Multi-layered market forecast framework for hotel revenue management by continuously learning market dynamics

Rimo Das¹ · Harshinder Chadha¹ · Somnath Banerjee¹

Received: 10 February 2020 / Accepted: 18 August 2020 / Published online: 30 March 2021
© The Author(s), under exclusive licence to Springer Nature Limited 2021

Abstract

With the rising wave of travelers and changing market landscape, understanding marketplace dynamics in commoditized accommodations in the hotel industry has never been more important. In this research, a machine learning approach is applied to build a framework that can forecast the unconstrained and constrained market demand (aggregated and segmented) by leveraging data from disparate sources. Several machine learning algorithms are explored to learn traveler's booking patterns and the latent progression of the booking curve. This solution can be leveraged by independent hoteliers in their revenue management strategy by comparing their behavior to the market.

Keywords Revenue management · Forecast · Unconstrained demand · Constrained demand · Market · Machine learning

Introduction

The revenue management process has three core principles: pricing, revenue management, and product distribution (Vinod 2004). All three principles should be in synergy and work together in a unified manner to maximize revenue and profits. In order to harmonize these three principles and a hotel's financial goals, accurate forecasting is critical for decision making.

In the last decade, with the maturity of the internet, economic growth and shift in consumer behavior have introduced millions of new travelers. The TripAdvisor and Oxford global travel market study stated that in 2017, the travel industry is valued at 5.29 trillion USD, and the global travel market has grown 41.9% in the last 10 years (Oxford 2018).

It is essential for hoteliers to broaden their revenue strategy to respond to these dynamic changes in the hospitality industry. In economics, a market is a place where buyers and

sellers come together to exchange some products or goods (Blackmar 1912). In the hotel industry, the idea of a market can be conceptualized by defining a geographical area, where sellers are an arbitrary set of hotels, and buyers are travelers looking for accommodation. Not only can a market be defined by the geographical area, but it can also be defined by a combination of hotels in the same star category having similar pricing or reservation trends.

A vast majority of legacy hospitality systems are dated and limited in their interaction capabilities, making the transfer of clean data challenging from the onset. Beyond system intelligence issues, revenue managers face the challenge of seemingly endless new streams of data being accessible ranging from market fluctuations, competitor pricing to impactful events, varying air traffic, and even minute factors like changing weather that impact demand and revenue. Managing these technical complications and balancing a near-limitless array of data points are not feasible to do manually in a timely and cost-effective way. Traditionally, revenue management forecasts are rigid which have tended to operate largely in broad strokes (business rules) or are otherwise based on conventional modeling techniques which often lack optimal predictive efficiency. These methods are primitive and inaccurate because they do not capture the exogenous factors affecting the hotel, nor are they capable of handling very large datasets that are exploding in volume and velocity. Here, a machine learning architecture is

✉ Rimo Das
rimo@lodgiq.com; rimo@outlook.com

Harshinder Chadha
harshinder@lodgiq.com

Somnath Banerjee
somnath@lodgiq.com

¹ LodgiQ, 440 N Wolfe Rd, Sunnyvale, CA 94085, USA

introduced that more accurately forecasts by capturing the hidden patterns in the massive data.

