ITCS498: Special Topics in Data Sciences
Data Science Midterm Exam: Start-up
Update: 11-March-2021

Allowed materials

- Open-book exam.
- Students can talk to each other, use any devices, and use the Internet during the exam.
    - If you use the code from the Internet, you **MUST** give the reference of where you get the code from.
- Note: We encourage you to discuss with your friends regarding techniques that you have tried. However, you must write code and answer by yourself.
- Copy your friends' answers will result in severe punishments.

Submission

- Part 1 (Morning Session @ 9:00 - 12:00)
    - Upload a Jupyter notebook to MyCourse (per group)
- Part 2 (Afternoon Session @ 13:00 - 16:00)
    - Upload a Jupyter notebook to MyCourse (per individual)

Grading Criteria: The grade will be based on the completeness of your work. The correctness of your flow, the selected models, the results, and the conclusion.

Part 1 (40%): “Analyzing the data is hard… really?”

Img Ref: Startup Jobs Asia, http://blog.startupjobs.asia/8-business-insights-can-learn-korean-drama-start/#sthash.GXPvQEUX.z2hdy6Sv.dpbs

“Running a business is hard, right? If you succeed, you are called a CEO. If you fail, you are called a fraud.” - Yoon Seon-hak (Start-up)

Introduction

From Lecture 1 to Lecture 4, we discussed problem formulation, data understanding and data preparation. Specifically, we talked about how to select the feature space that represents the problem. In a few case studies, the data from the insurance, retail business, etc. were analyzed and showed how to use exploratory analysis to understand the data. This information was used to create a visualization, prediction model, and clustering model. The insights gained from models help us better understand the problem. For this midterm, you are going to work on “start-up analytics”.

For this problem, you can think of yourself as a venture capital. You are analyzing the current start-up business to see where you should invest. The ultimate goal is to invest in a start-up that will be successful. The current data that your underlinks found in public is the list of start-up companies, and several pieces of information. The data is given the file called “startup-train.csv”. It contains 8 features and 1 target feature listed below.

1. `sub_category`: The subcategory of the product of this startup.
2. `category`: The main category of the product of this startup.
3. `deadline`: The startup should receive sufficient funds before this date.
4. `launched`: The startup was launched on this date.
5. `backers`: The number of venture capitalists which have pledged their money in this startup.
6. `country`: The country that the startup company is located in.
7. `pledged`: This is the amount of money given to the startup by the supporter. (We will pledge our money to successful startup companies.)
8. `goal`: This is the amount of money needed by the startup to launch their product.
9. `target`: There are four possible values: canceled, failed, suspended, and successful. This represents the current status of the startup. If it is successful, it implies that the startup launched its product successfully and maintains its growth. If it is considered as failed, the startup is out of business after the required money is met. If it is canceled, the startup decides to give up before money is met. If it is suspended, the startup violates at least one of the market rules and is forcefully suspended from the regulator.

For part 1, you should apply all your knowledge, all your tools, and all your skills to solve the following problem. As a group, you are to perform the exploratory analytics on the data. You are to prepare and clean the data as you see fit. Specifically, you should try to find insight on what is the key indicator for a successful startup, what makes the startup fail, etc. In addition, you should prepare the data for the classification task. Feel free to create your own feature based on the given data.

Additional Info:

1) You should apply the data science process to clean the data, understand the data, and find the insight from the current data.
2) At each step, you should document your work in Jupyter notebook. For example, if you want to encode the text information in the input data, explain why you want to encode it, follow by what your code is doing, and show the results using head() function. Or if you want to see the correlation between attributes, explain what test you are doing and why you choose it, follow by your parameter setting and your code, and finally discuss the results of the correlation scores. If you want to show the joint distribution of variables, you should plot them, and discuss the results.
3) Finally, you should write the conclusion, summarize your findings?

What to hand in for the Morning:

You are to submit your work on mycourse website. You should submit a single Jupyter notebook. At the beginning of the file, you should list all your names and student ids. Note that everyone in the group should have a copy of this file for the afternoon session.

Material:

For this exam, you may use the Internet, reuse your codes, and reuse the codes provided to you in this class. You may use any lecture notes, or textbooks. In addition, you may use the codes available online, but you must give the reference of where you get the code from. You may talk to your friends, but you may not use or copy your friends' work. You may not copy their codes. You may not copy their discussion.

Grading Criteria:

The grade will be based on the completeness of your work. The correctness of your flow, the selected models, the results, and the conclusion.