FACULTY
OF INFORMATION
TECHNOLOGY
CTU IN PRAGUE

Assignment of bachelor's thesis

Title:	Analysis of Ticket Sales Data and Development of a New Methodology for Data Collection and Analysis
Student:	Šimon Marinov
Supervisor:	Ing. Mgr. Ladislava Smítková Janků, Ph.D.
Study program:	Informatics
Branch / specialization:	Knowledge Engineering
Department:	Department of Applied Mathematics
Validity:	until the end of summer semester 2023/2024

Instructions

Analyze real ticket sales data from a commercial partner. Based on the analysis, propose an extension to the data collection system or a new data collection methodology. Using synthetic data, compare the characteristics of the original and the new data collection methodology.
1. Study the recommended literature and conduct a state-of-the-art survey.
2. Analyze the data set of ticket sales and process the conclusions of the analysis.
3. Design a new methodology for data collection and construct a model for artificial data generation.
4. Perform and evaluate experiments.

1. Ticket Sales DataSet (commercial, available to the student)
2. H. -I. Ahn and W. S. Spangler, "Sales Prediction with Social Media Analysis," 2014 Annual SRII Global Conference, 2014, pp. 213-222
3. Suher, J: Forecasting Event Ticket Sales, Ph.D. Thesis, Warton University
4. Y. Kaneko and K. Yada, "A Deep Learning Approach for the Prediction of Retail Store Sales," 2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW), 2016, pp. 531-537
5. Branda, F.; Marozzo, F.; Talia, D. Ticket Sales Prediction and Dynamic Pricing Strategies in Public Transport. Big Data Cogn. Comput. 2020, 4, 36.

Electronically approved by Ing. Magda Friedjungová, Ph.D. on 22 February 2023 in Prague.

Bachelor's thesis

ANALYSIS OF TICKET SALES DATA AND DEVELOPMENT OF A NEW METHODOLOGY FOR DATA COLLECTION AND ANALYSIS

Šimon Marinov

Faculty of Information Technology
Department of Applied Mathematics
Supervisor: Ing. Mgr. Ladislava Smítková Janků, Ph.D.
May 11, 2023

Czech Technical University in Prague
Faculty of Information Technology

Citation of this thesis: Marinov Šimon. *Analysis of Ticket Sales Data and Development of a New Methodology for Data Collection and Analysis*. Bachelor's thesis. Czech Technical University in Prague, Faculty of Information Technology, 2023.

Contents

List of Figures

List of Algorithms

List of Tables

Rád bych poděkoval vedoucí Ing. Magdě Friedjungové, Ph.D. za vedení této práce. Srdečně děkuji za Vaši odbornou pomoc, cenné rady a podporu, kterou jste mi poskytla během mé bakalářské práce a studia. Vaše zkušenosti a znalosti v oboru mi byly velkým přínosem.

Declaration

I hereby declare that the presented thesis is my own work and that I have cited all sources of information in accordance with the Guideline for adhering to ethical principles when elaborating an academic final thesis. I acknowledge that my thesis is subject to the rights and obligations stipulated by the Act No. 121/2000 Coll., the Copyright Act, as amended. In accordance with Section 2373(2) of Act No. 89/2012 Coll., the Civil Code, as amended, I hereby grant a non-exclusive authorization (licence) to utilize this thesis, including all computer programs that are part of it or attached to it and all documentation thereof (hereinafter collectively referred to as the ”Work”), to any and all persons who wish to use the Work. Such persons are entitled to use the Work in any manner that does not diminish the value of the Work and for any purpose (including use for profit). This authorisation is unlimited in time, territory and quantity.

In Prague on May 11, 2023 . . . .

Abstract

This thesis delves into the challenges of event-based ticket sales data, with the goal of enhancing data collection and proposing models for dynamic pricing. A literature review examines various approaches, while data analysis identifies key insights and obstacles in the dataset provided by a commercial partner.

A refined data collection strategy, incorporating diverse data sources, is proposed alongside a synthetic data generation model to test out the proposed methodology. Although significant improvements were not observed, valuable insights into the potential areas for further refinement were gained.

Proposed dynamic pricing models demonstrated good performance during low-demand periods, indicating their potential use for discounts and optimizing ticket pricing strategies. This research contributes to the field of event-based ticket sales by investigating different approaches, identifying challenges, and laying the groundwork for future advancements in effective and efficient pricing strategies in the field.

Keywords machine learning, data analysis, event-based ticket sales, demand prediction, dynamic pricing, synthetic data

Abstrakt

Tato práce se zabývá daty o prodeji lístků na akce s cílem zlepšit metodology shchromaždovaných dat a návrhem modelů pro dynamickou cenu. Přehled literatury zkoumá různé přístupy, zatímco analýza dat identifikuje klíčové poznatky a překážky v existující datesetu poskytnutým komerčním partnerem.

V práci se navrhuje strategie shromažďování dat, která zahrnuje různé zdroje dat, spolu se syntetickým modelem generování dat k testování navržené metodiky. Ačkoli nebylo dosaženo zlepšení, identifikovali se oblasti které mohou vést k zdokonalaní.

Navržené dynamické cenové modely prokázaly dobrou výkonnost v období nízké poptávky, což naznačuje jejich potenciální využití pro slevy a optimalizaci cenové strategie. Tento výzkum přispívá k oblasti prodeje lístků na akce tím, že zkoumá různé přístupy, identifikuje výzvy a pokláda základ pro budoucí pokroky v účinných a efektivních cenových strategiích v této oblasti.

Klíčová slova strojové učení, datová analýza, prodej lístků událostí, predikce poptávky, dynamická cena, syntetická data

Summary

Literature Review

In this literature review, we examined the latest developments in machine learning (ML) models and methods aimed at maximizing revenue in e-commerce, focusing on dynamic pricing and sales forecasting. We reviewed several prominent models and techniques, discussing their underlying principles, strengths, and weaknesses, as well as their effectiveness in the ticket sales domain. The remaining part provides an overview of some fundamental concepts in ML, focusing on classification, regression, and time series prediction.

Data Analysis

In this chapter, we analyzed different data sources provided by a third party. Positive findings included increased demand at the start and end of ticket sales periods and the impact of holidays on sales. However, the internal database had negative aspects such as low scanned tickets and an overestimated event capacity. Also, the chapter noted concerns about the inability to transition data from the older GA3 format to the new GA4 format and the poor performance of Google's tracking applications for data analysis and machine learning.

Enhancing data collection

This section showcased a refined data collection strategy, aimed at enhancing accuracy. We advocated for the inclusion of data from sources such as social media analytics, economic measures, meteorological data, and competitor event information, emphasizing Facebook's significant role in providing valuable insights into user engagement, event participation, and demographic data.

Synthetic Data Generation

This chapter describes the development of a synthetic data generation model to enhance the training dataset for machine learning models. We used a rolling window model with feature engineering, data preprocessing, to generate synthetic data from manually picked online ticket sales. This approach let us test some parts of the proposed new data collection methodology.

Dynamic Pricing

Experiments on price prediction using various classifiers and a two-step architecture are shown in this chapter. The models showed a stronger ability to predict price decreases during periods of lower demand, indicating their potential for optimizing ticket pricing through discounts and promotions. The comparison between data collection methodologies did not show significant improvement, suggesting the need for refining the synthetic data generation model. Overall, this chapter highlights the potential of the models for dynamic pricing while pointing to areas for future research.

list of abbreviations

ML	machine learning
API	Application Programming Interface
GA3	Universal Analytics
GA4	Google Analytics
SQL	Structured Query Language

Introduction

The surge in e-commerce has resulted in substantial reliance on data to inform business strategies. Specifically, data-driven sales forecasting has become crucial for organizations to optimize profits and expand their customer base. Although the event ticketing industry may not be the first to come to mind for ML-driven commercial applications, there is a significant opportunity to employ ML in developing methods that accurately forecast sales to minimize unsold inventory and prevent revenue loss.

The aim of this thesis is to examine a real-world ticket sales dataset provided by a commercial partner and suggest an enhancement to the existing data collection system. This proposal is informed by research published in leading e-commerce sectors, primarily the airline/transport and hotel industries, which have successfully utilized ML to maximize revenue and identify trends from collected data.

The initial phase of this thesis involves analyzing the ticket sales dataset to uncover valuable insights and discern patterns that may aid sales prediction. The findings from this analysis were employed in designing a methodology for data collection, which will ideally enhance the prospects for the creation of more precise models. To assess the efficacy of the newly developed data collection approach, a synthetic data generation process was constructed based on the original data's characteristics. Subsequently, an experiment was conducted in which the performance of models with newly collected features and without are compared.

In summary, this thesis aspires to augment the body of knowledge on sales prediction by introducing a data collection methodology aimed at enhancing the accuracy of ticket sales forecasting and laying the groundwork for ML models capable of predicting prices for specific entertainment products.

Objectives

The objectives of this thesis are organized into several key components. Firstly, we investigated and reviewed the existing literature methodologies employed in leading e-commerce sectors, specifically the airline/transport and hotel industries. By doing so, we hope to understand how ML methods have been effectively used for revenue optimization and trend identification in these sectors, and how similar techniques can be applied to the event ticketing industry.

Secondly, we aim to analyze the provided real-world ticket sales dataset to extract valuable insights and identify patterns that can be used for demand prediction models. This analysis formed the basis for our understanding of the industry-specific data and trends.

Next, we proposed an extension to the existing data collection methodology, informed by insights from the ticket sales dataset analysis and best practices in the airline/transport and hotel industries. This new approach aims to improve the accuracy and effectiveness of sales prediction models in the event ticketing domain.

Following the development of the new data collection methodology, we constructed a synthetic data generation process based on the characteristics of the original data. This process allows us to assess the efficacy of the newly developed data collection methodology in a controlled environment, without the need for additional real-world data.

Subsequently, we designed an experiment in which two models are trained: one on the synthetic data with old features and the other on newly added features. This experiment enables us to evaluate and compare the performance of the models in terms of accuracy, providing insights into the effectiveness of our proposed data collection methodology.

Ultimately, our thesis strives to establish a benchmark performance for dynamic pricing models in the entertainment industry and assess these models to recommend new data collection methodologies that could potentially enhance the accuracy of forecasts. By doing so, we hope to provide a foundation for the development and implementation of models that may be suitable for deployment and online testing in real-world scenarios.

Chapter 1

Literature Review and Fundamentals of Machine Learning

This chapter aims to provide a review of the latest developments in ML models and methods that focus on maximizing and forecasting within the broad context of e-commerce. The chapter begins by exploring the recommended literature, followed by a survey of the state-of-the-art technologies and approaches in the field. In the other half of the chapter fundamentals of ML are described This allows the reader to better understand the landscape of the thesis.

1.1 Machine Learning Models for Revenue Optimization

In recent years, various ML models have been developed and employed to maximize revenue in e-commerce. This chapter delves into the most prominent models and techniques which concentrate on dynamic pricing and sales forecasting. The focus of the chapter is on understanding their underlying principles, strengths, and weaknesses, as well as assessing their effectiveness in the ticket sales domain.

1.1.1 Sales forecasting

The motivation for forecasting ticket sales for cultural events has been a subject of interest for over a decade. One of the earliest attempts in this field can be traced back to 2008, with the publication of the paper titled *"Forecasting Event Ticket Sales"*[1]. In this work, the author analyzed ticket sales data and proposed a statistical model to predict weekly ticket sales.

While the introduction of such a model was a step forward in the industry, the paper, unfortunately, lacked a robust methodology for evaluating the performance of the proposed model. Furthermore, the dataset used in model construction consisted of a limited number of large ticket sales events, which may not be representative of the diverse range of events typically encountered in the industry.

