Digital Reputation Challenge

10 сентября - 10 октября 2019

Татьяна Некрасова

3 место

Задача:

- Предсказать 5 свойств характера человека
- Данные (анонимизированные):

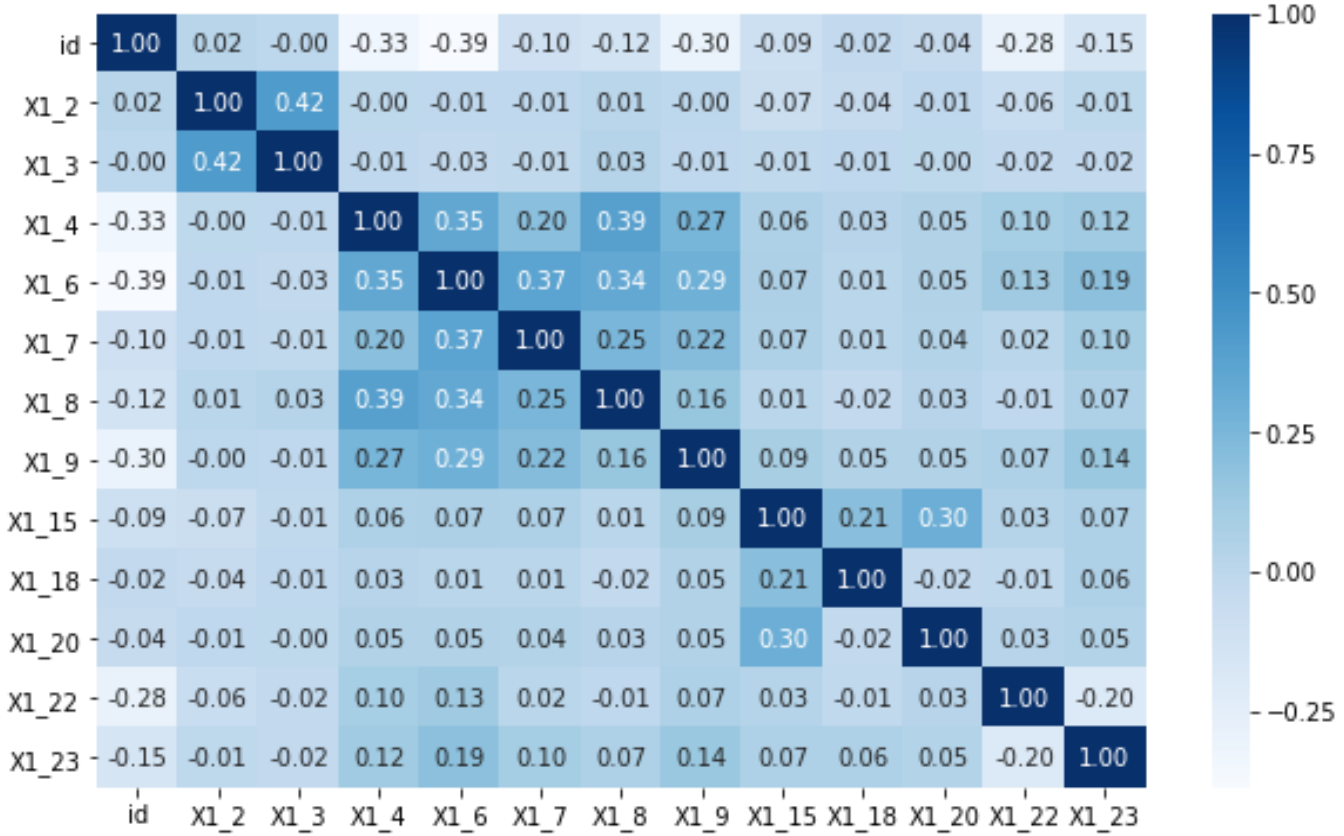
  X1: 25 переменных (количественных/категориальных + id)

  X2: каждому пользователю несколько A (id сайтов?)

