**Изображение выглядит как текст

Автоматически созданное описание**

МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ

Федеральное государственное бюджетное образовательное учреждение

высшего образования

**«КУБАНСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ»**

**(ФГБОУ ВО «КубГУ»)**

**Экономический факультет**

**КУРСОВАЯ РАБОТА**

**СИСТЕМА ПРЕДСКАЗАНИЯ СПРОСА НА БИЛЕТЫ И АДАПТИВНОГО ФОРМИРОВАНИЯ ЦЕНЫ С ПОМОЩЬЮ ИНСТРУМЕНТОВ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА**

Автор работы:

Кулягина Анастасия Павловна

студентка 3 курса,

направление подготовки «Инноватика» (бакалавриат)

Научный руководитель:

Васкевич Татьяна Владимировна,

доцент, канд. экон. наук.

Краснодар

2024

**Содержание**

Введение 4

1. Анализ факторов, влияющих на формирование цен и спроса на железнодорожные билеты в ОАО РЖД 6

1.1 Краткая характеристика ОАО РЖД 6

1.2 Сезонность и её влияние на спрос и предложение билетов 7

1.3 Влияние класса обслуживания на ценообразование 10

1.4 Расстояние и маршрут как детерминанты цены 12

2. Построение модели предсказания спроса на билеты 14

2.1 Описание набора данных и методологии анализа 14

2.2 Выбор метода прогнозирования спроса и построение модели в Loginom 16

3. Построение модели адаптивного формирования цены на билеты 21

3.1 Обзор возможностей и методов создания модели адаптивного ценообразования 21

3.2 Построение в Loginom модели адаптивного

ценообразования 47

Заключение 48

Список используемой литературы 49

Приложение А Настройки первой модели для предсказания спроса 51

Приложение Б Настройки второй модели для предсказания спроса 53

Приложение В Настройки третьей модели для предсказания спроса 55

Приложение Г Настройки первой модели для адаптации ценообразования 57

Приложение Д Настройки второй модели для адаптации ценообразования 59

Приложение Е Настройки третьей модели для адаптации ценообразования 61

Приложение Ж Настройки четвертой модели для адаптации ценообразования 63

Приложение З Настройки логической регрессии 65

Приложение И Настройки Цикла 67

**Введение**

В эпоху трансформации общества под воздействием информационных технологий, ключевым аспектом становится адаптация и инновация во всех сферах человеческой деятельности, включая транспорт и услуги. Разработка систем предсказания спроса и адаптивного формирования цены является повышением эффективности и удовлетворенности потребителей. В контексте железнодорожных перевозок, где динамика спроса и ценообразование играют решающую роль, применение инструментов искусственного интеллекта открывает новые горизонты для оптимизации этих процессов.

Проблематика формирования цен на железнодорожные билеты, зависящая от множества факторов, таких как сезонность, класс обслуживания, расстояние и маршрут, а также дополнительные услуги, требует комплексного подхода и проработки. Существующие модели ценообразования не всегда способны учитывать все аспекты спроса и предложения, что ведет к неоптимальному использованию ресурсов и уменьшению удовлетворенности клиентов.

В данной научной работе представлен подход к созданию интеллектуальной системы, способной анализировать большие объемы данных о продажах билетов и адаптировать цены для максимизации доходности и удовлетворенности потребителей. Исследование опирается на анализ данных продаж по маршруту Адлер - Москва Казанская, собранных в течение 2022 года и первого полугодия 2023 года, и предлагает модель в области искусственного интеллекта для предсказания спроса и адаптивного ценообразования.

Актуальность работы обусловлена необходимостью повышения эффективности управления ценами на железнодорожные билеты, учитывая колебания спроса и предложения, а также постоянно меняющиеся условия рынка.

Цель исследования заключается в разработке и тестировании модели, которая позволит ОАО «РЖД» управлять ценами на билеты более гибко, опираясь на актуальные данные.

Для достижения поставленной цели необходимо решить ряд задач, включая анализ факторов, влияющих на ценообразование, разработку модели предсказания спроса, интеграцию инструментов искусственного интеллекта для адаптивного управления ценами и оценку эффективности предложенной модели на основе реальных данных.

Объектом исследования являются железнодорожные перевозки по маршруту Адлер - Москва Казанская, а предметом - процессы формирования цен и предсказания спроса на билеты.

Курсовая работа включает 3 главы, введение, заключение и приложение. Работа представлена на 67 страницах, содержит 9 приложения. Список использованных источников включает 17 наименований.

1. **Анализ факторов, влияющих на формирование цен на железнодорожные билеты**
   1. **Краткая характеристика ОАО РЖД**

ОАО "РЖД" (Открытое Акционерное Общество "Российские Железные Дороги") является одним из крупнейших транспортных компаний в России и мире, играющим ключевую роль в обеспечении транспортной доступности и связности внутри страны и за её пределами. Компания осуществляет управление инфраструктурой железнодорожного транспорта, перевозку пассажиров и грузов, развитие и модернизацию железнодорожной сети.

ОАО "РЖД" управляет одной из самых протяжённых железнодорожных сетей в мире, охватывающей все крупные города России и обеспечивающей связь с портами, границами и промышленными центрами. Инфраструктура включает в себя множество станций, разветвлённую сеть путей, мостов и туннелей, обеспечивающих непрерывное движение поездов.

Основной вид деятельности — перевозка грузов и пассажиров. В сегменте грузоперевозок РЖД занимает лидирующие позиции по объёмам перевозок в России, обслуживая промышленные предприятия, порты и крупные распределительные центры. Пассажирские перевозки осуществляются по многочисленным маршрутам, включая дальнее следование, пригородное сообщение и международные рейсы.

ОАО "РЖД" активно инвестирует в расширение и модернизацию железнодорожной сети, строительство новых линий и развитие существующих маршрутов. Особое внимание уделяется развитию высокоскоростного и сверхвысокоскоростного железнодорожного сообщения, которое станет основой для создания нового качества пассажирских перевозок в России. Эти проекты призваны не только сократить время в пути, но и сделать путешествия более комфортными и доступными для широких слоев населения.

ОАО "РЖД" продолжает оставаться одним из столпов российской экономики, играя важную роль в развитии транспортной инфраструктуры и укреплении связей между различными регионами страны и мира. Стремление к инновациям, фокус на устойчивое развитие и социальную ответственность делают компанию важным участником в достижении целей национального развития и обеспечении благополучия будущих поколений.

* 1. **Сезонность и её влияние на спрос и предложение билетов**

Сезонность играет ключевую роль в формировании спроса и предложения на рынке железнодорожных перевозок. Она оказывает значительное влияние на ценообразование билетов и стратегии управления местами в поездах, что в свою очередь влияет на доходность и эффективность перевозочной деятельности. Этот фактор обусловлен множеством причин, включая изменения в потребительских предпочтениях, связанные с сезонами года, праздниками, каникулами и другими социально-экономическими явлениями.

Сезонность в контексте железнодорожных перевозок можно определить как периодические колебания в спросе на билеты, вызванные различными факторами, включая климатические условия, школьные каникулы, государственные и религиозные праздники. Эти колебания предсказуемы и имеют тенденцию повторяться каждый год, что позволяет операторам железных дорог заранее планировать свои ценовые и маркетинговые стратегии.

Спрос на железнодорожные билеты значительно увеличивается во время праздников и школьных каникул, когда люди стремятся посетить родных и друзей или отправиться в отпуск. Летний сезон и зимние праздники являются пиковыми периодами для железнодорожных перевозок, когда спрос может превышать предложение, ведя к увеличению цен на билеты и необходимости добавления дополнительных рейсов.

Для адаптации к колебаниям спроса, вызванным сезонностью, железнодорожные компании часто корректируют своё расписание и количество доступных мест в поездах. В пиковые периоды может быть увеличено количество рейсов или предложены поезда с большей вместимостью, в то время как в периоды низкого спроса количество рейсов может быть сокращено для снижения издержек.

Ценообразование является ключевым инструментом для управления спросом и предложением в условиях сезонных колебаний. В периоды высокого спроса цены на билеты могут быть повышены для максимизации доходов и управления загрузкой поездов. В межсезонье, для привлечения пассажиров, могут быть предложены скидки и специальные предложения. Эффективное применение динамического ценообразования позволяет железнодорожным компаниям адаптироваться к изменениям в спросе и оптимизировать свою прибыль.

Для адаптации к сезонным колебаниям и оптимизации стратегий ценообразования, железнодорожные компании активно используют анализ данных и предсказательные модели. Сбор и анализ исторических данных о продажах билетов, загрузке поездов, а также внешних факторов, таких как школьные каникулы и праздники, позволяют точно прогнозировать будущие тенденции спроса. Эти данные используются для адаптации расписания, планирования маршрутов и корректировки ценовой политики в соответствии с предполагаемыми колебаниями спроса.

Развитие технологий искусственного интеллекта и машинного обучения открывает новые возможности для оптимизации управления сезонными колебаниями в железнодорожной отрасли. Применение алгоритмов прогнозирования позволяет не только с высокой точностью оценивать будущий спрос, но и автоматически адаптировать цены, количество рейсов и состав поездов для максимизации доходности и удовлетворенности пассажиров. Интеграция этих технологий в системы управления перевозками позволяет значительно повысить эффективность операций в условиях меняющегося рынка.

В качестве иллюстративных примеров могут быть рассмотрены практики крупных железнодорожных операторов, таких как ОАО "РЖД" в России, Deutsche Bahn в Германии, и Amtrak в США. Эти компании используют сложные системы аналитики и предсказательные модели для управления сезонными колебаниями спроса и оптимизации своей деятельности. Регулярный анализ данных о продажах, пассажиропотоках и внешних факторах позволяет им не только эффективно распределять транспортные средства и корректировать цены, но и вносить изменения в маркетинговую стратегию, направленную на увеличение загрузки в межсезонье.

Таким образом сезонность является одним из основных факторов, определяющих динамику спроса и предложения на рынке железнодорожных перевозок. Эффективное управление сезонными колебаниями требует от операторов железных дорог не только глубокого понимания рынка и поведения потребителей, но и применения современных технологий анализа данных и искусственного интеллекта. Благодаря использованию данных подходов компании могут оптимизировать свою деятельность, увеличивать доходы и повышать уровень удовлетворенности клиентов, адаптируясь к изменениям в спросе и предлагая пассажирам более выгодные и удобные условия путешествия.

* 1. **Влияние класса обслуживания на ценообразование**

Класс обслуживания является одним из ключевых факторов, влияющих на ценообразование в сфере железнодорожных перевозок. Он определяет уровень комфорта и набор услуг, предоставляемых пассажирам во время поездки, что напрямую влияет на стоимость билета. Вариативность классов обслуживания позволяет пассажирам выбирать оптимальный вариант в соответствии со своими предпочтениями и финансовыми возможностями, в то время как железнодорожным компаниям — гибко управлять ценовой политикой и загрузкой поездов.

В железнодорожном транспорте традиционно выделяют несколько классов обслуживания, начиная от эконом-класса (плацкарт) до бизнес-класса (СВ) и первого класса. Каждый класс предлагает различный уровень комфорта и дополнительных услуг: больше пространства для ног, лучшее питание, повышенное внимание персонала, отдельные купе и т. д. Соответственно, чем выше класс обслуживания, тем выше и цена билета.

Ценообразование билетов сильно зависит от класса обслуживания, поскольку каждый класс направлен на определенную аудиторию пассажиров и предполагает разный уровень затрат на обслуживание. Как правило, билеты в высшие классы обслуживания стоят значительно дороже, что обусловлено не только более высокими издержками на предоставление услуг, но и стремлением железнодорожных компаний максимизировать прибыль с более состоятельных пассажиров.

Железнодорожные компании используют дифференциацию цен между классами обслуживания как инструмент управления спросом и предложением. Высокие цены на билеты в бизнес-классе и первый класс ограничивают доступ к ним широкого круга пассажиров, что позволяет сохранять эксклюзивность и высокий уровень комфорта. В то же время, доступные цены на билеты в эконом-классе привлекают большее количество пассажиров, что способствует повышению общей загрузки поезда и оптимизации доходов компании.

Стратегии ценообразования в зависимости от класса обслуживания могут включать предложение специальных акций, скидок и лояльностных программ для стимулирования продаж в определенных сегментах. Например, раннее бронирование билетов в высшие классы может предлагаться со скидкой, что стимулирует пассажиров заранее планировать свои поездки и обеспечивает железнодорожную компанию гарантированным доходом.

Класс обслуживания оказывает существенное влияние на формирование цен на железнодорожные билеты, обеспечивая компаниям гибкость в стратегии ценообразования и позволяя удовлетворять потребности различных категорий пассажиров. Дифференциация цен не только отражает различия в уровне комфорта и предлагаемых услугах, но и способствует оптимизации загрузки поездов и максимизации доходности перевозок.

Адаптация ценовой политики к потребностям и предпочтениям различных групп пассажиров, а также эффективное управление качеством обслуживания в каждом классе, способствуют повышению удовлетворенности клиентов и укреплению лояльности к бренду. В результате железнодорожные компании могут не только увеличивать свои доходы за счет дифференциации цен, но и эффективно конкурировать с другими видами транспорта, предлагая пассажирам оптимальное соотношение цены и качества услуг.

* 1. **Расстояние и маршрут как детерминанты цены**

Расстояние и маршрут являются фундаментальными факторами, влияющими на формирование цен на железнодорожные билеты. Эти параметры напрямую связаны с издержками на перевозку и определяют базовую стоимость путешествия, от которой затем отталкиваются при формировании конечной цены билета. Понимание влияния этих детерминантов на ценообразование позволяет железнодорожным компаниям оптимизировать свои тарифы и предложения для разных категорий пассажиров.

Общепринято, что стоимость перевозки напрямую зависит от расстояния: чем дальше нужно доставить пассажира, тем выше будут затраты на транспортировку и, соответственно, цена билета. Эта зависимость обусловлена необходимостью покрыть расходы на топливо, износ рельсового состава и другие операционные издержки, которые возрастают с увеличением пройденного пути.

Маршрут также существенно влияет на ценообразование, поскольку различные направления могут иметь разную степень популярности и загруженности. На популярных направлениях, где наблюдается высокий спрос, цены на билеты могут быть выше из-за желания железнодорожных компаний максимизировать доходы. Кроме того, на стоимость влияют сложность маршрута, наличие пересадок и проходимость через густонаселённые районы или труднопроходимые участки, такие как горы или мосты, требующие дополнительных затрат на обслуживание и техническое обеспечение.

Железнодорожные компании часто вводят дифференцированные тарифы в зависимости от расстояния и маршрута. Например, для дальних поездок могут предлагаться скидки при раннем бронировании или специальные тарифы для транзитных пассажиров. Такая дифференциация помогает оптимизировать загрузку поездов и увеличивать общую доступность железнодорожных перевозок.

В стратегиях ценообразования учитывается не только фактическое расстояние и сложность маршрута, но и конкурентная среда. На направлениях, где железнодорожные перевозки конкурируют с авиацией или автобусными перевозками, цены могут корректироваться в сторону уменьшения для сохранения привлекательности железной дороги. Аналогично, на эксклюзивных маршрутах, где железнодорожный транспорт предлагает уникальные преимущества, цены могут быть установлены выше, отражая эксклюзивность и уникальность предложения.

