San Francisco Crime

2 jun 2015 - 6 jun 2016.

Косяченко Семен. spiero@yandex.ru

Predict the category of crimes that occurred in the city by the bay

- > 2300 участников
- > 1000 “серьезных решений”
- 1.7 * 10^6 событий в тренировочной выборке
- это не задача прогноза
- (2 недели тренировочная выборка - неделя тестовой)

Raw data

данные

Dates	Category	Descript	DayOfWeek	PdDistrict	Resolution	Address	X	Y
13.05.2015 23:53	WARRANTS	WARRANT ARREST	Wednesday	NORTHERN	ARREST, BOOKED	OAK ST / LAGUNA ST	-122.465	37.774
13.05.2015 23:53	OTHER OFFENSES	TRAFFIC VIOLATION ARREST	Wednesday	NORTHERN	ARREST, BOOKED	OAK ST / LAGUNA ST	-122.426	37.774

submission

задание: всего 39 категорий. предсказание: предсказать вероятность принадлежности события к каждой из категорий.

Метрика - LogLoss.

Фичи

Date

время с даты первого преступления в днях
время дня (сдвинуто на 5.5 часов)
время дня зашумленное (дисперсия = 15 мин)
день года
день недели
будний день / выходной
кратность часу (получасу)
месяц, время года

PdDistrict (10 районов)

бинаризация на 10 фич
прибрежный район / парковый район / спальный район

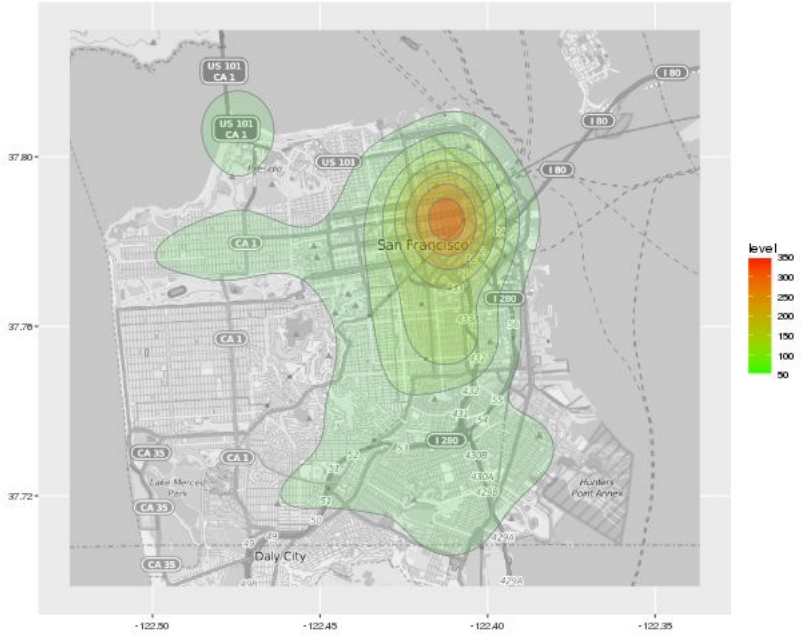
Adress (~1200 разных улиц)

бинаризация 80 фич x2 с разными hash
тип улицы (15 типов)
признак перекрестка

Фичи

Координаты

X, Y



Как использовать координаты на 100%?

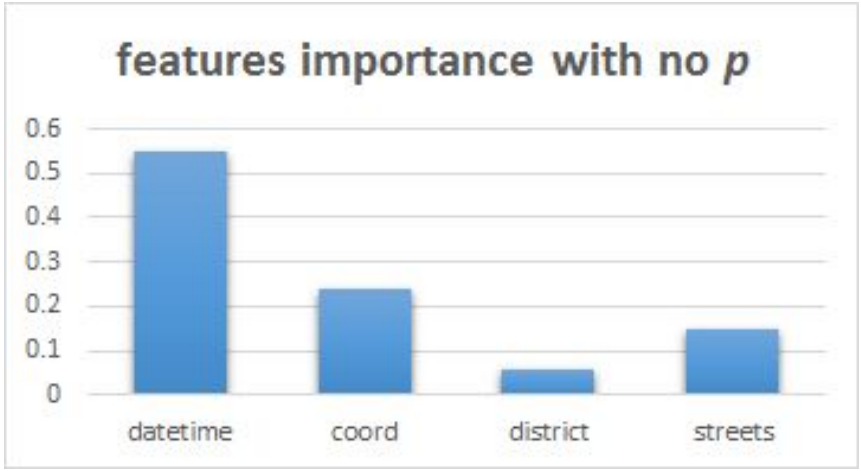
Полярные координаты (выбираем центр - downtown)

