Федеральное государственное образовательное учреждение

высшего профессионального образования

Санкт-Петербургский государственный университет

Институт «Высшая школа менеджмента»

ПОВЫШЕНИЕ ТОЧНОСТИ ПРОГНОЗИРОВАНИЯ АВИАЦИОННЫХ МЕТРИК С ПОМОЩЬЮ МОДЕЛЕЙ МАШИННОГО ОБУЧЕНИЯ ДЛЯ СНИЖЕНИЯ ОПЕРАЦИОННЫХ ЗАТРАТ ООО "ВОЗДУШНЫЕ ВОРОТА СЕВЕРНОЙ СТОЛИЦЫ"

Выпускная квалификационная работа студентки 4 курса, направление 38.03.02 – Менеджмент, профиль – Информационный менеджмент

Есауловой Натальи Андреевны

Научный руководитель:

Кандидат технических наук,

старший преподаватель кафедры информационных технологий в менеджменте

Лещева Ирина Анатольевна

Санкт-Петербург

2023

ЗАЯВЛЕНИЕ О САМОСТОЯТЕЛЬНОМ ВЫПОЛНЕНИИ ВЫПУСКНОЙ КВАЛИФИКАЦИОННОЙ РАБОТЫ

Я, Есаулова Наталья Андреевна, студентка 4 курса направления 080500 – Менеджмент (профиль подготовки – Информационный менеджмент), заявляю, что в моей выпускной квалификационной работе на тему «Повышение точности прогнозирования авиационных метрик с помощью моделей машинного обучения для снижения операционных затрат ООО "Воздушные Ворота Северной Столицы"», представленной в службу обеспечения программ бакалавриата для последующей передачи в государственную аттестационную комиссию для публичной защиты, не содержится элементов плагиата. Все прямые заимствования из печатных и электронных источников, а также из защищенных ранее курсовых и выпускных квалификационных работ, кандидатских и докторских диссертаций имеют соответствующие ссылки.

Мне известно содержание п. 9.7.1 Правил обучения по основным образовательным программам высшего и среднего профессионального образования в СПбГУ о том, что «ВКР выполняется индивидуально каждым студентом под руководством назначенного ему научного руководителя», и п. 51 Устава федерального государственного бюджетного образовательного учреждения высшего профессионального образования «Санкт- Петербургский государственный университет» о том, что «студент подлежит отчислению из Санкт-Петербургского университета за представление курсовой или выпускной квалификационной работы, выполненной другим лицом (лицами)».

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ (Подпись студента)



\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_24.05.2023\_\_\_\_\_\_\_\_\_\_\_\_\_ (Дата)

**ОГЛАВЛЕНИЕ**

[Введение 4](#_Toc135818489)

[Глава 1. Описание компании и действующего способа прогнозирования 7](#_Toc135818490)

[1.1 История аэропорта 7](#_Toc135818491)

[1.2 Общие сведения о компании 8](#_Toc135818492)

[1.3 Теоретические основания проогнозирования в авиационной индустрии 10](#_Toc135818493)

[1.4 Существующая модель прогнозирования 12](#_Toc135818494)

[1.5 Постановка задачи по методологии GRISP-DM 15](#_Toc135818495)

[Вывод по главе 1 17](#_Toc135818496)

[Глава 2. Обработка и анализ данных 20](#_Toc135818497)

[2.1 Предобработка данных 20](#_Toc135818498)

[2.2 Анализ данных 24](#_Toc135818499)

[Вывод по главе 2 29](#_Toc135818500)

[Глава 3. Обзор методов машинного обучения, выбор методов для построения моделей 31](#_Toc135818501)

[3.1 Определение машинного обучения 31](#_Toc135818502)

[3.2 Различия статистических методов и методов машинного обучения при анализе временных рядов 33](#_Toc135818503)

[3.3 Выбор методов для построения моделей 36](#_Toc135818504)

[3.4 Описание реализуемых методов 40](#_Toc135818505)

[Вывод по главе 3 45](#_Toc135818506)

[Глава 4. Модели прогнозирования авиационных метрик 46](#_Toc135818507)

[4.1 Обучение моделей 46](#_Toc135818508)

[4.2 Оценка точности, выбор наилучших моделей 56](#_Toc135818509)

[4.3 Рекомендации по внедрению, оценка влияния на операционные затраты 60](#_Toc135818510)

[Вывод по главе 4 63](#_Toc135818511)

[Заключение 64](#_Toc135818512)

[Список литературы 66](#_Toc135818513)

[Приложения 71](#_Toc135818514)

[Приложение 1. Графики прогнозов моделей LSTM по всем сегментам и метрикам 71](#_Toc135818515)

[Приложение 2. Графики прогнозов моделей XGBoost по всем сегментам и метрикам 74](#_Toc135818516)

# Введение

Данная работа выполнена в формате прикладного проекта для ООО «Воздушные Ворота Северной Столицы» – управляющей компании аэропорта Пулково, расположенного в Санкт-Петербурге. Пулково является одним из крупнейших и наиболее современных аэропортов в России, обслуживающим более 20 миллионов пассажиров в год. Он осуществляет как международные, так и внутренние авиарейсы, обладая широкой сетью направлений. Приоритетами работы аэропорта являются высокое качество обслуживания и безопасность пассажиров, а также оперативность и эффективность в работе. Основанный в 1932 году, аэропорт Пулково имеет более 90-летний опыт работы в сфере авиационного транспорта и туризма, привлекая клиентов из разных регионов России и всего мира.

Прогнозирование является одним из важнейших аспектов авиационной отрасли, особенно для аэропортов. Точные прогнозы позволяют администрации аэропортов принимать обоснованные решения, которые повышают эффективность работы, снижают затраты и улучшают общее качество обслуживания клиентов. В авиационной индустрии широко применяются статистические и экспертные модели для прогнозирования метрик, таких как количество пассажиров, грузовые объемы и другие операционные показатели. Однако, воздействие на деятельность аэропортов экономических кризисов, таких как пандемия COVID-19, подчеркивает необходимость использования более адаптивных методов прогнозирования.

Машинное обучение, благодаря доказанной эффективности в прогнозировании временных рядов, может улучшить точность прогнозирования в авиационной отрасли. Применение моделей машинного обучения может оказать помощь аэропорту в ведении бизнеса и сократить операционные расходы, включая расходы на трафикозависимый персонал. Внедрение машинного обучения является важным шагом для авиаиндустрии в улучшении прогнозирования метрик и повышении качества обслуживания пассажиров.

Высокая точность прогнозирования пассажиропотока и прочих авиационных метрик позволит повысить операционную эффективность аэропорта Пулково, снизить операционные издержки и повысить удовлетворенность пассажиров услугами, предоставляемыми аэропортом.

**Цели и задачи проекта**

Данная работа посвящена совершенствованию процесса прогнозирования авиационных метрик, повышению точности прогнозирования с помощью моделей машинного обучения и изучению возможности уменьшения затрат при эксплуатации аэропорта на основе улучшенных показателей аэропорта Пулково. На данный момент прогнозирование производится с помощью экспертной модели, позволяющей оценить метрики только в масштабе месяца. Данная модель основана на работе с историческими данными, но не дает достаточно информации для гибкого управления операционными затратами.

Для аэропортов повышение точности прогнозирования авиационных метрик является одним из ключевых факторов в улучшении бизнес-процессов и снижении операционных затрат. Создание новых моделей прогнозирования затрат может существенно повысить эффективность работы аэропортов. Благодаря улучшенной точности данных, связанных с авиарейсами, ООО «Воздушные Ворота Северной Столицы» смогут детально рассчитать сроки и общие затраты на проведение технического обслуживания, а также на закупку топлива, оборудования и других необходимых ресурсов для нормального функционирования.

Целью работы является повышение точности прогнозирования авиационных метрик и разработка рекомендаций на основе полученных улучшений точности прогнозирования, направленных на оптимизацию работы с операционными затратами. Для достижения поставленных целей предполагается решить следующие задачи:

1. Проанализировать деятельность ООО «Воздушные Ворота Северной Столицы», сформировать представление о действующем процессе прогнозирования авиационных метрик
2. Выявить недостатки и области развития в процессе прогнозирования;
3. Выявить пути совершенствования процесса прогнозирования с помощью методов машинного обучения
4. Разработать модели прогнозирования и определить наилучшую по показателям эффективности
5. Оценить возможности сокращения операционных затрат на результатах наилучшей модели

**Предполагаемый результат работы:**

* Описан действующий процесс прогнозирования авиационных метрик аэропортом Пулково;
* Выявлены области развития действующего процесса;
* Проведен выбор методов машинного обучения для прогнозирования временных рядов в данной задаче;
* Созданы модели краткосрочного прогнозирования;
* По результатам сравнения моделей выбрана модель с наилучшей точностью, на основе показателей эффективности модели оценены возможности для сокращения операционных затрат.

**Объект работы –** ООО «Воздушные Ворота Северной Столицы», **предмет** – процесс прогнозирования авиационных метрик.

В качестве инструментов используютcя среда Jupyter, язык программирования Python с набором необходимых для выполнения данной работы библиотек (pandas, numpy, matplotlib, seaborn, xgboost, tensorflow, math, statsmodels, copy, missingno, datetime, sklearn), методология CRISP-DM, методы анализа данных, методы построения моделей машинного обучения и их оценки.

В качестве основных источников информации использованы документы организации из открытых источников, интервью с представителями компании, научные труды в области прогнозирования временных рядов, научные труды в области прогнозирования в авиационной индустрии, индивидуальное наблюдение.

Текст данной работы содержит 4 главы. В первой предоставляется информация о компании и существующей модели прогнозирования, определяются ее недостатки, вводится методология CRISP-DM и ставится задача в рамках данной методологии. Во второй главе данные авиационных метрик обрабатываются и производится их анализ. В третьей главе дается определение машинному обучению и обоснование тому, что методы машинного обучения могут показать высокую точность на имеющихся данных, также проводится выбор методов машинного обучения. В четвертой главе на основе выбранных методов создаются модели, оценивается их точность, даются рекомендации по внедрению и производится оценка потенциального сокращения операционных затрат при использовании созданных моделей.

# Глава 1. Описание компании и действующего способа прогнозирования

## 1.1 История аэропорта

Аэропорт Пулково, расположенный в южной части Санкт-Петербурга, является одним из крупнейших аэропортов страны и важными воздушными воротами для внутренних и международных путешественников. Начало функционирования аэропорта датируется 24 июня 1932 года: именно тогда в эксплуатацию был сдан аэродром «Шоссейная», название которому было дано из-за ближайшей железнодорожной станции. В 30-е года происходило развитие региональных перевозок и из аэропорта стали выполняться полеты в Петрозаводск, Архангельск, Киев и другие города. Однако только в 1936 было заложено здание аэровокзала, произведено обновление авиапарка и выход новой техники на линии.

Во время Великой Отечественной Войны аэродром «Шоссейная» стал линией фронта и был закрыт, но после окончания войны он был быстро восстановлен как гражданский аэропорт. Мирные перевозки стали совершаться летчиками уже 27 января 1944, когда была снята 900-дневная блокада Ленинграда. В 1945 году авиаторы северного управления перевезли 12 000 пассажиров и сотни тонн грузов, а полностью возобновлена работа Ленинградского аэропорта была под конец 1940-х.

В 1950-х годах Пулково стал одним из крупнейших аэропортов Советского Союза и узловым аэропортом Аэрофлота, советской государственной авиакомпании. Строительство здания аэровокзала, начавшееся в 1936 году по проекту архитектора Генриха Майзеля, закончилось, и здание в 1951 году было открыто для пользования. В эксплуатацию поступила новая авиатехника, и с середины 50-х годов начался ввод реактивных самолетов в эксплуатацию. 15 марта 1959 был совершен первый пассажирский рейс на реактивном самолете Ту-104б с бортовым номером 42419.

В период с 1960 по 1970 года Ленинградский аэропорт продолжал свое активное развитие, в 1965 году начал использоваться новый вид пассажирского транспорта – скоростные самолеты большой вместимости, что позволило ленинградцам легче добираться до Петропавловска-Камчатского и Владивостока.

Благодаря своему развитию и производительности, в 1971 году аэропорт был награжден орденом Октябрьской Революции. В 1973 году был открыт новый терминал, который разработал Александр Жук, архитектор-авангардист, он содержал в себе множество инновационных решений, направленных на улучшение обслуживания пассажиров. В 1974 году Ленинградский аэропорт сменил название на «Пулково».

В 1980 году был открыт второй аэровокзальный комплекс «Пулково-2», который на тот момент обслуживал все международные направления. В 1986 году к аэровокзалу «Пулково-2» были пристроены два павильона для увеличения пассажиропотока.

В середине первого десятилетия ХХI века было принято решение о масштабной реконструкции аэропорта. В 2007 году аэропорт перешел в собственность Санкт-Петербурга и был объявлен международный конкурс на создание нового пассажирского терминала, который выиграло британское архитектурное бюро Grimshaw, а в 2009 году бюро вошло в консорциум «Воздушные ворота Северной Столицы», который стал оператором аэропорта Пулково до 2039 года.

В 2011 году началась реконструкция аэропорта и 4 декабря 2013 года новый терминал для пассажиров был открыт в тестовом порядке. В 2015 году здание «Пулково-1» было ещё раз реконструировано и соединено централизованным пассажирским терминалом. По завершении программы модернизации, аэропорт Пулково стал одним из наиболее современных и эффективных аэропортов в России.

В 2016, 2019 и 2020 годах международный совет аэропортов (ACI) признал Пулково лучшим аэропортом Европы по качеству обслуживания пассажиров в категории «Аэропорты с пассажиропотоком от 5 млн до 15 млн. чел. в год».

По итогам 2019 года аэропорт Санкт-Петербурга в пятый раз подряд стал победителем престижной отраслевой премии «Воздушные Ворота России». В течение 2019 года аэропорт Пулково показал стабильный рост пассажиропотока и по итогам года обслужил 19,6 млн человек, тем самым поставив новый исторический рекорд по пассажиропотоку. По итогам 2020 года инновационные проекты аэропорта стали лучшими в отрасли в рамках национальной премии «Воздушные ворота России».[[1]](#footnote-1)

Пулково — это не только транспортный узел, но и памятник богатой истории российской авиационной промышленности. Его путь от небольшого аэродрома до современного аэропорта, отмеченного наградами, является свидетельством упорного труда и преданности людей, которые внесли свой вклад в его успех.

## 1.2 Общие сведения о компании

Сегодня международный аэропорт Пулково обслуживает более 17 миллионов пассажиров в год и выполняет рейсы по более чем 100 направлениям в России и по всему миру. Также он является единственным на территории Санкт-Петербурга, обслуживающим официальные рейсы.

С 29 апреля 2010 года ООО «Воздушные Ворота Северной Столицы» является управляющей компанией аэропорта Пулково. Она осуществляет свою деятельность на основе проекта государственно-частного партнерства без использования бюджетных средств, что является первым в России проектом такого рода в авиационной отрасли.

За период с 2010 по 2015 год суммарный объем инвестиций в развитие аэропорта составил 1,2 млрд евро. Кроме управления операционной деятельностью Пулково, компания также занимается его реконструкцией и модернизацией. Владельцами ООО «Воздушные Ворота Северной Столицы» являются компании «ВТБ Капитал», «Фрапорт АГ», Суверенный фонд Катара и международный инвесторский консорциум.[[2]](#footnote-2)

Аэропорт по постановлению правительства раскрывает достаточно большое количество информации, релевантной для данной работы. Во-первых, это структура издержек.

| 1. Структура издержек Пулково за 2021 год | |
| --- | --- |
| **Наименование показателя** | **Уд. Вес, %** |
| Расходы, всего | 100.0% |
| Оплата труда всего состава | 31.5% |
| Отчисление на социальные нужды | 7.9% |
| Амортизация ОПФ | 24.5% |
| Содержание и эксплуатация зданий, сооружений и оборудования | 8.3% |
| Затраты на ремонт ОПФ | 2.6% |
| Аренда и услуги сторонних предприятий и организаций | 14.7% |
| Прочие производственные расходы | 5.9% |
| Общехозяйственные расходы | 4.6% |

Источник: Веб-сайт аэропорта Пулково [Электронный ресурс] — Режим доступа: <https://pulkovoairport.ru/docs/files/reports/cost_structure_12_21.pdf> (дата обращения: 07.03.2023).

По таблице 1 видно, что наибольшую долю в составе издержек занимают расходы на персонал, и, по словам сотрудников аналитического отдела среди всего персонала имеется весомая доля трафикозависимого персонала, издержки на который зависят от точности прогноза пассажиропотока и прочих метрик.

Также важной информацией являются источники дохода – согласно отчету Международного совета аэропортов, который отслеживает экономическую деятельность 919 воздушных портов по всему миру, в 2020 году общий доход составил $191,3 трлн. При этом большинство доходов (56%) приходится на авиационную деятельность, а 39,4% - на неавиационную (остальные 4,6% - на другие категории). Средний доход на одного пассажира был около $10 в авиационной сфере и чуть больше $7 в неавиационной. Неавиационный доход, полученный от торговли, является наиболее прибыльным пунктом (почти 29%), за которым следуют доходы от парковки (20,5%) и сдачи недвижимости в аренду (15%). Остальные категории составляют более 35% от общей суммы доходов.[[3]](#footnote-3)

## 1.3 Теоретические основания проогнозирования в авиационной индустрии

Прогнозирование – важнейший аспект авиационной отрасли, особенно для аэропортов. Точные прогнозы позволяют администрации аэропортов принимать обоснованные решения, которые повышают эффективность работы, снижают затраты и улучшают общее качество обслуживания клиентов. Далее перечислены причины, по которым прогнозирование важно для аэропортов.

