MUST BE THE WEATHER:

PREDICTING BOOKING DEMAND WITH THE WEATHER FORECAST

LEONARD SUGG

THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCE IN DATA SCIENCE & SOCIETY
AT THE SCHOOL OF HUMANITIES AND DIGITAL SCIENCES
OF TILBURG UNIVERSITY

STUDENT NUMBER

2039830

COMMITTEE

dr. Eric Postma
Marianne Kramer

LOCATION

Tilburg University
School of Humanities and Digital Sciences
Department of Cognitive Science &
Artificial Intelligence
Tilburg, The Netherlands

DATE

January 10th, 2025

WORD COUNT

8702

ACKNOWLEDGMENTS

I would like to thank Center Parcs and the revenue management department for supporting me, especially Tim van Fenema!
Furthermore, I would like to thank my supervisor Dr. Eric Postma for his support.
Last but not least, I would like to thank my friends and girlfriend for their support.

MUST BE THE WEATHER:
PREDICTING BOOKING DEMAND WITH THE WEATHER FORECAST

LEONARD SUGG

Abstract

This thesis investigates the role of weather forecasts in predicting occupancy levels at a Center Parcs location in Western Europe, addressing both the scientific challenge of enhancing predictive modeling and the practical need for accurate occupancy forecasting in the family resort industry. Previous research has emphasized the importance of including weather as a predictor for booking demand; however, limited studies have incorporated it so far, and none as extensively as this study. Three machine learning algorithms were evaluated for four different forecast horizons: Extreme Gradient Boosting (XGB) Regressor, Support Vector Regressor (SVR), and ridge regression. The analysis utilized four years of historical data, incorporating both capacity and weather forecast features. Various preprocessing and feature engineering techniques were explored, including robust scaling and principal component analysis (PCA) applied to weather and capacity features either separately or jointly. A comprehensive error analysis and model comparison showed that scaling improved the performance of XGBR and ridge regression models but reduced performance for SVR. Additionally, all implementations of PCA led to poorer model performance. Finally, weather-related features appeared to have minimal impact on occupancy prediction accuracy, highlighting the potential value of alternative predictors. These findings contribute to enhancing operational planning and resource allocation while emphasizing the need for expanded forecasting research across more diverse scenarios.

1 DATA SOURCE, ETHICS, CODE, AND TECHNOLOGY STATEMENT

The owners of the data are Center Parcs and Open Weather, which both allow sharing, modifying, and using the data for this thesis project. Center Parcs allows sharing the aggregated data, where no individual customer information is present. Center Parcs purchased the weather data from the Open Weather website, which sells under the Creative Commons Attribution-ShareAlike 4.0 International license and the Open Data Commons Open Database License (OpenWeather, n.d.). The code, final data, and Tableau sheet for this thesis are in the GitHub repository (Sugg, 2024). Tableau Prep version 2024.1.0 by Salesforce, and Visual Studio Code 1.95.3 by Microsoft with Python 3.12 were used 13. The Python libraries used are listed in this table 12. ChatGPT from OpenAI was used to correct spelling, for debugging code, and for paraphrasing. The provided Overleaf template was used for typesetting.

2 INTRODUCTION

Efficient planning of logistics, staffing, and discounts is crucial for Center Parcs, as they have 29 park locations (Center Parcs, 2024; Satu et al., 2020). It is important to anticipate bookings and cancellations to offer optimal service to the customer by planning sufficient staff and to make Center Parcs more profitable by optimally using the capacity (Gunaseelan, Alalmai, & Arun, 2020). Discounts are offered strategically to undecided customers to sway their behaviors toward booking a stay (Center Parcs, 2024). Center Parcs is constantly trying to improve its demand forecast to operate optimally (Center Parcs, 2024). There are several impactful factors already monitored in the revenue management systems, such as competition impact, price fluctuations, holiday pressure from main markets, and promotional strategy (Center Parcs, 2024). As certain activities are arguably more enjoyable with certain weather conditions, there is a need to research how weather conditions can affect booking demands (Bausch et al., 2021). For example, a Center Parcs location where most activities are hosted indoors might receive more bookings and guests during rainy periods. On the contrary, one would expect a park to receive more cancellations during rainy periods if the activities are mostly outdoor-focused. The weather generally seems to be a significant factor for tourism arrivals in alpine regions and less so for the felt satisfaction of staying in urban destinations (Bausch et al., 2021; McKercher et al., 2014). This discrepancy might be due to the different ways weather conditions affect the activities or settings offered. The influence of weather on the duration and satisfaction of a stay varies depending on the types of activities available. This tendency is evident in hotel pricing strategies, specifically hedonic pricing, where hospitality prices fluctuate based on expected weather conditions (Figini et al., 2019). For instance, hotels in sunny regions often increase their prices in anticipation of good weather (Figini et al., 2019). As the weather has hardly been addressed in the booking demand forecasting research, it is not clear what optimal training approaches are for models using weather features (Henriques & Nobre Pereira, 2024; Huang & Zheng, 2022).

Experimenting with the optimal use of various weather features for demand forecasting is a key focus of this thesis. Since most weather measurements tend to be inter-correlated, applying linear dimensionality reduction is anticipated to enhance the forecasting models (Farfan, Castillo, & Curilef, 2019). Additionally, scaling the features is expected to improve model performance by standardizing the different measurement units of weather data. Consequently, using principal component analysis for dimensionality reduction and scaling the features could provide a solid foundation for incorporating weather data into demand forecasting. However, principal components that combine weather and capacity features make model interpretation more difficult (Oshternian et al., 2024). Therefore, performing PCA separately on weather and capacity features to generate principal components specific to either capacity or weather simplifies interpretation (Oshternian et al., 2024). Finally, as shorter forecasting horizons typically yield more accurate predictions than longer ones, investigating whether this trend holds true can serve as a simple sanity check for the models, as observed in previous research (Henriques & Nobre Pereira, 2024; Huang & Zheng, 2022).

In conclusion, this thesis explores the influence of specific weather variables, including precipitation, temperature, and snow depth, on booking demand, with the goal of enhancing the accuracy of demand forecasting. Booking demand is commonly defined as the number of reservations or occupancy rate for a set amount of time (Huang & Zheng, 2022). To improve predictions, dimensionality reduction and scaling are applied to the features. To ease interpretation after applying PCA, separate principal components for weather and capacity features are generated (Oshternian et al., 2024). Accurate demand forecasts can help the business with revenue management and planning, which enables Center Parcs to give customers a better experience (Center Parcs, 2024). Previous research on demand forecasting in the hospitality industry focused on factors like pricing and historical booking data for mostly hotels (Henriques & Nobre Pereira, 2024). However, little research investigated the combined influence of weather conditions and booking data for recreational parks offering both indoor and outdoor activities (Henriques & Nobre Pereira, 2024; Huang & Zheng, 2022). Thus, the following research questions are addressed:

1. *How well can the booking demand be predicted for different forecast horizons (1, 5, 7, and 14 days before the visit)?*

1.1 *Does (relative) forecasting performance improve with shorter forecast horizons?*

1.2 *To what extent do the weather variables contribute to the forecast horizon-dependent performance?*

2. *What model and training procedure provide the best demand forecasting?*

2.1 *How does robust scaling affect the models?*

2.2 *Does performing a separate dimensionality reduction for weather and capacity variables hinder performance?*

3 RELATED WORK

Recent meta-reviews of hotel demand forecasts concluded that artifical intelligence-based models (AI) perform as well as, and in some cases better than, traditional statistical methods (Henriques & Nobre Pereira, 2024). As purely statistical models are commonly used for forecasting demand, they are often challenged by large datasets due to complex variable relationships (Henriques & Nobre Pereira, 2024; Huang & Zheng, 2022). As a result, research is shifting towards AI-based models or a mix of the two (Henriques & Nobre Pereira, 2024; Huang & Zheng, 2022). Notably, research is gradually incorporating reservation data alongside the commonly used historical data, making the combination of several data sources more common (Huang & Zheng, 2022). Furthermore, several reviewed studies even recommend using weather data, historical data, and their used methods in future research (Henriques & Nobre Pereira, 2024). So far, booking forecast models have primarily utilized features such as social media keywords, room cancellations, prices, online reviews, room reservations, and room occupancy (Henriques & Nobre Pereira, 2024). Thus, table 1 showcases used models and their best performances, which illustrate the state of the art for booking demand forecasting and are extracted from the original studies and meta reviews (Henriques & Nobre Pereira, 2024; Huang & Zheng, 2022).

This gives a quick overview; however, differences in data collection and analysis methods prevent a simple comparison of models (Henriques & Nobre Pereira, 2024). For example, the exclusion of environmental and economic factors adds complexity to making comparisons across different locations (Henriques & Nobre Pereira, 2024). Additionally, various forecasting methods are employed, such as point, quantile, density, and ensemble forecasts. Point forecasts predict a specific value at a future time; quantile forecasts provide predictions based on set quantile thresholds; density forecasts estimate a complete probability distribution for each time point; and ensemble forecasts generate predictions by combining previous predictions (Niemann & Schienle, 2023). Generally, point forecasts are common in booking demand research, computationally cheap, and are thus chosen over the other methods (Henriques & Nobre Pereira, 2024; Huang &

Zheng, 2022). Furthermore, in previous research, occupancy typically refers to rooms, whereas in this study, it refers to houses (Henriques & Nobre Pereira, 2024; Huang & Zheng, 2022). This distinction is important because prior studies have indicated that factors such as group size and travel purpose can significantly influence travel behavior (Ito, Kanemitsu, Kimura, & Omori, 2024). For a family resort like Center Parcs, this could result in fewer last-minute cancellations and less spontaneous bookings compared to booking an individual room (Ito, Kanemitsu, Kimura, & Omori, 2024). Another difference between studies is the inclusion of COVID-19 data, which made booking demand forecasting more challenging (Henriques & Nobre Pereira, 2024; Huang & Zheng, 2022). Currently, it seems that this was an exceptional and atypical time when weather data likely had minimal or no impact on booking behaviors (Golets, Farias, Pilati, & Costa, 2021). Thus, studies analyzing booking data from the COVID-19 pandemic are solely mentioned in the table 1, as travel restrictions and government regulations during this period were likely the dominant influencing factors.

A more detailed discussion of studies using machine learning and statistical methods follows to facilitate comparison and showcase state of the art: A study by Ampountolas and Legg (2021) used social media keywords to forecast occupancy for a major urban hotel chain in the USA, using ARIMA and XGBoost. As is common in forecasting research, a naive baseline model was created by using observations from the previous year to predict the current test time frame. This study found that XGBoost outperformed ARIMA by 70% for a 1-day forecast ($MAPE$ = 5%) and by 30% for a 14-day forecast window ($MAPE$ = 8.2%), using the MAPE metric (Ampountolas & Legg, 2021). The number of rooms occupied before the forecasted day was the strongest predictor (Ampountolas & Legg, 2021). Additionally, overfitting issues of XGBoost were mentioned (Ampountolas & Legg, 2021). Further limitations were that results might only apply to major hotel chains with popular social media (Ampountolas & Legg, 2021). Another study by Pereira and Cerqueira (2021) compared 22 forecasting models with past occupancy data from a South-European hotel. Machine learning models using dynamic ensemble methods outperformed statistical methods by 54% for a 1-day forecast ($RMSE$ = 9.3) and by 45% for a 14-day forecast ($RMSE$ = 13.8), based on the RMSE metric (Pereira & Cerqueira, 2021). The main limitation mentioned is that Covid-19 data was not included (Pereira & Cerqueira, 2021). Phumchusri and Ungtrakul (2019) used monsoon season, oil prices, holidays, exchange rates, and historical booking data. These features were used to predict the daily booking demand for a hotel in Thailand (Phumchusri & Ungtrakul, 2019). Artificial neural networks and support vector machines significantly outperformed statistical methods on the MAPE metric ($MAPE$ = 8.955%) (Phumchusri

& Ungtrakul, 2019). The main limitation mentioned was that the average room occupancy in the region was unavailable (Phumchusri & Ungtrakul, 2019).

Despite its use in pricing and as a tourist attraction, weather has recently not been used to predict booking demand (Bausch et al., 2021; McKercher et al., 2014). In past research of Pan and Yang (2016), weather forecast data and internet searches were used to predict weekly occupancy rates for the Charleston area (USA). From the several compared regression models, Autoregressive Moving Average with Exogenous Variables (ARMAX) performed best with a MAPE of 8.73% (Pan & Yang, 2016). Weather variables were the weekly maximum and minimum for temperature, average humidity, and counts of snowy and rainy days (Pan & Yang, 2016). The weather features did not significantly improve predictions in this study, as during feature selection, only snowy days were identified as statistically significant. This led to the exclusion of all other weather-related features (Pan & Yang, 2016). Furthermore, it only snowed 28 days in 10 years of data, which is not sufficient to assess the weather impact (Pan & Yang, 2016).

As the data includes many features with varying scales, scaling and dimensionality reduction are commonly used to improve machine learning performance (Ahsan, Mahmud, Saha, Gupta, & Siddique, 2021; Henriques & Nobre Pereira, 2024; Huang & Zheng, 2022). These techniques generally enhance the performance of SVR and ridge regression models, which are less robust to scaling issues and high dimensionality compared to tree-based algorithms like XGBR (Ahsan, Mahmud, Saha, Gupta, & Siddique, 2021). Among various scaling methods, scaling based on interquartile ranges, as employed in robust scaling, is generally more resilient to outliers than other common techniques, such as standard scaling, as demonstrated in previous research.

For dimensionality reduction a principal component analysis is well suited for weather data, as most weather measurements are intercorrelated (Farfan, Castillo, & Curilef, 2019). Capacity has been found to be suitable for PCA as well in previous research (Contessi, Viverit, Nobre Pereira, & Heo, 2024). Applying PCA to a dataset with two different feature categories, such as weather and capacity, typically reduces the dimensionality and computational costs, but it can also make interpretation more challenging (Oshternian et al., 2024). On the other hand, applying PCA separately to weather and capacity data could help attribute the effects of weather more clearly, though this approach might increase dimensionality and computational demands. However, this added complexity is often negligible when working with rich datasets, as previous research suggests that the benefits of improved interpretability outweigh the costs in many cases

(Oshternian et al., 2024). Finally, the number of principal components is often determined using scree plots and the "elbow rule" (Wicklin, 2017). This heuristic involves identifying the point where the variance explained by each component shows a sharp decline, followed by a plateau (Wicklin, 2017). Alternatively, a more sophisticated approach involves permutation testing, where observations are sampled and modified (Toledo Jr., 2022; Wicklin, 2017). However, the "elbow rule" is computationally more efficient and often yields similar results, making it the preferred method for this thesis (Toledo Jr., 2022; Wicklin, 2017).

In conclusion, previous research shows that machine learning methods outperform statistical methods for booking demand forecasting and that weather has not been thoroughly researched as a predictor. Furthermore, this study employs point forecasting rather than other forecasting methods, as it is commonly used in demand research and is computationally cheaper (Henriques & Nobre Pereira, 2024; Huang & Zheng, 2022; Niemann & Schienle, 2023). Since the COVID-19 pandemic significantly impacted lives and imposed travel restrictions, it is reasonable to exclude such data to ensure better comparisons and model performance (Golets, Farias, Pilati, & Costa, 2021). Finally, robust scaling and experimenting with the implementation of PCA are common and expected to improve performance (Ahsan, Mahmud, Saha, Gupta, & Siddique, 2021; Henriques & Nobre Pereira, 2024; Huang & Zheng, 2022). Additionally, the number of principal components will be selected with the "elbow rule" as it is computationally efficient and common (Wicklin, 2017). Thus, the main difference between this thesis and other studies is the weather forecast data, occupancy refering to houses, and that Center Parcs is a family resort instead of a hotel (Henriques & Nobre Pereira, 2024; Huang & Zheng, 2022). Therefore, the majority of performance differences will probably be due to the weather forecast data or the different type of accommodation (Henriques & Nobre Pereira, 2024; Huang & Zheng, 2022).

