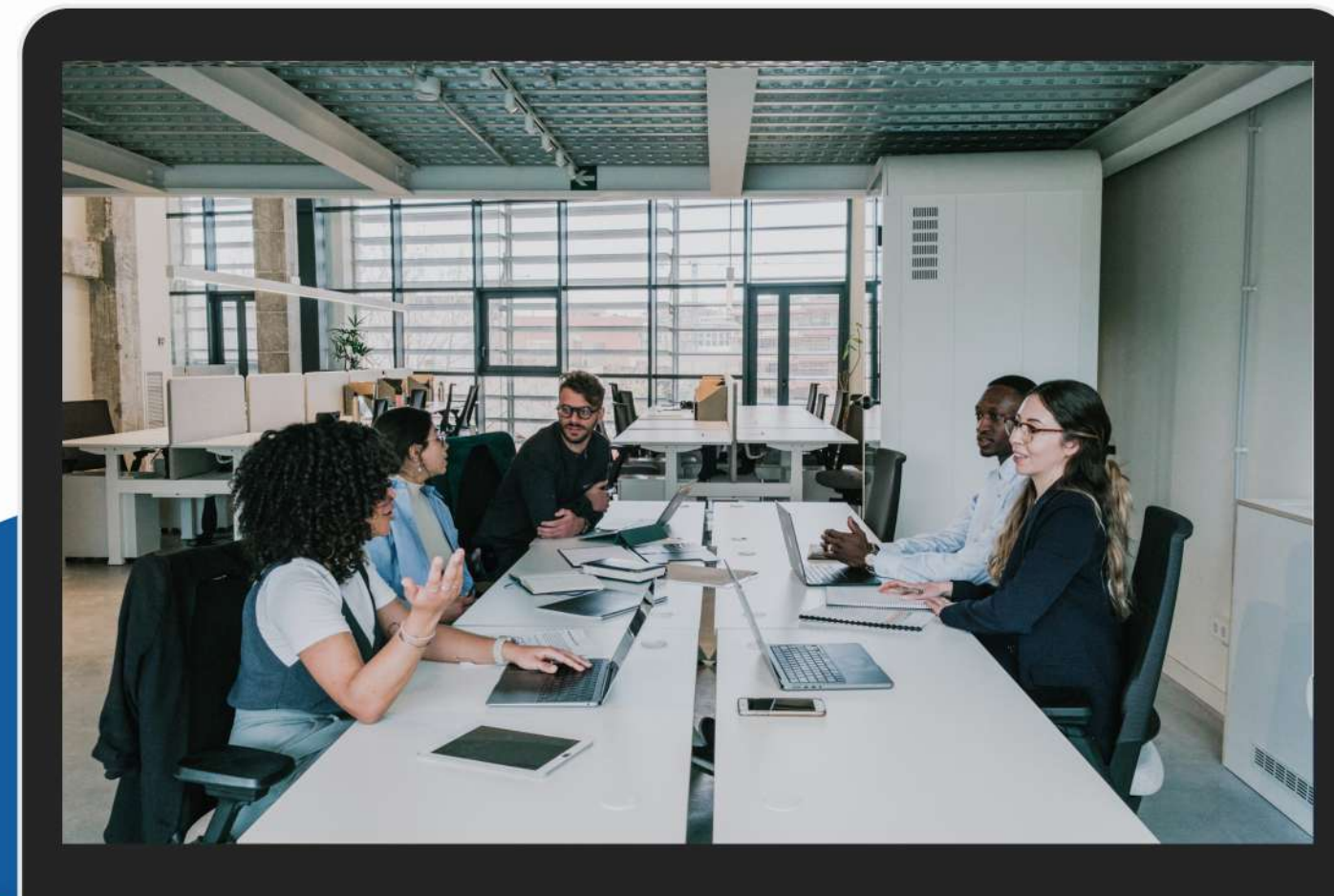




Выпускная Квалификационная Работа

Разработка информационной системы, предсказывающей увольнение

Студент: Галкин Андрей Андреевич



Описание проекта

Заказчик предоставил данные о сотрудниках, включая уровень их удовлетворённости работой, рассчитанный на основе опросов. Сбор таких данных в большой компании трудоёмок, но удовлетворённость может влиять на отток. Цель проекта — разработать модель, предсказывающую увольнение сотрудников на основе всех предоставленных данных.

Задачи проекта

01

Разработка модели, предсказывающей уровень удовлетворённости сотрудников работой.

02

Составление портрета «уволившегося сотрудника».

03

Проверка гипотезы о том, что уровень удовлетворённости сотрудника работой влияет на вероятность его увольнения.

04

Разработка модели, предсказывающей увольнение сотрудников, на основе данных заказчика, и предсказанного уровня удовлетворённости работой.

Пропуски

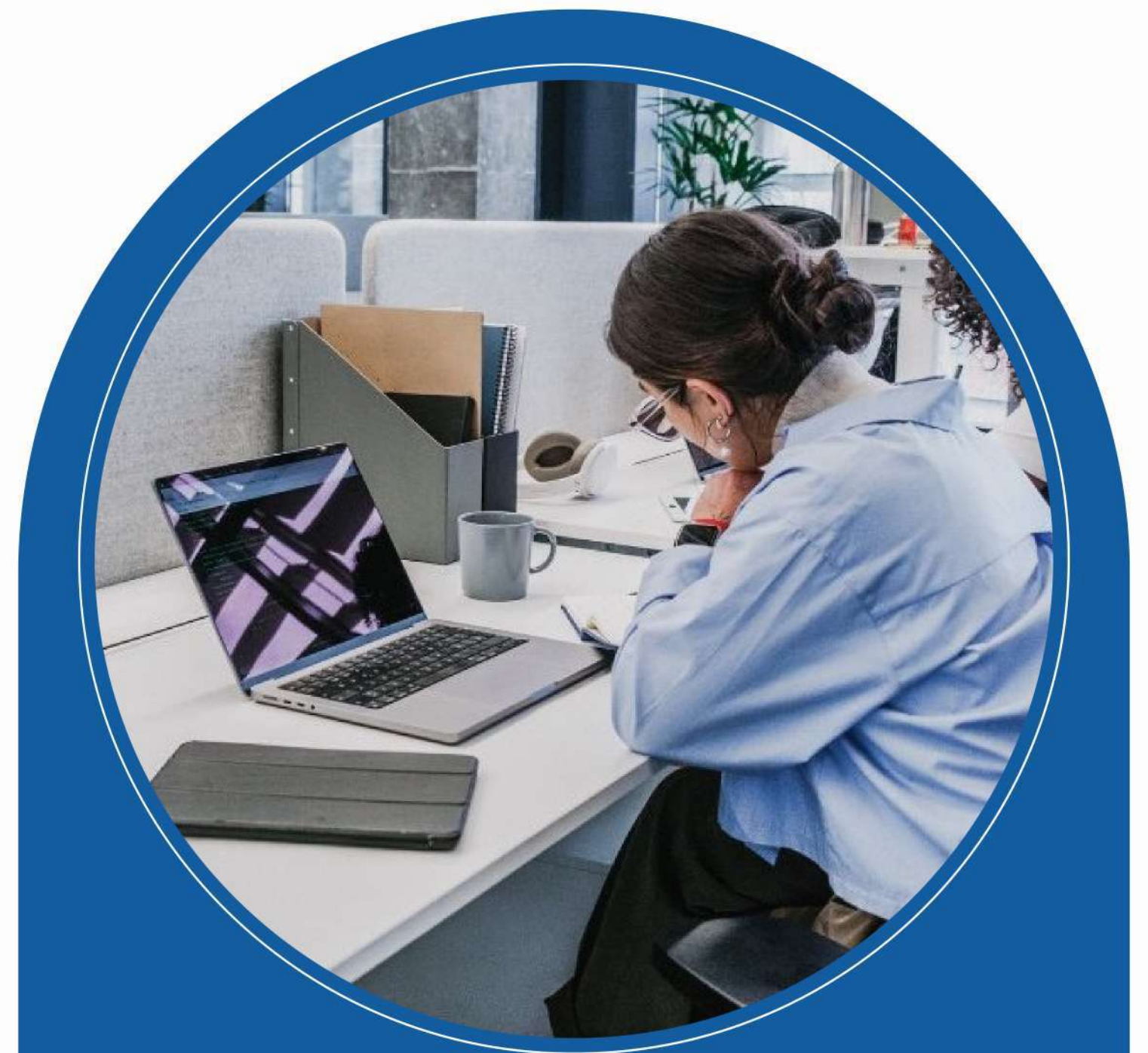
В категориальных входных признаках присутствуют редкие пропущенные значения (менее 1%). Пропуски могут быть как NAN, так и пробелами.

