Lumen Project Documentation

Boris Čuljak, Nenad Radović, Ivan Radman, Anja Kovačević
Faculty of Technical Sciences, Novi Sad, Serbia
{culjak.ra31.2020, radovic.ra18.2020, radman.ra68.2020, kovacevic.ra33.2020} @uns.ac.rs

Abstract

Predicting hotel occupancy plays an important role in hotel resource management. This paper aims to propose a one such method of designing a prediction machine learning model for the sole purpose of helping hospitality owners and resource managers in their day-to-day jobs. As such, the analysis proposed in this paper takes into account feature engineering, gathering events happening nearby the hotel of interest, as well as model engineering. Through conformal prediction, which represents a relatively new method accounting uncertainties of predictions (and thereby improving models), we present our solutions to mentioned problems.

Keywords: ***conformal prediction***, ***time series data***, ***hotel occupancy prediction***

1. Introduction

This project was developed as part of the Lumen data science competition, organized by Alfatec group and eStudent organization. We are grateful for the opportunity to participate and for the support provided by organizers and sponsors. This challenge not only allowed us to apply our skills in data science and software development, but also to contribute to resolving prediction problem in real world scenarios.

The hospitality industry is a cornerstone of the global economy, experiencing dynamic shifts influenced by a myriad of factors ranging from economic conditions to social trends. As such, precise forecasting of hotel occupancy rates has become paramount for operational and strategic planning. This research delves into the development of a predictive model designed to forecast hotel occupancy, leveraging historical data and machine learning techniques. The objective of this study is to enhance decision-making processes for hotel management, optimizing revenue management and improving guest experiences.

The complexity of occupancy prediction is underscored by the need to adapt to fluctuating demand patterns, which can be influenced by local events, seasonal trends, and market changes. Thus, the focus of this project is not only on achieving high accuracy in predictions but also on ensuring that the model can adapt and scale according to varying conditions and inputs. By integrating advanced data analytics and machine learning methodologies, this study aims to contribute a robust tool to the field of hospitality management, addressing both theoretical and practical aspects of occupancy forecasting.

This paper is organized as follows: after a comprehensive review of relevant literature and existing technologies, we introduce our methodological approach, followed by a detailed analysis of the dataset and features used. We then discuss the model development process, evaluation metrics, and results. Finally, the implications of our findings for the hospitality industry and areas for future research are explored.

2. Related Work

Recent advances in machine learning applications for hotel occupancy forecasting provide a compelling backdrop for our Lumen Project. The work by Kozlovskis et al. (2023) [4] serves as a significant point of reference. The paper mentions employing various machine learning algorithms, including bagged CART, bagged MARS, XGBoost, random forest, and SVM, to predict daily hotel occupancy rates. Then, models were tested against a traditional ARDL econometric model, which provided a robust comparison for evaluating the effectiveness of advanced machine learning techniques in real-world settings.

One notable finding from their study was the superior performance of the bagged CART model, which exhibited considerable predictive accuracy, although it did not outperform the ARDL model in all tested conditions. This highlights the potential and limitations of using machine learning for occupancy predictions in the hospitality industry, which is particularly relevant to our Lumen Project's focus on enhancing predictive analytics capabilities within the same sector.

Kozlovskis et al. (2023) [4] also emphasized the importance of integrating various data sources, such as weather conditions, public holidays, and online search trends (e.g.,

Baidu search index), into their predictive models. This approach aligns with our project's methodology, where we aim to incorporate diverse datasets to refine our forecasting models further.

Moreover, our research contributes to the field by implementing a hybrid model that combines traditional statistical methods machine learning techniques to provide more robust predictions, which are crucial for operational and strategic decision-making in the hospitality industry.

Overall, the insights from Kozlovskis et al. (2023) [4] validate the relevance of our research direction and underscore the importance of innovative data integration and algorithmic development in the ongoing enhancement of hotel occupancy forecasts.

