Spark + Python: DO Big Data Analytics & ML

Course Apply Project: Solution Hints

This document contains a proposed solution guide for the problems provided in the course APPLY project. This is just a guide. You are free try out and solve the problem in your own way. We recommend you do the following steps

1. Data Cleansing and Augmentation

In this part, you will clean and prepare the data for further analysis

1. Load the csv file into a data frame
2. Remove the header lines
3. The CSV file has junk characters in some rows. Remove them
4. The CSV file has double quotes around certain values. Remove them.
5. Write a conversion function that would convert this text RDD into a Row RDD of transformed data. Perform the following changes / transformations for the data
    a. Create a new age variable where the age is rounded off to 10s. The age would be 10, 20, 30 etc. Required for PR#06
    b. The Sex column contains both numeric (1, 2) and text representations (M, F). Normalize them to 1 and 2.
    c. Compute average Billed amount (optional). These are things you try out additionally.
    d. Compute average Pay amount (optional)
    e. Compute average Pay duration. Make sure the values are positive. The dataset has a lot of negative values. This is required for PR#04
    f. Compute Average Percentage paid as (average billed amount / average paid amount). This is to pursue a hypothesis that there is a possibility that this value might be able to predict defaulters. A low percentage paid “may” resulting in high defaulting. This is where you get creative with the solution. Feel free to try other ones too.
6. Add a new column SEXNAME that contains Male and Female as values. Create a Data frame with those IDs and values and then join them with the main data frame. Required for PR#02
7. Add a new column ED_STR that contains an actual string for education. Create a Data frame with those IDs and values and then join them with the main data frame. Required for PR#03

8. Add a new column MARR_DESC that contains a description for marital status. Create a Data frame with those IDs and values and then join them with the main data frame. Required for PR#04

2. Perform Analysis

1. Load the Data frame as a temp table /view
2. Query the temp table to solve PR#02
3. Query the temp table to solve PR#03
4. Query the temp table to solve PR#04
5. Perform correlation analysis

3. Predict Defaulters (PR#05)

1. Prepare the data in the standard manner for machine learning
    a. Convert to labeled point
    b. Add indexing
    c. Split into training and test data sets.
2. Run classification using 3 algorithms – namely Decision trees, Random Forests and Naïve Bayes. Find out which one gives the most accuracy on the test dataset.

4. Group Data based on Attributes (PR#06).

1. Create a filtered dataset with only the attributes required for grouping.
2. Perform centering and scaling on all the values
3. Use KMeans clustering to group the data into 4 clusters.

Compare your output with the solution provided. It is not necessary to match fully with the provided solution. It's just a guide.