ETH
Swiss Federal Institute of Technology Zurich

Seminar for Statistics

Department of Mathematics

Master Thesis Winter 2017/2018

Michael Heinzer

Predicting the Success of Hotels with Online Ratings

Submission Date: March 26th 2018

Co-Adviser: Prof. Dr. Nicolai Meinshausen
Advisers: Prof. Dr. Elgar Fleisch, Daniel Müller

Preface

First of all, I would like to thank Daniel Müller for his feedback, guidance, and enthusiasm during the entire thesis. Also, I would like to thank Dominik Bettler for the frequent discussions and helpful suggestions. Another thank you goes to Prof. Meinshausen, for making this collaboration with MTEC possible. And last but not least, I would like to thank my family for their support during my studies.

Abstract

This thesis addresses the problem of how to predict business performance, as measured by revenue, RevPAR and growth from online ratings. To conduct this study, we received multi-year revenue data from a Swiss insurance company. We collected online ratings from TripAdvisor, Booking.com and Google. Other online data of the hotels was retrieved from Swisshotels and Federal Statistics Office.
In a first step we built multiple models to predict the revenue. Secondly, we applied the same models towards revenue per available room (RevPAR) prediction. Thirdly, we created a model to classify growing and shrinking hotels.
Our advanced model outperformed baseline models in revenue prediction. The number of online reviews was among the most significant predictors. RevPAR predictions failed to outperform the baseline model significantly. Our model was able to classify growing and shrinking hotels better than the baseline. Moreover further model analysis indicated that changes in online ratings were among the most predictive attributes for growth in our data.

Contents

List of Figures

List of Tables

Notation

Abbreviations

ANN Artificial Neural Net

BT Boosted Trees

CLV Customer Lifetime Value

CRM Customer Relationship Management

CSS Cascading Style Sheets

CSV Comma Separated Values

eWOM Electronic Word of Mouth

FSP Financial Service Provider

GPS Global Positioning System

ID Identifier

IDSS Intelligent Decision Support System

MAE Mean Average Error

MCAR Missing Completely At Random

MLR Multiple Linear Regression

MP Mean Prediction

OLS Ordinary Least Squares

RevPAR Revenue Per Available Room

RMSE Root Mean Square Error

SME Small and Medium-Sized Enterprises

SQL Structured Query Language

SVM Support Vector Machine

URL Uniform Resource Locator

XGBoost eXtreme Gradient Boosting

XPATH XML Path Language

Chapter 1

Introduction

In this chapter, we first give an overall description of the problem from the perspective of an insurance company and rationalize why it is worth solving. Then, we explore the relevance of online data sources and demonstrate their use in various fields to explain or predict customer behavior and sales. Lastly, we outline how our contribution addresses the problem.

1.1 Motivation

In Switzerland, small and medium-sized enterprises (SMEs) comprise 99% of all enterprises and create two-thirds of jobs FSO (2017). Financial service providers (FSPs), such as banks and insurance companies, have robust relationships with many of these SME enterprises. Hence, it is a common interest of both societies and FSPs to understand drivers of business growth or a decline of SMEs.

For FSPs, maintaining a holistic customer perspective is a non-trivial task, as it requires considerable resources to acquire and process the necessary data. However, once such a system is achieved, it can provide considerable advantages. For example, it can automate manual administrative tasks and optimize portfolio structures. It could also allow for a ranking of the most valuable customers in order to better allocate the limited resources of the customer relationship/sales personnel. Nonetheless, many of those benefits would only be achievable through the automatic collection and processing of a wide range of information Cronin (1997).

An intelligent decision support system (IDSS) can provide such benefits and improve decision making Scott and Bruce (1987). An IDSS is a combination of traditional decision support systems, which provide aggregated information, and newer techniques, such as machine learning, which extracts new insights from the collected data. To achieve such an IDSS, FSPs need to collect and aggregate all available data sources that relate to their customers. Only in the last decade it has become possible, even for smaller companies, to acquire and process this information. Leveraging state-of-the-art methods in information technology is becoming essential for FSPs Soni and Duggal (2014). In the case of FSPs covering commercial liability risks, a key input factor of the underwriting process is

the revenue, which must be frequently updated to insure proper coverage. However, in a competitive market, resources to investigate are scarce, and self-reporting of customers is limited. Up-to-date revenue data is important since under-insured companies are particularly problematic for two reasons. First, in the case of an incident, the insurance payout might not fully cover the costs, which is likely to displease the customer. Second, under-insured companies lead to a loss of revenue for FSPs but lower premiums for customers, which offers another incentive to under-report changes in revenue.

This thesis investigates how FSPs can improve their customer management in the hotel industry through the use of online data that can help to categorize and prioritize customers. This problem is worthy of examination since the overall tourism industry is a significant sector in the Swiss economy, with 16.4 billion CHF gross value added and over 160,000 jobs in 2016 FSO (2016).

1.2 Problem Statement

The arrival of the Internet economy has drastically transformed many sectors, and the most famous is probably the advertisement industry. An important aspect of this transformation has been the harnessing of information that consumers provide on the Internet. Part of this information, which is publicly accessible, is consumer-generated reviews for a wide range of products. The relevant literature commonly refers to those reviews as electronic word of mouth (eWOM). Critical factors of eWOM are the rating, usually an average of ratings (valence), the volume of reviews, and the variance of ratings.

Electronic word of mouth evidently relates to sales for a range of products. In the movie industry, eWOM is essential for box office performance both as a precursor and for the outcome. While the rating has no significant impact on box offices revenues, the volume of online postings is positively related Duan, Gu, and Whinston (2008b). This relationship could be an indicator of the underlying intensity of eWOM for a specific movie Duan, Gu, and Whinston (2008a). In the book industry, an increase in online ratings on the Barnes & Noble website prompts an increase in relative sales on the site Chevalier and Mayzlin (2003). On Amazon, more reviews and recommendations also lead to additional sales S.-y. Chen Pei-Yu; Wu and Yoon (2004).]. Currently, customer recommendations generate one-third of sales McKinsey (2017). However, recommendations are more important for lesser-known books than they are for famous ones S.-y. Chen Pei-Yu; Wu and Yoon (2004), an effect that has also been demonstrated to apply to games Zhu and Zhang (2010). While the rating has no influence on absolute sales S.-y. Chen Pei-Yu; Wu and Yoon (2004), it does improve relative sales on the two aforementioned sites Chevalier and Mayzlin (2003). Still, consumers tend to pay more attention to the content of reviews than to summary ratings Chevalier and Mayzlin (2003).

This thesis focuses on the hospitality industry. In this industry, service is a crucial component of the business. A significant portion of potential customers research their destinations online before making any transactions. According to Toh, Raven, and Dekay (2011), 8 out of 10 leisure travelers search for hotels online, and 67% of them book online. In accordance with previous findings, Agius (2014) has found that 22% of consumers

always read reviews, while 43% read them most of the time. Sixty-seven percent indicated that they read at least four reviews. Reviews written by consumers are perceived as more helpful than reviews from experts M. Li, Huang, Tan, and Wei (2013) and marketer-generated content Goh, Heng, and Lin (2012). Perceived communicator characteristics often determine the strength of eWOM's influence Wangenheim and Bayón (2004). Research has suggested that consumer reviews, or eWOM, generate more trust compared to simple numerical ratings Pavlou and Dimoka (2006). This phenomenon also applies in general electronic marketplaces Ba and Pavlou (2002). Trust resulting from eWOM can in turn have a positive impact on booking intention Kim, Kim, and Park (2017) Moloi (2016) and can help explain price premiums for certain sellers Pavlou and Dimoka (2006). Nevertheless, it is not only the quality of reviews that influences booking behavior but also the quantity. Dipendra Singh (2015) has found that the number of reviews and online ratings positively impacts booking transactions. These findings have been validated for a variety of platforms, such as TripAdvisor Edwin N. Torres (2016) Tuominen (2011) Blal and Sturman (2014), TraveloCity.com C. Anderson (2012), trustyou.com Phillips, Barnes, Zigan, and Schegg (2016), Booking.com Viglia, Minazzi, and Buhalis (2016) Öğüt and Taş (2012), and Venere.com Viglia et al. (2016).

Another factor that influences booking behavior is the interplay between ranking and rating on search engines Anindya Ghose (2014), and this is disproportionally true for higher-class hotels. Moreover, it is more important to match average competitor ratings than to surpass them Karaman (2017). These factors can incite some hotels to fabricate reviews Mayzlin, Dover, and Chevalier (2012). Another problem with ratings is that early ratings tend to be more negative than later ones Melián-González, Bulchand-Gidumal, and López-Valcárcel (2013), and reviews in general tend to be overly positive. Thus, the average rating does not correspond to quality in every case Hu, Pavlou, and Zhang (2006).

Key determinants of room prices in New York include not only reviews that influence potential customers but also the room quality and location Zhang, Ye, and Law (2011). Research has demonstrated an influence of geographical region on ratings Molinillo, Ximénez-De-Sandoval, Fernández-Morales, and Coca-Stefaniak (2016). Nevertheless, customer preferences are changing rapidly; while it was once important to have a phone in the room, WiFi is now a top priority G. Li, Law, Vu, Rong, and Zhao (2015). Hence, it is necessary to have a holistic view of the hotel and collect as many attributes as possible.

To summarize, not rely solely customer review data is important, even if they evidently have a significant impact. It is imperative to incorporate diverse datasets in order to more accurately predict consumer behavior and hotel revenue.

1.3 Own Contribution

As outlined above, an important part of an IDSS for FSPs is the revenue of a business; for this element, up-to-date data are often missing and expensive to collect. Our first contribution is the creation of a model that uses data from only online sources to identify possible discrepancies in insured revenues and predicted revenues. To compare those findings with existing literature, we repeat the prediction process for revenue per available room (RevPAR). As a second contribution, we illustrate the ability of online data to sepa-

rate growing businesses from shrinking ones, which can help FSPs in identifying customers who most urgently need an adapted policy.

In pursuit of this goal, we received revenue data of hotels from a Swiss insurance company for the years 2008 to 2016. Moreover, we collected data from multiple online platforms, including TripAdvisor.com, Booking.com, and Google.com, that amass consumer feedback for the hotel industry. To account for hotel properties, such as location, size, and stars, data was acquired online from Hoterlleriesuisse, the Swiss hotel association. Finally, the Swiss Federal Statistical Office provided economic data for different regions of Switzerland.

To verify the accuracy of the received data. We plotted the yearly changes in revenue of all hotels in Switzerland, and those in our dataset in Figure 1.1. Overall, the insurance dataset has fluctuates less, most likely due to the under-reporting bias described earlier.

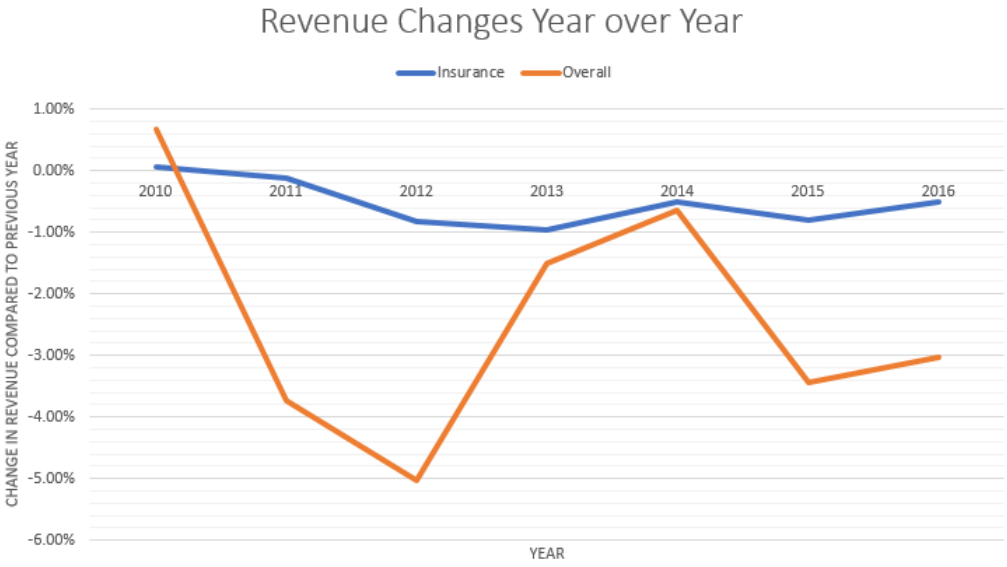

Figure 1.1: Yearly revenue changes for hotels, for the insurance dataset and for the entire hotel industry in Switzerland

However the general distribution of hotels within a specific revenue range is matching the distribution of all hotels in Switzerland. Figure 1.2 displays the general pattern. There is a subtle bias towards higher revenue hotels in the upper revenue ranges.

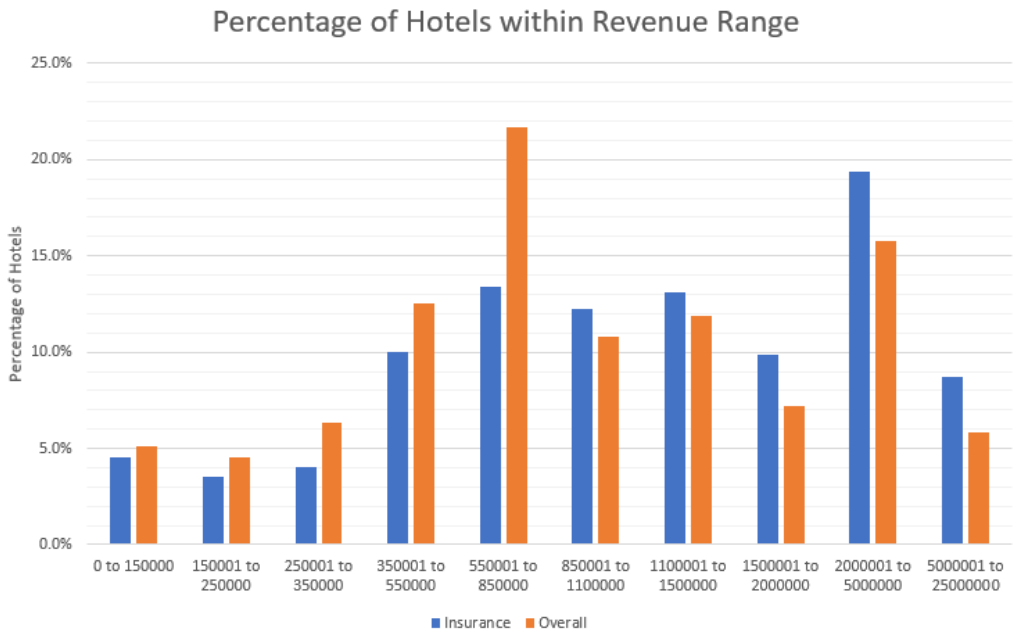

Figure 1.2: Percentages of hotels in Switzerland and in our insurance dataset ordered by revenue range

Figure 1.3 shows the geographical distribution of the insurance data. Clusters in larger cities, such as Zürich, Geneva and Basel are recognizable.

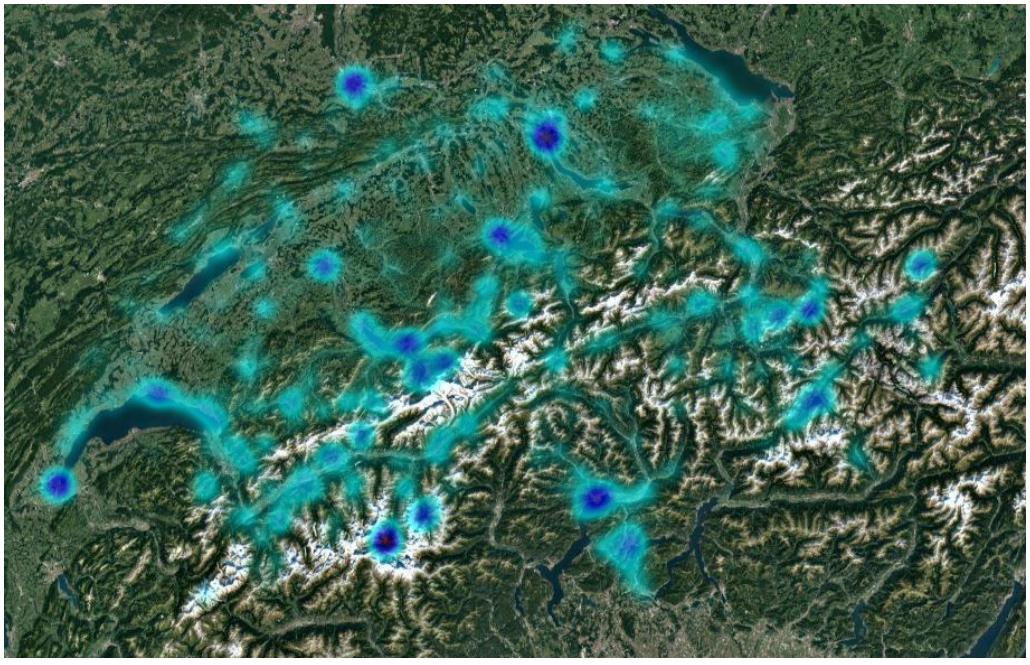

Figure 1.3: Geographical distribution of insurance data on Swiss map

We conclude our data is representative for Switzerland, with respect to geographical and revenue distribution.

The structure of this paper is as follows, after outlining the problem in this chapter, we present previous and related work on the topic. Then we describe the process of acquiring and processing the data. Thirdly, results of different models are given. Lastly, we interpret the data and give advice for improvement.

Chapter 2

Literature Review

This thesis is based on previous findings in multiple areas. This chapter reviews literature regarding eWOM and hotel sales, the point of view of financial institutions, risk predictions, and forecasting in tourism.

2.1 Predicting Business Performance

Muller, Te, Meyer, and Cvijikj (2016) have attempted to use quantitative models to predict business performance for SMEs in Switzerland. Their study classified 7,064 SMEs into revenue bins using a random forest classifier. Moreover, it fitted a multiple linear regression model to perform inference on company characteristics.

For financial performance, not only sales are relevant but also the risk of bankruptcy. Traditional methods include the creation of efficacy measures for businesses using non-parametric data and subsequently applying regression methods. These efficacy measures have significant explanatory power for predicting the likelihood of defaults in French Maria Psillaki (2010) and Indian Kumar (2017) markets. Bankruptcy prediction is intimately related to credit risk, for which machine-learning models based on support vector machines (SVMs) delivered the best performance among many models Yu (2008). By incorporating machine-learning algorithms into customer relationship management (CRM) software, enterprises can effectively manage those relationships from a risk perspective Lai, Yu, Wang, and Huang (2007). Bayesian-based SVM data-mining models have again exhibited a superior performance compared to testing samples Lai et al. (2007).

Social media and machine learning can also support public safety. Approximately one in six Americans becomes sick every year due to food poisoning. Of these cases, an estimated 60% are caused by insufficient hygiene in restaurants. In view of this, a restaurant inspection program was started in New York City. However, this program conducts inspections only once per year, which could allow restaurants to shirk their responsibilities. Jorge Mejia (2018) has suggested using social media data to analyze complaints and identify public health hazards in ”real time.”

2.2 eWOM in Tourism

For the hotel industry specifically, there has been a wealth of studies on online data and sales. This research ordered the studies by regional market and began with papers on Chinese hotels. Ye, Law, and Gu (2009) have used online data, which were collected with a web crawler, to make inferences about hotel room sales. They collected reviews from ctrip.com for 248 hotels in three cities. However, they estimated sales data according to the number of reviews. The study found a significant relationship between online consumer reviews and business performance, and this finding was repeated for 1,639 hotels in over 10 cities in China Ye, Law, Gu, and Chen (2011). Both studies used a log linear regression model for inference. With similar data and the same sales estimation method, Qi and Qiang (2013) and Lu, Ye, and Law (2014) have determined that the average rating and the rating variance have a significant impact on sales that is moderated by the star rating. Lu, Xiao, and Ye (2012) have contradicted this finding, however, in reporting that the average rating of customers is more important for hotels with higher star ratings.

For the U.S. market, C. Anderson (2012) has analyzed 13,341 reservations from an online travel agency using consumer data from TripAdvisor and comScore. The percentage of consumers who consulted reviews on TripAdvisor steadily increased during the time period of 2008 to 2010. Moreover, Travelocity data revealed that a one-point rating increase on a five-point scale would allow for a price increase of 11% without loss of market share. With logistic regression, C. Anderson (2012) calculated that a 1% increase in reputation would lead to a 1.42% increase in revenue per available room (RevPAR). The researcher later validated that finding in C. K. Anderson and Lawrence (2014) in illustrating that a 1% reputation increase leads to 0.99% increase in RevPAR, but the influence of eWOM monotonically decreases as the hotel class increases.

Various studies have also examined the European market context. For example,Öğüt and Taş (2012) have examined data from Booking.com for 388 hotels in London and 562 hotels in Paris for the beginning of 2009. They estimated sales based on the number of reviews per hotel. The researchers concluded that a 1% increase in the online customer rating increased RevPAR by 2.68% in Paris and 2.62% in London. Moreover, prices of high-star hotels are more sensitive to customer ratings in comparison to lower-star hotels. Blal and Sturman (2014) have validated this finding in a study of 319 hotels in London in 2011 that employed real sales data from STR Global and eWOM from TripAdvisor. In addition, they observed that the volume of reviews has a more substantial effect on low-tier hotels Blal and Sturman (2014). Other studies have identified correlations between hotel performance and both the number and ratings of reviews for hotels in Stockholm, Copenhagen, Oslo, Helsinki, and Tampere Tuominen (2011). Raguseo and Vitari (2017) have used performance data from STR Global to randomly select 221 French hotels with available revenue data for the period 2005 to 2013. They also collected eWOM data from TripAdvisor to examine effects on branded and non-branded hotels. They reported that the volume of reviews has no effect on RevPAR growth for branded chain hotels and a positive effect for non-branded hotels. Moreover, the effect of valence on RevPAR growth applied only to non-branded hotels. Viglia et al. (2016) have revealed, via regression analysis, that there is a positive effect of the review score on the occupancy rate in Rome.

One of the first studies to link online data to revenue in Switzerland has found that website adoption relates positively to RevPAR Scaglione, Schegg, and Murphy (2009). Hotels without an online presence exhibited a decline in revenue in between 1992 and 2003. This study was based on data from 147 hotels. Phillips et al. (2016) have used partial least squares modeling to test a number of hypotheses on data of 442 Swiss hotels. Results indicate that attributes such as quality of rooms, Internet provision, and building have the highest impact on hotel performance. Using a smaller sample of 235 Swiss hotels and online data from TrustYou, Phillips, Zigan, Silva, and Schegg (2015) have investigated relationships between eWOM, hotel characteristics, and RevPAR. Through the use of multiple linear regression (MLR) and an artificial neural net (ANN), Phillips et al. (2015) predicted RevPAR from the following 10 attributes: canton, region, stars, quality label, number of rooms, number of beds, TrustYou score, number of sources, number of reviews, and percentage of positive reviews. The reported root mean square error (RMSE) was 65.615 for the MLR model and 0.015 for the ANN, while the mean RevPAR was 112.28. The reported R2 value was 0.39 for the MLR model and 0.99 for the ANN. However, we believe that these results do not accurately reflect the predictive capabilities of the data, as the study did not split training and test data, which led to an over-fit of the ANN.

2.3 View of Financial Institutions

Modern economies have become increasingly service based and derive revenue from the creation and sustenance of long-term relationships with their customers Gupta, Hanssens, Hardie, and Kahn (2006). One metric to measure the value of a customer is customer lifetime value (CLV). Customers who are selected for their CLV provide higher profits in future periods compared to those who are selected on the basis of other metrics Venkatesan and Kumar (2004). Hence, corporate success depends on the maintenance of valued customer relationships, for which it is essential to adapt segment customers based on various characteristics. One such characteristic is growth: out of a hundred randomly selected companies, four firms will create 50% of job growth over a decade Storey (1994). Identifying and keeping these firms as customers is key for FSPs.

To summarize, the literature evidences that online data are related to sales for many markets and models. There have also been attempts to measure and explain hotel sales in Switzerland. We next seek to validate those findings by using similar data to predict revenue and RevPAR. Moreover, we demonstrate that eWOM data have the power to distinguish between hotels with growing and shrinking revenues.

Chapter 3

Methodology

This chapter specifies our data sources for both online data and data received by the insurance company. It then explains how we collected the data and pre-processed them for prediction. Since the data from diverse sources were not always uniquely related, we also describe how we matched the entries from different databases. Not all the hotels were present on all platforms, which led to missing data. We handled this case with a technique called imputation. Finally, the chapter provides some background information on the employed models and the error functions.

3.1 Sample Description

This section briefly overviews the diverse datasets we used in this work. For illustrative purposes, we present the case of the "Great Hotel," a fictional property in Zurich.

3.1.1 Insurance

We were able to extract 667 entries from the insurance company database, including the name, address, owner, and revenue of the customer. However, some of these businesses were duplicates, out of business, or shell companies. We then manually reviewed each entry to find the corresponding URL for TripAdvisor and Booking.com. This manual search resulted in 374 entries for TripAdvisor and 356 for Booking.com. The data did not fully overlap.

3.1.2 TripAdvisor

TripAdvisor was the most thoroughly explored data source. Visitors to this site can explore, book, and review hotels. The main page offers a quick overview of the hotel's rating by fellow travelers, the ranking in the city, and any special amenities it might offer. The rating values range from 1 to 5 in 0.5 increments, and they are rounded up or down from the raw value that is calculated from the reviews. We recorded 21 general attributes per hotel.

