Big Data Analytics Assessment 2

Overview

Write Python code (especially PySpark where possible) to implement the tasks below (Please explain your code in detail in your report). Where appropriate, the output of code execution should be presented that relates to your answer to the tasks into your report, as the evidence of working program (e.g., screenshots). Output included without supporting explanation or interpretation will not receive credit.

Where the tasks involve big data analytics with machine learning, you may need to formulate technical solution step by step, e.g., choose appropriate machine learning models / algorithms, justifying appropriateness of models/algorithms/techniques used, applying them in the context of the given task, and practising data visualisation techniques where appropriate. Test your solution and conduct experiments where appropriate. Evaluate the performance of implemented solution and analyse results where appropriate. Delve into deep technical explanations of the results and suggest possible improvement, etc. Make conclusions where appropriate.

You need to present your solution to the tasks into a technical report. The report should be submitted onto Blackboard.

Assignment tasks

Part (I)

You will need to download the data file "dataset1.csv" from the Blackboard. It is a synthetic dataset which contains anonymized features '*X1', 'X2', ..., 'X9'* and a binary target variable '*Y1'* (with label '1' or '2'). Please note there are null or missing values in some columns of the dataset.

1. Load the data file into a Spark DataFrame (1st DataFrame). Describe the structure of the DataFrame. (3 marks)

2. Create a new DataFrame (2nd DataFrame) by removing all the rows with null/missing values in the 1st DataFrame and calculate the number of rows removed. (3 marks)

3. Calculate summary statistics of the '*X1*' feature in the 2nd DataFrame, including its min value, max value, mean value, median value, variance and standard deviation. Generate a histogram for the '*X1'* feature and describe the distribution of the feature. (3 marks)

4. Display the quartile info of the '*X2'* feature in the 2nd DataFrame. Generate a boxplot for the '*X2*' feature and discuss the distribution of the feature based on the boxplot. (3 marks)

5. Use Spark DataFrame API (i.e., expression methods) to count the number of rows where '*X1*' is greater than 50 and '*Y1*' equals 1. (3 marks)

6. Use the '*Y1*' feature in the 2nd DataFrame as the target label, to build two classification models based on all other columns as predictors. Conduct performance evaluation for the two models and make conclusions. (15 marks)

Part (II)

You will need to download the data file "dataset2.csv" from the Blackboard. It is a synthetic dataset which contains anonymized features '*X1', 'X2', ..., 'X10'*.

1. Load the data file into a Spark DataFrame (1st DataFrame). Describe the structure of the created data frame. (3 marks)

2. Create a new DataFrame (2nd DataFrame) by removing the '*X10'* column. (3 marks)

3. Use a graph, explore and describe the relationship between '*X2*' feature and '*X8*' feature in the 2nd DataFrame. (3 marks)

4. Use Spark SQL query to display the '*X2*' and '*X8*' columns in the 2nd DataFrame where '*X2*' is great than 1.0 and '*X8*' is greater than 70. (3 marks)

5. Build a linear regression model to predict the 'X8' column in the 2nd DataFrame using the '*X2*' column as the predictor. Conduct performance evaluation for the model and make conclusions. (9 marks)

6. Build a Lasso regression model to predict the '*X8*' column in the 2nd DataFrame using all other columns as the predictor. Conduct performance evaluation for the model and make conclusions. (9 marks)