

Бинарная классификация движения цен по новостному потоку

Ахияров Валентин

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.