Анализ поездок на ж/д транспорте по России

Соломатин Макар
СПб ПУ
Санкт-Петербург, Россия
topdoggyp@gmail.com

Дамаскинский Константин
СПб ПУ
Санкт-Петербург, Россия
damaskinsk@mail.ru

Аннотация—Данный отчёт посвящён проекту о сборе и анализе данных о железнодорожных поездках по России. В ходе реализации проекта был произведён сбор и анализ данных о продажах билетов на поезда дальнего следования. Предложена система горизонтального масштабирования сбора и анализа данных. Выведены некоторые закономерности программы динамического ценообразования.

Ключевые слова – большие данные, поезд дальнего следования, динамическое ценообразование, РЖД

I. Введение

В настоящее время в России основным пассажирским перевозчиком в железнодорожном сообщении на дальние расстояния (более 200 км) является ПАО “РЖД”. Помимо государственной “РЖД” существует всего два частных перевозчика: ООО “Тверской экспресс”, владеющий курсирующим между двумя столицами “Мегаполисом”, и АО ТК “Гранд Сервис Экспресс”, которое управляет поездами “Таврия”, курсирующими из разных городов России в Крым, и поездом “Гранд Экспресс”, выполняющим рейсы из Петербурга в Москву. При этом информация о всех поездах, следующих по территории РФ, как принадлежащих РЖД, так и нет, размещена на сайте https://rzd.ru.

АО “РЖД” устанавливает цены на билеты с использованием программы динамического ценообразования. Это означает, что цена на билет постоянно меняется, начиная с дня открытия продаж (за 90 дней до отправления поезда) и заканчивая днём отправления поезда. Цена билета зависит от множества факторов: времени до отправления поезда, количества свободных мест, скорости покупки билетов, дня недели, близости праздников, времени года и многих других факторов.

Согласно статистике РЖД [1], за период с января по ноябрь 2021 года РЖД перевезено 85,3 млн человек в дальнем следовании, что составляет в среднем 7.8 млн человек в месяц или 258 тысяч человек в день.

Совокупность указанных фактов позволяет сделать вывод о том, что анализ ценовой политики “РЖД” является очень востребованным и будет полезен для огромного количества путешественников.

[1] https://company.rzd.ru/ru/9377

II. Цели и задачи проекта

A. Цели проекта

Данная работа, как учебный проект, в первую очередь преследует целью получение опыта в анализе больших объёмов данных. Производственные цели проекта – это, во-первых, исследование политики динамического ценообразования РЖД, и, во-вторых, исследование потребительских трендов в сфере ж/д перевозок.

B. Задачи проекта

Для достижения поставленных целей необходимо решить задачи:

- Получение доступа к источнику данных, отбор наиболее интересной для исследования информации
- Разработка системы горизонтального масштабирования сбора данных
- Анализ сбора данных
- Создание дистрибутива

C. Аналитические задачи

В рамках данного проекта необходимо проанализировать программу динамического ценообразования РЖД. В текущей версии проекта исследуется зависимость стоимости билетов на поезд от количества дней до отправления поезда и зависимость количества свободных мест от количества дней до отправления поезда. Анализ первой и второй зависимости в совокупности позволяет судить о том, как меняется стоимость билета на поезд в зависимости от потребительского спроса. Анализ второй зависимости позволяет понять, в какие дни пассажиры наиболее активно скупают билеты на поезд, а также как РЖД решает проблему дефицита мест: есть ли смысл ждать добавления дополнительного вагона, или же надо сразу покупать билет на более дорогой поезд либо в более дорогой класс вагона.

Наиболее интересными представляются зависимости стоимости от количества дней до отправления, в которых отношение максимальной и минимальной цены за весь период продаж наибольшее, поскольку именно в таких ситуациях наше исследование является наиболее востребованным. На данных графиках мы будем обращать внимание на дни, в которые стоимость билетов

максимальна и минимальна, и на характер изменения цены (в какие периоды растёт, в какие падает). Поскольку работа делается в канун нового года – периода крайне высокой пассажирской активности – обозначенные исследования будут проведены для дат отправления в непосредственной близости праздников (27-30 декабря), а также задолго до праздников и сразу после них. Такая совокупность исследований также позволит выбрать наиболее удачные дни непосредственно для совершения поездки.