Literature review

There has been a tremendous amount of research in demand forecasting in the field of revenue management. Rajopadhye et al. (2001) built an algorithm to forecast the unconstrained hotel room demand. The forecasting algorithm in the article was constructed by using two inputs: short-term and long-term forecasts. The long-term forecast uses a Holt–Winters model to forecast the room demand, and the short-term forecast uses the advanced reservations, cancellation rate, the net turndowns, and net demand booking profile. In Weatherford and Kimes (2003), the authors describe three types of forecasting methods: historical, advanced, and combined booking models (a combination of historical and advanced) for revenue management. Furthermore, the authors explain the issues in revenue management forecasting citing level of aggregation, forecast accuracy, breadth of data, and outliers. In the paper (Zakhary et al. 2009), the authors proposed a Monte Carlo simulation for forecasting hotel arrivals and occupancy. In their research, they address the revenue management forecasting issues (Weatherford and Kimes 2003) detailed by modeling the effects that seasonality, reservations, cancellations, length of stay, group reservations, and trends have on hotel arrivals and occupancy. The authors use hotel data to forecast arrivals and occupancy at an aggregated daily and weekly level (Zakhary et al. 2009). All of the research written above is focusing on demand forecasting at the property level. For this specific topic in market forecasting, there has been some research on how a market affects revenue management and tourism. In the paper (Sigala 2015), the author puts forward a “learning with the market” framework in revenue management and conducts several focus groups with revenue management professionals to ascertain the viability of a market-driven framework. Several concepts such as market dynamics, market engagement, and social media are explored to determine the impact on tourism markets (Sigala 2015).

In the paper (Cooper et al. 2006), the authors present an analysis of the underlying errors in revenue management models. The authors state that in practice, assumptions may be incorrect and model parameters are unknown. Moreover, in their research, they detail how controls based on erroneous assumptions can lead to poor forecasting results because the historical data depend on past protection levels, which inevitably leads to modeling errors. Cooper et al. (2006) proposed an iterative approach where assumptions and model parameters are updated as the observed data are collected and reflected in the forecasting-optimization process. Group forecasting accuracy has been an important area of research

within demand forecasting. Kimes (1999) researched group forecasting accuracy for hotels and wrote about the challenges in group forecasting. The author found that inaccurate group forecasting has more impact during high occupancy days and noted that the transient segment is severely impacted because the willingness to pay is higher for the transient segment. Through the analysis, Kimes found accuracy can be affected by (not limited to lead time), type and size of the hotel, dependence on group business, frequency of forecasting updates, or occupancy of the hotel. This paper aims to combine demand forecasting concepts, a market-driven approach to revenue management leveraging machine learning to build a model that forecasts market demand.

Research methodology

In this section, the overall objective and approach are introduced along with the factors that impact the machine learning framework. First, multi-textured data such as market reservations (powered by Amadeus Business Intelligence), market pricing records, and events are collected from different sources. Thereafter, these data are explored to understand the characteristics of the data. Then, various factors are considered which impact a market and would be crucial for the machine learning algorithms to be trained on them. After that, a set of machine learning algorithms are used to develop the market forecast model.

Data categories

In the previous sections, it was stated that a market is defined as a collection of hotels in a geographic region. For the purpose of this research, a total of 97 hotels are chosen to illustrate the New York City market with a total capacity of 38,745 rooms. The most critical metrics considered at the market level that impact daily demand are given in Table 1:

Table 1 Market KPIs

Metric	Definition	Expression
Capacity	Total available rooms in the market	$\sum \text{capacity}$
Occupancy	Percentage of occupied rooms in the market	$\frac{\sum \text{Rooms sold}}{\text{Capacity}}$
ADR	The average price paid per room in the market	$\frac{\text{Revenue}}{\text{Rooms sold}}$
RevPAR	The average daily rooms revenue generated per available rooms in the market	$\frac{\text{Revenue}}{\text{Capacity}}$
Revenue	Sum total revenue from the individual hotels comprising the market	$\sum \text{revenue}$

The purpose of the market forecast model is to forecast the unconstrained and constrained demand on a daily basis, in a defined market (aggregated and segmented).

Seasonality

Butler (1998) defines seasonality as “the temporal imbalance in the phenomenon of tourism” and that it can be determined by various factors such as the number of visitors, expenditure of visitors, transportation disturbances, and attractions. In “The Measurement of Seasonality and Its Economic Impacts” (Bar-On 1999), the author describes seasonality as a phenomenon that can be defined as effects occurring each year at a similar time span and magnitude, and these effects can be but are not limited to climate patterns, festivals, events, and local holidays. Figure 1 displays the occupancy (number of visitors who visited the New York City Market/capacity of the market) during the years, 2017 and 2018. From the graph, the cyclical nature of the occupancy is an example of seasonality.