4 METHOD

4.1 *Data preprocessing and Data cleaning*

The final dataset for this thesis is a merged dataset based on capacity measurements provided by Center Parcs and a historic weather forecast dataset from OpenWeather for a specific Center Parcs location. The Center Parcs location is in Western Europe and offers winter and summer activities. The OpenWeather dataset is a CSV file with 1,048,576 rows and 24 columns (OpenWeather, 2024). It was purchased and downloaded from the OpenWeather marketplace by a Center Parcs employee on the 02.09.2024

Table 1: *Forecasting Model Performance in Hotel Demand Studies*

Author (Year)	Metrics & Model	Features	Best performance
Viverit et al. (2023)	Metrics: MSE, RMSE, MAE, MAPE, sMAPE, MdAPE. Model: Cluster-based Pickup Method.	Historical booking curves grouped by patterns using a machine learning algorithm including Covid-19 data.	7, 14, 30, 50 days forecast horizons. Cluster-based Pickup showed: - Best performance during for 7 (RMSE = 2.45, MAE = 1.8, MAPE = 14.52%) & 14 (RMSE = 3.02, MAE = 2.26, MAPE = 21.53%) days
Wu et al. (2021)	Metrics: MAPE, RMSE, Theil's U Models: MIDAS-Almon, U-MIDAS	Monthly hotel occupancy rates during Covid-19, daily visitor arrivals, daily Baidu search query index	30 days forecast horizon (MAPE = 28.92%, RMSE = 8.3)
Wu et al. (2022)	Metrics: MAPE, RMSE Model: ARIMAX	Bullish Sentiment Index (BSI), Average Sentiment Index (ASI), Variance Sentiment Index (VSI) from daily online reviews; weekday dummy variables	1-7 days horizons, 1 (RMSE = 2.926, MAPE = 14.516%), 5 (RMSE = 2.856, MAPE = 16.442%), 7 (RMSE = 5.298, MAPE = 16.224%)
Kaya et al. (2022)	Metrics: MAE, MAPE Models: Random Forest Regression (RFR), Extra Trees Regression (ETR), Gradient Boosting Regression (GBR), XGBR, Gated Recurrent Unit (GRU), Long Short-Term Memory (LSTM), Attention-LSTM	Raw demand data, full hotel features, top-10 hotel features, 10D embedding	7 days forecast horizon RFR: MAE = 6.43, MAPE = 36.61; ETR: MAE = 6.39, MAPE = 6.64; GBR: MAE = 6.14, MAPE = 33.37; XGBR: MAE = 6.56, MAPE = 37.13; GRU: MAE = 5.57, MAPE = 26.9; LSTM: MAE = 5.23, MAPE = 26.72; Attention-LSTM: MAE = 4.99, MAPE = 26.53

(OpenWeather, 2024). Moreover, the capacity dataset from Center Parcs is a CSV file with 326,864 rows and 33 columns. It was extracted from the Center Parcs data storage for the specific Center Parcs location on 02.09.2024. As these datasets are merged in Tableau, the merged dataset is a CSV file with 311,413 rows and 67 columns. The increase of columns is due to aggregation measurements like minimum, maximum, and average temperature. Additionally, some duplicate columns were created to validate aggregations,which were removed later. From the merged dataset, all rows with a lead time greater than 15 days and rows corresponding to certain dates (e.g., closures or COVID-19 disruptions) were removed. Since the largest forecast horizon tested is 14 days, the final dataset consists of 22,272 rows and 51 columns. The reduction in the number of columns is due to the removal of duplicate capacity columns, which was carried out to validate aggregation changes, as well as the removal of columns related to Unix time, as ISO time was sufficient. Additionally, columns containing latitude and longitude were removed to maintain the confidentiality of the Center Parcs location. The final dataset includes current capacity measurements and weather forecasts for both the booking day and the first day of the stay. This approach simulates a realistic scenario where a customer checks the destination's weather forecast for the first day of their stay before making a booking (Diego R.-Toubes, Araújo-Vila, & Fraiz-Brea, 2020). The creation of the final dataset is described in more detail below.

First, all observations from the Center Parcs location are aggregated in Tableau Prep by a Center Parcs employee from the revenue management department. This aggregation is performed for each visit date and booking day, transforming the dataset from individual booking-level data to booking-day-level data. As a result, each observation represents summary information for an entire booking day, including metrics such as the total number of villas booked, the number of customers with travel insurance, and other relevant variables (El-Mhouti, 2023). This makes the problem simpler, less confidential, and more in line with previous research because of the loss of individual information (Henriques & Nobre Pereira, 2024; Huang & Zheng, 2022). Then for the weather dataset, aggregated variables were also created because multiple forecasts were made on a single forecasting day for the same future time. To summarize this information, the weather measurements of a booking day were consolidated into minimum, maximum, and average values (El-Mhouti, 2023). Finally, the features used are described in the appendix 11.

As the dataset included bookings made far in advance but lacked corresponding weather forecast data for those periods, only bookings made up to a maximum of 16 days before the visit were considered. Furthermore, previous research on demand forecasting focused on horizons ranging from 1 to 14 days. Therefore, despite weather data being available for up to 16 days, only data with a maximum horizon of 14 days was included. In this study, the selected horizons of 1, 5, 7, and 14 days align with the range of previous research, enabling meaningful comparisons with other short-term forecasting studies (Ampountolas & Legg, 2021; Pan & Yang, 2016; Pereira & Cerqueira, 2021). Additionally, as mentioned earlier, data from the peak of the COVID-19 pandemic (01.10.2019–01.10.2021) was excluded, aligning with the Center Parcs business year, which starts in October (Golets, Farias, Pilati, & Costa, 2021). Other excluded dates are between 19.11.2018 and 06.12.2018, due to a temporary closure of that Center Parcs location.

The weather forecast data were aligned so that the forecast available at the time of booking corresponds to the forecast for the visit date. For the capacity data, time lags were applied to ensure that the model only uses capacity measurements from days prior to the booking date. This setup reflects the realistic constraints of forecasting: for a forecast horizon of 14 days, the model uses weather and capacity data available up to the booking date to predict the booking demand for the subsequent 14 days. Importantly, the only variable without a time lag is the dependent variable, booking demand, which corresponds directly to the visit date. For some dates there are missing values for the price per person feature. Therefore, the missing values are filled with median values calculated from the training set to avoid data leakage (Brownlee, 2020). Median imputation is chosen over mean imputation as this is showed increased robustness in previous research (Jäger et al., 2021).

The training data ranges from 01.10.2018 until 01.10.2023, while the testing data ranges from 01.10.2023 until 01.09.2024. Thus, the training data contains approximately three years, and the testing data contains nearly a full year, which is a commonly used ratio in machine learning (Huang & Zheng, 2022). The validation data is part of the training data and starts from 01.10.2022 until 01.10.2023. The ratio of training, validation, and test data is thus 50:25:25. While this deviates slightly from the conventional 60:20:20 ratio, the full-year validation and test sets support generalizability by tuning and testing the model across a complete seasonal cycle, as done in previous research (Kaya et al., 2022; Maleki et al., 2022).

4.2 *Evaluation metrics*

The models' performances are compared with the Mean Absolute Error (MAE). Furthermore, Root Mean Square Error (RMSE) and Mean Absolute Percentage Error (MAPE) are calculated to enable comparison with previous research in case no MAE was reported. (Acharya, 2021; Makridakis et al., 1998; Hyndman & Athanasopoulos, 2018). MAE, RMSE, and MAPE are all commonly used in the forecasting literature (Henriques & Nobre Pereira, 2024; Huang & Zheng, 2022). Furthermore, as MAE and Mean Standard Error are commonly known metrics and RMSE 1 is simply the square root of MSE (Acharya, 2021), only MAPE is explained in detail. MAPE is used since it is well suited for comparing different models and their interpretability (Makridakis et al., 1998; Hyndman & Athanasopoulos, 2018) and is shown in equation 2. MAPE values start at 0 for perfect predictions and can exceed 1 for large prediction errors (Kim & Kim, 2016). In the equation, n is the number of observations, y_i is the true value, and $\hat{y}_i$ is the prediction for an observation (Makridakis et al., 1998).

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

$$\text{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \times 100 \tag{2}$$

This selection offers a wide range to measure and compare performance on absolute and relative levels with previous research (Acharya, 2021; Henriques & Nobre Pereira, 2024; Huang & Zheng, 2022; Hyndman, 2009; Kim & Kim, 2016). For the error analysis, predictions that deviate by more than two standard deviations (SD) from the true values are plotted (Zhang et al., 2018). A threshold of two standard deviations is selected because it strikes a balance between identifying extreme values and excluding those that are not overly extreme (Zhang et al., 2018). Furthermore, the mean and median residuals are used to assess model bias, with the mean indicating general overestimation or underestimation of predictions, and the median providing a more robust measure of central tendency, especially in the presence of outliers or skewed data (Gelman & Hill, 2007). Finally, checking whether shorter forecast horizons perform better than longer ones is a simple sanity check for the models, as this is typically the case in previous research (Henriques & Nobre Pereira, 2024; Huang & Zheng, 2022).

4.3 *Model selection*

As it is common to use a naive model as a baseline that predicts last year's occupancy rates as the test year, this study will do the same (Ampountolas & Legg, 2021). In previous research, ridge regression, Support Vector Regressor (SVR), and Extreme Gradient Boost Regressor (XGBR) consistently outperformed statistical methods (Henriques & Nobre Pereira, 2024; Huang & Zheng, 2022). Therefore, these machine learning algorithms are used in this study and are explained further. Firstly, ridge regression is a form of linear regression with an added cost function to reduce overfitting (Murel & Kavlakoglu, 2023). Like linear regression, ridge regression predicts by fitting a linear equation and minimizing the residual sum of squares between the true and predicted values. However, in high-dimensional or small datasets, linear regression can lead to overfitting (Murel & Kavlakoglu, 2023). Ridge regression mitigates this issue by adding a regularization term to the cost function, which penalizes large coefficients. This penalty term reduces the likelihood of overfitting by shrinking some coefficients toward zero, although it does not eliminate features (Murel & Kavlakoglu, 2023). The amount of regularization is determined by lambda (λ), and setting λ to zero results in a standard linear regression (Murel & Kavlakoglu, 2023). In the equation of the ridge regression cost function (equation 3), β is the coefficient of the feature value $\mathbf{x}_i^T$ and $\hat{\beta}$ is the estimated coefficient (Hoerl & Kennard, 1970). The hyperparameters are λ and whether an intercept is used (VanderPlas & Scikit-learn developers, n.d.).

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \left(\sum_{i=1}^{n} \left(y_i - \mathbf{x}_i^T \beta \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right) \tag{3}$$

The support vector regressor (SVR) performed well for demand forecasting and is thus tested as well (Henriques & Nobre Pereira, 2024; Huang & Zheng, 2022). Finally, the extreme gradient boost (XGBR) algorithm is an ensemble machine learning model that works with several decision tree models and regularization (Chen & Guestrin, 2016). Decision trees are limited in performance by overfitting and noisy data unless boosting, pruning, or bagging techniques are applied (Kotsiantis, 2013). In XGBR, each newly added decision tree model learns from the mistakes of the previous models and then gets merged with the previous model, which results in an improved model (Chen & Guestrin, 2016). New models are created and merged until specified or errors are minimized, also called boosting (Chen & Guestrin, 2016). Furthermore, the applied regularization incentivizes simple decision tree models, as complex models receive greater penalties (Chen & Guestrin, 2016). Thus, the decision tree depths are generally less deep for XGBR to avoid overfitting (Chen & Guestrin, 2016). Equation 4 illustrates the role of λ in regularization, as K is the number of leave nodes and $\|w_k\|^2$ is the weight of leave node k (Chen & Guestrin, 2016). Furthermore, $f(x_i)$ is defined as the prediction for observation i (Chen & Guestrin, 2016).

$$\hat{f} = \underset{f}{\operatorname{argmin}} \left(\sum_{i=1}^{n} (y_i - f(x_i))^2 + \lambda \sum_{k=1}^{K} \|w_k\|^2 \right) \tag{4}$$

4.4 *Experimental setup*

To answer the research questions of what the best-performing model from the selection is and what impact the weather features have, several experiments are conducted. For each experiment a model is trained in two slightly different ways to make the impact of the changes in training procedure clear. For all experiments, three machine learning algorithms are trained separately for forecast horizons of 1, 5, 7, and 14 days. This means that, for example, four XGBR models are created, one for each forecast horizon. Taking the 7-day forecast horizon as an example, the corresponding XGBR model is trained and tested exclusively on data for the 7-day horizon. Thus, for experiment 2 and 3, three algorithms are trained across four forecast horizons using two approaches, resulting in a total of 24 models per experiment. Since the first experiment focuses solely on one training approach, which is training the models without weather data, only 12 models are trained. To ensure optimal performance for each model, hyperparameter tuning is conducted on the validation set using a 5-fold time-sensitive grid search, following the approach used in previous research (Kaya et al., 2022; Tashman, 2000). The hyperparameters used are mentioned in the flowchart 1.

The first experiment involves training each model without weather features to establish a baseline for assessing the impact of weather features (Wu & Xue, 2024). This feature elimination approach provides a general estimate of the impact of weather features on model performance by comparing models that include weather features in subsequent experiments to those without weather features (Wu & Xue, 2024). Therefore, the following experiments test different model training approaches with the weather and capacity features. Experiment 2 compares models that are trained with robustly scaled or unscaled data. As mentioned earlier, the models are trained separately for each forecast horizon and machine learning algorithm, once with robustly scaled data and once with unscaled data. The robust scaling increases robustness to outliers by using the median and interquartile ranges and is commonly done before applying PCA (Sharma, 2022). The scaled models are expected to perform better since the metrics vary widely, especially SVR, which depends on distance (Ahsan, Mahmud, Saha, Gupta, & Siddique, 2021).

The third experiment tests how different applications of PCA impact model performance for each machine learning algorithm and forecast horizon. PCA can reduce the dimensionality of features by creating linear combinations of features, referred to as principal components, based on maximum variance (Kuhn & Johnson, 2013). The number of components is selected using scree plots and the "elbow rule" for its computational

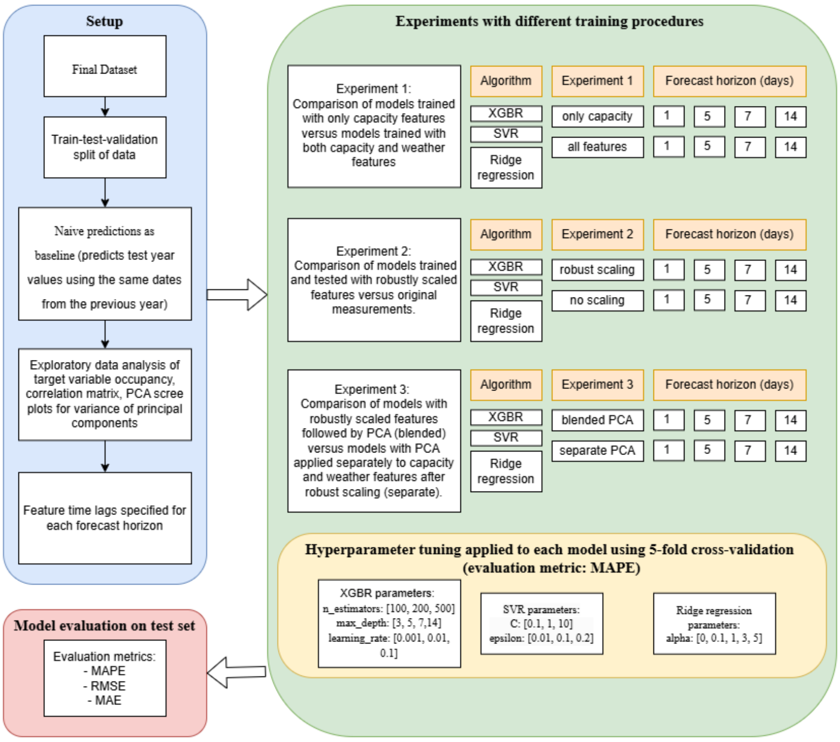

Figure 1: Research methodology flowchart

efficiency and comparable results (Toledo Jr., 2022; Wicklin, 2017). For this experiment the PCA is applied in two different ways before training the models: The blended approach, which combines weather and capacity data to create principal components, and the separated approach, which creates separate principal components for the weather and capacity data. Notably, categorical features are not included into any PCA transformation, as PCA is not suitable for categorical features (Kuhn & Johnson, 2013). An increase in performance is expected after dimensionality reduction for both approaches, mostly for SVR, as the data contains over 50 variables, and SVR struggles exceptionally with high dimensionality (Rastogi, Taterh, & Kumar, 2023).

