ULB Credit Card Fraud data

Background

Banks need to decide—rapidly and thus algorithmically—whether credit card transactions appear fraudulent and should be declined. The Machine Learning Group at Université Libre de Bruxelles (ULB) have assembled and published[1] a data set that bears on this issue. It includes information about 284,807 transactions made by European credit card users over a two day period in September 2013, but despite its large size, it includes only 492 cases of fraud. This extreme rareness is among the main challenges in genuine fraud-detection problems.

Problem & Data

The problem for the oral exam is to develop a way to test whether a given transaction is fraudulent. The extreme rarity of fraud makes this difficult using the methods we've studied this term, so you should work with certain artificially-constructed subsets of the data in which the fraction of fraudulent transactions has been enhanced to around 10%. There are thus two versions of the data for this problem:

- A collection of CSV files, `CreditCardFraud_1.csv` through `CreditCardFraud_5.csv`, each of which describes a sample of 5000 transactions that includes all 492 cases of fraud and a further 4508 examples sampled from the legitimate transactions.
- The full data set as a CSV file, `CreditCardFraud_AllData.csv`.

Table 1 describes the format of these files.

What to prepare

Develop a strategy to determine whether a given transaction is fraudulent or not, based only on the data provided by the ULB group (see below for details), then prepare a short presentation (no more than 5 or 6 slides) describing your approach and assessing its success.

[1] I got the data from Kaggle, a company now owned by Google that runs Machine Learning competitions and provides a platform for the sharing of data and code.

Time	Amount	Is.Fraud	V1	$\cdots$	V28
40	13.84	FALSE	1.11069200372208	$\cdots$	0.00494429617585088
41	2.67	FALSE	1.15431211678574	$\cdots$	0.0196676927362536
	$\vdots$		$\vdots$		$\vdots$
further legitimate transactions			$\vdots$		$\vdots$
	$\vdots$		$\vdots$		$\vdots$
406	0	TRUE	-2.3122265423263	$\cdots$	-0.143275874698919
472	529	TRUE	-3.0435406239976	$\cdots$	0.0357642251788156
	$\vdots$		$\vdots$		$\vdots$
further fraudulent transactions			$\vdots$		$\vdots$
	$\vdots$		$\vdots$		$\vdots$

Table 1: The format of the files `CreditCardFraud_*.csv`. The first column gives the time in second since the first transaction in the database, while the second gives the amount (in Euros? Kaggle doesn't say) and the third is a logical variable indicating whether the transaction was fraudulent. The remaining 28 columns are anonymised, real-valued scores derived from features of the transaction that the ULB group wish to obscure.

Marking scheme

The oral exam, which accounts for 40% of your overall mark, will be assessed against the following marking scheme, reproduced from the module handbook.

Mark	**Interpretation**
Technical Understanding (out of 15)	
0–7	Major error or ommission.
8–9	Understanding of technical material at expected level.
10-11	Deeper understanding of material than expected.
12–15	Outstanding grasp of all technical issues.
Presentation (out of 15)	
0–7	Significant lack of clarity.
8–9	Presentation at expected level.
10-11	Particularly clear presentation.
12–15	Outstanding presentation of material.
Response to Questioning (out of 10)	
0–4	Inability to answer key questions.
5	Acceptable responses to questions.
6	Good answers to questions.
7–10	Very strong response to questioning.