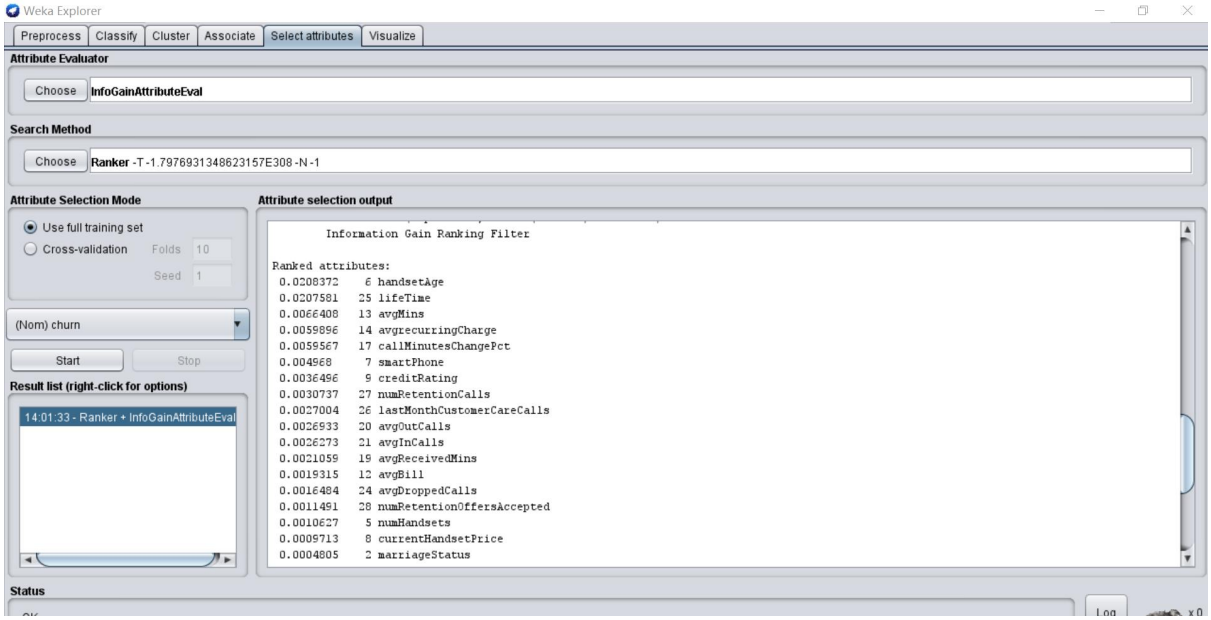
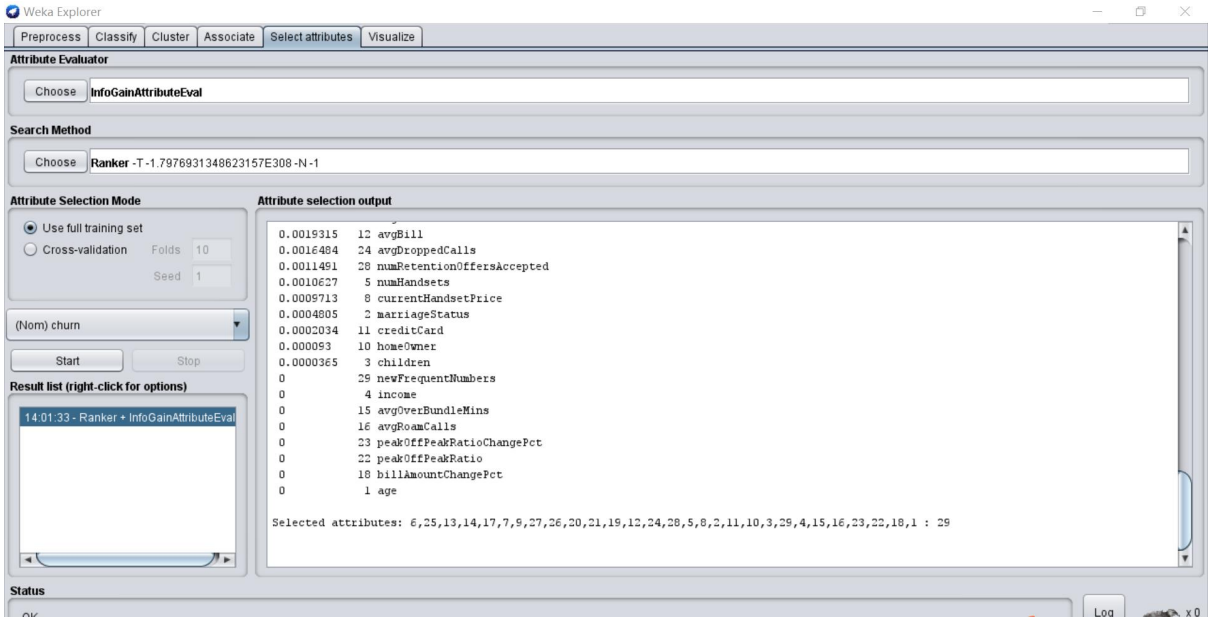
Question 1

