МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ
федеральное государственное бюджетное образовательное учреждение высшего образования
«Тольяттинский государственный университет»

Кафедра Прикладная математика и информатика
(наименование)

09.04.03 Прикладная информатика
(код и наименование направления подготовки)

Управление корпоративными информационными процессами
(направленность (профиль))

ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА (МАГИСТЕРСКАЯ ДИССЕРТАЦИЯ)

на тему «Методы и модели прогнозирования текучести кадров в ИТ-компании»

Обучающийся Э.М. Бикмаев
(Инициалы Фамилия) (личная подпись)

Научный руководитель д.т.н., доцент, С.В. Мкртычев
(ученая степень (при наличии), ученое звание (при наличии), Инициалы Фамилия)

Тольятти 2024

Оглавление

Введение

На сегодняшний день в виду сложившихся обстоятельств в мире текучесть кадров особенно негативного влияет на работу компаний в целом. ИТ-отрасль известна своей высокой конкуренцией за квалифицированных специалистов, а их уход связан с обучением новых сотрудников, что является дорогостоящим и времязатратным процессом, в тоже время постоянная текучесть кадров негативно сказывается на эффективности работы команды.

Проблема заключается в том, что текучесть кадров возникает из-за неэффективного использования имеющихся данных о сотрудниках. Большинство аналитических решений основаны на стандартных отчетах и интуиции сотрудников, а также принятии этих решений, когда проблема уже возникла, а не до его возникновения. Это может негативно сказываться на точности прогнозирования и возможностях удержания сотрудников.

Таким образом, актуальность темы исследования обусловлена необходимостью разработки модели, обеспечивающей эффективное использование данных о сотрудниках для поддержки принятия эффективных и своевременных управленческих решений.

Целью данной работы является исследование и разработка моделей и алгоритмов прогнозирования текучести кадров в ИТ-компании, обеспечивающих повышение эффективности принимаемых решений по управлению персоналом последней.

Объектом исследования является процесс управления персоналом ИТ-компании.

Предметом исследования является прогнозирование текучести кадров в ИТ-компании.

Задачи исследования:

- проанализировать современное состояние исследований в области прогнозирования текучести кадров в ИТ-компаниях;
- проанализировать методы прогнозирования текучести кадров в

коммерческих организациях;

- разработать модели и алгоритмы прогнозирования текучести кадров в ИТ-компании;
- выполнить апробацию проектных решений и оценить их эффективность.

В рамках представленной работы была сформулирована следующая гипотеза: эффективность принятия управленческих решений можно повысить, если использовать технологию машинного обучения для прогнозирования текучести кадров.

Методы исследования. В процессе исследования были использованы такие методы как теоретические при анализе научной литературы по рассматриваемой теме, выявлении проблемы исследования и приведении методологии разработки модели, эмпирические при апробации разработанной модели и сборе результатов экспериментов и статистические при проверке возможности применения разработанной модели посредством проверки выдвинутых статистических гипотез.

Научная новизна исследования:

- в данном исследовании был определен рекомендуемый метод оценки модели, заключающийся в использовании таких метрик, как accuracy, precision, recall, f1-score, AUC, а также матрицы ошибок в виду обнаруженной некорректной оценки моделей, разработанных другими авторами;
- разработана и апробирована модель для решения проблемы текучести кадров с оценкой возможности ее практического использования.

Практическая значимость состоит в том, что результаты работы могут быть применены в любых компаниях, где есть потребность в предсказании текучести кадров. Особенно это актуально для ИТ-компаний, где конкуренция за квалифицированных специалистов высока и потеря данного специалиста обойдется компании дороже, чем найм нового сотрудника. Также результаты

работы могут быть применены для улучшения условий труда за счет приведения важности признаков.

Теоретической основой диссертационного исследования являются научные работы российских и зарубежных ученых, занимающихся проблемами текучести кадров и управления персоналом.

Основные этапы исследования: исследование проводилось с 2022 по 2024 год в несколько этапов.

На первом (констатирующем) этапе формулировалась тема исследования, выполнялся сбор информации по теме исследования из различных источников, проводилась формулировка гипотезы, определялись постановка цели, задач, предмета исследования, объекта исследования и выполнялось определение проблематики данного исследования.

Второй этап (поисковый этап). В ходе проведения данного этапа была приведена методология работы с технологией машинного обучения, также были произведены анализ и подготовка данных, реализация возможных алгоритмов машинного обучения и выбор наиболее подходящего алгоритма машинного обучения.

На третьем этапе была апробирована предложенная модель и оценена ее эффективность, а также были сделаны выводы о полученных результатах по проведенному исследованию.

На защиту выносится:

- модель и алгоритмы прогнозирования текучести кадров ИТ-компании;
- результаты апробации проектных решений и оценка их эффективности.

По теме исследования опубликована 1 статья.

Бикмаев Э. М. Применение hr-аналитики для управления персоналом ит-компании // Сборник материалов IX Международной научно-практической конференции (школы-семинара) молодых ученых. Тольятти, 2023. С. 229-234.

Диссертация состоит из введения, четырех глав, заключения и списка

литературы.

Во введении обоснована актуальность темы исследования, представлены цель, объект, предмет, задачи и положения, выносимые на защиту диссертации.

В первой главе приведено ознакомление с существующей научной литературой и исследованиями по схожей теме.

Во второй главе приведена методология работы с технологией машинного обучения.

Третья глава посвящена анализу и подготовке данных для реализации алгоритмов, а также реализации возможных алгоритмов машинного обучения и выбору наиболее подходящего алгоритма машинного обучения.

В четвертой главе выполнены апробация предложенной модели и оценка ее эффективности.

В заключении приводятся результаты исследования.

Данная работа содержит 84 страниц, включая приложение А, 29 таблиц, 31 рисунок, 48 источников.

Глава 1 Анализ современного состояния исследований в области прогнозирования текучести кадров в ИТ-компаниях

1.1 Анализ научной литературы по теме применения прогнозирования в отношении увольнения сотрудников

В современной рыночной среде риски являются неотъемлемой частью жизни любой компании, поэтому факт конкуренции в нашем мире является неоспоримым, и мы, как и все остальные, сталкиваемся с этим явлением каждый день. Это особо хорошо заметно на примере компаний, ведь они существуют в конкурентной среде. Они ежедневно сталкиваются с множеством воздействующих факторов, оказывающих различное влияние на их деятельность. Устойчивость компании напрямую зависит от способности её руководства анализировать и контролировать как внутренние, так и внешние факторы. Неизбежно, разнообразные факторы, с которыми сталкивается компания, могут привести к возникновению негативных рисков.

Так, например, внутренние риски могут привести к увольнению ключевого сотрудника, а это, в свою очередь, может обернуться экономической проблемой для всей компании. Внешний риск, связанный с усилением конкуренции, чреват финансовыми проблемами, вследствие чего возможен крах компании [24]. Поэтому так важно уделить пристальное внимание прогнозированию как средству сведения рисков к минимуму.

Прогнозирование – это процесс предвосхищения результатов на основе прошлых и настоящих данных, то есть извлечение новой, ранее неизвестной информации.

Обычно под внутренними рисками, связанными с увольнением сотрудников, понимают:

- увольнение по собственному желанию;
- увольнение по инициативе работодателя.

Данная проблема как никогда актуальна, так как на сегодняшний день

современные компании часто сталкиваются с ней, что является преградой для их роста. Также не стоит забывать о положительных сторонах этой проблемы, но эти положительные стороны скорее являются негативными [34].

В данной работе нас интересуют внутренние риски, связанные с увольнением сотрудника по собственному желанию, так как это грозит серьёзными рисками для компании. Возможность компании спрогнозировать будущее даёт ей стимул для развития. Благодаря тому, что риски, связанные с текучестью персонала, находятся под контролем, компания обретает уверенность в завтрашнем дне, а значит, будет вкладываться в своих сотрудников, не боясь риска их скорого увольнения.

Следовательно, выявляется проблема увольнения сотрудников, которая является актуальной и достойной дальнейшего рассмотрения.

В статье [28] говорится о важности кадровой безопасности, так как она является основой для экономического развития предприятия.

Кадровая безопасность – процесс снижения вероятности возникновения нежелательных сценариев, связанных с персоналом и, соответственно, экономической безопасностью компании.

Сотрудник может продолжать, как и раньше ходить на работу, но никто даже не заподозрит, что он уже принял решение об уходе из компании. Это, в свою очередь, обостряет риски, связанные с кадровой безопасностью, такие как хищения, незаконное сотрудничество с конкурентами, разглашение коммерческой тайны, мошенничество, коррупция, нанесение ущерба бизнесу. Стоит отметить немаловажный факт, что процесс подбора и найма персонала является начальным этапом кадровой безопасности.

В статье [28] Троценко В.М. также отмечает, что в России наблюдается проблема с принятием управленческих решений, что не позволяет должным образом управлять ресурсами компании, что, в свою очередь, бьёт по эффективности и обеспечению стабильного роста данной компании.

Следовательно, раннее выявление риска увольнения сотрудника позволит снизить связанные с этим риски, а правильный подход к возникшей

ситуации повысит экономическую эффективность всей компании.

Важность кадровой безопасности также раскрывается в статьях [6], [10], [18], [9], [17].

Например, в статье [6] персонал рассматривается как стратегический ресурс любой компании, как фундамент. Именно эта «основа» обеспечивает такие виды безопасности, как:

– кадровая;
– информационная;
– экономическая;
– материально-техническая.

Особую актуальность кадровая безопасность возымела в государственных структурах.

Вантеева В. В. в статье [6] приводит исследование показывающее, что угрозы физическому имуществу компании в основном исходят от собственного персонала – около 80%, тогда как только 20% извне.

Поэтому становится так важно дать вовремя оценку кадровым рискам, так как она обеспечит будущую безопасность. В статье также говорится о важности создания инструментов для минимизации этих рисков.

В работе [12] приводится статистика, проанализировав которую становится ясно, что для предотвращения риска, связанного с утечкой информации, необходимо преждевременно выявить эту тенденцию. Автор указывает, что есть период, который является особенно опасным с точки зрения компрометации информации, а именно месяц перед увольнением, а не за неделю до него.

Следовательно, выявляется потребность в прогнозировании деятельности компании, а именно её персонала, что позволило бы минимизировать связанные с этим риски.

На сегодняшний день становится актуальным такой HR-тренд, как HR-аналитика. Это не удивительно, ведь данный вид аналитики активно используется в сфере управления персоналом.

HR-аналитика – это методы по повышению финансовой составляющей компании благодаря её персоналу.

В статье [31] автор приводит статистику, где наглядно показаны области применения HR-аналитики. Изобразим данную статистику на рисунке 1.

Рисунок 1 – Области применения HR-аналитики

Приведенная статистика показывает, что области применения HR-аналитики обширны. Она активно применяется в подборе и найме персонала, а значит, применяя данный вид аналитики, можно превентировать ошибки на этапе подбора и найма персонала, следовательно, снизить риски, связанные с кадровой безопасностью. Также этот вид аналитики используется в прогнозировании текучести и удержании персонала, что также благотворно скажется на самой компании в перспективе, как в плане безопасности, так и эффективности, ведь HR-аналитику используют и в управлении эффективностью.

В статье [32] раскрывается важность безопасности хранения данных и отношение сотрудников к сбору персональных данных, однако в данной работе вопросы безопасности HR-аналитики рассматриваться не будут, так как в данной исследовательской работе будут применяться

анонимизированные персональные данные.

Анонимизированные персональные данные – это информация о людях, которая была обработана или изменена таким образом, чтобы идентифицировать конкретное лицо было невозможно или чрезвычайно затруднено. Это позволяет использовать данные для исследований или анализа, не раскрывая личную информацию о конкретных индивидах.

В статье [19] Максимова К. А. приводит исследование, в котором участвовало более 1000 руководителей HR-отдела с разных стран. В этом исследовании было выявлено, что HR-аналитика на сегодняшний день является перспективной технологией, достойной того, чтобы в краткосрочной перспективе увеличить вложения в неё.

Такие большие компании, как:

– ПАО «Газпром нефть»;
– ПАО «МегаФон»;
– Xerox;
– Deloitte.

Уже используют HR-аналитику для решения своих проблем и повышения эффективности управленческих решений.

Автор статьи акцентирует внимание на том факте, что компании, причём российские, в основном держатся за традиционные методы управления персоналом, которые заключаются в решении задач здесь и сейчас и не занимаются составлением прогнозов на перспективу. Однако, как ни странно, сегодня, во время неустойчивого развития экономической составляющей бизнеса, это направление аналитики становится так популярно.

Следовательно, именно сейчас для бизнеса становится так важно знать, куда в перспективе придёт компания, а использование HR-аналитики для построения прогнозов сможет дать ответы на столь важные вопросы. Но чтобы достичь этого, необходимо определиться с признаками, на которые можно было бы опираться при принятии того или иного решения, коими являются измеримые индивидуальные черты личности.

HR-аналитика на сегодняшний день используется не только в HR, но и в других областях. В статьях [27], [33], [31] авторы провели исследование, благодаря чему были выявлены перспективные направления использования результатов HR-аналитики. Одной из таких областей является искусственный интеллект, в частности, его подраздел машинное обучение.

Искусственный интеллект (ИИ) – это термин, который вобрал в себя всё то, что является моделированием человеческого интеллекта с помощью компьютерных систем.

Машинное обучение (МО) – это одна из областей искусственного интеллекта, которая заключается в построении математических моделей на основе вычислительных алгоритмов, которые в дальнейшем могут обучаться и выявлять закономерности в данных.

В HR-аналитике выделяют 4 вида аналитики и начиная с предиктивной аналитики средством для достижения цели становится МО [27]. В таблице 1 приведены цели каждого вида аналитики, а также сферы их применения и методы анализа.

Таблица 1 – Виды HR-аналитики

Название	Цель	Сферы применения	Методы анализа
Дескриптивная аналитика	Выявление и мониторинг проблем и их диагностика на основе данных. Отвечает на вопрос «Что произошло?»	Структурированные данные и отчетность, например: – структура персонала; – нормы труда и нормативы численности; – обзоры заработных плат; – метрики эффективности процессов; – бенчмарки. Dashboard	Описательная статистика

Продолжение таблицы 1

Название	Цель	Сферы применения	Методы анализа
Прогнозная аналитика	Прогнозирование на основе подтвержденных статистических гипотез. Ответ на вопрос «Почему и Как»	– методы прогнозирования численности; – планирование загрузки; – формирование профиля успешного сотрудника; – план мероприятий по повышению вовлеченности сотрудников; – определение норм прохождения тестов.	– корреляционно регрессионный анализ; – кластерный анализ.
Предиктивная аналитика	Прогнозирование на основе выявления неочевидных зависимостей и мультивариантноти сценариев. Отвечает на вопрос «Что будет в будущем?»	– прогнозирование увольнения сотрудника; – валидизация модели компетенции; – прогнозирование успешности сотрудника в конкретной среде.	– машинное обучение; – деревья решений.
Прескриптивная аналитика	Предложение решений в динамической среде. Ответ на вопрос «Как изменить/улучшить то, что будет в будущем?»	Формирование решений для снижения вероятности увольнения конкретного сотрудника	Искусственные нейронные сети

В данном исследовании нас будет интересовать именно предиктивная аналитика, ведь именно она предоставляет возможность прогнозировать увольнение сотрудника, где средством служит МО.

Таким образом, было выяснено, что с потерей ключевого сотрудника обостряются риски, связанные с кадровой безопасностью, а также внутренние и внешние. Прогнозирование же дает возможность взять под контроль данные риски, что поспособствует тому, что у компании появится уверенность в завтрашнем дне, что позволит ей вкладываться в сотрудника, не боясь риска его скорого увольнения, предиктивная аналитика, использующая технологию МО как раз таки направлена на это. Детальнее эта технология и сферы её

применения будут рассматриваться в следующей главе.

1.2 Машинное обучение

Машинное обучение, как уже было упомянуто выше, является областью искусственного интеллекта. Интерес к искусственному интеллекту в последнее время возрастает.

Благодаря веб-сайту Google Trends, который был разработан корпорацией Google, можно отследить динамику популярности того или иного поискового запроса за какой-то период.

Отследим популярность запроса «искусственный интеллект» за последние 5 лет в России и по всему миру, проиллюстрировав это на рисунках 2 и 3.