Since the publication of this paper, the field of mathematical models has evolved considerably, with more advanced ML techniques being employed to enhance prediction accuracy. A more recent study focused on sales forecasting in the automobile industry, titled *"Sales Prediction with Social Media Analysis"*[2], demonstrated that the frequency of specific words in social media could improve the accuracy of ARIMA models, which predict monthly sales. Another paper, *"Social media marketing in the sales volume prediction for the Lolita fashion brand"*[3], analyzed the impact of social media factors such as comments, likes, and shares on sales volume. The study found a statistically significant correlation between social media factors and sales volume in fashion-specialized e-commerce websites and developed ML models based on these features, achieving respectable performance.

These studies highlight the potential of incorporating social media data sources into demand forecasting models for various industries, including event-based ticket sales. By leveraging the ML and the information available on social media platforms, researchers and developers can develop more accurate and actionable predictions, leading to improved revenue generation strategies.

1.1.2 Purchase Forecasting and Dynamic Pricing

In the previous sections, we provided an overview of models that focus on forecasting overall sales. However, another technique can be leveraged to increase revenue. That is dynamic pricing based on demand prediction. This approach involves adjusting prices in real-time to match demand, which can lead to higher revenue generation and more efficient resource allocation.

Dynamic pricing has been explored in various industries, such as hospitality and airline services. In the paper *"Customized Regression Model for Airbnb Dynamic Pricing"*[4], the authors proposed multi-step regression models to predict prices for Airbnb listings. Similarly, the study *"Dynamic Pricing for Airline Ancillaries with Customer Context"*[5] described the architecture of models designed for dynamic pricing in the airline industry, focusing on ancillary services such as baggage fees and seat upgrades.

Both papers developed evaluation metrics to assess the effectiveness of dynamic pricing models. These metrics help ensure that the models are achieving their primary objectives, such as maximizing revenue or improving customer satisfaction.

The models introduced in these papers consist of a multi-step architecture, in which the first step is a classifier that tries to predict whether a specified action is going to happen. In the context of the Airbnb paper, the classifier aimed to predict whether an accommodation would be booked for a specific day, while in the paper focusing on pricing ancillaries, the goal for the classifier was to predict whether the ancillary would be purchased. In both papers, the classification predictions were interpreted as demand estimations, which were then used in subsequent steps for price prediction.

This approach to predicting the demand based on classifiers predictions leads to a paper titled *"Ticket Sales Prediction and Dynamic Pricing Strategies in Public Transport"*[6], where the authors developed highly accurate classifiers that predicted whether users would buy or not buy a bus ticket with 95% accuracy and an F1 score. While this is a positive result, a closer examination of the work reveals potential issues with the model architecture. The underlying problem is that the model predicts based on the last two steps of the purchase process when the price of the product is already visible to the customer, which disables any possibility of using this model for dynamic pricing.

Given that this model is unfit for the architecture described in the Airbnb and airline ancillary pricing papers, it raises the question of whether a similar model that attempts to predict the purchase of products on an e-commerce website, based on customers' scrolling characteristics can be developed. If successful, this could lead to more accurate classifiers, ultimately resulting in better dynamic pricing strategies. Such an approach could prof universal across a broad spectrum of e-commerce websites and could lead to accurate dynamic pricing.

1.2 Fundamentals of Classification, Regression and Time Series Prediction

Before we delve into the analysis of real-world ticket sales data, it is essential to understand the foundations of ML, focusing on classification, regression, time series prediction, and the concept of rolling and expanding windows. This chapter provides an introduction to these concepts, which play a role in further chapters.

1.2.1 Supervised and Unsupervised Learning

ML algorithms are typically divided into two main categories: supervised and unsupervised learning. Understanding these categories is crucial for understanding ML.

1.2.1.1 Unsupervised Learning

Unsupervised learning deals with unlabeled data. The goal is not to predict an output but to explore the underlying structure and patterns of the input data. Since the data is unlabeled, the learning algorithm is left on its own to find structure in the input.[7]

Common tasks for unsupervised learning are:

Clustering: The aim is to group instances so that instances in the same cluster are more similar to each other than to instances in other clusters. A typical example might be segmenting customers into different groups based on their purchasing behaviour.

Dimensionality reduction: The goal is to simplify the input data without losing too much information. One way to do this is to merge several correlated features into one.

1.2.2 Supervised Learning

Supervised learning is the most prevalent type of ML. It involves training a model on a labelled dataset. In other words, the model learns from data that includes both the input (features) and the output (target variable).[7]

The 'supervised' aspect of this learning method comes from the idea that the learning algorithm is guided towards the right solution by providing it with correct answers at the start. Once the model is trained, it can be used to predict the output for new, unseen data based on the patterns it learned during the training phase.

Two common types of supervised learning tasks are:

Classification: The target variable consists of categories. The aim is to predict the category of unseen instances based on their features.

Regression: The target variable is continuous. The aim is to predict a numerical value for unseen instances based on their features.

1.2.2.0.1 Classification Classification is a supervised learning approach where the output is a category or class. In a classification problem, the algorithm learns from the training data and uses it to classify new observations into one of the predefined classes. Examples include email spam detection (spam or not spam) and image recognition (image categories).

1.2.2.0.2 Regression Regression, another form of supervised learning, differs from classification in the sense that the output is a continuous value, rather than a class label. Examples include predicting house prices, stock prices, or in our case, ticket sales.

1.2.2.0.3 Time Series Prediction Time series prediction is a specialized form of regression where the goal is to predict future values based on previously observed values. Time series data have a natural temporal ordering, making time series analysis distinct from cross-sectional studies in which there is no natural ordering of the observations.

Time series prediction is widely used for non-stationary data, such as predicting stock prices, weather forecasts, or ticket sales in our case. Autoregressive Integrated Moving Average (ARIMA), Long Short-Term Memory Networks (LSTM), and Facebook's Prophet are popular algorithms for time series prediction.[8]

1.2.2.1 Expanding and Rolling Windows Models

Time series forecasting often involves breaking down the data into windows of time, helping to organize and structure the data in a way that can improve the performance of a predictive model. Two common techniques are the use of expanding and rolling windows.

1.2.2.1.1 Expanding Windows An expanding window is a technique where the size of the window increases over time. This means that the model considers all prior data in its predictions. The starting point is fixed, and as new data points become available, they are added to the window.

1.2.2.1.2 Rolling Windows Rolling windows, also known as sliding or moving windows, are different in that they maintain a fixed size. As the window "rolls" forward in time, it drops the oldest data point and adds the most recent one.[9]

Rolling windows are particularly useful when the most recent data is more relevant for predictions, as they give equal weight to all observations in the window. This method can help to smooth out short-term fluctuations and highlight longer-term trends or cycles.

For instance, a 7-day rolling window model for predicting ticket sales would use the sales data from the previous seven days to predict sales for the eighth day. After this, the window would roll forward one day, dropping the first day's data and incorporating the eighth day's data, and so on.

"Expanding window is useful when the series has a strong seasonal pattern and stable trend as in this case, the first observations of the series contain potential information the future values. While the rolling window is useful when we have a rather volatile series or when the most recent history is more relevant for forecasting."[10]

1.3 Summary

In this chapter we examined the latest developments in ML models and methods aimed at maximizing revenue in e-commerce, focusing on dynamic pricing and sales forecasting. We reviewed several prominent models and techniques, discussing their underlying principles, strengths, and weaknesses, as well as their effectiveness in the ticket sales domain.

The section began by discussing the early attempts at forecasting ticket sales for cultural events and the evolution of statistical models predicting sales. We highlighted the importance of incorporating social media data sources into sales forecasting models for various industries.

Next, we explored purchase forecasting and dynamic pricing techniques used in various industries, such as the hospitality and transport industries. We discussed the multi-step model architectures and considered the potential for a more accurate classifier model based on customers' scrolling before the price of the product is shown, which could lead to better dynamic pricing strategies across a broad range of e-commerce websites.

The latter portion of the chapter provides an overview of fundamental concepts in ML, briefly describing classification, regression, and time series prediction.

Chapter 2
Data Analysis

In this chapter, we delve into the analysis of data obtained from a third-party online e-commerce website, which serves as a business-oriented platform for customers to create events and sell tickets to consumers. We provided specific details about the platform and its user base, offering an understanding of the context in which the data was generated and collected.

2.1 Technological Tools Employed

Data were transferred to a personal computer for subsequent analysis. Initially, customized SQL queries were employed to facilitate data analysis, followed by the download of the data in CSV format to the personal computer. This was made possible through the DataLore feature, which allows for query result downloads. Data from Google website tracking services were obtained using Looker studio and subsequently with the Google API for GA3 (Google Universal Analytics) Reporting API v4, utilizing the oauth2client Python library, and for GA4 (Google Analytics) Analytics Data API, employing Google analytics Python library. Lastly, there is a two additional data sources: the Meteostat API, which enables access to historical weather information recorded at meteorological stations and provides weather forecasts for a 14-day period. The Python version of the API was utilized in this case. Next OpenHolidays API which was used for public holiday information.

Data were analyzed in Python notebooks using libraries pandas and numpy and for visualizations libraries seaborn, and sublibrary matplotlib.pyplot. The source code for the visualizations can be seen in the enclosed material `src/analysis`.

2.2 Data Sources

The primary source of data for this study is the internal SQL database, which is designed to gather information for the website's internal operations and store data from customers and their users. Additionally, two other sources of data are derived from Google website tracking: the earlier version GA3, and its more recent counterpart GA4, which was developed to replace the old architecture of GA3. The analyzed data were collected from 2020 up to the present. The data analyzed consisted of 120,000 records of ordered tickets.

2.3 Overview of Website Consumer Architecture

As previously mentioned, the data sources originate from an online, business-to-business ticket sales platform that enables the selling of tickets through a platform, which also includes a cash desk feature for recording in-person ticket sales. The website's primary focus is on developing features for its customer businesses, resulting in consumer interactions with the platform occurring mainly through the process of purchasing tickets for specific events, which are promoted externally.

Consumers can engage with the platform via an integrated widget on the event's website or by being redirected through a link from the event's website to the platform. In both implementations, consumers can select tickets and proceed to provide the necessary payment information for the transaction. Smaller businesses often promote their events via social media where they provide the platform's link, while larger businesses usually have their own websites where they can incorporate the widget feature, allowing consumers to buy the tickets on the event website without being redirected.

2.4 Internal Database

The internal database is structured into 4 main tables organizers, events, orders, and tickets the next section takes a look at data analysis and description of the subject of each data structure. Then lastly we connected the chosen event sales with recorded weather from Meteostat API. The data structure of visualized elements can be seen in table 2.1.

■ **Table 2.1** Internal database data structure

Table: Event		**Table: Order**		**Table: ticket**	
Column Name	Data Type	Column Name	Data Type	Column Name	Data Type
Event_id	UUID	Order_id	UUID	Ticket_id	UUID
name	Varchar	customer_id	UUID	order_id	UUID
organizer_id	UUID	created	DateTime	price	Float
orderable_from	DateTime	status	Varchar	scanned	Boolean
orderable_until	DateTime	paid	Boolean	event_id	UUID
latitude	Float				
longitude	Float				
start	DateTime				
end	DateTime				

2.4.1 Organizers and Event Categories

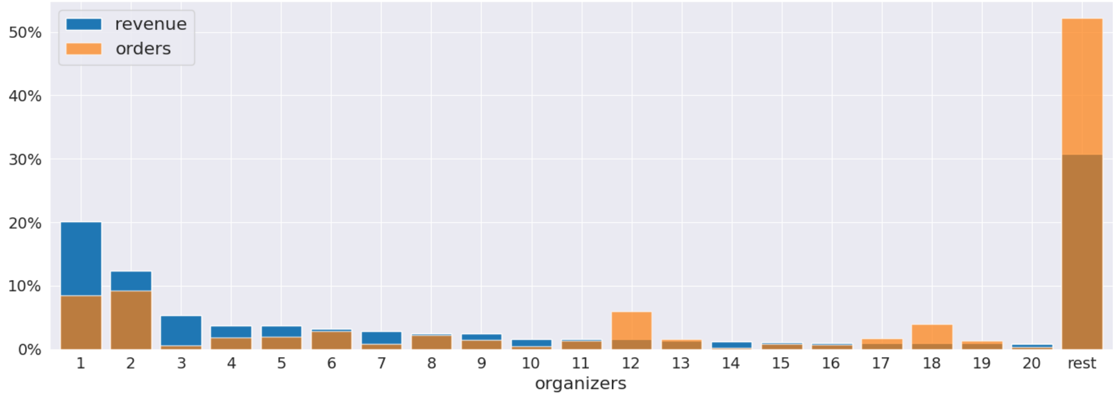
Various types of businesses utilize the platform to sell tickets for a wide range of events, catering to diverse customer preferences and needs. These events span from small theatre productions to natural parks, festivals, cultural events, and sports events. Below is a summary of the primary event categories:

- Sports
- Entertainment
- Live Music Performances
- Conferences

- Festivals
- Food Events
- Live online recordings

Regrettably, the internal database does not contain an organizer or event categorisation that would enable sector-specific analysis. As illustrated in the figure 2.1, there are two primary organizers, with one generating over 20% of the total revenue. The visualization indicates that a higher number of tickets sold does not always correspond to increased revenue, and the opposite is also true.

