

## Отчет по домашней работе №1

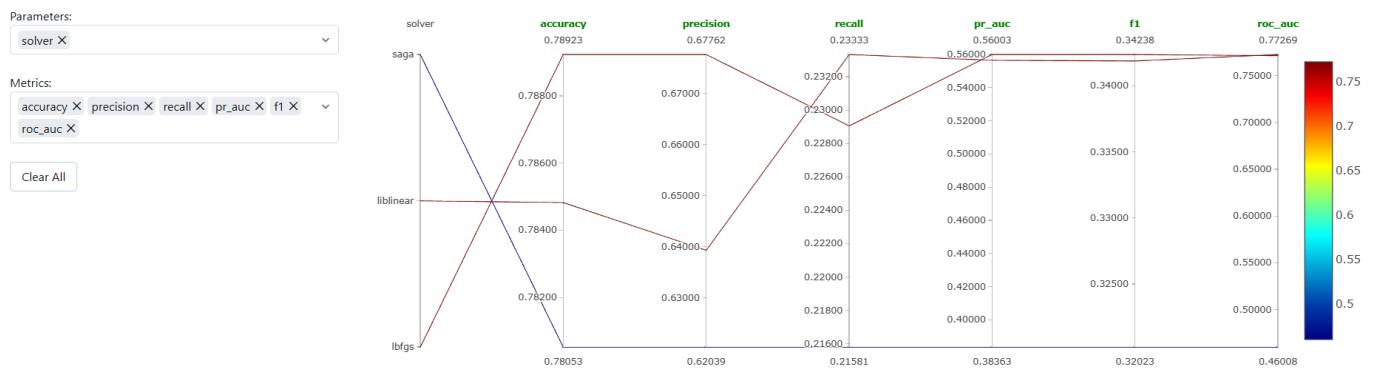
1. Разрез параметры модели для логистической регрессии – алгоритм оптимизации

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

Параметр – алгоритм оптимизации (solver)

меняется по сетке:

- lbfsgs
- liblinear
- saga



I1\_ratio

I1_ratio	0.5		
penalty	l2	I1	elasticnet
solver	lbfsgs	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

Artifacts

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

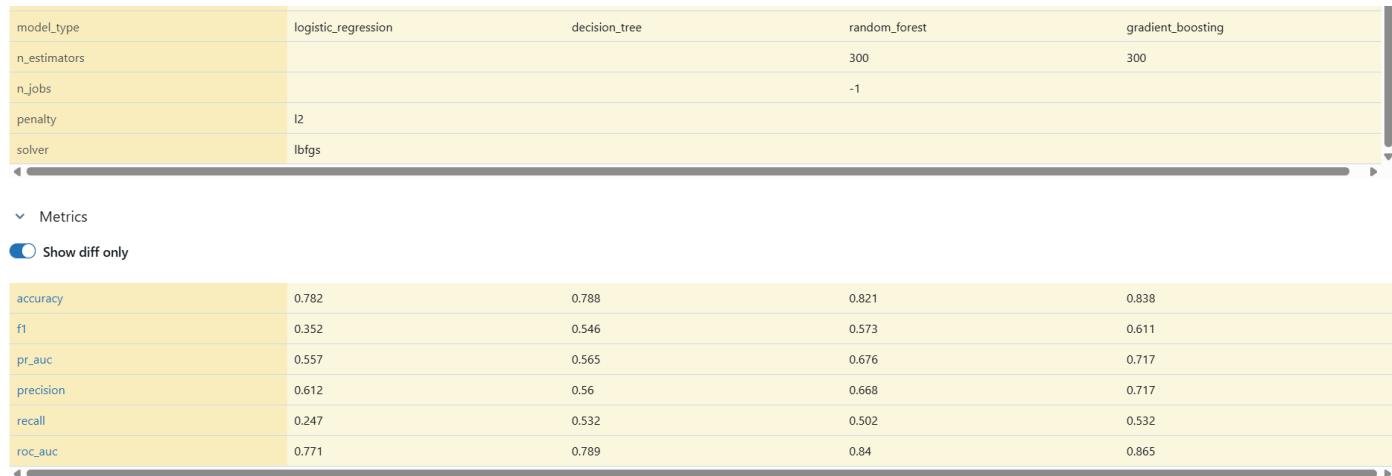
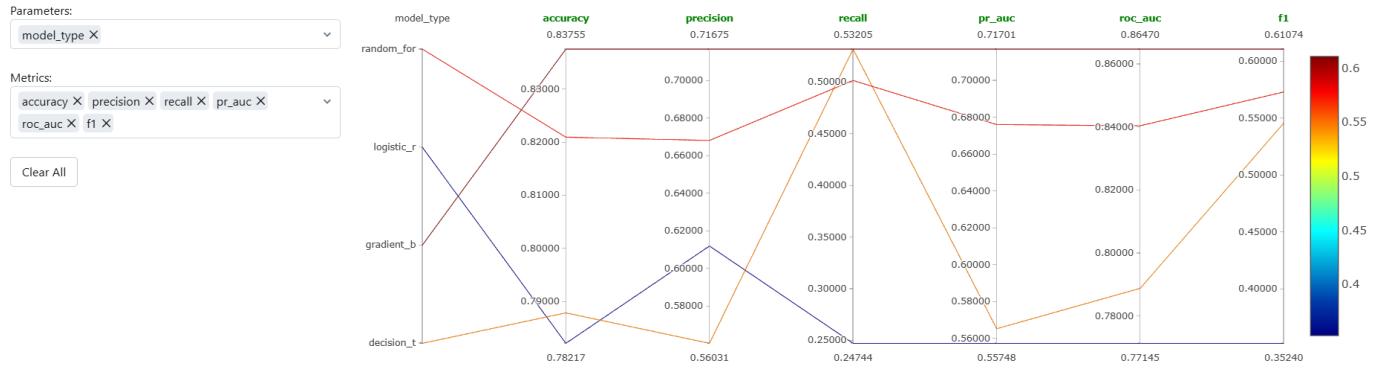
Гипотеза – более сложные нелинейные модели лучше улавливают зависимости в данных и повышают качество классификации