  X3: 452 количественных переменных (по X2?)
- Train: 4000 пользователей
- Test: 4058 пользователей
- Задача: классификации
- Метрика: ROC-AUC (~0,6)

Анализ данных

X1 — много связанных по группам признаков



X2 — уникальных значений „A“: всего 214 тысяч,

в train 134 тысячи, в test 132 тысячи

Предобработка данных

X1 — np.log для колонок ['4', '5', '6', '7', '9'];
StandardScaler, TruncatedSVD
(только для линейной модели)

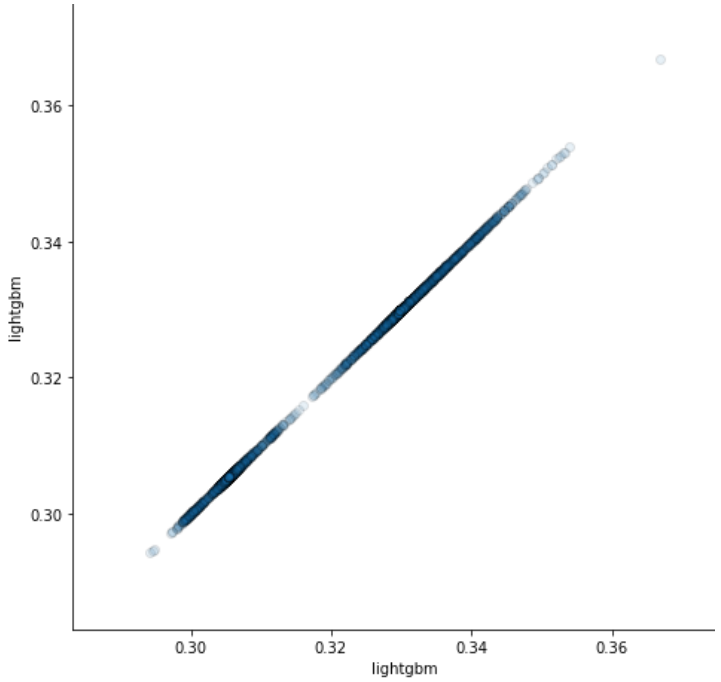
X2 — TfidfVectorizer (ngram (1,9), features 50000). PCA, UMAP и т.д. не давали результата.

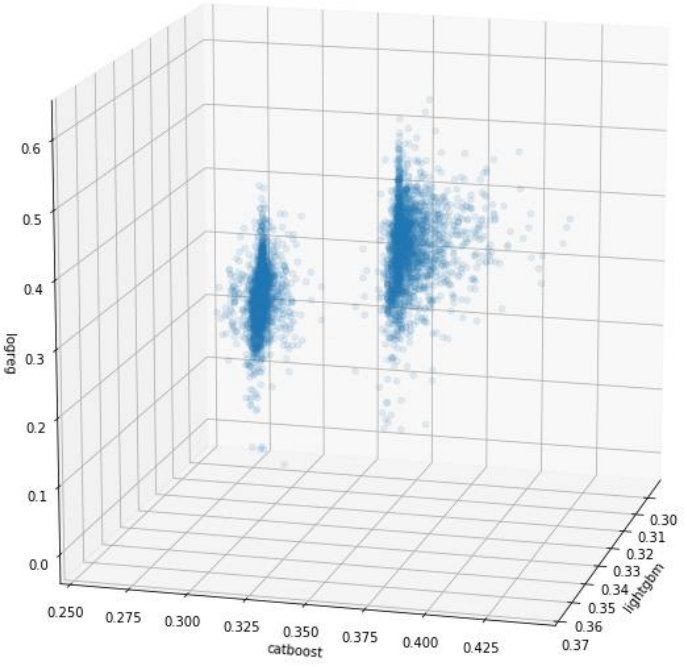
X3 — не использовала. Переобучение под публичный лидерборд.

Выбор моделей

- LogisticRegression — хорошо улавливает линейные связи. Отличие cv от pub ~ 0,01, priv 0,005.
- Catboost - симметричные неглубокие деревья (depth: 2). Отличие cv от pub ~ 0,005, priv 0,01.
- Lightgbm - глубокие деревья (depth без ограничений, регулирование по num_leaves). Отличие cv от pub ~ 0,0005, priv 0,001.

