

Introduction to Machine Learning (SS 2024)

Project: Programming Project

Author 1

Last name: Rieser
First name: David
Matrikel Nr.: 12141689

Author 2

Last name: Sillaber
First name: Cedric
Matrikel Nr.: 12211124

This template outlines the sections that your report must contain. Inside each section, we provide pointers to what you should write about in that section (in blue text).

Please remove all the text in blue in your report! Your report should be 2 pages for regular teams (excluding references!) and 3 pages for the three person team.

I. INTRODUCTION

The task of this project is to implement machine learning models to detect fraudulent transactions. The problem type is a binary classification task. Based on the timestamp, value of transactions and 28 features the model should predict if a transaction is fraudulent or not. The dataset consists of 227845 instances with 30 features and flag indicating if the transaction is fraudulent or not. The dataset is highly imbalanced with 0.15% of the transactions being fraudulent. There is no information about the 28 features, so it is not clear what they represent.

As there the dataset is labelled, we can use supervised learning to solve this problem. We decided to use two methods for this problem. The first choice was to use a neural network for this task. Neural networks can extract the most important features from the data and are able to learn complex patterns. The second choice was a decision tree. In order to achieve even better results, we also implemented a random forest classification.

- What is the nature of your task (regression/classification)? Is it about classifying types of birds, or deciding the number of cookies an employee receives?
- Describe the dataset (number of features, number of instances, types of features, missing data, data imbalances, or any other relevant information).

II. IMPLEMENTATION / ML PROCESS

- Did you need to pre-process the dataset (e.g. augmenting data points, extracting features, reducing the dimensionality, etc.)? If so, describe how you did this.
- Specify the method (e.g. linear regression, or neural network, etc.). You do not have to describe the algorithm in

detail, but rather the algorithm family and the properties of the algorithm within that family, e.g. which distance functions for a decision tree, what architecture (layers and activations) for a neural network, etc.

- State (in 2-5 lines) what makes the algorithm you chose suitable for this problem. What are the reasons for choosing your ML method over others?
- If you used a method that was not covered in the VO, describe how it is different from the closest method described in the VO.
- How did you choose hyperparameters (other design choices) and what are the values of the hyperparameters you chose for your final model? How did you make sure that the choice of hyperparameters works well?

III. RESULTS

- Describe the performance of your model (in terms of the metrics for your dataset) on the training and validation sets with the help of plots or/and tables.
- You must provide at least two separate visualizations (plot or tables) of different things, i.e. don't use a table and a bar plot of the same metrics. At least three visualizations are required for the 3 person team.

IV. DISCUSSION

- Analyze the results presented in the report (comment on what contributed to the good or bad results). If your method does not work well, try to analyze why this is the case.
- Describe very briefly what you tried but did not keep for your final implementation (e.g. things you tried but that did not work, discarded ideas, etc.).
- How could you try to improve your results? What else would you want to try?

V. CONCLUSION

- Finally, describe the test-set performance you achieved. Do not optimize your method based on the test set performance!
- Write a 5-10 line paragraph describing the main take-away of your project.