Решение

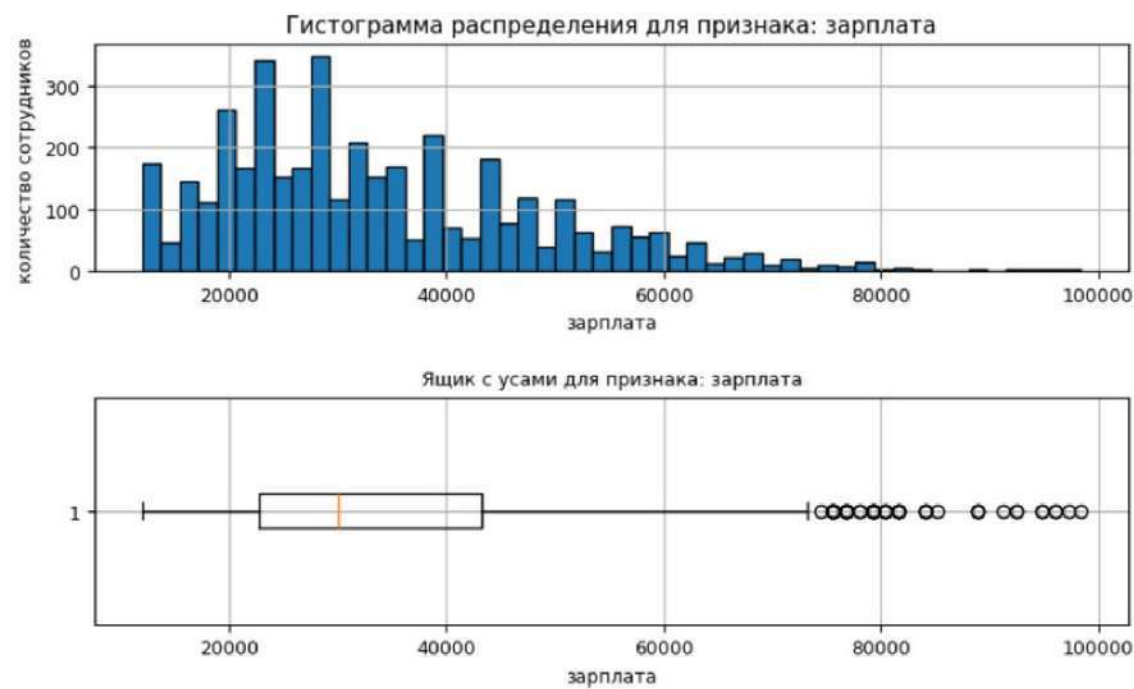
Пропуски также могут встретиться и в продакшн данных, поэтому они будут обработаны внутри пайплайна.

Вначале, с помощью вспомогательной функции, все пропуски будут приведены к единому виду, а потом заполнены с помощью SimpleImputer.

Использовать более сложные алгоритмы заполнения пропусков нецелесообразно, т.к. пропуски встречаются крайне редко.

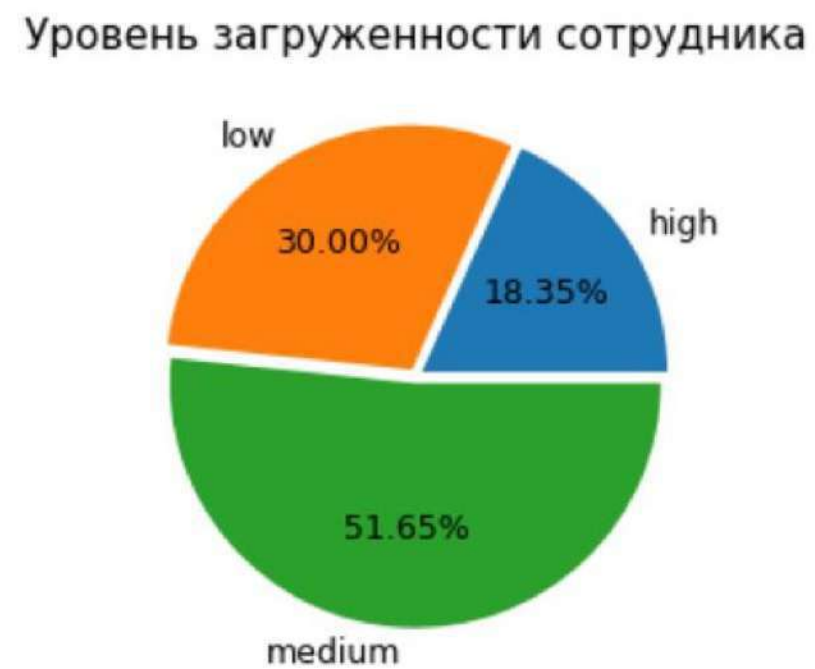


Исследовательский анализ



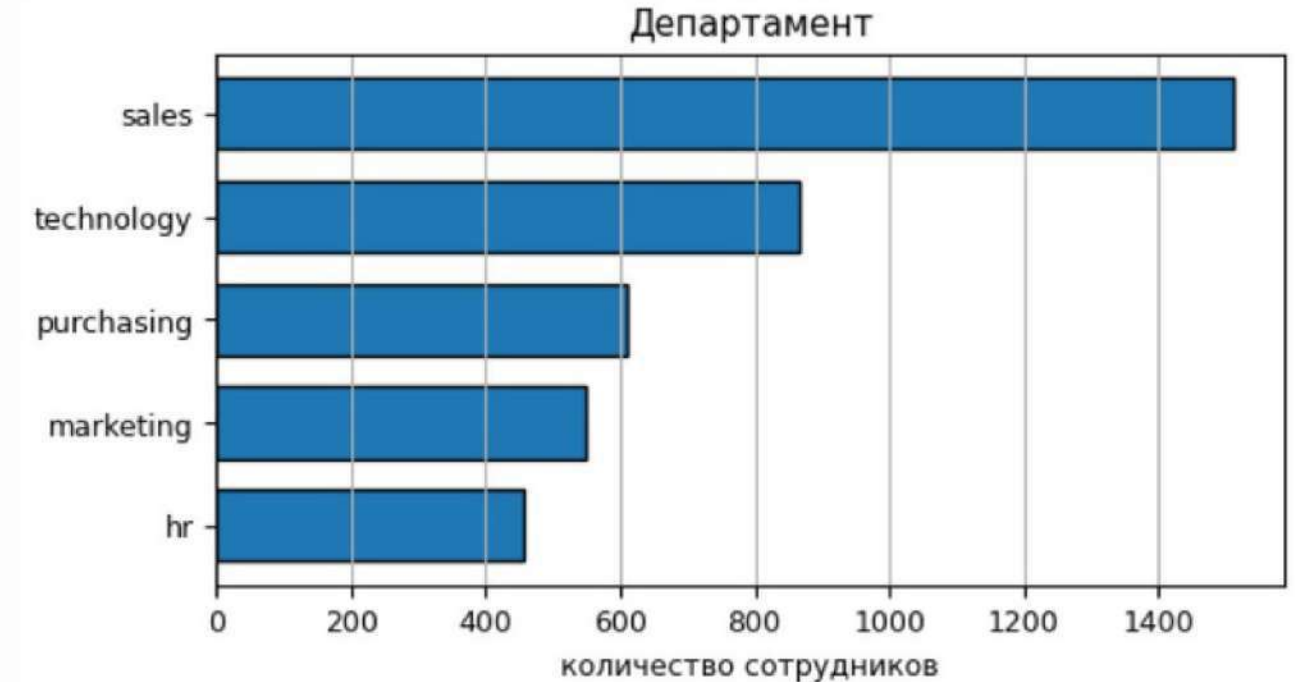
Зарплата сотрудников

Гистограмма распределения и «ящик с усами»