3. Our approach to the problem

We begin our analytical approach by extensively exploring the dataset consisting of historical hotel occupancy records. This initial phase of data exploration is crucial for understanding underlying patterns, detecting outliers, and identifying key factors influencing occupancy rates. Following this, we construct a comprehensive event table that captures significant events and public holidays in Croatia, which are known to affect tourist inflows and, consequently, hotel occupancy. This event table is meticulously integrated with the historical occupancy data to enrich the original dataset. The augmented dataset is then structured to align with **time series forecasting methods**(ili koji vec approach), allowing us to systematically analyze and predict future occupancy trends. This technique ensures that our model not only leverages historical data but also dynamically incorporates the impact of scheduled events and seasonal variations, providing a robust foundation for accurate occupancy forecasting.

3.1. Data exploration

Initial examination of the dataset provided by Alfatec revealed a structured framework suitable for in-depth analysis. The preliminary dataset consists of various attributes relevant to hotel bookings. A representative table of the dataset structure is presented below (Table 1), outlining key columns and sample data entries to illustrate the dataset's composition at a glance.

Column Name	Data Type	Description
reservation_id, guest_id	Integer	Unique IDs for reservation and guest
night_number, room_cnt	Integer	Number of nights and rooms
stay_date, date_from, date_to	datetime64[ns], Object	Dates related to the stay
guest_country_id, reservation_status	Object	Country ID and reservation status
price, price_tax, total_price	Float	Pricing details including taxes
room_category_id, sales_channel_id	Integer, Float	Room category and sales channel

Table 1. Dataset Structure

The dataset was thoroughly inspected for completeness, revealing no missing values (`None` or `NaN`) in critical columns, ensuring a robust foundation for further analysis. Duplicate entries and any negative values, particularly in columns quantifying the number of people, were meticulously checked and corrected to maintain data integrity.

A noteworthy observation was that the dataset predominantly contained bookings from solo travelers or pairs, typically adults, with five reservations including a child, suggesting the hotel's primary focus on business conferences, adult leisure, or senior excursions. Understanding the geographical origins of our guests is crucial, particularly for optimizing event table — a topic we will explore in greater detail later in this section. Presented below (Table 2) is a summary of the top five countries from which our hotel guests are visiting, even though there are guests from 66 countries in total.

Country Code	Reservations
HR	16061
I	6041
F	2057
GB	1712
NL	1052

Table 2. Countries with Most Reservations

To refine the dataset for effective model training and testing, all cancelled reservations and their associated data were removed. This cleanup was facilitated by dropping several non-essential columns, enhancing processing efficiency and model focus: `'guest-id'`, `'resort-id'`, `'price'`, `'price-tax'`, `'total-price'`, `'total-price-tax'`, `'food-price'`, `'food-price-tax'`, `'other-price'`, `'other-price-tax'`, `'sales-channel-id'`

Further, data from the year 2007 was excluded due to its incompleteness, with analyses confined to the years 2008 and 2009. This period was defined with a minimum stay date of `2008-01-01` and a maximum of `2009-12-31`. The unique room category IDs present were [4, 5, 7, 2, 6, 3, 11, 1], indicative of diverse accommodation offerings.

A seasonal analysis was conducted, supported by visualizations such as the total rooms booked per month (Figure 1) and a Fourier decomposition (Figure 2). These visual insights confirmed the seasonal demand variations, aligning with expectations of booking increases during spring and summer, tapering off during other months.

Conclusion of the seasonal analysis was that data was periodic by year.

