Description of Classifiers

Random Forests are an ensemble learning method that operates by constructing a multitude of decision trees at training time. For classification tasks, the output of the random forest is the class selected by most trees. It achieves better performance and generalization by averaging multiple deep decision trees, trained on different parts of the same training set, which reduces the variance.

SVMs are a set of supervised learning methods used for classification, regression, and outliers detection. The Gaussian kernel, also known as the Radial Basis Function (RBF), is used in SVM to transform the input space into a higher-dimensional space where a hyperplane can be used to separate classes. The RBF kernel is effective in cases where the relationship between class labels and attributes is nonlinear.

Training Methodology

For both classifiers, the training methodology involves:

- Loading the dataset in LibSVM format using load_svmlight_file
- Converting sparse matrices to dense arrays
- Splitting the training data further into training and validation sets using train_test_split
- Employing RandomizedSearchCV to optimize hyperparameters through a specified number of iterations over cross-validation, which ensures robust evaluation of each parameter combination
- Evaluating the final model on a separate test set to gauge generalization performance

List of Hyperparameters Tuned

Random Forest

- n_estimators: Number of trees in the forest
    - Default = 100
    - Tested = [100, 300]
- bootstrap: Whether bootstrap samples are used when building trees.
    - Default = True
    - Tested = [True, False]
- max_depth: Maximum depth of the tree.
    - Default = None, meaning nodes are expanded until all leaves are pure.
    - Tested = [None, 20]
- min_samples_leaf: Minimum number of samples required to be at a leaf node.
    - Default = 1
    - Tested = [1, 4]

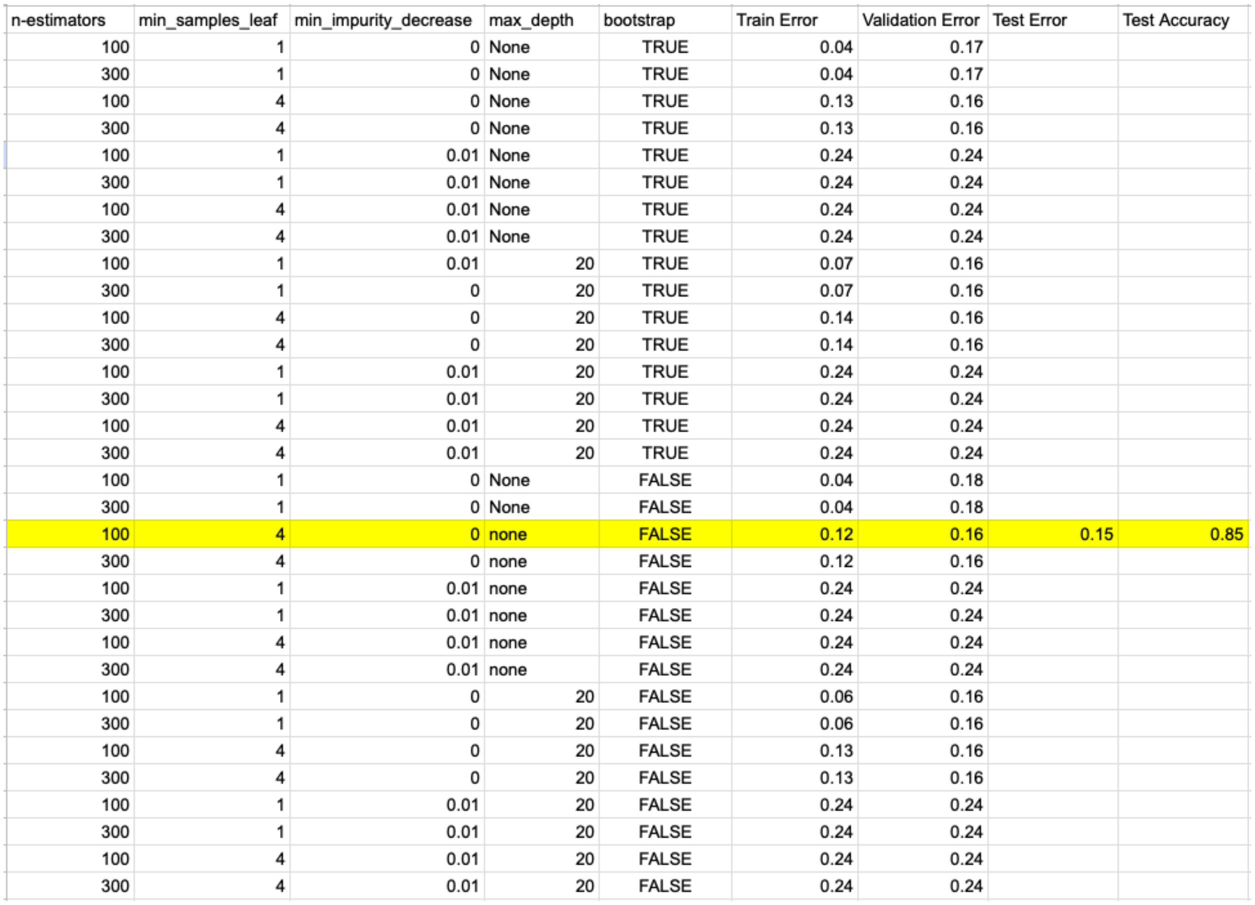
- min_impurity_decrease: A node will be split if this split induces a decrease of the impurity greater than or equal to this value
    - Default = 0.0
    - Tested = [0.0, 0.01]