Уровень загрузки

Круговая диаграмма

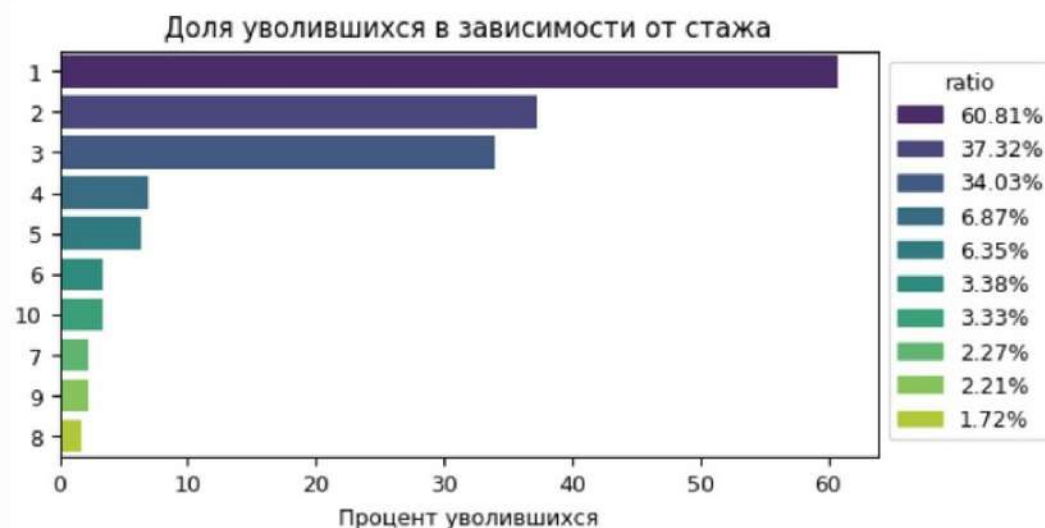


Отдел

Столбчатая диаграмма

Было проанализировано распределение всех признаков.
Скрытых дубликатов в категориальных признаках не выявлено.
Аномальных значений и выбросов не выявлено.

Портрет уволившегося сотрудника



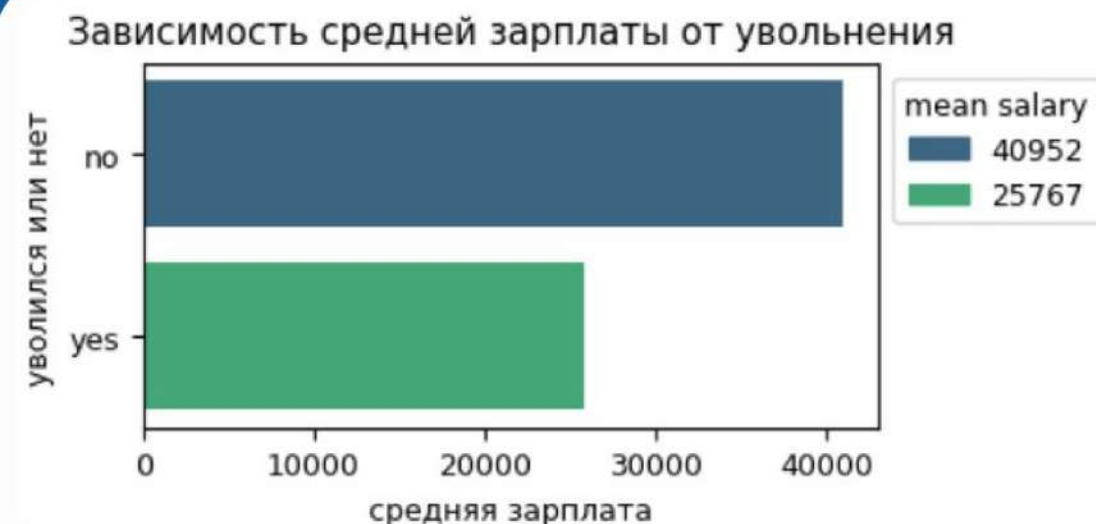
Стаж работы

В основном увольняются сотрудники, работающие менее 3-х лет.



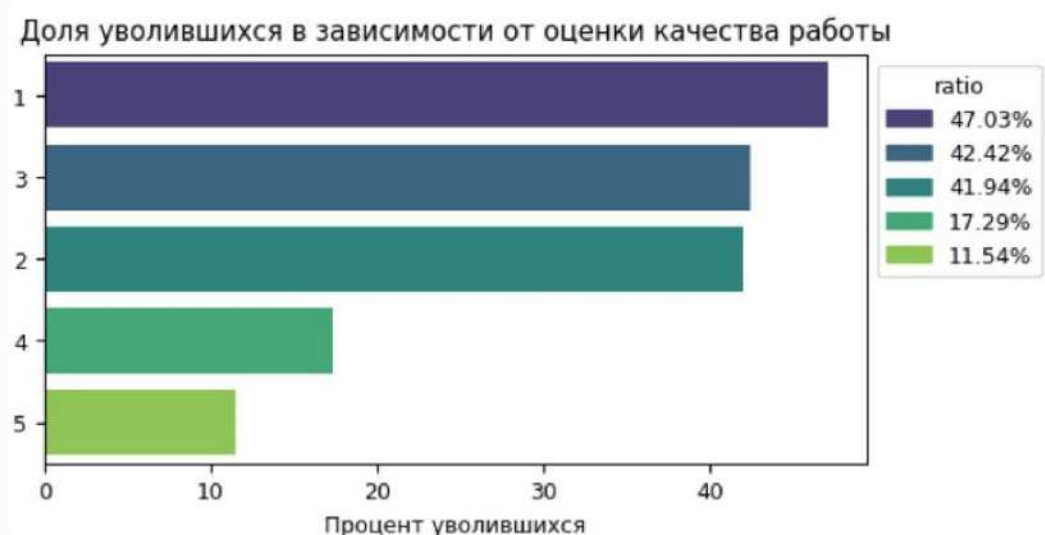
Загрузка работой

Сотрудника с низким уровнем загрузки увольняются чаще.



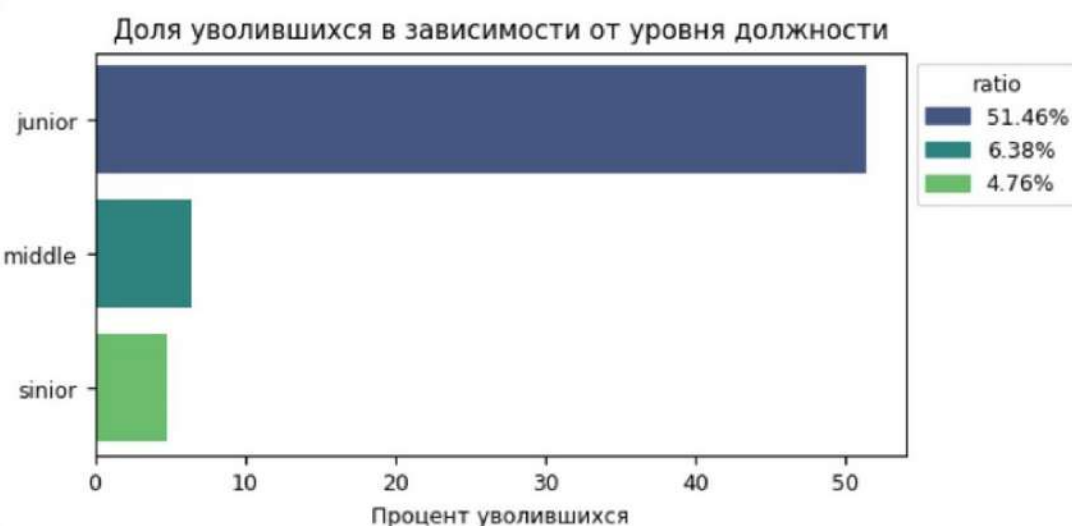
Зарплата

Средняя зарплата у уволившихся сотрудников ниже, чем у неуволившихся.