For a deeper view of the seasonality, Figs. 2 and 3 display the average occupancy by month and day of the week in the New York City market during 2017 and 2018. From Fig. 2, January and February are the softest periods, and the occupancy progressively climbs in the following months. Similarly, in Fig. 3, the occupancy on Sunday and Monday is significantly lower compared to the rest of the week

where higher demand can be exhibited. Even though these visitor trends are specific to the New York City market, the seasonality effects can be observed for any defined markets.

Market segments

Segmentation is defined as “the process of dividing consumers into distinct groups with similar sets of needs and wants to require the development of marketing mixes by the hotel” (Ivanov 2014). Identifying the right customer is a very important aspect of revenue management. To that end, hotel revenue managers must identify the booking patterns of a potential segment and then determine if the desired pricing strategy can achieve the hotel’s revenue goals.

At the market segmentation level, market data (powered by Amadeus Business Intelligence) are broadly classified into the following market segments:

1. Discount—customers from advanced purchase, Online Travel Agencies (OTAs), promotional/package rates, and frequent guest program redemptions.
2. Retail—customers from rack rates, best available rates, non-discounted, non-affiliated, and non-contracted customer demand.
3. Negotiated—corporate negotiated and consortia rates.
4. Qualified—customers from qualifying affiliation discounts, such as AAA, senior discounts, and employees.

Actual Occupancy in New York Market

Fig. 1 Actual occupancy from 2017 and 2018

Fig. 2 Monthly seasonality

Fig. 3 Weekday seasonality

5. Wholesale—customers from wholesalers, consolidators, and tour operators.
6. Group—customers from group tours, domestic and international groups, association, convention, and corporate groups.

In Fig. 4, a plot of the booking pace can be seen of all these segments.

From Fig. 4, the booking trends for the various market segments can be seen. The Discount and Retail segments

have an inflection point where the Discount segment overtakes the Retail segment.

Group segment

Among the several market segments, the Group segment is a highly differentiated segment as the trends are varied from the regular demand. The Group market segment can be split into two distinct categories: **Group-Sold** and **Group-Committed** segments. The Group-Committed segment

Fig. 4 Market segmented pace for 12/31/18

reflects customers who have signed a contract with the intent to commit to a block of rooms and have officially paid the terms of the contract. The Group-Sold segment reflects the customers who have reserved rooms but have not signed a contract or paid for the rooms. As the segments get closer to the lead time, the demand from the Group-Sold segment transfers over to the Group-Committed segment.

The difference between the final occupancy from Group-Sold and the Group-Committed blocks is known as the group

wash. From Fig. 5, it can be seen that as we move closer to the stay date, the demand block moves from sold segment to committed, while there remains an unused demand block or the group wash. These characteristics make that the segment behave a little more erratically in comparison to the other market segments. Moreover, the effect of seasonality causes a multiplicative effect.

Fig. 5 Group booking pace for 24/09/19

Holidays and events

In tourism and the lodging industry, in particular, seasonality has been the focus of several studies (Butler 1998; Bar-On 1999). In fact, the term “holiday,” within this context of seasonality in tourism, has typically assumed a period spanning several days, called “Shoulder Days”, (even weeks or months) compared to the shorter (single day) application in finance and economics (Demico et al. 2006).

Holidays introduce a year-to-year shift in arrivals when more than a month is involved. This occurs when travel related to the event or holiday takes place during the days immediately before and/or after the holiday and these days belong to the month prior to or just after the month of the holiday itself (Bureau of Transportation Statistics). Holidays in the US such as Labors Day, Easter, Memorial Day, Thanksgiving, and Christmas also introduce year-to-year variation when they affect more than one month (Bureau of Transportation Statistics).