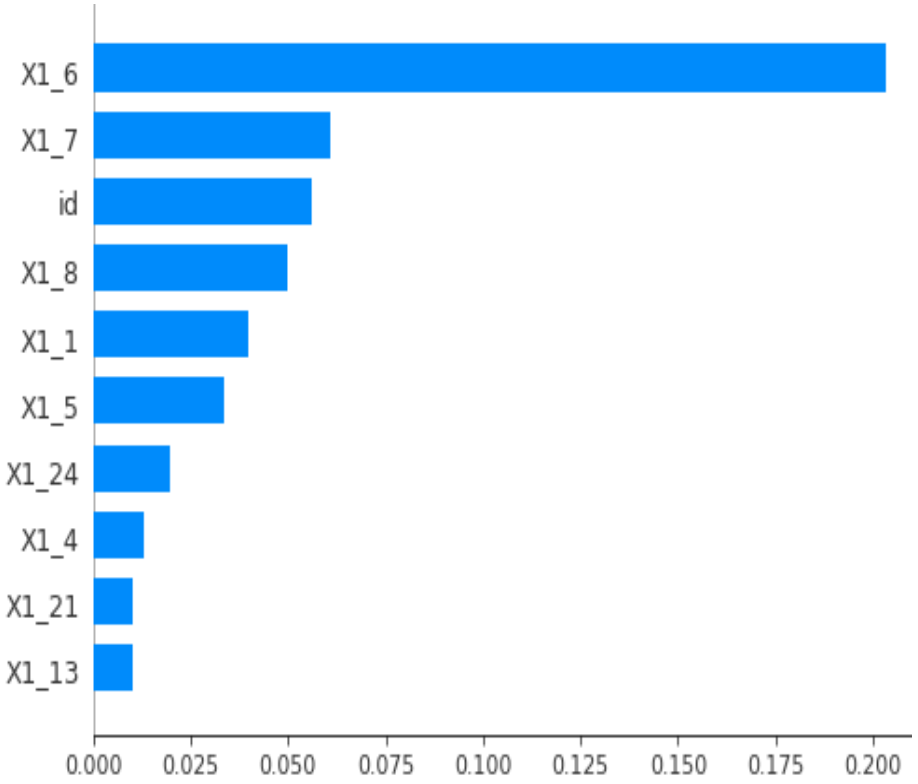
Усреднение трех моделей позволяет выделить два облака точек для y3 и y4. (на рисунке y4).

