

МИНОБРНАУКИ РОССИИ
федеральное государственное автономное образовательное
учреждение высшего образования
«Санкт-Петербургский политехнический университет Петра Великого»
Институт дополнительного образования
Высшая инженерная школа

ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА

«Разработка автоматизированного отчёта «Показатели рынка труда в РФ»»»

по программе профессиональной переподготовки: «Анализ данных на языке Python»

Выполнила: Михайлова Е.В.
Руководитель: Мещеряков А.О.

Санкт-Петербург
2022

Область исследования

Рынок труда представляет собой систему общественных отношений в согласовании интересов работодателей и наемной рабочей силы.



По определению А.И. Рофе: «Рынок труда — это составная часть структуры рыночной экономики, которая функционирует в ней наряду с другими рынками: сырья, материалов, товаров народного потребления, услуг, жилья, ценных бумаг и др.»

Цели и задачи исследования

- выгрузка первоначальных данных, преобразование и разметка;
- анализ данных, базовые статистики и распределения переменных;
- тестирование линейной регрессии, гребневой регрессии, регрессии Лассо, градиентного бустинга. Выбор модели для прогнозирования минимальной заработной платы ;
- разработка отчёта с интерфейсом для прогнозирования минимальной заработной платы по отрасли



Данные для исследования

Данные для исследования взяты на сайте
«Инфраструктура научных-исследовательских данных».

Первоисточник данных: Информационно-аналитическая система
Общероссийская база вакансий «Работа в России»



- **20 млн** записей

- данные по **33** отраслям

- за **4** года с 2018 по конец 2021 года

https://data-in.ru/data-catalog/datasets/186/#dataset-custom_tab_56

<https://trudvsem.ru/>

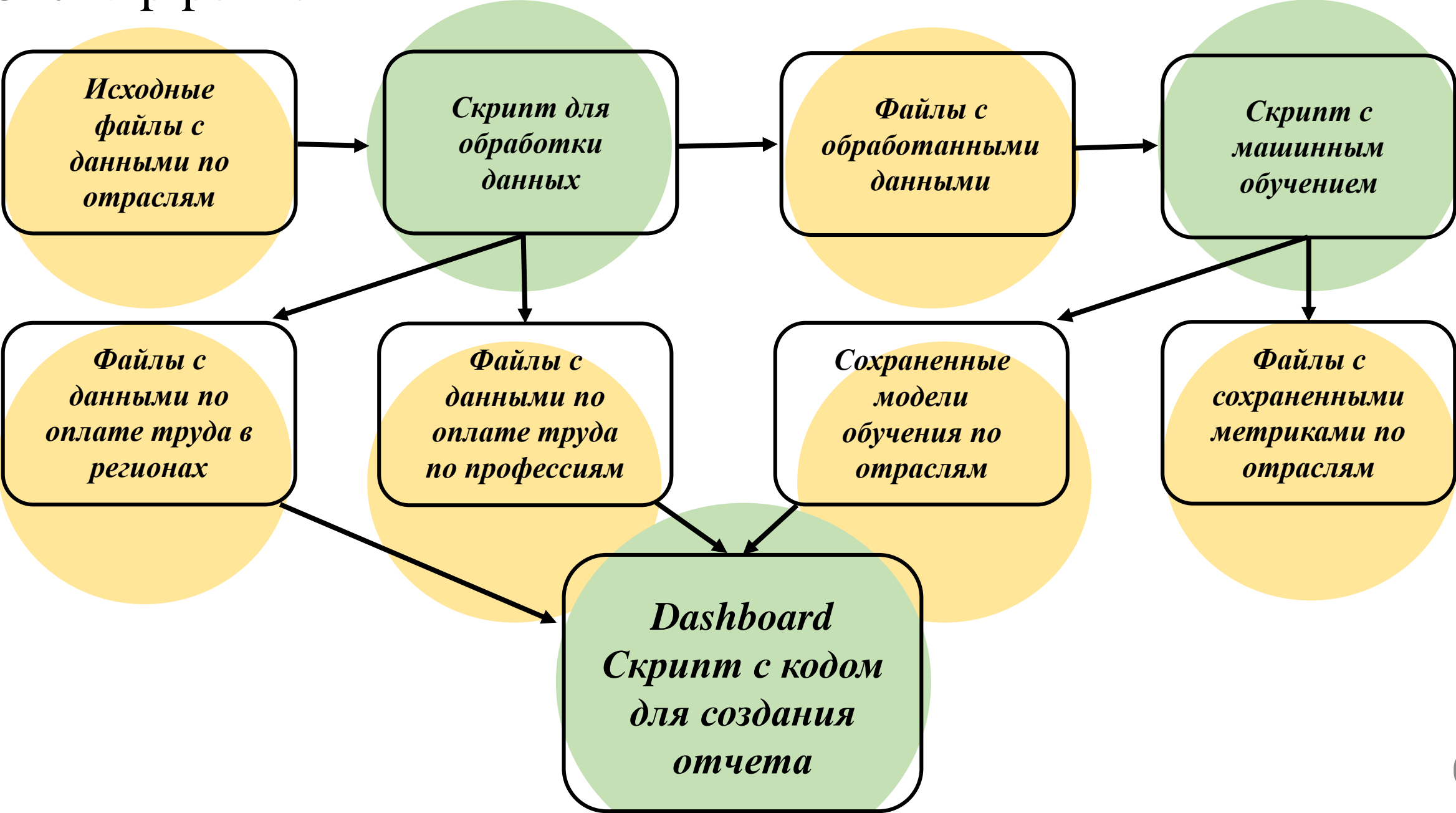
Инструменты

В рамках данной ВКР:

