МИНОБРНАУКИ РОССИИ

Федеральное государственное бюджетное образовательное учреждение
высшего образования
«САРАТОВСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н. Г. ЧЕРНЫШЕВСКОГО»

Кафедра дифференциальных уравнений и математической экономики

ВИЗУАЛЬНЫЙ АНАЛИЗ ДАННЫХ И РАЗРАБОТКА СИСТЕМЫ ПРОГНОЗИРОВАНИЯ ТЕКУЧЕСТИ КАДРОВ

БАКАЛАВРСКАЯ РАБОТА

студента 4 курса 451 группы
направления 38.03.05 — Бизнес-информатика

механико-математического факультета
Ковина Семёна Дмитриевича

Научный руководитель
доцент, к. ф.-м. н.,доцент _____ И. Ю. Выгодчикова

Заведующий кафедрой
зав.кафедрой, д.ф.-м.н., профессор _____ С. И. Дудов

Саратов 2022

СОДЕРЖАНИЕ

ОПРЕДЕЛЕНИЯ, ОБОЗНАЧЕНИЯ И СОКРАЩЕНИЯ

HR(Human Recources) - кадровая служба;

ML(Machine Learning) - машинное обучение;

ОКВЭД - общероссийский классификатор видов экономической деятельности;

EDA(Exploratory Data Analysis) - разведочный анализ данных;

Датасет, датафрейм - набор данных, таблица;

ВВЕДЕНИЕ

Сегодня эффективное и грамотное управление человеческими ресурсами является одним из главных условий успешного экономического развития строительных организаций. Ведь персонал – это главная движущая сила предприятия, выступающая в качестве трудовых ресурсов и являющаяся одновременно связующим звеном между техническими и экономическими факторами производственного процесса. Поэтому руководителям нельзя недооценивать значение человеческого фактора и его места в производственном процессе. Нельзя также пренебрегать психологическими и социологическими методами работы, так как современный HR-менеджмент имеет огромную методологическую базу, позволяющую подбор и управление кадрами сделать максимально эффективными [1].

Первоначальным этапом планирования кадровой работы считается прогнозирование, которое является базой при подготовке плановых решений (заданий). Кадровое прогнозирование в организациях применяют с целью предсказания изменения в структуре и динамике кадровой работы в перспективе, используя при этом анализ прошлого и настоящего опыта. Этот вид прогнозирования основывается на целях развития компании, которые нужно достичь на соответствующей перспективной стадии. Динамика развития предприятия (отрасли) обуславливается множеством причин, которые оказывают действие на данные экономические системы, а также существованием подвижной связи отдельных звеньев в структуре управления.

На практике методы кадрового прогнозирования и улучшения качественного состава работников на уровне предприятия (организации) пока не нашли широкого применения и используются эпизодически. При этом кадровая политика современности нуждается в их комплексном использовании в процессе разработки социально-экономических программ развития.

В данной работе проведен разведочный анализ данных о текучести кадров в компании АО «Неофлекс Консалтинг» и разработан веб-сервис для прогнозирования вероятности ухода сотрудника.

Актуальность темы. В современных условиях мало кто из руководителей российских предприятий задумывается о том, насколько дорого обходится

предприятию текучесть кадров. Так, по результатам проведенного в 2018 году крупнейшим в мире сообществом HR-профессионалов – «Общество по управлению человеческими ресурсами» (SHRM) исследования, было выявлено, что стоимость замены сотрудников низшего уровня (в работы входит: поиск кандидатов, проведение с ними собеседований, прием на работу, обучение, работа над повышением производительности) составляет 30 - 50 % от их годового оклада. Замена сотрудников среднего звена обходится предприятиям еще дороже - до 150 % от их годового оклада, а топ менеджмента - до 400 % [2].

В наши дни мало кто будет спорить с тем, что персонал – один из важнейших ресурсов организации. Большинство руководителей уже убедились в важности кадровой политики. Ведь только правильно подобранный трудовой коллектив, сплоченная команда способны эффективно реализовывать цели и задачи компании.

Обеспечение устойчивости кадрового состава на сегодняшний момент является одной из самых сложных проблем управления персоналом. Именно данный показатель в условиях экономического кризиса снижает негативные эффекты, действующие на организацию. Как правило, неустойчивость развития экономики в первую очередь проявляется как рост показателя текучести кадров в организации. Разумеется, текучесть кадров обусловлена множеством факторов, например, спецификой бизнеса, географическим фактором, стадией развития организации, а также структурным составом самих сотрудников. Поэтому для каждой компании важно определить свой уникальный уровень текучести кадров.

Теоретико-методологическая база. Проблемой текучести кадров занимались многие отечественными ученые. Г.В. Щекин [3] и А.В. Филиппов [24], например, в своих работах рассматривали основные причины текучести кадров и возможные пути их преодоления.

Проблематике управления в условиях высокой текучести персонала посвящены работы Ю.Г. Одегова, Т.Ю. Базарова, А.Я. Кибанова, Б.Л. Еремина, П.В. Журавлева, Л.В. Карташовой, В.Р. Веснина, В.И. Маслова. В.И. Кабалина, А.А. Никишов, И.М. Козина в своих исследованиях делали упор на внутрифирменную политику и управление процессом эффективности персо-

нала.

Но несмотря на то, что проблема текучести кадров достаточно освещена в научной литературе, новизна данного исследования состоит в том, что проблема текучести кадров изучается в конкретной организации, более того разрабатывается веб-сервис, который способен предсказать вероятность ухода сотрудника из данной компании.

Целью бакалаврской работы стали проведения визуального анализа данных о проблеме текучести кадров в АО «Неофлекс Консалтинг» и разработка системы прогнозирования вероятности ухода сотрудника из этой организации.

В рамках бакалаврской работы были поставлены следующие основные **задачи**:

1. Изучить теоретическое обоснование проблематики текучести кадров.

2. Выполнить предобработку предоставленных данных о сотрудниках и подготовить их для машинного обучения.

3. Провести визуальный анализ данных и построить графики по статистическим показателям.

4. Построить модель классификации сотрудников относительно вероятности утечки кадра.

5. Построить регрессионную модель прогнозирования количества лет и месяцев, которые проработает сотрудник в данной организации.

6. Реализовать веб-приложение для удобного использования разработанных моделей прогнозирования.

Объектом исследования является АО «Неофлекс Консалтинг».

Предметом исследования являются данные о сотрудниках и их текучести из организации АО «Неофлекс Консалтинг».

Структура и содержание бакалаврской работы. Работа состоит из введения, трех разделов, заключения, списка использованных источников и приложений.

1 Теоретические основы исследования текучести кадров

Кадры предприятия – это количество работников различных профессионально – квалификационных групп, которые заняты в организации и входят в ее списочный состав. В списочный состав включают всех работников, кто принят на работу, связанную с основной деятельностью и с не основной.Если не правильно подобрать кадры, то компания, предприятие или организация не смогут эффективно и качественно функционировать. Каждая организация при этом имеет ряд своих особенностей. В условиях рынка и конкуренции, которая ежедневно растет и усиливается, проблемы становятся наиболее актуальными, если они связаны с кадрами. Если у компании есть специалисты, которые могут решать задачи, поставленные перед ними, то данная компания может достичь самых высоких целей и любых прибылей.

1.1 Методы изучения текучести кадров в организации

Сегодня текучесть кадров - одна из наиболее важных проблем, с которыми сталкиваются современные предприятия. Главным внутренним ресурсом любой организации являются сотрудники, осуществляющие совместную трудовую деятельность, направленную на удовлетворение определенных потребностей человеческого общества [4]. Поскольку уровень сложности и качества человеческих потребностей на сегодня достаточно высок и постоянно увеличивается, то и сложность труда, направленного на их удовлетворение, со временем увеличивается.

Чтобы работа выполнялась качественно, необходим соответствующий уровень профессиональной подготовки сотрудников. Развивающаяся организация должна заботиться о повышении уровня квалификации своих сотрудников и не менее важным становится фактор сохранения и увеличения кадрового состава.

Подготовка хорошего специалиста это сложный и трудоемкий процесс, который требует больших временных, финансовых, материальных и духовных затрат. В этом процессе участвуют: сам человек со своими способностями, умениями, навыками и другими личными качествами; его семья; государство, которое создает условия для получения человеком фундаментальных основ необходимых обществу профессий; организации и предприятия, где

осуществляется непосредственная трудовая деятельность [5].

Исключая индивидуальный труд, человек получает специализацию своих навыков и умений именно на уровне организации или предприятия, именно здесь происходит его профессиональный рост. Даже если на предприятие приходит специалист высокого уровня, например, переходит с другого места работы, его необходимо ознакомить с регламентами работы предприятия, с проектами, над которыми ему предстоит работать, с применяемым программным обеспечением, с коллективом и т.д. Поэтому, даже в случае, когда квалифицированный сотрудник заранее предупреждает о своем увольнении, у предприятия возникает множество проблем.

Текучесть кадров - насущная проблема российской экономики и международного рынка труда в целом [10]. Если процесс найма и увольнения сотрудников носит постоянный характер, то это свидетельствует о наличии проблем в организации бизнес-процессов или об игнорировании руководством необходимости поддержания внутреннего микроклимата в компании. Высокий процент текучести кадров – показатель разрозненности трудового коллектива. Словарь управления персоналом [11] дает следующее определение этому явлению - «движение рабочей силы, обусловленное неудовлетворенностью работника рабочим местом или неудовлетворенностью организации конкретным работником».

Современные ученые делают упор на то, что текучесть возникает в первую очередь из-за неудовлетворѐнности работника условиями труда или рабочим местом. Немаловажно также учитывать удовлетворенность организации конкретным сотрудником. По мнению Л. Никифоровой [3, с. 89], текучесть кадров и ее влияние на результаты деятельности компании следует рассматривать как совокупность количественного и качественного аспектов. При рассмотрении количественного аспекта нужно обратить внимание на то, что существует два уровня текучести кадров: естественный и повышенный.

Естественный уровень колеблется в пределах от 3 до 5% от общей численности всех сотрудников организации. Повышенный же, как правило, ведет к значительным экономическим потерям.

Говоря о естественном уровне текучести кадров, нужно понимать, что данный показатель способствует обновлению коллектива предприятия. С эко-

номической точки зрения, это непрерывный и необходимый для организации процесс. Он не требует вмешательства со стороны руководства и кадровых служб. Вполне закономерно, что часть сотрудников может уйти на пенсию, а часть уволиться по другим причинам. Данная ситуация характерна для любой компании. Она позволяет открыть в коллективе возможности для ротации кадров и построения карьерных лестниц для сотрудников, что выступает еще одним инструментом мотивации и стимулирования.

В ситуации, когда текучесть кадров значительно превышает 5%, издержки становится критичными, поскольку возрастают с увеличением потока кадров. В этом случае предприятие терпит убытки. Они выражаются в виде упущенной прибыли, а также снижении производительности труда. Высокие показатели текучести кадров уменьшают укомплектованность штата. Такая ситуация требует от опытных, высококвалифицированных сотрудников временных затрат на обучение новичков, что отвлекает их от работы и снижает их продуктивность. Постоянная смена сотрудников в коллективе ухудшает морально-психологический климат, мешает созданию слаженной команды и существенно снижает производительность труда. Разумеется, коллектив организации не является постоянным по численному составу и уровни квалификации: одни работники увольняются, на их место приходят другие. Для того, чтобы объективно проанализировать изменение численности и состава сотрудников используются различные показатели.

Наиболее часто используются такие показатели, как [12]:

1. показатель среднесписочной численности сотрудников

$$\overline{P} = \frac{\frac{1}{2}P_1 + P_2 + ... + P_{11} + \frac{1}{2}P_{12}}{12}, \tag{1}$$

где $P_1, P_2 ... P_{11}, P_{12}$ - численность принятых работников.

2. коэффициент приема;

$$\mathrm{K_п} = \frac{P_\mathrm{П}}{\overline{P}}, \tag{2}$$

где $P_\mathrm{п}$ - численность принятых работников;
$\overline{P}$ - среднесписочная численность персонала.

3. коэффициент выбытия;

$$\mathrm{K_в} = \frac{P_{\text{ув}}}{\overline{P}}, \tag{3}$$

где $P_{\text{ув}}$ - численность уволенных работников;
$\overline{P}$ - среднесписочная численность персонала.

4. коэффициент стабильности;

$$\mathrm{K_с} = \frac{1 - P_{\text{ув}}}{\overline{P} + P_{\text{П}}}, \tag{4}$$

где $P_{\text{ув}}$ - численность работников, уволившихся с предприятия по собственному желанию и из-за нарушения трудовой дисциплины за отчетный период;
$\overline{P}$ - среднесписочная численность работающих на данном предприятии в период, предшествующий отчетному;
$P_{\text{п}}$ - численность вновь принятых за отчетный период работников.

Для того чтобы рассчитать коэффициент текучести кадров (идентично формуле 3) нужно определить отношение численности работников предприятия или отдельного структурного подразделения, выбывших за отчѐтный период, к среднесписочной численности сотрудников за тот же период.