Географическое положение маршрутов также играет значительную роль в ценообразовании. Маршруты, проходящие через районы с высоким уровнем жизни или туристические зоны, могут иметь более высокие цены из-за повышенного спроса и готовности пассажиров платить больше за путешествие в эти регионы. В то же время, направления, проходящие через менее развитые или отдалённые районы, могут предлагать более доступные цены для привлечения пассажиров.

Кроме расстояния и сложности маршрута, на ценообразование влияет и время в пути. Быстрые поезда, сокращающие время путешествия между городами, часто стоят дороже, отражая высокую ценность времени для пассажиров. Удобство маршрута, включая количество остановок и необходимость пересадок, также может влиять на стоимость билетов, поскольку прямые маршруты обычно предпочтительнее и, соответственно, могут стоить дороже.

В некоторых случаях государственное регулирование и субсидирование могут влиять на ценообразование железнодорожных перевозок, особенно на маршрутах социального значения. Цены на таких маршрутах могут быть искусственно снижены за счёт государственной поддержки, чтобы обеспечить доступность и мобильность населения.

Таким образом, расстояние и маршрут являются важными детерминантами в процессе формирования цен на железнодорожные билеты. Они напрямую влияют на издержки перевозок и определяют стратегии ценообразования, которые железнодорожные компании применяют для максимизации своих доходов и удовлетворения потребностей различных групп пассажиров. Учёт географических особенностей, времени в пути и конкурентной среды позволяет операторам железных дорог гибко управлять своими тарифами, предлагая пассажирам оптимальные условия путешествия.

1. **Построение модели предсказания спроса на билеты**
   1. **Описание набора данных и методологии анализа в Loginom**

Для демонстрации анализа данных на примере работы с набором данных о продажах билетов железнодорожного транспорта используется платформа Loginom, которая представляет собой мощный инструмент для обработки и анализа данных. Loginom обладает широким спектром функциональных возможностей для выполнения задач data mining, обработки больших данных, а также построения прогностических моделей на их основе.

Для анализа выбран набор данных, содержащий информацию о продажах билетов железнодорожного транспорта по маршруту Адлер - Москва Казанская за определённый период. Данные включают в себя следующие атрибуты:

- Дата покупки билета: позволяет анализировать динамику продаж во времени.

- Месяц отправления: месяц, когда осуществляется поездка. Это позволяет изучить сезонные колебания спроса.

- Класс обслуживания: указание на класс обслуживания (плацкарт, купе, СВ), что важно для анализа предпочтений пассажиров и ценообразования.

- Цена билета: фактическая стоимость купленного билета.

- Количество проданных мест: информация о количестве проданных мест позволяет оценить загрузку поездов и популярность конкретных рейсов.

Анализ данных в Loginom включает несколько ключевых этапов:

1. Предварительная обработка данных: на этом этапе производится очистка данных от ошибок и пропусков, а также их нормализация и преобразование для последующего анализа. В Loginom доступны различные инструменты для работы с пропущенными значениями, дубликатами, а также для преобразования типов данных.

2. Анализ данных: включает в себя изучение распределений значений, корреляционный анализ, анализ временных рядов и т. д. Loginom предоставляет возможности для визуализации данных, построения графиков и диаграмм, что позволяет глубже понять характеристики и закономерности в данных.

3. Моделирование: на основе анализа данных строятся прогностические модели для предсказания будущих тенденций. В Loginom реализованы различные алгоритмы машинного обучения, включая регрессию, классификацию и кластеризацию, которые могут быть использованы для создания моделей предсказания спроса, ценообразования и других ключевых показателей.

4. Оценка и интерпретация результатов: последний этап включает в себя оценку качества и точности построенных моделей, а также интерпретацию полученных результатов. Loginom позволяет оценивать эффективность моделей с помощью различных метрик и проводить их сравнительный анализ.

Платформа Loginom обеспечивает комплексный подход к анализу данных, предлагая интуитивно понятный интерфейс и широкий спектр инструментов для работы с данными на всех этапах анализа. Это делает её подходящей как для специалистов в области аналитики данных, так и для пользователей с ограниченным опытом в данной области.

* 1. **Выбор метода прогнозирования спроса и построение модели в Loginom**

Прогнозирование спроса является ключевым элементом в планировании ресурсов компании. Возможность точно предсказывать будущий спрос позволяет оптимизировать запасы, управлять производственными мощностями и эффективно распределять бюджет. В данной подглаве рассматриваются методы прогнозирования спроса, которые можно реализовать с помощью аналитической платформы Loginom, и обоснование выбора нейросетевой модели регрессии как наиболее подходящего инструмента для решения задачи.

Loginom предлагает широкий спектр инструментов для анализа данных, среди которых:

- Статистические методы: включают в себя классические временные ряды, такие как ARIMA, экспоненциальное сглаживание и методы на основе скользящего среднего.

- Машинное обучение: к этой категории относятся решающие деревья, случайные леса и градиентный бустинг, которые могут быть использованы для прогнозирования спроса на основе исторических данных.

- Нейросетевые модели: представляют собой сложные алгоритмы, способные выявлять нелинейные зависимости в данных и адаптироваться к изменяющимся условиям рынка.

Перед построением модели необходимо определить ключевые переменные, которые будут использоваться в прогнозировании. Важно учитывать не только исторические данные о продажах, но и внешние факторы, такие как сезонность и ценовые индикаторы.

Нейросетевые модели регрессии обладают рядом преимуществ перед традиционными методами. Они могут обрабатывать большие объемы данных и находить сложные зависимости, что делает их идеальными для задач прогнозирования в условиях неопределенности. Кроме того, Loginom предоставляет инструменты для автоматической настройки и оптимизации моделей, что значительно упрощает процесс моделирования.

Поэтому в ходе проведения научной работы была выбрана, именно нейросетевая модель регрессии для прогнозирования спроса в Loginom.

Для достижения наилучшего результата, было обучено 3 нейросетевые модели регрессии с различными параметрами.

Первая модель показала основные результаты

|  |  |
| --- | --- |
| Среднеквадратическая ошибка на тестовом множестве | 0,5560054826744364 |
| Средняя абсолютная ошибка на тестовом множестве | 0,25890195889855333 |

Таблица 1 – результаты первой модели

Вторая модель показала основные результаты

|  |  |
| --- | --- |
| Среднеквадратическая ошибка на тестовом множестве | 0,4792552205511237 |
| Средняя абсолютная ошибка на тестовом множестве | 0,2605903034080874 |

Таблица 2 – результаты второй модели

Третья модель показала основные результаты

|  |  |
| --- | --- |
| Среднеквадратическая ошибка на тестовом множестве | 0,37839388480016223 |
| Средняя абсолютная ошибка на тестовом множестве | 0,21721283515106957 |

Таблица 3 – результаты третьей модели

График 1– Результаты трех моделей

Из Графика 1 видно, что лучший результат показала третья модель нейросетевой регрессии. Это связано с тем, что входе настроек модели в Loginom в отличии от других моделей на вход были подано большее количество столбцов с данными, которые влияют на спрос, а также был применен метод Монте-Карло, суть которого заключается в то, что:

1. Определяется модель или алгоритм, которые лучше всего соответствуют данным, описывающим исследуемый объект или процесс.
2. Модель многократно применяется к данным, сгенерированным случайным образом в соответствии с распределением вероятности исходных данных, формируя каждый раз некоторую реализацию случайной выходной переменной.
3. Полученные реализации усредняются, и к результатам усреднения применяются статистические методы для оценки параметров вариации, доверительных интервалов и т.д.

Таким образом, смысл метода заключается в моделировании исследуемого процесса путем многократных повторений его случайных реализаций, называемых статистическими испытаниями. [3]

Дополнительно для подтверждения достоверности результатов можно сравнить результаты других показателей моделей.

Одним из важнейших показателей нейросети является показатель Кросс-валидация (CVRMSError, CVAvgError, CVAvgRelError) — это метод оценки точности модели, который заключается в разделении исходного набора данных на несколько поднаборов. Модель обучается на одном поднаборе данных, а затем оценивается на другом поднаборе данных. Этот процесс повторяется для каждого поднабора данных, и результаты объединяются для получения общей оценки точности модели.

Первая модель

|  |  |  |
| --- | --- | --- |
| CVRMSError | Среднеквадратическая ошибка при кросс-валидации | 1,675506527755417 |
| CVAvgError | Средняя абсолютная ошибка при кросс-валидации | 1,2261862007270319 |
| CVAvgRelError | Средняя относительная ошибка при кросс-валидации | 1,089853718227388 |

Таблица 4– результаты кросс-валидации первой модели

Вторая модель

|  |  |  |
| --- | --- | --- |
| CVRMSError | Среднеквадратическая ошибка при кросс-валидации | 0,575081402773625 |
| CVAvgError | Средняя абсолютная ошибка при кросс-валидации | 0,2789645316847782 |
| CVAvgRelError | Средняя относительная ошибка при кросс-валидации | 0,1915083364353341 |

Таблица 5–результаты кросс-валидации второй модели

Третья модель

|  |  |  |
| --- | --- | --- |
| CVRMSError | Среднеквадратическая ошибка при кросс-валидации | 0,5504014184510044 |
| CVAvgError | Средняя абсолютная ошибка при кросс-валидации | 0,20994359568602933 |
| CVAvgRelError | Средняя относительная ошибка при кросс-валидации | 0,12591607838752852 |

Таблица 6– результаты кросс-валидации третьей модели

График 2– Результаты трех моделей

Тем самым можно наглядно увидеть, что действительно третья модель показывает наилучшую точность, в отличии от других моделей, потому что значения CVRMSError, CVAvgError и CVAvgRelError указывают на то, что модель имеет достаточно высокую точность на независимом наборе данных.

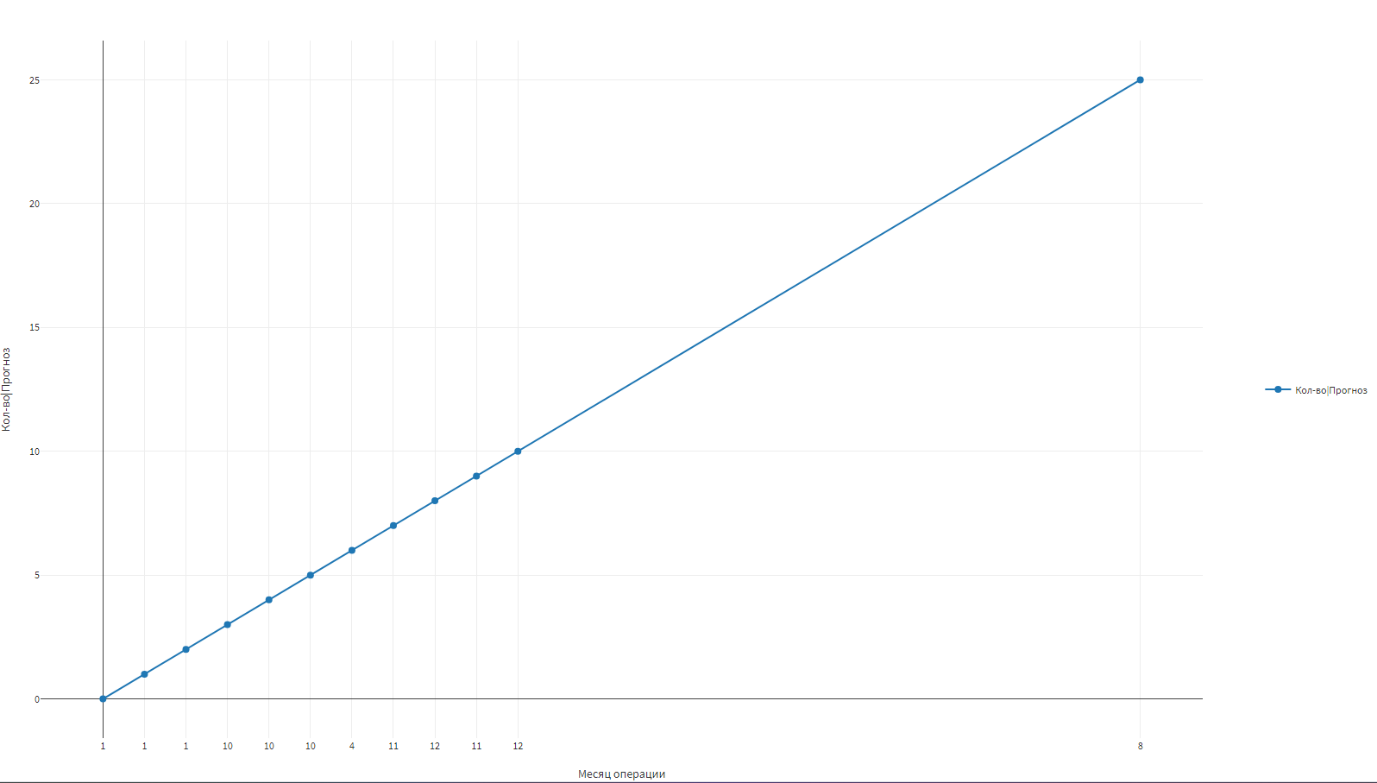


Рисунок 1–Спрогнозированный спрос в зависимости от сезона

Следовательно, если в третью модель загрузить входные данные, а именно второй датасет, который был предоставлен ОАО РЖД, в котором содержатся данные за 2022 год, можно выявить следующую закономерность, которая наглядно отражена на Рисунке 1, что больше всего спросом пользуется 8 месяц, но так же большой спрос покупки ЖД билетов можно увидеть в 11 и 12 месяцах. Тем самым перевозчик благодаря данной модели сможет спрогнозировать спрос исходя из входных данных.

1. **Построение модели адаптивного формирования цены на билеты**
   1. **Обзор возможностей и методов создания модели адаптивного ценообразования**

Адаптивное ценообразование играет важную роль в максимизации доходов и увеличении конкурентоспособности.

Адаптивное ценообразование — это динамичный подход, позволяющий компаниям быстро реагировать на изменения в спросе и предложении, конкуренции, затратах и других внешних факторах. Это требует анализа большого объема данных и применения сложных алгоритмов для определения оптимальных ценовых точек.

Loginom предоставляет различные инструменты для анализа данных и построения моделей ценообразования:

- Статистический анализ: Использование статистических методов для анализа исторических данных о ценах и продажах.

- Машинное обучение: Применение алгоритмов машинного обучения для прогнозирования оптимальных цен на основе текущих рыночных тенденций.

- Оптимизация: Разработка моделей для автоматической корректировки цен в соответствии с заданными бизнес-правилами и целями.

Но исходя из предоставленного набора данных лучше всего для решения данной задачи подойдет нейросетевая регрессия в Loginom.   
Выбор обусловлен несколькими ключевыми преимуществами:

1. Способность к обработке больших объемов данных: Нейросети эффективно работают с большими и сложными наборами данных, что является важным аспектом при анализе факторов, влияющих на ценообразование.
2. Выявление нелинейных зависимостей: Адаптивное ценообразование часто включает сложные взаимосвязи между переменными, которые традиционные линейные модели могут не уловить.
3. Гибкость и адаптивность: Нейросетевые модели могут адаптироваться к изменениям в данных и рыночных условиях, что позволяет поддерживать актуальность модели ценообразования в динамичной среде.
4. Автоматическая настройка параметров: Loginom предоставляет инструменты для автоматической настройки гиперпараметров нейросети, что упрощает процесс моделирования и сокращает время на подготовку модели.
5. Повышение точности прогнозов: Нейросети, благодаря своей структуре и способности к глубокому обучению, могут предоставлять более точные прогнозы, что критически важно для определения цен в реальном времени.

