ДИПЛОМНАЯ РАБОТА ПО ТЕМЕ: «АНАЛИЗ СОТРУДНИКОВ КОМПАНИИ И ФАКТОРОВ УВОЛЬНЕНИЙ (ПОИСК ИНСАЙТОВ, СОСТАВЛЕНИЕ РЕКОМЕНДАЦИЙ СТЕЙКХОЛДЕРАМ)»

Лаврушева Елена Викторовна

Профессия: «аналитик данных», DA-111

г. Москва, 2025

Анализ сотрудников компании и факторов увольнений является критически важным для успешного функционирования и развития организации

Он поможет компании понять причины увольнений и разработать стратегии для улучшения условий работы и удержания сотрудников

Позволит компании не только снизить финансовые потери, но и повысить производительность, удержать таланты и улучшить корпоративную культуру

Это ключевой инструмент для создания устойчивой и конкурентоспособной организации

Цель проекта:

проведение анализа данных сотрудников с целью оптимизации управления персоналом и бизнес-процессами компании

Бизнес-задачи:

1. Проведение анализа данных сотрудников и определение ключевых факторов, влияющих на увольнения
2. Разработка рекомендаций по улучшению кадровой политики компании с учетом выявленных рисков
3. Построение модели, предсказывающей, уволится ли сотрудник в ближайшее время

Стейкхолдеры:

1. Руководство компании (Топ-менеджмент)
2. Менеджеры среднего звена
3. Служба по управлению персоналом (HR-отдел)
4. Сотрудники компании
5. Внешние стейкхолдеры

Описание данных

Исследование проведено на данных датасета «IBM HR Analytics Employee Attrition & Performance», содержащих информацию о сотрудниках компании

- **Возраст работников**: средний – составляет 37 лет, минимум – 18 и максимум – 60
- **Расстояние от дома до работы** в среднем 9,19 миль, при этом некоторые сотрудники живут всего в 1 миле, а другие в 29 милях от своего дома
- Средняя **дневная ставка** – 802,49 с широким диапазоном от 102,00 до 1499,00.
- Средний **ежемесячный доход** составляет примерно 6502,93, минимально – 1009,00, максимально – 19999,00
- В среднем сотрудники имеют около 11,28 лет **общего рабочего стажа**, некоторые только начали, а другие имеют до 40 лет опыта
- **Сроки работы в компании**: в среднем 7 лет (от 3-х до 9-и)

Проверка на	Результат
соответствие типов признаков их смысловому содержанию	соответствуют, однако целесообразно некоторым категориальным параметрам изменить тип данных с целью улучшения совместимости
наличие пропущенных значений	пропуски значений отсутствуют
значение со знаком «-»	отрицательные значения отсутствуют
уникальность строк и наличие дубликатов	все строки уникальны, дубли отсутствуют
названий столбцов	некорректное название столбца «п»їAge»
выбросы в параметрах	наличие выбросов обнаружено в: MonthlyIncome, NumCompaniesWorked, PerformanceRating, StockOptionLevel, TotalWorkingYears, TrainingTimesLastYear, YearsAtCompany, YearsInCurrentRole, YearsSinceLastPromotion, YearsWithCurrManager
неинформативные признаки (более 95% строк признака содержат одно и то же значение)	- EmployeeCount - Over18 - StandardHours
нерелевантные признаки	- EmployeeNumber
возможные опечатки	опечатки символов в признаках имеющих тип данных текстовый или смешанный числовой и нечисловой не обнаружены

Преобразования и очистки данных

- Столбцам Attrition, OverTime изменён тип данных на числовой
- Оставлены только параметры, имеющие значение для анализа
- Столбцы расставлены в удобном для анализа порядке
- Выбросы в параметрах оставлены без изменений, т.к. не оказывают большого влияния. Однако данные приняты во внимание
- Тип данных в остальных категориальных столбцах остаются неизменны

Алгоритмы и техники, применяемые для решения задачи

Сравнение статистических показателей уволенных и работающих сотрудников

Уволенные сотрудники

	Attrition	JobLevel	Age	Education	TrainingTimesLastYear	DistanceFromHome	HourlyRate
count	237.00	237.00	237.00	237.00	237.00	237.00	237.00
mean	1.00	1.64	33.61	2.84	2.62	10.63	65.57
std	0.00	0.94	9.69	1.01	1.25	8.45	20.10
min	1.00	1.00	18.00	1.00	0.00	1.00	31.00
25%	1.00	1.00	28.00	2.00	2.00	3.00	50.00
50%	1.00	1.00	32.00	3.00	2.00	9.00	66.00
75%	1.00	2.00	39.00	4.00	3.00	17.00	84.00
max	1.00	5.00	58.00	5.00	6.00	29.00	100.00

...

Работающие сотрудники

	Attrition	JobLevel	Age	Education	TrainingTimesLastYear	DistanceFromHome	HourlyRate
count	1233.00	1233.00	1233.00	1233.00	1233.00	1233.00	1233.00
mean	0.00	2.15	37.56	2.93	2.83	8.92	65.95
std	0.00	1.12	8.89	1.03	1.29	8.01	20.38
min	0.00	1.00	18.00	1.00	0.00	1.00	30.00
25%	0.00	1.00	31.00	2.00	2.00	2.00	48.00
50%	0.00	2.00	36.00	3.00	3.00	7.00	66.00
75%	0.00	3.00	43.00	4.00	3.00	13.00	83.00
max	0.00	5.00	60.00	5.00	6.00	29.00	100.00

