**Документация**

**Дипломная работа по теме:**

**“Анализ сотрудников компании, факторов увольнений и построение предиктивной модели оттока сотрудников (поиск инсайтов, составление рекомендаций стейкхолдерам)”**

**Профессия “Аналитик данных”, DA-72**

**Макушкин А.Б.**

**г. Москва, 2025 г.**

Оглавление

[Введение 3](#_Toc155513088)

[Блок 1. Описание исходного датасета и типов данных (35 столбцов) 5](#_Toc155513089)

[Блок 2. Подготовка и преобразование данных 8](#_Toc155513090)

[Блок 3. Анализ данных для стейкхолдеров 8](#_Toc155513091)

[3.1 Общий анализ атрибутов. 8](#_Toc155513092)

[Рекомендации стейкхолдерам. 8](#_Toc155513093)

[3.2 Детальны анализ отдельных атрибутов. 9](#_Toc155513094)

[Рекомендации стейкхолдерам. 13](#_Toc155513095)

[3.3 Корреляция увольнений (Attrition). 14](#_Toc155513096)

[3.4 Тепловая карта. 15](#_Toc155513097)

[Блок 4. Построение модели логистической регрессии. 16](#_Toc155513098)

[Итоги проекта и заключение. 17](#_Toc155513099)

# Введение

Увольнение сотрудников неизбежный процесс практически для любых компаний вне зависимости от размера, количества сотрудников или оборотов.

Увольнение сотрудника всегда несет для компании риски и, как следствие, финансовые потери.

Риски, связанные с увольнением сотрудника:

-время на поиск замены и адаптацию нового сотрудника (в течении этого времени, функционал, выполняемый сотрудником, перераспределяется среди других сотрудников подразделение, что ведет к повышенной нагрузке на других сотрудником и, как следствие их выгоранию),

-уволившийся сотрудник может быть узким специалистом, поэтому время на его замену увеличивается,

-уволившийся сотрудник может занимать руководящую позицию, поэтому нужно заранее готовить ему замену из внутреннего кадрового резерва, или искать на рынке труда.

Время на поиск и адаптацию нового сотрудника может быть очень существенным, исходя из занимаемой должности и необходимых компетенциях.

Поэтому для любой компании важно минимизировать количество увольняемых сотрудников, и в случаях, когда увольнения избежать не удается, знать о потенциальном увольнении заранее. Компания должна не просто набирать квалифицированных специалистов, но и следить за тем, чтобы они оставались на месте.

Исходя из этого стейкхолдерами данного исследования является в первую очередь HR служба компании, а также руководители всех уровней и направлений.

Некоторые вопросы, на которые может помочь ответить данное исследование:

* Как влияет пол на увольнение?
* Как влияют деловые поездки на увольнение?
* Каково влияние расстояния от дома на увольнение?
* В каком департаменте больше всего случаев увольнения?
* Как влияют годы работы с текущим менеджером на увольнение?
* Какие должности больше всего подвержену увольнениям?
* Является ли зарплата одной из основных причин?
* Играет ли роль сфера образования?
* Играет ли роль уровень образования?
* Перегружает ли сверхурочная работа сотрудников и заставляет ли их увольняться?
* Какие факторы в большей степени способствуют увольнения?
* Каково влияние возраста на увольнение?
* Как самоудовлетворенность работой влияет на увольнение?
* Как баланс между работой и личной жизнью влияет на увольнение?
* Как опыт работы влияет на увольнение?
* Как продолжительность работы в текущей должности влияет на увольнение?
* Влияет ли процентный коэффициент увеличения заработной платы на увольнение?
* Сколько новых сотрудников уходит менее чем за год?

**Цели:**

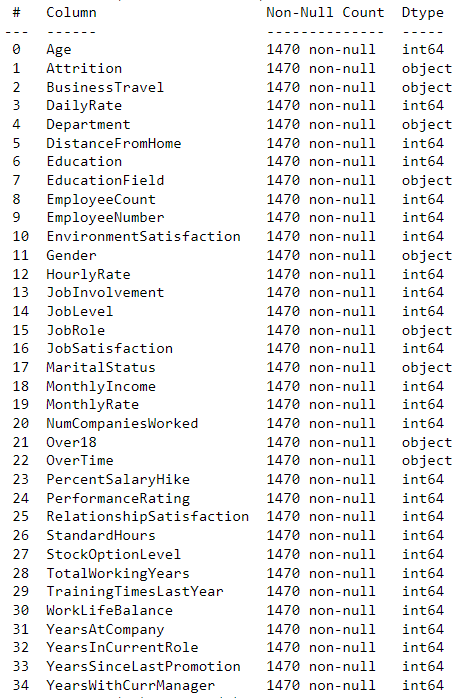
1. Поиск закономерностей и инсайдов по наиболее рисковым зонам (какие группы сотрудников наиболее подвержены риску увольнения). Данная информация поможет действовать на опережение и минимизировать причины высоких рисков. Это могут быть переработки, частые командировки, уровень заработанной платы и т.д.
2. Построение предиктивной модели, что позволит не только работать с профилактикой увольнений сотрудников, но и планировать работу по поиску и развитию персонала.

