Evaluated Exercise – Part III: Big Data

MPMD3.2 PÜ Advanced Data Mining Techniques, Databases and Big Data - WiSe2022/23

Points in Parts: Databases SQL: 20% // SciKit-Stack/Cloud: 20% // Big Data: 20% / Natural Language Processing: 20% / Presentation: 20%

Instructions

Report

- Please deliver all commands in your documentation (use a word-document and convert it later into a pdf or an html export of your jupyter-Notebook etc.).
- Also add screenshots of the various steps and of results if necessary.
- Add comments

Software / Platform

- Platforms are
    - Databricks Community Edition

Deadline

- Deadline is the day before the presentations at midnight: 18th of December 2022, midnight.
- Please use the drop-off zone in htw eMPMD system (if you experience technical difficulties, inform me!)

Task 3 – Big Data: Apache Spark / Databricks Community Edition

Points: 10%; Dataset: Log File data (see instructions below)

Task 3.1: Spark SQL: Typical Log-File Data

1. **Please download one R software log-file** from RStudio webpage (Daily R softwareDownloads section) (one weekday) of fourth quarter year 2021, unzip! Web: http://cran-logs.rstudio.com/
2. **Upload the data into your DBFS-system** (Databrick Community Edition)
3. **Import the files into Apache Spark jupyter notebook**
   - Set the schema – it should refer to the variable names on the webpage from RStudio (also set the correct storage types, e.g. integer, string etc.)
   - Register the dataset as Spark SQL
4. **Data Understanding**
   - Check the structure of the dataset, e.g. Print the schema; How many cases (rows); print out the first five rows; …
5. **Data Preparation**
   - Please run necessary data preparation steps, e.g. reduce the dataset to the necessary variables;
6. **Analyses:** Count the number of packages
7. **Display the Top-5-versions (e.g., R 4.1.2, 4.1.1) and Top-5-Operation Systems (e.g. Win etc.)**
   - Sort the dataset
   - Extract the Top-5 versions and operation systems (os)
   - Display the distribution using a barchart (use Apache Spark to do that, just play a little bit around), you can also use pySpark etc.

Task 3.2: pySpark; MLlib; Hyperparameter Tuning/Cross Validation: Lending Club Data

Points: 10%; Dataset: Lending Club Dataset

Please run the same task you run into Azure Machine Learning Studio in Apache Spark with pySpark and MLlib – Target variable is interest rate again

1. **Import File into HDFS**
   a. Load the file into DBFS system
2. **Data Understanding**
   a. Inspect every variable with pySpark or SparkSQL
   b. Use appropriate charts to show the distribution
3. **Data Preparation**
   a. Clean variables
   b. Filter the variables, generate new variables, etc.
   c. Build dummies out of the categorical variables
   d. Transform the data into the typical structure needed in Apache Spark MLlib to run analyses (label and features vectors)
   e. Generate the analysis data frame
4. **Split the file into train- and test-datasets**
   a. Split the file into a training-file (70% of cases) and a test-file (30% of cases)
5. **Conduct a Classical Regression Analysis**
   a. Run a classical linear regression on the training data set
   b. Check the model based on the test dataset
   c. Report the results

6. Conduct a Random Forest Model

a. Run a random forest on the training data set
b. Check the model based on the test dataset
c. Report the results
d. Use hyperparameter (plus Cross Validation) tuning and modify three relevant hyperparameter; report the best model