Документация

Дипломная работа по теме:
“Анализ сотрудников компании и факторов увольнений (поиск инсайтов, составление рекомендаций стейкхолдерам, построение предиктивной модели оттока сотрудников)”

Профессия “Аналитик данных”, DA-65
Минеева Елизавета Александровна

г. Москва, 2023 г.

Содержание

Введение

Цели проекта:

Используя фиктивный набор данных, построенный учеными IBM, провести исследование с целью выявления факторов, которые приводят к увольнению сотрудников, и построения моделей машинного обучения для прогнозирования увольнения сотрудников.

Бизнес-Задачи:

1. Определить, для кого будет проводиться анализ, кто стейкхолдеры, какие у них интересы, боли и ожидания, и также описать ценность, которую может принести данное исследование для очерченного круга стейкхолдеров.

2. Описать данные, провести их очистку и интерпретацию;

3. Проверить гипотезы и выявить закономерности в данных. По итогам необходимо сделать выводы, ценные для ранее очерченного круга стейкхолдеров;

4. Визуализировать результаты анализа для удобного и информативного представления результатов исследования бизнес-заказчику;

5. Подготовить отчет-презентацию о результатах исследования.

Блок 1. Описание исходного датасета и типов данных

Для исследования был взят датасет IBM HR Analytics Employee Attrition & Performance. Этот набор данных представляет собой опрос сотрудников IBM, показывающий, имеет ли место увольнение в случае с тем или иным сотрудником компании. Набор данных содержит 1470 записей.

Рассмотрим подробнее представленные в датасете данные:

№	Имя столбца	Описание	Тип данных
1	Age	Возраст сотрудника	int64
2	Attrition	Увольнение сотрудника	object
3	BusinessTravel	Частота командировок	object
4	DailyRate	Ежедневная ставка	int64
5	Department	Отдел, сфера деятельности	object
6	DistanceFromHome	Расстояние от дома до работы	int64
7	Education	Образование	int64
8	EducationField	Сфера образования, специальность	object
9	EmployeeCount	Количество сотрудников (одно уникальное значение - 1)	int64
10	EmployeeNumber	ID сотрудника	int64
11	EnvironmentSatisfaction	Удовлетворение окружающей средой на работе	int64
12	Gender	Пол сотрудника	object
13	HourlyRate	Часовая ставка сотрудника	int64
14	JobInvolvement	Вовлеченность сотрудника в работу	int64
15	JobLevel	Должностная позиция сотрудника (начинающий специалист, специалист среднего уровня, старший специалист, руководитель, управляющий)	int64
16	JobRole	Конкретная должность сотрудника	object
17	JobSatisfaction	Удовлетворенность работой	int64
18	MaritalStatus	Семейное положение	object
19	MonthlyIncome	Месячный доход	int64
20	MonthlyRate	Месячная ставка	int64
21	NumCompaniesWorked	Количество компаний, в которых работал сотрудник	int64
22	Over18	Старше ли сотрудник 18 лет (одно уникальное значение - “Y”, так как по законодательству США разрешенный возраст начала трудовой деятельности - 18	object

		лет)	
23	OverTime	Работает ли сотрудник сверхурочно	object
24	PercentSalaryHike	Процентное увеличение заработной платы	int64
25	PerformanceRating	Рейтинг производительности	int64
26	RelationshipSatisfaction	Удовлетворенность отношениями в рабочей среде	int64
27	StandardHours	Количество рабочих часов (одно уникальное значение - 80)	int64
28	StockOptionLevel	Опционы на акции сотрудников (вид компенсации акционерного капитала, предоставляемой компаниями своим сотрудникам и руководителям)	int64
29	TotalWorkingYears	Количество отработанных лет, трудовой стаж	int64
30	TrainingTimesLastYear	Количество часов, потраченных на обучение в прошлом году	int64
31	WorkLifeBalance	Баланс между работой и личной жизнью	int64
32	YearsAtCompany	Количество лет, отработанных в данной компании	int64
33	YearsInCurrentRole	Количество лет в данной должности	int64
34	YearsSinceLastPromotion	Количество лет, прошедших с момента последнего повышения в должности	int64
35	YearsWithCurrManager	Количество лет, проведенных с текущим руководителем	int64

Блок 2. Подготовка и преобразование данных

В ходе исследования качества данных были проведены следующие действия:

2.1 Получение сводной информации о данных

Получили информацию о количестве признаков, количестве пустых значений и типе данных (данная информация отражена в комментарии к данному шагу и в блоке 1).

