

Отчет по домашней работе №1

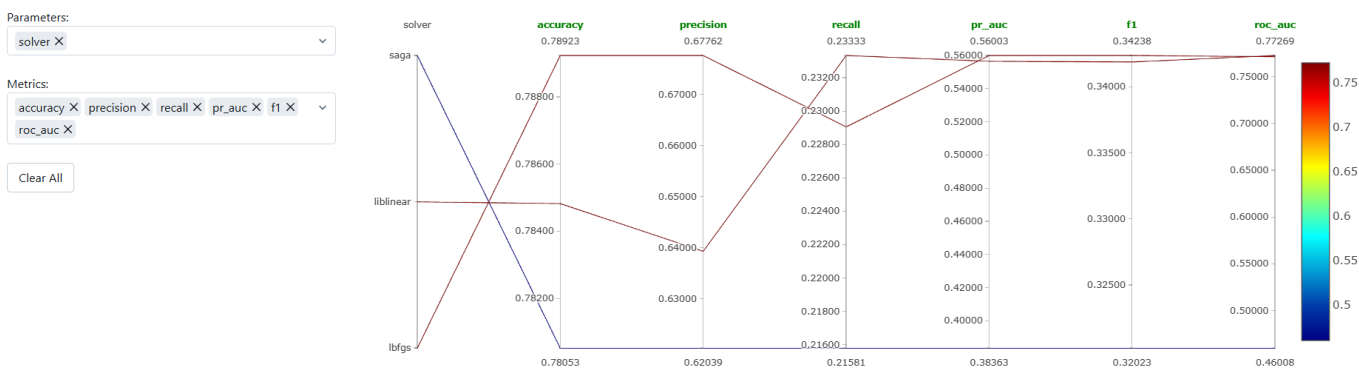
1. Разрез параметры модели для логистической регрессии — алгоритм оптимизации

Гипотеза — выбор алгоритма оптимизации влияет на скорость сходимости и итоговое качество модели логистической регрессии

Параметр — алгоритм оптимизации (solver)

меняется по сетке:

- lbfgs
- liblinear
- saga



Show diff only

l1_ratio	0.5		
penalty	l2	l1	elasticnet
solver	lbfgs	liblinear	saga
train_rows	1500	1000	1000

Metrics

Show diff only

accuracy	0.789	0.785	0.781
f1	0.342	0.342	0.32
pr_auc	0.56	0.557	0.384
precision	0.678	0.639	0.62
recall	0.229	0.233	0.216
roc_auc	0.771	0.773	0.46

Artifarte

Вывод: разные алгоритмы оптимизации показали различное качество на валидационной выборке. Наилучшие показатели метрик достигнуты при использовании liblinear, что подтверждает его эффективность для данной задачи и размера данных