Необходимо учитывать и специфические особенности бизнеса конкретной компании, анализируя состояние дел с текучестью персонала. Для большинства компаний можно подобрать свой индивидуальный порог процента естественной текучести. Данная норма определяется на основании следующих факторов:

1. специфика отрасли;
2. категория персонала (линейный персонал или же управленческий);
3. конкурентоспособность организации;
4. территориальное расположение организации;
5. кадровая политика;
6. стиль управления.

К текучести кадров нельзя подходить однозначно. Ее следует рассматривать с трех позиций: экономики отрасли (региона, страны), предприятия и человека. Хотя все они взаимосвязаны, каждая из них может быть рассмотрена

самостоятельно.

В настоящее время на текучесть уже не смотрят как на сугубо отрицательное явление, с которым нужно бороться вплоть до полной ликвидации. Как это происходило в период плановой, командной экономики [6]. Напротив, трудовая мобильность работников сегодня рассматривается, как условие осуществления процесса производства Проблема управления заключается в реорганизации неорганизованного перемещения трудящихся с одного предприятия на другое в организованное, регулируемое. Конечной целью анализа текучести кадров является поиск методов регулирования и управления этим процессом в желательном направлении. При таком подходе оценка экономической роли текучести может быть противоречивой.

С одной стороны, можно говорить, что текучесть кадров ведет к ряду отрицательных последствий, таких как:

• сбои в производственном механизме предприятия, приводящие к ряду экономических потерь;

• снижение качества его трудовых ресурсов;

• потери, вызванные простоями оборудования;

• излишние затраты на подбор кадров и адаптацию работников;

• проблемы работников, оторванных от привычной работы несущих определенные материальные и психологические потери;

С другой стороны, текучесть кадров можно рассматривать в качестве положительного явления, поскольку этот процесс выполняет ряд важных позитивных функций, так как способствует:

• межотраслевому и территориальному перераспределению рабочей силы;

• квалификационно-профессиональному продвижению кадров;

• повышению благосостояния и развитию людей, если перемещение идет по карьерной лестнице.

Кроме того, полное отсутствие трудовых перемещений в организации приводит к «окостенению» структуры коллектива. Для изучения текучести, как и других форм движения рабочей силы, необходимы количественные показатели, которые бы достаточно правильно характеризовали уровень и динамику рассматриваемого явления,

1.2 Классификация факторов, вызывающих текучесть кадров

Факторы, которые вызывают текучесть персонала, обладают разными источниками, различной силой влияния, которая достаточно изменчива и трудно поддается количественной оценке [7]. Наиболее популярной классификации факторов является деление их условно на три группы:

1. Группа внутренних факторов:

1.1 Низкая заработная плата. Ставки оплаты труда, которые неспособны конкурировать, не могут привлечь новых специалистов и заставляют сотрудников искать более выгодные предложения. Постоянные задержки заработной платы также влияют на текучесть. В этой ситуации не важна величина заработной платы, так как стоимость ее фактически снижается. Если в компании небольшая, но стабильная заработная плата, то текучесть может быть меньше, чем в организациях, где уровень выше, но существуют проблемы с своевременностью. Нарушается уверенность сотрудников в будущем. Невозможность заработать больше также может стать причиной увольнений сотрудников.

1.2 Несправедливость заработной платы. Ее величина несправедлива, то есть отсутствует связь результатов труда в соотношении зарплат сотрудников различных подразделений, резкой разницей зарплат сотрудников одной специальности в организации одного того же региона.

1.3 Нет карьерного роста. Это одна из самых распространенных причин ухода сотрудников. Проработав в компании около года многие сотрудники уже думает о повышении, а работодатель не может продвинуть его по служебной лестнице, даже если он того заслуживает. Повышение заработной платы может успокоить его, но не навсегда.

1.4 Отношения с коллективом и руководством, которые не складываются, постоянный дискомфорт на работе, так же зачастую становятся сильным побудительным мотивом к увольнению при высоком уровне зарплаты в престижной компании и на солидной должности.

1.5 Однообразная работа – частая причина высокой текучести персонала. К быстрому профессиональному выгоранию, усталости, приводят такие факторы как стресс, монотонность и скука. В конечном итоге, испытывая все это, у сотрудника возникает желание сменить место работы [8].

2. Группа внешних факторов:

2.1 Демографическаяситуация в регионе;

2.2 Экономическаяситуация в регионе;

2.3 Семейные обстоятельства;

2.4 Появление новых предприятий.

Одним из наиболее распространенным внешних факторов является демографическая ситуация в стране. К примеру, демографические ямы (провалы в рождаемости) негативно скажутся на состоянии рынка труда, снижая количество имеющихся на нѐм специалистов. Высокий дефицит персонала приводит к конкуренции, и, как следствие, повышает текучесть кадров, вынуждая работодателей вступать в борьбу за каждого квалифицированного специалиста. Это сказывается и на стоимости трудовых ресурсов, повышая издержки компании. Компании, которые не могут позволить себе повысить заработную плату, теряют свою конкурентоспособность.

3. Группа личностных факторов:

3.1 Возраст;

3.2 Опыт работы;

3.3 Квалификации;

3.4 Уровень их образования.

Самый пик переходов из одной организации в другую заканчивается в 25 – 30 лет. Чаще всех меняют работу сотрудники, у которых низкая квалификация, нет семьи, отсутствуют перспективы, небольшая заработная плата и те, кто живут территориально далеко от рабочего здания.

Если есть определенная неудовлетворенность демографической ситуацией и качеством трудовых ресурсов при экономическом росте, то это все рано или поздно приведет к дефициту всех и квалифицированных и неквалифицированных сотрудников.

Текучесть персонала, по мнению В. Свистунова и М. Тюленевой [9] классифицируется следующим образом:

1) активная;

2) пассивная;

3) внутриорганизационная;

4) внешняя.

Активная форма текучести возникает в случае неудовлетворѐнности

работника следующими факторами: рабочее место, содержание работы, условия труда. Пассивная же форма текучести сотрудников возникает в связи с неудовлетворенностью организации сотрудникам. Иными словами, пассивная текучесть включает в себя случаи выбытия сотрудников вследствие несоблюдения трудовой дисциплины или неудовлетворительных результатов труда. То есть, активные и пассивные текучести кадров различаются посредством причин увольнения сотрудников.

Ещѐ одна форма текучести – внутриорганизационная. Она возникает в связи с ротацией сотрудников внутри организации. При перемещении сотрудников из одной организации в другую следует говорить о внешней текучести персонала.

Существует множество классификаций факторов, влияющих на текучесть кадров. В качестве примера ниже изображены иные группы факторов, в соответствии с рисунком 1.1.

Рисунок 1.1 – Классификация мотивов текучести кадров

Все мотивы тесно взаимосвязаны. Их группировка, разумеется, может быть иной. В структуре мотивов выделяют главные, определяющие принятие решения об уходе, и второстепенные, подкрепляющие это решение. Мотивы текучести различаются по группам работников в зависимости от возраста,

образования, профессии, уровня квалификации.

1.3 Подходы к управлению текучестью кадров

Проблемы текучести кадров решаются исходя из того, что происходит на конкретном предприятии. Управление текучестью кадров предполагает поэтапное проведение мероприятий. Они должны быть выполнены кадровой службой предприятия. Управление текучестью кадров выполняется по хронологии следующих этапов [13]:

1. Установить уровень текучести кадров. При этом определяется количественный показатель и устанавливается его значение и отклонение от среднего. Выявляется вероятность необоснованных экономических потерь. Принято считать нормативным уровень в 3–5%. Но он не должен восприниматься как некий индикатор, поскольку движение кадров на конкретном предприятии осуществляется под воздействием совокупности факторов: отраслевой принадлежности, технологии производства, трудоемкости работ, наличия или отсутствия фактора сезонности в производственном цикле, стиля руководства, уровня и принципов корпоративной культуры. Поэтому при определении уровня текучести кадров следует провести анализ динамики трудовых показателей предприятия за возможно больший отрезок времени, выявить наличие и величину сезонных колебаний текучести.

2. Установить уровень, величину экономических потерь, к которым приведет текучесть кадров. Это один из наиболее трудоемких, но очень важных этапов. Для его осуществления требуются дополнительные расчеты. Эти данные, складываются из анализа ущерба и дополнительных затрат, связанных со следующими факторами:

- потеря времени в работе;
- затраты на обучение и переобучение новых работников;
- падение выработки у работников перед увольнением;
- низкий уровень выработки у вновь принятых работников;
- затраты на подбор персонала уволившегося по причинам, относимым текучести кадров;
- брак, вызванный недостаточной профессиональной адаптацией у вновь поступивших работников.

3. Выяснение причин увольнения работников вообще и, в том числе, по причинам текучести кадров. Текучесть кадров может быть вызвана отраслевой особенностью деятельности предприятия или несовершенством системы управления им. Затем необходимо определить структуру мотивов текучести кадров. Она основана на реальных причинах, вызывающих у работника желание принять решение об увольнении с предприятия.

4. Определение системы мероприятий, направленных на стабилизацию коллектива. Для этого принимаются три группы мер по сокращению текучести кадров:

- улучшение условий труда, совершенствование системы материального стимулирования, организации и управления производством и др., относящееся к технико-экономическим мерам;

- совершенствование технологий трудовой адаптации, системы профессионального продвижения и др., относящееся к организационным мерам;

- улучшение стилей и методов руководства и взаимоотношений в коллективе, системы морального поощрения и др. относящееся к социально-психологическим мерам.

5. Расчет показателей эффективности от внедрения разработанных мер (проведение мониторинга).

После этого проводится сравнительный анализ затрат на проведение мероприятий по оптимизации текучести кадров и убытка из-за высокого уровня текучести. Если затраты на решение проблемы превысят потери, вызванные высокой текучестью кадров, то необходимо найти новые, другие, оптимальные варианты работы с персоналом.

Итак, выяснив причину увольнения сотрудников, необходимо принимать меры по ее устранению. Например, если причина увольнения в уровне заработной платы, необходимо выяснить, есть ли финансовая возможность увеличения заработных плат, увеличения премиальных или другие варианты финансовой мотивации персонала. Если же часть сотрудников не устраивали условия труда, то необходимо принять решение о возможности их улучшения. Причины увольнения персонала и пути решения изображены ниже, в соответствии с рисунком 1.2.

Рисунок 1.2 – Классификация мотивов текучести кадров

Если оказывается, что большинство уволившихся имеют стаж работы до 6 месяцев, это указывает на ошибки в подборе персонала и его адаптации. В таком случае следует пересмотреть критерии отбора специалистов, улучшить процесс адаптации, возможно необходимо ввести кураторство опытного сотрудника, контролировать процесс адаптации каждого, провести тренинг и т. д. Бывают ситуации, когда текучка наблюдается у конкретного руководителя, в этом случае необходимо обсудить с ним причины увольнений его сотрудников, обучить его эффективному управлению и взаимодействию с

коллективом.

Таким образом, на основании рассмотренного теоретического материала можно сделать следующие **выводы:**

1) Текучесть - это движение работников как из штата организации, так и в штат. Основной способ отслеживания текучести – это регистрация уволившихся и принятых сотрудников.

2) Существует два уровня текучести кадров: естественный(3-5%) и повышенный(>5%).

3) Для оценки текучести кадров используются такие показатели, как:

1. показатель среднесписочной численности сотрудников;

2. коэффициент приема;

3. коэффициент выбытия;

4. коэффициент стабильности.

4) Наиболее часто встречаются следующие причины увольнения персонала:

- низкая заработная плата;
- задержки заработной платы;
- система оплаты труда;
- премирование труда;
- отсутствие карьерного роста;
- социально-психологический климат;
- однообразие выполняемой работы;
- экономическая и демографическая ситуация в стране.

5) Факторы увольнения сотрудников делятся на три группы. Они бывают внешними по отношению к организации, внутренним по отношению к организации, личностные, характеризующие сотрудника.

2 Анализ данных о текучести кадров в компании АО «Неофлекс Консалтинг»

Анализ ситуации с кадрами в подразделениях — это сбор и обработка информации о сотрудниках, которая помогает отладить работу конкретных подразделений или функциональных служб. Имея перед глазами результаты анализа, руководство сможет понять, к примеру, в каких отделах недопустимо высокая текучесть кадров и эту проблему нужно будет срочно решать, а в каких филиалах в этом смысле все благополучно. Сотрудники, занимающиеся кадровой аналитикой, объединяют данные об индивидуальных достижениях каждого по основным оцениваемым направлениям с показателями стоимости привлечения данного сотрудника и отношения специалистов к своему делу, готовности долго работать в компании.

2.1 Организацонно-экономическая характеристика

Полное наименование предприятия: Акционерное общество «Неофлекс Консалтинг». Сокращенное: АО «Неофлекс».

Место нахождения: 127015, г. Москва, ул. Вятская, 35 стр. 4.

Генеральный директор организации: Рубан Олег Викторович.