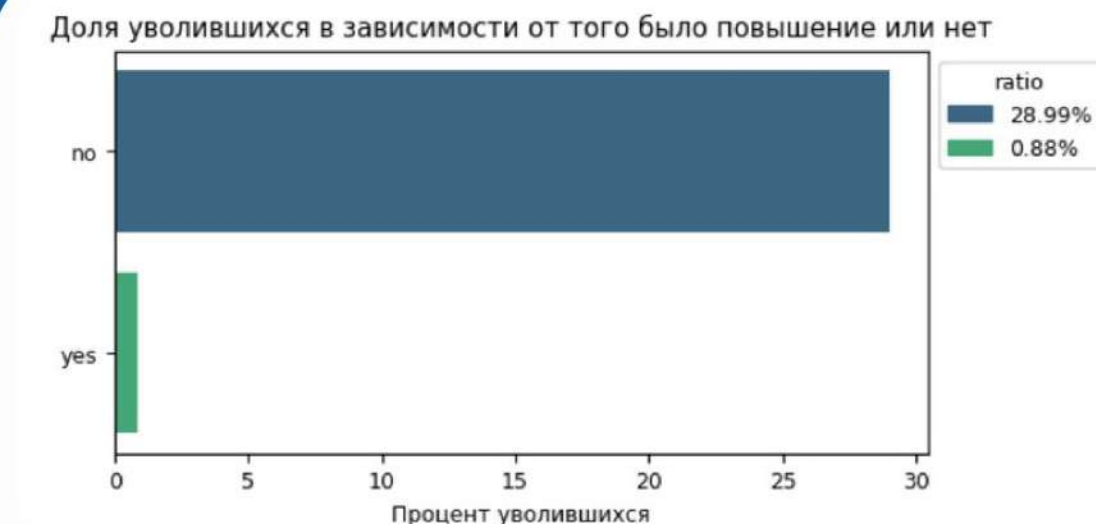
Оценка сотрудника

Чем ниже оценка работы сотрудника, тем более вероятно его увольнение.



Уровень должности

В основном увольняются сотрудники, занимающие должность «junior»

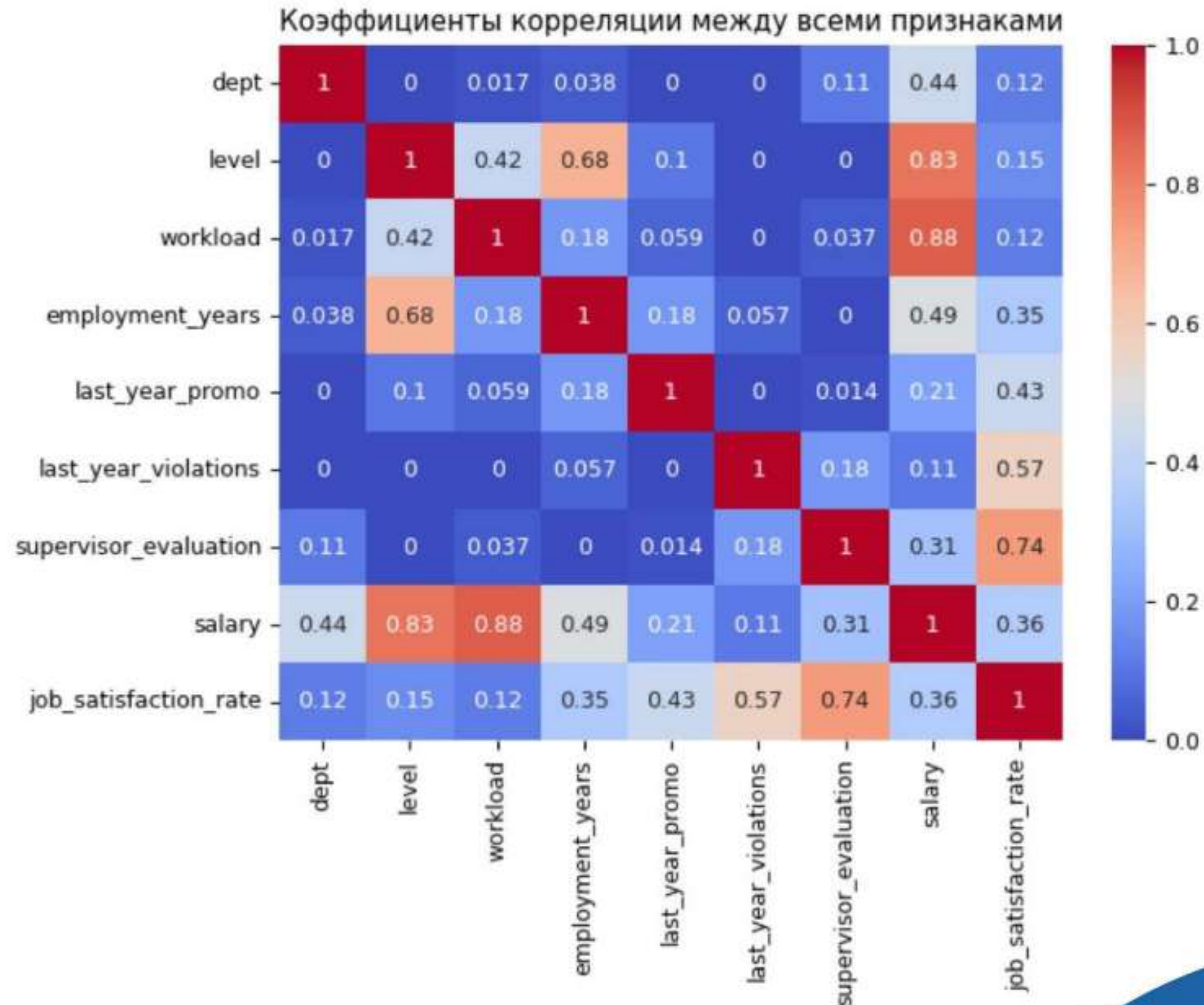


Повышение

Доля уволившихся после повышения очень низкая (менее 1%).

Корреляционный анализ

1. Сильнее всего целевой признак (`job_satisfaction_rate`) коррелирует с оценкой качества работы сотрудника (`supervisor_evaluation`).
2. Меньше всего целевой признак зависит от департамента, уровня должности и загруженности (`dept`, `level`, `workload`).
3. Наблюдается высокая корреляция между следующими входными признаками (возможно, их не стоит использовать для обучения модели, есть смысл протестировать все варианты):
 - salary и workload - 0.88
 - salary и level - 0.83



Тестирование моделей регрессии

Только модели, основанные на деревьях смогли уверенно пройти порог SMAPE < 15 %. Лучший результат показал CatBoostRegressor.

	model	SMAPE_cv	SMAPE_test
1	CatBoostRegressor	11.070393	10.357616
2	RandomForestRegressor	13.427484	12.372859
3	DecisionTreeRegressor	14.840427	13.586431
4	KNNRegressor	16.107809	14.347945
5	LinearRegression	24.945747	23.597329
6	Ridge	25.027972	23.561608

Best trial: 65. Best value: -24.9457: 100% 100/100 [00:41<00:00, 2.36it/s]

LinearRegression	SMAPE	duration	drop_workload_flag	drop_level_flag	num_encoding	cat_encoding	fit_intercept
1	24.945747	00:00:00.427884	False	False	MinMaxScaler	OneHotEncoder	False

Best trial: 93. Best value: -25.028: 100% 100/100 [01:05<00:00, 2.21it/s]

