Data classification

1. Introduction

A practical application that we have talked about is determining whether or not to make someone a loan. The goal is to develop predictive models that can determine someone's credit risk, where 0 is high risk, and 1 is low risk. Then loans could potentially be made to just the low risk people.

2. Approaches

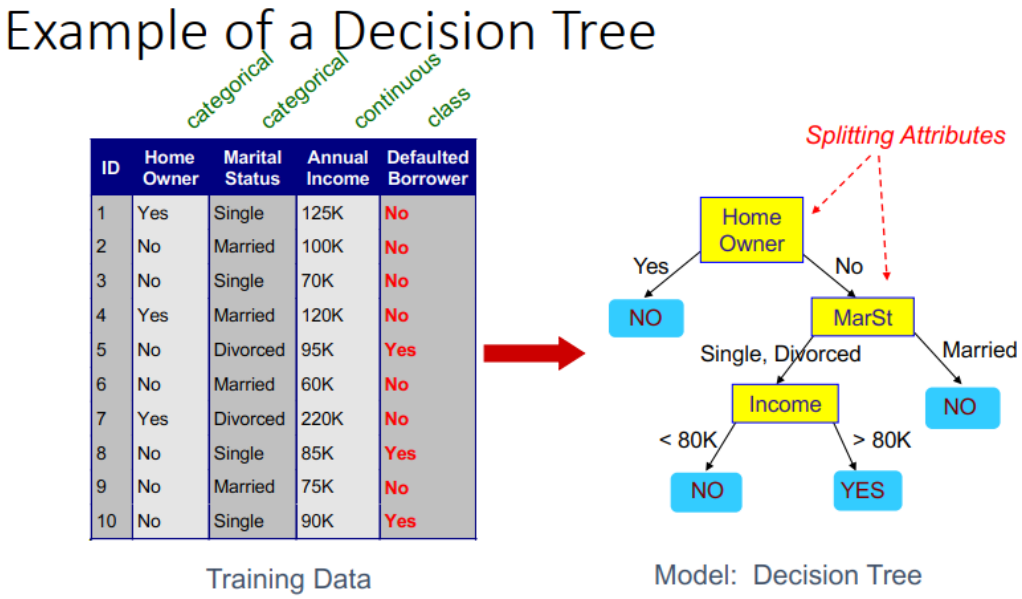

Figure 1. The example of a Decision Tree

The figure 1 shows the example of a Decision Tree.
There are 11 features in train data and test data.
Sklearn package of python is used. By training decision tree is formed and is used to predict credit.

K nearest neighbor is used for performing classification.
To classify the credit, the similarities from train data to test data are calculated and k nearest neighbors are obtained. By k nearest neighbors, the label of credit is obtained.

3. Experimental Results

There are 11 features F1,F2, F3,F4, F5,F6, F7,F8, F9,F10, F11 in dataset.
The strings of train data and test data are converted numbers, for example: Black->0, White->1, Male->0, Female->1.
ID is not considered because it is not feature.

To select best features, the model to select features according to the k highest scores is used. The selected features are F1,F2, F3,F4, F5,F6, F7.
To handle the class imbalance problem, cross validation using grid search is used.
Tuned parameter for decision tree is {'max_depth': 11}.
Tuned parameters for knn are {'n_neighbors': 8, 'weights': 'uniform'}

The figure 2 shows the confusion matrix of validation for decision tree and the figure 2 shows the confusion matrix of validation for knn.
The accuracy of validation can be calculated from confusion matrix.
The accuracy of validation for decision tree =(23547+4687)/(23547+4687+3154+1173)*100=86.7(%)
The accuracy of validation for knn =(23828+4344)/(23828+4344+3497+892)*100=86.5(%)