АО «Неофлекс» - поставщик услуг в области разработки программного обеспечения и внедрения сложных информационных систем. Компания создает ИТ-платформы для цифровой трансформации бизнеса, помогая заказчикам получать устойчивые конкурентные преимущества в цифровую эпоху. АО «Неофлекс» сотрудничает в основном с организациями из финансового сектора рынка, совершенствуя технологии банковской и финансовой деятельности клиентов, принося банкам не только готовые продукты, но также экспертизу, платформы и компоненты, на базе которых можно быстро построить сложные решения. При таком подходе тиражируются уже не продукты, а определенные паттерны, подходы, элементы архитектуры и программные компоненты.

Основной вид деятельности - «Разработка компьютерного программного обеспечения».

АО «Неофлекс Консалтинг» также осуществляет следующие виды деятельности(по ОКВЭД):

- 62.01 - Разработка компьютерного программного обеспечения;

- 62.09 - Деятельность, связанная с использованием вычислительной техники и информационных технологий;
- 63.11 - Деятельность по обработке данных, предоставление услуг по размещению информации и связанная с этим деятельность;
- 62.02 - Деятельность консультативная и работы в области компьютерных технологий;
- 63.11.1 - Деятельность по созданию и использованию баз данных и информационных ресурсов;
- 95.11 - Ремонт компьютеров и периферийного компьютерного оборудования
- 47.63.1 - Торговля розничная музыкальными записями, аудиолентами, компакт-дисками и кассетами в специализированных магазинах;
- 47.63.2 - Торговля розничная лентами и дисками без записей в специализированных магазинах;
- 46.15.4 - Деятельность агентов по оптовой торговле радио- и телеаппаратурой, техническими носителями информации;
- 73.20.1 - Исследование конъюнктуры рынка.

АО «Неофлекс Консалтинг» была создана в феврале 2005 года. В 2008 году сформулирована продуктовая стратегия Neoflex и начинается разработка первых собственных программных продуктов Neoflex FrontOffice и Neoflex Reporting. Официальную регистрацию компания получила 21 мая 2009. В 2011 году открывается филиал организации в городе Саратов. С каждым годом далее компания бурно развивалась.

У компании АО «Неофлекс Консалтинг» есть торговые марки, общее количество - 2, среди них «NEOFLEX». Первая торговая марка зарегистрирована 8 апреля 2010 г. - действительна до 8 апреля 2019 г. Последняя торговая марка зарегистрирована 12 апреля 2019 г. и действительна до 13 сентября 2028 г.

Для визуального представления развития компании с 2012 по 2018 года на рисунках 2.1 и 2.2 представлены графики по выручки и прибыли компании за этот период.

Рисунок 2.1 – Выручка АО «Неофлекс Консалтинг» с 2012 по 2018 г.

Рисунок 2.2 – Прибыль АО «Неофлекс Консалтинг» с 2012 по 2018 г.

По графикам выше можно заметить, что после падения выручки в 2013-2014 гг. наблюдается трёхлетний пероид стагнации в компании. Невзирая на это, с 2016 года явно виден восходящий тренд выручки и прибыли.

Для более подробного описания финансовых результатов компании ниже приведена таблица основных показателей бухгатерской отчётности.

Таблица 2.1 – Финансовая отчетность АО «Неофлекс Консалтинг» за 2013-2020 гг.

Код	2013	2014	2015	2016	2017	2018	2019	2020
2110 Выручка	**791 905**	**726 605**	**729 132**	**725 341**	**972 223**	**1 182 626**	**1 496 759**	**2 077 392**
2120 Себестоимость продаж	**485 594**	**458 765**	**475 733**	**469 365**	**488 802**	**552 306**	**716 628**	**947 118**
2100 Валовая прибыль (убыток)	**306 311**	**267 840**	**253 399**	**255 976**	**483 421**	**630 320**	**780 131**	**1 130 274**
2210 Коммерческие расходы	**8 983**	**8 327**	**6 797**	**14 893**	**10 054**	**3 172**	**1 571**	**1 466**
2220 Управленческие расходы	**214 453**	**232 693**	**205 545**	**249 198**	**337 712**	**477 672**	**589 568**	**716 620**
2200 Прибыль (убыток) от продаж	**82 875**	**26 820**	**41 057**	**-8 115**	**135 655**	**149 476**	**188 992**	**412 188**

На данный момент организация имеет филиалы в следующих городах Российской Федерации:

- Воронеж;
- Москва;
- Санкт-Петербург;
- Нижний Новгород;
- Саратов;
- Самара;
- Новосибирск;
- Краснодар;
- Пенза.

За свою историю существования АО «Неофлекс Консалтинг» разработало множество своих продуктов и участвовало в более чем 100 различных проектах, наиболее успешными и обширными по объему привлеченного персонала являются следующие бизнес направления:

- SOA;
- NFO;
- SOA-support;
- BigData Solutions;
- SOA-ВТБ;
- Neoflex Reporting;
- АРГО;
- Департамент аналитики;
- Финансовые рынки и риски;
- FastData;
- Инженерные практики;
- Департамент управления проектами;
- Департамент разработки хранилищ данных;
- Служба персонала;
- ФАС;
- Департамент тестирования;
- Менеджмент;
- Data Science;

- DataGram.

2.2 Методы предобработки данных

Для дальнейшего анализа и оценки необходимо провести предобработку исходных (необработанных) данных. Это процесс позволит привести необработанные данные к соответствующим требованиям для решения исходной задачи. Реальные данные собираются для последующей обработки из разных источников и процессов. Они могут содержать ошибки и повреждения, негативно влияющие на качество набора данных.

Предварительная обработка данных – это метод интеллектуального анализа данных, который является важнейшим этапом, от качества выполнения которого зависит возможность получения качественных результатов всего процесса [15]. Поскольку данные поступают из нескольких источников, исходные данные часто бывают неполными, непоследовательными или могут возникнуть проблемы из-за человеческих ошибок.

Предобработка данных включает два направления: очистку и оптимизацию.

Очистка производится с целью исключения различного рода факторов, снижающих качество данных и мешающих работе аналитических алгоритмов. Она включает обработку дубликатов, противоречий и фиктивных значений, восстановление и заполнение пропусков, сглаживание, подавление шума и редактирование аномальных значений. Кроме этого, в процессе очистки восстанавливаются нарушения структуры, полноты и целостности данных, преобразуются некорректные форматы.

Оптимизация данных как элемент предобработки включает снижение размерности, выявление и исключение незначащих признаков. Основное отличие оптимизации от очистки в том, что факторы, устраняемые в процессе очистки, существенно снижают точность решения задачи или делают работу аналитических алгоритмов невозможной. Проблемы, решаемые при оптимизации, адаптируют данные к конкретной задаче и повышают эффективность их анализа.

Основные задачи предварительной обработки данных:

- Очистка данных - заполнение отсутствующих значений, обнаружение и

удаление шума данных и выбросов;

- Преобразование данных - нормализация данных для уменьшения размеров и шума;
- Уменьшение данных - образцы записей данных или атрибутов для упрощения обработки данных;
- Дискретизация данных - преобразование непрерывных атрибутов в атрибуты категорий, чтобы упростить их использование с определенными методами машинного обучения;
- Очистка текста - удаление внедренных символов, которые могут нарушать выравнивание данных, например внедренных символов табуляции в файле с разделителем-табуляцией, внедренных новых линий, которые могут, например, разбивать записи.

Одним из первых этапов предобработки данных является обезличивание персональных данных. Для проведения статистических, социологических, исторических, медицинских и других научных и практических исследований держатель (обладатель) массива персональных данных обезличивает используемые данные, придавая им форму анонимных сведений. Режим конфиденциальности, установленный для персональных данных, снимается. Обезличивание должно исключать возможность идентификации субъекта персональных данных [16].

К методам обезличивания персональных данных относятся:

• метод введения идентификаторов – замена части сведений (значений персональных данных) идентификаторами с созданием таблицы соответствия идентификаторов исходным данным;

• метод изменения состава или семантики – изменение состава или семантики персональных данных путем замены результатами статистической обработки, обобщения или удаления части сведений;

• метод декомпозиции – разбиение множества (массива) персональных данных на несколько подмножеств (частей) с последующим раздельным хранением подмножеств;

• метод перемешивания – перестановка отдельных записей, а также групп записей в массиве персональных данных [17].

Очень часто в наборе данных есть пропущенные значения. Это могло

произойти во время сбора данных или объединения из нескольких наборов данных. Чтобы справиться с отсутствующими данными, можно использовать несколько подходов:

• устранить с помощью удаления полей с недостающими данными;

• заменить пропущенное значение вручную;

• игнорировать пропуски;

• заполнение их средним или медианным значением.

• подстановка по регрессии — использование регрессионной модели для замены пропущенных значений регрессионными данными.

Большой объем дополнительных бессмысленных данных называется шумом. Это могут быть:

• дубликаты записей данных;

• сегменты данных, не представляющие ценности для конкретного исследования;

• ненужные информационные поля для каждой из переменных. Для решения этой проблемы можно применить один из следующих методов:

• регрессионный анализ – помогает решить, какие переменные действительно имеют влияние. Он необходим для сглаживания больших объемов данных.

• применение алгоритмов кластеризации для группировки данных.

При этом необходимо уделить особое внимание выбросам. Выбросы – это особые точки данных, не похожие на остальную часть домена. Важно не подменять выбросы, воспринимая их как шум.

Данные могут содержать противоречивые значения. Это может быть связано с человеческой ошибкой или информация была неправильно прочитана при сканировании с рукописного бланка. Поэтому всегда рекомендуется выполнять оценку данных, знать, каким должен быть тип данных и является ли он одинаковым для всех объектов данных.

Помимо всего, набор данных может включать в себя объекты данных, которые дублируют друг друга. В большинстве случаев дубликаты удаляются, чтобы не дать этому конкретному объекту данных преимущества и повлиять на конечный результат.

Предобработка исходных данных – одна из наиболее актуальных задач

интеллектуального анализа. Плохое качество данных является одной из самых больших проблем при построении аналитических решений, так как на основе некорректной информации делаются неверные выводы. Нарушение полноты данных приводит к смещению основных статистических характеристик, таких как математическое ожидание или дисперсия, например, возрастает прямо пропорционально числу пропусков, и как следствие к искажению выводов, которые могут быть сделаны по результатам исследования и принятию неверных стратегических решений. Поэтому стоит уделить особое внимание данному этапу.

2.3 Исходные данные и подготовка их для анализа

Важнейшим этапом создания модели машинного обучения является этап сбора и обработки данных. На данном этапе необходимо подготовить набор данных, с помощью которого в дальнейшем будет обучена модель и произведена оценка еѐ качества. Набор данных – это таблица, строки которой содержат образцы для обучения, а столбцы – атрибуты.

Для анализа и визуализации данных был выбран язык программирования Python [18]. Python является отличным инструментом для работы с данными, их обработке и визуализации. Он прост в изучении, достаточно мощный и гибкий, позволяет работать с большими объемами данных. В нем уже существует огромное количество готовых библиотек для визуализации, машинного обучения и интеллектуального анализа данных. Так же его использует большое число людей и организаций, что способствует его быстрому развитию. Язык является кроссплатформенным и поддерживает почти все современные системы.

На языке Python реализованы библиотека NumPy [22], необходимая для обработки массивов, предназначенная для эффективной работы с большими многомерными массивами произвольных записей без потери слишком большой скорости и Pandas [19] – программная библиотека для обработки и анализа данных. Работа Pandas с данными строится поверх библиотеки NumPy. Для построения графиков использовалась библиотека Pyplot [21], которая предоставляет собой процедурный интерфейс к объектно-ориентированной библиотеке построения графиков Matplotlib [20].

Компания АО «Неофлекс Консалтинг» предоставила данные о своих сотрудниках в виде трёх «сырых» таблиц - Employees, Grades и Cities. Атрибуты каждой таблицы следующие:

1) Employees:
   - Табельный номер
   - Сотрудник
   - Дата рождения
   - Дата приема
   - Дата увольнения
   - Физическое лицо.Пол
2) Grades:
   - Дата
   - Сотрудник
   - Бизнес-направление
   - Грейд
3) Cities:
   - Дата
   - Сотрудник
   - Филиал
   - Город

Первая таблица содержит 2338 экземпляров, вторая - 7542, третья - 6296. Фрагменты каждой таблицы в первоначальном виде приведены в Приложении А.

На основе приведенных выше атрибутов предоставленных данных, можно заметить, что каждая таблица содержит столбец «Дата», это стоит описать подробнее. В таблице Grades «Дата» обозначает дату присвоения определенного грейда, другими словами «коэффициент компетентности специалиста». Из дальнейшего анализа данных, выясняется, что это может быть датой устройства сотрудника в компанию впервые, либо движение персонала внутри компании(например дата повышения в должности). Столбец «Дата» в таблице Cities обозначает очень схожую информацию, такую как дату устройства сотрудника в определенный филиал, дату его повышения, либо дату релокации в другой филиал. Столбец «Город» показывает родной город сотрудника,

откуда он приехал, а столбец «Филиал» показывает в какой из филиалов организации он устроился работать. Столбцы «Сотрудник» в каждой таблице, которые должны показывать ФИО сотрудников, зашифрованы. Обезличивание персональных данных выполнено путем введения идентификаторов.