From Fig. 6, the number of travelers shows a decrease from the Tuesday (November 20, 2018) before Thanksgiving Day (November 22, 2018) and remains high through Thanksgiving Day and Black Friday (November 23, 2018), and falls on Sunday (November 25, 2018). The days before Thanksgiving Day often fall in November while days after that occur in December. Thanksgiving Day, therefore, has the potential to impact both November and December (Fig. 7).

Events also potentially impact more than one month, because it may occur in one month this year but in another month the next year (DreamForce at San Francisco, for

instance). Moreover, there is also a change of locations of events, (SuperBowl, for instance) because it may occur in one location this year but in another location the next year. These kinds of fluctuations should be included in the model to control for the year-to-year variation which local holidays and events induce in seasonal movements.

During these days, the market demand has a higher deviation than the other normal days. This is critical as they are typically interested from a business perspective, to tailor strategies structured around the emerging opportunities (or lack thereof).

Span of data and sample size

The multi-textured data are divided into two groups. Figure 8 shows data separation for training and testing of the machine learning models (Fig. 9).

Machine learning

In this research, several machine learning models are utilized that combine historical market-level data and build features that deal with seasonality and events to more precisely forecast demand. Bootstrap Aggregate methods and Gradient Boosting machines are utilized to achieve improved forecasting performance and model interpretability. This work paved the way to interpret how the machine learning models work and determining what features would impact the forecast. A complete survey of the

Fig. 6 Thanksgiving 2018

Fig. 7 United nations general assembly 2018

Fig. 8 Training and test set

topic “model interpretability” is beyond the scope of this paper (please refer to SHAP (Lundberg and Lee 2017) for an overview).

Bootstrap aggregating

Bootstrap Aggregating (Breiman 1997) involves building each model in the ensemble using a randomly drawn sample of the data. The predictions are made by averaging the predictions of each decision tree. This method of random sampling elements of the predictor space adds more diversity which reduces the variance of the trees at the cost of equal or higher bias. Random Forest is the main representative of bagging (Amit et al. 1997) in which prediction by an ensemble of trees is more accurate than that of any individual tree.

GradientBoost

The idea behind ‘boosting’ in machine learning is to train weak learners in a gradual, additive, and sequential manner, each trying to improve over its prior learner. This method fits the new predictor of the residual errors made by the previous predictor for every incorrect observation at every iteration. Gradient Boosting Decision Tree (Friedman 2001) uses Gradient Descent to find the shortcomings in the previous learner’s predictions. The boosting algorithms used in this research are as follows:

1. XGBoost (Chen et al. 2016).
2. LightGBM (Ke et al. 2020).
3. CatBoost (Dorogush et al. 2018).

New York City Demand

Fig. 9 Training data with sliding window

Stacking

Stacking (Breiman 1996a, b) is an ensemble machine learning modeling technique where multiple unique models are created, and it uses their output as the input for the meta-classifier/regressor to make a prediction. Unlike bagging and boosting, stacking may be (and normally is) used to combine models of different types. One of the advantages of stacking is that it leverages the strength of each individual estimator by using their output as the input of a final estimator. This reduces the bias any individual model might have.

Feature engineering

As mentioned in previous sections, there are several factors that affect market demand: market reservations, seasonality, events, market pricing. In this section, raw data are transformed to extract meaningful features which directly influence our model.

The lead time or days before arrival (DBA) is the time between the date of engagement/reservation and the reservation check-in date. In this research, DBA is divided into groups (DBA buckets) of fixed width. The DBA buckets are formed according to the assumption that the booking pace in these buckets follows the same trend (DBA Bucket: 0–3, 4–7, 8–14, 15–30, 31–45, 46–60, 61–90, 91–365).

A binary feature is added which handles events data. If there is an event, then the variable indicates 1, otherwise it indicates 0. Also, as the shoulder dates are adjacent to event

dates, these dates are dynamically estimated based on the event type. These dates are then designated as 1.