2.2 Анализ на наличие пропущенных значений и дубликатов в данных

С помощью проведенного анализа пришли к выводу, что в представленном датасете не наблюдается пропущенных значений и дубликатов, поэтому дополнительные преобразования и очистку проводить не требуется.

2.3 Определение числовых и категориальных признаков

Разделили признаки на числовые и категориальные, привели категориальные признаки к числовой классификации. Результат данной работы сохранили в специальном словаре, который дает информацию о закодированном значении и его исходном значении.

2.4 Проверка числовых данных на аномалии

С помощью гистограммы проверили числовые данные аномалии и обнаружили следующее:

- В столбце 'EmployeeCount' одно уникальное значение - 1.0.
- В столбце 'StandardHours' все значения - 80.0.
- Столбец 'EmployeeNumber' - идентификационный номер сотрудника, который совпадает с индексом каждой строки.

Можно сделать вывод, что представленные выши признаки не являются статистически значимыми.

Также было обнаружено, что в столбцах 'Education', 'JobLevel' - пять уникальных значений (от 1 до 5), 'EnvironmentSatisfaction', 'JobInvolvement', 'JobSatisfaction', 'RelationshipSatisfaction', 'WorkLifeBalance' - четыре уникальных значения (от 1 до 4), 'StockOptionLevel' - четыре уникальных значения (от 0 до 3),'PerformanceRating' - два уникальных значения (3 и 4). Данное наблюдение может говорить о том, что данные признаки по сути являются категориальными, но предварительно уже были приведены в числовой вид.

Таким образом, представленные числовые данные не содержат никаких аномалий.

2.5 Проверка числовых данных на аномалии

index	Attrition	BusinessTravel	Department	EducationField	Gender	JobRole	MaritalStatus	Over18	OverTime
value_ratios	{'No': 83.87755102040816, 'Yes': 16.122448979591837}	{'Travel_Rarely': 70.95238095238095, 'Travel_Frequently': 18.843537414965986, 'Non-Travel': 10.204081632653061}	{'Research & Development': 65.37414965986395, 'Sales': 30.34013605442177, 'Human Resources': 4.285714285714286}	{'Life Sciences': 41.224489795918366, 'Medical': 31.564625850340132, 'Marketing': 10.816326530612246, 'Technical Degree': 8.979591836734693, 'Other': 5.578231292517007, 'Human Resources': 1.836734693877513}	{'Male': 60.0, 'Female': 40.0}	{'Sales Executive': 22.176870748299322, 'Research Scientist': 19.86394557823129, 'Laboratory Technician': 17.61904761904762, 'Manufacturing Director': 9.863945578231291, 'Healthcare Representative': 8.91156462585034, 'Manager': 6.938775510204081, 'Sales Representative': 5.646258503401361, 'Research Director': 5.442176870748299, 'Human Resources': 3.537414965986395}	{'Married': 45.78231292517007, 'Single': 31.97278911564626, 'Divorced': 22.244897959183675}	{'Y': 100.0}	{'No': 71.70068027210884, 'Yes': 28.29931972789116}

Данные согласно данному отчету распределены нормально. На гистограмме, представленной в ноутбуке, мы также не видим отклоняющихся от нормы значений.

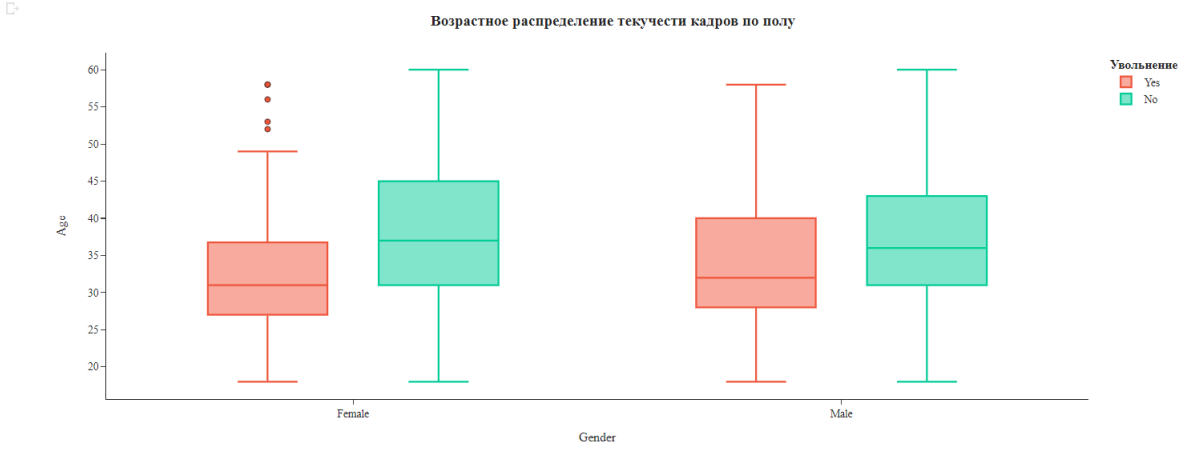
Признак "Over18" имеет только одно значение "Y", указывающее на то, что все сотрудники в наборе данных старше 18 лет, что соответствует законодательству США, где возраст начала трудовой деятельности составляет 18 лет.