В таблице Employees «Дата увольнения», обозначена только у тех сотрудников, которые увольнялись из компании, либо (снова как выяснилось из дальнейшего анализа данных) сотрудник релоцировался в другой филиал или перевелся в другой отдел(сменил бизнес-направление). Соответственно поле «Дата увольнения» пустое у сотрудников, продолжающих свою деятельность в компании. Табельный номер был дан далеко не у всех сотрудников, причина пропусков неизвестна.

Исходя из вышесказанного, можно понять, что предоставленные данные являются достаточно «грязными» и имеют много проблематичных ньюансов, такие как пропущенные и повторяющиеся значения(особенно с датами), дисбаланс количества экземпляров каждой таблицы. Предобработка первоначальных данных явно необходима для проведения корректного анализа данных и построения моделей ML.

Для удобной и эффективной дальнейшей работы с данными необходимо агрегировать 3 исходные таблицы в единый датафрейм(набор данных или таблица). Так как далее будет строиться модель ML, нужно чтобы все атрибуты имели числовые значения, иначе алгоритм машинного обучения не сможет обучаться на данных. Также добавлены новые атрибуты(столбцы) по имеющимся данным для дальнейшего обучения моделей ML по ним. Дубликаты были удалены. Реализации предобработки данных выполнена по следующим этапам:

1) Удалены повторяющиеся сотрудники с одинаковыми датами приема;

2) Добавлен бинарный атрибут «Ушел», обозначаемый 0, если сотрудник продолжает свою деятельность в компании и 1, если сотрудник ушел из компании;

3) Добавлен столбец «Стаж», вычисляется как «Дата увольнения» - «Дата приема», и «2021-05-17» - «Дата приема», если сотрудник продолжает работу в компании;

4) Все атрибуты с датами (изначально имели тип данных строки) пере-

ведены в тип данных «Datetime»;

5) Таблица Employees объединена с таблицей Grades путем «Left join» (возвращает все строки из левой таблицы, даже если в правой таблице нет совпадений) в новый датафрейм «df»;

6) Df объединен с таблицей Cities;

7) Удалены дублированные сотрудники с одинаковыми бизнес-направлениями и филиалами;

8) Добавлен атрибут возраста сотрудника, вычисляется как «2021-05-17» - «Дата рождения», после чего полученная дата переводится в количество пройденных часов и делится целочисленно на 8640;

9) Добавлен бинарный атрибут «Пол», мужской пол обозначен как 1, а женский как 0.

Код на языке Python выглядит следующим образом:

```
# 1
employees.drop_duplicates(subset=['Сотрудник','Дата
↪  приема'],ignore_index=True,keep = 'first',inplace=True)
# 2
def addleft(row):
    if pd.isnull(row['Дата увольнения']):
        val = 0
    else:
        val = 1
    return val
employees['Ушел'] = employees.apply(addleft,axis=1)
# 3
employees['Стаж'] =(employees['Дата увольнения'] - employees['Дата приема']) /
↪  np.timedelta64(1,'Y')
# 4
employees['Дата приема'] = pd.to_datetime(employees['Дата приема'],
↪  errors='coerce')
employees['Дата увольнения'] = pd.to_datetime(employees['Дата увольнения'],
↪  errors='coerce')
# 5
df = employees.merge(grades, on='Сотрудник', how='left')
# 6
df = df.merge(cities,on='Сотрудник',how='left')
# 7
df.drop_duplicates(subset=['Сотрудник','БН','Филиал'], ignore_index=True, keep
↪  = 'first',inplace=True)
# 8
def addage(row):
```

```
    if pd.notnull(row['Дата рождения']):
        val = np.datetime64('2021-05-17') - row['Дата рождения']
        val = pd.to_timedelta([val]).astype('timedelta64[h]')[0]
        val = int(val//8640)
    else:
        val = 1
    return val
df['Возраст'] = df.apply(addage,axis=1)
# 9
encode_sex = {"sex": {"Женский": 0, "Мужской": 1}}
df['Пол'] = df['Физическое лицо.Пол']
df = df.replace(encode_sex)
# Также необходимо обозначить возраст, у тех сотрудников, которые продолжают
# работать в компании
def fix_years_worked(row):
    if pd.isnull(row['years_worked']):
        val = np.datetime64('2021-05-17') - row['Дата приема']
        val = pd.to_timedelta([val]).astype('timedelta64[D]')[0]
        val = val / 365
    else:
        val = row['Стаж']
    return val
df['Стаж'] = df.apply(fix_years_worked,axis=1)
```

Далее стоит проверить каждый столбец нового агрегированного датафрейма на пропущенные значения, используя функцию df.isna().sum(), результат изображен ниже в таблице 2.2.

Таблица 2.2 – Количество пропущенных значений в каждом столбце датафрейма

Название атрибута	Кол-во пропущенных значений
Табельный номер	1167
Сотрудник	0
Дата рождения	2
Дата приема	0
Дата увольнения	1694
Физическое лицо.Пол	0
Ушел	0
Стаж	1694
Дата БН	2
БН	2
Грейд	2
Дата	938
Филиал	1243
Город	1763

Из таблицы выше видно, что столбцы «Дата БН», «БН» и «Грейд» имеют по 2 пропущенных значения, которые нужно удалить, остальные пропуски не критичны для поставленной задачи. Также видно, что у 1243 сотрудников не обозначены филиалы, их тоже необходимо будет опустить, так как алгоритмы ML не принимают пустые значения.

Итак, опустив ненужные столбцы («Сотрудник», «Дата рождения», «Физическое лицо.Пол», «Табельный номер», «Дата БН', «Дата», «Город») получаем итоговый датафрейм, как показано на рисунке 2.3.

	Дата приема	Дата увольнения	Ушел	Стаж	БН	Грейд	Филиал	Возраст	Пол
0	2013-07-10	2017-12-20	1	4.446361	Общее_Менеджеры	2.0	Москва	29	0
1	2013-07-10	2017-12-20	1	4.446361	Общее_Декрет	2.0	Москва	29	0
2	2020-11-24	NaT	0	0.476712	SOA	3.0	Воронеж	24	1
3	2017-03-01	NaT	0	4.213699	NFO	3.0	Саратов	29	0
4	2017-03-01	NaT	0	4.213699	Общее_Декрет	2.0	Саратов	29	0
...	...	...	...	...	...	...	...	...	...
2118	2013-06-10	2018-03-02	1	4.725627	Neoflex Reporting	2.0	Москва	27	0
2119	2013-06-10	2018-03-02	1	4.725627	Общее_Декрет	2.0	Москва	27	0
2120	2013-06-10	2018-03-02	1	4.725627	Neoflex Reporting ЦБ	2.0	Москва	27	0
2121	2013-06-10	2018-03-02	1	4.725627	Департамент аналитики	2.0	Москва	27	0
2122	2019-10-01	2020-11-27	1	1.158135	NFO	2.0	Саратов	19	1

Рисунок 2.3 – Фрагмент очищенной таблицы

По рисунку выше видно, что в подготовленном датафрейме в итоге 2123 экземпляров, что должно быть достаточным для эффективного обучения ML. Значения в стобце «Стаж» являются вещественными числами, так как года делятся на 12 (то есть количество проработанных лет известно до числа дней), но при необходимости можно будет преобразовать в целые числа в будущем.

Теперь рассмотрим данные бизнес-направлений с количеством сотрудников, работающих по каждому направлению, полный список представлен в приложении Б. По таблице Б.1 видно, что некоторые направления повторяются, но обозначены разными названиями, проблемы заключаются в появлении лишних пробелов и сокращений названий направлений. Было принято решение объединить «BigData» и «BigData » с «BigData Solutions», и «NR» с «Neoflex Reporting».

2.4 Разведочный анализ данных

Информативная визуализация – представление информации в графической форме, например в виде круговой диаграммы, графика или визуального представления другого типа. Она может быть частью процесса исследования, например, применяться для выявления выбросов, определения необходимых преобразований данных или поиска идей для построения моделей [23].Качественная визуализация данных имеет критическое значение для анализа данных и принятия решений на их основе. Визуализация позволяет быстро и легко замечать и интерпретировать связи и взаимоотношения, а также выявлять развивающиеся тенденции, которые не привлекли бы внимания в виде необработанных данных. В большинстве случаев для интерпретации графических представлений не требуется специальное обучение, что сокращает вероятность недопонимания.Продуманное графическое представление не только содержит информацию, но и повышает эффективность ее восприятия за счет наглядности, привлечения внимания и удержания интереса в отличие от таблиц и документов.

Разведочный анализ данных (EDA) играет важную роль в понимании исследуемого набора данных. В статистике под разведочным анализом данных понимается подход по обобщению основных характеристик данных, в том числе с их последующей визуализацией. В первую очередь EDA предназначен для того, чтобы исследователь мог «увидеть данные» и, возможно, построить первые гипотезы по ним. EDA был предложен Джоном Тьюки, чтобы побудить статистиков изучать данные, а также сформулировать гипотезы, которые могут привести к сбору новых данных и экспериментам [4].

В первую очередь для всех переменных из объединенного датасета с помощью функции общего назначения describe() рассмотрим основные параметры описательной статистики. Одной это функции достаточно для получения минимального (Min) и максимального (Max) значений, количество значений (Count), арифметической средней (Mean), стандартного отклонения(std), первого (25%), второго (50%) и третьего (75%) квартилей. Результаты представленны на рисунке 2.4.

	Ушел	Стаж	Грейд	Возраст	Пол
count	2123.000000	2123.000000	2123.000000	2123.000000	2123.000000
mean	0.309939	3.670789	2.589732	33.135657	0.609986
std	0.462577	3.489016	2.076716	7.299404	0.487868
min	0.000000	0.000000	-1.000000	1.000000	0.000000
25%	0.000000	0.704110	1.000000	28.000000	0.000000
50%	0.000000	2.224658	3.000000	32.000000	1.000000
75%	1.000000	6.175342	4.000000	37.000000	1.000000
max	1.000000	11.882192	9.000000	71.000000	1.000000

Рисунок 2.4 – Описательная статистика по данным

По рисунку выше видно, что в данных есть аномальные значения возраста - 1, от чего нужно избавиться. Было принято решение не удалять этих сотрудников, а подставить среднее значение возраста сотрудников по таблице - 33.

Для визуального представления распределения ушедших и не ушедших из компании сотрудников по возрасту ниже представлен график на рисунке 2.5.

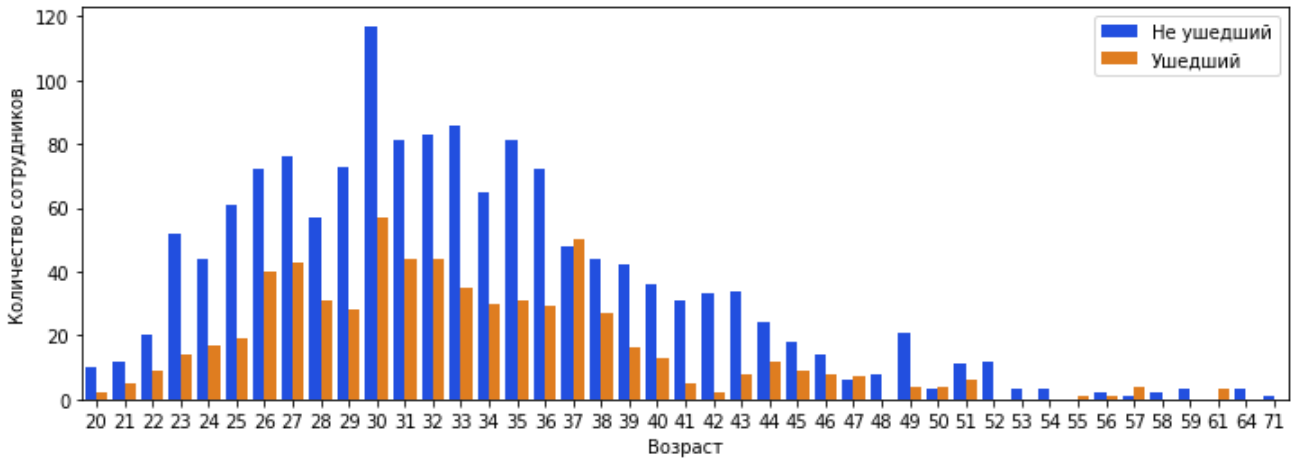

Рисунок 2.5 – Распределение ушедших и не ушедших сотрудников по возрасту

По рисунку выше можно выявить, что наибольшее количество сотрудников имеют возраст равному 26 годам. Причем в эту группу входит наибольшее количество как ушедших сотрудников, так и тех, кто продолжает свою деятельность в компании АО «Неофлекс Консалтинг». Также видно, что возраст

самых молодых кадров - 20 лет, имеются сотрудники возрастов 56, 57, 58, 59, 61, 64 и 71, тяжело определить являются ли эти значения настоящими или же это ошибки в данных. Более того, можно заметить примечательную статистику, что в возрасте 37 соотношение ушедших сотрудников больше, чем неушедших.