2. Разрез тип модели

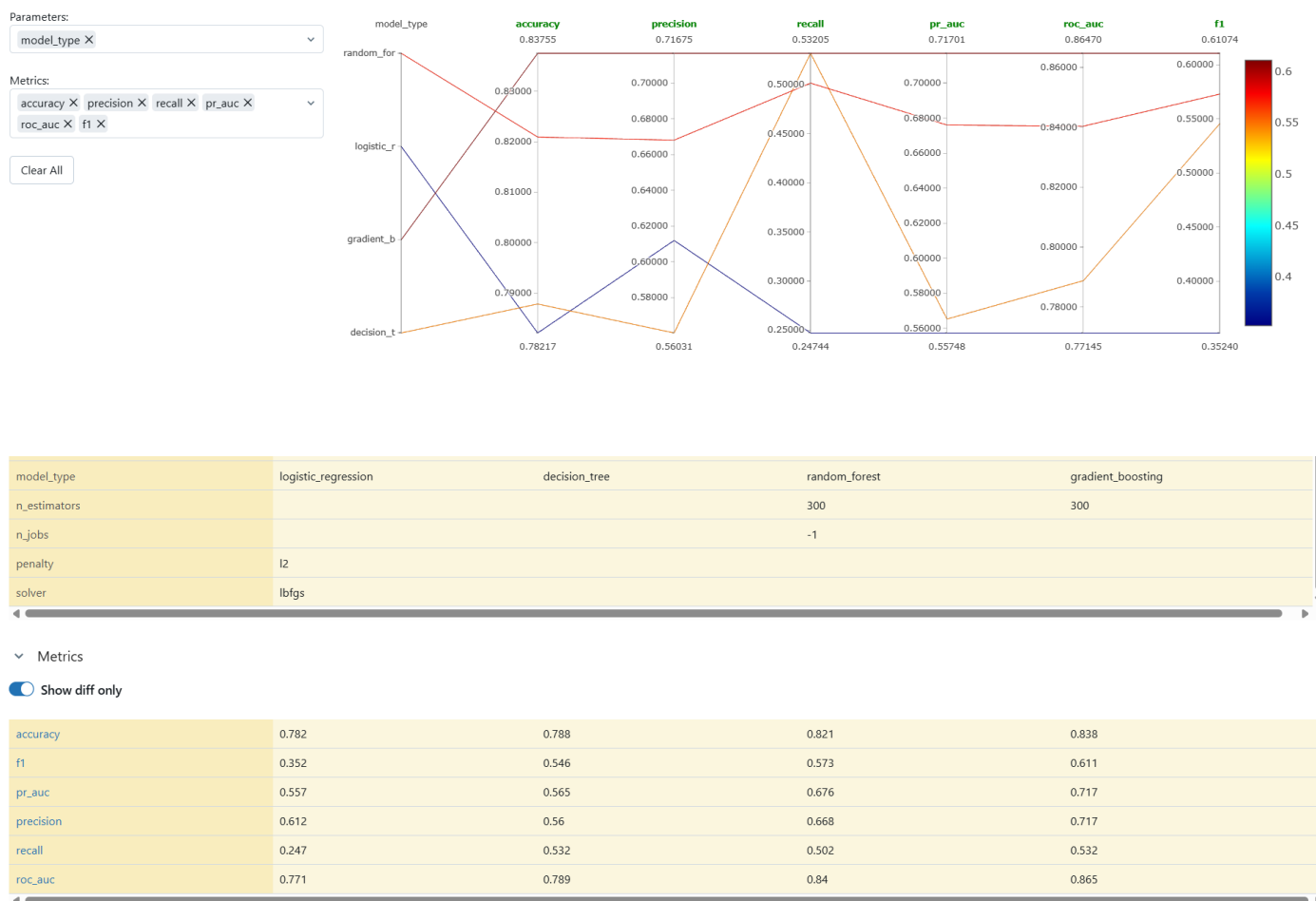
Гипотеза — более сложные нелинейные модели лучше улавливают зависимости в данных и повышают качество классификации

Параметр — тип модели (model_type)

меняется по сетке:

- logistic_regression
- decision_tree
- random_forest
- gradient_boosting

Примечание: предварительно для каждой модели были подобраны лучшие для нее параметры



Вывод: модели градиентного бустинга и случайного леса показали более высокое качество по сравнению с логистической регрессией и деревом решений. Это подтверждает, что нелинейные зависимости в данных существуют и ансамблевые методы успешно их улавливают.

3. Разрез набор фичей

Гипотеза — расширение набора признаков за счет добавления финансовых и демографических характеристик будет последовательно повышать качество модели

Параметр — набор признаков (feature)

меняется по сетке:

- набор 1 (6 признаков): age, workclass, sex, occupation, hours.per.week, education.num

um

- набор 2 (7 признаков): набор 1 + capital.gain

- набор 3 (7 признаков): набор 1 + native.country
- набор 4 (8 признаков): набор 1 + capital.gain + native.country



features				
	["age", "workclass", "sex", "native.country", "occupation", "hours.per.week", "education.num", "capital.gain"]	["age", "workclass", "sex", "occupation", "hours.per.week", "education.num"]	["age", "workclass", "sex", "occupation", "hours.per.week", "education", "capital.gain"]	["age", "workclass", "sex", "occupation", "hours.per.week", "education.num", "capital.gain"]
▼ Metrics				
⚙ Show diff only				
accuracy	0.835	0.802	0.834	0.838
f1	0.607	0.533	0.592	0.611
pr_auc	0.713	0.594	0.711	0.717
precision	0.706	0.613	0.717	0.717
recall	0.532	0.471	0.503	0.532
roc_auc	0.863	0.829	0.864	0.865

Вывод: набор 2 (с capital.gain) показал лучшее качество, что говорит о более высокой прогностической силе финансовых признаков по сравнению с демографическими для задачи предсказания дохода.

4. Ссылка на запуск с лучшим значением метрики ROC-AUC.

<http://158.160.2.37:5000/#/experiments/2/runs/04fb741eb5394148ab3b77a3ebfe24>