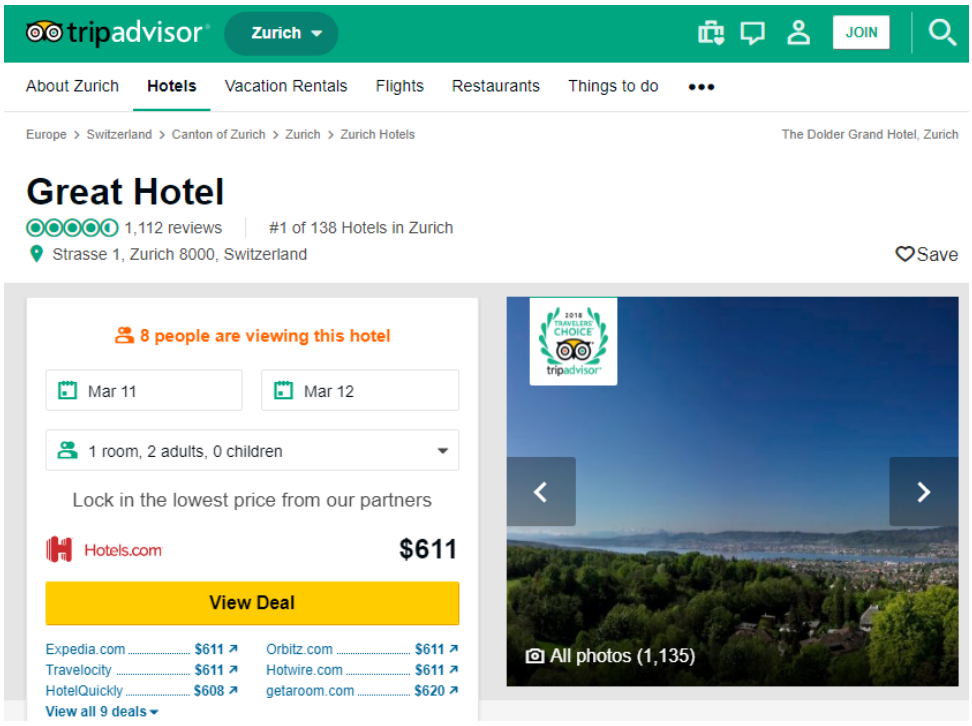

Figure 3.1: Overview of TripAdvisor site for the fictional hotel on 27.02.2018

Some of those attributes appear in Figure 3.2, such as the number of stars, the usual price range, and the number of rooms. In the case of the Great Hotel, there would be many more attributes available, but this is not usually true for smaller hotels.

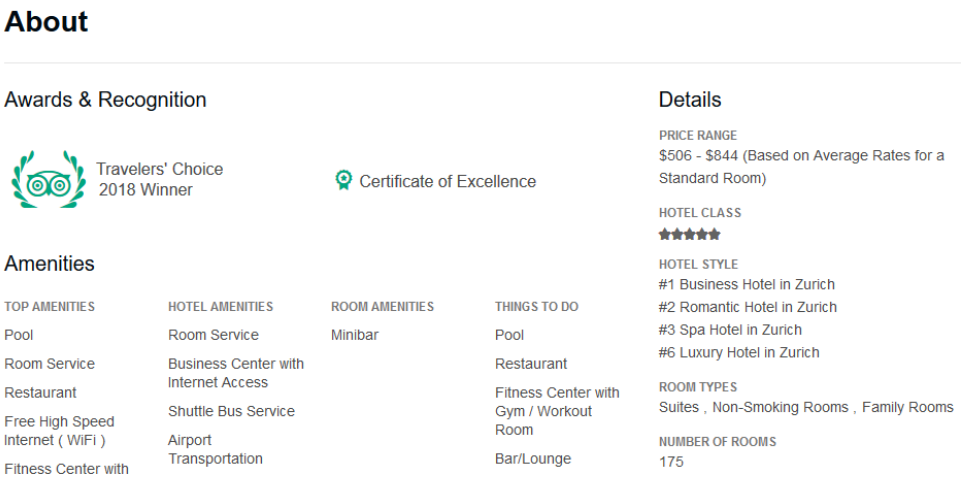

Figure 3.2: Overview of TripAdvisor details for the fictional hotel on 13.02.2018

We decided to crawl the entire accommodation database of TripAdvisor and collect the 21 attributes for all accommodation providers in Switzerland. This occurred in two phases. First, we crawled the site for every canton and created a list of all accommodation offers on the website. This yielded a list of 6,855 entries with not only hotels but also bed and breakfasts (B&Bs), hostels, and special lodging. Figure 3.3 indicates the three categories below.

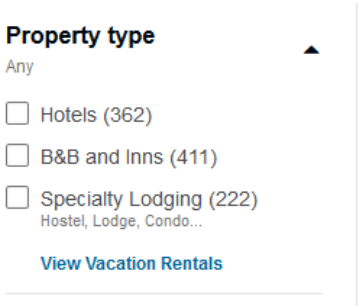

Figure 3.3: Overview of different Lodging Types available on 14.02.2018

For the 6,856 hotels in the database, we collected the summary review count in all languages, as Figure 3.4 depicts, but not the individual reviews themselves.

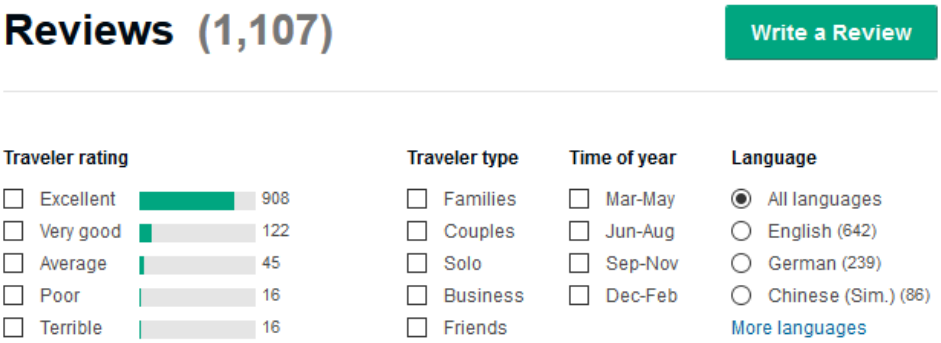

Figure 3.4: Review summary offered for the fictional hotel on 13.02.2018

For the 374 hotels that we could link with the insurance database, we collected every review. An example is given in the lower part of Figure 3.5. The amassed data included the rating, the title, the date of the review, and the beginning of the review text. This yielded a data set of 51,514 reviews in various languages.

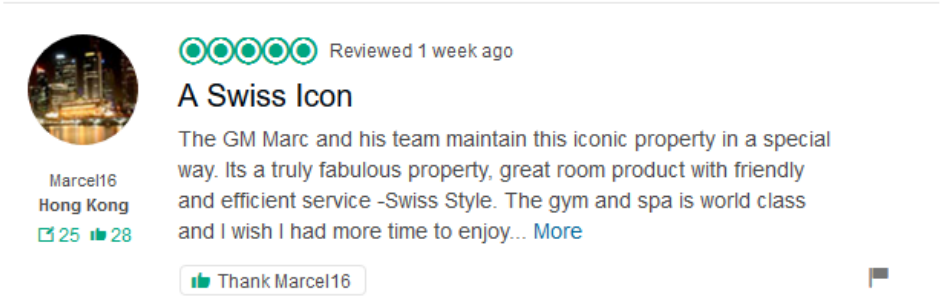

Figure 3.5: A sample review of the fictional hotel on 13.02.2018

3.1.3 Booking.com

As the name suggests, the Booking.com website is mainly for online booking of accommodation around the world. It features a rating from 1 to 10 in 0.1 increments. Furthermore, it allows for more nuanced ratings in seven sub-categories: cleanliness, comfort, facilities, staff, value for money, free WiFi, and location. It also features reviews written by clients.

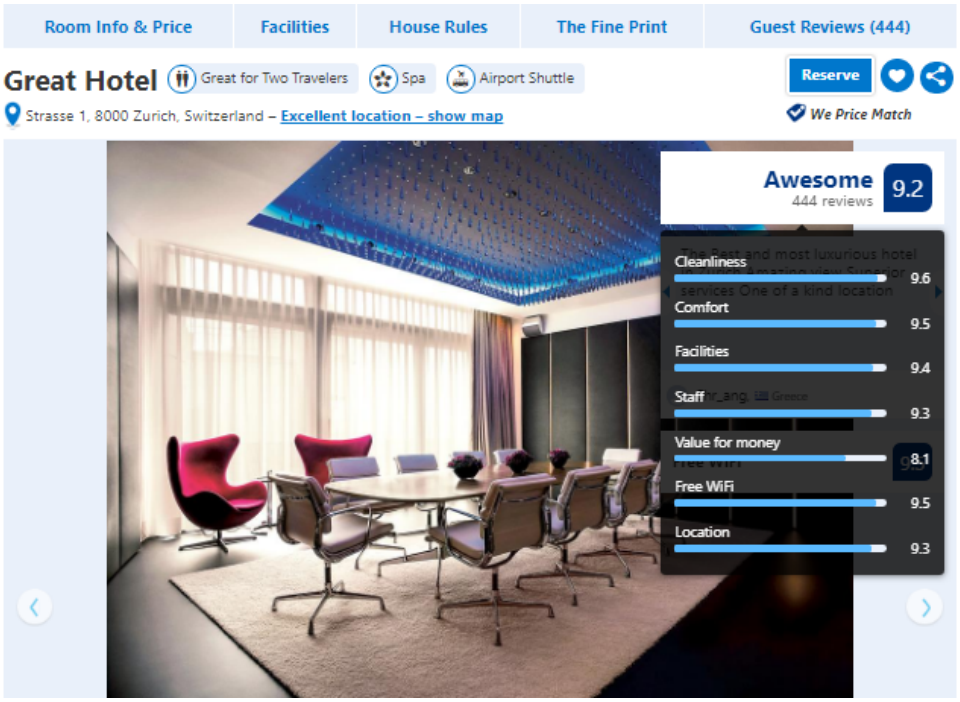

Figure 3.6: Overview of the Booking.com site for the fictional hotel on 27.02.2018

3.1.4 Google

Since it is integrated into Google Search and Maps, Google enables people to review certain places. For hotels, it features a rating from 1 to 5 in 0.1 increments (see Figure 3.7). It also offers the possibility to write a more detailed review. We searched for Google entries based on the names and cities collected from the previous crawls on TripAdvisor and Booking.com. In total, we were able to find 389 hotel ratings on Google that corresponded to an entry in the insurance database.

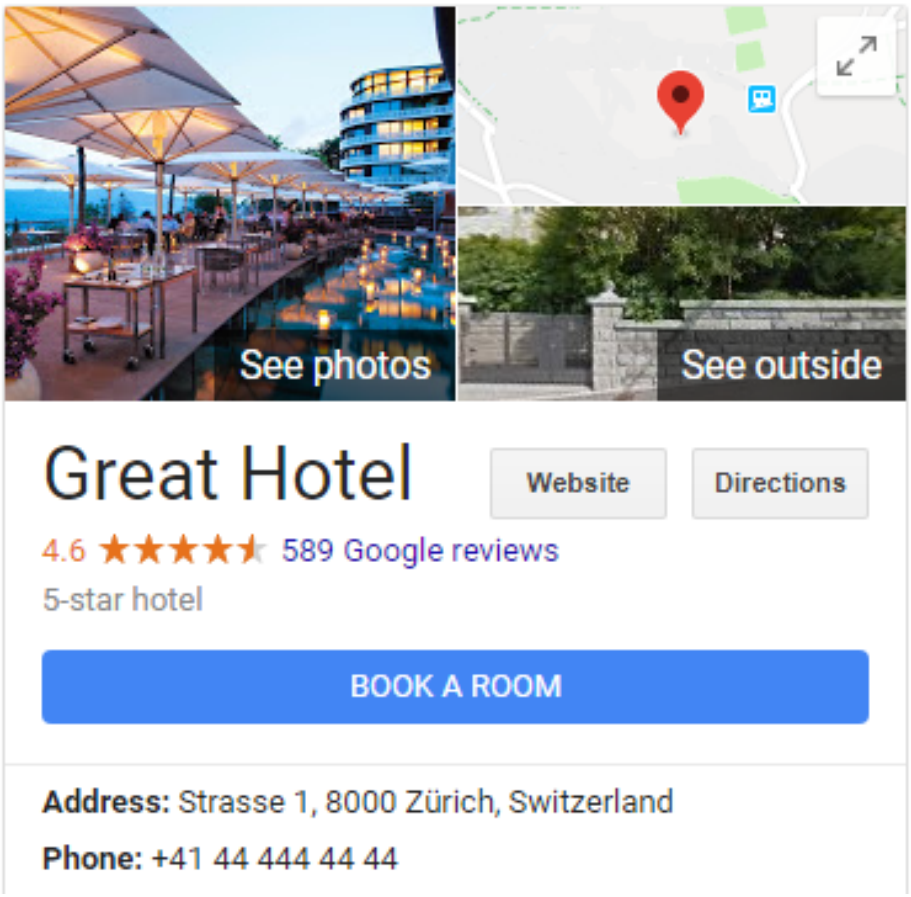

Figure 3.7: Overview of the Google site for the fictional hotel on 27.02.2018

3.1.5 Swisshotel

Swisshotel is a database of the Swiss Hotel Association. According to its own description, it lists all hotels in Switzerland alongside a portrait. It contains no user ratings, though it does feature far more details than the platforms mentioned above.

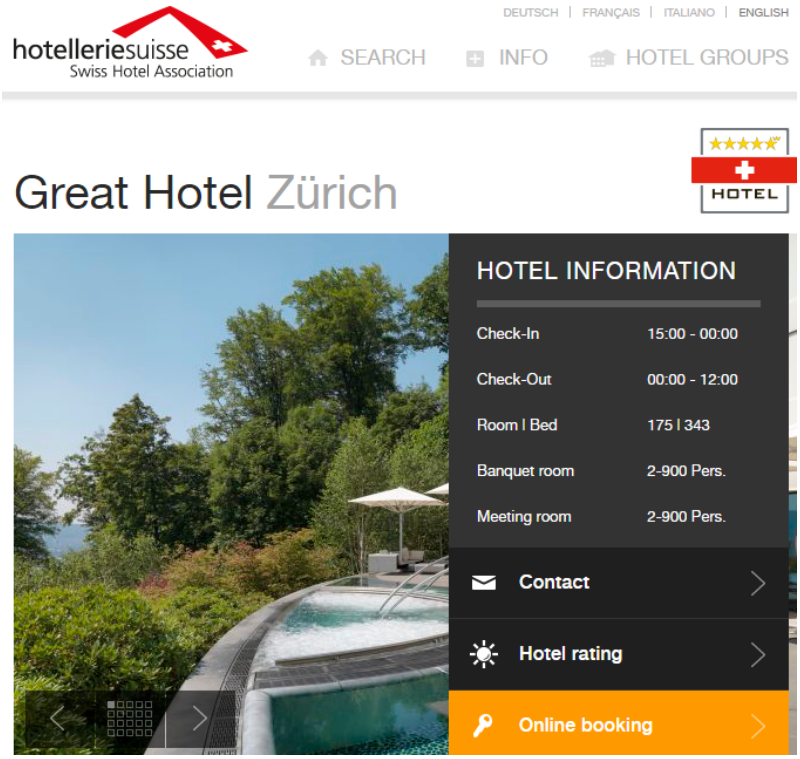

Figure 3.8: Overview of Swisshotel site for fictional hotel on 27.02.2018

A basic overview is followed by information about the hotel infrastructure, management, local infrastructure in a 10-kilometer radius, hotel classification, group/chain labels, and accepted payment methods.

We first crawled the overview of all hotels to collect links for each hotel. On a second crawl, we accessed the details of every hotel collected in the previous crawl, which resulted in data for 3,976 hotels.

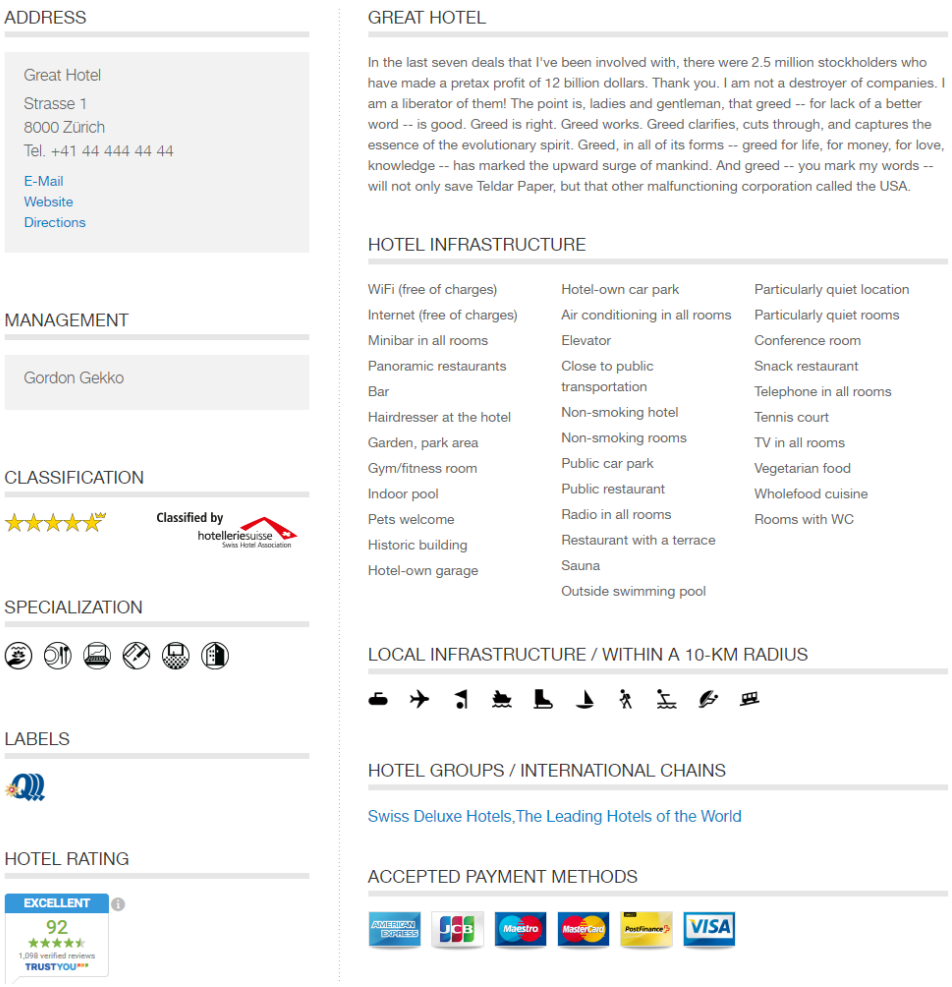

Figure 3.9: Overview of Swisshotel details for the fictional hotel on 27.02.2018

3.1.6 Relation of Data

Our main goal was to collect as much data as possible for the initial 667 hotels found in the customer database, as we have only revenue information for this dataset. The links we manually collected for TripAdvisor and Booking.com did not entirely overlap. However, we were able to find almost 400 of the hotels that were available on one of the two platforms on Google as well, as Figure 3.10 indicates.

The collection of data from all hotels in Switzerland from the TripAdvisor and Swisshotel pages resulted in much more data then we could feasibly connect to our insurance database. Since we already had the URLs for TripAdvisor entries, connecting them with the full TripAdvisor database was trivial, and we were able to connect all of the 374 entries. Matching the hotels from the insurance database with the Swisshotel dataset was less simple, however. The procedure is described later in the chapter on matching. We were able to connect 331 entries from the insurance database with Swisshotel. Also, the Swisshotel and full TripAdvisor databases overlap in 3,174 cases (according to the same matching procedure used for insurance and Swisshotel).

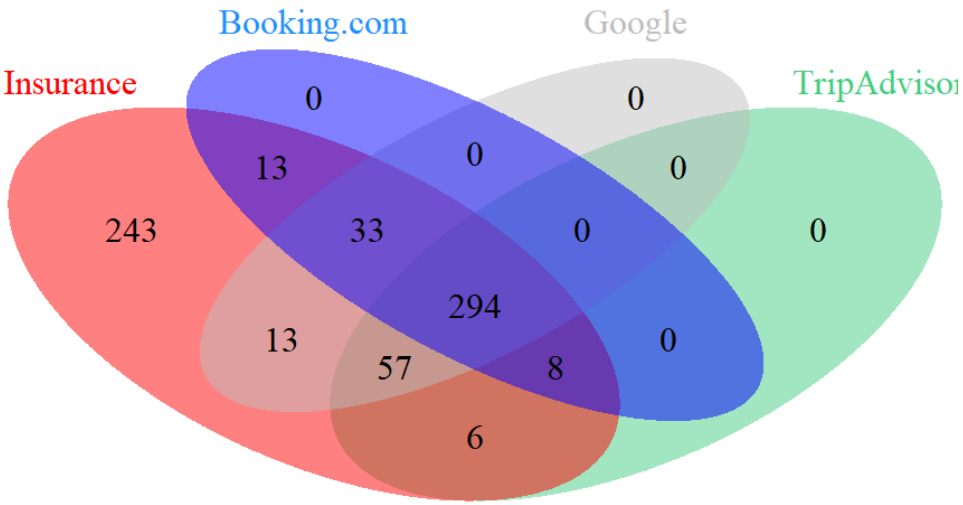

Figure 3.10: Overlap with the insurance database and the data we were able to collect online from TripAdvisor, Booking.com, and Google

3.2 Data Collection

For each dataset, we provided an overview with the name, a short explanation, and a count of how often the attribute was present. The first attribute was always an identifier (ID) that was either unique in the table or, in the case of TripAdvisor reviews, associated with a hotel in the insurance database. The IDs were used to connect data from multiple tables. In the case of TripAdvisor or Booking.com, we could also use the URL to connect them to the insurance dataset. This was only true for connecting customers with the full TripAdvisor dataset.

3.2.1 Customer Dataset

The attributes of street, nb (street number), plz (zip code), and city are copies from the database of the insurance company for which we received revenue data. In order to protect the internal ID, we replaced it with a temporary one called the tempid. The three remaining attributes were collected manually with Google to determine which hotel was associated with the customer. For this process, we had access to additional information, such as the name of the customer. Since not all entries could be connected to an online entry, the number of collected URLs is significantly lower than the size of the full dataset.

Attribute	Description	Count
tempid	Unique id for the insurance dataset	667
street	Street name	667
nb	Street number	535
plz	Zip code	667
city	City name	667
website	URL to the hotel if available, or comment	529
tripadvisor	URL to the website on TripAdvisor if available	374
booking	URL to the website on Booking.com if available	356

Table 3.1: Customer dataset

3.2.2 TripAdvisor Dataset

Two datasets were obtained from TripAdvisor. The first was a list of all accommodations in Switzerland, including hotels as well as B&Bs, pensions, and other sites. TripAdvisor specifies a type, but this is not always accurate or complete. While users can rate an accommodation with an integer from 1 to 5, the accumulated rating is rounded to the nearest multiple of 0.5. We retrieved the number of ratings for each integer and were able to calculate an exact rating for each entry. Global positioning system (GPS) data were gathered by another service following dataset collection.

Reviews were collected only for customers in view of the large number of reviews in general. Some larger hotels had accumulated over 1,000 reviews. When collecting reviews, one must be careful to select the option to display all languages, as the default option presents only reviews that are in the same language as the browser settings. We did not assign a unique ID to each review since they were used only in an accumulated way. When treating the review date field, it was necessary to translate special cases, such as ”three weeks ago,” into a proper date format.

We processed the raw reviews to extract historical ratings. There was not sufficient data for all hotels to produce a meaningful rating over the whole time period.

Attribute	Description	Count
taid	Unique TripAdvisor id for each entry	6,855
link-href	Unique URL to the website of the entry	6,855
ta_streetaddress	Street name and number	6,442
ta_rating	Accumulated and rounded rating	6,021
ta_postalcode	Zip code	6,711
ta_name	Full name of the property	6,855
ta_city	City name	5,763
ta_reviewcount	Accumulated number of reviews	6,021
ta_stars	Number of stars if available	3,335
ta_local_ranking	Nr. x out of y properties in city	6,019
x	GPS x coordinate	6,855
y	GPS y coordinate	6,855
ta_lower_price	Lower price of the average rate	4,805
ta_higher_price	Higher price of the average rate	4,805
ta_price	(ta_lower_price + ta_higher_price) divided by 2	4,902
ta_rooms	Number of rooms available at property	5,659
ta_fives	Number of five ratings the property has received	6,023
ta_fours	Number of four ratings the property has received	6,023
ta_threes	Number of three ratings the property has received	6,023
ta_twos	Number of two ratings the property has received	6,023
ta_ones	Number of one ratings the property has received	6,023
ta_rating_value_exact	The rating value calculated from previous five entries	6,023
ta_reviewcount_exact	The number of ratings calculated by category	6,023
ta_local_ranking_percentile	Calculate $\frac{x}{y}$ from ta_local_ranking	6,019
ta_type	Either {hotel/pension/other} or empty	6,093

Table 3.2: TripAdvisor dataset

Attribute	Description	Count
tempid	ID according to the insurance dataset	51,513
ta_reviews_reviewcount	Accumulated number of reviews for the hotel	51,491
ta_reviews_ratingvalue	Accumulated rating value for the hotel	51,367
ta_local_ranking	"Nr. x out of y hotels in this city"	51,491
ta_review_date	Date of the review	51,510
ta_review_score	Score of the review	51,510
ta_review_title	Title of the review	51,510
ta_review_text	Text of the review	51,491
ta_rooms	Number of rooms at the hotel	46,231
ta_local_ranking_value	x extracted from ta_local_ranking	51,491
ta_local_ranking_max	y extracted from ta_local_ranking	51,491

Table 3.3: Raw reviews from hotels on TripAdvisor

Attribute	Description	Count
tempid	ID according to the customer dataset	354
ta_local_ranking_max	y extracted from ta_local_ranking	349
ta_local_ranking_percentile	Calculate $\frac{x}{y}$ from ta_local_ranking	349
ta_local_ranking_value	x extracted from ta_local_ranking	349
ta_ratingvalue_on_2011_01_01	Rating value on 01.01.2011	239
ta_ratingvalue_on_2012_01_01	Rating value on 01.01.2012	264
ta_ratingvalue_on_2013_01_01	Rating value on 01.01.2013	285
ta_ratingvalue_on_2014_01_01	Rating value on 01.01.2014	306
ta_ratingvalue_on_2015_01_01	Rating value on 01.01.2015	330
ta_ratingvalue_on_2016_01_01	Rating value on 01.01.2016	338
ta_ratingvalue_on_2017_01_01	Rating value on 01.01.2017	345
ta_ratingvalue_on_2018_01_01	Rating value on 01.01.2018	351
ta_reviewcount_on_2011_01_01	Number of reviews on 01.01.2011	239
ta_reviewcount_on_2012_01_01	Number of reviews on 01.01.2012	264
ta_reviewcount_on_2013_01_01	Number of reviews on 01.01.2013	285
ta_reviewcount_on_2014_01_01	Number of reviews on 01.01.2014	306
ta_reviewcount_on_2015_01_01	Number of reviews on 01.01.2015	330
ta_reviewcount_on_2016_01_01	Number of reviews on 01.01.2016	338
ta_reviewcount_on_2017_01_01	Number of reviews on 01.01.2017	345
ta_reviewcount_on_2018_01_01	Number of reviews on 01.01.2018	351
ta_reviews_ratingvalue	Accumulated rating value at time of collection	336
ta_reviews_reviewcount	Accumulated number or reviews at time of collection	349
ta_rooms	Number of rooms at the hotel	275

Table 3.4: Reviews accumulated by hotel on TripAdvisor

3.2.3 Booking Dataset

The booking dataset was created by crawling the manually collected links with the Scrapy framework. There was not always a rating available; however, if the general rating was present, so were the specialized ratings.