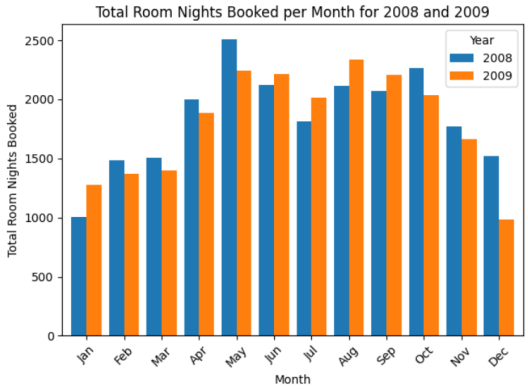

Figure 1. Total rooms booked per month for the years 2008 and 2009

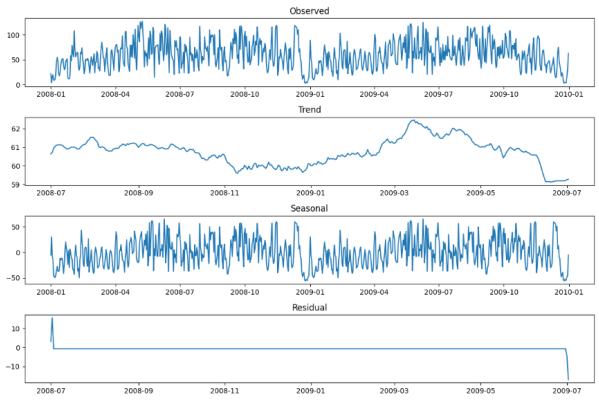

Figure 2. Analysis of room bookings

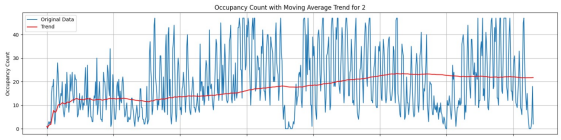

Figure 3. Analysis of one room booking

Trend was analyzed for each of the rooms (Table 3). The seasonal decomposition graph (Figure 2) shows the breakdown of a time series into trend, seasonal, and residual components. Here's an analysis of each component based on the plots:

Observed: This is the actual data we are analyzing. The plot shows fluctuations over time, and while there are patterns, it's not immediately clear what causes these fluctuations just from the observed data.

Trend: The trend component shows a slight upward trajectory followed by a downward slope later. It indicates an initial increase in room nights booked, reaching a peak, and then a gradual decrease. There might be many reasons for this, including changes in the hotel's popularity, competition, or general market conditions. The sharp decline at the end suggests a significant event or change in conditions that may need further investigation.

Seasonal : The seasonal plot displays consistent and regular patterns within each year, which indicates a clear seasonality in the data. These peaks and troughs repeat at similar intervals, suggesting that the hotel experiences regular periods of high and low demand, which could be tied to holiday seasons, local events, or other cyclical factors.

Residual: Residuals are the left-over part of the data after the trend and seasonal components are removed. Ideally, residuals should show random noise if the model has captured all the patterns. The residual plot here shows some variability, which could be due to random fluctuations, or it might indicate additional patterns that the seasonal and trend components haven't captured. The spike at the end could be due to an outlier or a specific event not accounted for by the seasonal or trend components.

3.2. Dataset creation

3.2.1 Event table

The event table is a pivotal component of our dataset, providing a comprehensive record of historical events that have a direct impact on hotel occupancy rates.

Event table attributes

- Event Name: The name of the event.
- Type: Possible types of events, which include cultural holiday, music festival, theatre festival, carnival, conference and sport event.
- Start Date and End Date: The duration of the event, indicating when it begins and concludes. These dates are crucial for aligning events with occupancy data.

Although original event table was written in JSON, here (Table 3) we present a table example of one event from the data.

Event Name	Type	Start Date	Finish Date
Uskrs	Cultural Holiday	2008-03-23	2008-03-26
Uskrs	Cultural Holiday	2009-04-12	2009-04-14
Uskrs	Cultural Holiday	2010-04-04	2010-04-06
Uskrs	Cultural Holiday	2011-04-24	2011-04-26
Rijecki karneval	Carnival	2008-01-17	2008-02-05
Rijecki karneval	Carnival	2009-01-17	2009-02-24
Rijecki karneval	Carnival	2010-01-17	2010-02-14
Rijecki karneval	Carnival	2011-01-21	2011-03-06

Table 3. Event Data Example

Importance of event table. The event table serves as a key predictor in our occupancy forecasting model. Historical data on events allows us to analyze patterns in how different types of events influence hotel demand. For instance, larger events such as national conferences or international sports events typically lead to significant spikes in occupancy rates due to the influx of attendees. Conversely, smaller local events might have a more subdued impact.

