

# Happy Data Year

22 ноября — 10 января

2018

2019



Татьяна Некрасова  
20 место



# Задача:

- Предсказать индекс популярности банкомата (-0,15 до 0,2)
- Данные: ID банка, ID банкомата, широта и долгота
- Трейн: 6261 банкомат
- Тест: 2504 банкомата
- Задача регрессии
- Метрика: RMSE



# Предобработка данных

- Пропущенных 420 адресов
- Неправильно распознанные (болото, реки, лес и т.д. Санкт-Петербург определялся как Москва)
- Ручная проверка адресов через [ahunter.ru](https://ahunter.ru) (проверка адресов по КЛАДР и ФИАС)

# Определение банков:

5478 ВТБ (Уралсиб)

1942 Альфабанк

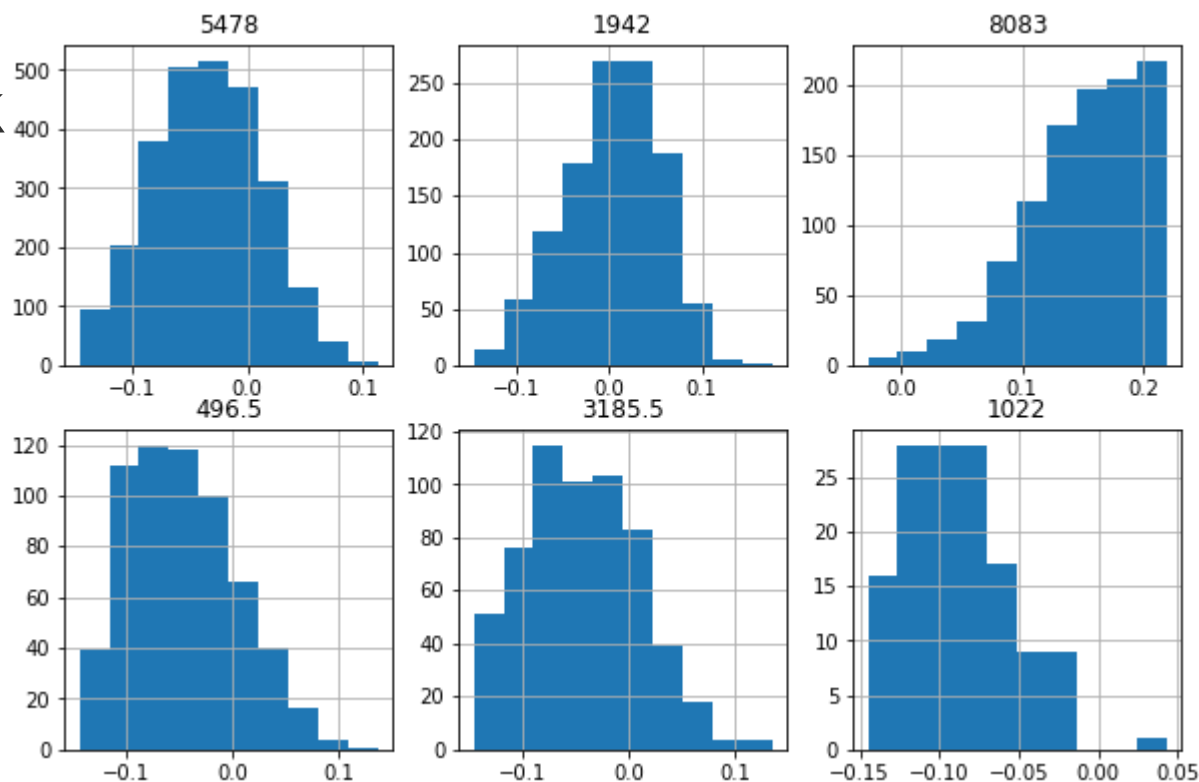
8083 Росбанк


496.5 Розсельхозбанк

3185.5 Газпромбанк

1022 Ак Барс Банк


32 Прочее





# Признаки на основе адресов (город и регион)

- Средняя сумма на текущем счете по региону (Москва и СПб как отдельные регионы)
- Плотность населения в регионе (Москва и СПб как отдельные регионы)

