Cyber Data Analytics Assignment 1

INTRODUCTION

One of the big tasks in cyber data analytics is detecting malicious or fraudulent data records. This is a machine learning task that is extra complicated due to several properties of the data and its domain:

1. The large majority (typically over 99%) of the data is benign
2. The data come from individuals, whose privacy needs to be protected
3. Malicious users actively try to hide
4. The amount of data to learn from is enormous
5. Often this data is unlabeled

In this assignment, our focus will be on the first property: how to deal with large class imbalances in machine learning. Developing methods for dealing with the last three properties are the topic of follow-up assignments. You are expected to work for 30 hours per person on this lab assignment.

LEARNING OUTCOMES

After completing this assignment, you will be able to:

1. Correctly apply machine learning methods to real data
2. Modify machine learning algorithms to deal with class imbalance
3. Manipulate data to protect people's privacy
4. Analyze the outcomes of machine learning for fraud detection

INSTRUCTIONS

For this lab assignment you will use the 'data_from_student_case.csv' data to predict credit card fraud. One of you will work on SMOTEing and the other on rank swapping, you may not work together on these parts! Together you will complete the data exploration and classification tasks.

To get started with the assignments it is useful to load the dataset into a numpy array, the pandas library could be useful to do this.

Data exploration (5 points)

- Explore the dataset and generate 1-3 plots that give insight into the dataset, i.e. highlight a difference between benign and malicious samples.
- Explain what interesting insights you got from the plots.
- Explain what features could be useful for discriminating between these two classes.

You may use any visualization method such as a Heat map, a Scatter plot, a Bar chart, a set of Box plots, etc. as long as they show all data points in each figure.

SMOTEing - individual (10 points)

- Implement the SMOTE algorithm.
- Split the data into a train / test set (e.g. 80-20%) and apply your SMOTEing algorithm.
- Train three classifiers using SMOTE and visualize their results using ROC curves.
- Explain which method works best. Can you explain the performance difference between different classifiers? Is using SMOTE a good idea? Why / why not?

Rank swapping – individual (10 points)

- Implement the rank swapping algorithm.
- Split the data into a train / test set (e.g. 80-20%) and apply your rank swapping algorithm.
- Train three classifiers using rank swapping and visualize their results using ROC curves, resample data if needed.
- Explain which method works best. Can you explain the performance difference between different classifiers? Is it advisable to protect people's privacy using rank-swapping? Why / why not?

Classification (10 points)

Resample/synthesize data to deal with class imbalance if needed.

- Set up 10-fold cross validation.
- Train a black-box model that performs very well on the dataset.
- Train a white-box model that predicts well but can still be explained.
- Compare the performance of the two algorithms, focusing on performance criteria that are relevant in practice. Explain your choice of metrics.
- Explain when your white-box classifier predicts transactions as fraudulent.

Explanations and code quality (5 points)

- Remember to write clear code and explain your results understandably. During peer-review, your fellow students will be asked to run and understand your code!
- More text is not always better!

Bonus! (5 points)

So far, we looked at classifying individual transactions but perhaps there is more useful information in groups of transactions, e.g., how often is the same card used? Or, what is the total amount spend by this card?

- Try extracting such 'aggregate features' (not just these two examples) and use these features in your classifiers to improve the predictive performance.
- What aggregates did you try?
- Determine which aggregate feature works well/best, why do you think it works? How much does it improve on your results from the classification exercise?

RESOURCES

Slides from Lectures 1 and 2

- "Learning from imbalanced data" paper, by He and Garcia.
- "An introduction to ROC analysis" paper, by Fawcett.
- "SMOTE: Synthetic Minority Oversampling Technique" paper, by Chawla et. al .
- "A survey of inference control methods for privacy-preserving data mining" by Josep Domingo-Ferrer

All are made available through Brightspace.

Realistic fraud detection data provided by Adyen, available through Brightspace.

The Jupyter Python notebook for processing the fraud data, available on Brightspace.

PRODUCTS

A zip containing:

- A Jupyter Python notebook for the collaborative parts of the assignment. The word count should not exceed 1000 words (see first cell). Include libraries used to run the code other than numpy, scipy, pandas, and scikit-learn.
- Separate Jupyter notebooks for the individual assignments, each not exceeding 300 words.
- The notebooks will be assessed using the below criteria.

ASSESSMENT CRITERIA

The assignment will be reviewed by your peers, and you are expected to individually review 2 reports. The estimated time you should spend on a review (including code review) is 1 hour. The login details will be provided in the week of the deadline.

Knockout criteria (will not be evaluated if unsatisfied):

Your code needs to execute successfully on computers/laptops of your fellow students (who will assess your work). You may assume the availability of 4GB RAM. Please test your code before submitting. In addition, the flow from data to prediction has to be highlighted, e.g., using inline comments.

Your report needs to satisfy the word count requirements for the different parts.

Submissions submitted after the deadline will not be graded, **deadlines are strict!**

The report/code will be assessed using these criteria:

Criteria	Description	Evaluation
Visualization	Shows an interesting relationship in the data. The relationship is relevant for fraud detection.	0-5 points
ML Workflow (Individual)	The data preprocessing is sound. The flow from data to prediction is correctly implemented. At least three different classifiers have been tested on the data.	0-5 points
Modification (Individual)	The machine learning algorithms have been modified at the correct point in the data-prediction flow. The modification and obtained ROC curves are sound.	0-5 points
Performance	At least 100 fraudulent cases are found in the test data, with at most 1000 false positives. Worse performance means fewer points.	0-5 points
Analysis	The analysis is correct and the conclusions are reasonable. The conclusions are relevant for fraud detection in practice.	0-5 points
Bonus	Creative solution, correctly implemented.	0-5 points
Report and code	The data-prediction flow is clearly described, including preprocessing and post-processing steps.	0-5 points

Your total score will be determined by summing up the points assigned to the individual criteria. Your report and code will be graded by the teacher and assistants, and the peer reviews are used as guidance.

In total 35 points (including bonus) can be obtained in each lab assignment, of which 10 are for the individual parts.

In total 140 points (including bonus) can be obtained in the 4 lab assignments, of which 40 are individual. The total number of obtained points will be divided by 120 to determine the final course grade.

You will receive a penalty of 5 points for each peer review not performed. Significantly different reviews will be subject to investigation. If deemed badly done by the teacher or TA, you will also receive 5 penalty points.

SUPERVISION AND HELP

We use Mattermost for this assignment. Under channel Lab1, you may ask questions to the teacher, TAs, and fellow students. It is wise to ask for help when encountering start-up problems related to loading the data or getting a machine learning platform to execute. Experience teaches that students typically answer within an hour, TAs within a day, and the teacher the next working day. When asking a question to a TA or teacher, your questions may be forwarded to the channel to get answers from fellow students. Important questions and issues may lead to discussions in class.

Lab sessions are online on Wednesday afternoons on gather.town. Please see Brightspace for details.

SUBMISSION AND FEEDBACK

Submit your work in Brightspace, under assignments. Also submit it on peer.tudelft.nl. Within a day after the deadline, you will receive several (typically two) reports to grade for peer review as well as access to the online peer review form. You have 5 days to complete these reviews. You will then receive the anonymous review forms for your groups report and code.

There is the possibility to question the review of your work, up to 3 days after receiving the completed forms. You should do so via the response function on peer.tudelft.nl.