III. Обзор и исследование аналогичных решений

Для исследования программы динамического ценообразования в первую очередь следует обратиться на сайт РЖД. Корпорация описывает только общие черты программы: "Система в реальном времени анализирует сотни различных факторов и в зависимости от полученных данных периодически производит перерасчет стоимости проезда по всему маршруту следования. В результате цена билета может измениться в течение часа и даже нескольких минут" [2]. Кроме того, РЖД демонстрирует ценообразование для пяти популярных маршрутов, связывающих Москву с Санкт-Петербургом, Адлером, Саратовом, Белгородом и Самарой. На приведённых диаграммах показано, за сколько дней до отправления лучше всего покупать билеты, как меняется средняя стоимость в зависимости от дня недели, а также в какое время года поездку можно совершить наиболее выгодно [3]. Из графиков видно, что наиболее выгодным сезоном для поездки является зима, а наиболее дорогим – лето. Также можно наблюдать, что цены значительно выше в выходные дни, а билеты лучше всего покупать заранее. При этом для южного направления понятие "заранее" является гораздо более жёстким, чем для остальных: покупая билет за целых 60 дней до отправления поезда, пассажир переплачивает 22% по сравнению со стоимостью за 90 дней до отправления, в то время как для остальных маршрутов существенно динамичным периодом ценообразования является последний месяц продажи билетов.

Альтернативных хоть сколько-нибудь строгих исследований найдено не было. Так, журнал "Тинькофф" [4] даёт лишь общие рекомендации по покупке билетов. Авторы статьи советуют следить за календарём тарифов, который отображает процент повышения цены в каждом месяце, и выбирать наименее популярное время и дату отправления для дополнительной экономии денег. Другие найденные статьи советуют производить аналогичные манипуляции при выборе билета.

[2]https://www.rzd.ru/ru/9330

[3]https://www.rzd.ru/api/media/resources/1572962

[4]https://journal.tinkoff.ru/list/train-economy-tips/

IV. Описание модели данных и характеристикаи датасета

A. Характеристики данных

Для сбора исследуемых данных авторы используют http-запросы к сайту РЖД. Собираются данные о поездах, следующих между всеми парами из 100 наиболее населённых городов, то есть запросы о 10000 парах городов. Важно заметить, что далеко не все упомянутые пары городов связаны железной дорогой, поэтому в итоге мы получаем значительно меньше, чем 10000 пар городов. Данные собираются каждый день, поскольку необходимо отслеживать динамику изменения цен на билеты. РЖД присылает данные о поездах в формате JSON. На данный момент собрано 10 гигабайт данных.

B. Особенности процесса сбора данных

1) API РЖД: Первая трудность, с которой столкнулись авторы, это полное отсутствие документации по API РЖД. Нами был найден проект энтузиаста, который создал API на PHP и описал последовательность запросов, которую необходимо совершить для получения требуемых данных [5].

Было установлено, что для получения информации о поездах между указанными парами городов необходимо совершить двухэтапный запрос. На первом этапе отправляются данные о дате отправления и кодах АСУ "Экспресс-3" станций отправления и прибытия. В ответ на данный запрос РЖД присылает уникальный идентификатор запроса RID, по которому можно один раз в течение некоторого периода времени совершить второй запрос и получить запрошенные данные. В результате экспериментов было выяснено, что для корректной работы данной схемы между запросами требуется выдержать "магический интервал" в 3 секунды. Кроме того, периодически сайт РЖД в ответ на второй запрос присылает новый RID вместо запрошенных данных.

В ходе сбора данных авторы выяснили, что РЖД отслеижвает число одновременных запросов с одного IP-адреса. В результате ряда экспериментов был сделан вывод о том, что после 15 одновременных запросов с одного IP адреса сайт РЖД отказывает в выполнении последующих запросов приблизительно на 10 минут. Максимально допустимое количество одновременных запросов меняется в зависимости от нагрузки на сайт и может составлять от 10 до 30 запросов.

Получив данные сведения, авторы приняли решение использовать proxy-сервера для увеличения количества возможных одновременных запросов. В текущей версии проекта нами используется 25 прокси-серверов, что позволяет исполнять в среднем 375 одновременных запросов.

