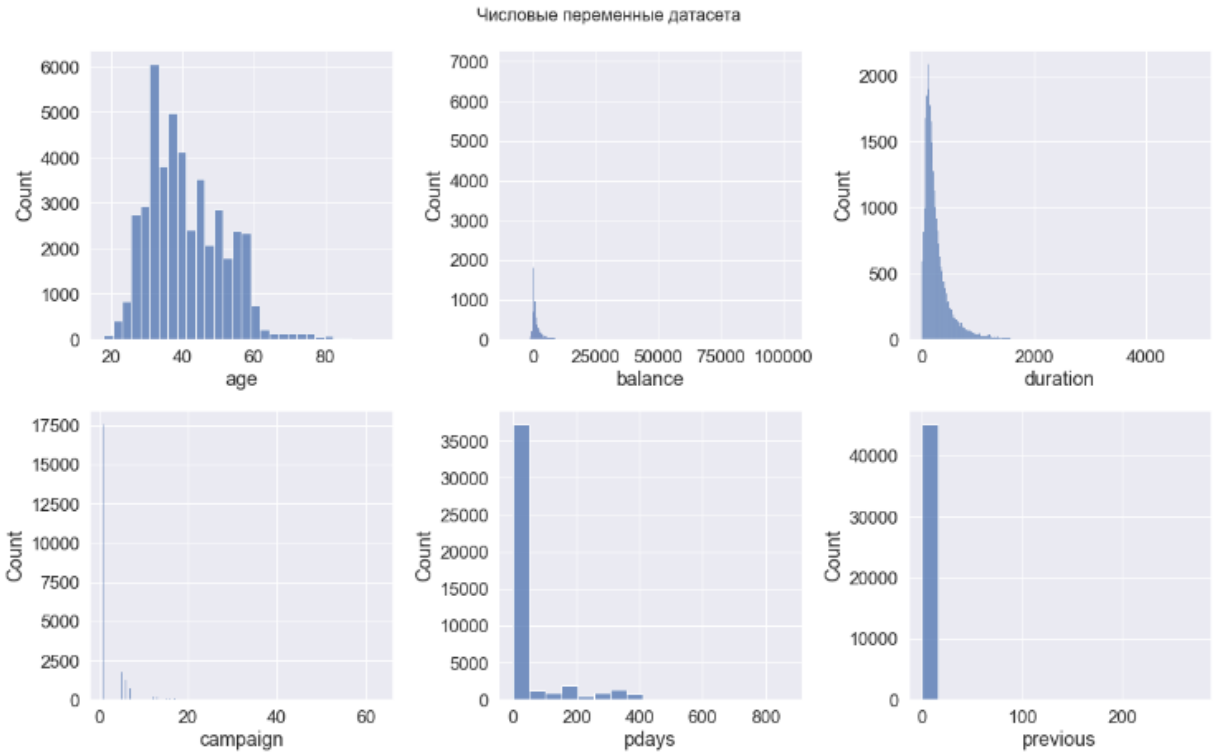
1 EXPLORATORY DATA ANALYSIS

2 IMPLEMENTED ML ALGORITHMS OUTLOOK

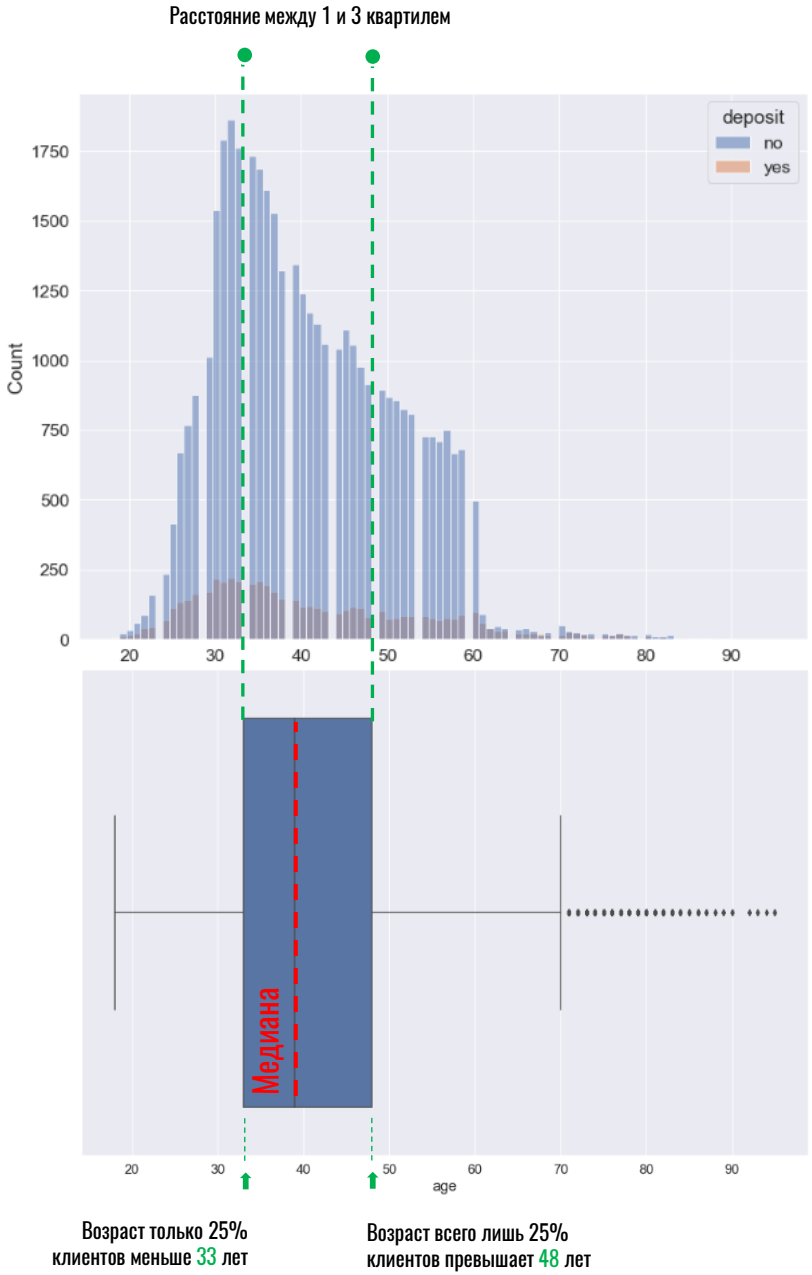
3 CONCLUSION



Выявление выбросов

В ряде числовых переменных присутствует существенное смещение плотности распределения.

Учитывая, что в рамках данной работы будут использоваться только Decision tree, Random forest, Logistic Regression алгоритмы, нормализация данных необязательна.