**Бизнес-Задачи:**

1. Определение категорий сотрудников, наиболее подверженных риску увольнения. Это позволит точечно проводить работы по сохранению персонала.
2. Построение предиктивной модели оттока сотрудников, что позволит периодически анализировать всех сотрудников компании на предмет риска увольнения.

# Блок 1. Описание исходного датасета и типов данных (35 столбцов)

Для исследования был взят датасет [IBM HR Analytics Employee Attrition & Performance](https://www.kaggle.com/datasets/pavansubhasht/ibm-hr-analytics-attrition-dataset/data) со статистикой увольнений сотрудников, созданный специалистами компании IBM.

****

**Атрибуты с информацией о сотрудниках можно разделить на несколько групп:**

1. Персональная информация о сотрудниках.

* Age - возраст сотрудника.
* Gender - пол сотрудника.
* MaritalStatus - семейное положение сотрудника.
* Education - уровень образования сотрудника.
* EducationField - область образования сотрудника.
* NumCompaniesWorked - количество компаний, в которых работал сотрудник.
* DistanceFromHome - расстояние от дома до работы.
* EmployeeNumber - уникальный идентификатор для каждой записи о сотруднике.
* Over18 - сотруднику старше 18 лет.

1. Сведения о работе.

* JobLevel - уровень работы сотрудника.
* JobRole - должность сотрудника.
* Department - департамент, в котором работает сотрудник.
* BusinessTravel - частота командировок сотрудника.

1. Показатели эффективности и удовлетворенности.

* PerformanceRating - производительность, повышающий рейтинг эффективности сотрудника.
* RelationshipSatisfaction - удовлетворенность сотрудника своими отношениями.
* EnvironmentSatisfaction - удовлетворенность сотрудника своей рабочей средой.
* JobSatisfaction - удовлетворенность сотрудника своей работой.
* WorkLifeBalance - Восприятие сотрудника своего баланса между работой и личной жизнью.
* JobInvolvement - уровень вовлеченности сотрудника.

1. Финансовые показатели.

* MonthlyIncome - ежемесячный доход сотрудника.
* MonthlyRate - ежемесячная ставка заработной платы сотрудника.
* DailyRate - ежедневная ставка заработанной платы сотрудника.
* HourlyRate - почасовая ставка заработанной платы сотрудника.
* StockOptionLevel - уровень опциона на акции сотрудника.
* PercentSalaryHike - процентный коэффициент увеличения заработной платы сотрудника.
* StandardHours - стандартное рабочее время сотрудника.

1. История работы в компании.

* EmployeeCount - общее количество сотрудников в организации.
* OverTime - работают ли сотрудник сверхурочно
* TotalWorkingYears - общее количество лет, отработанных сотрудником.
* TrainingTimesLastYear - количество раз, когда сотрудники проходили обучение за последний год.
* YearsSinceLastPromotion - количество лет с момента последнего повышения сотрудника.
* YearsWithCurrManager - количество лет, в течение которых сотрудник работает со своим текущим менеджером
* YearsAtCompany - количество лет, в течение которых сотрудник проработал в компании.
* YearsInCurrentRole - количество лет, в течение которых сотрудник занимал свою текущую должность.

Дополнительное описание к датасету. Приведенные ниже атрибуты являются дискретными.

