Data Mining – CS F415

Assignment – 2

Questions & Answers

Q1) Do you need data pre-processing? Justify your answer. Explain all the steps you took to pre-process the data. Justify each step.

Ans. Yes

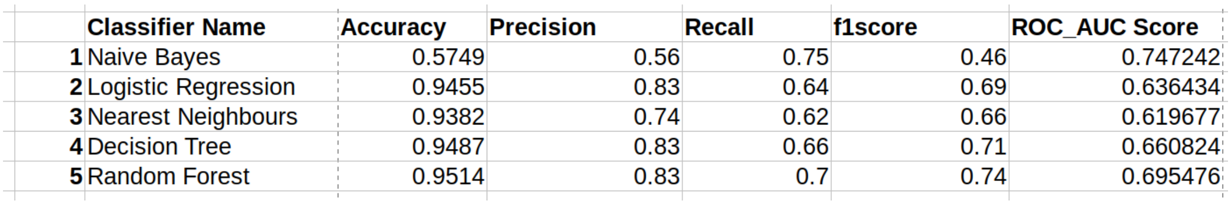
The provided data initially has irregularities in the form of data type mismatch (strings provided instead of numbers), and many missing values hence these was taken care as follows:

1. '?' was replaced by 'NaN'.
2. Many columns have nearly all of the value missing(eg. Enrolled, MLU, etc), so they were dropped
3. Some columns have a large number of unique categorical value, which is not good for OneHotEncoding as it increases the dimension a lot, so they were dropped(eg. Schooling, Married_life, etc)
4. Rows with remaining 'NaN' were dropped.
5. Check for duplicated-data was done, no duplicated data was found.
6. 'Class' and 'ID' attribute is dropped for processing.
7. The categorical value was converted by OneHotEncoding - A new column is added to the dataset for every category and corresponding matching values are set to 1, rest 0.
8. Normalization: The range of many attributes in the dataset is highly distinct, so to prevent from getting skewed results it was normalized using 'MinMaxScaler' class from 'sklearn'.
9. As in 'Classification', distance is a matrix, LabelEncoding can't be used.

Q2)Which classification algorithm yields the best results on the dataset and why?

Ans.

On the preprocessed data, Naive Bayes, Logistic Regression, Nearest Neighbours, Decision Tree, and Random Forest was applied, and the following result was obtained.



	Classifier Name	Accuracy	Precision	Recall	f1score	ROC_AUC Score
1	Naive Bayes	0.5749	0.56	0.75	0.46	0.747242
2	Logistic Regression	0.9455	0.83	0.64	0.69	0.636434
3	Nearest Neighbours	0.9382	0.74	0.62	0.66	0.619677
4	Decision Tree	0.9487	0.83	0.66	0.71	0.660824
5	Random Forest	0.9514	0.83	0.7	0.74	0.695476

Fig. 1 Classifier Comparision

As we can see, Random Forest gives better result according to Accuracy, Precision, and f1score but according to Recall and Roc_auc score Naive Bayes gives a better result. That is why Naive Bayes was chosen.

This is because in case of a skewed dataset i.e dataset with a class imbalance problem, as in our case where most of the data belong to Class '0', accuracy is not a good measure, as the model would predict the value of the majority class for all the predictions and still achieve a very high accuracy score because the number of correct classification would be still high. Thus this is misleading. Hence classification algorithm was chosen according to Recall and Roc_auc score. This is explained further in answer to question 3.

Q3) Which evaluation metric(s) will you use for evaluation of different performance measure of a classification model? Justify your answer.

Ans.

Auc-Roc Score and Recall is used as the evaluation matrix.

AUC(Area Under the Curve) is a metric to calculate the overall performance of a classification model based on area under the ROC(Receiver Operating Characteristic) curve. It shows how good the model is at separating the labels i.e predicting the 1’s as 1’s and the 0’s as 0’s. It basically signifies the relationship between precision and recall and how they vary at different values above which our given data point may be considered in a positive class. It determines a relation between sensitivity(true positive rate)(Fig. 3) and FPR(false positive rate)(Fig. 4). The area under the ROC curve is called AUC(Fig. 2). An AUC value of 1 indicates that the model is perfect. An AUC value of zero tells that it is predicting
the exact opposite of the actual label.
Hence because of the class imbalance problem as stated in answer to question 2 roc_auc score is a better evaluation metric.

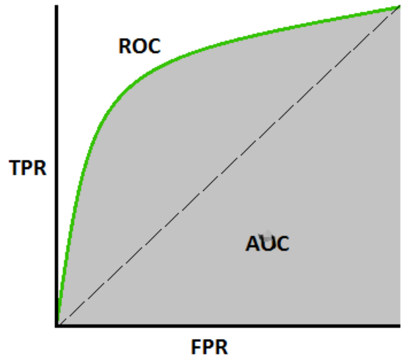
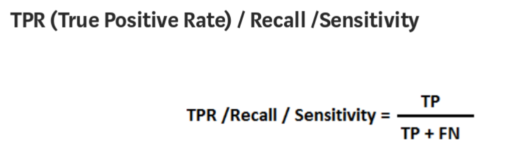


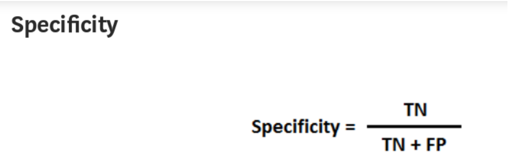
Fig. 2 ROC_AUC_CURVE



TPR (True Positive Rate) / Recall /Sensitivity

$$\text{TPR /Recall / Sensitivity} = \frac{TP}{TP + FN}$$

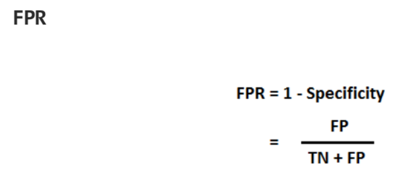
Fig. 3 (Sensitivity Formula)



Specificity

$$\text{Specificity} = \frac{TN}{TN + FP}$$

Fig. 4 (Specivity Formula)



FPR

$$\text{FPR} = 1 - \text{Specificity}$$
$$= \frac{FP}{TN + FP}$$

Fig. 5 (FPR Formula)

Result:

The final score is 75.42% using Naive Bayes Clasification Algorithm.

Submitted By:

Keshav Mittal
2016A7PS0080G