Одной из основных причин важности прогнозирования в авиационной отрасли является очень динамичный и сложный характер бизнеса. Авиакомпании и аэропорты должны постоянно адаптироваться к изменениям спроса, цен на топливо, погодных условий и многих других факторов. Точное прогнозирование помогает снизить риски, связанные с этими неопределенностями, и обеспечивает наиболее эффективное распределение ресурсов. Например, если аэропорт прогнозирует резкое увеличение пассажиропотока, он может обеспечить достаточное количество персонала и помещений для обслуживания возросшего числа пассажиров.[[4]](#footnote-4)

Еще одна причина, по которой прогнозирование так важно в авиационной отрасли, — это необходимость эффективного управления пропускной способностью. Авиакомпании и аэропорты должны сбалансировать спрос и предложение на рейсы и услуги, чтобы максимально эффективно использовать ресурсы. Прогнозирование позволяет администрации аэропорта предсказать спрос на рейсы, терминалы и взлетно-посадочные полосы, что позволяет им вести соответствующее планирование. Это помогает максимизировать получаемую прибыль за счет сотрудничества с различными авиакомпаниями и планирования рейсов на различные маршруты, а также предотвратить переполненность.[[5]](#footnote-5)

Прогнозирование также играет важную роль в финансовой деятельности аэропортов. Аэропорты вносят большой вклад в местную экономику, и их успех тесно связан с их способностью генерировать доходы. Благодаря точному прогнозированию руководство аэропорта может оценить доходы от пассажирских перевозок, розничных продаж и других источников. Эта информация важна для составления бюджета и планирования и помогает аэропортам обеспечить достаточные ресурсы для инвестиций в инфраструктуру и услуги.

Помимо этих эксплуатационных преимуществ, прогнозирование также важно для улучшения обслуживания клиентов. Точные прогнозы помогают аэропортам минимизировать задержки, обеспечить доступ пассажиров к необходимым им услугам, а также создать безопасную и надежную среду. Например, если аэропорт прогнозирует внезапное увеличение пассажиропотока, он может принять меры для обеспечения эффективного управления очередями, достаточного количества персонала для оказания помощи пассажирам и адекватных мер безопасности.[[6]](#footnote-6)

Наконец, стоит отметить, что прогнозирование также важно для снижения воздействия авиационной отрасли на окружающую среду. Аэропорты являются крупными потребителями энергии и выбрасывают значительное количество парниковых газов. Точное прогнозирование позволяет администрациям аэропортов сократить потребление энергии и минимизировать выбросы, обеспечивая наиболее эффективное использование ресурсов. Например, если аэропорт прогнозирует резкое увеличение пассажиропотока, он может принять меры по снижению энергопотребления путем оптимизации использования систем освещения и отопления/охлаждения.

Инвестируя в инструменты и методы прогнозирования, аэропорты могут гарантировать, что они хорошо подготовлены для решения возникающих проблем и использования появляющихся возможностей.

## 1.4 Существующая модель прогнозирования

Показатели, прогнозируемые аналитическим отделом и используемые в данной работе включают в себя:

* PAX – Passengers Approximately – Количество пассажиров
* ATM – Air Transport Movement – Взлетно-посадочная операция
* MTOW – Maximum Take-Off Weight – Максимальная взлетная масса
* LF – Load Factor – Коэффициент загрузки рейса (Пассажиры/компоновка)

Также прогноз производится не только в разрезе показателей, но и в разрезе следующих направлений:

* Domestic-Moscow (DOM-MOW) – рейсы на московском направлении
* Domestic-Regions (DOM-REG) – внутренние рейсы в регионы за исключением Москвы
* Domestic-International (DOM-INT) – рейсы российских авиакомпаний за границу
* International-Belavia (INT-BEL) – рейсы авиакомпании Белавиа
* International-Low Cost Carrier (INT-LCC) – рейсы, выполняемы иностранными перевозчиками по 7-й свободе воздуха
* International-Full Service Carrier (INT-OTH) – рейсы, выполняемые иностранными перевозчиками (кроме рейсов 7-й свободы)

Текущая модель, которую Пулково использует для прогноза связок показатель-направление, основывается на расписании полетов на будущие месяца и исторических данных и работает с помощью коэффициентов, полученных по экспертной оценке. Расписание же составляется по результатам бронирования авиакомпаниями слотов на полеты, но далеко не всегда оказывается верным, часто происходят изменения и многие слоты отменяются. Стоит отметить, что в предоставленной модели всего 5 сегментов, так как авиакомпании из сегмента INT-LCC прекратили сотрудничество с аэропортом Пулково в феврале 2022 года.

Существующая прогнозная модель работает в Excel-файлах, и предоставлена не была. Была предоставлена модель, использующая логику и алгоритм оригинальной модели, перенесенная на Python в формате «. ipynb» (тетради Jupyter Notebook), в ней реализуются следующие шаги.

Данные о расписании будущих полетов и информации об аэропортах находятся в эксель-файлах Schedule (расписание), Airports (коды аэропортов, города и названия аэропортов, страны, код региона), Airlines (код авиакомпании, название, код региона), LF (историческая информация по проценту загрузки воздушных суден) и MTOW (файл с типом воздушного судна и его максимальной взлетной массой.

После импорта необходимых библиотек и загрузки вышеперечисленных файлов из файла с расписанием полетов, оставляют наблюдения только по чартерным и регулярным рейсам, так как эти категории рейсов соответствуют пассажирским перевозкам. Далее происходит объединение отфильтрованного по типу перевозок файла с информацией о рейсах и файлов с расшифровками различных кодов и типов ВС.

Далее добавляется колонка «departure» с индикатором отправления рейса и показатель максимальной взлетной массы остается только у отбывающих рейсов. После этого на основе колонок направления перелета, кодов авиакомпаний и кодов регионов перелету в новой колонке «Сегмент» присваивается одно из шести направлений, в разрезе которых далее проводится анализ. Создается еще одна колонка «Авиакомпания - направление», по которой далее будет подтягиваться LF. Также для удобства прогнозирования далее создаются отдельные датафреймы с наблюдениями по каждому месяцу.

Само прогнозирование на метриках MTOW и ATM, проводится помесячно по тем месяцам, файл расписания с которыми был загружен в программу. Программа, работая в цикле, запрашивает по какому году и по какому номеру месяца необходимо сделать прогноз показателей для месяцев из расписания. Также в цикле подгружается файл с данными про Load factor (LF), в котором, как и в расписании создается колонка «Авиакомпания - направление» и в котором есть колонки года и месяца. Именно по данной колонке в датафрейм с расписанием в строки с месяцем, на котором находится цикл, подставляются исторические значения показателей Load factor.

Далее, если исторических значений Load factor по каким-либо причинам не было, этот параметр заполняется по следующим правилам:

* Если значение LF 100% и больше, присваивается значение 95%
* Для новых DOM-направлений присваивается значение 78%
* Для новых INT-направлений присваивается значение 55%
* Везде, где показатель меньше 40% проставляется нижний порог в 40%

Добавляется столбец с количеством пассажиров (PAX), который рассчитывается как Load factor перемноженный на компоновку (количество мест) самолета. Далее определяются переменные с коэффициентами (определенными экспертной оценкой) для ноября и декабря для всех шести направлений (здесь int\_fsc это сегмент INT-OTH).

Text, letter

Description automatically generated

1. Коэффициенты для ноября и декабря по направлениям

После этого создаются три сводные таблицы с показателями PAX, ATM и MTOW также по шести направлениями. Если в данных сводных таблицах нет спрогнозированных значений этих метрик по ноябрю и декабрю, то прогноз на ноябрь заполняется как значение метрики по направлению в октябре перемноженное на коэффициент, декабрь заполняется по аналогии, но на основе ноября.

Далее финальная таблица со спрогнозированным Load factor и сводные таблицы с PAX, ATM и MTOW выгружаются в эксель-файлы.

Стоит отметить, что по Load factor, как было описано ранее прогноз делается более детально – по связке авиакомпания + направление. При этом из такой детализации компания может перейти в прогноз просто по направлениям.

Одним из недостатков данной модели является прогноз двух метрик исключительно в связке месяц-сегмент или месяц-авиакомпания-сегмент, результаты модели выглядят как на рисунке 2. Такой вид прогнозирования ограничивает возможность компании гибко управлять ресурсами и персоналом.



1. Результаты прогнозирования метрики MTOW экспертной моделью

Горизонт прогноза не установлен программно, но данная модель используется для прогнозирования на месяц вперед.

## 1.5 Постановка задачи по методологии GRISP-DM

Существующее решение справляются с прогнозированием временных рядов, но по информации от аэропорта, для более точного планирования операций и сокращения издержек, как нового направления развития аэропорта, точности экспертной модели недостаточно. В дополнение к этому сама модель работает лучше в экономически стабильные периоды, после начала пандемии и с 2022 года модель не учитывает различные тренды в метриках и не имеет возможности спрогнозировать метрики, основываясь на данных трендах.

Для организации процесса постановки бизнес-задачи, обработки и анализа данных, обучения и внедрения моделей используется методология CRISP-DM (Cross-Industry Standard Process for Data Mining).

CRISP-DM — это методология интеллектуального анализа данных, которая охватывает шесть важных этапов:

1. Понимание бизнес-целей (Business Understanding), на котором аналитики изучают процессы, проводимые в компании, и определяют, как анализ данных может помочь в их улучшении.
2. Понимание данных (Data Understanding), где производится подробное изучение имеющихся данных, понимание их структуры и соответствия целям проекта.
3. Подготовка данных (Data Preparation), является одним из наиболее трудоемких и важных этапов. Аналитики должны привести данные в порядок, убрать шумы и дубли, а также провести агрегацию или обогащение данных. В зависимости от специфики компании и направления ее деятельности подготовка данных обычно включает:
   1. консолидацию данных;
   2. формирование выборок;
   3. агрегирование;
   4. обогащение данных;
   5. очистку данных;
   6. разделение данных на обучающие и тестовые.
4. Моделирование (Modeling), на данном этапе строятся и внедряются аналитические модели. Моделирование обычно проводится в несколько итераций. Сначала запускается несколько моделей с параметрами по умолчанию. Затем параметры настраиваются таким образом, чтобы модель выполняла требуемую обработку данных. Если это не удается, приходится возвращаться на этап подготовки данных и вносить изменения.
5. Оценка (Evaluation), где делается анализ того, насколько успешно выполнены поставленные цели.
6. Внедрение (Deployment), на последнем этапе полученные знания и идеи используются для увеличения эффективности компании. Здесь описывается процесс внедрения или даются рекомендации по внедрению готовых моделей в работу компании.[[7]](#footnote-7)

Схема работы данной методологии приведена на рисунке 3.

Изображение выглядит как текст, диаграмма, снимок экрана, круг

Автоматически созданное описание

1. Схема взаимодействия этапов разработки аналитического проекта по методологии CRISP-DM

Источник: Методология CRISP-DM // Веб-сайт компании Loginom. ─ URL: <https://wiki.loginom.ru/articles/crisp-dm.html> (дата обращения: 05.05.2023)

Первый этап CRISP-DM – понимание бизнес-целей, главная цель компании коррелирует с целью выполнения данной работы и заключается в повышении точности прогнозирования для снижения операционных затрат. В рамках первого этапа в разделе выше была описана действующая модель прогнозирования в компании и были определены ее недостатки:

* упрощенность модели, которая не позволяет работать со скрытыми в данных зависимостями
* агрегация прогнозируемых значений по месяцам для метрик MTOW и ATM
* открытый алгоритм, но недостаточная точность прогнозирования

Цель работы была обозначена ранее, в результате работы вышеперечисленные недостатки должны отсутствовать в созданных в данной работе моделях. Также по итогу проведения процесса разработки моделей должны быть реализованы следующие преимущества:

* точность прогнозирования выше по сравнению с существующей моделью
* единый формат прогнозирования по дням для всех сегментов и метрик
* рекомендации по тестированию и внедрению моделей
* оценка сокращения операционных затрат

Представители аналитического отдела аэропорта обозначили необходимую точность, при достижении которой моделями, данный проект будет считаться успешным и будет обсуждаться внедрение моделей в работу. Для того, чтобы модель была успешной, на тестовых данных она должна показать значение средней абсолютной процентной ошибки (MAPE) не превосходящее 10%.

Задача заключается в разработке модели прогнозирования авиационных метрик с повышенной точностью, которая будет динамически адаптироваться к изменениям в работе аэропорта для достижения наилучших результатов прогнозирования на месяц вперед. Для достижения этой цели необходимо опробовать и сравнить в работе статистические методы и методы машинного обучения, чтобы выбрать наилучшую модель. Прогнозирование четырех ключевых метрик – PAX, ATM, MTOW, LF – необходимо проводить в разрезе шести географических сегментов: DOM-MOW, DOM-REG, DOM-INT, INT-BEL, INT-LCC, INT-OTH. Также важно учесть пожелания сотрудников аналитического отдела к настройке модели, чтобы она удовлетворяла их требованиям. В результате выполнения задачи должна быть создана модель, которая позволит аналитическому отделу быстро и эффективно прогнозировать авиационные метрики и принимать на основе этого правильные бизнес-решения.

## Вывод по главе 1

ООО «Воздушные Ворота Северной Столицы» является оператором аэропорта Пулково с 2013 года, и занимается операционной деятельностью, а также модернизацией и реконструкцией аэропорта. В составе издержек большую часть издержек составляют затраты на персонал, значительная доля которого является трафикозависимым, рабочие часы которого можно скорректировать с помощью более точных прогнозов авиационных метрик. Помимо персонала, существуют также другие статьи операционных затрат, которые можно сократить с помощью точного прогнозирования.

Существующая в компании модель прогнозирования предсказывает следующие метрики:

* PAX – Passengers Approximately – Количество пассажиров
* ATM – Air Transport Movement – Взлетно-посадочная операция
* MTOW – Maximum Take-Off Weight – Максимальная взлетная масса
* LF – Load Factor – Коэффициент загрузки рейса (Пассажиры/компоновка)

Также предсказания производятся по шести сегментам, данную структуру прогнозирования необходимо сохранить в разрабатываемых моделях. Существующая модель входит в класс наивных моделей, основывается на расписании и исторических данных, а также использует экспертные коэффициенты для корректировки значений прогноза на некоторые месяца.

Для проведения данного проекта была выбрана методология ведения проектов по анализу данных CRISP-DM. В рамках первого этапа была описана действующая модель прогнозирования в компании и были определены ее недостатки:

* упрощенность модели, которая не позволяет работать со скрытыми в данных зависимостями
* агрегация прогнозируемых значений по месяцам для метрик MTOW и ATM
* открытый алгоритм, но недостаточная точность прогнозирования

Цель работы является повышением точности прогнозирования для снижения операционных затрат и будет достигаться с помощью моделей машинного обучения, в результате работы вышеперечисленные недостатки должны отсутствовать в созданных в данной работе моделях. Также по итогу проведения процесса разработки моделей должны быть реализованы следующие преимущества:

* точность прогнозирования выше по сравнению с существующей моделью
* единый формат прогнозирования по дням для всех сегментов и метрик
* рекомендации по тестированию и внедрению моделей
* оценка сокращения операционных затрат

Для того, чтобы модель была успешной, по запросу аналитического отдела аэрпорта, на тестовых данных она должна показать значение средней абсолютной процентной ошибки (MAPE) не превосходящее 10%.

В результате работы должны быть созданы модели с лучшей точностью, описано внедрение и проведена оценка сокращения операционных затрат, для этого, в следующей главе производится обработка и анализ данных.

# Глава 2. Обработка и анализ данных

## 2.1 Предобработка данных

Вторым и третьим этапами построения аналитического проекта по методологии CRISP-DM являются понимание и подготовка данных. Понимание данных было частично описано ранее (разбивка на сегменты и метрики) и далее этот этап будет продолжен в подразделе, посвященном анализу данных. Для этого, в рамках этапа подготовки данных, происходит их первичная обработка.

Для работы с данными используется Jupyter Notebook – среда, в которой написанный на Python код выполняется в ячейках, поэтапно. Версия Python, которая используется для данного проекта – 3.3.

**Загрузка данных**

Данные представлены по каждому рейсу и содержат следующие переменные:

1. Subsegment – сегмент
2. Flight number – номер рейса
3. IATA – код авиакомпании по кодификации IATA
4. Airline type – тип авиакомпании (российская, международная)
5. Airline short name – имя авиакомпании (сокращенное)
6. Destination – внутренний/международный рейс
7. RAP\_CODE – код аэропорта по кодификации IATA
8. City – название города
9. Country ENG – название страны
10. Region – географический регион
11. Service Type – категория полета (C – чартерный рейс, J – регулярный рейс)
12. Actual Time – фактическое время и дата прилета/вылета
13. Model – тип ВС
14. PAX Actual – фактический пассажиропоток
15. MTOW Actual – фактическая максимальная взлетная масса
16. Departure ATM Actual – количество ВПО на вылет
17. SEATS Actual – компоновка на рейсе

Данные загружаются из Excel файлов начиная с 2015 года и объединяются в один датафрейм для дальнейшей работы. Наблюдений на данном этапе более миллиона, в них нет дубликатов, но есть пропуски, которые необходимо рассмотреть ближе.

**Обработка данных**

Далее производится первичная обработка пропусков – изначально данные выглядят как на рисунке 4.

Изображение выглядит как снимок экрана, Прямоугольник, Параллельный, дизайн

Автоматически созданное описание

1. Визуализация пропусков в данных до обработки

В колонках MTOW и ATM показатели отображены только для рейсов на вылет, наблюдения с пропусками в данных столбцах необходимо удалить и обрабатывать пропуски далее в данных исключительно рейсов из Пулково. Аналогично с PAX – для построения прогноза необходимы наблюдения с известным количеством пассажиров.

Пропуски в IATA и прочих переменных не так важны – задача заключается в построении краткосрочного прогноза по направлениям, без связок с авиакомпаниями, данные колонки будут использованы только в предварительном анализе.

После удаления пропусков в прогнозируемых переменных данные выглядят как на рисунке 5:

Изображение выглядит как Прямоугольник, линия, Параллельный, текст

Автоматически созданное описание

1. Визуализация пропусков в данных после обработки

Далее создается столбец с рассчитанным LF, для этого нужно разделить реальное кол-во пассажиров (столбец PAX Actual) на кол-во мест на воздушном судне (SEATS Actual). Столбец с датой (Actual time) переводится в формат datetime для того, чтобы работать с ним не как со строкой, а как с данными даты, что полезно для построения графиков и в целом многих аспектов дальнейшей работы.