Attribute	Description	Count
tempid	ID according to the customer dataset	357
booking	URL to the booking website	357
bk_hotel_wifi	Rating value for the WiFi	327
bk_hotel_comfort	Rating value for comfort	333
bk_hotel_services	Rating value for the facilities	333
bk_hotel_clean	Rating value for cleanliness	333
bk_reviewcount	Number of reviews received	338
bk_hotel_location	Rating value for the location	333
bk_ratingvalue	Overall rating value	338
bk_hotel_value	Rating value for value for money	333
bk_hotel_staff	Rating value for staff	333
bk_name	Name of the hotel	339

Table 3.5: Data collected from Booking.com

3.2.4 Google Dataset

When we collected links for the customer hotels for TripAdvisor and Booking.com, we did not store the link to the Google search. Therefore, we had to use data derived from TripAdvisor and Booking.com to recreate a query for Google. Trials indicated that the best results were obtained by creating a query of the form:

https://www.google.ch/search?q=”hotel name”+”city”

When the data from TripAdvisor and Booking.com were overlapping, we prioritized the data from TripAdvisor.

Attribute	Description	Count
tempid	ID according to the customer dataset	357
go_name	Name of the hotel	386
go_ratingvalue	Aggregated rating value	380
go_reviewcount	Number of reviews	380
go_street	Street Name and number	382
go_postalcode	Zip code	381
go_city	Name of the city	382

Table 3.6: Data collected from Google on hotels

3.2.5 Swisshotel Dataset

The Swisshotel dataset has 364 attributes in its raw form. Of these, 155 are duplicates, which was predominantly due to attributes appearing in different languages for different

hotels. In cases where an attribute appeared in a language other than English, it was mapped to its English equivalent. The following is one example:

$$\text{sh\_infrastructure\_badewanne} \Rightarrow \text{sh\_infrastructure\_bath\_tub}$$

Appendix A.1 contains a complete list of the mapping. This mapping reduced the number of attributes to 254. Another problem was the length of the attribute names. For further storage in an SQL database, names had to be no longer than 30 characters but still unique. We fulfilled this criterion in two steps. First, we shortened reoccurring words, such as infrastructure and classification:

$$\text{sh\_infrastructure\_grill} \Rightarrow \text{sh\_in\_grill}$$
$$\text{sh\_classification\_5\_sterne\_superior} \Rightarrow \text{sh\_cl\_5\_sterne\_superior}$$

Second, we cut off any characters over the limit. Appendix A.2 presents the mapping from long to short names. Most of the attributes were binary; Table 3.7 describes only non-binary attributes. The binary attributes are self-explanatory.

Attribute	Description	Count
swissid	Unique ID of the hotel	3,976
swisshotel	URL to the website of the hotel	3,976
sh_rooms	Number of rooms	3,262
sh_name	Name of the hotel	3,976
sh_stars	Number of stars and category (Superior/Garni)	1,593
sh_check_out	Check-out time, format from - to or 24h	1,913
sh_meeting_room	Minimum and maximum size of meeting room	1,289
sh_code	Zip code of the city	3,976
sh_city	Name of the city	3,976
sh_street	Street name and number	3,747
trust_you	URL of the associated trust me website	3,949
sh_managers	Name(s) of the manager(s), separated by semicolon	2,946
sh_banquet_room	Minimum and maximum size of the banquet room	1,286
sh_check_in	Check-in time, format: from - to or 24h	1,977
sh_beds	Number of beds in hotel	2,910
sh_telephone	Telephone number of hotel	3,976
sh_google_name	Name of the hotel on Google	3,742
sh_google_ratingvalue	Rating value on Google	3,671
sh_google_reviewcount	Number of reviews on Google	3,648
sh_x	GPS x coordinate	3,975
sh_y	GPS y coordinate	3,975
sh_max_meeting_room_size	Max. size of the meeting room, in sh_meeting_room	1,289
sh_max_banquet_room_size	Max. size of the banquet room, in sh_banquet_room	1,286
sh_nb_stars	Numbers of stars as integer, extracted from sh_stars	1,593
sh_nb_managers	Numbers of managers as integer, from sh_managers	2,946

Table 3.7: Data collected from Swisshotel, non-binary attributes

3.2.6 Economic Dataset

Economic data for tourism regions was also collected. This information is freely available online on the website of the Federal Statistical Office of Switzerland. It offers data on 100 different regions.

Attribute	Description	Count
edid	Unique ID of the economic region	100
ed_city	Name of the region	100
ed_city_codes	List of zip codes in the region	100
ed_hotels_2013	Number of hotels in the region in 2013	100
ed_rooms_2013	Number of rooms in the region in 2013	100
ed_beds_2013	Number of beds in the region in 2013	100
ed_arrivals_2013	Number of people who stayed in the region in 2013	100
ed_stays_2013	Number of room nights in the region in 2013	100
ed_room_stays_2013	Number of booked rooms in region in 2013	100
ed_room_occupancy_2013	Average room occupancy rate in 2013	100
ed_bed_occupancy_2013	Average bed occupancy rate in 2013	100
ed_hotels_2014	Number of hotels in the region in 2014	100
ed_rooms_2014	Number of rooms in the region in 2014	100
ed_beds_2014	Number of beds in the region in 2014	100
ed_arrivals_2014	Number of people who stayed in the region in 2014	100
ed_stays_2014	Number of room nights in the region in 2014	100
ed_room_stays_2014	Number of booked rooms in region in 2014	100
ed_room_occupancy_2014	Average room occupancy rate in 2014	100
ed_bed_occupancy_2014	Average bed occupancy rate in 2014	100
2014	Change in bed occupancy rate, year over year	100
ed_hotels_2015	Number of hotels in the region in 2015	100
ed_rooms_2015	Number of rooms in the region in 2015	100
ed_beds_2015	Number of beds in the region in 2015	100
ed_arrivals_2015	Number of people who stayed in the region in 2015	100
ed_stays_2015	Number of room nights in the region in 2015	100
ed_room_stays_2015	Number of booked rooms in region in 2015	100
ed_room_occupancy_2015	Average room occupancy rate in 2015	100
ed_bed_occupancy_2015	Average bed occupancy rate in 2015	100
2015	Change in bed occupancy rate, year over year	100
ed_hotels_2016	Number of hotels in the region in 2016	100
ed_rooms_2016	Number of rooms in the region in 2016	100
ed_beds_2016	Number of beds in the region in 2016	100
ed_arrivals_2016	Number of people who stayed in the region in 2016	100
ed_stays_2016	Number of room nights in the region in 2016	100
ed_room_stays_2016	Number of booked rooms in region in 2016	100
ed_room_occupancy_2016	Average room occupancy rate in 2016	100
ed_bed_occupancy_2016	Average bed occupancy rate in 2016	100
2016	Change in bed occupancy rate, year over year	100

Table 3.8: Data collected from the Federal Statistics Office

3.2.7 Revenue

We received revenue data from the insurance company that were connected with 667 clients over multiple years (see Table 3.9). However, the data required treatment in order to have unique revenue entries for each year and for each company. The earliest revenues we received were from 2008, and the latest were from 2016. Not every company has an entry for every year. Some companies had multiple, contradictory entries for one year, which could be due to a change in insurance policy. Whenever there were multiple entries, we averaged the available values for that year. This procedure increased cases in which we noted a change in revenue from one year to the next (compared to a simple majority vote). The vast majority of companies had no reported change in revenue, which might be due to the self-reporting bias discussed in Chapter 1.

Attribute	Description	Count
tempid	Unique ID of the customer	15,716
year	Ranges from 2008 to 2016	15,716
month	Month of the insurance policy	15,716
revenue	Ranges from 10,000 to 19,500,000	15,716

Table 3.9: Revenue data received from the insurance company

Since data were sparser from TripAdvisor and the insurance dataset for earlier years, we decided to take into account only revenue from 2012 onward. From the incomplete and contradictory data in 3.9, we created a complete table with revenues from every company for the years 2012 to 2016. This was performed according to the following algorithm:

Name	2012	2013	2014	2015	2016
Company A		15'000	15'000	15'000	
Company B		18'000		18'000	18'000
Company C	15'000		20'000		30'000

Name	2012	2013	2014	2015	2016
Company A	15'000	15'000	15'000	15'000	
Company B	18'000	18'000		18'000	18'000
Company C	15'000		20'000		30'000

Name	2012	2013	2014	2015	2016
Company A	15'000	15'000	15'000	15'000	
Company B	18'000	18'000		18'000	18'000
Company C	15'000	17'500	20'000		30'000

... ...

Name	2012	2013	2014	2015	2016
Company A	15'000	15'000	15'000	15'000	15'000
Company B	18'000	18'000	18'000	18'000	18'000
Company C	15'000	17'500	20'000	25'000	30'000

Figure 3.11: Overview of the revenue completion process

3.2.8 Feature Generation

In order to improve the results of our algorithms, we experimented with additional features that we created from the gathered data. While less important in the case of revenue estimation, it is critical for the success of the growth prediction. For example, if we want to determine whether a hotel experienced growth between 2012 and 2013, it might be of interest to identify the change in rating value in that year as an attribute.

Attribute	Description
rooms	Number of rooms either from Swisshotel or TripAdvisor
stars	Number of stars either from Swisshotel or TripAdvisor
reviewcount	Sum of all review counts from all three sources
ratingvalue	Average of all rating values weighted and scaled
price	(upper_price + lower_price)/2 from TripAdvisor
sh_nb_managers	Counting the number of managers according to sh_managers
sh_manager_couple	Binary attribute if the manager is a couple (matching last names)
occupied_rooms	Rooms * ed_room_occupancy
stays_per_room	ed_room_stays divided by ed_rooms
reviews_per_room	go_reviewcount divided by rooms

Table 3.10: Features generated for revenue prediction

The following features only appear as predictors in the growth case (Table 3.11).

Attribute	Description
change_ratingvalue_yoy	TripAdvisor rating value - rating value one year ago
change_ratingvalue_2yoy	TripAdvisor rating value - rating value two years ago
change_reviewcount_yoy	TripAdvisor review count - review count one year ago
change_reviewcount_2yoy	TripAdvisor review count - review count two years ago
change_rating_variance_yoy	TripAdvisor variance - variance one year ago
change_rating_variance_2yoy	TripAdvisor variance - variance two years ago
change_rooms	Year over year change in number of available rooms
change_hotels	Year over year change in the number of available hotels
change_arrivals	Year over year change in arrivals
change_stays	Year over year change in room stays
change_room_stays	Year over year change in number of booked rooms
change_room_occupancy	Year over year change in room occupancy rate
change_bed_occupandcy	Year over year change in bed occupancy rate

Table 3.11: Attributes created for fuzzy matching

A full summary of the data which was used for revenue prediction can be found in the appendix (see Table A.3). The same applies to growth prediction data (see Table A.4).

3.3 Framework

This section briefly describes the frameworks for crawling the websites. It also provides a reference and a short explanation of the use case for both of them..

3.3.1 Scrapy

In order to collect data from the TripAdvisor and Booking.com websites, this study utilized a Python-based framework called Scrapy. This framework is open source and available online at https://scrapy.org/. It was integrated into the Python program, which processed all the data for the project. Once a website was retrieved, the content could be accessed with CSS or XPATH queries. However, since many websites change quickly, code is outdated in a matter of weeks. Scrapy functioned well with the above-mentioned websites because they are mainly static (low use of JavaScript) and do not obfuscate their code. Some websites, such as Google, deny access to web-crawlers such as Scrapy.

3.3.2 Webscraper.io

Webscraper.io is a closed source web scraper that works as an extension of the Chrome web browser. Therefore, it can load websites through Chrome and is hence not affected by the same restrictions as Scrapy and similar frameworks. Still, it is quite slow, as the minimal waiting time between requests is two seconds, which makes crawls of thousands of websites small. The items that should be collected can be selected in a graphical interface. It is also possible to click on elements, thereby enabling navigation through menus. This option was used to collect all reviews of selected TripAdvisor hotels.

Once the data collection was finished, the data was stored in a comma separated values (CSV) file for further use. The web-scraper plugin is free to use and can be downloaded at the official website, http://webscraper.io/.

3.3.3 GPS Data

The GPS data were collected by querying the Bing map service with the address of a hotel. We used a search string that was composed of the name, address, city, and the country. The collected data was useful for visualizing the location in Switzerland, as in Figure 1.3. The answer to a GPS query are always two numbers, x and y coordinates. If the exact address is not available, the service returns an approximate location. In most cases the city/town of the hotel.

3.4 Matching

The Swisshotel dataset had no identifier to link it to the other datasets. Since we wanted to use attributes from the Swisshotel dataset in our prediction file, we needed a way to relate it with the insurance dataset. The process of relating the entries from those two sources is called "matching." In theory, we could simply match the name and address of the hotel and obtain a unique identifier. However, due to the diverse sources of our data, those attributes often do not match completely, so we needed a way to decide when the attributes were "close enough" to be a match. Further complicating the problem was the fact that not every entry in the customer dataset has a corresponding entry in the Swisshotel dataset.

3.4.1 Approximate String Matching

Navarro (2001) has provided an overview of various approximate string matching techniques, which are sometimes also called "fuzzy string searching."

Ratcliff/Obershelp

Ratcliff and Metzener (1988) first proposed this algorithm. It computes the similarity of two strings as the number of matching characters divided by the total number of characters in the two strings. In this context, matching characters are those in the longest common sub-sequence in those two strings. Plus, it recursively matches the unmatched region on either side of the longest common sub-sequence Black (2004). The algorithm is implemented in the difflib module in Python under the name "SequenceMatcher." It returns a score between 0 and 1 (including boundaries).

Levenshtein Distance

Levenshtein (1965) originated this algorithm. The smallest number of insertions, deletions, and substitutions are required to change one string into another Pieterse and Black (2015). The algorithm is implemented in the Levenshtein module in Python and also returns a score between 0 and 1 (including boundaries).

Pre-Processing

The attributes we used were the name, address, city, and zip code. All letters were converted to lowercase, and special characters were removed. Furthermore, we removed all accents on the letters o, u, a, e, and o, such as é, è, ô, ö, ü and ä. The pre-processed attributes where then stored in new attributes (see Table 3.12).

Additionally all duplicate words were removed from each of the newly created attributes.

Attribute	Description
sh_fuzzy	name + city + street + zip code
sh_fuzzy_name	name of the hotel but removing ”hotel” and ”restaurant”
sh_fuzzy_street	name of the city

Table 3.12: Attributes created for fuzzy matching

Proposed Algorithms

Preliminary tests revealed that not all data matched optimally when using all attributes. For example, if the address is missing in one dataset but not the other, the score would be low even though the names might be a perfect match. However, if we would only match names, then there would be matches with hotels that have the same name but are in different locations. To address this problem, we tested different combinations of scores. In these combinations, we either added or multiplied scores obtained by comparing two strings. Table 3.13 contains all the proposed and tested algorithms based on a score function, which is either Levenshtein or Ratcliff/Obershelp.

Algorithm	Description
all	score(sh_fuzzy)
all_t_name	score(sh_fuzzy) * score(sh_fuzzy_name)
all_t_name_t_address	(score(sh_fuzzy) * score(sh_fuzzy_name) * score(sh_fuzzy_street))$^{\frac{1}{3}}$
all_p_name	(score(sh_fuzzy) + score(sh_fuzzy_name))/2
all_p_name_p_address	(score(sh_fuzzy) + score(sh_fuzzy_name) + score(sh_fuzzy_street))/3

Table 3.13: Proposed algorithms for fuzzy matching

To make the algorithms more comparable, we normalized the score by either taking a square/cubic root or dividing the result by 2 or 3.

3.4.2 Evaluation of Matching Algorithms

The algorithms described in Section 3.4.1 were evaluated on a sub-sample of the insurance dataset. The sub-sample contained 80 entries, or approximately 20% of the total available customer data. For each entry, we manually checked if there was a match in the Swisshotel dataset. Then, the algorithms were run with different settings to see which performed best on the test set. For each entry in the insurance dataset, we selected the entry in the Swisshotel data with the highest similarity score as a match. Furthermore, we defined a cutoff value, and all entries with a similarity score below the cutoff value were said to have no matching entry. By varying the cutoff score, we could then decide how closely we wanted the entries to be related in order to be considered a match. If we chose a cutoff value that was too high, then we might lose matches. If we selected too low of a cutoff score, we would have too many incorrect matches. The following graphs illustrate which percentage of the data would be correctly matched for a specific cutoff value. Figure 3.12 depicts how the algorithms compare to each other using the Ratcliff/Obershelp algorithm as the score function.

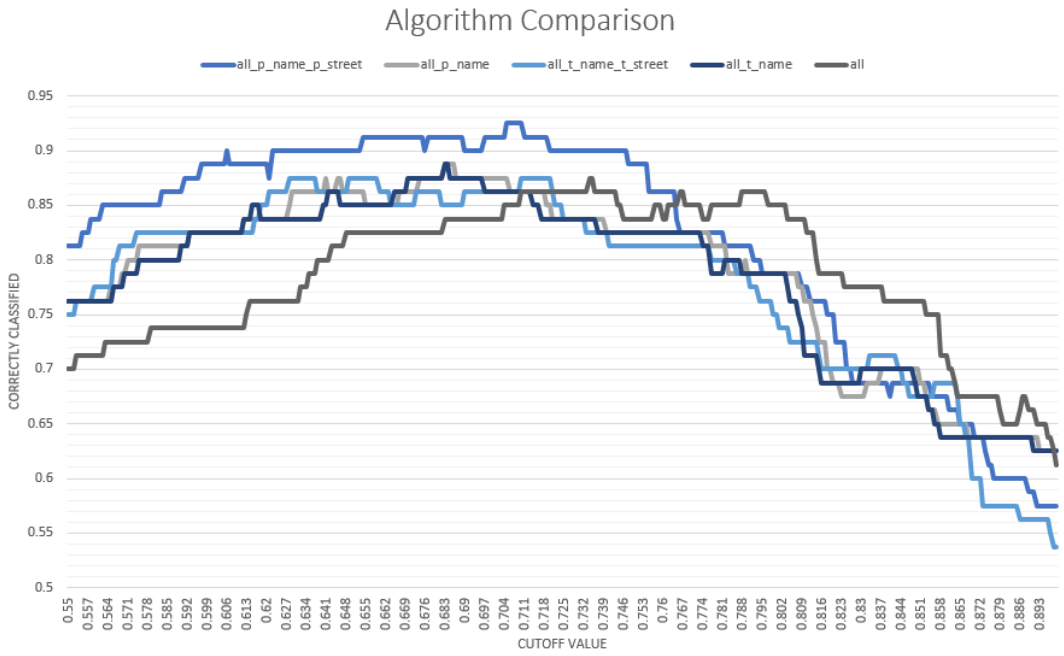

Figure 3.12: Proposed fuzzy algorithm performance

Evidently, the algorithm all_p_name_p_address outperformed the others throughout almost the whole range. More importantly, we achieved the highest overall precision as well. Next, we compare the performance of Levenshtein against Ratcliff/Obershelp (see Figure 3.13).

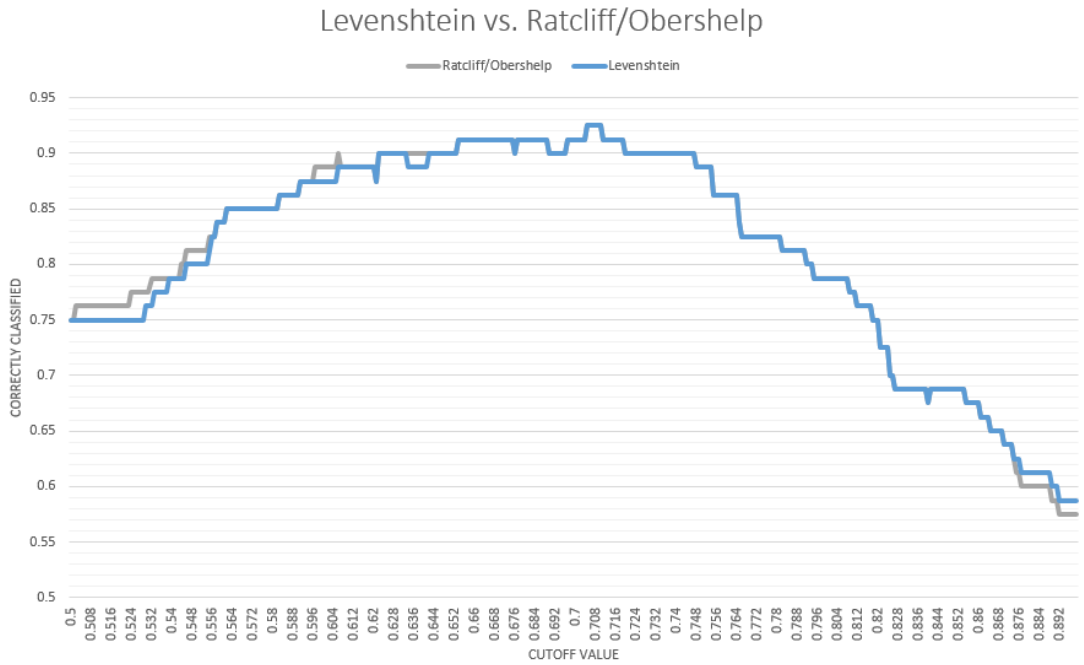

Figure 3.13: Comparison between Levenshtein and Ratcliff/Obershelp

The curves overlap over large parts of the graph. Especially in the crucial part where the

highest precision is achieved, there is no difference at all. At the fringes, Ratcliff/Obershelp marginally outperforms Levenshtein.
In view of these results, we chose to match the insurance dataset with Swisshotel using the _p_name_p_address algorithm with a Ratcliff/Obershelp score function and a cutoff value of 0.71. Resulting in 331 matches for those two datasets. Using the same approach, the full TripAdvisor data were also matched with Swisshotel, resulting in 3174 matches for those datasets.

3.5 Missing Data

As Figure 3.10 reveals, the datasets do not completely overlap. Therefore, not all entries are complete. The data analysis tool R excludes all entries with missing values by default. By excluding entries with missing data, we would reduce our dataset from 397 to 158 entries. One solution to this problem is multiple imputation, which is widely adopted in practice Mackinnon (2010).

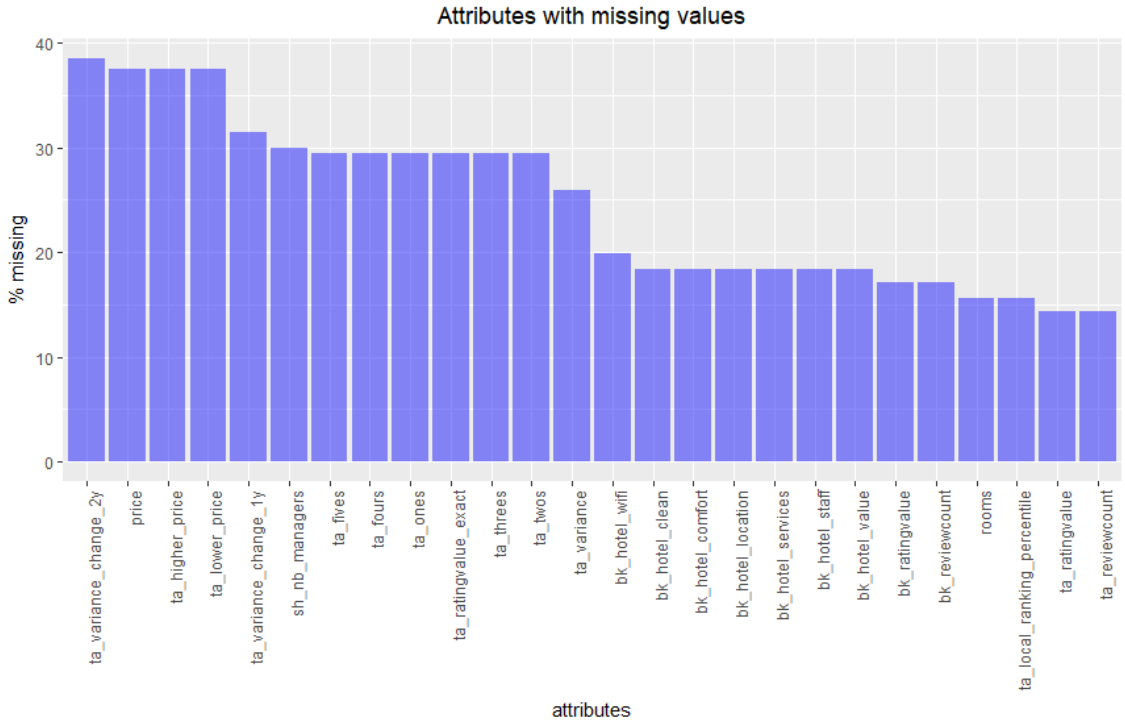

Figure 3.14: Percentage of missing data of attributes that are not complete

To impute missing values in the rooms and price attributes, we used custom-built models that were trained on the TripAdvisor data. For all other attributes, the package mice Buuren and Groothuis-Oudshoorn (2011) was used. Since the model of room and price data was trained on complete entries, we imputed the other attributes first and then applied those models to the imputed data to predict rooms and price. To avoid information leakage from the training to the test set, the imputation was executed separately for each set. An important question was how many imputations to perform. As a standard rule Bodner (2008) White, Royston, and Wood (2011), if an average of 20% of the data is missing, we should impute at least 20 times. As observable in Figure 3.14, 20% of missing

data seems to be a reasonable estimation in this case.