Education

1. 'Below College'
2. 'College'
3. 'Bachelor'
4. 'Master'
5. 'Doctor'

EnvironmentSatisfaction

1. 'Low'
2. 'Medium'
3. 'High'
4. 'Very High'

JobInvolvement

1. 'Low'
2. 'Medium'
3. 'High'
4. 'Very High'

JobSatisfaction

1. 'Low'
2. 'Medium'
3. 'High'
4. 'Very High'

PerformanceRating

1. 'Low'
2. 'Good'
3. 'Excellent'
4. 'Outstanding'

RelationshipSatisfaction

1. 'Low'
2. 'Medium'
3. 'High'
4. 'Very High'

WorkLifeBalance

1. 'Bad'
2. 'Good'
3. 'Better'
4. 'Best'

# Блок 2. Подготовка и преобразование данных

Датасет содержит 1470 записей. Имеет 35 атрибутов.

В датасете не выявлены пропуски и дубликаты данных.

Столбцы Over18, StandardHours, EmployeeCount, EmployeeNumber удалены, как не несущие информации для данного исследования.

Анализ проводился отдельно для категориальных и числовых атрибутов. Перед построение модели категориальные атрибуты были оцифрованы.

# Блок 3. Анализ данных для стейкхолдеров

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

## 3.1 Общий анализ атрибутов.

Проведен анализ распределения значений атрибутов и зависимости атрибутов и целевого атрибута увольнение (Attrition).

Более детально с графиками можно ознакомиться в Jupyter Notebook.

Поскольку показателей достаточно большое количество, можно с разных сторон смотреть на влияние увольнений и искать инсайты. Вот некоторые выводы из общих графиков:

* Увольнения являются самым высоким в возрасте от 18 до 35 лет и постепенно снижается.
* Средний возраст уволившихся сотрудников ниже, чем средний возраст оставшихся. Молодые более склонны к риску.
* С увеличением дохода увольнения уменьшаются.
* У одиноких людей увольнения происходят чаще.
* Увольнения сотрудников, которые обычно путешествуют, выше, чем у других.
* Увольнение является самым высоким среди сотрудников с показателем уровня работы равным 1.
* Больше всего увольняются на должности Laboratory Technician.
* Чем дольше сотрудник работает в компании, тем меньше вероятность его увольнения

### Рекомендации стейкхолдерам.

Выявлены рисковые зоны. Это сотрудники:

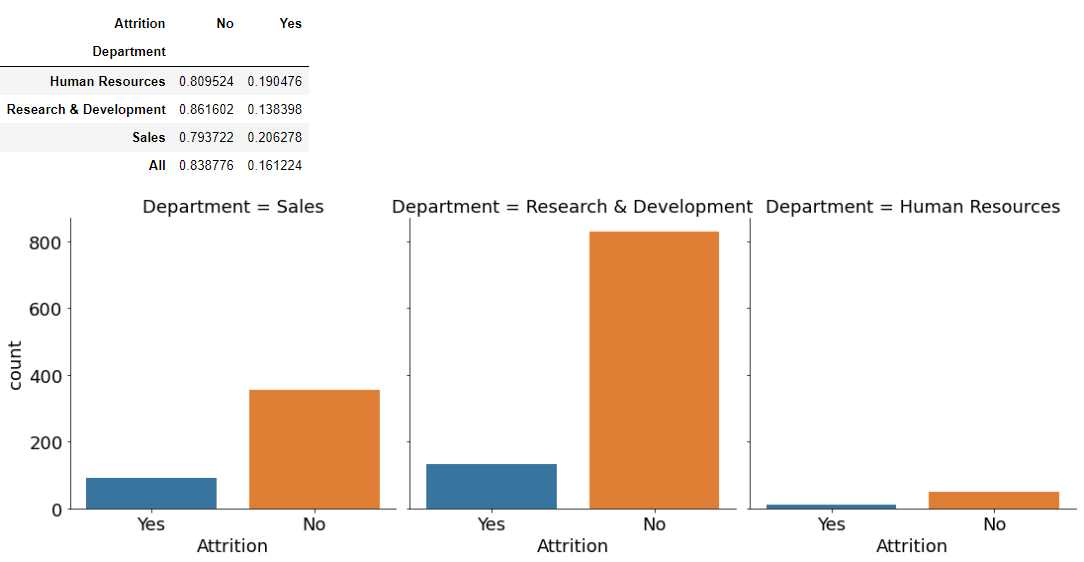
* В возрасте 18-35 лет.
* С доходом ниже среднего.
* Одинокие люди.
* Сотрудники, часто путешествующие.
* Сотрудники с уровнем работы равным 1.
* Сотрудники, отработавшие в компании менее трех лет
* Сотрудники с должностью Laboratory Technician.

Необходимо определить списочных состав сотрудников из данных категорий. Особое внимание уделить сотрудникам, с несколькими совпадениями (например, одинокий сотрудник с должностью Laboratory Technician в возрасте 20 лет).