5 RESULTS

5.1 *Exploratory Data Analysis*

The EDA begins with a correlation matrix of all variables, revealing strong intercorrelations among weather features and among booking-related features. To make it more understandable, only correlations (*r*) between weather and capacity variables are shown in figure 3. Furthermore, booking demand is more closely correlated with other capacity metrics, such as the number of villas rented before the booking date (*r*=.95) and whether there is a public holiday (*r*=.53). The highest correlated weather characteristics are the maximum temperature (*r*=.3) and the maximum snow depth (*r*=-.17). As mentioned previously, the naive model was created using the demand for bookings from last year as a forecast (*MAE*=.094, *MAPE*=.127, *RMSE*=.126). The occupancy distribution shows that there are strong deviations (*SD*=0.147) during the year, with peaks and lows especially around the beginning of the year. The PCA scree plots indicate that two components seem ideal for each PCA approach and each forecast horizon according to the "elbow rule" (Wicklin, 2017). The scree plots are included in the appendix and the model performance will solely be evaluated based on the MAE for the test set.

5.2 *RQ1: How well can the booking demand be predicted for different forecast horizons?*

This section evaluates the results of all experiments to identify the best-performing models, primarily using the MAE metric and, where applicable, by comparing the proportion of extreme errors. Additionally, model sanity is assessed by examining whether performance improves for shorter forecast horizons compared to longer ones. Finally, the impact of weather

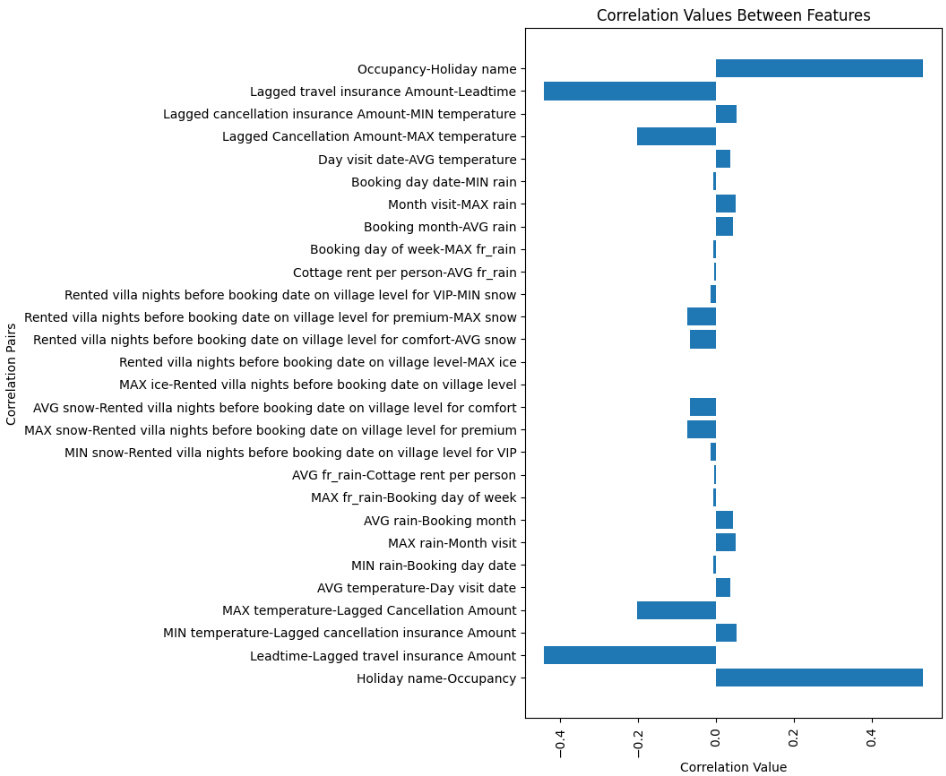

Figure 2: Correlations between capacity and weather features

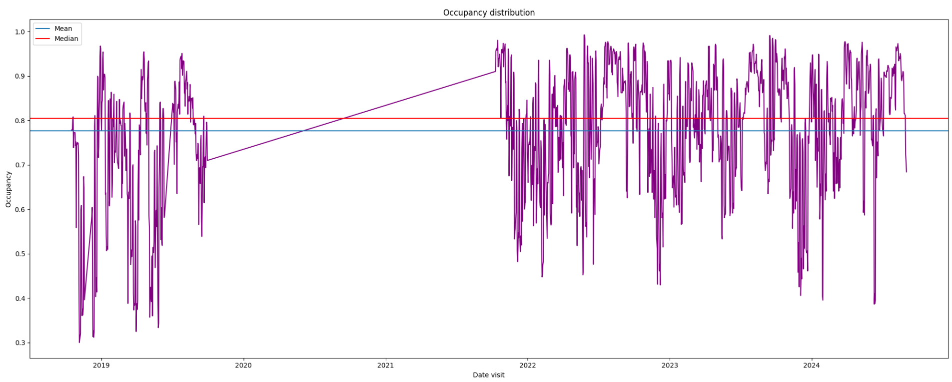

Figure 3: Distribution of booking demand

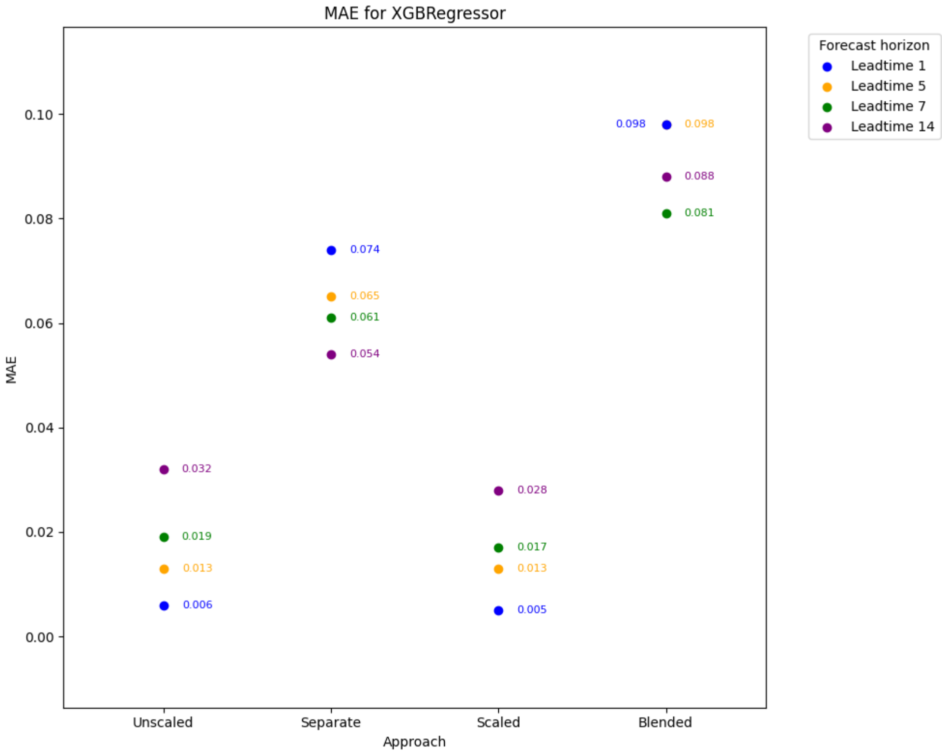

Figure 4: Experimental results for XGBR

features on model performance is analyzed. The results of all experiments are presented in figures 5, 6, and 4 for the MAE metric, with further explanations provided below. These figures illustrate the relationship between the training approach, shown on the X-axis, and model performance, represented on the Y-axis, across various forecast horizons. Each figure corresponds to a specific algorithm and evaluation metric. The labels on the X-axis reflect different experiments: "Separate" and "Blended" refer to models developed in Experiment 3, while "Unscaled" and "Scaled" correspond to models from Experiments 1 and 2, respectively. Similar figures for the MAPE and RMSE metrics can be found in the appendix. To address research question 1, which investigates how well booking demand can be predicted for different forecast horizons, the best-performing training approach and model are selected for each forecast horizon based on the experimental results. Subquestion 1.1, which examines whether performance improves with shorter forecast horizons, serves as a basic model sanity check across all experiments. Subquestion 1.2, regarding the impact of weather features, is addressed using the results from Experiment 1.

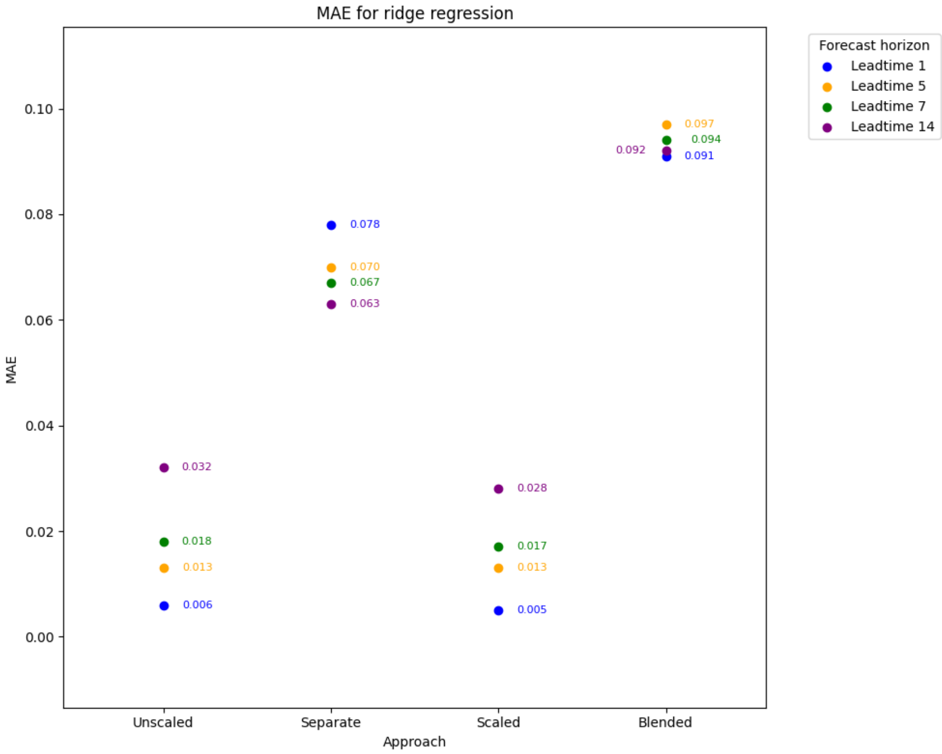

Figure 5: Experimental results for ridge regression

5.2.1 *Subquestion 1.1: Does (relative) forecasting performance improve with shorter forecast horizons?*

Subquestion 1.1 examines the performance improvement associated with shorter forecast horizons as a simple measure of model sanity. As shown in figures 5, 6, and 4, both unscaled and scaled models from experiments 1 and 2 perform better with shorter forecast horizons for XGBR and ridge regression. Similarly, for models trained without weather features (experiment 1), this trend holds true across all algorithms, as demonstrated in table2. In contrast, models with PCA (experiment 3) do not show improved performance with shorter forecast horizons, regardless of the algorithm. A similar pattern is observed for scaled SVR models (experiment 2). Thus, a detailed description and evaluation is provided in the answer to research question 2 regarding scaling and PCA.

5.2.2 *Subquestion 1.2: To what extent do the weather variables contribute to the forecast horizon-dependent performance?*

Subquestion 1.2, which examines the value of weather features for predicting booking demand, can be addressed by comparing models trained with

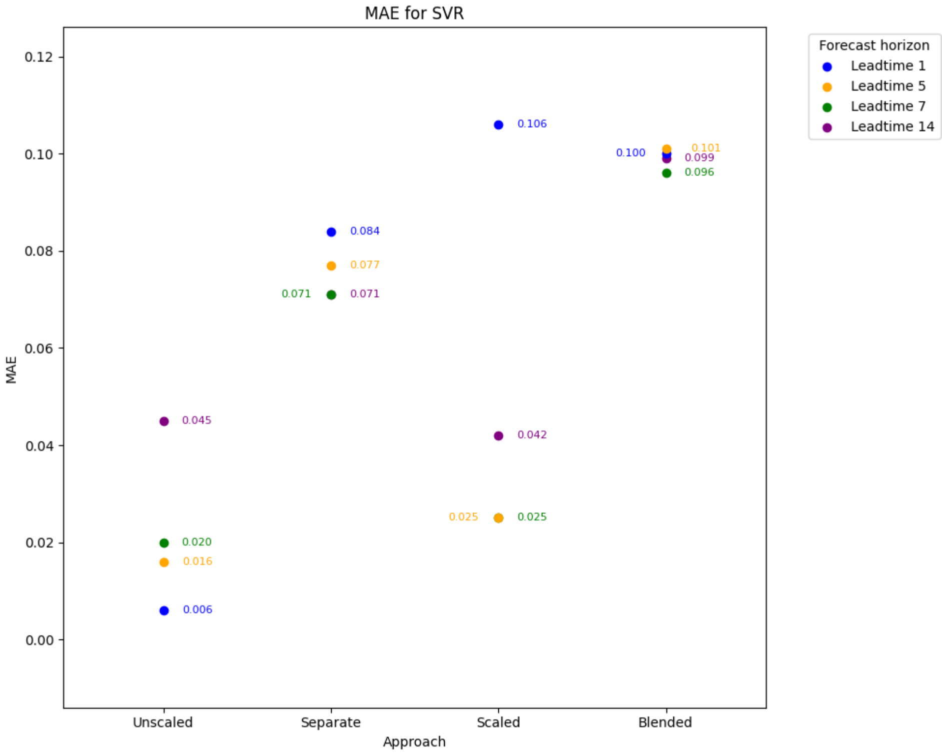

Figure 6: Experimental results for SVR

and without weather features. Models trained without weather features include only capacity-related features, with the inclusion or exclusion of weather features being the sole difference. The MAPE and RMSE tables can be found in the appendix 9 and 10, as MAE is used for the detailed comparison. The MAE metrics from table 2 indicate that XGBR models without weather features perform best for the 5-, 7-, and 14-day forecast horizons (figures 4 and 5). For the 1-day forecast horizon, ridge regression and SVR perform best. Shorter forecast horizons yield better performance overall, which supports the reasonableness of the models. Additionally, none of the predictions deviate by more than two standard deviations from the actual values. A comparison with unscaled models shows similar performance. Notably, the best-performing models without weather features outperform the unscaled best-performing models for the 1-, 7-, and 14-day forecast horizons. However, for the 5-day forecast horizon, models with weather features perform slightly better. As the performance differences are limited to the third decimal place of the MAE metric, these differences can be considered minimal. This comparison suggests that weather features do not have a significant impact on model performance.