После этого производится работа с переменными и обрабатываются выбросы и знаяения, не соответствующие действительности в прогнозируемых переменных. На рисунке 6 показаны ящики с усами прогнозируемых переменных и пассажирских мест.

Изображение выглядит как текст, снимок экрана, Параллельный, число

Автоматически созданное описание

1. Прогнозируемые переменные и переменная кол-ва пассажирских мест до обработки

На графиках можно заметить:

* Выбросы в значениях PAX, SEATS и MTOW (большие значения, которые не соответствуют реальности)
* Выбросы в LF (слишком маленькая загрузка)
* Ошибки в заполнении ATM (несколько наблюдений с цифрой 2 и 0)

Самый большой самолет в мире имеет 853 пассажирских места в компоновке c одним классом. В Пулково он был в 2018 году и после в трехклассной компоновке с максимальной вместимостью 519, также упоминалась 4-классная компоновка с 575 пассажирскими местами[[8]](#footnote-8), все значения выше 575 вероятно являются технической ошибкой. Возможно, за один рейс было посчитано несколько. Всего таких наблюдений 286 из более чем полумиллиона, стоит избавиться от них. Аналогично с реальной загрузкой пассажиров. Значений больше 575 всего 3, но их стоит убрать. И аналогично с MTOW: так как в данном проекте учитываются только пассажироперевозки, т.е. грузовые рейсы отсутствуют, то показатель стоит оставить максимум взлетной массы самого большого пассажирского судна. Максимальная взлётная масса Airbus A380 — 560 тонн, значения выше удаляются из датасета. Ошибок в индикаторе взлетно-посадочной операции всего 113, они также удаляются из датасета.

Наблюдения со значением LF >100 будут удалены, т. к. количество пассажиров превышающее количество посадочных мест невозможно и такие наблюдения скорее всего возникли по технической ошибке. LF <20 также маловероятен — это либо техническая ошибка, либо аномалия, т. к. обычно при экстремально низкой загрузке рейса он не производится. В итоге финальные ящики с усами изображены на рисунке 7. В данной ситуации, когда важна целостность временного ряда, по статистическим метрикам, например, межквартильному расстоянию, выбросы не удаляются. В итоге, до группировки по дням в датафрейме всего 502346 наблюдения.

Изображение выглядит как текст, снимок экрана, Параллельный, диаграмма

Автоматически созданное описание

1. Прогнозируемые переменные и переменная кол-ва пассажирских мест после обработки

**Деление на сегменты**

Так как данный проект требует прогнозирования в разрезе сегментов, создается структура из словарей, на первом уровне которой ключом является сегмент, а значением вложенный словарь. Во вложенном словаре сегмента ключом является метрика (например, Load factor), а значение – датафрейм с датами, и значениями данной метрики в соответствующие даты. Также, во время упаковки данных в эту структуру производится процедура ресемплинга – группировки значений метрики по дням. При группировке метрик PAX и MTOW в значение переменной в определенный день записывается сумма – сумма пассажиров в день и суммарная максимальная взлетная масса, LF – процент занятых пассажирских мест в определенный день, и ATM – количество взлетно-посадочных операций за день.

На рисунке 8 показано сколько всего наблюдений в словарях на пересечении метрик и сегментов после группировки по дням.

Изображение выглядит как текст, снимок экрана, Шрифт, число

Автоматически созданное описание

1. Количество наблюдений по сегментам и метрикам

## 2.2 Анализ данных

В машинном обучении большую важность имеет процедура подбора гиперпараметров модели – от правильно подобранных параметров зависит возникнет ли проблема пере- или недообучения, а также от них зависит точность предсказаний. [[9]](#footnote-9) В целом, понимание поведения данных может помочь в правильном выборе и настройке модели машинного обучения, а также в интерпретации результатов прогнозирования и корректировке параметров модели на дальнейших стадиях работы. Для анализа прогнозируемых переменных будут использованы описательная статистика, построение гистограмм и построение графиков декомпозированных временных рядов.

Построение гистограмм позволит оценить распределение значений переменной и оценить основные характеристики, такие как среднее значение, медиану, максимальное и минимальное значения и разброс значений, что в свою очередь может помочь в определении оптимальной установки параметров модели.

На рисунках 9 и 10 приведены гистограммы прогнозируемых переменных по сегментам DOM-INT (рейсы российских авиакомпаний заграницу) и DOM-MOW (рейсы в московском направлении), на которых можно увидеть разницу в распределениях переменных.

Изображение выглядит как диаграмма, График, линия

Автоматически созданное описание

1. Распределения значений прогнозируемых метрик для сегмента DOM-INT

Изображение выглядит как диаграмма, текст, График, линия

Автоматически созданное описание

1. Распределения значений прогнозируемых метрик для сегмента DOM-MOW

Например, распределение количества пассажиров по дням на московском сегменте больше напоминает нормальное, чем на интернациональном – там оно смещено в правую сторону, в Москву летает в основном количество людей в районе 4–8 тысяч в день, по интернациональному направлению разброс количества пассажиров гораздо больше и сложнее определить ожидаемое значение. Процент загрузки самолета также отличается: самолеты с загрузкой 70–80 процентов чаще летают в Москву, чем по интернациональным направлениям, это может быть связано с меньшими потерями при перелете на небольшие дистанции при неполностью загруженном воздушном судне. Также, рейсы в Москву более регулярные – в распределении есть отчетливый пик в 50 взлетно-посадочных операций (ATM), по интернациональным направлениям же большой разброс от 0 до 45 операций в день.

В дополнение к гистограммам, для определения точных значений характеристик можно обратиться к описательной статистике, которая приведена для двух сегментов на рисунке 11.

Изображение выглядит как текст, снимок экрана, Шрифт, число

Автоматически созданное описание

1. Описательная статистика для сегментов DOM-INT и DOM-MOW

По описательной статистике для двух данных сегментов можно заметить, что в некоторые дни пассажирских перевозок российскими компаниями из Пулково в аэропорты других стран не совершалось вообще (ATM равное 0), но в Москву летали каждый день на протяжении последних 7 лет. Подобный анализ был проведен по всем сегментам:

* DOM REG также имеет распределение нормальное и смещенное вправо, но отличается от DOM MOW разбросом значений, в основном по нему летают от 5 до 10 тысяч пассажиров в день, но бывают дни когда количество пассажиров превышает 25 тысяч. Вероятно, это праздники и летний период.
* INT BEL имеет большое количество дней с нулевым пассажиропотоком, в целом распределение метрик также напоминает нормальное. Это не самый крупный сегмент, так как он включает в себя только рейсы одной компании «Белавиа», здесь максимальное количество взлетно-посадочных операций за день равно 6.
* Сегмент INT OTH почти полностью повторяет гистограммы сегмента DOM INT – они различаются тем, что первый это рейсы заграницу российских компаний, а второй – иностранных (кроме 7-й свободы воздуха).
* INT LCC, деятельность по которому в настоящий момент приостановлена, имеет большое количество пропущенных значений и в целом является самым небольшим с максимальным количеством взлетно-посадочных операций 2. Он выделяется еще и загрузкой рейса – большинство значений LF находятся в диапазоне 90–100, вероятно из-за небольших цен перелета.

Для более точного подбора гиперпараметров обучаемых далее моделей полезно понять, как выглядят временные ряды в декомпозированном виде. Здесь будет рассмотрена самая очевидна в интерпретации метрика - PAX, которая в обработанных и сгруппированных данных представляет собой суммарное количество пассажиров, вылетающих из Пулково в день в период с 2015 по конец 2022 года.

Декомпозиция временного ряда помогает определить основные компоненты временного ряда, такие как тренд, сезонность и шум. Эти компоненты могут оказаться полезными для выбора и настройки параметров модели машинного обучения.

Например, если временной ряд имеет выраженную сезонность, то может быть использована модель машинного обучения, которая хорошо работает с данными с сезонностью. Декомпозиция временного ряда также может использоваться для выделения признаков, которые будут добавлены в модель машинного обучения. Например, признаком может быть сезонность или скорректированное значения тренда, что может улучшить качество модели.

Изображение выглядит как линия, График, скат, диаграмма

Автоматически созданное описание

1. Декомпозиция метрики PAX для сегмента DOM INT

Изображение выглядит как текст, линия, График, диаграмма

Автоматически созданное описание

1. Декомпозиция метрики PAX для сегмента DOM MOW

На рисунках 12 и 13 приведена декомпозиция метрики PAX для сегментов DOM-INT и DOM-MOW. На данных графиках можно заметить большую разницу в выраженности сезонности во временном ряде: рейсы в Москву имеют сезонность, но слабо выраженную, во временном ряду нет сильно выраженных систематических пиков и впадин, на интернациональном направлении же сезонность выражена сильно, пики приходятся на летние месяцы, в остальное время спрос гораздо ниже. Также сезонность сильно выражена на сегментах DOM REG, INT OTH и более слабо на сегментах INT BEL и INT LCC.

По тренду можно определить в каком состоянии находится сегмент: у обоих сегментов видна просадка в тренде в период пандемии и оба начали активно восстанавливать деятельность в конце 2020. Далее по временной шкале можно заметить, как санкционные ограничения повлияли на интернациональный сегмент, в 2022 году наблюдается падение пассажиропотока на интернациональном сегменте, в то время как на московском линия тренда постепенно приближается к значениям до пандемии. Сегмент DOM REG был наименее подвержен воздействию пандемии и восстановил тренд быстрее других, сегменты INT BEL и INT OTH же пострадали от пандемии больше всего, с большой вероятностью из-за ограничений в международных перелетах. Сегмент INT LCC также был задет пандемией, также можно заметить восстановление тренда в 2021 и прекращение работы сегмента в начале 2022 года.

Помимо годовой сезонности, во всех сегментах присутствует месячная и недельная сезонность, но в данном примере графики были построены именно с выносом годовой.

Декомпозиция также была проведена для всех сегментов по метрике PAX, а результаты анализов использованы в построении моделей.

## Вывод по главе 2

Для выбора инструментов прогнозирования и дальнейшей работы с данными и в рамках второго и третьего этапов CRISP-DM была проведена предобработка данных. Предобработка и анализ были выполнены на языке Python в среде для анализа данных Jupyter. При обработке данных были выполнены следующие шаги:

* Загрузка Excel-файлов с авиационными данными по годам и объединение их в один датафрейм
* Удаление строк с прибывающими рейсами, обработка и удаление пропущенных значений
* Создание расчетного столбца для параметра LF
* Обработка выбросов и значений, не соответствующих реальности

После обработки, так как прогноз необходимо проводить в разрезе сегментов и метрик, была создана структура из словарей, на первом уровне которой ключом является сегмент, а значением вложенный словарь. Во вложенном словаре сегмента ключом является метрика (например, Load factor), а значение – датафрейм с датами, и значениями данной метрики в соответствующие даты. Также, во время упаковки данных в эту структуру была произведена группировка значений метрики по дням. При группировке метрик PAX и MTOW в значение переменной в определенный день записывается сумма – сумма пассажиров в день и суммарная максимальная взлетная масса, LF – процент занятых пассажирских мест в определенный день, и ATM – количество взлетно-посадочных операций за день.

Далее, для лучшего понимания данных и правильного подбора параметров в обучаемых моделях, был произведен анализ данных с помощью следующих инструментов: описательная статистика, построение гистограмм и построение графиков декомпозированных временных рядов. В ходе анализа были оценены распределения значений переменных, основные характеристики и разброс значений по всем сегментам. По декомпозиции временного ряда были определены недельная, месячная и годовая сезонности на большинстве сегментов, различия в интенсивности данной сезонности и различия в поведении трендов. В следующей главе, основываясь в том числе и на информации, полученной из анализа данных, будет произведен выбор моделей, которые могут показать наилучшую точность в рамках задач данной работы.

# Глава 3. Обзор методов машинного обучения, выбор методов для построения моделей

Четвертым этапом в методологии CRISP-DM является моделирование, для этого этапа необходимо провести подготовку в виде выбора подходящих моделей машинного обучения для прогнозирования временных рядов на имеющихся данных.

## 3.1 Определение машинного обучения

Машинное обучение — это область знаний, принадлежащая к сфере искусственного интеллекта, которая включает в себя разработку алгоритмов, позволяющих компьютерам учиться на данных и улучшать свою производительность при выполнении определенных задач с течением времени. Алгоритмы машинного обучения обычно создаются с помощью фреймворков, ускоряющих разработку решений, таких как TensorFlow и PyTorch.

Машинное обучение может быть использовано в широком спектре отраслей, начиная от финансовых рынков и заканчивая диагностикой рака. Оно позволяет компьютерам создавать новые модели поведения на основе полученных из данных закономерностей и применяется для определения потребностей пользователей. Специалисты в области искусственного интеллекта и компьютерных наук в целом уделяют большое внимание машинному обучению как важному направлению исследования. Уже сейчас можно заметить популярность использования машинного и глубокого обучения (подкатегории ML) в различных сферах, и в ближайшие годы ожидается значительный рост его важности. [[10]](#footnote-10)

Существует три основных типа машинного обучения: обучение с учителем, обучение без учителя и обучение с подкреплением. Различия между ними состоят в том, как данные обрабатываются и используются для создания модели:

* Обучение с учителем: процесс, при котором алгоритму вводятся правильные ответы (метки) на задачу, чтобы он мог научиться предсказывать результаты. Например, в задаче классификации изображений алгоритм обучается распознавать определенные объекты на фотографии, изучая уже помеченные метками изображения.
* Обучение без учителя: процесс, в котором алгоритм работает с необработанными данными, не имея доступа к правильным ответам. Он производит кластеризацию данных в подходящую группу, исследуя их сходства и различия. Например, в задаче кластеризации алгоритм группирует схожие по свойствам объекты в определенные категории, не имея заранее предопределенных эталонов.
* Обучение с подкреплением: это процесс, основанный на стимулировании алгоритма при выполнении определенных действий. Алгоритм получает награду, если он выполняет правильное действие, и штраф, если выполняет неправильное. Это позволяет ему оптимизировать свой выбор и поведение, максимизируя награду и минимизируя штрафы. Например, алгоритм может учиться играть в шахматы, где он должен выбрать оптимальный ход, чтобы минимизировать победы соперника в итоге.

В зависимости от задачи и типа данных выбирают различные методы машинного обучения, которые наилучшим образом справятся с решением.

Также часть машинного обучения тесно связана со статистикой, потому что некоторые алгоритмы машинного обучения основаны на статистических методах анализа данных. Одним из примеров, является обучение на основе регрессионных моделей, которые используют статистические методы для предсказания результатов на основе ряда входных данных. Статистические методы также используются для проверки гипотез и оценки показателей точности и надежности моделей машинного обучения. [[11]](#footnote-11)

Решение задачи в рамках данной работы подразумевает применение машинного обучения к прогнозированию временных рядов. Эта задача принадлежит к обучению с учителем т. к. все данные размечены в парах дата – метрика.

В большинстве задач анализ временных рядов попадает под задачу обучения с учителем, где требуется на основе известных временных данных построить модель, способную предсказывать будущие значения. Для этого обычно используются методы регрессии, методы рекуррентных нейронных сетей, таких как LSTM (долгая краткосрочная память) или GRU (управляемые рекуррентные блоки), а также различные алгоритмы временных рядов, такие как ARIMA (модель авторегрессии интегрированного скользящего среднего), SARIMA (сезонная модель авторегрессии интегрированного скользящего среднего) или Prophet (инструмент от Meta, основанный на аддитивной регрессии).

Обучение без учителя также используется в анализе временных рядов, но в более сложных случаях, где размерность данных слишком велика или где задачей является именно обнаружение закономерностей. Одним из основных методов обучения без учителя для анализа временных рядов является кластерный анализ, который позволяет группировать временные ряды по их сходству в заданном пространстве признаков. Также в анализе временных рядов используются методы понижения размерности, например метод главных компонент, который помогает извлекать наиболее информативные признаки из временных данных.[[12]](#footnote-12) При анализе временных рядов без учителя основным преимуществом является возможность обнаружения неожиданных закономерностей и структур в данных, что может помочь в выявлении скрытых причин и предсказании будущих значений ряда. [[13]](#footnote-13)

## 3.2 Различия статистических методов и методов машинного обучения при анализе временных рядов

Поскольку в прогнозировании временных рядов широко используются статистические модели (например, линейная регрессия или ARIMA), необходимо определить разницу между использованием чистых статистических моделей и тех же моделей в машинном обучении. Основное различие между машинным обучением и статистикой заключается в их основной цели. Модели машинного обучения созданы для работы с большими объемами данных и точного прогнозирования будущих значений, в то время как статистические модели призваны делать выводы о взаимосвязях между переменными, используя меньшее количество данных. [[14]](#footnote-14)

Наглядный примером является случай линейной регрессии. Линейная регрессия — это статистический метод, но с помощью машинного обучения можно обучить линейный регрессор и получить тот же результат, что и статистическая регрессионная модель, цель которой - минимизировать квадратичную ошибку между точками данных.

В одном случае происходит обучение модели, которое предполагает использование части данных, называемых тренировочным датасетом. Насколько хорошо будет работать модель будет неизвестно до тестирования модели на дополнительных данных, которых не было в тренировочном датасете во время обучения. Целью машинного обучения в данном случае является получение наилучших результатов на тестовых данных.

Для статистической модели находят линию, которая минимизирует среднюю квадратичную ошибку по всем данным, предполагая, что данные являются линейным регрессором с добавлением случайного шума. Здесь нет необходимости в тренировочном и тестовом наборе. Во многих случаях цель такой модели - охарактеризовать связь между данными и итоговой переменной, а не делать прогнозы относительно будущих данных. Эта процедура называется статистическим выводом, а не прогнозированием. Эта модель может использоваться для прогнозирования, но способ оценки модели не будет включать тестовый набор, а будет включать оценку значимости параметров модели.