Использование нейросетевой модели регрессии в Loginom для адаптивного ценообразования позволяет не только повысить точность прогнозирования цен, но и обеспечить более глубокое понимание рыночных тенденций и поведения потребителей, что в конечном итоге способствует более эффективному управлению доходами и конкурентоспособности.

* 1. **Построение в Loginom модели адаптивного ценообразования**

В прошлой главе в ходе анализа возможностей реализации построения модели адаптивного ценообразования, была выбрана нейросеть (регрессия).

Как рассматривалось ранее на ценообразование влияют множество факторов. Поэтому для построения модели и повышения точности, нужно построить несколько моделей с разными параметрами и входными данными. Что и было сделано, в ходе проведения научной работы.

В ходе проведения научной работы было создано 4 модели с разными настройками и параметрами нейросетевой регрессии в Loginom.

Первая модель показала такие результаты

|  |  |
| --- | --- |
| Среднеквадратическая ошибка на тестовом множестве | 0,21645715739281354 |
| Средняя абсолютная ошибка на тестовом множестве | 0,13225759830119163 |

Таблица 7– результаты первой модели

Вторая модель показала такие результаты

|  |  |
| --- | --- |
| Среднеквадратическая ошибка на тестовом множестве | 7 152,640228245669 |
| Средняя абсолютная ошибка на тестовом множестве | 2 550,8449903955993 |

Таблица 8– результаты второй модели

Третья модель показала такие результаты

|  |  |
| --- | --- |
| Среднеквадратическая ошибка на тестовом множестве | 0,04653008152640707 |
| Средняя абсолютная ошибка на тестовом множестве | 0,025485042356069135 |

Таблица 9–результаты третьей модели

Четвертая модель показала такие результаты

|  |  |
| --- | --- |
| Среднеквадратическая ошибка на тестовом множестве | 0,09379014192883188 |
| Средняя абсолютная ошибка на тестовом множестве | 0,07822972423632561 |

Таблица 10 –результаты четверной модели

График 2– Результаты четырех моделей

Из Графика 2 видно, что лучший результат показала третья модель нейросетевой регрессии, однако и четвертая модель показала отличный результат. Это связано с тем, что входе настроек этих двух моделей в Loginom в отличии от других моделей в том, что была определена степень регуляции, а также число итераций. В третьей модели использовался метод Монте-Карло, который описывался выше, а в 4 модели использовался метод Кросс-валидации, который заключается в том, что:

В основе метода лежит разделение исходного множества данных на K примерно равных блоков, например k=5. Затем на k−1, т.е. на 4-х блоках, производится обучение модели, а 5-й блок используется для тестирования. Процедура повторяется K раз, при этом на каждом проходе для проверки выбирается новый блок, а обучение производится на оставшихся.

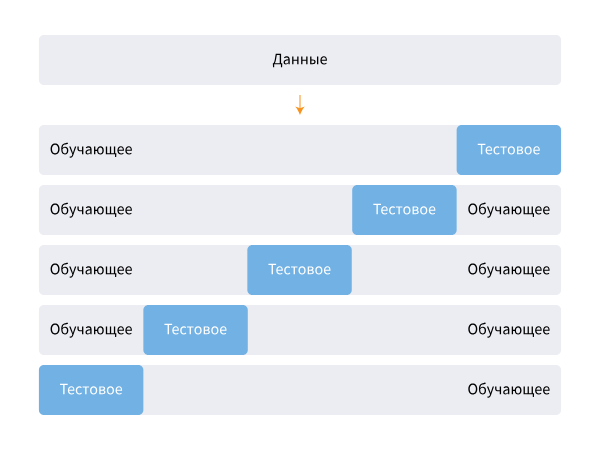
****

Рисунок 2–Принцип работы метода Кросс-валидации

Перекрестная проверка имеет важное преимущества перед применением одного множества для обучения и одного для тестирования модели: если при каждом проходе оценить выходную ошибку модели и усреднить ее по всем проходам, то полученная ее оценка будет более достоверной. [4]

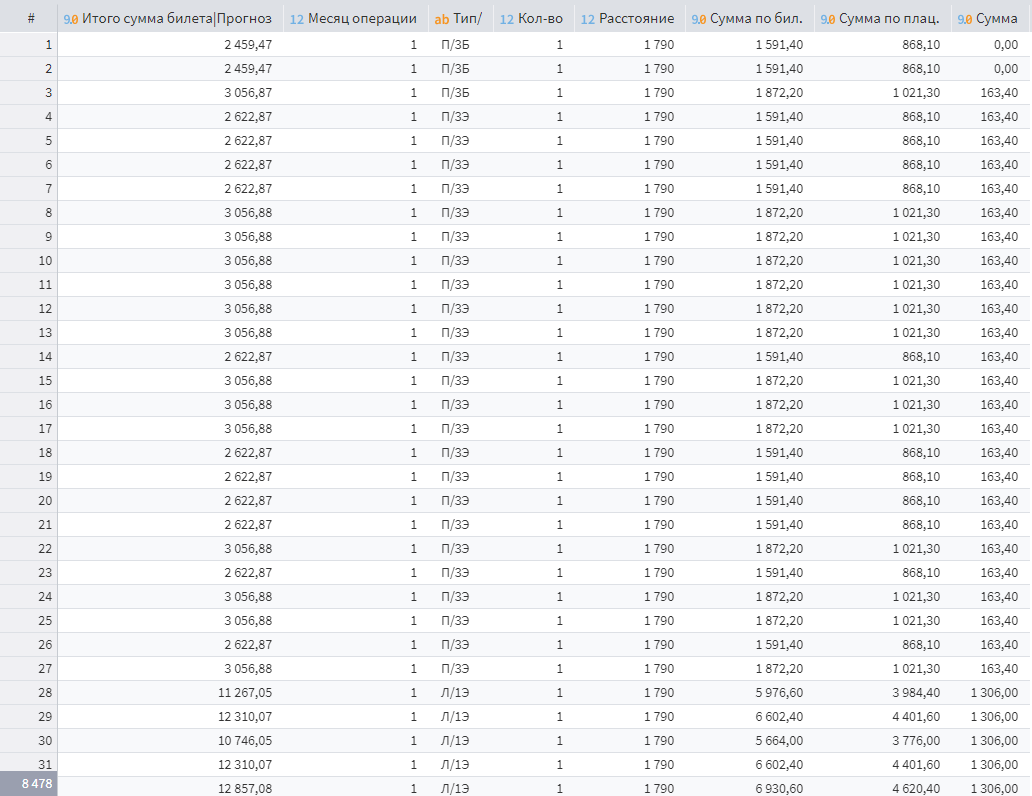
****

Рисунок 3–Принцип работы метода Кросс-валидации

На Рисунке 3 представлен фрагмент таблицы, в котором нейросеть сформировала цену на билет относительно входных данных. Таким образом используя данную модель, можно адаптивно менять цены на билеты, в зависимости от входных параметров. Так же можно к каждому месяцу отправления добавлять «вес» исходя из модели адаптации спроса в зависимости от сезона.

Дополнением к данному анализу может служить разделение ценового сегмента купленных билетов на «дешевые» и «премиум». Для того чтобы реализовать данный анализ, нужно выйти медианное значение в столбце Итоговой суммы, медианным значением является 5036,067. Для вычисления анализа можно использовать узел в Loginom Логическая регрессия. Логическая регрессия представляет собой разновидность множественной регрессии, общее назначение которой состоит в анализе связи между несколькими независимыми переменными и зависимой переменной. [11]

За выходной параметр была взят столбец Итого сумма билета\_1 с параметрами вид данных дискретный, который был выбран логический типа данных, для того чтобы получить нужный результат, где   
false– это билеты «премиум» класса, а значение true – это билеты «дешевого» класса.

Изображение выглядит как текст, снимок экрана, диаграмма, линия

Автоматически созданное описание

Рисунок 4 –ROC-кривая

Результаты, представленные на ROC-кривой, дают информацию о производительности классификатора. Вот основные моменты:

- TPR (True Positive Rate), также известный как чувствительность, показывает, какую долю реальных положительных случаев классификатор определяет правильно. Значение TPR равно 86,74%, что означает, что примерно 87% истинных положительных результатов были корректно идентифицированы моделью.

- FPR (False Positive Rate) отражает, какую долю реальных отрицательных случаев классификатор ошибочно считает положительными. Здесь FPR равно 13,26%, следовательно, около 13% отрицательных случаев были неверно классифицированы как положительные.

- Порог отсечения, обозначенный на графике, представляет собой значение, при котором достигается баланс между TPR и TNR. В данном случае порог равен 0,66. Это означает, что если выставить порог классификации на 0,66, то получится наиболее сбалансированное соотношение между чувствительностью и специфичностью.

Таким образом, классификатор работает достаточно хорошо, если считать оптимальным порог в 0,66 для баланса между обнаружением положительных случаев и избежанием ложных срабатываний. Но нужно учитывать, что некоторый уровень ошибок все еще присутствует.

Качество модели также можно оценить по близости кривой к левому верхнему углу графика. Чем ближе кривая к этому углу, тем выше TPR при низком FPR, что говорит о более высоком качестве классификации. [13]

Изображение выглядит как текст, снимок экрана, Шрифт, число

Автоматически созданное описание

Рисунок 5– Оценки классификации

На Рисунке 5 представлены метрики качества бинарного классификатора для обучающего и тестового наборов данных. Вот подробное описание:

- AUC ROC (Area Under the Curve - Receiver Operating Characteristic curve): это площадь под ROC-кривой, которая измеряет способность классификатора отличать два класса. Чем ближе значение к 1, тем лучше. Значения составляют 0,9373 для обучающего и 0,9385 для тестового набора данных, что указывает на высокую производительность классификатора.

- AUC PR (Area Under the Curve - Precision-Recall curve): аналогично AUC ROC, это площадь под кривой точности-полноты. Значения для обучающего и тестового наборов данных составляют 0,8373 и 0,8563 соответственно, что также указывает на хорошее качество классификации.

- Коэффициент Джини: рассчитывается как . Высокие значения указывают на лучшее различение классификатором классов. Здесь они составляют 0,8746 для обучающего набора и 0,8770 для тестового, что свидетельствует о сильной разделительной способности. [12]

- KS (Kolmogorov-Smirnov statistic): это статистика, используемая для определения порога, при котором различия между распределениями истинно положительных и истинно отрицательных результатов максимальны. Значения составляют 83,1683 и 83,0909, что может указывать на хороший порог разделения.

- TPR (True Positive Rate) и TNR (True Negative Rate): Оба показателя равны 0,8674 и 0,8667 для обучающего и тестового наборов данных соответственно, показывая, что классификатор хорошо идентифицирует как положительные, так и отрицательные случаи.

- FPR (False Positive Rate): Минимальные различия между обучающим (0,1326) и тестовым (0,1333) наборами данных указывают на стабильность классификатора в отношении ложноположительных результатов.

- PPV (Positive Predictive Value), также известный как точность, и F1 Score: Эти показатели качества комбинируют точность и полноту, где PPV равен 0,7797 и 0,7787, а F1 Score равен 0,8212 и 0,8203 для обучающего и тестового наборов данных соответственно. Это говорит о том, что классификатор дает высокий уровень точности и баланса между точностью и полнотой.

- MCC (Matthews Correlation Coefficient): это коэффициент корреляции, который учитывает все четыре квадранта матрицы ошибок. Значения 0,7189 и 0,7175 свидетельствуют о хорошем качестве классификации.

Обобщая, эти показатели свидетельствуют о том, что классификатор обладает высокой предсказательной способностью и стабильно работает как на обучающем, так и на тестовом наборах данных. [14]

Изображение выглядит как текст, снимок экрана, Шрифт, число

Автоматически созданное описание

Рисунок 6 – Матрица ошибок

На Рисунке 6 показаны две матрицы ошибок для обучающего и тестового наборов данных, а также два индикатора распознавания.

Матрица ошибок — это специальная таблица, которая используется для визуализации производительности алгоритма классификации. Она показывает количество верных и ошибочных предсказаний, сгруппированных по их фактическим и предсказанным классификациям.

Верхняя матрица относится к обучающему набору данных, а нижняя — к тестовому. Каждая матрица разделена на четыре части:

- Истинно положительные (True Positives, TP): Классификатор верно определил события. В обучающем наборе это 1,819 и в тестовом — 774.

- Истинно отрицательные (True Negatives, TN): Классификатор верно определил не-события. В обучающем наборе это 3,329 и в тестовом — 1,430.

- Ложно положительные (False Positives, FP): Классификатор ошибочно определил не-события как события. В обучающем наборе это 509 и в тестовом — 220.

- Ложно отрицательные (False Negatives, FN): Классификатор ошибочно определил события как не-события. В обучающем наборе это 278 и в тестовом — 119.

В нижней части изображения показаны индикаторы распознавания, где указано количество правильно классифицированных случаев (TP + TN) к общему количеству случаев (TP + TN + FP + FN) в процентном соотношении. Для обучающего набора это составляет 5,148 правильных из 5,935 возможных, а для тестового — 2,204 из 2,543. Эти значения показывают общую точность классификатора на каждом из наборов данных. [15]

Изображение выглядит как текст, диаграмма, линия, График

Автоматически созданное описание

Рисунок 7 –Базовые показатели

На Рисунке 7 изображена диаграмма с несколькими графиками, показывающими различные показатели эффективности классификационной модели в зависимости от порога отсечения. Это графики для TPR (True Positive Rate или чувствительность), TNR (True Negative Rate или специфичность), FPR (False Positive Rate или 1 - специфичность) и FNR (False Negative Rate или 1 - чувствительность).

- Синий график (TPR) быстро достигает высокого значения и затем остается постоянным, что означает, что модель успешно идентифицирует истинные положительные случаи при увеличении порога отсечения до определенной точки, после которой увеличение порога не влияет на TPR.

- Фиолетовый график (TNR) также растет и достигает высоты при определенном пороге, что означает улучшение способности модели отличать истинные отрицательные случаи от положительных.

- Зеленый график (FPR) показывает, что частота ложноположительных результатов уменьшается при увеличении порога отсечения.

- Красный график (FNR) демонстрирует уменьшение частоты ложноотрицательных результатов при повышении порога до определенной точки, после которой он начинает расти.

Оранжевая пунктирная линия обозначает порог отсечения, который, вероятно, выбран для балансировки между различными показателями. Это место, где модель находит компромисс между чувствительностью и специфичностью, максимизируя оба показателя.

В целом, диаграмма помогает определить оптимальный порог для классификации событий, который максимизирует истинно положительные и истинно отрицательные результаты, минимизируя при этом ложные срабатывания и пропуски. Это важно для разработки эффективной модели, способной точно разделять классы, что, в свою очередь, может влиять на бизнес-решения, такие как ценообразование билетов или их классификация на дешевые и премиальные. [14]

Изображение выглядит как текст, линия, График, диаграмма

Автоматически созданное описание

Рисунок 8 – Диаграмма % распознанных событий

На Рисунке 8 показан процент распознанных событий (кумулятивная диаграмма), относящийся к классификации итоговой суммы билета. Эта диаграмма используется для визуализации накопленного распределения и оценки эффективности классификационной модели.

