Kaggle Facebook V: Predicting Check Ins

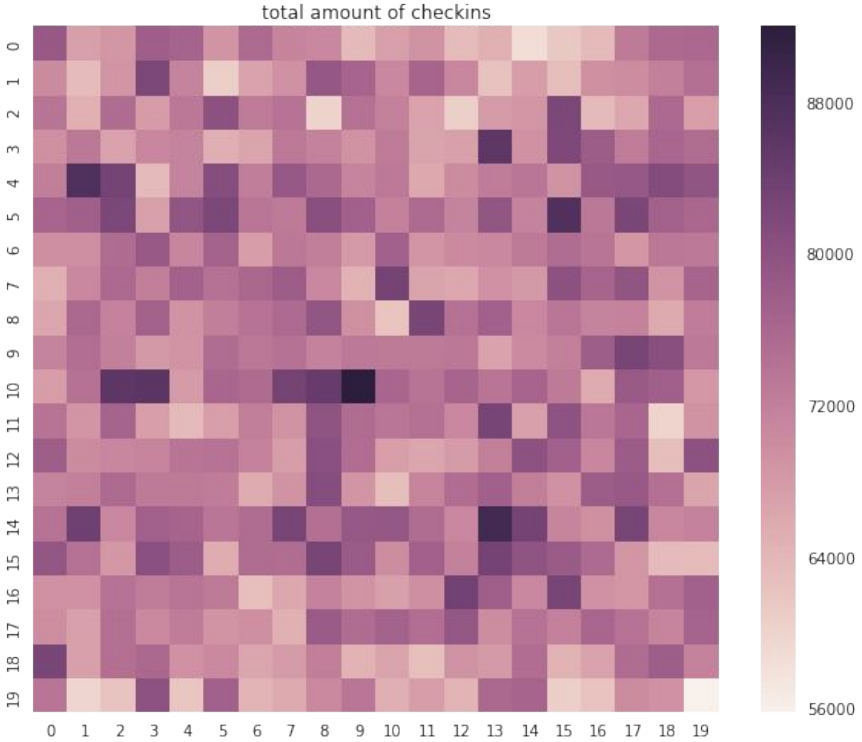
Першин Михаил

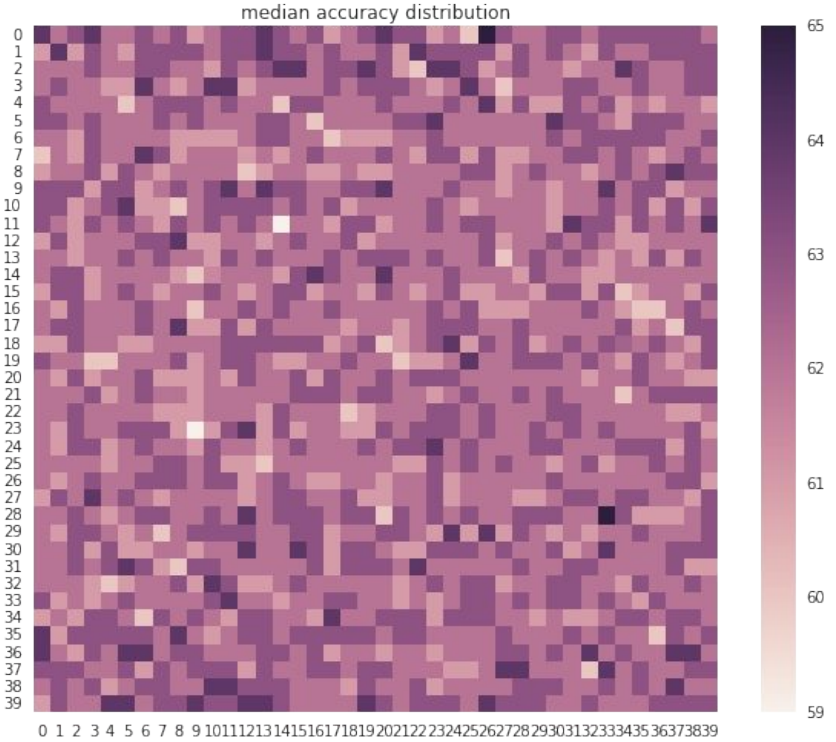
2016

Задача

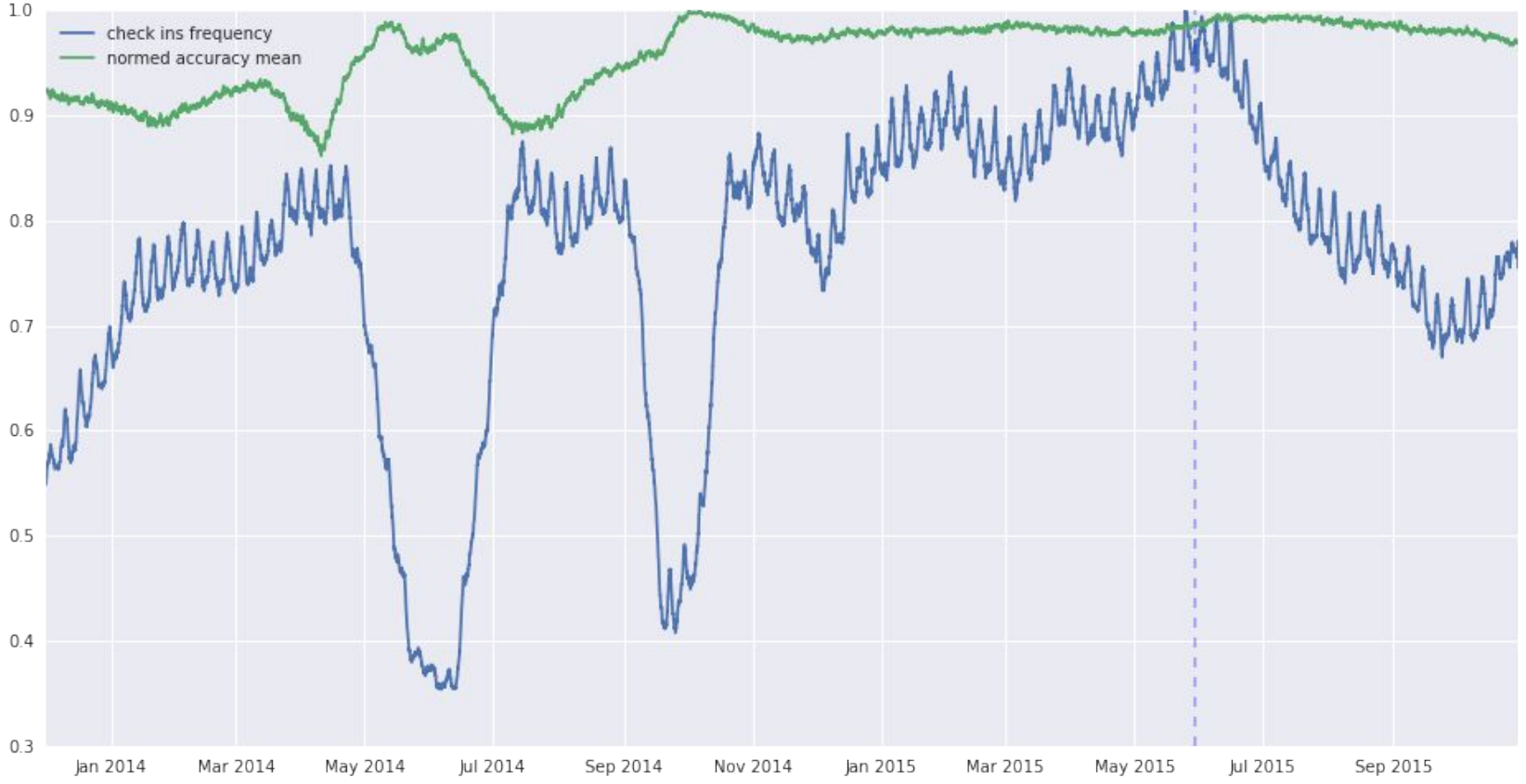
- Предсказать в каком месте пользователь совершил чек ин, основываясь на координатах, времени и точности измерения
- ~ 29 миллионов обучающих примеров, ~8 миллионов тестовых
- ~ 100 тысяч уникальных мест
- Метрика: Mean Average Precision @3 (MAP@3)

Данные


total amount of checkins
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19
88000
80000
72000
64000
56000


median accuracy distribution
65
64
63
62
61
60
59

Данные


check ins frequency
normed accuracy mean
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
Jan 2014
Mar 2014
May 2014
Jul 2014
Sep 2014
Nov 2014
Jan 2015
Mar 2015
May 2015
Jul 2015
Sep 2015

