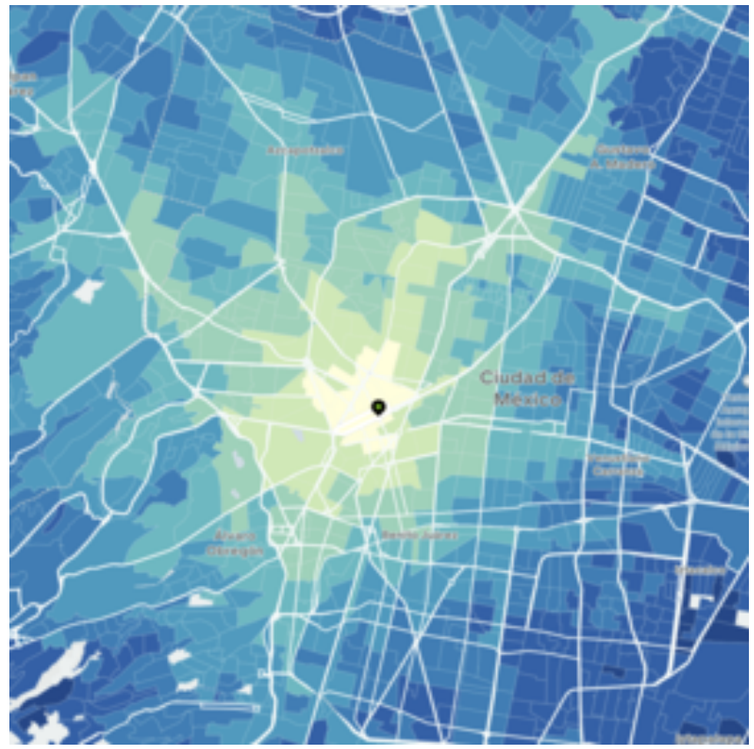
Uber Movement SANRAL Cape Town Challenge

Горячева Тоня
Тюкавин Андрей



Входные данные

1. Train

	EventId	Occurrence Local Date Time	Reporting Agency	Cause	Subcause	Status	longitude	latitude	road_segment_id
0	60558	01/01/16 00:53	Cam	Stationary Vehicle	Vehicle On Shoulder	Closed	18.5408955032	-33.888275	S0B3CGQ
1	60559	01/01/16 00:54	CAMERA	Accident	With A Fixed Object	Closed	18.9307563219	-34.140857	RYJYAPI
2	60560	01/01/16 02:26	Law Enforcement	Accident	Multi Vehicle	Closed	18.5533575029	-33.959154	U3KP57C
3	60561	01/01/16 02:56	CAMERA	Stationary Vehicle	Vehicle On Shoulder	Closed	18.6775561589	-33.895258	RY0TRQ8
4	60562	01/01/16 03:40	CAMERA	Accident	Multi Vehicle	Closed	18.8371319682	-34.087051	8LOVJZ3
5	60564	01/01/16 06:32	NaN	Stationary Vehicle	Vehicle On Shoulder	Closed	18.6384711081	-33.885498	X4UA382
6	60565	01/01/16 07:05	camera	Accident	Single Vehicle	Closed	18.4637854567	-33.943158	0QR8FDW
7	60567	01/01/16 07:39	camera	Police and Military	Road Rage	Closed	18.6359671258	-34.002366	DZABHQW
8	60568	01/01/16 08:00	camera	Stationary Vehicle	Vehicle On Shoulder	Closed	18.6350138684	-34.002237	EKZN1VM
9	60569	01/01/16 08:44	SAPS	Accident	Single Vehicle	Closed	18.4906240725	-33.949284	H9XYX9Q

2. Test

	datetime x segment_id	prediction
0	2019-01-01 01:00:00 x S0B3CGQ	NaN
1	2019-01-01 01:00:00 x RYJYAPI	NaN
2	2019-01-01 01:00:00 x U3KP57C	NaN
3	2019-01-01 01:00:00 x RY0TRQ8	NaN
4	2019-01-01 01:00:00 x 8LOVJZ3	NaN