Figure 2.1 Top 20 organizers



2.4.2 Orders

In this section, we examined the seasonality of orders, explore consumer behaviour patterns, and identify any significant trends in consumer actions. The analysis of orders covers various aspects, including the volume of orders, the impact of seasonal factors, and correlations between order volume and external factors, such as weather conditions.

2.4.2.1 Order Volume

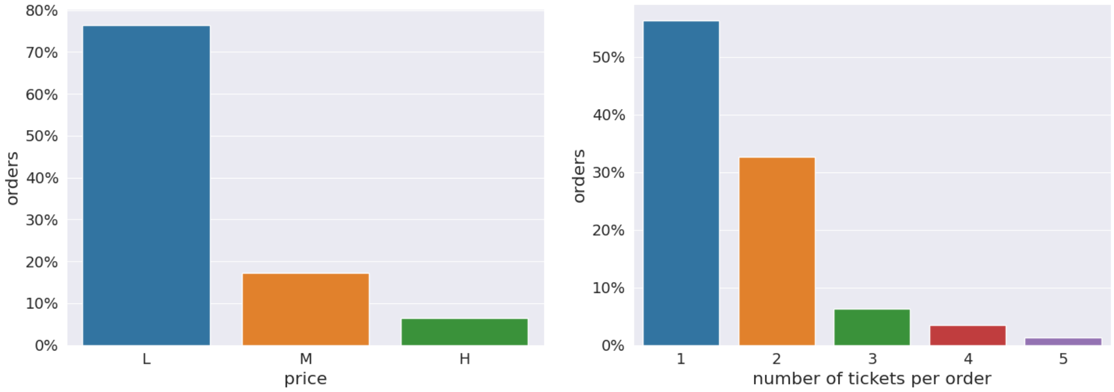
From the figure 2.2, we can observe that customers tend to place orders with lower prices and frequently purchase only one or two tickets per order. This insight implies that event organizers should consider offering a variety of ticket pricing options to accommodate different customer preferences and budgets.

2.4.2.2 Price Conversion

Initially, the forex-python API was used to obtain current and historical exchange rates. However, due to the slow speed of conversion caused by the API's inability to retrieve a list of historical conversions per single API call, a custom script was developed. This script downloads CSV files in a ZIP format from the European Central Bank's website[1], which updates exchange rates daily. From the downloaded ZIP file, the nearest date for conversion was selected, considering that some dates might be missing in the file. After converting all prices to the same currency, the orders' prices were classified into three categories.

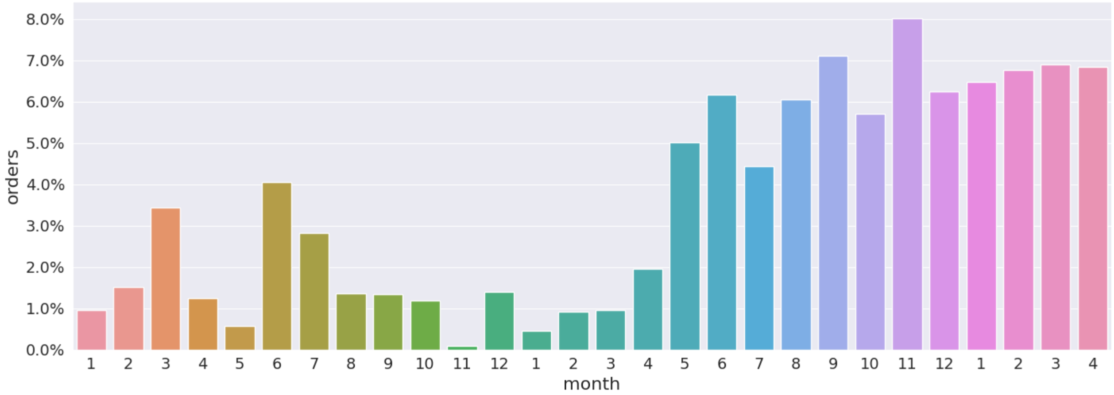
[1] https://www.ecb.europa.eu/stats/policy_and_exchange_rates/euro_reference_exchange_rates/html/index.en.html

Figure 2.2 Order volume



(a) Orders per price **(b)** Number of tickets bought per order

Figure 2.3 Monthly number of orders



To create these categories, the minimum and maximum prices for each event sales were first determined. Then, the price range was divided into three equal-length size groups: low, medium, and high.

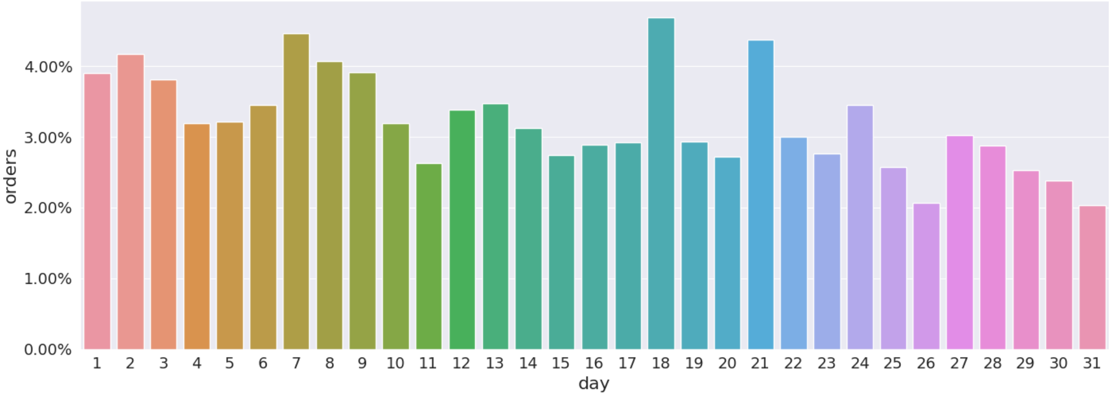
2.4.3 Seasonality

Next, we investigated the seasonality of orders and their impact on consumer behaviour. This involves analyzing the order volume over time and identifying patterns that may be attributed to seasonal factors, such as weather conditions, holidays, or cultural events. The integration of historical weather data from the Meteostat API helped us uncover potential correlations between weather patterns and ticket sales. Moreover, by studying the influence of seasonality on ticket sales, we can provide valuable insights into how businesses can optimize their event planning and marketing strategies according to seasonal trends.

From the figure 2.3 becomes apparent that the platform is in the process of acquiring new businesses, and no major information about monthly seasonal trends can be extracted. There are no major organizers with constant sales throughout the year, so analyzing the number of orders for individual organizers would not yield valuable insights.

In the figure 2.4, there is a little downward trend and then there are noticeable jumps in the 18th and 21st days of the month which exhibit a higher density of orders. Furthermore, the lower

Figure 2.4 Orders per day in month



density observed during the last days of the month can be attributed to the uneven number of days across different months, resulting in a less consistent distribution of orders during those days.

From the figure 2.5, it is evident that consumers tend to purchase tickets most frequently within a period of 5 to 20 days from the start of the sales. In terms of relative time before the end of sales, consumers are most likely to buy tickets during the final days of sales, particularly within the last 5 days. To provide context for these observations, the density in the last subfigure 2.5c demonstrates that events typically have around 25 days of sales.

From the figure 2.6a, we can observe an upward trend in ticket purchases on working days, with the number of purchases generally increasing as the week progresses. Interestingly, consumers are least likely to buy tickets on Saturdays. A similar pattern is also evident in the figure with events starting days 2.6b, with Saturdays being the exception.

This trend suggests that consumers may be more inclined to make ticket purchases during the workweek, potentially due to increased online activity or the influence of workplace discussions and social interactions. On the other hand, the decrease in ticket sales on Saturdays could be attributed to consumers engaging in weekend activities or being less focused on online event browsing and purchasing.

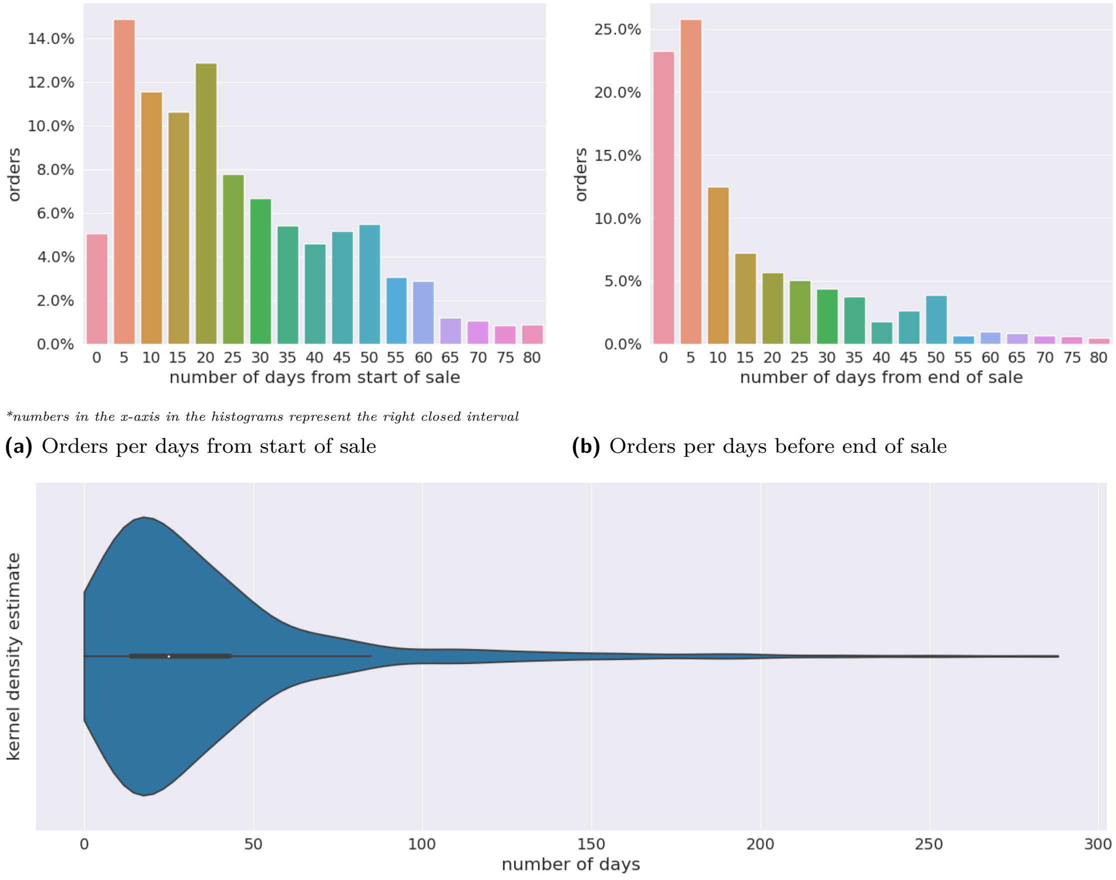
2.4.3.1 Recapitulation of Seasonality

These findings suggest that there are two critical periods in the ticket sales cycle: an initial period of heightened interest soon after the start of sales, and a last-minute surge in purchases just before the sales end. Businesses can capitalize on these trends by tailoring their marketing and promotional efforts to maximize engagement during these crucial periods.