- были использованы библиотеки Python – pandas, matplotlib, lightgbm, sklearn, streamlit;
- рассмотрены линейная регрессия, регрессия Лассо, гребневая регрессия и градиентный бустинг



Схема pipeline



Обработка данных

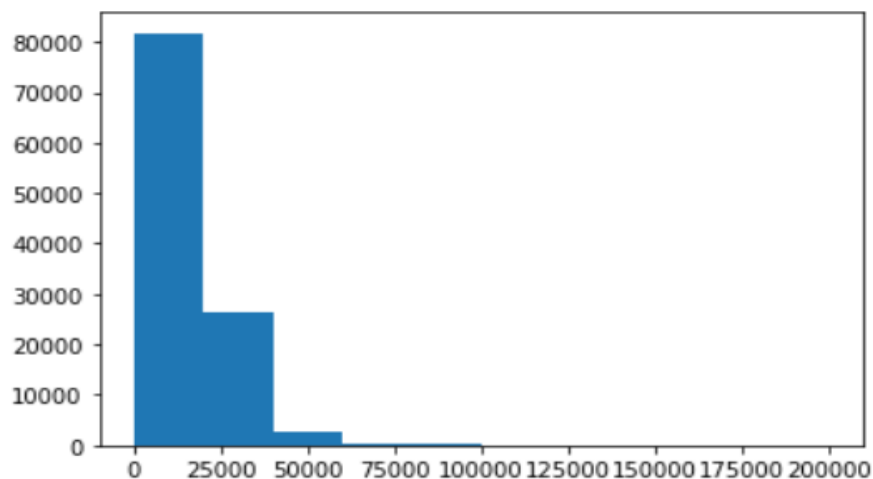
Предварительно файл размером 21 ГБ был разделен на файлы в разрезе отраслей для оптимизации обработки данных.

Удалены строки, в которых не указана заработная плата.

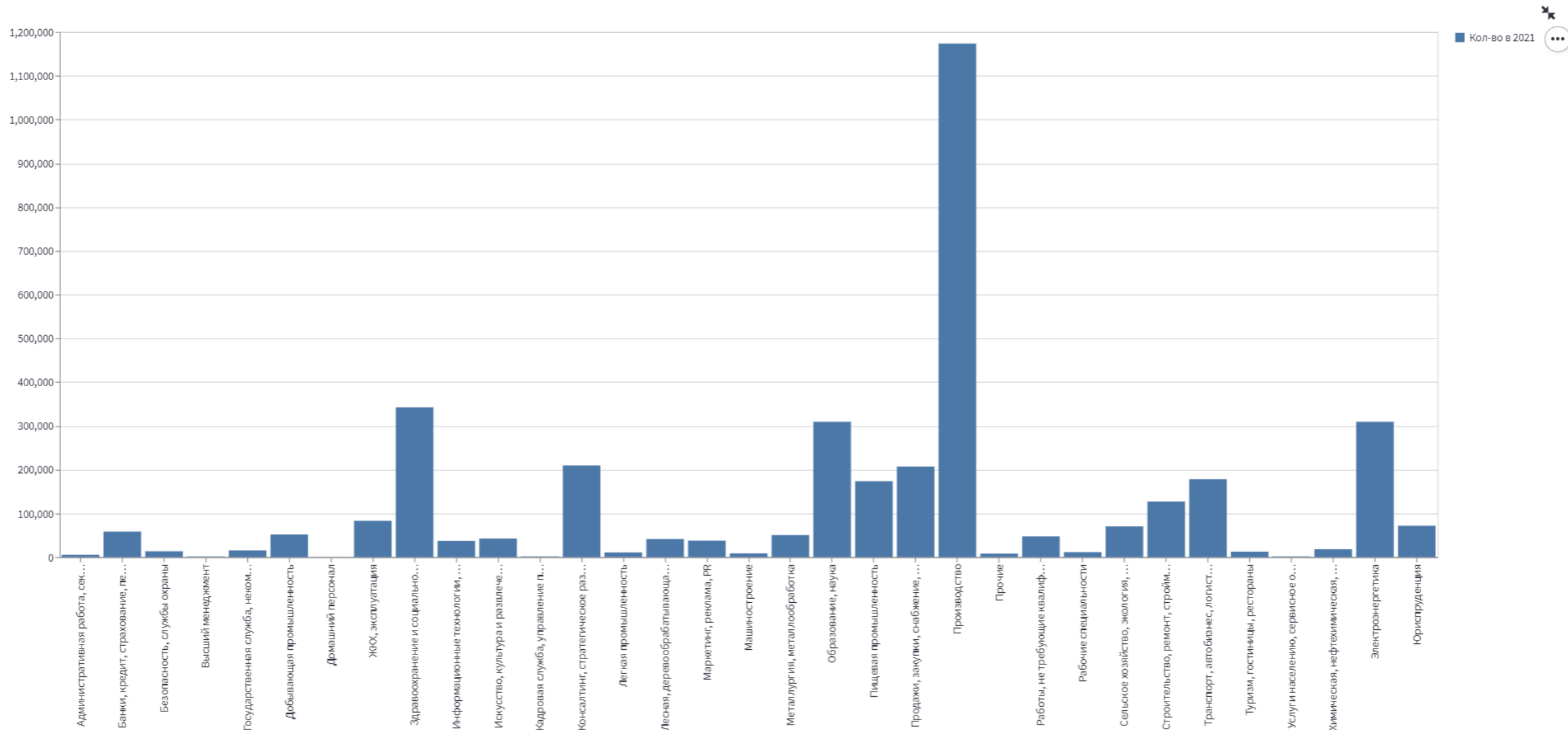
Удалены выбросы по оплате труда превышающие размер квантиля 99.