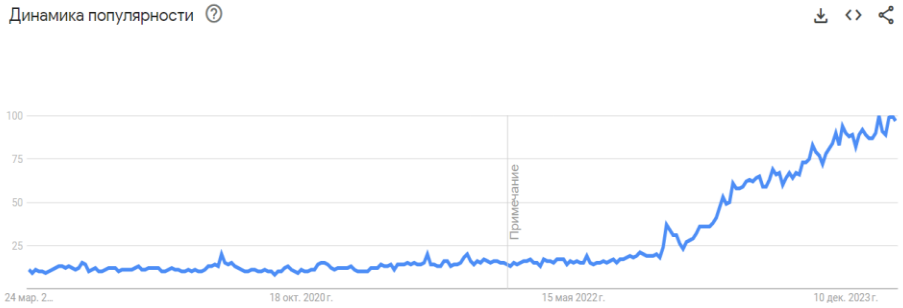


Рисунок 2 – Динамика популярности поискового запроса «искусственный интеллект» в России

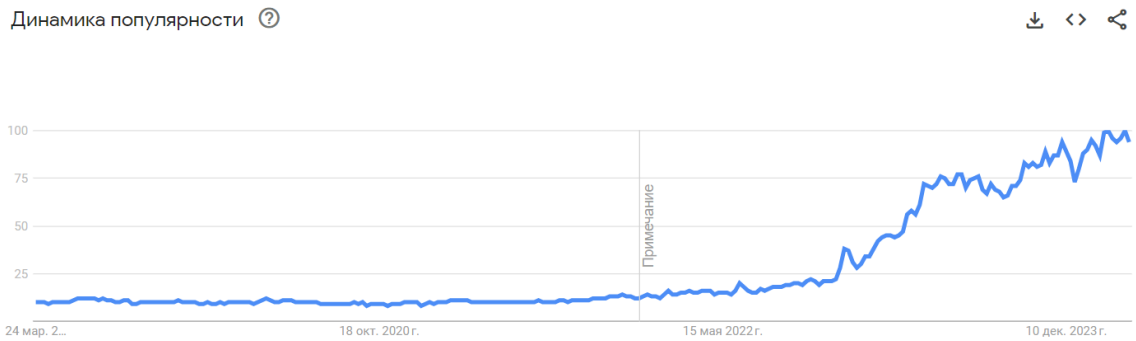


Рисунок 3 – Динамика популярности поискового запроса «искусственный интеллект» по всему миру

Числа от 0 до 100 показывают уровень интереса к теме, где 100 – высокий уровень популярности запроса, а 0 – низкий.

Как видим, популярность поискового запроса довольно высока.

Существует несколько методов машинного обучения. Сведем их в таблицу 2.

Таблица 2 – Основные методы машинного обучения

Метод	Описание
Обучение с учителем	Модель обучается на размеченных данных, где каждый пример имеет соответствующую метку или класс. Примеры алгоритмов: линейная регрессия, метод опорных векторов, случайный лес
Обучение без учителя	Модель обучается на неразмеченных данных, и сама находит внутренние закономерности и структуры. Примеры методов: кластеризация, глубокого обучения (нейронные сети)
Обучение с подкреплением	Модель обучается на основе взаимодействия с окружающей средой и получает обратную связь в виде награды или штрафа. Этот метод используется в задачах обучения агентов для принятия решений

Каждый метод обучения имеет свои особенности и применяются в зависимости от конкретной задачи и доступных данных. В данной работе нас будет интересовать обучение с учителем, так как предполагается, что компания уже имеет какие-то данные о сотруднике, и в случае его увольнения в столбце «увольнение» будет цифра 1, а у сотрудников, которые к данному времени не уволились, будет цифра 0.

Представим основные задачи решаемые МО в таблице 3.

Таблица 3 – Основные задачи машинного обучения

Задача	Описание
Классификация	Прогнозирование принадлежности объекта к одной из заранее определенных классов. Например, классификация электронной почты: спам и не спам

Продолжение таблицы 3

Задача	Описание
Регрессия	Предсказание количественного значения. Например, предсказание цены недвижимости
Кластеризация	Группировка объектов таким образом, чтобы объекты в одном кластере(классе) были более похожи друг на друга, чем на объекты в других кластерах за счет нахождения скрытых паттернов или групп. Например, сегментация персонала

Исследуем проявление машинного обучения в различных отраслях.

В работах [47], [45] машинное обучение показывает высокую точность прогнозов, что положительно сказывается на диагностике, классификации и прогнозировании выживаемости пациентов. В приведенной работе МО сравнивали с традиционными статистическими методами и выяснилось, что МО обладает большей гибкостью и масштабируемостью. В целом, МО помогает врачам принимать более информированные решения, основанные на анализе медицинских данных, что улучшает качество лечения пациентов.

В работах [39], [23] МО себя успешно зарекомендовало в финансах, прогнозируя рынки, управляя рисками и обнаруживая мошенничество.

В работе [38] были выявлены значительные преимущества применения машинного обучения в промышленном производстве. Некоторые из основных выгод включают большее количество инноваций, оптимизация процессов и ресурсов, а также улучшение качества продукции.

В статьях [35], [29] МО используется в образовании. Оно может адаптироваться к индивидуальным потребностям учащихся, предлагая персонализированные материалы и методы обучения, может эффективно оценивать знания студентов, обеспечивая быструю и объективную обратную связь, а также помогать в сокращении времени и затрат на достижение образовательных целей.

Таким образом, становится очевидно, что МО вносит значительный вклад в развитие различных отраслей, повышая их эффективность и конкурентоспособность, а также способствует общему развитию общества.

Несмотря на большие преимущества данной технологии, ей присущи и недостатки. Так, например, в работе [5] авторы М.В. Боброва и А.Е. Мастилин обсуждают проблемы и возможности использования МО в области кибербезопасности. Они отмечают, что хотя машинное обучение может помочь в выявлении аномалий, подозрительного поведения и уязвимостей, его применение также может создавать новые проблемы, такие как потеря рабочих мест и конфликты с правовой базой. Авторы подчеркивают, что необходимо помнить, что машинное обучение не является универсальным решением имеет свои ограничения и недостатки. Также отмечается, что внедрение систем, основанных на машинном обучении, требует особой осторожности и внимания к возможным последствиям.

В статье [8] перечислены некоторые проблемы, с которыми компании сталкиваются при внедрении машинного обучения. Вот некоторые из них:

- дефицит специалистов в области машинного обучения, что создает проблемы при поиске квалифицированных кадров;
- проблемы с обработкой данных и нехватка информации, так как некоторые компании сталкиваются с ограниченным объемом данных, что может привести к неэффективности проектов машинного обучения;
- сомнения в полезности и применимости методов машинного обучения к конкретному бизнесу и потребителям, что может вызвать неопределенность и затруднения в принятии решений.

Существуют и другие недостатки МО. Так, автор статьи [21] указывает на две основные проблемы методов машинного обучения: недообучение и переобучение моделей. Недообучение возникает, когда модель не может правильно обработать обучающие данные, а переобучение происходит, когда модель слишком точно подстраивается под обучающие данные, что затрудняет обобщение на новые данные, то есть:

- недообученная модель будет показывать плохие результаты как на данных обучения, так и на новых данных;

– переобученная модель демонстрирует очень низкую ошибку на обучающих данных, но плохо справляется с новыми данными.

Вышеописанные проблемы могут решаться следующим образом:

– правильно, что МО не является универсальным решением и может иметь свои ограничения и недостатки. Это говорит о том, что не следует бездумно использовать прогнозы МО, разумнее было бы сделать его своим помощником в принятии тех или иных решений;
– приведенная статистика популярности искусственного интеллекта говорит о том, что проблема дефицита специалистов вскоре будет решена;
– да, действительно, использование МО подразумевает наличие достаточного количество данных для более точных прогнозов. Ввиду этого технология МО для прогнозирования текучести кадров не будет доступна маленьким компаниям с небольшим коллективом, но и затраты она несет не столь большие в сравнении с большими компаниями. Однако, обучив модель однажды на нужном количестве данных, ее можно использовать и в маленькой компании, но появляется жесткая привязка к признакам, на которых обучалась модель;
– для решения проблемы сомнений в полезности и применимости методов машинного обучения к бизнесу и потребителям компании могут проводить обучающие мероприятия, запускать пилотные проекты, обращаться к консультантам, изучать опыт конкурентов, а также разрабатывать бизнес-кейсы для обоснования внедрения технологии. Очевидно, компания, использующая МО, будет развиваться эффективнее, и анализ научных работ показал это;
– проблема недообучения и переобучение моделей возникает ввиду плохо нормализованных данных (наличие шумов и искажений во входных данных) или в случае недостаточного количества данных для обучения.

Таким образом, в любой технологии есть свои плюсы и минусы, и МО в данном случае не является исключением. Но, несмотря на недостатки, которые не столь критичны, эту технологию продолжают использовать как большие, так и средние компании. Ввиду этого популярность технологии в последнее время растет.

Анализ [4], [11], [1], [26], [3], [16], [12], [25], [30], [20] статей позволил узнать, как МО показывает себя в прогнозировании оттока кадров и не только, и определиться с наиболее востребованными алгоритмами для задачи прогнозирования оттока сотрудников. В перечисленных работах именно они по сравнению с другими алгоритмами показали высочайшую точность.

Список данных алгоритмов:

– случайный лес;
– дерево решений;
– градиентный бустинг;
– логистическая регрессия.

Например, статья [4] посвящена использованию библиотек языка программирования Python для анализа оттока клиентов банка. В статье были использованы пять алгоритмов для расчета оттока клиентов: логистическая регрессия, наивный Байес, деревья решений, алгоритм ближайшего соседа и случайный лес. Из проведенного исследования выяснилось, что наилучший результат показал алгоритм случайный лес, достигнув точности предсказания на уровне 87,1%. Другой алгоритм, дерево решений, также показал хороший результат - 86%, но случайный лес оказался наиболее эффективным для решения задачи анализа оттока клиентов банка.

В другой работе [26] авторы провели сравнительный анализ классификационных алгоритмов для прогнозирования оттока кадрового состава в системе управления предприятием. Исследование акцентируется на актуальной проблеме текучести кадров в компаниях, где замена сотрудника обходится дороже, чем удержание текущего сотрудника. В ходе исследования были использованы различные алгоритмы МО, такие как случайный лес,

дерево принятия решений, логистическая регрессия, метод машинно-опорных векторов, наивный байесовский классификатор, метод к-ближайших соседей и градиентный бустинг. Особое внимание уделено логистической регрессии, которая показала хорошие результаты с точностью классификации 80.78%, затем идут градиентный бустинг (80%) и случайный лес (78%).

В работе [12] автор исследования провел анализ различных алгоритмов, что позволило выбрать наиболее оптимальный алгоритм для решения задачи прогнозирования оттока сотрудников. Из результатов исследования следует, что градиентный бустинг показал наилучшие показатели. Хотя использование дерева решений также продемонстрировало хорошие результаты (97%), градиентный бустинг (97%) оказался более гибким, менее подверженным переобучению и лучше работал "из коробки". Поэтому автор рекомендует использовать градиентный бустинг как наиболее оптимальный алгоритм для прогнозирования оттока сотрудников. Данное исследование позволило не только определить вероятность ухода сотрудника, но и предоставило возможность использовать обученную модель в HRM-системах для улучшения управления персоналом и снижения текучести кадров.

Таким образом, прогнозирование также способно помочь в ситуации, когда замена сотрудника обходится дороже, чем удержание текущего сотрудника. В этом случае компании важно знать, что позволит ей удержать данного сотрудника, а позволит ей это сделать предиктивный анализ данных сотрудника и последующее выявление фактора, который больше всего влияет на его решение уйти из компании. Выявляется средство для решения проблемы текучести кадров, коим является предиктивная аналитика, использующая технологию МО. Данная технология при правильном использовании данных способна более эффективно использовать данные HR-аналитики для удержания сотрудников, предоставляя прогнозы на основе данных о сотрудниках, которые можно использовать для принятия более эффективных управленческих решений.

1.3 Обзор алгоритмов машинного обучения

Для начала стоит разобраться с понятиями ансамбль, бэггинг, бустинг.

Ансамбль – это группа предсказателей, объединенных вместе для достижения более точного результата, чем у одиночного предсказателя. Это подход, при котором несколько предсказателей используется для предсказания одной переменной, и их прогнозы затем комбинируются для получения окончательного ответа.

Техники ансамблирования классифицируются на bagging и boosting.

Bagging – это техника построения независимых моделей, результаты которых комбинируются, используя, например, голосование большинства или усреднения.

Обычно генерируются случайные подмножества данных из исходной набора данных для каждой модели. Выборка происходит с возвращением, что означает, что объекты могут быть выбраны несколько раз или не выбраны вовсе. В итоге строится множество плохо коррелирующих моделей для последующего построения итоговой модели. Случайный лес в основе своей использует данную технику.

Способ создания различных подмножеств данных называется bootstrap. Например, если есть тысячи объектов, генерируется случайное число в диапазоне от 0 до 1000. Предположим, что получили число 134, тогда включаем объект с индексом 134 в обучающий набор. Затем снова генерируем случайное число в том же диапазоне и, например, получаем число 134 снова. Хотя это может показаться повторением, это не проблема, так как будет иметься столько же объектов, сколько было в исходной выборке. Если исходная выборка содержала 1000 объектов, также выбираем 1000 объектов для каждого дерева.

Boosting – это техника, при которой каждый последующий предсказатель строится таким образом, чтобы исправлять ошибки предыдущих. В отличие от bagging, где модели построены не зависимо,

boosting использует последовательность. Пример алгоритма, который использует данную технику, – градиентный бустинг.

Перейдем к обзору алгоритмов.

Градиентный бустинг – использует в основе своей технику boosting, то есть строится сложная модель прогнозирования путем комбинирования нескольких слабых моделей, основанных главным образом на деревьях решений. Последовательное применение этих слабых моделей позволяет минимизировать ошибку, которую каждая последующая модель исправляет по отношению к предыдущей. Таким образом, создается одна мощная модель, обладающая высокой эффективностью.

Для наглядного примера приведем график, изображенный на рисунке 4.

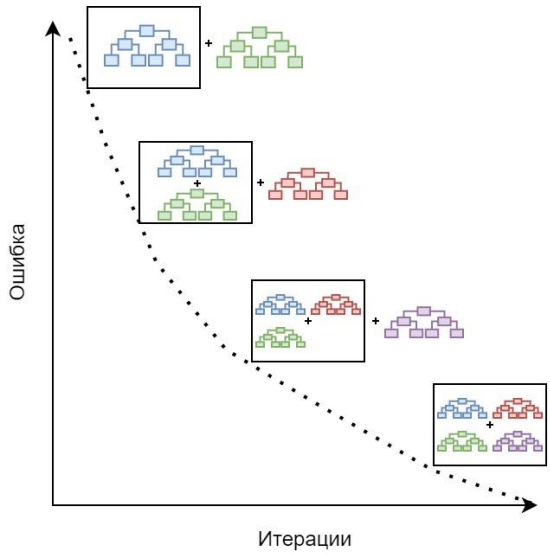


Рисунок 4 – График алгоритма градиентный бустинг

Градиентный бустинг обеспечивает высокую точность прогнозирования и способен обрабатывать различные типы данных и шумы в данных. Однако он также может быть подвержен переобучению, поэтому важно подобрать подходящие гиперпараметры и использовать методы регуляризации для контроля сложности модели.

Деревья решений – это алгоритм МО, который использует структуру

дерева для принятия решений на основе входных данных. Пример схемы алгоритма показан на рисунке 5.

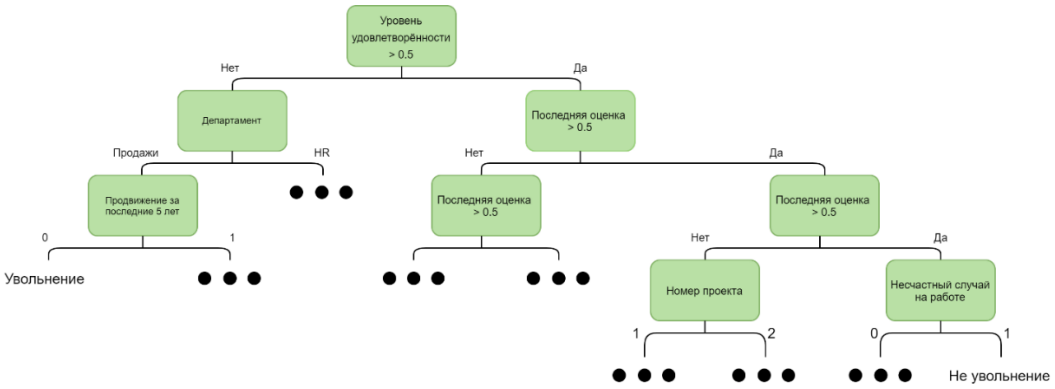


Рисунок 5 – Схема алгоритма дерева решений

Он состоит из серии вопросов, разбивающих данные на подгруппы и определяющих следующий шаг алгоритма. Алгоритм начинается с корневого узла, который содержит весь набор данных. Затем происходит разбиение данных на основе выбранного признака, который наилучшим образом разделяет данные по классам или прогнозируемой переменной. Процесс выбора признака и разбиения продолжается на каждом следующем уровне дерева, пока не достигнут критерии остановки, такие как достаточное количество примеров в каждом листовом узле или достижение определенной глубины дерева. Во время построения дерева решений используются различные метрики, такие как индекс Джини или энтропия, для определения оптимального разбиения данных на каждом шаге. Суть состоит в том, чтобы минимизировать неоднородность данных в каждом поддереве, чтобы получить более точные прогнозы. После построения дерева решений, оно может быть использовано для классификации новых данных или предсказания значения целевой переменной. Это происходит путем прохождения по дереву от корневого узла до соответствующего листового узла на основе значений признаков новых данных.

Алгоритм логистической регрессии основан на логистической функции,

также известной как сигмоида, которая преобразует линейную комбинацию признаков объекта в вероятность его принадлежности к одному из классов. График алгоритма изображен на рисунке 6.