By integrating the event data with our occupancy records, we can refine our predictive models to account for the variability introduced by these events. This integration enables the model to anticipate occupancy surges and dips effectively, thus allowing hotel managers to strategize their room pricing, staffing, and other logistical aspects more efficiently.

3.2.2 Weather forecasting data

We also explore the potential impact of weather conditions on hotel occupancy rates. Recognizing the influence of weather on tourist activities and preferences, we incorporate historical weather data into our study. This data is gathered from Visual Crossing Weather [1], which provides detailed weather reports for Rijeka, Croatia where the hotel from our data is situated. By integrating this weather data with our hotel occupancy records, we aim to uncover any significant correlations between weather patterns and occupancy trends.

Although weather data was not included in the final version of our dataset, we believe it is important to acknowledge its potential relevance. The impact of weather conditions on hotel occupancy may vary significantly across different locations and types of accommodations. Therefore, while it was not a determinant factor for our current analysis, weather data could prove to be more significant in studies involving other hotels or regions, offering valuable insights for tailored marketing and operational strategies.

4. Data Integration

Regarding integration and selection of data to be used as features for the model, there are several categories that have been used, and they are explained below.

4.1. Events

Events were filtered depending on the room type. Although the room type may not seem significant at first glance for event selection, the nature of the data and the fact that there are significant differences in occupancy rates among different room types and periods of occupancy lead to the necessity of filtering. Filtering was done by retaining only those events for which the occupancy was 1.5 times higher than the average value of the previous month. This ensured that only events showing a significant increase in occupancy were retained for future predictions. File containing selected events is in application.

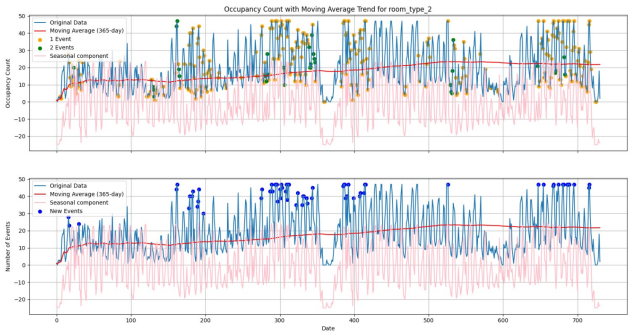

Figure 4. Analysis of room bookings

4.2. Statistical

Regarding statistical data, the indicators that have proven to be good for predictions are the mean values for the day of the week, mean values for the month, and day of the week. By integrating these, the model is fed with data about which days of the week are more important than others, as well as which months are more significant than others, introducing a monthly seasonal component into the data.

4.3. Lagged data

The lagged data is indeed the most crucial data used for predictions. Data on occupancy from the previous seven days was utilized, with each day serving as an individual input. What was additionally incorporated is the average occupancy value of the past seven days, serving as an additional statistical feature derived from the lagged data.

5. Experiments, Training and Evaluation methodology

Since long-term predictions cannot easily be made using traditional methods for solving time series problems due to lagged values [3], classic models were created to predict the next 7 days, along with a separate statistical model designed to predict any period and used for long-term seasonal forecasting.

Separated models were trained for each room type, multiple models were tested and evaluated for each dataset, resulting in different datasets ending up with different models.

5.1. Statistical data

Dataset for training of the statistical model is shown in (Table 4).

The `scaled_room_id` attribute is calculated to normalize the reservation counts across different room types. This is done by grouping rooms under

room_category_id and summing their reservations (room_cnt) for each category. The highest reservation count among these categories is used as a base to scale each category's reservations. The total reservations per category are divided by this maximum and rounded to three decimal places, producing a scaled_room_id that ranges between 0 and 1. A value of 1 indicates the category with the highest reservation activity. This normalization facilitates comparative analysis across room types by adjusting for their varying popularity.