Поставить задачу руководителям данных сотрудников по осуществлению мероприятий по сохранению персонала. Например, снизить количество командировок, пересмотреть финансовые условия и т.д.

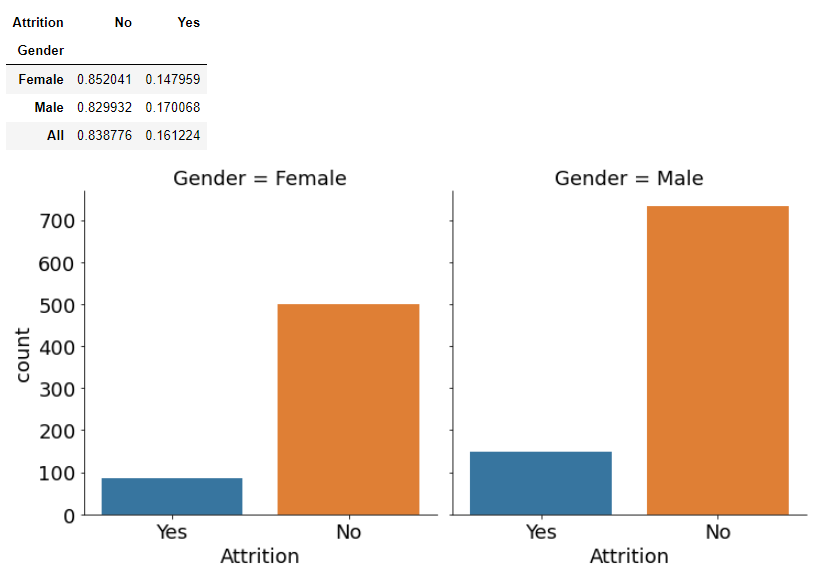
## 3.2 Детальны анализ отдельных атрибутов.

Увольнения по департаментам



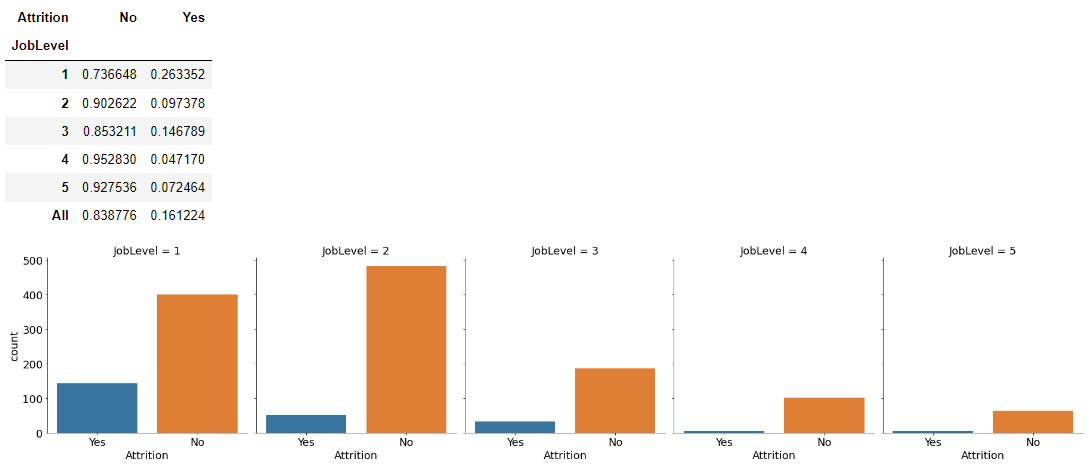
Вывод: наибольший процент увольнений наблюдается в Sales - более 20%

Влияние пола на увольнение



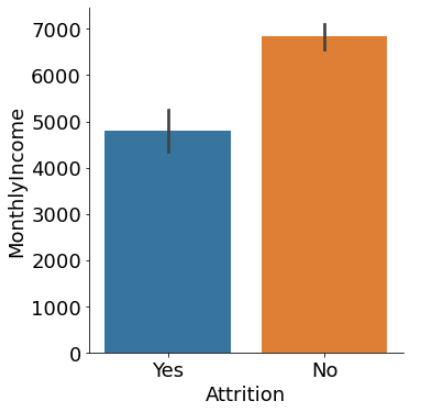
Вывод: мужчины более склонны к увольнениям, чем женщины