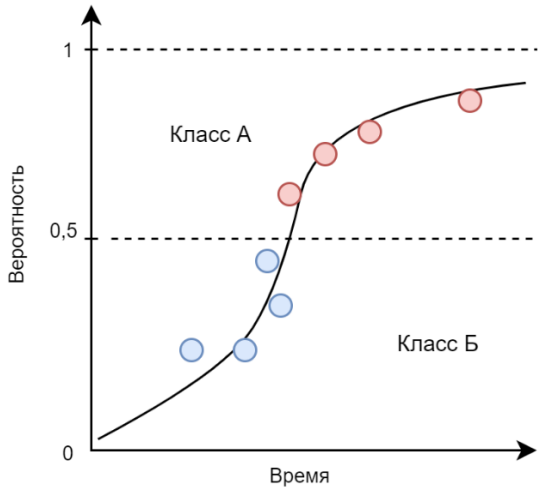


Рисунок 6 – График алгоритма логистической регрессии

Процесс обучения логистической регрессии включает подгонку модели к обучающим данным с использованием метода максимального правдоподобия или метода градиентного спуска. Веса модели оптимизируются таким образом, чтобы максимизировать вероятность правильной классификации обучающих примеров. Во время предсказания логистическая регрессия использует полученные веса для вычисления вероятности принадлежности объекта к каждому классу. В случае бинарной классификации, объект относится к классу, вероятность которого выше некоторого порога.

Случайный лес – это алгоритм машинного обучения, использующий технику bagging с bootstrap. В результате получается большое количество разнообразных деревьев, каждое из которых может быть немного переобученным, но в совокупности они дают хорошие результаты. Благодаря большому количеству деревьев и усреднению их предсказаний, ошибки, которые могут возникнуть в отдельных деревьях, компенсируются, и общая

модель становится более надежной и точной. Визуализацию графика алгоритма представим на рисунке 7.

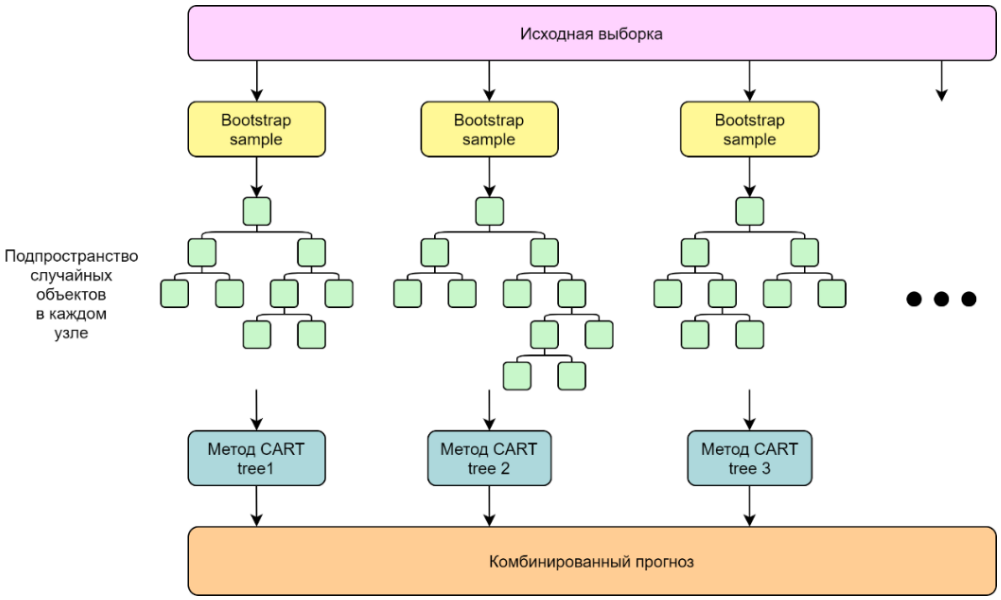


Рисунок 7 – Схема алгоритма случайного леса

На представленной визуальной иллюстрации показан процесс обучения дерева принятия решений. Исходная выборка данных используется для создания обучающих наборов с помощью процедуры bootstrap, а далее идет метод CART.

Метод CART (classification and regression trees - дерево классификации и регрессии) – это алгоритм, основанный на разбиении данных на более мелкие подгруппы с помощью определенных признаков. Алгоритм начинается с создания корневого узла, который представляет собой весь набор данных. Затем выбирается признак и пороговое значение, которые наилучшим образом разделяют данные на две подгруппы. Подгруппы создаются путем разбиения данных на основе выбранного признака и значения. Процесс разбиения повторяется для каждой созданной подгруппы, пока не будет выполнено определенное условие остановки. Условие остановки может быть достижение максимальной глубины дерева, минимального количества объектов в узле или других критериев. Когда достигнуто условие остановки, в каждом листовом

узле дерева принимается решение о классификации.

Когда есть множество предсказаний от каждого дерева выбирается предсказание с наибольшим количеством голосов. Получается усредненный результат, который основан на голосовании от множества деревьев.

Таким образом, были проанализированы различные научные публикации, связанные с задачами прогнозирования текучести кадров, а также смежные работы. Основываясь на полученных данных, были выделены наиболее часто упоминаемые алгоритмы машинного обучения. Итоговым результатом выбора стали 4 алгоритма, которые будут в дальнейшем реализованы.

Выводы по главе 1

В ходе выполнения данной главы были изучены научные работы, исследована литература по теме прогнозирования текучести кадров.

Было определено, что HR-аналитика, используя предиктивную аналитику вкупе с МО, которые зарекомендовали себя в различных сферах, в том числе в прогнозировании текучести кадров, может повысить эффективность принятия управленческих решений.

Были проанализированы различные виды МО, благодаря чему был выбран вид МО - обучение с учителем, который будет использоваться в исследовании, а также задачи, решаемые МО. В данном исследовании будет решаться задача классификации.

Глава 2 Анализ методов прогнозирования текучести кадров в коммерческих организациях

2.1 Инструментарий и предварительная обработка данных

Сначала необходимо определиться с инструментами. От правильно подобранных инструментов зависит конечный результат исследования. На данный момент во многих исследованиях по разработке моделей МО используются следующие инструменты:

- Python – язык программирования для написания программного кода;
- Pandas – библиотека, предназначенная для обработки и анализа данных;
- Numpy – библиотека, которая обеспечивает возможность работы с многомерными массивами и матрицами;
- Sklearn (Scikit-learn) – библиотека МО, предоставляющая алгоритмы и инструменты для анализа данных и построения моделей МО;
- Matplotlib – библиотека визуализации данных, позволяющая создавать разнообразные графики и диаграммы;
- Google Colab – представляет собой веб-сервис от Google, предназначенный для написания и выполнения кода на языке Python.

Python был выбран ввиду того, что он являются бесплатными, а также имеет необходимые библиотеки, например, sklearn, которая предоставляет необходимые алгоритмы для работы с МО.

Перейдем к анализу и нормализации данных.

Для создания модели потребуются данные о сотрудниках. Необходимо осуществить поиск для нахождения этих данных, и перед их использованием для обучения и тестирования алгоритма необходимо проверь их качество. В ходе этого процесса произведем непосредственно сам анализ данных и осуществим проверку на наличие пропущенных значений, аномалий и других ошибок в данных.

Нормализация включает в себя масштабирование и преобразование категориальных признаков в числовые значения.

Масштабирование – это процесс приведения значений признаков к одному и тому же диапазону. Это позволяет модели более эффективно использовать признаки и избежать возможных искажений результатов, которые могли бы возникнуть из-за различных масштабов признаков.

Также необходимо учитывать наличие категориальных признаков. Категориальные переменные представляют количественные данные, записанные в виде текста. Если в данных присутствуют категориальные переменные, их необходимо преобразовать в числовые значения, поскольку большинство алгоритмов МО принимают только числовые данные.

Предварительная обратка данных также включает в себя разделение данных на две части: обучающий набор данных и тестовый набор данных. Обучающий набор данных используется для обучения модели, а тестовый набор данных используется для оценки качества модели на новых, ранее не виденных данных. Обычно данные разделяют в соотношении, например, 80% для обучения и 20% для тестирования. Обучение модели включает передачу обучающих данных в модель и настройку ее параметров с использованием алгоритма обучения. Модель анализирует обучающие данные и «обучается» находить определенные закономерности или паттерны в данных [11].

2.2 Оценка модели

Оценка модели является неотъемлемой частью процесса разработки и применения моделей МО. Оценка модели позволяет оценить точность и эффективность модели. Для этого необходимо провести анализ различных метрик, чтобы получить объективное представление о качестве модели. Такая оценка помогает сделать обоснованные выводы и принять решения на основе результатов моделирования. Оценка модели также помогает сравнивать различные модели между собой, чтобы выбрать наиболее подходящую для

конкретной задачи. Кроме того, оценка модели помогает выявить проблемы в ее работе и улучшить качество модели.

В статьях [11], [30], [2], [4], [14], [15], [12] используются следующие метрики качества для оценки эффективности моделей прогнозирования:

– accuracy (правильность);
– precision (точность);
– recall (полнота);
– f1-score (f1-мера);
– AUC (area under curve – площадь под кривой).

Согласно проведенным исследованиям, эти метрики могут быть применены для оценки моделей прогнозирования текучести кадров. Приведем определение и описание каждой метрике, что поможет лучше понять их смысл.

Прежде чем перейти к объяснению метрик, важно познакомиться с концепцией. Эта концепция известна как матрица ошибок или матрица путаницы (confusion matrix).

Допустим, имеется два класса, и есть алгоритм, который предсказывает принадлежность каждого объекта к одному из этих классов. В результате получается матрица ошибок, которая отображает следующую информацию. В таблице 4 представлена матрица ошибок.

Таблица 4 – Матрица ошибок

Факт / Предсказание	Предсказанное не увольнение	Предсказанное увольнение
Фактическое не увольнение	Истинно позитивные (TP)	Ошибочно Позитивные (FP)
Фактическое увольнение	Ошибочно Негативные (FN)	Истинно Негативные (TN)

В данной матрице ошибок есть две фактические категории, а именно «увольнение» и «не увольнение», и две предсказанные категории

«увольнение» и «не увольнение».

- в ячейке TP (True Positive) записывается количество случаев, когда модель правильно предсказала, что сотрудник не будет уволен;
- в ячейке FN (False Negative) записывается количество случаев, когда модель ошибочно предсказала, что сотрудник не будет уволен, в то время как он фактически уволился;
- в ячейке FP (False Positive) записывается количество случаев, когда модель ошибочно предсказала, что сотрудник будет уволен, в то время как он фактически не уволился;
- в ячейке TN (True Negative) записывается количество случаев, когда модель правильно предсказала увольнение сотрудника.

Таким образом, матрица ошибок помогает наглядно представить, как модель классифицирует объекты и какие ошибки она делает. Это важный инструмент для анализа и оценки производительности модели.

После ознакомления с матрицей ошибок можно перейти к описанию метрик оценки модели.

2.4.1 Метрика «правильность»

Оценка модели с использованием данной метрики является одним из распространенных подходов при оценке эффективности моделей прогнозирования текучести кадров. Эта метрика измеряет долю правильных прогнозов модели относительно общего числа прогнозов.

Например, есть 100 случаев увольнения, и модель правильно предсказала 90 из них (TN = 90, FN = 10), и 10 случаев не увольнения, 5 из которых модель также предсказала правильно (TP = 5, FN = 5), для этой ситуации можем рассчитать значение метрики. Покажем это в формуле (1):

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN}, \quad (1)$$

$$accuracy = \frac{5 + 90}{5 + 90 + 10 + 5} = 86{,}4\%.$$

Однако, следует отметить, что метрика может быть несбалансированной в случае, если классы неравномерно представлены в данных. Например, если посчитать, что все сотрудники будут не уволены (TN = 100, FN= 0 и TP = 0, FN = 10), то показатель точности будет выше. Продемонстрируем это в выражении:

$$accuracy = \frac{0 + 100}{0 + 100 + 0 + 10} = 90{,}9\%$$

При использовании метрики важно учитывать контекст и баланс классов в данных. В случае несбалансированных классов, рекомендуется также рассмотреть другие метрики, такие как точность, полнота, и f1-мера, чтобы получить более полное представление о производительности модели прогнозирования текучести кадров.

2.4.2 Метрики «точность» и «полнота»

Метрики точности и полноты являются более надежными метриками, чем правильность, когда имеются несбалансированные данные. Данные метрики учитывают, насколько хорошо модель обнаруживает положительные примеры и избегает ложных срабатываний. Они не зависят от соотношения классов в данных, поэтому особенно полезны в случаях, когда один класс преобладает над другим.

Способы расчета метрик показаны в формулах (2) и (3).

$$precision = \frac{TP}{TP + FP}, \quad (2)$$

$$recall = \frac{TP}{TP + FN}. \quad (3)$$

Если точность модели высока, это означает, что большинство прогнозов, сделанных моделью как положительные, действительно являются положительными.

Если модель имеет высокую полноту, это означает, что она не пропускает много положительных объектов и способна их правильно идентифицировать.

2.4.3 Метрика f1-мера

F1-мера является метрикой, которая комбинирует точность и полноту в одну единую метрику, то есть f1-мера представляет собой гармоническое среднее между точностью и полнотой. Способ вычисления в общем виде показан в формуле (4).

$$F_\beta = (1 + \beta^2) \cdot \frac{precision \cdot recall}{(\beta^2 \cdot precision) + recall}, \quad (4)$$

где β – коэффициент бета, определяет важность точности по отношению к полноте.

В случае f1-мера, коэффициент бета равен 1, а это означает, что точность и полнота имеют одинаковый вес или важность при вычислении, поэтому формула (4) упрощается до формулы (5).

$$\text{f1 score} = \frac{2 \cdot precision \cdot recall}{precision + recall} \quad (5)$$

Как итог, f1-мера является полезной метрикой для оценки качества модели в случаях, когда важны как точность, так и полнота.

2.4.4 Метрика AUC

AUC является метрикой, которая позволяет оценить качество модели в общем виде, независимо от конкретного выбора порога.

Метрика AUC измеряет площадь под кривой ROC (Receiver Operating

Characteristic - рабочая характеристика приемника), которая представляет собой график, отображающий зависимость между чувствительностью (True Positive Rate) и специфичностью (False Positive Rate) классификатора при изменении порога вероятности принятия решения.

Важно отметить, что метрика AUC, не зависит от конкретного порога и позволяет оценить качество модели в целом, без привязки к конкретному значению порога.

Порог задает, какую вероятность или оценку принимать во внимание при принятии решения о классификации. Если предсказанная вероятность или оценка выше порога, объект будет отнесен к положительному классу, в противном случае - к отрицательному классу. Несмотря на то, что метрика AUC не привязана к конкретному порогу, знание о пороге помогает понять, какие решения будут приниматься на основе предсказаний модели.

Предположим, есть модель МО, которая анализирует данные о сотрудниках и предсказывает вероятность их увольнения. Модель может предсказывать вероятность увольнения для каждого сотрудника, например, от 0 до 1, где 0 - очень маленькая вероятность увольнения, а 1 - очень высокая вероятность увольнения.

Теперь можно выбрать порог, чтобы узнать о том, уволится ли сотрудник или останется на работе. Допустим, установим порог равным 0,5. Это означает, что если предсказанная вероятность увольнения для сотрудника выше 0,5, то сотрудник может уволиться и уже можно принимать какие-то меры по его удержанию, а если вероятность ниже 0,5, значит вероятность увольнения сотрудника крайне низка.

Как это выглядит:

Сотрудник А. Предсказанная вероятность увольнения - 0,6 (выше порога 0,5). Решение: увольнение.

Сотрудник В. Предсказанная вероятность увольнения – 0,3 (ниже порога 0,5). Решение: не увольнение.

Здесь порог 0.5 является произвольным выбором, и его можно

адаптировать в зависимости от конкретных требований и условий.

Вернемся к метрике AUC, а именно к True Positive Rate (TPR) и False Positive Rate (FPR). В формулах (6) и (7) показан способ их расчета.

$$TPR = \frac{TP}{TP + FN}, \tag{6}$$

$$FPR = \frac{FP}{FP + TN}. \tag{7}$$

– TPR показывает, какую часть положительных примеров модель правильно предсказала;
– False Positive Rate (FPR) показывает, какую часть отрицательных примеров модель неправильно предсказала.

Пример того, как выглядит график ROC, показан на рисунке 8.

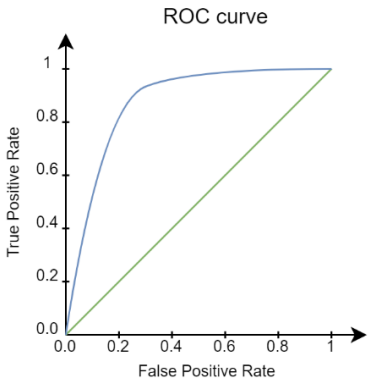


Рисунок 8 – График кривая-ROC

На графике ROC по оси X откладывается FPR, а по оси Y - TPR. Каждая точка на графике представляет собой различный порог принятия решений моделью.