For example, event organizers may consider offering discounts or special promotions in the middle of the selling period, particularly for events with longer sales durations, as these sales days tend to experience lower overall sales. Such incentives could potentially attract more customers and encourage purchases during this relatively slower period. Additionally, organizers could explore offering other benefits, such as exclusive merchandise or early access to the event, to incentivize early ticket purchases.

Likewise, organizers can employ strategies such as limited-time offers, flash sales, or countdown campaigns during the final days of sales to create a sense of urgency and stimulate last-minute ticket purchases. These targeted promotions can effectively capitalize on consumer behaviour trends and help drive sales during crucial timeframes.

Figure 2.5 Orders per period



**numbers in the x-axis in the histograms represent the right closed interval*

(a) Orders per days from start of sale

(b) Orders per days before end of sale

(c) Events orderable period distribution

By understanding and adapting to the dynamics of consumer ticket purchasing behaviour, businesses can enhance their promotional strategies, optimize ticket sales, and ultimately increase their overall revenue.

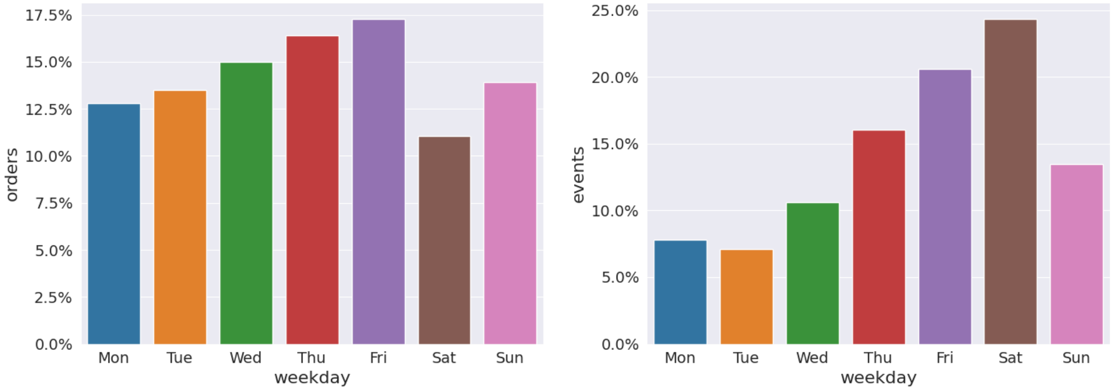
2.4.4 Occupancy

In this context, the occupancy of an event is defined as the number of tickets sold divided by the total number of tickets available for sale for each event.

From the figure 2.7, it is evident that event occupancy is generally low, although there are instances of events with full occupancy. This observation could suggest that event organizers might overestimate the potential number of attendees, resulting in low occupancy rates. Alternatively, they may not prioritize occupancy-based cutoff sales, instead relying on a combination of date-based cutoffs and chargebacks for tickets when the event's capacity exceeds the limit.

An accurate estimate of the number of viable attendees is crucial for an event's profitability. The figure indicates that organizers often overestimate this figure, which could result in higher expenditures for event venues and a subsequent loss of profits. To mitigate this issue, organizers should consider leveraging historical data, industry trends, and market research to make more informed predictions about attendance levels. By doing so, they can optimize their event planning, minimize financial risks, and maximise profitability.

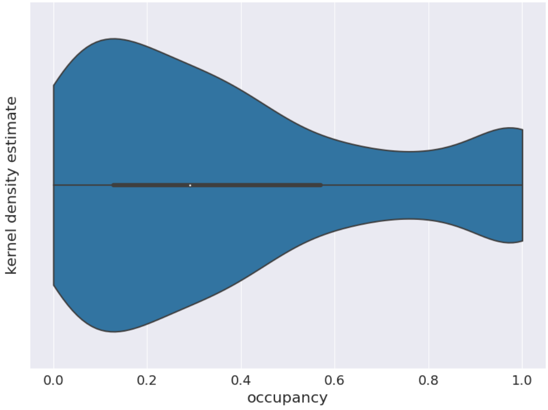
Figure 2.6 Orders per weekday with context



(a) Orders per weekday

(b) Events starting day

Figure 2.7 Occupancy of events



2.4.4.1 Scanned Tickets

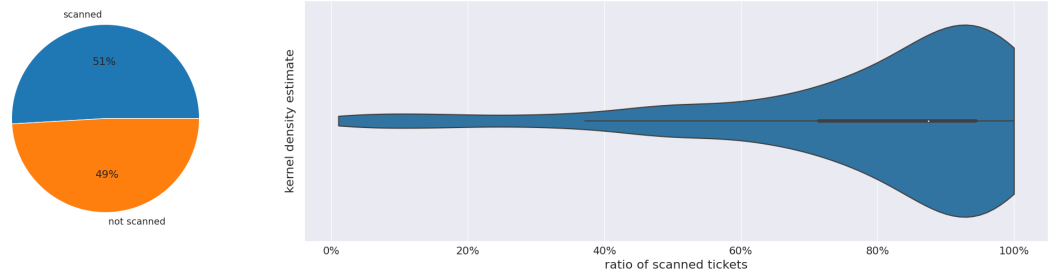
The pie chart 2.8a reveals a surprisingly high ratio of unscanned tickets. When we group tickets by their event and calculate the ratio for each event, visualizing the results in a violin plot 2.8b, the situation appears better. High ratio of unscanned tickets could be caused by unprofessional behaviour from event staff. This could lead to a significant number of illegal attendees without purchased tickets and limits the usefulness of this information for an informed strategy where organizers intentionally oversell the venue capacity, knowing that some ticket buyers won't attend the event. The relationship between scanned tickets and other factors was explored but did not yield any conclusive findings. Additional visualizations that investigate this phenomenon can be found in appendix A.2.

2.5 Holidays and Weather

For this analysis, we selected a specific organizer with steady sales for a recurring event, which is a natural monument. The rationale behind this choice was the assumption that this type of event would be most affected by weather conditions and holidays.

To compare sales data, the MinMaxScaler from the sklearn library was used to scale the sales figures.

Figure 2.8 Scanned tickets overview



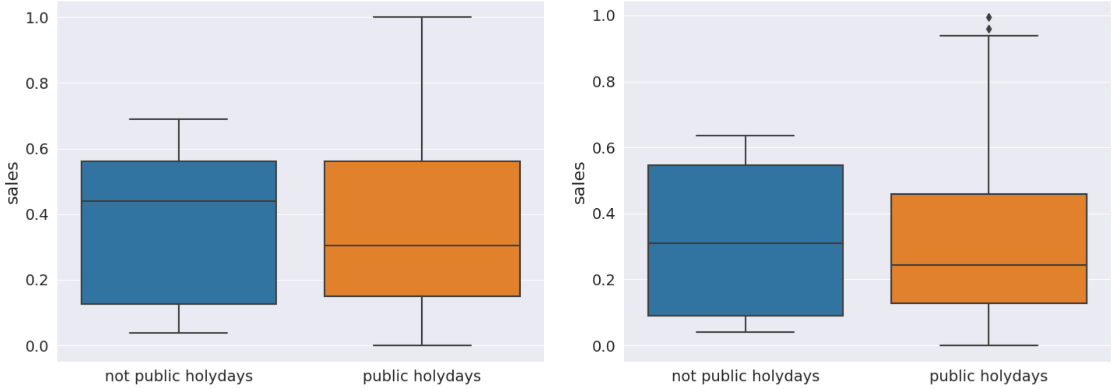
(a) Scanned tickets **(b)** Scanned tickets ratio per event

The Meteostat daily function[2] was employed to obtain weather data from meteorological stations located within a 20 km radius of the natural monument. Further information on the analyzed metrics can be found at Metostat documentation website[3]. For public holidays OpenHolidays API[4] was used as a data source, the analysis considered states and federation states within a 250 km range of the event venue. This distance was chosen as an influential range for the venue's potential customers based on the number of daily visitors.

2.5.1 Public Holidays

From the left subfigure in 2.9a, we can observe that, interestingly, the median sales value is lower on public holidays compared to regular days. However, the maximum sales value extends higher on holidays, suggesting that not all holidays impact sales equally. Some dominant public holidays might have a more significant influence on sales. This trend is further emphasized in the right subfigure 2.9b of the graph where only work days are taken into comparison.

Figure 2.9 Sales on holidays



(a) Daily sales on public holydays **(b)** Daily sales on public holidays on work days

[2] https://dev.meteostat.net/python/api/daily/#parameters

[3] https://dev.meteostat.net/formats.html#meteorological-data-units

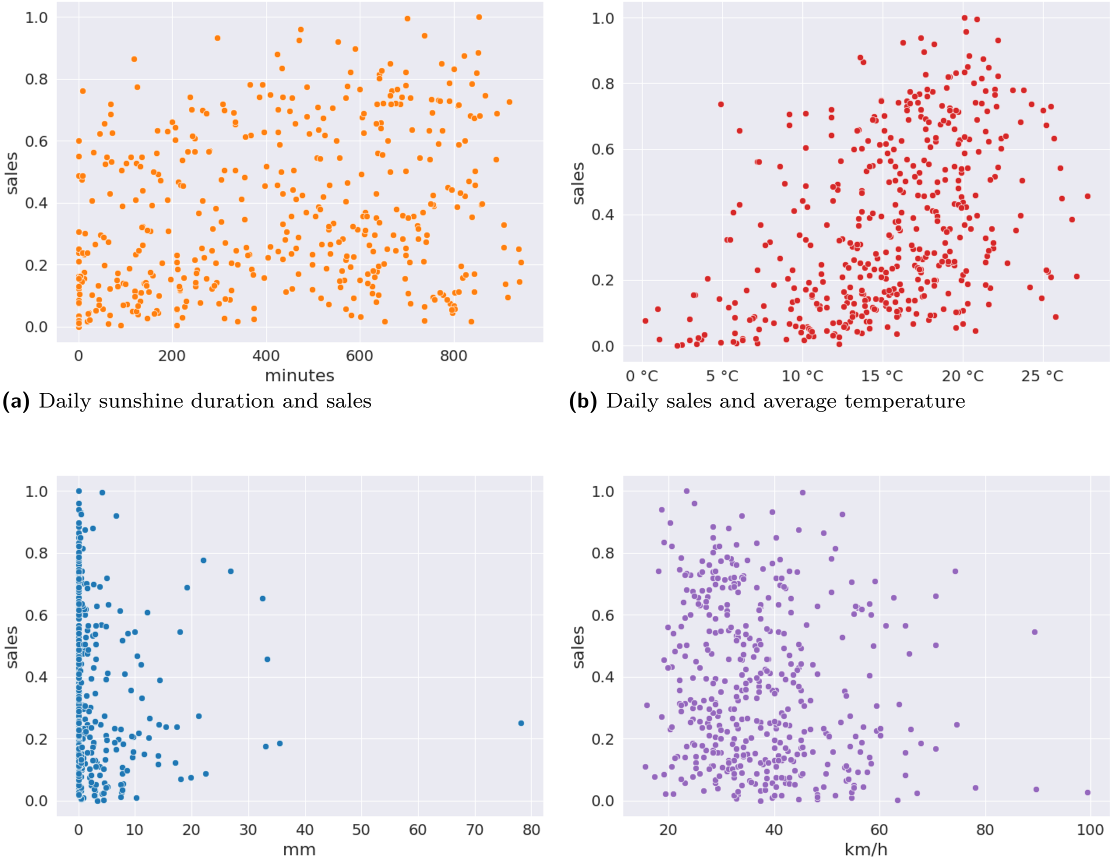
[4] https://www.openholidaysapi.org/en/

2.5.2 Weather

From the weather-related figure 2.10, we observe that the most significant positive impact on sales corresponds to the daily average temperature. Precipitation appears to have an overall negative influence on sales; however, due to the low frequency of days with substantial precipitation in the region, it is difficult to draw a definitive conclusion. Maximum daily peak wind speed has little effect on sales. However, it is important to highlight that sales tend to be significantly lower during extreme weather conditions, such as when wind speeds reach or exceed 80 km/h. There are no visible trends associated with daily sunshine duration.