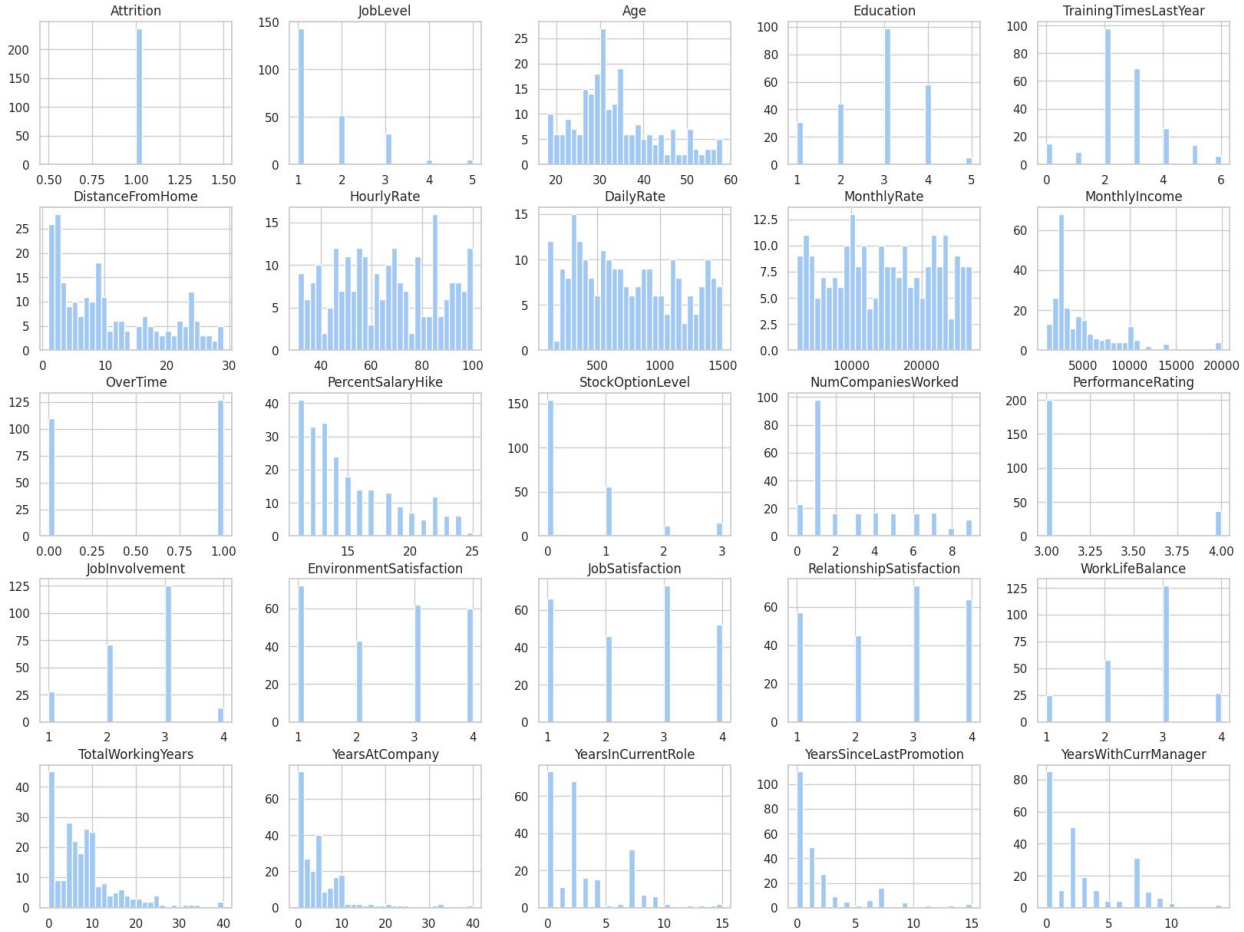
...

```
Всего = 1470
Количество уволенных сотрудников = 237
Процент уволенных сотрудников = 16.122448979591837 %
Количество работающих сотрудников = 1233
Процент работающих сотрудников = 83.87755102040816 %
```

Уволенные сотрудники	Работающие сотрудники
Более молодые, занимали более низкие должности	Более опытные, занимают выше должности
Менее удовлетворены работой и условиями	Больше удовлетворены работой
Чаще работали сверхурочно, но получали меньше	Реже перерабатывают, но имеют более высокий доход
Жили дальше от работы	Чаще получают бонусы и обучение
Имели меньший стаж и реже получали повышения	

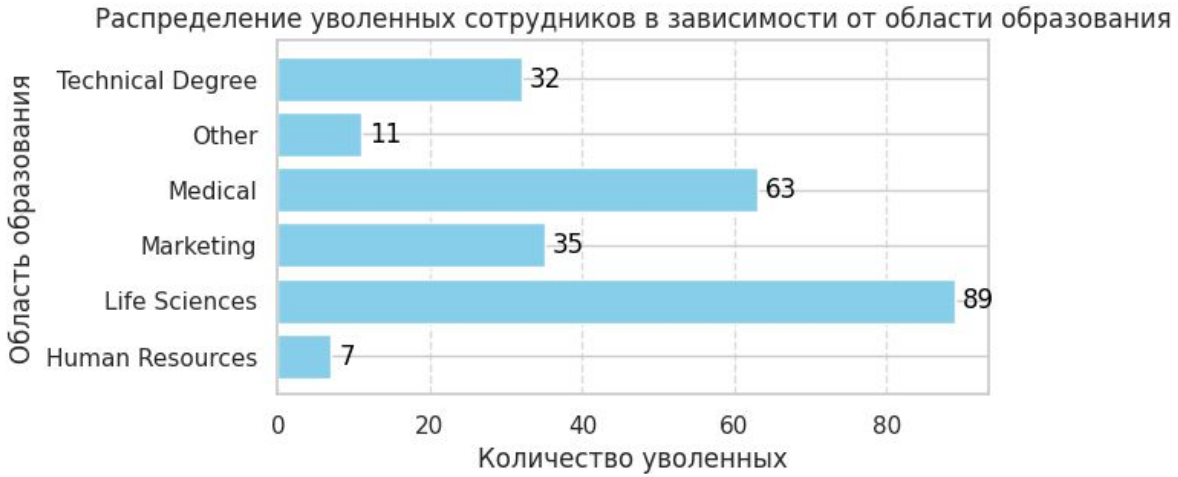
Сегментированный анализ уволенных сотрудников по каждому параметру позволяет выявить факторы наиболее сильно влияющие на увольнение