График ROC позволяет оценить, насколько хорошо модель отличает уволенных сотрудников от не уволенных, независимо от конкретного порога

принятия решений. Идеальная модель, которая делает все правильные предсказания, будет иметь точку в верхнем левом углу графика (TPR = 1, FPR = 0). Чем ближе кривая ROC к верхнему левому углу, тем лучше модель разделяет классы и делает более точные предсказания. Если кривая ROC приближается к диагонали (FPR = TPR), значит модель делает случайные предсказания, и ее производительность сопоставима с случайным угадыванием [22].

Выводы по главе 2

Во второй главе данной работы была приведена методология работы с технологией МО. Определены этапы разработки модели.

Анализ и нормализация включает в себя непосредственно анализ, который в свою очередь может включать проверку на наличие пропущенных значений, аномалий и других ошибок в данных, а нормализация в основном будет состоять из масштабирования и преобразования категориальных признаков.

Обучение и тестирование заключается в разбиении данных в соотношении 70% на 30%, где 30% данных будут использоваться для тестирования модели.

Оценка же модели осуществляется по следующим метрикам: accuracy, precision, recall, f1-score, AUC. Использование нескольких метрик даст более полную картину о качестве модели.

Применение этой методологии должно позволить разработать качественную модель машинного обучения.

Глава 3 Разработка моделей и алгоритмов прогнозирования текучести кадров в ИТ-компании

3.1 Инструментарий и предварительная обработка данных

Будем использовать уже ранее упомянутые инструменты:

- Python;
- Pandas;
- Numpy;
- Sklearn (Scikit-learn);
- Matplotlib;
- Google Colab.

Входные данные взяты на платформе Kaggle [41].

Входные данные представляют из себя cvs-файл с названием HR.csv.

В таблице 5 дается описание каждому признаку.

Таблица 5 – Описание признаков

Признак	Описание
satisfaction level	Уровень удовлетворенности сотрудника работой в компании. Значение находится в диапазоне от 0 до 1, где 0 - полное неудовлетворение, а 1 - полное удовлетворение
last evaluation	Оценка работы сотрудника на последней произведенной оценке. Значение находится в диапазоне от 0 до 1
number project	Количество проектов, над которыми работал сотрудник
average montly hours	Среднее количество часов, которое сотрудник тратит на работу в месяц
time spend company	Общее время, которое сотрудник уже проработал в компании в годах

Продолжение таблицы 5

Признак	Описание
work accident	Был ли у сотрудника несчастный случай на работе. Значение 1 указывает на наличие, а 0 - на ее отсутствие
left	Уволился ли сотрудник. Значение 1 указывает на увольнение, а 0 - на остающихся в компании
promotion last 5 years	Было ли продвижение у сотрудника за последние 5 лет. Значение 1 указывает на наличие продвижения, а 0 - на его отсутствие
department	Отдел, в котором работает сотрудник
salary	Уровень зарплаты сотрудника. Возможные значения: «low» (низкий), «medium» (средний), «high» (высокий)

Импортируем входные данные и для получения представления об используемых данных, выведем первые 10 записей. На рисунке 9 показаны первые 10 записей из файла с данными о сотрудниках.

	satisfaction_level	last_evaluation	number_project	average_montly_hours	time_spend_company	Work_accident	left	promotion_last_5years	department	salary
0	0.38	0.53	2	157	3	0	1	0	sales	low
1	0.80	0.86	5	262	6	0	1	0	sales	medium
2	0.11	0.88	7	272	4	0	1	0	sales	medium
3	0.72	0.87	5	223	5	0	1	0	sales	low
4	0.37	0.52	2	159	3	0	1	0	sales	low
5	0.41	0.50	2	153	3	0	1	0	sales	low
6	0.10	0.77	6	247	4	0	1	0	sales	low
7	0.92	0.85	5	259	5	0	1	0	sales	low
8	0.89	1.00	5	224	5	0	1	0	sales	low
9	0.42	0.53	2	142	3	0	1	0	sales	low

Рисунок 9 – Данные о сотрудниках

Можно заметить, что последние две колонки являются категориальными. Это означает, что данные в этих столбцах следует заменить на числовые либо провести для них бинаризацию.

Следующим шагом узнаем общее количество записей, количество ненулевых значений, типы данных для каждого столбца. Полученные данные проиллюстрируем на рисунке 10.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 14999 entries, 0 to 14998
Data columns (total 10 columns):
 #   Column                 Non-Null Count  Dtype
---  ------                 --------------  -----
 0   satisfaction_level     14999 non-null  float64
 1   last_evaluation        14999 non-null  float64
 2   number_project         14999 non-null  int64
 3   average_montly_hours   14999 non-null  int64
 4   time_spend_company     14999 non-null  int64
 5   work_accident          14999 non-null  int64
 6   left                   14999 non-null  int64
 7   promotion_last_5years  14999 non-null  int64
 8   department             14999 non-null  object
 9   salary                 14999 non-null  object
dtypes: float64(2), int64(6), object(2)
memory usage: 1.1+ MB
```

Рисунок 10 – Общая информация о наборе данных

Выведенная информация свидетельствует об отсутствии нулевых значений, что говорит о полноте данных и наличии 14999 записей в файле.

Для получения более полной картины о наборе данных также отобразим различные статистические метрики, такие как:

– mean – среднее значение;

– std – стандартное отклонение;

– 0,25, 0,5 и 0,75 – перцентили;

– min – минимум;

– max – максимум.

При этом если есть нечисловые столбцы, то они будут исключены из результата, так столбцы department, salary не попадут в вывод. Таблицу покажем на рисунке 11.

	satisfaction_level	last_evaluation	number_project	average_montly_hours	time_spend_company	work_accident	left	promotion_last_5years
count	14999.000000	14999.000000	14999.000000	14999.000000	14999.000000	14999.000000	14999.000000	14999.000000
mean	0.612834	0.716102	3.803054	201.050337	3.498233	0.144610	0.238083	0.021268
std	0.248631	0.171169	1.232592	49.943099	1.460136	0.351719	0.425924	0.144281
min	0.090000	0.360000	2.000000	96.000000	2.000000	0.000000	0.000000	0.000000
25%	0.440000	0.560000	3.000000	156.000000	3.000000	0.000000	0.000000	0.000000
50%	0.640000	0.720000	4.000000	200.000000	3.000000	0.000000	0.000000	0.000000
75%	0.820000	0.870000	5.000000	245.000000	4.000000	0.000000	0.000000	0.000000
max	1.000000	1.000000	7.000000	310.000000	10.000000	1.000000	1.000000	1.000000

Рисунок 11 – Статистические метрики

Исходя из предоставленной таблицы, можно сделать вывод, что данные не содержат аномалий, поскольку отсутствуют выбросы в данных.

Определим перечень отделов, представленных в столбце «department».

На рисунке 12 представлен перечень отделов для признака «department».

```
array(['sales', 'accounting', 'hr', 'technical', 'support', 'management',
       'IT', 'product_mng', 'marketing', 'RandD'], dtype=object)
```

Рисунок 12 – Перечень отделов

Как можно заметить, выявлено 10 уникальных отделов. В дальнейшем для этого признака будет проведена бинаризация.

Продолжая анализ данных, узнаем какое количество сотрудников уволилось и осталось в компании.

Покажем это на рисунке 13.

```
0    11428
1     3571
Name: left, dtype: int64
```

Рисунок 13 – Количество уволившихся и оставшихся сотрудников

Из 14999 сотрудников 3571 уволились из компании, а 11428 осталось.

Для этих двух групп выведем средние значения признаков, которые представим на рисунке 14. В результирующую таблицу попадут только числовые признаки.

left	satisfaction_level	last_evaluation	number_project	average_montly_hours	time_spend_company	work_accident	promotion_last_5years
0	0.666810	0.715473	3.786664	199.060203	3.380032	0.175009	0.026251
1	0.440098	0.718113	3.855503	207.419210	3.876505	0.047326	0.005321

Рисунок 14 – Средние значения признаков для двух групп, кто уволился и остался

Заметная разница для двух групп наблюдается в признаке «уровень удовлетворенности». Этот признак трудно выразить в числовом значении, поэтому, скорее всего, значение этого признака было выведено из результатов других тестов или оценок (которые нам не известны), направленных на количественную оценку производительности сотрудников.

Также можно отметить, что сотрудники, которым предоставляли повышение, чаще оставались в компании – разница почти в 5 раз.

Это подчеркивает значимость карьерного роста в контексте удержания сотрудников.

Интересно, что сотрудники, столкнувшиеся с несчастным случаем, почти в 4 раза чаще оставались в компании.

Ранее было выявлено 10 уникальных отделов, вычислим среднее значение признаков для каждого отдела.

Данные представим на рисунке 15.

department	satisfaction_level	last_evaluation	number_project	average_montly_hours	time_spend_company	work_accident	left	promotion_last_5years
IT	0.618142	0.716830	3.816626	202.215974	3.468623	0.133659	0.222494	0.002445
RandD	0.619822	0.712122	3.853875	200.800508	3.367217	0.170267	0.153748	0.034307
accounting	0.582151	0.717718	3.825293	201.162973	3.522816	0.125163	0.265971	0.018253
hr	0.598809	0.708850	3.654939	198.684709	3.355886	0.120433	0.290934	0.020298
management	0.621349	0.724000	3.860317	201.249206	4.303175	0.163492	0.144444	0.109524
marketing	0.618601	0.715886	3.687646	199.385781	3.569930	0.160839	0.236597	0.050117
product_mng	0.619634	0.714756	3.807095	199.965632	3.475610	0.146341	0.219512	0.000000
sales	0.614447	0.709717	3.776329	200.911353	3.534058	0.141787	0.244928	0.024155
support	0.618300	0.723109	3.803948	200.758188	3.393001	0.154778	0.248991	0.008973
technical	0.607897	0.721099	3.877941	202.497426	3.411397	0.140074	0.256250	0.010294

Рисунок 15 – Среднее значение признаков для каждого отдела

Можно заметить, что значения признаков для каждого отдела приблизительно одинаковые, и они находятся на хорошем уровне, что свидетельствует об успешном управлении. Это показывает, что управление отделом, в контексте удержания сотрудников, имеет важное значение.

Для наглядного отображения распределения сотрудников по отделам, создадим круговую диаграмму и покажем её на рисунке 16.

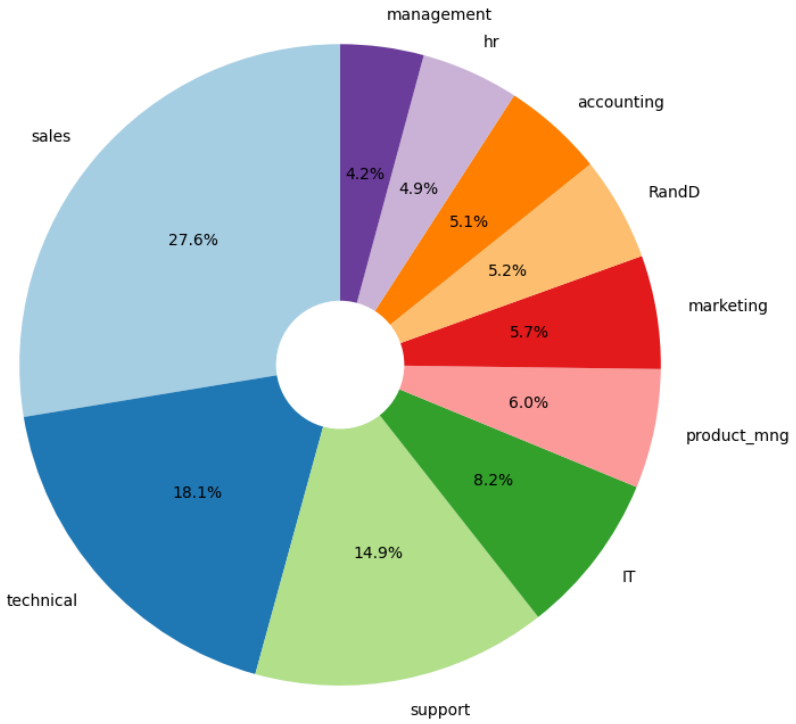


Рисунок 16 – Распределение работников в зависимости от отделов

По представленной диаграмме можно проанализировать компанию.

Заметим, что компания успешно развивает свою деятельность в области продаж благодаря активному участию большого числа сотрудников в клиентской поддержке.

Судя по всему, сотрудники в этой сфере отвечают на вопросы клиентов, в то время как технические специалисты фокусируются на поддержке программного обеспечения и его разработке.

После анализа одного из категориальных признаков переходим к следующему. Признак salary классифицирует сотрудников на три уровня зарплаты: high, middle, low.

Проанализируем данный признак так же, как и признак выше.

Начнем с вычисления средних значений признаков для каждого уровня зарплаты и представим это на рисунке 17.

salary	satisfaction_level	last_evaluation	number_project	average_montly_hours	time_spend_company	work_accident	left	promotion_last_5years
high	0.637470	0.704325	3.767179	199.867421	3.692805	0.155214	0.066289	0.058205
low	0.600753	0.717017	3.799891	200.996583	3.438218	0.142154	0.296884	0.009021
medium	0.621817	0.717322	3.813528	201.338349	3.529010	0.145361	0.204313	0.028079

Рисунок 17 – Среднее значение признаков для каждого уровня зарплаты

Заметим корреляцию, сотрудники, имеющие низкую заработную плату и отсутствующее повышение, чаще покидают компанию, о чем свидетельствует диаграмма, изображенная на рисунке 18.

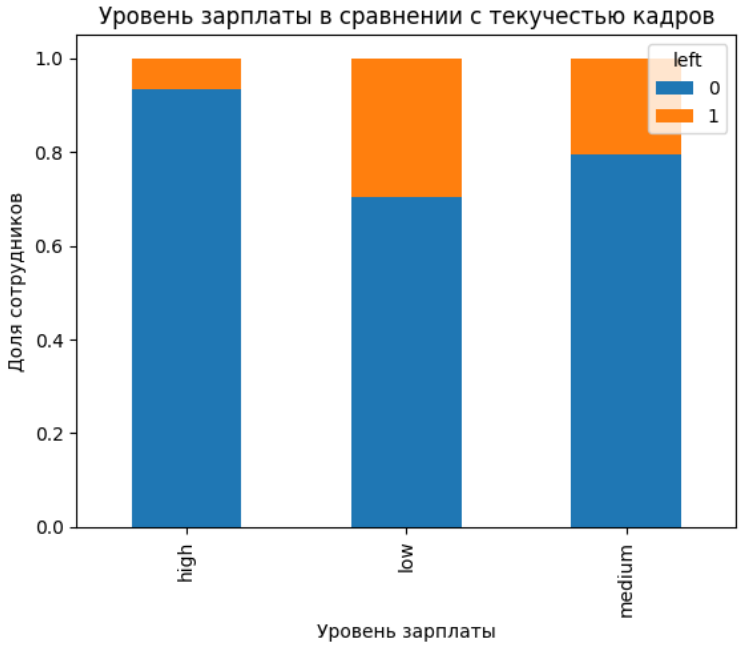


Рисунок 18 – Уровень зарплаты в сравнении с текучестью кадров

Вычислили относительные пропорции, разделив каждое значение в строке таблицы на сумму значений в этой строке. Это преобразование позволило получить долю уволившихся и оставшихся сотрудников для каждого уровня зарплаты.

Отобразим распределение работников в зависимости от уровня зарплаты на рисунке 19.

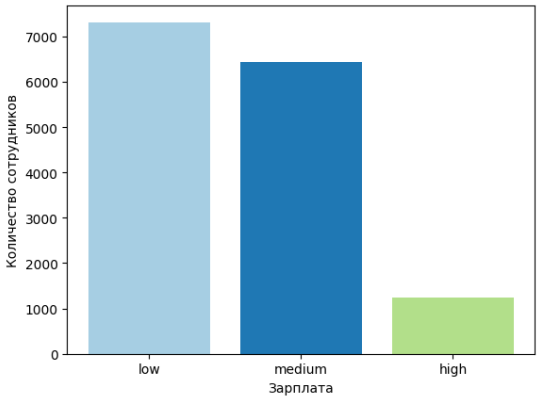


Рисунок 19 – Распределение работников в зависимости от уровня зарплаты

В компании преобладает низкий и средний уровень заработной платы, в то время как высокий уровень зарплаты предоставлен ограниченному числу сотрудников.

Построим столбчатую диаграмму на основе перекрестной таблицы, которая отразит количество сотрудников, уволившихся и оставшихся, в каждом отделе. Данные проиллюстрируем на рисунке 20.

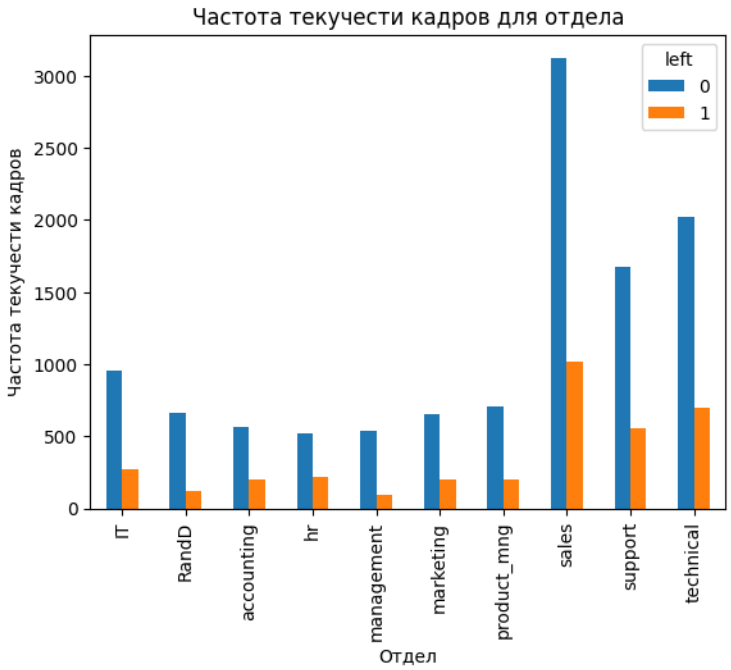


Рисунок 20 – Количество сотрудников, уволившихся и оставшихся в каждом отделе

Если рассчитать процент текучести кадров за все время, то выйдет, что она равна 23,8%. Это удалось высчитать благодаря формуле (2).

$$К_Т = \frac{Ч_{УВ} \cdot 100\%}{Ч_{СРСП}}, \quad (2)$$

где $К_Т$ – коэффициент текучести;

$Ч_{УВ}$ – число выбывших по причинам текучести;

$Ч_{СРСП}$ – среднее списочное число работников в течение того же периода.

Текучесть равная 23,8% представляет собой большой показатель, поскольку нормальным уровнем текучести кадров считается 3-5%. Понимая, что этот показатель может варьироваться в зависимости от сферы деятельности компании, в данном случае отмечается существенное превышение средних значений [13], [7].

Построим серию гистограмм для каждого числового столбца с целью визуализации распределения данных. На оси x будут отображаться значения признаков, а на оси y – количество сотрудников. Гистограммы показаны на рисунке 21.

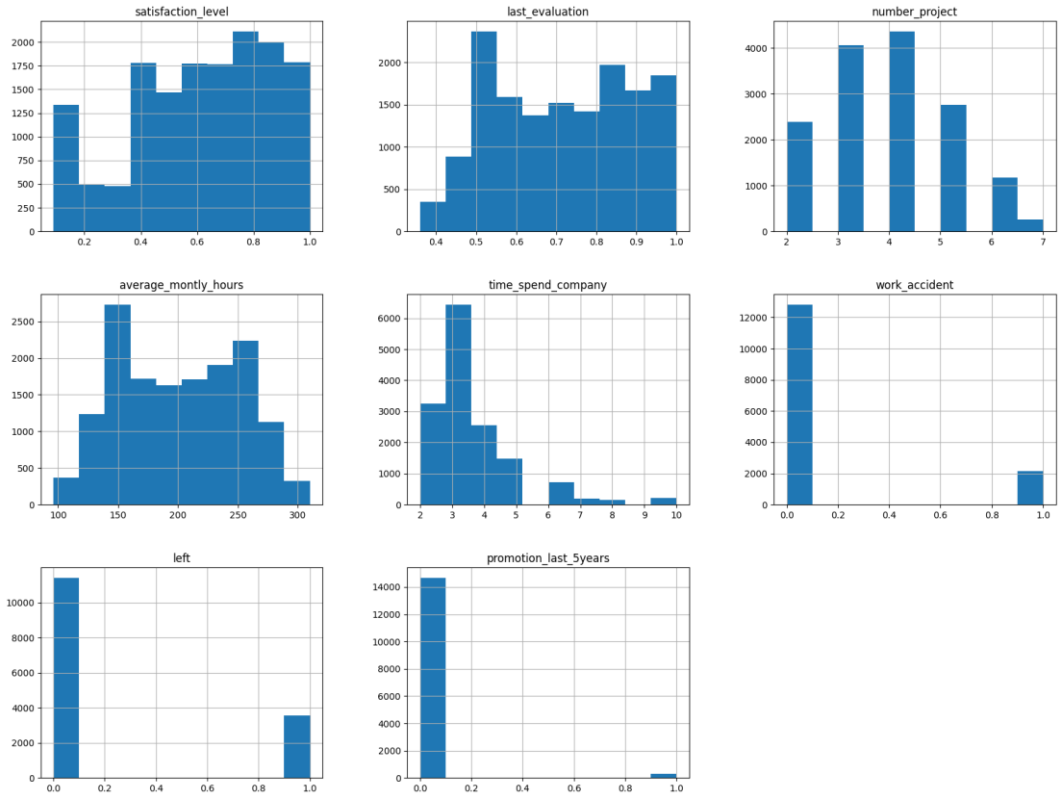
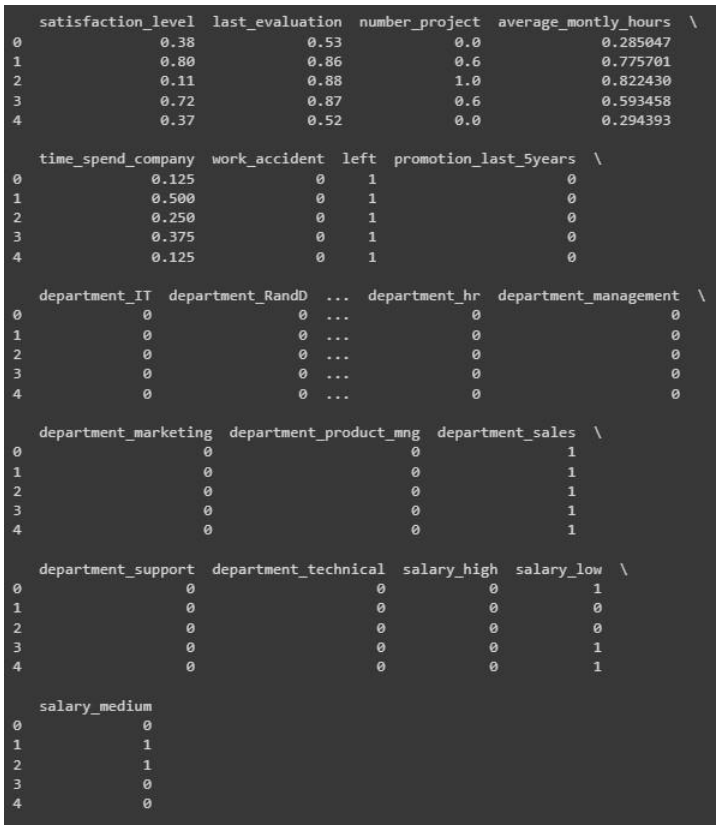


Рисунок 21 – Распределения данных для каждого числового признака

Поскольку в данных присутствуют категориальные признаки salary и department, их необходимо преобразовать в числовые перед моделированием. Для этого проведем бинаризацию, создав фиктивные переменные. Кроме того, выполним масштабирование в диапазоне от 0 до 1 для признаков

number_project, average_montly_hours, time_spend_company, воспользовавшись методом Min-Max Scaling из библиотеки scikit-learn. Применение масштабирования сделает значения признаков сопоставимыми, что повысит стабильность обучения алгоритма.

Выведем первые 5 записей, чтобы оценить результат преобразования данных. Покажем это на рисунке 22.



