CPS844 - Data Mining

**Report**

**Warren Pinto: 500396119**

**Krishna Motepalli: 500346743**

**Dr. Harley**

**April 10, 2014**

**TABLE OF CONTENTS**

Introduction……………………………..……………………………..…………………………**1**

Data Preprocessing……………………………..……………………………..……………….**1**

Data Classifications……………………………..……………………………..……………….**3**

Naive Bayes Classification……………………………..……………………………..……….**3**

J48 Tree Classification……………………………..……………………………..……………**6**

NBTree Classification……………………………..……………………………..……………**10**

Summary of Results……………………………..……………………………..……………..**11**

Summary of Models Used……………………………..……………………………..………**12**

References……………………………..……………………………..………………………..**13**

**Improving the classification accuracy using the Bank Marketing Data Set**

**Introduction**

Data Mining (DM) is an extraction process designed to retrieve useful information often hidden in large data sets. DM techniques are often used in marketing campaigns to promote a business’ success. For this report, various DM techniques were used in a Portuguese marketing campaign to increase their client’s bank deposit subscription. We will try and make an attempt to improve on the results presented by S. Moro, R. Laureano and P.. Cortez, the authors of *Using Data Mining for Bank Direct Marketing: An Application of the CRISP-DM Methodology*

To determine whether a customer will subscribe to the bank’s deposit, the data mining techniques would be more suited for a classification problem. Thus, this report will conduct data mining techniques that use classification models rather than clustering or nearest neighbour models to achieve the desired results.

**Data preprocessing phase**

The dataset that we used was a smaller subset (4521 instances) of the larger dataset (45000 instances). The reason as to why the smaller dataset was chosen is because the larger dataset needed larger computing power in comparison to the smaller dataset. The attributes that were used in the models are shown below:

1. Age (numeric)

2. Job (nominal)

3. Marital (nominal)

4. Education (nominal)

5. Default (nominal)

6. Balance (numeric)

7. Housing (nominal)

8. Loan (nominal)

9. Contact (nominal)

10. Day (numeric)

11. Month (nominal)

12. Duration (numeric)

13. Campaign (numeric)

14. PDays (numeric)

15. Previous (numeric)

16. POutcome (nominal)

17. Y (nominal) **class attribute.**

Attributes 1-8 shows us the general details about the customer, while attributes 9-16 are used as marketing data for each customer (such as when the customer was called or how long was the phone call). The class attribute is Y (yes/no). The instances will be used to create classification models that will eventually predict whether a customer with similar attributes will subscribe to the bank’s deposit.

The bank.csv file was initially in Excel format. To load the data into Weka, the bank.csv file had to be modified using a text editor program, specifically Notepad++. When editing the text in the bank.csv file, all semi-colons in the file were converted to commas. The file was then saved and loaded into Weka.

**Data Classifications**

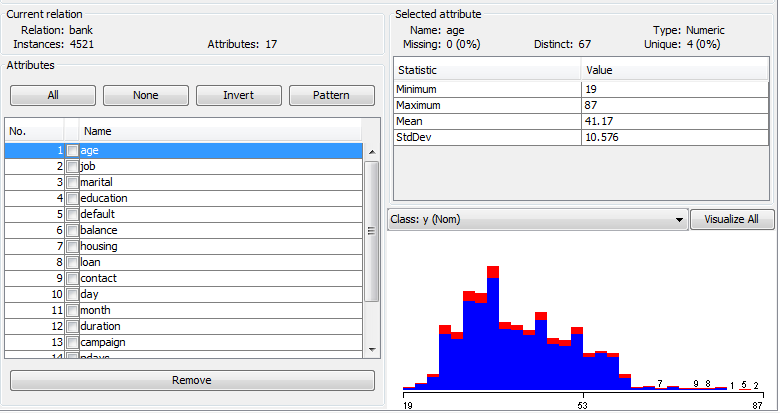
Classification methods in Data Mining are usually used to answer a “yes or no” problem. Because we are dealing with whether a customer would subscribe to the bank’s deposit or not, we use classification methods to derive a model that best predicts the customer’s final decision.