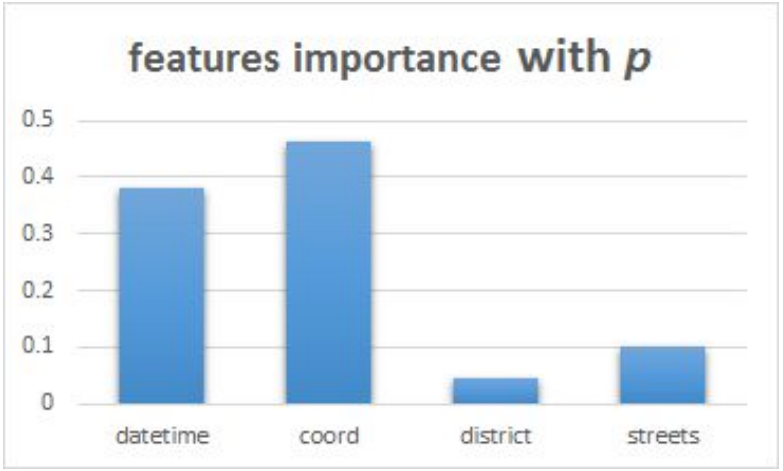
поделить город на небольшие квадраты и проставлять

Использовать X, Y как входные данные для построения функции плотности *p*

значение функции плотности *p*

Фичи



