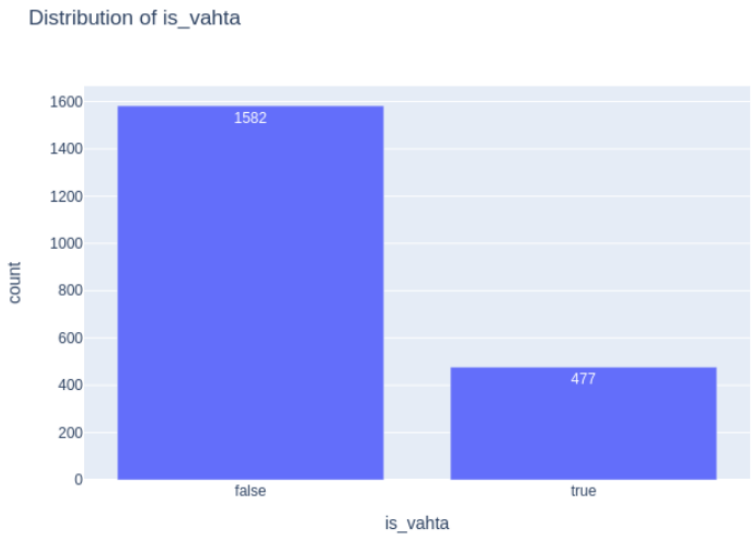
графики + shap

Изменения графиков

- На графики распределения region_name, is_vahta, is_parttime, experience_id были добавлены подписи к столбцам (численные значения)



Во время построения графиков распределения зарплаты на тестовых и тренировочных данных было сделано:

- зарплаты большее 1 миллиона обрезались, т.к. из-за этого страдала читабельность графика. Не было видно основного распределения зарплат.
- зарплаты **округлялись по бинам** (условно при округлении до 10к зарплата 18345 превращалась в 20000). За основу бралось деление на бины в объекте Axes (см. код), в случае если разница между уровнями зп был более 10к, то за разрыв принималось значение в 10к рублей).
- Высчитывался **коэффициент k** на который умножалось значение bins при построении графиков. Это было сделано для того, чтобы столбцы

распределения тренировочных и тестовых данных были примерно одной ширины.

```
koef_pred = max(df['predict_before2'])-min(df['predict_before2'])
koef_act = max(df['salary_new2'])-min(df['salary_new2'])

k = koef_act/koef_pred
print(k)


sns.histplot(data=df, x='salary_new2', bins=int(20*k), kde=True, label=f'salary_actual')
sns.histplot(data=df, x='predict_before2', bins=20, kde=True, label=f'salary_predicted', alpha=0.4)
```

КОД (см. вкладку “task 1, 2”):

tasks_1_2.ipynb

ГРАФИКИ:

https://drive.google.com/drive/folders/1LfG5LSBhSycYL1W70oHN9B3i7Z4aggMW?usp=sharing

Проблема с SHAP

Проблема: вклад фичи 'Руководство и контроль сварочного производства (999_3)’ высок, однако на графике SHAP красные точки очень прижаты к нормали (т.е. фича никак не влияет на зп).

Данный навык встречался в нескольких бундлах, однако такая проблема возникла только в 898. Ниже представлены распределения зарплат в зависимости от наличия навыка (999_3).