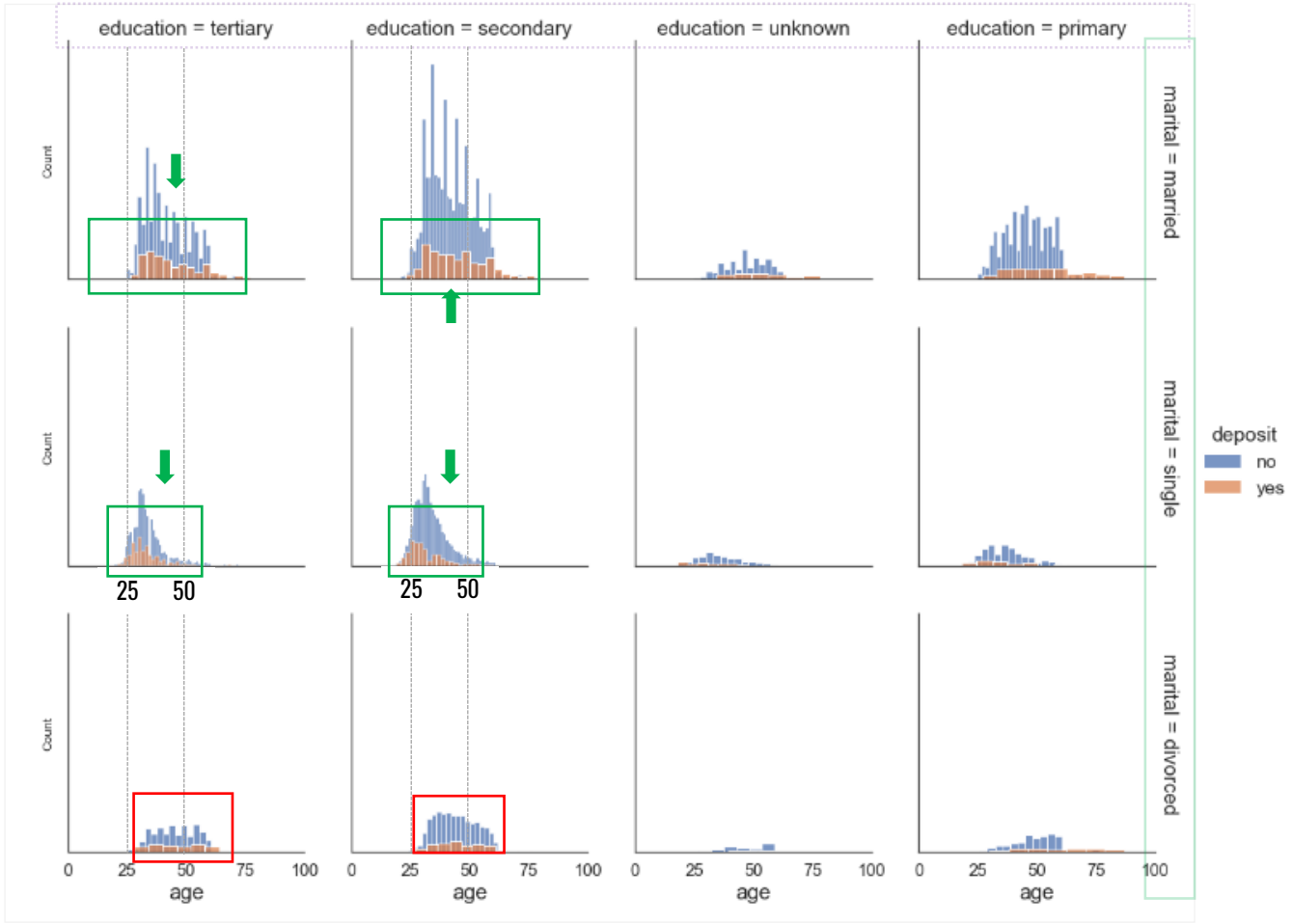
Возрастное распределение

С точки зрения возраста нет явного отличия между выборкой людей, которые согласились на депозит и генеральной совокупностью. Все статистические показатели для обеих групп пользователей практически идентичны:

- Межквартильный размах возрастов клиентов – между 33 и 48 годами
- Медианный возраст – 39 лет.

