**INFO 7390 Projects**

**Project selection guidelines and evaluation criteria**

**Project theme/topic selection deadline: Feb 20, 2022**

**Project change deadline: Feb 24, 2022**

**Project submission deadline: May 01, 2022**

**Maximum marks: 50**

1. Team size. A team can have a minimum of two persons (due to the sheer size of the class) and a maximum of four people. **Teaming up is highly recommended. Exceptions :** Reasons and background skills for doing a one-person project need to be discussed.
2. Suggested topics. These are suggested themes and are not in any manner binding ¾

(a) Use a Kaggle dataset. You could either come up with your own prediction goals or use Kaggle’s problem definition. Advantages of picking a Kaggle dataset include well curated, large datasets, problem is pre-defined and most importantly, you can see where your model performance stands against the leaderboard!

(b) You can gather and work on datasets from other sources like Google bigquery.

(c) You can work on your own problem in which case you need to look for the data yourselves. Maybe the dataset is lying hidden in some corner of the internet or, think scraping it off webpages!

(d ) OR, there are instructor led project datasets available. If you need them, please let me know !It will not affect their grade.

3. Project submission details. Please submit:

(a) code inside of a Jupyter notebook, that goes all the way from reading in the data to printing out model performance visualizations. Please make sure to submit commented code,

(b) a short write-up explaining the goals, and importance of the project and,

(c) Slide deck with not more than 6-7 slides summarizing the project, packages used, goal, etc.

The goal of the project is making sure you tried your very best in building a state-of-the-art predictive model while adhering to acceptable, god data science practices. A minimum of two different ML algorithms needs to be evaluated. Hyperparameter tuning differences DO NOT count as separate models. Data or feature Engineering also does not count as a separate algorithm. Please make sure you provide a detailed model comparison visual as well in your Jupyter notebook.

Point breakdown

Selection of performance metric of the algorithm, giving detailed reason = 5 points

Each algorithm with hyper-parameter tuning = 10 points each (10 times 3 = 30 total)

At least one comprehensive results visualization = 10 points, with 2 points for the legends, 2 points for clear labelling, 5 points for the right choice of data plot. We trust your creativity :

Data visualization WILL NOT be explicitly covered in this course and *you will need to come up with your own ones.*

A final short write-up (problem statement, importance and solution, all in a paragraph or two) and slides (the same section but with nice visuals) will carry 5 points

Code poorly commented will lose 5 marks. Comments are not a big stretch, is it ? Just put them in. This is to help yourself and the TA and me who will have to evaluate the cool projects.

Please submit code that runs from top to bottom of the Jupyter notebook when we click Run all. If the code fails and throws up an error, it will incur a penalty of 5 points (sorry).

Please submit on time because late submissions without previous explanation to me will accrue 5 point penalty for each day. The penalty is there to ensure fairness to everyone.

Good luck :-), we are here to help you do your best !