[5]https://github.com/visavi/rzd-api

2) Ограничения на доступ к данным: Список дат, на которые можно получить информацию о продажах билетов, ограничен. Так, нельзя получить сведения о поездах, которые отправились до даты сбора, либо позднее, чем 90 дней после неё, то есть информацию о поезде можно получить непосредственно в период продажи билетов на него. Из-за этого размер датасета, который могут собрать авторы, значительно ограничен, и его необходимо пополнять каждый день.

3) Неконсистентность данных: Часто оказывается, что название станции не совпадает с названием города. Поскольку исследование касается путешествий именно между городами, а не между станциями, необходимо сопоставлять название города и всех станций на его территории, с которых отправляются поезда. В решении этой проблемы помогает специальный сервис РЖД [6].

Неконсистентность данных также встречается, когда РЖД изменяет конечные станции следования поездов. Например, в 2021 году в Москве открылся новый вокзал "Восточный", и 12 декабря поезд 071В поменял начальную станцию с Москва-Курская на Москва-Восточная. Данный поезд идёт через Тулу и Курск, и в результате изменения расписания оказалось, что в один и тот же день между этими городами следует два поезда с одним номером, хотя предполагается, что один поезд идёт не чаще, чем раз в день.

C. Этапы сбора данных

Сбор данных начинается с получения списка интересующих городов. Для этого была взята статья на Википедии [7] , содержащая таблицу российских городов и их населения. Далее необходимо получить коды АСУ Экспресс-3 для станций в интересующих городах.

Затем ежедневно для каждой пары городов и для всех дат в 90-дневный период, начиная с текущего дня, необходимо сделать запросы списка поездов с билетами. Полученные данные сохраняются в JSON-формате, затем выделяются необходимые данные и отправляются в базу данных MongoDB.

V. Предлагаемое программное решение

Программные модули проекта реализованы на языке Python версии 3.10. Проект состоит из нескольких пакетов.

1) Пакет collect собирает данные по всем интересующим поездам на ближайшие 90 дней
2) Пакет export отправляет собранные данные в базу данных MongoDB
3) Пакет database совершает запросы к базе данных и локальному хранилищу. В пакете реализован общий интерфейс, перегрузив который можно при необходимости добавлять другие хранилища данных о билетах
4) Пакет process производит аналитику данных в двух режимах:
   - Локально с использованием многопоточности (используется количество потоков, равное количеству ядер процессора)
   - С подключением к кластеру Apache Spark (можно использовать любое количество worker-ов)

[6]https://pass.rzd.ru/suggester

[7]Кликабельная ссылка на статью

VI. Описание реализации и технологий

A. Стек технологий

Сбор данных производится с использованием Python 3.10 и брокера сообщений Rabbit MQ, в который посылаются задачи по сбору данных. Rabbit MQ необходим, когда сбор данных осуществляется не только с использованием нескольких proxy-серверов, но и на разных физических машинах (когда пропускной способности сетевой карты одного компьютера перестаёт хватать). Обработка данных также производится на Python 3.10 с использованием многопоточности или фреймворка Apache Spark. В качестве системы контроля версий используется git, удалённый репозиторий размещён на github https://github.com/slmtnm/rzd-analysis/. Система управления базами данных – MongoDB.

B. Горизонтальное масштабирование

Данный проект поддерживает горизонтальное масштабирование как при сборе, так и при обработке данных. Сбор данных можно осуществлять на нескольких физических машинах с использованием брокера сообщений RabbitMQ. При обработке данных горизонтальное масштабирование осуществляется посредством увеличения числа Spark worker-ов.

C. Состав дистрибутива

Дистрибутив представляет из себя набора Python пакетов. Для каждого пакета существует свой Dockerfile. Также в дистрибутив входит файл со списком кодов и названий станций, файл с настройками используемых proxy-сервеов (для каждого пользователя свой).

Для развёртывания всей инфраструктуры в локальных контейнерах используется Docker compose. Для развёртывания всего проекта используется Makefile.

Для получения дистрибутива необходимо клонирорвать репозиторий https://github.com/slmtnm/rzd-analysis/. Запуск и остановка осуществляются посредством команды make.