Далее текучесть кадров рассмотрена в разрезе грейдов сотрудников. Внутренний коэффициент уровня специалиста имеет значение -1 для интерна (т.е. еще нештатный сотрудник) и градацию от 1 до 9 для штатных сотрудников, где 1 это младший специалист, а 9 - директор. Визуальное представление распределения текучести кадров по их грейдам на основе подготовленного датафрейма представлено ниже на рисунке 2.6.

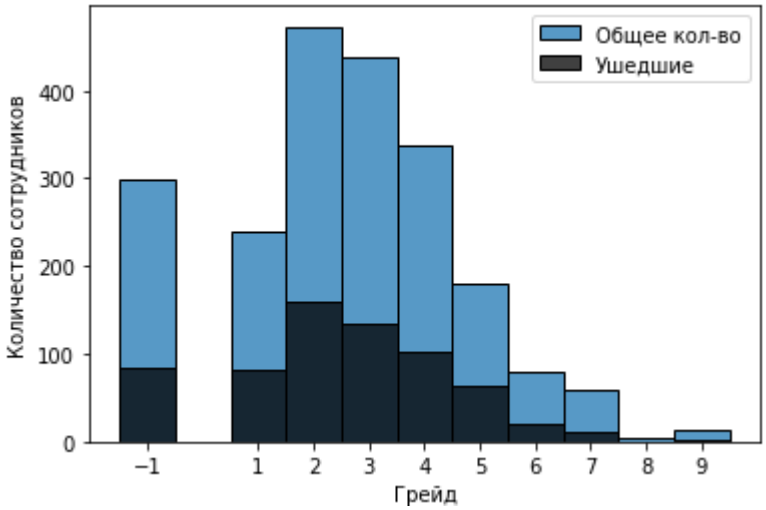

Рисунок 2.6 – Распределение сотрудников по грейду

На графике выше черным цветом отмечено количество ушедших сотрудников. В среднем можно отметить показатель текучести приблизительно в 35 %. Можно сделать вывод, что большинство сотрудников компании АО «Неофлекс Консалтинг» имеют грейды 2, 3 и 4, стоит заметить, что процент текучести по всем грейдам приблизительно равный. К сожалению, более примечательных выводов сделать не удалось.

На следующем этапе EDA текучесть кадров рассмотрена по каждому году деятельности компании АО «Неофлекс Консалтинг» в периоде с 2009 по 2021. Для построения визуализации считалось соотношение количества пришедших и ушедших сотрудников из компании за каждый год периода. График представлен в соответствии с рисунком 2.7.

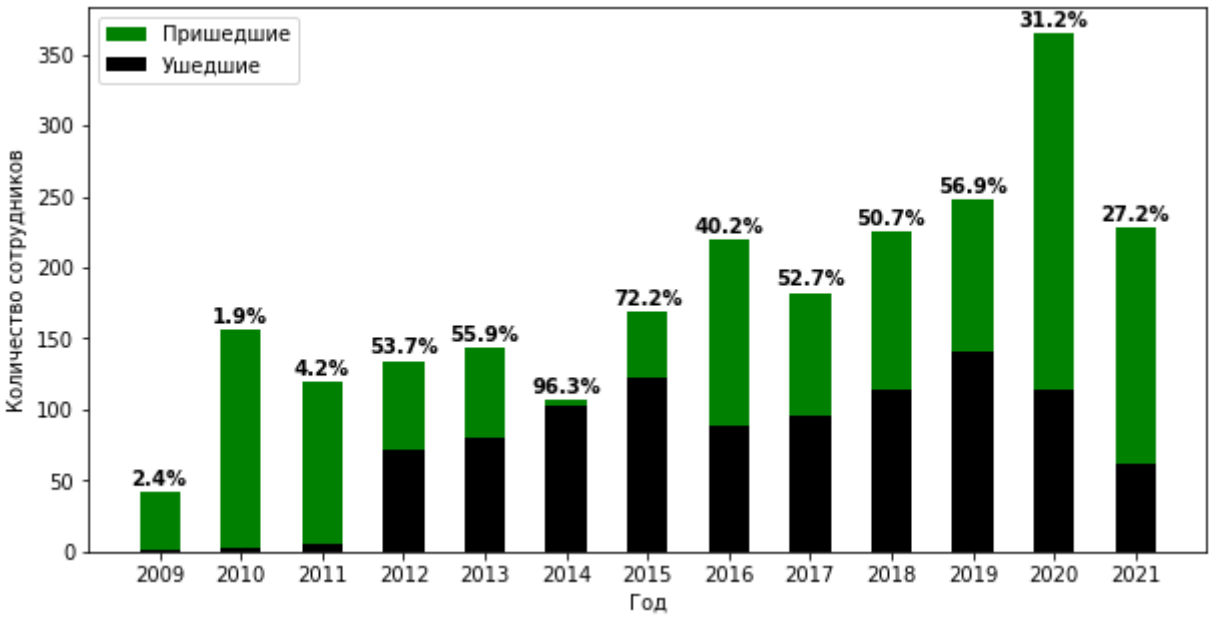

Рисунок 2.7 – Распределение текучести кадров по годам

На графике выше значение в 100% означает, что за n год из компании ушло столько же сотрудников, сколько пришло новых. По рисунку 2.7 явно видно, что в 2014 году компания потеряла сотрудников, практически равному числу пришедших (96.3%) за тот год. Было вычислено, что данная организация за период с 2009 по 2021 год в среднем теряла 42% персонала от количества пришедших сотрудников ежегодно. Также по графику видно, что компания количественно потеряла больше всего сотрудников в 2019 году ($\approx$ 140 чел.), а приобрела больше всего новых сотрудников в 2020 году ($\approx$ 350 чел.).

Далее текучесть кадров в исследуемой компании рассматривается относительно городов проживания сотрудников и филиалов, в которых они введут свою рабочую деятельность. Первый график, в соответствии с рисунком 2.8, показывает территориальное распределение сотрудников по 5 филиалам.

Рисунок 2.8 – Распределение городов проживания сотрудников

Второй график, в соответствии с рисунком 2.9, показывает все города откуда сотрудники родом.

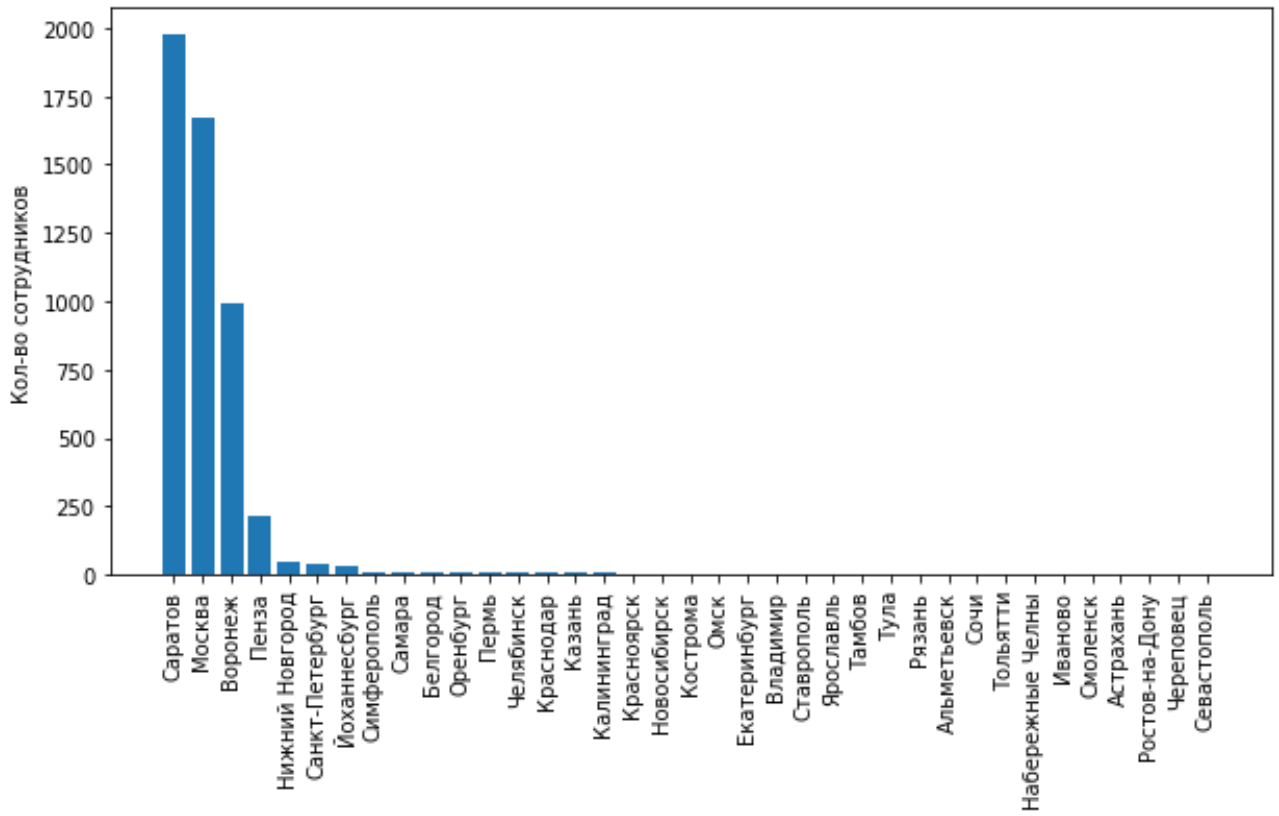

Рисунок 2.9 – Распределение городов проживания сотрудников

По графикам выше видно, что большинство сотрудников компании АО «Неофлекс Консалтинг» родом из города Саратов, несмотря на то что в Московском филиале работают на 11% больше персонала. В дополнение к этому, по рисунку 2.9 можно сделать вывод, что заметное количество сотрудников приезжают из городов Нижний Новгород и Санкт-Петербург, возможно данной компании стоит рассмотреть вариант открытия своих филиалов в перечисленных выше городах. Опираясь на статистические показатели проведенного анализа, можно добавить, что 1% общей численности кадров (в сумме) релоцируются из следующих городов: Симферополь, Самара, Оренбург, Белгород, Пермь, Калининград, Челябинск и Казань. Выявленная информация возможно будет полезной для руководства исследуемой компании.

Последний этап EDA заключается в анализе текучести кадров в зависимости от бизнес-направления сотрудников. На графике изображены 10 бизнес-направлений с наибольшим количеством ушедших сотрудников, в со-

ответствии с рисунком 2.10.

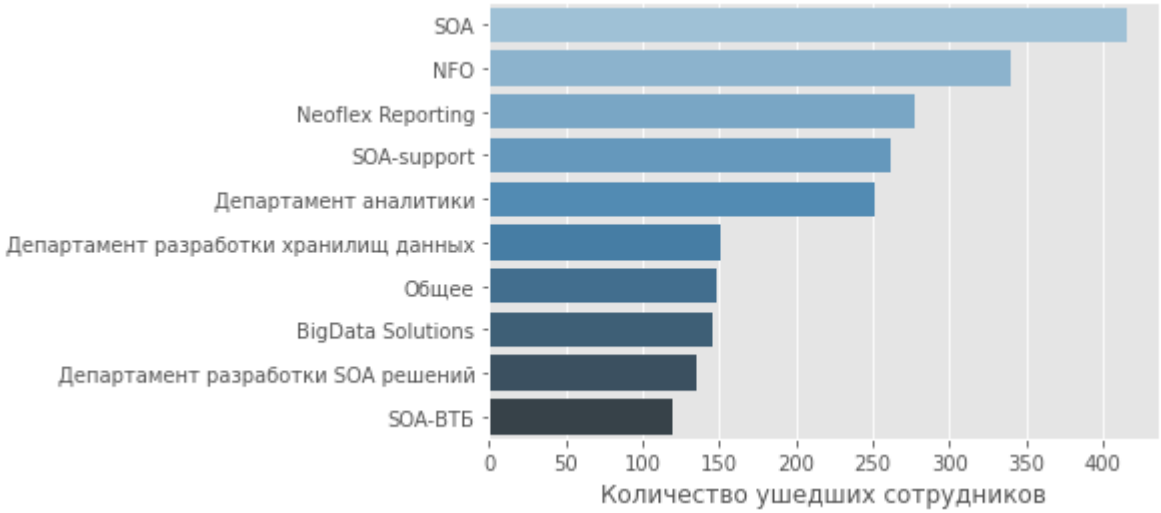

Рисунок 2.10 – Бизнес направления с наибольшим объёмом текучести

В силу специфики данных, за «текучесть сотрудника» может подразумеваться смена его специальности или повышение, ввиду этого. Наибольший показатель «движения рабочей силы» выявлен у бизнес-направления «SOA», неудивительно, так как это направление является самым обширным по количеству назначенных сотрудников.

В качестве заключения по данному разделу, можно сделать следующие **выводы**:

1) Рассматриваемая компания АО «Неофлекс Консалтинг» существует на рынке уже 17 лет, имеет 9 филиалов на территории РФ, более 20 бизнес-направлений и увеличила количество сотрудников со 100 до 1160.

