REPORT

CSE-472

Machine Learning Sessional

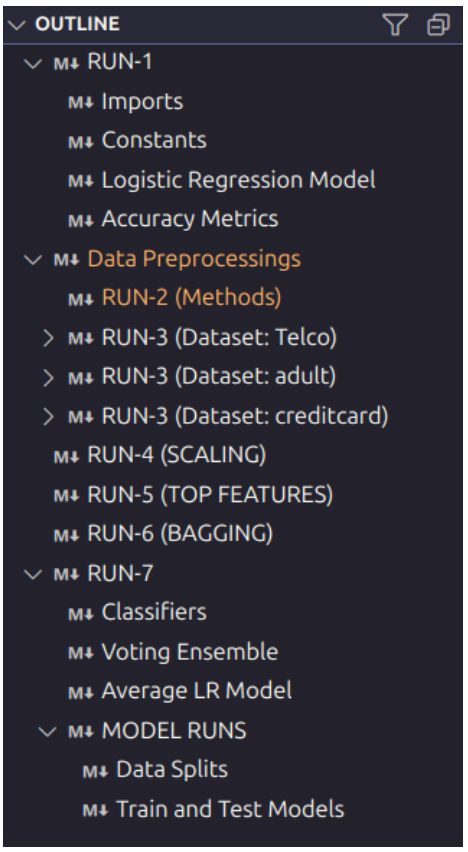
Apurbo Banik Turjo
1905096

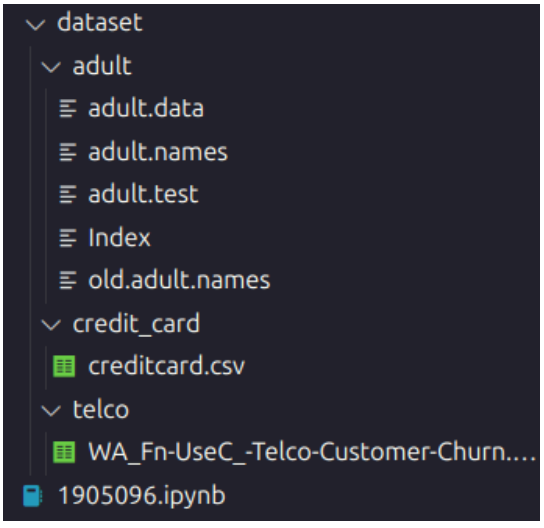
Offline on Stacking and Bagging
September 20, 2024

Instructions

Code Structure

- **OUTLINE**
  - The positional information of the full ipynb file can be found in the outline (vscode / colab)

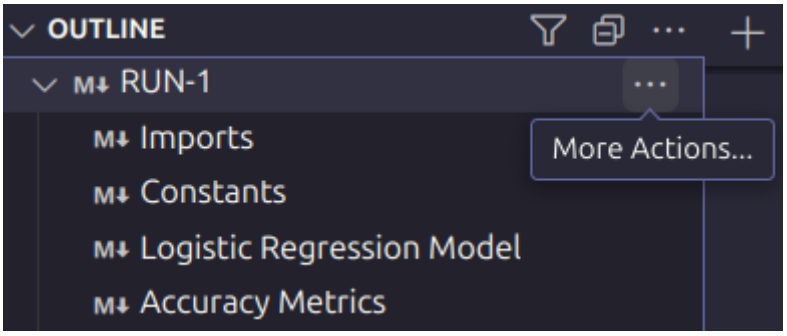


- **DATASET DIRECTORY STRUCTURE**
  - 

  - Where **telco == Dataset-1** mentioned in the spec &
  - where **adult == Dataset-2** mentioned in the spec &
  - where **credit_card == Dataset-3** mentioned in the spec
  - **The ipynb and the Dataset folder need to be in the same directory**

Running the code

- You can see that, sections that simply need to be run sequentially are denoted by **RUN-No**.
- In the VS-Code, run sections by using **More Actions => Run Cells in Section**



- For (multiple) **RUN-3** sections (3 datasets)
  - Select any one-of-the datasets and run that section
- **RUN-6, RUN-7** sections contains code related to **stacking and bagging**
  - **RUN-7** contains models LR (avg), Majority Voting Ensemble, Stacking Ensemble and generates evaluation metric upon running
- **Tweaking Hyperparameters**
  -
```
def train_model(x_train, y_train, epochs, batch_size):
    MODEL_DIM = x_train.shape[1]
    # Finetuning required for meta classifier training
    wt = np.zeros(MODEL_DIM + 1)
    alpha, beta, regularizer_type = 0.06, 0.00001, 'l1'
    print(f'alpha: {alpha}, beta: {beta}, regularizer_type: {regularizer_type}')
```
  - This code is listed under
    - **RUN-1**
      - **Logistic Regression Model**
        - Function **train_model**