Ridge	SMAPE	duration	drop_workload_flag	drop_level_flag	num_encoding	cat_encoding	alpha	fit_intercept	solver
1	25.027972	00:00:00.399300	False	False	passthrough	OneHotEncoder	1.727598	True	cholesky

Best trial: 51. Best value: -16.1078: 100% 100/100 [01:02<00:00, 1.18it/s]

KNNRegressor	SMAPE	duration	drop_workload_flag	drop_level_flag	num_encoding	cat_encoding	algorithm	n_neighbors	weights
1	16.107809	00:00:00.639448	False	False	RobustScaler	OneHotEncoder	auto	9	distance

Best trial: 97. Best value: -14.8404: 100% 100/100 [01:05<00:00, 2.21it/s]

DecisionTreeRegressor	SMAPE	duration	drop_workload_flag	drop_level_flag	num_encoding	cat_encoding	criterion	max_depth	max_features	min_samples_split
1	14.840427	00:00:00.445303	False	False	RobustScaler	OrdinalEncoder	squared_error	17	11	9

Best trial: 58. Best value: -13.4275: 100% 100/100 [08:50<00:00, 6.98s/it]

RandomForestRegressor	SMAPE	duration	drop_workload_flag	drop_level_flag	num_encoding	cat_encoding	criterion	max_depth	max_features
1	13.427484	00:00:04.517818	False	False	RobustScaler	OneHotEncoder	friedman_mse	16	5

Best trial: 69. Best value: -11.0704: 100% 100/100 [1:59:19<00:00, 123.18s/it]

CatBoostRegressor	SMAPE	duration	drop_workload_flag	drop_level_flag	num_encoding	cat_encoding	bagging_temperature	depth	embedding_dim	l2_leaf_reg	learning_rate	loss_function	random_strength	iterations
1	11.070393	00:01:02.084974	False	False	passthrough	EmbeddingEncoder	0.405326	8	9.0	0.014822	0.039292	RMSE	0.472629	741

3

DecisionTreeRegressor

14.840427

13.586431

4

KNNRegressor

16.107809

14.347945

5

LinearRegression

24.945747

23.597329

6

Ridge

25.027972

23.561608

<

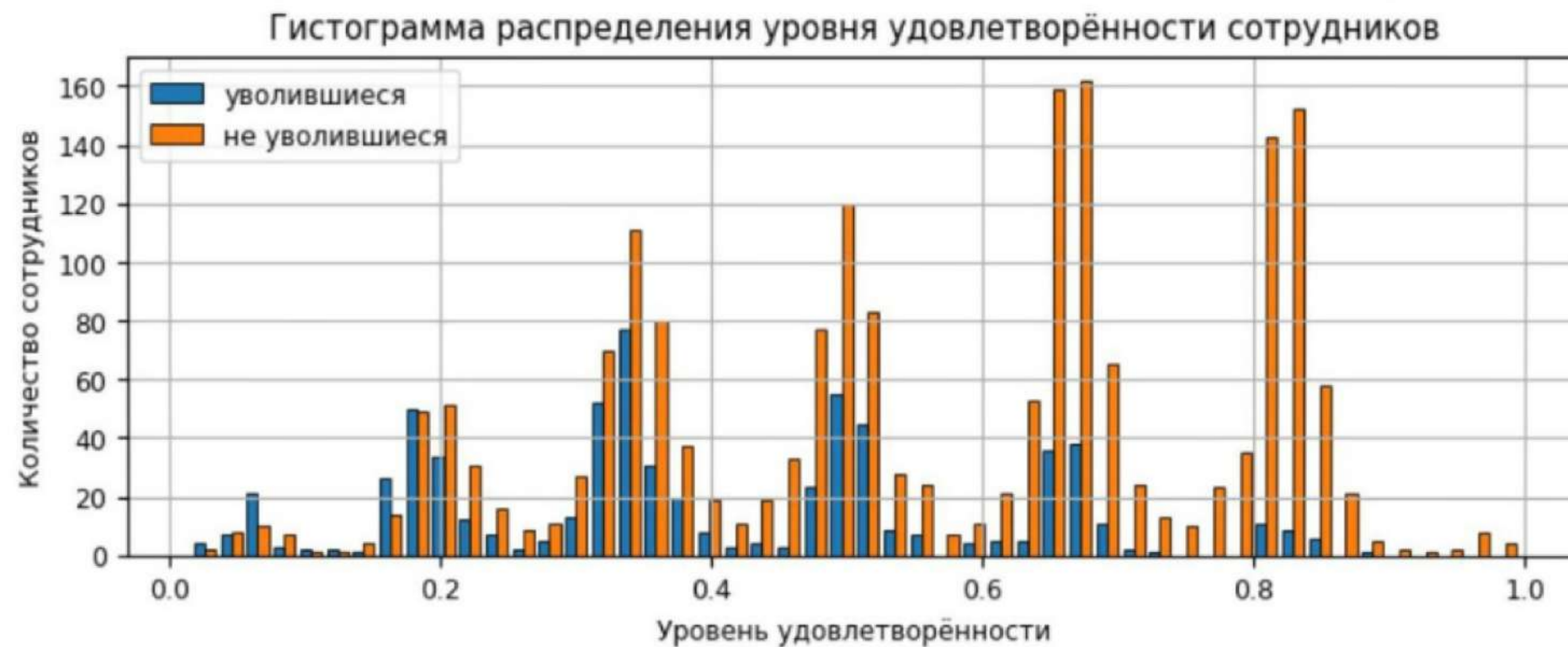
>

Проверка гипотезы

Нулевая гипотеза: У уволившихся и у не уволившихся сотрудников одинаковый уровень удовлетворённости работой.

Альтернативная гипотеза: Не уволившиеся сотрудники удовлетворены работой больше, чем уволившиеся.

Вывод: Гипотеза подтвердилась: Уволившиеся сотрудники в среднем удовлетворены работой меньше, чем не уволившиеся.