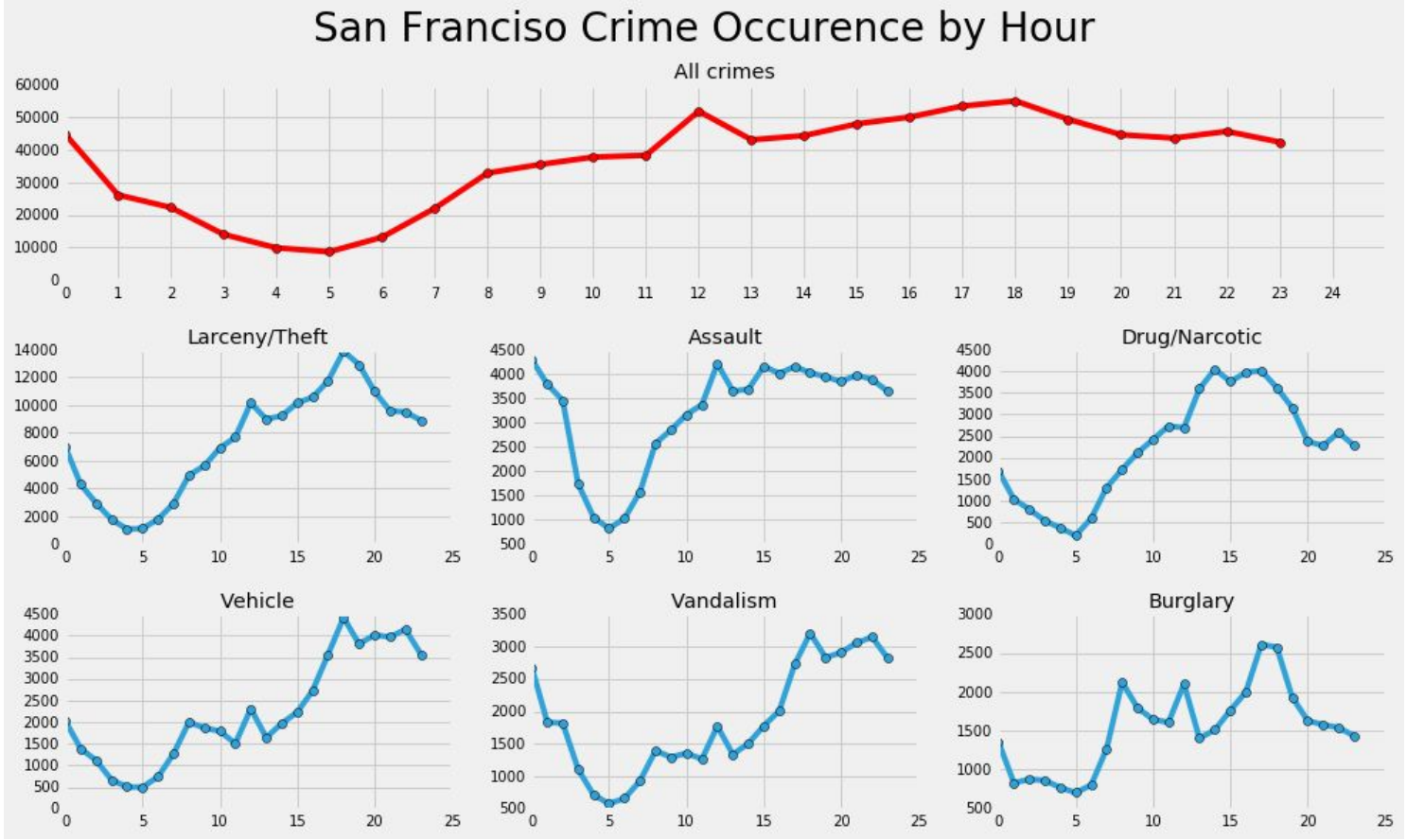


Фильтрация

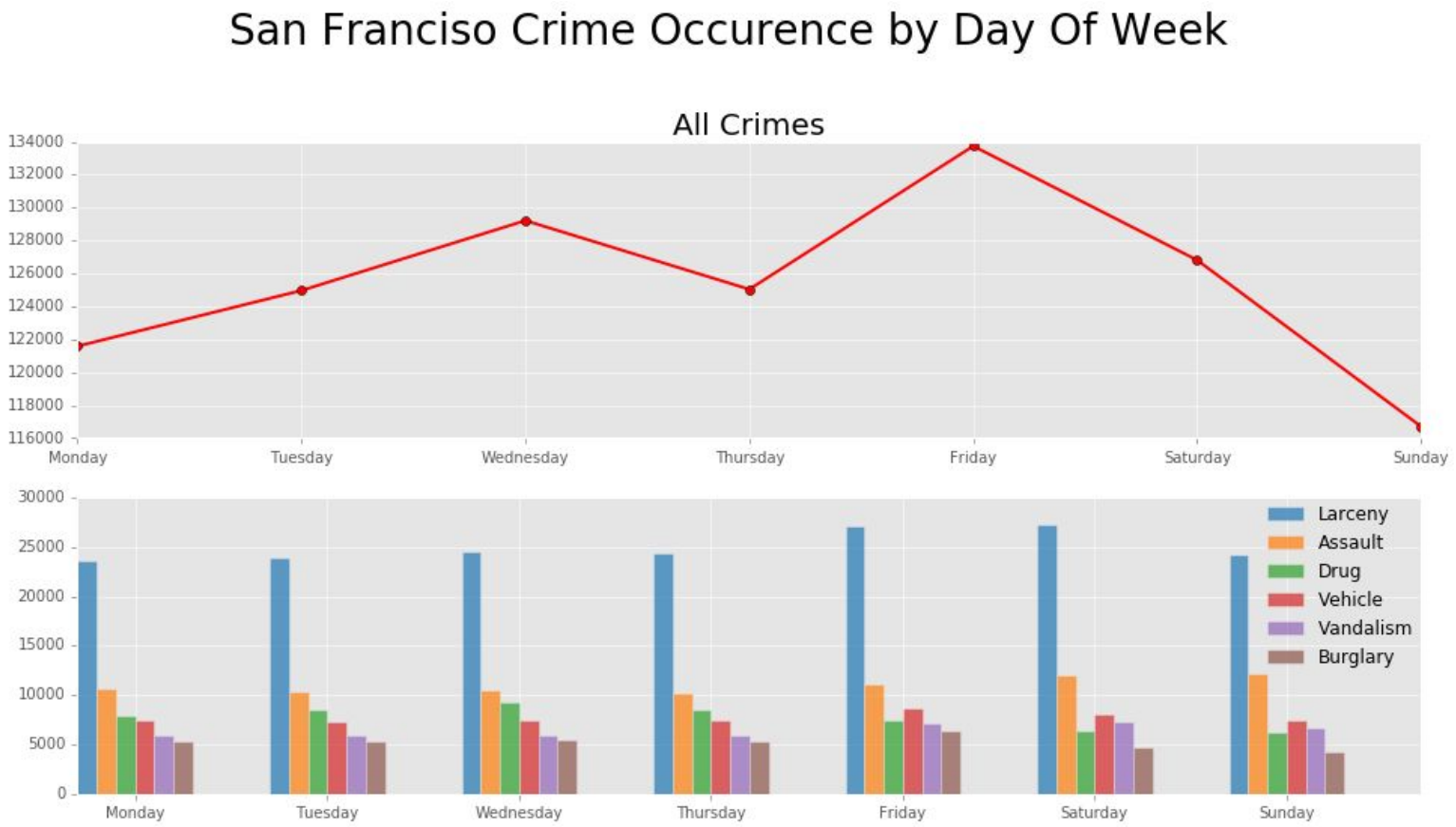
- не для всех событий существуют координаты, но нельзя брать средние!
- выкидывать события, которые слабо коррелируют со своим же предсказанием.

