

# Вакансии с портала HeadHunter

Куратор: Надежда Гераськина

Команда 3:

Аладинский Георгий Александрович

Дмитрий Тапанович Мандал

Панов Артём Сергеевич

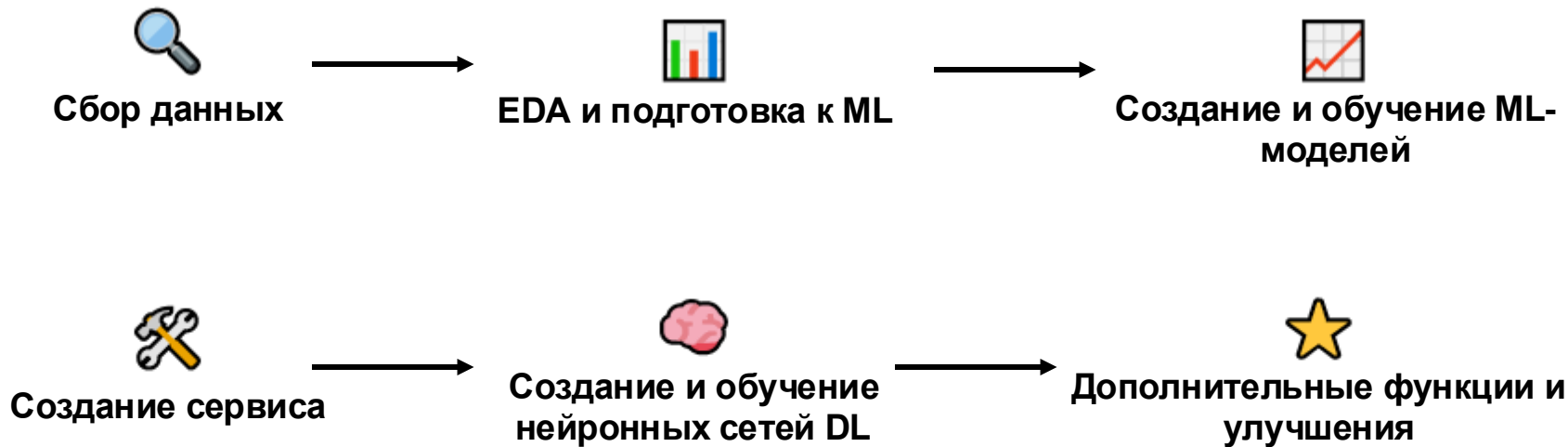
Больбот Елизавета Владимировна



# Задачи и цели

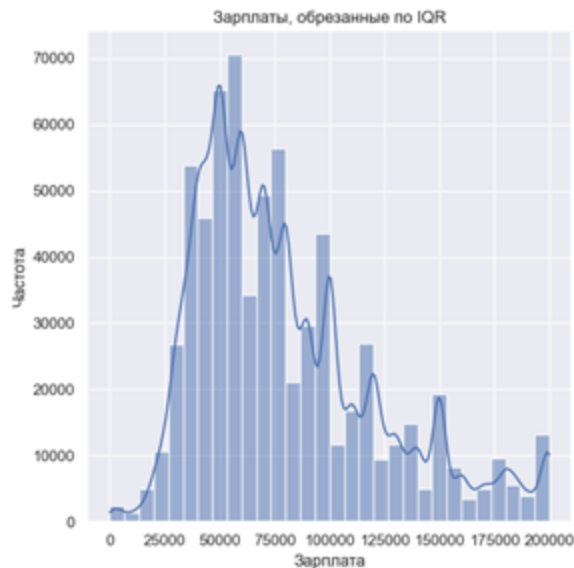
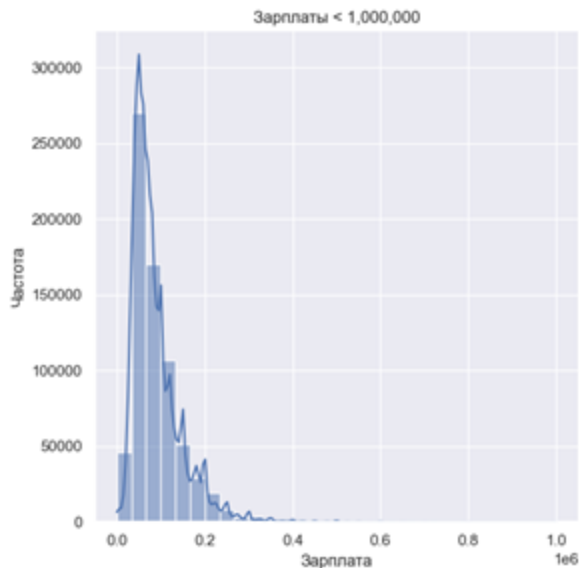
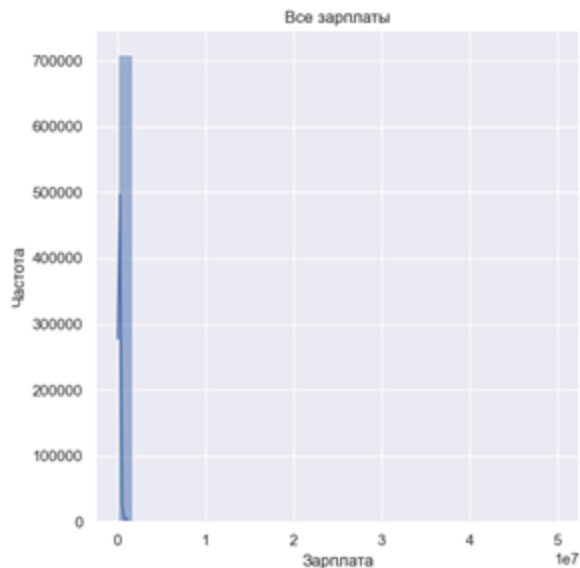
**Цель проекта:** Создать сервис, который предсказывает уровень заработной платы на основе анализа вакансий с помощью методов машинного и глубокого обучения.

