Group Projects

FH University of Applied Sciences TECHNIKUM WIEN

Lecturer: NA

Authors: Isabel Dregely, Bernhard Knapp, Stefan Lackner

CONTENTS

- **Introduction**
- **Project Requirements**
- **Project Presentation Requirements**
- **Project Notebook Requirements**
- **Project Idea Pitch Requirements**
- **Schedule**

CONTENTS

- **Introduction**
- **Project Requirements**
- **Project Presentation Requirements**
- **Project Notebook Requirements**
- **Project Idea Pitch Requirements**
- **Schedule**

Introduction

The project includes 3 uploads

- **#1, Slides for a project idea pitch**: A preliminary and very short explanation of your project idea. Purpose: obtain feedback and possibly change your project before running into trouble
- **#2, Slides for project presentation:** Needed for in class presentations at the end of the course.
- **#3, Python code** of your analysis (.py): This is the meat of your project. Must include code, comments, short interpretations and results. Make sure the code is structured s.t. a 3rd party can understand it!

CONTENTS

- **Introduction**
- **Project Requirements**
- **Project Presentation Requirements**
- **Project Notebook Requirements**
- **Project Idea Pitch Requirements**
- **Schedule**

FH University of Applied Sciences TECHNIKUM WIEN

Project Requirements

Your project should include the whole (supervised*) ML-workflow, including clean code and slides as a final project report

- Data acquisition
- Data exploration and cleaning
- Data preprocessing
- Model building, including performance assessment and hyperparameter tuning
- Interpretation & critical discussion (including a discussion about deployment)

*you can choose unsupervised settings (e.g. customer segmentation) also!

ML-workflow

Data acquisition

- **Sources:**
  - real data (e.g. from your work)
  - Find Open Datasets and Machine Learning Projects | Kaggle
  - UCI Machine Learning Repository
  - Open Datasets | Microsoft Azure
  - Offene Daten Österreich | data.gv.at
  - … you‘ll find many other sources
- **Remarks**:
  - Please note that **data acquisition takes time** & is a **major part of the ML-workflow**
  - Choose **a dataset you are intereted in!**

ML-workflow

Data acquisition

- **Requirements:**
  - Instances (rows) at least in the 1000s
  - At least ~10 features
  - Mixed features
  - Suitable for classification/regression
- **Remarks:**
  - If you want to **try specially structured data** (e.g. time series, geospatial, graphs, text, images …) please talk to the course lead
  - If you want to do something **unsupervised**, talk to the course lead!
  - Don‘t take a dataset which is too easy, too small, too well known (e.g. iris, …)

ML-workflow

Data exploration and cleaning

- **Requirements:**
  - Perform an **in-depth data exploration** including **numeric summaries** and - most of all - **visualizations** (scatter plots, density plots, …).
- **Remarks:**
  - The **quality of your visualizations will be part of the grade** so please care about colliding annotations, meaningful visibility, …
  - If you are interested in **interactive visualization**, you can look into **Bokeh**
  - In any case, understand exploration/visualization as a **vital part of your analysis** that you can use when presenting results!
  - **Please note: deep insights are possible by „simple“ descriptive analysis!**

ML-workflow

Data preprocessing

- **Requirements:**
  - Try at least one step of data preprocessing (e.g. dimensionality reduction, scaling, encoding, ...)
- **Remarks:**
  - Please note that **you should try**! If the results don‘t help your model present this fact and don‘t use this step for your final model.
  - **The aim is not to do everything, but to try a lot and to deploy the best model!**

ML-workflow

Model building

- **Requirements:**
  - Clean performance assessment and HP-tuning as presented during the course
- **Remarks:**
  - Consider your **computational ressources** when planning your pipelines!
  - If you have a **very large dataset**, consider doing feature engineering, performance assessment and HP-tuning **with a sample**, but train your final model on the whole dataset!

ML-workflow

Interpretation and critical discussion

- **Requirements:**
  - Discuss **model deployment** from a **technical point of view**
  - Try to interpret your model from the **domain expert view**. Is the model good enough? What are possible consequences of deploying the model? What needs to be monitored after deployment?
  - Discuss **model deployment** from a **societal point of view.** What does it mean for society if such a model would be deployed at large scale? Are there any dangers?
- **Remarks:**
  - One important point is if you have any biases in your dataset (e.g. if the data was predominantely white middle age males in it). Take a look at **Open data and data bias | data.europa.eu**

FH University of Applied Sciences TECHNIKUM WIEN

CONTENTS

- **Introduction**
- **Project Requirements**
- **Project Presentation Requirements**
- **Project Notebook Requirements**
- **Project Idea Pitch Requirements**
- **Schedule**

FH University of Applied Sciences TECHNIKUM WIEN

Project Presentation

Requirements

- Slides: 10‘ + 5-10‘ peer discussion

Remarks

- **Don‘t just scroll through your code, we need a presentation**!
- Include **many visualizations and think about how you can make your point visually**! Please avoid long „bulletpoint-keyword-only“ lists. The presentation should be a final report and it should be understandable in a stand-alone fashion
- **Slides as .pdf**

CONTENTS

- **Introduction**
- **Project Requirements**
- **Project Presentation Requirements**
- **Project Code Requirements**
- **Project Idea Pitch Requirements**
- **Schedule**

Project Notebooks

Requirements

- **Python3 Script (.py) – nothing else is accepted!**
- Consider organising you code into several files. **There must be one (and only one) file which clearly documents the analysis workflow in a step-bystep, linear fashion.**
- Your code should be clearly structured, commented and should be readable like a report (e.g. it should incude short conclusions, interpretations, findings).
- Anaconda **environment .ymal file** for reproducibility of your environment!
- Your **project must be reproducible**, meaning that **references to data downloads** and **all preprocessing steps** to „start from zero“ must be included!

FH University of Applied Sciences TECHNIKUM WIEN

CONTENTS

- **Introduction**
- **Project Requirements**
- **Project Presentation Requirements**
- **Project Notebook Requirements**
- **Project Idea Pitch Requirements**
- **Schedule**

FH University of Applied Sciences TECHNIKUM WIEN

Project Idea Pitch

Requirements:

- 4 slides, 4 minutes
- Must include: (1) **The chosen dataset**, (2) A discussion of **features and targets** (3) your Hypothesis – what do you want to predict?, (4) a **preliminary schedule** and work distribution
- **Slides as .pdf**

Remarks:

- Please invest enough time in choosing a dataset - this must already include a first exploration of the data!
- Please clearly state what problem you are tackling (regression, classification, …)

CONTENTS

- **Introduction**
- **Project Requirements**
- **Project Presentation Requirements**
- **Project Notebook Requirements**
- **Project Idea Pitch Requirements**
- **Schedule**

Schedule

Project Idea Pitch

- Upload **first version** until class 5
- Present and discuss your idea during the in-class session.

Project Presentation

- Upload **slides** until class 14 for all groups
- Presentation during classes 14/15

Code Upload

- Upload until the **end of the course** – include changes after feedback

FH University of Applied Sciences TECHNIKUM WIEN