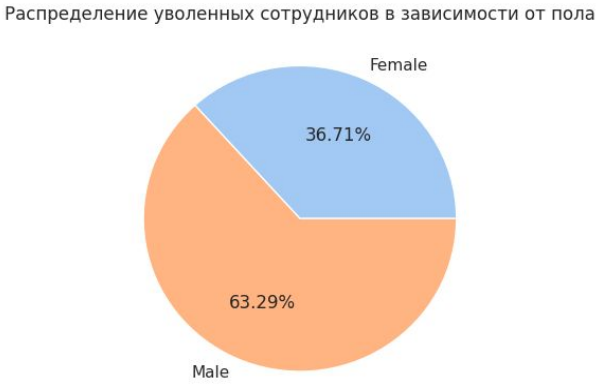
По числовым параметрам



Основные зоны риска: младшие должности, молодые сотрудники (25–35 лет), средний уровень обучения, отсутствие бонусов / повышений, небольшой стаж в компании

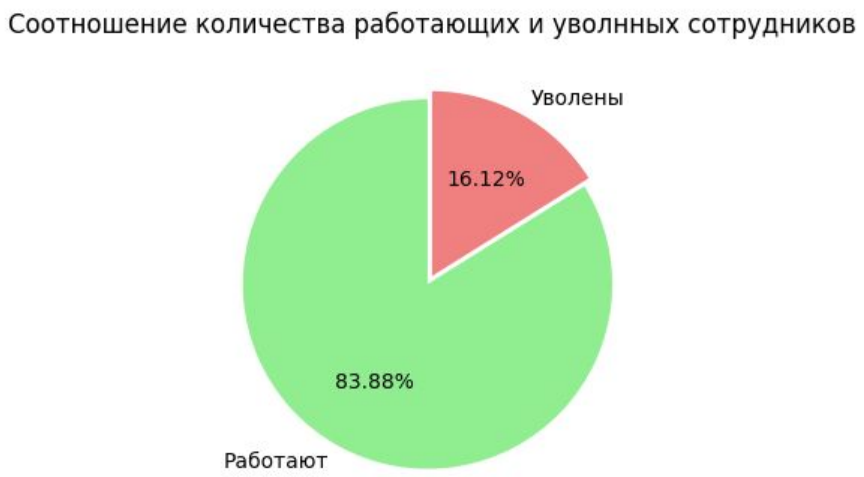
По категориальным параметрам

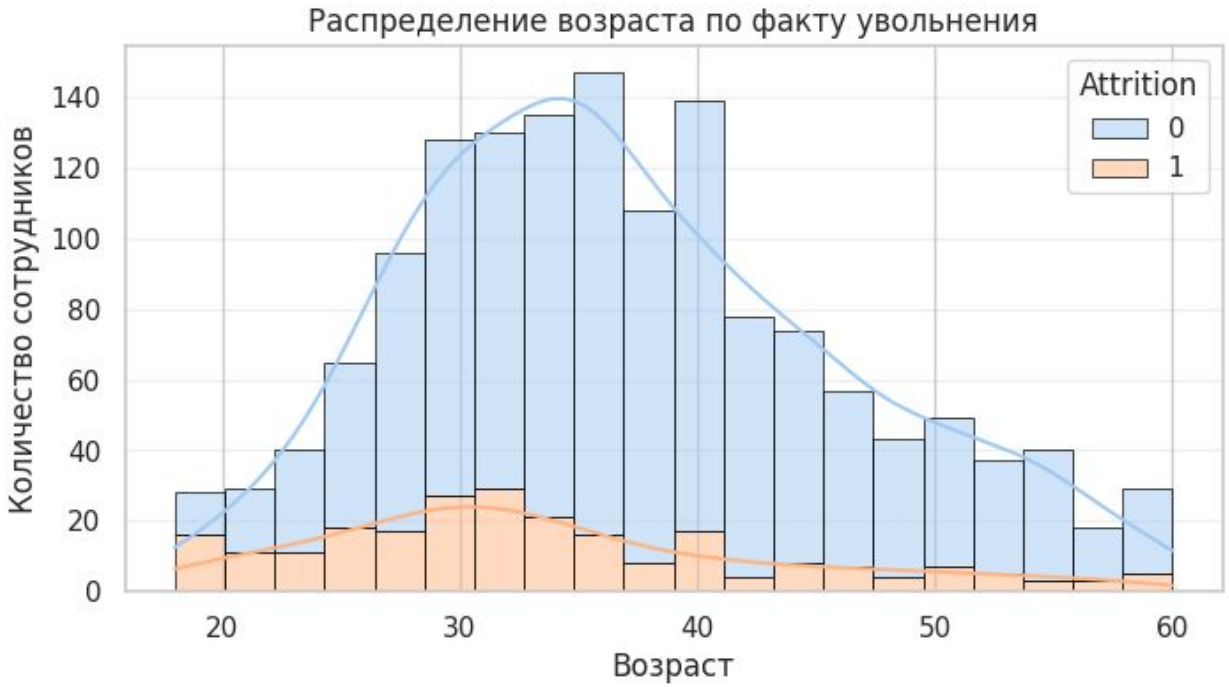




Основные зоны риска: текучесть кадров сильнее всего среди людей имеющих образование «Наука о жизни», в отделе R&D и продажах, среди техников и менеджеров, чаще у мужчин и холостых.