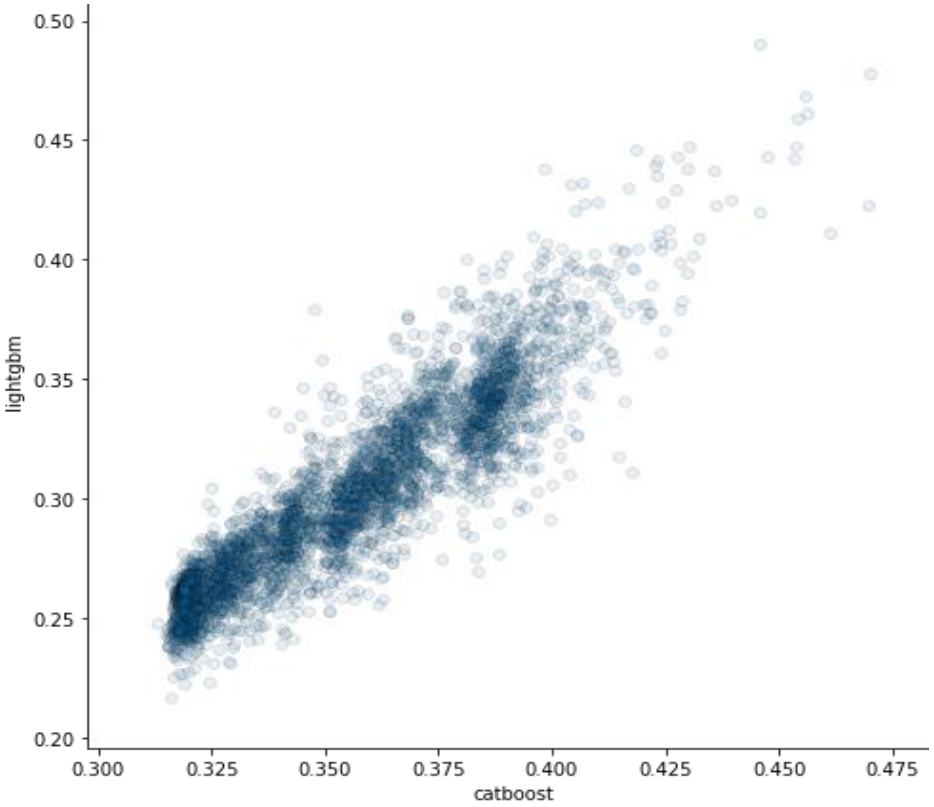




Черта характера у 1

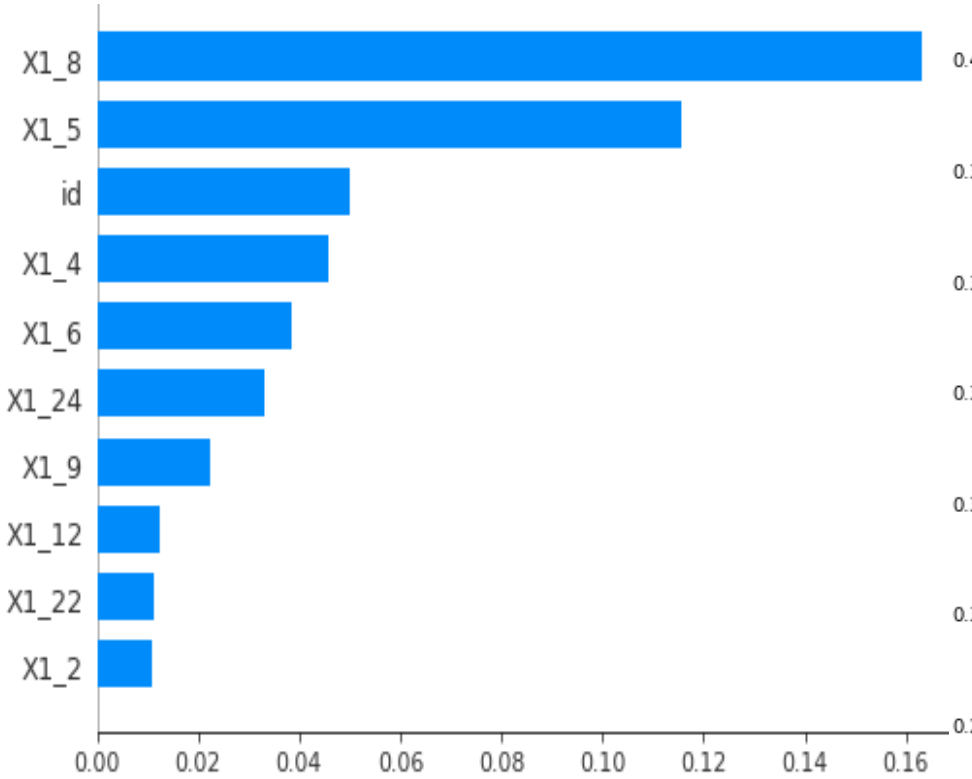
- Хорошо определяется по X1 (фичи 6,7 и id) + X2(tfidf)
- Хорошо предсказывается моделями деревьев