```
   satisfaction_level  last_evaluation  number_project  average_montly_hours  \
0                0.38             0.53             0.0              0.285047
1                0.80             0.86             0.6              0.775701
2                0.11             0.88             1.0              0.822430
3                0.72             0.87             0.6              0.593458
4                0.37             0.52             0.0              0.294393

   time_spend_company  work_accident  left  promotion_last_5years  \
0               0.125              0     1                      0
1               0.500              0     1                      0
2               0.250              0     1                      0
3               0.375              0     1                      0
4               0.125              0     1                      0

   department_IT  department_RandD  ...  department_hr  department_management  \
0              0                 0  ...              0                      0
1              0                 0  ...              0                      0
2              0                 0  ...              0                      0
3              0                 0  ...              0                      0
4              0                 0  ...              0                      0

   department_marketing  department_product_mng  department_sales  \
0                     0                       0                 1
1                     0                       0                 1
2                     0                       0                 1
3                     0                       0                 1
4                     0                       0                 1

   department_support  department_technical  salary_high  salary_low  \
0                   0                     0            0           1
1                   0                     0            0           0
2                   0                     0            0           0
3                   0                     0            0           1
4                   0                     0            0           1

   salary_medium
0              0
1              1
2              1
3              0
4              0
```

Рисунок 22 – Результат преобразования данных

В конечном итоге, получилась таблица с данными в диапазоне от 0 до 1, и в ней присутствует 21 столбец (признак).

Нормализация данных для моделирования выполнена.

Следующим шагом осуществим разбиение данных на обучающую и тестовую выборку.

Разделим данные в соотношении 80% для обучения и 20% для тестирования. Обучение алгоритма включает передачу обучающих данных и

настройку его параметров. Алгоритм анализирует обучающие данные и «обучается» находить определенные закономерности или паттерны в данных.

3.2 Реализация алгоритмов

3.2.1 Случайный лес

Перед реализацией алгоритма случайный лес необходимо определиться со значениями гиперпараметров. Настройка производится с целью достичь оптимального баланса между обобщающей способностью и предотвращением переобучения. В данной работе было решено сравнить как работают алгоритмы «из коробки», так как в документации scikit-learn написано, что хорошие результаты часто достигаются именно при такой настройке. К каждому алгоритму будут приписаны значения по умолчанию, которые использовались при обучении [46].

После создания алгоритма и его обучения сделаем прогноз на тестовом наборе данных. Данные сведем в таблицу 6.

Таблица 6 – Результат предсказания для алгоритма случайный лес

Метрики	Precision	Recall	F-score	Accuracy
0	0,99	1,00	0,99	0,99
1	0,99	0,97	0,98	

Даже без настройки гиперпараметров точность предсказания алгоритма оказалась очень высокой. Нелишним будет упомянуть, что достичь этого получилось благодаря эффективной нормализации данных.

Значения гиперпараметров покажем на рисунке 23.

```
def __init__(n_estimators=100, *, criterion='gini', max_depth=None, min_samples_split=2, min_samples_leaf=1,
min_weight_fraction_leaf=0.0, max_features='sqrt', max_leaf_nodes=None, min_impurity_decrease=0.0, bootstrap=True, oob_score=False,
n_jobs=None, random_state=None, verbose=0, warm_start=False, class_weight=None, ccp_alpha=0.0, max_samples=None)
```

Рисунок 23 – Значения гиперпараметров для алгоритма случайный лес

Матрица ошибок для алгоритма случайный лес приведена в таблице 7.

Таблица 7 – Матрица ошибок для алгоритма случайный лес

Факт / Предсказание	Предсказанное не увольнение	Предсказанное увольнение
Фактическое не увольнение	TP – 3450	FP – 34
Фактическое увольнение	FN – 12	TN – 1004

Здесь:

- TN (кол-во правильно предсказанных положительных случаев): 3450;
- TP (кол-во правильно предсказанных отрицательных случаев): 1004;
- FP (кол-во ошибочно предсказанных положительных случаев):12;
- FN (кол-во ошибочно предсказанных отрицательных случаев): 34.

Алгоритм обладает очень хорошей способностью классификации и обнаруживает как положительные, так и отрицательные случаи с высокой точностью.

3.2.2 Дерево решений

Следующим алгоритмом будет являться дерево решений. Сделаем предсказание на тестовом наборе данных и сведем результаты в таблицу 8.

Таблица 8 – Результат предсказания для алгоритма дерево решений

Метрики	Precision	Recall	F-score	Accuracy
0	0,99	0,98	0,99	0,98
1	0,94	0,97	0,95	

Данный алгоритм демонстрирует высокую точность предсказания, но несмотря на это, он проигрывает по результатам случайному лесу по всем метрикам, хоть и незначительно.

Значения гиперпараметров покажем на рисунке 24.

```
def __init__(*, criterion='gini', splitter='best', max_depth=None, min_samples_split=2, min_samples_leaf=1, min_weight_fraction_leaf=0.0,
max_features=None, random_state=None, max_leaf_nodes=None, min_impurity_decrease=0.0, class_weight=None, ccp_alpha=0.0)
```

Рисунок 24 – Значения гиперпараметров для алгоритма дерево решений

Представим последующие таблицы матрицы ошибок в упрощенном варианте, так как объяснение было дано выше. В таблице 9 приведена матрица ошибок для данного алгоритма.

Таблица 9 – Матрица ошибок для алгоритма дерево решений

Факт / Предсказание	Предсказание	
Факт	3395	32
	67	1006

Алгоритм дерево решений делает больше ошибок при предсказывании отрицательных случаев, то есть алгоритм предсказал, что сотрудник не будет уволен, в то время как он фактически уволился, по сравнению с случайным лесом – 67 против 12.

3.2.3 Градиентный бустинг

Следующий в реализации алгоритм – градиентный бустинг.

Результаты предсказания сведем в таблицу 10.

Таблица 10 – Результат предсказания для алгоритма градиентный бустинг

Метрики	Precision	Recall	F-score	Accuracy
0	0,98	0,99	0,99	0,98
1	0,97	0,93	0,95	

Градиентный бустинг также демонстрирует высокую точность прогнозирования. Тем не менее, минимальная точность прогноза, оцениваемая

по рассмотренным метрикам, составляет 93%, что является наименьшим показателем среди ранее проанализированных алгоритмов.

В остальном градиентный бустинг и дерево решений демонстрируют схожие результаты проигрывая при этом случайному лесу.

Значения гиперпараметров покажем на рисунке 25.

```
def __init__(*, loss='log_loss', learning_rate=0.1, n_estimators=100, subsample=1.0, criterion='friedman_mse', min_samples_split=2,
min_samples_leaf=1, min_weight_fraction_leaf=0.0, max_depth=3, min_impurity_decrease=0.0, init=None, random_state=None,
max_features=None, verbose=0, max_leaf_nodes=None, warm_start=False, validation_fraction=0.1, n_iter_no_change=None, tol=0.0001,
ccp_alpha=0.0)
```

Рисунок 25 – Значения гиперпараметров для алгоритма градиентный бустинг

Данные сведем в таблицу 11.

Таблица 11 – Матрица ошибок для алгоритма градиентный бустинг

Факт / Предсказание	Предсказание	
Факт	3431	71
	31	967

Действительно, градиентный бустинг чаще ошибается в предсказании увольнения сотрудника, когда на самом деле он не уволился, по сравнению с другими рассмотренными алгоритмами.

3.2.4 Логическая регрессия

Последний алгоритм, который будет рассмотрен в данной работе, это логическая регрессия.

Данные, полученные в результате прогнозирования, сведем в таблицу 12.

Таблица 12 – Результат предсказания для алгоритма логическая регрессия

Метрики	Precision	Recall	F-score	Accuracy
0	0,83	0,93	0,87	0,79
1	0,59	0,35	0,43	

Алгоритм показал наихудшие результаты по сравнению с предыдущими алгоритмами. Выведем коэффициенты логистической регрессии, которые были получены из обученного алгоритма и оценим их влияние на результат. График показан на рисунке 26.

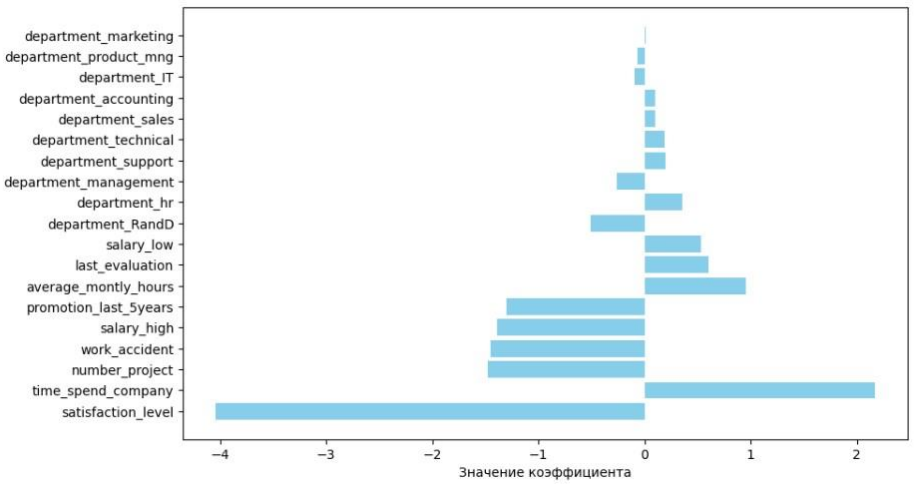


Рисунок 26 – График коэффициентов логистической регрессии

Полученные коэффициенты логистической регрессии являются показателями интенсивности влияния каждого признака на результат классификации. Анализ показывает, что уровень удовлетворенности существенно влияет на принятие решения отнести сотрудника к классу 0 (не уволится). Это объясняет наблюдаемые низкие значения точности и полноты алгоритма, подчеркивая, что алгоритм совершает значительное количество ошибок в предсказаниях, основываясь на данном конкретном признаке.

Значения гиперпараметров покажем на рисунке 27.

