INFSCI 2160: Homework #3

22 Points

This homework will cover

- XGBoost
- LightGBM
- CatBoost
- Gridsearch Techniques
- LIME and Shapley Additive Explanations (SHAP)

Part I: XGBoost, LightGBM, and CatBoost

We will be using the diabetes rehospitalizations dataset provided by https://archive.ics.uci.edu/ml/datasets/diabetes+130-us+hospitals+for+years+1999-2008 (https://archive.ics.uci.edu/ml/datasets/diabetes+130-us+hospitals+for+years+1999-2008).

Question #1: Use XGBoost, LightGBM, and CatBoost to predict 'readmitted' where readmitted == <30 or not (binary classification). We will assume that each row is a unique hospitalization. Which model performs best? Keep any hyperparameter tuning that you do to each model.

Hint: For CatBoost, categorical, independent variables are:

- race
- gender
- age
- weight
- admission_type_id
- discharge_disposition_id
- admission_source_id
- payer_code
- medical_specialty
- number_outpatient
- number_inpatient
- number emergency
- diag_1, diag_2, diag_3
- remaining columns after diag_3 that are 'object' datatypes

11 points:

- 4 points for train vs. test split
- 3 points for successful XGBoost, LightGBM, and CatBoost model (1 point per model)
- 4 points for identifying best model:
    - 1 point for using acceptable evaluation metric
    - 1 point for correctly identifying best model
    - 1 point for comparing all models on test data for same metric
    - 1 point for identifying models that may be overfit or underfit (if applicable)

Part II: SHAP Values

For your best model, create a summary plot of the SHAP values on your training set and testing set. Which variable drives the model the most? Next, create a variable interaction plot for any two independent variables from your test or training set. What does their interaction tell you?

5 points:

- 2 points for successful SHAP summary plots (1 point per plot)
- 1 point for identifying the most important variable
- 1 point for variable interaction plot
- 1 point for description of variable interaction plot

Part III Gridsearch

Choose either hyperband or hyperopt to perform gridsearch on one of your models. What are the best parameters? Train and test this model with the best parameters. How does your model perform on train and test in terms of AUC? Is it overfit, underfit, or fit well? If it's overfit or underfit, what would you do to make your model perform better?

6 points:

- 1 point for successful gridsearch with hyperopt or hyperband
- 1 point for identifying best parameters
- 1 point for successfully training and testing a new model with these parameters
- 1 point for evaluating AUC on train and test with new model
- 1 point for accurately identifying fit
- 1 point for explaining how you would change your model is underfit or overfit