Emre Doğan

Proposed to: Hande Alemdar

CENG790, Assignment 3

[Date]

TECHNICAL REPORT

This is the technical report of 3rd Assignment of the course, CENG790 Big Data Analytics. In the assignment, a random forest classifier is implemented for the credit risk prediction task. There are 20 different attributes and one binary output label for creditability (-1 for creditable, 0 for not creditable).

1. After reading data into a dataframe, I used VectorAssembler to combine all attributes into a single vector column. To do this, first I created a vectorAssembler object so that I can introduce column names.

val columnNames = new VectorAssembler().setInputCols(Array("balance", "duration", "history", "purpose", "amount", "savings", "employment", "instPercent", "sexMarried", "guarantors", "residenceDuration", "assets", "age", "concCredit", "apartment","credits", "occupation", "dependents", "hasPhone", "foreign")).setOutputCol("features")

Then, with a simple transform method, I applied this transformation on my data.

val assembledFeatures = columnNames.transform(creditDF)

Assembled features can be seen in Figure 1 below. 

Figure 1. Assembled Features.