Table 2: *Demand forecasting without weather features (MAE)*

Forecast horizon	MAE score		
	XGBR	Ridge regression	SVR
1	.006	**.005**	**.005**
5	**.014**	.016	.016
7	**.018**	.021	.019
14	**.031**	.034	.042

5.2.3 *Findings for Research Question 1*

The best-performing models from all experiments are showcased below and give an answer to research question 1. First, the robustly scaled XGBR models from experiment 2 performed best for the 7-day and 14-day forecast horizons. For the 1-day forecast horizon, the scaled XGBR model (experiment 2) and the scaled ridge regression model (experiment 1) both performed the best. Furthermore, for the 5-day forecast horizon, three models performed equally well: the scaled ridge regression, unscaled ridge regression, and scaled XGBR models. All of these top-performing approaches outperform the naive baseline model, and shorter forecast horizons generally yielded better results than longer ones, indicating that the models perform reasonably well. Figures 5 and 4 illustrate the model performances across forecast horizons, algorithms, and training approaches. These figures also highlight that both unscaled and scaled models perform nearly equally well and outperform the PCA-based approaches from experiment 3 for both XGBR and ridge regression.

5.3 *RQ2: What model and training procedure provide the best demand forecasting?*

This section further evaluates the results of all experiments to identify the model that provides the best demand forecasting, focusing on error analysis, the effects of scaling, and PCA. Where applicable, the proportion of extreme errors is compared. Additionally, model sanity is assessed by examining whether performance improves for shorter forecast horizons compared to longer ones. The results of all experiments are presented in Figures 5, 6, and 4 for the MAE metric, with further explanations provided below. All unscaled, scaled, and separate XGBR and ridge regression models outperformed the naive baseline model across the MAE, RMSE, and MAPE metrics, indicating better performance than using last year's occupancy as predictions. For SVR, the unscaled, separate models, and most unscaled models outperformed the naive model. However, the scaled model on the 1-day forecast horizon and all models with the blended PCA approach performed worse than the naive model. Additionally, only the blended XGBR models with the 7- and 14-day forecast horizons outperformed the naive model. Finally, for ridge regression, only the blended models with the 1- and 14-day forecast horizons outperformed the naive model. However, evaluating performance alone is insufficient, as some errors, while small in general, can significantly impact practical applications if they occur in critical scenarios. To address this, an error analysis is conducted for the best-performing models and training approaches identified in 5.2. These include the scaled XGBR model (experiment 2) and the unscaled ridge regression model (experiments 1 and 2) for the 1-day forecast horizon. For the 5-day, 7-day, and 14-day forecast horizons, the scaled XGBR models from experiment 2 performed best. Finally, while the models with PCA did not perform as well as expected, they are evaluated further to investigate the potential issues, as they were anticipated to perform better. Additionally, the impact of using separate PCA versus blended PCA will be compared. To assess model bias, the median and mean residuals are reported. Additionally, prediction errors exceeding two standard deviations from the actual values are analyzed.

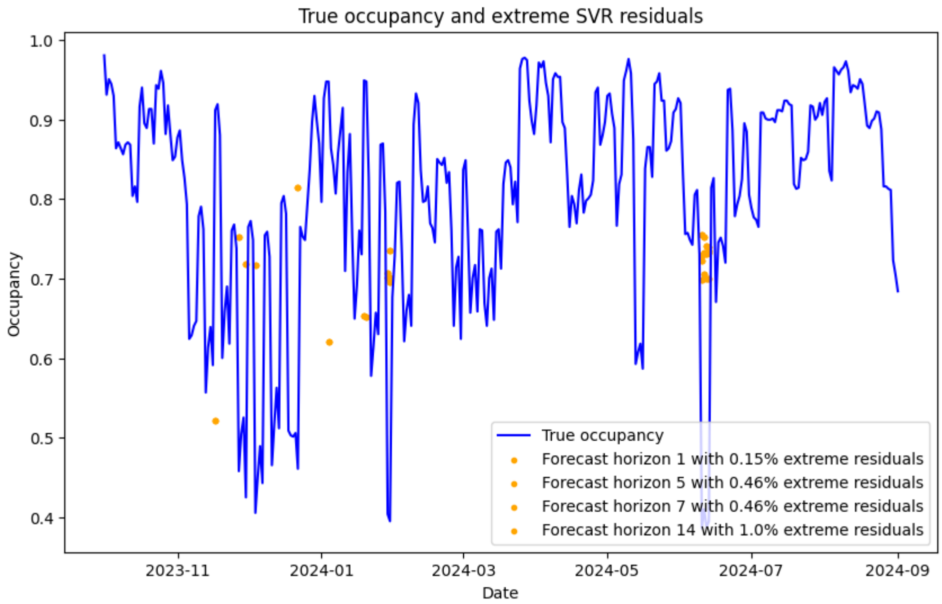

Figure 7: Extreme predictions by SVR with the blended training approach and true occupancy

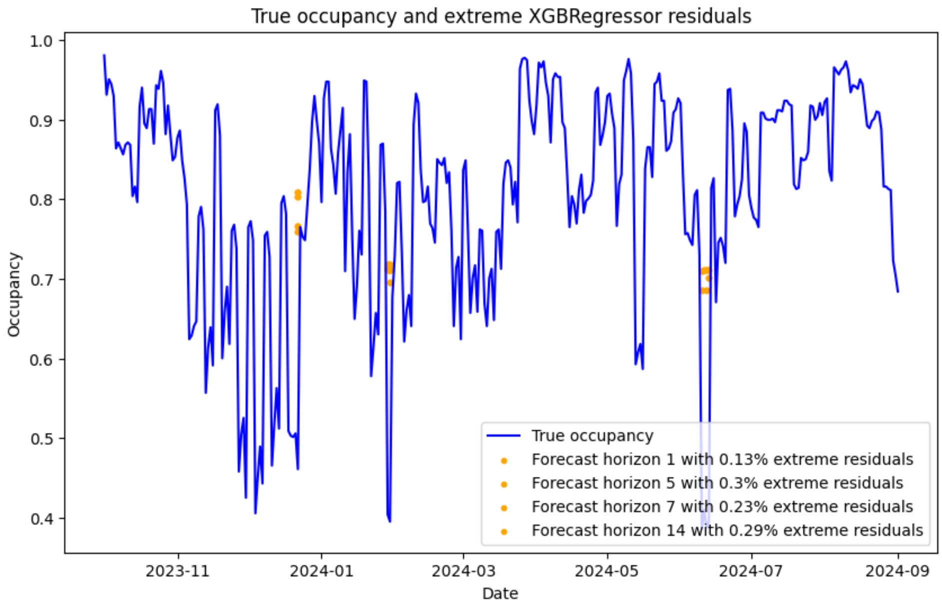

Figure 8: Extreme predictions by XGBR with the blended training approach and true occupancy

5.3.1 *Subquestion 2.1: How does robust scaling affect the models?*

Subquestion 2.1 is about the impact of scaling the features on model performances compared to the unscaled features (experiment 2). Therefore, the performance and the residuals are compared in an error analysis. This paragraph is especially relevant, as the scaled XGBR models were among the best performing models for the 5-, 7-, and 14- forecast horizons. Robust scaling improved the performance of both ridge regression and XGBR models for the 1-, 7-, and 14-day forecast horizons (figure 5 and 4). However, the performance on the 5-day horizon remained unchanged for both algorithms. For SVR, robust scaling resulted in worse performance for the 1-, 5-, and 7-day horizons, while performance improved for the 14-day horizon (figure 6). Notably, the SVR model for the 1-day horizon performed worse than those for the other horizons. Additionally, the SVR models for the 5- and 7-day horizons performed similarly, further reinforcing the observation that scaling does not work well for this algorithm. Furthermore, .58% of predictions made by the scaled SVR models differ at least 2 standard deviations from the truth. These extreme errors predominantly occur in December 2023, a period marked by exceptionally variable occupancy. Especially for the 1-day forecast horizon, where some predictions were even exceeding the possible occupancy. Since the hyperparameter tuning set the model complexity parameter C to 10 for the 1-day forecast horizon, the performance dip is probably due to overfitting. Furthermore, scaled

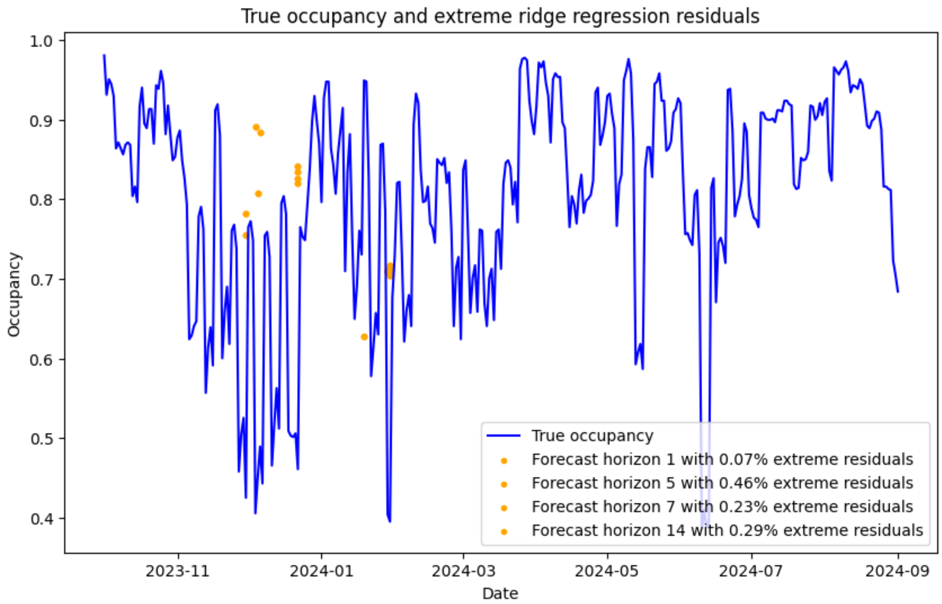

Figure 9: Extreme predictions by ridge regression with the blended training approach and true occupancy

ridge regression and XGBR both perform better with stronger regularization for the same horizon. This mistake in hyperparameter tuning could be mitigated by increasing the amount of folds in the cross-validation to improve generalizability.

Table 3: *Distribution of predictions differing from the truth above 2 standard deviations for the separate training approach*

Forecast horizon	Extreme predictions (%)		
	XGBR	Ridge regression	SVR
1	0.07	**.05**	.1
5	**.22**	.38	.23
7	.26	.39	**.12**
14	**0**	.41	.15
Total	**.55**	1.23	.6

Now proceeding with error analysis, focusing on the mean of the residuals. The mean residuals for the XGBR and ridge regression models indicate that scaling increased the underestimation of predictions for the 5-, 7-, and 14-day forecast horizons (tables 5 and 7). For the 1-day forecast horizon, there were no changes in the mean residuals for either XGBR or ridge regression, but both models slightly underestimated predictions. In

Table 4: *Distribution of predictions differing from the truth above 2 standard deviations for the blended training approach*

Forecast horizon	Extreme predictions (%)		
	XGBR	Ridge regression	SVR
1	0.11	**.07**	.14
5	**.3**	.46	.46
7	**.22**	.23	.46
14	**.29**	**.29**	1
Total	**.92**	1.04	2.06

contrast, the unscaled XGBR and ridge regression models slightly overestimated predictions (experiment 1). For SVR, the mean residuals increased after scaling for the 5- and 7-day forecast horizons, indicating a slight increase in underestimating predictions. For the 14-day forecast horizon, scaling resulted in a larger decrease in model underestimations. For the 1-day forecast horizon, scaling caused a shift from slightly overestimating to underestimating predictions.

Table 5: *Mean residuals (Truth-Prediction) for the models with weather features of experiment 1*

Forecast horizon	Mean residuals		
	XGBR	Ridge regression	SVR
1	.002	.002	**-.001**
5	**.003**	**.003**	.004
7	**.005**	**.005**	**.005**
14	**-.006**	**-.006**	.026

The median residuals for XGBR and ridge regression show that underestimations increased for the 5- and 7-day forecast horizons, albeit to a smaller extent than the mean residuals, indicating larger overall underestimations. However, these differences are minor, falling in the third decimal. For the 1-day forecast horizon, scaling reduced underestimations in the median residuals. For the 14-day forecast horizon, scaling shifted the models from overestimating to underestimating predictions. For SVR, the median residuals increased in underestimations after scaling was applied for the 1-, 5-, and 7-day forecast horizons. Notably, scaling also resulted in a significant decrease in underestimations for the 14-day forecast horizon. Comparing the mean and median residuals, the 1-day forecast horizon

Table 6: *Median residuals (Truth-Prediction) for the models with weather features of experiment 1*

Forecast horizon	Median residuals		
	XGBR	Ridge regression	SVR
1	**.002**	**.002**	**.002**
5	.002	.002	**.001**
7	**.001**	.002	**.001**
14	**-.009**	**-.009**	.026

shows slight underestimations in the median residuals but slight overestimations in the mean residuals, indicating the presence of some stronger overestimating predictions. For the 5- and 7-day forecast horizons, the median residuals are smaller than the mean residuals, suggesting stronger underestimations. For the 14-day forecast horizon, the mean and median residuals are equal. When comparing the scaled median and mean residuals for SVR, they are generally consistent in underestimating predictions, except for the 7-day forecast horizon.

Table 7: *Mean residuals (Truth-Prediction) for the scaled models of experiment 2*

Forecast horizon	Mean residuals		
	XGBR	Ridge regression	SVR
1	**.002**	**.002**	.004
5	**.006**	**.006**	.008
7	.007	**.006**	.008
14	.005	**.002**	.007

Table 8: *Median residuals (Truth-Prediction) for the scaled models of experiment 2*

Forecast horizon	Median residuals		
	XGBR	Ridge regression	SVR
1	**.001**	**.001**	.004
5	.004	**.003**	.007
7	.004	**.003**	.007
14	.004	**.002**	.007

5.3.2 *Subquestion 2.2: Does performing a separate dimensionality reduction for weather and capacity variables hinder performance?*

Subquestion 2.2 examines the impact of performing separate PCA for capacity and weather features compared to applying PCA on all features together (experiment 3). Overall, applying PCA resulted in worse performance regardless of the forecast horizon or algorithm, as shown in figures 5, 6, and 4. Surprisingly, the separate PCA outperformed the blended PCA. These unexpected results, along with potential reasons, are explored further below separately for each algorithm.

Ridge regression and XGBR produced their worst-performing models with the blended PCA. The same applied to SVR, except for the 1-day forecast horizon, where SVR performed the worst with scaling (experiment 1). Notably for ridge regression, the 1-day forecast horizon model performed only slightly better than the 14-day horizon model, while the 7-day model outperformed the 5-day model, which performed the worst. For XGBR and SVR the models with 7-days forecast horizons perform best, followed by the 14-days horizons. The worst performing model horizons for XGBR and SVR are the 5- and 1-days horizons. Since the models do not consistently perform better with shorter forecast horizons and sometimes perform worse, it suggests that the model's reliability is questionable. Additionally, the proportion of extreme predictions, as shown in table 4, indicates that the blended PCA contributes to these inaccuracies. For ridge regression and XGBR models, the separate PCA resulted in worse performance for shorter forecast horizons (figures 5 and 4). The same trend applies to SVR, except for the 7- and 14-day forecast horizons, which performed equally well. As previously mentioned, the decreasing performance for shorter forecast horizons suggests that the separate PCA produced unreliable models.

Both PCA approaches led to extreme predictions, as detailed in tables 3 and 4. For the separate PCA models combined with ridge regression, 1.23% of predictions deviated by at least two standard deviations from the truth, compared to 1.04% for the blended PCA models. Furthermore, extreme predictions increased with longer forecast horizons for the separate PCA models. A similar trend was observed for the blended PCA models, except for the 5-day horizon, which exhibited the highest proportion of extreme predictions among all forecast horizons. Among the blended models, XGBR achieved the lowest proportion of extreme predictions (0.92%), followed by ridge regression (1.04%), while SVR had the highest (2.06%). figures 7,8, and9 illustrate the distribution of extreme errors for each algorithm under the blended approach, highlighting consistent error patterns across models. Notably, most extreme errors occurred during volatile months such as December 2023, January 2024, and July 2024.

