Онлайн-хакатон
Pandemic Data Hack <data> <code>

# Решение команды НеИИ

Онлайн-хакатон
Pandemic Data Hack <data> <code>

# Workflow

Онлайн-хакатон
Pandemic Data Hack <data> <code>

# Генерация признаков

1) Агрегации по данным об опыте и образовании:
   - Общий стаж
   - Количество мест работы
   - Количество высших, средних, аспирантур

2) Признаки о пандемии:
   - Через сколько дней после локдауна создано резюме
   - Через сколько дней после локдауна изменено резюме
   - Через сколько дней после локдауна размещено резюме

3) Признаки по текстовым данным:
   - Процент грамматических ошибок в тексте резюме
   - Средняя длина  текста в поле “Обязанности”
   - Среднее кол-во слов с заглавной буквы в поле “Обязанности”
   - Среднее кол-во знаков препинания в поле “Обязанности”

# Генерация признаков из текста

## Кластеризация резюме на основе TF-IDF векторов по тексту “Обязанности”

```
train_salary_cluster[train_salary_cluster.cluster==33]
```

| | id | salary | position | employer | achievements | responsibilities | start_date | finish_date | text | cluster |
|---|---|---|---|---|---|---|---|---|---|---|
| **10812** | 7166 | 27749 | Ассистент | Костромской государственный университет | NaN | <p>Организация образовательного процесса (преп... | 2017-09-01 | NaN | организац образовательн процесс преподаван дис... | 33 |
| **103704** | 71853 | 19905 | преподаватель | Вельский экономический техникум | NaN | <p>Преподавание дисциплины "Информационные тех... | 1993-10-01 | 2008-10-01 | преподаван дисциплин информацион технолог студ... | 33 |
| **111475** | 77235 | 30000 | Преподаватель информатики | ГПОУ "Кемеровский техникум индустрии питания и... | NaN | <p>обучение студентов по дисциплине Информатик... | 2020-04-01 | 2020-07-01 | обучен студент дисциплин информатик информацио... | 33 |

# Генерация признаков из текста

## Агрегация средней зарплаты по кластеру -- доп.фича

```
df[df.cluster==0].tail()
```

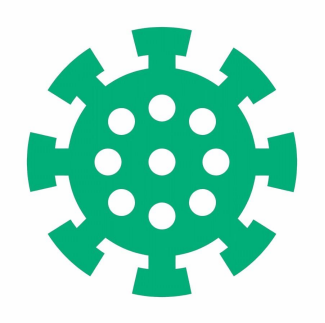
| | id | position | cluster | salary_desired | salary_mean_by_cluster | target |
|---|---|---|---|---|---|---|
| **304933** | 435599 | бухгалтер | 0.0 | 25000 | 44833.370096 | 40000 |
| **305062** | 435770 | бухгалтер | 0.0 | 20760 | 44833.370096 | 22920 |
| **305120** | 435843 | бухгалтер | 0.0 | 30000 | 44833.370096 | 31209 |
| **305231** | 436022 | бухгалтер | 0.0 | 134940 | 44833.370096 | 72882 |
| **305252** | 436046 | бухгалтер | 0.0 | 20760 | 44833.370096 | 22836 |
| **305268** | 436068 | менеджер | 0.0 | 30000 | 44833.370096 | 40000 |
| **305514** | 436429 | бухгалтер | 0.0 | 50000 | 44833.370096 | 60000 |
| **305532** | 436459 | бухгалтер | 0.0 | 25000 | 44833.370096 | 20000 |

# Кодирование категориальных признаков

| Поле | Значение | Тип кодирования |
|---|---|---|
| schedule | график работы | OneHotEncoding |
| drive_licences | категория прав водителей | |
| employement_type | тип занятости | |
| position_cat | набор профессий, выделенный из position (должность) | |
| | | Средняя желаемая ЗП |
| education_type | тип образования | OrderEncoding (осмысленно) |
| citizenship | гражданство | OrderEncoding (по ВВП) |
| region | регион | Средняя ЗП по региону |
| institution | университет | Рейтинг вузов РФ |
| | | Баллы ЕГЭ абитуриентов |