Attribute	Description
room_cnt	The number of rooms booked
day_of_week	Day of the week as an integer (e.g., 1 for Monday)
day_of_year	Day of the year, ranging from 1 to 365/366
scaled_room_id	An identifier for each room, scaled between 0 and 1
isEvent	A binary indicator (0 or 1)

Table 4. Description of Dataset Attributes

5.2. Training - Time series data

Dataset for training of separated non statistical models is shown in next table:

Attribute	Description
day_of_week	Day of the week as an integer (e.g., 1 for Monday)
week_day_avg	Average value of occupancy for that day
month_avg	Average value of occupancy for that month
week_day_importance	Inverse rank of that day importance based on occupancy
event	Indicator for events
occupancy_lag_1	Occupancy of day before
occupancy_lag_2	Occupancy of day before 2 days
occupancy_lag_3	Occupancy of day before 3 days
occupancy_lag_4	Occupancy of day before 4 days
occupancy_lag_5	Occupancy of day before 5 days
occupancy_lag_6	Occupancy of day before 6 days
occupancy_lag_7	Occupancy of day before 7 days
mean_last_7	Mean value of occupancy in last 7 days

Table 5. Description of Dataset Attributes

The training and testing data were divided in an 80-20 ratio. Naturally, considering the nature of the problem, the training data comprises information up to a certain date, while the testing data is after that date. Seven columns representing occupancy for the next 7 days were set as the target. The following models were trained and later in model selection part, the one with best metrics was selected:

- Linear Regression
- Ridge Regression
- Gradient Boosting
- XGBoost (Extreme gradient boosting)
- Random Forest
- SARIMA

5.3. Evaluation metrics

As metrics for comparing models, the mean squared error and mean absolute error were used. Additionally, R-squared was considered in the analysis, but it wasn't used for selection; rather, it served more as an indicator during testing.

In addition to their clear basic functions, interpreting the values of metrics was crucial in making decisions about further work in this problem. After training and testing the initial model that predicted occupancy for all rooms together, the metric values seemed impressive. However, due to significant differences in occupancy values among rooms, the metrics gave a false impression of good predictions, as explained below.

Rooms with a small number of guests were mostly empty, with only a few occupied, while for some room types, there were on average several dozen occupied rooms. These empty rooms were generally well predicted, thus reducing the average errors for room types with higher occupancy. For these reasons, although the predictions were not accurate, the metrics indicated good results. This is why it was decided to use different models for different room types.

5.4. Results and model selection

After training each model for every dataset, the conclusion was easily drawn. For rooms with generally low occupancy, a linear regression model was used, while for others, a random forest model was employed. The results are shown below (Table 6 and Table 7).

Room type	Model	Mean Squared Error
1	Linear Regression	0.020408
2	Random Forest	95.88046
3	Random Forest	29.15257
4	Linear Regression	1.2031
5	Random Forest	377.96282
6	Random Forest	9.98347
7	Linear Regression	0.47619
11	Linear Regression	0.03236

Table 6. Training metrics - Time series data

Regarding metrics, cross-validation was further conducted on the selected models, followed by comfortable prediction afterward.

5.5. Conformal Prediction

Conformal prediction [5] is a relatively new method aimed at improving machine learning model predictions with calculating (un)certainties of predictions. In the context of time series problems, the goal is to obtain a

Room type	Model	Mean Squared Error
1	Linear Regression	0.70200
2	Random Forest	72.75144
3	Random Forest	44.3836
4	Random Forest	1.21868
5	Random Forest	312.36123
6	Random Forest	10.53876
7	Random Forest	0.29298
11	Random Forest	0.214151

Table 7. Training metrics - Statistical data

range of values within which the model's prediction falls, in addition to the specific predicted value. Conformal prediction models are obtained by taking a previously trained and optimized model and conducting additional training to determine the range of values. Visualization for easier understanding is presented in the following graph.