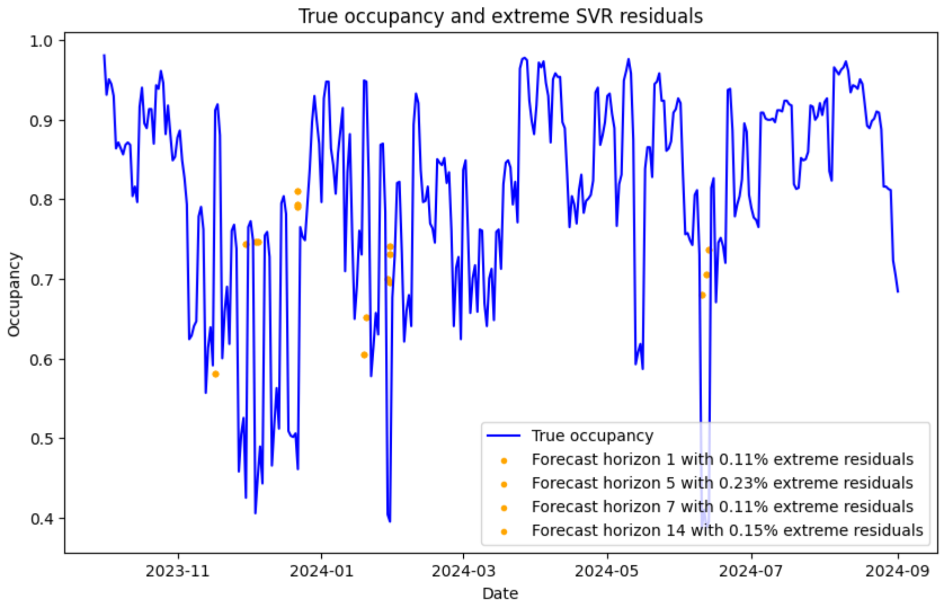

Figure 10: Extreme predictions by SVR with the separate training approach and true occupancy

For the separate PCA approach, the proportion of extreme predictions as shown in Table 3 indicates that XGBR and SVR models produced fewer extreme errors compared to the blended approach. In contrast, ridge regression generated a higher proportion of extreme errors with the separate PCA approach. Separate XGBR models achieved the fewest extreme predictions (0.55%), followed by SVR (0.6%), while ridge regression had the highest proportion (1.23%). Additionally, XGBR and ridge regression models exhibited an increase in extreme prediction errors with longer forecast horizons, as shown in figures 10,11, and12. For SVR, however, this trend was not observed, with the largest proportion of extreme errors occurring in the 5-day forecast horizon. Interestingly, under the blended approach, the proportion of extreme prediction errors did not consistently relate to forecast horizons, as indicated in table 4.

As mentioned earlier, the performance decrease may be attributed to the emphasis placed on the rented villa nights before the booking date features in the capacity-specific PCA. Notably, the performance difference between forecast horizons is smaller for the blended approach compared to other approaches, and excluding the rented villa nights before the booking date features from the PCA might improve performance. However, as shown in Figure 6, SVR models trained with scaling or PCA performed worse than unscaled models. This result is unexpected, given that SVR typically benefits from scaling and dimensionality reduction due to its

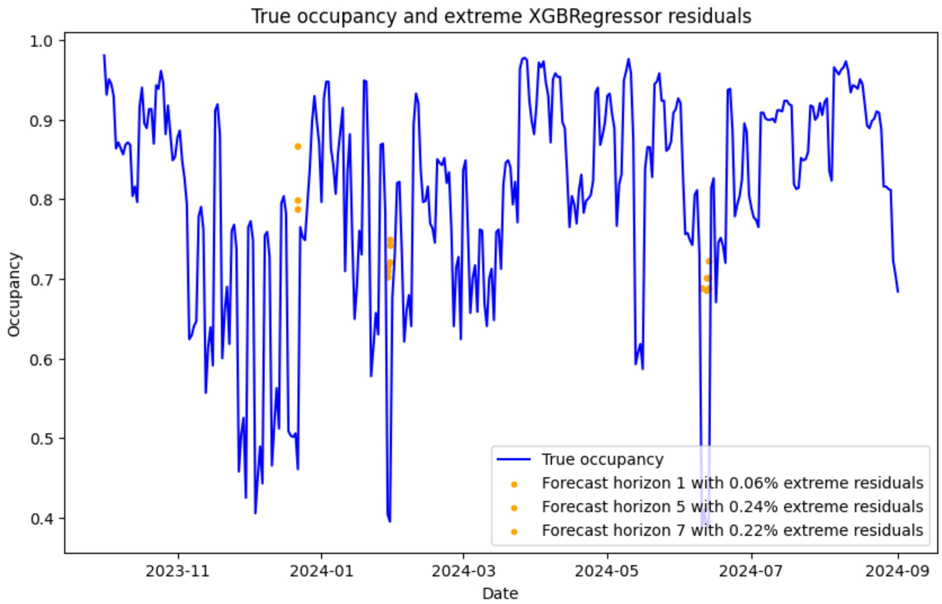

Figure 11: Extreme predictions by XGBR with the separate training approach and true occupancy

sensitivity to both high dimensionality and feature scaling (Ahsan, Mahmud, Saha, Gupta, & Siddique, 2021; Rastogi, Taterh, & Kumar, 2023). Additionally, performance generally decreases with shorter forecast horizons for both the separate and blended training approaches. This trend, coupled with the observed performance drop, further suggests that the models are not operating optimally, possibly due to the reduced impact of the rented villa nights before the booking date features when included in the PCA. The median and mean residuals further reveal that, for the blended approach, most models tend to overestimate, whereas the separate approach demonstrates a more balanced performance. In conclusion, the separate PCA improves performance compared to the blended PCA.

5.3.3 *Findings for Research Question 2*

Finally, the best-performing models are evaluated to address research question 2. None of the models discussed in this section produced prediction errors exceeding two standard deviations from the actual values. The error analysis reveals that the best-performing models exhibit slight underestimations across all forecast horizons. These findings align with the conclusions from research question 1 (5.2.3), confirming that the models identified there were indeed the best-performing. Specifically, the scaled XGBR models perform best across all forecast horizons, except for the 1- and 5-days forecast horizon, where they match the performance of the unscaled and

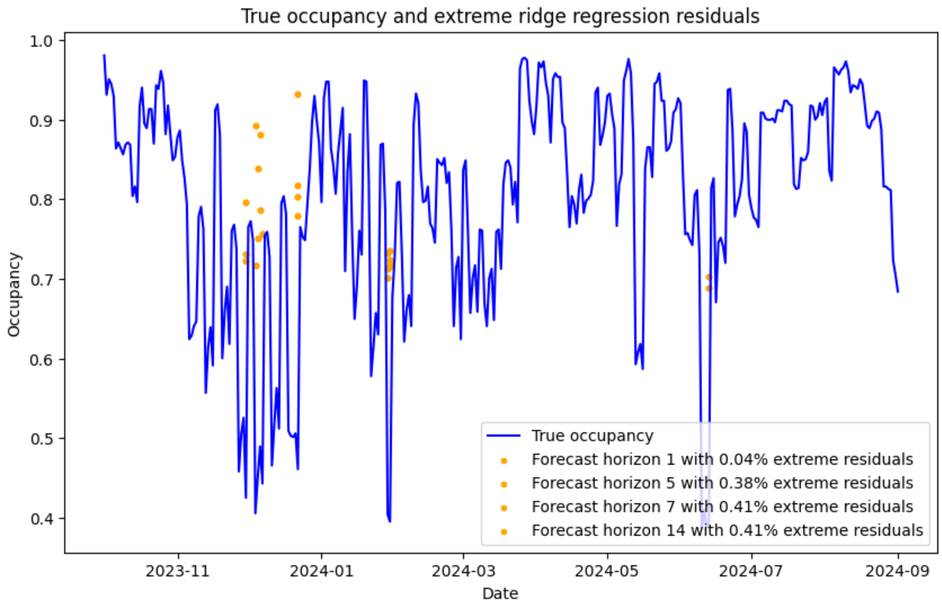

Figure 12: Extreme predictions by ridge regression with the separate training approach and true occupancy

scaled ridge regression models. Minimal differences between the mean and median residuals, along with the absence of deviant predictions, further validate these models as the best-performing.

6 DISCUSSION

This study aimed to establish a foundation for integrating various weather metrics with historical booking data to enhance machine learning-based booking demand forecasting. This was recommended by various previous literature reviews (Henriques & Nobre Pereira, 2024; Huang & Zheng, 2022). More accurate booking demand forecasting has societal relevance, as it allows the hotels to optimize planning staff, catering, activities, and other factors better (Gunaseelan, Alalmai, & Arun, 2020). Improved planning allows for better service and a better recreational experience, while also increasing revenue (Gunaseelan, Alalmai, & Arun, 2020). Given the limited prior research on weather features used for predicting booking demand (Pan & Yang, 2016), rather basic approaches have been tested. The following training approaches were tested for XGBR, SVR, and ridge regression: Robust scaling of continuous features versus no scaling, blending all continuous features in a PCA, and separating continuous features in capacity and weather-related principal components. Furthermore, the impact of weather features on predicting short-term occupancy was investigated by training models without weather features. For the error analysis, the mean and median residuals were examined in addition to predictions that differ over two standard deviations from the true values (Zhang et al., 2018).

6.1 *Findings to RQ1: How well can the booking demand be predicted for different forecast horizons?*

The first research question examines how effectively booking demand can be forecasted across various horizons. Subquestions investigate whether model performance improves with shorter forecast horizons (as a sanity check) and the impact of weather features on performance.

The best-performing models for addressing this research question are highlighted below. All models outperform the naive baseline, with shorter horizons yielding better results. Figures 5 and 4 demonstrate that scaled and unscaled models consistently outperform PCA-based approaches from experiment 3. Weather features did not appear to impact performance, as models trained without them performed similarly to those trained with weather features. While training models exclusively with weather features could provide further insights in future research, this specific study found no significant improvement in performance that justifies their inclusion. The study suggests that, for the selected Center Parcs location, weather features are less effective than historical capacity features.

Among the best-performing models, scaled XGBR models excelled for the 7- and 14-day horizons, while both scaled XGBR and ridge regression models performed best for the 1-day horizon. For the 5-day horizon, scaled and unscaled ridge regression and XGBR models performed equally well.

6.2 *Findings to RQ2: What model and training procedure provide the best demand forecasting?*

The second research question evaluated the best demand forecasting models, focusing on error analysis, scaling, and PCA effects. Methods include assessing model performance, comparing blended versus separate PCA approaches, and analyzing residuals for bias and extreme errors. In the blended PCA approach, principal components are calculated using all forecast horizons combined, while in the separate PCA approach, components are derived independently for each forecast horizon. Prediction errors exceeding two standard deviations are also examined to ensure robustness.

The results for the scaled models were somewhat surprising, as SVR models performed worse with robustly scaled features, while XGBR and ridge regression showed similar or even improved performance (Ahsan, Mahmud, Saha, Gupta, & Siddique, 2021). The performance decrease for

SVR with scaled features appears to be due to poorly chosen hyperparameters during cross-validation, which could be addressed by increasing the number of folds. Error analysis for the scaled models shows that scaling generally increased underestimations for XGBR and ridge regression at longer forecast horizons (5 and 7 days), although the median residuals showed smaller changes than the mean, indicating stronger overall underestimations. At the 1-day horizon, scaling reduced underestimations in the median residuals, but the mean residuals remained slightly underestimated, suggesting some stronger overestimations. For the 14-day horizon, scaling shifted XGBR and ridge regression models from overestimating to underestimating predictions slightly.

For SVR, scaling increased underestimations at shorter horizons (1, 5, and 7 days) but reduced them significantly at the 14-day horizon. Comparing mean and median residuals, smaller median values for longer horizons indicate stronger underestimations, while consistency between mean and median values at the 14-day horizon reflects balanced predictions. Overall, SVR's residuals remained largely consistent in underestimating predictions, with exceptions at the 7-day horizon.

Both PCA approaches resulted in extreme predictions, with the separate PCA models (ridge regression) showing more frequent significant deviations than the blended PCA models. Extreme errors increased with forecast horizons in the separate PCA approach, while the blended PCA approach displayed inconsistent trends, with a noticeable spike at the 5-day horizon. Among blended models, XGBR produced the fewest extreme errors, followed by ridge regression, while SVR had the most. For separate PCA models, XGBR also had the fewest extreme errors, followed by SVR, with ridge regression producing the most. Performance variations may stem from the inclusion of the rented villa nights feature in PCA, which seems to reduce model effectiveness, as it is highly correlated with the target variable. SVR performed worse with scaling or PCA, contrary to expectations given its sensitivity to feature scaling. Across both PCA approaches, shorter forecast horizons generally resulted in poorer performance. Blended PCA models tended to overestimate, whereas separate PCA models demonstrated more balanced predictions. Overall, the separate PCA approach proved superior to the blended approach. However, the error analysis indicates that PCA reduced both performance and model reliability compared to other approaches.

The best-performing models were evaluated for research question 2, and confirm the findings from research question 1, with scaled XGBR models performing best overall, except for the 1-day horizon, where they match unscaled ridge regression. Moreover, the best-performing models for the 7-day horizon (MAE = 0.017, $MAPE$ = 0.022, $RMSE$ = 0.024), and 14-day horizon (*MAE* = 0.028, *MAPE* = 0.036, *RMSE* = 0.039) were scaled XGBR models. For the 1-day forecast horizon, both the scaled XGBR and scaled ridge regression models performed equally well (*MAE* = 0.005, *MAPE* = 0.007, *RMSE* = 0.007). Finally, for the 5-day forecast horizon (*MAE* = 0.013, *MAPE* = 0.017, *RMSE* = 0.019), ridge regression and XGBR models performed equally well, regardless of scaling. The best models show slight underestimations across all forecast horizons without significant prediction errors. Minimal residual differences in the third decimal and no deviant predictions validate these models' performance.

6.3 *Comparison of results with previous research*

The best models of this study were obtained with the scaled and unscaled training approaches, which will be used to compare with previous research. Only performance for the same forecast horizons of 1, 5, 7, and 14 days will be compared. Furthermore, the same applies for the metrics, which are MAE, RMSE, and MAPE.

This study outperformed prior research using Covid-19 data, as shown in Table 1 (Viverit et al., 2023; Wu et al., 2021). Viverit et al. (2023) clustered booking curves and applied autoregression to predict occupancy for three European hotels using three years of data, including Covid-19. Their models, tested on 7- to 50-day horizons, had higher RMSE, MAE, and MAPE for 7- and 14-day horizons compared to this study. Limitations included reliance on Covid-19 and hotel-specific data. Wu et al. (2021) forecasted monthly occupancy during COVID-19 times, making performance comparisons with this study difficult.

The best models of this study outperformed those from previous research without Covid-19 data across comparable forecast horizons (Ampountolas & Legg, 2021; Pereira & Cerqueira, 2021; Phumchusri & Ungtrakul, 2019; Kaya et al., 2022; Pan & Yang, 2016; Wu et al., 2022). First, Wu et al. (2022) forecasted 1- to 7-day horizons using online reviews for Macau hotels, but their models had higher RMSE and MAPE for all these horizons. Kaya et al. (2022) forecasted weekly demand for Turkish hotels using XGBR, but their models underperformed for comparable weekly forecasts. Ampountolas and Legg (2021) forecasted 1-, 7-, and 14-day horizons using social media data for a U.S. hotel chain, achieving lower accuracy than this study. Similarly, Pereira and Cerqueira (2021) tested multiple algorithms on South European hotel data for 1-, 5-, 7-, and 14-day horizons, all with higher RMSE. Phumchusri and Ungtrakul (2019) forecasted daily demand for a Thai hotel, but their models underperformed in comparison on the 1-day horizon. Lastly, Pan and Yang (2016) used ARIMAX and weather data to predict weekly U.S. hotel occupancy but achieved worse MAPE for weekly forecasts. These studies faced limitations such as regional bias, restricted data sources, or exclusion of relevant features.

