Практикум

Лоскутова Софья, 615 группа, химический факультет

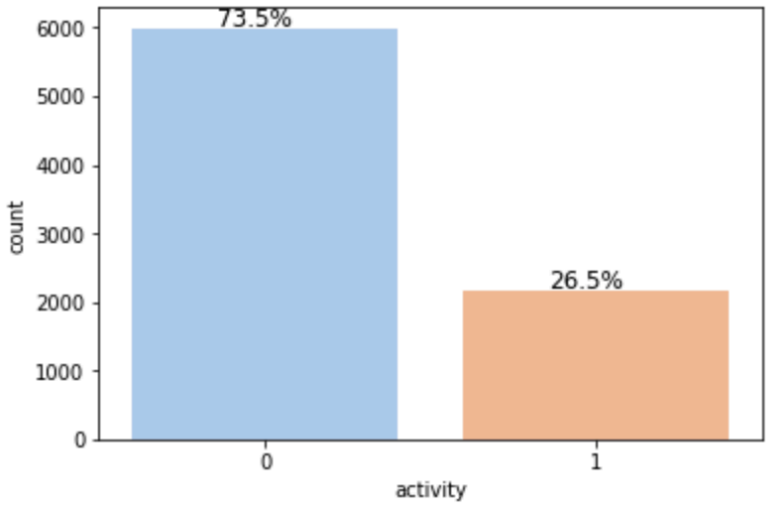
https://colab.research.google.com/drive/1iluXPpJPdqZJ9KTewgowWVNEIw2Gj72d?usp=sharing

Обзор данных

Общий размер датасета = 8154

Количество молекул, для которых дана активность по отношению к hERG гену = 8153

Распределение по активности (1 — молекула активна, 0 — молекула неактивна)



Из графика видно, что классы не сбалансированы.

Выбор модели и дескриптора

Сначала будем сравнивать модели по метрике ***accuracy_score***

	GetAvalonFP	GetHashedAtomPairFingerprint	GetMACCSKeysFingerprint
GaussianNB	0.6309	0.4304	0.4727
ComplementNB	0.7376	0.7223	0.6891
RandomForestClassifier	0.8921	**0.9007**	0.8847
BaggingClassifier	0.8743	0.8811	0.8780
XGBClassifier	0.8614	0.8872	0.8565

Наибольшее значение соответсвует дескриптору `GetHashedAtomPairFingerprint` и модели `RandomForestClassifier`, остановимся на них и будем варьировать параметры.

Параметры

Сразу зададим `class_weight='balanced_subsample'`, так как данные несбалансированные

Проверялись следующие значения гиперпараметров:

```
criterion=«entropy» / default=”gini”
n_estimators=50, 100, 200, 500, 1000, 1500
max_features='log2' / default=”auto”
```

`warm_start=True`

Для всех случаев метрики отличались на сотые доли, например:

class_weight=‘balanced'		
accuracy_score	f1_score	matthews_corrcoef
0.8976	0.7951	0.7342

class_weight='balanced_subsample',criterion="entropy",n_estimators=500, max_features='log2'		
accuracy_score	f1_score	matthews_corrcoef
0.8958	0.7870	0.7289

class_weight='balanced_subsample',criterion="entropy",n_estimators=500, warm_start=True		
accuracy_score	f1_score	matthews_corrcoef
0.9037	0.8088	0.7508

Кажется, что значение `n_estimators` больше 100 и `warm_start=True`, дает чуть лучше результат. Остановимся на модели

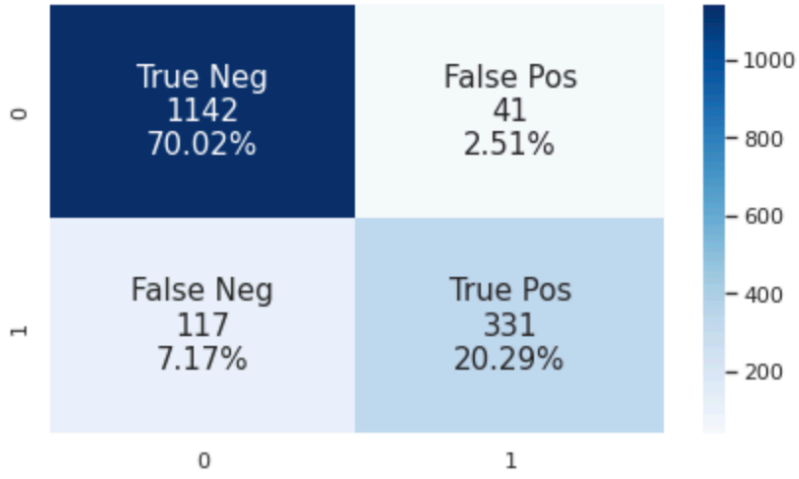
```
RandomForestClassifier(class_weight='balanced_subsample',criterion
="entropy",n_estimators=500, warm_start=True)
```

Посчитанные метрики для нее:

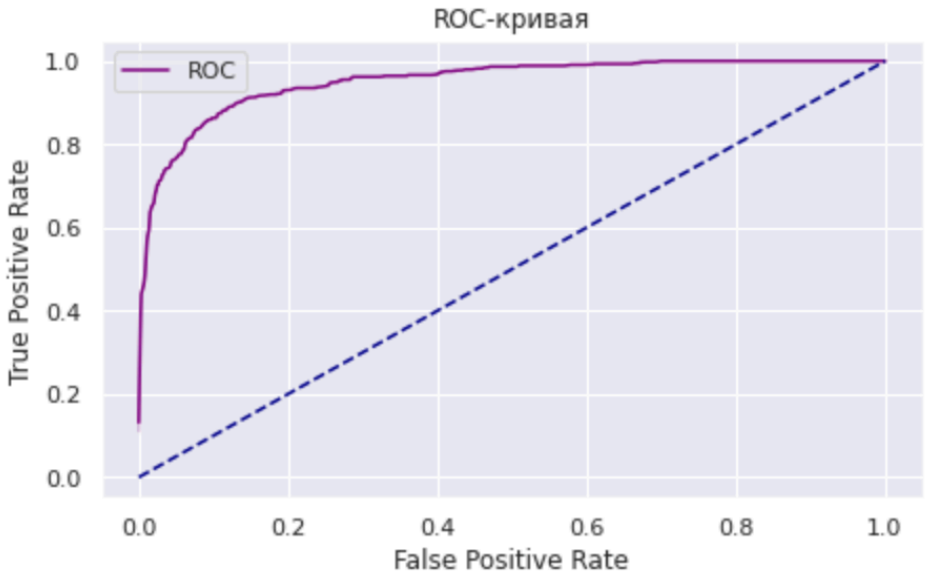
```
accuracy_score = 0.9031
precision_score = 0.8898
recall_score = 0.7388
f1_score = 0.8073
matthews_corrcoef = 0.7491
```

Confusion_matrix и roc-кривая

Тепловая карта confusion_matrix:



Roc-кривая:


ROC-кривая
ROC
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0