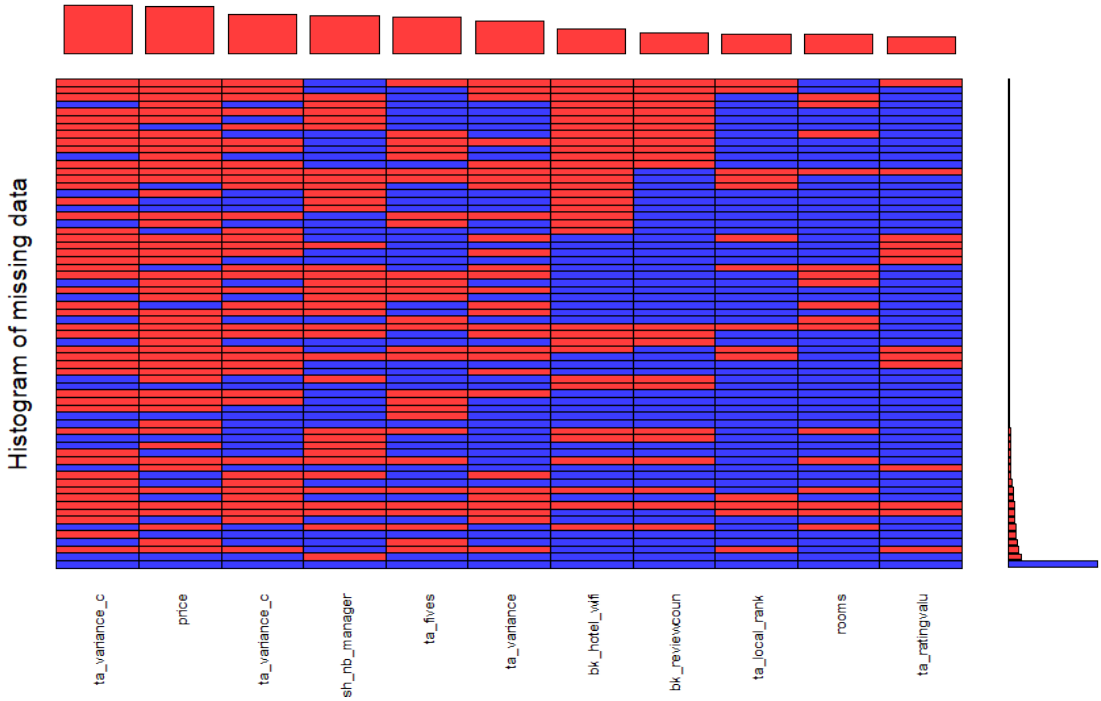

Figure 3.15: Histogram of missing values, red means missing, blue means available. Only selected attributes are shown

Figure 3.15 only shows attributes which differ in the percentage of missing data. For example, Figure 3.14 shows that the percentage of missing rating values from TripAdvisor, is the same as that of the TripAdvisor review count. They are both missing if a hotel is not in the database, or has not yet received any rating. Hence only uncorrelated attributes appear in the histogram, to avoid a misleading pattern.

When data is missing, one should ask why it is absent. The cause is called the "missing data mechanism," and it can have a serious effect on results. The simplest case is called missing completely at random (MCAR). Meaning there is no relationship between missing values and any other variables in the dataset. However, this is rarely the case in reality, and the histogram in Figure 3.15 reflects that the data are not missing in a random pattern. Another option is called missing at random ("mar), which means data is missing conditionally on other data, which may or may not be observed. It is not statistically verifiable, but it is a reasonable assumption with which we can work Gelman and Hill (2016).

One check for the imputation is plotting the imputed values in a density plot in Figure 3.16.

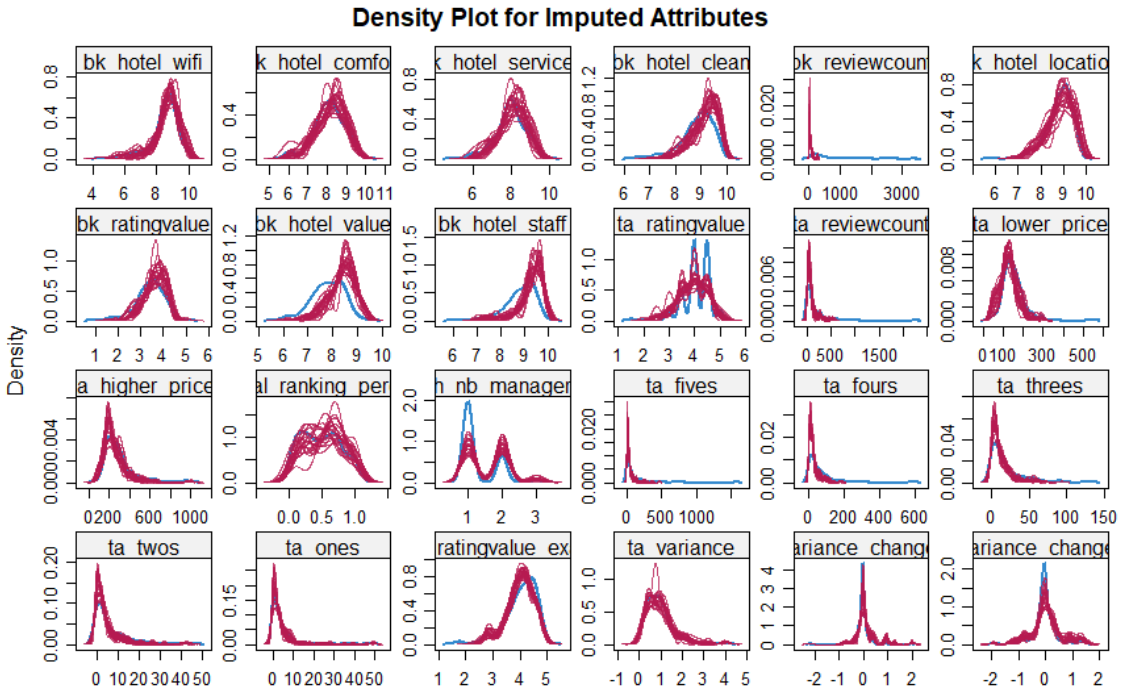

Figure 3.16: Density plot of imputed (red) and available (blue) attributes

The densities of the imputed and available attributes match in almost all cases with some notable exceptions. The number of reviews for Booking.com and TripAdvisor is skewed towards the lower end of the spectrum (see Figure 3.16). This makes sense in case where the missing hotels are smaller on average than the rest, which is a reasonable assumption. The imputation also misses the discrete steps of 0.5 for the rating value of TripAdvisor hotels.

3.6 Prediction Methods and Error Functions

In this section, we concisely explain the theoretical background of the methods that were used to predict hotel revenue and growth. As a baseline and simple model, we employed a linear regression model. For greater accuracy, we used support vector machines and boosted trees.

3.6.1 Linear Regression

Suppose we have a dataset with n entries and m attributes each. We denote these by (x_{ij}) for $i \in \{1, \dots, n\}, j \in \{1, \dots, m\}$, or more compactly by $X \in \mathbb{R}^{n \times m}$. Furthermore, for each entry, we have a response y_i, (y_i) with $i \in \{1, \dots, n\}$. We write y_i more compactly as $y \in \mathbb{R}^n$. Suppose as well that we have a noise vector $\epsilon \in \mathbb{R}^n$. Our goal is now to search for a vector $\beta \in \mathbb{R}^n$ such that

$$y = X\beta + \epsilon$$

The most commonly used method to estimate the vector is ordinary least squares (OLS), with which we try to solve the following problem

$$\hat{\beta} = \arg\min_{\beta \in \mathbb{R}^n} ||X\beta - y||_2^2.$$

After some algebra, this reduces to

$$\hat{\beta} = (X^T X)^{-1} X^T y.$$

However, linear regression is a relatively simple method and should not be expected to perform that well out of the box. Moreover, it entails a couple of assumptions, such as homoscedasticity and joint normality for errors. These assumptions are important when performing inference, but since we engaged in prediction, we did not check them.

3.6.2 Support Vector Machines

Support Vector Machines are supervised learning models, which construct a set of hyperplanes in a high dimensional space. Trying to separate inputs into classes in the case of classification. In this case we are interested in regression. Our data is in the form of N labeled data points $\{(x_1, y_1), \ldots, (x_N, y_N)\}$ where $x_i \in \mathbb{R}^p$ and $y_i \in \mathbb{R}$. Consider the case where the regression function f is approximated by a set of basis functions $\{h_1, \ldots, h_m\}$ with

$$f(x) = \sum_{m=1}^{M} \beta_m h_m(x) + \beta_0.$$

Estimating β and β_0 by

$$H(\beta, \beta_0) = \sum_{i=1}^{N} V(y_i - f(x_i)) + \frac{\lambda}{2} \sum \beta_m^2$$

for some general error measure $V(r)$, for example $V(r) = r^2$. For any choice of $V(r)$ the solution $\hat{f}(x) = \sum \hat{\beta} h_m(x) + \hat{\beta}_0$ has the form

$$\hat{f}(x) = \sum_{i=1}^{N} \hat{\alpha}_i K(x, x_i),$$

where $K(x, y) = \sum_{m=1}^{M} h_m(x) h_m(y)$ Hastie, Tibshirani, and Friedman (2009). In the case of this thesis a support vector machine with radial basis functions as kernels is used, these kernels are of the form

$$K(x, y) = \exp(-\gamma ||x - y||^2) \quad \gamma > 0.$$

3.6.3 Boosted Trees

Suppose we have N labeled data points $\{(x_1, y_1), \ldots, (x_N, y_N)\}$ where $x_i \in \mathbb{R}^p$ and $y_i \in \mathbb{R}$. A tree ensemble using K additive functions to predict the output is

$$\hat{y}_i = \phi(x_i) = \sum_{k=1}^{K} f_k(x_i) \quad f_k \in \mathbb{F}.$$

Where $\mathbb{F} = \{f(x) = w_{q(x)}\}$ with $q : \mathbb{R}^{\text{l}} \longmapsto T, w \in \mathbb{R}^T$ is the space of regression trees and q represents the structure of each tree that maps an example to the corresponding leaf. To learn the set of functions used in the above model, a regularized objective is minimized

$$L(\phi) = \sum_i l(\hat{y}_i, y_i) + \sum_k \Omega(f_k)$$
$$\text{where } \Omega(f) = \tau T + \frac{1}{2}\lambda ||w||^2$$

l is a differentiable convex loss and Ω penalizes the complexity of the model T. Chen and Guestrin (2016). As that equation contains functions as parameters it can not be optimized using traditional minimization methods, instead it is trained in an additive manner. Let $\hat{y}_i^{(t)}$ be the prediction of the i-th instance at the t-th iteration. To minimize, one adds greedily f_t to the following objective function

$$L^{(t)} = \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)) + \Omega(f_t).$$

Boosted trees are extensively used in practice due to their high performance in many real world tasks DMLC (2018).

3.6.4 Error Functions

When making predictions, we have to somehow measure the quality of our retrieved values. We denote the true value as y_i for $i \in \{1, \ldots, n\}$, and the estimated value as $\hat{y}_i$ for $i \in \{1, \ldots, n\}$.The root mean squared error (RMSE) is defined as

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2},$$

whereas the mean average error (MAE) is defined as

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|.$$

The difference between these two error functions is the sensitivity of the RMSE to outliers. Large errors have a disproportionately heavy effect on the RMSE compared to the MAE.

3.7 Data Treatment for Prediction

We have previously described how we processed and imputed the data. Now, we elaborate on the full cycle in Figure 3.17. The split in the beginning is important as otherwise we would have information flowing from the test to the training dataset, leading to a bias in results. Also, all binary Swisshotel attributes with less than where less than 5% was present, were excluded. The train-test split in Figure 3.17 was repeated five times, in a process called cross validation.

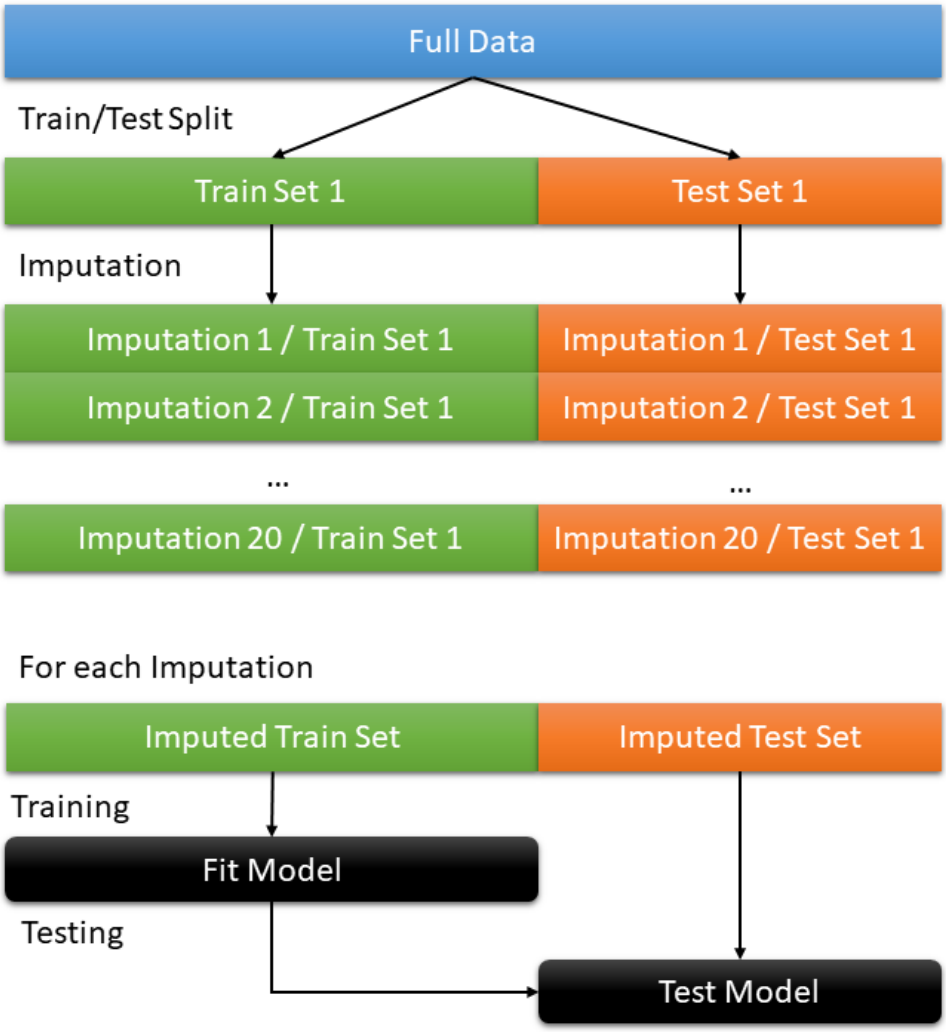

Figure 3.17: How data were processed for prediction

Chapter 4

Results

We now evaluate our approach by running multiple models on the data we collected. We explore two settings. First, we try to predict the revenue of hotels and then the RevPAR, and lastly we try to separate growing from shrinking hotels. For the case of revenue and RevPAR prediction, we use an estimation by mean (MP) as a baseline model. The models we evaluate are based on linear regression, SVMs, and random forests. We specify different error measurements for all three models. For the case of growth prediction, we evaluate only a random forest model. All the data were prepared according to the process described in the previous chapter.

4.1 Revenue

In this sub-section, we predict revenue for hotels based on the data we collected online. All the cleaning and imputation procedures were performed as described in the previous chapter. We test three methods: linear regression, random forest, and support vector machine. All of these are compared against a baseline model of estimating the revenue by the mean revenue of the training set.
We performed a five-fold cross validation for every model Kuhn (2018). For each train and test set, we imputed missing values separately 20 times, which led to 100 slightly different datasets. The model was separately trained on each training set, while the errors were measured on the associated test set. The model parameters were previously selected for performance, but have not been optimized on the training set.
Table 4.1 reports the averages for the RMSE, MAE and the R^2 for each method as well as the standard deviation from those averages.

Model	RMSE	MAE	R^2 train	R^2 test
MP	1,679,633 (153,517)	1,251,279 (86,468)		
LR	1,774,202 (302,481)	132,8747 (233,502)	0.82 (0.01)	-0.18 (0.54)
BT	**1,114,597** (102,674)	**773,954** (74,149)	**0.75** (0.02)	**0.53** (0.14)
SVM	1,201,932 (63,325)	797,173 (45,437)	0.67 (0.03)	0.46 (0.10)

Table 4.1: Revenue prediction errors

Boosted trees deliver the best performance, but with a higher variance than SVM (Table 4.1). The R^2 test value can be negative, due to the fact that the mean value on the test set differs from the training set. Not only has the linear regression model the worst error, but it also has the highest variance among all models for all measures.

4.1.1 Linear Regression

We predicted revenues from a linear regression. The performance was compared to a baseline model estimation that simply guesses the mean of the training set. We demonstrate the performance across a range of cross-validation and imputation measures. Figure 4.1 illustrates the RMSE across all datasets.

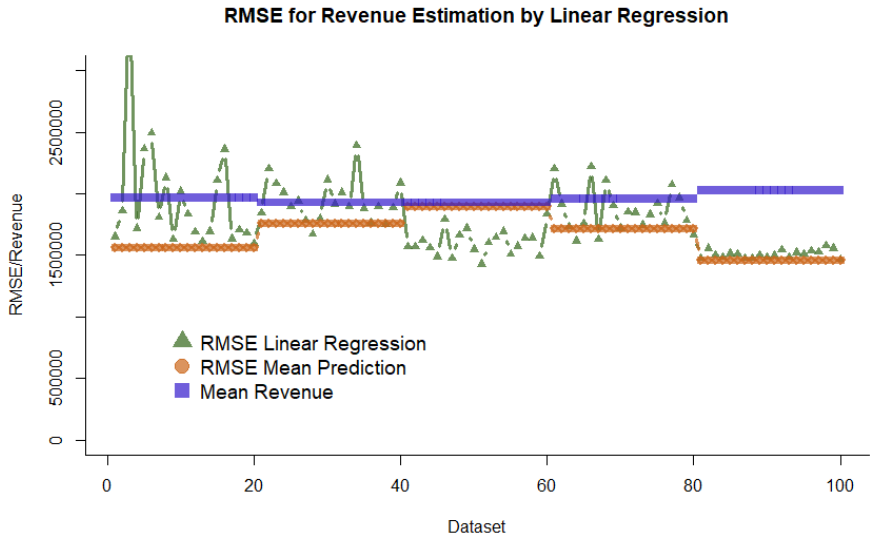

Figure 4.1: RMSE for linear regression revenue prediction

The variation induced by cross validation is visible by the five different means of mean prediction algorithm. Figure 4.1 shows that the linear regression method is susceptible to large swings from the imputed data. Next, Figure 4.2 depicts how the MAE evolves over datasets. The error in general is lower than with the RMSE, indicating that the errors are heavily influenced by outliers.

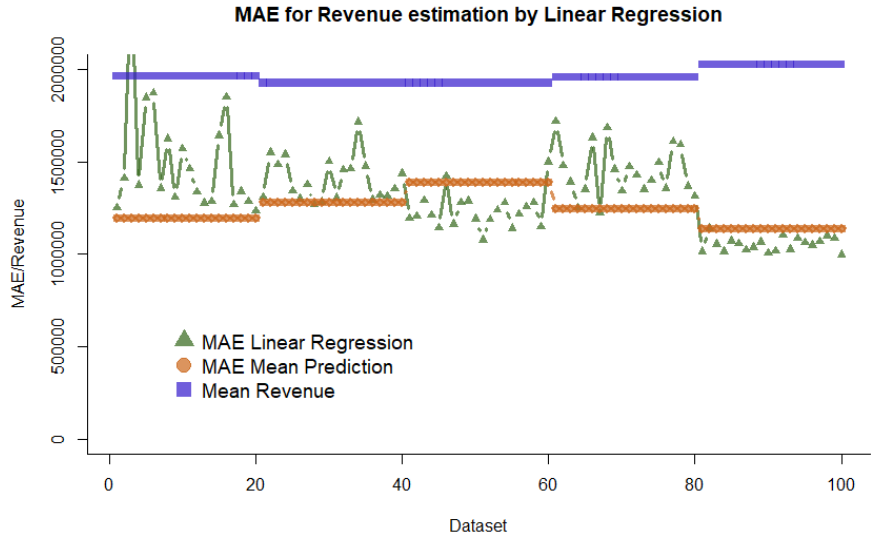

Figure 4.2: MAE for linear regression revenue prediction

Figure 4.2 reports the evolution of the training and test error. The training error is almost constant across all datasets, as indicated by the blue line in Figure 4.2.

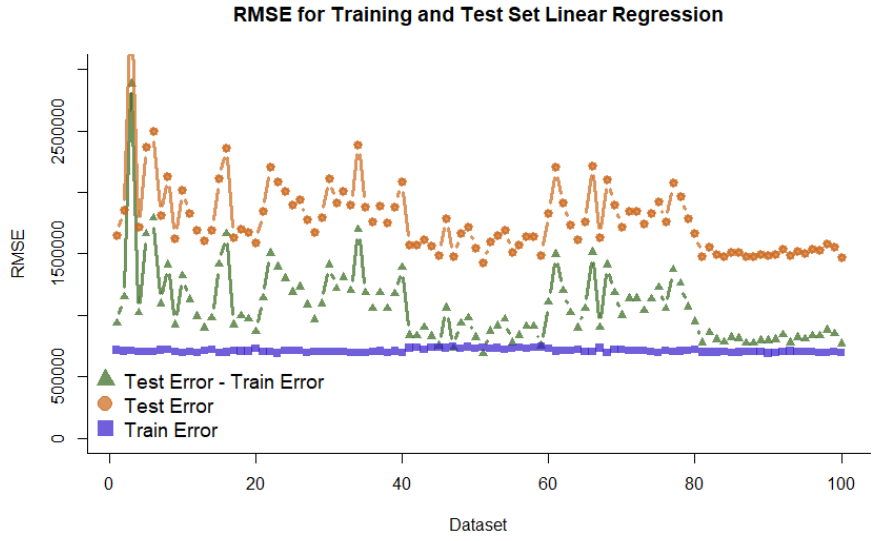

Figure 4.3: RMSE for train and test set

Due to the low predictive performance of the linear regression, we do not present any further details about the model.

4.1.2 Boosted Trees

We also estimated the revenue using a boosted trees model. The model was implemented with the eXtreme gradient boosting (XGBoost) library. The prediction for each test set

was compared to the base model. For the RMSE as error functions, we achieved the results in Figure 4.4. The mean prediction method varies again with the train-test split, however the boosted trees model shows less sensibility than linear regression (see Figure 4.1).

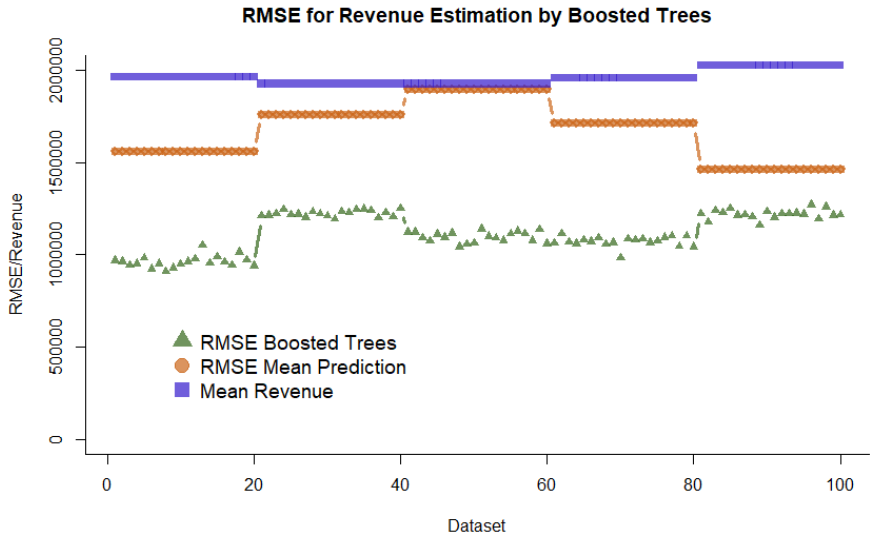

Figure 4.4: RMSE for boosted trees revenue prediction

Figure 4.5 reports the same computation for the MAE. The MAE shows a similar pattern, but lower overall error, indicative for large outliers in the data.

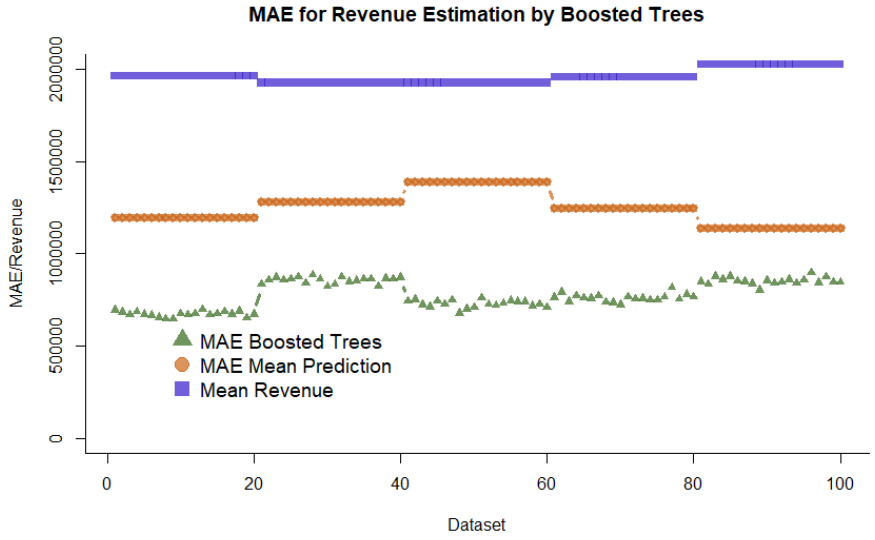

Figure 4.5: MAE for boosted trees revenue prediction

In order to demonstrate that our models did not over-fit as we suspected for Phillips et al. (2015), we illustrate the training RMSE in comparison with the test RMSE over all datasets in Figure 4.6. The train error is only slightly above the test error, suggesting a

good fit of the model. However some variation is introduced by cross validation.