несколько событий в одно время с одинаковыми координатами

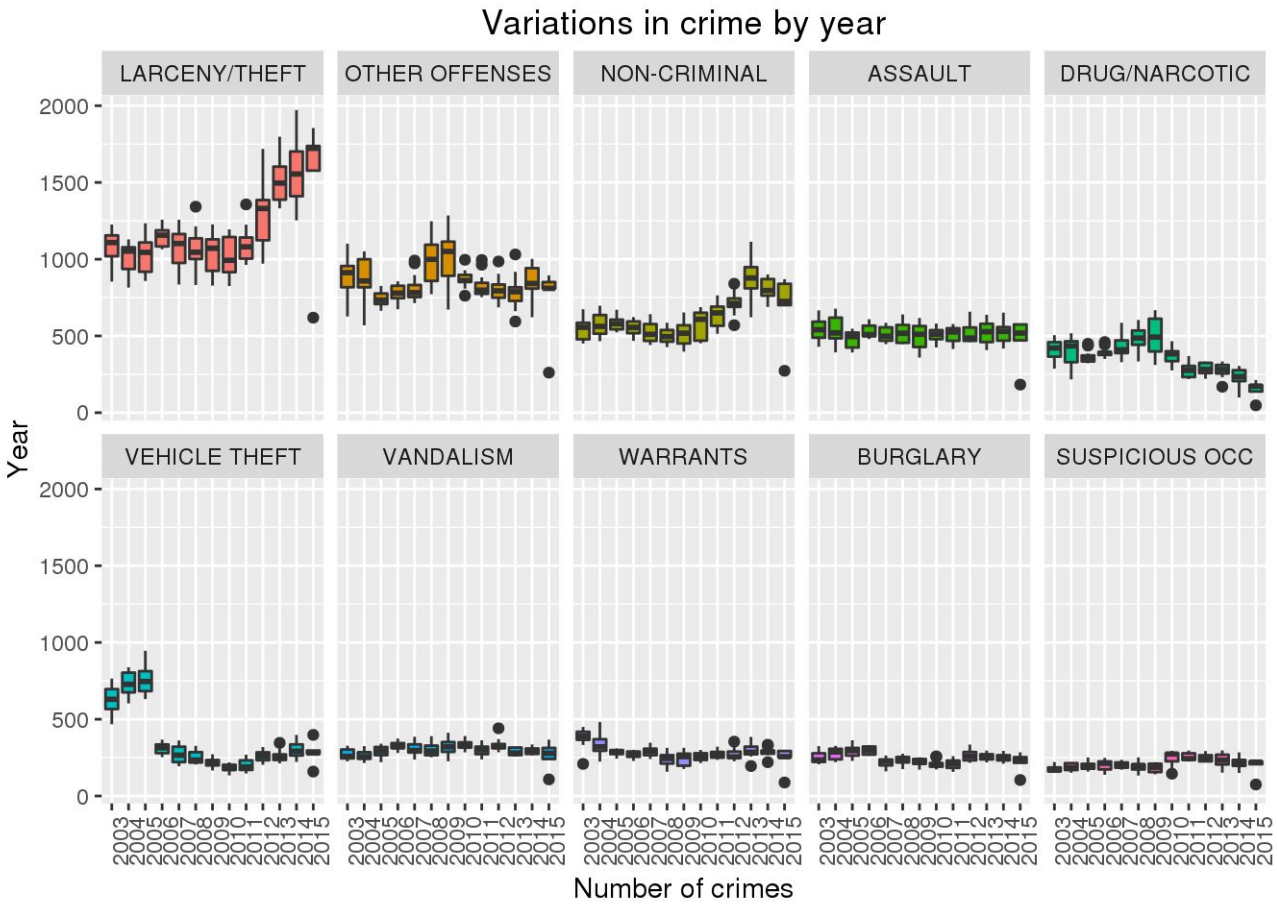
Важность фич



Важность фич



Важность фич



Что пробовал, чем пользовался

Чем пользовался:

- python 2.7
- scikitlearn, hyperopt, seaborn, theano, lasagne

Что пробовал:

- SVC
- BayesClassifier
- GradientBoosting
- RandomForestClassifier
- AdaBoostClassifier
- NeuralNet (Lasaghe)

Лучшее решение одной модели:
SVC `kernel='rbf'` ~ 2.66.