Таким образом, представленные данные не содержат пропущенных значений, дубликатов и аномалий, поэтому предварительная обработка данных и очистка не были проведены.

Блок 3. Анализ данных для стейкхолдеров

Целью блока является поиск тенденций и инсайтов для составления рекомендаций стейкхолдерам.

3.1 Возрастное распределение текучести кадров по полу



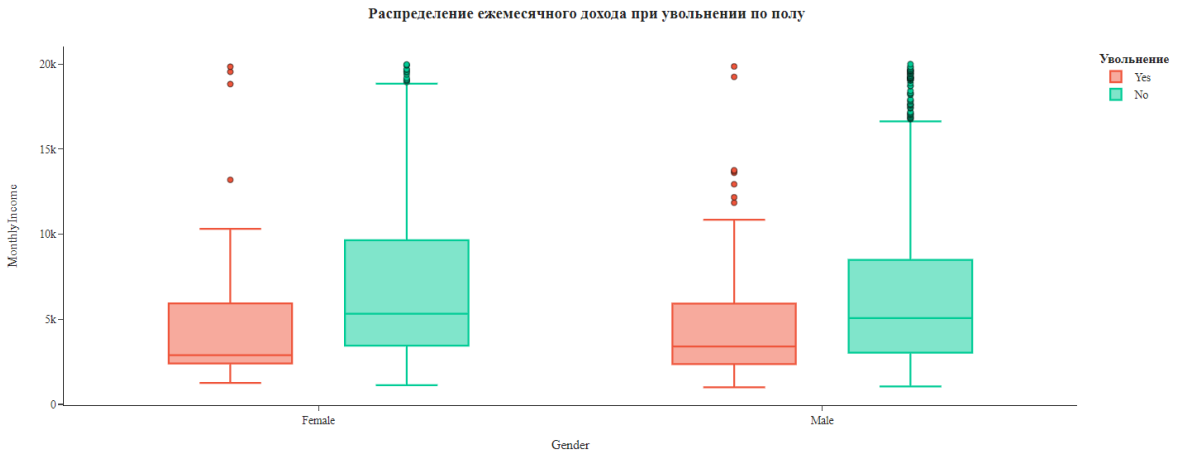
Вывод:

По данному графику видим, что женщины и мужчины в среднем увольняются примерно в одно и то же время - женщины в 31 год, мужчины в 32.

Рекомендация стейкхолдерам:

Руководителям HR-департамента следует уделить особое внимание данной группе сотрудников. Например, предложить дополнительные тренинги для карьерного роста, при возможности, дополнительные обязанности и повышение заработной платы для поощрения сотрудников и минимизации возможности увольнения. Также можно предоставить сотрудникам возможность проводить сессии с корпоративным психологом, так как вполне вероятно, что к этому возрасту многие сотрудники получили “эмоциональное выгорание” от текущего рода деятельности.