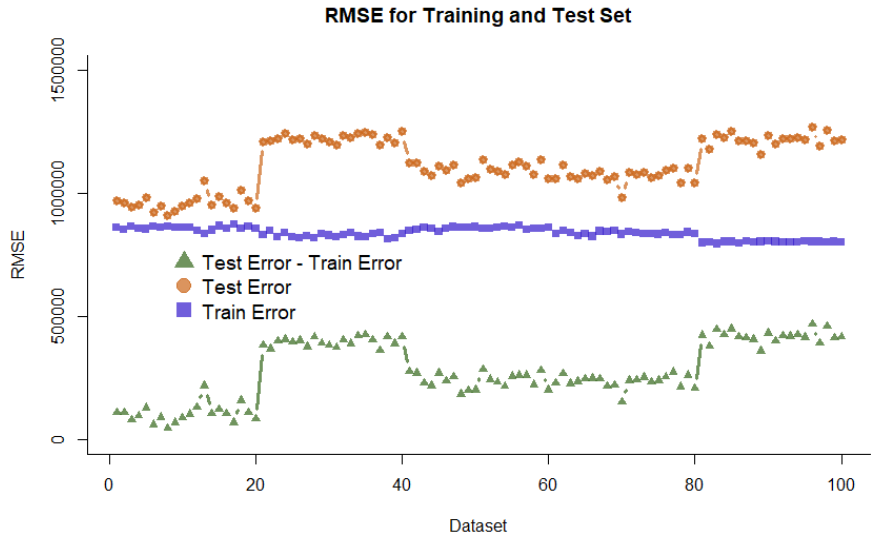

Figure 4.6: RMSE for train and test set

We further indicate which attributes were the most influential for the model based on Figure 4.7. The most important variables are the number of occupied rooms, a feature we introduced in Section 3.2.8. In accordance with literature, the number of reviews is an important feature as well. The class of the hotel, as indicated by the number of stars is the fourth most influential attribute.

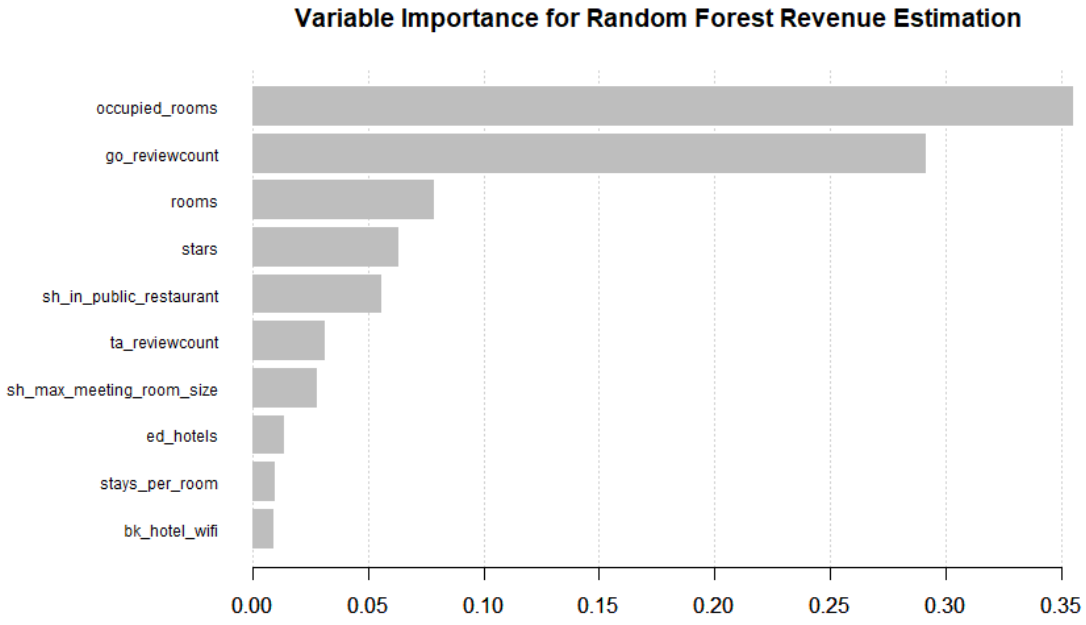

Figure 4.7: Variable importance for boosted trees revenue prediction

4.1.3 Support Vector Machine

We estimate the revenue using a SVM model with radial basis functions. The base model RMSE with the SVM model was compared on all datasets (see Figure 4.8). The SVM model shows less variance than the boosted trees model (see Figure 4.4). Comparing to the other two, it is least susceptible to perturbations from imputations.

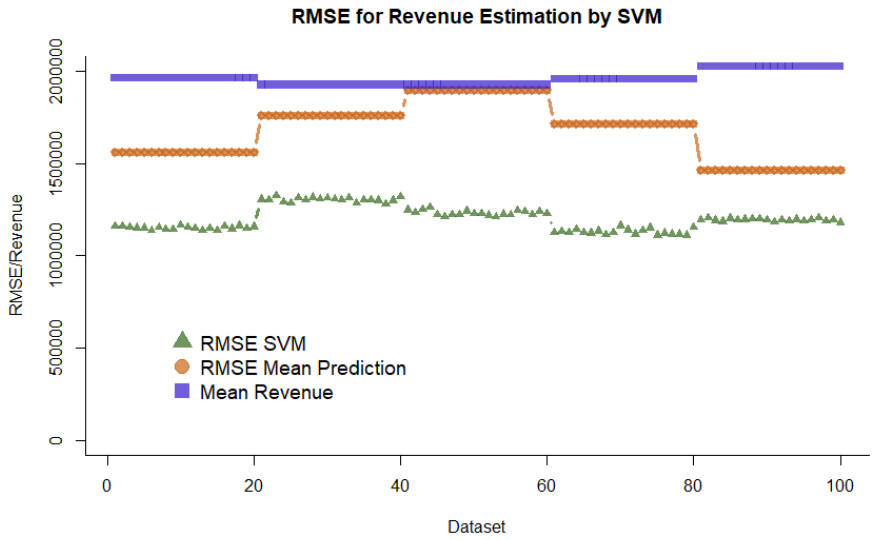

Figure 4.8: RMSE for SVM prediction

Figure 4.9 presents the MAE for the same models and the same datasets. The MAE evolution follows a similar path than the RMSE, only for a lower value. Again suggesting the presence of outliers.

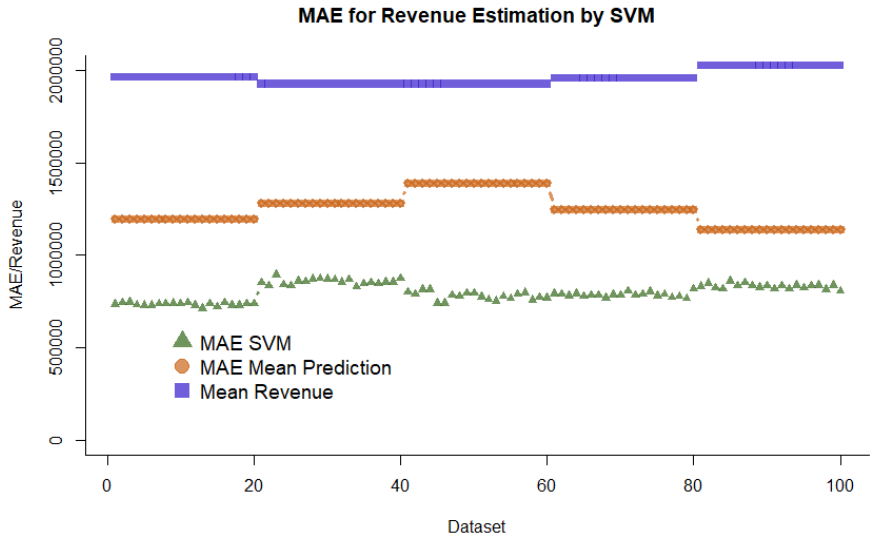

Figure 4.9: MAE for SVM prediction

For further information, Figure 4.10 evidences that our model does not over-fit on the training sets. As in the boosted trees case, the test set error is close to the training error and moves only with the cross validation.

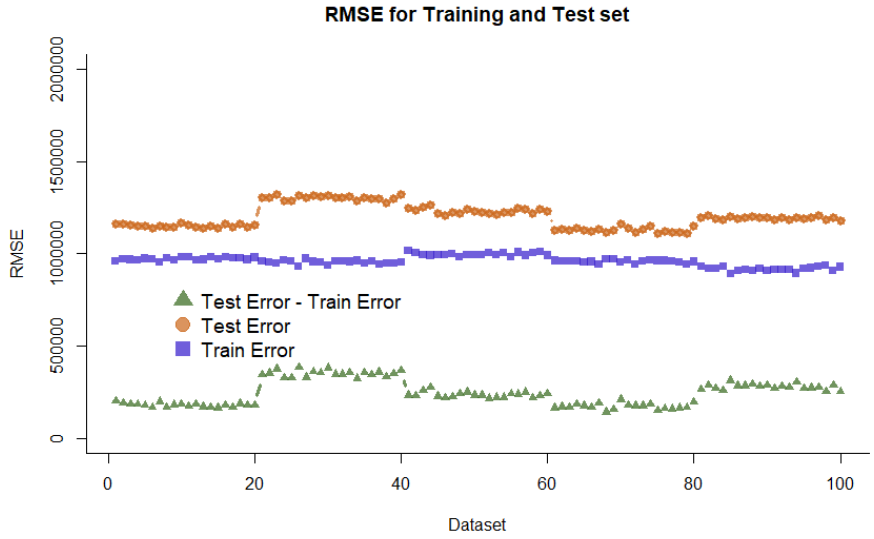

Figure 4.10: RMSE of train and test set for revenue SVM revenue prediction

For further discussion, Figure 4.11 provides an overview of the most important variables in the model. The result is comparable with Figure 4.7 from the boosted trees model, number of rooms are the most important attributes, followed by the number of reviews per hotel. Google is again more important than other platforms.

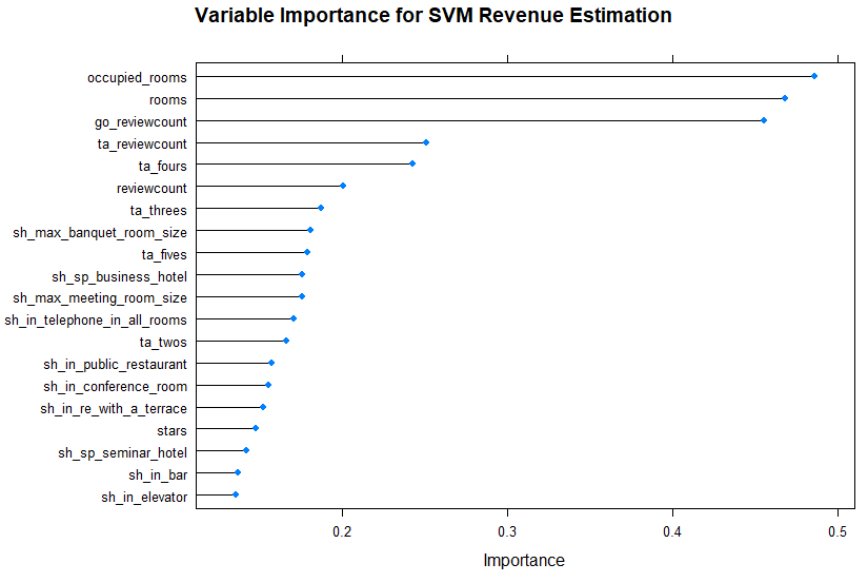

Figure 4.11: Variable importance for SVM revenue prediction

4.2 RevPAR

As in the case of revenue prediction, we generated 100 datasets for which we fit our linear regression, boosted trees, and SVM models. Table 4.2 summarizes the overall results. The boosted trees model delivers again the best performance, however with a significantly lower margin compared to revenue case (see Table 4.1). Another problem is the high variance in all of the models, suggestive of unstable predictions.

Model	RMSE	MAE	R^2 train	R^2 test
MP	186 (107.5)	106 (10.1)		
LR	324 (112.1)	208 (55.9)	0.41 (0.09)	-3.93 (4.39)
BT	**171** (112.4)	**87** (11.7)	**0.73** (0.09)	**0.16** (0.10)
SVM	173 (114.9)	87 (12.6)	0.22 (0.16)	0.13 (0.11)

Table 4.2: RevPAR prediction errors

4.2.1 Linear Regression

We first evaluated the linear regression model estimates of the RevPAR over the different datasets (see Figure 4.12). The first train-test split induces a large misrepresentation of hotels with extreme revenue values in the test set, leading to a comparatively large error in both models. As in the revenue case, linear regression is susceptible to variations from imputation, as indicated by the jumps inside the cross validation split.

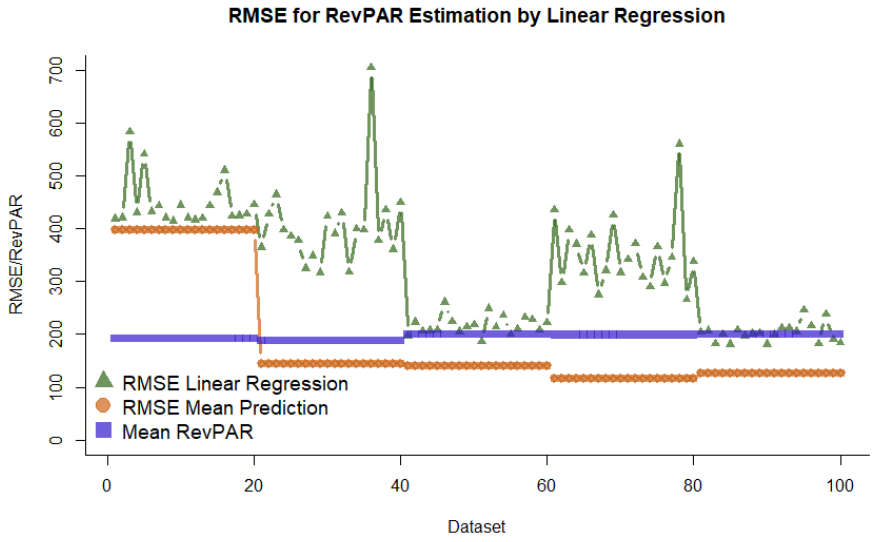

Figure 4.12: RMSE of the RevPAR prediction using linear regression

Figure 4.13 depicts the same fit but with the MAE error function. MAE delivers a similar image as RMSE, however with a larger advantage for the MP model.

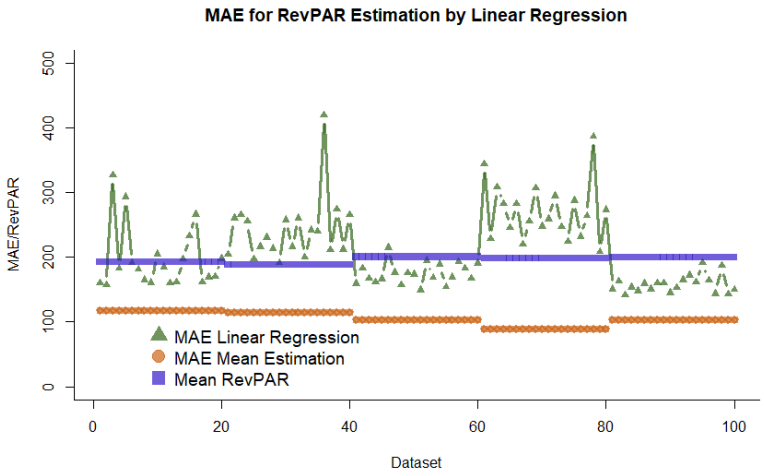

Figure 4.13: MAE of the RevPAR prediction using linear regression

In view of the poor fit, we refrain from reporting further details for this model.

4.2.2 Boosted Trees

We estimated the RevPAR with a boosted trees model. Figure 4.14 reports the RMSE of a mean guess and the results of the random forest model. The boosted tree model follows the MP model closely, suggesting it learns a similar prediction function. In that case to predict the mean of the training set, without regard for most attributes. Indicating almost no predictive value for all available attributes.

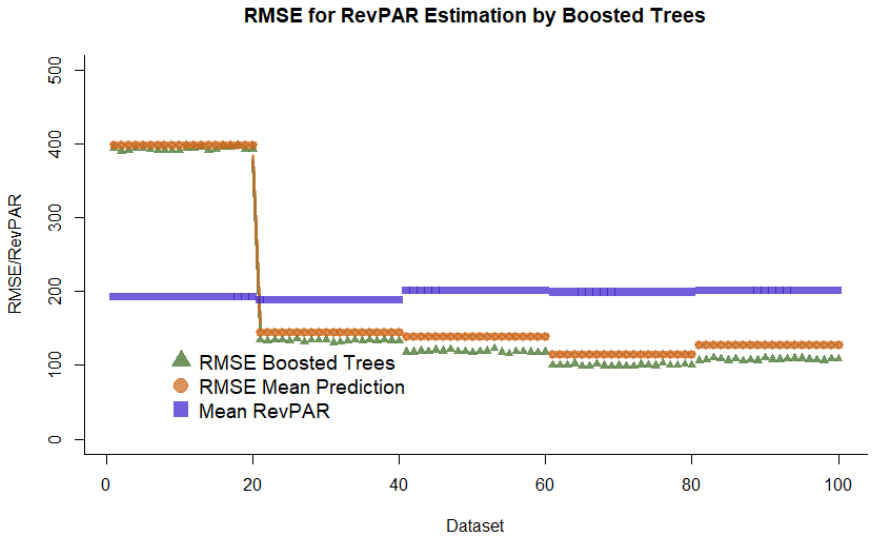

Figure 4.14: RMSE of the RevPAR prediction using a boosted trees

Figure 4.15 presents the MAE for the baseline and the boosted trees model. The same

case as for the RMSE error can be made, only a small improvement over the base model is made.

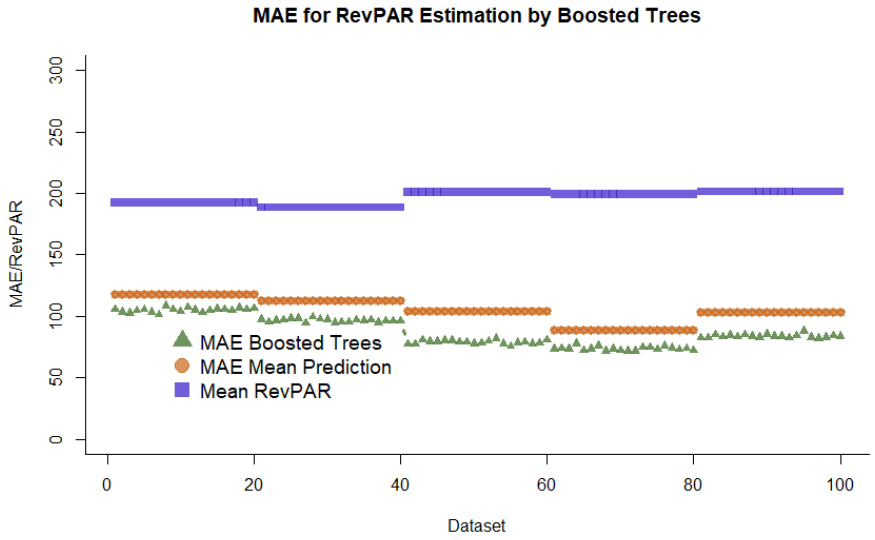

Figure 4.15: MAE of the RevPAR prediction using a boosted trees

For further discussion, Figure 4.16 includes a graph of the most important attributes for the RevPAR prediction. The only three attributes with some influence are those which have information about the numbers of rooms encoded, which includes the reviews per room or the number of occupied rooms.

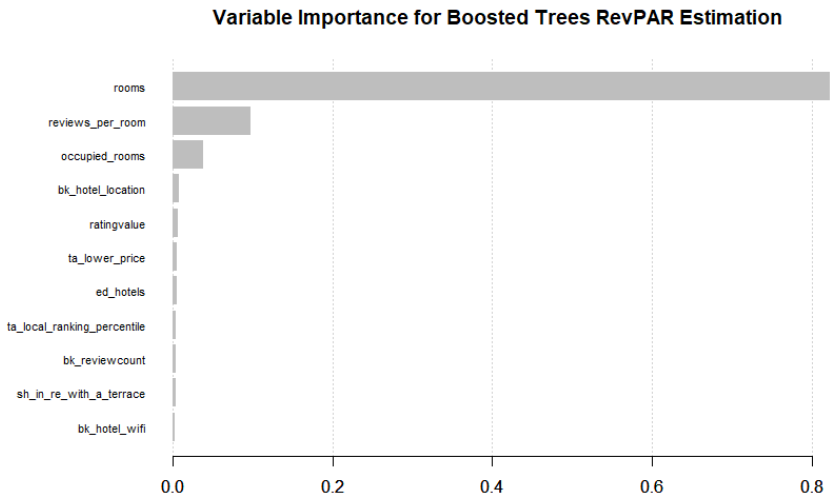

Figure 4.16: Variable importance for RevPAR prediction using a boosted trees

4.2.3 Support Vector Machine

We additionally estimated the RevPAR using a SVM model and report the RMSE in Figure 4.17. Similar as in the revenue prediction case, SVM follows the example of boosted trees in Figure 4.14. The baseline model is not outperformed in a significant way.

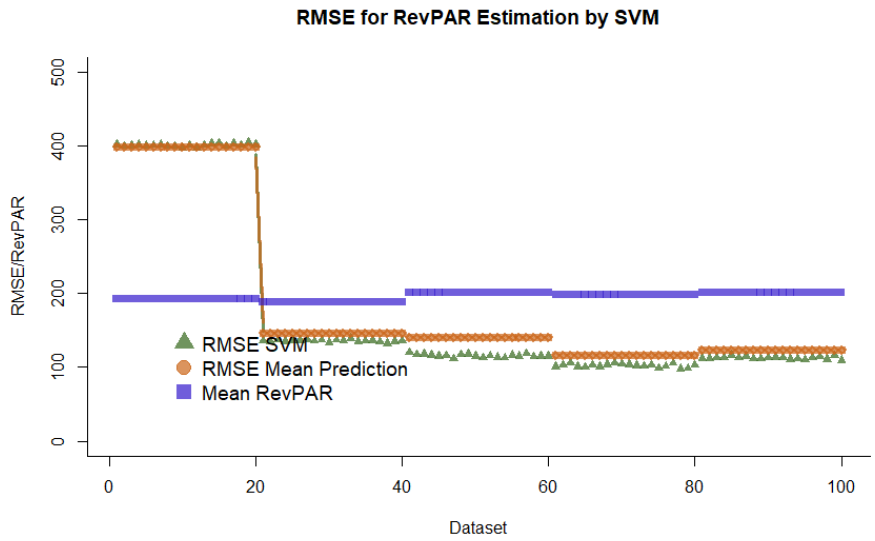

Figure 4.17: RMSE for the RevPAR prediction using a SVM

We did the same for the MAE (see Figure 4.18). The MAE depicts slightly better performance than in the RMSE case, but strongly depending on the train-test split by cross validation.

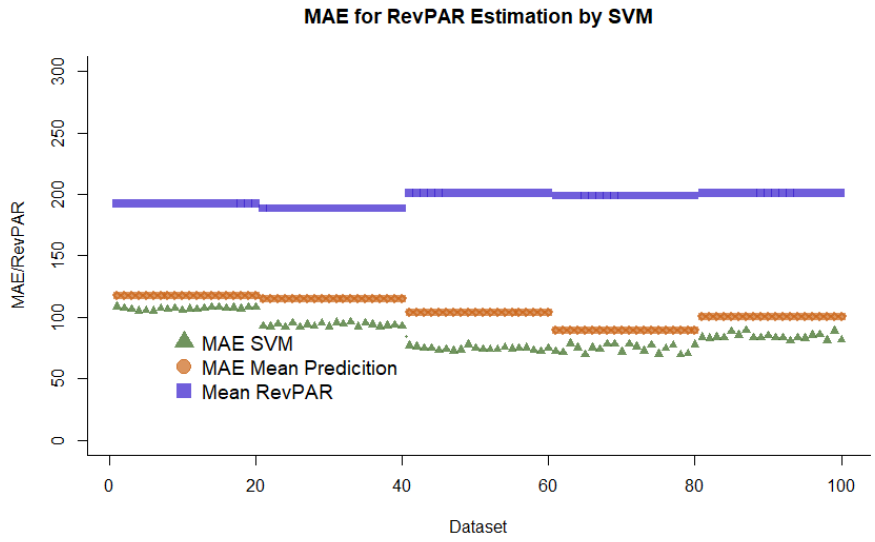

Figure 4.18: MAE for the RevPAR prediction using a SVM

Figure 4.19 depicts the importance of attributes for RevPAR prediction. Similarly as in Figure 4.16 for boosted trees, attributes related to the number of rooms are the most important. However online ratings related data is next, such as the overall rating value and Booking.com ratings. But in light of the overall performance in Figures 4.14 and 4.18, they most likely do not contribute in a significant way to an improved prediction.

