Big Data Analytics Techniques and Applications

Homework IV

Due Date: 2020/12/22 23:59:59

Goal

Analyze the Airline Dataset by using Spark MLlib on Hadoop platform.
You may choose either one language from Java, Scala, Python, or R to implement it.
Build a predictive framework for predicting whether each flight in 2005 will be cancelled or not by using the data from 2000 to 2004 as training data.
Items to be delivered:

1. Show the predictive framework you designed.
   Hint: What features do you extract? What algorithms do you use in the framework?
2. Explain the validation method you use.
   Hint: Leave-one-out, Holdout, k-fold, or other methods?
3. Explain the evaluation metric you use.
   Hint: Don't just show the prediction results, you should show the effectiveness of your framework using like confusion matrix.
4. Show the validation results and give a summary of results.

Dataset

**Airline on-time performance dataset
The datasets are on New E3.

Requirements

1. Submit a zip file named “Hw4_StudentID.zip” that includes the following items:

- Source codes (including code comments)
- A report in PDF or Word file
  - Show the predictive framework, validation method, evaluation metric, and results
  - Program workflow
  - Execution commands
  - Anything else worth mentioning (e.g. insightful observations)