D. Инфраструктура разработки

Для организации процесса распределённого сбора и обработки данных используется Яндекс.Облако. Зайдействованы 3 виртуальные машины, содержащие 2 ядра Intel Cascad Lake с гарантированной долей CPU 10%, 4 Гб оперативной памяти и жёсткий диск на 20 Гб. Данные сохраняются на Яндекс.Диск. Для улучшения качества кода используется средство непрерывной

интеграции Github Actions CI. В CI-конвейер входят линтеры кода, также код частично покрыт тестами. Кроме этого, CI-конвейер собирает docker образы.

VII. Особенности программного решения

Данное программное решение применяет горизонтальное и вертикальное масштабирование как сбора, так и обработки данных. Также предоставлена возможность запускать модуль аналитики как из JSON-файлов, находящихся в локальном хранилище, так и из базы данных MongoDB.

Очень важной особенностью с точки зрения исследовательской ценности данного проекта является то, что данный проект собирает невоспроизводимые данные.

VIII. Анализ данных

В рамках данного проекта реализуется построение зависимости стоимости билета от времени до отправления. Для этого строится зависимость для каждой доступной даты отправления поезда для всех пар городов.

Приведём графики для маршрутов с наиболее сильной динамикой изменения цены на 9, 12, 27 и 30 декабря 2021 года и 2 января 2022 года.

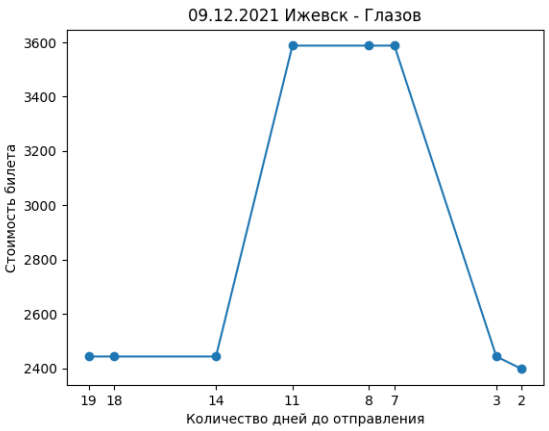

Рис. 1. Поезд с наибольшей динамикой ценообразования 9 декабря 2021 года

Проанализировав данные графики, можно заметить следующее:

- В даты сильно повышенного спроса цены повышаются значительно раньше и сильнее, чем в даты с небольшим спросом
- Стоимость билета на поезд не обязана повышаться даже задолго до даты отправления. Можно заметить, что билеты на поезд, отправившийся 9 декабря, дорожали за 2 недели до отправления, а вот на поезд, отправившийся 12 декабря, наоборот подешевели
- За несколько дней до отправления (обычно за три дня), когда спрос спадает практически до нуля, а

