Cyber Data Analytics Assignment 1

Study the "Learning from imbalanced data" paper, by He and Garcia.
Study the "An introduction to ROC analysis" paper, by Fawcett.
Study the "SMOTE: Synthetic Minority Oversampling Technique" paper, by Chawla et. al .
Study the "MetaCost: A General Method for Making Classifiers Cost-Sensitive" paper, by Domingos.

Visualization task (5pt) – 1 A4

Load the fraud data into your favorite analysis platform (R, Matlab, Python, Weka, KNIME, …) and make a visualization showing an interesting relationship in the data when comparing the fraudulent from the benign credit card transactions. 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 one figure. What feature(s) and relation to show is entirely up to you. Describe the plot briefly.

Imbalance task (5 pt) – 1 A4

Process the data such that you can apply SMOTE to it. SMOTE is included in most analysis platforms, if not you can write the scripts for it yourself. Analyze the performance of at least three classifiers on the SMOTEd and UNSMOTEd data using ROC analysis. Provide the obtained ROC curves and explain which method performs best. Is using SMOTE a good idea? Why (not)?

Classification task (10 pt) – 2 A4, prices from Adyen for the top performers!

Build two classifiers for the fraud detection data as well as you can manage:

1. A black-box algorithm, ignoring its inner workings: it is the performance that counts.
2. A white-box algorithm, making sure that we can explain why a transaction is labeled as being fraudulent.

Explain the applied data pre-processing steps, learning algorithms, and post-processing steps or ensemble methods. Compare the performance of the two algorithms, focusing on performance criteria that are relevant in practice, use 10-fold cross-validation. Write clear code/scripts for this task, for peer-review your fellow students will be asked to run and understand your code!

Bonus task (3 pt) – 1 A4

Try to add context to your classifier by linking/grouping the transactions based on IP, or card number, or transaction date, or country code, etc. For example, you can first aggregate the data before learning a classifier, or post-process the classifier decisions into improved estimates. Do whatever you can to improve your classifiers performance by not seeing every table row as an individual record, they are linked to others in many different ways.