Ось абсцисс (горизонтальная) показывает долю выборки в процентах — это то, какую часть всего набора данных составляют рассматриваемые случаи. Ось ординат (вертикальная) отображает процент распознанных событий, также в процентах.

На диаграмме представлены две линии:

- Оранжевая линия относится к обучающему набору данных. Она показывает, как процент правильно распознанных событий увеличивается по мере рассмотрения большего количества данных.

- Синяя линия относится к тестовому набору данных и демонстрирует аналогичное поведение для тестовых данных.

Пунктирные линии показывают базовую линию и идеальную модель:

- Базовая линия (коричневая пунктирная линия) представляет собой диагональ, которая показывает, что процент распознанных событий равен доле выборки — это означает случайное угадывание без какой-либо предсказательной способности.

- Идеальная линия показывает гипотетическую ситуацию, когда модель идеально классифицирует все события без ошибок.

Из диаграммы видно, что и обучающая, и тестовая линии выше базовой, что говорит о том, что модель обладает предсказательной способностью и работает лучше, чем случайное угадывание. Однако они не достигают идеальной производительности, что является нормальным для реальных задач классификации. [14]

Изображение выглядит как текст, диаграмма, снимок экрана, График

Автоматически созданное описание

Рисунок 9 – Precision-Recall кривая

На Рисунке 9 представлена Precision-Recall кривая (кривая точности и полноты), которая является важным инструментом для оценки качества бинарных классификаторов, особенно в случаях, когда классы распределены неравномерно (например, когда один класс существенно реже другого).

Ось X (горизонтальная ось) показывает TPR (True Positive Rate), также известный как полнота или recall. Ось Y (вертикальная ось) показывает PPV (Positive Predictive Value), известный также как точность или precision.

- Оранжевая линия отображает Precision-Recall кривую для обучающего набора данных.

- Синяя линия отображает Precision-Recall кривую для тестового набора данных.

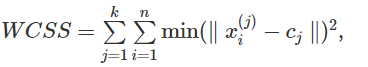
Оранжевая пунктирная линия показывает выбранный порог отсечения, который в данном случае обеспечивает баланс между TPR и PPV со значениями 0,66 (порог), 86,74% (TPR) и 77,97% (PPV). Это означает, что при данном пороге классификатор достигает указанной полноты и точности.

Precision-Recall кривые помогают определить, на каком уровне порога отсечения классификатор дает лучшее сочетание точности и полноты. Это особенно важно в приложениях, где важнее избегать ложноположительных результатов (для поддержания высокой точности) или важнее выявлять все положительные случаи (для поддержания высокой полноты). [16**]**

Так же для создания ценообразования можно воспользоваться методом «локтя». Метод локтя (Elbow method) — инструмент анализа данных, направленный на оптимизацию числа кластеров в алгоритмах кластеризации. Впервые был предложен Робертом Л. Торндайком в 1953 году.

Правильно подобранное количество кластеров в алгоритмах позволяет найти баланс между погрешностью вычисляемой дисперсии и сложностью модели. Использование метода позволяет избежать недообучения или переобучения алгоритма кластеризации.

На первом этапе для различных значений числа кластеров K вычисляется сумма квадратов расстояний каждой точки данных до их центроида (центра тяжести) (WCSS) по формуле:



**Второй этап** содержит построение графика зависимости WCSS от количества кластеров, где по оси X откладывается число кластеров k, а по оси Y — соответствующая сумма квадратов расстояний.

**Третий этап** заключается в поиске точки излома («локтя») на графике, которая указывает на оптимальное число кластеров. Оптимальным k будет то, при котором ошибка перестает существенно уменьшаться, т.е. начинает сглаживаться.

График может иметь следующий вид:

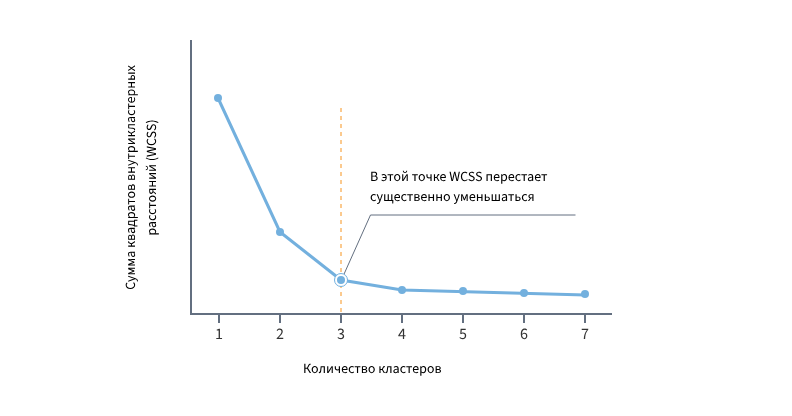


Рисунок – 10 график метода локтя

На основание данного графика можно определить, что оптимальным будет использование трех кластеров.

В методе локтя основной акцент делается на визуальный анализ. Если линейный график выглядит как рука, то «локоть» (точка перегиба на кривой) является наилучшим значением k. Притом «рука» может быть направлена как вверх, так и вниз. [17]

Для нашей задачи, как говорилось выше мы будем использовать метод «локтя». Для начала нужно создать кластеризацию данных. На вход были поданы следующие столбцы:



Рисунок –11 данные, подаваемые на вход кластеризации

На выходе мы получили разделение на кластеры

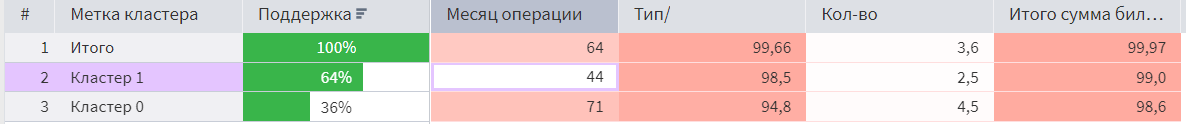


Рисунок – 12 результаты кластеризации

Из результатов кластеризации можно сделать небольшой вывод по каждому кластеру:

1. Кластер 1 - Этот кластер включает 64% всех данных. Средний месяц операции - 44, что указывает на то, что продажи в этом кластере происходят в среднем в начале года. Средний тип билета - 98,5, среднее количество билетов - 2,5, и средняя итоговая сумма билета - 99,0. Этот кластер может представлять группу клиентов, покупающих билеты заранее и в меньших количествах.
2. Кластер 2 - Этот кластер включает 36% всех данных. Средний месяц операции - 71, что указывает на то, что продажи в этом кластере происходят в среднем ближе к концу года. Средний тип билета - 94,8, среднее количество билетов - 4,5, и средняя итоговая сумма билета - 98,6. Этот кластер может представлять группу клиентов, покупающих билеты ближе к дате отправления и в больших количествах.

Стоит так же отметить, что также на вход кластеризации были поданы переменные

Изображение выглядит как текст, снимок экрана, Шрифт, число

Автоматически созданное описание

Рисунок–13 переменные, поданные на вход кластеризации

Они нужны для того, чтобы обозначить минимальное число кластеров и число итераций, для правильного результата метода «локтя».

Для того чтобы построить график по методу «локтя» была создана подмодель, которая имеет следующий вид:

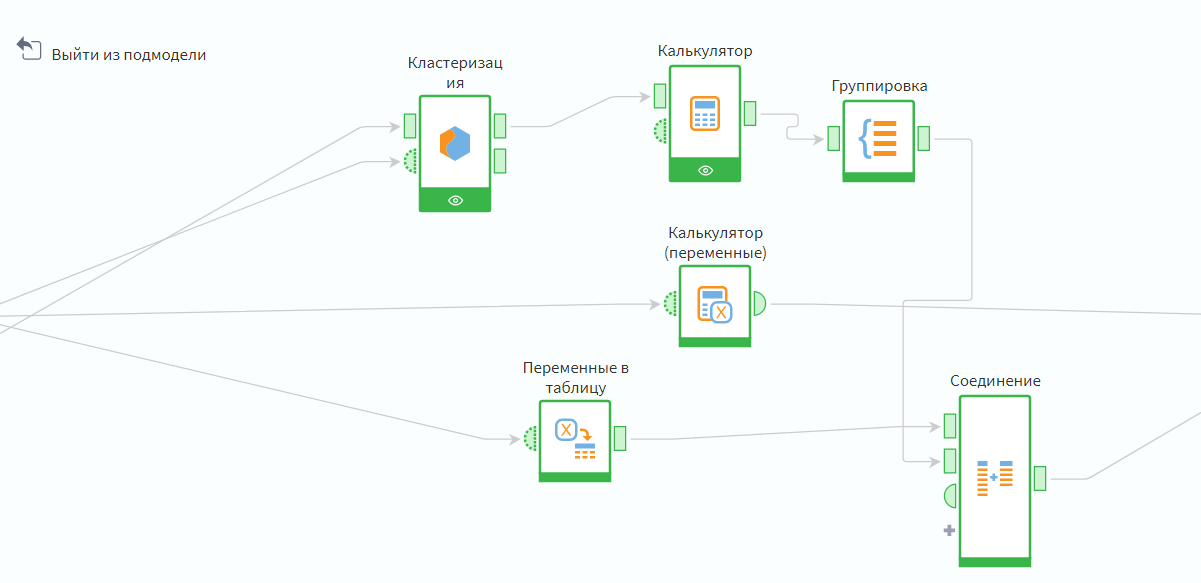


Рисунок 14–структура подмодели

Краткое описание структуры подмодели, показанной на Рисунке выше:

* 1. Кластеризация:
* Узел Кластеризация использует метод k-средних для разделения данных на кластеры.
* Входные параметры включают начальное количество кластеров и число итераций.
  1. Калькулятор:
* Узел Калькулятор используется для вычисления необходимых метрик, таких как сумма квадратов расстояний от точек до центров кластеров.
* Эти метрики помогут определить разброс данных внутри каждого кластера.
  1. Группировка:
* Узел Группировка собирает данные по каждому кластеру и вычисляет средние значения и другие статистические показатели для каждого кластера.
* Это позволяет оценить качество кластеризации.
  1. Калькулятор (переменные):
* Дополнительный узел Калькулятор (переменные) используется для вычисления вспомогательных переменных, необходимых для анализа и визуализации.
  1. Переменные в таблицу:
* Узел Переменные в таблицу преобразует вычисленные переменные в таблицу для последующего анализа.
* Это необходимо для удобства визуализации и интерпретации результатов.
  1. Соединение:
* Узел Соединение объединяет результаты кластеризации с исходными данными или с дополнительными метриками.
* Это позволяет получить полную картину и оценить, как данные распределены по кластерам.

На выходе подмодели мы получили следующие значения

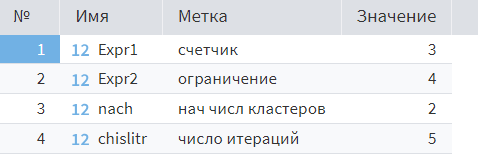
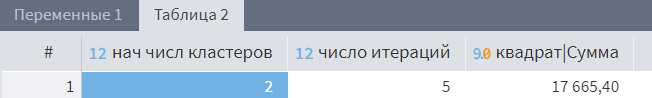


Рисунок – 15 данные полученные на выходе подмодели



Рисунок– 16 данные полученные на выходе подмодели

Следующим и заключительным шалом является добавление компонента Цикл.

Зачем нужен компонент "Цикл"

Метод локтя требует проведения кластеризации с различным числом кластеров, чтобы определить оптимальное количество. Компонент "Цикл" позволяет автоматизировать этот процесс, выполняя кластеризацию многократно с разными параметрами и собирая результаты для дальнейшего анализа.

Настройки компонента "Цикл"

1. Вид цикла: Выбран "Цикл с постусловием", что означает, что цикл будет выполняться до тех пор, пока не будет выполнено заданное условие.
2. Переменная: Указана переменная 12 Переменные 1.число итераций, которая используется для отслеживания числа итераций.
3. Условие завершения: Цикл завершается, когда переменная 12 Переменные 1.число итераций равна нулю.

Процесс выполнения цикла

1. Начальная инициализация переменных:
   * Устанавливаются начальные значения переменных, такие как начальное количество кластеров и число итераций.
2. Кластеризация:
   * В каждой итерации цикла выполняется кластеризация данных с текущим числом кластеров.
3. Вычисление метрик:
   * После каждой кластеризации вычисляется сумма квадратов расстояний от точек до центров кластеров.
4. Сохранение результатов:
   * Результаты каждой итерации сохраняются в таблицу для дальнейшего анализа.
5. Увеличение числа кластеров:
   * Перед переходом к следующей итерации число кластеров увеличивается на 1.
6. Проверка условия завершения:
   * Цикл продолжается до тех пор, пока не будет достигнуто максимальное число кластеров или пока переменная не достигнет значения, определенного в условии завершения.

С подробной настройкой компонента Цикл, можно ознакомиться в Приложении

Результатом цикла является график

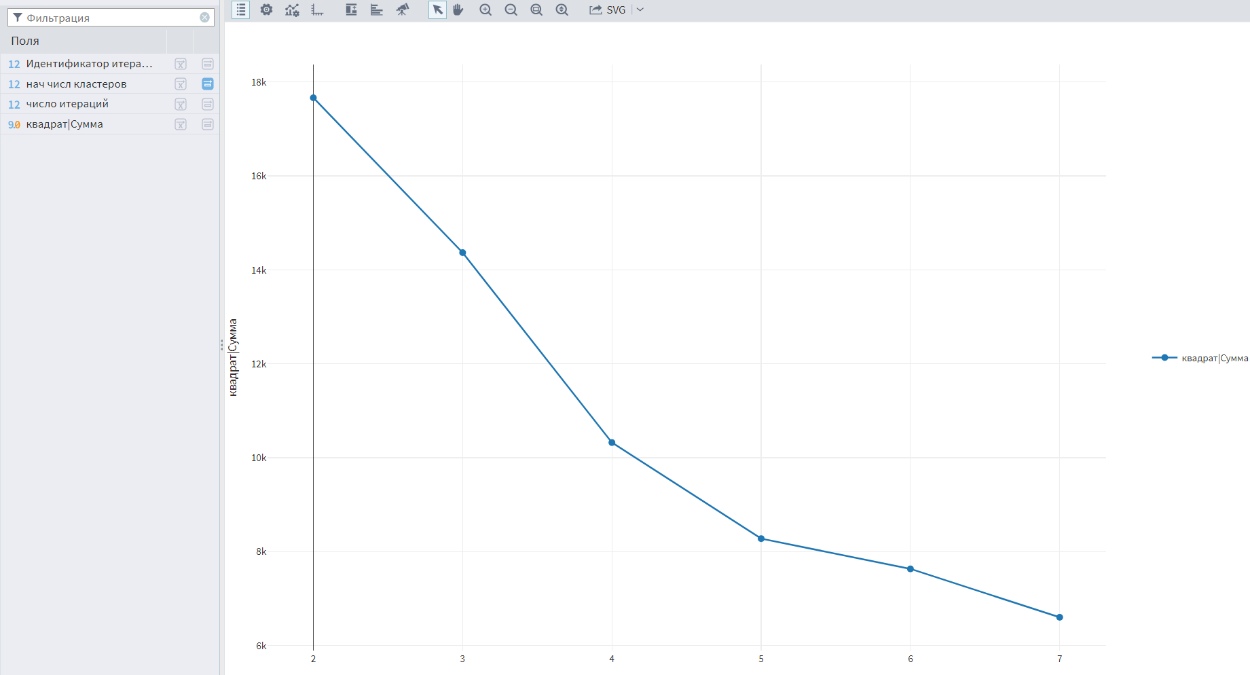


Рисунок – 17 график метода локтя

График локтя, представленный на Рисунке, иллюстрирует зависимость суммы квадратов расстояний от точек до центров кластеров от числа кластеров.