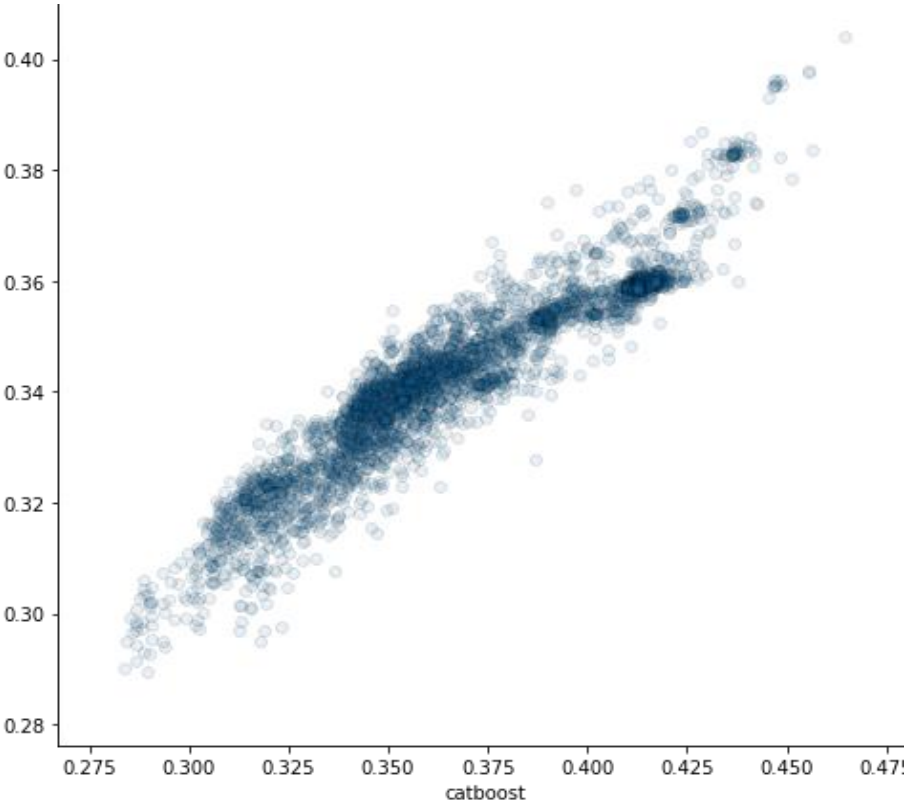




Черта характера у 2

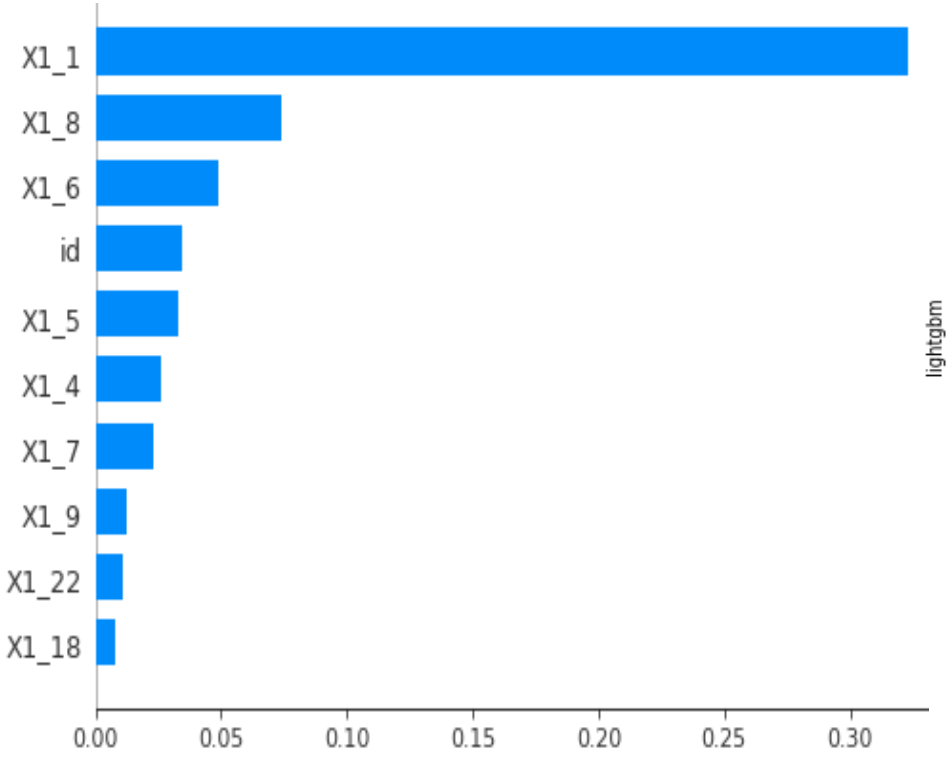
- Хорошо определяется по X1 (фичи 8,5 и id) + X2(tfidf)
- Хорошо предсказывается моделями деревьев