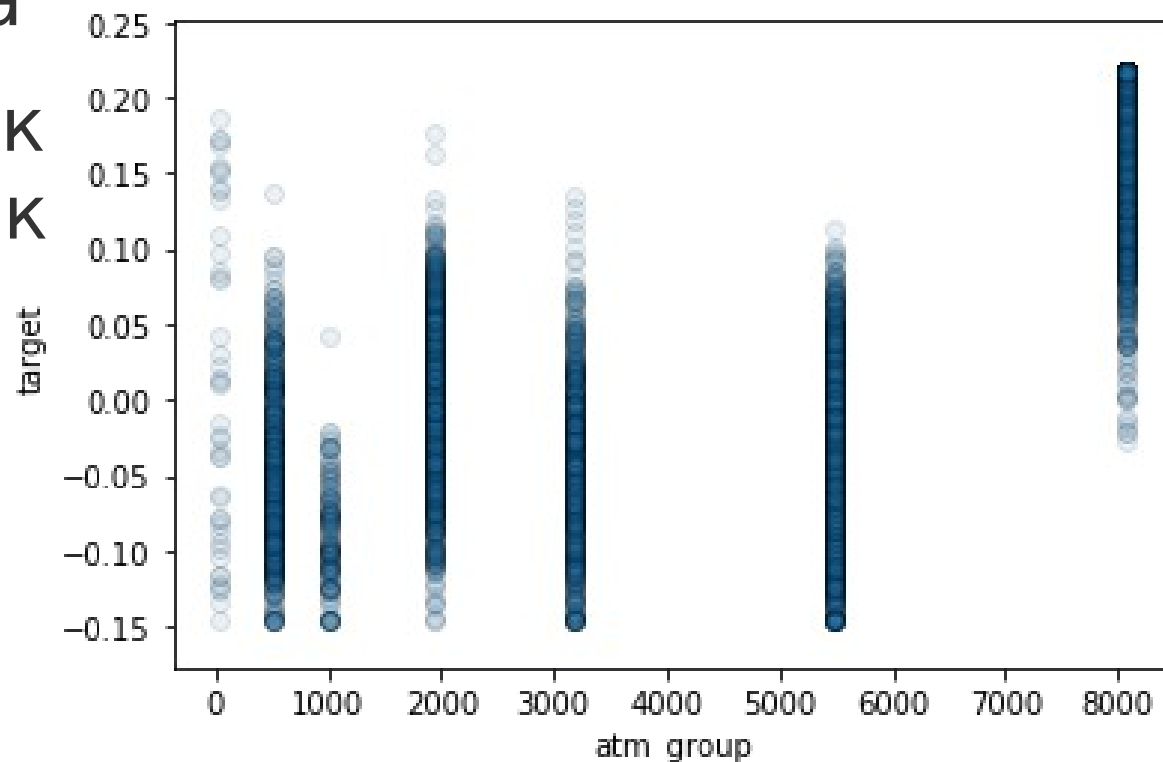


# Признаки на основе координат

- Расстояние до 4,5 банкоматов
- Индекс 5 ближайшего банкомата
- Вариация дистанции до 0-6 ближайших банкоматов
- Количество банкоматов в радиусе 0,1 и 1 градус (примерно 10 и 100 км)
- Расстояние до 4 по дальности отделения Росбанка и его индекс
- Расстояние до центра населенного пункта

# Признаки на основе банка

- Идентификатор банка
- Среднее по группе банкоматов банка
- Бинарный признак принадлежности к банку



# Модели

	cv	public	private
XGBRegressor	0.042771	0.0443625	0.042968
Усреднение (LGBMRegressor, CatBoostRegressor, XGBRegressor)	-	0.0441712	0.0426609
StackingRegressor ((gbm, cat, xgbr), meta_regressor=ridge)	0.042925	0.0436585	0.0425493