График показывает, как изменяется внутренняя дисперсия (сумма квадратов расстояний) при увеличении числа кластеров.

Резкое снижение: на графике видно резкое снижение суммы квадратов при увеличении числа кластеров от 2 до 4. Это означает, что добавление новых кластеров значительно уменьшает разброс данных внутри кластеров.

Точка локтя: после 5 кластеров уменьшение суммы квадратов становится менее значительным. Точка локтя на графике, где линия начинает сглаживаться, указывает на оптимальное число кластеров. В данном случае точка локтя находится в районе 5 кластеров.

Так же хочется отметить, что в начале создания метода «локтя» число итераций ровнялось 10 и график имел следующий вид:

Изображение выглядит как линия, диаграмма, График, Параллельный

Автоматически созданное описание

Рисунок – 18 график метода локтя

Где отчетливо видно, что после значения 5 точки не имеют резкого перепада в значениях, что говорит о том, что оптимальным числом итераций является 5.

Помимо графика результатом выполнения цикла является таблица с данными:

Изображение выглядит как текст, число, Шрифт, линия

Автоматически созданное описание

Рисунок – 19 результаты выполнения цикла

Таблица результатов цикла содержит следующие данные:

* нач числ кластеров: Начальное число кластеров.
* число итераций: Количество итераций, выполненных для достижения сходимости.
* квадрат(Сумма): Сумма квадратов расстояний от точек до центров кластеров.
* Идентификатор итерации: Номер итерации.

Детальный анализ строк таблицы результата цикла:

Строка 1

* нач числ кластеров: 2
* число итераций: 5
* квадрат(Сумма): 17 665,40
* Идентификатор итерации: 0
* Анализ: При начальном количестве кластеров 2 и 5 итерациях алгоритм достиг суммы квадратов 17 665,40. Это значение указывает на довольно большой разброс данных внутри кластеров, что свидетельствует о том, что два кластера недостаточны для адекватного представления структуры данных.

Строка 2

* нач числ кластеров: 3
* число итераций: 4
* квадрат(Сумма): 14 367,44
* Идентификатор итерации: 1
* Анализ: При увеличении числа кластеров до 3 и проведении 4 итераций, сумма квадратов снизилась до 14 367,44. Это указывает на улучшение сегментации данных, хотя разброс данных внутри кластеров всё ещё достаточно велик.

Строка 3

* нач числ кластеров: 4
* число итераций: 3
* квадрат(Сумма): 10 322,45
* Идентификатор итерации: 2
* Анализ: При использовании 4 кластеров и 3 итераций, сумма квадратов снизилась до 10 322,45. Это значительное улучшение по сравнению с предыдущими значениями, что указывает на более точное разделение данных на кластеры.

Строка 4

* нач числ кластеров: 5
* число итераций: 2
* квадрат(Сумма): 8 276,35
* Идентификатор итерации: 3
* Анализ: При начальном количестве кластеров 5 и всего 2 итерациях сумма квадратов составила 8 276,35. Это показывает, что 5 кластеров являются достаточно хорошим числом для сегментации данных, так как разброс внутри кластеров значительно снизился.

Строка 5

* нач числ кластеров: 6
* число итераций: 1
* квадрат(Сумма): 7 631,84
* Идентификатор итерации: 4
* Анализ: Увеличение числа кластеров до 6 и одна итерация привели к сумме квадратов 7 631,84. Снижение разброса внутри кластеров продолжилось, но улучшение по сравнению с 5 кластерами уже не такое значительное.

Строка 6

* нач числ кластеров: 7
* число итераций: 0
* квадрат(Сумма): 6 599,89
* Идентификатор итерации: 5
* Анализ: При использовании 7 кластеров и отсутствии дополнительных итераций, сумма квадратов составила 6 599,89. Это показывает дальнейшее снижение разброса внутри кластеров, но улучшение становится всё менее значительным.

Этот анализ помогает понять структуру данных и выбрать оптимальное количество кластеров для модели адаптивного формирования цены на билеты.

**Заключение**

На основе проведённого анализа данных и тестирования ряда нейросетевых моделей регрессии, направленных на адаптацию и оптимизацию процессов ценообразования в сфере железнодорожных перевозок, были получены значимые результаты. Исследование подтвердило эффективность применения инструментов искусственного интеллекта для предсказания спроса и адаптивного формирования цен, что открывает новые перспективы для повышения уровня удовлетворенности потребителей и оптимизации использования ресурсов. Особенно значимым стало внедрение методов Монте-Карло и кросс-валидации в процессе обучения и тестирования моделей, что позволило значительно увеличить точность предсказаний.

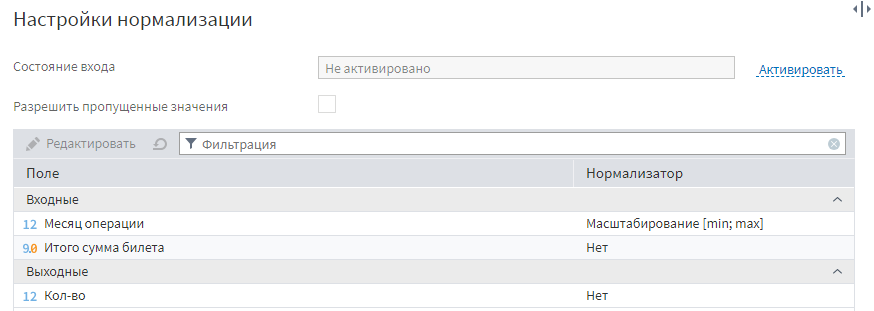
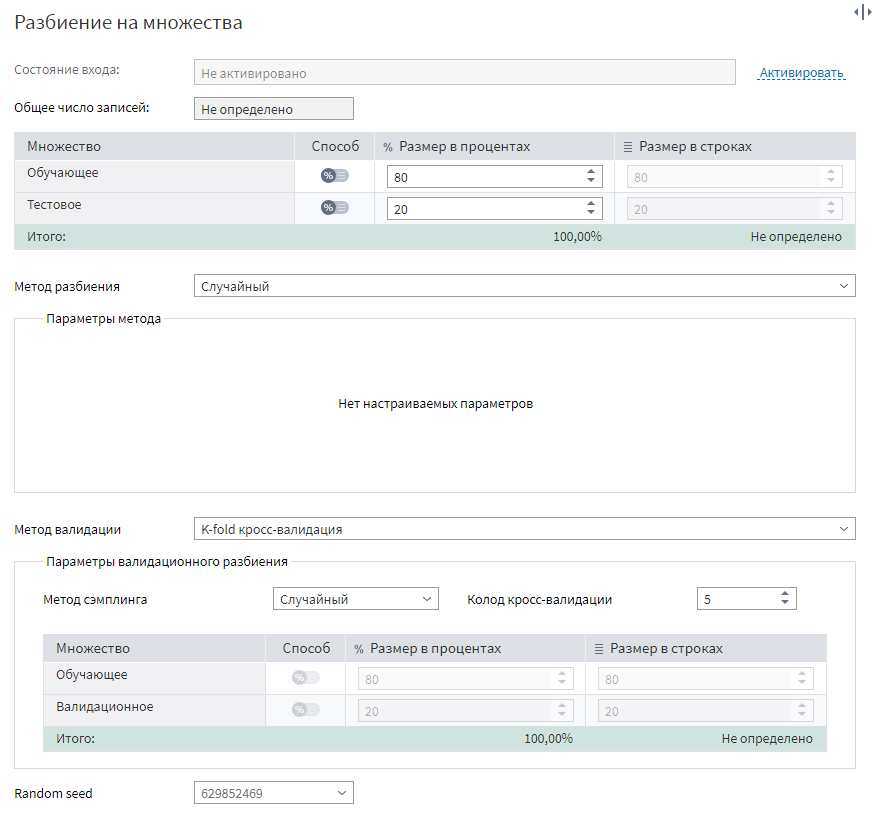
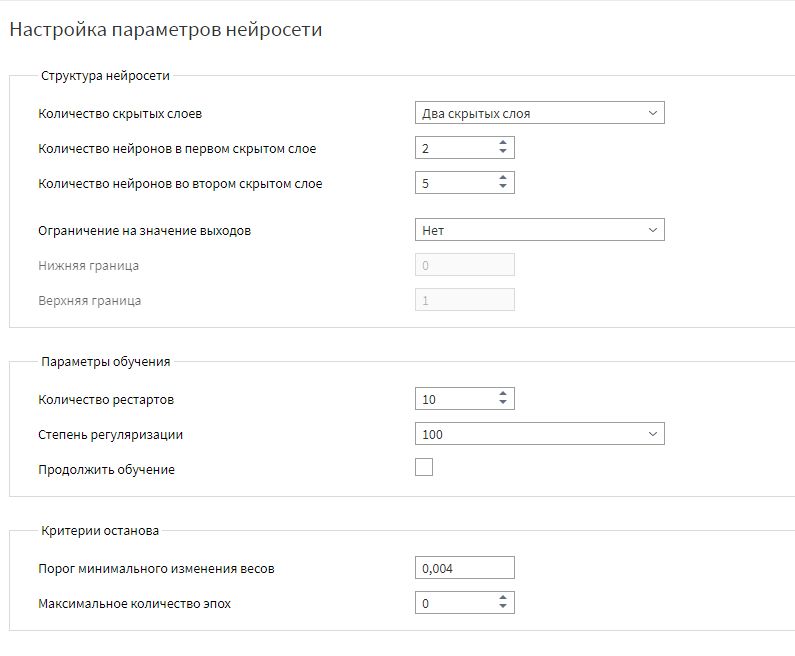
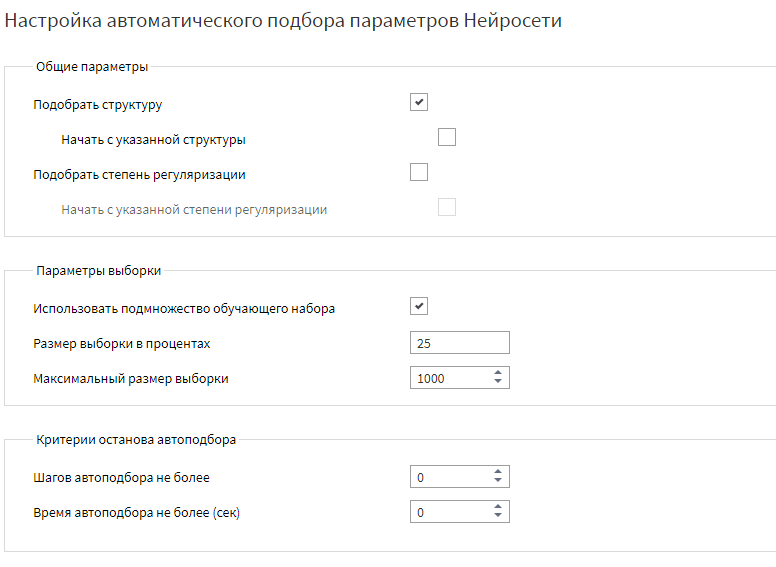
Результаты исследования подтверждают первоначальную гипотезу о значительном потенциале применения инструментов искусственного интеллекта для оптимизации процессов в сфере железнодорожных перевозок. В частности, предложенные модели позволят не только повысить эффективность управления ценами, но и адаптировать их в соответствии с реальным спросом и другими влияющими факторами. Это создает предпосылки для разработки более гибких и эффективных стратегий ценообразования, что в конечном итоге способствует повышению удовлетворенности клиентов и оптимизации доходности для ОАО «РЖД».

В заключение следует отметить, что достигнутые результаты исследования открывают новые возможности для дальнейших разработок в данной области. Важным аспектом будущих исследований является углубленный анализ взаимодействия различных факторов, влияющих на спрос и ценообразование, а также разработка более сложных и точных моделей искусственного интеллекта.