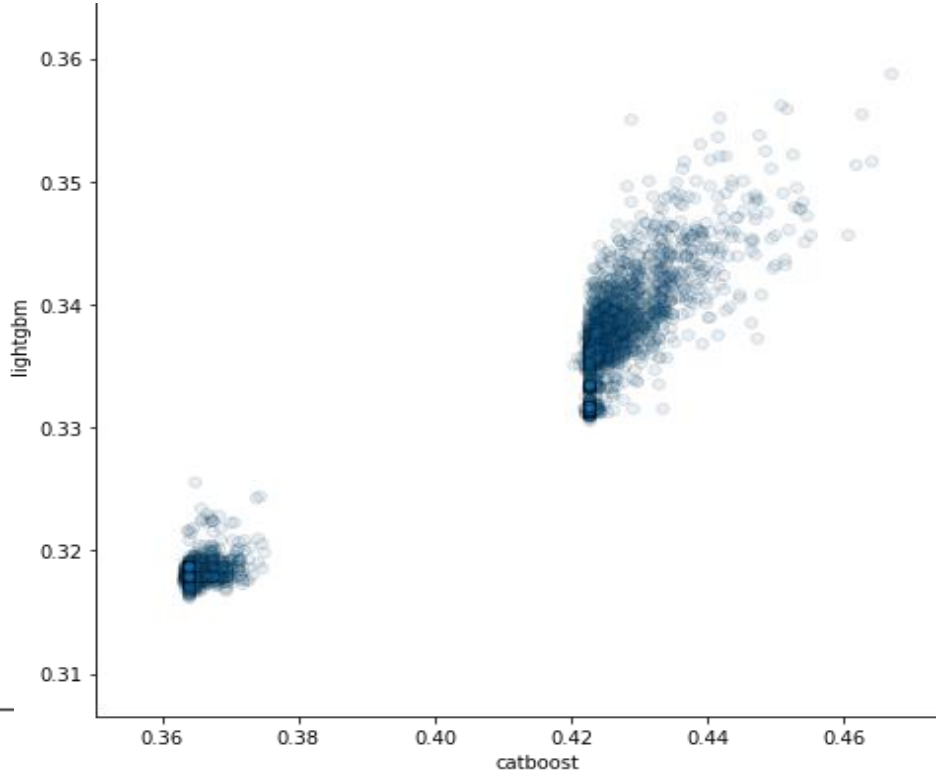




Черта характера у 3

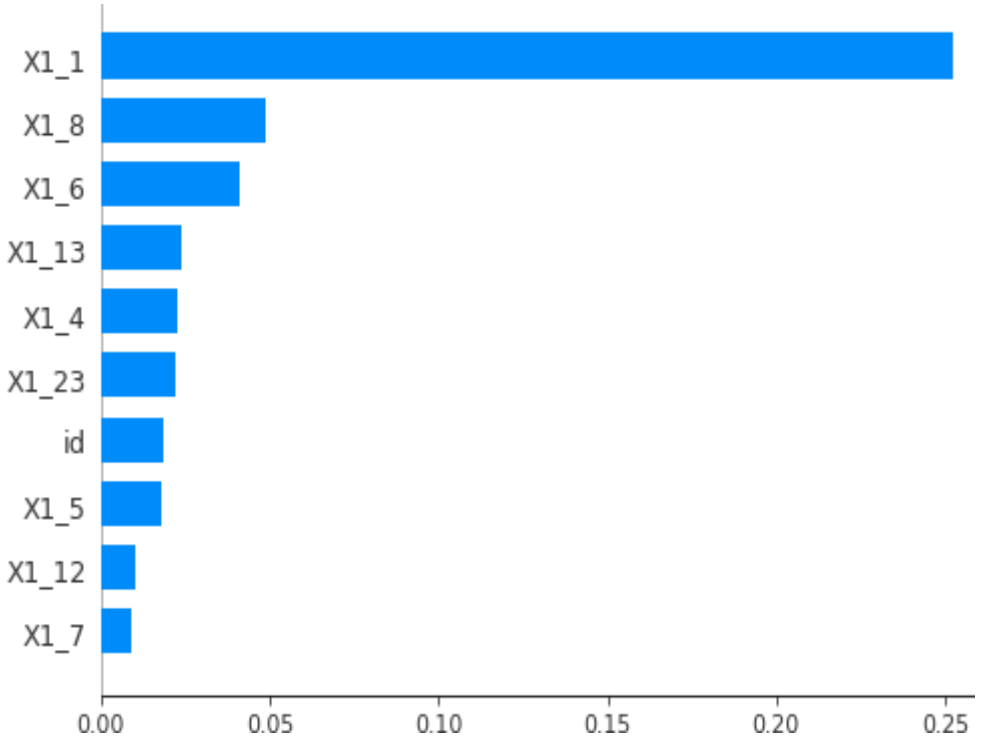
- Хорошо определяется по X1 (фичи 1,8 и 6) + X2(tfidf)
- Хорошо предсказывается линейной моделью
- Модели деревьев разделяют пользователей на 2 класса