Results

- The final result would be saved in a file (filename: Log.txt) upon the successful completion of the sub-section named **Model RUNS**
- **Log.txt**
  - **Would append results of all the models**
- **NB:**
  Running the Model RUNS section would take quite a time

Performance on Test-Set

Dataset : Telco

	Accuracy	Sensitivity	Specificity	Precision	F_1-score	AUROC	AUPR
LR*	0.8004 +- 0.0039	0.5498 +- 0.0603	0.8896 +- 0.0235	0.6434 +- 0.0272	0.5896 +- 0.0271	0.8458 +- 0.0016	0.6466 +- 0.0060
Voting ensemble	0.8021	0.5474	0.8929	0.6454	0.5924	0.8474	0.6501
Stacking ensemble	0.8050	0.5203	0.9064	0.6644	0.5836	0.8459	0.6549

Dataset : Adult

	Accuracy	Sensitivity	Specificity	Precision	F_1-score	AUROC	AUPR
LR*	0.8428 +- 0.0033	0.5812 +- 0.0602	0.9253 +- 0.0228	0.7171 +- 0.0406	0.6380 +- 0.0201	0.8946 +- 0.0004	0.7512 +- 0.0008
Voting ensemble	0.8450	0.5842	0.9272	0.7170	0.6438	0.8956	0.7532
Stacking ensemble	0.8417	0.6220	0.9111	0.6882	0.6534	0.8892	0.7303

Dataset : Credit Card

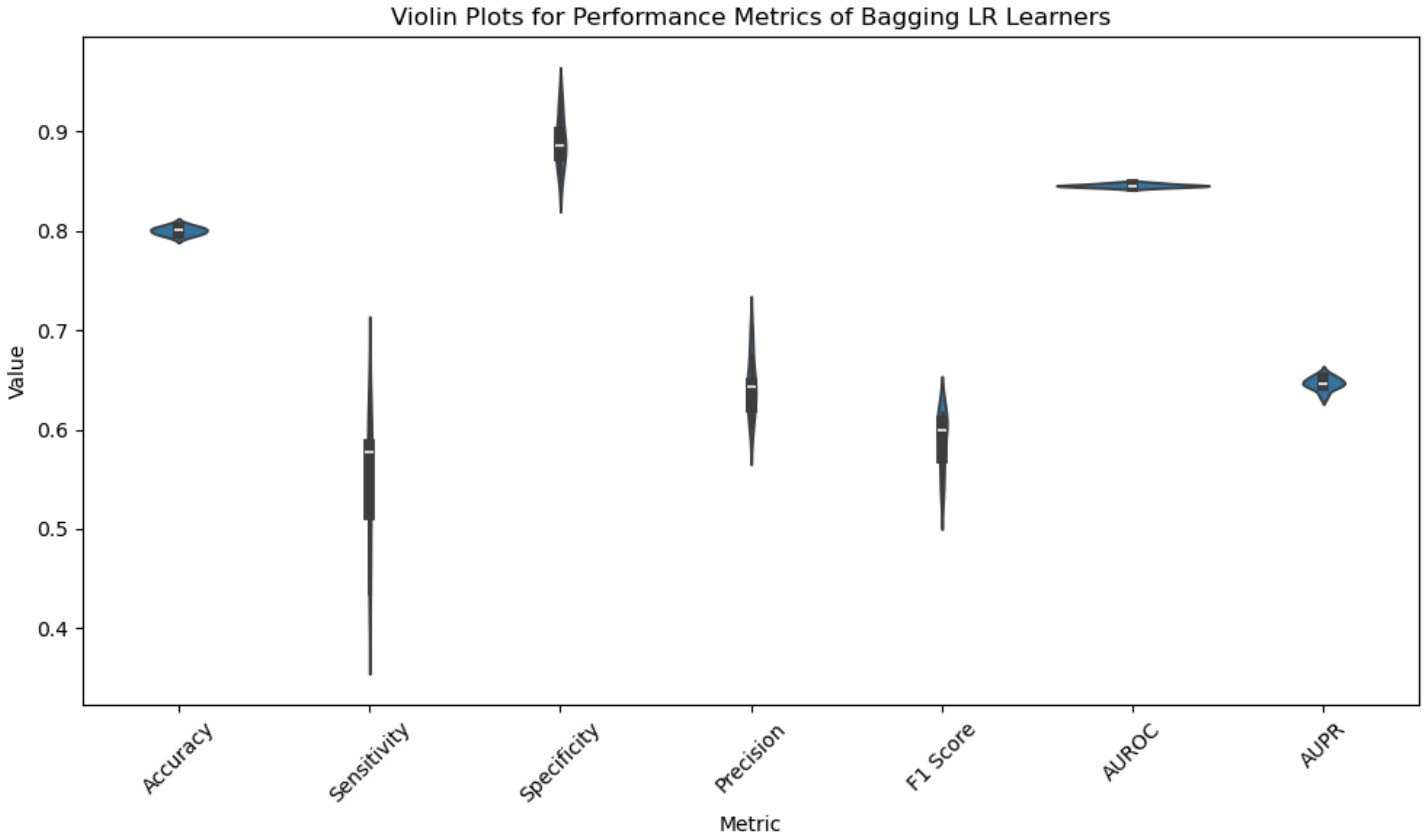
	Accuracy	Sensitivity	Specificity	Precision	F_1-score	AUROC	AUPR
LR*	0.9931 +- 0.0003	0.7269 +- 0.0137	0.9995 +- 0.0001	0.9707 +- 0.0074	0.8312 +- 0.0082	0.9673 +- 0.0026	0.8548 +- 0.0048
Voting ensemble	0.9934	0.7396	0.9995	0.9726	0.8402	0.9681	0.8560
Stacking ensemble	0.9932	0.7396	0.9992	0.9595	0.8353	0.9781	0.8748