Сравнительный анализ между уволенными и работающими сотрудниками наглядно показывает соотношение этих групп



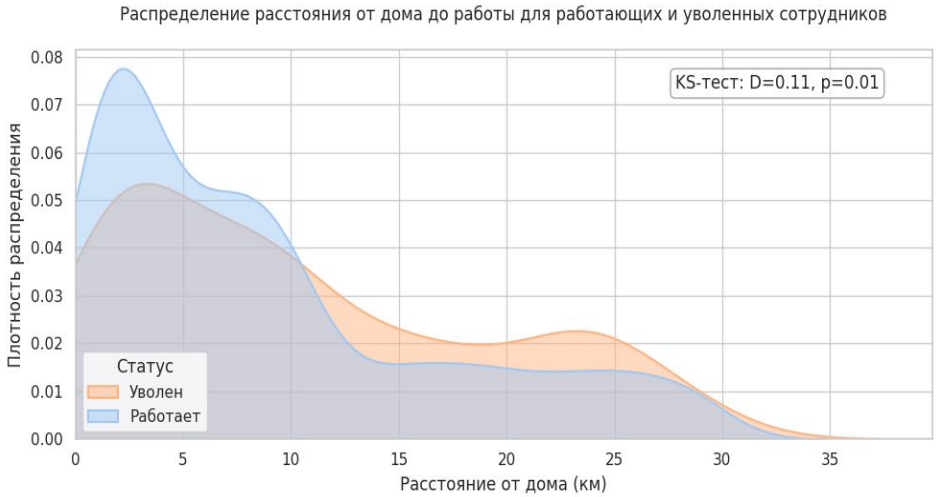


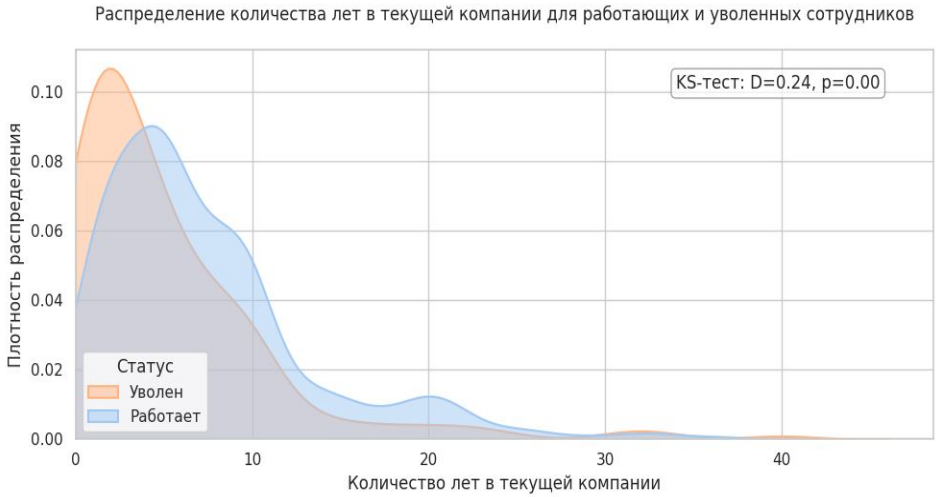
Коэффициент текучести кадров = 16,12

В сфере торгово-промышленного производства коэффициент нормы текучести кадров в пределах нормы. Что является показателем стабильного функционирования организации

Наибольшая концентрация сотрудников наблюдается в возрасте около 30-40 лет. Это типично для большинства компаний

Анализ с помощью KDE (Kernel Density Estimate) позволил провести проверку гипотез и оценить значимость влияния различных факторов на увольнение