```
def __init__(penalty='l2', *, dual=False, tol=0.0001, C=1.0, fit_intercept=True, intercept_scaling=1, class_weight=None,
random_state=None, solver='lbfgs', max_iter=100, multi_class='auto', verbose=0, warm_start=False, n_jobs=None, l1_ratio=None)
```

Рисунок 27 – Значения гиперпараметров для алгоритма логистической регрессии

Данные для матрицы ошибок сведем в таблицу 13.

Таблица 13 – Матрица ошибок для алгоритма логистической регрессии

Факт / Предсказание	Предсказание	
Факт	3208	254
	679	359

Как можно заметить алгоритм делаем большое количество ошибок, как при определении сотрудников, которые могут уволиться, так и тех, которые останутся. Данный алгоритм явно не подходит для решения задачи прогнозирования.

Покажем ROC-кривую для 4 алгоритмов на рисунке 28.

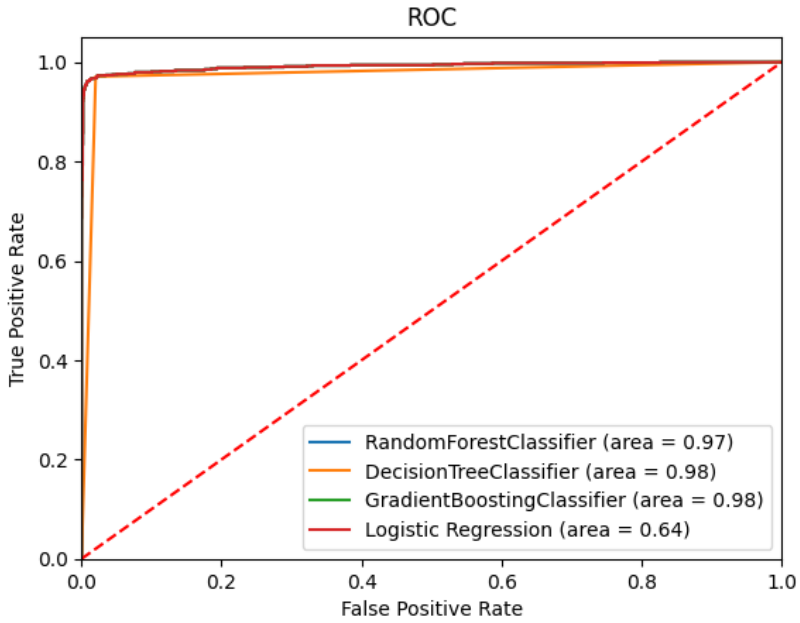


Рисунок 28 – ROC-кривая для 4 алгоритмов

График показывает, что все 4 алгоритма демонстрируют почти идеальные показатели, ведь идеальный алгоритм – это алгоритм, который делает все правильные предсказания и благодаря этому будет иметь точку в верхнем левом углу графика (TPR = 1, FPR = 0).

То есть, чем ближе кривая ROC к верхнему левому углу, тем лучше алгоритм разделяет классы и делает более точные предсказания.

3.3 Сравнение характеристик алгоритмов

Сделаем сравнительный анализ, чтобы определить какой алгоритм подходит больше всего для решения задачи прогнозирования текучести кадров.

Приведем сравнительную таблицу 14.

Таблица 14 – Сравнительная таблица по метрикам

Алгоритм	Классификация	Precision	Recall	F-score	Accuracy
Случайный лес	0	0,99	1,00	0,99	0,99
	1	0,99	0,97	0,98	
Дерево решений	0	0,99	0,98	0,99	0,98
	1	0,94	0,97	0,95	
Градиентный бустинг	0	0,98	0,99	0,99	0,98
	1	0,97	0,93	0,95	
Логическая регрессия	0	0,83	0,93	0,87	0,79
	1	0,59	0,35	0,43	

Сведем данные матрицы ошибок для 4 алгоритмов в таблицу 15.

Таблица 15 – Сравнительная таблица по матрице ошибок

Алгоритм	True Positive	False Positive	False Negative	True Negative
Случайный лес	3450	34	12	1004
Дерево решений	3395	32	67	1006
Градиентный бустинг	3431	31	71	967
Логическая регрессия	3208	254	679	359

На основании анализа метрик, связанных с четырьмя рассмотренными алгоритмами, можно сделать вывод о превосходстве случайного леса. Данный

алгоритм демонстрирует высокие показатели по большинству метрик, что подтверждает его эффективность. Также отметим, что случайный лес делает меньше ошибок в предсказаниях по сравнению с другими алгоритмами. Отобразим важность признаков для этого алгоритма на рисунке 29.

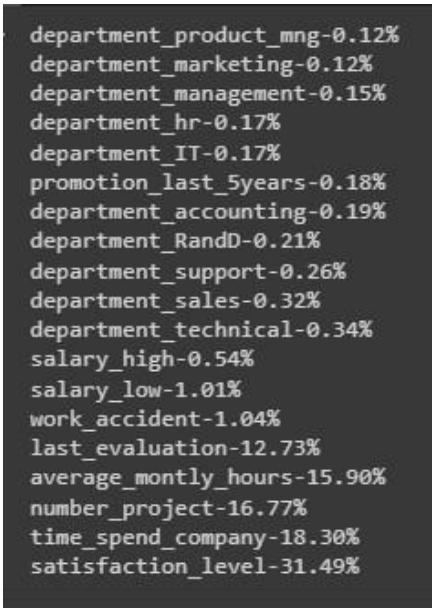


Рисунок 29 – Важность признаков

Согласно алгоритму, единственным наиболее важным признаком, определяющим, останется сотрудник или уйдет, является удовлетворенность.

Если сотрудник удовлетворен своей работой, вероятность того, что он покинет компанию, маловероятна.

Полный код программы представлен в Приложении А.

Приведем схему, на которой покажем путь разработки модели от определения набора данных для прогнозирования до важности признаков.

Схема показана на рисунке 30.

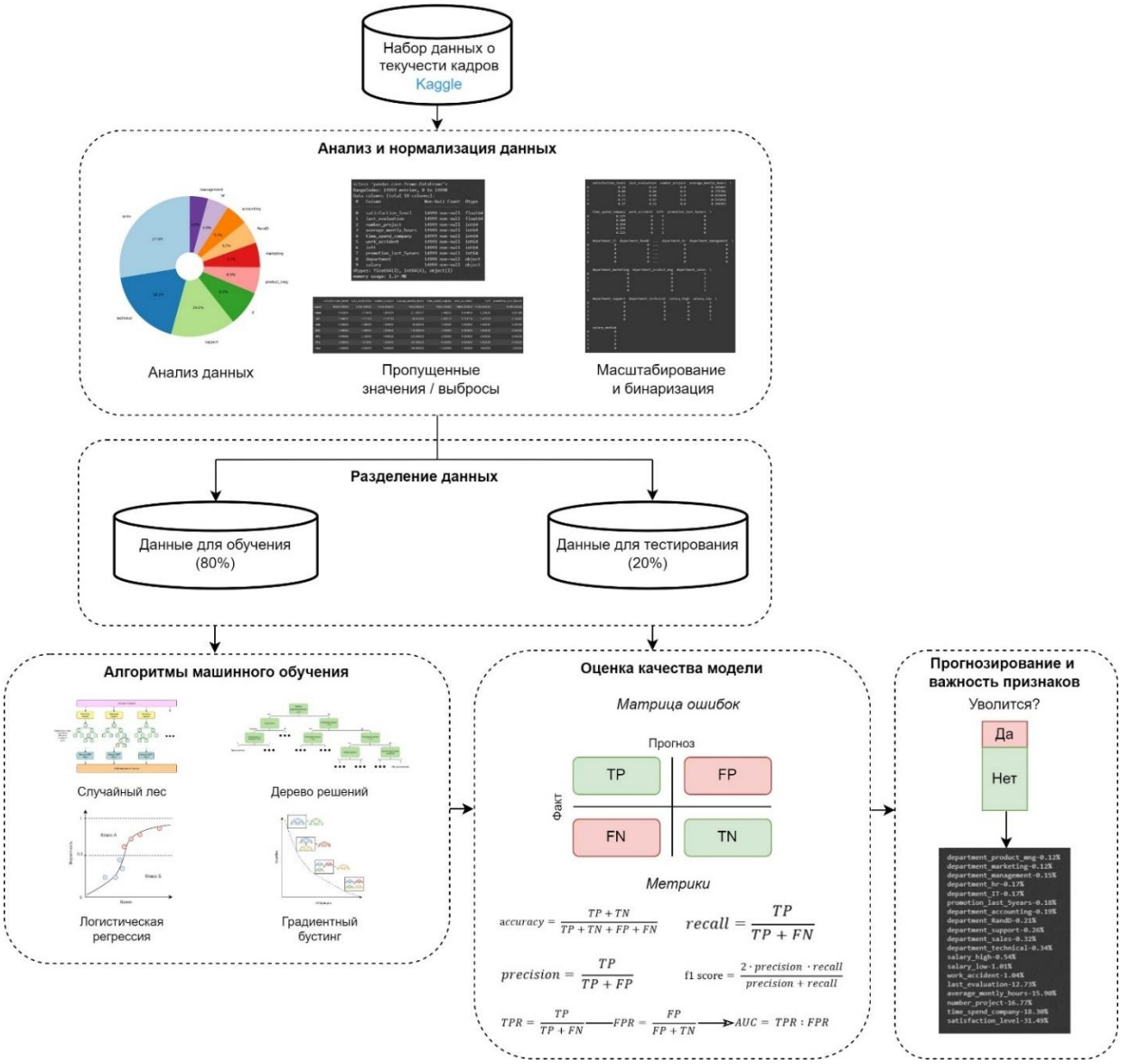


Рисунок 30 – Схема разработки модели прогнозирования

Выводы по главе 3

В ходе выполнения данной части работы был проведен тщательный анализ данных с использованием инструментария. Для анализа использовались данные с платформы Kaggle, включающие 10 признаков. После изучения набора данных было выявлено, что данные не содержат нулевых значений и аномалий, что показало их надежность. Была успешно произведена нормализация данных, которая включала в себя бинаризацию категориальных значений для признаков department и salary. Дополнительно применение Min-Max Scaling для масштабирования данных в диапазоне от 0

до 1 помогло стандартизировать данные перед обучением модели.

Вывод о значимости карьерного роста для удержания сотрудников подчеркивает важность развития внутренних возможностей для персонала. Дополнительно, наблюдение о том, что сотрудники с низкой заработной платой и отсутствием повышения чаще покидают компанию, подчеркивает важность заботы о вознаграждении и развитии персонала.

Разделение данных на обучающий и тестовый наборы в соотношении 80/20 обеспечило адекватную оценку производительности модели.

В результате сравнительного анализа четырех алгоритмов было обнаружено, что случайный лес оказался наилучшим алгоритмом для решения задачи классификации. Это говорит о его эффективности в данной задаче. Кроме того, была выделена важность признаков для случайного леса показывающая, что удовлетворенность является наиболее важным фактором, определяющим, останется ли сотрудник в компании или уйдет. Это практическое открытие, которое может быть использовано для улучшения стратегий удержания персонала.

Глава 4 Апробация проектных решений и оценка их эффективности

4.1 Анализ эффективности предложенной модели

В данной главе приведем экспериментальную апробацию результатов и сделаем оценку гипотезы исследования, а также приведем результаты апробации исследования с описанием оценки возможности практического использования полученных в рамках исследования научных результатов.

Для проведения экспериментальной апробации результатов обратимся к работам, где использовался алгоритм случайный лес и решалась та же задача, что в данном исследовании, то есть задача классификации. Сравниваться две модели будут по сопоставимым метрикам оценки модели, а именно:

- accuracy (правильность);
- precision (точность);
- recall (полнота);
- f1-score (f1-мера);
- AUC (area under curve – площадь под кривой).

В работе [40] автор использует набор данных и алгоритм, как и в данном исследовании. Результаты работы оказались впечатляющими. Автору удалось достичь точности предсказания, равной 98,64%, однако автор не приводит другие метрики оценки модели, кроме accuracy. Это объясняется тем, что для тестирования модели используется кросс-валидация, в частности, 10-кратная перекрестная валидация. Это позволило поднять точность предсказания до 99,40%, однако 10-кратная перекрестная валидация является недетерминированным методом, то есть нет гарантии того, что в следующий раз при применении данного метода результат повторится.

Стоит отметить, что в данной диссертационной работе использовалась кросс-валидация hold-out, которая заключалась в том, что набор данных разделялся на обучающие и тестовые случайным образом, а значение 0 указанное в параметрах дало бы возможность воспроизвести результаты, тем

самым нивелировать недостаток метода hold-out, заключающийся в том, что результат зависит от разбиения, то есть как в случае с 10-кратной перекрестной валидацией.

Таким образом, приведенное исследование является важным для сравнения двух моделей. Да, автору удалось превзойти результаты текущей работы, однако нет убежденности в том, что в следующий раз при тестировании модели используя метод 10-кратной перекрестной валидации, получится такой же результат. Также ввиду схожести исследований можно утверждать, что метрики precision, recall, f-score, accuracy приведенной работы имеют такие же высокие значения, как и в данном исследовании.

Говорить что-либо про эффективность модели пока преждевременно, проанализируем другие работы, приведенные в научных исследованиях.

Чтобы понять, действительно ли модель может прогнозировать с такой высокой точностью как в текущем исследовании, обратимся к работе [12], где автор использовал тот же набор данных, что и в данном исследовании. Однако автор не использовался алгоритм случайный лес, и в самой работе отсутствует причина того, почему не был выбран данный алгоритм. Были использованы следующие алгоритмы (справа указана точность прогноза):

- дерево решений – 97%;
- логистическая регрессия – 80%;
- градиентный бустинг – 97%.

Основываясь на приведенной работе, можно констатировать, что такая высокая точность предсказания является приемлемой, так как и в данном исследовании алгоритмы дерево решений и градиентный бустинг имели схожую точность предсказания.

В работе [4] предсказывается отток клиентов банка. Наилучший результат предсказания с вероятностью 87,1% показывает модель, которая использует алгоритм обучения случайный лес. Метрики оценки отобразим в таблице 16.

Таблица 16 – Метрики оценки для второй анализируемой работы

Class	Precision	Recall	F-score	Accuracy
0	0,89	0,96	0,92	87,1%
1	0,77	0,52	0,62	

Для полноты картины также рассмотрим матрицу ошибок в упрощенном виде в таблице 17.

Таблица 17 – Матрица ошибок для второй анализируемой работы

Факт / Предсказание	Предсказание	
Факт	1531	64
	194	211

Приведем метрики, полученные в настоящем исследовании в таблице 18.

Таблица 18 – Метрики оценки полученные в настоящем исследовании

Class	Precision	Recall	F-score	Accuracy
0	0,99	1,00	0,99	99%
1	0,99	0,97	0,98	

Также придем матрицу ошибок, полученную в настоящем исследовании в таблице 19.

Таблица 19 – Матрица ошибок полученная в настоящем исследовании

Факт / Предсказание	Предсказание	
Факт	3450	12
	38	1000

Проведём сравнительный анализ между двумя моделями. Обозначим результаты другой работы как модель 1, а модель, реализованную в данном

исследовании, – как модель 2, и будем этого придерживаться при последующем анализе работ.

- по точности модель 2 демонстрирует более высокую точность в обоих классах;
- по полноте модель 2 также показывает лучшую полноту в обоих классах;
- по F-мере модель 2 также имеет высокие значения F, особенно по сравнению с отрицательным классом;
- по общей точности модель 2 превосходит модель 1 на 12%;
- по матрице ошибок модель 2 показывает значительно меньшее количество ошибок по сравнению с моделью 1.

Таким образом, модель 2 демонстрирует высокую точность, полноту и F-меру для обоих классов, а также высокую общую точность и меньшее количество ошибок. Эти результаты свидетельствуют о том, что модель 2 будет надежнее в практическом применении и, соответственно, эффективнее.

В работе [11] прогнозируется отток клиентов банка в телекоммуникационной компании. Исследование показало, что наиболее эффективной моделью оказалась модель, использующая случайный лес и набравшая 83,7%, что меньше, чем в другой ранее проанализированной работе, на 3,4%. В данной работе автор приводит просто значения метрик без разбиения их на классы 0 и 1, что снижает полезность результатов и их интерпретируемость. Однако все же приведем метрики данной работы для сравнительного анализа в таблице 20.

Таблица 20 – Метрики оценки для третьей анализируемой работы

Precision	Recall	F-score	Accuracy
0,84	0,84	0,84	83,7%

Модель 1 показывает хорошую точность предсказания, но по метрике accuracy она на ~15% хуже справляется с задачей, чем модель 2. Стоит также

отметить, что это результаты с учетом применения техники SMOTE. До применения данной техники результаты были и того меньше – 79,2%.

SMOTE (Synthetic Minority Over-sampling Technique - метод увеличения числа примеров миноритарного класса) – это алгоритм, использующийся для предварительной обработки данных в случае, если есть необходимость устранения дисбаланса классов в представленном наборе данных. Происходит это путем генерации новых примеров для класса с меньшим количеством выборок, используя методы интерполяции между соседними примерами этого класса.

Результат точности, равный 79,2%, объясняется автором как имеющийся дисбаланс классов в датасете, то есть на 7043 записи о клиентах приходилось 1869 тех, кто покинул, и 5163 тех, кто не покинул компанию. Ввиду этого алгоритмы прогнозирования не смогли дать высокую точность, поэтому была применена техника недостаточной выборки (SMOTE), что улучшило значения прогнозов. В данной диссертационной работе данная техника не применялась ввиду достаточной точности прогнозирования в отличие от результатов приведенной работы. Но все же стоит учесть данный опыт исследования и в будущем использовать данную технику.

Таким образом, модель 2 имеет интерпретируемые данные, не нуждалась в применении техники SMOTE и показывает лучшие результаты по приведенным метрикам, следовательно, модель 2 эффективнее.

В работах [26, 3] методы оценки модели отличаются от стандартных способов. В большинстве своем сравнение полученных алгоритмов происходит по метрике accuracy с учетом других метрик с целью получить более надежные и обоснованные результаты, но в первой работе модель оценивается только по двум метрикам: ROC AUC и accuracy, а в другой работе – только по метрике accuracy. Этого явно недостаточно, чтобы констатировать, что модели надежны и справляются со своей задачей. Попробуем это доказать в анализе следующей работы.

Анализ работ показывает, что в исследованиях используются не совсем

правильные методы оценки моделей, что уже было подмечено ранее. Однако стоит отметить, что модель первой работы прогнозирует с вероятностью 85,11%, а другая – 77,92%. Также во второй работе автор пренебрег предварительной обработкой данных, что и привело к невысокому показателю точности.

Таким образом, модель 2 справляется со своей задачей лучше представленных 2-х моделей. Преимущество 2 модели заключается в том, что в данном исследовании использовались стандартные методы оценки, что помогает в дальнейшем интерпретировать результаты данной работы. Использование нескольких метрик оценки повышает надежность модели. Ввиду вышесказанного можно констатировать, что модель 2 является эффективной моделью.

Представим показатели метрик следующей работы [36] в таблице 21.

Таблица 21 – Метрики оценки для шестой анализируемой работы

Class	Precision	Recall	F-score	Accuracy
0	0,99	0,98	0,93	99,20%
1	0,90	0,40	0,26	

В случае, когда положительный класс предсказывается лучше, чем отрицательный, это может указывать на некоторые проблемы с моделью или данными. Возможно, что в данных существует дисбаланс между классами. То есть в датасете положительных примеров гораздо больше, чем отрицательных. Соответственно, результаты модели могут сильно смещаться в сторону предсказания положительного класса, что и приводит к высокой точности, но низкой полноте и F-мере для положительного класса.

Действительно, в данной работе использовался набор данных, составленный IBM, в котором содержалось 1470 записей о сотрудниках с 32 признаками. В общей сложности 1233 сотрудника относились к категории «нет», в то время как остальные 237 сотрудников относились к категории «да».

В данной работе имел бы смысл использовать технологию SMOTE, чтобы повысить точность предсказания. Ввиду ее отсутствия полученная модель показывает низкие результаты.

Если бы автор исследования решил использовать одну метрику, например, accuracy, или привел бы результаты оценки модели без разделения на классы, как будет показано в анализе последующей работы, то получилась бы, что модель делает прогнозы с высокой точностью, тогда как на деле построенная модель являлась бы ненадежной для решения задачи.