These observations indicate that normal weather does have an impact on ticket sales for specific natural monument venues analyzed in the analysis. To be able to draw definitive conclusions about the effects of weather and holidays on event sales, further analysis should be conducted on a broader range of event types.

■ **Figure 2.10** Weather and sales



(a) Daily sunshine duration and sales

(b) Daily sales and average temperature

(c) Daily precipitation and sales

(d) Maximum daily peak wind speed and sales

2.6 Google Website Tracking

The e-commerce website under analysis utilizes Google services to track and record website traffic. Google provides a vast array of tools for analyzing the gathered data in GA3 and GA4 and configuring the data collection process. However, it lacks tools for accessing granular data, which limits in-depth data analysis. Throughout our examination, we identified several inconsistencies, which are detailed in this subsection. Google's extensive yet superficial documentation contributes to confusion and hinders the drawing of precise conclusions. Due to the end of life of GA3 administrators of a websites need to transition to GA4 if they want to continue using Google website tracking service more on the topic on Google documentation website[5].

2.6.1 Data Structure

GA3 and GA4 structure their data into dimensions and metrics, both of which are minimally documented and explained. Google's definitions of dimensions and metrics are as follows:

"A dimension is an attribute of your data. It describes your data and it's usually text as opposed to numbers. An example of a dimension is Event name, which shows the name of an event that someone triggers on your website or application (such as "click")."[11]

"A metric is a quantitative measurement, such as an average, ratio, percentage, and so on. It's always a number as opposed to text. One way to think about metrics is that you can apply mathematical operations to them. An example of a metric is Event count, which shows the total number of times an event is triggered."[12] Description of available dimensions and metrics can be found on GA3[6] and for GA4[7] documentation websites.

2.6.2 Architecture

The architecture of Google website tracking consists of four primary components: Collection, Configuration, Processing, and Reporting. These components facilitate data collection from user interactions, management of data processing settings, data processing using configuration data, and access to processed data through reports, respectively.

The Core Reporting API serves as a crucial component, enabling access to Google Analytics reporting infrastructure. It allows data export for dimensions and metrics, supports calculations, and provides segmentation options.[13]

In summary, while Google website tracking presents an array of tools and features, its architecture and API possess limitations that may influence the quality and depth of analysis achievable. It is vital to be aware of these constraints when relying on Google website tracking for your data requirements, and to consider alternative solutions if the limitations prove to be an obstacle.

2.6.3 Comparison between GA3 and GA4

While there are only a few changes concerning data, key elements result in distinct characteristics and challenges when transitioning from GA3 to GA4, despite Google's intent for a seamless upgrade process. The main changes include:

- Event-based architecture: GA4 is event-based, meaning each event, such as visiting the website or clicking on an item, is stored separately. In contrast, GA3 used a session-based architecture that grouped a series of events into one session. This difference affects data and should be considered when comparing the two versions. Some events are collected by default, while others must be specifically configured to align with the website architecture, such as for example adding an item to a shopping cart.
- Introduction of new dimensions and removal of old GA3 dimensions: The architectural shift from GA3 to GA4 led to the introduction of new dimensions and the removal of some older ones.

[5] https://support.google.com/analytics/answer/11583528?hl=en

[6] https://ga-dev-tools.google/dimensions-metrics-explorer/

[7] https://support.google.com/analytics/answer/9143382?sjid=7597815771812991532-EU

- Historical data limit: GA3 had no expiry date for data storage, while GA4 offers two options 2 months and 14 months.[14]
- Conversion count: In GA3, a conversion is counted only once per user session. In GA4, a conversion can be counted multiple times per user session. Conversions are events which are defined by the administrator as conversions, with some events being set up as conversions by default.[14]
- Data threshold: GA4 introduces a data threshold, which is vaguely defined as *"If your report or exploration is missing data, it may be because GA4 has applied a data threshold. Data thresholds are applied to prevent anyone viewing a report or exploration from inferring the identity of individual users based on demographics, interests, or other signals present in the data."*.[15]

There is no concrete definition of the limit at which the threshold is applied, such as the number of customers visiting the website within a specific time frame. Furthermore, there is no specification on the returned value when the threshold is applied. From our experience, no null values are returned, making it difficult to determine the accuracy of the requested value. Some tools can visualize when the threshold is applied, but these tools are not suitable for data analysis and extraction. For more information on this topic, visit GA4 documentation website[8].

2.6.4 Data Comparison

During the 3-month transition to GA4, the website simultaneously utilized both GA3 and GA4, which enables direct comparison between the data sets. The analysis concentrated on comparing metrics specified in the Google article[9], which were considered to be comparable and try to capture the same theoretical value. To examine the differences in data characteristics, visualizations were created to explore these disparities.

The objective of this analysis was to investigate the feasibility of converting the data to a new format in order to expand data availability for long-term trend analysis and the potential development of ML models for new website tracking.

The rationale behind this data conversion effort was inspired by a successful comparison of website tracking tools in article[10]. In the article, authors compared long-term tracking of different tracking tools with one of them being GA4, without a major difference between website tracking data.

For the comparison of metrics, dimensions page path and daily traffic were selected, as these were deemed to be of particular interest to event organizers and essential for the development of ML models aimed at predicting ticket demand.

The comparison 2.11 revealed significant differences in values and characteristics between GA3 and GA4 data, as illustrated in the figure. This outcome hindered the transformation of old data to the new format. One potential cause of these discrepancies could be the impact of GA4's data thresholds. However, due to the limited explanation provided in the GA4 documentation, it is challenging to reach a satisfactory conclusion. Another possible cause might be attributed to the incorrect setup of either GA3 or GA4 on the event website, although this eventuality was explored and setup was controlled. This possibility is unlikely, given the default setup of GA3 and GA4, is designed to be simple and deployed without extensive technical experience in frontend development. The process primarily requires registration on Google's GA3 or GA4 website and deploying a code snippet to the JavaScript frontend of the website. No articles or documentation were found regarding tracking conflicts when GA3 and GA4 run simultaneously, and Google does not discourage administrators from doing so. Additional comparisons between metrics can be found in appendix A.1.

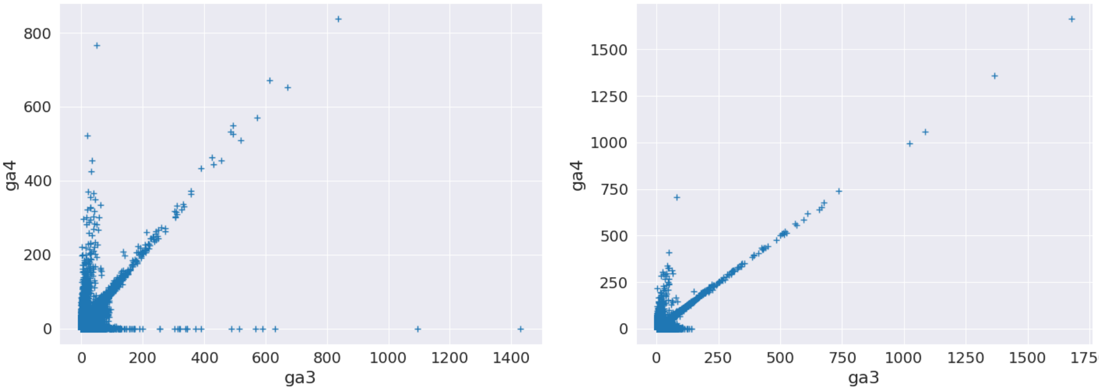
[8] https://support.google.com/analytics/answer/9383630?hl=en

[9] `https://support.google.com/analytics/answer/11986666`

[10] `https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0268212`

Taking into account the effects of data thresholds and the short period of data collection with GA4, the focus of this study is primarily be on working with data from GA3.

Figure 2.11 GA3 metrics compared to GA4



(a) GA3 ga:sessions comparison to GA4 sessions **(b)** GA3 ga:users comparison to GA4 totalUsers

2.6.5 Inconsistencies Found

When extracting the data some inconsistencies in the received data were found. When analyzing ticket sales for each event separately, there were instances when tickets were purchased through the website but no traffic was recorded in GA3.

When sending a request for GA4 data through the GA4 API for analysis, noncomplying behaviour was observed. When specifying a large date range for data requests, the API returned a smaller number of data points than when dividing the date range into smaller segments. Interestingly, this behaviour was also observed in Looker Studio, a tool for data analysis of Google utilities.

2.7 Data Analysis Summary

Throughout this chapter, we examined various data sources available provided by a third party. Key positive findings include the identification of increased demand at the beginning and end of ticket sales periods and the influence of holidays on sales.

However, we also observed several negative aspects within the internal database, such as a low number of scanned tickets and an overestimation of event capacity. Throughout the findings, we incentivize the third party to modify their processes and address/communicate these issues with event organizers to maximize potential benefits.

Another notable concern was the inability to transition data from the older GA3 format to the new GA4 format, as well as the overall poor performance of Google's tracking applications for comprehensive data analysis and ML model development.

Chapter 3

Enhancing data collection

In the previous chapter, we analyzed the ticket sales data provided by our commercial partner and identified areas where improvements could be made. Based on our analysis, we recognized that the current data collection system could benefit from an extension or a new methodology to gather more accurate, relevant, and comprehensive information. In this chapter, we presented the design of a new data collection methodology.

3.1 Objectives of the New Data Collection Methodology

The primary objective of our new data collection methodology is to increase the data collected, allowing for a more detailed analysis of ticket sales patterns and customer behaviour which can provide a more reliable basis for decision-making and enables analysis of long-term trends and changes in customer preferences. Facilitate the integration of external data sources, such as social media and demographic information, to further enrich the analysis.

3.2 Expanded Data Sources

In addition to the primary ticket sales data, we propose integrating additional data sources to enrich data for possible analysis and provide more comprehensive insights into factors that may influence ticket sales. These sources include social media metrics, economic indicators, weather data, information about competing events, localization information and types of events.

3.2.1 Social Media Metrics

Social media platforms, such as Facebook, Twitter, and Instagram, offer a rich source of data on user engagement and sentiment related to events, artists, and venues. We can gauge the level of interest and popularity surrounding specific events or artists by analyzing metrics like likes, shares, comments, retweets, and hashtag usage. This information can indicate potential ticket sales and help identify trends or shifts in customer preferences.

3.2.1.1 Facebook as a Critical Information Source

Facebook, in particular, can be a vital source of information accessible through its API [1]. As a social media platform, it provides valuable insights connected to its event feature. This feature allows organizers to register private or public events on the Facebook website and offers additional functionalities to help with marketing and enhance brand or organizer awareness.

The type of data that can be accessed through Facebook on registered events includes the number of people interested in the event and their locations. This information can be crucial for understanding the potential audience for a specific event. When combined with long-term characteristics and sales data, this metric can be highly efficient in predicting demand for a particular event.

Furthermore, Facebook event data can also provide insights into attendees' demographics, such as age, gender, and interests. This information can help event organizers tailor their marketing efforts to target specific audience segments more effectively. Additionally, analyzing user interactions with the event page, such as likes, comments, and shares, can help gauge the overall sentiment and excitement surrounding the event, further informing marketing strategies and potential ticket sales projections.

3.2.2 Economic Indicators

Macroeconomic factors can significantly influence the entertainment and culture industry. Key indicators, such as Gross Domestic Product (GDP), unemployment rate, and disposable income levels, can provide valuable insights into a region's overall economic health and potential effect on ticket sales. For instance, in a robust economy with higher disposable income levels, consumers may be more inclined to spend on entertainment and cultural events, increasing ticket sales.