p-значение: 2.4192827253155234e-159

Отвергаем нулевую гипотезу

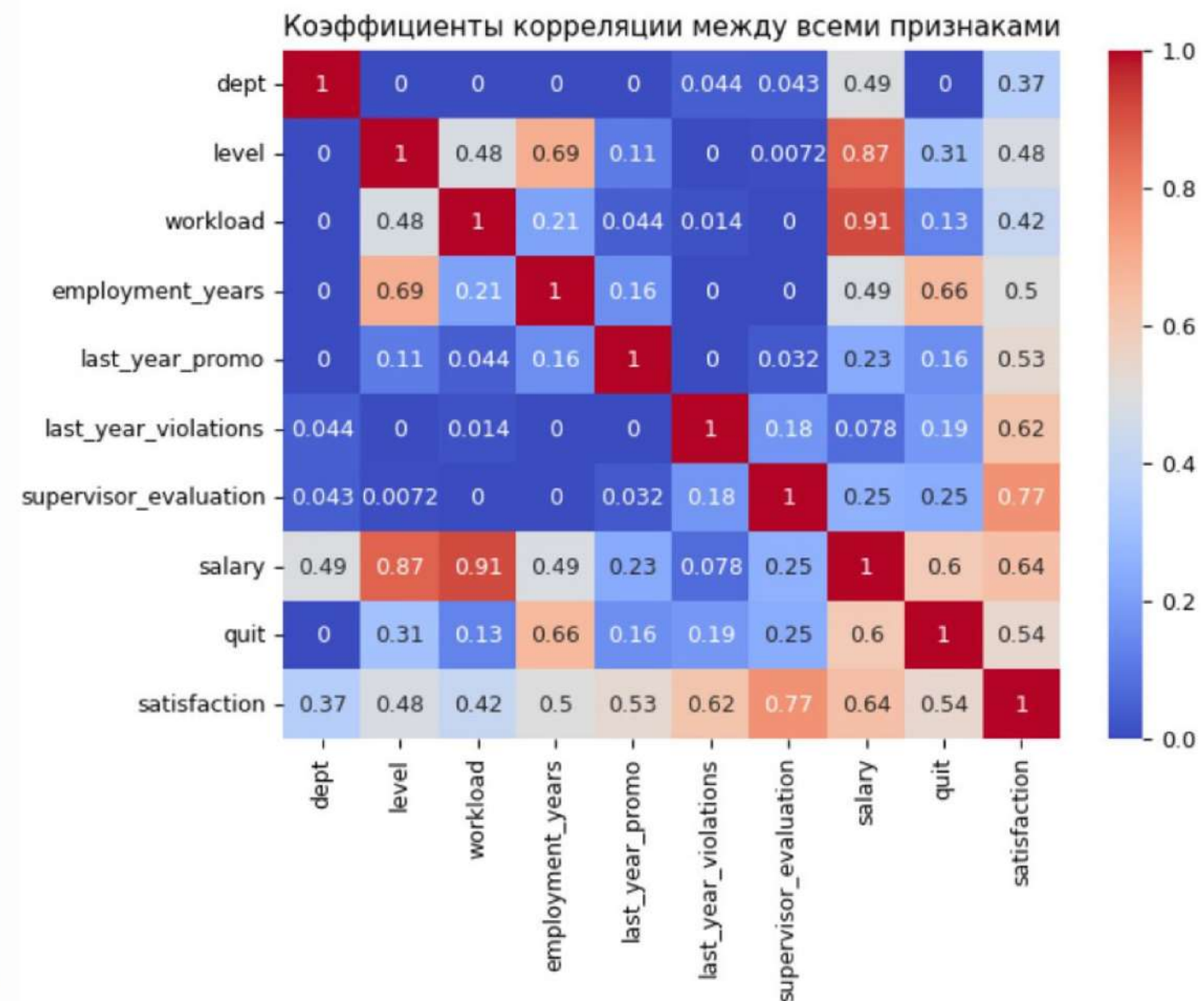
```
# Задаём минимальную вероятность получить такую выборку, при условии, что гипотеза верна
alpha = 0.01
# Передаём в метод имеющиеся выборки и необходимые настройки
results = st.ttest_ind(quit_no['satisfaction'], quit_yes['satisfaction'],
                      equal_var=False,
                      alternative='greater')
# Выводим на экран p-значение
print('p-значение:', results.pvalue)
# Сравниваем p-значение с alpha и делаем вывод
if results.pvalue < alpha:
    print('Отвергаем нулевую гипотезу')
else:
    print('Не отвергаем нулевую гипотезу')
```



Повторный корреляционный анализ

Показатель удовлетворённости работой был предсказан лучшей регрессионной моделью. После этого был проведён повторный корреляционный анализ для изучения зависимости между увольнением сотрудников и входными признаками.