Влияние уровня работы на увольнение



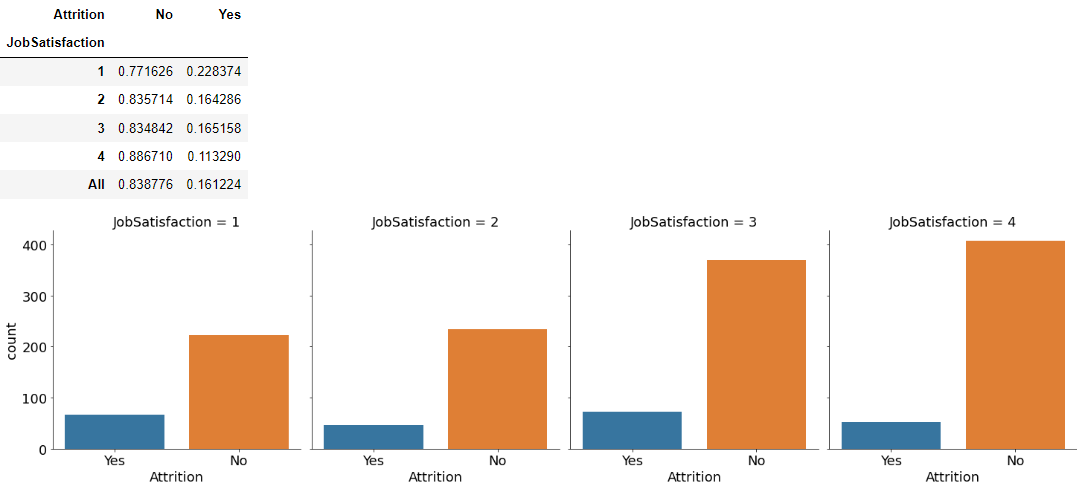
Вывод: люди на 4-м уровне работы менее всего увольняются, при этом на 1-м и 3-м уровне работы максимальные процент увольнений

Влияние ежемесячного дохода



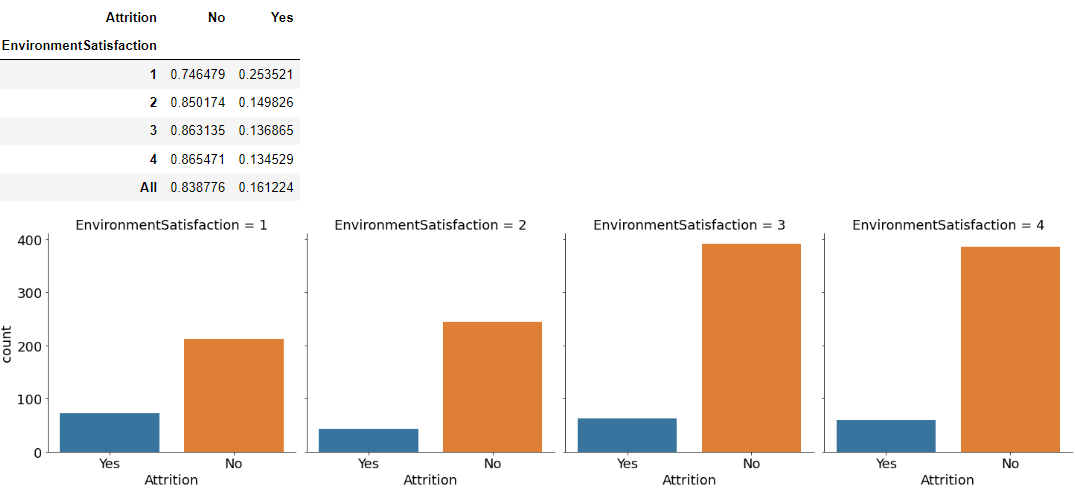
Вывод: Средний доход тех, кто не хочет увольняться существенно выше, чем у тех, кто увольняется. Заработанная плата оказывает существенное влияние на увольнение сотрудников.