- бундл 954

- бундл 833

- **проблемный бундл 898**

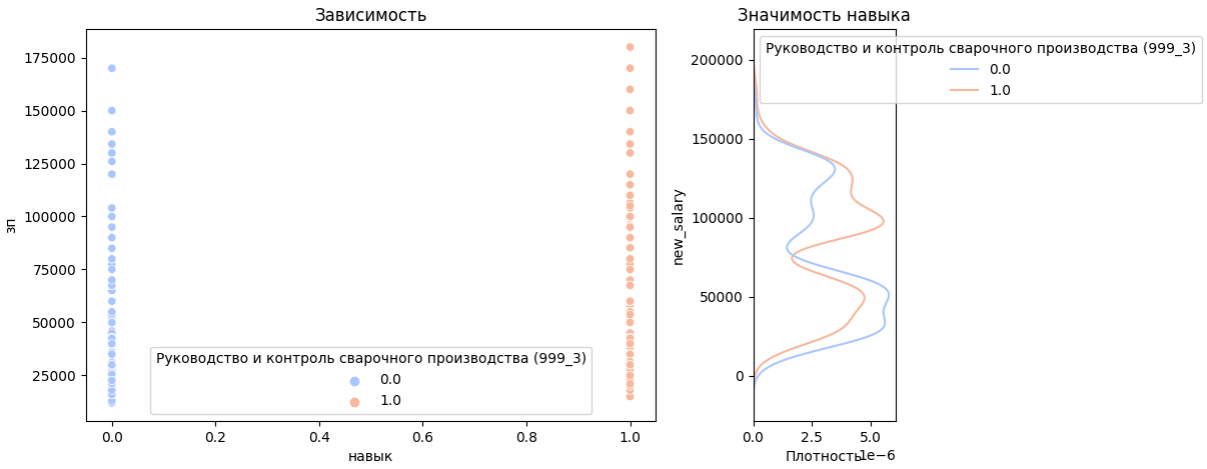

```
1.0    2515
0.0    2192
Name: Руководство и контроль сварочного производства (999_3), dtype: int64
```

Сразу заметно, что в бундле 898 отсутствие и наличие навыка 999_3 примерно одинаково. Также важно, что навык 898 равен 1 на всех уровнях зарплат (хотя по логике, если вклад навыка большой, то встречаться этот навык должен только у высоких зарплат).

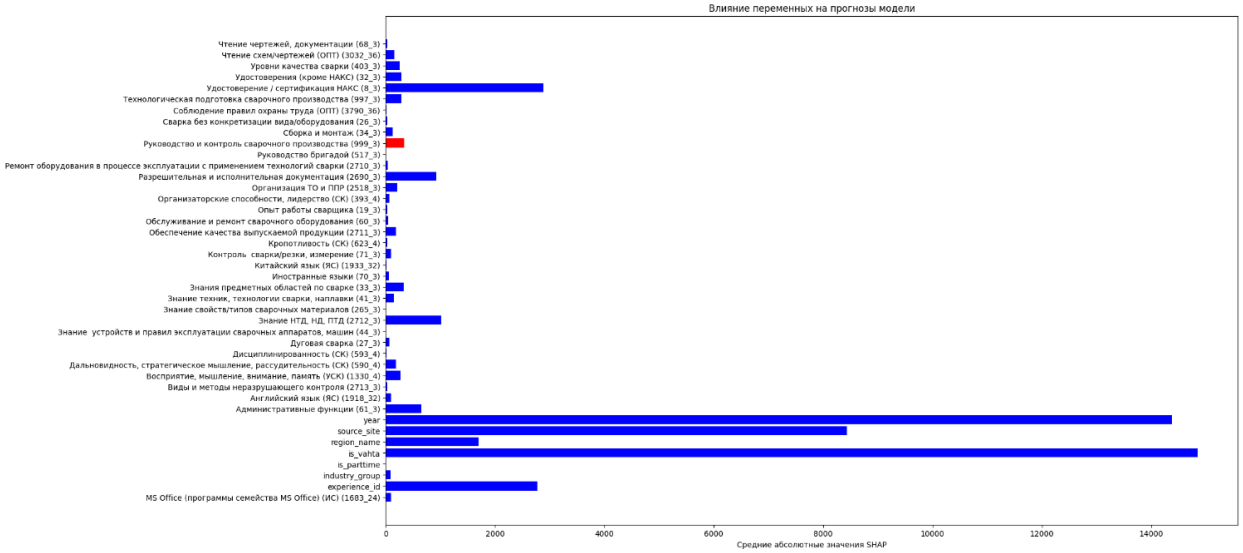
КОД: (см. вкладку task 4))

tasks_1_2_4.ipynb

Для подтверждения предположения для бундла 898 была заново обучена модель catboost. Далее были построены таблицы, дополняющие результаты SHAP. На графике ниже (правый) видно, что плотности распределения зарплат при разных значениях навыка (999_3) примерно одинаковые, заметны лишь небольшие различия для зарплат от 100к до 150к.



Поэтому навык 999_3 оказывает незначительное влияние на зарплату. На графике ниже представлены средние абсолютные значения SHAP для разных фич.



код (вкладка (task 4))

Baseline_4task.ipynb