Project Kick-off

Machine Learning for Information Systems Students

02.06.2021

Simon Fuchs – Omar Shouman

Plan

Agenda Item	Slot
General Feedback Homework	**13:00 – 13:15**
Intro to Data Projects	**13:15 – 14:00**
Break	**14:00 – 14:15**
Problem Formulation	**14:15 – 14:45**
Project Groups and Deliverables	**14:45 – 15:00**
Data Overview (Jupyter Notebook)	**15:00 – 15:15**
Q&A	**15:15 – 16:00**

Homework Feedback

- Feedback
  - How hard it was for different backgrounds?
  - How much time spent on average?
- Discussion
  - Ideas
  - Extenstions

Data Projects

- Start from the business outcome
- Figure out stakeholders
- Measure current standing if exists
- Reflect on the value of your solution and where it should bring you

Data Projects

- Formulate your learning problem
  - Inputs
  - Output
  - Optimization objective
  - Evaluation metrics

- Check available data
  - Get to know the data
  - Identify data problems

Data Projects

- List the major milestones you have, try to include checkpoints
  - Will help you as well when splitting the tasks among the team members
- Work iteratively, test small pieces before connecting everything (Data preprocessing, model training, …)
- Build a simple pipeline, test, iterate to extend and improve

Data Projects

- Organize your code
  - Do not put all code in one file
  - Wrap code into functions when used multiple times
  - Make good use of pandas and sklearn built-in features

- Make it easy for people to read and extend your code

Break

Project Idea and Problem Formulation

- Improvement of a Ticketing System using machine learning

- Goals and business outcomes:
  - Faster processing/resolution of tickets
  - Better customer experience and satisfaction
  - Others …

Project Description

- Example problem formulation:
  - Input: Ticket text and meta-data (e.g. logs, status change)
  - Output (your formulation):
    - Classification (support level, product, etc…)
    - Clustering
    - ……
- Think thoroughly how you can use the meta-data available
- Business outcome and proposed solution should be clearly linked

Project Description

- Feel free to use additional python library, also ones for natural language processing
- **Formulation of the problem** and **applying machine learning** are two essential parts that **cannot be missing**
- Scope:
  - Be innovative, extend the scope and the task
  - Explore ideas from literature
  - Explore the dataset well
  - Think of complementary modules and services
  - Think of how to transform your work into a product

Project Groups, Timeline, Deliverables

- Groups of 3-4 students
- Distribute the workload, everyone should contribute

Milestone	Expectations	Date
Project Proposal 10%	- Clear project roadmap and scope - Task distribution among team members - Initial overview of the data - Q&A	15.06.2021
Final Presentations 20%	- Overview of the whole process - Results & discussion - What would you do if you had more time - Lessons learnt (Problems, solutions) - Delivery of the project	13.07.2021

Deliverables (aside from presentations) 10%

1. Git repository with everything ☺ (gitlab or github)
   - If Gitlab, use TUM LRZ, https://gitlab.lrz.de/users/sign_in
   - If Github, make it private until you finish the work (but up to you)

- Requirements for the repo:
  - well-written readme
  - reasonable directory structure
  - Requirements and environment to ease reproducibility
  - modular code design (classes, files, etc…)
  - scripts to run train-test

Think of this as part of building your online profile

Deliverables (aside from presentations) 10%

2. A write-up for the openpower website, a good example is here: https://openpower.ucc.in.tum.de/music-genre-detection-web-system/

- Use material from your proposal, presentation, readme
- Make some good-looking diagrams
- Make it look good and professional, this will be online and you can refernece it later

Deliverables (aside from presentations)