МАХ вакансий «Производство» - 1 173 592,
MIN вакансий «Домашний персонал» - 225

МАХ заработная плата - 5 533 757 руб.
MIN заработная – 0 руб.



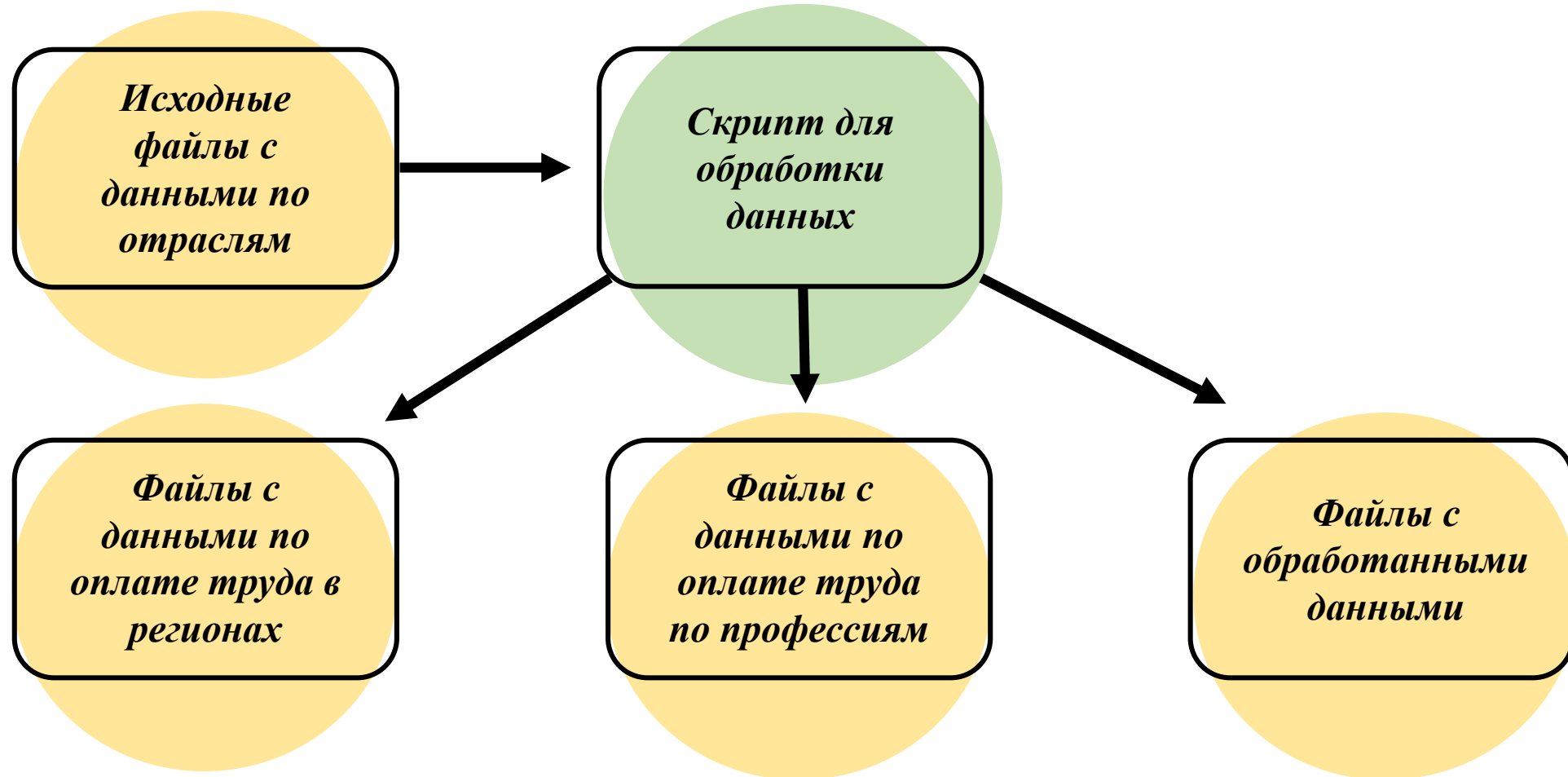
Статистические данные



Статистические данные

Отрасль	Кол-во в 2018	Кол-во в 2019	Кол-во в 2020	Кол-во в 2021	Средняя зарплата 2018	Средняя зарплата 2019	Средняя зарплата 2020	Средняя зарплата 2021	Медиана зарплат 2018	Медиана зарплат 2019	Медиана зарплат 2020	Медиана зарплат 2021	Минимальная зарплата	Максимальная зарплата	Квартиль 25	Квартиль 75
Работы, не требующие	84477	49944	44146	47542	14423,88	15704,75	17728,84	19289,55	12837	22000	25000	25000	0	200000	12000	18048
Образование, наука	250431	200343	275531	309236	15592,40	16575,89	17889,17	18816,76	13000	20000	21781	25000	0	2474004	12792	20000
Искусство, культура и	39072	29838	37969	42552	15722,67	16964,14	18240,40	19316,49	13000	20000	22000	22500	0	450000	12130	20000
Пищевая промышленность	139699	106572	132876	173463	16520,33	17848,94	19206,60	20828,60	15000	14664	15163	16000	0	360000	13000	22000
Продажи, закупки, снабжение	204200	146735	171521	206834	17303,91	18673,07	20266,90	22443,34	15000	14186,5	15000	16000	0	2155307	13500	23000
Прочие	3914	1562	807	8107	13832,00	18295,77	17091,03	21752,16	11163	16000	18000	20000	0	180000	12420	22000
Банки, кредит, страхование	64951	47328	54258	58370	19703,63	21421,76	22574,27	24175,20	18000	35000	30000	27000	0	1813435	15000	25000
Легкая промышленность	8719	6532	9357	10735	18774,88	19955,88	21591,56	23210,23	17000	18000	20000	21000	0	150000	15000	25000
ЖКХ, эксплуатация	66222	54382	68985	82988	18196,56	19430,49	21440,52	22949,17	15326	19900	20000	21312	0	1437872	14000	25000
Лесная, деревообработка	41982	30972	39479	41337	19801,08	21605,62	23072,92	24664,84	18000	23000	31500	35000	0	372000	15000	26000
Туризм, гостиницы, рестораны	16697	11584	10490	12547	17853,30	19065,62	20994,16	22778,41	15000	30000	30325	35000	0	305001	13000	24000
Сельское хозяйство, охота и рыболовство	72333	53800	71283	70398	17498,90	18943,63	20863,88	23128,16	15000	20000	20000	22542	0	1471080	13950	25000
Производство	1109707	813177	1011434	1173592	19460,74	20919,07	22383,68	24090,62	16100	16500	19408	20000	0	3163734	13500	25330
Безопасность, службы	16024	11673	11834	13260	19250,94	20322,76	22831,18	24079,97	16910	13256	15000	15500	0	1116300	13000	25000
Административная работа	8679	5336	4976	5228	23034,18	22587,20	23417,33	24900,27	22000	20000	20000	20000	0	999999	16000	28000