## Цели на год:



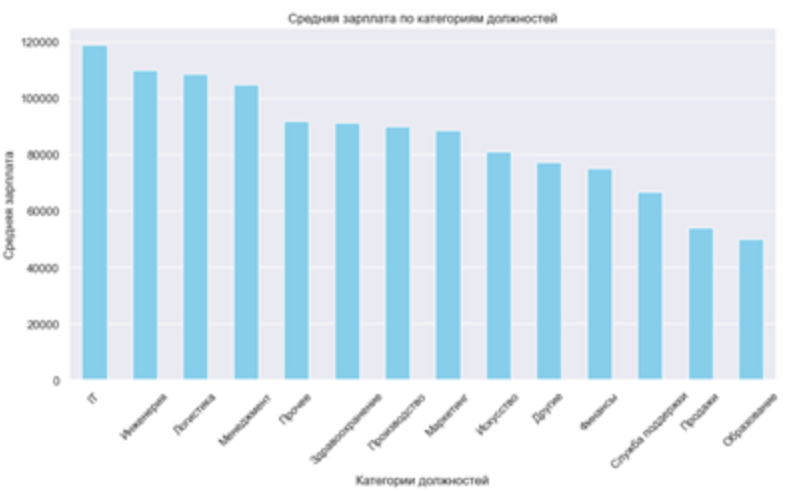
# Описание данных

- Размер датасета: 709 572 строк, 122 признака.
- Некоторые признаки содержат до 99% пропущенных значений.
- Признаки описывают вакансии, включая зарплаты, навыки, тип занятости, график работы, регион и другие категории.
- Диапазон зарплат варьируется от минимальных до очень высоких.



# EDA

- Обработаны пропущенные значения, удалены нерелевантные признаки
- Объем данных после обработки: 709 524 записей, 43 признака
- Больше всего вакансий ориентированы на следующие города: Москва, Санкт-Петербург, Екатеринбург, Новосибирск, Красноярск
- Самая высокая средняя заработная плата в IT и инженерии, а самая маленькая в образовании



# Метрики качества

 Абсолютные

ошибки

 RMSE - (Root Mean Squared Error)

 MAE (Mean Absolute Error)

 Ошибки в процентах

 SMAPE (Symmetric Mean Absolute Percentage Error)

 Качество интерпретации

данных

  $R^2$  (Коэффициент детерминации)

 Устойчивость к

выбросам

 MedAE (Median Absolute Error)

# Бейзлайн-модель

Бейзлайн был построен используя **среднее значение** зарплаты как предсказание для всех записей.

Полученные метрики:

- RMSE – 99590.31
- MAE – 42519.632
- $R^2$  – 0.0
- SMAPE – 46.64%
- MedAE – 34089.938

Дальнейшие шаги: применение более сложных подходов

# Улучшение бейзлайна

Были приняты попытки улучшения бейзлайна, путем предобработки данных с помощью: StandardScaler, OneHotEncoder, LabelEncoder. Так же были обучены различные модели.