Целью машинного обучения с учителем же является получение модели, которая может делать повторяющиеся предсказания. Во многих случаях интерпретируемость модели является не таким важным показателем как в статистике. Машинное обучение ставит акцент на результатах, в то время как статистическое моделирование больше направлено на поиск взаимосвязей между переменными и значимости этих взаимосвязей, хоть и может использоваться для прогнозирования.[[15]](#footnote-15)

В данной работе рассматриваются временные ряды: значения различных авиационных показателей и необходимых для прогнозирования метрик на временной шкале. Прогнозирование временных рядов является важнейшим аспектом многих областей бизнеса, таких как финансы, экономика и маркетинг. Традиционные статистические методы, такие как ARIMA и экспоненциальное сглаживание, являются наиболее популярными подходами для прогнозирования временных рядов. Однако, алгоритмы машинного обучения в последние годы пользуются большим вниманием благодаря своей способности обрабатывать сложные взаимосвязи и нелинейные закономерности в данных.

В статье «Machine Learning vs Statistical Methods for Time Series Forecasting: Size Matters» рассматривается эффективность алгоритмов машинного обучения по сравнению с традиционными статистическими методами для прогнозирования временных рядов. [[16]](#footnote-16)

Авторы поставили перед собой цель сравнить эффективность алгоритмов машинного обучения и традиционных статистических методов для прогнозирования временных рядов. Они отмечают, что предыдущие исследования сравнивали эти два подхода, но не принимали во внимание размер используемого набора данных. Они утверждают, что размер набора данных является важным фактором, который может повлиять на производительность алгоритмов машинного обучения, и поэтому крайне важно учитывать его в таких сравнениях.

Чтобы проверить свою гипотезу, авторы проводят эксперименты на четырех наборах данных разного размера. Для сравнения они используют пять различных алгоритмов машинного обучения и два статистических метода. Используемые алгоритмы машинного обучения: нейронные сети, машины опорных векторов, случайные леса, k-ближайших соседей и градиентный бустинг. В качестве статистических методов используются ARIMA и экспоненциальное сглаживание. Авторы оценивают производительность каждого алгоритма с помощью нескольких показателей, включая среднюю абсолютную процентную ошибку (MAPE), среднюю абсолютную ошибку (MAE) и среднюю квадратичную ошибку (MSE). Они также проводят статистические тесты для определения значимости различий между производительностью каждого алгоритма.

Результаты показывают, что алгоритмы машинного обучения работают лучше традиционных статистических методов для больших наборов данных, но не для малых. Разница в производительности между двумя подходами становится значительной, когда размер набора данных превышает 1000 наблюдений. Кроме того, было обнаружено, что нейронные сети и машины опорных векторов превосходят другие алгоритмы машинного обучения при прогнозировании временных рядов.

Авторы также утверждают, что алгоритмы машинного обучения могут давать более точные прогнозы, чем традиционные статистические методы, но только при наличии достаточного количества данных. Таким образом, предприятия, собирающие большие объемы данных, могут извлечь выгоду из использования алгоритмов машинного обучения для прогнозирования временных рядов.

Также существует статья с немного иным выводом по результатам использования методов машинного обучения в прогнозировании. Статья «Statistical and Machine Learning forecasting methods: Concerns and ways forward»[[17]](#footnote-17) обсуждает проблемы, возникающие при использовании статистических и методов машинного обучения для прогнозирования. В ней проведен анализ 100 наиболее цитируемых статей на тему прогнозирования и сделан вывод, что традиционные статистические методы все еще используются в большинстве исследований.

Авторы статьи признают, что методы машинного обучения могут быть полезными в определенных случаях, например, когда данные имеют сложную структуру или когда необходимо обрабатывать большие объемы информации. Однако, они рекомендуют использовать методы машинного обучения в сочетании со статистическими методами и экспертным мнением, чтобы получить наилучшие результаты прогнозирования.

В статье также обсуждаются вопросы, связанные с выбором метрик для оценки качества прогнозов, выбором моделей и методов обучения, а также с валидацией результатов. Авторы подчеркивают, что важно учитывать контекст и цель прогнозирования, чтобы выбрать наиболее подходящие методы. В целом, статья подчеркивает необходимость сбалансированного подхода к прогнозированию, который учитывает как статистические, так и методы машинного обучения, и принимает во внимание контекст и цель прогнозирования.

В итоге можно сказать, что не существует единственного наилучшего решения для всех бизнес-задач. Однако, модели машинного обучения могут быть в некоторых случаях даже лучше статистических подходов для прогнозирования временных рядов, хоть их эффективность и зависит от конкретной задачи и набора данных. Учитывая большой объем данных для обучения моделей, предоставленный аэропортом, можно предположить о повышении точности прогнозирования необходимых метрик при использовании машинного обучения и, в частности, глубокого обучения. Далее необходимо определить набор методов для получения краткосрочного прогноза.

## 3.3 Выбор методов для построения моделей

Для составления прогноза в рамках задачи данной работы будут использоваться модели машинного и глубокого обучения. После сравнения моделей по максимальной точности прогноза, будет выбрана лидирующая модель, которая будет положена в основу рекомендаций по снижению операционных затрат.

Научная литература показывает, что прогнозы составляются различными способами в зависимости от периода прогнозирования и объема данных, а также от вопросов, ответы на которые необходимо получить из прогноза. Методы прогнозирования варьируются от простых «наивных» прогнозов, более сложных причинно-следственных прогнозов до прогнозов, основанных на статистических моделях и моделях машинного обучения. Наивные прогнозы полностью полагаются на исторические значения того, что прогнозируется. Например, такой прогноз может предполагать, что рост числа пассажиров в следующие пять лет будет таким же, как рост числа пассажиров за последние пять лет. Наивные прогнозы слишком упрощены, чтобы иметь возможность построить понимание того, что можно ожидать в будущем в ответ на изменения политики или экономики страны или региона. Однако они способны быстрее реагировать на изменения в спросе и поэтому иногда используются в качестве эталонов, с которыми можно сравнивать более сложные модели.[[18]](#footnote-18) В Пулково в настоящий момент используется такой вид прогнозирования с экспертными поправками.

Выбор более сложных моделей будет производиться исходя из объема и структуры данных, похожих кейсов по прогнозированию авиационных метрик и различных статей про прогнозирование временных рядов.

Данные представлены с 2007 по 2023 год, но при моделировании по рекомендации представителей аналитического отдела аэропорта, будут использованы данные, начиная с 2015 года. Говоря именно об исторических записях, в каждом эксель-файле с рейсами и метриками за определенный год от 120 тысяч записей – предполагается группировать их по дате по дням для более удобной работы и более наглядного формата прогноза, но даже в этом случае наблюдений будет достаточно для качественного обучения модели и выдвижения гипотезы о лучшей точности именно у моделей машинного обучения.

Обращаясь к научным трудам, в частности в сфере авиационного прогнозирования, одним из наиболее популярных статистических методов и дающим наиболее точные предсказания является ARIMA[[19]](#footnote-19) и эта же модель с сезонным компонентом – SARIMA. SARIMA - единственный статистический метод, способный превзойти, хоть и без статистической разницы, некоторые модели машинного обучения, но точность прогнозирования в таком случае достигается за счет большего числа параметров.[[20]](#footnote-20) Также в кейсах прогнозирования пассажиропотока используются гибридные модели, включающие в себя SVR (регрессия опорных векторов), являющийся алгоритмом машинного обучения, таких примеров достаточно много, гибриды создаются как со статистическими моделями[[21]](#footnote-21), так и с более современными инструментами такими как Prophet[[22]](#footnote-22).

Возвращаясь к статьям, в которых рассматриваются методы машинного обучения, задача данной работы относится к прогнозированию одномерных временных рядов, поэтому будут рассмотрены методы, применимые к имеющимся данным. Выше были приведены две статьи, сравнивающие статистические методы и методы машинного обучения для прогнозирования временных рядов, но в одной из них (Makridakis, Spiliotis)[[23]](#footnote-23) был сделан акцент на гибридных моделях и их точности как самой наилучшей. Во второй статье (Cerqueira, Torgo, Soares)[[24]](#footnote-24) результаты исследования показали, что эффективность различных методов варьировалась в зависимости от размера набора данных.

* Для небольших наборов данных статистические методы, включая ARIMA, превзошли модели машинного обучения.
* Для средних наборов данных случайный лес и градиентный бустинг превзошли статистические модели.
* Для больших наборов данных нейронные сети LSTM превзошли все остальные методы.

Не смотря на то, что для больших наборов данных более эффективны модели машинного обучения, такие как нейронные сети LSTM (этот метод также самый распространенный среди методов машинного обучения для данной задачи[[25]](#footnote-25)), с учетом возможной группировки данных для увеличения скорости вычислений, стоит обратить внимание и на методы, показавшие себя наилучшими для данных среднего объема.

Сравнение различных моделей машинного обучения для прогнозирования одномерных временных рядов, как уже упоминалось, приводит к различным результатам в зависимости от данных. В различных статьях на первые места по точности выходят различные модели, поэтому ориентироваться стоит также на относительную простоту использования модели сотрудниками аэропорта и подтверждение результатов использования в научной литературе. Из анализа литературы по теме можно выделить следующие методы машинного обучения, которые могут подойти под задачу:

* Рекуррентные нейронные сети (RNN): RNN разработаны для обработки последовательных данных и использовались для решения различных задач прогнозирования временных рядов[[26]](#footnote-26)
* Нейронные сети с длинной кратковременной памятью (LSTM): LSTM — это тип RNN, который может изучать долгосрочные зависимости в данных временного ряда
* Регрессия опорных векторов (SVR): SVR — это подход машинного обучения, который использовался для прогнозирования временных рядов и в некоторых случаях показал хорошие результаты
* eXtreme Gradient Boosting (XGBoost): XGBoost — это метод ансамблевого обучения, который показал хорошие результаты прогнозирования для некоторых временных рядов[[27]](#footnote-27)

Эти методы применялись для решения различных реальных задач и продемонстрировали свою эффективность во многих областях. [[28]](#footnote-28)

Помимо обычных методов машинного обучения существуют готовые инструменты такие как Prophet, DeepAR и N-BEATS, созданные большими компаниями и основанные на различных моделях глубокого обучения.[[29]](#footnote-29)

В итоге набор моделей будет следующим: изначальная экспертная модель и методы машинного обучения: LSTM и XGBoost. LSTM показала себя наилучшей на больших датасетах из всех RNN-моделей и подтвердила высокую точность в прогнозировании временных рядов в различных научных работах. Метод градиентного бустинга XGBoost вместе с алгоритмом случайного леса показали хорошие результаты на датасетах средних размеров, несмотря на отсутствие экстраполяции тренда. XGBoost использует деревья решений, как и случайный лес, но в отличие от случайного леса новые деревья не строятся отдельно, а добавляются к результатам работы предыдущих для уменьшения ошибки.

В данной задаче XGBoost является более подходящим вариантом, чем случайный лес, благодаря своей способности обрабатывать пропущенные данные (которые есть в большинстве сегментов и метрик), обнаруживать и использовать краткосрочные закономерности (что важно для краткосрочного прогноза), гибкости при работе с различными типами переменных и масштабируемости при работе с большими наборами данных.[[30]](#footnote-30)

## 3.4 Описание реализуемых методов

В данном разделе приведено описание принципов работы выбранных методов машинного обучения для дальнейшей реализации на данных компании.

**LSTM**

LSTM (long short-term memory, дословно (долгая краткосрочная память) – это один из типов рекуррентных нейронных сетей, который способен обучаться на долгосрочных зависимостях. LSTM модели были представлены в работе Hochreiter, Schmidhuber (1997), а позднее были усовершенствованы и распространены другими исследователями. LSTM хорошо подходят для многих задач и до сих пор широко применяются.

Рекуррентные нейронные сети (Recurrent Neural Networks, RNN). RNN — это тип нейронных сетей, которые содержат обратные связи и могут сохранять информацию.

Нейронные сети с прямой связью столкнулись с несколькими проблемами, такими как невозможность обработки последовательных данных, возможность учитывать лишь текущие данные и отсутствие способности запомнить предыдущий вход. В результате появились рекуррентные нейронные сети, которые могут обрабатывать данные по порядку, используя как текущие, так и прошлые входные данные, а также запоминать результаты прошлых вычислений благодаря встроенной памяти.

Изображение выглядит как снимок экрана

Автоматически созданное описание

1. Рекуррентная нейронная сеть в развернутом виде

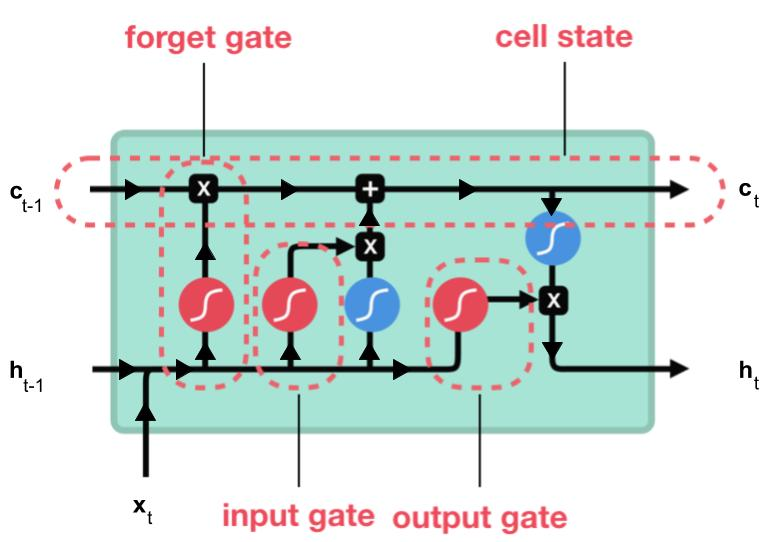
Источник: Understanding LSTMs // Веб-сайт Github. ─ URL: <https://colah.github.io/posts/2015-08-Understanding-LSTMs/> (дата обращения: 30.04.2023).

Рекуррентная нейронная сеть может рассматриваться, как несколько копий одной и той же нейронной сети, где каждая копия передает информацию последующей копии, что проиллюстрировано на рисунке 14.

На каждом цикле работы внутренний слой нейронов, который находится в блоке А, получает набор входных данных Х(t) и информацию о предыдущем состоянии внутреннего слоя А, на основании чего генерирует ответ h(t). Главное свойство этой сети проявляется в возможности обратиться из точки t к информации, сохраненной в блоках до блока t. [[31]](#footnote-31)

За последние несколько лет RNN с успехом применили к целому ряду задач, в числе которых прогнозирование временных рядов. Главный недостаток RNN в том, что при больших временных разрывах между необходимой для предсказания информацией и точкой выполнения предсказания, RNN теряют способность связывать информацию.

LSTM же разработаны специально, чтобы избежать проблемы долговременной зависимости. Любая рекуррентная нейронная сеть имеет форму цепочки повторяющихся модулей нейронной сети, LSTM также имеют эту структуру, похожую на цепочку, но повторяющиеся модули имеет структуру, отличающуюся от RNN. Вместо одного слоя нейронной сети здесь четыре, взаимодействующих особым образом.



1. Компоненты блока LSTM

Источник: Illustrated Guide to LSTM’s and GRU’s: A step by step explanation // Электронный портал, посвященный анализу данных Towards Data Science. ─ URL: <https://towardsdatascience.com/illustrated-guide-to-lstms-and-gru-s-a-step-by-step-explanation-44e9eb85bf21> (дата обращения: 03.05.2023).

LSTM состоит из следующих компонентов, которые указаны на рисунке 15:

1. Forget gate (ворота забывания) – определяют, какую информацию в памяти LSTM необходимо удалить. Эти ворота контролируют, должен ли LSTM сохранять предыдущее состояние или нет. То есть, этот компонент удаляет из памяти LSTM информацию, которая может не быть важна для решения текущей задачи.
2. Input gate (входные ворота) – контролируют, какую информацию следует добавить в память LSTM. Этот компонент решает, какие новые значения следует добавить в текущее состояние LSTM, что происходит путем применения определенных функций к входным данным и сравнения результата применения функций с текущим значением состояния LSTM с помощью математических преобразований.
3. Output gate (выходные ворота) – контролируют, какую информацию следует передать дальше из памяти LSTM. Они используют значение клеточного состояния LSTM, чтобы решить, какую информацию нужно передать дальше.
4. Cell state (клеточное состояние) – это основная часть LSTM, которая отвечает за хранение информации, которая должна передаваться через разные шаги времени. Этот компонент с помощью ворот добавляет или удаляет соответствующую информацию из текущего состояния LSTM. С помощью клеточного состояния информация с более ранних временных шагов может пробиться к более поздним временным шагам, так как оно может нести соответствующую информацию на протяжении всей обработки последовательности блоков LSTM.[[32]](#footnote-32)

**XGBoost**

Для начала стоит ввести несколько базовых понятий. Дерево решений в широком смысле — это метод представления решающих правил в иерархической структуре. Дерево решений в машинном обучении, это алгоритм, который предсказывает значение целевой переменной с помощью применения последовательности простых решающих правил.

Ансамблем в машинном обучении называется комбинация нескольких алгоритмов обучения, которые, работая вместе, позволяют построить модель более эффективную и точную, чем любая из моделей, построенная с помощью отдельного алгоритма

Бустинг — это техника построения ансамблей, в которой предсказатели (например, регрессии или деревья решений) построены не независимо, а последовательно. Эта техника использует идею о том, что следующая модель будет учиться на ошибках предыдущих, каждый следующий алгоритм старается уменьшить ошибку текущего ансамбля. Из-за того, что предсказатели обучаются на ошибках, совершенных предыдущими, требуется меньше времени для того, чтобы получить необходимую точность предсказаний.

Бустинг, использующий деревья решений в качестве базовых алгоритмов, называется градиентным бустингом над решающими деревьями, Gradient Boosting on Decision Trees. Градиентный бустинг — это техника машинного обучения для задач классификации и регрессии, которая строит модель предсказания в форме ансамбля слабых предсказывающих моделей, обычно деревьев решений.

Цель любого алгоритма обучения с учителем — определить функцию потерь и минимизировать её. Для того, чтобы рассмотреть принцип работы градиентного бустинга в качестве функции потерь будет, например, среднеквадратичная ошибка (MSE).

