CSE472 (Machine Learning Sessional)

Assignment# 2: Logistic Regression with Bagging and Stacking

Student Id : 1905038

Learning Rate : 0.01

Epoch : 1000

**Constant SEED used: 40**

How to run:

- First, make sure that numpy, pandas, fetch_ucirepo, and sklearn is installed in your device
- To install, you can write the command "pip install numpy"
- After that, to read the datasets, you need to download the 1_{st} and 3_{rd} dataset and absolute path of the dataset needs to be added to pd.read_csv
- For dataset 2 no need to download, it will work just as given
- Now, to run dataset 1, we need to uncomment **cell 13**
- Now, to run dataset 1, we need to uncomment **cell 14**
- Now, to run dataset 1, we need to uncomment **cell 15**
- **When data is calculated, dataset 1 and 3 is calculated without feature selection and dataset 2 is calculated with top 50 correlated features selected.**

Telco Customer Churn Dataset:

Performance on Test set

	Accuracy	Sensitivity	Specificity	Precision	F1-score	AUROC	AUPR
LR*	0.7926±.0047	0.679±.0107	.8331±.0094	.592±.0106	.6324±.004	.8423±.0009	.6428±.003
Voting ensemble	0.795	0.6694	0.8398	0.5981	0.6317	0.8422	0.6452
Stacking ensemble	0.7879	0.6856	0.8243	0.5816	0.6294	0.8434	0.65

Violin Plot:

violin plots
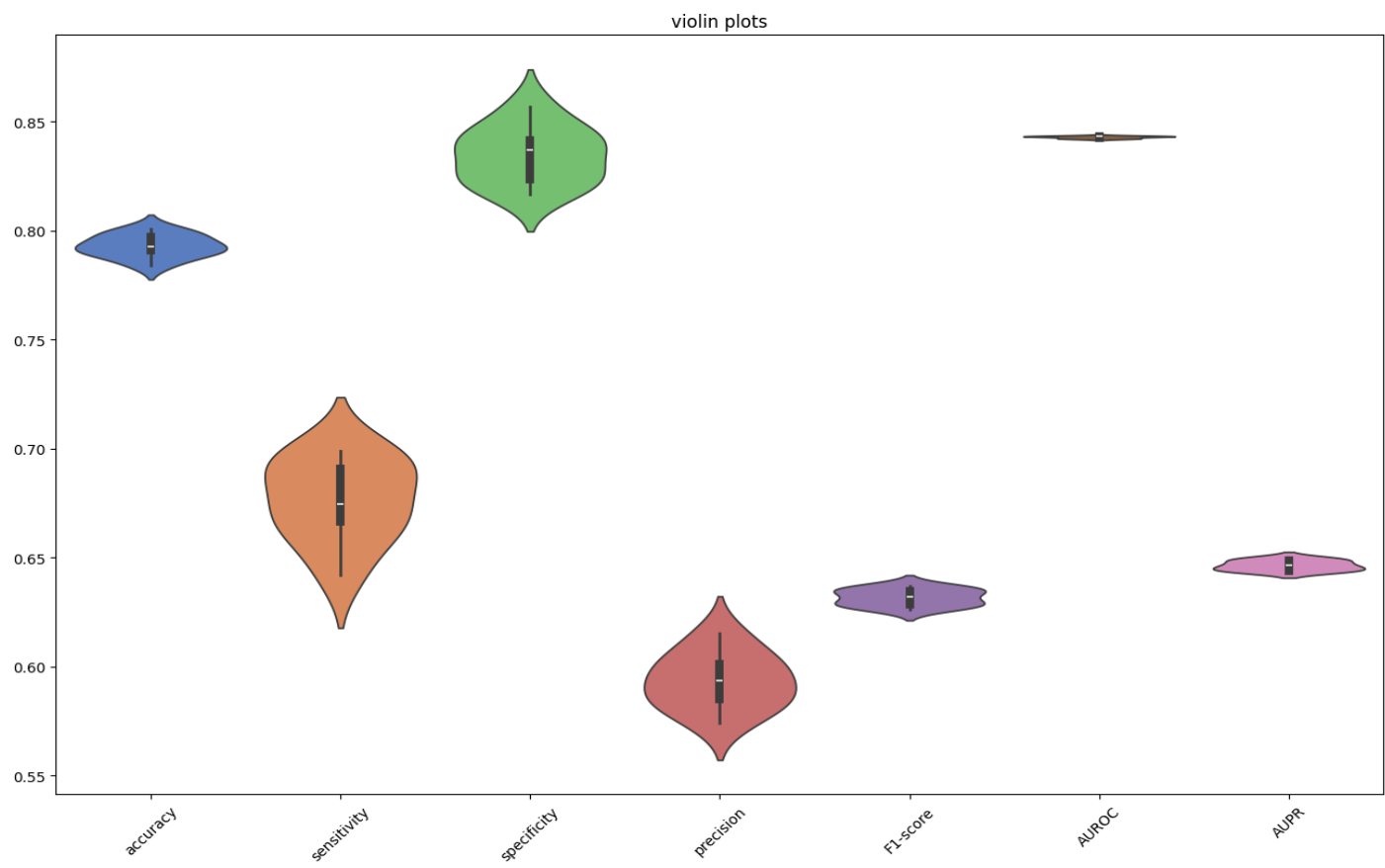
Adult Dataset:

Performance on Test set

	Accuracy	Sensitivity	Specificity	Precision	F1-score	AUROC	AUPR
LR*	0.8247±.001	0.6167±.0102	0.8934±.0045	0.6563±.006	0.6358±.0004	0.8837±.0004	0.7077±.0009
Voting ensemble	0.8253	0.6185	0.8935	0.657	0.6372	0.8837	0.7074
Stacking ensemble	0.824	0.6449	0.8831	0.6455	0.6452	0.8857	0.7104

violin plots
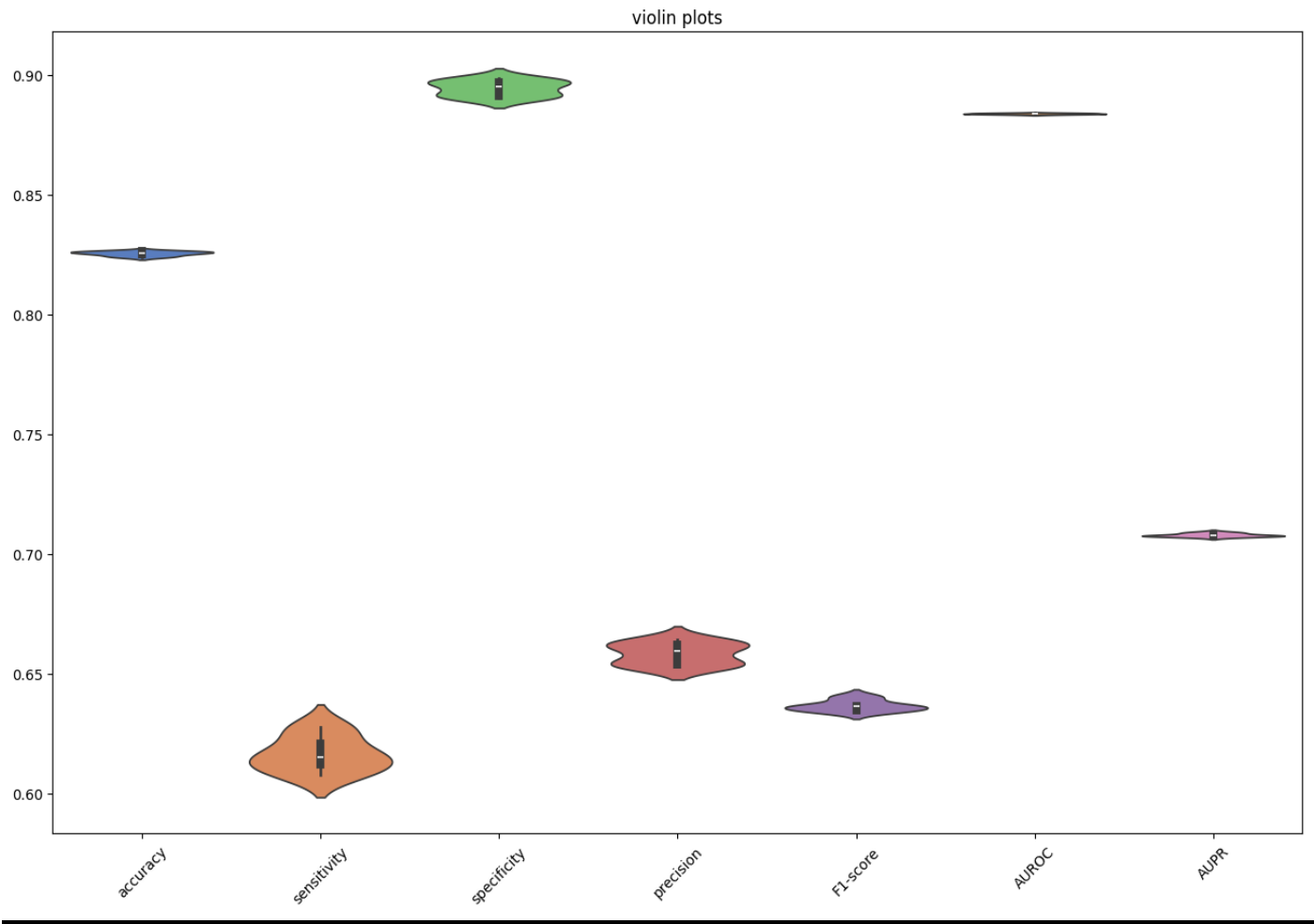
Credit Card Fraud Dataset:

Performance on Test set

	Accuracy	Sensitivity	Specificity	Precision	F1-score	AUROC	AUPR
LR*	0.9958±0.0	0.84±0.0	0.9997±0.0	0.9882±0.0	0.9081±0.0	0.9903±0.0	0.9233±0.003
Voting ensemble	0.9958	0.84	0.9997	0.9882	0.9081	0.9917	0.9264
Stacking ensemble	0.9958	0.84	0.9997	0.9882	0.9081	0.9897	0.9181

Violin Plots
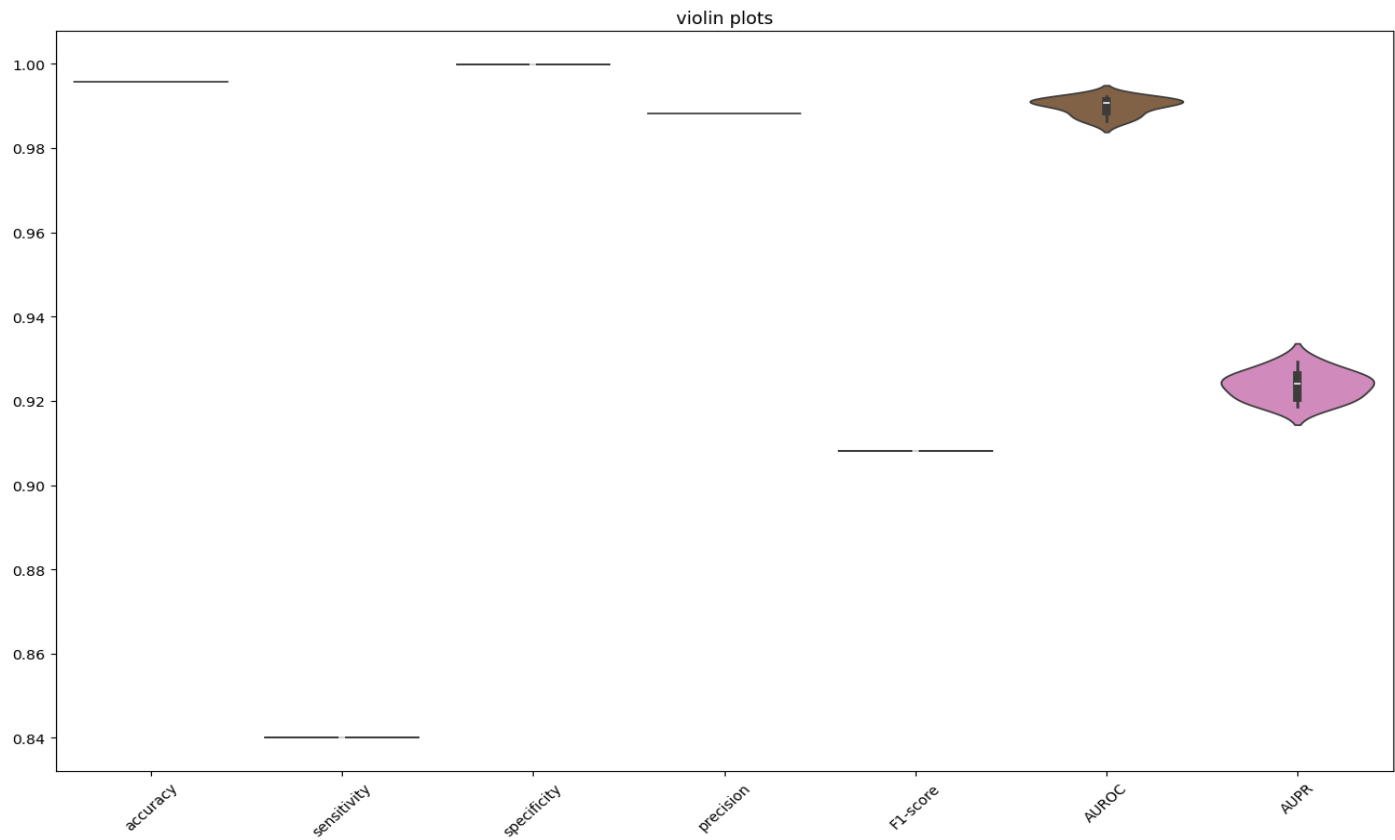
Observation:

- **LR*, Voting Ensemble , Stacking Ensemble give almost same accuracy for 3 dataset**
- For dataset2, without feature selection much time is needed as dataset size is huge