## DecisionTreeRegressor

- RMSE – 51413.942
- MAE – 26090.648
- $R^2$  – 0.4947
- SMAPE – 29.31%
- MedAE – 15788.492

## LinearRegression

- RMSE – 64151.253
- MAE – 32800.14
- $R^2$  – 0.2134
- SMAPE – 36.36%
- MedAE – 23599.383

## RandomForestRegressor

- RMSE – 44484.304
- MAE – 21484.424
- $R^2$  – 0.6218
- SMAPE – 24.32%
- MedAE – 12514.5

## CatBoost

- RMSE – 54271.69
- MAE – 29581.659
- $R^2$  – 0.4818
- SMAPE – 33.25%
- MedAE – 21173.342

# Улучшение бейзлайна

## AdaBoost

- RMSE – 52624.323
- MAE – 31336.274
- $R^2$  – 0.4707
- SMAPE – 35.46%
- MedAE – 22492.852

## XGBoost

- RMSE – 55360.588
- MAE – 30357.436
- $R^2$  – 0.4608
- SMAPE – 34.26%
- MedAE – 21872.023

## LightGBM

- RMSE – 53473.192
- MAE – 26745.593
- $R^2$  – 0.4969
- SMAPE – 29.52%
- MedAE – 17557.4

## Вывод

Если посмотреть на результаты всех моделей то можно сделать вывод, что лучшей моделью можно, смело, считать **RandomForestRegressor**.



# Улучшение бейзлайна

## Новые признаки:

- snippet\_requirement\_length — длина описания требований.
- employer\_quality — показатель качества работодателя.
- selectivity — показатель избирательности работодателя.
- Seniority Level (Уровень квалификации).
- is\_central\_metro - находится ли станция метро в центре города.

## Снижение размерности

N	Время	Результат
20	8 мин 17 сек	<span>●</span> RMSE – 44484.304 <span>●</span> MAE – 22287.332 <span>●</span> R <sup>2</sup> – 0.4793 <span>●</span> SMAPE – 25% <span>●</span> MedAE – 12941.199
25	9 мин 8 сек	<span>●</span> RMSE – 44846.4 <span>●</span> MAE – 21712.37 <span>●</span> R <sup>2</sup> – 0.6156 <span>●</span> SMAPE – 24.57% <span>●</span> MedAE – 12686.5
30	10 мин 8 сек	<span>●</span> RMSE – 44679.432 <span>●</span> MAE – 21522.357 <span>●</span> R <sup>2</sup> – 0.6236 <span>●</span> SMAPE – 24.37% <span>●</span> MedAE – 12560.01
35	10 мин 22 сек	<span>●</span> RMSE – 44375.096 <span>●</span> MAE – 21542.446 <span>●</span> R <sup>2</sup> – 0.6184 <span>●</span> SMAPE – 24.39% <span>●</span> MedAE – 12582.5





















# Улучшение бейзлайна с помощью DL

В рамках данной работы мной была предпринята серия экспериментов по улучшению качества модели за счёт применения различных конфигураций **глубоких нейронных сетей (DNN)**.

На первом этапе использовали базовую архитектуру сети с двумя скрытыми слоями — DNN [64,32]. Для повышения качества я последовательно модифицировал модель по следующим направлениям:

- **Оптимизаторы:** тестировались SGD и Adam
- **Нормализация:** добавление Batch Normalization (BN) и Layer Normalization (LN)
- **Инициализация весов:** использовались стратегии инициализации He и Glorot
- **Глубина и ширина сети:** пробовались более глубокие архитектуры [128, 64, 32], [256, 128, 64, 32], [64x5], [512, 256]

## Топ результатов

Конфигурация	Результат
Обычный RF	 RMSE – 44484.304  MAE – 21484.424  $R^2$ – 0.6218  SMAPE – 24.32%  MedAE – 12514.5
RF с эмбедингами	 RMSE – 45317.9  MAE – 23649.4  $R^2$ – 0.6074  SMAPE – 26.64%  MedAE – 14132.4
TabNet + Embeddings	 RMSE – 49171.3  MAE – 26366.7  $R^2$ – 0.5379  SMAPE – 29.62%  MedAE – 17047.4
DNN [64,32] + Embeddings	 RMSE – 50941.3  MAE – 29103.1  $R^2$ – 0.504  SMAPE – 32.33%  MedAE – 20343.1