Необходимо построить прогноз таким образом, чтобы MSE была минимальна. Используя градиентный спуск (в данном случае как метод нахождения локального минимума функции) и обновляя предсказания, основанные на скорости обучения (learning rate), производится поиск значений, на которых MSE становится наименьшей. Предсказания обновляются таким образом, что сумма отклонений стремилась к нулю и предсказанные значения были близки к значениям целевой переменной, схема работы градиентного бустинга изображена на рисунке 16.[[33]](#footnote-33)

Изображение выглядит как текст, диаграмма, снимок экрана, линия

Автоматически созданное описание

1. Схематичное изображение алгоритма градиентного бустинга

Источник: Top 10 Interview Questions On Gradient Boosting // Веб-сайт ИТ-ресурса Analytics Vidhya. ─ URL: <https://www.analyticsvidhya.com/blog/2022/11/top-10-interview-questions-on-gradient-boosting/> (дата обращения: 07.05.2023).

В итоге, общие шаги работы алгоритма можно описать так:

1. Сначала создаются и обучаются простые модели, сохраняются и анализируются ошибки.
2. Определяются наблюдения, которые не вписываются в простые модели.
3. Создаются новые модели, которые обрабатывают выдающиеся случаи, выявленные на прошлом этапе. Это повторяется до достижения минимального значения функции потерь.
4. Все построенные модели с весами, которые привели к минимизации ошибки, сохраняются в конечном обученном алгоритме.

Существуют различные инструменты с реализацией алгоритма градиентного бустинга. Одним из них является XGBoost, который расшифровывается как экстремальный градиентный бустинг, это библиотека, написанная на C++, которая оптимизирует обучение для градиентного бустинга.[[34]](#footnote-34)

Алгоритм XGBoost был разработан с учетом оптимизации использования вычислительных ресурсов времени и памяти. Основная задача проекта заключалась в максимально эффективном использовании имеющихся ресурсов для обучения модели. XGBoost поддерживает три основные формы градиентного бустинга, включая стандартный градиентный бустинг с возможностью регулировки скорости обучения (learning rate), стохастический градиентный бустинг и регуляризованный градиентный бустинг. В данной работе используется первая форма.

## Вывод по главе 3

В данной главе было дано понятие машинного обучения, проведено различие между статистическими методами и методами машинного обучения. Было дано обоснование тому, что методы машинного обучения могут показать более высокую точность на имеющихся данных. Также был проведен выбор моделей машинного обучения и дано описание принципов работы выбранных методов. Все это было выполнено в рамках этапа моделирования по методологии CRISP-DM.

Решение задачи в рамках данной работы подразумевает применение машинного обучения к прогнозированию временных рядов. Эта задача принадлежит к обучению с учителем т. к. все данные размечены в парах дата – метрика.

Так как традиционные статистические методы, такие как ARIMA и экспоненциальное сглаживание, являются наиболее популярными подходами для прогнозирования временных рядов, было описано различие использования чистых статистических методов и статистических методов в машинном обучении (и использования машинного обучения для данной цели в принципе). Также, были приведены результаты исследований, доказывающих, что машинное обучение может показывать лучшие результаты, чем статистические методы на средних и больших объемах данных, что в свою очередь подтверждает целесообразность выбора именно машинного обучения для выполнения задачи данной работы.

Далее, на основе научной литературы, результатов анализа данных по авиационным метрикам и прочих источников информации был проведен обзор моделей для прогнозирования временных рядов и выбраны два метода, которые лучше всего подходят к задаче: LSTM (один из типов рекуррентных нейронных сетей, который способен обучаться на долгосрочных зависимостях) и XGBoost (реализация градиентного бустинга, написанная на C++). Для выбранных методов были описаны принципы работы. В следующей главе по данным методам создаются и обучаются модели машинного обучения, проводится оценка точности моделей, а также даются рекомендации по внедрению и производится оценка сокращения операционных затрат.

# Глава 4. Модели прогнозирования авиационных метрик

После выбора моделей можно полноценно приступить к четвертому этапу CRISP-DM – моделированию. Далее будет описана обработка данных для передачи их в модель, деление на выборки для обучения, подбор параметров модели, а также будут приведены графики результатов моделирования.

## 4.1 Обучение моделей

**LSTM**

**Подготовка данных**

Для начала создается глубокая копия lstm\_all изначального словаря со вложенными словарями dict\_all. Так как изначальный словарь является составным объектом и в качестве значений у него изменяемые объекты – словари, необходима именно глубокая копия. Глубокая копия создает новый составной объект, и затем рекурсивно вставляет в него копии объектов, находящихся в оригинале. Она отличается от поверхностной копии, которая вставляет не копии внутренних объектов, а ссылки на них, т. е. при изменении словаря-копии изменяется изначальный словарь, что в данном случае нежелательно.

Перед подачей данных на вход LSTM необходимо произвести их масштабирование, чтобы все значения лежали в определенном диапазоне (обычно [0, 1] или [-1, 1]). При этом масштабирование осуществляется только на обучающей выборке и соответствующим образом применяется к тестовой выборке. Далее данные масштабируются с помощью инструмента scaler и пропущенные значения в них заменяются нулями.

Стоит отметить, что даже после обработки пропусков в изначальном датафрейме, они остались в датафреймах по сегментам. В некоторые дни полеты могли не совершаться в определенных сегментах, тогда значения метрик в эти даты будут отсутствовать. Чтобы сохранить целостность временного ряда, данные пропуски необходимо оставить и заменить нулями для того, чтобы модель могла с ними работать.

Для подготовки данных используется метод скользящего окна (sliding window), который позволит преобразовать исходный временной ряд в последовательность входных и выходных значений.

Определение подходящего размера окна для модели LSTM не имеет универсальной формулы и может происходить в формате эксперимента. Однако в научной литературе есть несколько общих рекомендаций, которые могут быть полезны.

Один из распространенных подходов заключается в том, чтобы выбрать размер окна, по крайней мере, такой же длины, как самую короткую повторяющуюся закономерность в данных, в случае авиационных метрик – это сезоность на различных временных промежутках. Это связано с тем, что модели LSTM предназначены для изучения долгосрочных зависимостей, и использование окна меньшего размера, чем самой короткой повторяющейся закономерности, может затруднить точное отражение этих паттернов моделью. В данных по авиационным метрикам присутствует еженедельная сезонность – это самая наименьшая по длине закономерность, то есть размер окна должен быть не менее 7.

Другой подход заключается в выборе размера окна, достаточно длинного, чтобы уловить как краткосрочные, так и долгосрочные тенденции в данных, но при этом достаточно маленького, чтобы не добавить слишком много шума и не перегрузить модель. В зависимости от сложности данных это может потребовать экспериментов с различными размерами окна, а также настройки других гиперпараметров модели LSTM.[[35]](#footnote-35)

В данной работе рассматриваются ежедневные данных с ежемесячной, еженедельной и ежегодной сезонностью, для которых может быть подходящим размер окна 30–90. Это позволит модели улавливать еженедельные и ежемесячные закономерности, последние из которых особенно важны для прогнозирования на месяц вперед, и в то же время учитывать некоторые повторяющиеся годовые тенденции в данных. В процессе обучения модели с различными размерами окон был определен оптимальный размер 62.

Важным является то, что прогнозирование значений производится на временном отрезке, длина которого равна разнице длин тестовой выборки и размера окна. То есть чтобы получить прогноз на последние 31 день всего датасета, и, если размер окна будет равен 62, необходимо взять размер тестовой выборки равным 93.

Далее необходимо разбить данные на три части: обучающую, валидационную и тестовую выборки. Валидационная выборка будет использоваться для проверки качества прогноза модели на наборе данных, не участвовавшем в обучении, и с помощью результатов, полученных на ней, будут совершенствоваться параметры модели. Тестовая выборка покажет, как на самом деле себя поведет наилучшая модель и какую ошибку покажет. В данном случае разбивка делается на тренировочную, валидационную и тестовую выборки соответственно в соотношении 93.4:3.3:3.3 для первых пяти сегментов с количеством наблюдений 2903 на всех метриках, и в соотношении 88.6:5.57:5.57. Выбор такого соотношения обусловлен краткосрочным форматом прогнозирования и выбранным размером окна 62, прогноз будет совершаться в пределах 31-40 дней. На рисунке 17 показано итоговое количество наблюдений в выборках по сегментам для LSTM.

Изображение выглядит как текст, Шрифт, снимок экрана, рукописный текст

Автоматически созданное описание

1. Итоговое количество наблюдений в выборках по сегментам для LSTM

**Подбор параметров**

Подбор параметров осуществляется с помощью инструмента GridSearchCV библиотеки sklearn. В работе были использованы другие методы подбора параметров такие ка к HalvingGridSearch (полный подбор параметров с обучением моделей не на полном датасете, а на различных его долях в различных итерациях) и RandomizedGridSearch (подбор случайных комбинаций параметров), но они демонстрировали показатели эффективности значительно хуже, из-за чего выбор был сделан именно в пользу полного подбора параметров с помощью GridSearchCV.

GridSearchCV – инструмент поиска лучших параметров из переданного набора возможных значений с помощью перекрёстной проверки (кросс-валидации), метода, который контролирует переобучение модели. При установке в GridSearchCV количества сегментов n, на которые делится датасет для кросс-валидации, модель прогоняется на всех этих сегментах, т. е. n разныx моделей обучаются на n разных выборках. Если все n результатов не имеют больших ошибок, значит модель не переобучена и не недообучена. Этому также способствует передаваемый параметр validation\_data, в котором в модель, при обучении ее на тренировочных данных, передаются валидационные для проверки и контроля переобучения.

Для инициализации модели LSTM используется библиотека tensorflow и подбиблиотеки, где реализованы функции для работы с LSTM.

У модели существует большое количество гиперпараметров, но есть те, правильный подбор которых может оказать большее влияние на точность прогноза:

* num\_nodes: обозначает количество нейронов в памяти ячейки. Когда данных много, увеличение сложности памяти ячейки может дать лучшую производительность, но это замедляет вычисления.
* batch\_size: объем данных, обрабатываемых за один шаг. Увеличение этого параметра повышает производительность, но предъявляет более высокие требования к памяти.
* num\_unrollings: количество временных шагов. Увеличение повышает как производительность, так и требования к памяти и время вычислений.

Также можно подбирать параметр epochs – это гиперпараметр, определяющий количество раз, которое алгоритм обучения прорабатывает весь набор данных для обучения.[[36]](#footnote-36)

В данной работе подбор параметров был реализован следующим образом: были написаны функции для создания моделей LSTM с окном 65 дней, далее в четырех ячейках были реализованы функции подбора параметров и поставлен таймер, замеряющий время выполнения ячейки. Подбор осуществлялся по параметрам batch\_size и epochs, остальные параметры были установлены по умолчанию. В итоге для каждой метрики и сегмента выводились все RMSE и параметры, при которых он был получен, пример на рисунке 18.

Изображение выглядит как текст, снимок экрана, Шрифт, число

Автоматически созданное описание

1. Отображение полученных результатов подбора параметров для сегмента DOM-INT метрики PAX

На рисунке 18 показан итог подбора параметров для одной метрики и одного сегмента. Здесь RMSE отрицательная т. к. GridSearchCV подбирает параметры считая наибольшее значение наилучшим. Также ошибку нельзя интерпретировать, потому что данные нормализованы, и RMSE, так как выражается в тех же единицах, тоже.

Подбор необходимых параметров для перебора производился с ориентировкой на итоговые лучшие параметры: если, например, модели определенного сегмента показывали лучший результат при параметре epochs больше 20 на всех значениях batch\_size, то значения меньше 20 убирались из списка передаваемых в GridSearchCV параметров, и далее добавлялись значения от 20 до 35 с шагом 5. Так, несколько раз, окно передаваемых параметров было смещено для различных метрик. В среднем, на подбор параметров для одной метрики и всех сегментов (одна из четырех ячеек) уходит около 50 минут, в целом он занимает около 3.5 часа.

**Обучение моделей**

После подбора параметров происходит обучение моделей LSTM на тех параметрах, которые показали наименьшие ошибки и были записаны в объект лучших параметров результата GridSearchCV, который был записан в аналогичную используемой для данных структуру из словарей. В качестве входа в LSTM используется последовательность входных значений (скользящее окно) и предсказывается следующее значение в выходных данных.

После обучения модели с оценкой на валидационной выборке, предсказываются значения для временного ряда тестовой выборки и оцениваются путем сравнения с настоящими историческими значениями. При оценке как метрика качества используется корень из среднеквадратичной ошибки (RMSE) и среднеквадратичная процентная ошибка (MAPE).

Так как данные были масштабированы для обучения модели, сами предсказанные значения также были масштабированы. Для интерпретации прогнозированных значений эти значения преобразованы обратно из масштабированного вида в исходный масштаб. Причем, так как для каждого сегмента изначальный scaler был свой и подстраивался именно под данные этого сегмента, обратное масштабирование проводится с помощью ранее внесенных в словарь сегмент-метрика-scaler скейлеров.

Далее был произведен прогноз на тестовых данных. Ниже, на рисунках 19 и 20 приведены графики исторических и предсказанных значений по двум сегментам – один с высоким значением MAPE, другой с низким.

Изображение выглядит как текст, снимок экрана, линия, Шрифт

Автоматически созданное описание

1. Исторические и предсказанные LSTM значения на сегменте DOM INT, метрике PAX

Изображение выглядит как текст, снимок экрана, График, Шрифт

Автоматически созданное описание

1. Исторические и предсказанные LSTM значения на сегменте DOM REG, метрике PAX

Графики предсказаний модели LSTM по всем сегментам и метриками приведены в Приложении 1.

**XGBoost**

**Подготовка данных**

Аналогично с LSTM, первым шагом к подготовке данных является создание xgb\_all – глубокой копии словаря dict\_all. Далее для всех сегментов и метрик в качестве индекса, устанавливается время, а пропуски в данных заполняются нулями.

Один из способов работы с временными рядами в XGBoost подразумевает разложение дат временного ряда на несколько признаков и передачу этих признаков в модель для предсказания зависимой переменной (значений метрики в данном случае). Для этого была написана простая функция, которая из индекса в формате datetime создает пять признаков: год, месяц, день года (например, 365), день месяца (например, 30) и день недели.

Далее ко всем метрикам и сегментам в цикле применяется данная функция, вместе с делением на выборки для обучения и тестирования. В данном случае для деления на тренировочную, валидационную и тестовую выборки не нужно оставлять «запас» под размер скользящего окна. Поэтому соотношения другие, они приведены ниже на рисунке 21.

Изображение выглядит как текст, Шрифт, снимок экрана, рукописный текст

Автоматически созданное описание

1. Итоговое количество наблюдений в выборках по сегментам для XGBoost

После деления на выборки все сегменты делятся на признаки и зависимую переменную, которые также записываются в словарные структуры. Датафрейм с тренировочной выборкой по сегменту INT-BEL метрике PAX до отделения зависимой переменной выглядит как на рисунке 22.

Изображение выглядит как текст, снимок экрана, число, Шрифт

Автоматически созданное описание

1. Зависимая переменная PAX и временные признаки в тренировочной выборке сегмента INT-BEL

**Подбор параметров**

Для подбора параметров для XGBoost был использован GidSearchCV, после экспериментов со случайным поиском и поиском на различных размерах датасета, здесь, как и с LSTM, наивысшую точность показал именно полный подбор. У XGBoost есть много параметров, которые описаны в официальной документации, но главные, которые могут значительно влиять на качество модели и скорость обучения, включают в себя:

* learning\_rate (также известный как шаг градиентного спуска) — это гиперпараметр, который определяет величину обновления весов на каждом шаге. Высокая скорость обучения может привести к переобучению, а слишком низкая - к слабым моделям. Необходимо подобрать оптимальное значение, которое максимизирует качество модели и сокращает время обучения.
* max\_depth — это максимальная глубина каждого дерева решений. Установка слишком большой глубины может привести к переобучению, а слишком маленькой - к слабым моделям. Этот параметр также может быть регулирован с помощью других параметров, таких как min\_child\_weight и gamma.
* min\_child\_weight — это минимальная сумма весов, необходимая для создания нового листа в дереве решений. Большие значения этого параметра помогают предотвратить переобучение.
* gamma — это параметр, который контролирует степень сокращения узлов дерева решений. Он также помогает предотвратить переобучение, уменьшая количество созданных узлов дерева.
* subsample и colsample\_bytree — это параметры, которые определяют долю обучающих данных и признаков, используемых для обучения каждого дерева. Они помогают предотвратить переобучение и ускорить обучение.
* n\_estimators — это количество деревьев решений, которые будут создаваться в процессе обучения. Оптимальное значение зависит от размера и качества данных и может быть подобрано с помощью кросс-валидации.[[37]](#footnote-37)

Важно заметить, что в данном случае, как и с LSTM списки с параметрами редактировались по мере обучения моделей и получения результатов. Экспериментально было выяснено, что с различными комбинациями параметров на всех сегментах и метриках модель дает лучшие результаты при значении n\_estimators равным 800, colsample\_bytree равным 0.5 и min\_child\_weight равным 10. Остальные параметры – subsample, learning\_rate и gamma – в финальных моделях были подобраны под сегменты и метрики с помощью GidSearchCV.

**Обучение моделей**

После подбора параметров, которые были аналогично с параметрами LSTM сохранены в словарную структуру, происходит обучение моделей на лучших параметрах c с использованием валидационных данных для отимизации. После обучения моделей с оценкой на валидационной выборке, предсказываются значения для временного ряда тестовой выборки и оцениваются путем сравнения с настоящими историческими значениями. При оценке используются те же метрики – корень из среднеквадратичной ошибки и среднеквадратичная процентная ошибка.

Также при обучении моделей было учтено пожелание представителей аэропорта и установлен параметр, ограничивающий количество итераций которые могут пройти без улучшения метрики на валидационных данных до преждевременной остановки обучения - early\_stopping\_rounds. Во время обучения модели XGBoost, как было описано в главе 3, каждый шаг производится на основе оценки функции потерь на обучающих данных и проверяется метрика на валидационных данных. Если в течение early\_stopping\_rounds количество шагов не увеличивает метрику на валидационной выборке, то обучение прекращается. Этот параметр помогает избежать переобучения модели на тренировочных данных и улучшить ее обобщающие способности.