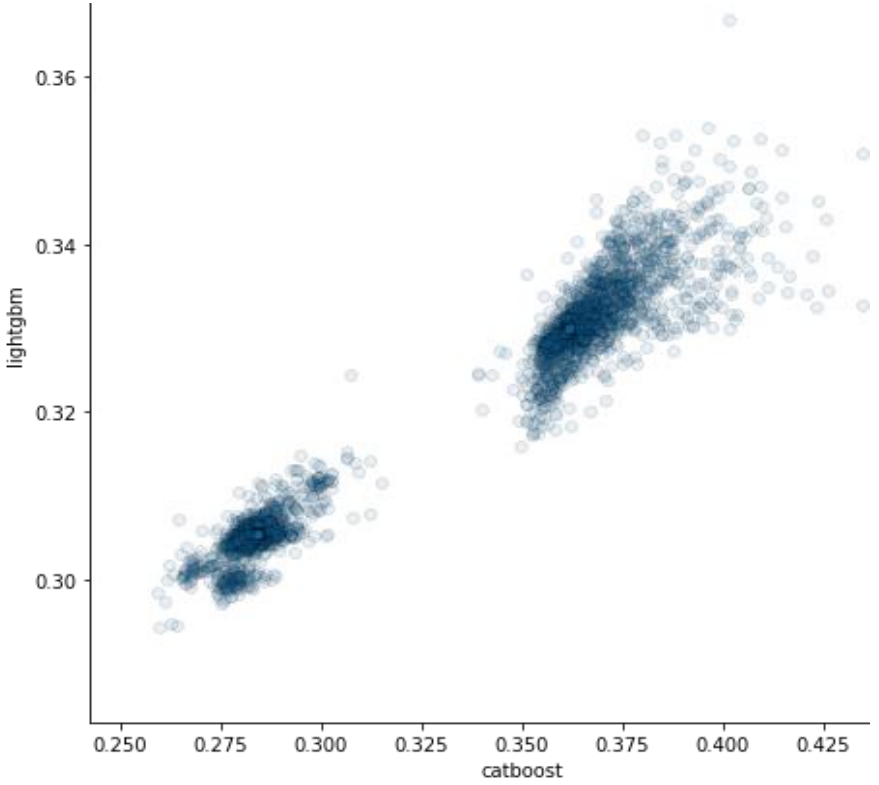




Черта характера у 4

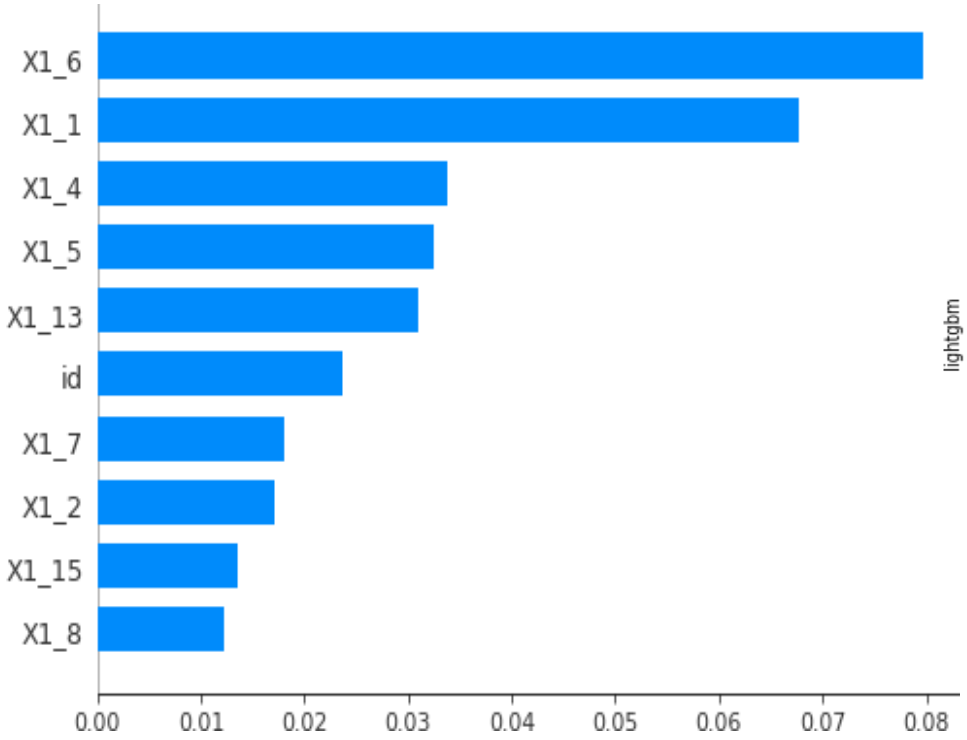
- Хорошо определяется по X1 (фичи 1,8 и 6) + X2(tfidf)
- Хорошо предсказывается линейной моделью
- Модели деревьев разделяют пользователей на 2 класса

