



Методика корректировки расчетного времени поездки на основе анализа  
исторических и географических данных

# Zvezdochka

26 апреля 2020

# Команда



Олег Бартов,  
Data Scientist:  
анализ данных, кластеризация,  
исследование аномалий



Олег Черемисин,  
Data Scientist:  
анализ данных,  
пайплайн обучения

# Инструментарий

- Sklearn KMeans clustering (координаты выезда и прибытия)
- Lightgbm с DART (Dropouts meet Multiple Additive Regression Trees)
- Суши и бургеры по промокоду

Предложения по выводу в продуктив:

- Python 3.7
- Сохраненные предобученные модели KMeans и lightgbm
- Docker
- Веб-сервис на основе Tornado

# Фичи

- **кластеризация координат начала и окончания поездки**
- день недели поездки
- час начала поездки













Фичи, которыми можно ограничиться практически без потерь в score:

**ETA      EDA      dayofweek   hour      clust      del\_clust      new\_clust**

## Фичи, которые не зашли :(

- OpenStreetMap (расстояния до объектов, количество объектов определенного типа в заданном радиусе)
- Количество точек и агрегация углов в route
- Выходные/праздничные/рабочие дни
- Расстояние до центра города

# Метрика MAPE (Kaggle Public LB)

#	Team Name	Notebook	Team Members	Score ?	Entries	Last
1	Magic Clty			13.90496	17	3h
2	Zvezdochka			14.15008	4	12m
Your Best Entry 						
Your submission scored 14.15008, which is an improvement of your previous score of 14.16952. Great job!				 Tweet this!		
3	команда			14.22083	25	7m
4	На результат			14.27910	18	7m
5	Натуральный логарифм			14.30537	23	1m
6	84			14.39697	23	19m
7	Wuld Duck			14.56061	18	~10s
8	Ольга Цветкова			14.92965	5	4m
9	diht_hackers_mipt			15.09877	3	14m
10	DEVILS			15.21001	2	1h

# Дальнейшие улучшения

- Использование алгоритмов других видов кластеризации:
  - непараметрической
  - иерархической
  - спектральной
- Кластеризация маршрута отдельно от точек посадки и высадки
- Классификация временных признаков по соответствию расчетного и фактического времени
- Добавление внешней информации о погодных условиях
- Построение интерпретируемых моделей для управленческих рекомендаций

# Состав команды

## Бартов Олег Борисович

- автоматизация производственных предприятий, преподавание, научная деятельность
- руководитель проектов АО “Группа “СВЭЛ”
- приглашенный преподаватель НИУ ВШЭ

## Черемисин Олег Александрович

- маркетинговая аналитика, разработка ПО для подбора и оптимизации работы нефтедобывающего оборудования
- ведущий программист АО “Новомет-Пермь”
- студент Академии больших данных Mail.ru Group (MADE)