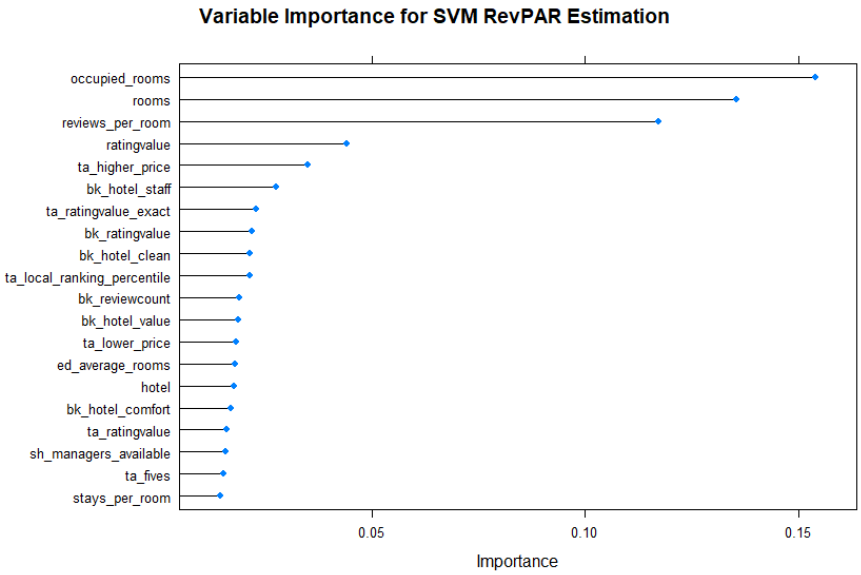

Figure 4.19: Variable importance for RevPAR prediction using SVM

Due to the low performance of all models, we will not show a train-test split evaluation.

4.3 Growth

The data was imputed 50 times and then split into a training and a test set. A boosted trees model was applied to predict into which of the two groups a hotel would be classified: growing (1) or shrinking (-1). We summarize the result in Table 4.3, giving the mean and standard deviation of our experiments. The true positive rate, or precision is over 72%. While the accuracy is at 68%.

Model	Accuracy	Precision	Recall	F1 Score
BT	0.681 (0.055)	0.729 (0.043)	0.789 (0.077)	0.756 (0.047)

Table 4.3: Growth prediction metrics

The data is slightly imbalanced, having 72 growing hotels compared to 43 shrinking ones. This gives us a baseline accuracy of 62%, if we always predict a positive result. This is also the case for the train-test split. However the high standard error in 4.3 shows that this result is not significant, without further variance reduction. We report the average confusion matrix in Figure 4.20 over 50 imputations.

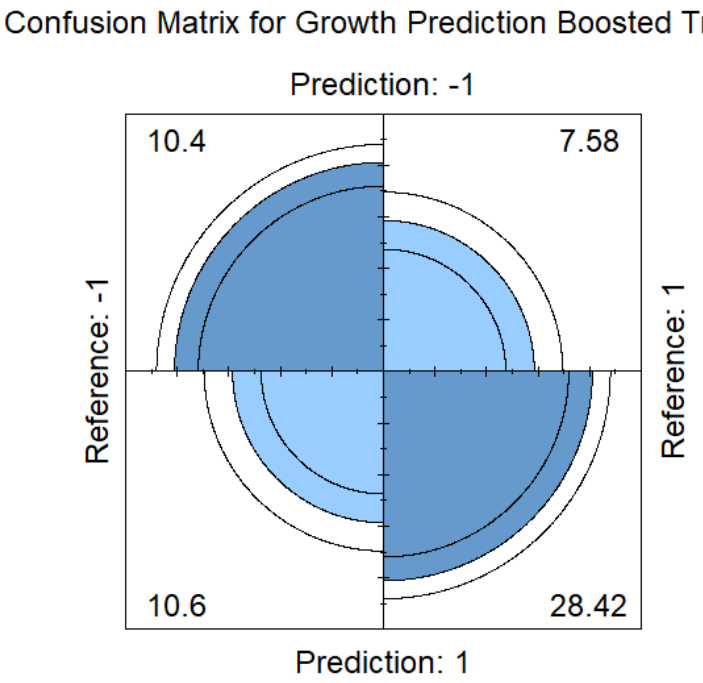

Figure 4.20: Confusion matrix for growth prediction using a boosted trees

Figure 4.21 displays an importance plot for the model to prove the usefulness of the generated variables. Curiously, the lateral position in Switzerland according to the GPS system is the most influential predictor. Followed by attributes which are derived from online ratings, the change in number of available rooms in the region, the change in variance in the hotel rating compared to two years ago, the change in the overall rating and the change in the number of reviews.

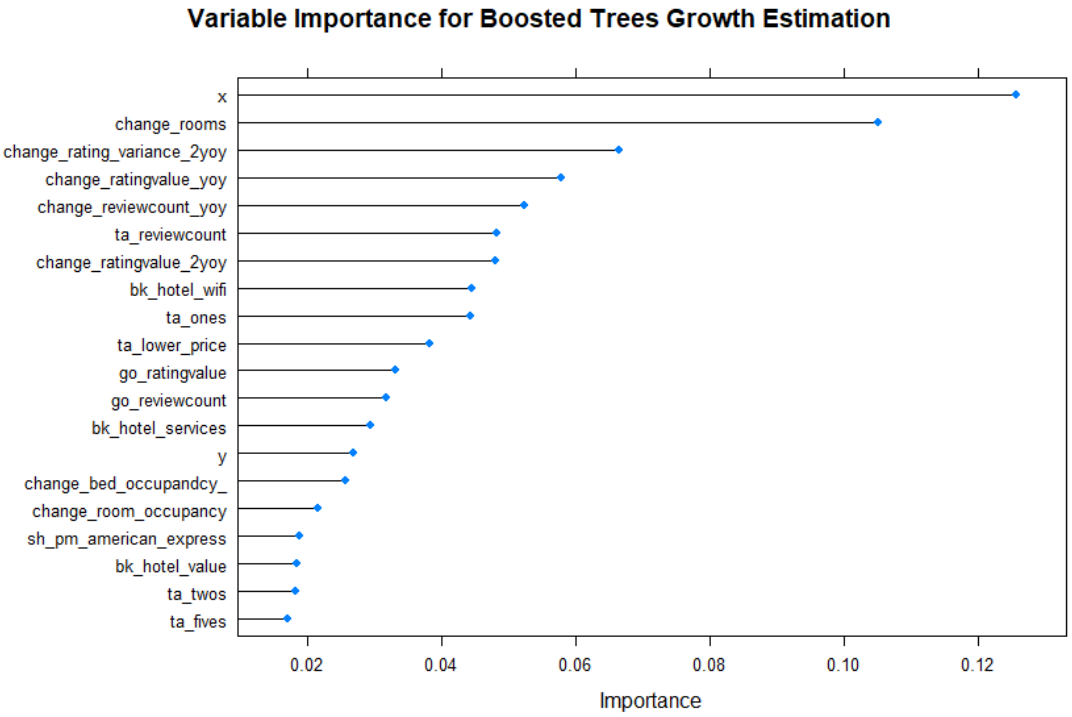

Figure 4.21: Variable importance for growth prediction using a boosted trees

Chapter 5

Discussion

Our approach of using online data to predict revenue, RevPAR, and growth has yielded mixed results. This chapter elaborates on these results and interprets the tables and graphs in the previous chapter.

5.1 Revenue

We demonstrated that boosted trees and SVM models produce a better revenue estimate than the baseline mean prediction model. Thus, our online data contain some additional information. The error is reduced by approximately one-third compared to the baseline (see Figures 4.5 and 4.9). However large outliers lead to a higher RMSE compared to MAE (compare Figures 4.9 and 4.8).
Moreover our models do not overfit the data (see Figures 4.6 and 4.10). Analysis of the most important features for the models shows the size of the hotel, i.e. the number of rooms is the most important predictor for revenue. In accordance with literature (Section 2.2), the number of reviews on different platforms is good proxy for sales (see Figures 4.7 and 4.11). Also influential are the number of higher ratings, five, four and three stars on TripAdvisor (see Figure 4.11) . However the number of specific star ratings could simply be an indicator for the number of reviews. Other predictive attributes for hotel revenue are more classical attributes, such as maximal size of the banquet room or the availability of public restaurants. These attributes are most likely proxies for the size of hotel, and hence indicators of larger absolute revenue.

5.2 RevPAR

For the RevPAR estimation, the models can barely improve on the baseline model (see Table 4.2). That means the more complicated models are only as good as a mean guess, or even worse for the linear regression case. The performance issue persists irrespective of the error measure (see Figures 4.15 and 4.17). This implies that much of what our models learn, is encoded in the number of rooms in a hotel. In the case of RevPAR, this information is already included in the response by the following definition:

$$\text{RevPAR} = \frac{\text{yearly revenue}}{360 \times \text{number of rooms}}$$

A closer look on the variable importance plot in Figures 4.19 and 4.16 shows that they dominated by the number of rooms. The online rating data has a comparatively lower power. However we should not give those figures too much consideration, as the predictions are not relevant.

5.3 Growth

We have evidenced that our online data have the potential to detect if a firm is growing or shrinking 4.3. Even though the confusion matrix shows a tendency towards false positives (see Figure 4.20). The variable importance plot in Figure 4.21 reveals surprisingly that the most important predictor is the lateral position in Switzerland. However it is closely followed by features generated from online ratings. The most important ones being, the change in available room in the region, the change in rating value variance over a two years, the change in rating value over a year and the change in the number of reviews over a year. Overall the Figure 4.21 is dominated by online rating attributes, in contrast to the plots from revenue prediction (see Figure 4.11). These findings suggest that at least part of the changes in business performance can be explained by eWOM. Hence FSP should consider monitoring eWOM of their clients, to better anticipate changes in their business needs.

5.4 Comparison

The paper with the most similar purpose Phillips et al. (2015), namely to predict RevPAR for hotels with similar online data, has reported a lower RMSE. However, since it did not feature a train/test split, we cannot directly compare the results.
Still, we can validate findings that indicate that the number of reviews is a valid proxy for sales in Switzerland (see Figure 4.7), although the number of Google reviews is seemingly a more accurate source of sales than the number of reviews on comparable platforms, such as Booking.com and TripAdvisor.

Chapter 6

Conclusion

We have proposed a method for estimating revenue and growth based on online data for hotels. Our method provides estimates of varying quality for those benchmarks. Growth seems to be more easily predictable than revenue or RevPAR. We have also established which attributes of eWOM are more reliable predictors for those use cases, and we have explained how to gather and process online data to make it usable for such purposes. Our results suggest that the prediction capacity of linear regression models is too low for the purpose of revenue prediction (Figure 4.1), whereas SVMs and random forests exceed baseline prediction abilities (Figures 4.4 and 4.8). Nevertheless, all of these models occasionally produce large outliers when predicting revenue and RevPAR.

6.1 Implications

This section addresses the implications of our results with regard to the research gap that was identified in Section 1.3. We specifically outline how these results influence the targeted fields.

6.1.1 Revenue and RevPAR

Our approach has evidenced that it is possible for FSPs to obtain rough estimates of hotel revenue from online data (Table 4.1). However, the results should be received with caution. In most cases, our tool is too crude for industry purposes, and the most important predictors are still classical attributes (Figure 4.7), such as the number of rooms or the presence of a restaurant. It would also be difficult to estimate the revenue of a new hotel, which would not yet have an online presence for eWOM.
From the academic perspective, we were able to validate results that link the number of customer reviews to sales. Moreover, we could cast doubt on the results of Phillips et al. (2015) (Section 2.2), which reported a particularly low RMSE of RevPAR estimates using a similar approach.

6.1.2 Growth

We have illustrated that our data have the capability to distinguish between growing and shrinking hotels (Table 4.3) – an avenue that was previously unexplored in the literature. This is especially helpful for FSPs that want to prioritize their resources based on such characteristics. Moreover our results show that among the most important predictors, for the development of a hotel, are changes in online ratings (Figure 4.21). Such as the change in rating variance, rating value and the number of reviews.

6.2 Limitations

Our predictions were limited by the amount of data we could collect in such a short time frame. For some sources, we had to use up-to-date summary ratings because of the difficulty or impossibility of collecting single reviews to create historical data. Hence, in some cases, there was information that should be disregarded.
A further avenue for improvement would be diversifying the review platforms or using meta-review data from platforms such as TrustYou. Moreover, a more detailed analysis of the text of reviews could further enhance accuracy by employing techniques from fields such as sentiment analysis to develop a more accurate image of customer feelings toward the hotel. However, this would be a challenging task in a country where multiple languages are spoken, such as in Switzerland. We also utilized data for only the limited time frame of 2012 to 2016. In that period, the hotel industry contended with currency shocks emanating from global markets and the Swiss National Bank. Such abrupt changes surely prompted behavioral changes from international customers, who comprise a key part of the business. The case of classifying growing and shrinking businesses seems especially promising, our study was not sensitive to rate of change. A hotel was classified as growing irrespective of size or rate. One possibility would be more strict on what classifies as changing, for example only use businesses with rates of change above 2%.

References

Agius, D. (2014, September). *The power of complaints.* Retrieved from http://hotelintel.co/2014/09/10/power-complaints/ (Accessed March 2018)

Anderson, C. (2012). The impact of social media on lodging performance. *Cornell Hospitality Report*. Retrieved from https://scholarship.sha.cornell.edu/chrpubs/5/

Anderson, C. K., & Lawrence, B. (2014). The influence of online reputation and product heterogeneity on service firm financial performance. *Service Science*, *6*(4), 217–228. doi: 10.1287/serv.2014.0080

Anindya Ghose, B. L., Panagiotis G. Ipeirotis. (2014, July). Examining the impact of ranking on consumer behavior and search engine revenue. *Management Science*. doi: 10.1287/mnsc.2013.1828

Ba, S., & Pavlou, P. A. (2002). Evidence of the effect of trust building technology in electronic markets: Price premiums and buyer behavior. *MIS Quarterly*, *26*(3), 243. doi: 10.2307/4132332

Black, P. E. (2004, December). *Ratcliff/obershelp pattern recognition.* Retrieved from https://www.nist.gov/dads/HTML/ratcliffObershelp.html (Accessed March 2018)

Blal, I., & Sturman, M. C. (2014, Jun). The differential effects of the quality and quantity of online reviews on hotel room sales. *Cornell Hospitality Quarterly*, *55*(4), 365–375. doi: 10.1177/1938965514533419

Bodner, T. E. (2008). What improves with increased missing data imputations. *PsycEXTRA Dataset*. doi: 10.1037/e645052007-001

Buuren, S. V., & Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations inr. *Journal of Statistical Software*, *45*(3). doi: 10.18637/jss.v045.i03

Chen, S.-y., Pei-Yu; Wu, & Yoon, J. (2004). Te impact of online recommendations and consumer feedback on sales. *ICIS 2004 Proceedings. 58*. Retrieved from http://aisel.aisnet.org/icis2004/58

Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD 16*. doi: 10.1145/2939672.2939785

Chevalier, J., & Mayzlin, D. (2003, August). The effect of word of mouth on sales: Online book reviews. *Journal of Marketing Research*, *43*, 345-354. doi: 10.3386/w10148

Cronin, B. (1997). Information management for the intelligent organization: The art of scanning the environment. *Information Processing & Management*, *33*(3), 405–406. doi: 10.1016/s0306-4573(97)88272-6

Dipendra Singh, E. T. (2015, January). Hotel online reviews and their impact on booking transaction value. In *Xvi annual conference proceedings.*

DMLC. (2018). *xgboost.* Retrieved from https://github.com/dmlc/xgboost/tree/master/demo#machine-learning-challenge-winning-solutions (Acces-

sed March 2018)

Duan, W., Gu, B., & Whinston, A. (2008b). The dynamics of online word-of-mouth and product sales—an empirical investigation of the movie industry. *Journal of Retailing*, *84*(2), 233–242. doi: 10.1016/j.jretai.2008.04.005

Duan, W., Gu, B., & Whinston, A. B. (2008a). Do online reviews matter? — an empirical investigation of panel data. *Decision Support Systems*, *45*(4), 1007–1016. doi: 10.1016/j.dss.2008.04.001

Edwin N. Torres, D. S. (2016, October). Towards a model of electronic word-of-mouth and its impact on the hotel industry. *International Journal of Hospitality & Tourism Administration*, 472-489. Retrieved from https://doi.org/10.1080/15256480.2016.1226155 doi: 10.1080/15256480.2016.1226155

FSO. (2016). *Tourism 2016.* Retrieved from https://www.bfs.admin.ch/bfs/en/home/statistics/tourism.html (Accessed March 2018)

FSO. (2017). *Kmu in zahlen: Firmen und beschäftigte.* Retrieved from https://www.kmu.admin.ch/kmu/de/home/kmu-politik/kmu-politik-zahlen-und-fakten/kmu-in-zahlen/firmen-und-beschaeftigte.html (Accessed March 2018)

Gelman, A., & Hill, J. (2016). *Data analysis using regression and multilevel/hierarchical models.* Cambridge University Press.

Goh, K. Y., Heng, C. S., & Lin, Z. (2012). Social media brand community and consumer behavior: Quantifying the relative impact of user- and marketer-generated content. *SSRN Electronic Journal*. doi: 10.2139/ssrn.2048614

Gupta, S., Hanssens, D., Hardie, B., & Kahn, W. (2006, November). Modeling customer lifetime value. *Journal of Service Research*, *9*(2), 139 - 155. doi: https://doi.org/10.1177/1094670506293810

Hastie, T. J., Tibshirani, R. J., & Friedman, J. H. (2009). *The elements of statistical learning* (Vol. 2). Springer.

Hu, N., Pavlou, P. A., & Zhang, J. (2006). Can online reviews reveal a products true quality? *Proceedings of the 7th ACM conference on Electronic commerce - EC 06*. doi: 10.1145/1134707.1134743

Jorge Mejia, A. G., Shawn Mankad. (2018). *A for effort? using the crowd to identify moral hazard in nyc restaurant hygiene inspections.* Retrieved from http://misrc.umn.edu/workshops/2018/spring/Hygiene_Circulation.pdf (Workshop, Accessed March 2018)

Karaman, H. (2017). *Competition and the impact of online reviews on product financial performance: Evidence from the hotel industry* (Unpublished doctoral dissertation). Goizueta Business School, Emory University.

Kim, S., Kim, J., & Park, S. (2017, Jul). The effects of perceived value, website trust and hotel trust on online hotel booking intention. *Sustainability*, *9*(12), 2262. doi: 10.3390/su9122262

Kuhn, M. (2018). *Training function in caret.* Retrieved from https://www.rdocumentation.org/packages/caret/versions/6.0-78/topics/train (Accessed March 2018)

Kumar, M. (2017). Investigation of credit risk based on indian firm performance. *International Journal of Risk and Contingency Management*, *6*(2), 35–46. doi: 10.4018/ijrcm.2017040103

Lai, K. K., Yu, L., Wang, S., & Huang, W. (2007). An intelligent crm system for identifying high-risk customers: An ensemble data mining approach. *Computational Science – ICCS 2007 Lecture Notes in Computer Science*, 486–489. doi: 10.1007/

978-3-540-72586-2_70

Levenshtein, V. I. (1965). Binary codes capable of correcting deletions, insertions, and reversals. *Doklady Akademii Nauk SSSR*.

Li, G., Law, R., Vu, H. Q., Rong, J., & Zhao, X. R. (2015). Identifying emerging hotel preferences using emerging pattern mining technique. *Tourism Management*, *46*, 311–321. doi: 10.1016/j.tourman.2014.06.015

Li, M., Huang, L., Tan, C.-H., & Wei, K.-K. (2013, Jul). Helpfulness of online product reviews as seen by consumers: Source and content features. *International Journal of Electronic Commerce*, *17*(4), 101–136. doi: 10.2753/jec1086-4415170404

Lu, Q., Xiao, L., & Ye, Q. (2012). Investigating the impact of online word-of-mouth on hotel sales with panel data. *2012 International Conference on Management Science & Engineering 19th Annual Conference Proceedings*. doi: 10.1109/icmse.2012.6414153

Lu, Q., Ye, Q., & Law, R. (2014). moderating effects of product heterogeneity between online word-of-mouth and hotel sales. *Journal of Electronic Commerce Research*, *15*(1).

Mackinnon, A. (2010, Oct). The use and reporting of multiple imputation in medical research - a review. *Journal of Internal Medicine*, *268*(6), 586–593. doi: 10.1111/j.1365-2796.2010.02274.x

Maria Psillaki, D. M., Ioannis E. Tsolas. (2010). Evaluation of credit risk based on firm performance. *European Journal of Operational Research*. doi: https://doi.org/10.1016/j.ejor.2009.03.032

Mayzlin, D., Dover, Y., & Chevalier, J. A. (2012, August). *Promotional reviews: An empirical investigation of online review manipulation* (Tech. Rep. No. 18340). National Bureau of Economic Research. Retrieved from http://www.nber.org/papers/w18340 doi: 10.3386/w18340

McKinsey. (2017, October). *Advanced analytics in hospitality.* Retrieved from https://www.mckinsey.com/business-functions/digital-mckinsey/our-insights/advanced-analytics-in-hospitality (Accessed March 2018)

Melián-González, S., Bulchand-Gidumal, J., & López-Valcárcel, B. G. (2013). Online customer reviews of hotels. *Cornell Hospitality Quarterly*, *54*(3), 274–283. doi: 10.1177/1938965513481498

Molinillo, S., Ximénez-De-Sandoval, J. L., Fernández-Morales, A., & Coca-Stefaniak, A. (2016). Hotel assessment through social media: The case of tripadvisor. *Tourism & Management Studies*, *12*(1), 15–24. doi: 10.18089/tms.2016.12102

Moloi, M. (2016). *The influence of online consumer reviews on purchasing intent* (Unpublished master's thesis). Wits Business School.

Muller, D., Te, F., Meyer, F., & Cvijikj, I. P. (2016). Towards data driven decision support for financial institutions: Predicting small companies business volume in switzerland. *2016 7th International Conference on Computer Science and Information Technology (CSIT)*. doi: 10.1109/csit.2016.7549449

Navarro, G. (2001, Jan). A guided tour to approximate string matching. *ACM Computing Surveys*, *33*(1), 31–88. doi: 10.1145/375360.375365

Pavlou, P. A., & Dimoka, A. (2006). The nature and role of feedback text comments in online marketplaces: Implications for trust building, price premiums, and seller differentiation. *Information Systems Research*, *17*(4), 392–414. doi: 10.1287/isre.1060.0106

Phillips, P., Barnes, S., Zigan, K., & Schegg, R. (2016, Apr). Understanding the impact of online reviews on hotel performance. *Journal of Travel Research*, *56*(2), 235–249. doi: 10.1177/0047287516636481

Phillips, P., Zigan, K., Silva, M. M. S., & Schegg, R. (2015). The interactive effects of online reviews on the determinants of swiss hotel performance: A neural network analysis. *Tourism Management*, *50*, 130–141. doi: 10.1016/j.tourman.2015.01.028

Pieterse, V., & Black, P. E. (2015, June). *Levenshtein distance.* Retrieved from https://www.nist.gov/dads/HTML/Levenshtein.html (Accessed March 2018)

Qi, L., & Qiang, Y. (2013). How hotel star rating moderates online word-of-mouth effect: A difference-in-difference approach. *2013 International Conference on Management Science and Engineering 20th Annual Conference Proceedings*. doi: 10.1109/icmse.2013.6586254

Raguseo, E., & Vitari, C. (2017, Dec). The effect of brand on the impact of e-wom on hotels' financial performance. *International Journal of Electronic Commerce*, *21*(2), 249–269. doi: 10.1080/10864415.2016.1234287

Ratcliff, J. W., & Metzener, D. (1988, July). Pattern matching: The gestalt approach. *Dr. Dobb's Journal*, 46. Retrieved from http://collaboration.cmc.ec.gc.ca/science/rpn/biblio/ddj/Website/articles/DDJ/1988/8807/8807c/8807c.htm

Scaglione, M., Schegg, R., & Murphy, J. (2009). Website adoption and sales performance in valais' hospitality industry. *Technovation*, *29*(9), 625–631. doi: 10.1016/j.technovation.2009.05.011

Scott, M., & Bruce, R. (1987). Five stages of growth in small business. *Long Range Planning*, *20*(3), 45–52. doi: 10.1016/0024-6301(87)90071-9

Soni, A., & Duggal, R. (2014). Reducing risk in kyc (know your customer) for large indian banks using big data analytics. *International Journal of Computer Applications*, *97*(9), 49–53. doi: 10.5120/17039-7347

Storey, D. (1994). *Understanding the small business sector.* Routledge.

Öğüt, H., & Taş, B. K. O. (2012). The influence of internet customer reviews on the online sales and prices in hotel industry. *The Service Industries Journal*, *32*(2), 197–214. doi: 10.1080/02642069.2010.529436

Toh, R. S., Raven, P., & Dekay, F. (2011, Oct). Selling rooms: Hotels vs. third-party websites. *Cornell Hospitality Quarterly*, *52*(2), 181–189. doi: 10.1177/1938965511400409

Tuominen, P. (2011). The influence of tripadvisor consumer-generated travel reviews on hotel performance. *UH Business School Working Paper*. Retrieved from http://hdl.handle.net/2299/7612

Venkatesan, R., & Kumar, V. (2004). A customer lifetime value framework for customer selection and resource allocation strategy. *Journal of Marketing*, *68*(4), 106–125. doi: 10.1509/jmkg.68.4.106.42728

Viglia, G., Minazzi, R., & Buhalis, D. (2016, Dec). The influence of e-word-of-mouth on hotel occupancy rate. *International Journal of Contemporary Hospitality Management*, *28*(9), 2035–2051. doi: 10.1108/ijchm-05-2015-0238

Wangenheim, F. V., & Bayón, T. (2004). The effect of word of mouth on services switching. *European Journal of Marketing*, *38*(9/10), 1173–1185. doi: 10.1108/03090560410548924

White, I. R., Royston, P., & Wood, A. M. (2011). Multiple imputation using chained equations: Issues and guidance for practice. *Statistics in Medicine*, *30*(4), 377-399. doi: https://doi.org/10.1002/sim.4067

Ye, Q., Law, R., & Gu, B. (2009). The impact of online user reviews on hotel room sales. *International Journal of Hospitality Management*, *28*(1), 180–182. doi: 10.1016/j.ijhm.2008.06.011

Ye, Q., Law, R., Gu, B., & Chen, W. (2011). The influence of user-generated content

on traveler behavior: An empirical investigation on the effects of e-word-of-mouth to hotel online bookings. *Computers in Human Behavior*, *27*(2), 634–639. doi: 10.1016/j.chb.2010.04.014

Yu, L. (2008). Credit risk analysis with a svm-based metamodeling ensemble approach. *Bio-Inspired Credit Risk Analysis*, 157–177. doi: 10.1007/978-3-540-77803-5_9

Zhang, Z., Ye, Q., & Law, R. (2011, Apr). Determinants of hotel room price. *International Journal of Contemporary Hospitality Management*, *23*(7), 972–981. doi: 10.1108/09596111111167551

Zhu, F., & Zhang, X. M. (2010). Impact of online consumer reviews on sales: The moderating role of product and consumer characteristics. *Journal of Marketing*, *74*(2), 133–148. doi: 10.1509/jmkg.74.2.133

Appendix A

Complementary information

A.1 Additional Tables

Tables which were too long to be displayed in the main part.