GradientBoosting

GradientBooting - строится отдельно для каждого типа событий отдельно. т. о. строим отдельно 39 gradientBoosting моделей.

+	-
1. Возможность настройки каждой модели отдельно hyperopt (estimators_num, max_depth, min_samples_split).	1. скорость работы.
2. Высокая точность (результат выше на 10-12%).	2. необходимость постобработки.
3. Эффективно делится на потоки.	
4. Возможность обработки части типов событий	

Постобработка

1. Фильтрация а минимальное значение вероятности для каждого события.
2. Склейка результатов всех классификаторов в один массив.
3. Проверка событий по количеству

Кроссвалидация

Кроссвалидация: 2 фолда. Корреляция с лидерборд ~1.

Perfomance

Мои ресурсы:

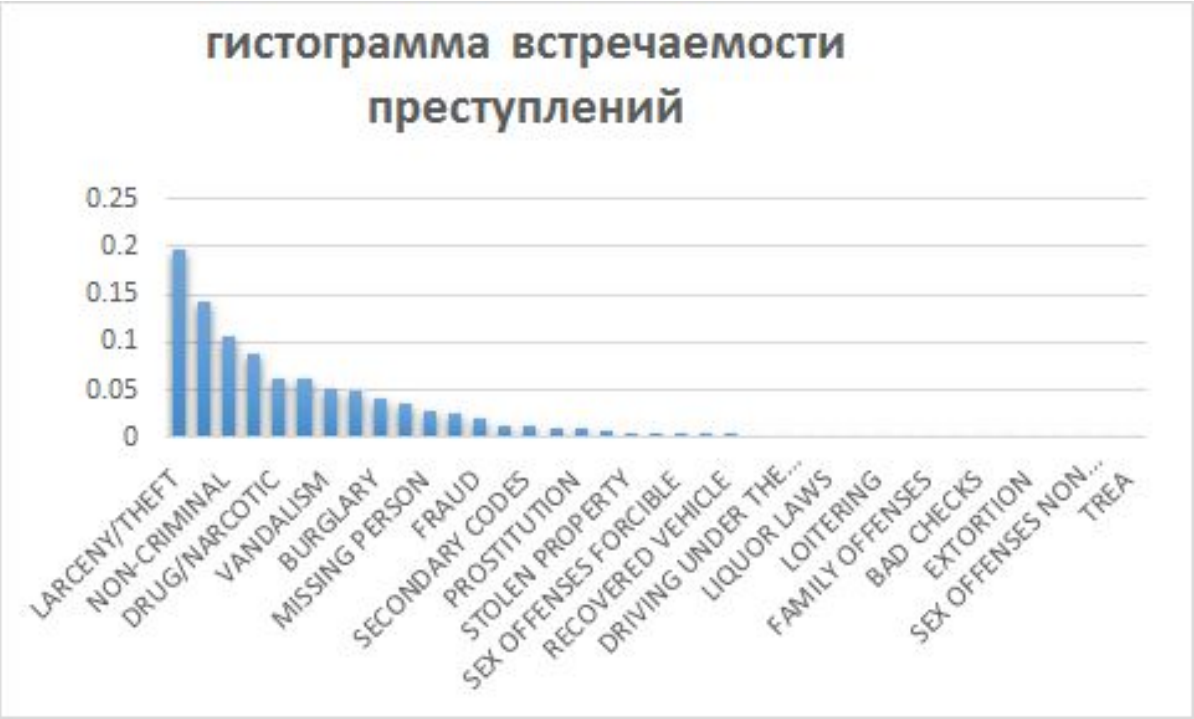
core i3, 4gb ram, интегрированная видеокарта

gb n_estimators ~ 200 занимает 8 часов.

FlyElephant (

- очень быстро есть вычислительные часы
- невозможность дать на вход более одного файла данных.

Perfomance



число событий по 13-ти типам преступлений составляют менее 1% от числа всех преступлений.
число событий по 22-м типам преступлений составляет менее 5%.

Общий план решения

- Чтение входных данных
- Генерация фич
- Канонизация
- Фильтрация
- Обучение моделей
- Кроссвалидация
- Постобработка

code: https://github.com/piero10/KaggleCrime

spiero@yandex.ru

Сюда не ходите

“Какие интересные наблюдения можно было сделать по данным?”