Далее производится прогнозирование на тестовых данных. Ниже, на рисунках 23 и 24, приведены графики исторических и прогнозных значений для двух сегментов метрики PAX, тех же, что приведены в LSTM. На графиках также видно, что модель успешнее предсказала значения на сегменте DOM REG

Изображение выглядит как текст, снимок экрана, линия, График

Автоматически созданное описание

1. Исторические и предсказанные XGBoost значения на сегменте DOM INT, метрике PAX

Изображение выглядит как текст, снимок экрана, График, диаграмма

Автоматически созданное описание

1. Исторические и предсказанные XGBoost значения на сегменте DOM REG, метрике PAX

Графики предсказаний модели XGBoost по всем сегментам и метриками приведены в Приложении 2.

## 4.2 Оценка точности, выбор наилучших моделей

Имея готовые обученные модели и прогноз на тестовых временных диапазонах, стоит приступить к пятому этапу CRISP-DM – оценке полученных моделей.

Точность прогноза наивной модели по сегментам можно оценить с помощью вычисления MAPE для LF и PAX. Для этого берутся ежедневные исторические данные LF и PAX за октябрь 2022 года (последний месяц в предоставленном расписании) и с помощью модели также производится прогноз на октябрь 2022. Так как для MTOW и ATM нет ежедневной разбивки, и в данных за 2022 год отсутствует сегмент INT-LCC, то для выполнения задачи данной работы – повышения точности прогнозирования – по остальным метрикам необходимо ориентироваться на запрос аналитического отдела по пороговому значению MAPE 10%.

Для того, чтобы избежать повторного процесса подготовки и очистки данных, значения, которые были предсказаны действующей в аэропорте моделью, сохраняются в словарной структуре данных, и с помощью команды %store данная структура сохраняется в памяти среды Jupyter. Далее, с помощью команды %store -r эта структура переносится в файл с данными и моделями этой работы и там же вычисляется MAPE.

При поиске наименьшей ошибки были выполнены несколько прогнозов на различных базовых месяцах для октября и была выбрана наилучшая для каждого сегмента и метрики. Ниже в таблице 2 представлены ошибки по 5 сегментам при прогнозе LF действующей в аэропорте моделью.

| 1. Значения MAPE для прогноза наивной моделью | | |
| --- | --- | --- |
| **Сегмент** | **Метрика** | **MAPE, %** |
| DOM-INT | PAX | 21 |
| LF | 3 |
| DOM-MOW | PAX | 51 |
| LF | 7 |
| DOM-REG | PAX | 82 |
| LF | 4 |
| INT-BEL | PAX | 35 |
| LF | 3 |
| INT-OTH | PAX | 9 |
| LF | 16 |

Источник: составлено автором

Данная точность, особенно для PAX не выглядит конечной – вероятно прогнозы корректируют в аналитическом отделе далее, но базовая модель показывает данные значения. Также, колебания в точности могли возникнуть из-за агрегации результатов по дням, что в изначальной модели не предусматривалось. Результаты оценки кратковременного прогноза моделей с помощью параметров эффективности приведены в таблицах 3 и 4 ниже.

| 1. Значения показателей эффективности RMSE и MAPE моделей LSTM | | | |
| --- | --- | --- | --- |
| **Сегмент** | **Метрика** | **RMSE** | **MAPE, %** |
| DOM-INT | PAX | 304.83 | 25 |
| MTOW | 142.93 | 26 |
| LF | 7.22 | 7 |
| ATM | 1.27 | 18 |
| DOM-MOW | PAX | 440.71 | 7 |
| MTOW | 236.60 | 7 |
| LF | 2.99 | 3 |
| ATM | 2.44 | 5 |
| DOM-REG | PAX | 741.83 | 5 |
| MTOW | 345.29 | 4 |
| LF | 2.11 | 2 |
| ATM | 4.54 | 4 |
| INT-BEL | PAX | 50.14 | 15 |
| MTOW | 31.02 | 16 |
| LF | 9.02 | 9 |
| ATM | 0.37 | 8 |
| INT-LCC | PAX | 94.12 | - |
| MTOW | 44.54 | - |
| LF | 34.11 | - |
| ATM | 0.53 | - |
| INT-OTH | PAX | 395.39 | 23 |
| MTOW | 335.73 | 32 |
| LF | 5.45 | 5 |
| ATM | 1.88 | 18 |

Источник: составлено автором

Как можно увидеть, не все модели, даже после многократной настройки размеров окон и подбора параметров дают достаточный показатель метрики MAPE на тестовых данных. Значение MAPE на сегменте INT-LCC отсутствует из-за того, что в структуре данных по этому сегменту много пропусков, оценить его по данной метрике невозможно.

| 1. Значения показателей эффективности RMSE и MAPE моделей XGBoost | | | |
| --- | --- | --- | --- |
| **Сегмент** | **Метрика** | **RMSE** | **MAPE, %** |
| DOM-INT | PAX | 397.14 | 29 |
| MTOW | 166.57 | 24 |
| LF | 11.78 | 13 |
| ATM | 1.90 | 23 |
| DOM-MOW | PAX | 889.81 | 15 |
| MTOW | 693.62 | 23 |
| LF | 8.95 | 10 |
| ATM | 9.95 | 25 |
| DOM-REG | PAX | 1007.90 | 7 |
| MTOW | 830.99 | 12 |
| LF | 4.55 | 5 |
| ATM | 13.03 | 13 |
| INT-BEL | PAX | 59.03 | 15 |
| MTOW | 33.69 | 15 |
| LF | 11.89 | 12 |
| ATM | 0.36 | 10 |
| INT-LCC | PAX | 142.63 | - |
| MTOW | 61.93 | - |
| LF | 52.02 | - |
| ATM | 0.68 | - |
| INT-OTH | PAX | 432.60 | 23 |
| MTOW | 408.72 | 29 |
| LF | 5.02 | 5 |
| ATM | 2.50 | 22 |

Источник: составлено автором

Из сравнения таблиц 3 и 4 становится видно, что модели LSTM показали результат значительно лучше, чем модели XGBoost на соответствующих сегментах. Одной из возможных причин такого различия может быть то, что даже с алгоритмом, направленным на корректировку результатов и уменьшение ошибок прогнозирования, при сильных изменениях в тренде в авиационных метриках в последние два года отсутствие возможность экстраполировать тренд помешало градиентному бустингу достичь хороших результатов.

Из 20 моделей XGBoost, для которых была рассчитана MAPE, ее значение выше 10% показали только 3, превысили показатели существующей модели на доступных сегментах и метриках – 4 из 10 возможных. Из 20 моделей LSTM MAPE выше 10% показали 12, превысили показатели существующей модели на доступных сегментах и метриках – 6 из 10 возможных.

## 4.3 Рекомендации по внедрению, оценка влияния на операционные затраты

**Рекомендации по внедрению**

На последнем этапе CRISP-DM, внедрении нового решения по работе с данными, необходимо описать процесс внедрения моделей LSTM, даже тех, которые показали не лучшую точность или дать компании рекомендации по внедрению и улучшению созданного решения. Также, необходимо произвести оценку того, насколько созданная модель может помочь в сокращении операционных затрат. Далее описана стратегия внедрения модели в работу аэропорта, основанная на лучших практиках внедрения[[38]](#footnote-38). Ключевыми шагами в ней являются:

1. Коммуникация. Необходимо объяснить всем заинтересованным сторонам, что является моделями машинного обучения, как их использовать, какие преимущества они могут дать и как изменят бизнес-процессы
2. Подготовка оборудования. Необходимо обеспечить сотрудников средой Jupyter и установить необходимые библиотеки для того, чтобы файл с расчетами мог быть воспроизведен на ПК сотрудников без ошибок. Также, для следующего этапа необходимо объяснить, как оценивать точность созданных моделей и сравнивать с точностью действующей модели.
3. Тестирование. Внедрение моделей машинного обучения в работу организации должно начинаться с тестового периода, например нескольких месяцев. В данном случае, когда точность предсказаний созданных моделей машинного обучения не всегда превосходит существующие, необходимо отследить как будет меняться точность на соответствующих метриках и сегментах во время тестового периода. В случае, если на определенном сегменте и метрике точность созданных моделей будет стабильно превышать точность действующих в аэропорте, ее стоит внедрить в работу аналитического отдела. В случае, если модель для сегмента-метрики показывала лучшую точность не на всем тестовом периоде, ее стоит дообучить и повторить тестирование.
4. Мониторинг. Для мониторинга и управления результатами тестирования и работы моделей стоит создать систему отчетности, которая будет предоставлять данные о точности прогнозирования и анализировать причины несоответствия прогноза фактическим результатам. Она поможет анализировать результаты и вносить предложения по улучшению моделей.
5. Обучение персонала. Как только тестирование успешно завершено, необходимо обучить всех работников, которые будут использовать модель, включая аналитиков и менеджеров, как ее использовать.
6. Улучшение моделей. Необходимо постоянно работать над улучшением моделей, даже показывающих хорошую точность, анализируя новые данные и совершенствуя методы, используемые для обработки и анализа информации. Это позволит создать систему, которая максимально точно прогнозирует авиационные метрики и улучшит работу аэропорта в целом.[[39]](#footnote-39)
7. Интеграция. Последний шаг — это полная интеграция моделей в рабочие бизнес-процессы после проверки моделей на тестовом периоде, создания системы отчетности, введения практики постоянного улучшения моделей и обучения персонала работе с данными моделями. Это может потребовать изменения различных процедур, изменение рутинных задач аналитического отдела по настройке инструментов прогнозирования и включения модели в бизнес-план.

В настоящий момент модель готова для периода тестирования и плавного внедрения, но требует работы по увеличению точности на некоторых из сегментов.

**Оценка влияния на операционные затраты**

Операционные затраты – это издержки, включающие множество повседневно производимых расходов, связанных как с изготовлением и продажей произведенной продукции, так и с управлением предприятием. Подробнее об их составе можно узнать из нашей статьи. Оценка влияния на операционные затраты проведена на основе сегмента DOM MOW и метрики PAX, на данном сегменте наблюдается одно из наибольших улучшений точности и достаточно адекватное предсказание PAX предоставленной моделью. Для расчетов будет взято пороговое значение точности, установленное аналитическим отделом для всех сегментов, так как ошибка в 51% слишком большая для реального значения, и при экспертном использовании данной модели она доходит до поставленного порога.

Научной литературы по теме влияния повышения точности прогнозов на снижение операционных затрат в открытом доступе немного, но есть исследования, которые показывают, как увеличивается выручка при повышении точности прогнозирования. Например, одно из исследований показало, что на каждый 1% повышения точности прогноза можно получить примерно 0,17% выручки в виде экономии при занижении прогноза и 0,20% выручки в виде экономии при завышении прогноза. То есть, если компания с объемом продаж в $100 млн завысила прогноз на 10%, ее издержки упущенных возможностей составят примерно $2 млн, причем чем больше объем продаж, тем больше влияние. [[40]](#footnote-40)

Понижение MAPE может косвенно повлиять на затраты, поскольку позволяет принимать более эффективные решения и делать более точные прогнозы, эта связь не всегда линейная, но можно сделать определенные предположения. Для расчета сокращения операционных затрат на функционирование аэропорта необходимо знать текущие затраты и ожидаемое улучшение прогнозирования пассажиропотока. Последняя бухгалтерская отчетность в открытом доступе датирована 2022 годом, данные об операционных затратах также будут взяты за 2022 год.

Операционные затраты аэропорта оцениваются по информации из опубликованной на сайте audit-it. По отчету о финансовых результатах, расходы за 2022 год составили 12 554 584 тыс. руб. (себестоимость продаж, управленческие и прочие расходы, исключаются налоги и проценты к уплате)[[41]](#footnote-41).

Доля операционных расходов в общей структуре расходов аэропортов может варьироваться в зависимости от размера и функционала аэропорта, а также от страны и региона, где находится аэропорт. Однако, как правило, операционные расходы являются основным расходным статьям у аэропортов.

Согласно исследованию "Airport Economics Report 2019" компании Airports Council International (ACI)[[42]](#footnote-42), операционные расходы составляют около 66% от всех расходов аэропортов. Оставшиеся 34% расходов аэропортов связаны с инвестиционными расходами на разработку инфраструктуры аэропорта, обновление и модернизацию оборудования и другие капитальные затраты. В итоге получаем 8 286 025.44 тыс. руб.

Далее необходимо оценить долю операционных затрат, приходящихся на сегмент DOM MOW ориентируясь на долю пассажиропотока среди всех сегментов. В 2022 году количество пассажиров, отбывших из Пулково по всем сегментам, составило 8 382 427 человек, количество пассажиров, отбывших из Пулково в Москву, составило 5 245 849 человек. Доля пассажиропотока на сегменте DOM MOW составила 62,6%. Далее, при условии того, что операционные расходы распределяются единым образом между сегментами по количеству пассажиров, рассчитаны операционные затраты на сегмент DOM MOW как 62,6 \* 8 286 025.44 / 100, что в итоге дает 5 178 052 тыс. руб.

С помощью созданных в данной работе моделей было получено улучшение точности прогнозирования пассажиропотока на 3%, оно должно сократить ошибку прогноза с установленной 10% до полученной 7%. Для примерной оценки сокращения затрат можно предположить, что операционные затраты сократятся на ту величину процентов, на которую может повыситься выручка, опираясь на информацию из научной литературы. Т.е предположив, что при повышении точности на 1%, затраты снижаются на 0,2%, то при повышении точности на 3% затраты снизятся на 0,6%. Это даст сокращение затрат в денежном объеме в 31 068.31 тыс. р., что скорее является верхней границей для возможного понижения затрат.

То есть, использование новой модели на сегменте DOM MOW помогло снизить процентную ошибку прогнозирования на 3%, что, в свою очередь, может привести к сокращению операционных расходов на 31 068.31 тыс.р. Этот пример не является точным расчетом, поскольку не учитываются конкретные статьи расходов и их доли в общих затратах, но он дает представление об экономической эффективности новой модели.

## Вывод по главе 4

В данной главе были реализованы три последних этапа методологии ведения проектов по анализу данных CRISP-DM: моделирование, оценка и внедрение.

Для обоих методов – LSTM и XGBoost – этап моделирования был разделен на три шага: подготовка данных, подбор гиперпараметров и обучение моделей. При подготовке данных были совершены необходимые преобразования над данными чтобы передать их в модели, а также данные были разделены на тренировочную, валидационную и тестовую выборки. Были выделены параметры, которые необходимо настроить для того, чтобы модель продемонстрировала хорошие результаты. Подбор параметров в обоих случаях производился с помощью инструмента GridSearchCV, точность моделей при подбое параметров оценивалась на валидационной выборке. Далее, параметры, с использованием которых была получена наименьшая ошибка, были сохранены в аналогичной созданной для данных словарной структуре, и модели были обучены на данных параметрах.

Результаты прогноза были визуализированы с помощью графиков, а также для обеих кластеров моделей были замерены ошибки RMSE и MAPE по сегментам. По итогам прогнозирования, из 20 моделей LSTM (на которых возможно замерить MAPE) MAPE выше 10% показали 12, превысили показатели существующей модели на доступных сегментах и метриках – 6 из 10 возможных. XGBoost же показал худшие результаты – MAPE выше 10% показали только 3 модели из 20, превысили показатели существующей модели на доступных сегментах и метриках – 4 из 10 возможных.

Далее были даны рекомендации по внедрению LSTM моделей в работу аэропорта и оценено потенциальное сокращение операционных затрат при повышении точности прогнозирования от пороговой 10% до 7% на сегменте DOM MOW – оно составило более 30 млн рублей. Этот пример не является точным расчетом, но дает представление об экономической эффективности новой модели.

Заключение

Данная работа выполнена в формате прикладного проекта для ООО «Воздушные Ворота Северной Столицы» – управляющей компании аэропорта Пулково, расположенного в Санкт-Петербурге. В то время как статистические и экспертные модели широко используются для прогнозирования метрик, таких как пассажиропоток и грузовые объемы, наступление экономических кризисов и пандемия COVID-19 подчеркивают необходимость использования более адаптивных методов прогнозирования, в том числе машинного обучения. Внедрение машинного обучения в авиаиндустрии – важный шаг для улучшения прогнозирования метрик и повышения качества обслуживания пассажиров. Для аэропорта Пулково это означает повышение операционной эффективности, сокращение операционных расходов и увеличение удовлетворенности пассажиров услугами, предоставляемыми аэропортом.

Проект проводился в соответствии с методологией для ведения проектов по анализу данных CRISP-DM. В ходе работы над проектом, в целях совершенствования процесса прогнозирования авиационных метрик и оценки влияния повышения точности на снижение операционных затрат следующие задачи:

1. Проанализирована деятельность ООО «Воздушные Ворота Северной Столицы», сформировано представление о действующем процессе прогнозирования авиационных метрик
2. Выявлены недостатки и области развития в процессе прогнозирования;
3. Выявлены пути совершенствования процесса прогнозирования с помощью методов машинного обучения
4. Разработаны модели прогнозирования и определить наилучшую по показателям эффективности
5. Оценены возможности сокращения операционных затрат на результатах наилучшей модели

На первом этапе CRISP-DM и в первой главе данной работы было дано описание компании, ее структуры издержек и выяснено что значительную долю составляют операционные издержки, на которые влияет точность прогнозирования. Также была описана существующая наивная модель с экспертными коэффициентами, определены ее главные недостатки и требования к разрабатываемым моделям машинного обучения. Помимо этого, была описана сама методология CRISP-DM.

На втором и третьем этапах CRISP-DM и во второй главе работы была проведена предобработка данных. Также был проведен анализ данных по сегментам, в ходе него были оценены распределения значений переменных, основные характеристики и разброс значений по всем сегментам, сезонности и тренды. Эта информация была использована понимания данных, выбора моделей и подбора параметров моделей.

В главе 3 был затронут этап моделирования CRISP-DM, дано определение машинному обучению проведено различие между статистическими методами и методами машинного обучения. Было дано обоснование тому, что методы машинного обучения могут показать более высокую точность на имеющихся данных. Также был проведен выбор моделей машинного обучения, итогом которого стал выбор методов LSTM и XGBoost и дано описание принципов работы выбранных методов.