3.2 Распределение ежемесячного дохода при увольнении по полу



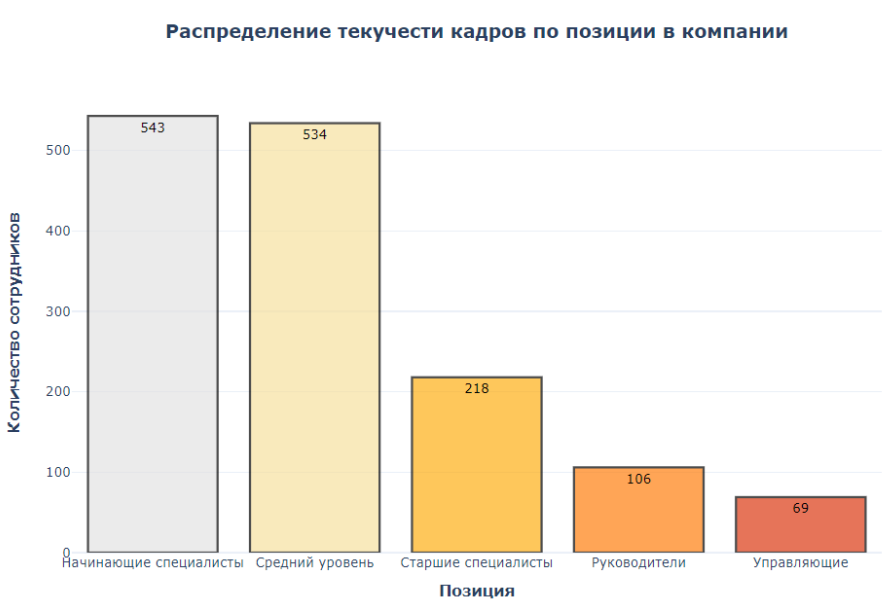
Вывод:

По данному графику мы видим, что и мужчины, и женщины чаще всего увольняются тогда, когда их ежемесячный доход меньше 5000. Также мы видим, что медианное значение при увольнении чуть меньше, чем у мужчин (2886 у женщин против 3407.5 у мужчин). Возможно, это связано с тем, что заработок женщины обычно ниже заработка мужчин.

Рекомендация стейкхолдерам:

Стоит обратить внимание на данную группу сотрудников, которые, как правило, являются начинающими специалистами. Нужно поощрять участие в различных тренингах, проводимых компанией, горизонтальный рост внутри компании, если сотрудника не устраивает текущая должность и он/она хочет попробовать себя в новой сфере.

3.3 Распределение текучести кадров по позиции в компании



Вывод:

Большая доля увольняющихся сотрудников приходится на уровень 1 (начинающие специалисты), которые в силу возраста склонны чаще менять работу. Это может быть обусловлено тем, что, набираясь опыта, молодые люди уходят в другие компании на более выгодных условиях.

Сотрудники уровня 2 (средний уровень) также имеют относительно высокую текучесть кадров, что обусловлено возрастом сотрудников и приобретением ими опыта.

Сотрудники, достигшие уровня 3 (старшие специалисты), 4 (руководители) и 5 (управляющие), увольняются гораздо реже.

Таким образом, из полученных наблюдений можно сделать вывод, что молодые сотрудники, только что пришедшие в компанию, имеют наибольшую склонность к увольнению в сравнению с другими позициями.

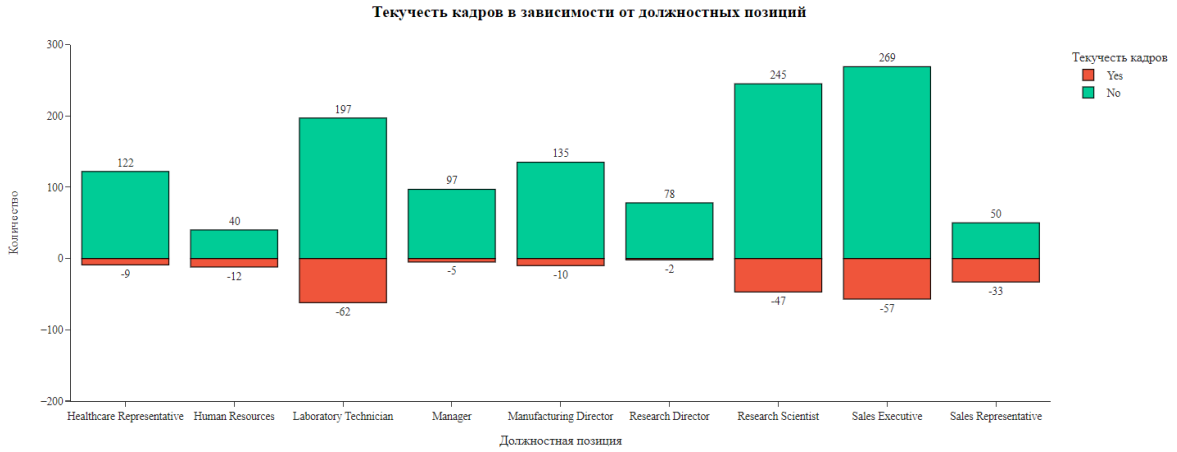
Рекомендация стейкхолдерам:

Рекомендация в данном пункте может повторить рекомендацию, приведенную выше. Нужно поощрять участие начинающих специалистов и специалистов среднего уровня в различных тренингах, проводимых компанией, горизонтальный рост внутри компании, если сотрудника не устраивает текущая должность и он/она хочет попробовать себя в новой сфере.

