Data Mining Coursework

Outline Solution

https://machinelearningmastery.com/use-ensemble-machine-learning-algorithms-weka/ (voting commitee)

**Abstract**

In this report we provide a solution to a….. Using data from a marketing campaigns of a banking institution…

**1. Introduction**

This document presents an outline solution to the coursework. It is not meant to be a complete report, but highlights some of the issues and results that arise when following a systematic data mining process.

The goal is to develop models to classify…. under both equal and unequal costs. In the following sections, we report on data exploration, data pre-processing, model building, and results.

…

**2. Data exploration**

The dataset was supplied as an ARFF file ready for use with the Weka data mining toolkit. Initial exploration of the training dataset showed the following features.

* The 'no' class made up roughly 88% of the data. This is the accuracy of the default classifier
* The data set has no missing values.
* The data set has 17 attributes and 36,000 instances. Since we have a large amount of data, reducing the number of attributes might not be necessary. We would try to reduce the data set anyway…
* Some attributes have imbalanced distributions of values:
  + Default: More than 95% of the data have value ‘yes’
  + Loan: Most of the data have value ‘no’
* Looking at the histograms of each attribute, there were no attribute that were strongly predictive of the class. *Viewing the histograms for each variable showed that there were no variables that were strongly predictive of the class*. (In the histogram of the 'default' attribute, it might look like that a 'yes' label of this attribute would belong to class 'no'. But inspecting the scatter plot showing that there are small number of instances with label 'yes' belong to class 'yes'

**3. Data pre-processing**

There are 7 numeric attributes: age (a1), balance (a6), day (a10), duration (a12), campaign (a13), pdays (a14) and previous (a15).

Principal component for in attribute selection to plot scatter plot.  
add expression

**4. Classification Models**

**4.1**

**4.2 Attribute selection**

**4.3**

**5. Evaluation and Conclusions**

The maximum achieved by anyone was 82%. The test dataset contains 34 good examples and 16 bad examples, so the default rule (classifying every example as good) has an accuracy of 68%

The performance of….