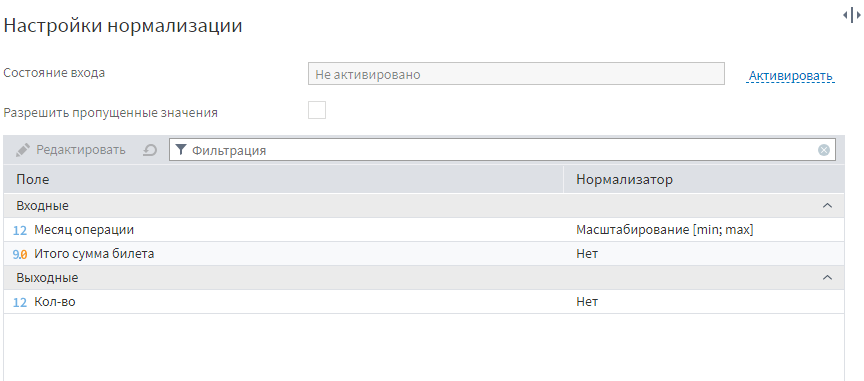
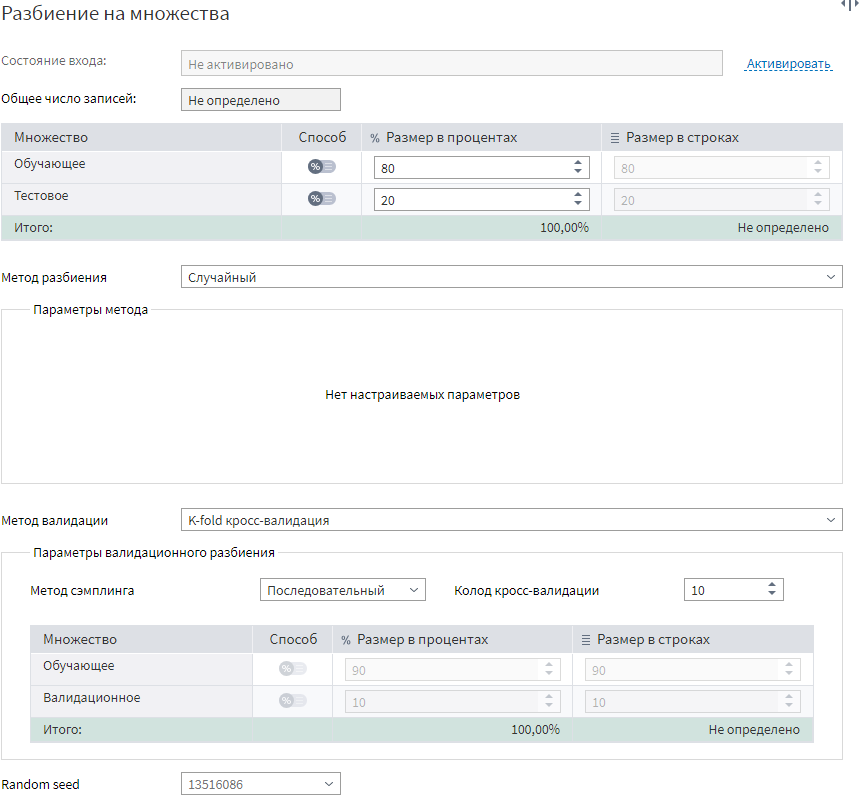
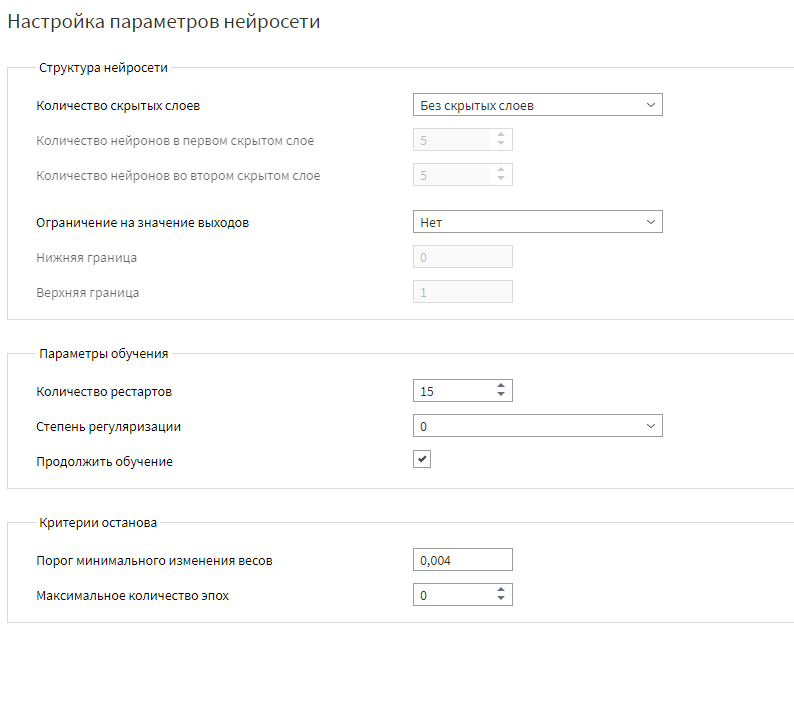
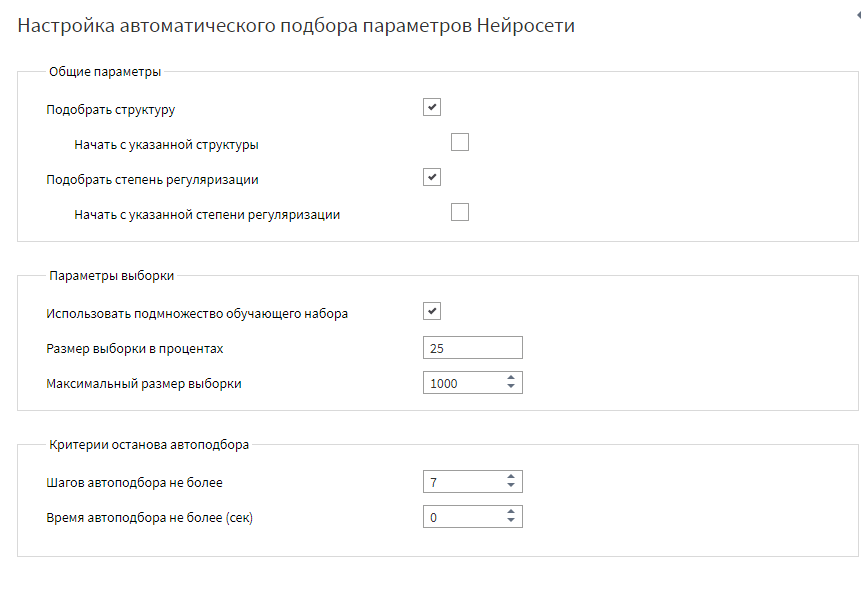
**Список использованной литературы.**

1. Описание проблемы кейса от ОАО РЖД в рамках хакатона. – URL: <https://friendly-possum-13e.notion.site/aa5c0691236241d49573e6eaad1c9f4f> (дата обращения 23.03.2024).
2. Официальный датасет от ОАО РЖД. – URL: https://drive.google.com/drive/folders/1CzVPgzko3DdmnCeBATbbOb3SLnGgT6vH (дата обращения 23.03.2024).
3. Метод Монте-Карло (Monte-Carlo technique). – URL: <https://wiki.loginom.ru/articles/monte-carlo-technique.html> (дата обращения 23.03.2024).
4. Кросс-валидация (Cross-validation). – URL: https://wiki.loginom.ru/articles/cross-validation.html (дата обращения 23.03.2024).
5. Нормализация непрерывных данных. – URL: <https://help.loginom.ru/userguide/processors/normalization/normalization-continuous.html> (дата обращения 23.03.2024).
6. Динамическое ценообразование. – URL: <https://www.rzd.ru/ru/9330> (дата обращения 23.03.2024).
7. Формирование тарифов. – URL: <https://ar2019.fpc.ru/ru/results/marketing/pricing-policy> (дата обращения 23.03.2024).
8. Дерево решений (Decision Trees). – URL: <https://wiki.loginom.ru/articles/decision-trees.html> (дата обращения 23.03.2024).
9. Деревья решений: общие принципы. – URL: <https://loginom.ru/blog/decision-tree-p1> (дата обращения 23.03.2024).
10. Основные методы ценообразования: что это, механизмы и способы. – URL: <https://www.cleverence.ru/articles/finansy/metod-tsenoobrazovaniya-osnovnye-sposoby-i-mekhanizmy-obrazovaniya-i-rascheta/> (дата обращения 23.03.2024).
11. Логистическая регрессия. – URL: <https://help.loginom.ru/userguide/processors/datamining/logistic-regression/index.html> (дата обращения 25.03.2024).
12. Коэффициент Джини. Из экономики в машинное обучение. – URL: <https://habr.com/ru/companies/ods/articles/350440/> (дата обращения 27.03.2024).
13. Анализ с помощью характеристической кривой (ROC-analisys). – URL: <https://wiki.loginom.ru/articles/roc-analysis.html> (дата обращения 27.03.2024).
14. Качество бинарной классификации. – URL: <https://help.loginom.ru/userguide/visualization/binary-classification/index.html> (дата обращения 27.03.2024).
15. Матрица ошибок. – URL: <https://help.loginom.ru/6.4/userguide/visualization/binary-classification/error-matrix.html> (дата обращения 27.03.2024).
16. Метрики качества моделей бинарной классификации. – URL: <https://loginom.ru/blog/classification-quality> (дата обращения 27.03.2024).
17. Метод локтя (Elbow method) – URL: https://wiki.loginom.ru/articles/elbow-method.html (дата обращения 04.06.2024).

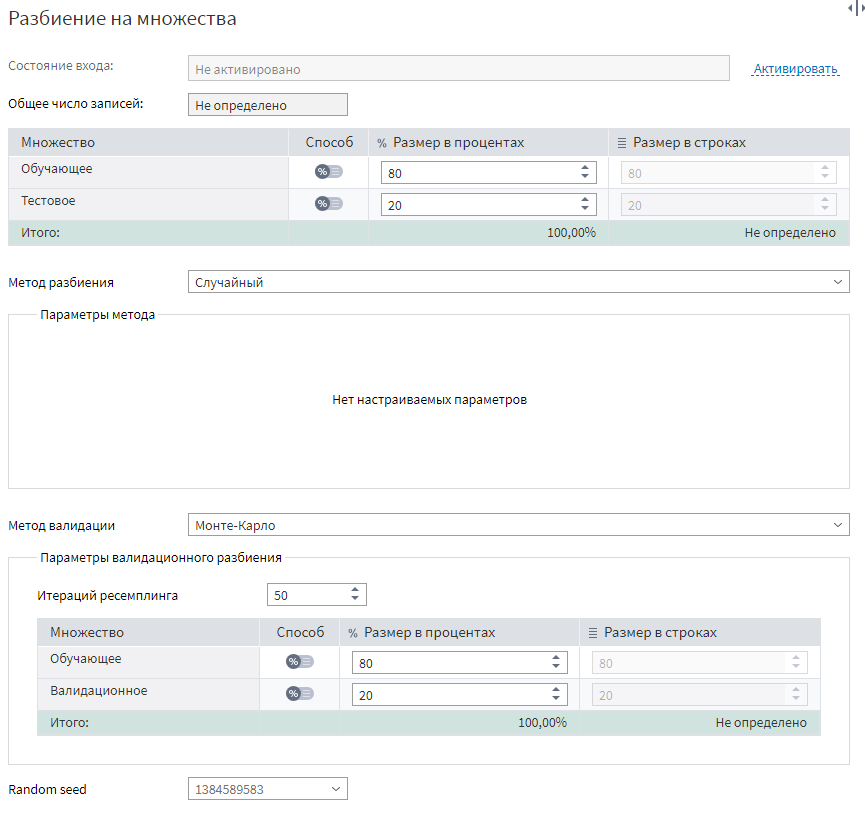
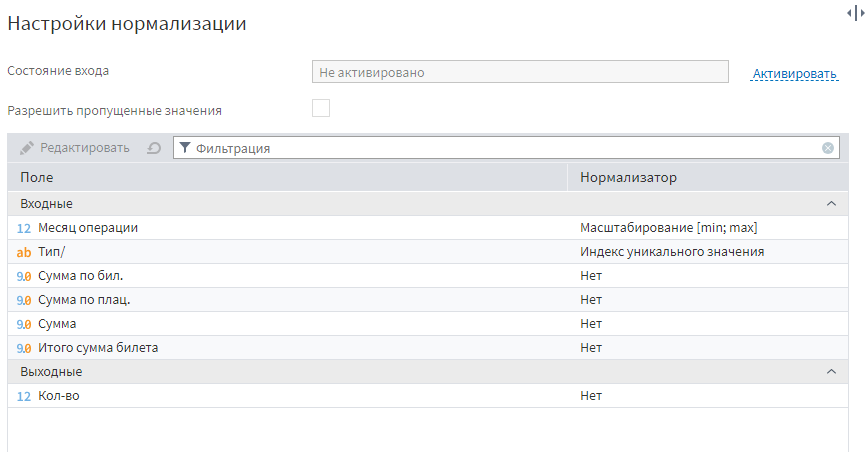
**Приложение А Настройки первой модели для предсказания спроса**

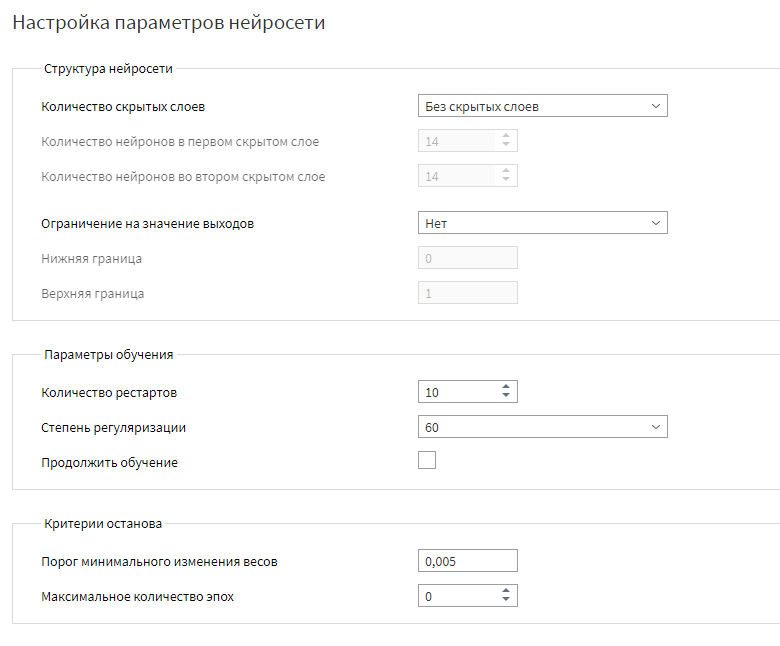
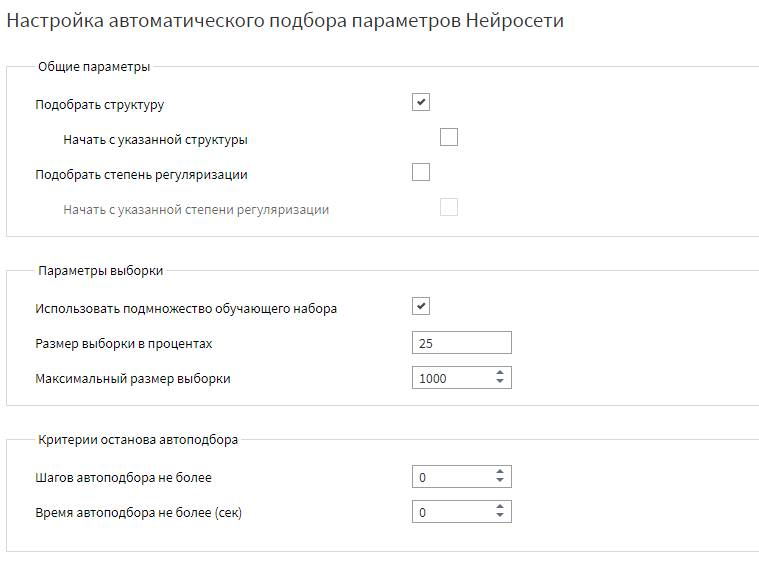
   