1. Сильнее всего целевой признак (quit) коррелирует с трудовым стажем в компании (employment_years) и зарплатой (salary).
2. Целевой признак не зависит от департамента (dept). Лучше не учитывать его для обучения модели классификации.
3. Наблюдается высокая корреляция между следующими входящими признаками (возможно, их не стоит использовать для обучения модели):
 - salary и workload – 0.91
 - salary и level – 0.87
4. Целевой признак (quit) коррелирует с предсказанным признаком (satisfaction), коэффициент корреляции 0.55, поэтому стоит использовать его для обучения модели.



Тестирование моделей классификации

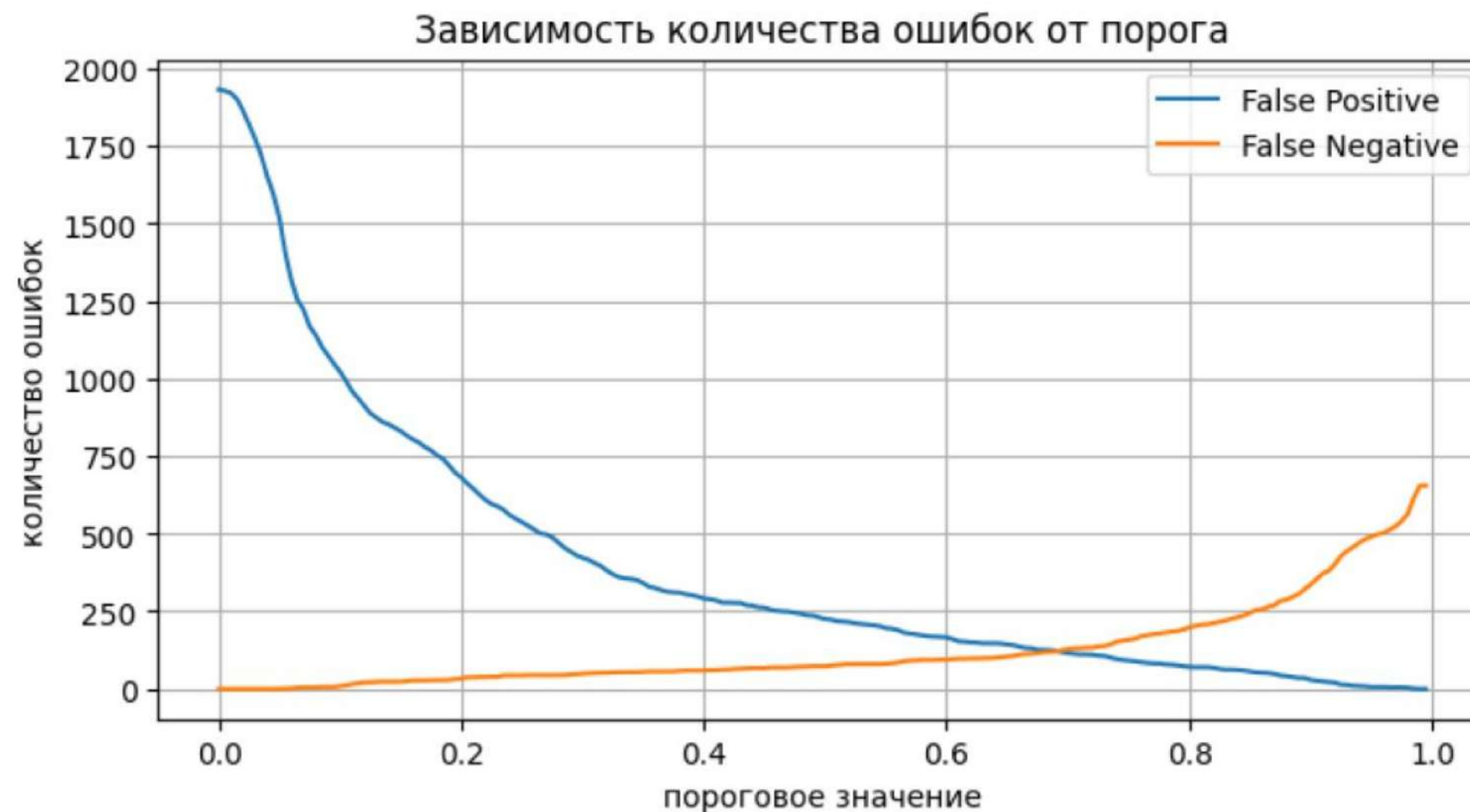
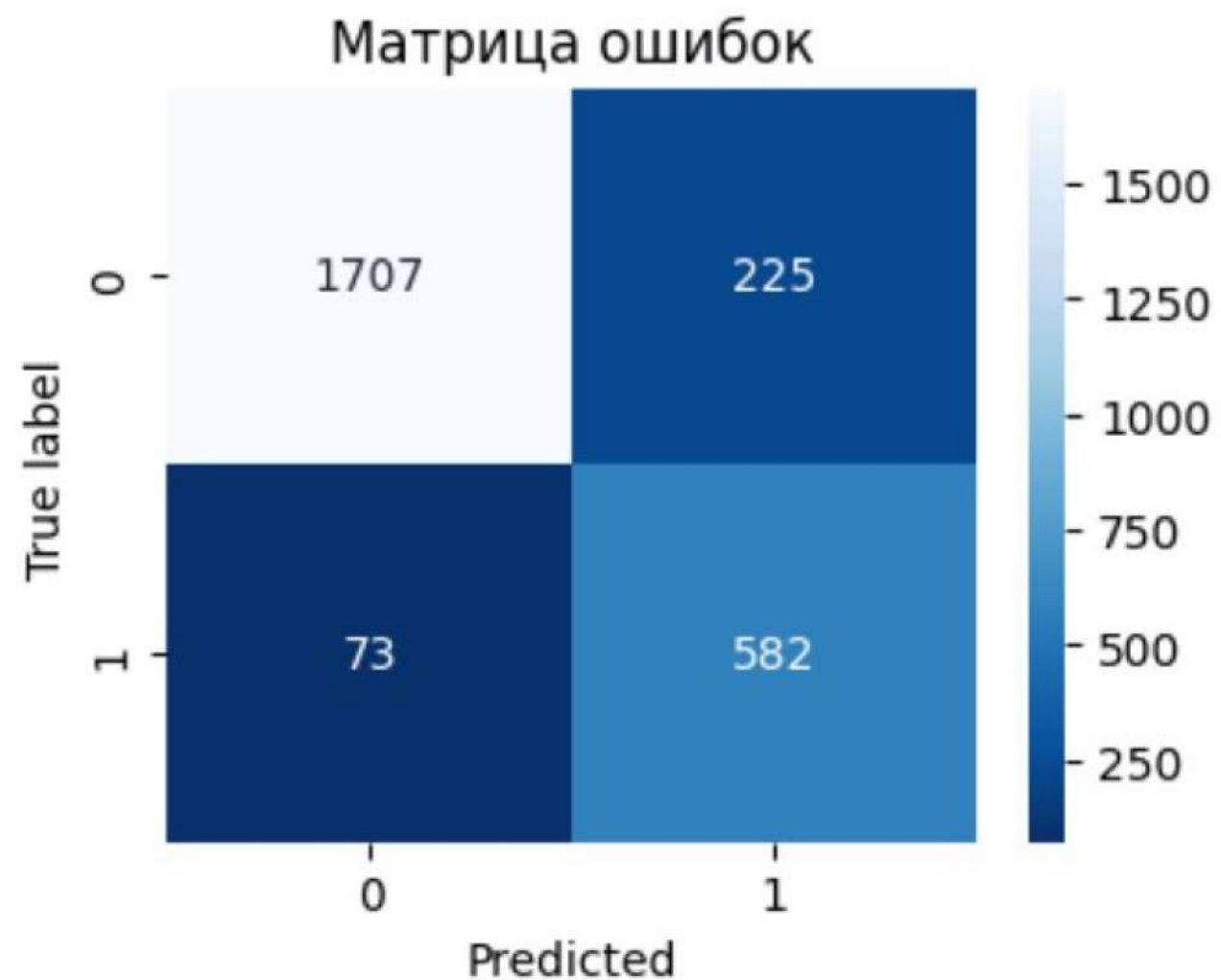
Все модели прошли порог на
тестовой выборке ROC-AUC > 91.
Лучший результат показал
CatboostClassifier.