A.1.1 Swisshotel Data Treatment

A comprehensive list of which attribute was mapped to which to eliminate duplicates.

attribute name from	attribute name to
sh_infrastructure_internet_(gratuit)	sh_infrastructure_internet_(free_of_charges)
sh_infrastructure_radio_dans_la_chambre	sh_infrastructure_radio_in_all_rooms
sh_infrastructure_wifi_(gratuit)	sh_infrastructure_wifi_(free_of_charges)
sh_local_location_de_voiliers_bateaux	sh_local_yacht_and_boat_hire
sh_local_sentiers_pédestres	sh_local_hiking_trails
sh_infrastructure_restaurant_public	sh_infrastructure_public_restaurant
sh_infrastructure_climatisation_dans_la_chambre	sh_infrastructure_air_conditioning_in_all_rooms
sh_infrastructure_téléphone_dans_la_chambre	sh_infrastructure_telephone_in_all_rooms
sh_local_parcours_bike	sh_local_riding
sh_infrastructure_tv_dans_la_chambre	sh_infrastructure_tv_in_all_rooms
sh_local_canoë	sh_local_canoeing
sh_infrastructure_minibar_dans_la_chambre	sh_infrastructure_minibar_in_all_rooms
sh_local_navigation_à_moteur	sh_local_motor_boats
sh_infrastructure_ascenseur	sh_infrastructure_elevator
sh_local_ski_nautique	sh_local_water_skiing
sh_local_parcours_de_fitness	sh_local_keep-fit_circuit
sh_infrastructure_situation_centrale	sh_infrastructure_central_location
sh_local_equitation	sh_local_riding
sh_local_planches_à_voile	sh_local_windsurfing
sh_infrastructure_bâtiment_historique	sh_infrastructure_historic_building
sh_infrastructure_chambres_non-fumeurs	sh_infrastructure_non-smoking_rooms
sh_local_patinage	sh_local_skating
sh_infrastructure_proche_des_transports_publics	sh_infrastructure_close_to_public_transportation
sh_infrastructure_chambres_avec_wc	sh_infrastructure_rooms_with_wc

sh_infrastructure_garage_privé	sh_infrastructure_hotel-own_garage
sh_infrastructure_salle_de_réunion	sh_infrastructure_conference_room
sh_local_aérodrome	sh_local_airport
sh_infrastructure_chambres_très_calmes	sh_infrastructure_particularly_quiet_rooms
sh_salle_de_réunion	sh_infrastructure_conference_room
sh_chain_accorhotels_ibis_styles	sh_chain_accorhotels_ibis
sh_infrastructure_hoteleigene_garage	sh_infrastructure_hotel-own_garage
sh_infrastructure_haustiere_willkommen	sh_infrastructure_pets_welcome
sh_infrastructure_fumoir	sh_infrastructure_smoking_lounge
sh_infrastructure_hoteleigene_parkplätze	sh_infrastructure_hotel-own_car_park
sh_bankettraum	sh_salle_de_banquet
sh_specialization_seminarhotel	sh_specialization_seminar_hotel
sh_infrastructure_Öffentliches_restaurant	sh_infrastructure_public_restaurant
sh_infrastructure_restaurant_mit_terrasse	sh_infrastructure_restaurant_with_a_terrace
sh_infrastructure_lift	sh_infrastructure_elevator
sh_infrastructure_klimaanlage_im_zimmer	sh_infrastructure_air_conditioning_in_all_rooms
sh_infrastructure_tv_im_zimmer	sh_infrastructure_tv_in_all_rooms
sh_local_segel-_bootsvermietung	sh_local_yacht_and_boat_hire
sh_infrastructure_Öffentliche_parkplätze	sh_infrastructure_public_car_park
sh_infrastructure_telefon_im_zimmer	sh_infrastructure_telephone_in_all_rooms
sh_infrastructure_minibar_im_zimmer	sh_infrastructure_minibar_in_all_rooms
sh_infrastructure_nichtraucherzimmer	sh_infrastructure_non-smoking_rooms
sh_local_motorschifffahrt	sh_local_motor_boats
sh_infrastructure_snackrestaurant	sh_infrastructure_snack_restaurant
sh_infrastructure_wlan_(kostenlos)	sh_infrastructure_wifi_(free_of_charges)
sh_infrastructure_zentrale_lage	sh_infrastructure_central_location
sh_local_kanu	sh_local_canoeing
sh_infrastructure_zimmer_mit_wc	sh_infrastructure_rooms_with_wc
sh_infrastructure_gymnastik_fitnessraum	sh_infrastructure_gym_fitness_room
sh_infrastructure_sitzungszimmer	sh_infrastructure_conference_room
sh_seminarraum	sh_infrastructure_conference_room
sh_infrastructure_nähe_Öv	sh_infrastructure_close_to_public_transportation
sh_specialization_ausgezeichnete_küche	sh_specialization_excellent_cuisine
sh_local_skischule	sh_local_ski_school
sh_local_fitness-parcours	sh_local_keep-fit_circuit
sh_infrastructure_sehr_ruhige_lage	sh_infrastructure_particularly_quiet_location
sh_infrastructure_internet_(kostenlos)	sh_infrastructure_internet_(free_of_charges)
sh_specialization_bikehotel_(2011)	sh_specialization_bike_hotel_(2011)
sh_infrastructure_nichtraucherhotel	sh_infrastructure_non-smoking_hotel
sh_infrastructure_kinderspielplatz	sh_infrastructure_children's_playground
sh_local_langlaufloipen	sh_local_cross-country_skiing
sh_infrastructure_ladestation_elektroauto	sh_infrastructure_charging_station_electric_cars
sh_infrastructure_aussichtsrestaurant	sh_infrastructure_panoramic_restaurants
sh_infrastructure_sehr_ruhige_zimmer	sh_infrastructure_particularly_quiet_rooms
sh_infrastructure_spielzimmer	sh_infrastructure_play_room
sh_local_skilifte	sh_local_ski-lifts
sh_infrastructure_radio_im_zimmer	sh_infrastructure_radio_in_all_rooms
sh_specialization_wanderhotel	sh_specialization_hiking_hotel
sh_local_wanderwege	sh_local_hiking_trails

sh_infrastructure_kinderbetreuung	sh_infrastructure_child_care
sh_infrastructure_coiffeur_im_hotel	sh_infrastructure_hairdresser_at_the_hotel
sh_local_schlittschuhlaufen	sh_local_skating
sh_infrastructure_grillrestaurant	sh_infrastructure_grill_restaurant
sh_infrastructure_hallenbad	sh_infrastructure_indoor_pool
sh_infrastructure_cuisine_végétarienne	sh_infrastructure_vegetarian_food
sh_infrastructure_gymnastique_salle_de_fitness	sh_infrastructure_gym_fitness_room
sh_infrastructure_parkings_publics	sh_infrastructure_public_car_park
sh_specialization_excellente_cuisine	sh_specialization_excellent_cuisine
sh_infrastructure_jardin,_parc	sh_infrastructure_garden,_park_area
sh_infrastructure_restaurant_avec_terrasse	sh_infrastructure_restaurant_with_a_terrace
sh_infrastructure_coiffeur_à_l'hôtel	sh_infrastructure_hairdresser_at_the_hotel
sh_infrastructure_animaux_domestiques_bienvenus	sh_infrastructure_pets_welcome
sh_local_chemin_de_fer_à_crémaillère_funiculaire	sh_local_zahnrad-_standseilbahn
sh_infrastructure_parkings_privés	sh_infrastructure_hotel-own_garage
sh_infrastructure_hôtel_pour_non-fumeurs	sh_infrastructure_non-smoking_hotel
sh_specialization_hôtel_de_congrès	sh_specialization_conference_hotel
sh_infrastructure_piscine_extérieure	sh_infrastructure_outside_swimming_pool
sh_infrastructure_piscine_couverte	sh_infrastructure_indoor_pool
sh_specialization_hôtel_d'affaires	sh_specialization_business_hotel
sh_infrastructure_restaurant_panoramique	sh_infrastructure_panoramic_restaurants
sh_infrastructure_situation_très_calme	sh_infrastructure_particularly_quiet_location
sh_infrastructure_court_de_tennis	sh_infrastructure_tennis_court
sh_specialization_hôtel_de_séminaire	sh_specialization_seminar_hotel
sh_infrastructure_cuisine_à_base_d'aliments_co...	sh_infrastructure_wholefood_cuisine
sh_infrastructure_barrierefreies_badezimmer	sh_infrastructure_wheelchair_accessible_washbasin
sh_infrastructure_haltegriffe	sh_infrastructure_handles
sh_infrastructure_angebote_für_gäste_mit_sehbe...	sh_infrastructure_offer_for_guests_with_visual...
sh_infrastructure_betthöhe_45-50_cm	sh_infrastructure_bed_height_45-50cm
sh_infrastructure_rollstuhlgängige_toilette	sh_infrastructure_wheelchair_accessible_toilet
sh_infrastructure_bett_unterfahrbar	sh_infrastructure_wheelchair_accessible_elevator
sh_infrastructure_lift_bedingt_barrierefrei	sh_infrastructure_lift_partly_accessible
sh_infrastructure_bedingt_rollstuhlgängig_toil...	sh_infrastructure_public_areas_partly_accessible
sh_infrastructure_bedingt_barrierefreie_zimmer	sh_infrastructure_partly_accessible_rooms
sh_infrastructure_Öffentliche_bereiche_bedingt...	sh_infrastructure_public_areas_partly_accessible
sh_infrastructure_unterfahrbares_waschbecken	sh_infrastructure_wheelchair_accessible_washbasin
sh_specialization_kongresshotel	sh_specialization_conference_hotel
sh_local_reiten	sh_local_riding
sh_infrastructure_barrierefreier_frühstücksber...	sh_infrastructure_accessible_breakfast_area
sh_infrastructure_rollstuhl-shuttle__taxi	sh_infrastructure_wheelchair_accessible_hotel_...
sh_infrastructure_vegetarische_küche	sh_infrastructure_vegetarian_food
sh_infrastructure_Öffentliche_bereiche_barrier...	sh_infrastructure_public_areas_accessible
sh_infrastructure_lift_barrierefrei	sh_infrastructure_wheelchair_accessible_elevator
sh_infrastructure_barrierefreie_zimmer	sh_infrastructure_accessible_rooms
sh_local_flugplatz	sh_local_airport
sh_local_ecole_d'alpinisme	sh_local_mountaineering_school
sh_infrastructure_garderie_d'enfants	sh_infrastructure_child_care
sh_specialization_vélo_vtt_(2011)	sh_specialization_bike_hotel_(2011)
sh_specialization_hôtel_top_familles	sh_specialization_top_family_hotels

sh_local_pistes_de_ski_de_fond	sh_local_cross-country_skiing
sh_local_téléskis	sh_local_cross-country_skiing
sh_specialization_hôtel_golf	sh_specialization_golf_hotel
sh_infrastructure_régime_repas_diététiques	sh_infrastructure_light_diet_food
sh_infrastructure_espace_de_jeux_pour_enfants	sh_infrastructure_children's_playground
sh_local_ecole_de_ski	sh_local_ski_school
sh_infrastructure_tennis_couvert	sh_infrastructure_indoor_tennis_courts
sh_infrastructure_schonkost_diät	sh_infrastructure_light_diet_food
sh_infrastructure_historisches_gebäude	sh_infrastructure_historic_building
sh_infrastructure_seeanstoss	sh_infrastructure_lake_border
sh_infrastructure_garten_parkanlage	sh_infrastructure_garden,_park_area
sh_infrastructure_aussenschwimmbad	sh_infrastructure_outside_swimming_pool
sh_local_bergsteigerschule	sh_local_mountaineering_school
sh_local_autofreier_ort	sh_local_car-free_locality
sh_infrastructure_spa_behandlungen	sh_infrastructure_spa_treatments
sh_infrastructure_dampfbad	sh_local_thermal_bath
sh_local_bikewege	sh_local_riding
sh_specialization_schneesporthotel	sh_specialization_snow_sport_hotel
sh_specialization_familienfreundliches_hotel	sh_specialization_family_friendly_hotel
sh_infrastructure_traitements_spa	sh_infrastructure_spa_treatments
sh_infrastructure_station_de_recharge_pour_voi...	sh_infrastructure_charging_station_electric_cars
sh_local_bains_thermaux	sh_local_thermal_bath
sh_local_localité_sans_autos	sh_local_car-free_locality
sh_infrastructure_chambre_de_jeux	sh_infrastructure_play_room
sh_specialization_hôtel_pour_familles_(2011)	sh_specialization_family_friendly_hotel
sh_classification_5_étoiles	sh_classification_5_sterne
sh_specialization_bikehotel	sh_specialization_bike_hotel
sh_infrastructure_elektrisch_höhenverstellbare...	sh_infrastructure_height-adjustable_electric_bed
sh_infrastructure_badewanne	sh_infrastructure_bath_tub
sh_local_thermalbad	sh_local_thermal_bath
sh_specialization_hôtel_familles_bienvenues	sh_specialization_family_friendly_hotel
sh_specialization_hôtel_de_sport_d'hiver	sh_specialization_snow_sport_hotel
sh_specialization_suites	sh_specialization_suite

Table A.1: Data mapping for duplicate swisshotel attributes

A comprehensive list of how attribute names were shortened.

attribute name from	attribute name to
original_attribute	shortened_attribute
swissid	swissid
swisshotel	swisshotel
sh_classification_classification_hs	sh_cl_cl_hs
sh_rooms	sh_rooms
sh_infrastructure_close_to_public_transportation	sh_in_close_to_public_transpor
sh_local_hiking_trails	sh_lo_hiking_trails
sh_payment_method_american_express	sh_pm_american_express
sh_name	sh_name
sh_infrastructure_non_smoking_hotel	sh_in_non_smoking_hotel

sh_payment_method_postfinance_card	sh_pm_postfinance_card
sh_stars	sh_stars
sh_specialization_excellent_cuisine	sh_sp_excellent_cuisine
sh_infrastructure_wifi_free_of_charges	sh_in_wifi_free_of_charges
sh_infrastructure_internet_free_of_charges	sh_in_internet_free_of_charges
sh_infrastructure_tv_in_all_rooms	sh_in_tv_in_all_rooms
sh_check_out	sh_check_out
sh_infrastructure_vegetarian_food	sh_in_vegetarian_food
sh_payment_method_visa	sh_pm_visa
sh_meeting_room	sh_meeting_room
sh_infrastructure_non_smoking_rooms	sh_in_non_smoking_rooms
sh_infrastructure_pets_welcome	sh_in_pets_welcome
sh_infrastructure_public_car_park	sh_in_public_car_park
sh_code	sh_code
sh_city	sh_city
sh_specialization_country_inn	sh_sp_country_inn
sh_payment_method_maestro	sh_pm_maestro
sh_infrastructure_public_restaurant	sh_in_public_restaurant
sh_infrastructure_restaurant_with_a_terrace	sh_in_re_with_a_terrace
sh_payment_method_reka	sh_pm_reka
sh_street	sh_street
trust_you	trust_you
sh_managers	sh_managers
sh_banquet_room	sh_banquet_room
sh_infrastructure_conference_room	sh_in_conference_room
sh_check_in	sh_check_in
sh_infrastructure_particularly_quiet_location	sh_in_particularly_quiet_locat
sh_payment_method_mastercard	sh_pm_mastercard
sh_infrastructure_wifi_subject_to_charges	sh_in_wifi_subject_to_charges
sh_beds	sh_beds
sh_telephone	sh_telephone
sh_infrastructure_light_diet_food	sh_in_light_diet_food
sh_infrastructure_hotel_own_garage	sh_in_hotel_own_garage
sh_payment_method_diners_club	sh_pm_diners_club
sh_infrastructure_radio_in_all_rooms	sh_in_radio_in_all_rooms
sh_infrastructure_elevator	sh_in_elevator
sh_infrastructure_telephone_in_all_rooms	sh_in_telephone_in_all_rooms
sh_infrastructure_panoramic_restaurants	sh_in_panoramic_restaurants
sh_infrastructure_central_location	sh_in_central_location
sh_classification_swiss_lodge	sh_cl_swiss_lodge
sh_specialization_bike_hotel_2011	sh_sp_bike_hotel_2011
sh_infrastructure_lake_border	sh_in_lake_border
sh_infrastructure_particularly_quiet_rooms	sh_in_particularly_quiet_rooms
sh_infrastructure_rooms_with_wc	sh_in_rooms_with_wc
sh_chain_minotel	sh_ch_minotel
sh_local_fishing	sh_lo_fishing
sh_infrastructure_childrens_playground	sh_in_childrens_playground
sh_local_cross_country_skiing	sh_lo_cross_country_skiing
sh_local_golf	sh_lo_golf

sh_local_skating	sh_lo_skating
sh_local_cog_railway_funicular	sh_lo_cog_railway_funicular
sh_infrastructure_hotel_own_car_park	sh_in_hotel_own_car_park
sh_local_riding	sh_lo_riding
sh_local_windsurfing	sh_lo_windsurfing
sh_local_keep_fit_circuit	sh_lo_keep_fit_circuit
sh_local_motor_boats	sh_lo_motor_boats
sh_local_ski_lifts	sh_lo_ski_lifts
sh_local_ski_school	sh_lo_ski_school
sh_local_yacht_and_boat_hire	sh_lo_yacht_and_boat_hire
sh_local_casino	sh_lo_casino
sh_infrastructure_garden_park_area	sh_in_garden_park_area
sh_local_water_skiing	sh_lo_water_skiing
sh_infrastructure_wholefood_cuisine	sh_in_wholefood_cuisine
sh_local_mountaineering_school	sh_lo_mountaineering_school
sh_infrastructure_air_conditioning_in_all_rooms	sh_in_air_cond_in_all_rooms
sh_local_canoeing	sh_lo_canoeing
sh_payment_method_jcb	sh_pm_jcb
sh_local_curling	sh_lo_curling
sh_local_car_free_locality	sh_lo_car_free_loity
sh_infrastructure_bed_usable_for_hoists	sh_in_bed_usable_for_hoists
sh_infrastructure_wheelchair_accessible_washbasin	sh_in_wa_washbasin
sh_infrastructure_wheelchair_accessible_hotel_...	sh_in_wa_hotel_shuttle
sh_infrastructure_offer_for_guests_with_visual...	sh_in_guests_with_visual_impai
sh_infrastructure_bed_height_45_50cm	sh_in_bed_height_45_50cm
sh_infrastructure_accessible_bathroom_with_rol...	sh_in_ac_bath_with_roll_in_sho
sh_infrastructure_public_areas_partly_accessible	sh_in_public_areas_partly_ac
sh_specialization_business_hotel	sh_sp_business_hotel
sh_infrastructure_snack_restaurant	sh_in_snack_restaurant
sh_infrastructure_minibar_in_all_rooms	sh_in_minibar_in_all_rooms
sh_local_airport	sh_lo_airport
sh_specialization_seminar_hotel	sh_sp_seminar_hotel
sh_infrastructure_smoking_lounge	sh_in_smoking_lounge
sh_specialization_green_living	sh_sp_green_living
sh_infrastructure_bar	sh_in_bar
sh_payment_method_myone	sh_pm_myone
sh_infrastructure_historic_building	sh_in_historic_building
sh_payment_method_unionpay	sh_pm_unionpay
sh_chain_sorell_hotels	sh_ch_sorell_hotels
sh_payment_method_discover	sh_pm_discover
sh_infrastructure_motorway_hotel	sh_in_motorway_hotel
sh_infrastructure_indoor_pool	sh_in_indoor_pool
sh_infrastructure_sauna	sh_in_sauna
sh_chain_swiss_quality_hotels	sh_ch_swiss_quality_hotels
sh_infrastructure_internet_subject_to_charges	sh_in_internet_subject_to_char
sh_infrastructure_play_room	sh_in_play_room
sh_chain_ramada	sh_ch_ramada
sh_infrastructure_gym_fitness_room	sh_in_gym_fitness_room
sh_infrastructure_whirlpool	sh_in_whirlpool

sh_specialization_wellness	sh_sp_wellness
sh_infrastructure_outside_swimming_pool	sh_in_outside_swimming_pool
sh_infrastructure_grill_restaurant	sh_in_grill_restaurant
sh_infrastructure_partly_accessible_rooms	sh_in_partly_accessible_rooms
sh_infrastructure_entry_via_side_entrance	sh_in_entry_via_side_entrance
sh_infrastructure_lift_partly_accessible	sh_in_lift_partly_accessible
sh_infrastructure_wheelchair_accessible_elevator	sh_in_wa_elevator
sh_infrastructure_accessible_breakfast_area	sh_in_ac_breakfast_area
sh_chain_schweizer_jugendherbergen	sh_ch_schweizer_jugendherberge
sh_infrastructure_tennis_court	sh_in_tennis_court
sh_chain_accorhotels_ibis	sh_ch_accorhotels_ibis
sh_specialization_hiking_hotel	sh_sp_hiking_hotel
sh_local_mountain_biking	sh_lo_mountain_biking
sh_specialization_design_lifestyle	sh_sp_design_lifestyle
sh_infrastructure_suitable_for_groups	sh_in_suitable_for_groups
sh_infrastructure_wheelchair_accessible_toilet	sh_in_wa_toilet
sh_local_bowling	sh_lo_bowling
sh_infrastructure_partly_wheelchair_accessible...	sh_in_partly_wa_toilet
sh_chain_ambassador_swiss_hotels	sh_ch_ambassador_swiss_hotels
sh_infrastructure_massage	sh_in_massage
sh_infrastructure_charging_station_electric_cars	sh_in_charging_station_electri
sh_infrastructure_accessible_rooms	sh_in_accessible_rooms
sh_local_thermal_bath	sh_lo_thermal_bath
sh_infrastructure_public_areas_accessible	sh_in_public_areas_accessible
sh_specialization_bike_hotel	sh_sp_bike_hotel
sh_specialization_historic_hotel	sh_sp_historic_hotel
sh_chain_swiss_historic_hotels	sh_ch_swiss_historic_hotels
sh_chain_romantik_hotels_restaurants	sh_ch_romantik_hos_res
sh_chain_hotels_with_a_bookmark	sh_ch_hotels_with_a_bookmark
sh_classification_swiss_lodge_garni	sh_cl_swiss_lodge_garni
sh_infrastructure_wheelchair_accessible_parkin...	sh_in_wa_parking_space
sh_chain_claire_george_hotelspitex	sh_ch_claire_george_hospitex
sh_infrastructure_hotelspitexch	sh_in_hotelspitexch
sh_infrastructure_hairdresser_at_the_hotel	sh_in_hairdresser_at_the_hotel
sh_specialization_health_hotel	sh_sp_health_hotel
sh_chain_the_leading_hotels_of_the_world	sh_ch_the_leading_hos_of_the_w
sh_specialization_wellness_i	sh_sp_wellness_i
sh_chain_swiss_deluxe_hotels	sh_ch_swiss_deluxe_hotels
sh_specialization_conference_hotel	sh_sp_conference_hotel
sh_infrastructure_handles	sh_in_handles
sh_infrastructure_shower_chair	sh_in_shower_chair
sh_infrastructure_indoor_tennis_courts	sh_in_indoor_tennis_courts
sh_classification_in_progress	sh_cl_in_progress
sh_chain_swiss_charme_hotels	sh_ch_swiss_charme_hotels
sh_infrastructure_bath_tub	sh_in_bath_tub
sh_infrastructure_height_adjustable_electric_bed	sh_in_height_adj_electric_bed
sh_specialization_golf_hotel	sh_sp_golf_hotel
sh_specialization_wellness_spa	sh_sp_wellness_spa
sh_chain_club_grand_hotel_palace	sh_ch_club_grand_hotel_palace

sh_specialization_family_friendly_hotel	sh_sp_family_friendly_hotel
sh_specialization_snow_sport_hotel	sh_sp_snow_sport_hotel
sh_infrastructure_child_care	sh_in_child_care
sh_infrastructure_steam_bath	sh_in_steam_bath
sh_infrastructure_spa_treatments	sh_in_spa_treatments
sh_chain_relais_du_silence	sh_ch_relais_du_silence
sh_specialization_sustainable_living	sh_sp_sustainable_living
sh_specialization_wellness_ii	sh_sp_wellness_ii
sh_chain_best_3_star_hotels	sh_ch_best_3_star_hotels
sh_chain_swiss_premium_hotels	sh_ch_swiss_premium_hotels
sh_infrastructure_solarium	sh_in_solarium
sh_chain_zurich_city_hotels	sh_ch_zurich_city_hotels
sh_chain_private_selection_hotels	sh_ch_private_selection_hotels
sh_specialization_unique	sh_sp_unique
sh_chain_hauenstein_hotels	sh_ch_hauenstein_hotels
sh_infrastructure_offer_for_guests_with_hearin...	sh_in_guests_with_hearing_impa
sh_classification_0_star	sh_cl_0_star
sh_chain_accorhotels_novotel	sh_ch_accorhotels_novotel
sh_infrastructure_accessible_indoor_pool	sh_in_accessible_indoor_pool
sh_specialization_international_chain_hotel	sh_sp_international_ch_hotel
sh_chain_premium_swiss_family_hotels	sh_ch_premium_swiss_family_hos
sh_specialization_top_family_hotels	sh_sp_top_family_hotels
sh_classification_5_etoiles_superior	sh_cl_5_etoiles_superior
sh_payment_method_twint	sh_pm_twint
sh_salle_de_banquet	sh_salle_de_banquet
sh_chain_relais_chateaux	sh_ch_relais_chateaux
sh_infrastructure_water_sports_offer	sh_in_water_sports_offer
sh_local_zahnrad_standseilbahn	sh_lo_zahnrad_standseilbahn
sh_classification_4_sterne	sh_cl_4_sterne
sh_specialization_businesshotel	sh_sp_businesshotel
sh_chain_accorhotels_mercure	sh_ch_accorhotels_mercure
sh_classification_4_sterne_superior	sh_cl_4_sterne_superior
sh_infrastructure_ski_in_ski_out_hotel	sh_in_ski_in_ski_out_hotel
sh_chain_manotel	sh_ch_manotel
sh_chain_best_western_hotels	sh_ch_best_western_hotels
sh_chain_nh_switzerland	sh_ch_nh_switzerland
sh_chain_movenpick_hotels_resorts	sh_ch_movenpick_hotels_resorts
sh_chain_sunstar_hotels	sh_ch_sunstar_hotels
sh_specialization_family_friendly_hotel_2011	sh_sp_family_friendly_ho_2011
sh_chain_sv_hotel	sh_ch_sv_hotel
sh_classification_international_chain_hotel	sh_cl_international_ch_hotel
sh_chain_accorhotels_ibis_budget	sh_ch_accorhotels_ibis_budget
sh_classification_1_star_superior_garni	sh_cl_1_star_superior_garni
sh_specialization_apartment_hotel	sh_sp_apartment_hotel
sh_specialization_medical_wellness_spa	sh_sp_medical_wellness_spa
sh_infrastructure_pool_lift	sh_in_pool_lift
sh_infrastructure_accessible_outdoor_swimming_...	sh_in_ac_outdoor_swimming_pool
sh_chain_club_med	sh_ch_club_med
sh_infrastructure_snack_bar	sh_in_snack_bar

sh_infrastructure_grill	sh_in_grill
sh_chain_swissotel	sh_ch_swissotel
sh_specialization_medical_wellness	sh_sp_medical_wellness
sh_chain_accorhotels_pullman	sh_ch_accorhotels_pullman
sh_chain_accorhotels_adagio	sh_ch_accorhotels_adagio
sh_classification_5_sterne	sh_cl_5_sterne
sh_specialization_gite_detape	sh_sp_gite_detape
sh_classification_5_sterne_superior	sh_cl_5_sterne_superior
sh_classification_1_star_garni	sh_cl_1_star_garni
sh_chain_accorhotels_mgallery	sh_ch_accorhotels_mgallery
sh_classification_1_star_superior	sh_cl_1_star_superior
sh_payment_method_apple_pay	sh_pm_apple_pay
sh_classification_2_sterne	sh_cl_2_sterne
sh_classification_5_etoiles_international_chai...	sh_cl_5_etoiles_international_
sh_chain_accorhotels_suitehotel	sh_ch_accorhotels_suitehotel
sh_specialization_suite	sh_sp_suite
sh_classification_0_star_garni	sh_cl_0_star_garni
sh_classification_3_sterne	sh_cl_3_sterne
sh_chain_zg_hotels	sh_ch_zg_hotels
sh_payment_method_samsung_pay	sh_pm_samsung_pay
sh_google_name	sh_google_name
sh_google_ratingvalue	sh_google_ratingvalue
sh_google_reviewcount	sh_google_reviewcount
sh_coordinates	sh_coordinates
sh_x	sh_x
sh_y	sh_y
sh_check_in_specified	sh_check_in_specified
sh_24_hours_check_in	sh_24_hours_check_in
sh_max_meeting_room_size	sh_max_meeting_room_size
sh_max_banquet_room_size	sh_max_banquet_room_size
sh_nb_stars	sh_nb_stars
sh_managers_available	sh_managers_available
sh_nb_managers	sh_nb_managers
sh_manager_couple	sh_manager_couple

