

Бинарная классификация движения цен по новостному потоку

Борисов Александр

9 декабря 2018 г.

0.1 Сравнение классификаторов на разных призна- ках

| Classifier | Features | Data Set | F1 Score | AUC ROC | Average | |
|------------|----------|----------|----------|---------|---------------|---------------|
| | | | | | F1 Score | AUC ROC |
| RF | Unigrams | 1 | 0.7811 | 0.7181 | 0.8093 | 0.5565 |
| | Unigrams | 2 | 0.8061 | 0.4973 | | |
| | Unigrams | 3 | 0.8408 | 0.4541 | | |
| | NMF 50 | 1 | 0.7397 | 0.6080 | 0.7647 | 0.5394 |
| | NMF 50 | 2 | 0.8087 | 0.5000 | | |
| | NMF 50 | 3 | 0.7458 | 0.5102 | | |
| | NMF 100 | 1 | 0.7602 | 0.5841 | 0.7984 | 0.5487 |
| | NMF 100 | 2 | 0.8061 | 0.4973 | | |
| | NMF 100 | 3 | 0.8288 | 0.5648 | | |
| | NMF 200 | 1 | 0.7720 | 0.7235 | 0.7996 | 0.5838 |
| | NMF 200 | 2 | 0.8087 | 0.5000 | | |
| | NMF 200 | 3 | 0.8180 | 0.5278 | | |
| | Ensemble | 1 | 0.7907 | 0.7198 | 0.8045 | 0.5617 |
| | Ensemble | 2 | 0.8018 | 0.5006 | | |
| | Ensemble | 3 | 0.821 | 0.4648 | | |

Таблица 1: RandomForestClassifier on 3 data sets

| Classifier | Features | Data Set | F1 Score | AUC ROC | Average | |
|------------|----------|----------|----------|---------|---------------|---------------|
| | | | | | F1 Score | AUC ROC |
| XGB | Unigrams | 1 | 0.8371 | 0.7623 | 0.835 | 0.5805 |
| | Unigrams | 2 | 0.8035 | 0.4946 | | |
| | Unigrams | 3 | 0.8643 | 0.4846 | | |
| | NMF 50 | 1 | 0.8239 | 0.7508 | 0.8284 | 0.5716 |
| | NMF 50 | 2 | 0.8035 | 0.4946 | | |
| | NMF 50 | 3 | 0.8577 | 0.4693 | | |
| | NMF 100 | 1 | 0.7989 | 0.7054 | 0.8257 | 0.5586 |
| | NMF 100 | 2 | 0.8035 | 0.4946 | | |
| | NMF 100 | 3 | 0.8747 | 0.4759 | | |
| | NMF 200 | 1 | 0.7923 | 0.6815 | 0.8221 | 0.5617 |
| | NMF 200 | 2 | 0.8061 | 0.4973 | | |
| | NMF 200 | 3 | 0.8679 | 0.5063 | | |
| | Ensemble | 1 | 0.8046 | 0.7314 | 0.8217 | 0.5680 |
| | Ensemble | 2 | 0.8035 | 0.4946 | | |
| | Ensemble | 3 | 0.8571 | 0.4780 | | |
| LR | Unigrams | 1 | 0.8217 | 0.6873 | 0.8464 | 0.5624 |
| | Unigrams | 2 | 0.8087 | 0.5000 | | |
| | Unigrams | 3 | 0.9087 | 0.5000 | | |
| | NMF 50 | 1 | 0.8235 | 0.6831 | 0.8470 | 0.5610 |
| | NMF 50 | 2 | 0.8087 | 0.5000 | | |
| | NMF 50 | 3 | 0.9087 | 0.5000 | | |
| | NMF 100 | 1 | 0.8154 | 0.6724 | 0.8443 | 0.5575 |
| | NMF 100 | 2 | 0.8087 | 0.5000 | | |
| | NMF 100 | 3 | 0.9087 | 0.5000 | | |
| | NMF 200 | 1 | 0.8244 | 0.6811 | 0.8473 | 0.5604 |
| | NMF 200 | 2 | 0.8087 | 0.5000 | | |
| | NMF 200 | 3 | 0.9087 | 0.5000 | | |
| | Ensemble | 1 | 0.8214 | 0.6782 | 0.8463 | 0.5594 |
| | Ensemble | 2 | 0.8087 | 0.5000 | | |
| | Ensemble | 3 | 0.9087 | 0.5000 | | |