Фичи

- Координаты простые и с ними много не придумаешь
- Точность непонятная
- Время в сыром виде бесполезно - нужно выделять фичи для учёта переодичности событий

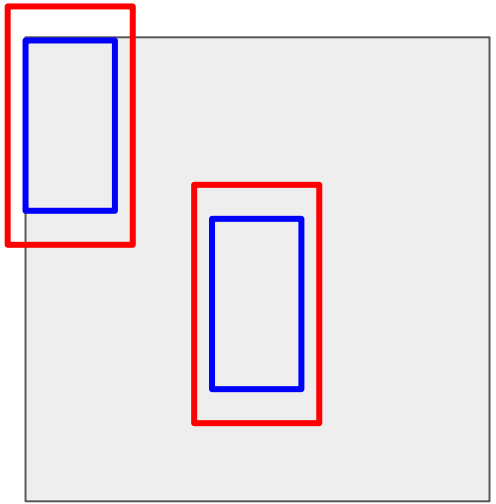
Модель 1

- Что у меня работало плохо: Naive Bayes
- Что работало неплохо: RF, KNN
- Что работало хорошо: xgboost

Модель 1

Если пытаться кормить модель данными “в лоб” - требования к ресурсам становятся абсурдными.

Что делаем: вспоминаем про природу данных и разбиваем их сеткой



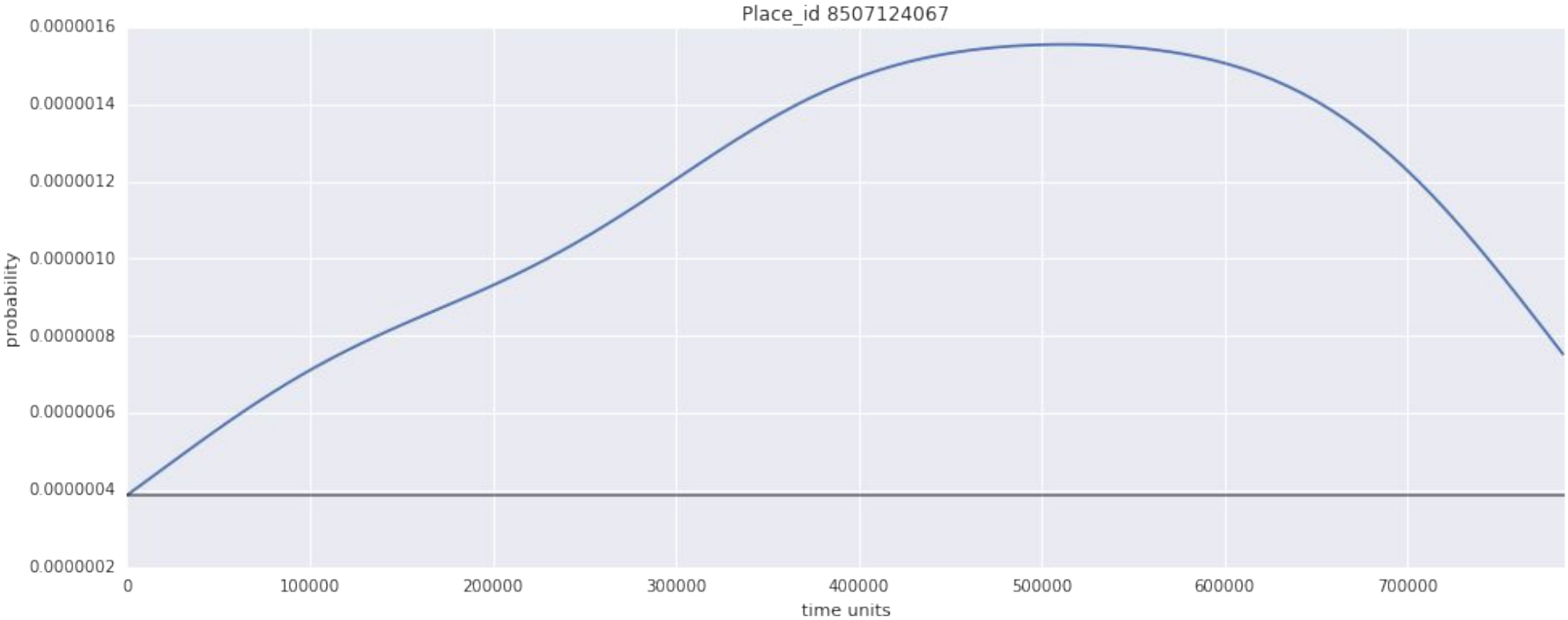
Модель 1

После аккуратного подбора всех возможных гиперпараметров:

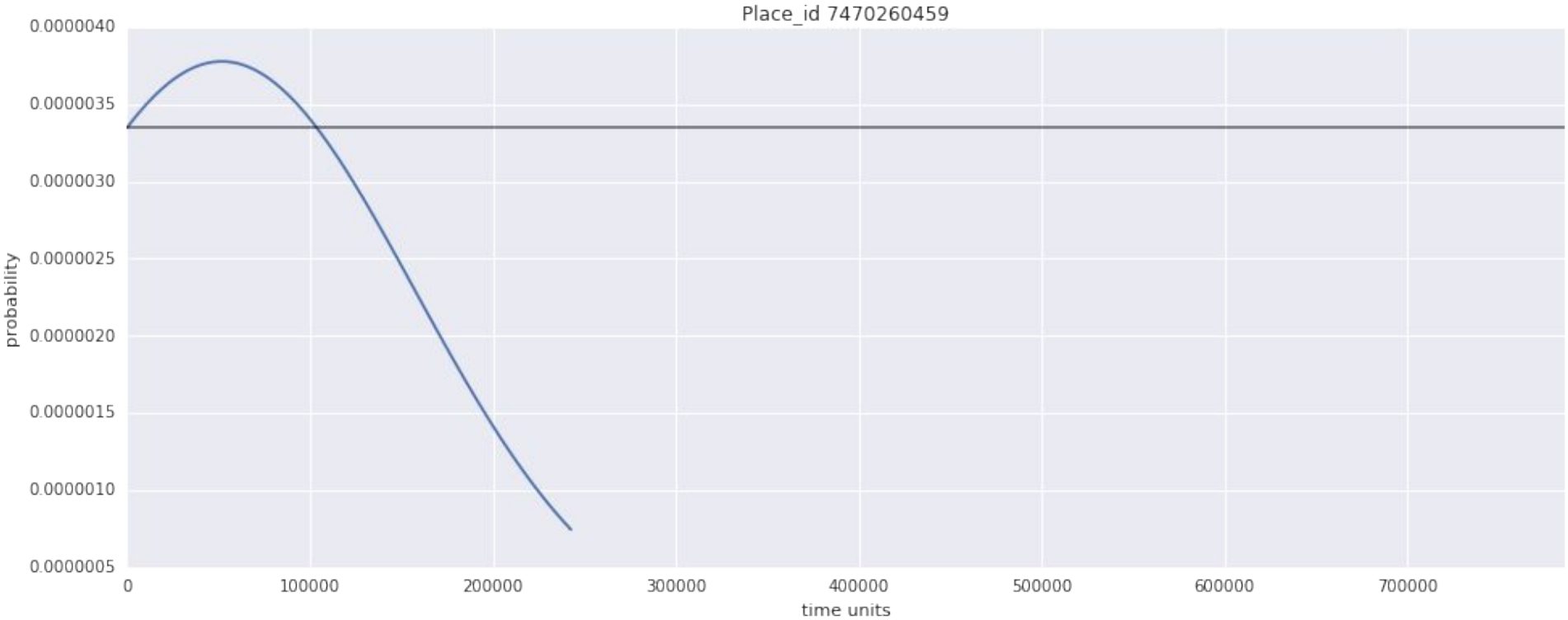
~0.589 MAP@3

~ 86е место на Private Leaderboard-е (топ 7% !)

Киллер-фича: kde



Киллер-фича: kde



Киллер-фича: kde

Зачем останавливаться, если хорошо пошло? Построим распределения для остальных фичей, в том числе и периодичных!

Что получим:

~ 0.606 MAP@3

~ 12е место на Private Leaderboard-е

Как же без ансамблей

Ансамблировать что-то внутри модели слишком дорого, попробуем использовать уже посчитанные ранее файлы (а может быть даже утянем публичные)

Используя несколько файлов решений, переранжируем place_id: в каждом файле будем в каждой строчке назначать каждому предсказанному place_id баллы: 3 за первое место, 2 за второе, 1 за третье. Суммируем по файлам, и делаем коллективное предсказание.

В итоге: +0.0025, 8е место.
(на модели 1 вплоть до +0.0075 для ансабля из 10+ решений)

Спасибо за внимание!

Код появится тут: https://github.com/pershinmr/kaggle_facebook_v