Assignment 2: Logistic Regression with Bagging and Stacking

1905113 - Anamul Hoque Emtiaj

September 2024

1 How to Run the Code

To run the notebook, please follow the steps below:

1. You will need to place the following files in the same folder structure:
   - `1905113.ipynb` (Notebook)
   - `datasets/creditcard.csv` (Credit Card dataset)
   - `datasets/WA_Fn-UseC_-Telco-Customer-Churn.csv` (Telco Customer Churn dataset)
   - `datasets/adult/adult.data` (Adult dataset - training data)
   - `datasets/adult/adult.test` (Adult dataset - test data)
2. The folder structure should look like this:

```
1905113.ipynb
datasets/
   creditcard.csv
   WA_Fn-UseC_-Telco-Customer-Churn.csv
   adult/
       adult.data
       adult.test
```

3. Once your setup is done, open `1905113.ipynb` with vscode, select kernel and run all cells to execute the code.

2 Experiment

The function `experiment` is designed to run different models on three datasets: Telco, Adult, and Credit Card. Depending on the choice of the dataset and experiment number, different techniques such as Logistic Regression (LR), Majority Voting, or Stacking with Bagging can be used. Below are the details of the function parameters and how to run experiments:

2.1 Function Definition

The function `experiment` has the following signature:

```
experiment(dataset, exp_no, withTopFeatures=False, top_features=20, withCorrelation=False)
```

2.2 Parameters

The function takes the following input parameters:

- `dataset`: This parameter selects the dataset for the experiment. It accepts the following values:
  - 1: Telco Dataset
  - 2: Adult Dataset

  - 3: Credit Card Dataset

- `exp-no`: This parameter selects the type of model to run. It accepts the following values:
  - 1: Bagging Logistic Regression (`LRStar`)
  - 2: Majority Voting of the Base Models
  - 3: Stacking with Bagging
- `withTopFeatures`: A boolean value to indicate whether to perform feature selection. The default is `False`. When set to `True`, the function selects the top `k` features based on either information gain or correlation, depending on the next parameter.
- `top-features`: The number of top features to select when `withTopFeatures=True`. The default value is 20. If feature selection is not enabled, this value is ignored.
- `withCorrelation`: A boolean value that determines the method of feature selection when `withTopFeatures=True`. The default is `False`, meaning feature selection is performed using information gain. When set to `True`, feature selection is performed based on correlation.

2.3 Example Usage

Here is an example of how to run the model for the Telco dataset with Bagging Logistic Regression using all features:

```
report = experiment(dataset=1, exp_no=1)
```

For running the model on the Adult dataset with top 10 features selected using information gain:

```
report = experiment(dataset=2, exp_no=1, withTopFeatures=True, top_features=10)
```

For running the model on the Credit Card dataset with the Stacking ensemble model and the top 15 features selected using correlation:

```
experiment(dataset=3, exp_no=3, withTopFeatures=True, top_features=15, withCorrelation=True)
```

2.4 Output

The function returns a report, which contains performance metrics such as Accuracy, Precision, Recall, F1 Score, AUROC, and AUPRC for the selected model and dataset.

3 Violin Plots for Performance Metrics

The function `violin_plot` is designed to generate violin plots for the performance metrics (Accuracy, Precision, Recall, F1 Score, AUROC, and AUPRC) of Bagging Logistic Regression models trained on three datasets: Telco, Adult, and Credit Card. The plots help visualize the distribution of performance metrics across the models.

3.1 Function Definition

The function `violin_plot` has the following signature:

```
violin_plot(dataset, withTopFeatures=False, top_features=20, withCorrelation=False)
```

3.2 Parameters

The function takes the following input parameters:

- `dataset`: This parameter selects the dataset for which violin plots will be generated. It accepts the following values:
  - 1: Telco Dataset
  - 2: Adult Dataset
  - 3: Credit Card Dataset
- `withTopFeatures`: A boolean value indicating whether to perform feature selection. The default is `False`. When set to `True`, the function selects the top `k` features based on either information gain or correlation.
- `top-features`: The number of top features to select when `withTopFeatures=True`. The default value is 20. This value is ignored when feature selection is not enabled.
- `withCorrelation`: A boolean value determining the method of feature selection when `withTopFeatures=True`. The default is `False`, meaning feature selection is based on information gain. When set to `True`, feature selection is based on correlation.