Hyper Parameters:

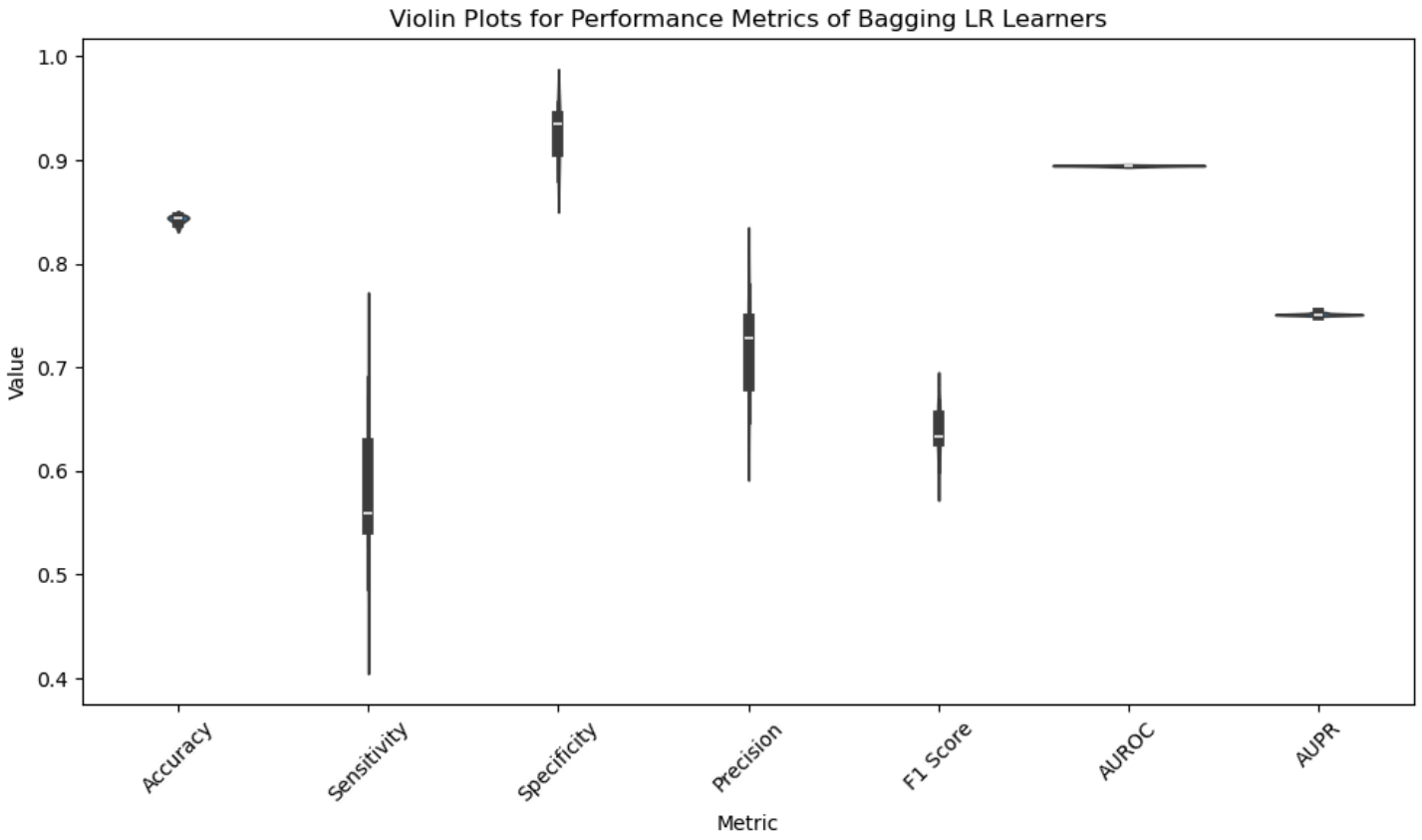
Dataset	alpha	beta (lambda)	regularizer	epochs	mini-batch size
Telco	0.01	0.01	L1	100	200
Adult	0.05 0.02 (stacking)	0.00001	L1	100	200
Credit Card	0.06	0.00001	L1	100	200

