Experiments with protein binding data

Experiment 1:

- Trained an AdaBoost Classifier on a balanced training set (4000 positive and 4000 negative samples).
- The data was initially split into 80% training data and 20% test data.
- Using Variance Threshold as a feature selection strategy, 101 features out of 188 were selected with a threshold of 1.0
- AdaBoost classifier trained on these 101 features with default hyper parameters produces the below result:
    - Precision: 0.922
    - Recall: 0.898
    - F1 Score: 0.909
    - Accuracy: 0.910

Experiment 2:

- Used Random Forest Classifier to determine the feature importance of each feature. Since Random Forest Classifier builds trees on random features, the classifier was run multiple times to determine the features that appeared the most in multiple runs.
- The feature selection on 188 features(protein, ligand and protein-ligand features), produced a minimal set of 8 features with results similar to those using all 188 features.

  Features: ['nPyrimidines', 'nImidazoles', 'nArCONHR', 'PCR', 'Wi_D', 'TPSA(Tot)', 'nCbH', 'nCb-']
- Using the above features AdaBoost classifier was trained on individual protein cluster. The decision to train the classifier on individual clusters was motivated by the curiosity to determine which specific protein conformation acts as a good binding site for the ligands.
- The results from different protein clusters are shown in next page:

- Cluster 1
  - Precision: 0.912
  - Recall: 0.912
  - F1 Score: 0.912
  - AUC: 0.97
- Cluster 2
  - Precision: 0.936
  - Recall: 0.928
  - F1 Score: 0.932
  - AUC: 0.97
- Cluster 3
  - Precision: 0.941
  - Recall: 0.943
  - F1 Score: 0.941
  - AUC: 0.98
- Cluster 4
  - Precision: 0.90
  - Recall: 0.923
  - F1 Score: 0.911
  - AUC: 0.96
- Cluster 5
  - Precision: 0.922
  - Recall: 0.941
  - F1 Score: 0.931
  - AUC: 0.97
- Cluster 6
  - Precision: 0.932
  - Recall: 0.929
  - F1 Score: 0.930
  - AUC: 0.97
- Cluster 7
  - Precision: 0.90
  - Recall: 0.925
  - F1 Score: 0.912
  - AUC: 0.96
- All Cluster
  - Precision: 0.989
  - Recall: 0.990
  - F1 Score: 0.990
  - AUC: 0.99

- From the above results, it is clear that classifiers trained on all the protein clusters perform better than the classifier trained on individual clusters. Since the number of samples in individual cluster is far less than the samples for all the clusters (680 vs 8000) and the fact that the data available for training is lesser in individual clusters is reflected in the results.

- These results were obtained from 5-fold cross validation set.

- One of the hyper parameters in AdaBoost Classifier is *"n_estimators".* This parameter specifies the number of weak learners to be used in the training process. In order to help us determine the ideal value for this parameter, we plot the F-score vs n_estimators. For different values of n_estimators, we calculate the F-score and the variation of F-score is shown(next page):