Рабочие специальности	7603	5945	9006	11406	28674,72	30592,07	32182,76	30665,56	23000	17600	19000	20000	0	90000	13000	25000
Металлургия, металлообработка	41143	32675	43535	50322	24600,53	26163,13	27569,99	29981,98	22000	18000	19500	20000	0	313000	19408	32000
Транспорт, автобизнес	156898	123335	147268	178219	22890,93	24450,58	26546,32	28889,84	20000	22000	22000	23000	0	710000	15000	30000
Юриспруденция	61816	55491	70634	71759	21241,93	22534,44	24335,95	24795,37	20000	15000	17000	18063	0	450000	15000	30000
Услуги населению, сервис	1287	859	1762	1819	26797,94	22883,51	23338,01	22300,40	21000	18451	22123	23100	0	190000	15000	30000
Химическая, нефтехимия	12257	9922	13949	18075	21008,21	22839,11	25091,41	27439,23	18415	20000	21834	22600	0	175000	15000	30000
Здравоохранение и социальное обеспечение	358886	277150	347854	342238	21388,95	24126,89	26336,48	27040,98	18000	25000	28000	30000	0	2317815	14665	30000
Государственная служба	20834	13208	16182	15360	19777,78	21731,18	25752,61	27315,07	17000	25000	25000	28700	0	2200039	14000	30000
Электроэнергетика	250431	200343	275531	309236	15592,40	16575,89	17889,17	18816,76	13000	17000	19131	20000	0	370000	19000	35000
Консалтинг, стратегические исследования	191082	153656	190093	209515	24890,68	26457,81	28103,75	29092,72	20000	21000	23000	25000	0	3067337	15000	32000
Информационные технологии	31461	26443	35112	36878	25500,11	27167,70	29592,83	32616,99	20000	25000	27000	30000	0	687469	15000	35000
Машиностроение	6889	5308	7538	8630	28173,19	29994,05	31955,67	35332,36	25000	18000	20000	20000	0	240000	20000	40000

Схема работы скрипта для обработки данных



Прогнозирование минимальной заработной платы

Отбор моделей для прогнозирования заработной платы осуществлялся на основе следующих решений:

- Линейная регрессия (LinearRegression)
- Регрессия Лассо (Lasso)
- Гребневая регрессия (Ridge)
- Градиентный бустинг (lightgbm)



Тестирование и выбор модели

```
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
```

```
X_train, X_test, Y_train, Y_test = train_test_split(X, Y, test_size=0.2, random_state=10, shuffle=True)
X_train, X_val, Y_train, Y_val = train_test_split(X_train, Y_train, train_size=0.2, random_state=10)
```

```
regressor = LinearRegression()
regressor.fit(X_train, Y_train)
```

```
y_pred_train = regressor.predict(X_train)
y_pred_val = regressor.predict(X_val)
y_pred_test = regressor.predict(X_test)
```