**Приложение Б Настройки второй модели для предсказания спроса**

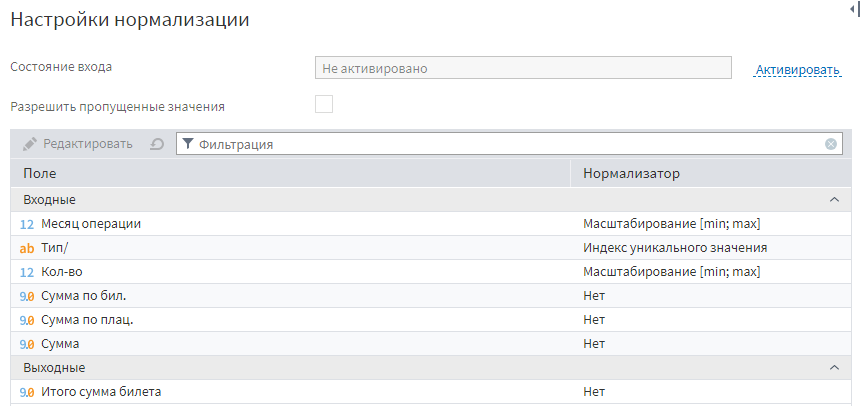
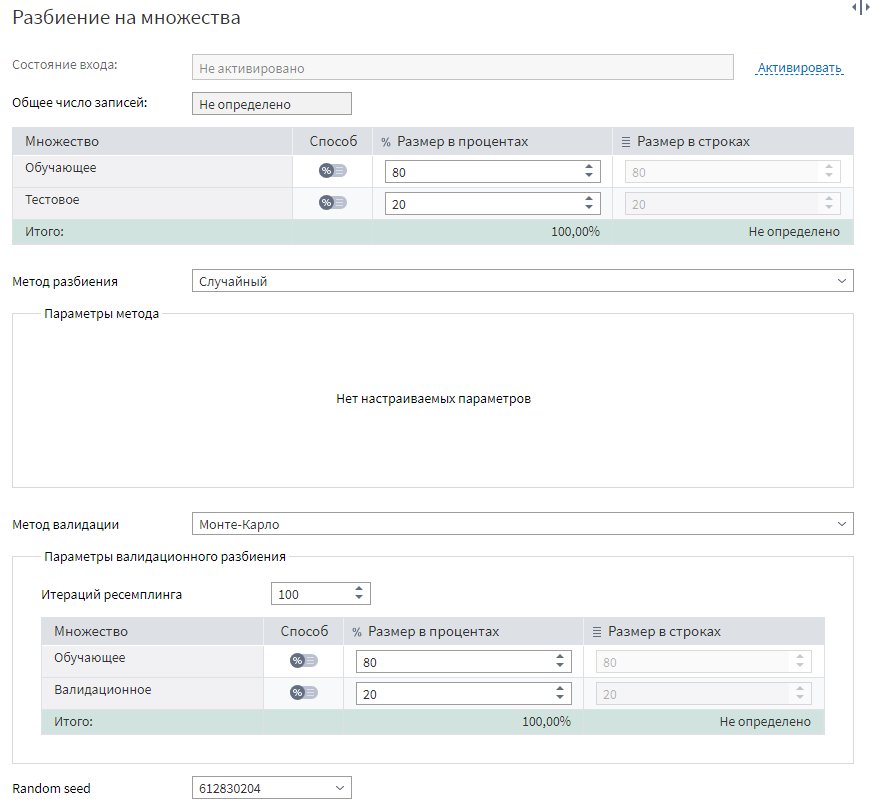
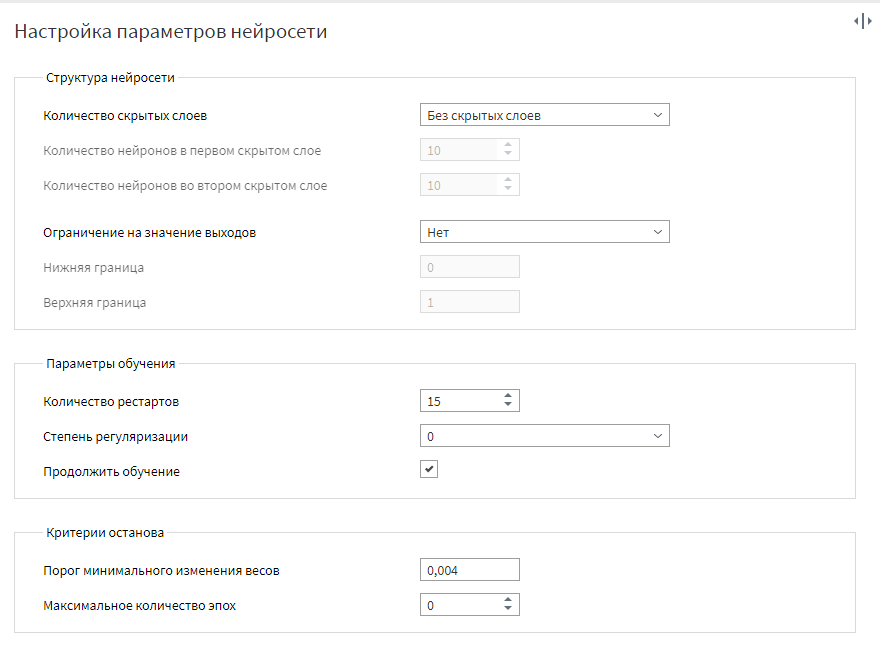
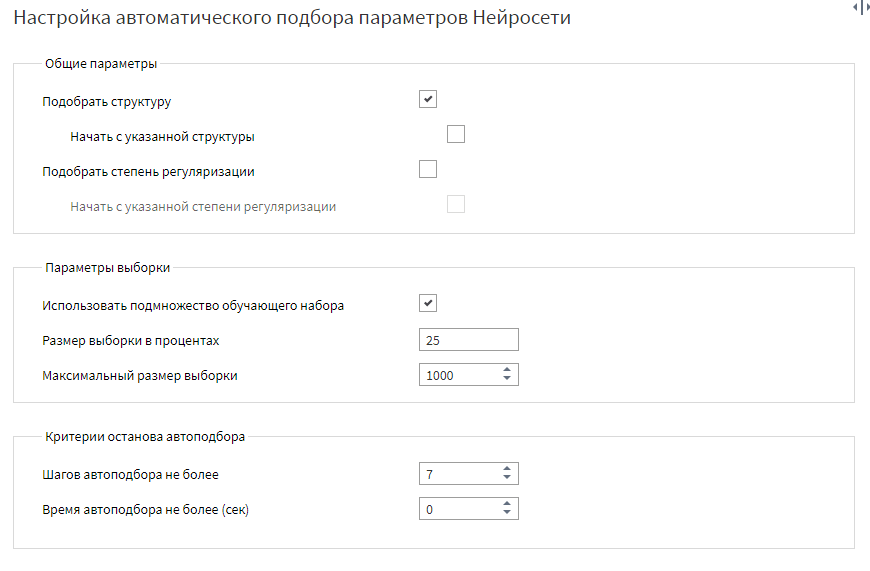
**** ****  

**Приложение В Настройки третьей модели для предсказания спроса**

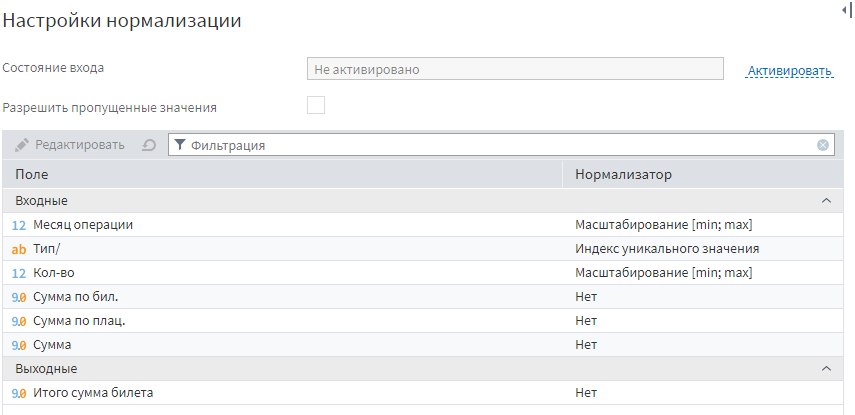
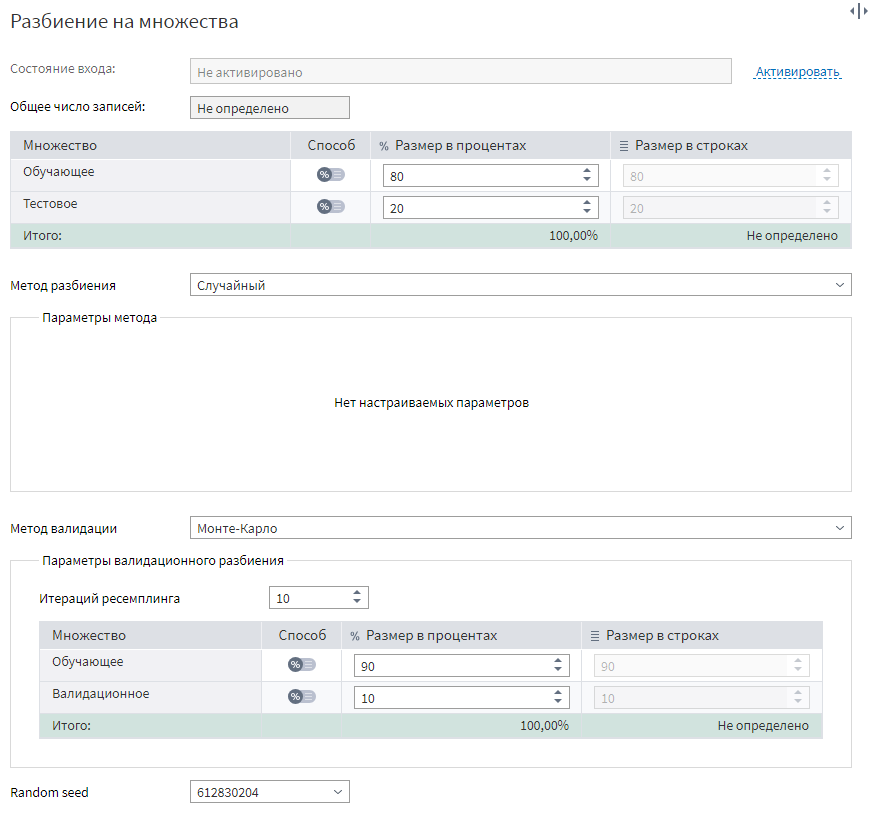
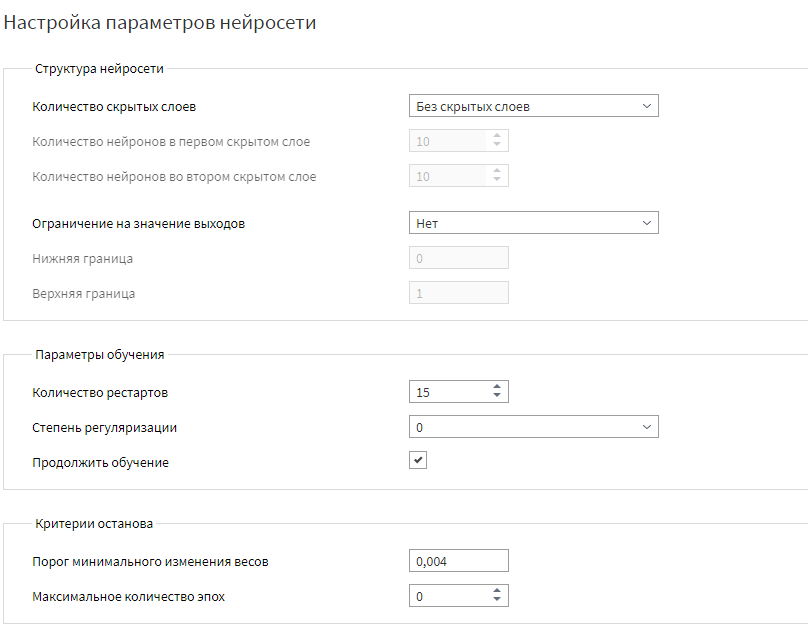
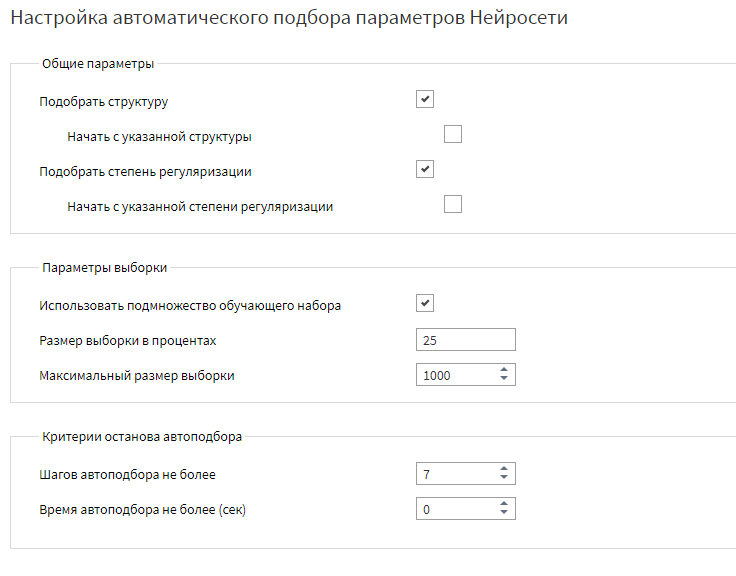
****

**** 

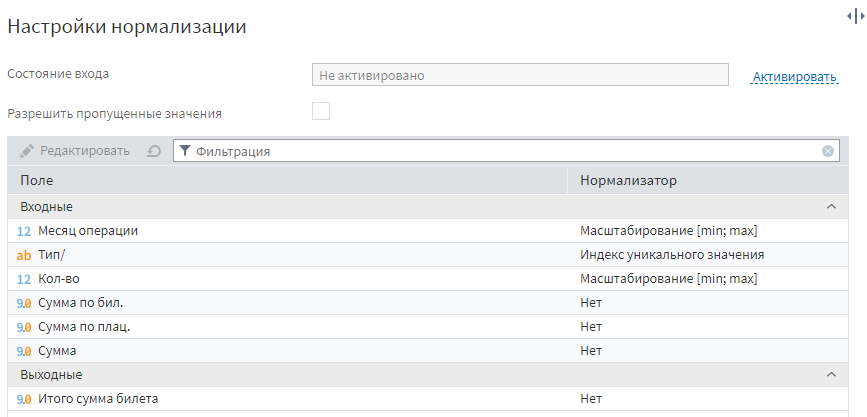
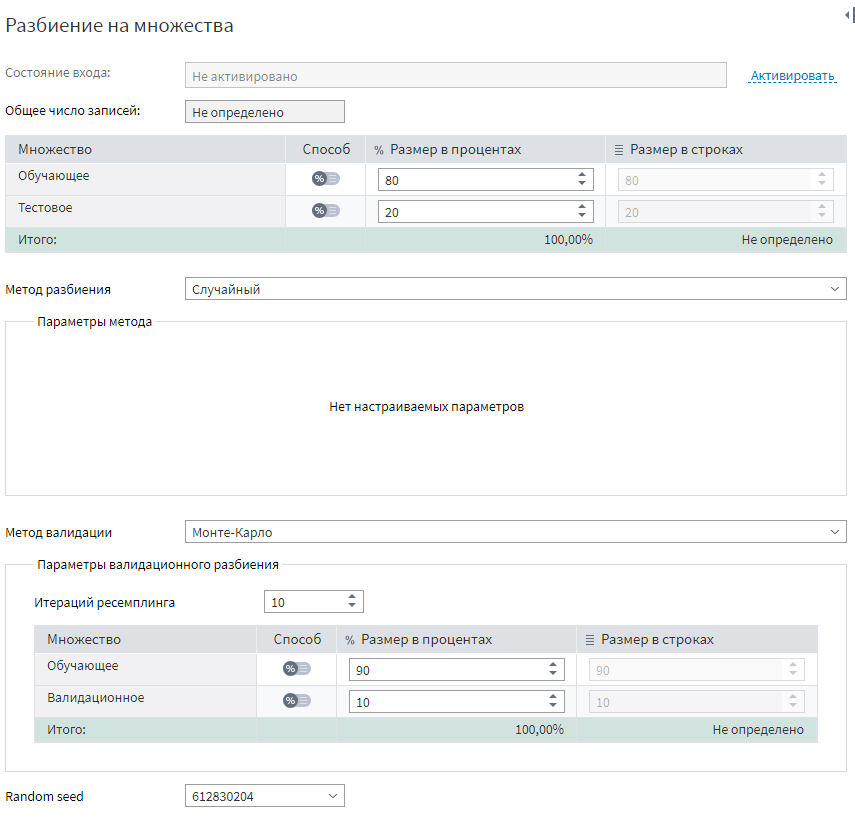
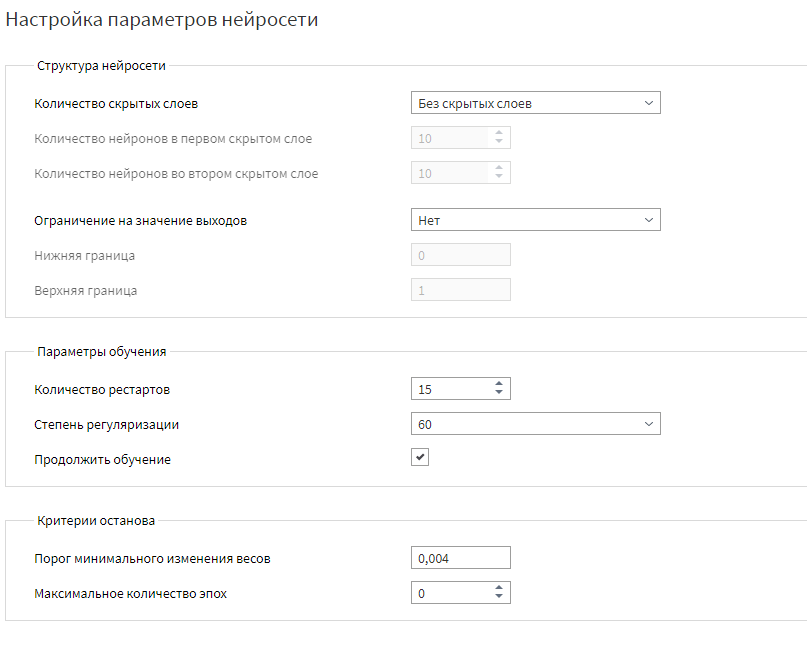
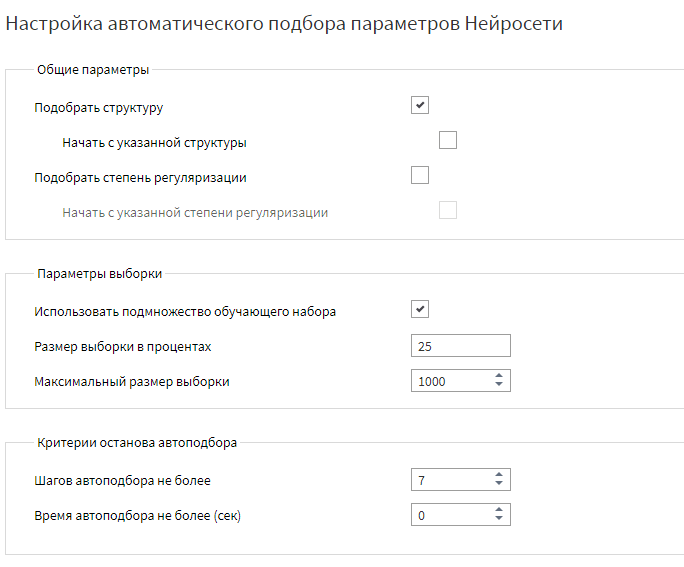
**Приложение Г Настройки первой модели для адаптации ценообразования**

