CS 4622 - Machine Learning
Lab 01 - Report

Name : Jayakody J.A.N.K
Index number : 190253K

Python notebook :
https://colab.research.google.com/drive/1OlE_Go6uT1hdbpOhcCohDBu3HnRyg-fw?usp=sharing

Introduction

The main task of this lab exercise is to practice feature engineering and feature selection techniques such as feature scoring , PCA, etc. The training data set has 28,520 rows and columns with 256 features and 4 target labels. The label 2 had some missing values and label 4 was not equally distributed. For label 2, Xgboost was used and RandomForestClassifier for the remaining labels.

Method

For the data preprocessing process, all the missing values were dropped and features were rescaled using the StandardScaler() of scikit-learn.

Label 01:

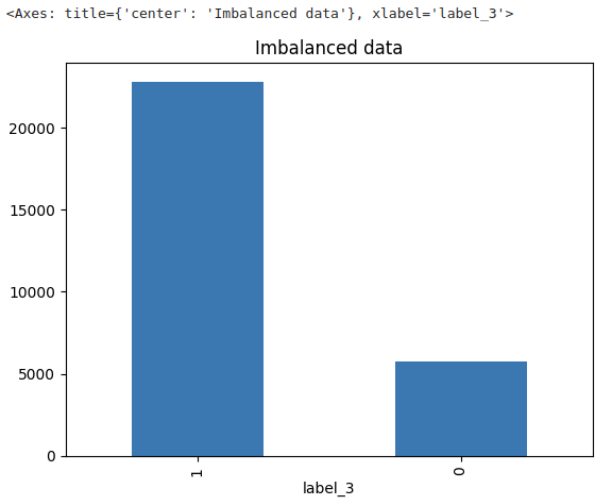
- PCA with 0.97 as the variance was used to reduce the features at the first phase.
- Then feature importance values were extracted from the model and all the features with the score less than 0.009 were removed to reduce the features furthermore. This reduced the whole feature count to 65 without a high loss of accuracy.

Label 02:

- Since this is a prediction, XGBoost was used and PCA was used to reduce the feature count.
- Finally, 31 features were extracted with a mean error of 3.62.

Label 03:

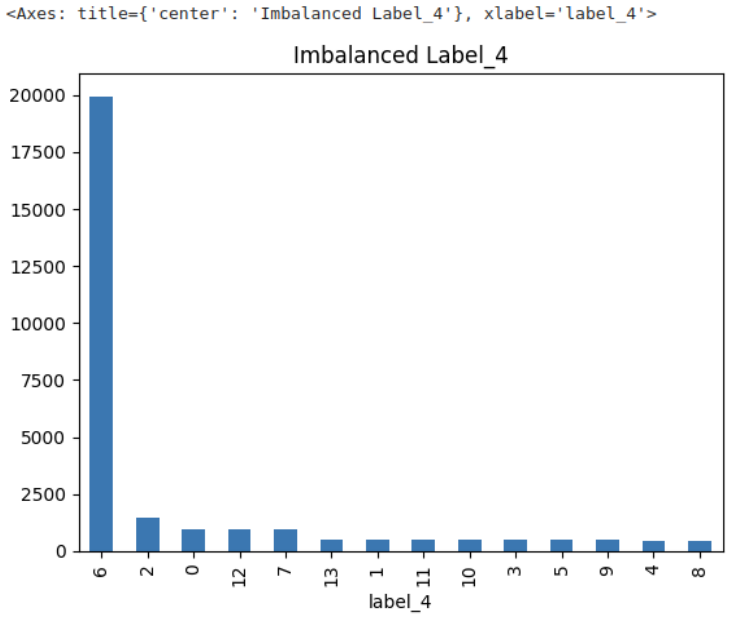
- Since the values in this label were imbalanced as follows,



  Python's imbalanced-learn package was used to resample the data.
- Then PCA was used with variance value, 0.98 and got 87 features left.
- Finally, using 0.008 as the threshold value for feature importances of the resulting features, another 60 features were removed and got 17 features at the end.

Label 04:

- First PCA was used to reduce some features with the variance value of 0.97.
- Since this label was also imbalanced as follows,



The data was resampled and then the features were eliminated further using feature importance values.

- At the end of the process, 22 features were found remaining.