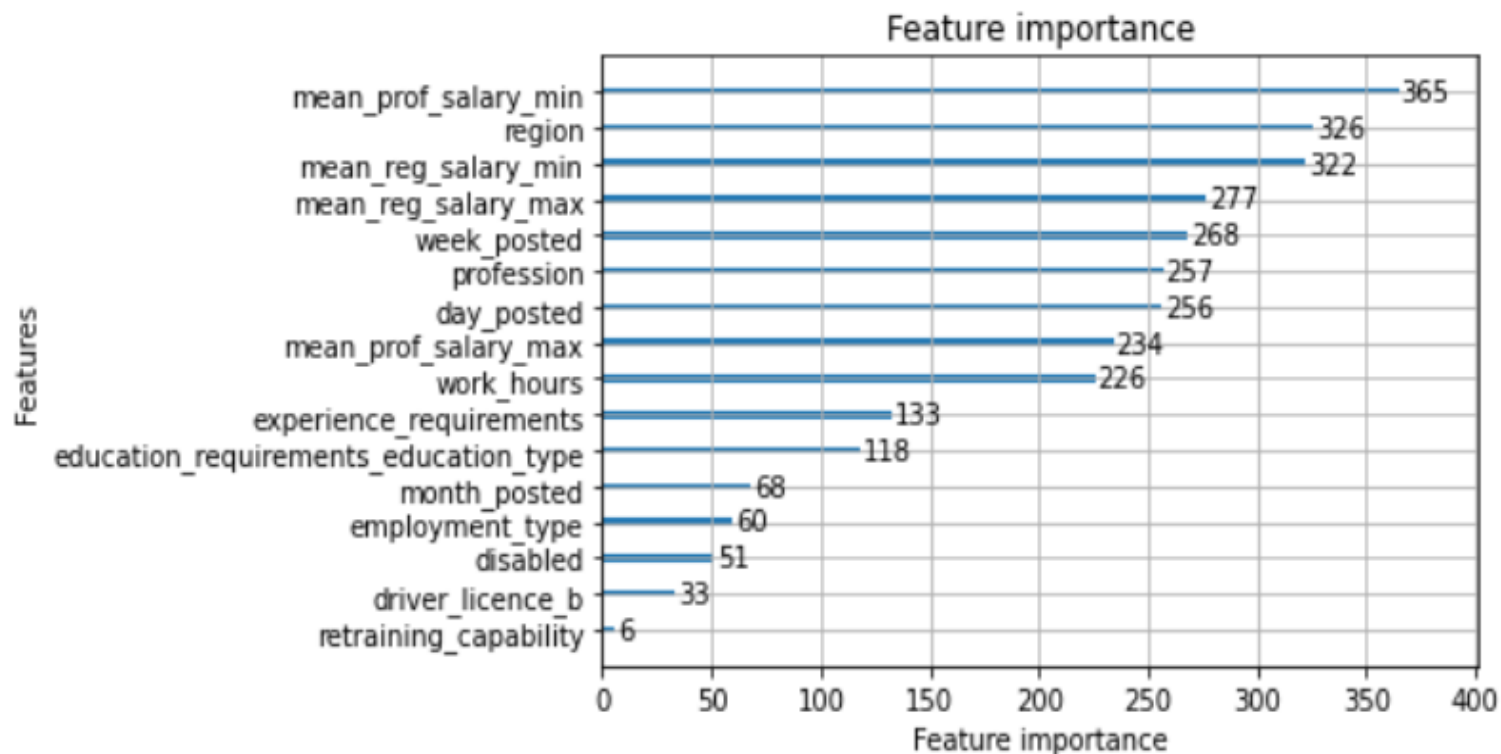
```
print('Mean Absolute Error:', metrics.mean_absolute_error(Y_test, y_test_pred))
print('Mean Squared Error:', metrics.mean_squared_error(Y_test, y_test_pred))
print('Root Mean Squared Error:', np.sqrt(metrics.mean_squared_error(Y_test, y_test_pred)))
print('R2_score Error:', metrics.r2_score(Y_test, y_test_pred))
print('mean absolute percentage error:', metrics.mean_absolute_percentage_error(Y_test, y_test_pred))
```

```
Mean Absolute Error: 2955.3779394213075
Mean Squared Error: 19966612.30084107
Root Mean Squared Error: 4468.401537556922
R2_score Error: 0.6253768997205226
mean absolute percentage error: 0.17994761824886052
```

Тестирование и выбор модели

```
lgbm.plot_importance(lgbm_model)
```

```
<AxesSubplot:title={'center':'Feature importance'}, xlabel='Feature importance', ylabel='Features'>
```