Влияние удовлетворенности работой на увольнение



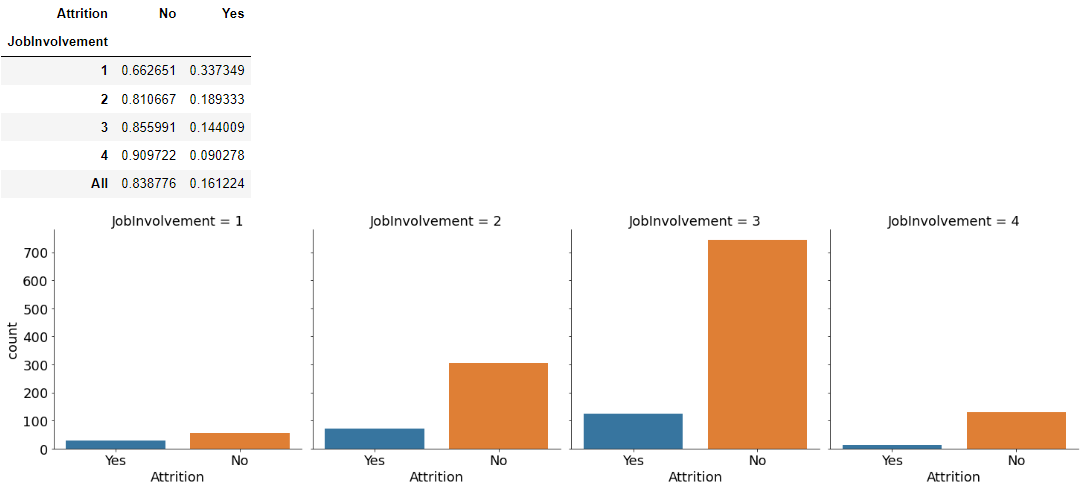
Вывод: при высокой удовлетворенности работой люди меньше увольняются, при этом уровни 2 и 3 сопоставимы.

Влияние удовлетворенности окружающей средой на увольнения

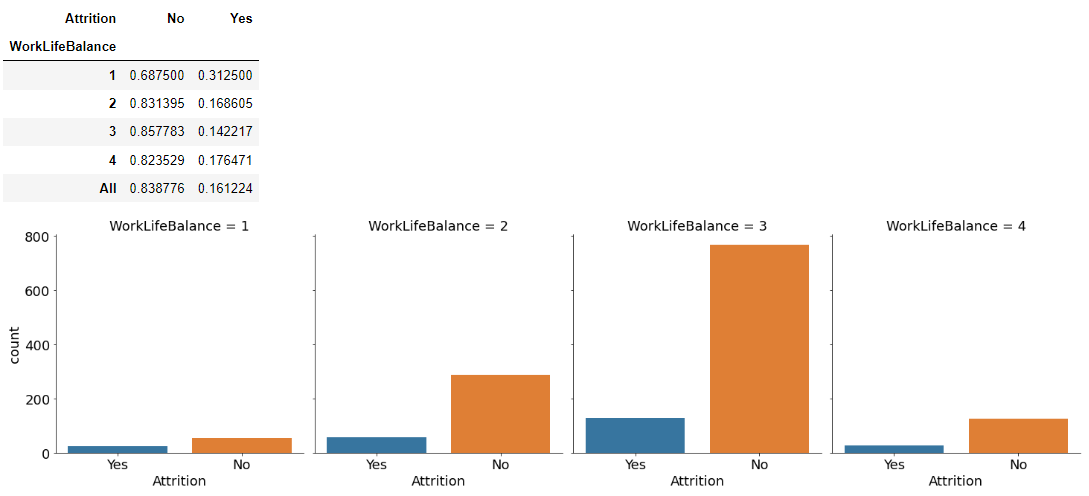


Вывод: больше всего людей увольняется с низким уровнем удовлетворенности, уровни 2, 3, 4 сопоставимы