Судя по значениям метрик, полученным в результате построения модели и ее оценки, можно констатировать, что модель 2 будет эффективнее справляться с задачей классификации.

В другой работе [44] автор разрабатывает модель для того, чтобы повысить эффективность отдела HR за счет автоматизации рутинных задач, улучшения процесса принятия решений, а также повышения вовлеченности и удержания сотрудников. Приведем результаты работы в таблице 22.

Таблица 22 – Метрики оценки для седьмой анализируемой работы

Precision	Recall	F-score	Accuracy
0,83	0,85	0,82	85,51%

Можно заметить, что в данной работе не показано, к какому конкретно классу принадлежит значение. Как было показано ранее, приведение одного значения может говорить о том, что прогнозируется хорошо только какой-то определенный класс, тогда как точность предсказания для другого класса может быть крайне низка. Действительно, последующий анализ показал, что в работе используются данные IBM, про которые уже шла речь выше, имеющие дисбаланс в классах. В результате выходит, что данные, скорее всего, были предоставлены для положительного класса (0), тогда как для отрицательного класса (1) значения были не представлены. Ввиду всего вышеизложенного можно сказать, что модель 2 по сравнению с приведенной работой

эффективнее.

Вышеописанная проблема также наблюдается и в работе [43], где используются данные IBM. Значения метрик покажем в таблице 23.

Таблица 23 – Метрики оценки для восьмой анализируемой работы

Precision	Recall	F-score	Accuracy
0,47	-	0,32	86,89%

Хоть и значение метрики accuracy приемлемая, остальные метрики свидетельствуют о том, что модель явно не подходит для решения той задачи, для которой была разработана. К тому же автор не стал приводить метрику оценки модели recall.

В работе [37] автор использует набор данных, который объединяет в себе информацию о карьере сотрудников, собранный с помощью веб-сайта поиска вакансий Glassboor и данные сообщества по корпоративным обзорам и поиску работы в США. Итого набор данных состоял из 17724 уже очищенных данных.

Для тестирования использовалось 7090 записей. Автор приводит значения метрик, однако для одного класса, а также матрицу ошибок. Приведем данные метрик в таблице 24.

Таблица 24 – Метрики оценки для девятой анализируемой работы

Precision	Recall	F-score	Accuracy
0,55	0,28	-	63%

Точность модели низкая, что говорит о неэффективности модели. Также не приведена метрика F-score.

Отобразим значения матрицы ошибок для каждой работы в таблице 25.

Таблица 25 – Сравнение значений матрицы ошибок девятой анализируемой работы с результатами настоящего исследования

Данные анализируемого исследования	Данные настоящего исследования
[[3615 651] [2022 802]]	[[3450 12] [38 1000]]

По матрице ошибок модель 2 показывает значительно меньшее количество ошибок по сравнению с моделью 1, что говорит об эффективности модели 2.

В другой работе [48] оценивает уровень текучести медсестер в США. Набор данных состоял из 43 987 записей. В данных наблюдался дисбаланс классов – 89% (1) против 11% (0), однако автор использовал технику SMOTE, благодаря чему смог исправить изначальный дисбаланс в классах. Как итог, дисбаланс классов был устранен – 57% (1) на 43% (0). Покажем результат в таблице 26.

Таблица 26 – Метрики оценки для десятой анализируемой работы

Precision	Recall	F-score	Accuracy
0,82	0,91	0,88	82,21%

Значения матрицы ошибок приведем в таблице 27.

Таблица 27 – Сравнение значений матрицы ошибок десятой анализируемой работы с результатами настоящего исследования

Данные анализируемого исследования	Данные настоящего исследования
[2714 561] [322 1708]	[3450 12] [38 1000]

Хорошая работа. Учтены недостатки других работ. За исключением представления значений для двух классов. Сама модель демонстрирует хорошую точность, представленная матрица ошибок тому доказательство, однако модель 2 будет эффективнее за счет более точных прогнозов.

Автор исследования [42] разрабатывал модель прогнозирования текучести кадров медсестер в Корее с использованием МО. Использовались данные для обучения и тестирования 630 вышедших на пенсию и 780 работающих медсестер в отделении сестринского ухода больницы общего профиля, то есть 1410 записей. Данные собирались с 1 января 2011 года по 31 июля 2021 года. Признаками служили: возраст, пол, район проживания, пользование общежитием, семейное положение, отдел, год приема на работу, год увольнения, зарплату и стаж работы. Автор не приводит значения для положительного и отрицательного класса. Данные были разделены 80 на 20, то есть 282 записи для тестирования модели. Сильного дисбаланса классов не наблюдалось. Приведем результаты исследования в таблице 28 и 29.

Таблица 28 – Метрики оценки для одиннадцатой анализируемой работы

Precision	Recall	F-score	Accuracy
0,92	0,91	0,92	92%

Таблица 29 – Матрица ошибок для одиннадцатой анализируемой работы

Факт / Предсказание	Предсказание	
Факт	146	5
	18	113

Метрики оценки показали, что два алгоритма, случайный лес и дерево решений, показали одинаковую точность – 92%, но следовало выбрать алгоритм, который лучше другого справлялся бы с задачей, и для этого было проведено дополнительное сравнение алгоритмов, используя метрику AUC-ROC, где случайный лес показал себя лучше со значением 97% против 96% у дерева решений. Затем автор использовал оптимизацию гиперпараметров, и за счет этого точность алгоритма случайный лес повысилась до 98,6%, что является очень хорошим показателем.

Таким образом, исследование показывает, что можно при

необходимости использовать оптимизацию гиперпараметров, которая может улучшить точность алгоритма. Можно заметить, что значения метрик были приведены только для какого-то конкретного класса, что, конечно же, является недостатком исследования. Также можно отметить, что при разработке использовались реальные данные сотрудников, что говорит о том, что модель прогнозирования эффективна и в решении реальных задач.

4.2 Результаты проведенного анализа

Для наглядного представления результатов построим график, в котором отобразим полученные результаты проанализированных работ. Для сравнения будем использовать метрику accuracy, однако при анализе работ было выяснено, что некоторые модели не совсем правильно оценивались. К данным моделям, а именно к названиям, добавим символ «*». Покажем график на рисунке 31.

Рисунок 31 – Точность предсказания моделей

Красный маркер показывает точность прогнозирования модели, разработанной в данной работе.

M1, M2 и т.д. – это сокращение слова «модель», а число указывает на порядковый номер модели. Первая модель указывает на работу, которая была

проанализирована в самом начале работы, а одиннадцатая – на самую последнюю проанализированную работу. Двенадцатая же модель – это модель, которая была получена в настоящем исследовании. Как можем заметить, из 11 проанализированных работ 5 справляются с решением задачи прогнозирования ввиду того, что полученные результаты работы были обоснованы несколькими метриками или матрицей ошибок, в отличие от других 6, результаты которых вызывают сомнения. В итоге имеются 4 модели с высокими показателями и результаты модели, разработанной в данной работе, среди них.

Выводы по главе 4

Апробация эффективности разработанной модели, показала, что модель получилась достаточно надежной и эффективной для прогнозирования текучести кадров, и вот почему:

Это также говорит о правильном подходе к разработке, то есть желательно приводить несколько метрик для лучшей обоснованности точности модели, причем приводить результаты метрик для обоих классов, так как в ином случае это чревато тем, что модель может хорошо предсказать положительный класс, тогда как отрицательный класс – неточно.

В случае если используется несколько метрик оценки, выше шанс обнаружить проблемы с моделью, которые можно будет проанализировать, а затем найти способ решения возникших проблем, например, использовав технику SMOTE или в случае отсутствия проблем, можно просто попытаться улучшить точность, используя, например, оптимизацию гиперпараметров.

Было приведено исследование, где брались реальные данные сотрудников, и модель, использующая алгоритм случайный лес, показала высокие результаты, благодаря чему делается вывод, что и модель, полученная в настоящем исследовании, может быть применена для решения реальных бизнес-задач.

Заключение

В ходе выполнения данной работы были изучены научные работы и исследована литература по теме исследования, благодаря чему выяснилось, что потеря ключевого сотрудника обостряет кадровые риски, тогда как прогнозирование позволяет контролировать эти риски, укрепляя уверенность компании в будущем и позволяя безопасно инвестировать в сотрудников. Определено, что предиктивная аналитика, использующая машинное обучение, способна прогнозировать вероятность увольнения сотрудников и тем самым помогает эффективнее удерживать персонал.

Была приведена методология работы с технологией МО, представляющая собой четыре этапа: выбор инструментария, подготовка данных, обучение и тестирование модели, оценка модели, что позволило систематизировать процесс разработки модели и обеспечить ее качество.

Был проведен анализ данных с платформы Kaggle, включающих 10 признаков. Данные оказались надежными, не содержали нулевых значений и аномалий. Нормализация включала бинаризацию категориальных значений и масштабирование, что стандартизировало данные перед обучением модели. Также анализ данных позволил прийти к выводу о важности карьерного роста и вознаграждения для удержания сотрудников.

Благодаря проведенному анализу научных работ были отобраны те алгоритмы, которые наилучшим образом себя зарекомендовали в прогнозировании текучести кадров, а именно случайный лес, дерево решений, градиентный бустинг и логистическая регрессия. В дальнейшем данные алгоритмы были реализованы и сравнены между собой по таким метрикам, как accuracy, precision, recall, f1-score, AUC, вследствие чего был определен наиболее подходящий для предсказания текучести кадров алгоритм – случайный лес. Для данного алгоритма также была приведена важность признаков, что может использоваться для улучшения стратегий удержания персонала.

Разработанная модель была успешно апробирована и оценена с использованием различных метрик и матрицы ошибок, что подтвердило её надёжность и эффективность. Сравнение с аналогичными моделями подтвердило высокую точность и применимость данной модели для реальных бизнес-задач.

Результаты настоящего исследования могут быть применены в любых компаниях, где есть потребность в предсказании текучести кадров, особенно в ИТ-компаниях, где уровень зарплаты квалифицированного сотрудника может быть довольно высок, а найм нового сотрудника может обойтись компании дороже. Также результаты могут использоваться для улучшения условий труда за счёт анализа важности признаков.

Гипотеза о том, что эффективность принятия управленческих решений можно повысить, если использовать такую технологию, как машинное обучение для прогнозирования увольнения сотрудников, была подтверждена.

Таким образом, все задачи, поставленные в данной работе, были выполнены.

Список используемой литературы и используемых источников

1. Абзалилова Л. Р. Моделирование оттока кадров в крупной компании с применением технологий интеллектуального анализа данных //Экономика и управление: научно-практический журнал. – 2021. – №. 3. – С. 152.

2. Арзамасцев С. А., Бгатов М. В., Картышева Е. Н., Деркунский В. А., Семенчиков Д. Н. Предсказание оттока абонентов: сравнение методов машинного обучения // КИО. 2018. №5. URL: https://cyberleninka.ru/article/n/predskazanie-ottoka-abonentov-sravnenie-metodov-mashinnogo-obucheniya (дата обращения: 02.06.2023).

3. Бабаев А.М., Шемякина М.А. Обзор классических методов машинного обучения в контексте решения задач классификации // Форум молодых ученых. 2018. №11-1 (27). URL: https://cyberleninka.ru/article/n/obzor-klassicheskih-metodov-mashinnogo-obucheniya-v-kontekste-resheniya-zadach-klassifikatsii (дата обращения: 26.04.2024).

4. Белова Е. Е., Толстель О. В. Использование библиотек языка программирования Python для анализа оттока клиентов банка // Вестник Балтийского федерального университета им. И. Канта. Серия: Физико-математические и технические науки. 2019. №4. URL: https://cyberleninka.ru/article/n/ispolzovanie-bibliotek-yazyka-programmirovaniya-python-dlya-analiza-ottoka-klientov-banka (дата обращения: 26.04.2024).

5. Боброва М. В., Мастилин А. Е. Машинное обучение в кибербезопасности // Научные междисциплинарные исследования. 2021. №2. URL: https://cyberleninka.ru/article/n/mashinnoe-obuchenie-v-kiberbezopasnosti (дата обращения: 19.03.2024).

6. Вантеева В. В. Кадровая безопасность предприятия // Скиф. 2021. №1 (53). URL: https://cyberleninka.ru/article/n/kadrovaya-bezopasnost-predpriyatiya (дата обращения: 03.10.2022).

7. Герасимов К. Б., Браева Т.С. Определение уровня текучести

персонала в организации // Экономика и бизнес: теория и практика. 2016. №12. URL: https://cyberleninka.ru/article/n/opredelenie-urovnya-tekuchesti-personala-v-organizatsii (дата обращения: 24.12.2023).

8. Давлетов А. Р. Главные трудности при интеграции машинного обучения в коммерческую эксплуатацию // Инновации и инвестиции. 2023. №10. URL: https://cyberleninka.ru/article/n/glavnye-trudnosti-pri-integratsii-mashinnogo-obucheniya-v-kommercheskuyu-ekspluatatsiyu (дата обращения: 24.03.2024).

9. Дубкова Е. В., Дубков В. А. Кадровая составляющая экономической безопасности организации // Финансовые рынки и банки. 2021. №7. URL: https://cyberleninka.ru/article/n/kadrovaya-sostavlyayuschaya-ekonomicheskoy-bezopasnosti-organizatsii (дата обращения: 18.10.2022).

10. Ибрагимова П. А., Гусайниева Х. Г. Кадровая безопасность: риски, угрозы, пути совершенствования // РППЭ. 2021. №5 (127). URL: https://cyberleninka.ru/article/n/kadrovaya-bezopasnost-riski-ugrozy-puti-sovershenstvovaniya (дата обращения: 06.11.2022).

11. Камалходжаева Н., Шиков А.Н. Анализ алгоритмов машинного обучения для прогнозирования оттока клиентов в телекоммуникационной компании // МНИЖ. 2022. №7-1 (121). URL: https://cyberleninka.ru/article/n/issledovanie-i-analiz-algoritmov-mashinnogo-obucheniya-dlya-prognozirovaniya-ottoka-klientov-v-telekommunikatsionnoy-kompanii (дата обращения: 26.04.2024).

12. Колесников В. Д. Построение модели прогнозирования оттока сотрудников // Вестник Балтийского федерального университета им. И. Канта. Серия: Физико-математические и технические науки. 2019. №1. URL: https://cyberleninka.ru/article/n/postroenie-modeli-prognozirovaniya-ottoka-sotrudnikov (дата обращения: 26.04.2024).