3.3 Example Usage

Here is an example of how to generate violin plots for the Telco dataset using all features:

```
violin_plot(dataset=1)
```

For generating violin plots on the Adult dataset with the top 10 features selected using information gain:

```
violin_plot(dataset=2, withTopFeatures=True, top_features=10)
```

For generating violin plots on the Credit Card dataset with the top 15 features selected using correlation:

```
violin_plot(dataset=3, withTopFeatures=True, top_features=15, withCorrelation=True)
```

3.4 Visual Output

The function generates violin plots for each performance metric. The **X-axis** represents the different models, while the **Y-axis** represents the performance metric values.

4 Performance on Test Sets

In this section, we present the performance of different models (Logistic Regression, Majority Voting, and Stacking Ensemble) on the Telco, Adult, and Credit Card datasets. The following metrics are reported: Accuracy, Precision, Sensitivity (Recall), Specificity, F1 Score, AUROC, and AUPRC.

4.1 Telco Dataset with All Features

Model	Accuracy	Precision	Sensitivity	Specificity	F1 Score	AUROC	AUPRC
LR*	0.7889 ± 0.0030	0.6244 ± 0.0093	0.5175 ± 0.0113	0.8872 ± 0.0060	0.5658 ± 0.0061	0.7023 ± 0.0039	0.6351 ± 0.0047
Voting Ensemble	0.7868	0.6194	0.5134	0.8858	0.5614	0.6996	0.6310
Stacking Ensemble	0.7925	0.6464	0.4840	0.9042	0.5535	0.6941	0.6338

Table 1: Performance on the Telco dataset with all features.

4.2 Adult Dataset with All Features

Model	Accuracy	Precision	Sensitivity	Specificity	F1 Score	AUROC	AUPRC
LR*	0.8531 ± 0.0004	0.7326 ± 0.0033	0.5959 ± 0.0046	0.9327 ± 0.0016	0.6572 ± 0.0017	0.7643 ± 0.0016	0.7120 ± 0.0008
Voting Ensemble	0.8535	0.7360	0.5928	0.9342	0.6567	0.7635	0.7125
Stacking Ensemble	0.8529	0.7369	0.5871	0.9352	0.6535	0.7611	0.7108

Table 2: Performance on the Adult dataset with all features.

4.3 Credit Card Dataset with All Features

Model	Accuracy	Precision	Sensitivity	Specificity	F1 Score	AUROC	AUPRC
LR*	0.9949 ± 0.0004	0.9694 ± 0.0149	0.8214 ± 0.0041	0.9993 ± 0.0003	0.8892 ± 0.0070	0.9103 ± 0.0021	0.8976 ± 0.0080
Voting Ensemble	0.9951	0.9767	0.8235	0.9995	0.8936	0.9115	0.9023
Stacking Ensemble	0.9954	0.9882	0.8235	0.9997	0.8984	0.9116	0.9081

Table 3: Performance on the Credit Card dataset with all features.

5 Violin Plot for Performance Metrics

In this section, we present the violin plots for the performance metrics across different models for each dataset (Telco, Adult, and Credit Card). The violin plots show the distribution of the performance metric values, allowing for a better understanding of model performance variability.