В последней главе были описаны три последних этапа методологии CRISP-DM: моделирование, оценка и внедрение. По итогам обучения моделей и построения прогнозов лучших результат прогнозирования показали модели LSTM - из 20 моделей (на которых возможно замерить MAPE) MAPE выше 10% показали 12, превысили показатели существующей модели на доступных сегментах и метриках – 6 из 10 возможных. Самая высокая точность была показана на сегменте DOM REG и метрике Load Factor – MAPE там составила 2%. Также были и не самые точные результаты с ошибкой более 20%. Рекомендации по внедрению были даны с расчетом на тестирование моделей в работе компании и включение их модернизации в работу аналитического отдела. Также было оценено сокращение операционных затрат при повышении точности прогнозирования от пороговой 10% до 7% на сегменте DOM MOW – оно составило более 15 млн рублей.

В данный момент итогом работы является готовое для тестирования, модернизации и последующего внедрения решение по прогнозированию. Единый формат прогноза на месяц вперед позволит распределять ресурсы аэропорта основываясь на всех прогнозируемых метриках, а ежедневная разбивка позволит включить элементы гибкости в управление аэропортом. Цель по повышению точности была достигнута не на всех сегментах, но на более чем половине из них модели машинного обучения показали результат лучше поставленного компанией порога.

Для компании, после проведения всех шагов по внедрению моделей, их использование может значительно сократить операционные затраты, которые зависят от точности прогнозирования. Как было указано в работе, большая часть затрат у аэропортов приходится на операционные затраты. При развитии прогнозирования с помощью машинного обучения, введении в работу и дальнейшей настройке созданных моделей для получения наивысшей точности на всех рассматриваемых сегментах и метриках, аэропорт Пулково и его управляющая компания смогут улучшить свою операционную деятельность, снизить затраты и увеличить лояльность клиентов.

# Список литературы

1. A. Parmezan, Evaluation of statistical and machine learning models for time series prediction: Identifying the state-of-the-art and the best conditions for the use of each model / Antonio Parmezan, Vinicius Souza, Gustavo Batista, // Information Sciences Journal. ─ 2019. ─ Vol. 484. ─ P. 302-337. ─ DOI: 10.1016/j.ins.2019.01.076. ─ URL: [https://www.researchgate.net/publication/330742498\_Evaluation\_of\_statistical](https://www.researchgate.net/publication/330742498_Evaluation_of_statistical_and_machine_learning) (дата обращения: 26.04.2023).
2. Abbasi R., Short Term Load Forecasting Using XGBoost / Abbasi, R.A., Javaid, N., Ghuman, M.N.J., Khan, Z.A., Ur Rehman, S., Amanullah // Web, Artificial Intelligence and Network Applications. ─ 2018. ─ DOI: 0.1007/978-3-030-15035-8\_108. ─ URL: [https://www.researchgate.net/publication/331746834\_Short\_Term\_Load](https://www.researchgate.net/publication/331746834_Short_Term_Load_Forecasting_Using_XGBoost) (дата обращения: 30.04.2023).
3. Aileen Nielsen, Practical Time Series Analysis: Prediction with Statistics and Machine Learning. ─ O'Reilly Media, Inc., 2019. ─ 504 p. ─ ISBN 9781492041658.
4. Airport Economics Report, 2019 // Веб-сайт Airports Council International. ─ URL: <https://store.aci.aero/product/airport-economics-report-2019/> (дата обращения: 20.05.2023).

Always be prepared: The importance of forecasting during winter // Веб-сайт интернет-газеты International Airport Review. ─ URL: <https://www.internationalairportreview.com/article/76669/always-be-prepared-winter-weather/> (дата обращения: 09.04.2023).

Benefits to Better Financial Forecasting // Веб-сайт FPAExperts. ─ <https://www.fpaexperts.com/studies-show-increasing-financial-forecast-accuracy-will-save-your-company-millions-heres-how/> (дата обращения: 23.05.2023).

1. Cerqueira V. Machine Learning vs Statistical Methods for Time Series Forecasting: Size Matters / Vitor Cerqueira, Luis Torgo, Carlos Soares // Электронное издание ArXiv. ─ 2019. ─ Vol. abs/1909.13316. ─ URL: <https://arxiv.org/pdf/1909.13316.pdf> (дата обращения: 16.04.2023).
2. E. M. Shishkov, A. V. Pronichev, Applying Machine Learning Methods for Power Plant Generation Time Series Forecasting // 2022 International Conference on Industrial Engineering, Applications and Manufacturing (ICIEAM). ─ 2018. ─ P. 67–71. ─ URL: <https://www.semanticscholar.org/paper/Applying-Machine-Learning-Methods-for-Power-Plant> (дата обращения: 28.04.2023).

Exploring the LSTM Neural Network Model for Time Series // Электронный портал, посвященный анализу данных Towards Data Science. ─ URL: <https://towardsdatascience.com/exploring-the-lstm-neural-network-model-for-time-series-8b7685aa8cf> (дата обращения: 09.04.2023).

1. Forecasting time series: using statistics vs machine learning // Веб-сайт блога Data Warehousing and Data Science. ─ URL: <https://dwbi1.wordpress.com/2022/02/16/forecasting-time-series-using-statistics-vs-machine-learning/> (дата обращения: 28.04.2023).
2. Henrik Heymann, Guideline for Deployment of Machine Learning Models for Predictive Quality in Production / Henrik Heymann, Alexander D. Kies, Maik Frye, Robert H. Schmitt, Andrés Boza // Procedia CIRP. ─ 2018. ─ Vol. 107. ─ P. 815–820. ─ DOI: 10.1016/j.procir.2022.05.068. ─ URL: <https://www.sciencedirect.com/science/article/pii/S2212827122003523> (дата обращения: 18.05.2023).
3. Illustrated Guide to LSTM’s and GRU’s: A step by step explanation // Электронный портал, посвященный анализу данных Towards Data Science. ─ URL: <https://towardsdatascience.com/illustrated-guide-to-lstms-and-gru-s-a-step-by-step-explanation-44e9eb85bf21> (дата обращения: 03.05.2023).
4. Incredibly Useful Time Series Forecasting Algorithms// Веб-сайт аналитического агентства Advancing Analytics. ─ URL: <https://www.advancinganalytics.co.uk/blog/2021/06/22/10-incredibly-useful-time-series-forecasting-algorithms> (дата обращения: 30.04.2023).
5. Kiefer D., Demand Forecasting Intermittent and Lumpy Time Series: Comparing Statistical, Machine Learning and Deep Learning Methods / Kiefer, D., Grimm, F., Bauer, M., & Dinther, C.V. // Hawaii International Conference on System Sciences. ─ 2021. ─ DOI: 10.24251/HICSS.2021.172. ─ URL: [https://www.semanticscholar.org/paper/Demand-Forecasting-Intermittent](https://www.semanticscholar.org/paper/Demand-Forecasting-Intermittent-and-Lumpy-Time-and-Kiefer-Grimm/89739506fc40ac3307555b21f883220e24ba060e) (дата обращения: 30.04.2023).
6. Liang Guo, The hybrid PROPHET-SVR approach for forecasting product time series demand with seasonality / Liang Guo, Weiguo Fang, Qiuhong Zhao, Xu Wang // Computers & Industrial Engineering ─ 2021. ─ Vol. 161. ─ DOI: 10.1016/j.cie.2021.107598. ─ URL: <https://www.sciencedirect.com/science/article/abs/pii/S0360835221005027> (дата обращения: 26.04.2023).
7. Linsey Pang, Applied Machine Learning Methods for Time Series Forecasting / Linsey Pang, Wei Liu, Lingfei Wu, Kexin Xie, Stephen Guo, Raghav Chalapathy, and Musen Wen // 31st ACM International Conference on Information & Knowledge Management. ─ 2022. ─ DOI: 10.1145/3511808.3557492. ─ URL: [https://www.semanticscholar.org/paper/Applied-Machine-Learning](https://www.semanticscholar.org/paper/Applied-Machine-Learning-Methods-for-Time-Series-Pang-Liu/298f95a128033ce3346d6f509233b9a7729a8f27) (дата обращения: 30.04.2023).

Machine learning, explained // Веб-сайт MIT Sloan School of Management. ─ URL: <https://mitsloan.mit.edu/ideas-made-to-matter/machine-learning-explained> (дата обращения: 11.04.2023).

1. Makridakis S. Statistical and Machine Learning forecasting methods: Concerns and ways forward / Makridakis S, Spiliotis E, Assimakopoulos V // Электронное издание PLoS ONE. ─ 2018. ─ Vol. 13, Issue 3. ─ DOI: 10.1371/journal.pone.0194889. ─ URL: <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0194889> (дата обращения: 19.04.2023).
2. Robin W. Long-term and Short-term Forecasting Techniques for Regional Airport Planning : Master’s Thesis in Mathematical Statistics // KTH Royal institute of technology, 2016. ─ URL: <https://www.diva-portal.org/smash/get/diva2:954470/FULLTEXT02> (дата обращения: 26.04.2023).
3. Shuojiang Xu, Forecasting the demand of the aviation industry using hybrid time series SARIMA-SVR approach / Shuojiang Xu, Hing Kai Chan, Tiantian Zhang // Transportation Research Part E: Logistics and Transportation Review ─ 2019. ─ Vol. 122. ─ P. 169-180. ─ DOI: 10.1016/j.tre.2018.12.005. ─ URL: [https://www.researchgate.net/publication/329625988\_Forecasting\_the\_demand](https://www.researchgate.net/publication/329625988_Forecasting_the_demand_of_the_aviation_industry) (дата обращения: 26.04.2023).
4. The Actual Difference Between Statistics and Machine Learning // Электронный портал, посвященный анализу данных Towards Data Science. ─ URL: <https://towardsdatascience.com/the-actual-difference-between-statistics-and-machine-learning-64b49f07ea3> (дата обращения: 16.04.2023).

The Importance of Accurate Forecasting for Aviation // Веб-сайт аналитического агентства DTN. ─ URL: <https://www.dtn.com/the-importance-of-accurate-forecasting-for-aviation/> (дата обращения: 09.04.2023).

1. The Most Successful Approaches to Leading Organizational Change // Веб-сайт издания Harvard Business Review. ─ URL: <https://hbr.org/2023/04/the-most-successful-approaches> (дата обращения: 20.05.2023).
2. Time Series Forecasting: Statistical Model to Predict Future Values // Веб-сайт портала Code Algorithms Pvt. Ltd. ─ URL: <https://www.enjoyalgorithms.com/blog/time-series-forecasting-using-machine-learning> (дата обращения: 13.04.2023).
3. Top 10 Interview Questions On Gradient Boosting // Веб-сайт ИТ-ресурса Analytics Vidhya. ─ URL: <https://www.analyticsvidhya.com/blog/2022/11/top-10-interview-questions-on-gradient-boosting/> (дата обращения: 07.05.2023).
4. Understanding LSTMs // Веб-сайт Github. ─ URL: <https://colah.github.io/posts/2015-08-Understanding-LSTMs/> (дата обращения: 30.04.2023).

Using data to enhance forecast and demand planning for airports // Веб-сайт аналитического агентства авиационных данных Cirium. ─ URL: <https://www.cirium.com/thoughtcloud/data-enhance-forecast-and-demand-planning-airports/> (дата обращения: 09.04.2023).

What is machine learning? // Веб-сайт компании IBM. ─ URL: <https://www.ibm.com/topics/machine-learning> (дата обращения: 11.04.2023).

1. XGBoost // Веб-сайт ИТ-ресурса GeeksforGeeks. ─ URL: <https://www.geeksforgeeks.org/xgboost/> (дата обращения: 07.05.2023).
2. XGBoost Parameters // Веб-сайт с официальной документацией XGBoost. ─ URL: <https://xgboost.readthedocs.io/en/latest/parameter.html> (дата обращения: 14.05.2023).
3. Бухгалтерская отчетность ООО «Воздушны Ворота Северной Столицы» // Веб-сайт Audit-it. ─ URL: <https://www.audit-it.ru/buh_otchet/7703590927> (дата обращения: 20.05.2023).
4. Градиентый бустинг — просто о сложном // Электронный портал, посвященный анализу данных Neurohive. ─ URL: <https://neurohive.io/ru/osnovy-data-science/gradientyj-busting/> (дата обращения: 07.05.2023).

Деньги из воздуха. Аэропорт «Пулково» расширит центральный терминал // Веб-сайт газеты Ведомости Санкт-Петербург. ─ URL: <https://spbvedomosti.ru/news/financy/dengi_iz_vozdukha_aeroport_pulkovo_rasshirit_tsentralnyy_terminal/> (дата обращения: 07.03.2023).

История аэропорта // Веб-сайт аэропорта Пулково. ─ URL: <https://pulkovoairport.ru/about/about_pulkovo/history/> (дата обращения: 07.03.2023).

1. Метод главных компонент // Веб-сайт аналитического агентства АСУ-Аналитика. ─ URL: <https://asu-analitika.ru/metod-glavnyh-komponent> (дата обращения: 13.04.2023).

Методология CRISP-DM // Веб-сайт компании Loginom. ─ URL:<https://wiki.loginom.ru/articles/crisp-dm.html> (дата обращения: 05.05.2023).

Оператор аэропорта // Веб-сайт аэропорта Пулково. ─ URL: <https://pulkovoairport.ru/about/about_pulkovo/operator/> (дата обращения: 07.03.2023).

Пулково встретил самый большой пассажирский самолет Airbus A380 // Веб-сайт аэропорта Пулково. ─ URL: [https://pulkovoairport.ru/about/press\_center/news/9657/#:~:text](https://pulkovoairport.ru/about/press_center/news/9657/#:~:text=%D0%90%D1%8D%D1%80%D0%BE%D0%BF%D0%BE%D1%80%D1%82%20%D0%9F%D1%83%D0%BB%D0%BA%D0%BE%D0%B2%D0%BE%2025%20%D0%BE%D0%BA%D1%82%D1%8F%D0%B1%D1%80%D1%8F%202018,Airbus%20A380%2D800%20%D0%B0%D0%B2%D0%B8%D0%B0%D0%BA%D0%BE%D0%BC%D0%BF%D0%B0%D0%BD%D0%B8%D0%B8%20%D0%AD%D0%BC%D0%B8%D1%80%D0%B5%D0%B9%D1%82%D1%81) (дата обращения: 09.04.2023).

1. Сказки от TENSORFLOW и LSTM // Веб-сайт ИТ-портала Хабр. ─ URL: <https://habr.com/ru/articles/656635/> (дата обращения: 14.05.2023).
2. Чем машинное обучение отличается от статистики // Веб-сайт ИТ-портала Хабр. ─ URL: <https://habr.com/ru/companies/skillfactory/articles/570736/> (дата обращения: 13.04.2023).

# Приложения

## Приложение 1. Графики прогнозов моделей LSTM по всем сегментам и метрикам

Изображение выглядит как текст, снимок экрана, линия, Шрифт

Автоматически созданное описаниеИзображение выглядит как текст, снимок экрана, линия, График

Автоматически созданное описание

Изображение выглядит как текст, снимок экрана, Шрифт, линия

Автоматически созданное описаниеИзображение выглядит как текст, снимок экрана, линия, График

Автоматически созданное описание

Изображение выглядит как текст, снимок экрана, График, линия

Автоматически созданное описаниеИзображение выглядит как текст, снимок экрана, График, линия

Автоматически созданное описание

Изображение выглядит как текст, снимок экрана, График, линия

Автоматически созданное описаниеИзображение выглядит как текст, снимок экрана, линия, График

Автоматически созданное описание

Изображение выглядит как текст, снимок экрана, График, Шрифт

Автоматически созданное описаниеИзображение выглядит как текст, снимок экрана, График, линия

Автоматически созданное описаниеИзображение выглядит как текст, снимок экрана, Шрифт, График

Автоматически созданное описаниеИзображение выглядит как текст, снимок экрана, График, Шрифт

Автоматически созданное описание

Изображение выглядит как текст, снимок экрана, линия, Шрифт

Автоматически созданное описаниеИзображение выглядит как текст, снимок экрана, График, линия

Автоматически созданное описание

Изображение выглядит как текст, снимок экрана, График, линия

Автоматически созданное описаниеИзображение выглядит как текст, снимок экрана, линия, График

Автоматически созданное описание

Изображение выглядит как текст, снимок экрана, линия, График

Автоматически созданное описаниеИзображение выглядит как текст, снимок экрана, линия, График

Автоматически созданное описаниеИзображение выглядит как текст, снимок экрана, Шрифт, линия

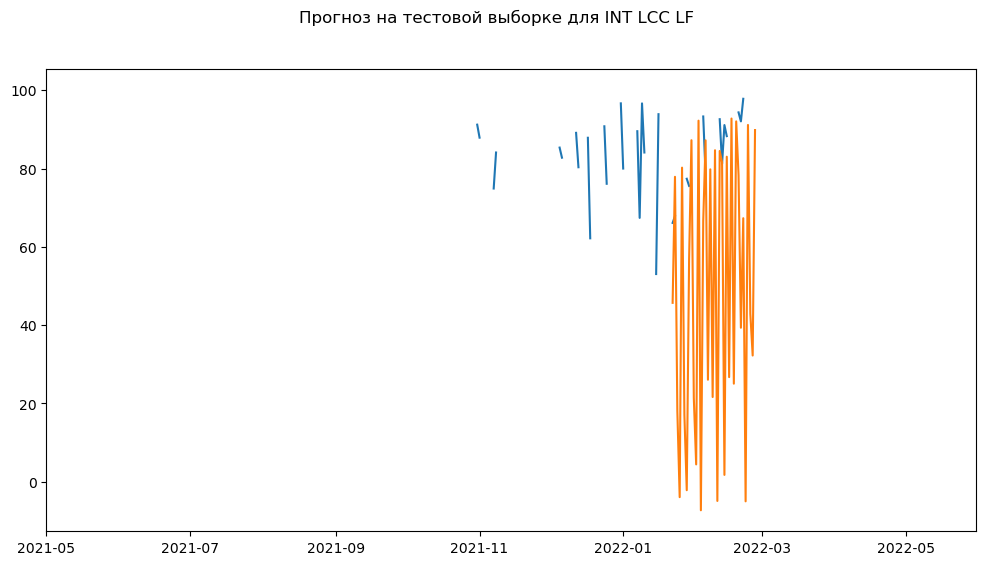
Автоматически созданное описаниеИзображение выглядит как текст, снимок экрана, линия, График

