COMP 3610 - Big Data Analytics

The University of the West Indies, St. Augustine

Assignment 3

Date Distributed: March 27th, 2020 **Date Due:** April 10th, 2020 (11:59pm)

Introduction

You will use the UCI Credit Approval Dataset where each record is a credit card application. All attribute names and values have been changed to meaningless symbols to maintain confidentiality. The dataset has been cleaned to remove missing attributes. The data is stored in a comma-separated file (csv) here. Each line describes an instance using 16 columns: the first 15 columns represent the attributes of the application, and the last column is the ground truth label for credit card approval.

Part A

You are required to load, explore and clean the dataset. This section includes whether or not you choose to use scaling and PCA. If you use PCA set the variance to 95%. Explain each of your steps and choices using markdown code.

Part B

Section 1

You will perform binary classification on the dataset to determine if a credit card application is safe to approve or not. You are required to use 5 classifiers and compare the results using appropriate graphs and performance metrics. Your classifiers should include the Random Forest Classifier and the KNN classifier. For KNN, use cross-validation to determine an appropriate value for the number of neighbours. Use markdown code to explain your steps, choices and results.

Section 2

In random forests, it is not necessary to perform explicit cross-validation or use a separate test set for performance evaluation. Out-of-bag (OOB) error estimate has shown to be reasonably accurate and unbiased.

Each tree in the forest is constructed using a different bootstrap sample from the original data. Each bootstrap sample is constructed by randomly sampling from the original dataset with replacement (usually, a bootstrap sample has the same size as the original dataset). Statistically, about one-third of the cases are left out of the bootstrap sample and not used in the construction of the kth tree. For each record left out in the construction of the kth tree, it can be assigned a class by the kth tree. As a result, each record will have a "test set" classification by the subset of trees that treat the record as an out-of-bag sample. The majority vote for that record will be its predicted class. The proportion of times that a predicted class is not equal to the true class of a record averaged over all records is the OOB error estimate.

Using the RandomizedSearchCV module provided by the sklearn library

1. Do parameter tuning to obtain the optimal parameters to initialize the RandomForest Classifier. The parameters to tune are as follow:

    i. n_estimators
    ii. max_features
    iii. max_depth
    iv. min_samples_split
    v. min_sample_leaf
    vi. bootstrap

2. Determine the recall score of the classifier

Part C

You are required to perform clustering on the dataset using the KMeans algorithm. Your solution should include steps to find a suitable value for "k", as well as graphs showing the results. Use markdown code to explain your steps and results. The last column of the dataset i not needed in this section.

Submission Details:

1. Ensure that your name and ID and present in the notebook.
2. Export/download your file from Jupyter notebook.
3. Email your file to ssooklal27@gmail.com. Put as the title of the email COMP 3610 A3.
4. There would be a 10% penalty per day for late submissions, up to 5 days.