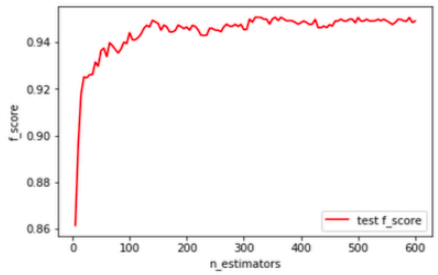

Figure: Protein Cluster 1

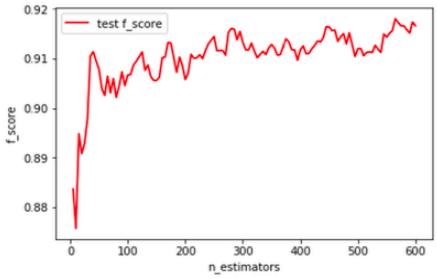

Figure: Protein Cluster 2

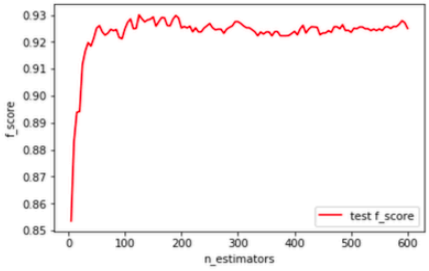

Figure: Protein Cluster 3

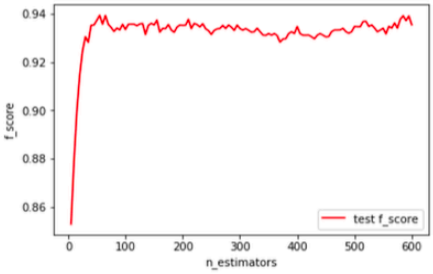

Figure: Protein Cluster 4

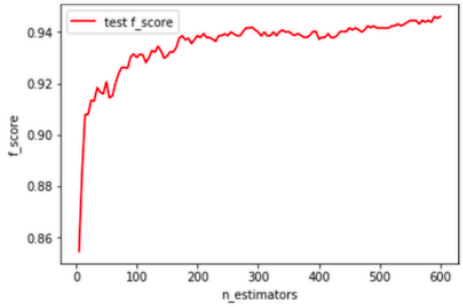

Figure: Protein Cluster 5

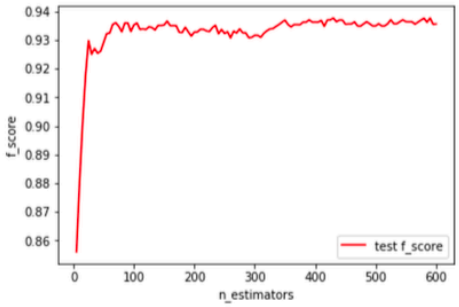

Figure: Protein Cluster 6

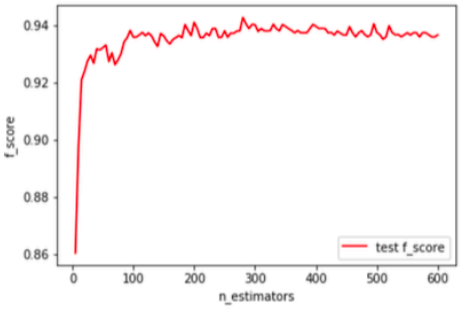

Figure: Protein Cluster 7

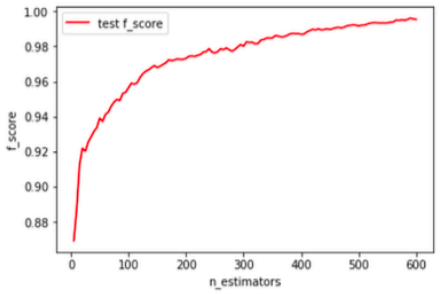

Figure: All Protein Clusters

Experiment 3:

- In order to validate these results further, the classifier trained on all protein clusters was tested on a blind test set consisting of 95 active ligands.

- Since, the protein-ligand features were not available at the time of testing, the feature space was restricted to ligand features.

- The classifier was evaluated in 2 parts: one with feature selection and the other without feature selection.

Part 1:

- All the protein features were used in the feature selection process, and out of 3850 features, 10 features were used in the training process.

  Features: ['H-049', 'P_VSA_MR_8', 'F03[N-Cl]', 'F01[C-N]', 'B03[N-Cl]', 'T(N..N)', 'PCR', 'SaaN', 'SsssCH', 'Eig02_EA(dm)']

- With this approach, the following results were obtained.

  - Accuracy: 0.872
  - Precision: 1.0
  - Recall: 0.872
  - Fscore: 0.932

Part 2:

- All features were used in training the classifier and the results obtained with this approach are as follows:

  - Accuracy: 0.840
  - Precision: 1.0
  - Recall: 0.840
  - Fscore: 0.913