German University in Cairo
Faculty of Media Engineering and Technology
Mervat AbuElkheir

CSEN1095 Data Engineering

Winter Term 2019

Course Project – Engineer Ahead!

For your final assignment in this course you will work on a month-long data science project. The goal of the project is to go through the complete data engineering process to answer questions you have about one **topic of your own choosing**. Topics and associated datasets are:

- The Android App Market on Google Play (potential dataset: https://www.datacamp.com/projects/619)
- Visual History of Nobel Prize Winners (potential dataset: https://www.datacamp.com/projects/441)
- European Soccer Database (potential dataset: https://www.kaggle.com/hugomathien/soccer)

You will acquire the data, design your visualizations, run exploratory and statistical analysis, preprocess the data as needed, and communicate the results. You can augment the data with data from external sources if you want to enrich your project.

Project Dates

A rough timeline of your project is:

- *Week of Nov 2 - 9*: Group creation and project theme selection
- *Week of Nov 16 - 21*: Data collection and cleaning finished
- *Week of Nov 23 - 28*: Exploratory analysis finished, some visualizations, start notebook and GitHub page
- *Week of Nov 30 - Dec 5*: pipeline with basic prediction finished, Github page/notebook/visualizations finalized

The **Hard Deadlines** for your projects are:

- *December 7*: Final project submission due (11:59 pm)
- *December 9*: Presentations shown in class, location/exact times to be updated.

Any changes that you make to your GitHub repositories and webpages **after the due date will be ignored**. Please have all your work submitted and tested (websites, screencasts, etc.) before the deadline.

Project Team

You will work closely with other classmates in a **3-5 person project team**. You can come up with your own teams and register the information on this link. If you can't find partners by the deadline I will team you up randomly after the group creation. If individual schedules or other constraints might limit your ability to work in a team, ask me for permission to work alone. In general, I do not anticipate that the grades for each group member will be different. However, I reserve the right to assign different grades to each group member based on peer assessments (see below).

Project Milestones

It is critical to note that **no extensions will be given** for any of the project submission due date declared above. Presentation sessions may be rescheduled in the revision week, subject to your own loads. Projects submitted after the final due date will not be graded. If you anticipate any issues (e.g., due to travel plans) you need to send me an email at least one week in advance.

There are several deliverables for your project that will be graded individually to make up your final project score:

Deliverable 1: Jupyter Notebook

An important part of your project is your Jupyter notebook. Your notebook details your steps in developing your solution, including how you collected the data, alternative solutions you tried, describing statistical methods you used, and the insights you got. Equally important to your final results is how you got there! Your notebook is the place you describe and document the space of possibilities you explored at each step of your project. I strongly advise you to include many visualizations in your notebook. Your notebook should include the following topics. Depending on your project type the amount of discussion you devote to each of them will vary:

- *Overview and Motivation:* Provide an overview of the project goals and the motivation for it.
- *Related Work:* Anything that inspired you, such as a paper, a web site, or something we discussed in class. You must properly credit sources.
- *Initial Questions:* What questions are you trying to answer? How did these questions evolve over the course of the project? What new questions did you consider in the course of your analysis?
- *Data:* Source, scraping method, cleanup, storage, etc.
- *Exploratory Data Analysis:* What visualizations did you use to look at your data in different ways? What are the different statistical methods you considered? Justify the decisions you made, and show any major changes to your ideas. How did you reach these conclusions?
- *Final Analysis:* What did you learn about the data? How did you answer the questions? How can you justify your answers?
- *Presentation:* Summarize your results, the strengths and short-comings of your results, and speculate on how you might address these short-comings if given more time. Present your final results in a compelling and engaging way using text, visualizations, images, and videos on your project web site.

Describe the storytelling elements and goals in your process notebook and show sketches and screenshots of different work iterations. As this will be your only chance to describe your project in detail make sure that your notebook is a standalone document that fully describes your process and results.

Deliverable 2: Code

You are expected to write high-quality and readable Python code in your notebook. You should strive for doing things the right way and think about aspects such as reusability, error handling, etc. You are also expected to document your code.

Deliverable 3: Project Webpage or Github Page

You will create a public website for your project using Github Pages or any other web hosting service of your choice. The web site should effectively summarize the main results of your project and tell a story. Consider your audience (the site is public) and keep the level of discussion at the appropriate level. Your Jupyter notebook and data should be linked to the page as well, either using a zip file, GitHub, bitbucket, Dropbox, Drive, or another code hosting site. Also embed your main visualizations and your presentation in your website.

Deliverable 4: Project Presentation

Each team will create a **five minute** presentation with narration showing a demo of you Jupyter notebook. Embed your presentation into your project page. Use principles of good presentations to get your key points across. Focus the majority of your presentation on your main contributions rather than on technical details. What do you feel is the best part of your project? What insights did you gain? What is the single most important thing you would like your audience to take away? Make sure it is up front and center rather than at the end.

Peer Assessment

It is important to provide positive feedback to people who truly worked hard for the good of the team and to also make suggestions to those you perceived not to be working as effectively on team tasks. I ask you to provide an honest assessment of the contributions of the members of your team, including yourself. The feedback you provide should reflect your judgment of each team member's:

- *Preparation*: were they prepared during team meetings?
- *Contribution*: did they contribute productively to the team discussion and work?
- *Respect for others' ideas*: did they encourage others to contribute their ideas?
- *Flexibility*: were they flexible when disagreements occurred?

Your teammate's assessment of your contributions and the accuracy of your self-assessment will be considered as part of your overall project score.

Submission Instructions

Submission will be handled through GitHub. All team members must use a single shared GitHub repository. *If I cannot access your work because these directions are not followed correctly, I will not grade your work.* You will need to specify your project GitHub URL in the team registration form after you create the page (*Week of Nov 23 - 28*). Store the following in your GitHub repository:

- Jupyter Notebook - Your project notebook.
- Data - Include all the data that you used in your project. If the data is too large for GitHub, store it on a cloud storage provider, such as Dropbox, and link to project page.
- README - The README file must give an overview of what you are handing in: your project notebook, any non-standard Python libraries you used, and so on. **The README must contain URLs to your project websites and presentation/demos**.

Projects submitted after the final due date will not be graded. Late days do not apply. After your team submits the project, you will fill out a peer evaluation (5 points), which will be published on Google Forms on Dec

8th at midnight. The evaluation will be short and must be completed by Dec 9th by midnight. Again, late days do not apply.

Grading Criteria

The project is worth 20% of your total course grade. These 20% will be divided as follows:

- *Project Scope* (10%) - Did you choose the appropriate complexity and level of difficulty of your project?
- *Notebook* (40%) - Did you follow the data science process and is it well documented in your notebook?
- *Solution* (10%) - Is your analysis effective and correct in answering your intended questions?
- *Implementation* (10%) - What is the quality of your code? Is it appropriately polished, robust, and reliable?
- *Presentation* (20%) – Is your presentation clear, engaging, and effective?
- *Peer Evaluations* (10%) - Your individual project score will also be influenced by your peer evaluations. Distribution is ideally 2%-4%-2%-2% for each of the four elements mentioned in the peer assessment.