Violin Plot

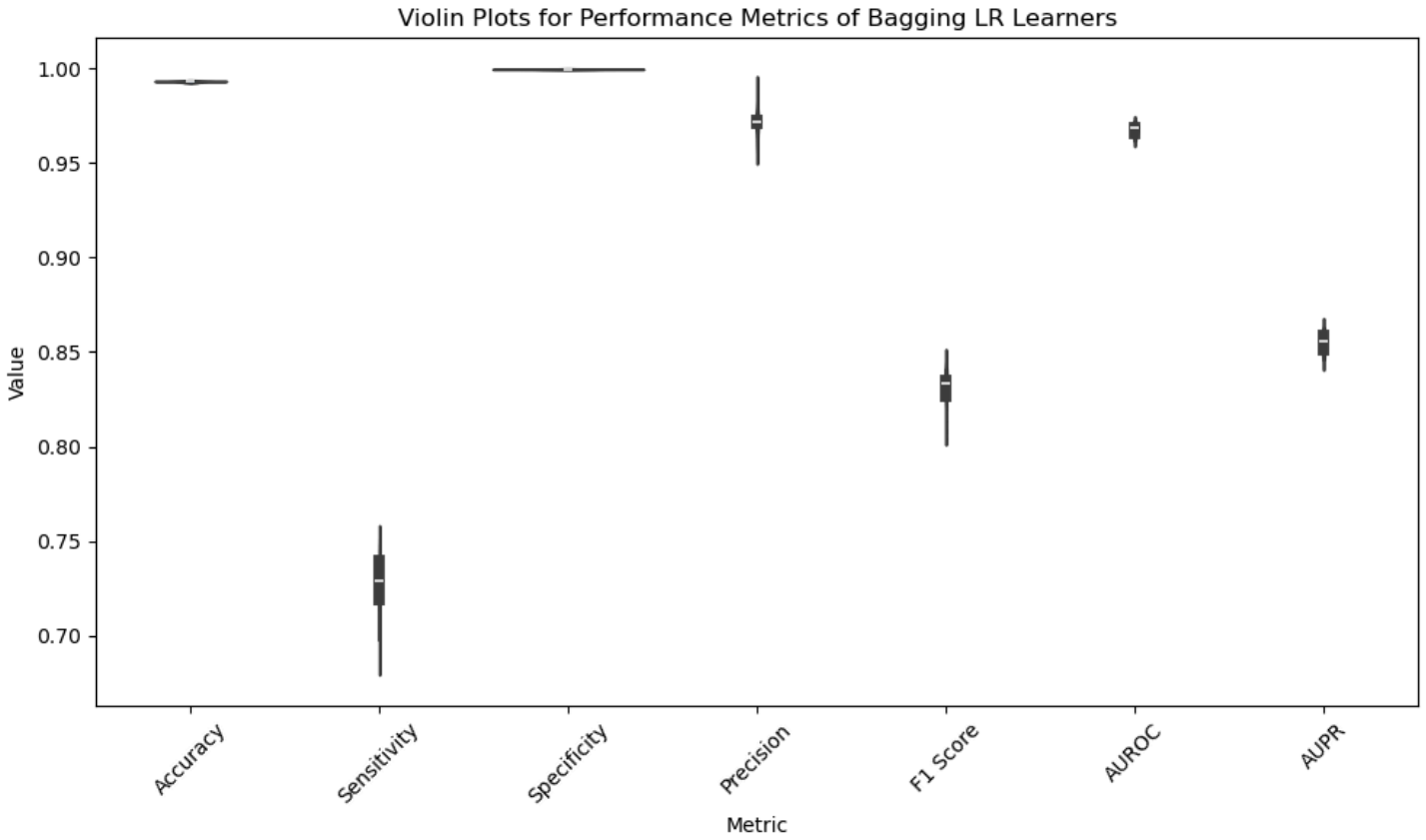
Dataset : Telco



Dataset : Adult



Dataset : Credit Card



Observations

Dataset Related Observation

- The **"telco"** dataset contains a column "CustomerID" which contains data that are all unique. Which in terms with other columns would add no additional value to the model learning. So, this column is dropped in the data cleaning.

Performance Related Observation

- As the learning rate goes up, the accuracy seems to be increasing for credit_card dataset.
- Increasing number of epochs also contribute to the performance enhancement of the models.