13. Кондакова А. А. Текучесть кадров: подходы и классификация понятий // Концепт. 2017. №S1. URL: https://cyberleninka.ru/article/n/tekuchest-kadrov-podhody-i-klassifikatsiya-ponyatiy (дата обращения: 24.12.2023).

14. Кремкова Д. Д., Сафонов И. А. Модели машинного обучения для идентификации потенциально уходящих абонентов на примере телекоммуникационной компании tele2 // Научные записки молодых исследователей. 2020. №5. URL: https://cyberleninka.ru/article/n/modeli-mashinnogo-obucheniya-dlya-identifikatsii-potentsialno-uhodyaschih-abonentov-na-primere-telekommunikatsionnoy-kompanii-tele2 (дата обращения: 01.06.2023).

15. Кричевский М. Л., Дмитриева С. В., Мартынова Ю. А. Выбор модели оценки текучести персонала // Лидерство и менеджмент. – 2022. – Том 9. – № 2. – С. 391–404. doi: 10.18334/lim.9.2.114741.

16. Крюкова Я. Э., Кручинин И. И. Обзор способов применения методов машинного обучения для прогнозирования поведения пользователей //Международный студенческий научный вестник. – 2019. – №. 2. – С. 25-25.

17. Кузнецов А. Д. Кадровая составляющая экономической безопасности предприятия // Вестник науки. 2022. №7 (52). URL: https://cyberleninka.ru/article/n/kadrovaya-sostavlyayuschaya-ekonomicheskoy-bezopasnosti-predpriyatiya (дата обращения: 12.11.2022).

18. Логвинова И. В., Прохачева О. И. Кадровая безопасность в системе обеспечения экономической безопасности // Евразийский Союз Ученых. 2021. №3-8 (84). URL: https://cyberleninka.ru/article/n/kadrovaya-bezopasnost-v-sisteme-obespecheniya-ekonomicheskoy-bezopasnosti (дата обращения: 03.10.2022).

19. Максимова К. А. Применение HR-аналитики для принятия эффективных управленческих решений // Телескоп. 2021. №4. URL: https://cyberleninka.ru/article/n/primenenie-hr-analitiki-dlya-prinyatiya-effektivnyh-upravlencheskih-resheniy (дата обращения: 05.10.2022).

20. Марков А. Б. Прогнозирование методами машинного обучения оттока клиентов компании, занимающейся онлайн торговлей / А. Б. Марков // Хроники цифровых трансформаций: Сборник научных тезисов и статей по материалам межкафедральных круглых столов, Москва, 03–29 марта 2022

года. Том Выпуск 2. – Волгоград: ИП ЧЕРНЯЕВА ЮЛИЯ ИГОРЕВНА (Издательский дом "Сириус"), 2022. – С. 55-59. – EDN NASYEX.

21. Митяков С. Н., Митяков Е. С. Машинное обучение в задачах исследования инновационных процессов // Журнал прикладных исследований. 2020. №4. URL: https://cyberleninka.ru/article/n/mashinnoe-obuchenie-v-zadachah-issledovaniya-innovatsionnyh-protsessov (дата обращения: 03.11.2022).

22. Михайличенко А. А. Аналитический обзор методов оценки качества алгоритмов классификации в задачах машинного обучения // Вестник Адыгейского государственного университета. Серия 4: Естественно-математические и технические науки. 2022. №4 (311). URL: https://cyberleninka.ru/article/n/analiticheskiy-obzor-metodov-otsenki-kachestva-algoritmov-klassifikatsii-v-zadachah-mashinnogo-obucheniya (дата обращения: 28.05.2023).

23. Михайлов А. Н. Использование машинного обучения в бизнесе // Современные инновации. 2023. №1 (42). URL: https://cyberleninka.ru/article/n/ispolzovanie-mashinnogo-obucheniya-v-biznese (дата обращения: 20.03.2024).

24. Пугачева К. С. риски при увольнении персонала. Пути их предотвращения // Научные известия. 2021. №22. URL: https://cyberleninka.ru/article/n/riski-pri-uvolnenii-personala-puti-ih-predotvrascheniya (дата обращения: 02.11.2022).

25. Русских Т. Н., Ивашечкин Д. Применение методов машинного обучения к решению задачи прогнозирования оттока клиентов //Тренды развития современного общества: управленческие, правовые, экономические и социальные аспекты. – 2023. – С. 347-349.

26. Саранцев С. О., Вайтекунене Е. Л. Обзор алгоритмов модели прогнозирования кадрового состава в системе управления предприятием //актуальные проблемы авиации и космонавтики. – 2022. – С. 488-490.

27. Свиридова О. П., Чуланова О. Л. Программа реализации hr -

аналитики как цифрового тренда // Материалы Афанасьевских чтений. 2020. №3 (32). URL: https://cyberleninka.ru/article/n/programma-realizatsii-hr-analitiki-kak-tsifrovogo-trenda (дата обращения: 05.10.2022).

28. Троценко В. М. Обеспечение кадровой безопасности как инструмент обеспечения экономической безопасности организации // Московский экономический журнал. 2019. №9. URL: https://cyberleninka.ru/article/n/obespechenie-kadrovoi-bezopasnosti-kak-instrument-obespecheniya-ekonomicheskoi-bezopasnosti-organizatsii (дата обращения: 28.10.2022).

29. Черемисин Д. Г., Мкртчян В.Р. Методы машинного обучения // Символ науки. 2023. №6-2. URL: https://cyberleninka.ru/article/n/metody-mashinnogo-obucheniya (дата обращения: 19.03.2024).

30. Чечнев А. А. Прогнозирование оттока клиентов телекоммуникационной компании // StudNet. 2021. №6. URL: https://cyberleninka.ru/article/n/prognozirovanie-ottoka-klientov-telekommunikatsionnoy-kompanii (дата обращения: 21.05.2023).

31. Чуланова О. Л. Возможности применения дескриптивной, прогнозной, предиктивной и прескриптивной hr -аналитики как цифровых трендов // Материалы Афанасьевских чтений. 2020. №1 (30). URL: https://cyberleninka.ru/article/n/vozmozhnosti-primeneniya-deskriptivnoy-prognoznoy-prediktivnoy-i-preskriptivnoy-hr-analitiki-kak-tsifrovyh-trendov (дата обращения: 02.11.2022).

32. Чуланова О. Л., Свиридова О. П. Бенчмаркинг рисков применения HR-аналитики в цифровой экономике // Журнал исследований по управлению. 2020. №. 1. С. 25-31. URL: https://naukaru.ru/ru/nauka/article/36587/view (дата обращения: 10.11.2022).

33. Чуланова О. Л., Свиридова О. П. Проектирование внедрения системы HR-аналитики как цифрового тренда для оптимизации процессов управления персоналом // Журнал исследований по управлению. 2020. №. 1. С. 32-42. URL: https://naukaru.ru/ru/nauka/article/36588/view (дата обращения:

21.03.2020). (дата обращения: 04.11.2022).

34. Щукина Е. А., Оглоблин В. А. Текучесть персонала и ее причины // Экономика и бизнес: теория и практика. 2020. №10-2. URL: https://cyberleninka.ru/article/n/tekuchest-personala-i-ee-prichiny (дата обращения: 13.11.2022).

35. Alenezi H. S., Faisal M. H. Utilizing crowdsourcing and machine learning in education: Literature review //Education and Information Technologies. – 2020. – Т. 25. – №. 4. – С. 2971-2986.

36. Chakraborty R. et al. Study and prediction analysis of the employee turnover using machine learning approaches //2021 IEEE 4th International Conference on Computing, Power and Communication Technologies (GUCON). – IEEE, 2021. – С. 1-6.

37. Chang V. et al. Job satisfaction and turnover decision of employees in the Internet sector in the US //Enterprise Information Systems. – 2023. – Т. 17. – №. 8. – С. 2130013.

38. Cioffi R. et al. Artificial intelligence and machine learning applications in smart production: Progress, trends, and directions //Sustainability. – 2020. – Т. 12. – №. 2. – С. 492.

39. Goodell J. W. et al. Artificial intelligence and machine learning in finance: Identifying foundations, themes, and research clusters from bibliometric analysis //Journal of Behavioral and Experimental Finance. – 2021. – Т. 32. – С. 100577.

40. Hossen M. A. et al. Ensemble method based architecture using random forest importance to predict employee's turn over //Journal of Physics: Conference Series. – IOP Publishing, 2021. – Т. 1755. – №. 1. – С. 012039.

41. Kaggle. URL: https://www.kaggle.com/datasets/kuniowu/human-resource (дата обращения: 01.06.2023).

42. Kim S. K. et al. Development of a nurse turnover prediction model in Korea using machine learning //Healthcare. – MDPI, 2023. – Т. 11. – №. 11. – С. 1583.

43. Olufunke B., Akinode J. L. Employee attrition prediction using machine

learning algorithms //International Conference. – 2022.

44. Paigude S. et al. Potential of artificial intelligence in boosting employee retention in the human resource industry //International Journal on Recent and Innovation Trends in Computing and Communication. – 2023. – Т. 11. – С. 01-10.

45. Rajula H. S. R. et al. Comparison of conventional statistical methods with machine learning in medicine: diagnosis, drug development, and treatment //Medicina. – 2020. – Т. 56. – №. 9. – С. 455.

46. Scikit-learn. URL: https://scikit-learn.ru/1-11-ensemble-methods/ (дата обращения: 01.06.2023).

47. Shehab M. et al. Machine learning in medical applications: A review of state-of-the-art methods //Computers in Biology and Medicine. – 2022. – Т. 145. – С. 105458.

48. Xu Y. et al. Predicting Nurse Turnover for Highly Imbalanced Data Using the Synthetic Minority Over-Sampling Technique and Machine Learning Algorithms //Healthcare. – MDPI, 2023. – Т. 11. – №. 24. – С. 3173.

Приложение А

Программный код разработанной модели

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import matplotlib.pyplot as plt
from sklearn.preprocessing import MinMaxScaler
from sklearn.model_selection import train_test_split
from sklearn.ensemble import  RandomForestClassifier
from sklearn.metrics import confusion_matrix, classification_report, accuracy_score, roc_auc_score, auc
from sklearn.tree import DecisionTreeClassifier
from sklearn.metrics import roc_curve, auc
from sklearn.ensemble import GradientBoostingClassifier
import numpy as np

dataset = pd.read_csv('HR_data.csv')
dataset.head(10)
dataset.info()
dataset.describe()
dataset['department'].unique()
dataset['left'].value_counts()

# Группируем данные по признаку "department" и подсчитываем количество сотрудников в каждом отделе
department_counts = dataset['department'].value_counts()
# Устанавливаем размер графика
plt.figure(figsize=(8, 8))
```

```
# Создаем круговой график
plt.pie(department_counts, labels=department_counts.index, autopct='%1.1f%%', startangle=90, colors=plt.cm.Paired.colors)

# Добавляем пустой круг в середину
centre_circle = plt.Circle((0, 0), 0.2, fc='white')
fig = plt.gcf()
fig.gca().add_artist(centre_circle)
# Устанавливаем аспект соотношения осей, чтобы круг был круглым
plt.axis('equal')
# Отображаем график
plt.show()

# Группируем данные по признаку "salary" и подсчитываем количество сотрудников в каждой категории
salary_counts = dataset['salary'].value_counts()
# Создаем гистограмму
plt.bar(salary_counts.index, salary_counts.values, color=plt.cm.Paired.colors)
# Добавляем подписи осей
plt.xlabel('Зарплата')
plt.ylabel('Количество сотрудников')
# Добавляем заголовок
# plt.title('Distribution of Employees by Salary')
# Отображаем гистограмму
plt.show()
pd.crosstab(dataset.department, dataset.left).plot(kind='bar')
```

```
plt.title('Частота текучести кадров для отдела')
plt.xlabel('Отдел')
plt.ylabel('Частота текучести кадров')
pd.crosstab(dataset.department, dataset.left)
table=pd.crosstab(dataset.salary, dataset.left)
table.div(table.sum(1).astype(float), axis=0).plot(kind='bar', stacked=True)
plt.title('Уровень зарплаты в сравнении с текучестью кадров')
plt.xlabel('Уровень зарплаты')
plt.ylabel('Доля сотрудников')
num_bins = 10

dataset.hist(bins=num_bins, figsize=(20,15))
plt.show()

# Применим one-hot encoding к категориальному признаку "department"
dataset = pd.get_dummies(dataset, columns=['department'], prefix='department')
dataset = pd.get_dummies(dataset, columns=['salary'], prefix='salary')
print(dataset.head(5))
# Выбираем признаки, которые требуют масштабирования
features_to_scale = ['number_project', 'average_montly_hours', 'time_spend_company']
# Создаем объект MinMaxScaler
scaler = MinMaxScaler()
# Применяем масштабирование только к выбранным признакам
dataset[features_to_scale] = scaler.fit_transform(dataset[features_to_scale])
# Результат: признаки теперь находятся в диапазоне от 0 до 1
dataset.head(5)
print(dataset.head(5))
```

```
print(dataset.describe())
dataset.info()
left_column = dataset.pop("left")
dataset["left"] = left_column
dataset.info()

X = dataset.iloc[:, 0:20].values # На каких данных обучаемся
y = dataset.iloc[:, 20].values # Целевой столбец
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.3, random_state = 0)
rf = RandomForestClassifier()
rf.fit(X_train, y_train)
print('Случайный лес:')
predictions_rf = rf.predict(X_test)
print(classification_report(y_test, predictions_rf))
print(confusion_matrix(y_test, predictions_rf))
print('Случайный лес =', accuracy_score(y_test, predictions_rf))
dt = DecisionTreeClassifier()
dt.fit(X_train, y_train)

print('Decision Trees:')
predictions_dt = dt.predict(X_test)
print(classification_report(y_test, predictions_dt))
print(confusion_matrix(y_test, predictions_dt))
print('Decision Trees =', accuracy_score(y_test, predictions_dt))

# Создаем модель градиентного бустинга
gb = GradientBoostingClassifier()
# Обучаем модель на обучающих данных
```

```
gb.fit(X_train, y_train)
print('Gradient Boosting:')
predictions_gb = gb.predict(X_test)
print(classification_report(y_test, predictions_gb))
print(confusion_matrix(y_test, predictions_gb))
print('Gradient Boosting =', accuracy_score(y_test, predictions_gb))

# LogisticRegression
from sklearn.linear_model import LogisticRegression
from sklearn import metrics
lr = LogisticRegression()
lr.fit(X_train, y_train)
print('Logistic Regression:')
predictions_lr = lr.predict(X_test)
print(classification_report(y_test, predictions_lr))
print(confusion_matrix(y_test, predictions_lr))
print('Logistic Regression =', accuracy_score(y_test, predictions_lr))

rf_roc_auc = roc_auc_score(y_test, rf.predict(X_test))
rf_fpr, rf_tpr, rf_thresholds = roc_curve(y_test, rf.predict_proba(X_test)[:,1])

dt_roc_auc = roc_auc_score(y_test, dt.predict(X_test))
dt_fpr, dt_tpr, dt_thresholds = roc_curve(y_test,
dt.predict_proba(X_test)[:,1])

gb_roc_auc = roc_auc_score(y_test, gb.predict(X_test))
gb_fpr, gb_tpr, gb_thresholds = roc_curve(y_test,
gb.predict_proba(X_test)[:,1])
```

```
lr_roc_auc = roc_auc_score(y_test, lr.predict(X_test))
lr_fpr, lr_tpr, lr_thresholds = roc_curve(y_test, lr.predict_proba(X_test)[:,1])

plt.figure()
plt.plot(rf_fpr, rf_tpr, label='RandomForestClassifier (area = %0.2f)' % rf_roc_auc)
plt.plot(dt_fpr, dt_tpr, label='DecisionTreeClassifier (area = %0.2f)' % dt_roc_auc)
plt.plot(rf_fpr, rf_tpr, label='GradientBoostingClassifier (area = %0.2f)' % gb_roc_auc)
plt.plot(rf_fpr, rf_tpr, label='Logistic Regression (area = %0.2f)' % lr_roc_auc)
plt.plot([0, 1], [0, 1],'r--')
plt.xlim([0.0, 1.0])
plt.ylim([0.0, 1.05])
plt.xlabel('False Positive Rate')
plt.ylabel('True Positive Rate')
plt.title('ROC')
plt.legend(loc="lower right")

feature_labels = dataset.columns.values[:19]
importance = rf.feature_importances_
feature_indexes_by_importance = importance.argsort()
for index in feature_indexes_by_importance:
    print('{}-{:.2f}%'.format(feature_labels[index], (importance[index] *100.0)))
```