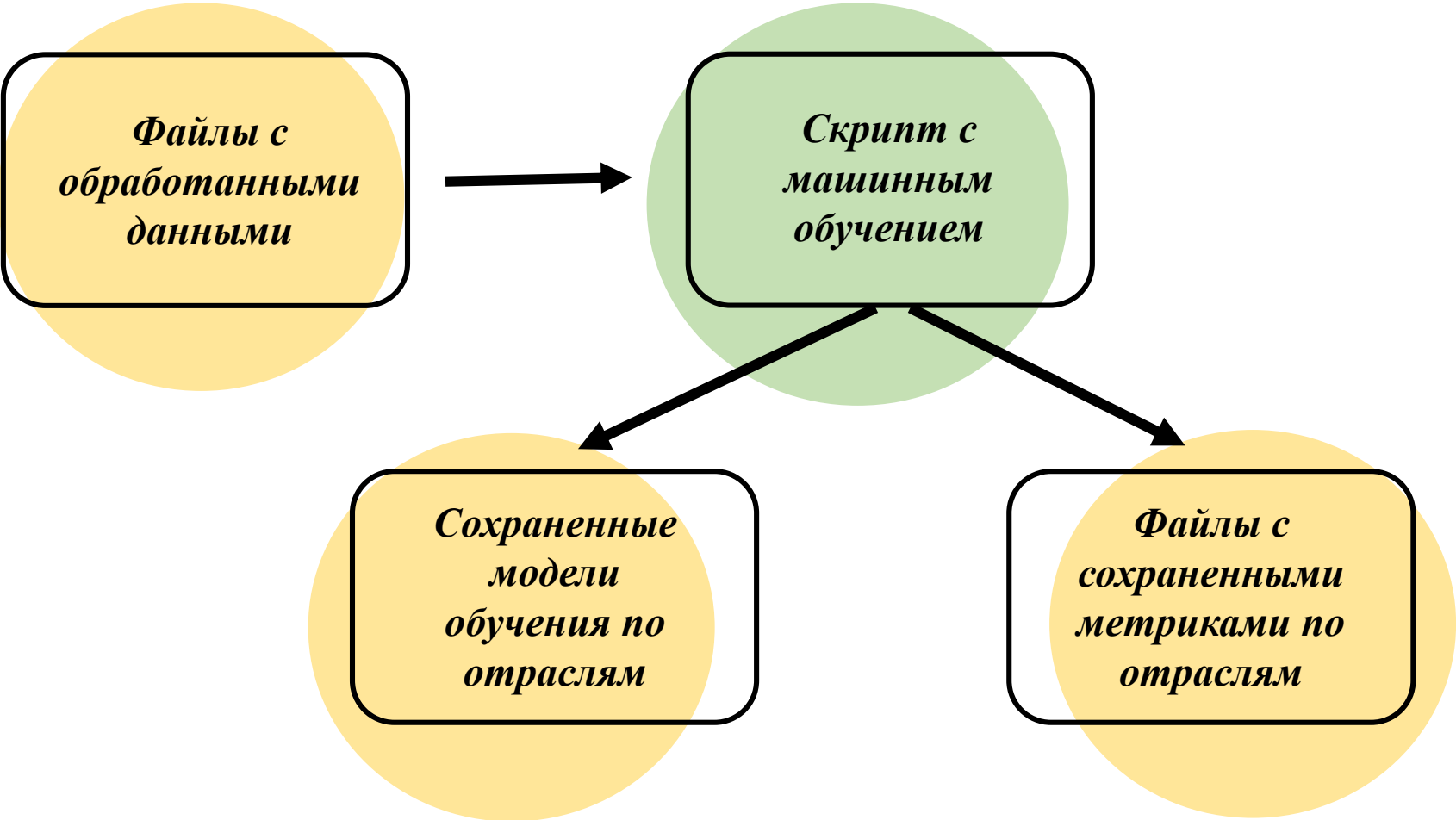


По результатам тестирования методом обучения был выбран **градиентный бустинг**.

Метрики	Линейная регрессия	Регрессия Лассо	Гребневая регрессия	Градиентный бустинг
Mean Absolute Error	3933,17	3933,41	3933,23	2932,2
Mean Squared Error	51604329,26	51596173,38	51592704,67	22499014,87
Root Mean Squared Error	7183,61	7183,047	7182,8	4743,31
Коэффициент детерминации (R2_scor)	0,467251	0,4673361	0,4673719	0,770978
Mean absolute percentage error	0,232492	0,2324967	0,2325051	0,1791843