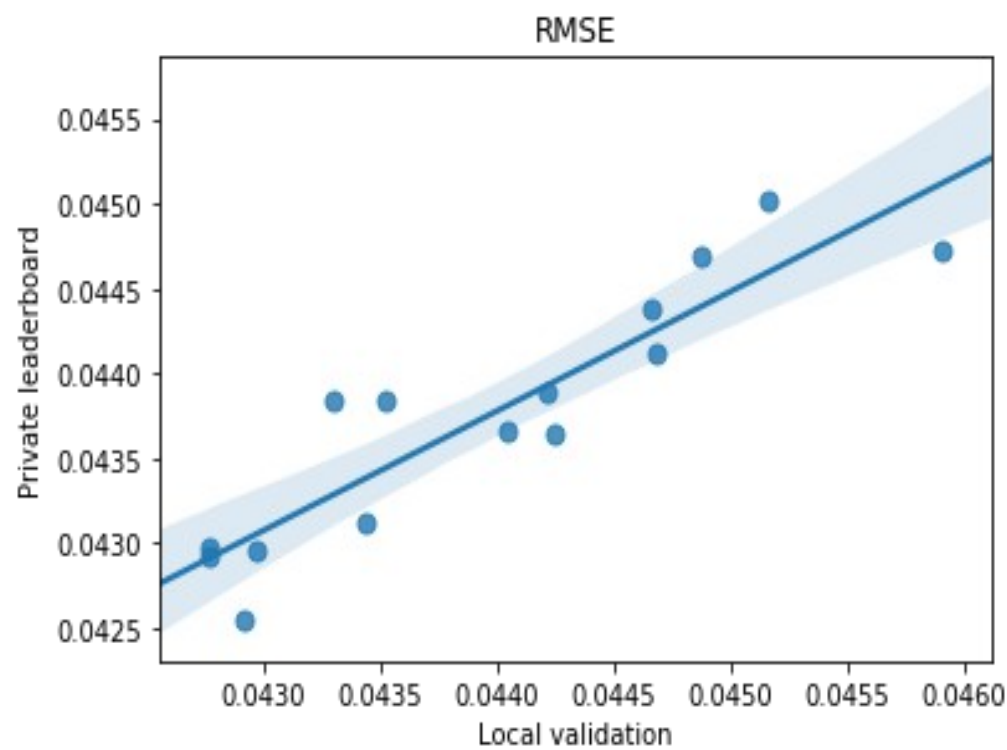
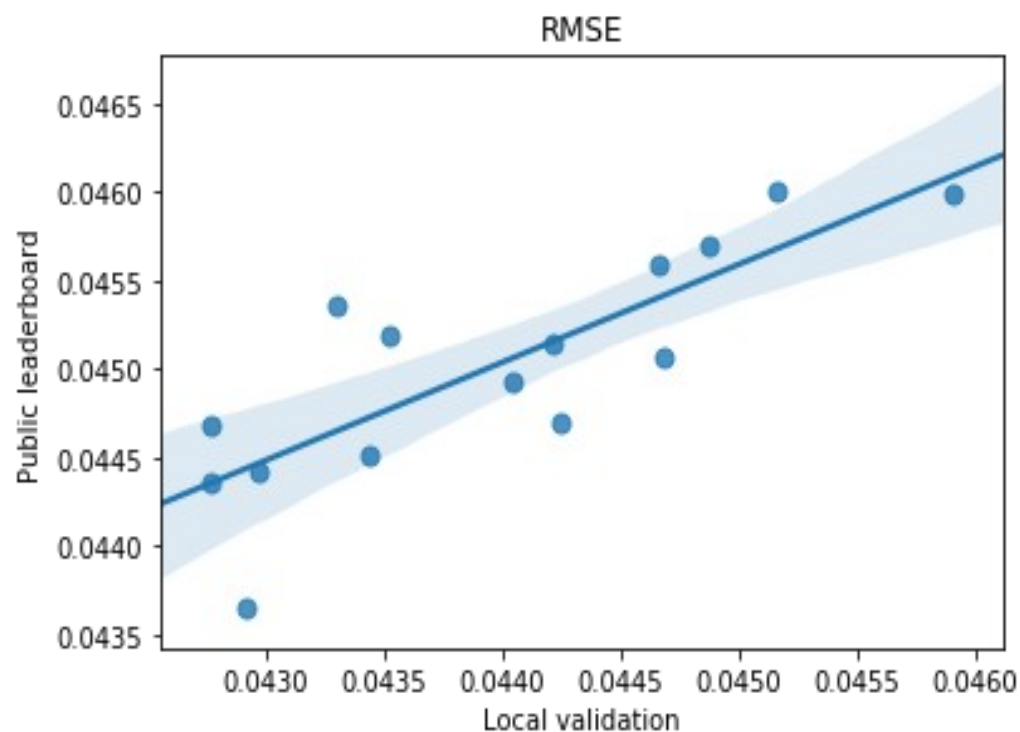


# Кросс-валидация

(cross\_val\_score)

Kfold — 10 фолдов (только catboost)

RepeatedKfold — 50 фолдов



# Важность признаков

LightGBM	eli5
var	atm_group
indexes_5	median
distance_4	var
distance_5	distance_5
distance_o_4	rub_chet
long	indexes_o_4
indexes_o_4	long
rub_chet	distance_4
plot_nas_17	distance_o_4
atm_group	distance_c

Xgboost	eli5
distance_4	atm_group
var	median
distance_c	var
indexes_5	distance_5
lat	long
long	distance_o_4
distance_5	plot_nas_17
distance_o_4	distance_4
rub_chet	lat
atm_group	indexes_o_4

LightGBM	eli5
atm_group	atm_group_3185
atm_group_8083	atm_group_1022
median	indexes_5
long	lat
rub_chet	atm_group_496
var	distance_c
distance_o_4	indexes_r_10
distance_4	indexes_r_1
distance_5	indexes_o_4
lat	distance_4