2) Основными задачами предварительной обработки данных являются: «Очистка данных», «Преобразование данных», «Уменьшение данных», «Дискретизация данных», «Очистка текста». Более того, при необходимости, нужно обезличить данные, решить проблему дублированных и пропущенных значений.

3) Предобработка данных нужна для того, чтобы не допустить ложных выводов по первичным данным. Этот процесс делает наборы данных более целостными и эффективными для выполнения анализа данных и машинного обучения. Также это повышает точность и скорость алгоритмов машинного обучения, которые обучаются на данных. Качественные данные — это

необходимое условие для создания качественных моделей прогнозирования.

4) Исходные данные оказались достаточно «грязными», предобработка данных оказалась необходимой, были выявлены такие проблематичные нюансы, как повторяющиеся и пропущенные значения. Более того, выяснилось, что под «текучестью кадра» данные могут подразумевать не только увольнение сотрудника, но и его релокация или повышение.

5) В исследуемой компании наибольшее количество рабочей силы было потеряно в 2019 году, а по соотношению «пришедших/ушедших сотрудников» худший год для компании относительно этого показателя был 2014 год, с 96.3% ушедших к пришедшим. Самый распространенный грейд среди сотрудников - 2, соответственно уровень текучести наивысший у персонала этого грейда. Возможно данной компании стоит рассмотреть вариант открытия своих филиалов в городах Нижний Новгород и Санкт-Петербург, так как почти 2% общей численности кадров проживают в этих городах. Самый надежный возраст сотрудников, относительно вероятности ухода, является 42 года. Напротив, самый ненадежный возраст трудоустройства для данной компании, является 37 лет. Среднее значение продолжительности работы сотрудников в АО «Неофлекс Консалтинг» - 3 года.

3 Разработка системы прогнозирования текучести кадров

Кадровое прогнозирование — это анализ текучести кадров, наличия преемников у высших руководителей и перспектив разных видов деятельности компании — такой анализ нужен, чтобы в будущем не возникло дефицита или избытка тех или иных специалистов [24].

Кадровое прогнозирование основывается на переработке необходимой информации, при этом выделяют следующие стадии этой процедуры:

1) Ретроспекция, включающая исследование состояния работы с кадрами, структуры персонала в прошлом (последние 10 лет);

2) Диагноз, определяющий характер, состояние кадровой работы и структуру персонала на основе всестороннего их исследования. На этой стадии происходит выявление тенденций для развития самой структуры персонала и определение путей улучшения работы с кадрами.

3) Выбор метода, которые могут быть как формализованными (использование математической статистики), так и неформализованными (применение экспертных оценок и предложений качественного характера);

4) Разработка прогноза, который представляет собой предвидение всех изменений и сдвигов в структуре профессионально-квалификационного и социально-демографического состава персонала. На этой стадии также происходит организация и проведение работы с кадрами в сфере производства.

3.1 Теория об использованных методах машинного обучения

Машинное обучение – обширный подраздел искусственного интеллекта, изучающий методы построения алгоритмов, способных обучаться. Различают два типа классического обучения: с учителем и без учителя. Обучение с учителем — когда у машины есть некий учитель, который знает, какой ответ правильный. Это значит, что исходные данные уже размечены (отсортированы) нужным образом, и машине остается лишь определить объект с нужным признаком или вычислить результат. Обучение без учителя — когда машина сама должна найти среди хаотичных данных верное решение и отсортировать объекты по неизвестным признакам. Например, определить, где на фото собака. К первому типу относят задачи классификации и регрессии. Ко второму типу относят такие задачи как кластеризация, поиск правил и уменьшение

размерности.

Помимо описанных выше методов, также существуют ансамбли - это группы алгоритмов, которые используют сразу несколько методов машинного обучения и исправляют ошибки друг друга. В рамках данной работы используется такой вид ансамбля, как беггинг - когда один алгоритм многократно обучают на случайных выборках, а потом усредняют ответы.

Первая модель подразумевает задачу классификации, где класс 0 - сотрудник не уйдет, а класс 1 - сотрудник уйдет. Для определения вероятности текучки кадра, вычисляется вероятность его отношения к классу 1. В конечном итоге наиболее эффективным алгоритмом для этой задачи оказался «случайный лес». Данный метод представляет собой алгоритм машинного обучения, заключающийся в использовании множества деревьев принятия решений. Каждый из этих классификаторов строится на случайном подмножестве объектов и признаков. Алгоритм сочетает в себе метод бэггинга и метод случайных подпространств.

Алгоритм обучения классификатора:

1) Из обучающей выборки генерируется m случайных подвыборок, размера n.

2) Строится m деревьев, которые классифицирует объекты данной подвыборки.

3) Дерево строится до полного исчерпания подвыборки.

Классификация объектов проводится путем голосования: каждое дерево относит объект к одному из классов, и побеждает класс, которому отдает предпочтение большинство деревьев из набора.

Достоинства метода:

- способность обрабатывать большое число признаков и классов;
- нечувствительность к масштабированию;
- хорошая обработка непрерывных и дискретных признаков;
- возможность построения деревьев по данным с пропущенными значениями.

Недостатки метода:

- склонность к переобучению;
- обучение может занимать длительное время;

• требуется большой объем памяти для хранения модели.

Вторая модель подразумевает задачу регрессии, в которой нужно определить количество лет и месяцев, сколько сотрудник предположительно проработает в компании. Аналогично предыдущей модели, наиболее эффективным алгоритмом для данной задачи оказался «Случайный лес». Модель обучалась на следующих атрибутах сотрудников: возраст, грейд, город, бизнес-направление, пол и текучесть. По сути алгоритм прогнозирует значение стажа.

В машинном обучении гиперпараметрами называют параметры алгоритмов, значения которых устанавливаются перед запуском процесса обучения. В этом смысле они и отличаются от обычных параметров, вычисляемых в процессе обучения. Гиперпараметры используются для управления процессом обучения. Основные гиперпараметры алгоритма «Случайный лес»:

• max_features – количество учитываемых признаков при поиске лучшего расщепления. Чем меньше, тем меньше дисперсия и больше смещение;

• max_depth – максимальная глубина дерева;

• min_samples_leaf – минимальное количество объектов в листовом (терминальном) узле;

• min_samples_split – минимальное количество объектов, требуемое для расщепления узла. Вместе с предыдущим позволяет контролировать максимальную глубину дерева в целях предотвращения переобучения.

3.2 Реализация моделей и их сравнение

Для написания моделей машинного обучения был выбран язык программирования Python, использовалась библиотека Scikit-Learn, которая реализует множество методов интеллектуального анализа данных для визуализации, классификации, кластеризации и т.д., а также вспомогательные классы наподобие предобработки, поиска гиперпараметров модели, оценки качества моделей различными методиками, вычисления значений метрик и другие.

В подготовленном наборе данных всё же остались категориальные переменные, такие как филиалы и бизнес-направления. По причине того что алгоритмы ML не могут обучаться на категориальных переменных, необходимо преобразовать их в числовые. Города Москва, Саратов, Воронеж, Йоханнесбург и Пенза были закодированы в числа 2, 4, 0, 1 и 3 соответствен-

но. Все бизнес-направления были закодированы в числа в диапазоне от 0 до 70. В конечном итоге получился готовый для обучения набор данных, в соответствии с рисунком 3.1.

	Возраст	БН	Пол	Грейд	Филиал	Ушел	Стаж
0	32	52	0	2.0	2	1	4.446361
1	32	51	0	2.0	2	1	4.446361
2	27	17	1	3.0	0	0	0.476712
3	32	9	0	3.0	4	0	4.213699
4	32	51	0	2.0	4	0	4.213699
...	...	...	...	...	...	...	...
2118	30	11	0	2.0	2	1	4.725627
2119	30	51	0	2.0	2	1	4.725627
2120	30	14	0	2.0	2	1	4.725627
2121	30	30	0	2.0	2	1	4.725627
2122	22	9	1	2.0	4	1	1.158135

Рисунок 3.1 – Фрагмент подготовленного набора данных

В результате предобработки данных, таблица имеет 2123 образцов и 7 числовых признаков, по которым модели ML будут обучаться далее.

Для подбора наиболее эффективного алгоритма ML, с помощью которого будут обучаться модели, была использована кросс-валлидация по 6 самым распространенным алгоритмам. В качестве оценочных метрик были использованы Roc Auc, F1 и Accuracy. Результат изображен ниже, в соответствии с рисунком 3.2.

	Algorithm	ROC AUC Mean	F1 Mean	Accuracy Mean
1	Random Forest	76.00	45.61	72.74
3	KNN	65.10	40.06	67.09
2	SVM	64.93	28.58	68.91
4	Decision Tree Classifier	63.18	49.11	67.96
5	Gaussian NB	60.28	6.47	68.22
0	Logistic Regression	60.09	45.91	55.15

Рисунок 3.2 – Оценка различных алгоритмов обучения

По рисунку выше видно, что наивысшую точность имеет алгоритм

Random Forest (случайный лес). По этой причине был выбран данный алгоритм для обучения обеих моделей.

Первая модель классификации, обучается на следующих параметрах: возраст, грейд, город, бизнес-направление, пол и стаж. После успешного обучения модели, можно получить вероятность отношения данного сотрудника к классу 1(ушел), это значение и будет спрогнозированная вероятность ухода сотрудника из компании.

После оптимизации гиперпараметров, модель получила следующие оценки качества:

Accuracy: 0.7796610169491526,

Precision: 0.6621621621621622,

Recall: 0.593939393939394,

F1: 0.6261980830670928,

Roc_auc: 0.7286636860407352.

Вторая модель регрессии, обучается на аналогичных параметров, за исключением последнего, вместо стажа, что модель выдаст на выходе, добавляется бинарный атрибут «Ушел». Для оценки качества регрессионной модели, используются следующие метрики:

R2(точность): 0.6611945277399738,

MAE(средняя абсолютная ошибка): 1.4226785772879065,

MSE(среднеквадратическая ошибка): 1.9936977708647168.

Опираясь на значение MAE, можно сделать вывод, что модель в спрогнозированном значение ошибается в среднем на 1,42(года).

3.3 Преобразование в веб-сервис

Для удобного взаимодействия с построенными моделями было разработано веб-приложение с пользовательским интерфейсом для ввода параметров сотрудника и выводом результатов оценки реализованной модели. В качестве основного инструмента реализации веб-приложения, был использован фреймворк Streamlit на языке Python.

В результате при переходе по ссылке веб-приложения, загружается окно с формой для ввода данных и после нажатия кнопки «Рассчитать» введенные данные передаются модели, после чего выводится результат. Для иллюстра-

ции работы системы, ниже приведены два примера работы приложения. На первом, в соответствии с рисунком 3.3, введены данные молодого специалиста, соответственно с низким значением грейда и стажа.

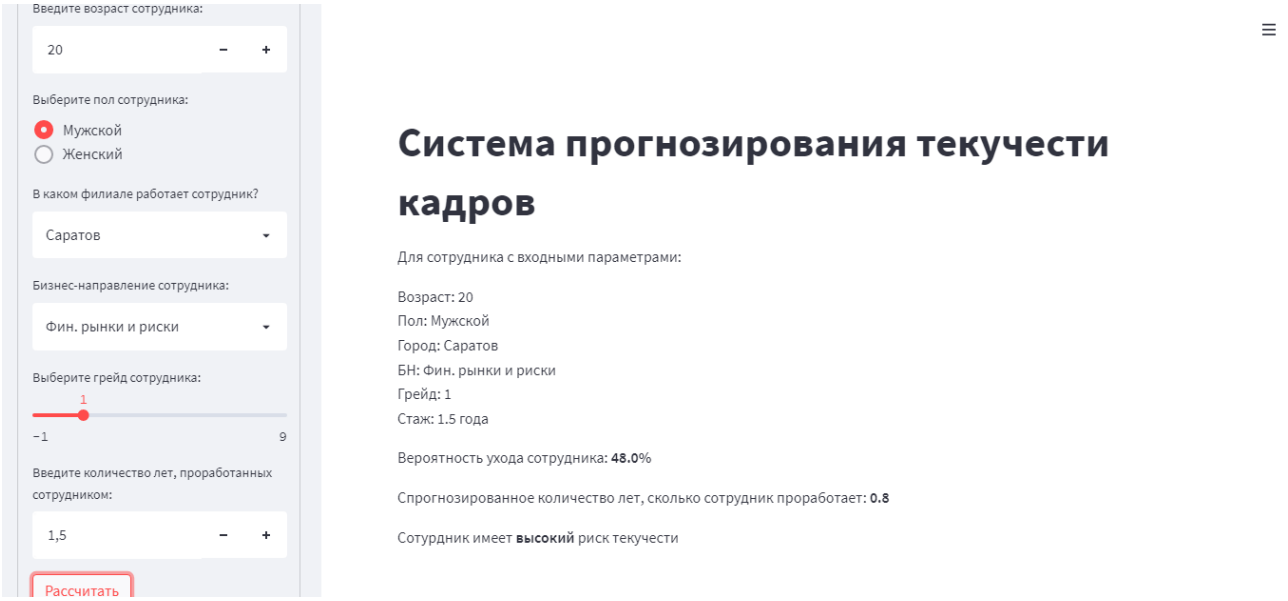

Рисунок 3.3 – Пример работы веб-сервиса 1

По рисунку выше видно, что сотрудник с низкими значениями возраста, грейда и стажа имеет высокий уровень риска ухода из компании. По прогнозу модели данный сотрудник проработает 10 месяцев в компании.