Влияние вовлеченности в работу на увольнения

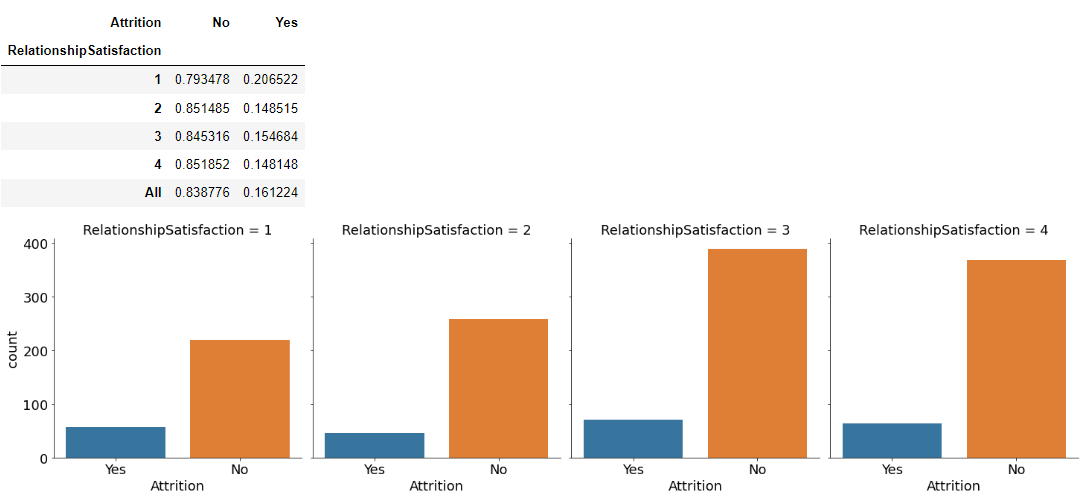


Вывод: наблюдается прямая линейная зависимость увольнений от вовлеченности в работу.

Влияние баланса между работой и личной жизнью на увольнения

Вывод: больше всего людей увольняется на самом низком показателе, при этом между значениями 3 и 4 наблюдается рост увольнений

Влияние удовлетворенности отношениями на увольнения



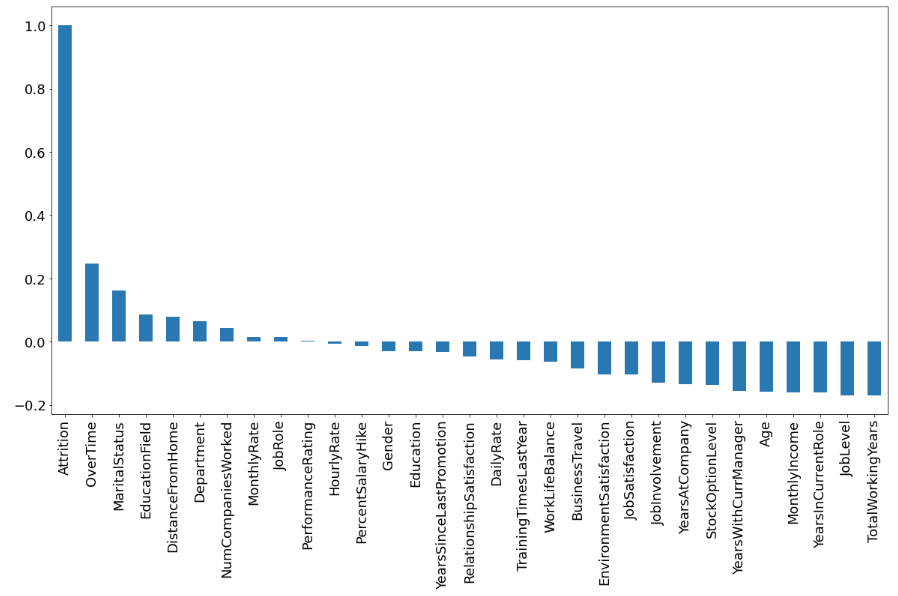
Вывод: одинокие люди существенно чаще увольняются.

### Рекомендации стейкхолдерам.

Детальный анализ отдельных показателей подтверждает выводы из общего анализа.