Онлайн-хакатон
Pandemic Data Hack <data> <code>

# Сбор данных из открытых источников



*Скрепинг*

1. Рейтинг университетов
2. Средний балл ЕГЭ университетов
3. Данные о куммулятивной сумме заболевших, умерших, выздоровевших по датам по России
4. Средняя зарплата по позиции
5. Средняя зарплата по городу
6. Средняя зарплата по регионам за 2019 и 2020 год

*Росстат*

Онлайн-хакатон
Pandemic Data Hack <data> <code>

# Заполнение пропусков

Модой - для исходных признаков и категориальных сгенерированных признаков

Чаще люди не готовы переезжать, мода = 0

Чаще люди  готовы обучаться, мода = 1

Специфической константой - для признаков, у которых нет возможности выделить логичное значение, так как это бустинг

ru_name из world skills : без навыка

Средним - для численных признаков, для которых не определена константа

Средние по зарплатам по городам

# Моделирование

**Опробованные алгоритмы:** линейная регрессия, бустинг, стэккинг.

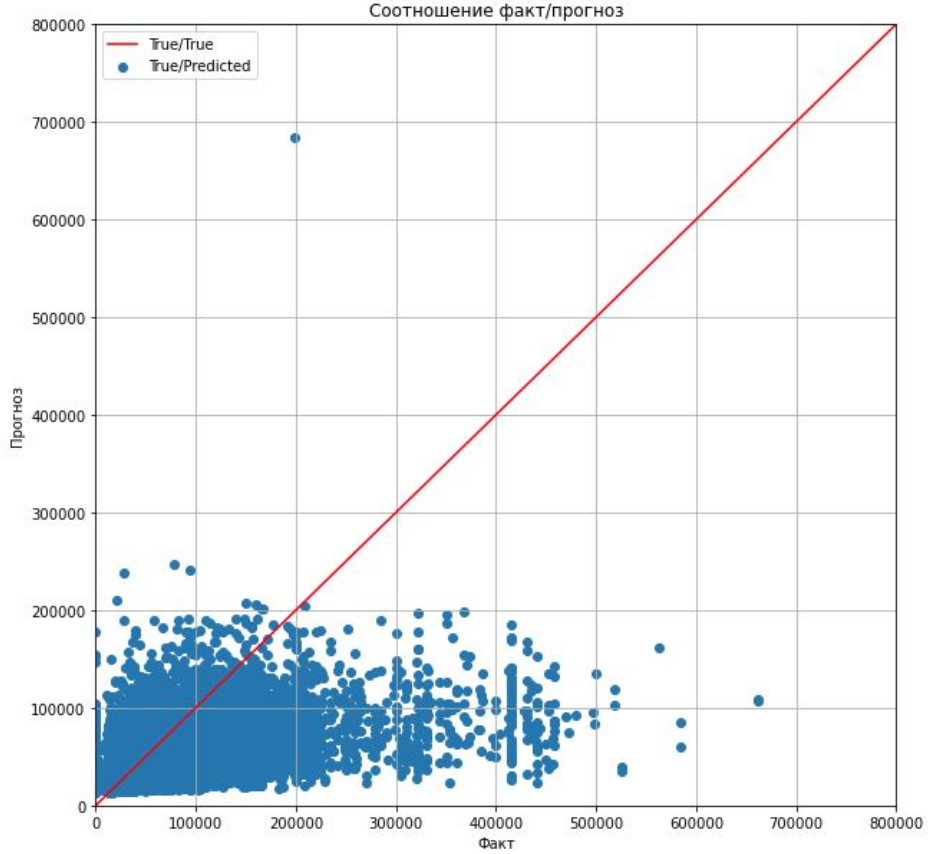
**Выбранная модель:** Бустинг

**Почему именно бустинг:**

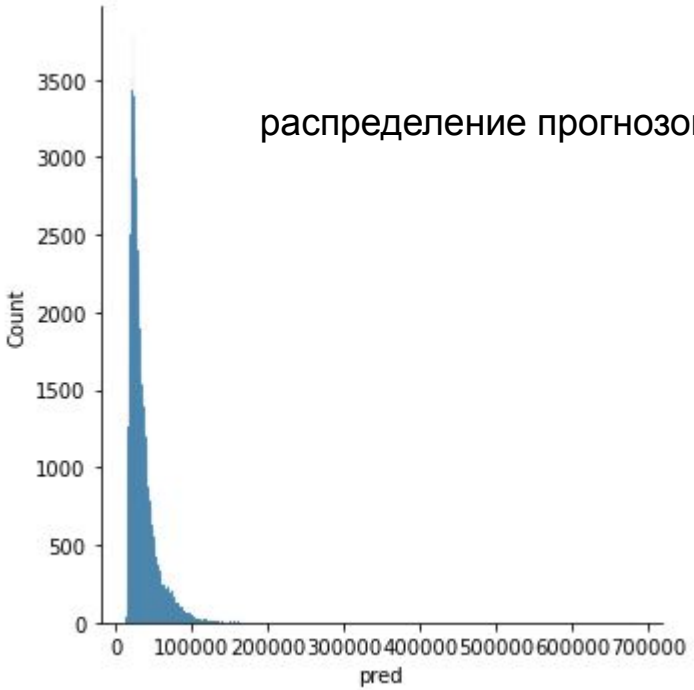
- Древесная модель -- хорошо работает с категориальными признаками (особенно с One-Hot)
- Бустинг -- это ансамбль моделей → меньше переобучения
- В бустинге каждый следующий алгоритм старается исправить ошибки предыдущего.