Возрастное распределение – с привязкой к конкретным признакам

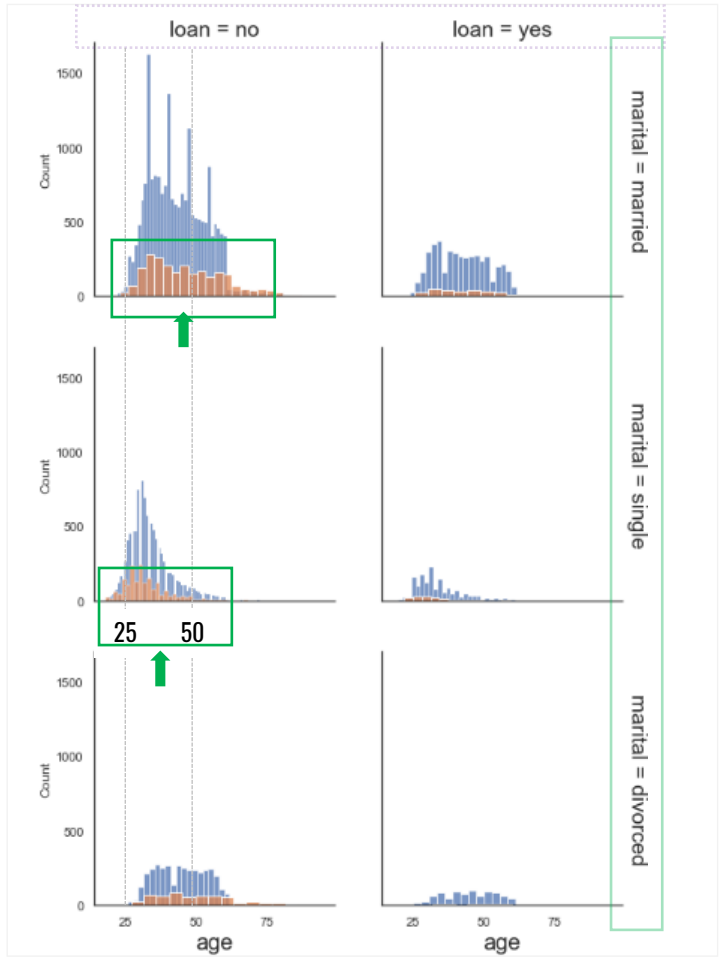
Уровень образования



Семейное положение

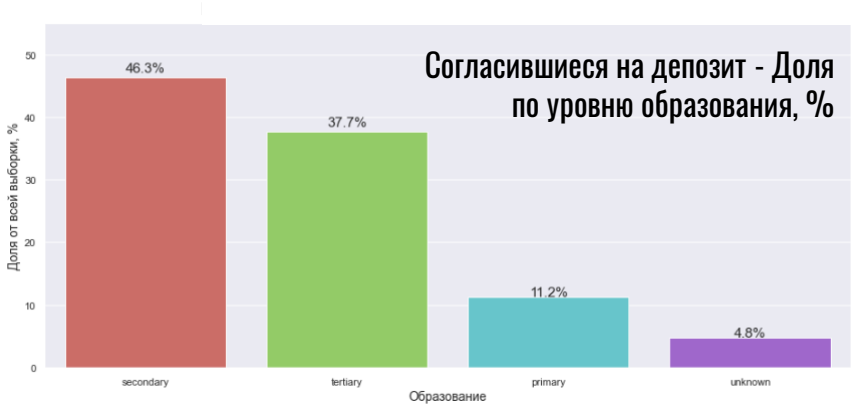
- Высокая концентрация согласившихся на депозит среди Married, Single,
- В первую очередь – для людей с education tertiary, secondary.
- Для Married согласившиеся на кредит – преимущественно люди в возрасте 27 - 55 лет.
- Для Single – 25 – 40 лет.
- Для divorced распределение возраста согласившихся на депозит смещено правее – на интервал 30 – 65 лет.

Есть ли займ



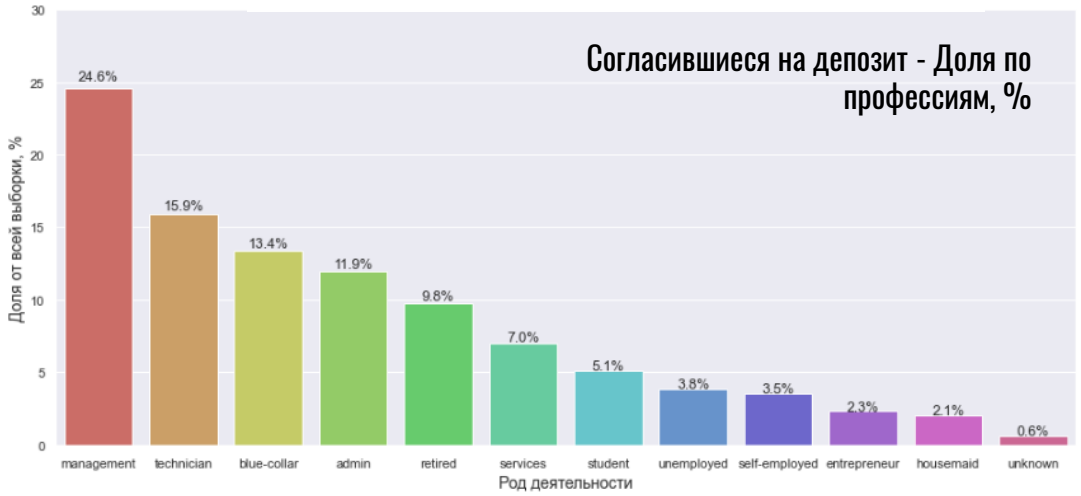
Семейное положение

Абсолютное большинство клиентов, соасившихся на депозит, не имели кредита.