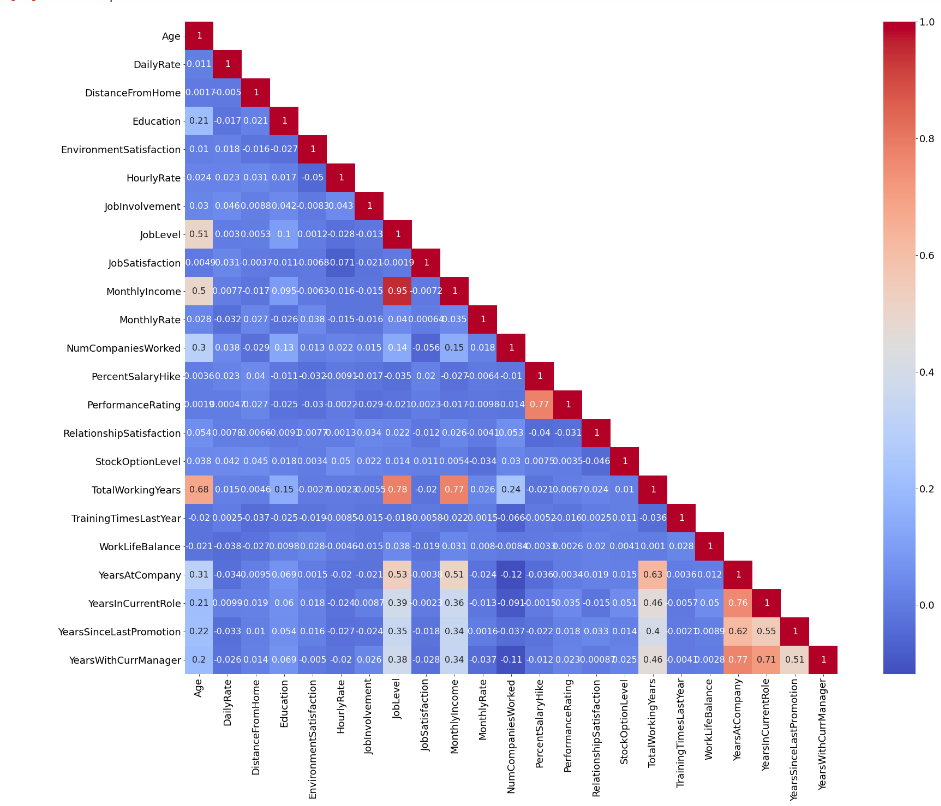
* Взять на контроль департамент Sales, т.к. там наблюдается увеличенный процент увольнений.
* Пусть на немного, но регулярно индексировать заработанную плату.
* Провести отдельно исследования для выявления факторов неудовлетворенности работой.
* Регулярно проводить мероприятия по увеличению вовлеченности сотрудников в работу.
* Сократить переработки для проблемных сегментов. Если переработки неизбежны, пересмотреть дополнительные надбавки к зарплате.

## 3.3 Корреляция увольнений (Attrition).



Вывод: как видно из графика наибольшая корреляция с увольнениями у переработок и семейного положения. В свою очередь количество лет, отработанных в компании и уровень работы имеют наибольшую отрицательную корреляцию.

## 3.4 Тепловая карта.



На синие оттенки не обращаем внимания, т.к. здесь имеем слабую корреляцию. Интерес представляют только красные оттенки тепловой карты. Их мало и можно разобрать детально.

* Возраст, что логично, коррелирует с годами, отработанными в компании.
* Уровень работы имеет сильную корреляцию с месячным доходом.
* Рейтинг эффективности сотрудника коррелирует с процентным коэффициентом увеличения заработной платы сотрудника.
* Количество лет, отработанных в компании коррелирует не только с возрастом, но и с уровнем работы и месячным доходом. А также с годами, отработанными с текущим менеджером, годами на текущей должности.

# Блок 4. Построение модели логистической регрессии.

Логистическая регрессия - это метод статистического анализа для прогнозирования бинарного результата, такого как "да" или "нет", на основе предыдущих наблюдений за набором данных. Модель логистической регрессии предсказывает зависимую переменную данных путем анализа взаимосвязи между одной или несколькими существующими независимыми переменными.

Получена точность модели 0,86, т.е. модель дает верное предсказание в 86% случаях. Таким образом, с помощью данной модели можно периодически анализировать данные сотрудников для выявления потенциальных увольнений.

Нельзя не сказать об особенности модели. Данная модель обучалась на наблюдениях достаточно сильно смещенных в сторону \ Attrition = No, что логично, т.к. большинство сотрудников продолжают работать и не увольняются. Поэтому, чтобы увеличить точность модели можно искусственно синтезировать данные, чтобы устранить данных перекос. Но это цель отдельного исследования.

# Итоги проекта и заключение.

Область исследования актуальна для любой компании.

Стейкхолдерами являются руководители разных уровней и направлений. из этого следует, что актуальность данного исследования будет сохранятся всегда.

По бизнес-задачам:

1. Определены наиболее рисковые категории сотрудников. Это позволит точечно проводить работы по сохранению персонала.
2. Построена предиктивной модели оттока сотрудников на основе логистической регрессии с точностью 86%. Это позволит периодически анализировать всех сотрудников компании на предмет риска увольнения.

.

Рекомендации:

* Продолжить сбор информации о сотрудниках для выявления дополнительных атрибутов, которые могут влиять на увольнения.
* Максимально полно собирать информацию о новых сотрудниках.
* При появлении новых атрибутов перестраивать предиктивную модель с учетом новых данных
* Переобучать модель при появлении новых данных.
* Регулярно проводить анализ с помощью предиктивной модели.
* Осуществлять регулярные мероприятия по сохранению персонала, особенно в рисковых зонах.