На втором примере, в соответствии с рисунком 3.4, введены данные опытного специалиста, соответственно с высоким значением грейда и стажа.

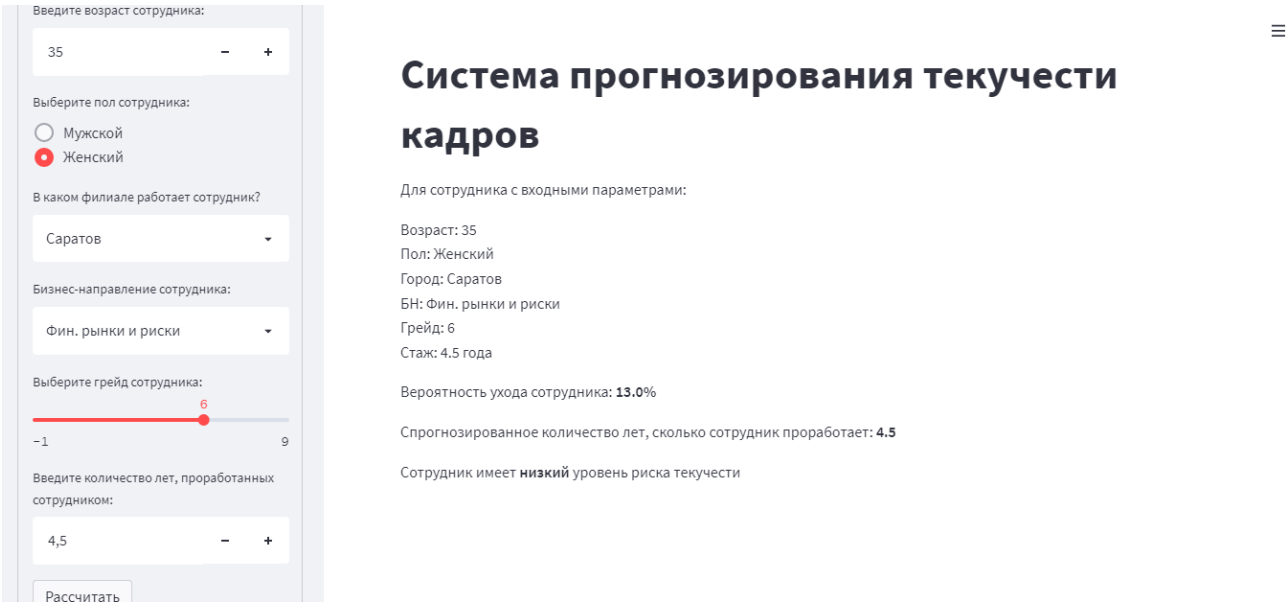

Рисунок 3.4 – Пример работы веб-сервиса 2

Так как у данного сотрудника значения грейда, возраста и стажа выше, то соответственно ниже вероятность его ухода из компании и проработает он дольше чем сотрудник из первого примера.

Полный программный код разработанной системы представлен в приложении В.

Выводы по разделу:

В работе рассмотрено два класса задач: задача классификации сотрудника относительно вероятности утечки кадра и задача построения модели прогнозирования количества лет и месяцев, которое сотрудник проработает.

Для обеих моделей наиболее эффективным алгоритмом машинного обучения оказался «Случайный лес».

Модель классификации имеет точность $\approx 75\%$.

Модель регрессии имеет точность $\approx 62\%$.

ЗАКЛЮЧЕНИЕ

Текучесть кадров - это движение работников как из штата организации, так и в штат. Следует различать ее естественный уровень, который способствует обновлению производственных коллективов, происходит непрерывно и не требует каких-либо чрезвычайных мер со стороны кадровых служб и руководства, и повышенный, вызывающий значительные экономические потери.

Потеря высококвалифицированного персонала для любой компании является большой проблемой, ведь обучение одного хорошего специалиста длится от нескольких месяцев до года, и за это время работник не приносит прибыли. Более того, компания вкладывает в этого работника финансы и другие ресурсы. Даже если в компанию приходит специалист высокого уровня, этот человек должен изучить программы, применяемые в компании, ознакомиться с проектом, над которым он будет работать. Множество проблем возникают, когда сотрудник увольняется из компании. Например, большие сложности с делегированием его обязанностей другим специалистам. Такая ситуация увеличит нагрузку на этих специалистов, а также замедлит работу над проектами, над которыми работал уволенный сотрудник. Кроме того, текучесть кадров негативно влияет на моральный дух компании.

От уровня текучести персонала предприятия зависит, в том числе, его конечная прибыль, а также становление корпоративной культуры и успех бизнеса в целом. Безусловно, ликвидировать это явление в деятельности предприятия полностью невозможно, но его можно в какой-то степени контролировать и минимизировать. Поэтому руководство компании заинтересовано в снижении текучести кадров.

В целях экономии убытков, связанных с текучестью кадров, в данной работе был проведен визуальный анализ данных по сотрудникам компании АО «Неофлекс Консалтинг», более того разработано веб-приложение для прогнозирования вероятности ухода персонала.

Во втором разделе был выполнен этап предварительный подготовки данных и разведочного анализа данных на предмет выбросов, дубликатов, аномалий и ошибок в исходных данных. Выявлено множество проблематичных

нюансов в исходных данных.

В третьем разделе, на основе подготовленных данных, были реализованы две модели. Первая представляет собой задачу классификации сотрудника относительно вероятности утечки кадра или другими словами, расчет вероятности ухода сотрудника. Целью модели является общая оценка текучести кадров в организации и выявление наиболее вероятных на смену работы сотрудников. Вторая модель основана на задаче построения регрессии по предоставленным данным о рабочем персонале. Цель данной модели предсказать количество лет и месяцев, которые предположительно проработает новый сотрудник в компании, исходя из его пола, возраста, филиала, бизнес направления и внутреннего коэффициента уровня специалиста.

Разработанная система ML может помочь отделу кадров в предоставлении информацию о потенциальном решении сотрудника покинуть организацию. В зависимости от специфики введенных данных о сотруднике, реализованный сервис предсказывает вероятность ухода данного сотрудника и вычисляет примерное количество лет, сколько этот человек проработает в организации, более того определяется степень потенциального риска оттока сотрудника. Так как приложение было разработано для АО «Неофлекс Консалтинг» и модели ML обучены на данных, предоставленными этой компании, использование данной системы не рекомендуется для посторонних организаций.

Основные результаты проведенной работы:

1. Изучено теоретическое обоснование проблематики текучести кадров.

2. Рассмотрены реальные данные компании АО «Неофлекс Консалтинг» с точки зрения проблематики текучести кадров.

3. Проведена предобработка предоставленных данных и их визуальный анализ по статистическим показателям.

4. Построена модель классификации сотрудников относительно вероятности утечки кадра.

5. Построена модель прогнозирования количества лет и месяцев, которые проработает сотрудника в данной организации.

6. Реализовано веб-приложение для удобного использования разработанных моделей прогнозирования.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1 Рыбалкина, З.М. / Региональная экономика: теория и практика, 2017, т. 15, вып. 3 - 495 с.

2 Мансуров, Р. Как своевременно укомплектовать штат квалифицированными кадрами // Кадровик. Рекрутинг для кадровика. 2018. № 6. – 203 с.

3 Щекин, Г. В. Организация и психология управления персоналом. — К.: МАУП, 2002. – Т. 726.

4 Чижов, Н.А. Управление корпоративными кадрами. – СПб.: Питер, 2005. – 352 с.

5 Борисова, А.А. Регулирование текучести кадров на основе оценки экономического ущерба предприятия // Российское предпринимательство. – 2017. — № 11 (18). – 1692 с.

6 Синяева, Л. П. Социально-экономические проблемы формирования стабильных кадров на Куйбышевской железной дороге // Социально-экономические проблемы труда в развитом социалистическом обществе. Межвузовский сборник. Вып. 3. – Куйбышев: Куйбышевский государственный университет, 1977.

7 Долбунов, А. А. Текучесть кадров - основная проблема предприятий // Маркетинг. - 2006. - № 12. - 64 с.

8 Гусаров, А.В. Определение миссии организации //Менеджмент в России и за рубежом. - 2013. - №3 - 21 с.

9 Гордиенко, Ю.Ф. Управление персоналом : Серия «Высшее образование». – Ростов н/Д: Феникс, 2007. – 352 с.

10 Рогозина, А. Снижение текучести кадров за счет внедрения системы адаптации и обучения производственного персонала // Кадровик.ру. 2017. № 7. – 69 с.

11 Словарь управления персоналом [Электронный ресурс] - URL: https://psyfactor.org/personal0.htm/ (дата обращения 10.04.2022). - Загл. с экрана. - Яз. рус.

12 Гольцов, А. В. Методы снижения текучести кадров // Маркетинг. - 2006. - № 2. - 43 с.

13 Синяева, Л. П. Текучесть кадров как индикатор адекватности управления предприятием / Л. П. Синяева, И. В. Додорина, Е. А. Герасимова // Концепт. – 2013. – Спецвыпуск № 4. – 35 с.

14 Официальный веб-сайт компании АО «Неофлекс Консалтинг» [Электронный ресурс] - URL: https://www.neoflex.ru/ (дата обращения 20.04.2022). - Загл. с экрана. - Яз. рус.

15 Чубукова, И.А. Data Mining: учебное пособие. — М.: Интернет-университет информационных технологий: БИНОМ: Лаборатория знаний, 2006. – 208 с.

16 Правовое регулирование информационных отношений в области государственной и коммерческой тайны, персональных данных : учебное пособие / О. В. Ахрамеева, И. Ф. Дедюхина, О. В. Жданова, Н. В. Мирошниченко. — Ставрополь : СтГАУ, 2015. — 59 с.

17 Приказ Роскомнадзора от 5.09.2013 № 996 «Об утверждении требований и методов по обезличиванию персональных данных».

18 Документация языка программирования Python [Электронный ресурс] - URL: https://docs.python.org/3/ (дата обращения: 25.04.2022). - Загл. с экрана. - Яз. англ.

19 Документация библиотеки pandas [Электронный ресурс] - URL: https://pandas.pydata.org/docs/ (дата обращения: 25.04.2022). - Загл. с экрана. - Яз. англ.

20 Документация библиотеки matplotlib [Электронный ресурс] - URL: https://matplotlib.org/ (дата обращения: 25.04.2022). - Загл. с экрана. - Яз. англ.

21 Документация библиотеки pyplot [Электронный ресурс] - URL: https://matplotlib.org/stable/api/_as_gen/matplotlib.pyplot.html (дата обращения: 25.04.2022). - Загл. с экрана. - Яз. англ.

22 Документация библиотеки NumPy [Электронный ресурс] - URL: https://numpy.org/doc/ (дата обращения: 25.04.2022). - Загл. с экрана. - Яз. англ.

23 Маккинни, У. Python и анализ данных / У. Маккинни ; перевод с английского А. А. Слинкина. — 2-ое изд., испр. и доп. — Москва : ДМК Пресс, 2020. — 273 с.

24 Филиппов, А.В. Работа с кадрами[Текст]/А.В. Филипов.-М.: Экономика, 2010. - 348 с.

ПРИЛОЖЕНИЕ А

Исходные данные

Компанией АО «Неофлекс Консалтинг» были предоставлены данные о сотрудниках в виде трех таблиц. Фрагменты каждой таблицы представлены ниже. Данные о текучести персонала, в соответствии с рисунком А.1.