The best models of this study outperformed previous research in predicting occupancy, as outlined in Table 1 and the related work section 3. The results are robust, as train-test-validation splits were used for cross-validation, and the models were tested on a year of unseen data. However, for generalization, it is important to note that occupancy in this study refers to houses rather than hotel rooms (Henriques & Nobre Pereira, 2024; Huang & Zheng, 2022). This distinction is crucial for comparing performance, as the houses at Center Parcs are likely less prone to short-term cancellations or bookings due to the customer base, which predominantly consists of families or groups rather than individuals (Ito, Kanemitsu, Kimura, & Omori, 2024). Generally, travel reasons and group size can influence short-term occupancy variance, as found in previous research (Ito, Kanemitsu, Kimura, & Omori, 2024), so this should be investigated further. Further limitations for comparisons are that this study only used data from four years for a family resort in Western Europe. Thus, it might not be generalizable to a hotel or resort elsewhere. Additional differences include some studies relying on COVID-19 data (Viverit et al., 2023; Wu et al., 2021), while others did not incorporate capacity features (Wu et al., 2022). Furthermore, weather data only marginally influenced occupancy predictions, which aligns with findings from earlier studies using weather features (Pan & Yang, 2016).

6.4 *Future research recommendations*

In case future research is based on these experiments, there should be changes with the implementation of PCA. Specifically, a sophisticated selection of which features used for dimensionality reduction should be implemented. In addition to that, a more reliable measure for the ideal component number should be experimented with. To get a better understanding of noise and signal in the data, the "elbow rule" and other PCA methods such as permutation tests should be compared (Toledo Jr., 2022; Wicklin, 2017). Additionally, to further investigate whether the weather has an impact, different locations and different time frames should be used to improve generalizability. Furthermore, models trained solely on weather data could provide further insights into if the weather data can be used to predict occupancy. As the weather features were expected to predict occupancy because of weather-dependent activities, it might be better to directly predict activity bookings (Bausch et al., 2021; McKercher et al., 2014). Also, a wider selection of hyperparameters and an increased number of folds in cross-validation could further increase the performance.

Generally, future demand forecasting research could benefit from this study as motivation to investigate the impact of the type of accommodation on occupancy forecasting. Moreover, the potential impact of weather should be investigated further because there is a lack of research, and this study cannot be fully generalized, as it only used one location over four years (Henriques & Nobre Pereira, 2024; Huang & Zheng, 2022).

7 CONCLUSION

This study investigated the impact of weather forecast data alongside historical booking data for predicting the booking demand of a Center Parcs location in Western Europe. The research questions addressed and compared the forecasting performance for various forecast horizons, machine learning algorithms, and model training procedures. Notably, this study outperformed previous occupancy predictions with mostly capacity data features, of which the number of rented villas before the forecast horizon was the most impactful feature. In general shorter forecast horizons led to better performance, scaled training data led to small performance improvements in some cases, and applying a PCA to the training data led to a performance decrease. Comparison of models trained without weather features to the models trained with weather features showed that the weather features only led to minor improvements in some cases. This led to the conclusion, that the weather appears to have only a negligible impact on forecasting booking demand. Thus, weather shall not be included in the revenue management systems of Center Parcs or similar family resorts until more research is conducted. Instead, available historical data shall be used for demand forecasting, as those were found to perform nearly equally well. Society, Center Parcs, and to some extent other family resorts can benefit from this research, since less money and labor need to be invested into revenue management to include weather features. Those saved resources can be redirected to other areas of Center Parcs to optimize the recreational experience. The best-performing models of this study showed an increase in performance compared to previous research, based on MAPE, RMSE, and MAE. XGBR, SVR, and ridge regression were the machine learning algorithms used. The limitations of this study were that occupancy in this study referred to houses rather than hotel rooms, and Center Parcs is a family-oriented resort, which differs significantly from traditional hotels. Additionally, the sampled data lacks diversity in terms of location and time, and a large number of capacity features were included alongside weather features, which may have influenced the results. In case dimensionality reduction is conducted via PCA, the number of components and features used in the PCA should be optimized. Generally, a wider selection of hyperparameters and a larger number of folds in the cross-validation could improve performance. Thus, future research should investigate occupancy for similar family-oriented resorts in different locations and across diverse time frames.

8 REFERENCES

Ahsan, M. M., Mahmud, M. A. P., Saha, P. K., Gupta, K. D., & Siddique, Z. (2021). Effect of data scaling methods on machine learning algorithms and model performance. *Technologies*, 9(3), 52. https://doi.org/10.3390/technologies9030052

Ampountolas, A., & Legg, M. P. (2021). A segmented machine learning modeling approach of social media for predicting occupancy. *International Journal of Contemporary Hospitality Management*, 33(6), 2001–2021. https://doi.org/10.1108/ijchm-06-2020-0611

Bausch, T., Gartner, W. C., & Humpe, A. (2021). How weather conditions affect guest arrivals and duration of stay: An alpine destination case. *International Journal of Tourism Research*, 23(6), 1006–1026. https://doi.org/10.1002/jtr.2459

Brownlee, J. (2020, August 18). Statistical imputation for missing values in machine learning. *Machine Learning Mastery*. https://machinelearningmastery.com/statistical-imputation-for-missing-values-in-machine-learning/

Center Parcs. (2024, August 21). Kom werken bij Center Parcs. https://www.workatcenterparcs.com/nl/kom-werken-bij-center-parcs/

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794). https://doi.org/10.1145/2939672.2939785

Contessi, D., Viverit, L., Nobre Pereira, L., & Heo, C. Y. (2024). Decoding the future: Proposing an interpretable machine learning model for hotel occupancy forecasting using principal component analysis. *Journal of Forecasting*, 43(8), 1484-1498. https://doi.org/10.1016/j.jforecast.2024.08.004

Diego R.-Toubes, Noelia Araújo-Vila, & José Antonio Fraiz-Brea. (2020). Influence of Weather on the Behaviour of Tourists in a Beach Destination. *Atmosphere*, 11(1), 121. https://doi.org/10.3390/atmos11010121

El-Mhouti, A., Fahim, M., Bahbah, A., El-Borji, Y., Souf, A., Aoulalay, A., & Ouazri, C. (2023). A machine learning-based approach for meteorological big data analysis to improve weather forecast.

International Journal of Computing and Digital Systems, 14(1), 863–874. https://doi.org/10.12785/ijcds/140167

Farfan, H., Castillo, A., & Curilef, S. (2019). Statistical analysis of weather variables of Antofagasta. *Proceedings of SOCHIFI 2018 (Sociedad Chilena de Física)*. arXiv. https://doi.org/10.48550/arXiv.1911.04232

Figini, P., Cicognani, S., & Zirulia, L. (2019). Booking in the rain: Testing the impact of public information on prices. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3481738

Géron, A. (2019). *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow* (2nd ed.). O'Reilly Media.

Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press.

Golets, A., Farias, J., Pilati, R., & Costa, H. (2021). COVID-19 pandemic and tourism: The impact of health risk perception and intolerance of uncertainty on travel intentions. *Tourism Management*, *42*, 2500–2513. https://doi.org/10.1007/s12351-021-00678-2

Gunaseelan, D., Alalmai, A., & Arun, A. (2020). Operational need and importance of capacity management in the hotel industry: A review. *International Journal of Advanced Science and Technology*, *7*, 122–130.

Gurucharan, M. K. (2018, October 9). Machine learning basics: Support vector regression. *Towards Data Science*. https://towardsdatascience.com/machine-learning-basics-support-vector-regression-660306ac5226

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., ... & Oliphant, T. E. (2020). Array programming with NumPy. *Nature*, *585*(7825), 357–362 [Library]. https://doi.org/10.1038/s41586-020-2649-2

Henriques, H., & Nobre Pereira, L. (2024). Hotel demand forecasting models and methods using artificial intelligence: A systematic literature review. *Tourism & Management Studies*, *20*(3), 39–51. https://doi.org/10.18089/tms.20240304

Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, *12*(1), 55–67. https://doi.org/10.1080/00401706.1970.10488634

Huang, L., & Zheng, W. (2022). Hotel demand forecasting: A comprehensive literature review. *Tourism Review*, *78*(1), 218–244. https://doi.org/10.1108/tr-07-2022-0367

Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. *Computing in Science & Engineering*, *9*(3), 90–95. https://doi.org/10.1109/MCSE.2007.55

Hyndman, R. J. (2009). Forecasting for the longer term. *Foresight: The International Journal of Applied Forecasting*. https://robjhyndman.com/papers/foresight.pdf

Hyndman, R. J., & Athanasopoulos, G. (2018). *Forecasting: Principles and practice* (2nd ed.). OTexts. https://otexts.com/fpp2/

Ito, K., Kanemitsu, S., Kimura, R., & Omori, R. (2024). Time changes of customer behavior on accommodation reservation: A case study of Japan. *Japan Journal of Industrial and Applied Mathematics, 41*, 881–902. https://doi.org/10.1007/s13160-023-00455-5

Jäger, S., Allhorn, A., & Bießmann, F. (2021). A benchmark for data imputation methods. *Frontiers in Big Data, 4*, Article 693674. https://doi.org/10.3389/fdata.2021.693674

Kim, S., & Kim, H. (2016). A new metric of absolute percentage error for intermittent demand forecasts. *International Journal of Forecasting, 32*(3), 669–679. https://doi.org/10.1016/j.ijforecast.2015.12.003

Kotsiantis, S. B. (2013). Decision trees: A recent overview. *Artificial Intelligence Review, 39*(3), 261–283. https://doi.org/10.1007/s10462-011-9272-4

Kuhn, M., & Johnson, K. (2013). *Applied predictive modeling.* Springer. https://doi.org/10.1007/978-1-4614-6849-3

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems (NIPS), 30*. https://doi.org/10.48550/arXiv.1705.07874

Makridakis, S., Wheelwright, S. C., & Hyndman, R. J. (1998). *Forecasting: Methods and applications* (3rd ed.). Wiley.

Maleki, F., Ovens, K., Gupta, R., Reinhold, C., Spatz, A., & Forghani, R. (2022). Generalizability of machine learning models: Quantitative evaluation of three methodological pitfalls. *arXiv.* https://doi.org/10.48550/arXiv.2202.01337

McKercher, B., Shoval, N., Park, E., & Kahani, A. (2014). The [limited] impact of weather on tourist behavior in an urban destination. *Journal of Travel Research, 54*(4), 442–455. https://doi.org/10.1177/0047287514522880

McKinney, W. (2010). Data structures for statistical computing in Python. *Proceedings of the 9th Python in Science Conference*, 51–56. https://doi.org/10.25080/Majora-92bf1922-00a

Microsoft. (n.d.). *Visual Studio Code* [Computer software]. https://code.visualstudio.com

Murel, J., & Kavlakoglu, E. (2023, November 21). Ridge regression. *IBM.* https://www.ibm.com/topics/ridge-regression

Niemann, M., & Schienle, M. (2023). Forecast horizon and point forecast. In *Forecasting in Logistics and Supply Chains* (pp. 113–122). Springer. https://doi.org/10.1007/978-3-031-27852-5_5

OpenWeather. (n.d.). Terms and conditions of sale. *OpenWeather*. https://openweather.co.uk/storage/app/media/Terms/Openweather_terms_and_conditions_of_sale.pdf

OpenWeather. (2024, September 2). *Historical weather data: Bulk* [Data set]. Retrieved from https://openweathermap.org/api/history-forecast-bulk

Oshternian, S. R., Loipfinger, S., Bhattacharya, A., & Fehrmann, R. S. N. (2024). Exploring combinations of dimensionality reduction, transfer learning, and regularization methods for predicting binary phenotypes with transcriptomic data. *BMC Bioinformatics, 25*, Article 167. https://doi.org/10.1186/s12859-024-05567-9

Overleaf. (n.d.). *Overleaf* [Online LaTeX editor]. TeX Live 2023. https://www.overleaf.com

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research, 12*, 2825–2830. https://www.jmlr.org/papers/v12/pedregosa11a.html

Pereira, L. N., & Cerqueira, V. (2021). Forecasting hotel demand for revenue management using machine learning regression methods. *Current Issues in Tourism, 25*(17), 2733–2750. https://doi.org/10.1080/13683500.2021.1999397

Phumchusri, N., & Ungtrakul, P. (2019). Hotel daily demand forecasting for high-frequency and complex seasonality data: A case study in Thailand. *Journal of Revenue and Pricing Management, 19*(1), 8–25. https://doi.org/10.1057/s41272-019-00221-6

Python Software Foundation. (2023). *Python* (Version 3.12) [Programming language]. https://www.python.org

Rastogi, A. K., Taterh, S., & Kumar, B. S. (2023, October 4–5). Dimensionality reduction algorithms in machine learning: A theoretical and experimental comparison. Paper presented at the *International Conference on Recent Advances in Science and Engineering*, Dubai, United Arab Emirates.

Sánchez-Medina, A. J., & C-Sánchez, E. (2020). Using machine learning and big data for efficient forecasting of hotel booking cancellations. *International Journal of Hospitality Management, 89*, 102546. https://doi.org/10.1016/j.ijhm.2020.102546

Satu, Md. S., Ahammed, K., & Abedin, M. Z. (2020). Performance analysis of machine learning techniques to predict hotel booking cancellations in hospitality industry. *2020 23rd International Conference on Computer and Information Technology (ICCIT)*. https://doi.org/10.1109/iccit51783.2020.9392648

Sharma, V. (2022). A study on data scaling methods for machine learning. *International Journal for Global Academic & Scientific Research, 1*(1), 31–42. https://doi.org/10.55938/ijgasr.v1i1.4
Sugg, L. (2024). *DSS Thesis 2024*. GitHub. Retrieved December 23, 2024, from https://github.com/Leo55555nard/DSS-Thesis-2024-
Tableau. (2024). *Tableau Prep* (Version 2024.1.0) [Computer software]. Salesforce. https://www.tableau.com
Tashman, L. J. (2000). Out-of-sample tests of forecasting accuracy: An analysis and review. *International Journal of Forecasting, 16*(4), 437–450. https://doi.org/10.1016/S0169-2070(00)00065-0
Toledo Jr., T. (2022, January 4). PCA 102: Should you use PCA? How many components to use? How to interpret them? A dive on some more intermediate concepts on the PCA analysis. *Towards Data Science*. Retrieved from https://towardsdatascience.com
VanderPlas, J., & Scikit-learn developers. (n.d.). Ridge regression. In *Scikit-learn: Machine learning in Python*. https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.Ridge.html
VanderPlas, J., & Scikit-learn developers. (n.d.). Support vector regression. In *Scikit-learn: Machine learning in Python*. https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVR.html
Wicklin, R. (2017, August 2). Dimension reduction: Guidelines for retaining principal components. *The DO Loop*. https://blogs.sas.com/content/iml/2017/08/02/dimension-reduction-guidelines-retaining-pcs.html
Wu, Y., & Xue, W. (2024). Data-driven weather forecasting and climate modeling from the perspective of development. *Atmosphere, 15*(6), 689. https://doi.org/10.3390/atmos15060689
Zhang, M., Li, J., Pan, B., & Zhang, G. (2018). Weekly hotel occupancy forecasting of a tourism destination. *Sustainability, 10*(12), Article 4351. https://doi.org/10.3390/su10124351

APPENDIX A

Table 9: *Demand forecasting without weather features (MAPE)*

Forecast horizon	MAPE score		
	XGBR	Ridge regression	SVR
1	**.007**	**.007**	**.007**
5	**.017**	.020	.020
7	**.023**	.026	.024
14	**.042**	.046	.060

Table 10: *Demand forecasting without weather features (RMSE)*

Forecast horizon	RMSE score		
	XGBR	Ridge regression	SVR
1	**.007**	.008	**.007**
5	**.018**	.023	.022
7	**.024**	.029	.026
14	**.040**	.043	.051

