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.

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", совершающим 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 (https://company.rzd.ru/ru/9377), 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.

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 отправ-

ления 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”. Кроме того, РЖД демонстрирует ценообразование для пяти популярных маршрутов, связывающих Москву с Санкт-Петербургом, Адлером, 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. Из приведённых 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 “Тинькоффа” 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.

IV. Описание модели данных и характеристикаи датасета

A. Характеристики данных

Для сбора исследуемых данных авторы используют web-запросы к сайту РЖД. Собираются данные о поездах, следующих между всеми парами из топ 100 городов по численности населения, то есть запросы о 10000 парах городах. 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 гигабайт данных.

B. Особенности процесса сбора данных

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.

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.

В ходе сбора данных авторы выяснили, что РЖД отслеижвает число одновременных запросов с одного IP-адреса. В результате ряда экспериментов был сделанвывод 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.

Кроме того, фундаментальным ограничением является список дат, в течение которых можно получить информацию о продажах билетов. Так, нельзя получить сведения о поездах, которые отправились до даты сборы, либо позднее, чем 90 дней после неё, то есть информацию о поезде можно получить непосредственно в период продажи билетов на него. Из-за этого размеры датасета, который могут собрать авторы, значительно ограничен, и его необходимо пополнять каждый день.

C. Units

- Use either SI (MKS) or CGS as primary units. (SI units are encouraged.) English units may be used as secondary units (in parentheses). An exception would be the use of English units as identifiers in trade, such as “3.5-inch disk drive”.

- Avoid combining SI and CGS units, such as current in amperes and magnetic field in oersteds. This often leads to confusion because equations do not balance dimensionally. If you must use mixed units, clearly state the units for each quantity that you use in an equation.
- Do not mix complete spellings and abbreviations of units: "Wb/m^2" or "webers per square meter", not "webers/m^2". Spell out units when they appear in text: ". . . a few henries", not ". . . a few H".
- Use a zero before decimal points: "0.25", not ".25". Use "cm^3", not "cc".)

D. Equations

Number equations consecutively. To make your equations more compact, you may use the solidus (/), the exp function, or appropriate exponents. Italicize Roman symbols for quantities and variables, but not Greek symbols. Use a long dash rather than a hyphen for a minus sign. Punctuate equations with commas or periods when they are part of a sentence, as in:

$$a + b = \gamma \tag{1}$$

Be sure that the symbols in your equation have been defined before or immediately following the equation. Use "(1)", not "Eq. (1)" or "equation (1)", except at the beginning of a sentence: "Equation (1) is . . .'.

E. LaTeX-Specific Advice

Please use "soft" (e.g., \eqref{Eq}) cross references instead of "hard" references (e.g., (1)). That will make it possible to combine sections, add equations, or change the order of figures or citations without having to go through the file line by line.

Please don't use the {eqnarray} equation environment. Use {align} or {IEEEeqnarray} instead. The {eqnarray} environment leaves unsightly spaces around relation symbols.

Please note that the {subequations} environment in LaTeX will increment the main equation counter even when there are no equation numbers displayed. If you forget that, you might write an article in which the equation numbers skip from (17) to (20), causing the copy editors to wonder if you've discovered a new method of counting.

BibTeX does not work by magic. It doesn't get the bibliographic data from thin air but from .bib files. If you use BibTeX to produce a bibliography you must send the .bib files.

LaTeX can't read your mind. If you assign the same label to a subsubsection and a table, you might find that Table I has been cross referenced as Table IV-B3.

LaTeX does not have precognitive abilities. If you put a \label command before the command that updates the counter it's supposed to be using, the label will pick up the last counter to be cross referenced instead. In particular, a \label command should not go before the caption of a figure or a table.

Do not use \nonumber inside the {array} environment. It will not stop equation numbers inside {array} (there won't be any anyway) and it might stop a wanted equation number in the surrounding equation.

F. Some Common Mistakes

- The word "data" is plural, not singular.
- The subscript for the permeability of vacuum μ_0, and other common scientific constants, is zero with subscript formatting, not a lowercase letter "o".
- In American English, commas, semicolons, periods, question and exclamation marks are located within quotation marks only when a complete thought or name is cited, such as a title or full quotation. When quotation marks are used, instead of a bold or italic typeface, to highlight a word or phrase, punctuation should appear outside of the quotation marks. A parenthetical phrase or statement at the end of a sentence is punctuated outside of the closing parenthesis (like this). (A parenthetical sentence is punctuated within the parentheses.)
- A graph within a graph is an "inset", not an "insert". The word alternatively is preferred to the word "alternately" (unless you really mean something that alternates).
- Do not use the word "essentially" to mean "approximately" or "effectively".
- In your paper title, if the words "that uses" can accurately replace the word "using", capitalize the "u"; if not, keep using lower-cased.
- Be aware of the different meanings of the homophones "affect" and "effect", "complement" and "compliment", "discreet" and "discrete", "principal" and "principle".
- Do not confuse "imply" and "infer".
- The prefix "non" is not a word; it should be joined to the word it modifies, usually without a hyphen.
- There is no period after the "et" in the Latin abbreviation "et al.".
- The abbreviation "i.e." means "that is", and the abbreviation "e.g." means "for example".