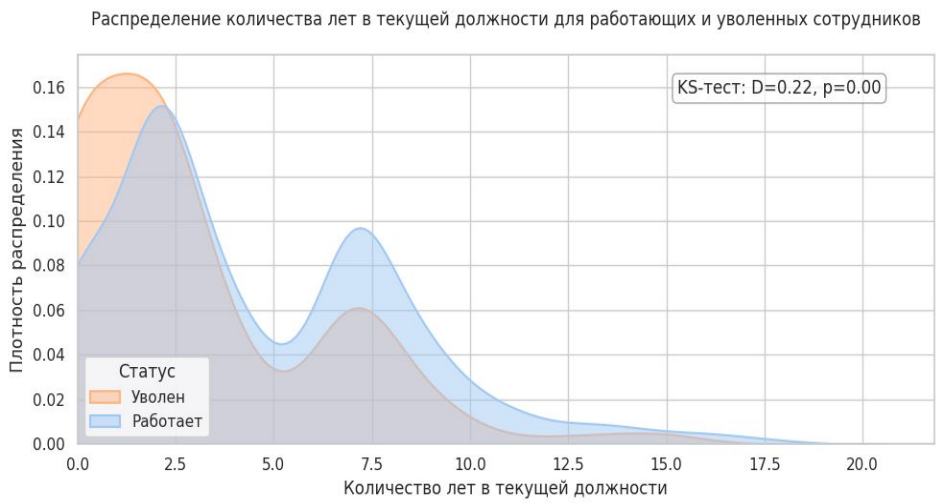


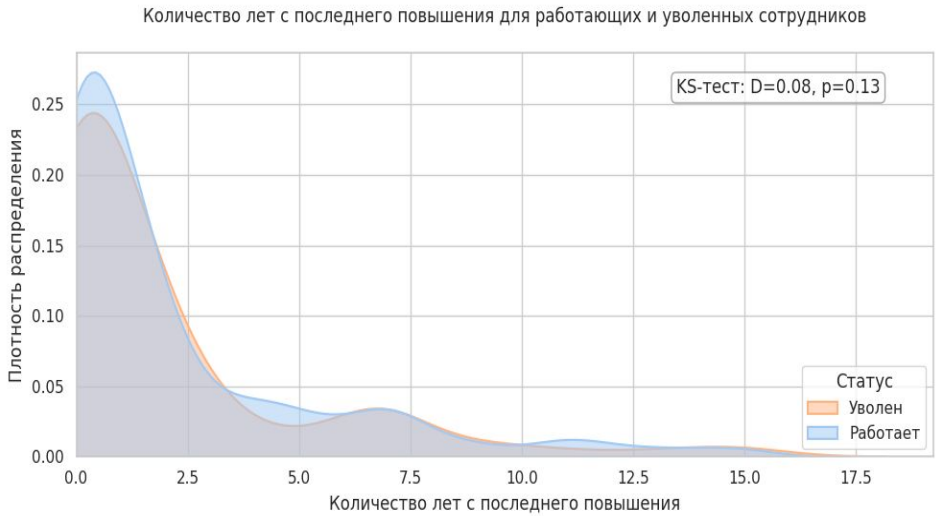


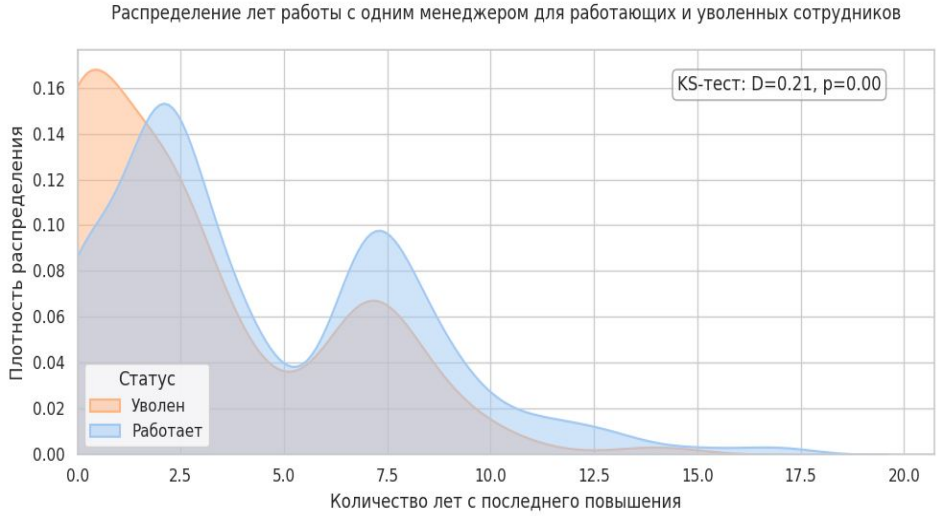
Вывод: гипотезы о значимости влияния на увольнение расстояния от дома до работы, общего стажа, количества лет в текущей компании и должности, а также работы с одним менеджером подтвердились

Это действительно значимые параметры

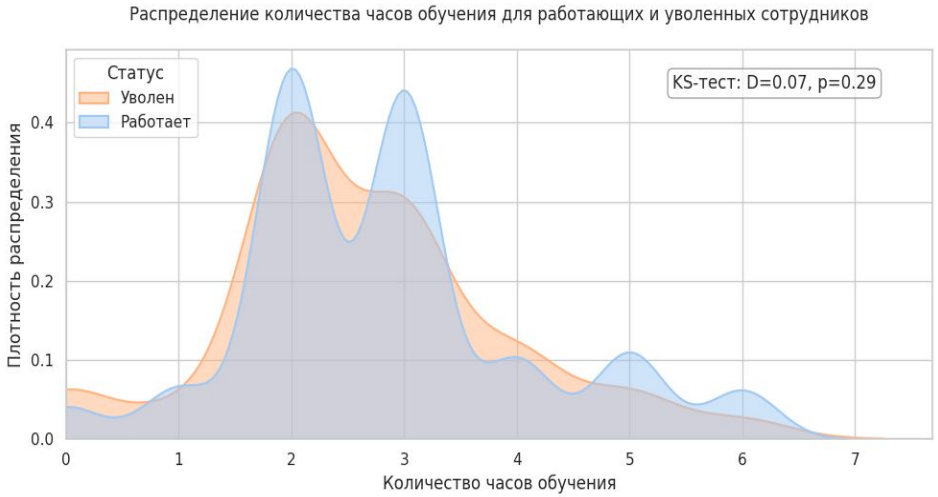
Значимые параметры:

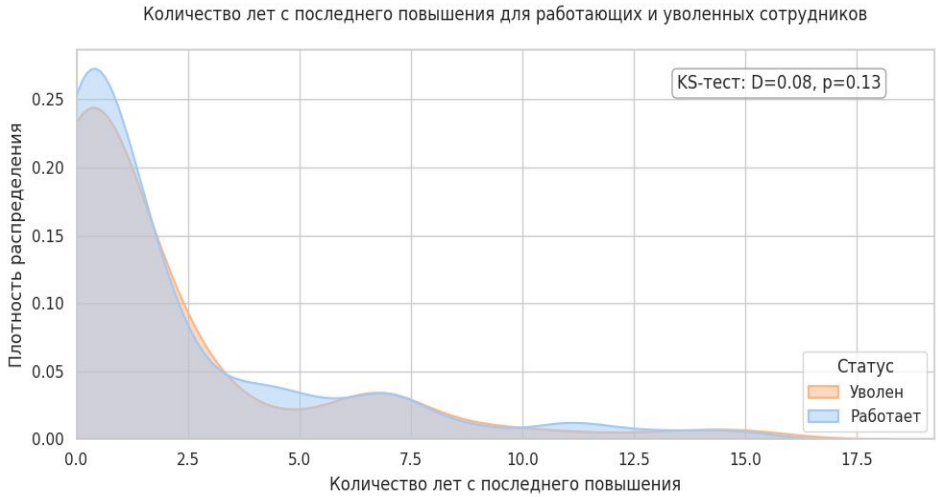






Незначимые параметры:





Вывод: в отношении часов обучения и количества лет с последнего повышения анализ показал, что распределения не различаются статистически значимо, а это значит, что данные параметры не имеют особого влияния на увольнения