Автоматически созданное описание

Изображение выглядит как текст, снимок экрана, График, линия

Автоматически созданное описаниеИзображение выглядит как текст, снимок экрана, График, линия

Автоматически созданное описание

Изображение выглядит как текст, снимок экрана, линия, График

Автоматически созданное описание

## Приложение 2. Графики прогнозов моделей XGBoost по всем сегментам и метрикам

Изображение выглядит как текст, снимок экрана, линия, График

Автоматически созданное описаниеИзображение выглядит как текст, снимок экрана, линия, График

Автоматически созданное описаниеИзображение выглядит как текст, снимок экрана, Шрифт, линия

Автоматически созданное описаниеИзображение выглядит как текст, линия, График, Шрифт

Автоматически созданное описание

Изображение выглядит как текст, снимок экрана, График, линия

Автоматически созданное описаниеИзображение выглядит как текст, снимок экрана, График, диаграмма

Автоматически созданное описаниеИзображение выглядит как текст, снимок экрана, линия, График

Автоматически созданное описание

Изображение выглядит как текст, снимок экрана, линия, График

Автоматически созданное описание

Изображение выглядит как текст, снимок экрана, График, диаграмма

Автоматически созданное описаниеИзображение выглядит как текст, снимок экрана, График, диаграмма

Автоматически созданное описание

Изображение выглядит как текст, снимок экрана, Шрифт, График

Автоматически созданное описаниеИзображение выглядит как текст, снимок экрана, График, линия

Автоматически созданное описание

Изображение выглядит как текст, снимок экрана, График, линия

Автоматически созданное описание Изображение выглядит как текст, снимок экрана, линия, График

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

Автоматически созданное описаниеИзображение выглядит как текст, снимок экрана, линия, График

Автоматически созданное описание

Изображение выглядит как текст, снимок экрана, линия, График

Автоматически созданное описаниеИзображение выглядит как текст, снимок экрана, линия, График

Автоматически созданное описаниеИзображение выглядит как текст, снимок экрана, линия, График

Автоматически созданное описаниеИзображение выглядит как текст, снимок экрана, линия, График

Автоматически созданное описание

Изображение выглядит как текст, снимок экрана, График, линия

Автоматически созданное описаниеИзображение выглядит как текст, снимок экрана, График, линия

Автоматически созданное описаниеИзображение выглядит как снимок экрана, текст, График, линия

Автоматически созданное описаниеИзображение выглядит как снимок экрана, текст, линия, График

Автоматически созданное описание

1. История аэропорта // Веб-сайт аэропорта Пулково. ─ URL: <https://pulkovoairport.ru/about/about_pulkovo/history/> (дата обращения: 07.03.2023). [↑](#footnote-ref-1)
2. Оператор аэропорта // Веб-сайт аэропорта Пулково. ─ URL: <https://pulkovoairport.ru/about/about_pulkovo/operator/> (дата обращения: 07.03.2023). [↑](#footnote-ref-2)
3. Деньги из воздуха. Аэропорт «Пулково» расширит центральный терминал // Веб-сайт газеты Ведомости Санкт-Петербург. ─ URL: <https://spbvedomosti.ru/news/financy/dengi_iz_vozdukha_aeroport_pulkovo_rasshirit_tsentralnyy_terminal/> (дата обращения: 07.03.2023). [↑](#footnote-ref-3)
4. Always be prepared: The importance of forecasting during winter // Веб-сайт интернет-газеты International Airport Review. ─ URL: <https://www.internationalairportreview.com/article/76669/always-be-prepared-winter-weather/> (дата обращения: 09.04.2023). [↑](#footnote-ref-4)
5. Using data to enhance forecast and demand planning for airports // Веб-сайт аналитического агентства авиационных данных Cirium. ─ URL: <https://www.cirium.com/thoughtcloud/data-enhance-forecast-and-demand-planning-airports/> (дата обращения: 09.04.2023). [↑](#footnote-ref-5)
6. The Importance of Accurate Forecasting for Aviation // Веб-сайт аналитического агентства DTN. ─ URL: <https://www.dtn.com/the-importance-of-accurate-forecasting-for-aviation/> (дата обращения: 09.04.2023). [↑](#footnote-ref-6)
7. Методология CRISP-DM // Веб-сайт компании Loginom. ─ URL:<https://wiki.loginom.ru/articles/crisp-dm.html> (дата обращения: 05.05.2023). [↑](#footnote-ref-7)
8. Пулково встретил самый большой пассажирский самолет Airbus A380 // Веб-сайт аэропорта Пулково. ─ URL: [https://pulkovoairport.ru/about/press\_center/news/9657/#:~:text](https://pulkovoairport.ru/about/press_center/news/9657/" \l ":~:text=%D0%90%D1%8D%D1%80%D0%BE%D0%BF%D0%BE%D1%80%D1%82%20%D0%9F%D1%83%D0%BB%D0%BA%D0%BE%D0%B2%D0%BE%2025%20%D0%BE%D0%BA%D1%82%D1%8F%D0%B1%D1%80%D1%8F%202018,Airbus%20A380%2D800%20%D0%B0%D0%B2%D0%B8%D0%B0%D0%BA%D0%BE%D0%BC%D0%BF%D0%B0%D0%BD%D0%B8%D0%B8%20%D0%AD%D0%BC%D0%B8%D1%80%D0%B5%D0%B9%D1%82%D1%81) (дата обращения: 09.04.2023). [↑](#footnote-ref-8)
9. Exploring the LSTM Neural Network Model for Time Series // Электронный портал, посвященный анализу данных Towards Data Science. ─ URL: <https://towardsdatascience.com/exploring-the-lstm-neural-network-model-for-time-series-8b7685aa8cf> (дата обращения: 09.04.2023). [↑](#footnote-ref-9)
10. Machine learning, explained // Веб-сайт MIT Sloan School of Management. ─ URL: <https://mitsloan.mit.edu/ideas-made-to-matter/machine-learning-explained> (дата обращения: 11.04.2023). [↑](#footnote-ref-10)
11. What is machine learning? // Веб-сайт компании IBM. ─ URL: <https://www.ibm.com/topics/machine-learning> (дата обращения: 11.04.2023). [↑](#footnote-ref-11)
12. Метод главных компонент // Веб-сайт аналитического агентства АСУ-Аналитика. ─ URL: <https://asu-analitika.ru/metod-glavnyh-komponent> (дата обращения: 13.04.2023). [↑](#footnote-ref-12)
13. Time Series Forecasting: Statistical Model to Predict Future Values // Веб-сайт портала Code Algorithms Pvt. Ltd. ─ URL: <https://www.enjoyalgorithms.com/blog/time-series-forecasting-using-machine-learning> (дата обращения: 13.04.2023). [↑](#footnote-ref-13)
14. Чем машинное обучение отличается от статистики // Веб-сайт ИТ-портала Хабр. ─ URL: <https://habr.com/ru/companies/skillfactory/articles/570736/> (дата обращения: 13.04.2023). [↑](#footnote-ref-14)
15. The Actual Difference Between Statistics and Machine Learning // Электронный портал, посвященный анализу данных Towards Data Science. ─ URL: <https://towardsdatascience.com/the-actual-difference-between-statistics-and-machine-learning-64b49f07ea3> (дата обращения: 16.04.2023). [↑](#footnote-ref-15)
16. Cerqueira V. Machine Learning vs Statistical Methods for Time Series Forecasting: Size Matters / Vitor Cerqueira, Luis Torgo, Carlos Soares // Электронное издание ArXiv. ─ 2019. ─ Vol. abs/1909.13316. ─ URL: <https://arxiv.org/pdf/1909.13316.pdf> (дата обращения: 16.04.2023). [↑](#footnote-ref-16)
17. Makridakis S. Statistical and Machine Learning forecasting methods: Concerns and ways forward / Makridakis S, Spiliotis E, Assimakopoulos V // Электронное издание PLoS ONE. ─ 2018. ─ Vol. 13, Issue 3. ─ DOI: 10.1371/journal.pone.0194889. ─ URL: <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0194889> (дата обращения: 19.04.2023). [↑](#footnote-ref-17)
18. Discussion Paper 01: Aviation Demand Forecasting // Сайт правительства Великобритании. ─ UK Government Aviation Commission, 2013. ─ URL: <https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/73143/aviation-demand-forecasting.pdf> (дата обращения: 19.04.2023). [↑](#footnote-ref-18)
19. Robin W. Long-term and Short-term Forecasting Techniques for Regional Airport Planning : Master’s Thesis in Mathematical Statistics // KTH Royal institute of technology, 2016. ─ URL: <https://www.diva-portal.org/smash/get/diva2:954470/FULLTEXT02> (дата обращения: 26.04.2023). [↑](#footnote-ref-19)
20. A. Parmezan, Evaluation of statistical and machine learning models for time series prediction: Identifying the state-of-the-art and the best conditions for the use of each model / Antonio Parmezan, Vinicius Souza, Gustavo Batista, // Information Sciences Journal. ─ 2019. ─ Vol. 484. ─ P. 302-337. ─ DOI: 10.1016/j.ins.2019.01.076. ─ URL: [https://www.researchgate.net/publication/330742498\_Evaluation\_of\_statistical](https://www.researchgate.net/publication/330742498_Evaluation_of_statistical_and_machine_learning) (дата обращения: 26.04.2023). [↑](#footnote-ref-20)
21. Shuojiang Xu, Forecasting the demand of the aviation industry using hybrid time series SARIMA-SVR approach / Shuojiang Xu, Hing Kai Chan, Tiantian Zhang // Transportation Research Part E: Logistics and Transportation Review ─ 2019. ─ Vol. 122. ─ P. 169-180. ─ DOI: 10.1016/j.tre.2018.12.005. ─ URL: [https://www.researchgate.net/publication/329625988\_Forecasting\_the\_demand](https://www.researchgate.net/publication/329625988_Forecasting_the_demand_of_the_aviation_industry) (дата обращения: 26.04.2023). [↑](#footnote-ref-21)
22. Liang Guo, The hybrid PROPHET-SVR approach for forecasting product time series demand with seasonality / Liang Guo, Weiguo Fang, Qiuhong Zhao, Xu Wang // Computers & Industrial Engineering ─ 2021. ─ Vol. 161. ─ DOI: 10.1016/j.cie.2021.107598. ─ URL: <https://www.sciencedirect.com/science/article/abs/pii/S0360835221005027> (дата обращения: 26.04.2023). [↑](#footnote-ref-22)
23. Cerqueira V. Machine Learning vs Statistical Methods for Time Series Forecasting: Size Matters / Vitor Cerqueira, Luis Torgo, Carlos Soares // Электронное издание ArXiv. ─ 2019. ─ Vol. abs/1909.13316. ─ URL: <https://arxiv.org/pdf/1909.13316.pdf> (дата обращения: 26.04.2023). [↑](#footnote-ref-23)
24. Makridakis S. Statistical and Machine Learning forecasting methods: Concerns and ways forward / Makridakis S, Spiliotis E, Assimakopoulos V // Электронное издание PLoS ONE. ─ 2018. ─ Vol. 13, Issue 3. ─ DOI: 10.1371/journal.pone.0194889. ─ URL: <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0194889> (дата обращения: 28.04.2023). [↑](#footnote-ref-24)
25. Forecasting time series: using statistics vs machine learning // Веб-сайт блога Data Warehousing and Data Science. ─ URL: <https://dwbi1.wordpress.com/2022/02/16/forecasting-time-series-using-statistics-vs-machine-learning/> (дата обращения: 28.04.2023). [↑](#footnote-ref-25)
26. E. M. Shishkov, A. V. Pronichev, Applying Machine Learning Methods for Power Plant Generation Time Series Forecasting // 2022 International Conference on Industrial Engineering, Applications and Manufacturing (ICIEAM). ─ 2018. ─ P. 67–71. ─ URL: <https://www.semanticscholar.org/paper/Applying-Machine-Learning-Methods-for-Power-Plant> (дата обращения: 28.04.2023). [↑](#footnote-ref-26)
27. Kiefer D., Demand Forecasting Intermittent and Lumpy Time Series: Comparing Statistical, Machine Learning and Deep Learning Methods / Kiefer, D., Grimm, F., Bauer, M., & Dinther, C.V. // Hawaii International Conference on System Sciences. ─ 2021. ─ DOI: 10.24251/HICSS.2021.172. ─ URL: [https://www.semanticscholar.org/paper/Demand-Forecasting-Intermittent](https://www.semanticscholar.org/paper/Demand-Forecasting-Intermittent-and-Lumpy-Time-and-Kiefer-Grimm/89739506fc40ac3307555b21f883220e24ba060e) (дата обращения: 30.04.2023). [↑](#footnote-ref-27)
28. Linsey Pang, Applied Machine Learning Methods for Time Series Forecasting / Linsey Pang, Wei Liu, Lingfei Wu, Kexin Xie, Stephen Guo, Raghav Chalapathy, and Musen Wen // 31st ACM International Conference on Information & Knowledge Management. ─ 2022. ─ DOI: 10.1145/3511808.3557492. ─ URL: [https://www.semanticscholar.org/paper/Applied-Machine-Learning](https://www.semanticscholar.org/paper/Applied-Machine-Learning-Methods-for-Time-Series-Pang-Liu/298f95a128033ce3346d6f509233b9a7729a8f27) (дата обращения: 30.04.2023). [↑](#footnote-ref-28)
29. Incredibly Useful Time Series Forecasting Algorithms// Веб-сайт аналитического агентства Advancing Analytics. ─ URL: <https://www.advancinganalytics.co.uk/blog/2021/06/22/10-incredibly-useful-time-series-forecasting-algorithms> (дата обращения: 30.04.2023). [↑](#footnote-ref-29)
30. Abbasi R., Short Term Load Forecasting Using XGBoost / Abbasi, R.A., Javaid, N., Ghuman, M.N.J., Khan, Z.A., Ur Rehman, S., Amanullah // Web, Artificial Intelligence and Network Applications. ─ 2018. ─ DOI: 0.1007/978-3-030-15035-8\_108. ─ URL: [https://www.researchgate.net/publication/331746834\_Short\_Term\_Load](https://www.researchgate.net/publication/331746834_Short_Term_Load_Forecasting_Using_XGBoost) (дата обращения: 30.04.2023). [↑](#footnote-ref-30)
31. Understanding LSTMs // Веб-сайт Github. ─ URL: <https://colah.github.io/posts/2015-08-Understanding-LSTMs/> (дата обращения: 30.04.2023). [↑](#footnote-ref-31)
32. Illustrated Guide to LSTM’s and GRU’s: A step by step explanation // Электронный портал, посвященный анализу данных Towards Data Science. ─ URL: <https://towardsdatascience.com/illustrated-guide-to-lstms-and-gru-s-a-step-by-step-explanation-44e9eb85bf21> (дата обращения: 03.05.2023). [↑](#footnote-ref-32)
33. Градиентый бустинг — просто о сложном // Электронный портал, посвященный анализу данных Neurohive. ─ URL: <https://neurohive.io/ru/osnovy-data-science/gradientyj-busting/> (дата обращения: 07.05.2023). [↑](#footnote-ref-33)
34. XGBoost // Веб-сайт ИТ-ресурса GeeksforGeeks. ─ URL: <https://www.geeksforgeeks.org/xgboost/> (дата обращения: 07.05.2023). [↑](#footnote-ref-34)
35. Aileen Nielsen, Practical Time Series Analysis: Prediction with Statistics and Machine Learning. ─ O'Reilly Media, Inc., 2019. ─ 504 p. ─ ISBN 9781492041658. [↑](#footnote-ref-35)
36. Сказки от TENSORFLOW и LSTM // Веб-сайт ИТ-портала Хабр. ─ URL: <https://habr.com/ru/articles/656635/> (дата обращения: 14.05.2023). [↑](#footnote-ref-36)
37. XGBoost Parameters // Веб-сайт с официальной документацией XGBoost. ─ URL: <https://xgboost.readthedocs.io/en/latest/parameter.html> (дата обращения: 14.05.2023). [↑](#footnote-ref-37)
38. Henrik Heymann, Guideline for Deployment of Machine Learning Models for Predictive Quality in Production / Henrik Heymann, Alexander D. Kies, Maik Frye, Robert H. Schmitt, Andrés Boza // Procedia CIRP. ─ 2018. ─ Vol. 107. ─ P. 815–820. ─ DOI: 10.1016/j.procir.2022.05.068. ─ URL: <https://www.sciencedirect.com/science/article/pii/S2212827122003523> (дата обращения: 18.05.2023). [↑](#footnote-ref-38)
39. The Most Successful Approaches to Leading Organizational Change // Веб-сайт издания Harvard Business Review. ─ URL: [https://hbr.org/2023/04/the-most-successful-approaches](https://hbr.org/2023/04/the-most-successful-approaches-to-leading-organizational-change) (дата обращения: 20.05.2023). [↑](#footnote-ref-39)
40. Benefits to Better Financial Forecasting // Веб-сайт FPAExperts. ─ <https://www.fpaexperts.com/studies-show-increasing-financial-forecast-accuracy-will-save-your-company-millions-heres-how/> (дата обращения: 23.05.2023). [↑](#footnote-ref-40)
41. URL: [https://www.audit-it.ru/buh\_otchet/7703590927](https://www.audit-it.ru/buh_otchet/7703590927_ooo-vozdushnye-vorota-severnoy-stolitsy) [↑](#footnote-ref-41)
42. Airport Economics Report, 2019 // Веб-сайт Airports Council International. ─ URL: <https://store.aci.aero/product/airport-economics-report-2019/> (дата обращения: 20.05.2023). [↑](#footnote-ref-42)