SVM

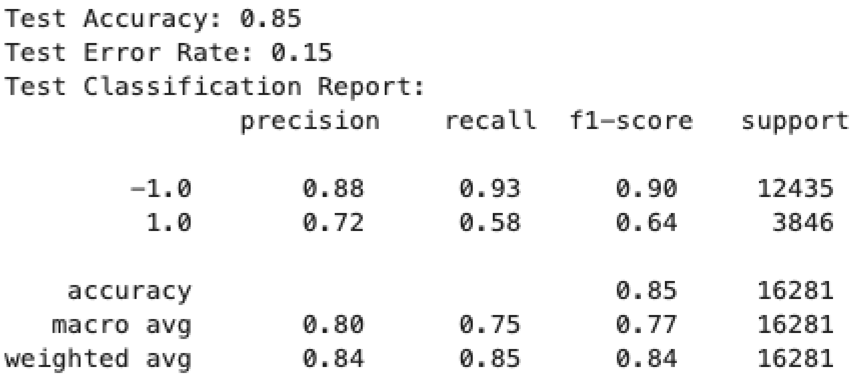
- C: Regularization parameter. The strength of the regularization is inversely proportional to C.
    - Default = 1.0
    - Tested = [1, 10]
- gamma: Kernel coefficient for 'rbf'.
    - Default is 'scale' which uses 1 / (n_features * X.var()) as the value of gamma
    - Tested = [0.1, 0.01]
- kernel: Specifies the kernel type to be used in the algorithm.
    - Tested = 'rbf' (Gaussian)

Error Rates and Accuracy

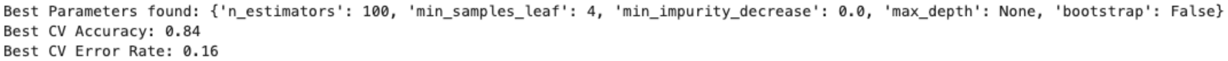
Random Forest



n-estimators	min_samples_leaf	min_impurity_decrease	max_depth	bootstrap	Train Error	Validation Error	Test Error	Test Accuracy
100	1	0	None	TRUE	0.04	0.17		
300	1	0	None	TRUE	0.04	0.17		
100	4	0	None	TRUE	0.13	0.16		
300	4	0	None	TRUE	0.13	0.16		
100	1	0.01	None	TRUE	0.24	0.24		
300	1	0.01	None	TRUE	0.24	0.24		
100	4	0.01	None	TRUE	0.24	0.24		
300	4	0.01	None	TRUE	0.24	0.24		
100	1	0.01	20	TRUE	0.07	0.16		
300	1	0	20	TRUE	0.07	0.16		
100	4	0	20	TRUE	0.14	0.16		
300	4	0	20	TRUE	0.14	0.16		
100	1	0.01	20	TRUE	0.24	0.24		
300	1	0.01	20	TRUE	0.24	0.24		
100	4	0.01	20	TRUE	0.24	0.24		
300	4	0.01	20	TRUE	0.24	0.24		
100	1	0	None	FALSE	0.04	0.18		
300	1	0	None	FALSE	0.04	0.18		
100	4	0	none	FALSE	0.12	0.16	0.15	0.85
300	4	0	none	FALSE	0.12	0.16		
100	1	0.01	none	FALSE	0.24	0.24		
300	1	0.01	none	FALSE	0.24	0.24		
100	4	0.01	none	FALSE	0.24	0.24		
300	4	0.01	none	FALSE	0.24	0.24		
100	1	0	20	FALSE	0.06	0.16		
300	1	0	20	FALSE	0.06	0.16		
100	4	0	20	FALSE	0.13	0.16		
300	4	0	20	FALSE	0.13	0.16		
100	1	0.01	20	FALSE	0.24	0.24		
300	1	0.01	20	FALSE	0.24	0.24		
100	4	0.01	20	FALSE	0.24	0.24		
300	4	0.01	20	FALSE	0.24	0.24		

