ISM 5136: Data Analytics and Mining for Business

Project Guidelines

Summary:

The purpose of this project is to demonstrate the data mining and programming knowledge you acquired through the semester using data science competitions. Specifically, you will identify and attend a competition currently on-going on the website: kaggle.com. Kaggle is subsidiary of Google and it is a platform for providing datasets and running competitions sponsored by real-world corporations.

Competition Information:

- In general, there are 10 – 15 competitions that are active on Kaggle at any given time. A majority of these competitions tend to be supervised machine learning / classification type. You are free to choose any one among these competitions. However, when making your selection, pay attention to data characteristics and the difficulty of the competitions.
    - Certain data structures such as video, images, audio, etc. are out of context for our course and might be too difficult to handle.
    - Conversely, there are newcomer level competitions that are pinned and solely for introductory purposes (e.g., Titanic). By default, these competitions are too trivial and I would expect you to include additional analyses, if you pick any of these.
- Read the competition instructions carefully. In general, these competitions will ask you to upload a dataset that includes your predictions to a target attribute. This prediction dataset will be provided by the data sponsoring company. Competitors are ranked by one or more performance evaluation metrics (accuracy, precision, recall, etc.) and this information is provided in the Leaderboard section. You are free to submit as many entries as you'd like to the real competition. For the purposes of this project, only a single submission is sufficient.
- A different type of competition is kernel-type, where each entry to the competition is the entire code script.
- Note that making a submission to a competition is required, but not sufficient for this project. You will also need to submit a "kernel" – i.e., a thoroughly annotated and documented code script as part of your deliverable for the project. Please see below for additional information.

Teams: 1 to 2 people. You can work on this project by yourself or with another team member. If working as a team, only one submission is sufficient; however, make sure to include both names in the deliverables.

Deliverables: There are three required deliverables for the project:

1. A single Jupyter notebook (published as a Secret Gist on Github) and the Gist address. This notebook is your project "kernel". Make sure that the cells have been run and the outputs are visible before you submit this document.
2. Confirmation that you have participated in the competition (i.e., you have submitted and entry to an on-going competition). Confirmation can be of any type – an e-mail, a screenshot, etc.
3. The dataset that you have used for the project (collected from Kaggle).

Structure of the Kernel: The kernel should provide your entire Python code script that you developed for this project. It should also include definitions and discussions for various items (see below). These can be entered as markdown cells or comments in the script. A typical kernel should include the following sections:

1. Introduction (in text form)
    - A brief summary of the competition and the problem
    - A brief summary of the data
    - Your objective and the data mining approach that you followed (in a few sentences).
2. Exploratory Analysis Stage (code + output + text)
    - Summary statistics
    - Visualizations
    - Brief discussion on interesting findings
3. Data Preparation and Pre-processing Stage (code and text)
4. Modeling Stage (code and text)
5. Performance Evaluation Stage (code + output + text)
    - Performance results
    - Visualizations, if applicable

Additional Notes:

- You can find many examples of high-quality kernels developed for various competitions / projects on the Kaggle website. It is acceptable to look at these kernels for inspiration for your analysis. However, you are <u>restricted</u> from copying / using any code from these kernels. In other terms, all of your code needs to be developed by yourself. I will run a plagiarism checker on your kernel to compare it with the ones available on the Kaggle website.