Applied Competitive Lab in Data Science

Exercise 3

Deadline: 24.11.2022 at 23:55

0 Submission Guidelines

Your grade would be made from the top 3 questions you've answered.

Submit a single Jupyter notebook file containing your answers(including your plots, code, and textual answers). Your notebook should be run from beginning to end without errors, and conform to the style shown in class.

Over the course of this exercise, you will be required to use trained models in questions 3 and 4. You may re-train models, or use the same models and hyperparameters you find in questions 1 and 2. The data used will be the data provided for exercise 2, with any added features you created at the exercise (make sure to add them in this exercise's notebook).

As an overarching issue, make sure you understand which features are 'safe' to use in your model, and which could incur leakage- avoid using those. We'll be trying to predict the n_killed column.

Some features you should avoid using are - participant_type, participant_status, incident_characteristics, note, n_injured, n_killed

1 Question 1

In class, we've discussed the processes and importance of choosing the correct metric to optimize for and producing robust and statistically significant results. Choose any two models you know and an evaluation metric to try and predict using the n_killed column as a target. Explain your choice of metric and models, what are their potential advantages and disadvantages, describe your process and final results.

2 Question 2

In class, we've discussed methods of performing hyperparameter selection. Using a machine learning model and evaluation metric of your choosing and, use two of the methods shown in class to explore the feature space, find the best performing hyperparameters, and report the performance. You don't have to optimize over all existing hyperparameters (as this could take a while), but make sure to select up to 4 of them. If using Bayesian HPO, we encourage you to use the Optuna library seen in class- but any other framework that performs this task is okay to use.

3 Question 3

(a) In class, we'll talk about discerning feature importance. Train a model to predict the n_killed column. Report the model's performance and visualize the most important features using any one method shown in class.

(b) Now, train two models- one without the single most important feature, and one with only the top 4. Report both models' performance and show their new feature importances. What changes occurred? Are they surprising?

4 Question 4

(a) Using one of your trained models, discern whether 'problematic' data points exist. Perform a short exploration and try to find differentiating patterns in its features to understand how they differ from data points with lower loss values. Visualize your results, and attempt to give an explanation for them.