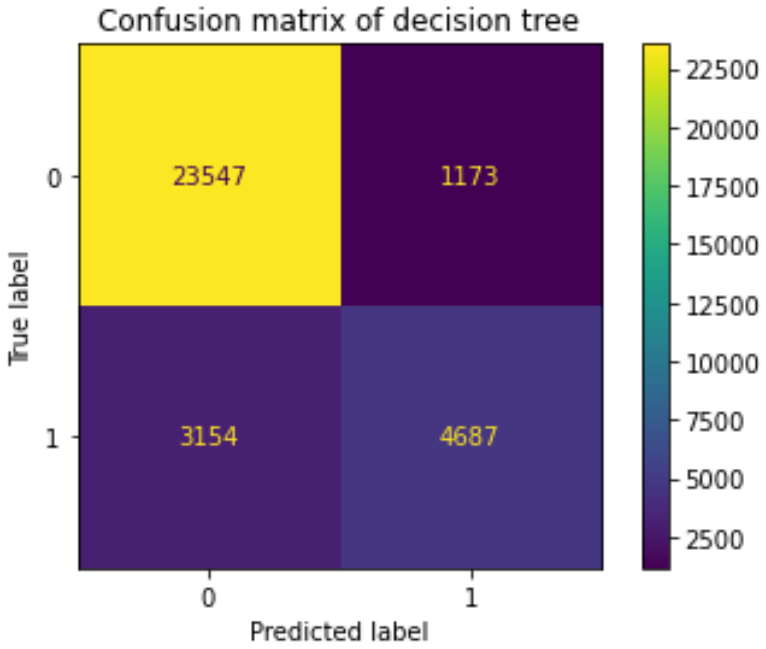

Figure 2. confusion matrix of decision tree

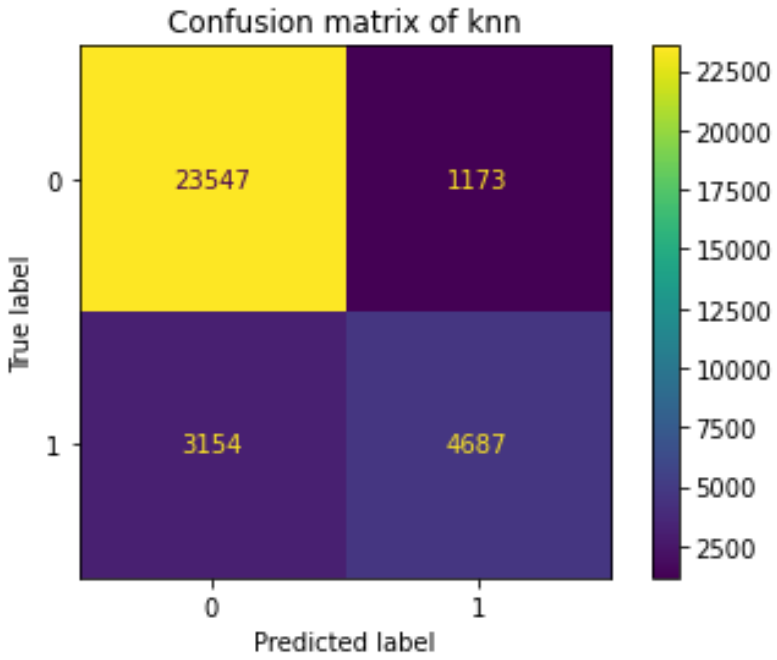

Figure 3. confusion matrix of knn

The accuracy of decision tree for validation data is 86.7% and the accuracy of knn for validation data is 86.5%.
The decision tree is used to predict for test data because it is high.

4. Conclusion

The strings of train data and test data are converted numbers, for example: Black->0, White->1. The features F1,F2, F3,F4, F5,F6, F7 are selected using best selection model. To handle the class imbalance problem, cross validation using grid search is used.
The parameters for decision tree and knn are obtained using grid search

The accuracy of validation for decision tree is high than the accuracy of knn. It is 86.7%.

The prediction label of test data is saved in Format File.txt