A Decision Tree is a simple structure where the nodes in the trees represent test cases and the leaves of the tree provides the results [1]. Hence, Decision Trees can provide effective classifications in predicting whether the customer will subscribe to the bank’s deposit or not. Probabilistic models also offer good predictions on the outcomes when given data on previous events. Therefore to cover the Decision Trees model and the Probabilistic model, some of the classification methods that we will be using are Naive Bayes, J48 Tree, NBTree and OneR. The test options that were used in the report were 20-fold cross validations, however, we will mostly use “use training set” option for the test option.

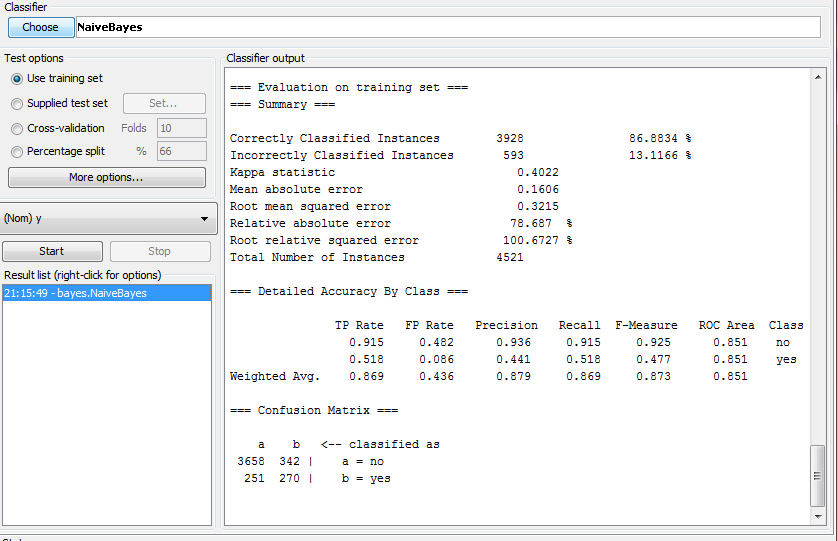
**Naïve Bayes Classification**

The Naïve Bayes classification uses a probabilistic model to classify classes, where the outcome of an event can be predicted based on other events that have been observed. Since we are dealing with a classification problem, the Naïve Bayes classification model could be used to predict the outcome of the client’s decision to subscribe to the bank’s deposit given the clients events/evidence such as age, call duration, balance etc.

To test the Naive Bayes classification model, we load the bank.csv file data into WEKA.

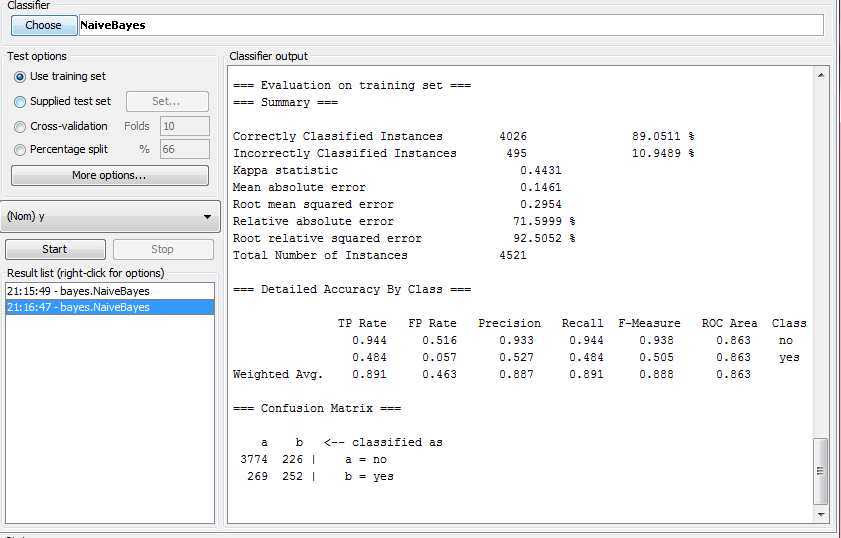


After loading the data, we witch to the Classify tab. We then select Naïve Bayes classifier with “use training set” as our Test Option. The results are computed and shown below.