**** ****  

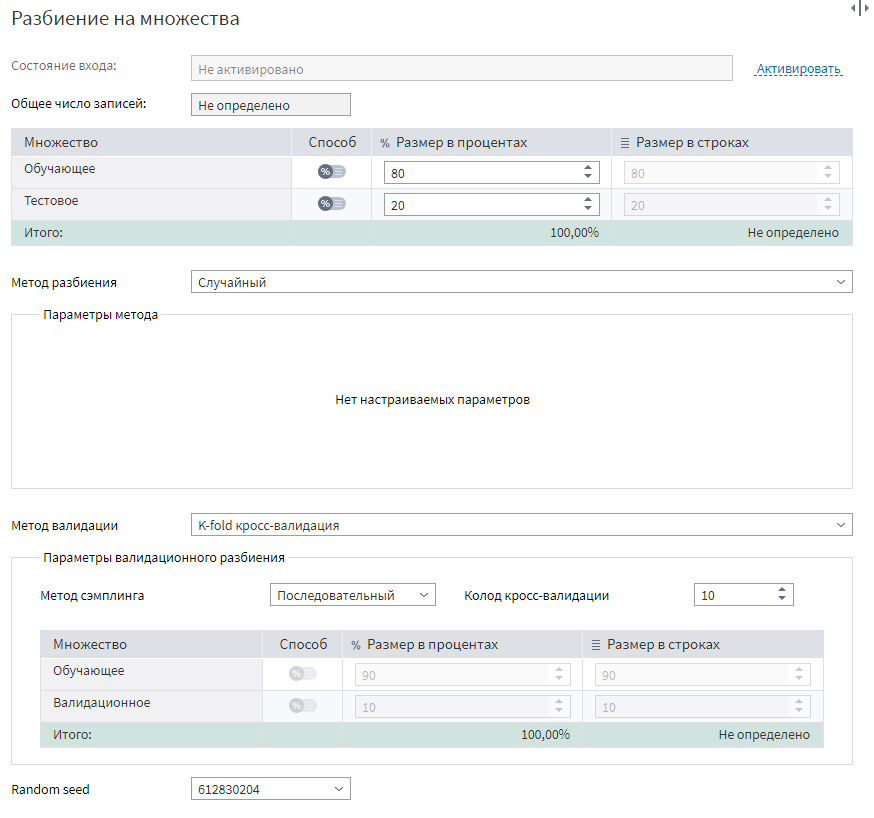
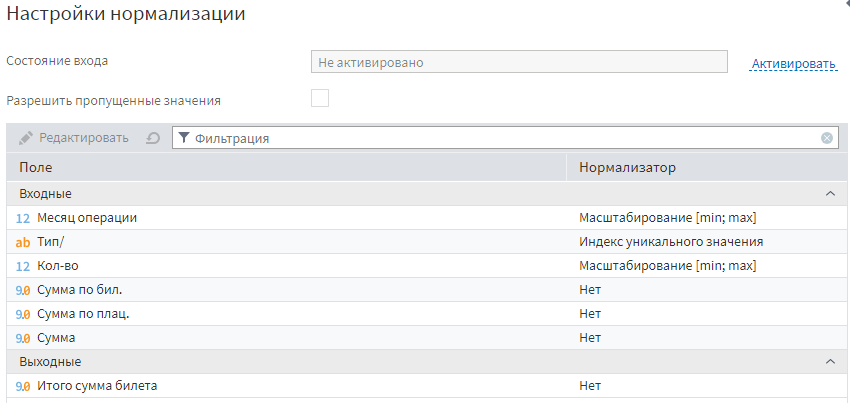
**Приложение Д Настройки второй модели для адаптации ценообразования**

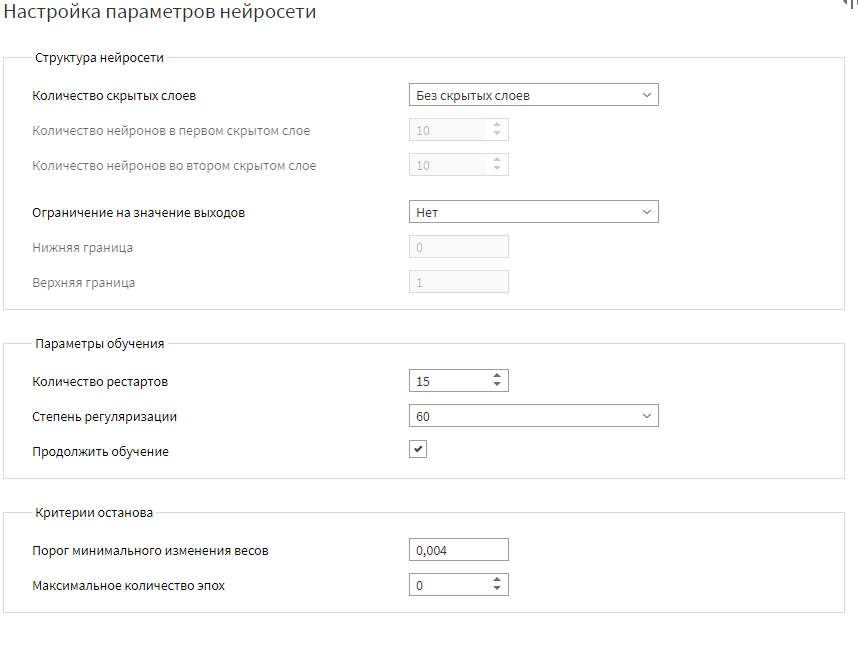
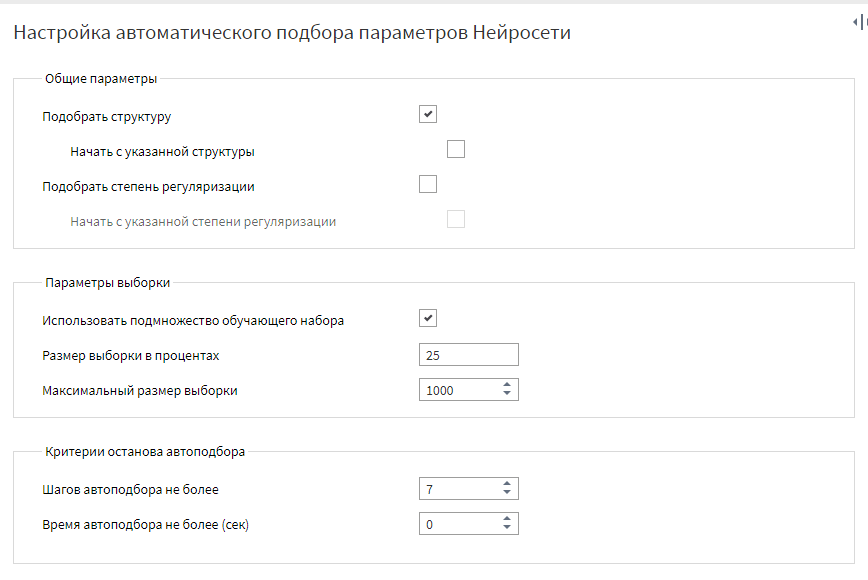
**** ****  

**Приложение Е Настройки третьей модели для адаптации ценообразования**

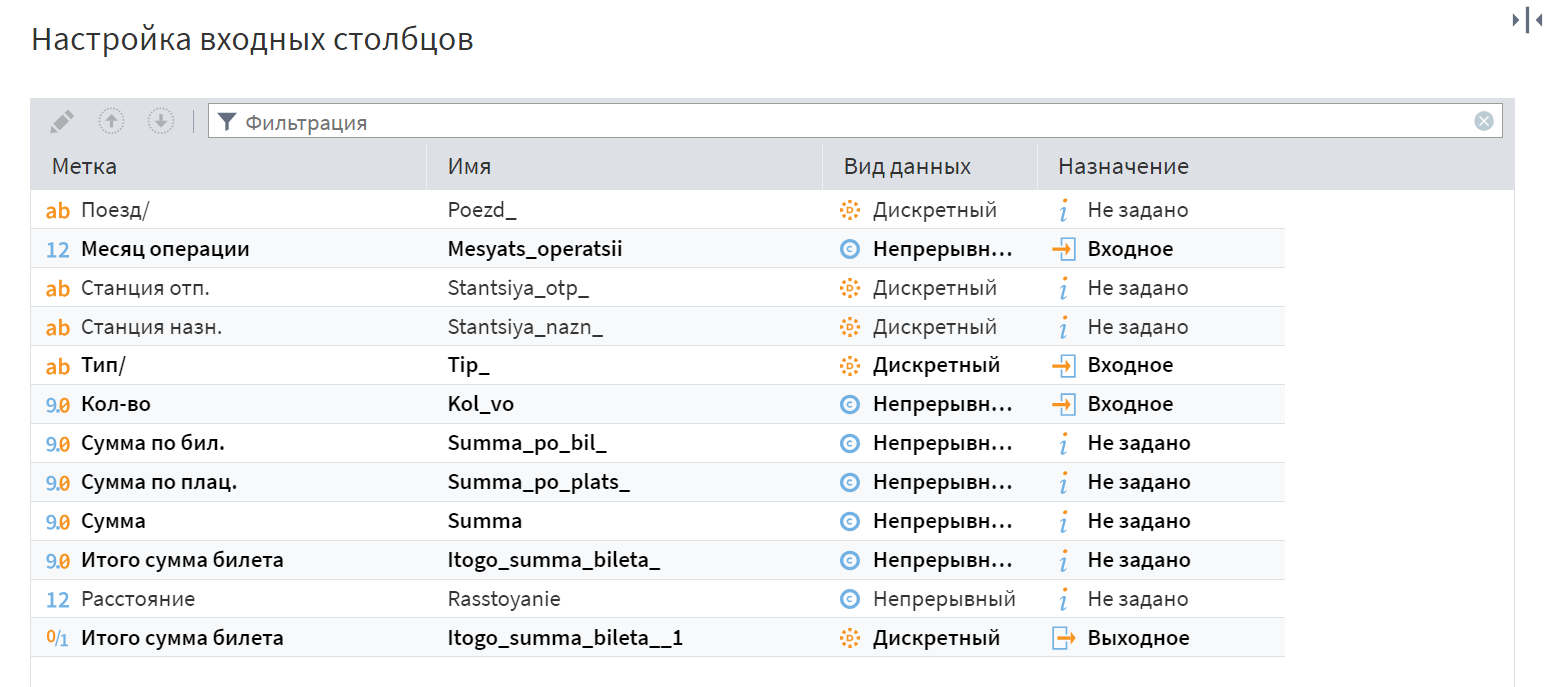
**** ****  

**Приложение Ж Настройки четвертой модели для адаптации ценообразования**

****

**** 

**Приложение З Настройки логической регрессии**

**** **Изображение выглядит как текст, Шрифт, число, линия

Автоматически созданное описание** Изображение выглядит как текст, снимок экрана, программное обеспечение, веб-страница

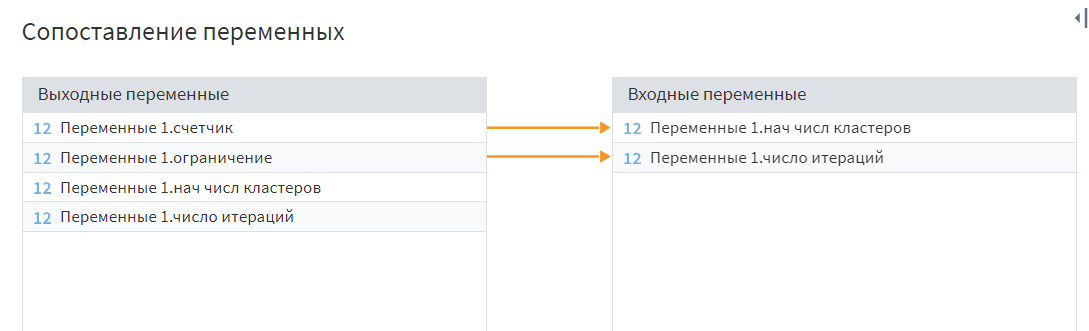
Автоматически созданное описание Изображение выглядит как текст, снимок экрана, число, Шрифт

Автоматически созданное описание Изображение выглядит как текст, снимок экрана, документ, Шрифт

Автоматически созданное описание

**Приложение И Настройки Цикла**

**Изображение выглядит как текст, снимок экрана, Шрифт, число

Автоматически созданное описание**