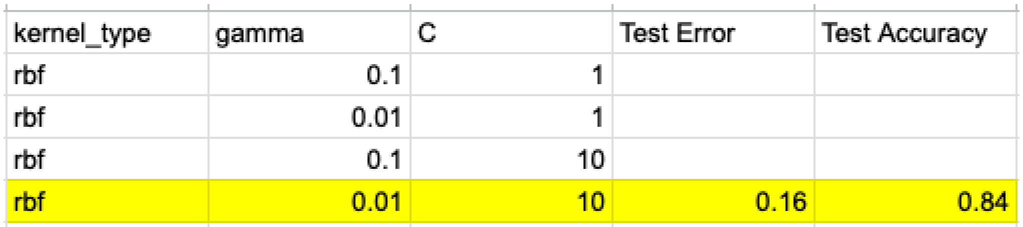


```
Test Accuracy: 0.85
Test Error Rate: 0.15
Test Classification Report:
              precision    recall  f1-score   support

        -1.0       0.88      0.93      0.90     12435
         1.0       0.72      0.58      0.64      3846

    accuracy                           0.85     16281
   macro avg       0.80      0.75      0.77     16281
weighted avg       0.84      0.85      0.84     16281
```

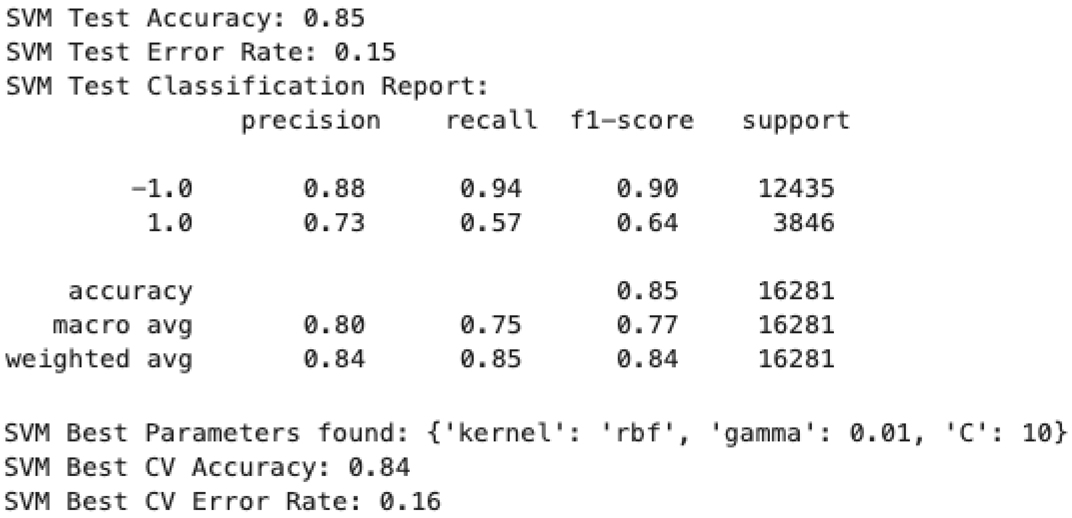


```
Best Parameters found: {'n_estimators': 100, 'min_samples_leaf': 4, 'min_impurity_decrease': 0.0, 'max_depth': None, 'bootstrap': False}
Best CV Accuracy: 0.84
Best CV Error Rate: 0.16
```

SVM



kernel_type	gamma	C	Test Error	Test Accuracy
rbf	0.1	1		
rbf	0.01	1		
rbf	0.1	10		
rbf	0.01	10	0.16	0.84



```
SVM Test Accuracy: 0.85
SVM Test Error Rate: 0.15
SVM Test Classification Report:
              precision    recall  f1-score   support

        -1.0       0.88      0.94      0.90     12435
         1.0       0.73      0.57      0.64      3846

    accuracy                           0.85     16281
   macro avg       0.80      0.75      0.77     16281
weighted avg       0.84      0.85      0.84     16281

SVM Best Parameters found: {'kernel': 'rbf', 'gamma': 0.01, 'C': 10}
SVM Best CV Accuracy: 0.84
SVM Best CV Error Rate: 0.16
```