Table A.2: Shortening swisshotel attributes to 30 characters per attribute

A.1.2 Summary of the Revenue Prediction Table

Attribute	Min	Median	Mean	Max	NAs
bk_hotel_wifi	4.50	8.70	8.53	10.00	73
bk_hotel_comfort	5.800	8.100	8.104	9.900	67
bk_hotel_services	5.100	8.000	7.985	9.800	67
bk_hotel_clean	6.400	9.000	8.887	10.000	67
bk_reviewcount	10.0	269.0	423.3	3327.0	62
bk_hotel_location	5.800	8.900	8.836	9.800	67
bk_ratingvalue	1.000	3.500	3.427	5.000	62

bk_hotel_value	5.800	7.900	7.850	9.300	67
bk_hotel_staff	6.100	9.000	8.862	10.000	67
ta_ratingvalue	1.500	4.000	4.065	5.000	52
ta_reviewcount	1.0	74.0	146.5	2247.0	52
ta_lower_price	34.0	145.0	153.6	542.0	141
ta_higher_price	78.0	253.0	287.0	998.0	141
ta_local_ranking_percentile	0.0100	0.5000	0.4859	1.0000	56
x	45.99	46.69	46.71	47.72	
y	6.140	8.040	8.301	10.364	
go_ratingvalue	1.000	4.200	4.154	5.000	
go_reviewcount	0.00	33.00	43.85	395.00	
sh_cl_cl_hs	0.0000	1.0000	0.5703	1.0000	
sh_in_close_to_public_transpor	0.0000	0.0000	0.3289	1.0000	
sh_lo_hiking_trails	0.0000	0.0000	0.4164	1.0000	
sh_pm_american_express	0.0000	0.0000	0.4695	1.0000	
sh_in_non_smoking_hotel	0.0000	0.0000	0.3793	1.0000	
sh_pm_postfinance_card	0.000	1.000	0.504	1.000	
sh_sp_excellent_cuisine	0.0000	0.0000	0.0504	1.0000	
sh_in_wifi_free_of_charges	0.0000	0.0000	0.4324	1.0000	
sh_in_internet_free_of_charges	0.0000	0.0000	0.1936	1.0000	
sh_in_tv_in_all_rooms	0.0000	1.0000	0.5676	1.0000	
sh_in_vegetarian_food	0.0000	0.0000	0.2308	1.0000	
sh_pm_visa	0.0000	1.0000	0.5782	1.0000	
sh_in_non_smoking_rooms	0.0000	0.0000	0.4721	1.0000	
sh_in_pets_welcome	0.0000	0.0000	0.3077	1.0000	
sh_in_public_car_park	0.0000	0.0000	0.2414	1.0000	
sh_pm_maestro	0.0000	1.0000	0.5597	1.0000	
sh_in_public_restaurant	0.0000	0.0000	0.4324	1.0000	
sh_in_re_with_a_terrace	0.0000	0.0000	0.3714	1.0000	
sh_pm_reka	0.0000	0.0000	0.2759	1.0000	
sh_in_conference_room	0.0000	0.0000	0.3369	1.0000	
sh_in_particularly_quiet_locat	0.0000	0.0000	0.1989	1.0000	
sh_pm_mastercard	0.0000	1.0000	0.5809	1.0000	
sh_in_light_diet_food	0.0000	0.0000	0.1088	1.0000	
sh_in_hotel_own_garage	0.000	0.000	0.191	1.000	
sh_pm_diners_club	0.0000	0.0000	0.3024	1.0000	
sh_in_radio_in_all_rooms	0.0000	0.0000	0.4695	1.0000	
sh_in_elevator	0.000	0.000	0.496	1.000	
sh_in_telephone_in_all_rooms	0.0000	0.0000	0.4801	1.0000	
sh_in_panoramic_restaurants	0.0000	0.0000	0.1406	1.0000	
sh_in_central_location	0.0000	0.0000	0.4854	1.0000	
sh_sp_bike_hotel_2011	0.00000	0.00000	0.09019	1.00000	
sh_in_particularly_quiet_rooms	0.0000	0.0000	0.2202	1.0000	
sh_in_rooms_with_wc	0.0000	0.0000	0.4775	1.0000	
sh_lo_fishing	0.0000	0.0000	0.1432	1.0000	
sh_in_childrens_playground	0.00000	0.00000	0.06897	1.00000	
sh_lo_cross_country_skiing	0.0000	0.0000	0.2838	1.0000	
sh_lo_golf	0.0000	0.0000	0.2016	1.0000	
sh_lo_skating	0.0000	0.0000	0.3395	1.0000	

sh_lo_cog_railway_funicular	0.0000	0.0000	0.2838	1.0000
sh_in_hotel_own_car_park	0.0000	0.0000	0.3873	1.0000
sh_lo_riding	0.0000	0.0000	0.2228	1.0000
sh_lo_windsurfing	0.0000	0.0000	0.1141	1.0000
sh_lo_keep_fit_circuit	0.0000	0.0000	0.3183	1.0000
sh_lo_motor_boats	0.0000	0.0000	0.1485	1.0000
sh_lo_ski_lifts	0.0000	0.0000	0.3077	1.0000
sh_lo_ski_school	0.0000	0.0000	0.2944	1.0000
sh_lo_yacht_and_boat_hire	0.0000	0.0000	0.1485	1.0000
sh_lo_casino	0.0000	0.0000	0.1194	1.0000
sh_in_garden_park_area	0.0000	0.0000	0.2308	1.0000
sh_lo_water_skiing	0.0000	0.0000	0.1247	1.0000
sh_in_wholefood_cuisine	0.00000	0.00000	0.09019	1.00000
sh_lo_mountaineering_school	0.0000	0.0000	0.1936	1.0000
sh_lo_canoeing	0.0000	0.0000	0.1459	1.0000
sh_pm_jcb	0.0000	0.0000	0.2016	1.0000
sh_lo_curling	0.0000	0.0000	0.2812	1.0000
sh_lo_car_free_loity	0.00000	0.00000	0.07692	1.00000
sh_in_bed_usable_for_hoists	0.00000	0.00000	0.06631	1.00000
sh_in_wa_washbasin	0.00000	0.00000	0.09814	1.00000
sh_in_wa_hotel_shuttle	0.00000	0.00000	0.06101	1.00000
sh_in_guests_with_visual_impai	0.00000	0.00000	0.06101	1.00000
sh_in_bed_height_45_50cm	0.00000	0.00000	0.05836	1.00000
sh_in_public_areas_partly_ac	0.00000	0.00000	0.08223	1.00000
sh_sp_business_hotel	0.00000	0.00000	0.05305	1.00000
sh_in_snack_restaurant	0.0000	0.0000	0.2016	1.0000
sh_in_minibar_in_all_rooms	0.0000	0.0000	0.2308	1.0000
sh_lo_airport	0.00000	0.00000	0.07958	1.00000
sh_sp_seminar_hotel	0.00000	0.00000	0.09284	1.00000
sh_in_smoking_lounge	0.0000	0.0000	0.0557	1.0000
sh_in_bar	0.0000	0.0000	0.2891	1.0000
sh_in_historic_building	0.000	0.000	0.122	1.000
sh_pm_unionpay	0.00000	0.00000	0.09284	1.00000
sh_in_indoor_pool	0.00000	0.00000	0.09019	1.00000
sh_in_sauna	0.0000	0.0000	0.2546	1.0000
sh_in_play_room	0.00000	0.00000	0.07162	1.00000
sh_in_gym_fitness_room	0.0000	0.0000	0.1273	1.0000
sh_in_outside_swimming_pool	0.00000	0.00000	0.06101	1.00000
sh_in_grill_restaurant	0.00000	0.00000	0.07692	1.00000
sh_in_partly_accessible_rooms	0.00000	0.00000	0.07427	1.00000
sh_in_lift_partly_accessible	0.00000	0.00000	0.08223	1.00000
sh_in_wa_elevator	0.00000	0.00000	0.08488	1.00000
sh_in_ac_breakfast_area	0.00000	0.00000	0.07958	1.00000
sh_sp_hiking_hotel	0.0000	0.0000	0.2095	1.0000
sh_lo_mountain_biking	0.00000	0.00000	0.06101	1.00000
sh_in_partly_wa_toilet	0.00000	0.00000	0.06366	1.00000
sh_in_charging_station_electri	0.00000	0.00000	0.05305	1.00000
sh_in_public_areas_accessible	0.00000	0.00000	0.07958	1.00000
sh_sp_bike_hotel	0.0000	0.0000	0.0504	1.0000

sh_check_in_specified	0.0000	1.0000	0.5517	1.0000	
sh_24_hours_check_in	0.00000	0.00000	0.04509	1.00000	
sh_max_meeting_room_size	0.00	0.00	27.87	500.00	
sh_max_banquet_room_size	0.00	0.00	37.63	500.00	
sh_managers_available	0.0000	1.0000	0.7029	1.0000	
sh_nb_managers	1.000	1.000	1.272	3.000	112
sh_manager_couple	0.00000	0.00000	0.08753	1.00000	
ta_fives	0.00	33.00	87.82	1576.00	111
ta_fours	0.00	28.50	53.76	587.00	111
ta_threes	0.00	10.00	17.67	129.00	111
ta_twos	0.000	3.000	5.929	45.000	111
ta_ones	0.000	2.000	4.462	49.000	111
ta_ratingvalue_exact	1.647	4.176	4.073	5.000	111
rooms	4.00	28.00	35.13	169.00	57
stars	0.000	3.000	2.204	5.000	
reviewcount	1.0	337.0	523.8	3678.0	
ratingvalue	1.046	3.802	3.713	5.000	
price	64.0	199.8	220.3	739.5	141
hotel	0.0000	1.0000	0.6525	1.0000	
other	0.000	0.000	0.122	1.000	
pension	0.0000	0.0000	0.2255	1.0000	
revenue	200000	1400000	1960309	8300000	
ed_hotels	4.50	25.17	34.25	121.67	
ed_rooms	142.6	835.7	1547.6	8013.0	
ed_arrivals	13008	92251	262629	1668112	
ed_room_stays	16914	142775	341776	2050531	
ed_room_occupancy	30.67	55.97	56.21	72.31	
ta_variance	0.0000	0.8000	0.9126	4.2000	96
ta_variance_change_1y	-2.2500	-0.0200	-0.0056	2.0000	118
ta_variance_change_2y	-1.93000	-0.03000	-0.01648	1.48000	144
ed_average_rooms	16.32	32.97	39.92	179.27	

Table A.3: Summary of the revenue prediction table

A.1.3 A Summary of the Growth Prediction Table

attribute	Min	Med	Mea	Max	NA
class	-1.0000	1.0000	0.2522	1.0000	
bk_hotel_wifi	5.000	8.800	8.551	10.000	16
bk_hotel_comfort	6.900	8.200	8.227	9.600	14
bk_hotel_services	6.500	8.100	8.112	9.400	14
bk_hotel_clean	7.2	9.0	9.0	10.0	14
bk_reviewcount	30	227	422	1927	14
bk_hotel_location	7.300	8.900	8.791	9.700	14
bk_ratingvalue	2.200	3.500	3.482	4.500	14
bk_hotel_value	6.700	7.800	7.844	9.300	14
bk_hotel_staff	7.200	8.900	8.934	9.800	14
ta_ratingvalue	3.000	4.000	4.054	5.000	4
ta_reviewcount	2.0	67.0	195.4	2247.0	4

ta_lower_price	79.0	147.0	157.1	300.0	33
ta_higher_price	119.0	247.0	275.4	647.0	33
ta_local_ranking_percentile	0.020	0.500	0.477	1.000	
x	46.02	46.73	46.81	47.72	
y	6.145	8.186	8.175	10.364	
go_ratingvalue	3.000	4.200	4.194	5.000	
go_reviewcount	0.00	42.00	56.77	357.00	
sh_cl_cl_hs	0.0000	1.0000	0.5478	1.0000	
sh_in_close_to_public_transpor	0.0000	0.0000	0.3304	1.0000	
sh_lo_hiking_trails	0.0000	0.0000	0.4261	1.0000	
sh_pm_american_express	0.0000	0.0000	0.4348	1.0000	
sh_in_non_smoking_hotel	0.0000	0.0000	0.3913	1.0000	
sh_pm_postfinance_card	0.0000	0.0000	0.4783	1.0000	
sh_sp_excellent_cuisine	0.0000	0.0000	0.1043	1.0000	
sh_in_wifi_free_of_charges	0.0000	0.0000	0.4435	1.0000	
sh_in_internet_free_of_charges	0.0000	0.0000	0.2261	1.0000	
sh_in_tv_in_all_rooms	0.0000	1.0000	0.5652	1.0000	
sh_in_vegetarian_food	0.0	0.0	0.2	1.0	
sh_pm_visa	0.0000	1.0000	0.5304	1.0000	
sh_in_non_smoking_rooms	0.000	0.000	0.487	1.000	
sh_in_pets_welcome	0.0000	0.0000	0.2957	1.0000	
sh_in_public_car_park	0.0000	0.0000	0.2522	1.0000	
sh_pm_maestro	0.0000	1.0000	0.5217	1.0000	
sh_in_public_restaurant	0.0000	0.0000	0.4348	1.0000	
sh_in_re_with_a_terrace	0.0000	0.0000	0.4087	1.0000	
sh_pm_reka	0.0000	0.0000	0.2957	1.0000	
sh_in_conference_room	0.0000	0.0000	0.3913	1.0000	
sh_in_particularly_quiet_locat	0.0000	0.0000	0.1652	1.0000	
sh_pm_mastercard	0.0000	1.0000	0.5304	1.0000	
sh_in_light_diet_food	0.00000	0.00000	0.03478	1.00000	
sh_in_hotel_own_garage	0.0000	0.0000	0.2174	1.0000	
sh_pm_diners_club	0.0000	0.0000	0.3565	1.0000	
sh_in_radio_in_all_rooms	0.0000	0.0000	0.4261	1.0000	
sh_in_elevator	0.0000	0.0000	0.4957	1.0000	
sh_in_telephone_in_all_rooms	0.0000	0.0000	0.4522	1.0000	
sh_in_panoramic_restaurants	0.00000	0.00000	0.08696	1.00000	
sh_in_central_location	0.0000	0.0000	0.4957	1.0000	
sh_sp_bike_hotel_2011	0.00000	0.00000	0.08696	1.00000	
sh_in_particularly_quiet_rooms	0.0000	0.0000	0.2174	1.0000	
sh_in_rooms_with_wc	0.0000	0.0000	0.4261	1.0000	
sh_lo_fishing	0.0000	0.0000	0.1739	1.0000	
sh_in_childrens_playground	0.00000	0.00000	0.05217	1.00000	
sh_lo_cross_country_skiing	0.0000	0.0000	0.2435	1.0000	
sh_lo_golf	0.0000	0.0000	0.1652	1.0000	
sh_lo_skating	0.0000	0.0000	0.3826	1.0000	
sh_lo_cog_railway_funicular	0.0000	0.0000	0.2957	1.0000	
sh_in_hotel_own_car_park	0.0000	0.0000	0.4261	1.0000	
sh_lo_riding	0.0000	0.0000	0.2957	1.0000	
sh_lo_windsurfing	0.0000	0.0000	0.1304	1.0000	

sh_lo_keep_fit_circuit	0.000	0.000	0.313	1.000	
sh_lo_motor_boats	0.0000	0.0000	0.2609	1.0000	
sh_lo_ski_lifts	0.0000	0.0000	0.2522	1.0000	
sh_lo_ski_school	0.0000	0.0000	0.2261	1.0000	
sh_lo_yacht_and_boat_hire	0.0000	0.0000	0.2522	1.0000	
sh_lo_casino	0.0000	0.0000	0.1739	1.0000	
sh_in_garden_park_area	0.0000	0.0000	0.2609	1.0000	
sh_lo_water_skiing	0.0000	0.0000	0.1913	1.0000	
sh_in_wholefood_cuisine	0.00000	0.00000	0.03478	1.00000	
sh_lo_mountaineering_school	0.0000	0.0000	0.1913	1.0000	
sh_lo_canoeing	0.0000	0.0000	0.2087	1.0000	
sh_pm_jcb	0.0000	0.0000	0.2435	1.0000	
sh_lo_curling	0.0000	0.0000	0.2348	1.0000	
sh_lo_car_free_loity	0.00000	0.00000	0.04348	1.00000	
sh_in_bed_usable_for_hoists	0.00000	0.00000	0.07826	1.00000	
sh_in_wa_washbasin	0.00000	0.00000	0.09565	1.00000	
sh_in_wa_hotel_shuttle	0.00000	0.00000	0.06957	1.00000	
sh_in_guests_with_visual_impai	0.00000	0.00000	0.05217	1.00000	
sh_in_bed_height_45_50cm	0.00000	0.00000	0.06087	1.00000	
sh_in_public_areas_partly_ac	0.00000	0.00000	0.09565	1.00000	
sh_sp_business_hotel	0.0000	0.0000	0.1043	1.0000	
sh_in_snack_restaurant	0.0000	0.0000	0.1826	1.0000	
sh_in_minibar_in_all_rooms	0.0000	0.0000	0.2783	1.0000	
sh_lo_airport	0.00000	0.00000	0.07826	1.00000	
sh_sp_seminar_hotel	0.0000	0.0000	0.1391	1.0000	
sh_in_smoking_lounge	0.00000	0.00000	0.05217	1.00000	
sh_in_bar	0.0000	0.0000	0.2435	1.0000	
sh_in_historic_building	0.0000	0.0000	0.1304	1.0000	
sh_pm_unionpay	0.0000	0.0000	0.1043	1.0000	
sh_in_indoor_pool	0.00000	0.00000	0.07826	1.00000	
sh_in_sauna	0.0	0.0	0.2	1.0	
sh_in_play_room	0.00000	0.00000	0.04348	1.00000	
sh_in_gym_fitness_room	0.0000	0.0000	0.1391	1.0000	
sh_in_outside_swimming_pool	0.00000	0.00000	0.06087	1.00000	
sh_in_grill_restaurant	0.000	0.000	0.113	1.000	
sh_in_partly_accessible_rooms	0.00000	0.00000	0.09565	1.00000	
sh_in_lift_partly_accessible	0.00000	0.00000	0.09565	1.00000	
sh_in_wa_elevator	0.0000	0.0000	0.0885	1.0000	2
sh_in_ac_breakfast_area	0.0000	0.0000	0.0708	1.0000	2
sh_sp_hiking_hotel	0.0000	0.0000	0.2523	1.0000	8
sh_lo_mountain_biking	0.00000	0.00000	0.05319	1.00000	21
sh_in_partly_wa_toilet	0.00000	0.00000	0.07778	1.00000	25
sh_in_charging_station_electri	0.00000	0.00000	0.04598	1.00000	28
sh_in_public_areas_accessible	0.00000	0.00000	0.09195	1.00000	28
sh_sp_bike_hotel	0.00000	0.00000	0.05747	1.00000	28
sh_check_in_specified	0.0000	1.0000	0.5181	1.0000	32
sh_24_hours_check_in	0.00000	0.00000	0.02469	1.00000	34
sh_max_meeting_room_size	0.0	0.0	38.3	350.0	34
sh_max_banquet_room_size	0.00	0.00	51.67	350.00	36

sh_managers_available	0.0000	1.0000	0.6456	1.0000	36
sh_nb_managers	1.000	1.000	1.353	2.000	64
sh_manager_couple	0.0000	0.0000	0.1013	1.0000	36
ta_fives	2.0	36.0	159.3	1576.0	58
ta_fours	3.0	38.0	88.6	587.0	58
ta_threes	0.00	18.00	26.81	93.00	58
ta_twos	0.000	4.000	8.491	45.000	60
ta_ones	0.000	4.000	6.945	46.000	60
ta_ratingvalue_exact	2.953	4.031	4.051	4.844	60
rooms	6.00	37.00	43.81	169.00	52
stars	0.000	3.000	2.507	5.000	40
reviewcount	9.0	359.0	700.8	3435.0	40
ratingvalue	2.369	3.742	3.733	4.771	43
price	123.5	185.0	205.6	423.5	61
hotel	0.0000	1.0000	0.7639	1.0000	43
other	0.00000	0.00000	0.05556	1.00000	43
pension	0.0000	0.0000	0.1806	1.0000	43
change_ratingvalue_yoy	-1.00000	-0.01000	-0.01076	0.45000	49
change_ratingvalue_2yoy	-1.00000	0.00000	0.02356	0.95000	56
change_reviewcount_yoy	0.00	15.00	33.23	289.00	49
change_reviewcount_2yoy	2.0	30.0	60.8	527.0	56
change_rating_variance_yoy	-0.93000	0.00000	0.03369	2.00000	50
change_rating_variance_2yoy	-3.43000	0.00000	-0.01345	1.73000	57
change_rooms	-67.660	2.490	12.832	162.180	43
change_hotels	-2.09000	-0.04000	0.07069	3.17000	43
change_arrivals	-15572.00	5146.00	7977.85	44421.00	43
change_stays	-27844	5293	11078	86217	43
change_room_stays	-15116	2313	5812	57264	43
change_room_occupancy	-6.4700	0.6300	0.3549	4.2500	43
change_bed_occupandcy_	-9.1200	0.8050	0.2188	4.1100	43

Table A.4: Summary of the growth prediction table

A.1.4 Additional Material

The scripts used for webscraping, the Python program that handled data processing, and R scripts which created the predictions can all be found on github:

https://github.com/heinzermch

Declaration of Originality

The signed declaration of originality is a component of every semester paper, Bachelor's thesis, Master's thesis and any other degree paper undertaken during the course of studies, including the respective electronic versions.

Lecturers may also require a declaration of originality for other written papers compiled for their courses.

I hereby confirm that I am the sole author of the written work here enclosed and that I have compiled it in my own words. Parts excepted are corrections of form and content by the supervisor .

Title of work (in block letters):

Predicting the Success of Hotels with Online Ratings

Authored by (in block letters):
For papers written by groups the names of all authors are required.

Name(s):	**First name(s):**
Heinzer	Michael

With my signature I confirm that

- I have committed none of the forms of plagiarism described in the Citation etiquette information sheet.
- I have documented all methods, data and processes truthfully.
- I have not manipulated any data.
- I have mentioned all persons who were significant facilitators of the work .
- I am aware that the work may be screened electronically for plagiarism.
- I have understood and followed the guidelines in the document *Scientific Works in Mathematics*.

Place, date:	**Signature(s):**
Mountain View, CA, 26.03.2018	

For papers written by groups the names of all authors are required. Their signatures collectively guarantee the entire content of the written paper.