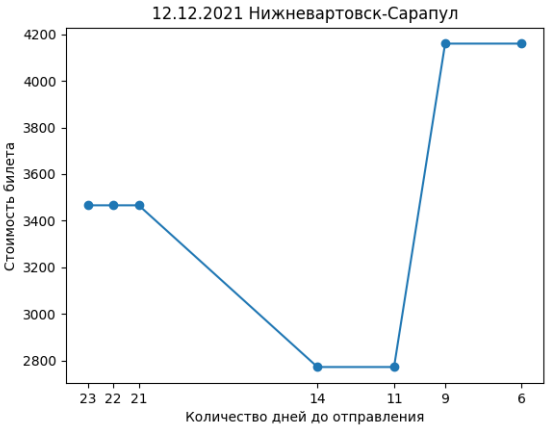

Рис. 2. Поезд с наибольшей динамикой ценообразования 12 декабря 2021 года

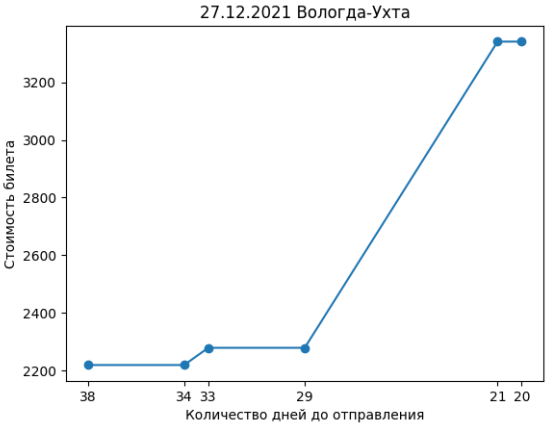

Рис. 3. Поезд с наибольшей динамикой ценообразования 27 декабря 2021 года

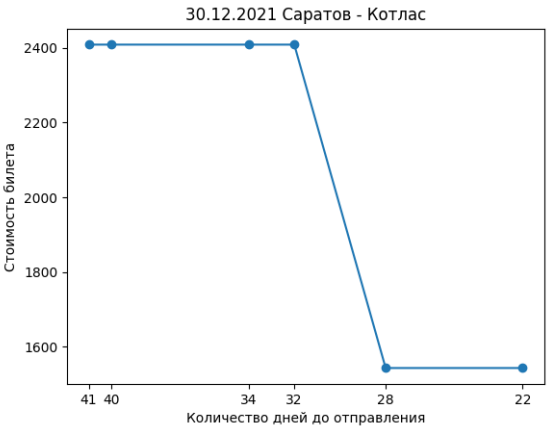

Рис. 4. Поезд с наибольшей динамикой ценообразования 30 декабря 2021 года

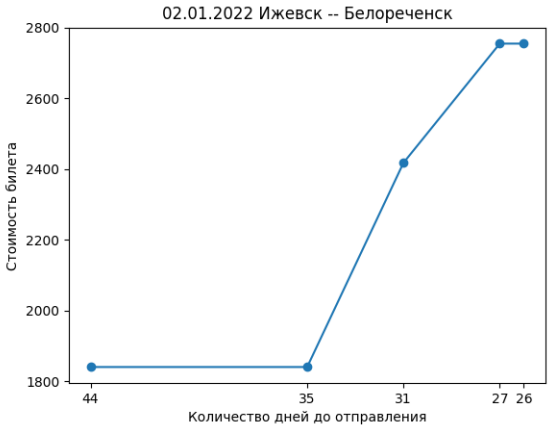

Рис. 5. Поезд с наибольшей динамикой ценообразования 30 декабря 2021 года

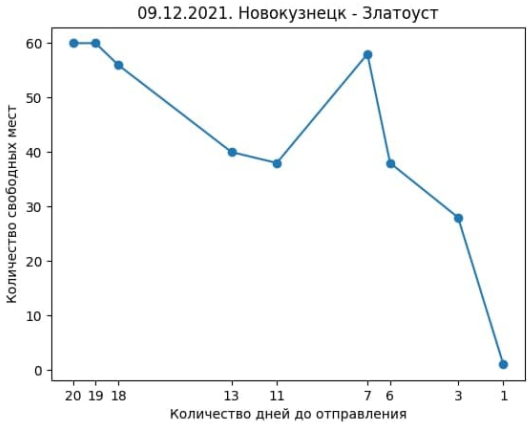

Рис. 7. Поезд с наибольшей динамикой раскупаемости 9 декабря 2021 года

люди начинают сдавать билеты, эти самые билеты дешевеют. Данную особенность можно заметить, сопоставив приведённые выше графики с графиками зависимости количества свободных мест от числа дней до отправления.

Рассмотрим теперь графики зависимости числа свободных мест от количества дней до отправления.

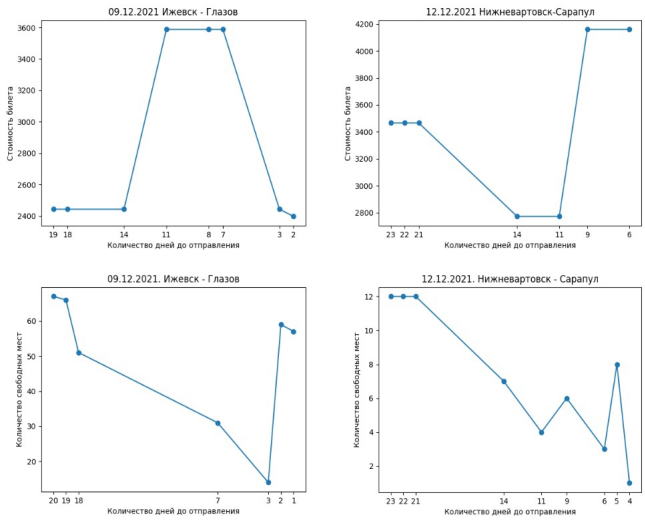

Рис. 6. Прямой связи между количеством свободных мест и стоимости билетов не наблюдается: в одном случае при высвобождении мест билеты подешевели, в другой ситуации при изменении спроса цена не менялась