- Обучается быстрее, чем нейронная сеть или стэккинг.
- Не требует вычислительных мощностей в отличии от нейронной сети.
- Находит нелинейные зависимости в отличие от линейной регрессии
- Не требует нормировать данные перед подачей в модель и подбирать подходящий тип масштабирования.

Онлайн-хакатон
Pandemic Data Hack <data> <code>

# Моделирование




RMSLE = 0.9979:, MAE = 14709.42



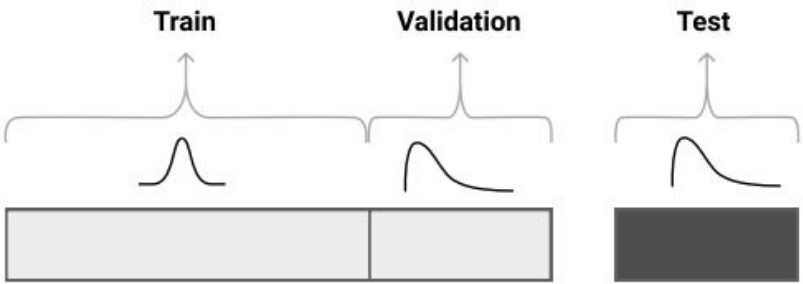

распределение прогнозов

Среднее RMSLE = 1.0180411364466664
Среднее MAE = 14677.922664581725

# Моделирование

**Adversarial validation** -- Создание валидационного датасета похожего на контрольную выборку.

1) Объединяем трейн и тест. Создаем метку “source” (0-train,1-test)
2) Обучаем классификатор, разделяющий трейн и тест.
3) Сортируем сэмплы из трейна в порядке их “похожести” на тест сет.
4) Создаем валидационный сет -- n сэмплов наиболее похожих на тест
5) Проводим кросс-валидацию на созданном валидационном сете



# Нереализованные идеи

1. Поиск данных для компаний: рейтинг, годовой оборот, количество сотрудников
2. Обучение нейронной сети
3. Заполнить пропуски в значении пола по тексту в признаках опыта работы
4. Выделить отрасли промышленности для компаний
5. Выделить в достижениях звания валидные для всего рынка труда (Заслуженный работник машиностроения)