Анализ сотрудников в процентном и долевом отношениях с помощью формулы

Формула calculate_attrition_stats предоставляет возможность анализировать сотрудников в сравнении (уволенных и работающих) по интересующим параметрам в процентах и долях. Дополнительно в функцию заложена возможность применения различных метрик и фильтрации. Такая вариативность позволяет ответь практически на любой вопрос заказчиков и/или стейкхолдеров

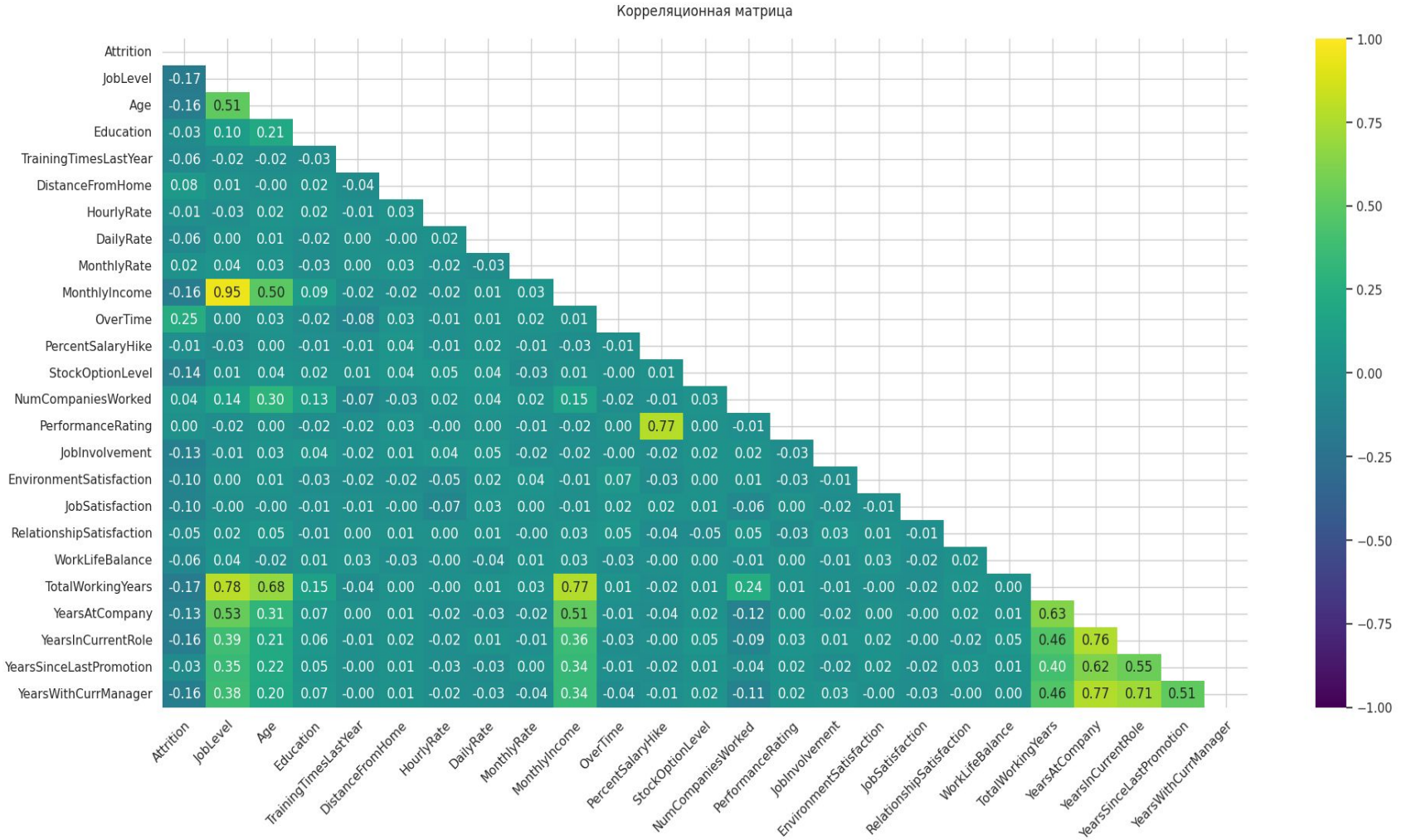
Увольнения в разрезе отделов и гендера

	Department	Gender	Attrition	count	count_share	count_share_pct
0	Human Resources	Female	0	14	0.70	70.00
1	Human Resources	Female	1	6	0.30	30.00
2	Human Resources	Male	0	37	0.86	86.05
3	Human Resources	Male	1	6	0.14	13.95
4	Research & Development	Female	0	336	0.89	88.65
5	Research & Development	Female	1	43	0.11	11.35
6	Research & Development	Male	0	492	0.85	84.54
7	Research & Development	Male	1	90	0.15	15.46
8	Sales	Female	0	151	0.80	79.89
9	Sales	Female	1	38	0.20	20.11
10	Sales	Male	0	203	0.79	78.99
11	Sales	Male	1	54	0.21	21.01

Вывод:

Отдел	Женщины (Attrition)	Мужчины (Attrition)	Общий вывод
HR	30% (высокий)	13,95% (средний)	Женщины уходят в 2 раза чаще
R&D	11,35% (низкий)	15,46% (средний)	Самый стабильный отдел, мужчины уходят чуть чаще
Sales	20,11% (высокий)	21,01% (высокий)	Максимальная текучесть, гендерный паритет

Анализ с помощью корреляционной матрицы помог увидеть связи переменных и оценить их значимость



Вывод: **очень высокая корреляция между** «Уровнем работы» и «Ежемесячным доходом» - 0,95

высокая корреляция между
«Общим стажем работы» и «Уровнем работы» - 0,78
«Общим стажем работы» и «Ежемесячным доходом» - 0,77
«Рейтингом производительности» и «Процентом повышения зарплаты» - 0,77