5.1 Violin Plot for Telco Dataset

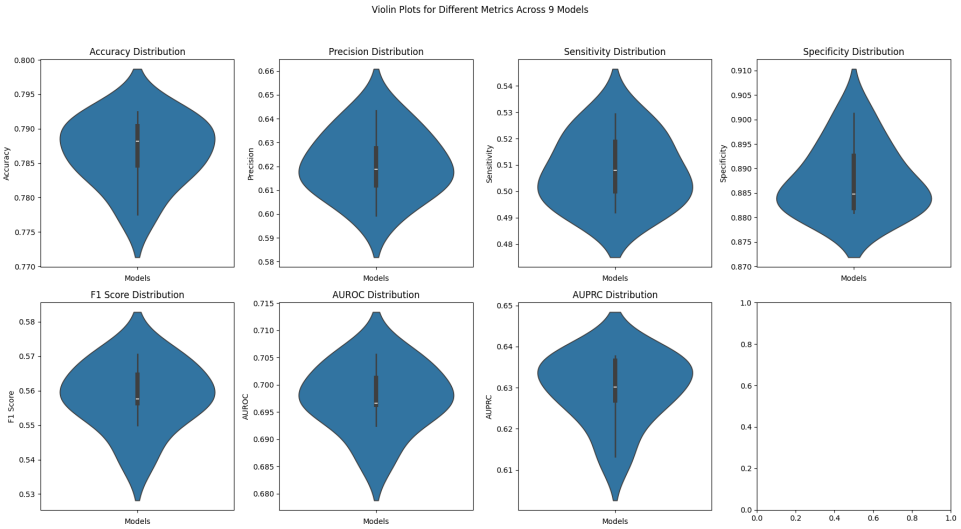

Figure 1: Violin plot for performance metrics of Telco Dataset

5.2 Violin Plot for Adult Dataset

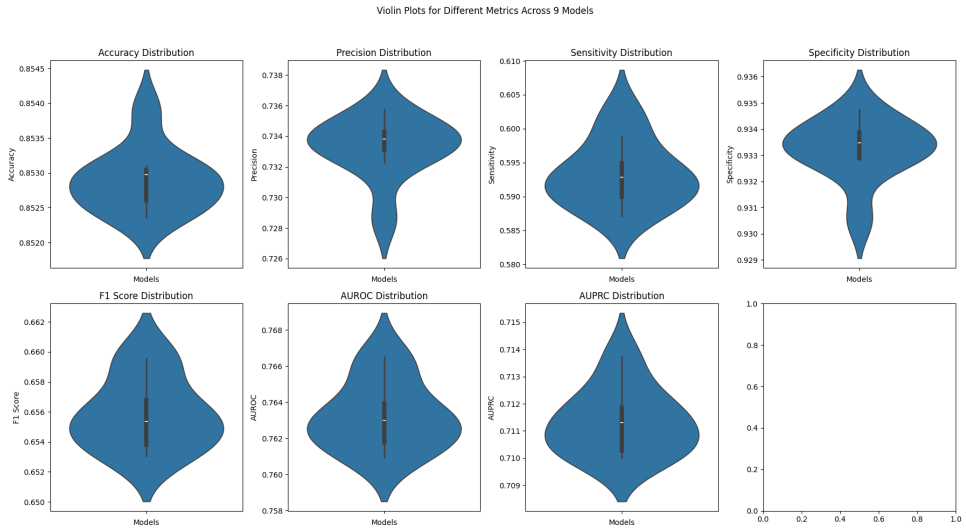

Figure 2: Violin plot for performance metrics of Adult Dataset

5.3 Violin Plot for Credit Card Dataset

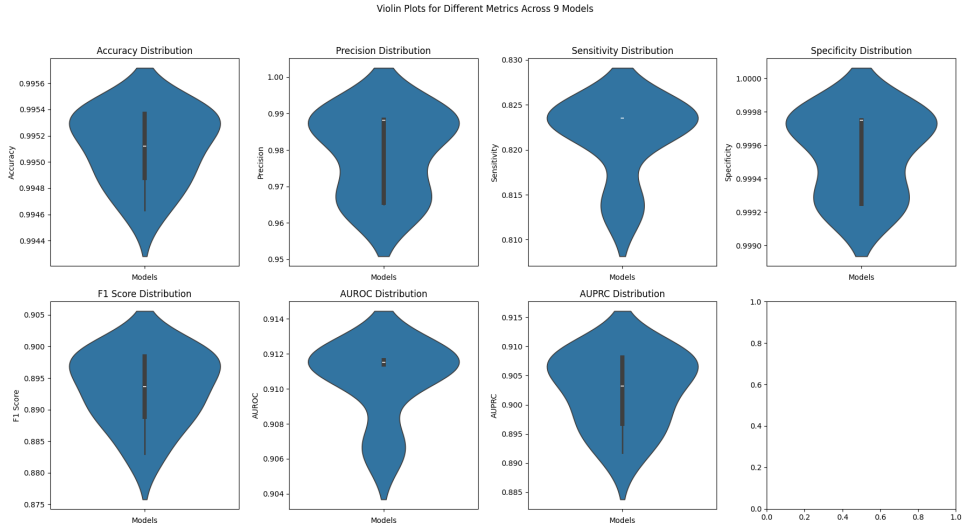

Figure 3: Violin plot for performance metrics of Credit Card Dataset