Входные данные

3. Weather Data - https://rp5.ru/

	local_time	temperature	p0	p	humidity	mean_wind_direction	mean_wind_speed	special_weather_phenomena	clouds	horizontal_visibility	Td
0	31.03.2019 23:00	16.0	761.2	765.0	83	Wind blowing from the east-southeast	2	NaN	Few clouds (10-30%) 720 m; broken clouds (60-9...	10.0 and more	13.0
1	31.03.2019 22:00	16.0	760.5	764.3	83	variable wind direction	2	NaN	Scattered clouds (40-50%) 690 m; broken clouds...	10.0 and more	13.0
2	31.03.2019 21:03	15.0	760.5	764.3	88	variable wind direction	1	NaN	Few clouds (10-30%) 300 m; broken clouds (60-9...	10.0 and more	13.0

4. Road Segments Info

	ROADNO	CLASS	WIDTH	LANES	SURFTYPE	PAVETYPE	CONDITION	length_1	segment_id
0	R300	Primary	20.2	2	Paved	FLEX	Good	471.207	D1U6OOF
1	R300	Primary	20.2	2	Paved	FLEX	Good	471.207	NG4X2MD
2	R300	Primary	20.2	2	Paved	FLEX	Good	471.207	792705Z
3	R300	Primary	20.2	2	Paved	FLEX	Good	471.207	IK67XHB
4	R300	Primary	20.2	2	Paved	FLEX	Good	471.207	OWCF2MH

5. Uber Movement Data

Когда-нибудь мы до них доберемся :)

Формирование обучающей выборки

Sample submission:

	datetime x segment_id	prediction
0	2019-01-01 01:00:00 x S0B3CGQ	NaN
1	2019-01-01 01:00:00 x RYJYAPI	NaN
2	2019-01-01 01:00:00 x U3KP57C	NaN
3	2019-01-01 01:00:00 x RY0TRQ8	NaN
4	2019-01-01 01:00:00 x 8LOVJZ3	NaN

- Всего 549 road segments
- Данные по авариям за 01.01.2016 - 31.12.2018 с точностью до минут

=> сгенерить выборку за 3 года с шагом в **1 час X segment_id**

New Features

- **Weather**
  - mean_wind_direction - почистили, оставили как есть (всего 18 категорий)
  - Clouds - 4 признака, regexp
  - Special_weather_phenomena - 3 признака, regex
  - Время суток - 3 признака (утро/день/вечер)

```
1 weather['clouds']
```

```
0                 No Significant Clouds
1            Few clouds (10-30%) 600 m
2            Few clouds (10-30%) 450 m
3            Few clouds (10-30%) 450 m
4          Broken clouds (60-90%) 450 m
                      ...
28367             No Significant Clouds
28368             No Significant Clouds
28369             No Significant Clouds
28370             No Significant Clouds
28371             No Significant Clouds
Name: clouds, Length: 28372, dtype: object
```

- **Road Segments INFO**

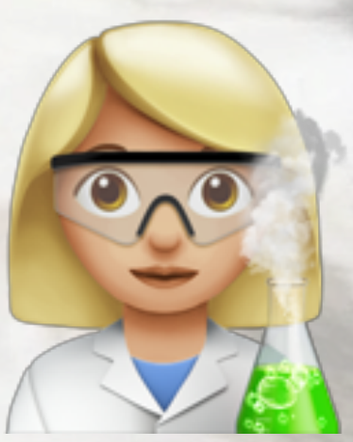
	ROADNO	CLASS	WIDTH	LANES	SURFTYPE	PAVETYPE	CONDITION	length_1	segment_id
0	R300	Primary	20.2	2	Paved	FLEX	Good	471.207	D1U6OOF
1	R300	Primary	20.2	2	Paved	FLEX	Good	471.207	NG4X2MD
2	R300	Primary	20.2	2	Paved	FLEX	Good	471.207	792705Z
3	R300	Primary	20.2	2	Paved	FLEX	Good	471.207	IK67XHB
4	R300	Primary	20.2	2	Paved	FLEX	Good	471.207	OWCF2MH