Таблица 2: XGBClassifier & LogisticRegression on 3 data sets

| Classifier | Features | Data Set | F1 Score | AUC ROC | Average | |
|------------|----------|----------|----------|---------|---------------|---------------|
| | | | | | F1 Score | AUC ROC |
| LSVC | Unigrams | 1 | 0.7952 | 0.5957 | 0.8309 | 0.5406 |
| | Unigrams | 2 | 0.8087 | 0.5000 | | |
| | Unigrams | 3 | 0.8889 | 0.5260 | | |
| | NMF 50 | 1 | 0.8049 | 0.6204 | 0.8349 | 0.5495 |
| | NMF 50 | 2 | 0.8087 | 0.5000 | | |
| | NMF 50 | 3 | 0.8912 | 0.5281 | | |
| | NMF 100 | 1 | 0.7933 | 0.5907 | 0.8310 | 0.5396 |
| | NMF 100 | 2 | 0.8087 | 0.5000 | | |
| | NMF 100 | 3 | 0.8912 | 0.5281 | | |
| | NMF 200 | 1 | 0.7962 | 0.5936 | 0.8312 | 0.5399 |
| | NMF 200 | 2 | 0.8087 | 0.5000 | | |
| | NMF 200 | 3 | 0.8889 | 0.5260 | | |
| | Ensemble | 1 | 0.8029 | 0.6155 | 0.8343 | 0.5479 |
| | Ensemble | 2 | 0.8087 | 0.5000 | | |
| | Ensemble | 3 | 0.8912 | 0.5281 | | |

Таблица 3: LinearSVC on 3 data sets

Будем обозначать модели Random Forest Classifier, XGB Classifier, Logistic Regression и Linear SVC как RF, XGB, LR и LSVC, соответственно. Сравнив модели (RF, XGB, LR, LSVC) на разных признаках (Unigrams, NMF 50, NMF 100, NMF 200, Ensemble) выберем признаки, на которых модели давали лучший результат по F1 Score. Такими оказались: Unigrams (для моделей RF и XGB), NMF 50 (для модели LSVC) и NMF 200 (для модели LR). Далее будем оптимизировать модели с этими признаками (RF с Unigrams, XGB с Unigrams, LSVC с NMF 50 и LR с NMF 200) по гиперпараметрам.

0.2 RandomForestClassifier с Unigrams

| | | | | | | | Average | | |
|-----------|------------------|-------------------|--------------|----------|----------|---------|----------|---------|--|
| max_depth | min_samples_leaf | min_samples_split | n_estimators | Data Set | F1 Score | AUC ROC | F1 Score | AUC ROC | |
| None | 3 | 5 | 2000 | 1 | 0.8677 | 0.7710 | 0.8615 | 0.5961 | |
| | | | | 2 | 0.8087 | 0.5000 | | | |
| | | | | 3 | 0.9080 | 0.5174 | | | |
| None | 3 | 2 | 2000 | 1 | 0.8663 | 0.7751 | 0.8611 | 0.5946 | |
| | | | | 2 | 0.8087 | 0.5000 | | | |
| | | | | 3 | 0.9084 | 0.5087 | | | |
| 10 | 3 | 2 | 1000 | 1 | 0.8661 | 0.7640 | 0.8611 | 0.5909 | |
| | | | | 2 | 0.8087 | 0.5000 | | | |
| | | | | 3 | 0.9084 | 0.5087 | | | |
| None | 3 | 2 | 1000 | 1 | 0.8647 | 0.7681 | 0.8605 | 0.5952 | |
| | | | | 2 | 0.8087 | 0.5000 | | | |
| | | | | 3 | 0.9080 | 0.5174 | | | |
| 50 | 3 | 5 | 2000 | 1 | 0.8640 | 0.7702 | 0.8604 | 0.5930 | |
| | | | | 2 | 0.8087 | 0.5000 | | | |
| | | | | 3 | 0.9084 | 0.5087 | | | |
| 10 | 3 | 5 | 2000 | 1 | 0.8639 | 0.7591 | 0.8603 | 0.5892 | |
| | | | | 2 | 0.8087 | 0.5000 | | | |
| | | | | 3 | 0.9084 | 0.5087 | | | |
| 50 | 3 | 2 | 2000 | 1 | 0.8647 | 0.7681 | 0.8599 | 0.5915 | |
| | | | | 2 | 0.8087 | 0.5000 | | | |
| | | | | 3 | 0.9062 | 0.5065 | | | |
| 50 | 3 | 2 | 1000 | 1 | 0.8624 | 0.7632 | 0.8597 | 0.5935 | |
| | | | | 2 | 0.8087 | 0.5000 | | | |
| | | | | 3 | 0.9080 | 0.5174 | | | |
| None | 3 | 5 | 1000 | 1 | 0.8624 | 0.7632 | 0.8597 | 0.5935 | |
| | | | | 2 | 0.8087 | 0.5000 | | | |
| | | | | 3 | 0.9080 | 0.5174 | | | |