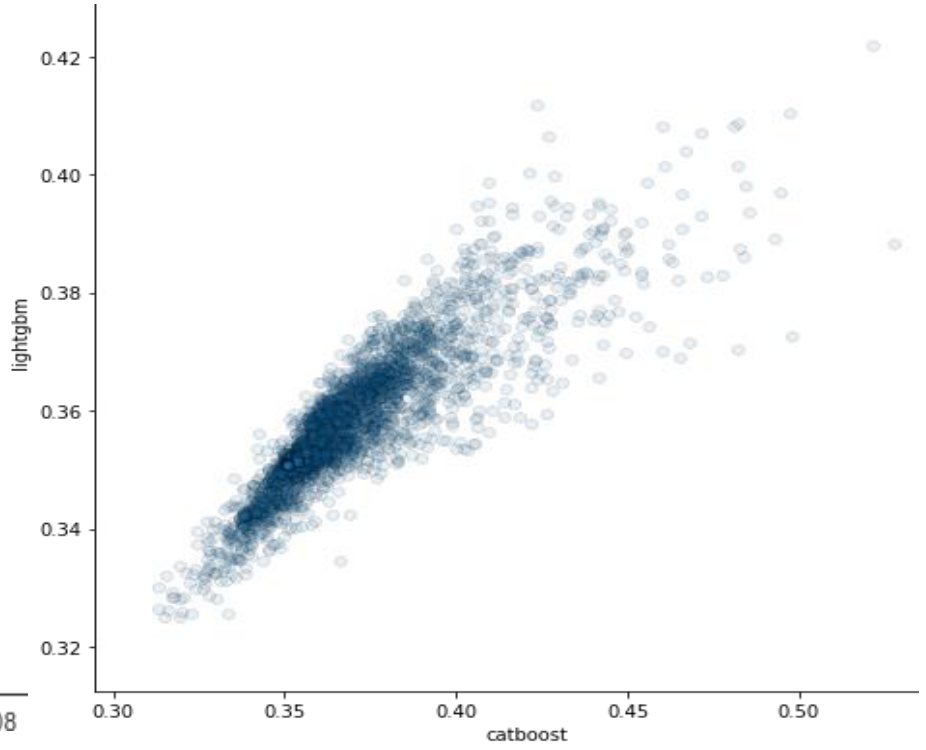




Черта характера у 5

- Хорошо определяется по X2 (tfidf)
- X1 ухудшает прогноз
- Хорошо предсказывается линейной моделью





Сравнение результатов моделей

	cv*	public	private
LogisticRegression	0.616	0.607	0.612
CatBoostClassifier	0.603	-	-
LGBMClassifier	0.610	-	-
Итоговая	0.610	0.617	0.619

*cv только по 20% данных

Кросс-валидация

Для локальной валидации использовался cross_val_score по 5 фолдам. Для выделения фолдов использовался ShuffleSplit (наименьшим std score по фолдам).

Разница между pablic score и privat score ~ 0,0016

Зависимость score от использованных данных:

X1 и X3 — pablic score лучше, чем privat score

X1 и X2 — privat score лучше, чем pablic score