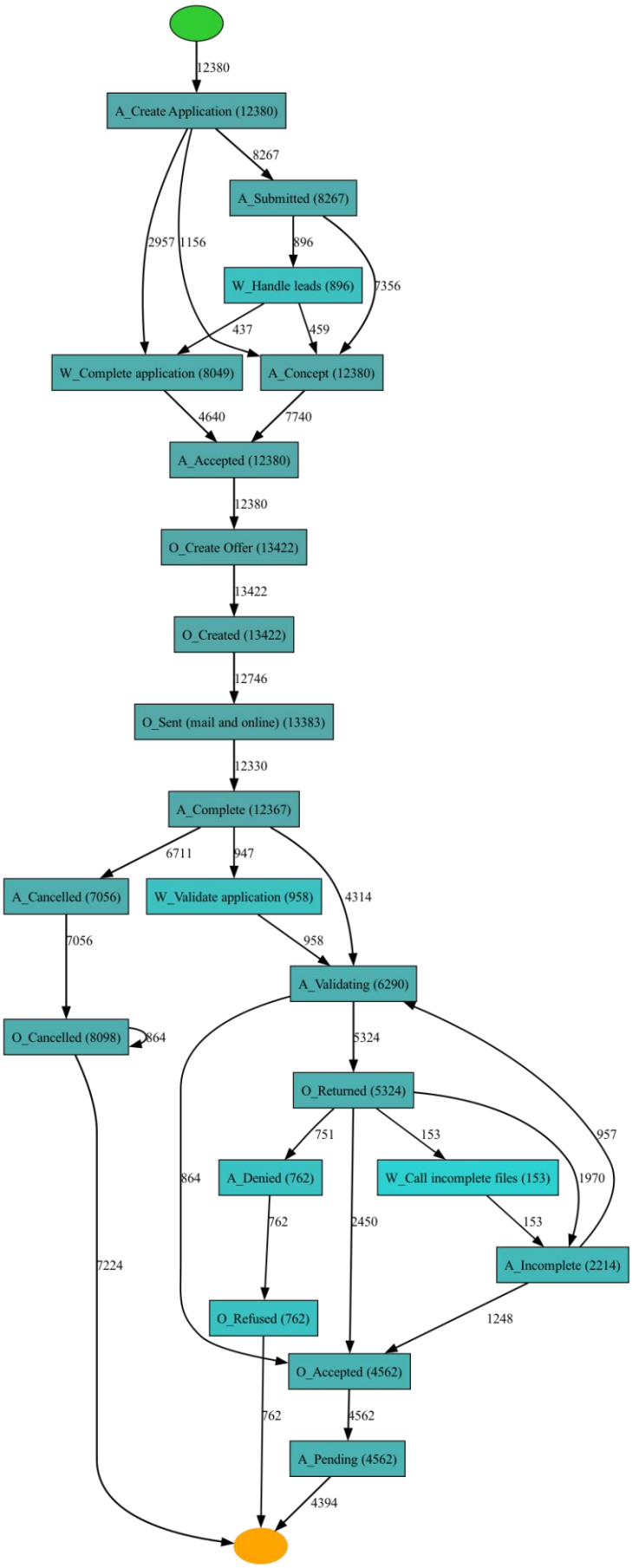

Figure 2 - Process Model discovered with Heuristic Miner

1.3.2 Piece of evidence 2: process model / figure / table

maximum of ½ page excluding figure/table

Include the visualization/table with a short description of how it was created.

As discussed in section B, “CreditScore” seems a reliable feature for process outcome prediction, but upon inspection it reveals its inconsistency: 70% of the dataset (and therefore of the client base) has a CreditScore of zero.

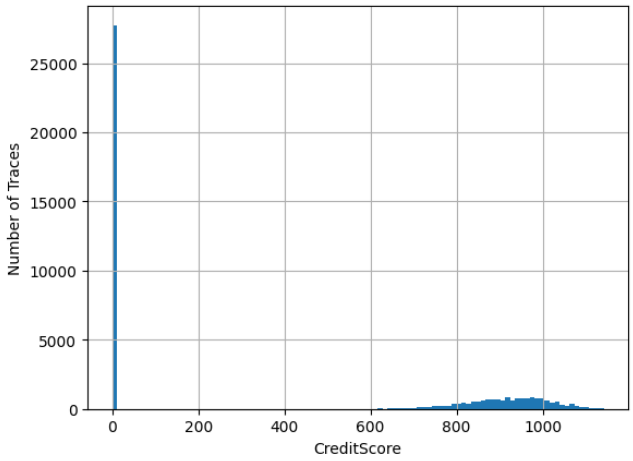

Figure 3 - Histogram of CreditScore feature

Moreover, you can verify that 100% of the credit scores of the cases that did not select any offer are equal to zero. You just need to add the following filters to the dataset on Disco (other than the one specified in the pre-processing):

1. Attribute > filter by: *CreditScore* > Keep Selected > deselect empty row
2. Attribute > filter by: *Selected* > Forbidden > *true*

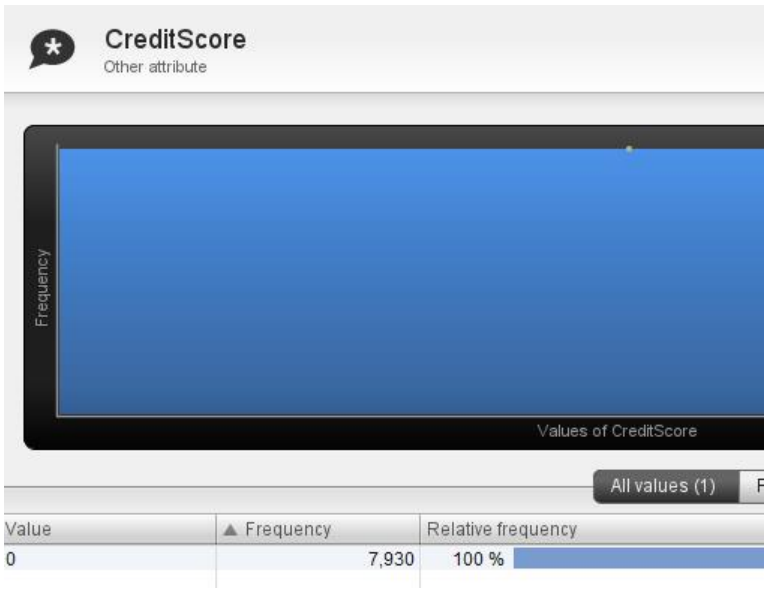

Figure 4 - Disco screenshot (relative frequency)