- В целом спрос на билеты либо линеен, то есть не заметны периоды наибольшей и наименьшей раскупаемости, либо же спрос лавинообразный, то есть билеты раскупают сразу же, как только появляются места. Такая динамика наиболее характерна для дат ажиотажного спроса, как в период праздников

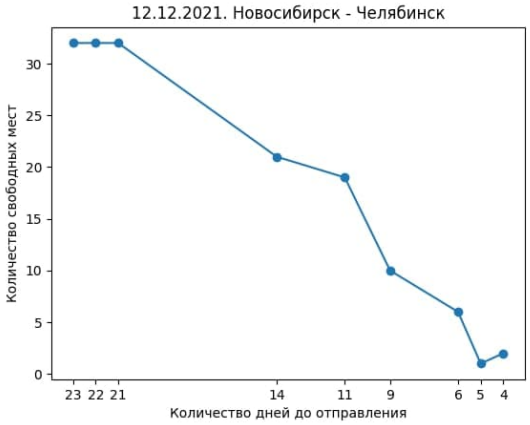

Рис. 8. Поезд с наибольшей динамикой раскупаемости 12 декабря 2021 года

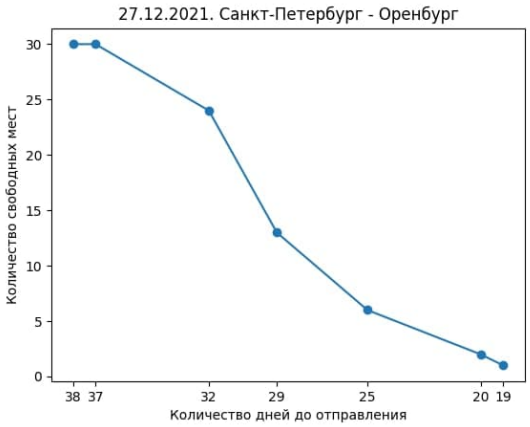

Рис. 9. Поезд с наибольшей динамикой раскупаемости 27 декабря 2021 года

- Зачастую за несколько дней до конца продаж би-

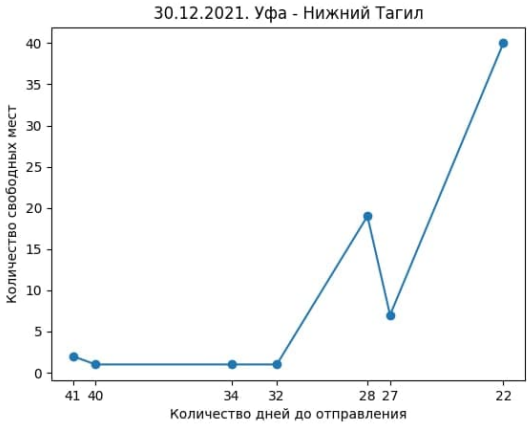

Рис. 10. Поезд с наибольшей динамикой раскупаемости 30 декабря 2021 года

летов люди начинают сдавать билеты, что даёт некоторый шанс на покупку билета на дефицитный поезд перед отправлением

- При высоком спросе в ажиотажные даты РЖД добавляет дополнительные вагоны к поездам, причём иногда даже в несколько этапов. Это означает, что если в какой-то момент билетов на более дешёвый поезд или класс вагона нет, то не стоит сразу же покупать более дорогие билеты, а есть смысл подождать, пока РЖД добавит дополнительные вагоны

Список литературы

[1] Сайт ПАО “РЖД” https://rzd.ru/.
[2] Сайт RabbitMQ https://www.rabbitmq.com/.
[3] Документация docker https://docs.docker.com/.
[4] Документация Python https://docs.python.org/3/index.html.
[5] Документация Apache Spark https://spark.apache.org/docs/3.2.0/.