3.2.3 Weather

Weather conditions play a critical role in determining the success of outdoor events. By incorporating historical and forecasted weather data into our analysis, we can assess the potential impact of weather conditions on event attendance and ticket sales. For example, adverse weather conditions, such as heavy rainfall or extreme temperatures, may deter people from attending outdoor events, resulting in lower ticket sales and reduced revenue. Exploring how popular weather forecasting applications affect sales and their relationship to actual recorded weather is also essential.

3.2.4 Competing Events

It is crucial to consider other events occurring concurrently in the same geographical region or within the same industry segment, as they can directly impact ticket sales for a particular event. By monitoring and tracking competing events, we can identify potential overlaps in target audiences and evaluate the effect on ticket demand. For instance, two popular concerts on the same day in the same city may lead to lower ticket sales for both events due to audience fragmentation and shared interests.

[1] `https://developers.facebook.com/docs/graph-api`

3.2.5 Event Type Data Collection

Understanding the nature of the event is crucial for determining the factors that influence ticket sales. The data collection methodology should incorporate information about various event types, such as concerts, sports events, theatre performances, festivals, and conferences. By categorizing events based on their type, we can identify trends, preferences, and specific challenges associated with each category. This information aids in understanding the demands of different types of events, ultimately leading to more effective decision-making.

3.2.6 Localization

Localization data plays a significant role in assessing the potential success of an event. By incorporating information about the event location, such as the city, region, or country, we can better understand the local market and its impact on ticket sales. Additionally, gathering data on venue characteristics, such as size, accessibility, and amenities, can help identify the factors contributing to an event's success in a specific area.

3.3 Artificial Data Generation Model

To evaluate the performance of our proposed data collection methodology and to enlarge the training data set, we developed a synthetic data generation model that simulates newly collected data. This model incorporated the original data set's characteristics while accounting for the changes introduced by the new methodology. The architecture and process of the synthetic data generation model are discussed in the next chapter.

3.4 Data Collection Summary

In this chapter, we presented the design of a new data collection methodology, focusing on improving the accuracy, relevance, and comprehensiveness of the collected data. Our primary objectives were to increase the amount of data collected and facilitate the integration of external data sources.

We proposed expanding the data sources to include social media metrics, economic indicators, weather data, and information about competing events. We highlighted the importance of platforms like Facebook as a key source of information on user engagement, event attendance, and demographic insights.

Chapter 4

Synthetic Data Generation

In this chapter, we discuss the process of generating synthetic data, which was essential for creating training and validation datasets for ML model training due to the small size of the original dataset. A total of 181 daily events of online ticket sales were handpicked from the original third-party data source, amounting to 9,388 days of ticket sales. These online sales were specifically chosen due to their suitable size for training ML models to predict demand and their consistent web tracking records in GA3.

4.1 Technological Tools Employed

To facilitate data manipulation, we utilized Python libraries such as Pandas and NumPy. For prediction models, we employed Scikit-Learn's DecisionTreeRegressor and the XGBoost library. Finally, the tabgan.sampler GANGenerator was used to generate synthetic events starts with static features.

4.2 Overview of Synthetic Data Generation Process

The process of generating synthetic data involved several steps, as outlined below:

Feature Engineering: Relevant features were extracted from selected daily events of online ticket sales to create a comprehensive set of variables that could effectively represent the underlying patterns and trends within the ticket sales data. This step aimed to simplify the problem of synthetic data generation and ensure that the trends within the data would be accurately represented in the created synthetic data.

Data Preprocessing: The selected features were preprocessed to ensure consistency and remove any missing data points in website tracking. This step was crucial in preparing the dataset for the subsequent stages, as it laid the foundation for effective synthetic data generation.

Model Training: With the features engineered, we trained various models:

- Train the tabgan.sampler GANGenerator to generate synthetic events starts with static features.

- Train two rolling window models on predicting the next day's time-series data.

Combine Trained Models and Generate Synthetic Data: We generated over 500 synthetic events starts with static features and used the trained rolling window models to create synthetic sales data. These models were evaluated and fine-tuned to ensure optimal performance in predicting ticket sales demand. The synthetic data mimicked the characteristics of the original dataset while maintaining the privacy of individual data points.

4.3 Feature Engineering

To create a more accurate synthetic data generator that captures event sales as a whole and not just one-time orders without event sales-specific connotations, we selected daily variables of online sales and the number of orders on the websites along with a list of event sales-specific variables.

The selected event sales-specific variables include:

- Daily number of orders for a particular event
- Daily number of visitors for event order website
- Weather data recorded from the nearest weather station
- Absolute and relative time information
- Localization cluster
- Manually inputted values, such as:
  - GDP of the event location
  - Event type
  - Social media occurrence
  - Competition score
  - Type of event

4.3.1 Methodology

For weather data, we utilized the Meteostat hourly function, specifically the weather codes, which were then grouped into three ordered categories: good, normal, and bad. The list of weather codes can be found at the Meteostat documentation site[1].

The reason we have solely chosen the daily number of visitors for the event orders site, despite GA3 offering a larger number of metrics, was due to concerns about the quality and reliability of the information provided by those variables. By concentrating on a single, more trustworthy metric, we could ensure that models are based on precise data that accurately reflect customer behaviour and preferences in relation to event sales.

Relative time variables, such as the number of days from the start of sales and the number of days until the event, were included in the analysis. Manually inputted values were obtained through web search information and added to the data set to provide a more comprehensive and accurate representation of the event sales.

[1] https://dev.meteostat.net/formats.html#meteorological-data-units

4.4 Preprocessing

Although we chose to use only the number of visitors from GA3 web tracking, there were instances where days of online sales had no recorded number of visitors. To address this issue, we imputed the missing values by calculating the mean conversion rate and multiplying it by the number of orders. To introduce robustness into the equation, we further adjusted the predicted number of visitors by multiplying by random fraction.

To simplify the prediction task for the rolling window models, we decided to have them predict relative demand rather than absolute demand, represented by the number of sold tickets and visitors. To achieve this, we normalized each event's sales data using a unique MinMaxScaler for each event. This approach allowed the models to focus on predicting relative demand.

We created a map (a collection of key-value pairs) of scalers, where the key was the relative size of the event. By doing this, we encoded the event size information into the synthetic data generation process concretely into static features, which enabled us to reverse-transform the relative number of users and sold tickets back into their absolute values later on while preserving the features specific to each event size. This method not only simplified the problem for the time-series generation models but also ensured that the generated synthetic data was more accurate.

4.5 Model for Generating Synthetic Sales

We decided to implement a rolling window model due to the unique characteristics of the data. Implementation of the rolling window model was considered a better solution for this problem rather than using frameworks and libraries developed for time-series forecasting. A window length of 7 days of sales proved to be the best parameter for training the rolling window model.

We trained two separate models because having only one model for generating synthetic data resulted in constant-like time-series variables. For this reason, we trained an XGBoost model with optimized hyperparameters to best capture trends along the different events, such as sales at the end of the sales interval. Additionally, we trained another overfit model a Decision Tree with a relatively large depth to introduce more variability into the generated time series. We combined these two models by first predicting the target variable for each model separately and then choosing a random point on the line defined by those predictions.

By employing these two models, we were trying to generate synthetic data that accurately reflects the underlying patterns in the original data while introducing enough variability to produce meaningful results for our analysis.

This approach allowed us to produce meaningful results for our experiments by reflecting the underlying patterns in the original data while introducing sufficient variability through the combined use of two models. Algorithm 1 summarizes the entire process of generating synthetic sales data.

4.6 Evaluating and Validating Synthetic Data

Once the synthetic data was generated, we evaluate and validate the data by visualizing time series data. This process confirmed the successful generation of synthetic data that could be used for training and validation purposes while maintaining the privacy and security of the original data.

Algorithm 1 Synthetic sales generation algorithm

```
events_start
syn_events_start ← GAN.GENERATE(events_start)
syn_sales ← empty list
for each event in syn_events_start do
    syn_event_sales ← empty list
    X ← event.reg_input
    day ← event.start_date
    while $day \leq event.end\_date$ do
        pred ← COMBINEREG(X)
        append pred to syn_event_sales
        X ← ROLLWINDOW(X, pred, day, event.static_features)
        day ← day.tomorrow()
    end while
    syn_event_sales ← scaler_map[event.id].REVERSETRANSFORM(syn_event_sales)
end for
```

4.7 Synthetic data Generation Summary

In this chapter, we presented the development of a synthetic data generation model to enhance our training dataset. We selected 181 daily events of online ticket sales from the original third-party data source, amounting to 9,388 days of ticket sales.

The synthetic data generation process consisted of multiple steps, including feature engineering, data preprocessing, model training, and combining the trained models to generate synthetic data. We utilized a rolling window model for generating synthetic data, which involved training two separate models, an XGBoost model and an overfit Decision Tree model, to capture trends and introduce variability in the generated time series data.

The generated synthetic data successfully mimicked event sales with a lower number of sale days. providing valuable insights for training and validation purposes while maintaining data privacy and security. This chapter highlights the importance of synthetic data generation in enhancing ML model training and validating new data collection methodologies.

Chapter 5

Dynamic Pricing

In this chapter, we focus on predicting ticket prices based on synthetic data generated. We also discuss the performance of the original and proposed data collection methodologies in the context of price prediction. Finally, we evaluate performance on real test data. Throughout the development of the price predicting model, We decided to add Gaussian random noise to the synthetic data to achieve more robust results and to mitigate errors made during the synthetic data generation process.

5.1 Architecture of Models

The model architecture was inspired by the price predicting model tested in real-world scenarios described in the paper concerning airline ancillaries price prediction, concretely the model named *"Ancillary Purchase Prediction with Logistic Mappin"*[5]. The model comprises a two-step architecture. The first step is a classifier that decides, based on context X, whether it is the right situation for increasing the price. The second step consists of a tuned logistic function which, uses the returned probability from the classifier, to determine the price.

The classifier described in the paper was trained to predict whether the customer will or will not buy the ancillary. In contrast, the developed classifiers in this thesis are focused on determining if the number of orders will be higher than the mean number of orders for an event. This strategy aims to lower the price when sales are lower than the mean and raise the price when sales are higher than the mean. We decided to use this strategy due to the unavailability of data for training a classifier that would predict if a customer ordered tickets for an event, and based on the characteristics of event ticket sales.

5.1.1 Classifiers

We evaluated several classifiers from the Scikit-learn library for the classification task. To ensure optimal performance, we performed k-fold cross-validation for hyperparameter tuning and scaled the data using the MinMaxScaler. The scoring function used for hyperparameter tuning was Scikit-learn's f1_macro, which takes into consideration the F1 scores of both classes, in contrast to the default scoring functions for Scikit-learn classifiers accuracy which does not take recall into account which could lead to an imbalanced classifier.

Table 5.1 Classifiers score methodology comparison

	Average F1 score old methodology	Average F1 score new methodology
Logistic Regression	0.649434	0.658113
Decision Tree Classifier	0.739811	0.707925
Random Forest Classifier	0.734717	0.740943
Gradient Boosting Classifier	0.751509	0.749623
Gaussian Naive Bayes	0.610943	0.600943
K-Nearest Neighbors	0.627736	0.616226
SGD Classifier	0.644906	0.657925

5.1.1.1 Interpretation of Classifiers Prediction

To obtain the probability estimates, we used the Scikit-learn classifier method, predict_proba, which returns the probability for each class for dependent variables. In this case, c_0 represents the probability of sales being lower than the mean, and c_1 represents the probability of sales being higher than the mean. We selected the maximum value from c_0 and c_1 and proceeded accordingly, as described in Algorithm 2.

Algorithm 2 Classifier probability interpretation

```
$c_0, c_1 \leftarrow$ classifier.predict_proba(X)
$m \leftarrow \max(c_0, c_1)$
if $c_1 = m$ then
    return $c_1$
else if $c_0 = m$ then
    return $1 - c_0$
end if
```

5.1.1.2 Classifier Scores

From the table with 5.1 average f1 score on the synthetic testing dataset, we can see that the forests classifier have the best scores among the tested classifiers. ROC curves[1] can be seen in the appendix. However, it is crucial to thoroughly analyze each model's behaviour before online testing to determine if the classifier is compliant with an organizer's requirements and suitable for probability prediction in the context of our two-step model architecture, as this step indirectly influences the ticket price.