3. That is it :)
   No more

Evaluation

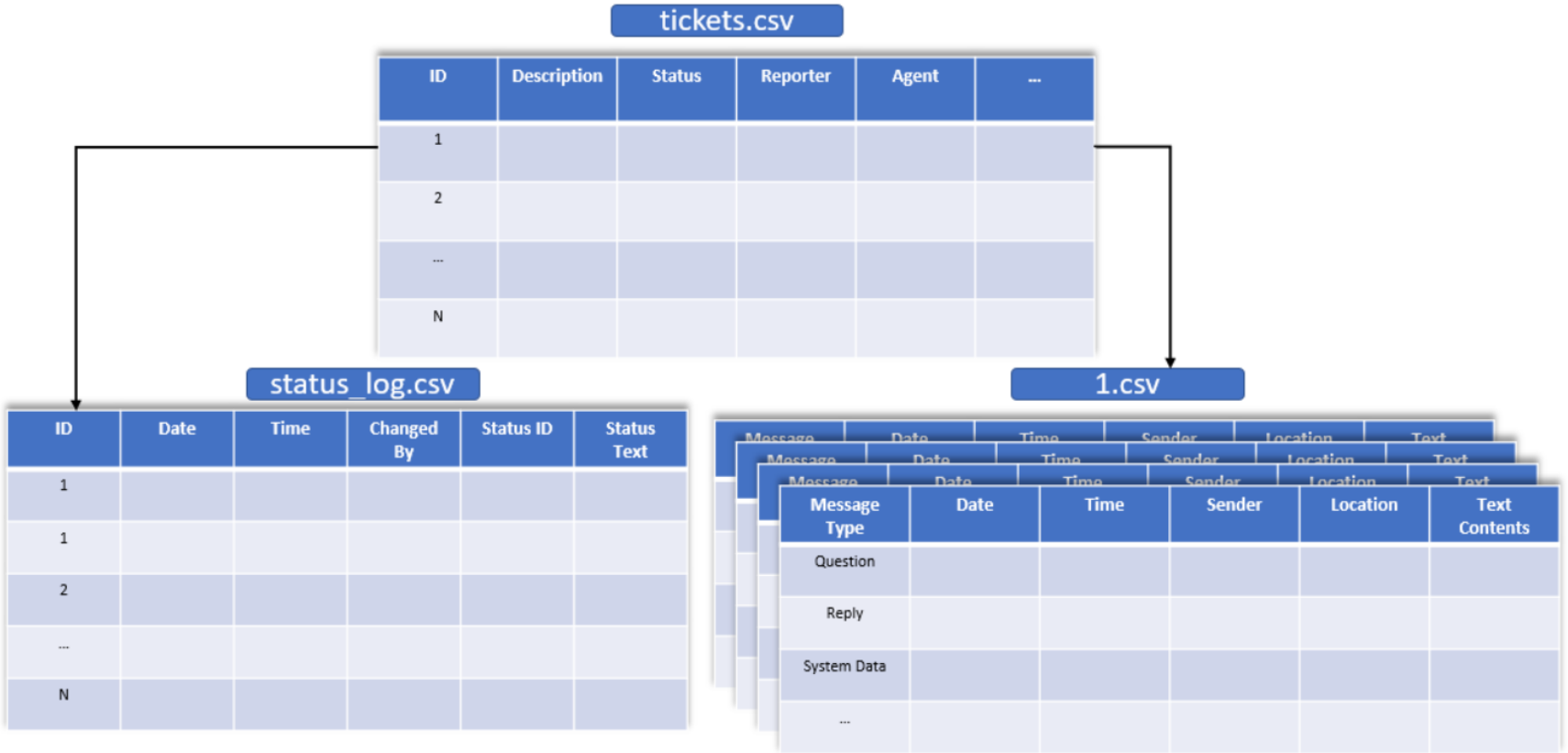
- General focus of our evaluation would be on:
  - the effort you exerted in the work
  - the quality of your work (not performance measures)

- Performance and code evaluation
  - Clear business outcome
  - Correct ML setup
  - Implementation
  - Reproducibility
  - Conceptual Errors

Support from Our Side

- Weekly office hours for follow-up and questions → annouced via Moodle
- Please always reach out if you think we can help
- If you happen to need resources, we can organize it
- If you have more points/ideas, please discuss with us

Data Overview



Credit goes to **Valeryia Andraichuk**

Data Overview

5.2_Project_data_overview.ipynb

Recommended Starting Points

- Literature Review from Simon (attached in Moodle). Please do not share it outside the course.

- Scikit-learn Text tutorial: https://scikit-learn.org/stable/tutorial/text_analytics/working_with_text_data.html

- Machine learning NLP libraries:
  - SpaCy: https://spacy.io/
  - NLTK: https://www.nltk.org/

- Introductory NLP courses
  - Courses 1 and 2 from: https://www.coursera.org/specializations/natural-language-processing

- Building Demo or WebApp for your Machine learning project:
  - Streamlit: https://streamlit.io/
  - Gradio: https://www.gradio.app/

References

- Géron, A. (2019). Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems. O'Reilly Media.
- Provost, F., & Fawcett, T. (2013). Data Science for Business: What you need to know about data mining and data-analytic thinking. " O'Reilly Media, Inc.".
- VanderPlas, J. (2016). *Python data science handbook: Essential tools for working with data*. " O'Reilly Media, Inc.".