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 2 (60%): “Data is sw… really?”

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

“Why look for answers that don’t exist? So, instead of looking for answers, make choices.” - Han Ji-Pyeong, (Start-up)

Introduction

For part 2, you are given “startup-test.csv”. You are to perform the prediction on this dataset. There are two classification tasks that you have to perform this afternoon.

Task 1: You should use the “startup-train.csv” to train the prediction model for multiclass classification problem: “Given the data point of 8 features, predict the target of this data point, whether it should be a “successful”, “canceled”, “failed”, or “suspended” startup”. Once you find your “best” model for this multiclass classification problem, you use the model to make the prediction on “startup-test.csv”. Then, append the prediction results to your original data frame, in column called “pred-1” as shown in Figure 1. Again, you should explain all the features you use for this task. How and Why you select your “best” model for the prediction task. Again, for pred-1, the column can only contain four possible values as mentioned above. Anything other than this will be considered as a not-available.

Task 2: For this task, you are to define your own classification problem. Specifically, you should think about how to solve the original problem “which startup should I, as the venture capital, invest my money in” as the classification task. You may reduce the number of possible values in the target to that of the binary if you think it fits better for the problem. For whatever you pick, you explain why you want to solve the problem this way. What features are you using and why are you using them. Explain the best model that you train for your part 2 problem. Append the prediction results to your original data frame, in column called “pred-2” as shown in Figure 1. Finally, save the dataframe into a file called “prediction.csv”. Note also that, for pred-2, you may have different names for the prediction if you have different labels. Since it is an open-ended problem, your prediction may be 0, 1, 2 for a three-class classification problem. Kindly, explain what 0, 1, 2 implies.

Note that you MUST not shuffle the order of the data so that the original order of data in startup-test.csv is preserved in prediction.csv.

	A	B	C	D	E	F	G	H	I	J
1	sub_category	category	deadline	launched	backers	country	pledged	goal	pred-1	pred-2
2	Music	Music	7/16/10 0:00	6/11/10 5:43	0	US	0	300	canceled	0
3	Painting	Art	5/29/17 0:00	4/29/17 16:41	1	US	25	500	suspended	2
4	Gadgets	Technology	7/1/14 0:00	6/3/14 6:16	2	US	155	386000	canceled	1
5	Experimental	Film & Video	2/23/15 0:00	1/24/15 17:01	1	US	20	500	suspended	1
6	Musical	Theater	12/31/15 0:00	12/11/15 6:08	14	AU	489.66	4385	canceled	1
7	Indie Rock	Music	3/24/14 0:00	2/22/14 17:30	56	US	2206	8000	successful	0
8	Indie Rock	Music	6/1/11 0:00	4/17/11 16:24	48	US	2966	2000	successful	0
9	Ready-to-wear	Fashion	3/5/15 0:00	2/3/15 17:44	2	DK	66.21	11770.07	successful	1

Figure 1: Example of the final file that you should save as the output of your notebook.

You may use the data preparation method in the morning as the starter. You can also create additional features. You may use any technique that you think will be useful. You must do your work in Jupyter Notebook.

Additional Info:

1) You should apply the data science process for two predictive model tasks. You may use the data preprocessing from your group in the morning for this part.
2) Similarly, at each step, you should document your work in Jupyter notebook. For example, if you want to use a decision tree, explain for what purpose, follow by your parameter setting and your code, and finally discuss the results of the tree (model accuracy, number of tree depths, etc.)
3) Finally, you should write your conclusion, whether it is possible to solve this prediction task. What are the key features of the models?

What to hand in for the Afternoon:

You are to submit your individual work on mycourse website. You should submit a single Jupyter notebook. Note that your Jupyter notebook will create one csv file called `predict.csv` with the column names shown in Figure 1. Again, pred-1 and

pred-2 are the results of the predictions for problem part 1 and problem part 2 respectively.

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:

Your scores will come from both the individual work (60%). The grade will be based on the completeness of your work. The correctness of your flow, the selected models, the results, and the conclusion. Finally, the label