Assignment 2: Start-up succeed or fail?

Assigned: 15/11/2021; **Due: 24/12/2021**

The main focus of INT303, the class, is to give you fundamental knowledge of big data such that you can tackle a variety of situations yourself, but you shouldn't always need to reinvent the wheel from the basics when others have been perfecting the wheel you need potentially for years or decades.

Goals

- Programming language Python and its libraries NumPy (to perform matrix operations) and SciKit-Learn (to apply machine learning algorithms)
- Practice summarizing a potential complex topic into usable information, distilling it down to the important points.
- Determining which modern big data libraries and tools are available for their project goals.
- Several machine learning algorithms (decision tree, random forests, extra trees, linear regression).
- Feature Engineering techniques.

Problem

The goal of the project was to predict if a start-up will succeed or fail based on a set of data. We used Kaggle competition "Can you predict if a start-up will succeed or fail? " (see https://www.kaggle.com/c/int303-big-data-analysis/data) to retrieve necessary data and evaluate accuracy of our predictions. The historical data has been split into two groups, a 'training set' and a 'test set'. For the training set, we are provided with the outcome (whether or not a start-up survived). We used this set to build our model to generate predictions for the test set. For each start-up in the test set, we had to predict whether or not a startup is 'acquired' by some other organization, means the startup succeed. Our score was the percentage of correctly predictions.

Competition Entrance

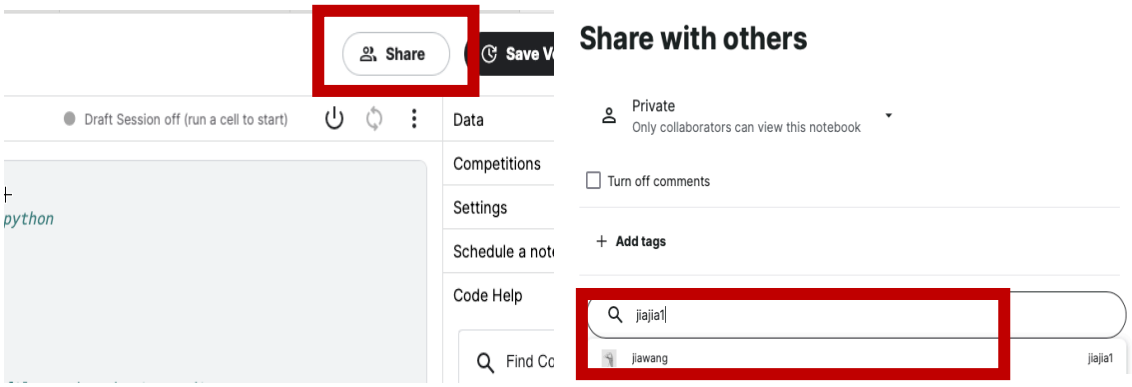
https://www.kaggle.com/c/int303-big-data-analysis/overview

Tasks 1 (30 Marks)

1. Create an account on https://www.kaggle.com/. You MUST create an ID with the format in (Name_ID). Example: JiaWANG_09211013.

2. Create a notebook on Kaggle and submit your code there. (20 Marks)

   - Name your notebook with Name_student number.
   - Share your notebook only with Dr. Jia WANG (Her ID is: jiajia1).



3. Submit your predictions for the test solution to Kaggle. Also, you are required to include your Kaggle scores in your report (see below in Task2). (20 Marks)

Tasks 2 (70 Marks)

Write a 1-page report, which **must contain** 2 or 3 number of tables or figures. The report must cover:

- **Introduction:** (*12* Marks)
  Why should we care about this technology? How is it related to Big Data?
- **Approach:** (*12* Marks)
  How does it work? Explain the algorithm or framework.
- **Results:** (*12* Marks)
  Are there benchmarks for its use? How does it compare to similar technology?
- **Pro-Cons:** (*12* Marks)
  What are the good aspects, what are the bad aspects? Be sure to add a sentence on "**contributor thoughts:**" What are your own unique thoughts on the

pros and cons of the technology? Do you envision an extension that might be helpful?

- **Conclusion:** **(***12* Marks)
  Summarize the 2 to 4 points you think are most important

Concise, information rich content. For each of the sections above you will not simply be graded on having content but on the quality of the content and how well it answers the questions in concise, clear, and engaging terms.

Style. (10 Marks)
In order to make your report consistent and visually appealing, as well as to make evaluation of your work fairly, each page was conformed to the following specifications:

- Margins: approx. 0.5" on all 4 sides.
- Columns: 2 with approx. 0.3in margin; justified text
- Fonts:
  - Body text: Times New Roman, 11pt.
  - Section headings: Calibri 13pt bold-Italic
  - Within captions, tables, figures, or images: Calibri 9 - 11pt.
- Line Spacing:
  - Body text: Single (1.0)
  - Section headings: 6pt spacing above heading

Academic Honesty. Copying chunks of code or problem-solving answers from other students, online or other resources is prohibited. You are responsible for both (1) not copying others work, and (2) making sure your work is not accessible to others. Assignments will be extensively checked for copying of others' work. Problem solving solutions are expected to be original using concepts discussed in the book, class, or supplemental materials but not using any direct code or answers. Please see the syllabus for additional policies.