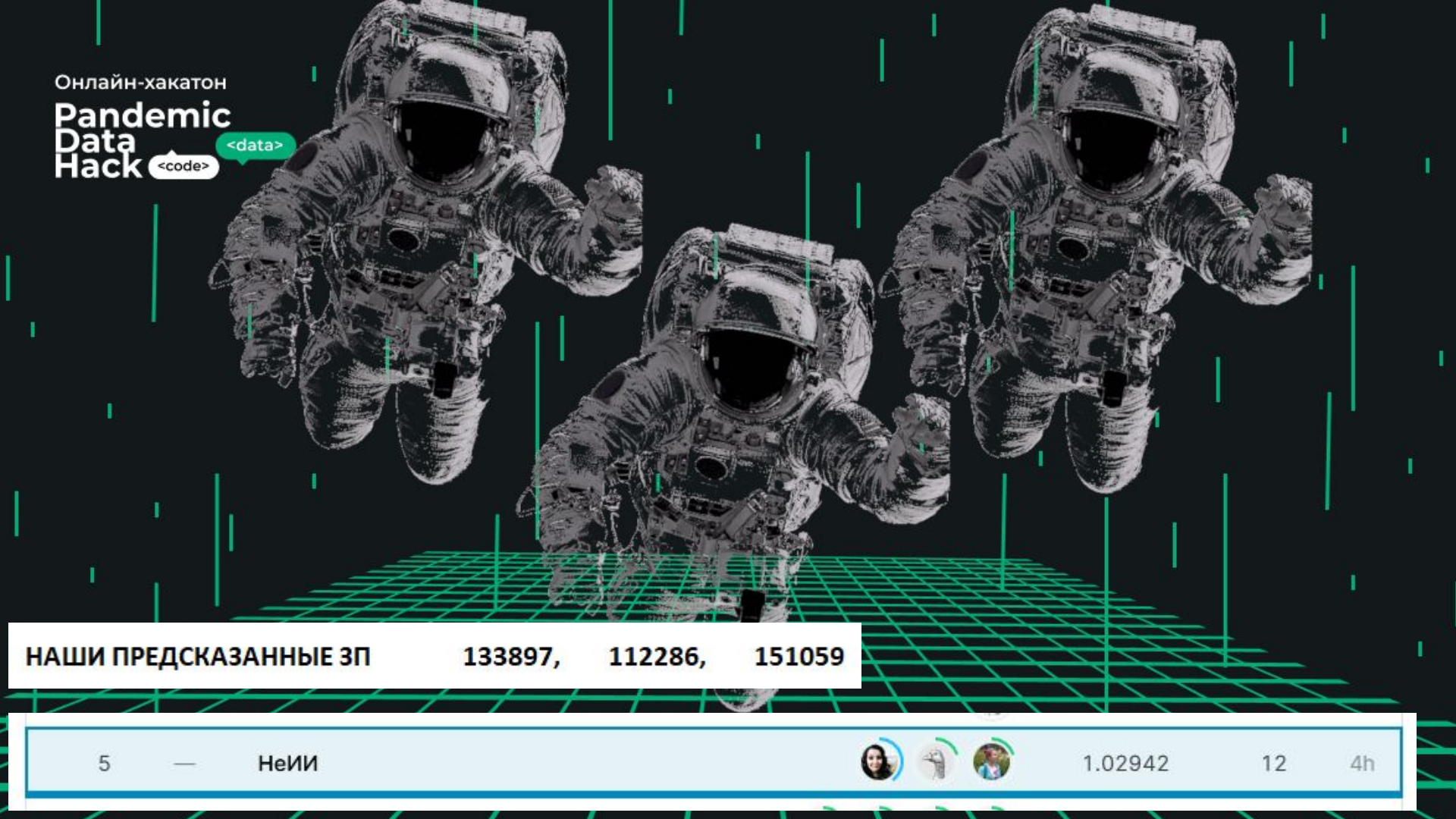
Онлайн-хакатон
Pandemic Data Hack <data> <code>

# Команда

Майя
maya-ami

Екатерина
Harunatsuko

Марина
PinkUrchin




НАШИ ПРЕДСКАЗАННЫЕ ЗП 133897, 112286, 151059

| | | | | | |
|---|---|---|---|---|---|
| 5 | — | НеИИ | 1.02942 | 12 | 4h |

# Команда

Выражаем благодарность организаторам хакатона и желаем им организовать ещё множество таких крутых мероприятий!