Табельный номер	Сотрудник	Дата рождения	Дата приема	Дата увольнения	Физическое лицо.Пол
	06c8118e2c3726b66efb4b8d113e2635	15.03.1976	21.05.2009	30.06.2009	Мужской
6	06c8118e2c3726b66efb4b8d113e2635	15.03.1976	01.07.2009	28.06.2019	Мужской
	06c8118e2c3726b66efb4b8d113e2635	15.03.1976	01.07.2009		Мужской
7	4e24fb1750996e0e8f91861df9ff65e7	01.05.1979	02.07.2009		Женский
	4e24fb1750996e0e8f91861df9ff65e7	01.05.1979	02.07.2009		Женский
	80af072e4e86c60e06fede7a8de6fb79	11.04.1969	14.07.2009	14.09.2012	Мужской
11	23ec45eb079c9ef3fa719b144067ac66	01.06.1960	14.07.2009	29.11.2019	Мужской
	23ec45eb079c9ef3fa719b144067ac66	01.06.1960	14.07.2009		Мужской
	5e7b2efe12281cee893cb63580a64c77	20.10.1975	17.07.2009	07.04.2011	Мужской
	b97a0a84585008373abe3b888e86b359	09.12.1983	21.07.2009	25.10.2013	Мужской
	2d77f813f4db00eef993aca0642a13f0	05.05.1984	21.07.2009	20.03.2014	Мужской
	198643fbcec1dadd44bba2d09d80df71	02.05.1983	21.07.2009	27.02.2015	Мужской
	2ee0b7faeafbd6c0eac0577b7d819469	19.11.1984	21.07.2009	29.05.2015	Женский
20	cc2e17ac3e5f8bd371e26a70c0ddc88f	24.11.1983	22.07.2009		Мужской
	cc2e17ac3e5f8bd371e26a70c0ddc88f	24.11.1983	22.07.2009		Мужской
	35e9db6439531c57043e45beee0d2cdc	12.02.1985	28.07.2009	14.01.2011	Мужской
	0573f5a2f09a64ac8cebabf50af4e8b8	29.03.1978	28.07.2009	18.11.2016	Мужской
	944bb5ee61fdf3abc29020b8da5c3f79	11.11.1985	30.07.2009	16.01.2015	Мужской
	1f9e8f1c2348efebe8513c5672e22872	18.11.1987	04.08.2009	15.01.2013	Мужской
	a0bc671ff2f46bc854a410f7214c338b	25.09.1986	20.08.2009	28.01.2015	Женский
	48f4f0cc2e170758fd1840fabf3953ea	23.06.1984	15.09.2009	10.01.2014	Мужской

Рисунок А.1 – Таблица «Employees»

Данные о филиалах компании, в соответствии с рисунком А.2.

Дата	Сотрудник	Филиал	Город
08.02.2005	60af5ad43f95c071e5f3f3588281834c	Москва	Москва
28.02.2005	e1e061c61daee1d518a9d8181e64c196	Москва	Москва
28.02.2005	8dcd51df55acd1afc191d263848e73d1	Москва	Москва
01.03.2005	16e30d6f664c5e4085b1e41e77b5aac0	Москва	Москва
02.08.2005	f66dac3272ab7d2398c79cbf9da03363	Москва	Москва
16.01.2006	7bd61b9428039f706f7c1e0602a5d5b1	Москва	Москва
07.12.2006	b2a67e7335e3072c1c6a01d5e7004c72	Москва	Москва
11.12.2006	82ac8834435110f81bb4e2dc2b12b567	Москва	Москва
02.04.2007	bead7915fc661af66f0862fff75eaa56	Москва	Москва
26.04.2007	2a8662c5d555b73331b45012e71e6b20	Москва	Москва
03.07.2007	1948c3b85be4caefd86d1cad0b8039d0	Москва	Москва
01.08.2007	9e25e42c243c8196768a74d2ecbc661f	Москва	Москва
30.10.2007	e75f54c484d022e8ef77c4780ce5ddd9	Москва	Москва
09.01.2008	dab3b281513447b57bc7a5dee370ed01	Москва	Москва
26.02.2008	4722d70b32de398f3e338452be305055	Москва	Москва
18.03.2008	aec526a464ad6813d6a6c00bbdf00f8d	Москва	Москва
21.07.2008	901efb68151020c160b9663c18a850aa	Москва	Москва
04.08.2008	0581ace14c9042669649774adf9f00ee	Москва	Москва
18.08.2008	09decd2b58fbcbdbe9bc2c70ef5491c7	Москва	Москва
25.08.2008	cc2e17ac3e5f8bd371e26a70c0ddc88f	Москва	Москва
11.12.2008	e1bc2eb75426666e817cdf797f25dd23	Москва	Москва

Рисунок А.2 – Таблица «Cities»

Данные о грейдах сотрудников, в соответствии с рисунком А.3.

Дата	Сотрудник	БН	grade
01.01.2019	8a195bab54f69a76d6be02a70692dc8e	ФАС	-1
01.01.2019	60af5ad43f95c071e5f3f3588281834c	Руководитель	9
01.01.2019	2e32f3195c2fd2d6660a4e48f5dbd40a	SOA-ВТБ	3
01.01.2019	c338b675efc7c06ca9f54c03c312ead1	BigData Solutions	2
01.01.2019	6f216663ad0259436e16b60c5202ce86	ФАС	-1
01.01.2019	68915ee9dd402ea9ab8745d3f1f1eca9	ФАС	-1
01.01.2019	b6c7af6f64f2e92c160caa1b4425247a	Водитель	-1
01.01.2019	1f3e490038d8dc794aaefa8802b34c0a	NFO	3
01.01.2019	bad50bad803a978392acd9d6922b7877	Департамент развития бизнеса	6
01.01.2019	b7eac6d9b1c222f3579036123b2b2256	SOA-ВТБ	2
01.01.2019	b7e281fa1e924b529d2917ab0c31091c	SOA-ВТБ	2
01.01.2019	66c677b4fd8615725fc90215db6ac2f9	SOA-ВТБ	3
01.01.2019	bd9b3c3f046542cb5d1fe520cde799c1	SOA-ВТБ	6
01.01.2019	786812d6349132b15312420e3c102a2f	SOA-ВТБ	2
01.01.2019	5a5a8e991404d751e6739cde7ec01ceb	SOA-ВТБ	3
01.01.2019	1f37a1d187ebf9822e7775fb11ea6174	SOA-ВТБ	1
01.01.2019	96270b54525e045688d1dfec3330dfdb	SOA-ВТБ	4
01.01.2019	c04e5bc0a3afbb5ca2365c558708d6ef	SOA-ВТБ	2
01.01.2019	e75f54c484d022e8ef77c4780ce5ddd9	SOA-ВТБ	5
01.01.2019	a5cc17f5ed657b28ed7c7886e98906d2	SOA-ВТБ	2
01.01.2019	2070bc632fdde2c9a3e5ba771708ec56	SOA-ВТБ	3

Рисунок А.3 – Таблица «Grades»

ПРИЛОЖЕНИЕ Б

Полный список бизнес-направлений

Полный список бизнес-направлений по данным до предварительной обработки представлен в таблице Б.1.

Таблица Б.1

Бизнес-направления	Кол-во сотрудников
SOA	192
NFO	185
BigData Solutions	184
SOA-ВТБ	176
SOA-support	172
Neoflex Reporting	87
АРГО	84
FastData	77
Инженерные практики	76
NR	71
Фин. рынки и риски	69
Общее	63
Департамент управления проектами	54
Общее по доходам	51
ФАС	51
Общее Декрет	46
Служба персонала	38
Саратовский филиал АО "Неофлекс Консалтинг"	34
Общее Менеджеры	34
Департамент аналитики	31
Департамент разработки SOA решений	28
Декрет	24
IT-разработка	20
Общее IT	19
Neoflex Reporting ЦБ	18

Продолжение таблицы Б.1

Служба ИТ	17
Рынки капитала	14
Развитие UI-UX	14
Департамент развития бизнеса	12
Служба маркетинга	10
Общее HR	10
Data Science	9
Департамент разработки NFO	9
Направление Банковская отчетность	8
Руководитель	8
Направление SOA-support	8
Финансово-административная служба	7
BigData	6
DataGram	6
Управление рисками	6
Администрация Саратов	6
Neoflex Reporting-support	5
Департамент тестирования	5
Департамент разработки хранилищ данных	5
Центр развития компетенций	5
Служба главного аналитика	5
Технологический офис	5
Администрация Воронеж	4
Общее Без группы	4
Инструментальная группа	4
Calypso	4
Совместитель	4
Neoflex Reporting 2.0	4
Направление управление рисками	4
Neoflex Reporting УО	3
Кластер молодых специалистов	3

Продолжение таблицы Б.1

RND	3
CloudSolutions	3
BigData	3
Администрация Пенза	3
ПО & АО	2
Общее по численности	2
Направление Localization	1
Водитель	1
Администрация Нижний Новгород	1
Служба по работе с партнерами	1
Администрация Йоханнесбург	1
Руководство	1
Sure24	1
Заказные проекты DWH	1
Администрация Санкт-Петербург	1

ПРИЛОЖЕНИЕ В

Исходный код веб-приложения

```
import streamlit as st
import numpy as np
import pandas as pd

import seaborn as sns
import matplotlib
from matplotlib.figure import Figure
import matplotlib.pyplot as plt

from sklearn.model_selection import train_test_split

from sklearn.linear_model import LogisticRegression
from sklearn.ensemble import RandomForestClassifier
from sklearn.ensemble import RandomForestRegressor

from sklearn import metrics
from sklearn.metrics import
↪  accuracy_score,mean_squared_error,r2_score,mean_absolute_error

# Ввод данных
df = pd.read_csv('databn.csv')

X = df[['age',"sex",'Филиал','БН','grade','years_worked']].values
y = df['Left']

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.25,
↪  random_state=7,stratify=y)

X2 = df[['age',"sex",'Филиал','БН','grade','Left']].values
y2 = df['years_worked']

X_train2, X_test2, y_train2, y_test2 = train_test_split(X2, y2, test_size=0.25,
↪  random_state=7)

# Обучение моделей

model = RandomForestClassifier(random_state=7)
model.fit(X_train, y_train)

model2 = RandomForestRegressor(random_state=0)
model2.fit(X_train2, y_train2)


bs = np.empty((0,0))
```

```
bs2 = np.empty((0,0))

st.set_page_config(page_title="Neoflex")
st.title("Система прогнозирования текучести кадров")

form = st.sidebar.form(key="input_form")
with form:

    # Возраст

    age= st.number_input('Введите возраст сотрудника:', 17, max_value=70)
    age= np.array(age)

    bs = np.append(bs,age)
    bs2 = np.append(bs2,age)

    # Пол

    #st.write("Выберите пол сотрудника:")
    sex = st.radio("Выберите пол сотрудника:", ('Мужской', 'Женский'))
    if sex == 'Мужской':
        sex = 1
    else:
        sex = 0
    sex=int(sex)
    bs = np.append(bs,sex)
    bs2 = np.append(bs2,sex)

    # Филиал

    city = st.selectbox("В каком филиале работает
    ↪  сотрудник?",['Москва','Саратов','Воронеж', 'Йоханнесбург', 'Пенза'])
    city_list = []
    city_list.append(city)
    city_pd = pd.DataFrame(city_list)
    city_n = city_pd.replace(['Москва','Саратов','Воронеж', 'Йоханнесбург',
    ↪  'Пенза'],['2','4','0','1','3'])
    bs = np.append(bs,city_n)
    bs2 = np.append(bs2,city_n)

    # Бизнес-направление


    bn = st.selectbox("Бизнес-направление сотрудника:", df['BN'].unique())
    bn_list = []
    bn_list.append(bn)
```

```
        bn_pd = pd.DataFrame(bn_list)
        global BN_name
        global BN_number
        BN_name = df["BN"].sort_values().unique()
        BN_number = df["БН"].sort_values().unique()
        bn_n = bn_pd.replace(BN_name,BN_number)

        bs = np.append(bs,bn_n)
        bs2 = np.append(bs2,bn_n)

        # Грейд

        #grade = st.slider('Выберите грейд сотрудника:', -1,9)
        grade = st.select_slider('Выберите грейд сотрудника: ',options=['-1', '1',
        ↪  '2', '3', '4', '5', '6', '7', '8', '9'])
        bs = np.append(bs,grade)
        bs2 = np.append(bs2,grade)

        # Стаж

        years = st.number_input("Введите количество лет, проработанных
        ↪  сотрудником:", 0.0, max_value=13.0,step=0.5,format="%.1f")
        bs = np.append(bs,years)

        bs = bs.reshape(1,-1)

        bs2 = np.append(bs2,0)
        bs2 = bs2.reshape(1,-1)

        # Вывод

        prediction = np.array(model.predict_proba(bs))
        prediction = round(float(prediction[:,1])*100,2)
        out = "Вероятность ухода сотрудника:  **{}%**".format(prediction)

        prediction2 = np.array(model2.predict(bs2))
        prediction2 = round(float(prediction2),1)
        out2 = "Спрогнозированное количество лет, сколько сотрудник проработает:
        ↪  **{}**".format(prediction2)

        submitted = st.form_submit_button(label="Рассчитать")


    if submitted:
         #st.success("Спасибо!")
```

```
def string(sex):
    sexstr = ""
    if sex == 0:
        sexstr = 'Женский'
    else:
        sexstr = 'Мужской'
    return sexstr

def addyears(years):
    yearstr = ""
    if years == 1:
        yearstr = 'год'
    elif years > 1 and years < 5:
        yearstr = 'года'
    else:
        yearstr = 'лет'
    return yearstr

st.write('Для сотрудника с входными параметрами:')
st.write('Возраст:    ', str(age), '  \n Пол: ', string(sex), '  \n Город:
↪  ', city,
    '  \n БН: ', bn, '  \n Грейд: ', str(grade), '  \n Стаж: ',
    ↪  str(years),addyears(years))
st.write(out)
st.write(out2)

if prediction2 < 1:
    st.write("Сотрудник имеет **высокий** уровень риска текучести!")
elif prediction2 >= 1 and prediction2 <=3:
    st.write("Сотрудник имеет **средний** уровень риска текучести")
else:
    st.write("Сотрудник имеет **низкий** уровень риска текучести :thumbsup:
    ↪  ")
```