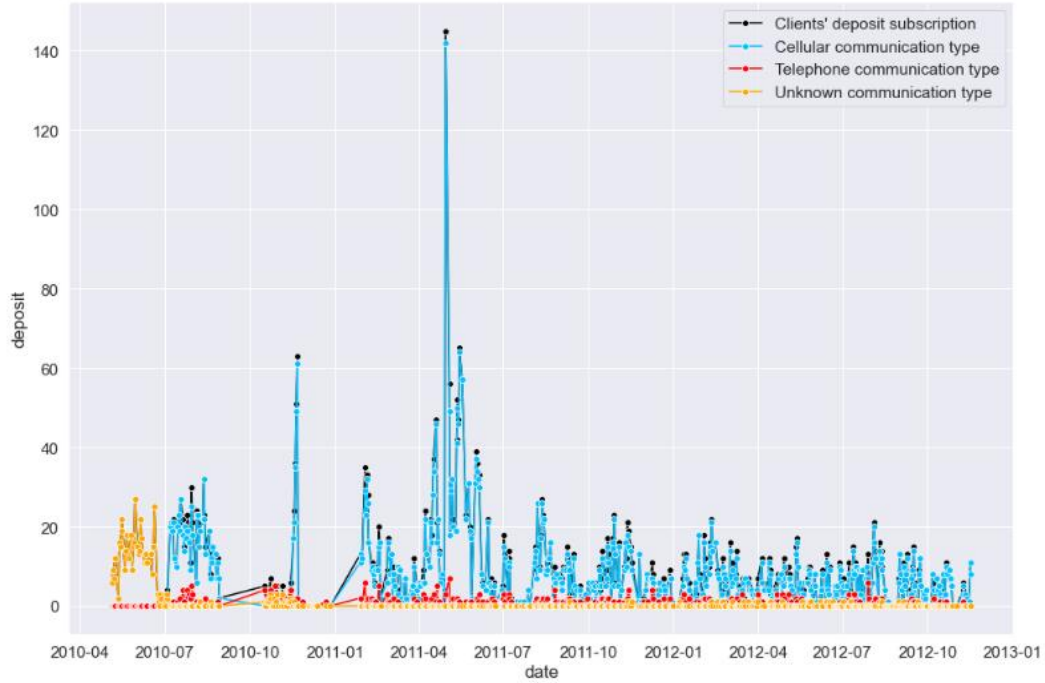
Согласившиеся на депозит:

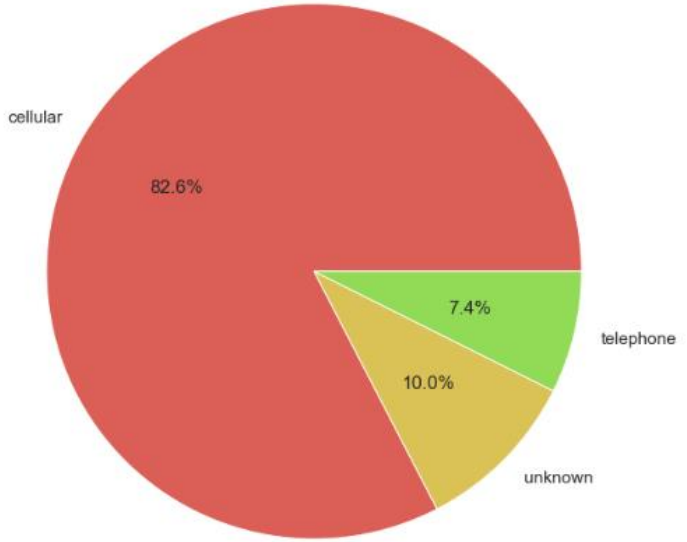
- 57.5% – люди без высшего образования. Это люди преимущественно 4 профессий: admin, blue-collar, technician, services.
- Абсолютное большинство клиентов, согласившихся на депозит и имевших высшее образование, - работали в сфере management, technician.