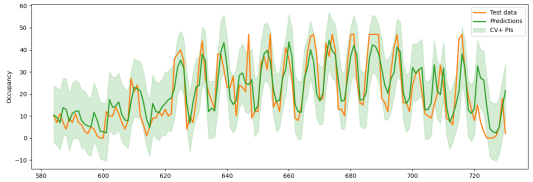

Figure 5. Predictions with conformal prediction

For comfortable prediction, the MAPIE library (Model Agnostic Prediction Interval Estimator) [6] was used, which provides a straightforward determination of value ranges. After training the "regular" models, MAPIE models were trained as an enhancement to determine the value ranges.

6. Conclusion

Although the prediction results are not at the highest level, the models manage to predict occupancy values and can anticipate normal fluctuations, with exceptions at extreme values, which generally pose a significant challenge in the field of machine learning and data science. Of course, it should also be considered that the problem is not such that it's crucial to predict the exact value, but rather to observe general fluctuations caused by seasonality and events over time, to help people working for hotel.

For short-term predictions, time-series models are forecasting occupancy with certainty boundaries derived from conformal predictions, while statistical methods can be employed for long-term predictions, subsequently utilized for seasonal and yearly planning. Naturally, with more data available and further analysis, model performances could be improved, so that's definitely something to consider.

The application is designed to be intuitive and easy to use. Users are only expected to input data and select prediction dates. Therefore, it is suitable for all individuals, with no restrictions on who can use it, from professionals to those with no experience in the field.

7. Future work

In future iterations of our hotel occupancy prediction model, we plan to enrich our dataset by incorporating search trend data from Google. Specifically, we aim to analyze the volume and frequency of search queries related to travel and accommodations in Croatia. This addition is expected to provide valuable insights into when potential tourists are most interested in visiting the region, which directly correlates with increased demand for hotel accommodations. By integrating Google Trends data, our model can better predict spikes in occupancy tied to rising online interest, allowing hotel managers to proactively adjust their offerings. This enhancement will not only improve the accuracy of our predictions but also enable a more dynamic response to market trends, ultimately contributing to optimized revenue management and improved guest satisfaction.

However, it is essential to note that our current dataset, encompassing the years 2008 and 2009, predates the widespread adoption of modern internet-based predictive tools and behaviors. This temporal gap means that while the proposed approach of integrating Google Trends data is well-suited for contemporary datasets, its application to our current dataset would be anachronistic and likely not yield accurate or meaningful insights. For truly effective implementation of this modern approach, it would be necessary to acquire more recent data that reflects the current dynamics of internet usage and search trends in the context of travel and hospitality.

For the future iterations, we would also like to propose a conduction of survey among hotel residents, to enrich our search query data by simply questioning them about the reasons they came, how did they found out about hotel, about the city, about the event etc. This will lead us to better understanding of what data we're looking for, as well as better usage of *Google Trends API* [2].

Moving forward, securing updated data will be a critical step in ensuring that our model remains relevant and capable of leveraging the latest advancements in data analytics and online behavior analysis.

References

[1] Visual Crossing Corporation. Visual crossing. 4

[2] Google. Google trends. 6

[3] Rob J Hyndman and George Athanasopoulos. *Forecasting: principles and practice*. OTexts, 2018. 4

[4] N. Lace K. Kozlovskis, Y. Liu and Y. Meng. Application

of machine learning algorithms to predict hotel occupancy. *JBEM*, 24(3):594–613, 2023. 1, 2

[5] Ryan Tibshirani. Conformal prediction. 5

[6] Nicolas BRUNEL Issam IBNOUHSEIN François DE-HEEGER Vianney TAQUET, Grégoire MARTINON and MAPIE contributors. MAPIE: Model-Agnostic Prediction Interval Estimator. https://github.com/scikit-learn-contrib/MAPIE, 2021. 6