(a) Apply one filter and one wrapper feature selection strategy from those available in Weka and report the feature subsets that they select. In the case of a filter, you must propose a way to choose a subset of the ranked features, rather than using the entire original set of features. You should justify your choice.

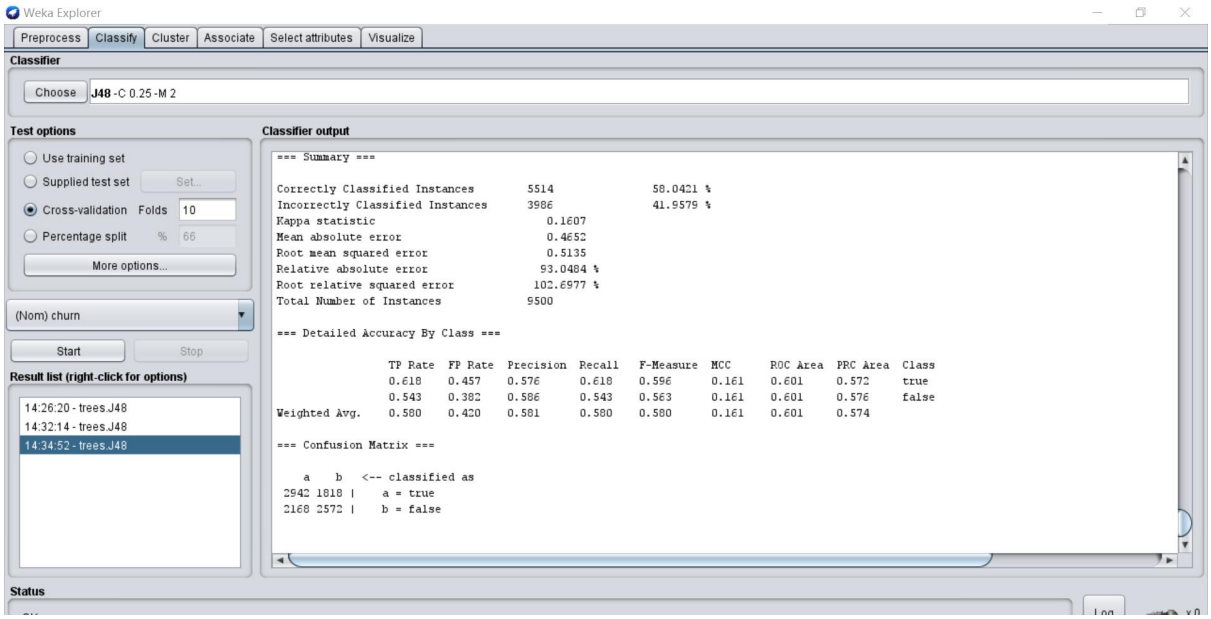
1. Filter feature selection strategy:

In Weka Select attributes tab, I choose InfoGainAttributeEval as the evaluator, Ranker as the method, and the ranked attributes are :

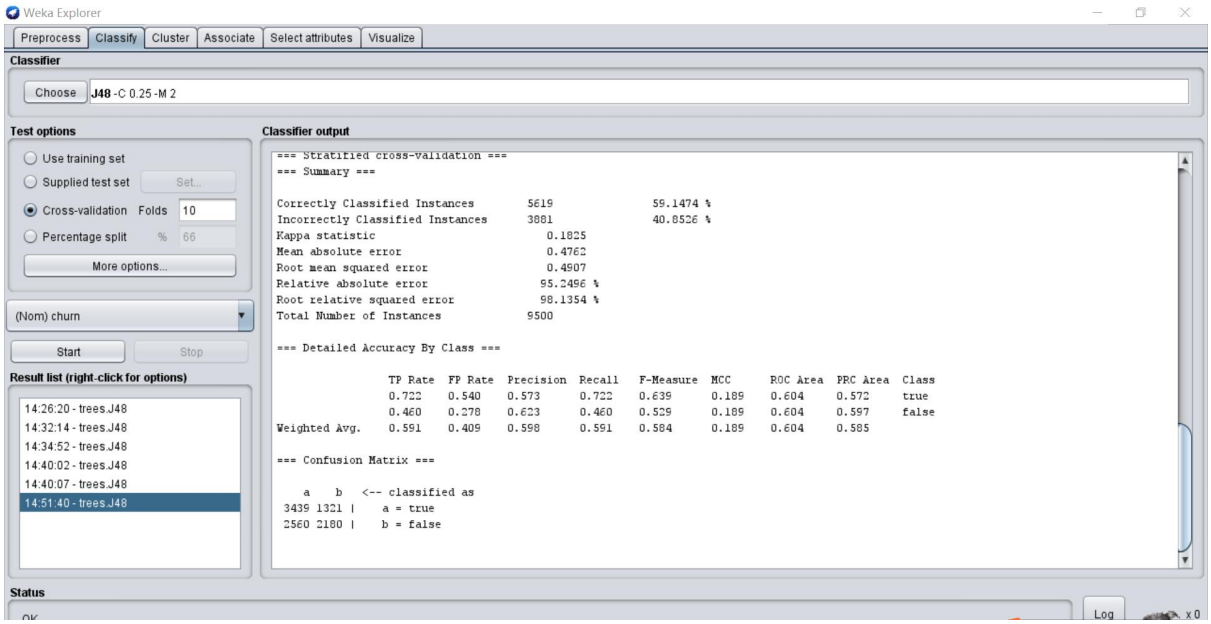




After I remove the the attributes which has the value of 0 and I select the top 50% of the rest ranked list of features, I run the J48 classifier with the new feature subset, the accuracy is :



Then I tried to use the top 4 most discriminating features, I got the accuracy :