Схема работы скрипта машинного обучения



Скрипт на обучение и сохранение моделей

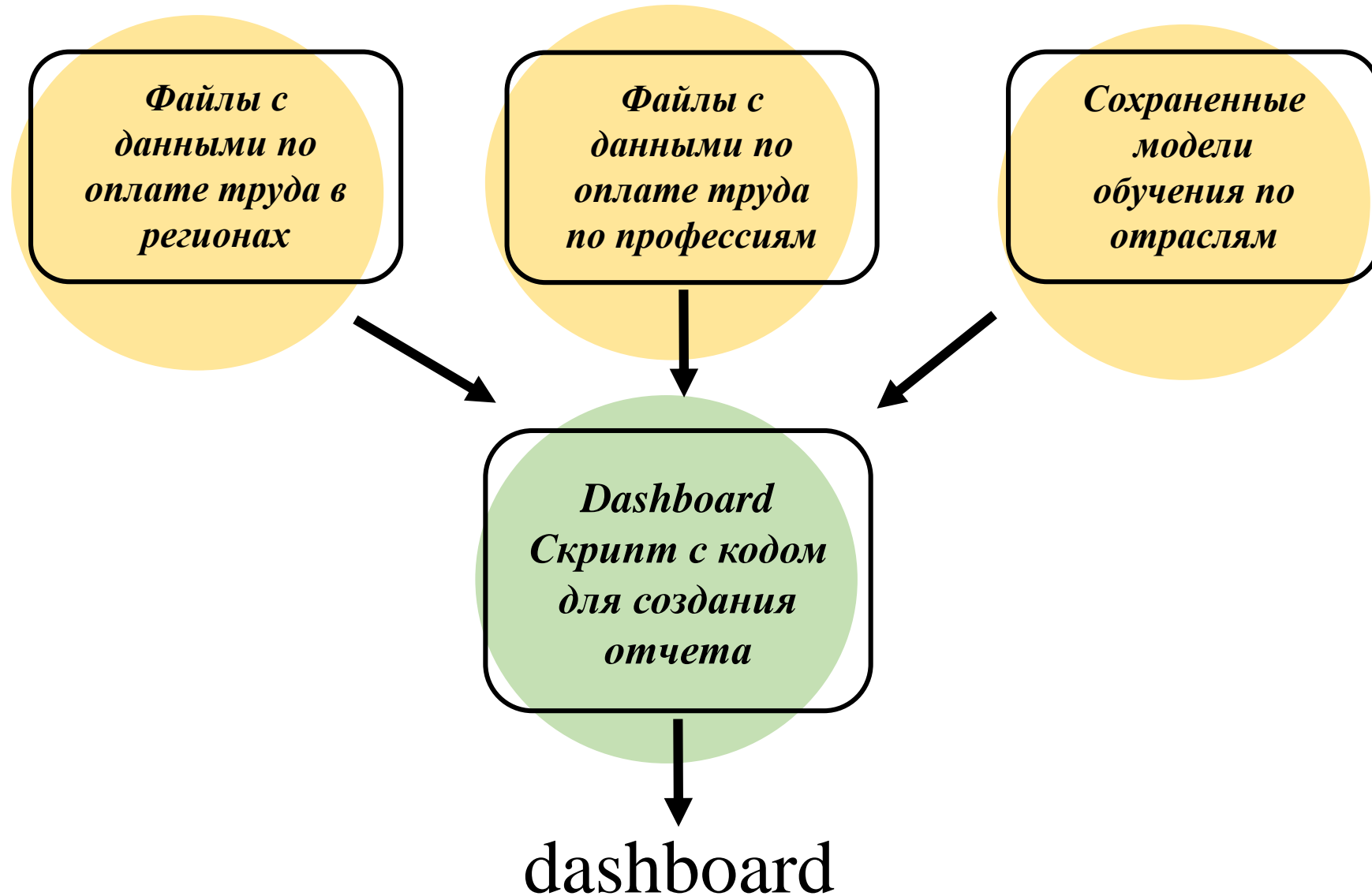
```
def training(name):
    df = pd.read_csv(name)
    X = df[['caring_workers',
           'disabled', 'dms', 'driver_licence_a', 'driver_licence_b', 'driver_licence_c',
           'driver_licence_d', 'driver_licence_e', 'education_academic_degree',
           'education_requirements_education_type', 'employment_type',
           'experience_requirements', 'large_families', 'minor_workers',
           'need_medcard', 'payment_meals', 'payment_sports_activities',
           'premium_size', 'profession', 'region', 'released_persons',
           'retraining_capability', 'single_parent', 'work_hours', 'day_posted',
           'month_posted', 'week_posted', 'mean_reg_salary_max', 'mean_reg_salary_min',
           'mean_prof_salary_max', 'mean_prof_salary_min']]
    Y = df[['base_salary_min']]
    X_train, X_test, Y_train, Y_test=train_test_split(X, Y, test_size=0.2, random_state=10, shuffle=True)
    X_train, X_val, Y_train, Y_val = train_test_split(X_train, Y_train, train_size=0.2, random_state=10)
    params = {
        "objective": "regression",
        "verbosity": -1,
    }
    lgbm_model = train_lgb(X_train, Y_train, params)
    y_train_pred = lgbm_model.predict(X_train)
    y_val_pred = lgbm_model.predict(X_val)
    y_test_pred = lgbm_model.predict(X_test)
    ind = 'models\\' + name.replace('finish_files\\', '').replace('.csv', '')
    lgbm_model.save_model(f'{ind}_mode.txt')
    model_metrics = pd.DataFrame({'Mean Absolute Error': [metrics.mean_absolute_error(Y_test, y_test_pred)],
                                  'Mean Squared Error': [metrics.mean_squared_error(Y_test, y_test_pred)],
                                  'Root Mean Squared Error': [np.sqrt(metrics.mean_squared_error(Y_test, y_test_pred))],
                                  'R2_scor Error': [metrics.r2_score(Y_test, y_test_pred)],
                                  'mean absolute percentage error': [metrics.mean_absolute_percentage_error(Y_test, y_test_pred)]})
    ind2 = 'metrics\\' + name.replace('finish_files\\', '').replace('.csv', '')
    model_metrics.to_excel(f'{ind2}_metrics.xlsx')
```

```
for i in files_name:
    training(i)
```

- Agricultural_mode
- BuldindRealty_mode
- ChemicalAndFuelIndustry_mode
- Communal_mode
- Consulting_mode
- Culture_mode
- DeskWork_mode
- Education_mode
- ElectricpowerIndustry_mode
- Finances_mode
- Food_mode
- Forest_mode
- HomePersonal_mode
- HumanRecruitment_mode
- Industry_mode

Mean Absolute Error	Mean Squared Error	Root Mean Squared Error	R2_scor Error	ean absolute percentage error
3412,705354	23880224,51	4886,73966	0,594712989	0,218530838

Схема работы скрипта dashboard



Dashboard

Отчет предназначен для соискателей. Основная задача ознакомить соискателя с основными показателями и спрогнозировать минимальную заработную плату на которую может претендовать соискатель на основе своих данных.