средняя корреляция:
«Количеством лет в компании» и «Уровнем работы» - 0,53
«Возрастом» и «Уровнем работы» - 0,51
«Ежемесячным доходом» и «Возрастом» - 0,50
«Количеством лет в компании» и «Ежемесячным доходом» - 0,51

Модель машинного обучения. Логистическая регрессия

Цель построения данной модели — прогнозирование вероятности увольнения сотрудников (Attrition) на основе их характеристик и факторов рабочей среды. Прогнозирование риска увольнения

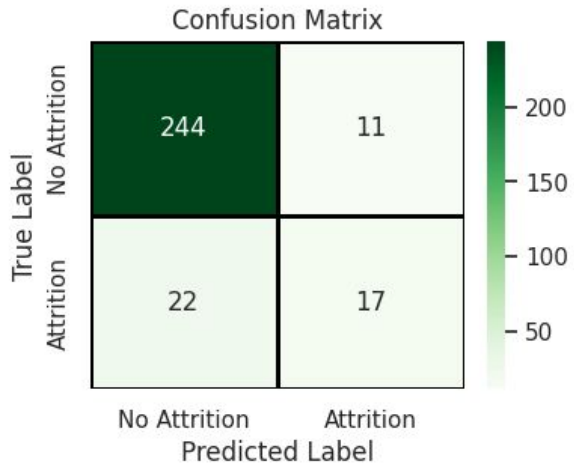
Признаки категориальные:			
EducationField	MaritalStatus	JobRole	Department
OverTime	Gender	BusinessTravel	

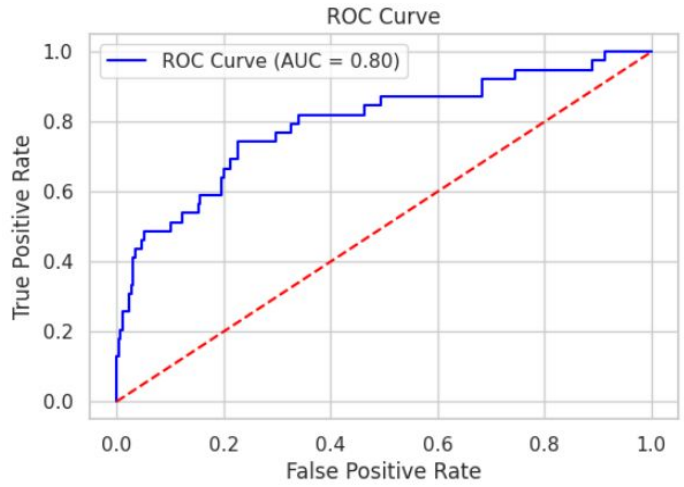
Признаки числовые:							
Age	DistanceFromHome	MonthlyIncome	YearsAtCompany	YearsSinceLastPromotion	JobLevel	Education	TrainingTimesLastYear
HourlyRate	DailyRate	MonthlyRate	PercentSalaryHike	StockOptionLevel	NumCompaniesWorkedм	PerformanceRating	JobInvolvement
EnvironmentSatisfaction	JobSatisfaction	RelationshipSatisfaction	WorkLifeBalance	TotalWorkingYears	YearsInCurrentRole	YearsWithCurrManager	

Целевая переменная:	**Attrition (бинарная: Yes/No)**

Модель машинного обучения. Логистическая регрессия

	Metric	Value
0	Train Accuracy	0.89
1	Test Accuracy	0.89
2	Precision (class 1)	0.61
3	Recall (class 1)	0.44
4	F1 (class 1)	0.51
5	ROC-AUC Score	0.80





Сильные стороны модели: высокий ROC-AUC (0,80) указывает на хорошую способность модели различать классы. Умеренная точность (precision = 0,61) означает, что, когда модель предсказывает увольнение, она часто права

Слабые стороны: низкий recall (0,44) — модель пропускает много реальных увольнений (FN = 22). Низкий F1 Score (0,51) указывает на несбалансированность precision и recall

Итог: модель имеет приемлемое качество (AUC = 0,81), но требует доработки для снижения количества пропущенных увольнений (улучшение recall)

На основе модели определены важные факторы увольнений:

Определены группы риска среди работающих сотрудников:

Risk_Group	
Low Risk	94.0%
Medium Risk	5.8%
High Risk	0.2%

Группа высокого риска High Risk (0,2%):

Количество сотрудников в зоне высокого риска увольнения: 3
Средняя вероятность увольнения: 76.32%
Средний уровень удовлетворенности работой: 3.0

Отделы:	Research & Development	2
	Sales	1
Должности:	Research Director	1
	Sales Representative	1
	Research Scientist	1

