Evaluated Exercise

MPMD3.2 PÜ Advanced Data Mining Techniques, Databases and Big Data - WiSe2021/22

Table of Contents

TABLE OF CONTENTS ... 1

INSTRUCTIONS ... 2

REPORT ... 2
SOFTWARE / PLATFORMS ... 2
DEADLINE ... 2
CONSULTATION HOURS ... 2

TASK 1 – DATABASES / SQL: DATABASES WITH MS AZURE ... 3

TASK 2 – ADVANCED DATA SCIENCE TECHNIQUES / MS AZURE MACHINE LEARNING STUDIO ... 4

RELEVANT VARIABLES IN DATASET ... 4
STEPS ... 5

TASK 3 – BIG DATA: APACHE SPARK / DATABRICKS COMMUNITY EDITION ... 6

TASK 3.1: SPARK SQL: TYPICAL LOG-FILE DATA ... 6
TASK 3.2: PYSPARK; MLLIB; HYPERPARAMETER TUNING/CROSS VALIDATION: LENDING CLUB DATA ... 6

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
- Executive Summary: If an Executive Summary is requested, please note that you should address management: technical details are not important here, focus on results, which are relevant for management.

Software / Platforms

- Platforms are
  - MS Azure: SQL + jupyter Notebooks
  - Google Collab: jupyter Notebooks
  - Databricks Community Edition: pySPark / jupyter Notebooks
- Instead of a platform, you can run all tasks *on-premise* (on your computer):
  - If you want to perform the SQL task on your computer, just download the MS SQL Express version (which is free to use) and MSSQL Management Studio (which is also free).
  - Download Anaconda to run the Advanced Data Science Techniques Tasks
  - Download Apache Spark to run the Big Data Tasks
- Please note: If you want to run the tasks *on-premise*, I cannot offer broad support. Cloud based techniques are selected here, because they are state of the art and broadly used in companies today and technical support often not necessary.

Deadline

- Deadline is the day before the presentations at midnight: 24th of January 2022, midnight.
- Please use the drop-off zone in htw moodle (if you experience technical difficulties, inform me!)

Consultation Hours

- Approx. 1 hour
- Zoom
- 22nd of December 2021, 12 o'clock
- 07th of January 2021, 12 o'clock
- 19th of January 2021, 12 o'clock

Task 1 – Databases / SQL: Databases with MS Azure

Points: 15%; Database: Microsoft SalesLT database

Data Science department wants to increase the sell-figures of the top-selling product. Predict which features in the database are relevant to predict if a customer will buy the product or not.

Your job is to build a dataset for the analyses – *it is not necessary to run the analyses here*. You first must identify the product which is sold most. Then find out if a customer bought the product or not. (Target variable: Customer in database bought the product (yes/no). Then identify relevant features in the dataset and add them. <u>Please add at least five different features (like location etc.).</u>

Use SQL to perform the job and generate a CSV export.

There is no one valid solution – so please add comments why you have chosen an approach and selected a feature or not. It is possible to delete customers which have no valid data etc.

Task 2 – Advanced Data Science Techniques / MS Azure Machine Learning Studio

Points: 30%; Database: Lending Club dataset

You are working as Data Scientist in a project for Lending Club (please check their web page). The company wants to offer potential customers an online tool to predict potential interest rates based on the purpose and other variables:

- [25% of 30%] Your job is to set up a model to look for possible influences on interest rates (variable **int_rate**) and to set up a multivariate model to predict it.
- [5% of 30%] In the last step you must prepare a management presentation with core findings.

Relevant Variables in Dataset

Please note: Target Variable int_rate

Name of the dataset: LoanStats.csv

No.	Variable-Name	Role	Description
1	int_rate	Target variable	Interest Rate on the loan
2	loan_amnt	Feature	The listed amount of the loan applied for by the borrower. If at some point in time, the credit department reduces the loan amount, then it will be reflected in this value.
3	term	Feature	The number of payments on the loan. Values are in months and can be either 36 or 60
4	grade	Feature	Assigned loan grade
5	home_ownership	Feature	The home ownership status provided by the borrower during registration or obtained from the credit report. Our values are: RENT, OWN, MORTGAGE, OTHER
6	annual_inc	Feature	The self-reported annual income provided by the borrower during registration
7	purpose	Feature	A category provided by the borrower for the loan request

Steps

1. **Preliminary Steps**
   - Set up a jupyter notebook in MS Azure
   - Set up the necessary compute instances in MS Azure
   - Upload the dataset in MS Azure (play a little bit around with Azure datastores and read the available help)
   - Clean the dataset
   - Select the variables shown in the table above

2. **Data Understanding**
   - Analyze the variables in dataset you selected (Schema, First rows, Descriptive Statistics / Frequency Tables, (Charts), …

3. **Data Preparation**
   - Missing Values
   - Transformation of all categorical variables
   - Split into Test and Training Dataset
   - …

4. **Modeling**
   - **Model with target variable: Interest rate**
     - Model 1: Multiple Linear Regression
     - Model 2: Hyperparameter Tuning / Grid Search
       - Run at least 10 models with different hyperparameters
       - Identify the most promising model
   - **Evaluate all Model Fits**
     - Core Parameter is Coefficient of Determination R^2
     - Always control for overfitting (just compare training and test datasets and reduce the complexity of the models if necessary)
     - Check the distribution of the error part
   - **Report a final model which fits best to the data (due to R^2 and overfitting)**.

5. **Management Presentation**
   - Present the core finding on a maximum of 5 slides (only for this task!). Summarize core findings. You are addressing General Management!

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

Task 3.1: Spark SQL: Typical Log-File Data

Points: 10%; Data: RStudio LogFiles

1. **Please download one R-package log-file** from RStudio webpage (Daily Package Downloads section) (one weekday) of third 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-10-packages in the week**
   - Sort the dataset
   - Extract the Top-10 packages
   - 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: 25%; Dataset: Lending Club Dataset

Please run the same task you run into Azure Machine Learning Studio in Apache Spark with pySpark and MLlib

1. **Import File into HDFS**
   a. Download the file from Moodle – File: lc.2017q3.EvalExer.csv
   b. Unzip the file
   c. Check the structure and the variables in the file
   d. 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 Decision Tree Regression**
    a. Run a decision tree 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