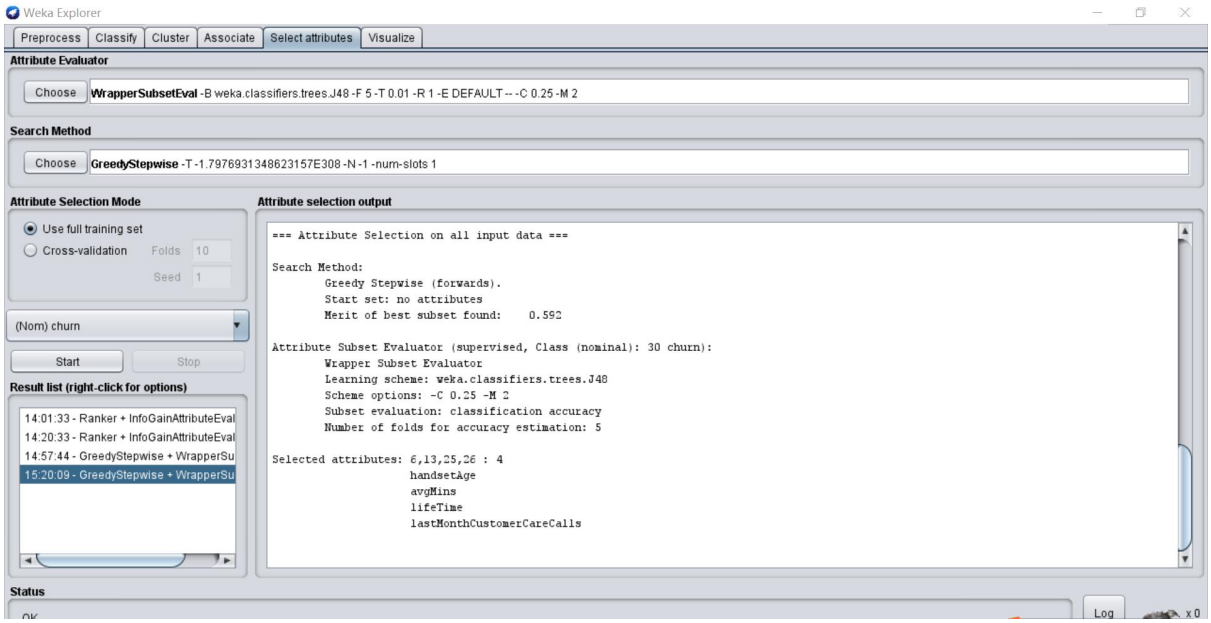


There are several different ways of choosing feature subset when using filter. Because the features are ranked based on their scores or Information Gain (IG), so normally, the higher the score is the more important the feature is . So I Select a subset of the top ranked features to use for classification in two ways:

Select the top 50% and select 4 features , and the 4 features can get more accurate classification outcome.

2. Wrapper feature selection strategy:

I used forward Sequential search: In Select attributes tab, choose WrapperSubsetEval as evaluator and J48 as classifier.I Choose GreedyStepwise as the search method, with option SearchBackwards = False. The outcome feature subset is :



(b) Report and discuss the differences between the feature subsets produced by the filter and wrapper techniques from Task (a). Provide explanations for why the two techniques can potentially produce different results.

The top 4 features in filter method is :handsetAge, lifeTime, avgMins, avgrecurringCharge

The top 4 features in wrapper method is :handsetAge, lifeTime, avgMins, lastMonthCustomerCareCalls

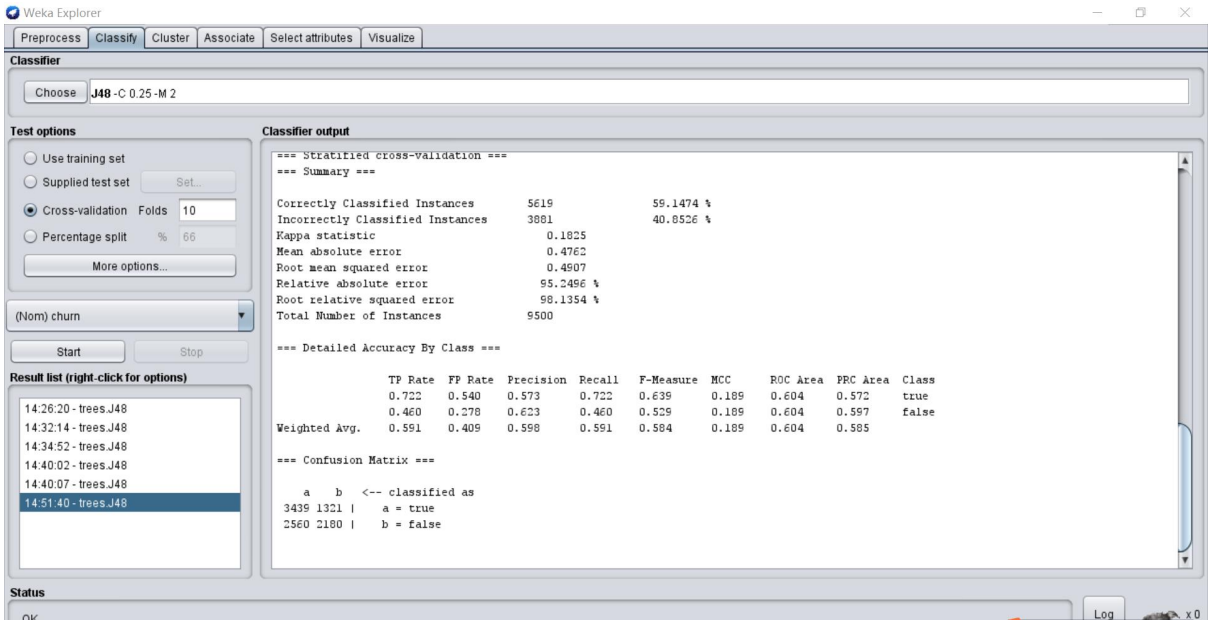
There is a few differences between the selected features between these two methods.The possible reasons are:

