Data Science Project Spring 2021

This project is designed as applied research and experimental development of a data science implementation solution. The project activities involve systematic and creative work of finding novel, uncertain, and reproducible results by applying the modern data science technologies, experienced in class, in a business context.
The journey goes through the five stages and milestones, each of which has an objective, tasks, Peergrade deliverables, deadline terms, and brings 20 study points:

Stage 1: Business Case Foundation

1. Collect ideas and define a business or social domain, where data science can bring a value.
2. Search Internet and get inspired by sources of information, related to your ideas.
3. Formulate context, purpose, questions, and hypotheses for a data science experiment.
4. Technology
    a. Select and install software tools and development environments.
    b. Create a Github repository, which will host all project components during all stages of the development and implementation process.
    c. Create and upload a .md file as an initial release of the project, which contains brief annotation of your ideas, in minimum four sentences, telling:
        - what is in the focus of your interest?
        - why is it interesting?
        - which outcome do you expect from your research?
        - who may be a user of the results?

Deadline: 27th April 2021

Stage 2: Business Data Storytelling

Based on the ideas and assumptions defined at the previous stage of your business case analysis, create the first prototype of your solution implementation.

1. Searching Internet and other media, find relevant data sources that can be used in your experiment.
2. Integrate the sources in shared repository by either ETL or ELT process (you can use public software, own code, or integration of tools)
3. Design a data story or data processing scenario (can be done manually on paper, but the use of software platforms is recommended).
    a. decide on data processing parameters and methods
    b. choose data visualisation techniques
    c. create visual representations
    d. create dashboards
    e. create a prototype of data story
4. Export your solution in a file and upload it to your git repository.

Deadline: 4th May 2021

Stage 3: Integrating AI

Upgrade your business data story by implementing AI analysis in it.

1. Select relevant ML methods and development tools.
2. Create the AI module, which includes:
    a. data preparation, which has to ensure that the data is:
        - meaningful – describes relevant and correctly measured features and observations
        - sufficient – describes various cases and feature occurrences, decided by testing
        - shaped – presented in a structure, appropriate for processing by machine learning algorithms
        - cleaned – repaired from missing values and outliners
        - scaled – transform data distributions in comparable scales, when necessary
        - engineered – all features are analysed and the most informative are selected for further processing
    b. training appropriate models that can solve the problem
        - select at least two or more machine learning algorithms to train a model based on a training data set
        - choose between supervised and unsupervised methods, or implement neural networks and deep learning
    c. test the models created by the learners
        - implement the models on the test data sets to predict the output
        - test them on new data, as well
    d. validate the models
        - select appropriate measures for assessing the quality of your models
        - calculate appropriate parameters to prove the validity of the models
    e. iterate the operations listed above at least twice, trying to improve the quality of the models, as much as possible
    f. compare the results of the validation and select the best performing model.
3. Store the trained model in a file for further implementation.
4. Integrate it with the other modules in your data story, as appropriate, in a new AI prototype of your product.
5. Export your solution files to your git repository.

Deadline: 18th May 2021

Stage 4: Immersive Analytics and Visualisation

At this stage you will extend your solution with a proposal for applying innovative methods for immersive analysis and visualisation of business data scenarios.

1. Consider applying 3D visualisation and VR/AR/MR techniques. Draw your ideas on paper or digital surface and add the image to your project assets.
2. Elaborate on the benefits of applying better visualisation techniques for data analytics.

Deadline: 25th May 2021

Stage 5: Exam Project

The final product of your design and development work during the semester is presented and discussed at the exam.

As a pre-requisite, you need to

- have collected 80% of the study points, given for fulfilment of the assignments, listed above, according the curriculum
- have submitted to Wiseflow one page of text in pdf format, on which stay
    - the title of the project
    - the names of the team members developing it
    - extended resume of the project statement and solution
    - the link to the Github repository, where the project solution can be found

The project solution hosted on Github includes problem objectives, motivation, theoretical foundation, argumentation of choices, code, artefacts, outcomes, and implementation instructions, as appropriate.

It is important to enable reproduction of the project functionality and results.

The quality of the solution is assessed according various criteria, such as:

- data science relevance
- compliance with the requirements
- problem statement and work hypotheses correctness
- argumentation of choices
- sufficient documenting of procedures and results
- proper use of resources
- applying data pre-processing, exploration, and integration techniques
- proper implementation of statistics and machine learning methods
- variety and usability of visualisations
- interpretation of results

Delivery date: 7th June 2021

Notes

This is a group project.

The exam, where it is presented, is individual.

The optimal group size is 2-3 students, individual project work is also acceptable.

Peergrade delivery deadlines are fixed, not extendable.

Exam dates: 14th and 15th June 2021

the instructors

{tdi, mrv}@cphbusiness.dk