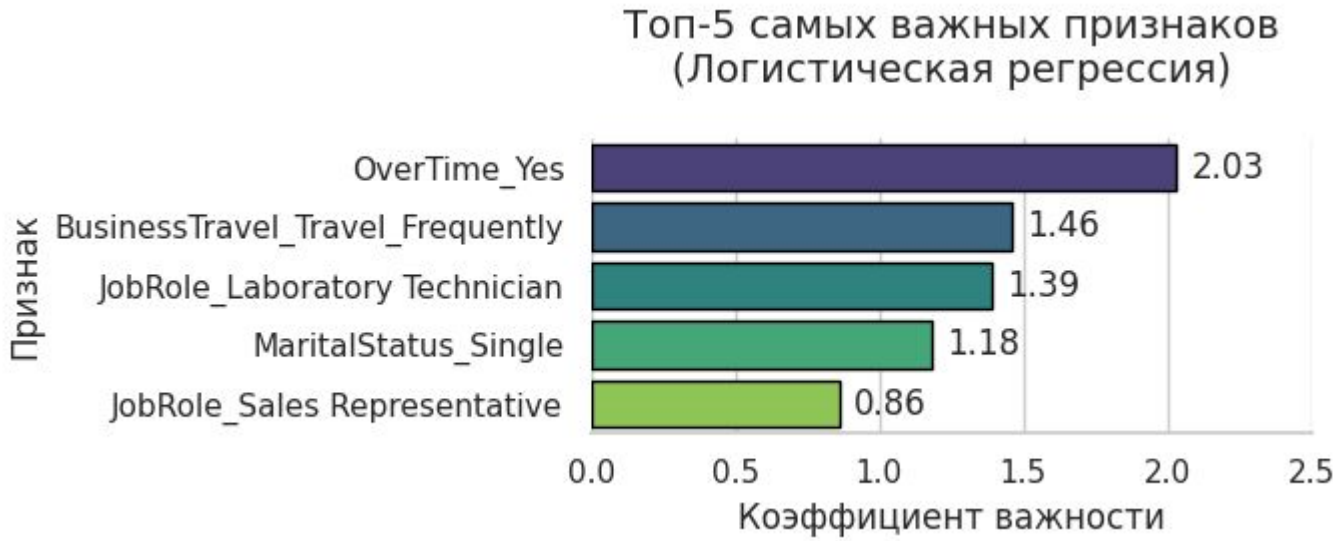
Группа среднего риска Medium Risk (5,8%):

Количество сотрудников в зоне среднего риска увольнения: 71
Средняя вероятность увольнения: 50.60%
Средний уровень удовлетворенности работой: 2.380281690140845

Отделы:	Research & Development	40
	Sales	28
	Human Resources	3
Должности:	Sales Executive	19
	Research Scientist	17
	Laboratory Technician	16
	Sales Representative	8
	Manufacturing Director	5
	Human Resources	3
	Manager	2
	Healthcare Representative	1

Топ-5 наиболее важных признаков для увольнений:

	Feature	Coefficient
43	OverTime_Yes	1.79
33	BusinessTravel_Travel_Frequently	1.55
26	JobRole_Laboratory Technician	1.31
32	JobRole_Sales Representative	0.96
42	MaritalStatus_Single	0.68



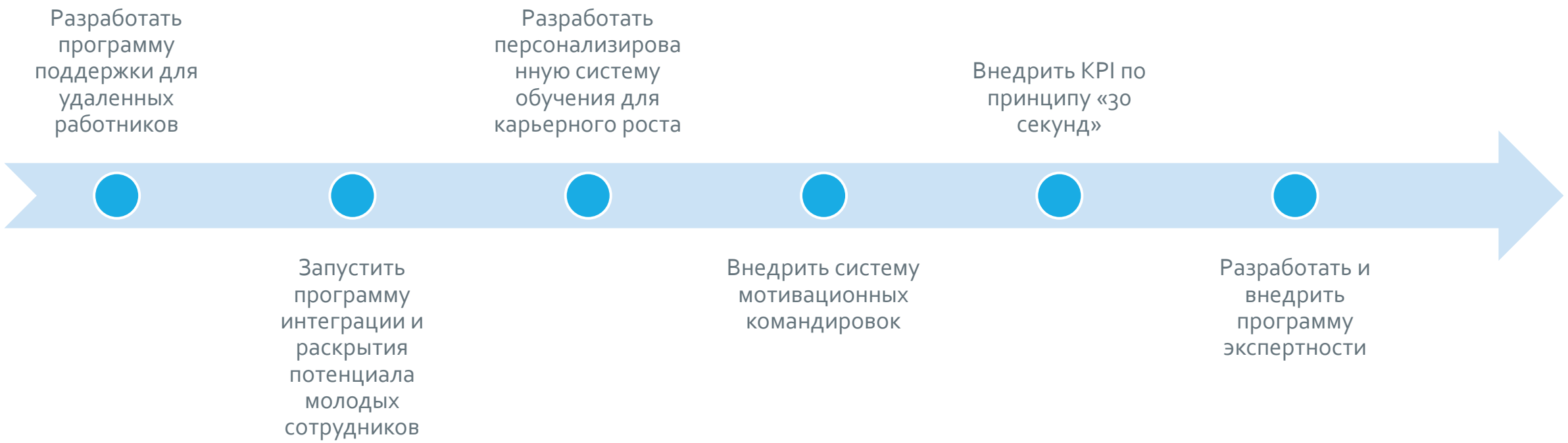
Данная модель логистической регрессии может быть эффективно использована для единичного нового наблюдения

При указании значений параметров модель предсказывает будет ли увольнение или нет и рассчитывает процент риска

Итоги и заключения по проекту

ПОРТРЕТ наиболее явного кандидата на увольнение: молодой человек (до 40 лет), занимающий невысокую должность, выполняющий однообразную работу, требующую большого расхода эмоционально-физических ресурсов. У данного индивида явно не высокая заработная плата и имеются сложности в личной жизни

РЕКОМЕНДАЦИИ СТЕЙКХОЛДЕРАМ Обобщая рекомендации по всем параметрам, можно предложить ряд стратегических мер позволяющих компании достичь желаемых результатов, за счет снижения текучести кадров и перенаправить коллектив в русло развития и достижений:



Для улучшения качества данных датафрейма необходимо доработать отчет, дополнив параметрами:

- дата принятия на работу и дата увольнения – даст возможность отследить наиболее пиковые моменты, вероятность сезонности. И на основе этого подготовить необходимые меры
- наличие иждивенцев – данный показатель может влиять на работоспособность сотрудника, это позволит оптимизировать баланс работы и личной жизни и улучшить адаптацию
- даты повышения квалификации (обучения) – позволит сбалансировать план обучения без ущерба для компании.
- причина увольнения в двух проекциях: со слов работника, по мнению организации – позволит выявить причины увольнения с субъективной точки зрения сотрудников, а также можно будет установить корреляционные связи с иными параметрами датасета. Анализируя мнение организации возможно будут выявлены новые метрики, влияющие на удержание работников в компании