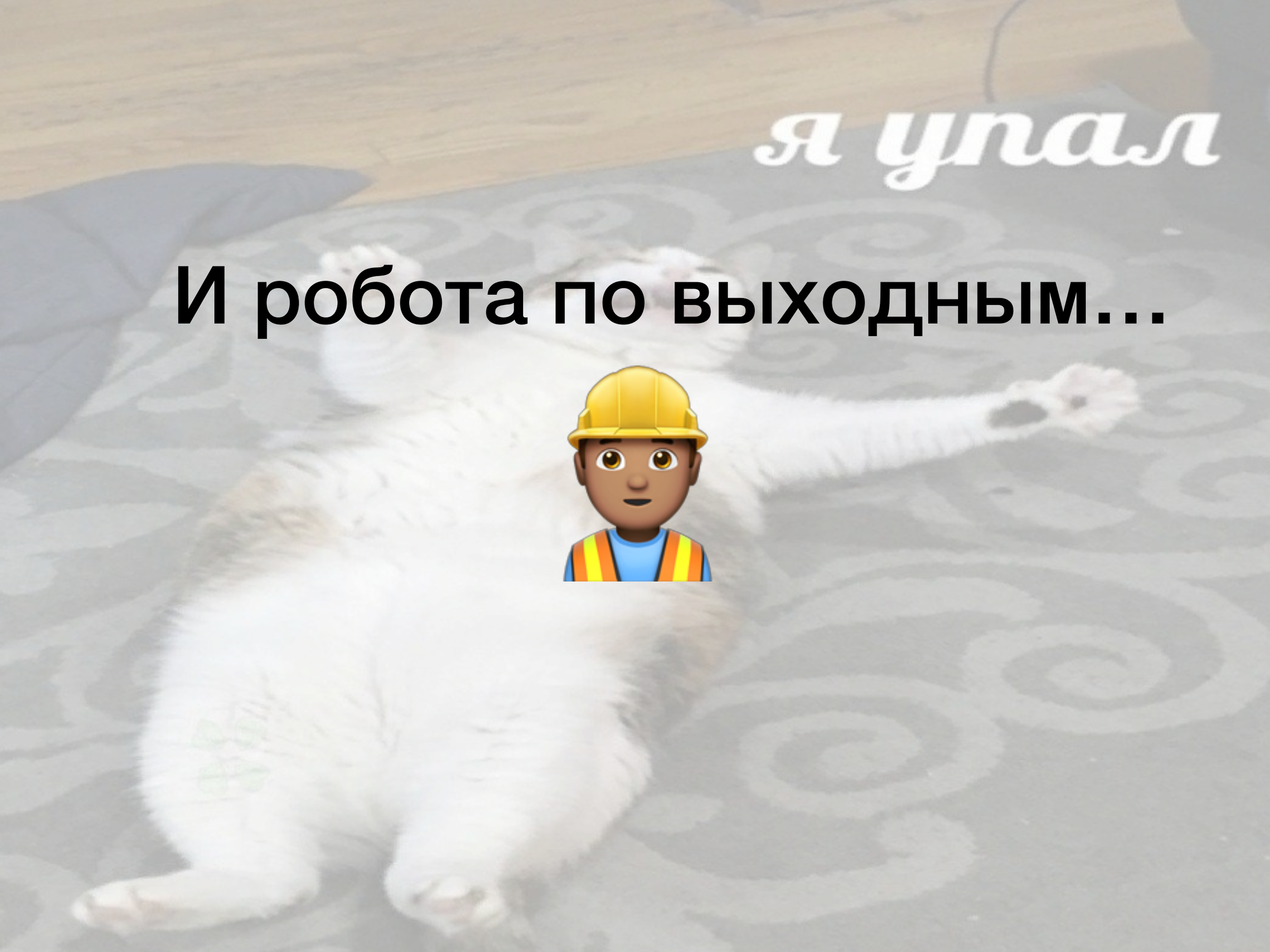
- **Target encoding**
  - **accident_count** - Число аварий на каждом сегменте дороги за все время, JOIN: segment_id
  - **accident_count_by_year** - Число аварий на сегменте дороги за год, JOIN: segment_id + year
  - **accident_count_by_hour** - Число на сегменте дороги за час, JOIN: segment_id+hour

Этапы

№	Features	Model	Test	Доля таргета	F1 score
1	+ Базовые фичи	RandomForest	10.2018 - 12.2018	0.3%	0.039
2	+ Базовые фичи + Погода + Фичи по road_segments	RandomForest	10.2018 - 12.2018	0.3%	0.045
3	+ Базовые фичи + Погода + Фичи по road_segments + Target encoding	CatBoost + CV	10.2018 - 12.2018	50% (все “1” и sample “0”)	0.082
4	+ Базовые фичи + Погода + Фичи по road_segments + Target encoding	CatBoost + CV	01.2018 - 03.2018	50% (все “1” и sample “0”)	0.018

А потом случилась сессия...





И робота по выходным...