2021AI511ML Project Rules

Teams

You will be expected to form a team of 2 students. There is no restriction on the pairings -- you may choose any teammate as you wish.

In extremely rare cases (e.g one person is unable to find a partner) we will allow a team of 3 but we will expect more work from them. We expect there (in the ideal case) to be not more than one team of 3 in each project. **You need to clear this with Prof Dinesh first.**

Project Format and Deliverables

There will be nine projects given as 2-month-long Kaggle contests (from Sep 10 to Dec 7). You may choose exactly one project to work on. The selection of projects is first-come-first-serve. There will be **20** slots available for each of the ten projects and once these slots are filled with teams no other team can choose this project. The logistics of this will be communicated with you separately.

The project will be evaluated based on:

1. Kaggle metrics
2. Final report and code
3. Intermediate evaluations and Slack discussions

There will be one TA allotted for each project and they will be in charge of evaluation for that particular project. **Which TA is responsible for which project will only be communicated to you after the project selection deadline.**

Intermediate Evaluations

There will be three intermediate evaluations -- the schedule is given below.

These will be in person.

1. Present the work done so far using slides **while mentioning the contribution of each teammate.**
2. Answer questions that are given by the TAs.
3. Tell us what your next steps are.

In addition, in the final meeting, **you will be expected to answer questions given by other teams. Marks will be awarded to the teams asking questions as well.**

Allowed software, models, and external data

Neural networks are allowed -- but only multilayer perceptrons. LSTMs, RNNs, transformers, CNNs, etc. are not allowed.

External pre-trained models are completely banned -- SOTA language models like BERT and GPT are also banned.

External datasets other than the given Kaggle dataset may not be used whatsoever -- however, word embeddings like GLoVe, Word2Vec, etc will be allowed.

You are allowed to use your own laptop, Kaggle Notebooks, Colab Notebooks. For the sake of fairness, no other platforms will be allowed.

Plagiarism note

1. Discussion with other teams of the same project is allowed and encouraged **provided you give due attribution to the same in the report and intermediate evaluations.**
2. Discussion with the TAs is allowed and encouraged.
3. **Discussion of the specific problem statement with other students of IIITB/IBAB or other communities outside is forbidden.** Asking general ML doubts online is allowed and encouraged as always.

   An example which is allowed:

   “ I’m dealing with a large dataset and I’m unable to efficiently apply XYZ operations on the Pandas data frame. How do I do this? “

   An example that violates academic honesty:

   “ I’m working on XYZ problem. Which features are most useful? Which model gave you the highest results?”

Summary of timeline

Project list announced: 19 Sep

Project and team selection deadline: 21 Sep

Competitions launched: 21 Sep

First intermediate eval: Week of Oct 7

Second intermediate eval: Week of Nov 4

Third intermediate eval: Week of Dec 2

Kaggle competition ends: Week of Dec 2

Report and code submission: Week of Dec 7

Project Descriptions

1: Yahoo Troll Questions Detection:

As a last-ditch attempt to give Yahoo Answers some legitimacy in the era of Quora and Reddit, Yahoo's CEO Merissa Meyer has tasked you with creating a machine learning model to detect spam and troll questions so that they can be removed. To this end, you are given a hand-labeled dataset consisting of questions with unique IDs and whether they are troll questions or not.

Tags: NLP

2. Home Loan Default Risk Prediction:

Many people struggle to get loans due to insufficient or non-existent credit histories. And, unfortunately, this population is often taken advantage of by untrustworthy lenders.

Home Credit strives to broaden financial inclusion for the unbanked population by providing a positive and safe borrowing experience. In order to make sure this underserved population has a positive loan experience, Home Credit makes use of a variety of alternative data--including telco and transactional information--to predict their clients' repayment abilities.

Tags: Tabular Data

3: Aid Escalating Internet Coverage

Think of yourself as a counsellor for an online marketing firm. The agency invests a lot of time and resources in researching the top websites to post its ads. They choose websites that would attract steady online traffic so that their adverts can be seen for a long time. Wouldn't it be fantastic if you could automate this procedure and conserve resources for the business? The agency has generated a dataset of raw HTML, metadata, and a binary label for each webpage to help with this. The binary label indicates whether or not the website was chosen for ad placement. In order to determine which online pages are worthy of posing an advertisement on, this assignment aims to select the pertinent, high-quality websites from a pool of user-curated web pages. Building large-scale, end-to-end machine learning models that can categorise websites as "relevant" or "irrelevant" based on factors like the alchemy category and its score, meta-data about the web pages, and a one-line summary of each page's content is required for the challenge. By having you translate the dataset's textual properties into some kind of numerical data and then build your machine learning models using this numerical data, this job intends to acquaint you with the field of NLP.