One potential issue with some classifiers could be drastic fluctuations in daily prices, which may lead to adverse consequences for ticket sales and customer satisfaction. Therefore, each model should be examined before online testing not only in terms of its overall performance metrics but also by taking into account the smoothness and stability of the predicted price adjustments.

5.1.2 Logistic Function

We modified the logistic function described in the paper on dynamic pricing of airline ancillaries[5] to better suit our model's requirements. The modified function returns a multiplier of the original price, which is then applied to adjust the ticket price. We introduced an offset parameter, o, to

[1] https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc

ensure that the returned multiplier remains within a predefined range from o to L, where o is the minimal price multiplier and L is the maximal multiplier of the original price. The addition of the offset parameter o ensures that the function's output, i.e., the price multiplier, always lies within the defined range $[o,\ L]$. This modification prevents the price multiplier from dropping below the minimum acceptable limit. The defined price function with parameters can be seen in figure 5.1

■ **Figure 5.1** The Logistic function

$$p^{rec\_mult}(x) = o + \frac{L - o}{1 + e^{-k(x - x_0)}}$$

$p^{rec\_mult}(x)$ = predicted multiplier of original price
o = minimal price multiplier
L = maximal price multiplier
k = steepness of the curve
x_0 = the x value of the sigmoid midpoint
x = probability

5.1.2.1 Parameter Tuning

To find the optimal parameters x_0 and k for the logistic function (see Figure 5.1), we employed Scipy's minimize function to search within the possible parameter ranges. We implemented two cost functions to be minimized on the validation dataset, aiming to fine-tune the logistic function parameters according to the classifier's behaviour.

5.1.2.1.1 Cost Function 1: Direction-Based Approach The first cost function takes into consideration whether the price returned by the logistic function was higher or equal to the original price when daily sales were higher than the mean sales, and lower or equal to the original price when daily sales were lower than the mean sales. This cost function ensures that the price adjustments are in the expected direction based on the demand for tickets.

5.1.2.1.2 Cost Function 2: Distance-Based Approach The second cost function builds upon the first one by taking into account both the distance from the mean sales and the distance from the original price. It calculates the distance between the actual daily sales for the dependent variable and the mean sales and multiplies this distance by the $p^{rec\_mult}(x)$. This cost function considers the magnitude of the deviation from the mean sales, theoretically allowing for more fine-grained control over the parameters.

5.1.2.1.3 Exhaustive Search Approach Lastly, we explored an exhaustive search approach, where we went through a finite number of possible parameters and selected the ones with the highest average F1 score on the validation dataset. This approach proved to return the best results among all models and when considering all evaluation metrics.

5.1.2.1.4 Tuning Parameters for Each Classifier We tuned the logistic function parameters separately for each classifier to optimize the parameters according to each classifier's unique behaviour. By doing so, we ensured that the logistic function is well-adapted to the specific characteristics of each classifier, resulting in more accurate price predictions.

5.1.2.2 Parameter Optimization Results

The parameter tuning process involving the direction-based approach, distance-based approach, and exhaustive search approach allowed us to identify optimal logistic function parameters for each classifier. This tailored approach to parameter tuning ensures that the logistic function is well-suited to each classifier's unique characteristics, ultimately leading to possibly more accurate price predictions and a better ticket pricing strategy for event organizers.

5.2 Comparison of Methodology on Synthetic Data

We conducted a performance comparison between the original data collection methodology and the newly proposed methodology in terms of classifier performance and overall performance. The comparison results are essential for understanding the effectiveness of the proposed methodology in addressing the limitations of the original data collection system.

5.2.1 Comparison

As shown in Table 5.1, the comparison indicates that the new data collection methodology does not outperform the original one in terms of classifier performance. This observation is consistent with the overall performance trend. This performance discrepancy suggests two potential explanations:

1. The proposed methodology may not effectively address the limitations of the original data collection system. In other words, the newly designed methodology might not be capturing the necessary information to improve the price prediction model's performance.

2. The synthetic data generator may not be creating data in accordance with the distribution of the new methodology's features. It is essential to ensure that the synthetic data accurately represents the characteristics of the data collected using the new methodology. In our case, the new methodology features were imputed manually, making them susceptible to errors.

5.2.2 Implications and Further Investigation

The performance comparison results underscore the importance of further investigation to determine the underlying cause of the observed performance. If the issue lies with the proposed methodology itself, it may be necessary to identify alternative strategies to address the original system's limitations.

On the other hand, if the synthetic data generator is not accurately representing the new methodology's features, we need to refine the data generation process to ensure that the synthetic data captures the true characteristics of the data collected using the new methodology. This may involve employing more advanced techniques for generating synthetic data that aligns with the distribution of the new methodology's features.

5.3 Evaluation

To evaluate the models we use scores introduced in paper *"Customized Regression Model for Airbnb Dynamic Pricing"*[4], which takes into consideration the predicted price and the situation when the price was predicted. For simplifications and clarification of next section we define P as original price and $P_{rec} = P * p^{rec\_mult}(x)$

5.3.1 Evaluation Metrics

Table 5.2 Defined values for metrics

	Sales higher than mean	Sales lower than mean
Psug $\geq$ P	a	b
Psug $<$ P	c	d

We define a set of metrics according to the defined values in table 5.2.

- Price Decrease Recall $PDR = \frac{d}{b+d}$
- Price Decrease Precision $PDP = \frac{d}{c+d}$
- Price Increase Recall $PIR = \frac{a}{a+c}$
- Price Increase Precision $PIP = \frac{a}{a+b}$
- Price Decrease F1 $PDF1 = \frac{2 * PDR * PDP}{PDR + PDP}$
- Price Increse F1 $PIF1 = \frac{2 * PIR * PIP}{PIR + PIP}$

5.3.2 Results

The results of our experiments, as shown in Table 5.3, indicate that the dynamic pricing model with Random Forest as a classifier delivers the best overall performance among all tested models by a small margin. In particular, the performance of the price decrease aspect is notably better than the price increase aspect among all developed models. This suggests that the models are more adept at offering lower prices during periods of lower demand. Note that the performance of models is evaluated on real test data. That is data that the synthetic data generator was not trained on implying that synthetic data have sufficient quality for training price predicting models.

Table 5.3 Dynamic pricing results

	PDR	PDP	PIR	PIP	PDF1	PIF1
random_forest_classifier	0.927	0.776	0.301	0.612	0.845	0.403
decision_tree_classifier	0.897	0.773	0.314	0.539	0.831	0.397
gradient_boosting_classifier	0.893	0.773	0.317	0.533	0.829	0.398
logistic_regression	0.821	0.763	0.333	0.417	0.791	0.371
k_nearest_neighbors	0.690	0.768	0.456	0.361	0.727	0.403
sgd_classifier	0.847	0.753	0.275	0.409	0.797	0.329
gaussian_naive_bayes	0.862	0.743	0.223	0.383	0.798	0.282

5.3.2.1 Implications for Dynamic Pricing Strategy

The observed results indicate that the developed models are more suitable for adjusting prices downward in response to lower demand. This finding suggests that the models can be effectively used to optimize ticket pricing by offering discounts during periods of reduced demand, potentially leading to increased ticket sales and improved customer satisfaction.

However, the models' relatively weaker performance in predicting price increases suggests that further research and refinements may be necessary to optimize the dynamic pricing strategy during periods of higher demand. Improvements in this area could help event organizers better capture potential revenue opportunities and enhance their overall pricing strategy.

5.3.2.2 Future Research Directions

The results of our experiments highlight the potential of the developed models for dynamic pricing, particularly in the context of price decreases during periods of lower demand. However, there is room for improvement in the models' performance when predicting price increases. Future research could focus on refining the existing models or exploring alternative approaches to better capture the dynamics of price adjustments during periods of higher demand, ultimately leading to more effective and well-rounded dynamic pricing strategies for event organizers.

5.4 Dynamic Pricing Summary

In this chapter, we focused on price prediction using various classifiers and a two-step architecture. The developed models showed a stronger ability to predict price decreases during periods of lower demand, suggesting their suitability for optimizing ticket pricing through discounts and promotions in such situations.

We also explored the parameter tuning process, implementing two cost functions and an exhaustive search approach to find the optimal logistic function parameters for each classifier. This allowed us to ensure that the logistic function is well-adapted to the specific characteristics of each classifier, resulting in more accurate price predictions.

However, the models demonstrated weaker performance in predicting price increases during periods of higher demand, indicating a need for further research and refinements to optimize dynamic pricing strategies in such situations. The performance comparison between the original data collection methodology and the proposed new methodology did not show significant improvement, suggesting that either the proposed methodology was not effective in addressing the original system's limitations or the synthetic data generation process did not accurately represent the new methodology's features.

Overall, this chapter highlights the potential of the developed models for dynamic pricing, particularly in the context of price decreases during periods of lower demand, while pointing to areas where future research could focus on refining the models or exploring alternative approaches to better capture the dynamics of price adjustments during periods of higher demand.

Chapter 6
Conclusion

This thesis provides an exploration of event-based ticket sales data, covering a range of aspects from literature surveys to practical data analysis. Throughout the research, insights into the potential challenges associated with event-based ticket sales were identified, serving as a foundation for future advancements.

The literature review chapter established an understanding of the latest developments in dynamic pricing and sales forecasting. By examining prominent models and techniques, we gained insights into their strengths, weaknesses, and effectiveness in the ticket sales domain. The comparison of supervised and unsupervised learning further illuminated the various approaches that can be applied in this context.

The data analysis unveiled both opportunities and obstacles within the available data sources. While we identified positive findings, such as increased demand at the start and end of ticket sales periods and the impact of holidays on sales, we also encountered challenges like low scanned tickets, overestimated event capacity, and issues with data Google website tracking application.

To address these challenges, we proposed a refined data collection strategy aimed at enhancing accuracy by incorporating data from various sources, such as social media, economic measures, meteorological data, and competitor event information. We also developed a synthetic data generation model to improve the training dataset for ML models. Although the comparison between data collection methodologies did not show significant improvement, this approach offered insights into the potential of synthetic data and identified areas for further refinement.

In our exploration of dynamic pricing, we experimented with various classifiers with two-step architecture for price prediction. The models demonstrated a stronger ability to predict price decreases during periods of lower demand, suggesting their potential for optimizing ticket pricing through discounts and promotions. We also highlighted areas for future research, such as refining the synthetic data generation model and exploring the integration of broader and more robust data sources.

Overall, this research contributes to the field of event-based ticket pricing by investigating a range of approaches, identifying challenges, and proposing potential solutions. Although some limitations were encountered, the work can serve as a foundation for further advancements in event-based ticket price prediction.

Future Work

Several opportunities for future work were identified throughout the thesis, we outline the ones which we consider to be the most critical.

Improving Data Collection A noticeable challenge encountered during the research was the quality of the website tracking information. Future work can focus substitution of the information which describes customers' online activity with data from social media information preferably Facebook Graph API [1].

Enhancing Synthetic Data Generation While the synthetic data generation model developed was sufficient for dynamic pricing model training, there is room for improvement. Future studies can explore more complex synthetic data generation methods, such as generative adversarial networks or variational autoencoders, to produce more realistic and diverse synthetic datasets.

Measuring the Impact of Dynamic Pricing An important aspect to consider in future work is the evaluation of the actual impact of dynamic pricing strategies on event-based ticket sales. This can be achieved by implementing a controlled testing environment, such as an A/B testing framework, where different pricing strategies can be compared and analyzed for their effectiveness in increasing sales, revenue, and overall customer satisfaction.

[1] `https://developers.facebook.com/docs/graph-api`

Appendix A

appendix

In this appendix, we are going to look at additional visualizations.

