Machine Learning & Data Mining
Miniproject 1

Caltech CS/CNS/EE 155
Released February 3, 2021

1 Overview

In miniproject 1, you will use data about the size, location, and time of reported wildfires to predict their cause. These wildfires were reported from the states of California and Georgia from 1992 to 2015, from small flames to massive forest fires.

You will participate in a competition on Kaggle, a site for data science competitions. There can be 10 submissions per day per team and submission window will close on Wednesday February 10th at 2:00 PM PST. **You may not use any additional data sources.**

In this competition, use the training data (`WILDFIRES_TRAIN/WILDFIRES_TRAIN.csv`) to come up with predictions for the test data (`WILDFIRES_TEST/WILDFIRES_TEST.csv`). There will be a public leaderboard that will show your performance, but it only consists of half of the test set (and you don't know which half). The private leaderboard ranking with the other half of the test set will only be revealed when the whole competition ends. The competition will end on Wednesday, February 10th at 2:00 PM PST.

The links to the competitions, which includes the datasets and guidebooks describing the datasets, can be found here: https://www.kaggle.com/c/caltech-cs-155-2021-mp1-part-1

There will be a benchmark submission added by the TAs. You should try to beat this benchmark.

Your task:

Each row in `WILDFIRES_TRAIN.csv` represents a fire event. The last column in `WILDFIRES_TRAIN.csv` is "`LABEL`". Each label is a description of the (statistical) cause of the fire.

- `LABEL` of 1: the fire originated from a natural cause (Lightning).
- `LABEL` of 2: the fire originated from an accidental cause (in order of decreasing frequency: Debris Burning, Equipment Use, Children, Campfire, Smoking, Railroad, Powerline, Fireworks, Structure).
- `LABEL` of 3: the fire originated from a malicious cause (Arson).
- `LABEL` of 4: the fire originated from another cause (Miscellaneous, or Missing/Undefined).

Your task is to predict the target in `WILDFIRES_TEST.csv` to the best of your ability. It is generally encouraged to submit probabilities from your models instead of 0/1 predictions, as you get rewarded for having a non-zero probability of the correct class even if it is not the highest probability for that sample. This also represents real world conditions, where you would like to know the relative likelihoods of each class being the right one.

Please follow the format in the sample submission files (`sample_submission.csv`) when generating your submissions to Kaggle.

Performance metric:
The metric on which your model performance is tested is AUC, namely, the **A**rea **U**nder the receiver operating characteristic **C**urve.

2 Key Notes

- The competitions end on Wednesday, February 10th at 2:00 PM PST.
- The report is due on Thursday, February 11th at 9:00 PM PST, via Gradescope. See below for the report guidelines. The report should explain your process and results in a thorough manner.
- You can work in groups of up to four people, but must make submissions from a single account.
- You can make up to 10 submissions a day. However, at the end, you need to select the 2 submissions that you think will perform the best on the private test sets for both competitions.
- If you have questions, please ask on Piazza! As with any Kaggle competition, it's best to get started early since you are only allowed to make 10 submissions a day.
- You can use any open-source tools, using both concepts you learned in class as well as any other techniques you find online (except for existing code written to model this particular wildfire dataset), to get the best score that you can.
- **You may collaborate fully within your team, but no collaboration is allowed between teams.**
- **You may not search for additional data related to this task; you may only train your models using the provided training set.)**

3 Report and Colab Demo Guidelines

- **Due date:** Thursday, February 11th at 9:00 PM PST
- **Report (75 points):** The report should be written exactly to the length specifications given in this document. If a section of your report is too long for that section, please try to be more concise - there is an extra credit section for you to discuss other interesting insights/approaches that you tried that you can use as overflow. You are encouraged to use graphs in your report and Colab demo, as visualization is very helpful!
- **Colab Demo (15 points):** You should write a Colab notebook that presents one or more interesting approaches / insights in a runnable and clearly written manner, so the class can learn from each other's work. This can include data exploration, feature engineering, model regularization, model ensembling. The notebook should be thoroughly annotated with markdown cells explaining what your code is doing and what you point you're making. Here is a nice example. Given that you have a 1 page limit for each section of your report, you can dive deeper on your approach/model selection/etc. Please try to include visualizations as well. To submit this, please share the public, read-only Colab link on Piazza in a public note, and attach the Piazza post link and the Colab link in your report.