The filter method is based on the fixed mathematical formula and there is no model bias in feature selection, furthermore, In filters, the features are considered in isolation from one another, and are not considered in context.

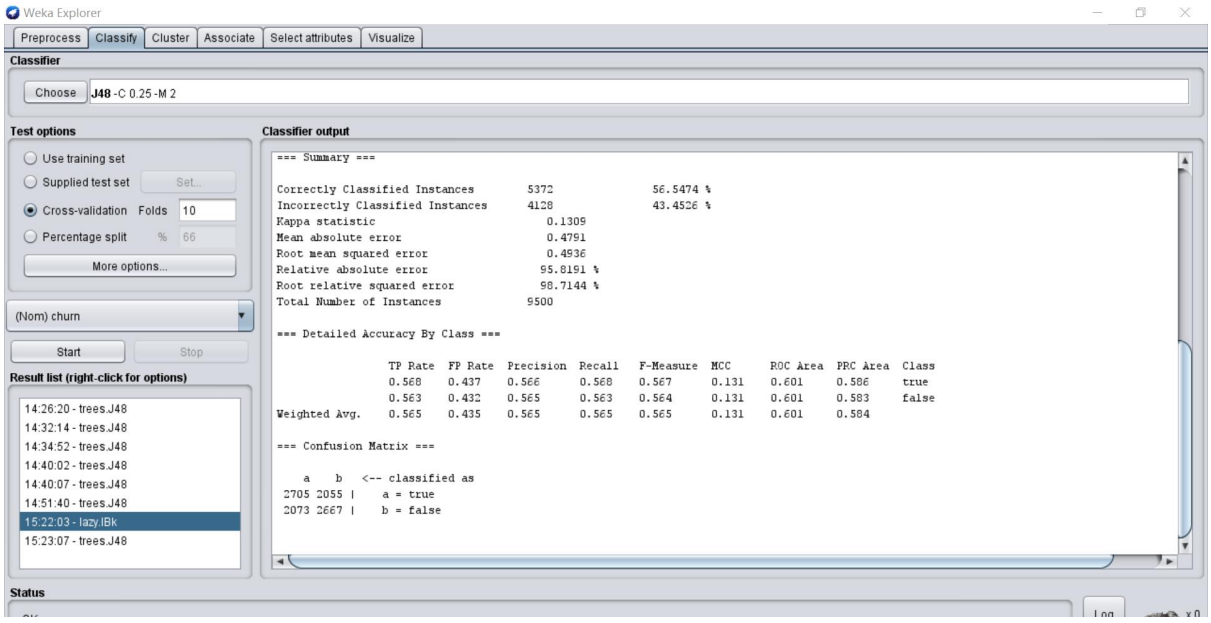
While in terms of wrapper method, it will takes bias of specific learning algorithm into account and considers features in context. However it will cost much running time because it will make a repetitive calls to the classifier, on the other hand it will provide more accuracy.