Также многие начинающие специалисты склонны перерабатывать и эмоционально вкладываться в работу, поэтому сессии с корпоративным психологом, организация зоны отдыха сотрудников, различные корпоративные мероприятия для поднятия командного духа и, конечно же, материальные поощрения в виде премии за

качественно проделанную работу будут стимулировать сотрудников к продолжению трудовой деятельности в рамках текущей компании.

3.4 Текучесть кадров в зависимости от должностных позиций



Вывод:

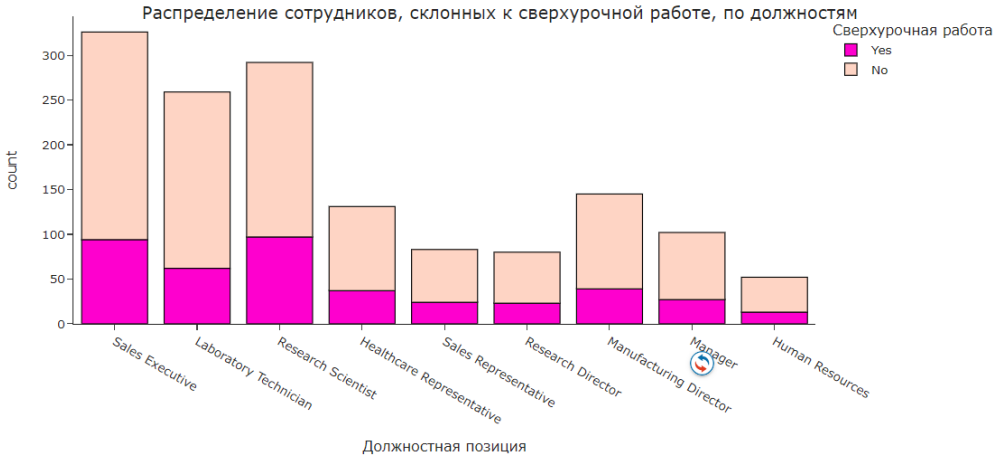
Согласно столбчатой диаграмме выше, наиболее представлены в компании руководители отдела продаж (Sales Executive), научные сотрудники (Research Scientist) и техники-лаборанты (Laboratory Technician).

Должностные позиции, наиболее склонные к увольнению — руководители отдела продаж (Sales Executive), торговые представители (Sales Representative), техники-лаборанты (Laboratory Technician) и научные сотрудники (Research Scientist). Должностные позиции, наименее склонные к увольнению — директора по исследованиям (Research Director), менеджеры (Manager) и представители здравоохранения (Healthcare Representative).

Рекомендация стейкхолдерам:

Следует разобраться, в чем причина увольнения сотрудников, занимающих те или иные должности (например, можно опросить сотрудников, уже решивших уволиться). Возможно, в данных отделах неполный штат и на сотрудников распределяется очень много обязанностей. Если дело в самом роде деятельности и стрессовых факторах, сопутствующих ему, стоит изменить систему премирования, организовать тренинги личностного роста и сессии с корпоративным психологом, а также организовать комфортную зону отдыха и предоставить различные корпоративные скидки, например, на массаж или на медицину.

3.5 Распределение сотрудников, склонных к сверхурочной работе, по должностям



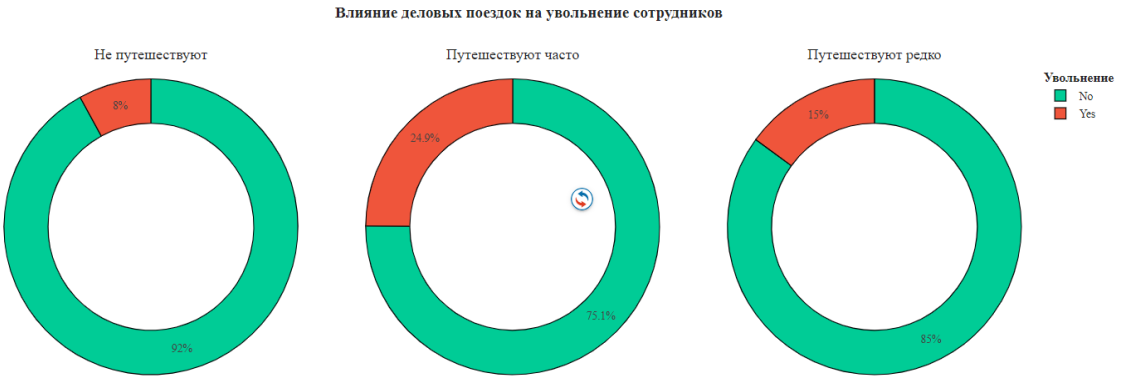
Вывод:

Согласно данному графику мы видим, что около трети всех сотрудников привыкли работать сверхурочно. Наибольший процент (около 28%) у должностных позиций, наиболее склонных к увольнению — руководители отдела продаж (Sales Executive), торговые представители (Sales Representative), техники-лаборанты (Laboratory Technician) и научные сотрудники (Research Scientist). Таким образом, корреляция между уровнем сверхурочной работы и увольнением точно имеется.

Рекомендация стейкхолдерам:

Рекомендации те же, что и в пункте 3.4 касательно отдыха.

3.6 Влияние деловых поездок на увольнение сотрудников



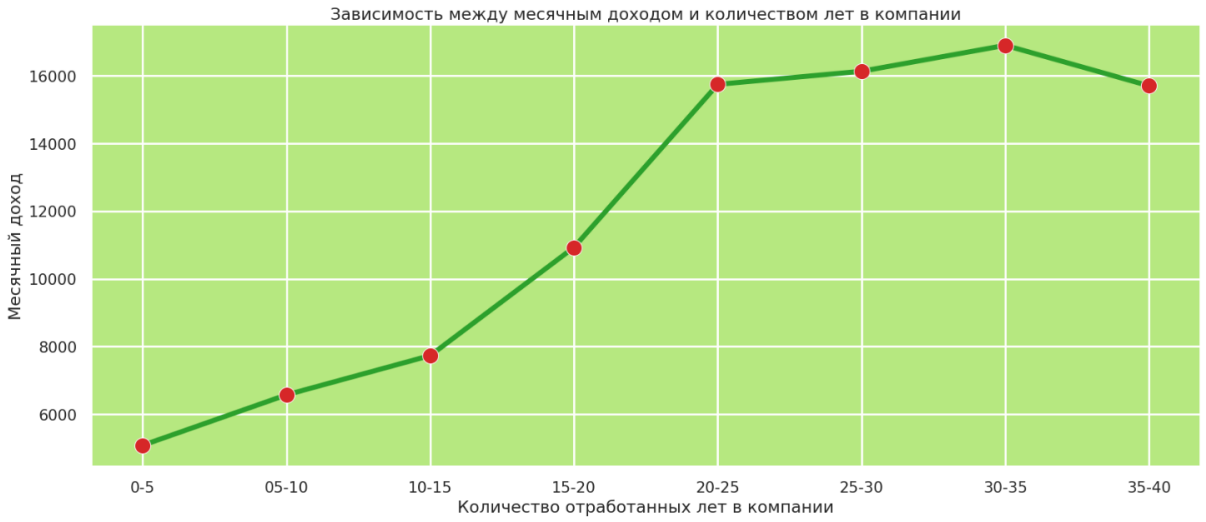
Вывод:

На данных графиках мы видим четкую корреляцию между количеством деловых поездок и степенью текучести кадров. Возможно, это связано с тем, что сотрудники, имеющие множество деловых поездок, меньше бывают дома и больше устают, что приводит к выгоранию и демотивации.

Рекомендация стейкхолдерам:

Обеспечить равномерную нагрузку сотрудников касательно деловых поездок. Не допускать вероятности того, что в деловые поездки будут отправляться одни и те же сотрудники.