An excellent style manual for science writers is [7].

G. Authors and Affiliations

The class file is designed for, but not limited to, six authors. A minimum of one author is required for all conference articles. Author names should be listed starting from left to right and then moving down to the next line. This is the author sequence that will be used in future citations and by indexing services. Names should not be listed in columns nor group by affiliation. Please keep your affiliations as succinct as possible (for example, do not differentiate among departments of the same organization).

H. Identify the Headings

Headings, or heads, are organizational devices that guide the reader through your paper. There are two types: component heads and text heads.

Component heads identify the different components of your paper and are not topically subordinate to each other. Examples include Acknowledgments and References and, for these, the correct style to use is "Heading 5". Use "figure caption" for your Figure captions, and "table head" for your table title. Run-in heads, such as "Abstract", will require you to apply a style (in this case, italic) in addition to the style provided by the drop down menu to differentiate the head from the text.

Text heads organize the topics on a relational, hierarchical basis. For example, the paper title is the primary text head because all subsequent material relates and elaborates on this one topic. If there are two or more sub-topics, the next level head (uppercase Roman numerals) should be used and, conversely, if there are not at least two sub-topics, then no subheads should be introduced.

I. Figures and Tables

a) Positioning Figures and Tables: Place figures and tables at the top and bottom of columns. Avoid placing them in the middle of columns. Large figures and tables may span across both columns. Figure captions should be below the figures; table heads should appear above the tables. Insert figures and tables after they are cited in the text. Use the abbreviation "Fig. ??", even at the beginning of a sentence.

Таблица I
Table Type Styles

Table Head	Table Column Head		
	Table column subhead	Subhead	Subhead
copy	More table copy[a]		

[a]Sample of a Table footnote.

Figure Labels: Use 8 point Times New Roman for Figure labels. Use words rather than symbols or abbreviations when writing Figure axis labels to avoid confusing the reader. As an example, write the quantity "Magnetization", or "Magnetization, M", not just "M". If including units in the label, present them within parentheses. Do not label axes only with units. In the example, write "Magnetization (A/m)" or "Magnetization {A[m(1)]}", not just "A/m". Do not label axes with a ratio of quantities and units. For example, write "Temperature (K)", not "Temperature/K".

Acknowledgment

The preferred spelling of the word "acknowledgment" in America is without an "e" after the "g". Avoid the stilted expression "one of us (R. B. G.) thanks ...". Instead, try "R. B. G. thanks...". Put sponsor acknowledgments in the unnumbered footnote on the first page.

References

Please number citations consecutively within brackets [1]. The sentence punctuation follows the bracket [2]. Refer simply to the reference number, as in [3]—do not use "Ref. [3]" or "reference [3]" except at the beginning of a sentence: "Reference [3] was the first ..."

Number footnotes separately in superscripts. Place the actual footnote at the bottom of the column in which it was cited. Do not put footnotes in the abstract or reference list. Use letters for table footnotes.

Unless there are six authors or more give all authors' names; do not use "et al.". Papers that have not been published, even if they have been submitted for publication, should be cited as "unpublished" [4]. Papers that have been accepted for publication should be cited as "in press" [5]. Capitalize only the first word in a paper title, except for proper nouns and element symbols.

For papers published in translation journals, please give the English citation first, followed by the original foreign-language citation [6].

Список литературы

[1] G. Eason, B. Noble, and I. N. Sneddon, "On certain integrals of Lipschitz-Hankel type involving products of Bessel functions," Phil. Trans. Roy. Soc. London, vol. A247, pp. 529–551, April 1955.

[2] J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

[3] I. S. Jacobs and C. P. Bean, "Fine particles, thin films and exchange anisotropy," in Magnetism, vol. III, G. T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271–350.

[4] K. Elissa, "Title of paper if known," unpublished.

[5] R. Nicole, "Title of paper with only first word capitalized," J. Name Stand. Abbrev., in press.

[6] Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, "Electron spectroscopy studies on magneto-optical media and plastic substrate interface," IEEE Transl. J. Magn. Japan, vol. 2, pp. 740–741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].

[7] M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.