(c) Evaluate and discuss the performance of both of the above feature selection techniques, when each one is combined with two different classifiers of your choice available in Weka (i.e. there will be four experimental combinations). Which combination do you believe is most suitable for this dataset?

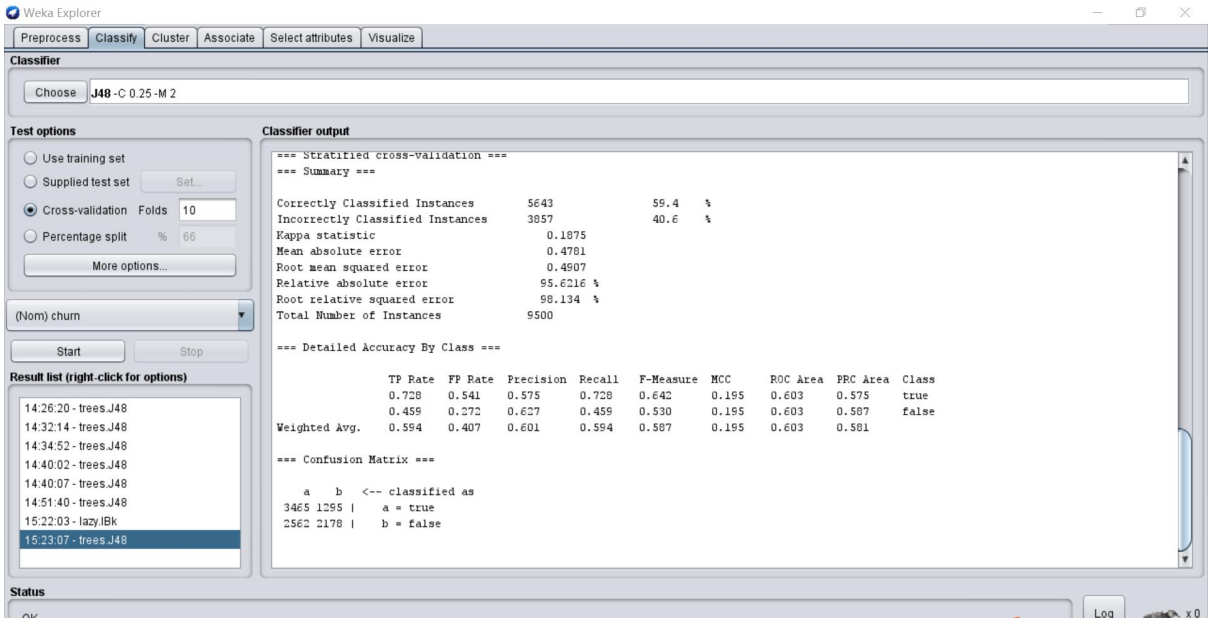
1.filter with top four features(handsetAge, lifeTime, avgMins, avgrecurringCharge) in classifier J48:



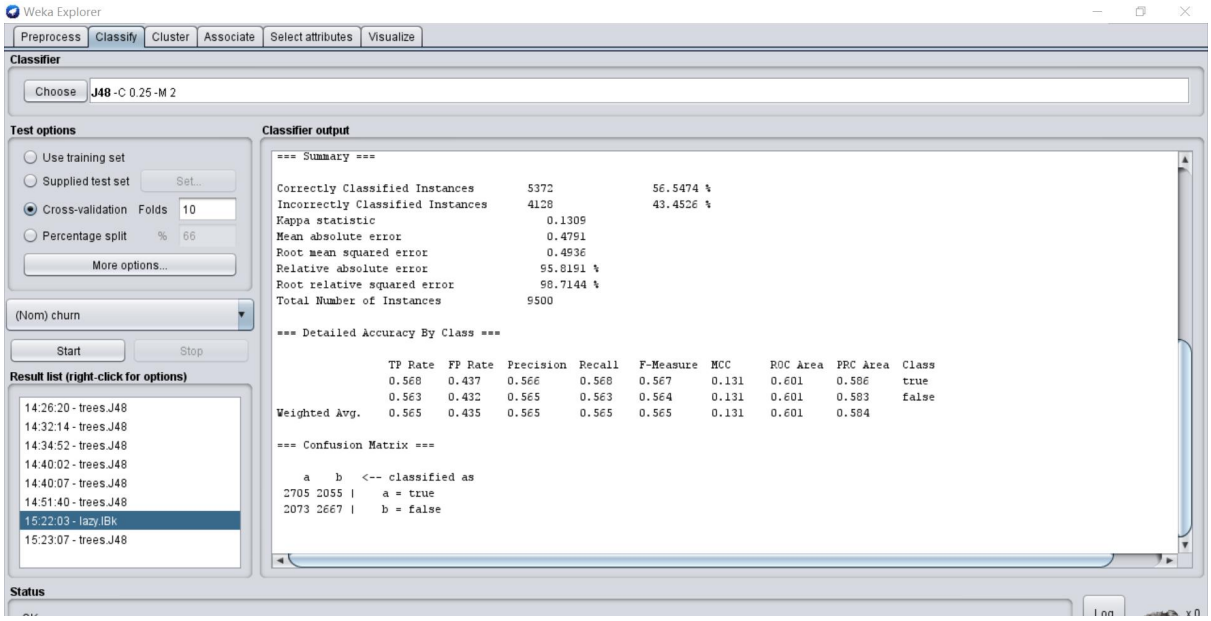
filter with top four features(handsetAge, lifeTime, avgMins, avgrecurringCharge) in classifier IBK(k=3):



2.wrapper using J48 as classifier with option SearchBackwards = False.



wrapper using IBK(k=3) as classifier with option SearchBackwards = False.



In conclusion, the wrapper using J48 as classifier with option SearchBackwards = False can perform the best accuracy.

Question 2

(a) What is a ROC curve and what does it represent?

ROC curve is the abbreviation of receiver operating characteristic curve. It is used to to see how any predictive model can distinguish between the true positives and negatives. It can show the strength of the classifier , the strength of the classifier increases as the ROC curve moves further from the line.

(b) Explain the difference between lazy learning and eager learning approaches in classification. Give an example of a classifier for each approach.

lazy learning approach: Classifier keeps all the training examples for later use and focus on the local space around the examples.

Example: K-Nearest Neighbour, Case -Based Reasoning.

eager learning approach: Classifier builds a full model during an initial training phase, to use later when new query examples arrive and generalise before seeing the query example which is different from lazy learning approach.

Example: Decision Tree, Naive Bayes.

(c) Describe the conditional independence assumption in Naïve Bayes.

Naive Bayes provides a way of calculating the posterior probability, P(c|x), from P(c), P(x), and P(x|c). And the conditional independence assumption in Naive Bayes is that the Naive Bayes classifier assumes that the effect of the value of a predictor (x) on a given class (c) is independent of the values of other predictors.

(d) Explain why classification accuracy may not be a good measure for classification problems with imbalanced data (e.g., fraud detection). Which evaluation measures are better suited for dealing with skewed class sizes?

Imbalanced data often have the features that vast majority of the data is positive, taking fraud detection for example, vast majority of data is legitimate, and only a small part of the data need to be found out. In the circumstance, some evaluation measures can be misleading because high accuracy can be achieved by trivial classifiers which just predict the majority class.

Balance Accuracy Rate and Balance Error Rate are better suited for dealing with skewed class sizes.

(e) Explain why is it not a good idea to select credit card number and name as features to split in a decision tree, even if these features result in the highest information gain?

Even if these features have highest information gain, there still are some problem :

Firstly the tree will result in having to many branches and leaves, and it cannot generalise to customers we haven't seen before if using credit card number and name these unique features .And it will have the problem of overfitting.