3.7 Изменение заработной платы в зависимости от срока работы в компании и ежемесячный доход по общему количеству лет работы и занимаемой должности



Вывод:

На основании первого линейного графика можно сделать вывод, что в период работы в компании от 0 до 35 лет средняя заработная плата сотрудников постепенно

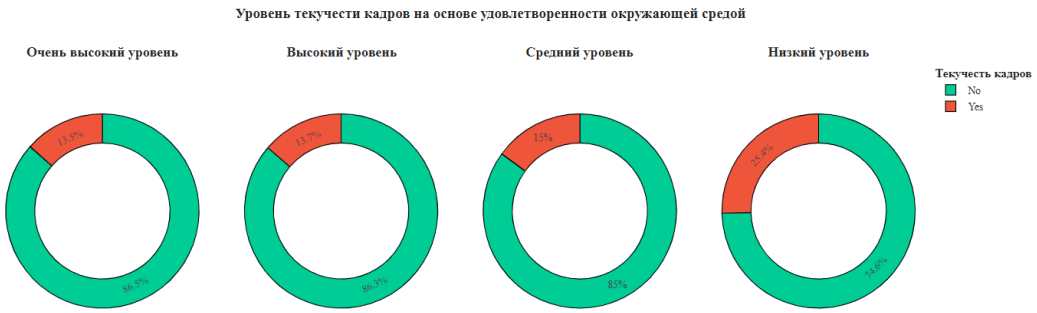
увеличивается с течением времени. Это объясняется тем, что при длительной работе в компании сотрудники набираются опыта, что поощряется компанией. Однако можно заметить еще одну интересную тенденцию - заработная плата в период работы с 35-40 лет немного уменьшается. Причина такого снижения может заключаться в том, что при стаже работы в компании 35-40 лет эти сотрудники приближаются к пенсионному возрасту, поэтому их производительность снижается или они уступают руководящие должности более молодым сотрудникам.

На основании второго графика можно сделать простой вывод: чем выше должность, тем больше ежемесячный доход. Начинающие специалисты и специалисты на среднем уровне обычно работают от 0 до 20 лет. Чтобы иметь возможность занимать высшую руководящую должность (уровень называется "Управляющие"), необходимо проработать более 20 лет, даже более 35 лет.

Рекомендация стейкхолдерам:

Данный отчет может использоваться для поднятия духа начинающих специалистов, которые склонны увольняться в первые годы работы из-за низкого дохода.

3.8 Удовлетворенность сотрудников различными сферами в рабочей среде



Вывод:

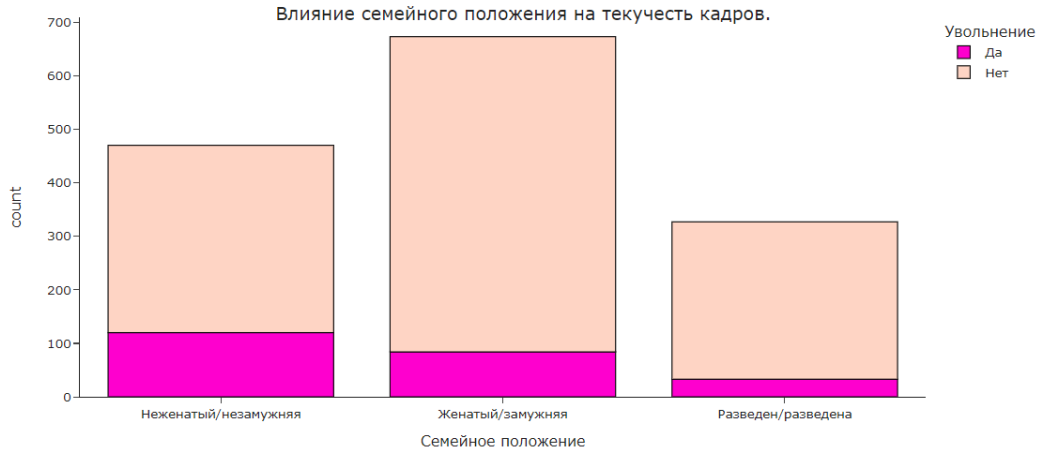
На данных графиках видим отчетливую корреляцию между текучестью кадров и уровнями удовлетворенности работой, окружающей средой и взаимоотношениями, что доказывает тот факт, что заработная плата и карьерный рост не являются единственными факторами, способствующими увольнению.

Рекомендация стейкхолдерам:

Для повышения уровня удовлетворенности работой нужно составить прозрачную систему премирования, чтобы сотрудники стремились к определенным результатам и были удовлетворены своими достижениями.

Касательно удовлетворенности окружающей средой и взаимоотношениями нужно провести следующие шаги. Во-первых, организовать пространство в офисе таким образом, чтобы у сотрудников была возможность отдохнуть в тишине и комфорте. Во-вторых, для улучшения взаимоотношений между сотрудниками нужны регулярные встречи для обсуждения итогов и дальнейших планов (например, по методике Scrum) и различные тимбилдинги.