# Сервисная часть

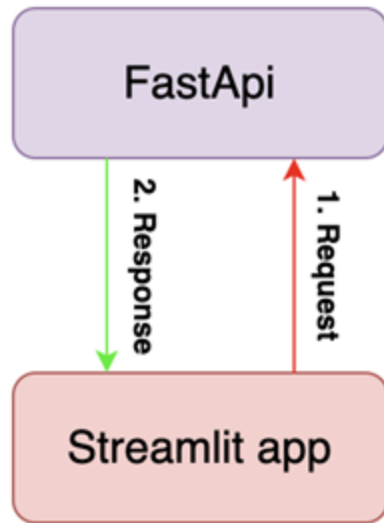
## Back-end часть

Основана на FastApi. На данный момент реализовано 11 ручек, подробнее про них можно почитать [тут](#).



## Front-end часть

Выполнена на библиотеке Streamlit. На данный момент реализовано 5 страниц, подробнее о каждой можно почитать [тут](#).



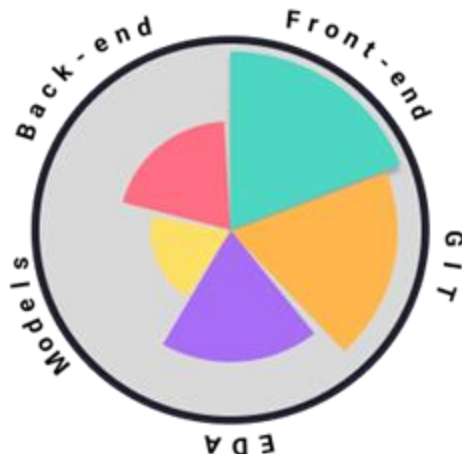
\* Работа приложения будет показана в конце презентации