Tags: Tabular Data, [slightly] Big Data

4: SnapIt - Always get the best price!

SnapIt is an online reselling platform where local sellers can list their products for sale and obtain the right amount of money from items they might not need anymore.

You are the new employee hired by SnapIt to solve their main issue- deciding the selling price of their products. Most sellers do not know how to reasonably price their products and would like to be able to choose an optimum selling point.

A user can list any item for sale or even a group of items from various categories along with product descriptions. Based on data provided about each item by the seller, you must predict a selling price for the specified product.

For example: Details might include things like - Name: ‘Madagascar Kids blu ray’, Category: ‘Media/Blu-Ray’, Description: ‘26Plays great. Tested. Watched once.’, Quality: ‘4’, etc.

Tags: Tabular Data, [slightly] Big Data

5: Why So Harsh?

Freedom of speech is one of the fundamental human rights. Keeping this in mind, you have created a blog on UpBlog where you would like to share content about your favorite things, especially movies and sports where the conflicts among the fan groups are enormously high. Since you stand by this “Freedom of Speech” right, you let people comment on your posts to get to know how the other fans feel about your opinion. However, this goes south too soon. The reason is that there has been a lot of abuse, hate, repugnant, and harsh content in the comments. Some are so hard to even read let alone how the targeted community feels like. Luckily you have started learning about machine learning and you see a clear application of it in this scenario.

A poor soul has earlier been in the exact same position as you and has done all the hard work to take every comment and judge them individually. To help any future bloggers, she has taken every solitary comment and labeled them into 6 mutually independent classes of being harsh. They are “harsh”, “extremely harsh”, “vulgar”, “threatening”, “disrespect”, “targeted hate”. Since you know that she faced a similar issue, you talked to her and asked for all of the data. She happily agreed and asked you to create a model to classify any new comment. Once the model is ready, UpBlog is ready to pay you a handsome fee and use the model so that such hate cannot be spread in the future. If the comment is harsh, notice the commenter and suspend their account for a few days.

Tags: NLP

6: It's a Fraud

Given details about online transactions, detect whether the transaction is fraudulent or not.

Tags: Tabular Data, [moderately] Big Data

7: PUBG ...Let's Play

In a PUBG game, up to 100 players start in each match (matchId). Players can be on teams (groupId) which get ranked at the end of the game (winPlacePerc) based on how many other teams are still alive when they are eliminated. In game, players can pick up different munitions, revive downed-but-not-out (knocked) teammates, drive vehicles, swim, run, shoot, and experience all of the consequences -- such as falling too far or running themselves over and eliminating themselves.

You are provided with a large number of anonymized PUBG game stats, formatted so that each row contains one player's post-game stats. The data comes from matches of all types: solos, duos, squads, and custom; there is no guarantee of there being 100 players per match, nor at most 4 player per group.

You must create a model which predicts players' finishing placement based on their final stats, on a scale from 1 (first place) to 0 (last place).

Tags: Tabular Data, [moderately] Big Data

8: FastHEAL Malware Detection:

FastHEAL is a firm which builds solutions to detect malware as well as to remove it. They have been working on building a tool which would let customers determine whether or not their system could be affected by malware. If yes, they provide the solutions they have developed so far. As apparent, this tool could help boost their revenues too.

You are a part of their Analytics team and are working on the problem: given different details about a system, predict whether or not it is prone to a malware attack.

They have also announced a raise for an employee who builds a reliable model for the purpose. Are you up for the challenge?

Tags: Tabular Data, Big Data

9: DDoS Attack Identification:

You are given various features regarding events in a network. You need to identify if those events were related to a DDoS attack, or was the network intrusion benign.

This is a binary classification problem involving approximately 80 features and 16 million training rows.

The metric to be used is the F1 score. Be warned: The dataset is imbalanced.

Tags: Tabular Data, [VERY] Big Data