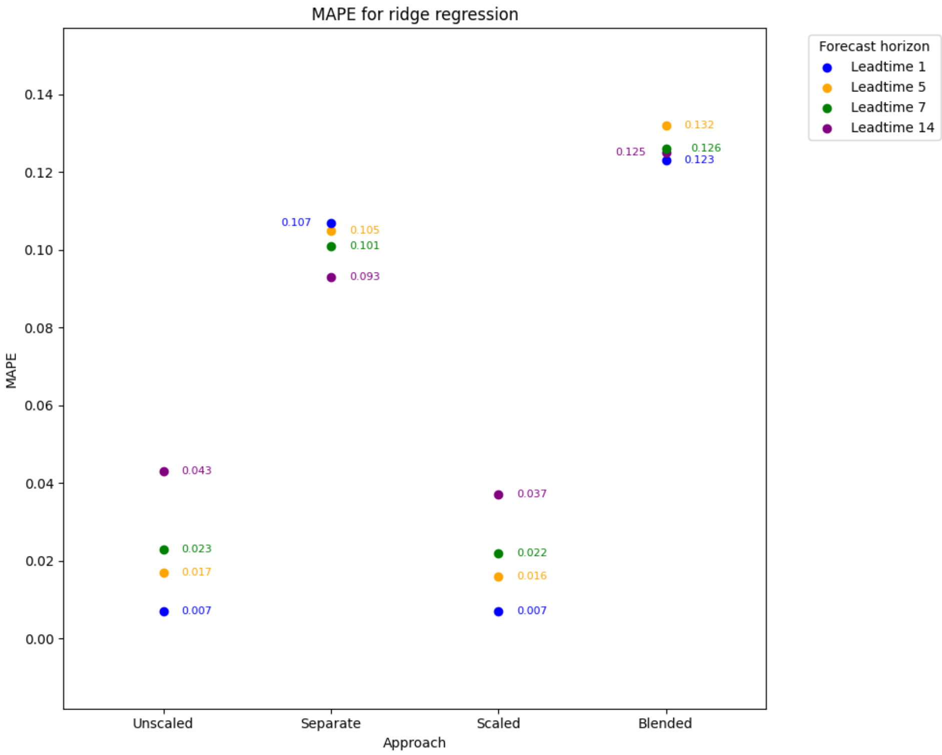

Figure 13: Experimental results for ridge regression

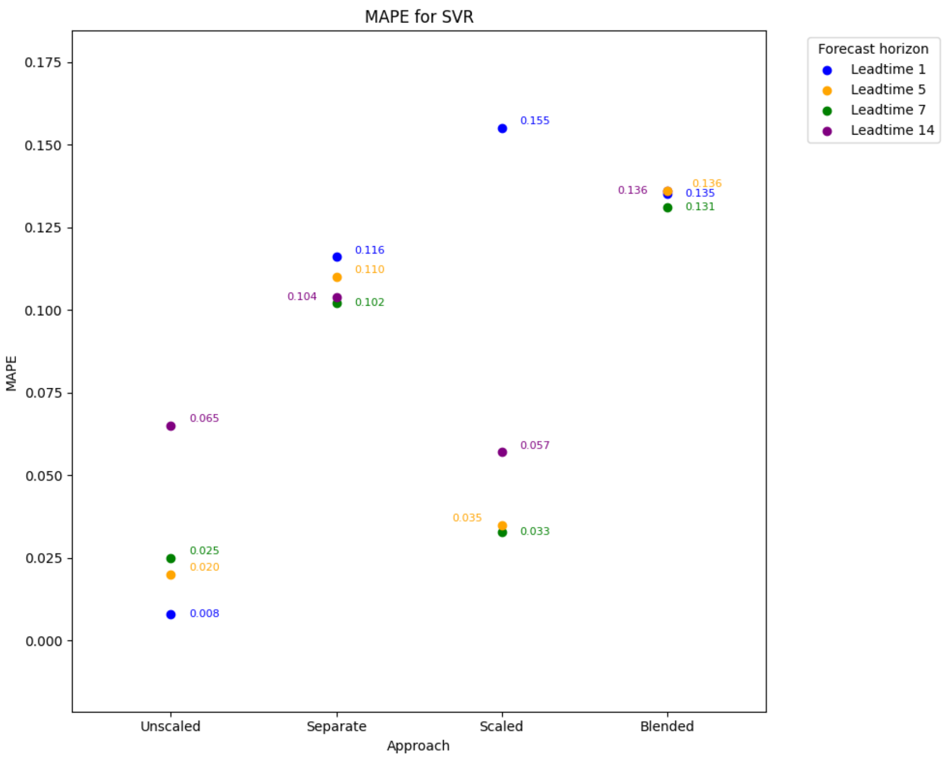

Figure 14: Experimental results for SVR

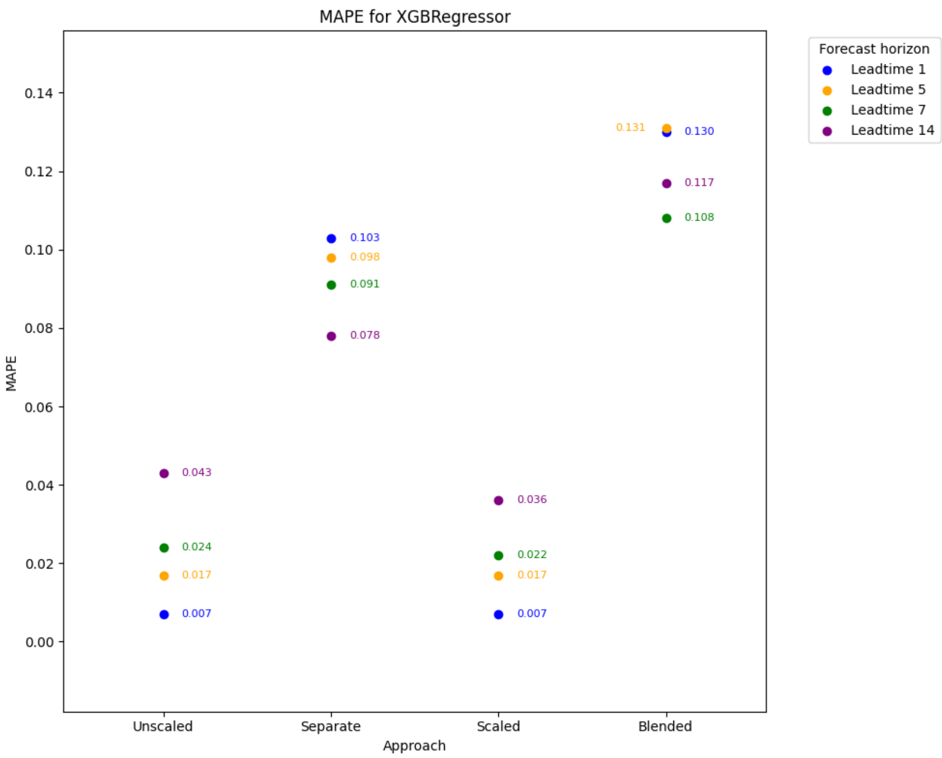

Figure 15: Experimental results for XGBR

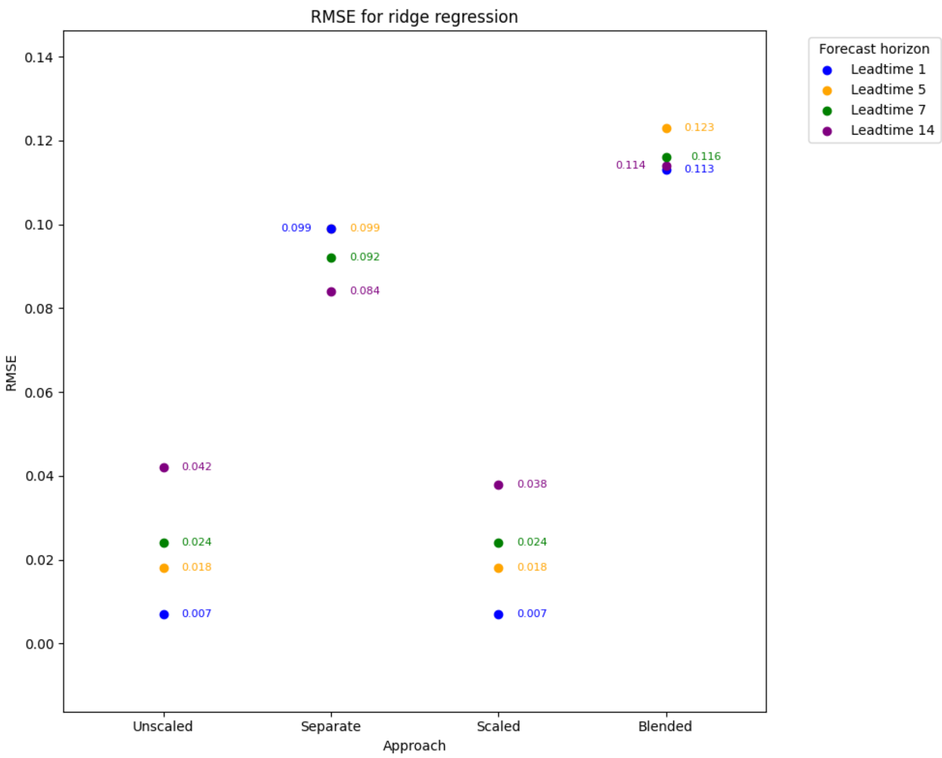

Figure 16: Experimental results for ridge regression

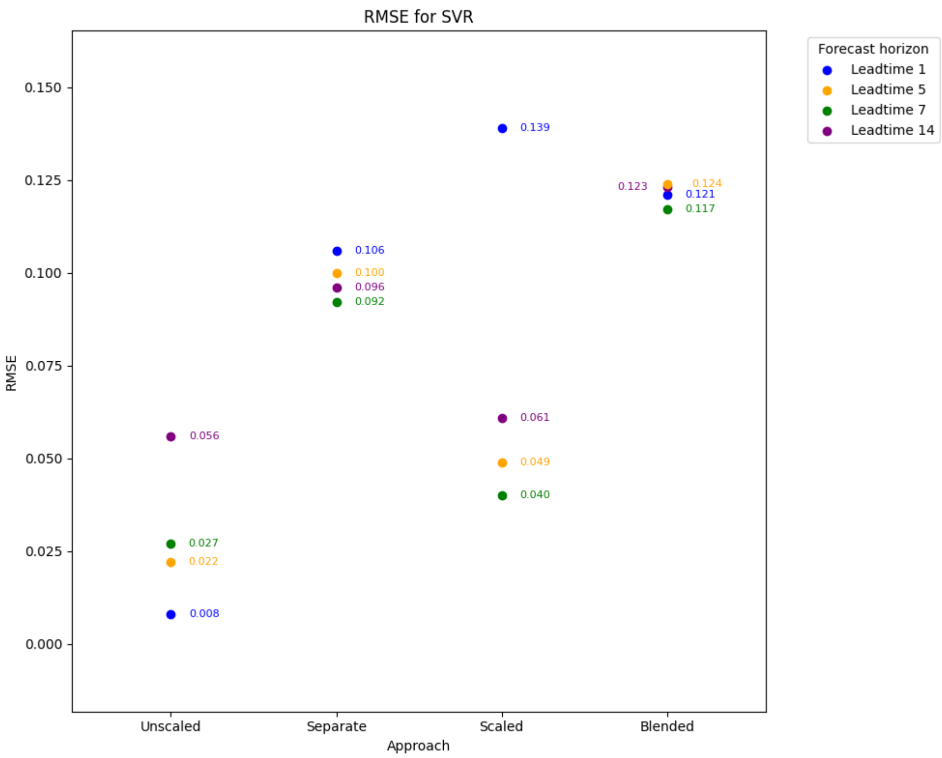

Figure 17: Experimental results for SVR

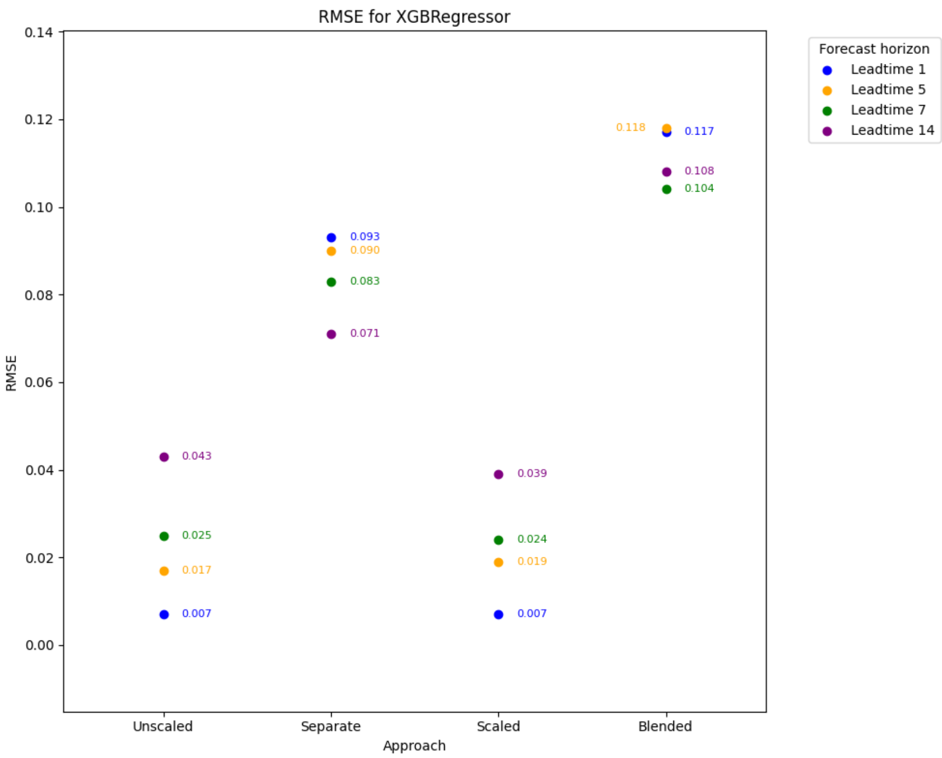

Figure 18: Experimental results for XGBR

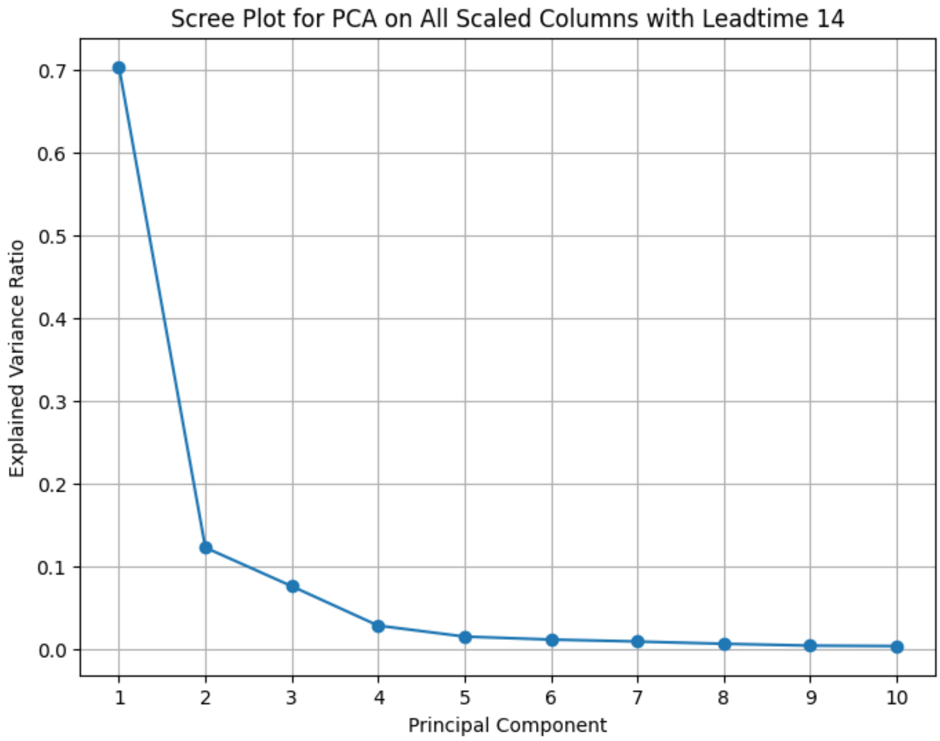

Figure 19: Scree plot for 14-day forecast horizon with all continuous features

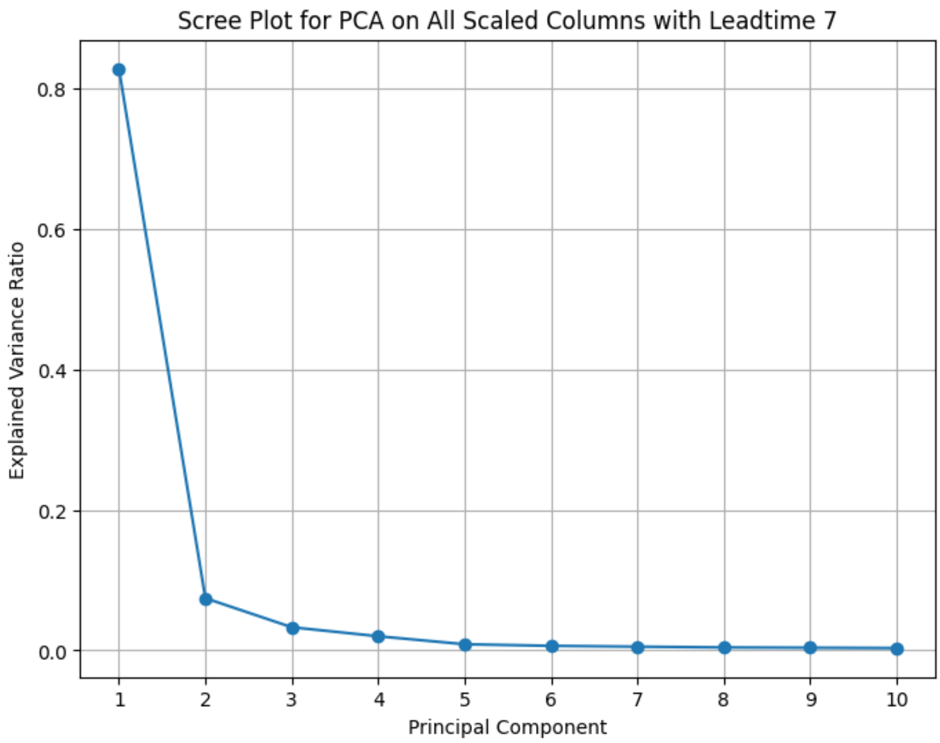

Figure 20: Scree plot for 7-day forecast horizon with all continuous features

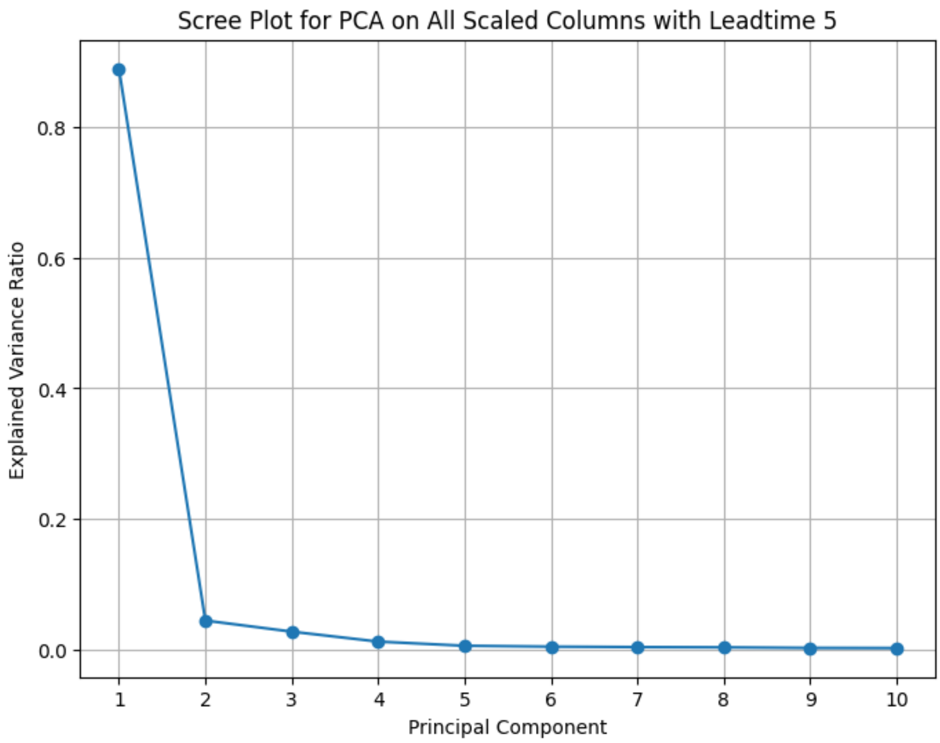

Figure 21: Scree plot for 5-day forecast horizon with all continuous features

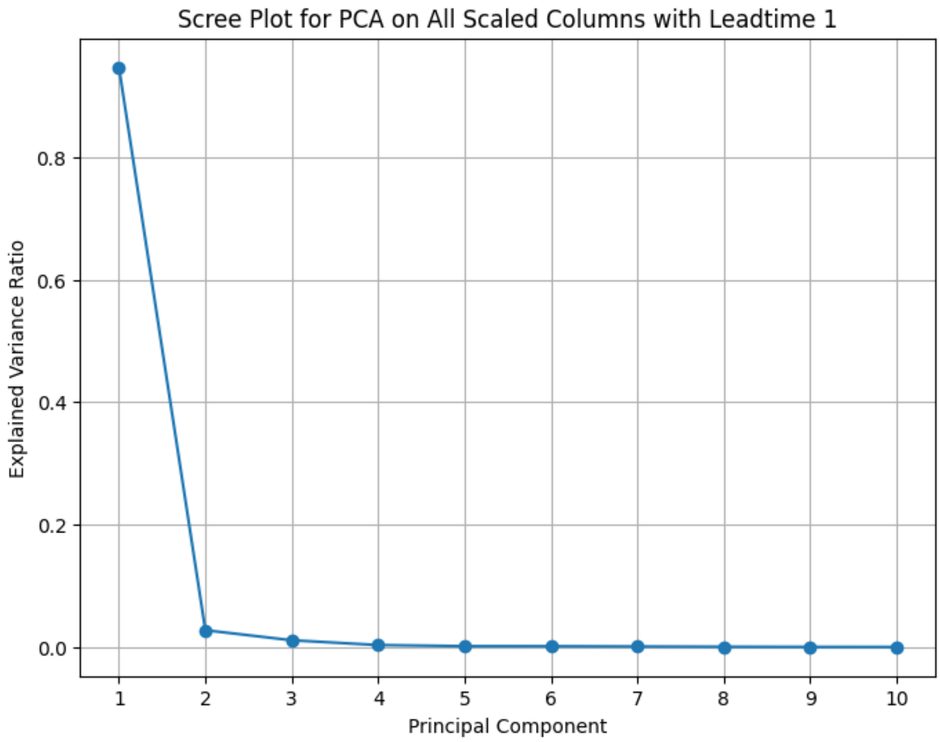

Figure 22: Scree plot for 1-day forecast horizon with all continuous features

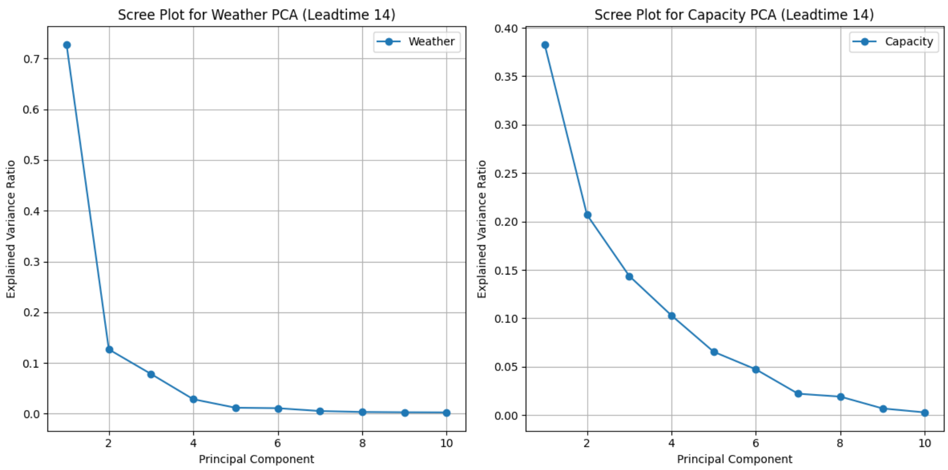

Figure 23: Scree plot for 14-day forecast horizon with weather and capacity separated

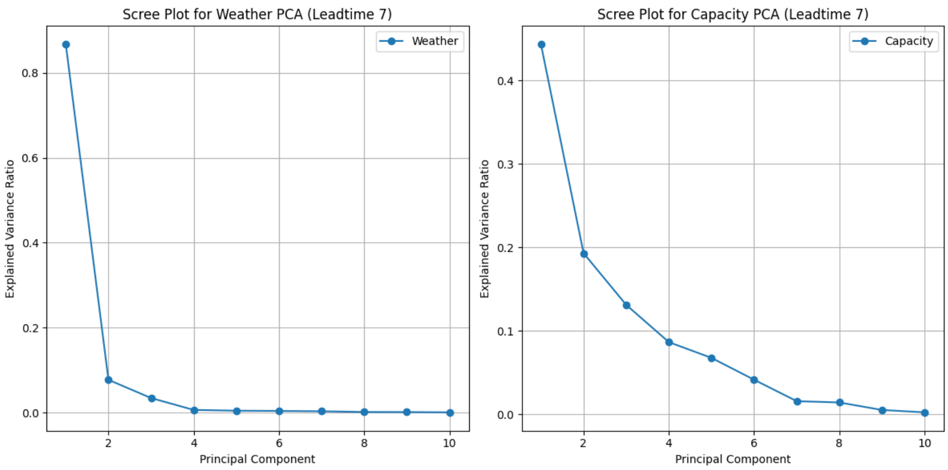

Figure 24: Scree plot for 7-day forecast horizon with weather and capacity separated

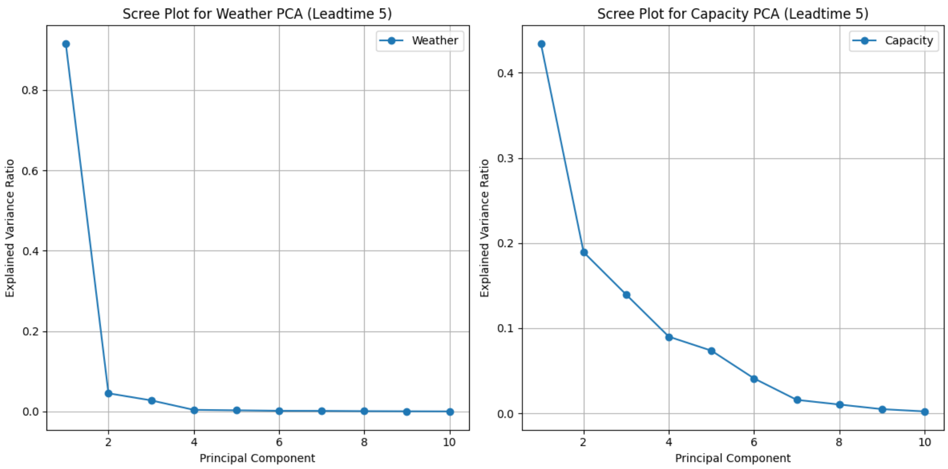

Figure 25: Scree plot for 5-day forecast horizon with weather and capacity separated

Table 11: *Variable list with definitions*

Variable	Definition
Occupancy	Target variable, calculated as the total number of rented villas divided by total physical capacity.
Holiday name	Indicates whether the day of visit is a holiday (1 = holiday, 0 = non-holiday).
Leadtime	Number of days between booking and the first day of stay (range: 0–16).
Min, Max, Avg temperature	Minimum, maximum, and average daily temperature in Celsius.
Min, Max, Avg rain	Minimum, maximum, and average daily rainfall in millimeters.
Min, Max, Avg fr_rain	Minimum, maximum, and average daily frozen rain in millimeters.