3.9 Влияние семейного положения на текучесть кадров



Вывод:

Процент сотрудников, состоящих в браке, составляет наибольшее количество, что является логичным, когда средний возраст сотрудников в этой компании составляет около 36 лет.

У одиноких сотрудников наблюдается наиболее высокий коэффициент текучести. Молодые люди, у которых нет семьи и определенных связанных с данным обстоятельством обязательств, склонны чаще менять работу и "искать себя".

Рекомендация стейкхолдерам:

Сотрудникам HR-отдела следует обратить внимание на неженатых/незамужних сотрудников. Как уже было сказано выше, как правило, это начинающие и молодые специалисты, поэтому нужно поощрять их участие в различных тренингах, проводимых компанией, горизонтальный рост внутри компании, если сотрудника не устраивает текущая должность и он/она хочет попробовать себя в новой сфере.

Блок 4. Построение модели логистической регрессии

Основная задача, которую необходимо решить, - предсказать, есть ли риск того, что тот или иной сотрудник покинет компанию.

Для решения данной задачи было решено использовать логистическую регрессию, где целевой переменной будут увольнения (признак "*Attrition*"). Данный метод был выбран по той причине, что требуется доказать бинарную величину (уволится ли сотрудник: 1 (да)/0 (нет)).

Для оценки качества модели была выбрана метрика *accuracy*, так как перед нами стоит задача бинарной классификации.

В ходе построения модели логистической регрессии были получены следующие результаты, что говорит о том, что точность модели составляет 88%. Данный коэффициент является довольно высоким, учитывая ограниченность вводных данных.

```
[63] accuracy_train = sklearn.metrics.accuracy_score(predictions_train, y_train)
     print(accuracy_train)
```

```
0.8928571428571429
```

```
accuracy_test = sklearn.metrics.accuracy_score(predictions_test, y_test)
print(accuracy_test)
```

```
0.8843537414965986
```

Итоги проекта и заключение

Текучесть кадров - это проблема, которая затрагивает все предприятия, независимо от географии, отрасли и размера компании. Увольнение сотрудников приводит к значительным затратам для бизнеса, включая затраты на срыв бизнеса, найм нового персонала и обучение нового персонала. Таким образом, были найдены нужные инсайты и рекомендации, которые смогут помочь понять причины текучести кадров и ее минимизировать.

По бизнес-задачам:

1. Анализ основных тенденций помог определить, что представленный отчет будет полезен высшему руководству компании (исполнительному, операционному директорам), а также руководителю HR-департамента, что поможет в результате снизить отток сотрудников и расходы, связанные с ним.

2. В результате проведенного анализа данных было выяснено, что представленный датасет является довольно качественным: в нем отсутствуют пропуски и дубликаты, аномалии же, напротив, отсутствуют.

3. Благодаря анализу основных тенденций можно дать следующие рекомендации стейкхолдерам:

- поощрять участие начинающих специалистов и специалистов среднего уровня в различных тренингах, проводимых компанией, горизонтальный рост внутри компании;
- создать прозрачную систему премирования для поощрения сотрудников за проделанную работу и стимулирования к дальнейшей деятельности;
- организовать зоны отдыха сотрудников, предоставить возможность проведения сессий с корпоративным психологом и предоставления различных корпоративных скидок;
- разобраться, в чем причина увольнения сотрудников, занимающих те или иные должности, исключить возможность неполного штата и неравномерной нагрузки на сотрудников;
- обеспечить равномерную нагрузку сотрудников касательно деловых поездок;
- организация тимбилдингов для поднятия командного духа.

4. Проведенный анализ тенденций содержит большое количество визуализаций для удобного и информативного представления результатов исследования бизнес-заказчику.

5. На этапе построения модели машинного обучения было достигнуто следующее: :

- была проведена предварительная обработка данных, категориальные признаки были приведены к числовой классификации;
- были удалены признаки, которые не являются статистически значимыми;

- для бинарной классификации (для предсказания принадлежности одному из двух классов (0 или 1) зависимой переменной "Attrition") была использована логистическая регрессия;
- для оценки качества модели была использована метрика accuracy, которая на тренировочных и тестовых данных составила 0.88 и 0.89 соответственно, что говорит об успешности данной модели.

6. Отчет-презентация был подготовлен отдельным файлом, который приложен к данной работе.

Таким образом, полученные в результате исследования данные могут быть использованы для поиска причин увольнения, их устранения и обеспечения бизнес-процессов внутри компании с минимальным коэффициентом текучести кадров.