Evaluated Exercise – Part II: SciKit-Stack

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
    - MS Azure: ML Lab
    - Instead of a platform, you can run all tasks *on-premise* (on your computer): (Alternatively, you can use any other jupyter-notebook system, but I cannot grant any support)

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

Points: 20%; 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 int_rate based on the purpose and other variables:

- 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.
- 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!