Min, Max, Avg snow	Minimum, maximum, and average daily snowfall in millimeters.
Min, Max, Avg ice	Minimum, maximum, and average daily ice accumulation in millimeters.
Min, Max, Avg snow_depth	Minimum, maximum, and average daily snow depth in millimeters.
Min, Max, Avg accumulated	Minimum, maximum, and average daily accumulated precipitation in millimeters.
Min, Max, Avg probability	Minimum, maximum, and average daily precipitation probability (0–1).
Min, Max, Avg rate	Minimum, maximum, and average hourly precipitation rate in millimeters.
Avg clouds	Average daily cloud cover percentage (0–100%).
Cancellation amount	Total number of cancellations for the day visit.
Cancellation insurance	Total number of cancellation insurances sold for the day visit.
Travel insurance	Total number of travel insurances sold for the day visit.
Booking day and Booking month	Day of the week and month when the booking was made (0 = Sunday, 6 = Saturday, 1 = January, 12 = December).
Day visit and month visit	Day of the week and month for the visit
Physical capacity	Physical capacity at the village level for hotel, comfort, premium, and VIP categories.
Rented villas before booking date	Total number of rented villas before the booking date and categories for hotel, comfort, premium, and VIP.
Cottage rent per person	Average price per person per villa night.

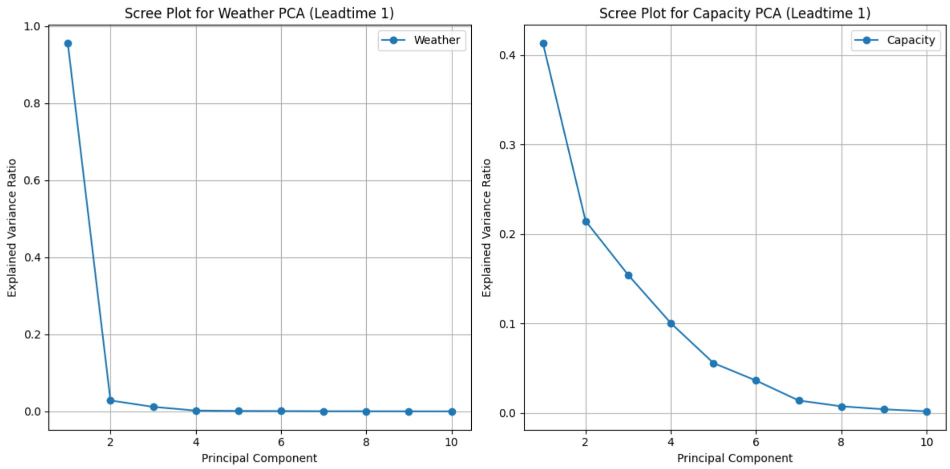

Figure 26: Scree plot for 1-day forecast horizon with weather and capacity separated

Table 12: *Python libraries used with short references*

Library	Creators
Matplotlib	Hunter (2007)
NumPy	Harris et al. (2020)
pandas	McKinney (2010)
scikit-learn	Pedregosa et al. (2011)
XGBoost	Chen and Guestrin (2016)

Table 13: *Software used with short references*

Software	Creators
Tableau Prep 2024.1.0	Tableau, (2024)
Python 3.12	Python Software Foundation, (2023)
Visual Studio Code 1.95.3	Microsoft (2024)
Overleaf TeX Live 2023	Overleaf, (n.d.)