After computing the Naïve Bayes model, we get 86.8834% accuracy. Although this accuracy is not in the 90s range, it is still a good estimate of the client’s final decision to subscribe. When the final results are being computed, we notice that “duration” attribute has a higher mean than any of the other attributes, making duration the prime factor for increasing the model’s accuracy. We also notice that “campaign” and “previous” attributes have a very low mean score. Therefore, we can ignore those two attributes and it should therefore increase the accuracy.

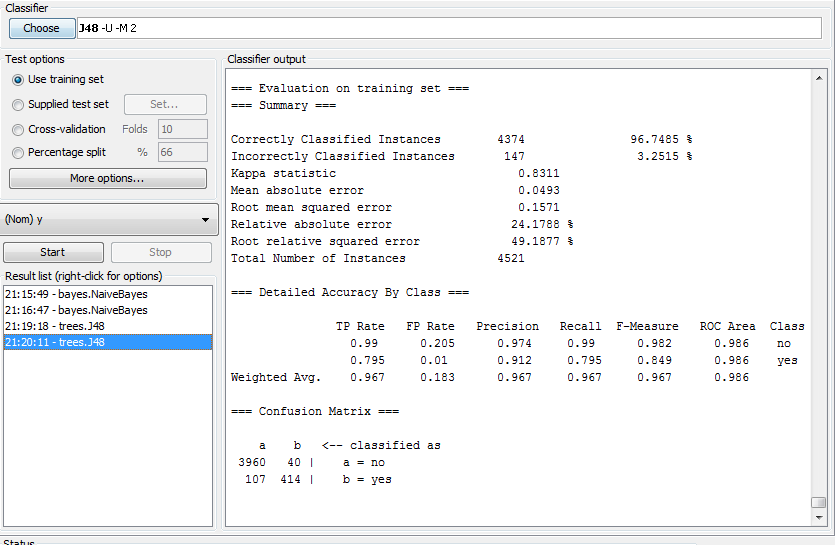
After removing the “campaign” and “previous” attributes, we notice a rise in accuracy from 86.8834% to 89.051%:

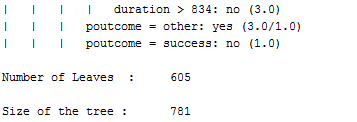


**J48 Tree Classification**

After testing the Probabilistic model using Naive Bayes classification, we will use the J48 classification model to test the accuracy of the data for Decision Trees. The J48 is a decision tree and it is the slightly modified C4.5 tree in WEKA [1]. The J48 Tree model is used to determine the unknown attribute y (yes/no) given a set of known attributes.

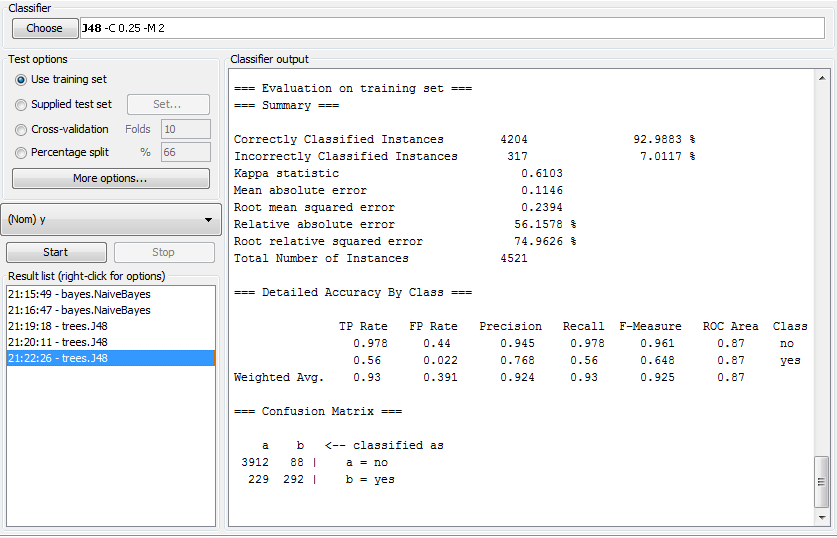
To test the J48 classification model, we re-load the bank.csv file data into WEKA.After loading the data, we switch to the Classify tab. We then select J48 classifier with “use training set” as our Test Option. We also set the J48 classifier to be unpruned.The results are computed and shown below:

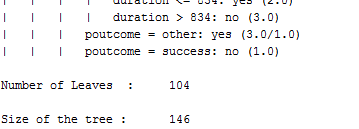




We notice that when we used the j48 classifier method, the model produces 96.7485% correctly classified instances. This models shows to be very accurate in classifying data; however, because this tree is unpruned, there are many nodes onto which a test instance may have to traverse through. Also an unpruned tree may be biased, that is, it may only work for this particular dataset, thereby giving a higher accuracy.

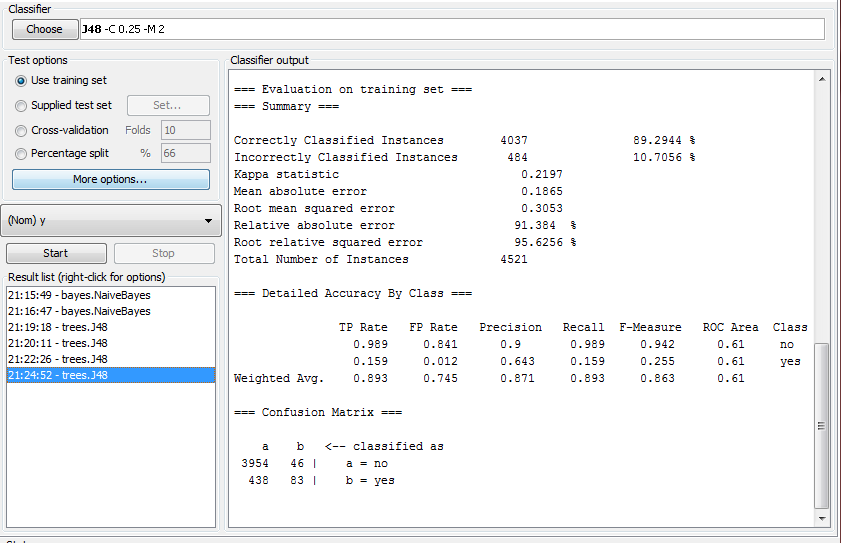
To avoid being biased, we set the J48 decision tree to be pruned, hence eliminating more leaves, making the tree smaller, and introducing more variety. The results are shown below





Analysing this data, it seems that we have dropped in accuracy (92.9883%) when comparing to the unpruned tree, however it is still in the 90s% range, therefore proving to be quite an accurate model that is unbiased. The size of the tree has been significantly reduced from 781 to 146, as well as the number of leaves from 605 to 104.

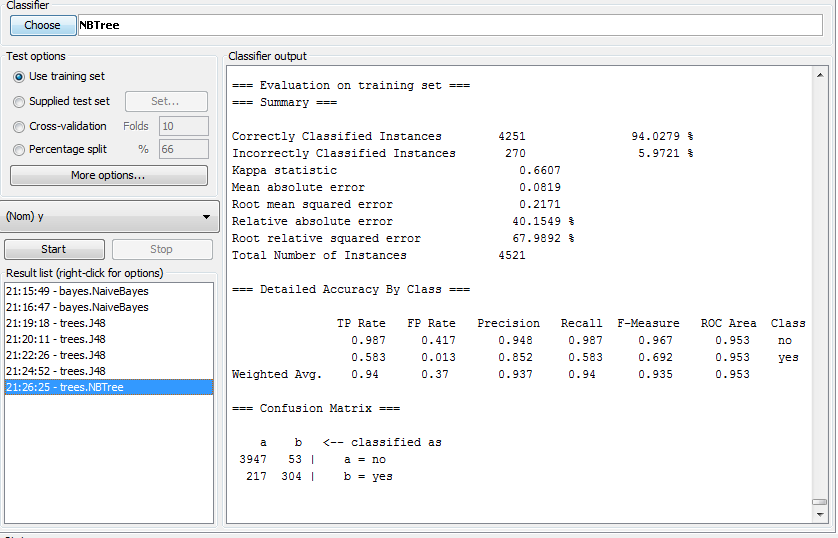
The J48 Tree algorithm determines its first node through the information gain of an attribute. The highest information gain attribute is the root of the tree. In our case, “duration” proved to have the highest information gain. To test whether the duration of the call has a high impact on the model, we removed the “duration” attribute and recomputed the J48 Decision Tree. The results are shown below:



The duration attribute had a high impact on the model as its accuracy has now dropped from 92.99% to 89.29%. This shows that the duration of a call can achieve a higher accuracy of success for the client subscribing to the model.

**NBTree**

Although Naïve Bayes has easy and fast computation, it can generally be slow if we had a large dataset. Naïve Bayes also assumes that the classes are independent of each other and therefore we may have loss of accuracy. NBTree is a combination of the decision tree with Naïve Bayes, where the Naïve Bayes algorithm is computed at the leaves of the decision tree, and a test instance is classified using the local Naïve Bayes on the leaf that the instance falls under [1]. Therefore, the NBTree can be more effective than the Naïve Bayes and the J48 decision tree model.

We are going to test the NBTree classifier onto the bank dataset. The bank.csv file data was reloaded into WEKA. After loading the data, we switch to the classify tab and select NBTree as our model. We also set the testing option as “use training set”. The NBTree is then computed and the results are shown below. 

The NBTree model with training set as our test option gave us 94.0279% correctly classified instances. In comparison with the j48 model, this model is a better classifier where the j48 model only has 92.99% correctly classified instances. This shows that the NBTree model can be effectively used to classify whether a customer would subscribe to the bank’s deposit or not as it is an accurate classifier than all the previous classifiers that we tested.

**Summary of results**

The NBTree model proved to be the most effective in classifying the instances of the Bank Marketing Data Set with an accuracy of 94%. J48 Decision Tree is the next best classification model with an accuracy of 92%. The unpruned and pruned versions of the decision tree had an impact on the J48 model, where the unpruned tree gave higher accuracy but was also biased to the training dataset. We also observed that the duration of the call has a large impact of the classification of the instances where if the duration was removed, accuracy would decrease. Naive Bayes has the least correctly classified instances accuracy, while the decision trees (J48 and NBTree) was much higher. Therefore, for this dataset, we can assume that Decision Trees provide the best accuracy in determining whether the final decision of the customer.

**Summary of classification methods used**

|  |  |  |  |
| --- | --- | --- | --- |
| **CLASSIFIER** | **TEST OPTIONS** | **ATTRIBUTES REMOVED** | **ACCURACY** |
| (Bayes) NaiveBayes | use training set | none | 86.8834 % |
| (Bayes) NaiveBayes | use training set | duration | 83.455 % |
| (Bayes) NaiveBayes | use training set | previous, campaign | 89.0511 % |
| (Trees) J48 | use training set | none | 92.9883% |
| (Trees) J48 (unpruned) | use training set |  | 96.7485% |
| (Trees) J48 (unpruned) | use training set | duration | 89.2944% |
| (Trees) NBTree | use training set | none | 89.0953 % |
| (Rules) OneR | use training set | none | 90.0243 % |

**References**

[1] [Zhao, Y, and Y Zhang. "Comparison Of Decision Tree Methods For Finding Active Objects." *Advances in Space Research* 41.12 (2008): 1955-1959. Print.](http://www.bibme.org/)

[2] S. Moro, R. Laureano and P. Cortez. Using Data Mining for Bank Direct Marketing: An Application of the CRISP-DM Methodology. In P. Novais et al. (Eds.), Proceedings of the European Simulation and Modelling Conference - ESM'2011, pp. 117-121, Guimaraes, Portugal, October, 2011. EUROSIS.