A.1 Additional Google web Tracking Comparison

in this section, we will look at additional comparisons of GA3 versus GA4 metrics. The methodology of comparison can be found in the Enclosed material path `src/data_transformation_and_sources/google/quiries_params`.

Figure A.1 ga3-ga:bounces comparison to ga4-bounces

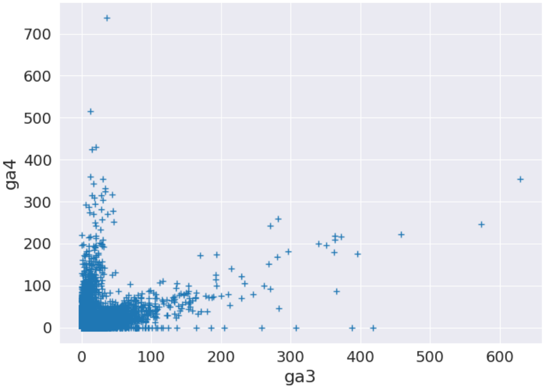


Figure A.2 ga3-ga:bounceRate comparison to ga4-bounceRate

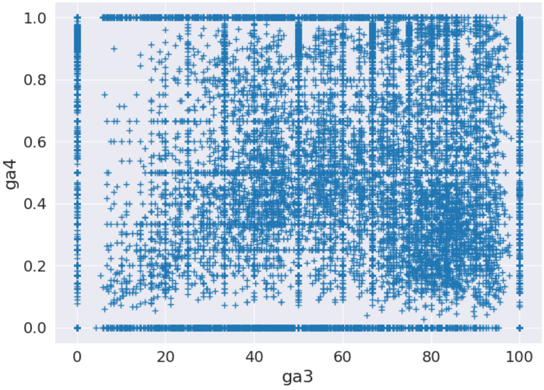


Figure A.3 ga3-ga:pageviews comparison to ga4-screenPageViews

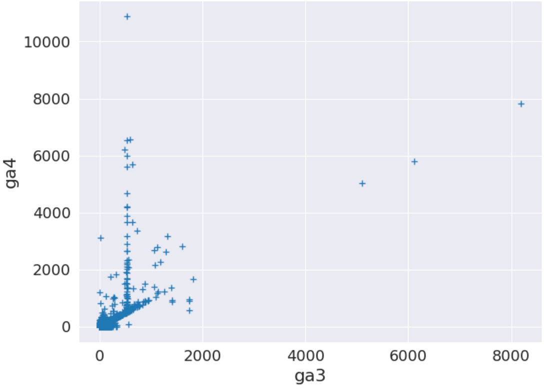
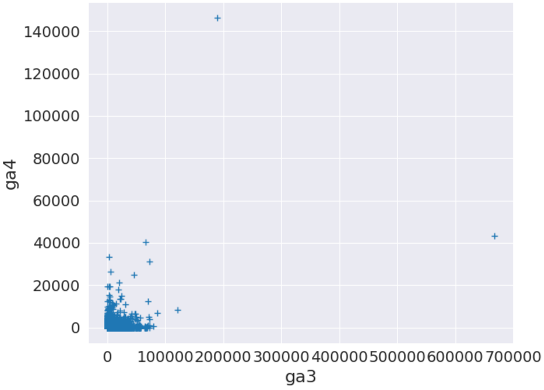
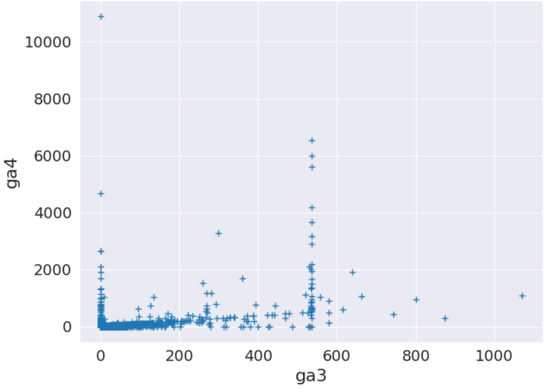


Figure A.4 ga3-ga:sessionDuration comparison to ga4-userEngagementDuration



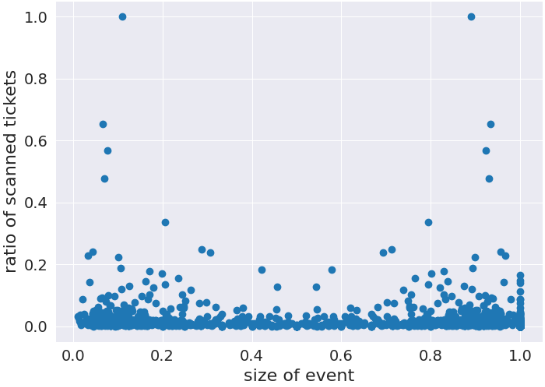
■ **Figure A.5** ga3-ga:pageviewsPerSession comparison to ga4-screenPageViewsPerSession



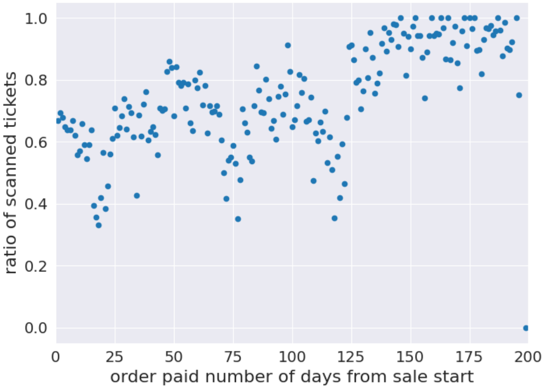
A.2 Additional Scanned Tickets Visualizations

These visualizations were made in purpose to explore the phenomenon of large quantities of unscanned tickets that can be caused. Tickets can be scanned for two reasons appropriate event worker did not scan the ticket or the customer did not use the ticket.

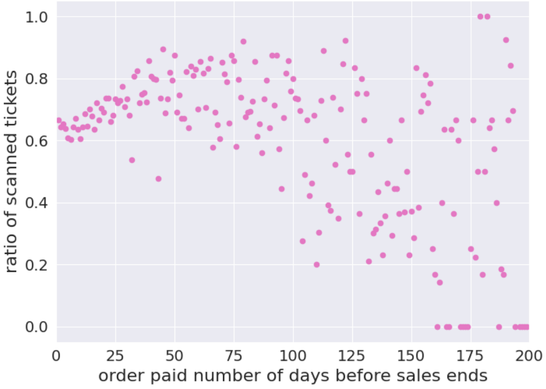
■ **Figure A.6** scanned tickets ratio per size of the event size



■ **Figure A.7** scanned tickets ratio per day from orderable

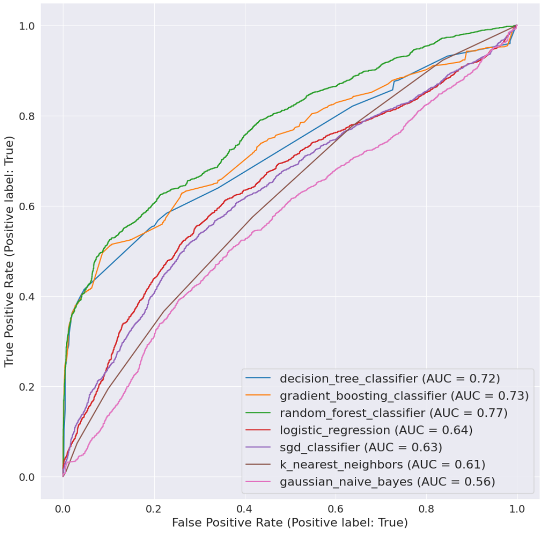


■ **Figure A.8** scanned tickets ratio per day until orderable



A.3 Additional Visualizations Classifiers

■ **Figure A.9** Classifier ROC Curves Old Methodology



Bibliography

1. SUHER, Jacob. Forecasting Event Ticket Sales. 2008. Available also from: https://repository.upenn.edu/wharton_research_scholars/49/.
2. AHN, Hyung-Il; SPANGLER, W. Scott. Sales Prediction with Social Media Analysis. In: *2014 Annual SRII Global Conference*. 2014, pp. 213–222. Available from DOI: 10.1109/SRII.2014.37.
3. CHAING, Ton; RAU, Hsin; SHIANG, Jung Wei; CHIANG, Luen Jon. Social media marketing in the sales volume prediction for the Lolita fashion brand. 2021. Available also from: https://www.preprints.org/manuscript/202111.0227/v1.
4. YE, Peng; QIAN, Julian; CHEN, Jieying; WU, Chen-hung; ZHOU, Yitong; DE MARS, Spencer; YANG, Frank; ZHANG, Li. Customized regression model for airbnb dynamic pricing. In: *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*. 2018, pp. 932–940. Available also from: https://dl.acm.org/doi/10.1145/3219819.3219830.
5. SHUKLA, Naman; KOLBEINSSON, Arinbjörn; OTWELL, Ken; MARLA, Lavanya; YELLEPEDDI, Kartik. Dynamic pricing for airline ancillaries with customer context. In: *Proceedings of the 25th ACM SIGKDD International Conference on knowledge discovery & data mining*. 2019, pp. 2174–2182. Available also from: https://arxiv.org/abs/1902.02236.
6. BRANDA, Francesco; MAROZZO, Fabrizio; TALIA, Domenico. Ticket sales prediction and dynamic pricing strategies in public transport. *Big data and cognitive computing*. 2020, vol. 4, no. 4, p. 36. Available also from: https://www.mdpi.com/2504-2289/4/4/36.
7. JULIANNA, Delua. *Supervised vs. Unsupervised Learning: What's the Difference?* [online]. 2021-03. [visited on 2023-05-01]. Available from: https://www.ibm.com/cloud/blog/supervised-vs-unsupervised-learning.
8. PEIXEIRO, Marco. *The Complete Guide to Time Series Analysis and Forecasting* [online]. 2023-04. [visited on 2023-04-20]. Available from: https://towardsdatascience.com/the-complete-guide-to-time-series-analysis-and-forecasting-70d476bfe775.
9. IVAN, Despot [online]. 2022-09. [visited on 2023-05-02]. Available from: https://www.timescale.com/blog/what-is-time-series-forecasting/.
10. GUREJA, SRISHTI. *Difference between use cases of expanding and rolling window in backtesting* [online]. [visited on 2023-05-01]. Available from: https://stats.stackexchange.com/questions/568814/difference-between-use-cases-of-expanding-and-rolling-window-in-backtesting. (version: 2022-12-24).
11. *[ga4] analytics dimensions and metrics - analytics help* [online]. Google [visited on 2023-04-04]. Available from: https://support.google.com/analytics/answer/9143382?hl=en..

12. *[ga4] metric - analytics help* [online]. Google [visited on 2023-04-12]. Available from: `https://support.google.com/analytics/answer/9355664?sjid=11838825629560389013-EU`.

13. *Google analytics architecture - a look under the hood to explore apis - yasen lilov: Blog* [online]. 2018-08. [visited on 2023-04-04]. Available from: `https://jsndesign.co.uk/blog/google-analytics-architecture-apis/`.

14. MARNEWICK, Ghia. *10 key differences between google analytics 4 (GA4) and Universal Analytics (GA3)* [online]. Aumcore, 2023-01 [visited on 2023-03-04]. Available from: `https://www.aumcore.com/blog/google-analytics-4-and-universal-analytics/`.

15. *[ga4] data thresholds - analytics help* [online]. Google [visited on 2023-04-05]. Available from: `https://support.google.com/analytics/answer/9383630?hl=en`.

Enclosed Material

- `readme.md` .... sources documentation in Markdown format
- `readme.pdf` .... sources documentation in PDF format
- `src` .... directory with implementation
  - `analysis` .... source codes for visualizations
  - `data` .... directory with synthetic data and original directory for classified data
  - `data_transformation_and_sources` .... data transformation and sources
  - `models` .... source codes of developed models
  - `tex` .... directory with latex source codes
- `thesis.pdf` .... thesis in pdf format