- **Please submit your report in groups rather than submitting it once per student!** You can see how to submit in groups here:
  https://www.gradescope.com/help#help-center-item-student-group-members

We highly recommend that you use the LaTeX template provided to you and simply fill in the blanks. To collaborate on the report writing, we recommend using Overleaf (https://www.overleaf.com/edu/caltech), an online LaTeX editor. Caltech students can get a pro account for free using caltech.edu emails. See our example file for guidelines. The structure is as follows:

1. **Introduction (15 points):** This section is purely for the TAs and should be brief. Maximum of 1 page.
   - Group members
   - Team name (needs to match your team name on Kaggle)
   - What place you got on the private leaderboard for both competitions.
   - What AUC score you got on the private leaderboard for both competitions.
   - Division of labor: Your team must ensure that each member has an equal amount of workload during the competition. If there is a noticeable discrepancy in the division of labor, team members may receive differing grades.
2. **Overview (15 points):** This section should be a concise summary of your attempts. More detailed explanations should go in the next section. Maximum of 1 page.
   - Models and techniques tried: What models did you try? What techniques did you use along with your models? Did you implement anything out of the ordinary?

     Descriptions should be concise, at most 1-2 sentences. Again, more details can be included in the next section. However, this section is meant to be a more general overview.
   - Work timeline: What did your timeline look like for the competition?
3. **Approach (15 points):** This section should be a more detailed explanation of how you approached the competition. Maximum of 1 page.
   - Data exploration, processing and manipulation: Did you manipulate the data or the features in any way, such as data cleaning or feature engineering? What techniques and libraries did you use to accomplish such manipulation? Please justify your methodologies.
   - Details of models and techniques: Why did you try the models and techniques that you used? What was that process like? What are the advantages and disadvantages of using such methods?
4. **Model Selection (15 points):** This section should outline how you chose the best models. Maximum of 1 page.
   - Scoring: What optimization objectives did you use, and why? How did you score your models, and why? Which models scored the best?
   - Validation and test: How did you split your data? Did you use validation techniques? How did you test your models? What were the results of these tests, and what did the results tell you?

5. **Colab and Piazza link (15 points):** Please paste your Colab link and Piazza post link in a page on your report. Your piazza post only needs to contain your team name, team members names, and your Colab link. Maximum of 1 page.
6. **Conclusion (15 points):** This section should be used to summarize the report, as well as to include any additional details. Maximum of 1 page.
   - Insights: Please answer the following questions
     - Among all the features in the data, which features have the most influence on the prediction target? Why? List top 10 features. (Bonus points if you can analyze whether these 10 features positively or negatively influence the prediction target.)
     - Overall, what did you learn from this project?
   - Challenges: What could you have done differently? What obstacles did you encounter during the process?
7. **Extra Credit (10 points):** This section should be used to mention additional interesting insights and make concluding remarks. You can be creative!
   - Examples
     - Why do we use AUC as our Kaggle competition metric? Do you think there is a better metric for this project? Why, or why not?
     - Among the machine learning methods/pipelines that your group uses, are there any methods/pipelines that are parallelizable? If so, how can they be parallelized? If not, why not? (You are not required to actually parallelize your codes)

4 Grading metrics

For the competition, you will be scored on the test set. You will see results of the public leaderboard (results of your model on half of the test set) for the duration of the competition, and the private leaderboard results (results on the other half of the test set) will be released after the deadline.

The report and Colab is worth the majority of your grade. That is, we care more about the process and thoughts behind your results rather than the scores.