# Распределение работы

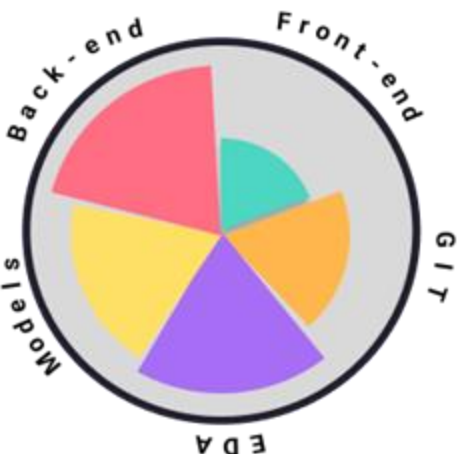
Мандал.Д.Т



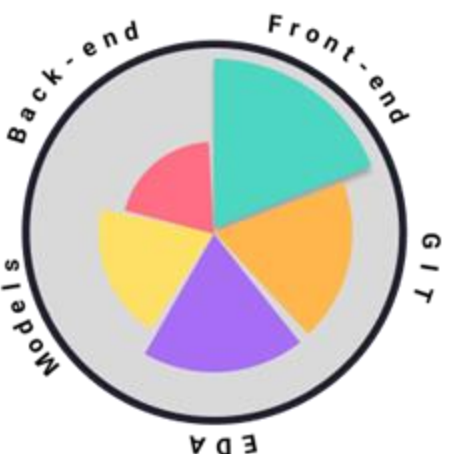
Аладинский.Г.А



Панов.А.С



Большот.Е.В



# Дальнейшие планы

## Улучшение визуальной части:

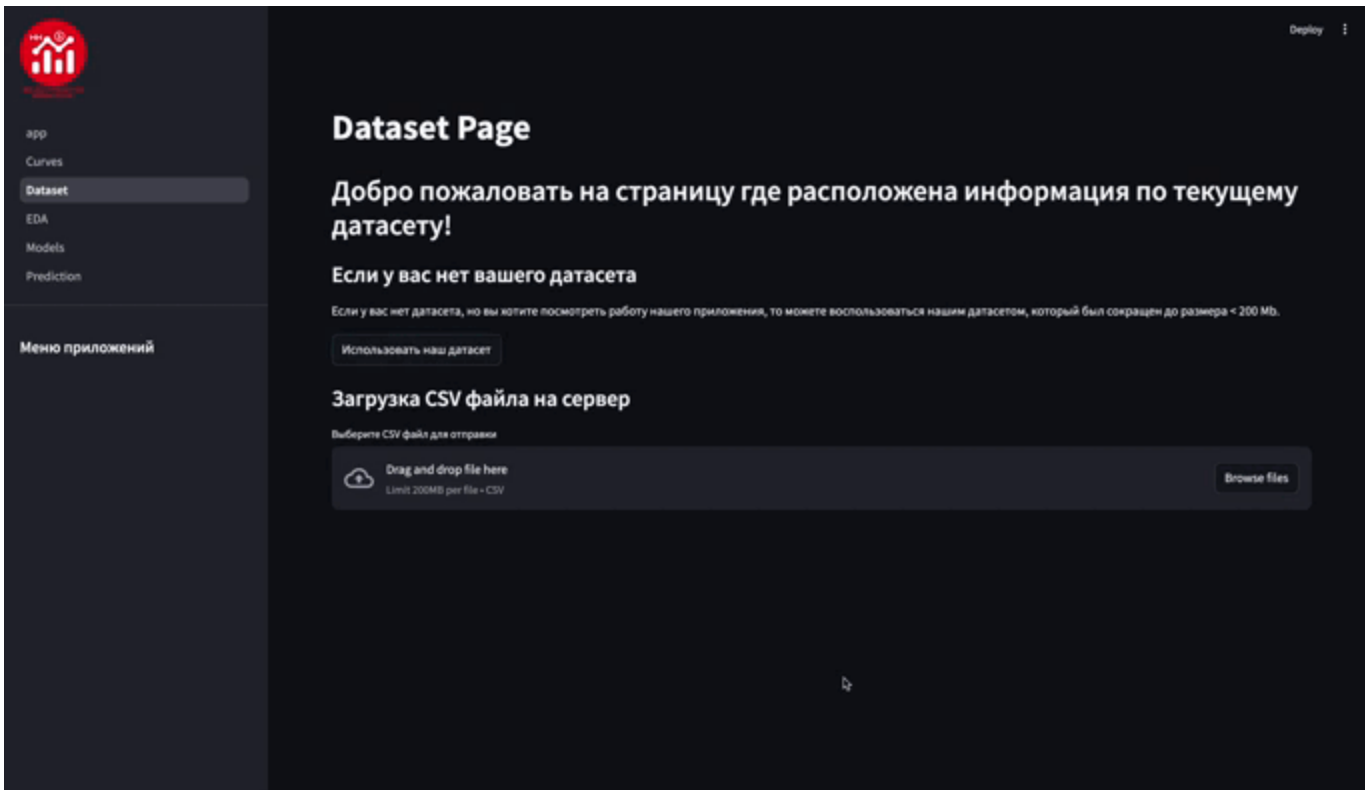
- Добавление новых страниц
- Улучшение интерфейса приложения
- Добавление новых возможностей работы с моделями

## Интеграция:

- Создание своего TG-бота
- Деплой Streamlit приложения



# Работа приложения



The screenshot shows the MLflow web interface. On the left is a dark sidebar with the MLflow logo and a menu. The main content area has a dark background with white text. At the top right of the main area is a 'Deploy' button. The page title is 'Dataset Page'. Below it is a large heading 'Добро пожаловать на страницу где расположена информация по текущему датасету!' (Welcome to the page where information about the current dataset is located!). This is followed by a sub-heading 'Если у вас нет вашего датасета' (If you don't have your dataset). A paragraph explains that if no dataset is present, users can use the provided dataset, which is a reduced version of the original (under 200 Mb). Below this is a button 'Использовать наш датасет' (Use our dataset). The next section is 'Загрузка CSV файла на сервер' (Upload CSV file to server), with a sub-instruction 'Выберите CSV файл для отправки' (Select CSV file for sending). A large dark box contains a cloud icon, the text 'Drag and drop file here', and 'Limit 200MB per file - CSV'. To the right of this box is a 'Browse files' button.

Deploy

app  
Curves  
Dataset  
EDA  
Models  
Prediction

Меню приложений

## Dataset Page

### Добро пожаловать на страницу где расположена информация по текущему датасету!

#### Если у вас нет вашего датасета

Если у вас нет датасета, но вы хотите посмотреть работу нашего приложения, то можете воспользоваться нашим датасетом, который был сокращен до размера < 200 Mb.

Использовать наш датасет

#### Загрузка CSV файла на сервер

Выберите CSV файл для отправки

Drag and drop file here  
Limit 200MB per file - CSV

Browse files

# Работа приложения