Слайд 10/14

model	roc_auc_cv	roc_auc_test
CatBoostClassifier	0.926640	0.930313
RandomForestClassifier	0.920908	0.927554
KNNClassifier	0.914381	0.916734
SVC	0.913683	0.915791
LogisticRegression	0.911645	0.919140
DecisionTreeClassifier	0.906847	0.925381

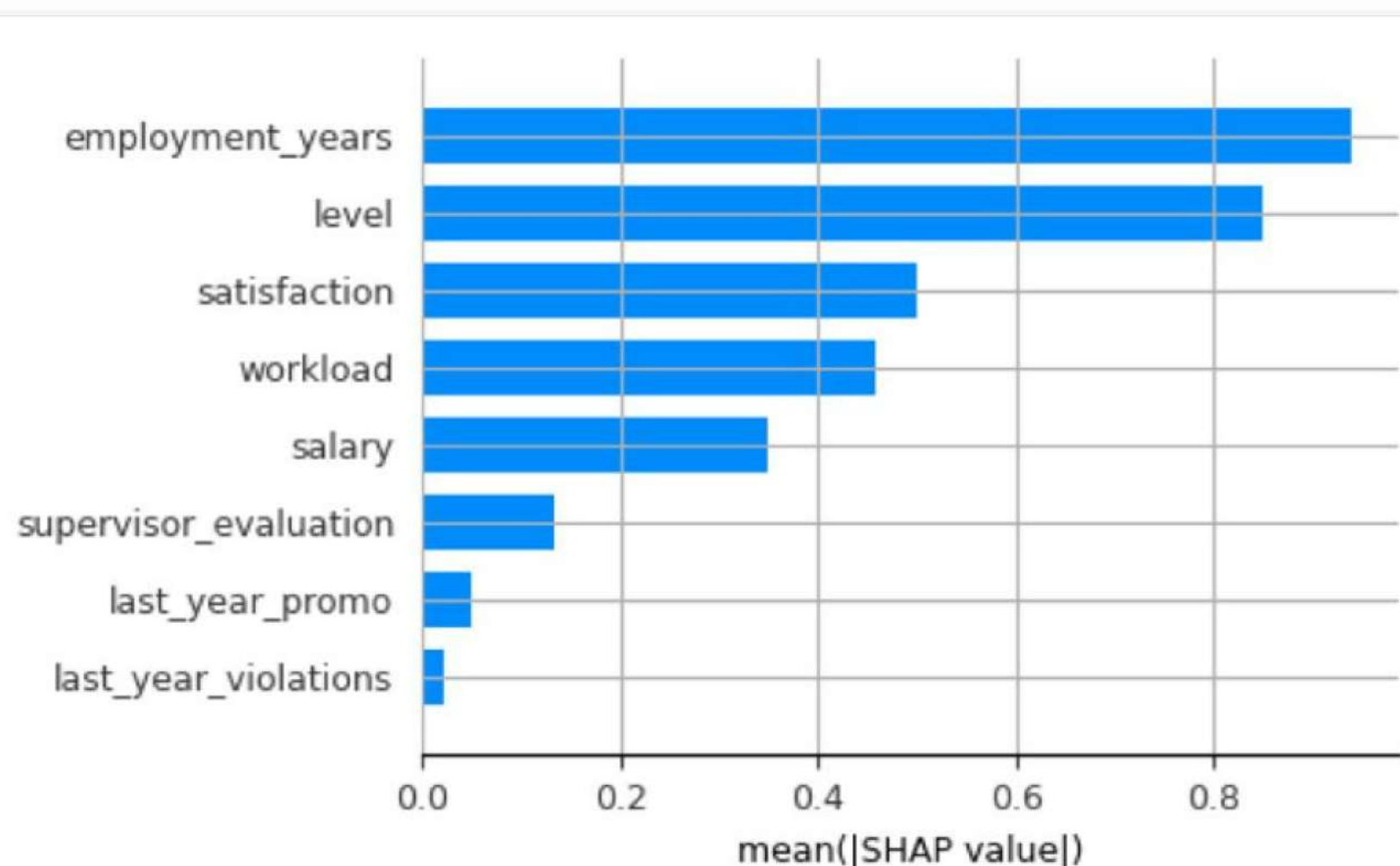


Установка порогового значения

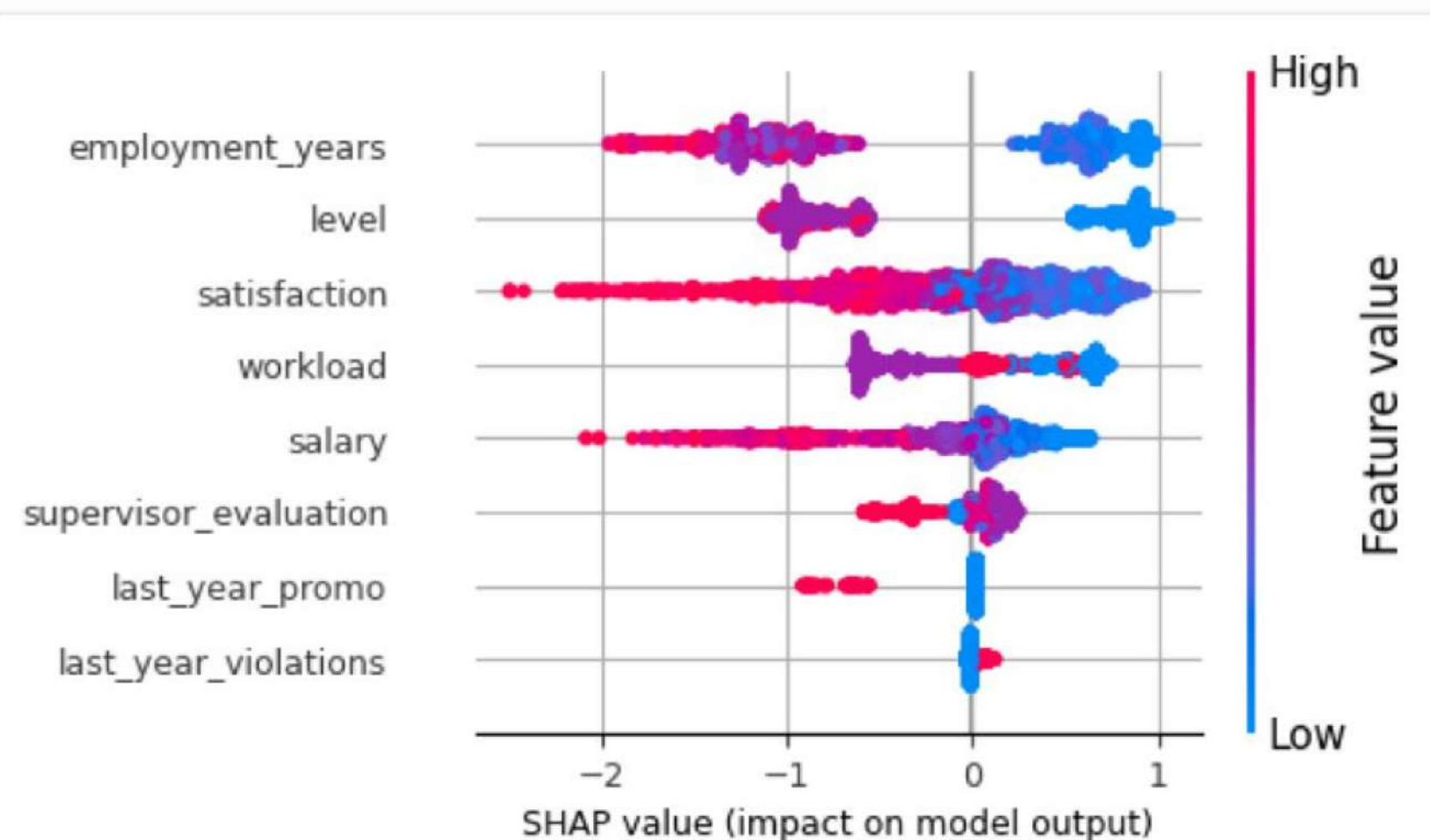


В зависимости от того, насколько для компании важно вовремя заметить, что сотрудник собирается увольняться, и какие затраты она понесёт на его удержание, можно выставить оптимальное пороговое значение.

Интерпретация лучшей модели



Важность входных признаков



Влияние входных признаков

Данные графики соответствуют портрету уволившегося сотрудника, полученному во время исследовательского анализа.

Расчёт вероятности увольнения сотрудника

Отдел
sales

Уровень занимаемой должности
junior

Уровень загрузки работой
low

Нарушения трудового договора за последний год
no

Повышение за последний год
no

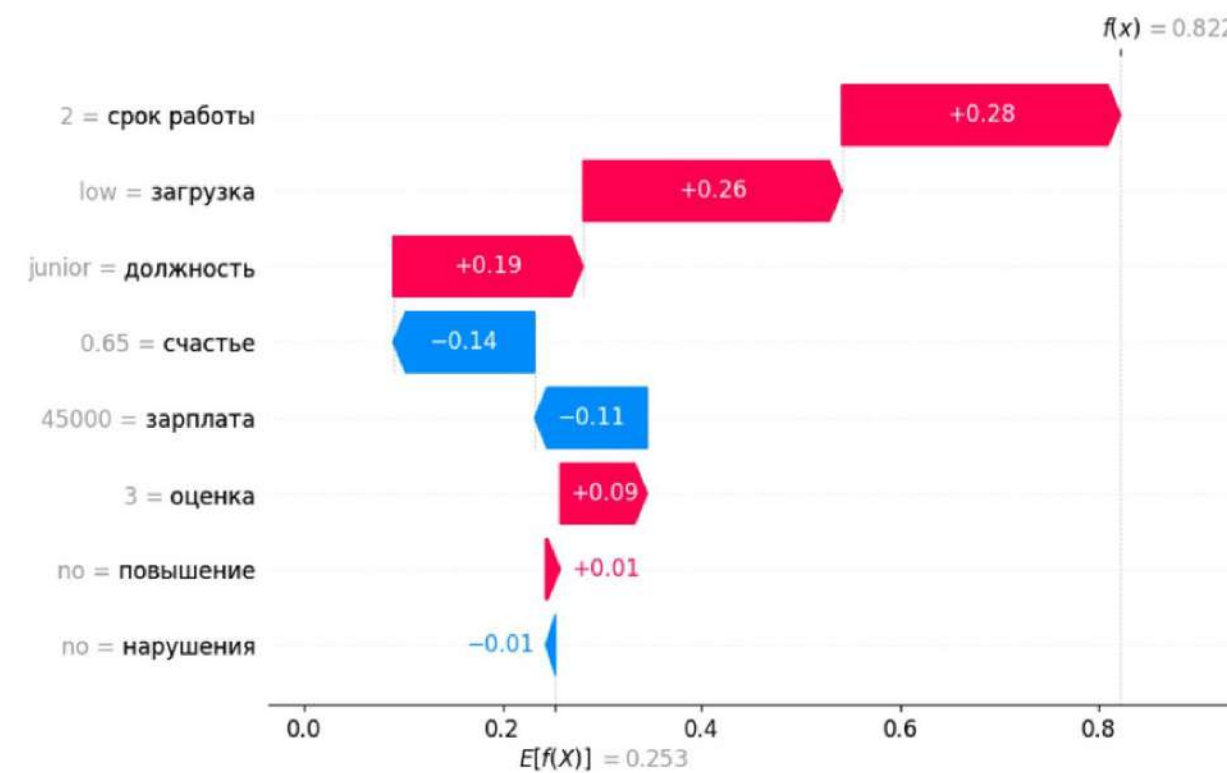
Зарплата
45000

Срок работы в компании
2

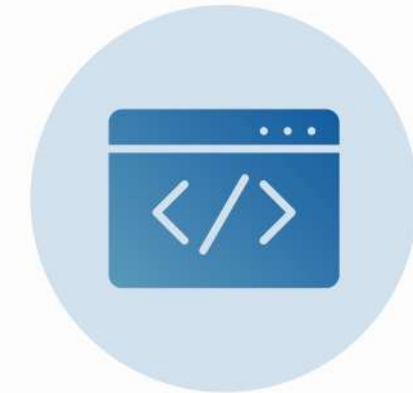
Оценка работы сотрудника
3

Удовлетворённость работой
0.65

Рассчитать вероятность



Вероятность увольнения: 0.82



Интерфейс

Для демонстрации модели заказчику был разработан интерфейс, позволяющий получать предсказания и просматривать график «водопада» для каждого сотрудника. Это наглядно показывает, почему модель сделала определенное предсказание.




Санкт-Петербургский
Государственный
Лесотехнический
Университет
им. С.М. Кирова

Спасибо за внимание!



Галкин Андрей Андреевич

 +7(951)-672-20-21

 taxi-ehe@inbox.ru

 <https://github.com/nightcarpenter>