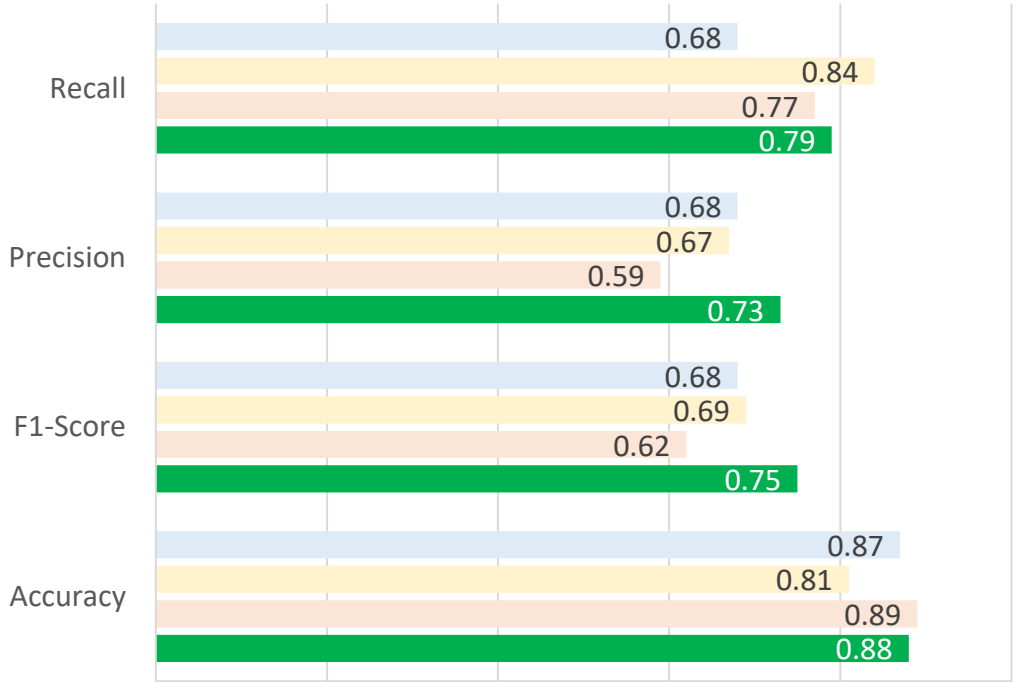
ТИП КОНТАКТА С КЛИЕНТАМИ, КОТОРЫЕ СОГЛАСИЛИСЬ НА ДЕПОЗИТ

- Из всех клиентов, согласившихся на депозит, 82.6% имели контакт через cellular тип связи и только 7.4% - через telephone.
- Отчётливо виден всплеск активности Cellular communication type в мае 2011 года, который повлёк за собой самый массовый приток клиентов, согласившихся на депозит



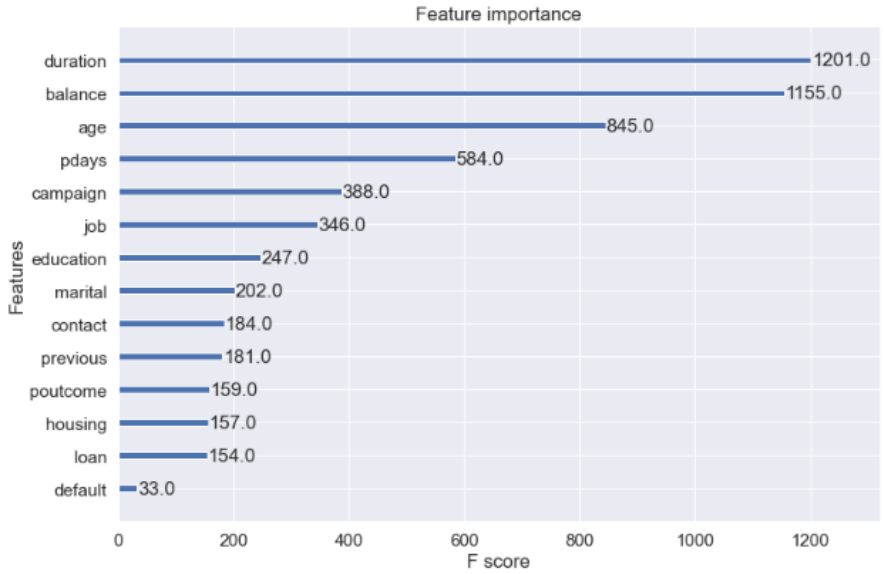


Сравнение целевых показателей точности моделей



	Accuracy	F1-Score	Precision	Recall
Decision Tree	0.87	0.68	0.68	0.68
Random Forest	0.81	0.69	0.67	0.84
Logistic	0.89	0.62	0.59	0.77
XGBoost	0.88	0.75	0.73	0.79

Несмотря на то, что XGBoost проигрывает Random Forest по показателю Recall (То есть хуже фильтрует ложноположительные результаты), в целом именно XGBoost является самой сбалансированной моделью по суммарному зачёту всех 4 метрик.



На какие метрики чаще всего опиралась модель для снижения уровня энтропии в процессе ветвления обучающих деревьев

Точность модели при threshold = 0.67 (вся прогнозная вероятность выше 67% считается положительным исходом)