Панель параметров

Выберите отрасль

Банки, кредит, страхование, пенс... ▾

Регион

г. Санкт-Петербург ▾

Профессия

Агент банка ▾

Академическая степень

Без степени ▾

Образование

Высшее ▾

Показатели рынка труда в РФ

Предполагаемая минимальная заработная плата по вашим данным

36636.45

Сводные данные по отраслям с 2018 по 2021 года

Dashboard

Образование

Высшее ▾

Тип занятости

Полная ▾

График работы

Полный рабочий день ▾

Требуемый опыт работы

5

0 30

Работники, осуществляющие уход за больными членами своих семей

Инвалид

ДМС

Сводные данные по отраслям с 2018 по 2021 года

	Отрасль	Кол-во в 2018	Кол-во в 2019	Кол-во в 2020	Кол-во в 2021
0	Работы, не требующие квалификации	84477	49944	44146	
1	Образование, наука	250431	200343	275531	
2	Искусство, культура и развлечения	39072	29838	37969	
3	Пищевая промышленность	139699	106572	132876	
4	Продажи, закупки, снабжение, торговля	204200	146735	171521	
5	Прочие	3914	1562	807	
6	Банки, кредит, страхование, пенсионное обеспечение	64951	47328	54258	
7	Легкая промышленность	8719	6532	9357	
8	ЖКХ, эксплуатация	66222	54382	68985	
9	Лесная, деревообрабатывающая, целлюлозно-бумажная	41982	30972	39479	

Dashboard

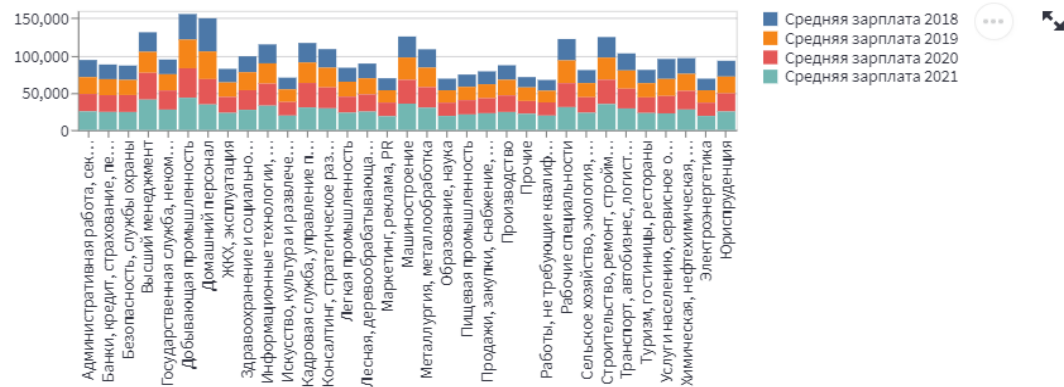
Сменный ▾

Требуемый опыт работы

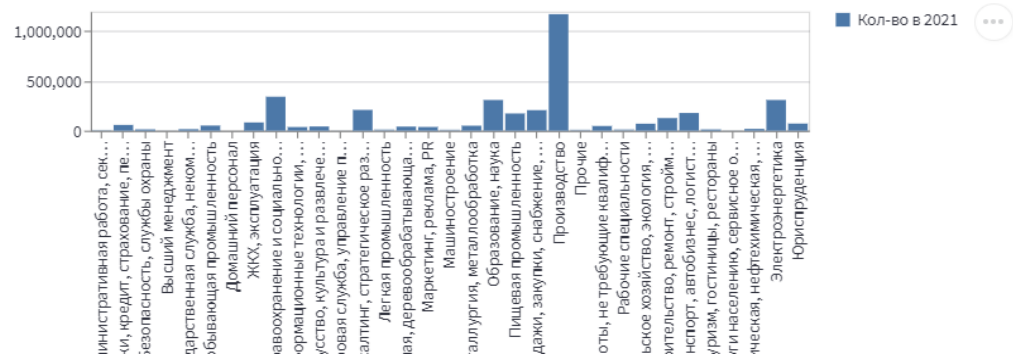
0 30

- Работники, осуществляющие уход за больными членами своих семей
- Инвалид
- ДМС
- Водительские права категории «А»
- Водительские права категории «В»
- Водительские права категории «С»
- Водительские права категории «D»
- Водительские права категории «Е»
- Многодетные семьи
- Несовершеннолетние работники
- Медицинская книжка
- Оплата питания
- Оплата спортивных занятий
- Судимость
- Готовность к переобучению

Средняя оплата труда по годам



Распределение вакансий на 2021 год



Вывод:

В рамках данной выпускной квалификационной работы были:

- проанализированы данные с 2018 по конец 2021 года;
- выведены и сохранены основные статистические данные;
- разработан скрипт для автоматической обработки данных;
- протестированы четыре модели обучения – линейная регрессия, регрессия Лассо, гребневая регрессия, градиентный бустинг;
- в качестве модели обучения выбран градиентный бустинг;
- разработан скрипт для обучения и сохранения моделей по предобработанным данным;
- разработан скрипт для визуализированного отчета по показателям рынка труда для соискателей.

Все поставленные задачи и цели в рамках данной выпускной квалификационной работы были выполнены.