Таблица 4: RandomForestClassifier with Unigrams

0.3 XGBClassifier c Unigrams

| gamma | learning_rate | max_depth | n_estimators | Data Set | Average | | |
|-------|---------------|-----------|--------------|----------|----------|---------|---------|
| | | | | | F1 Score | AUC ROC | AUC ROC |
| None | | | | 1 | | | |
| | | | | 2 | | | |
| | | | | 3 | | | |
| None | | | | 1 | | | |
| | | | | 2 | | | |
| | | | | 3 | | | |
| None | | | | 1 | | | |
| | | | | 2 | | | |
| | | | | 3 | | | |
| None | | | | 1 | | | |
| | | | | 2 | | | |
| | | | | 3 | | | |
| None | | | | 1 | | | |
| | | | | 2 | | | |
| | | | | 3 | | | |
| None | | | | 1 | | | |
| | | | | 2 | | | |
| | | | | 3 | | | |
| None | | | | 1 | | | |
| | | | | 2 | | | |
| | | | | 3 | | | |
| None | | | | 1 | | | |
| | | | | 2 | | | |
| | | | | 3 | | | |
| None | | | | 1 | | | |
| | | | | 2 | | | |
| | | | | 3 | | | |

Таблица 5: XGBClassifier with Unigrams

0.4 SVC c NMF

| | | | | | | Average | | |
|-------|---------------|-----------|--------------|----------|----------|---------|----------|---------|
| gamma | learning_rate | max_depth | n_estimators | Data Set | F1 Score | AUC ROC | F1 Score | AUC ROC |
| 0 | 0.1 | 3 | 50 | 1 | 0.8291 | 0.7495 | 0.8419 | 0.5824 |
| | | | | 2 | | | | |
| | | | | 3 | 0.8880 | 0.4977 | | |
| 0.2 | 0.1 | 9 | 50 | 1 | 0.8391 | 0.7269 | 0.8407 | 0.5705 |
| | | | | 2 | 0.8087 | 0.5000 | | |
| | | | | 3 | 0.8742 | 0.4846 | | |
| 0.2 | 0.1 | 9 | 100 | 1 | 0.8457 | 0.7417 | 0.8398 | 0.5725 |
| | | | | 2 | 0.8087 | 0.5000 | | |
| | | | | 3 | 0.8649 | 0.4759 | | |
| 0.2 | 0.01 | 9 | 1000 | 1 | 0.8387 | 0.7380 | 0.8396 | 0.5764 |
| | | | | 2 | 0.8087 | 0.5000 | | |
| | | | | 3 | 0.8714 | 0.4911 | | |
| 0 | 0.1 | 9 | 500 | 1 | 0.8415 | 0.7520 | 0.8389 | 0.5794 |
| | | | | 2 | 0.8061 | 0.4973 | | |
| | | | | 3 | 0.8690 | 0.4889 | | |
| 0 | 0.1 | 9 | 1000 | 1 | 0.8343 | 0.7483 | 0.8388 | 0.5804 |
| | | | | 2 | 0.8061 | 0.4973 | | |
| | | | | 3 | 0.8760 | 0.4955 | | |
| None | | | | 1 | | | | |
| | | | | 2 | | | | |
| | | | | 3 | | | | |
| None | | | | 1 | | | | |
| | | | | 2 | | | | |
| | | | | 3 | | | | |
| None | | | | 1 | | | | |
| | | | | 2 | | | | |
| | | | | 3 | | | | |

Таблица 6: XGBClassifier with Unigrams