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

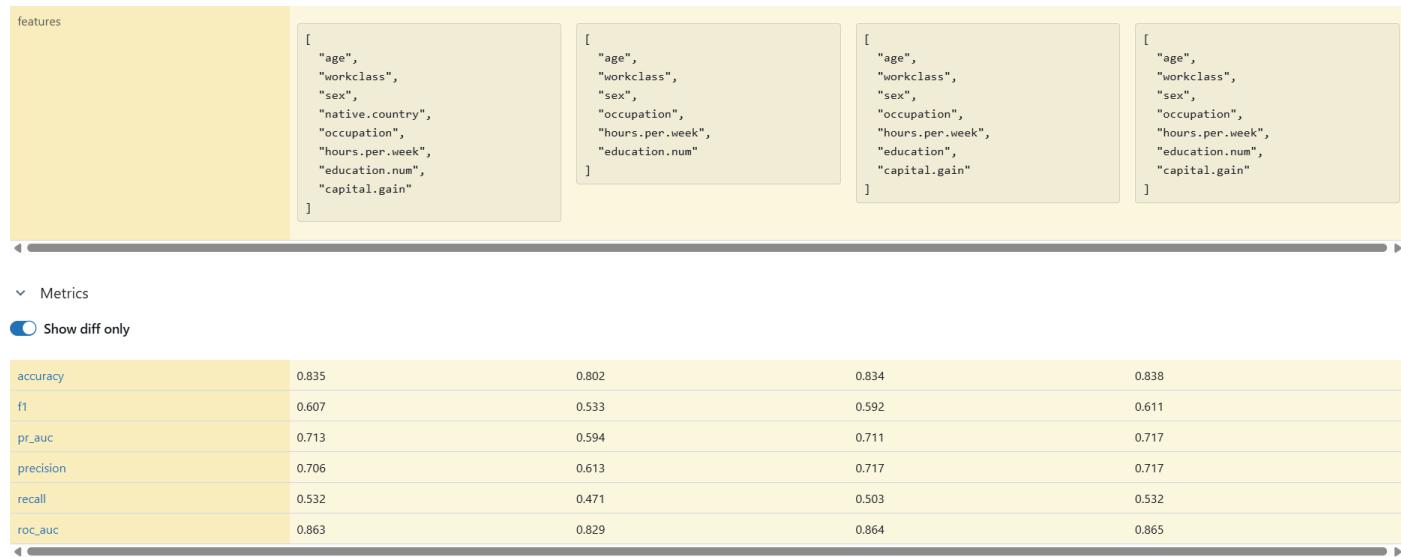
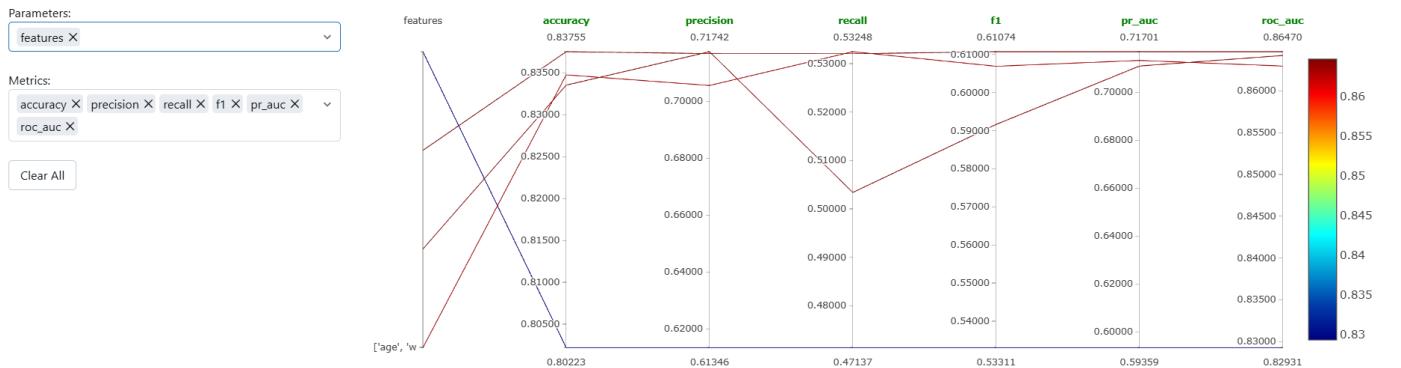
Гипотеза – расширение набора признаков за счет добавления финансовых и демографических характеристик будет последовательно повышать качество модели

Параметр – набор признаков (feature)

меняется по сетке:

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



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>