Several features are used to account for seasonality patterns. Due to periodic nature, a series of sine and cosine functions are employed in which frequencies are integer multiples of the fundamental seasonal frequency.

$$\begin{aligned} WS_{\cos} &= \cos \frac{nD\pi}{7} \\ WS_{\sin} &= \sin \frac{nD\pi}{7} \\ MS_{\cos} &= \cos \frac{nM\pi}{12} \\ WS_{\sin} &= \sin \frac{nM\pi}{12} \end{aligned}$$

where WS and MS are defined as the weekday and monthly seasonality, respectively. n is an integer multiple that can either be 2 or 4. D is the numerical representation of the day of the week and is defined by the interval $[0,6]$ where 0 and 6 correspond to Monday and Sunday, respectively. M is the numerical representation of the month of the year and is defined by the interval $[1,12]$ where 1 and 12 correspond to January and December, respectively.

Moreover, demand categories are created by sorting data values into equal bins of the market segmented and overall actual market occupancy over historical data time periods.

$$\text{demand}_{\text{buckets}} = \begin{cases} 1, & \min(X_t) \leq x < \min(X_t) + w \\ 2, & \min(X_t) + w \leq x < \min(X_t) + 2w \\ \dots & \\ N, & x \geq \min(X_t) + (N-1)w \end{cases}$$

where w is the width of an interval. X_t is an actual occupancy array over a specified time period. For example, this could correspond but not limited to the actual occupancy weekly, monthly, or quarterly level. N is the number ins. x represents an element in the array X_t . The piecewise function represents what category the element x belongs to with 1 representing the weakest demand category and N representing the strongest demand category.

Market prices are determined by taking the median of the lowest OTA prices for a single-day LOS grouped by hotels star rating in a market. Length of stay (LOS) is defined as the number of room nights a guest stays in a hotel. Online travel agencies (OTA) are travel websites selling travel products to consumers.

$$price(star) = [\min(OTA_{star,1}, \dots, \min(OTA_{star,n})]$$

$$MP = median(price(star)),$$

where $star$ represents the hotel star rating of interest. OTA represents the listed price offerings from a hotel in the channel. The variable n represents the number of hotels that comprise the market. MP represents the median of the lowest OTA prices for a given hotel star rating. Features such as moving averages and pickup (which is a subtraction based on a rolling window) are created. Pickup in the hospitality industry is the net change of a quantity during a defined time span. The moving average is calculated as below:

$$MA = \frac{MP_1 + MP_2 + MP_3 + \dots + MP_n}{n},$$

where MP is the market price, n is the number of days, and MA is the moving average.

Another technical indicator used is plus-minus which gives the number of days the market price has increased or decreased over a rolling window. Mathematically,

$$\Delta MP = \frac{MP_i - MP_{i-1}}{MP_{i-1}},$$

$$sign(i) = \begin{cases} -1, & \Delta MP < 0 \\ 1, & \Delta MP \geq 0 \end{cases},$$

$$plus-minus(i, n) = \sum_{t=i}^n sign(i),$$

where i and n are the specified date range. ΔMP is the percentage change in the market price. "plus-minus" is the sum of sign changes over a span of dates. With these transformed pricing features, the underlying trends of the market pricing can be uncovered.

Table 2 has information on the possible feature set used in the market forecast model.

The entire feature set in Table 2 is not necessarily needed for a market. Hundreds and thousands of markets can be defined with unique hotels—each with its own defining characteristics and features. With that in mind, each market can have a different set of features. This can also be described as global and local feature sets. Local features belonging to a distinct market can be determined by a multi-step process. The first step involves using a Recursive Feature Elimination algorithm. To begin with, the target variable (demand) is first trained on the whole feature set, and the importance of each feature is obtained. Then, the least important features are removed from the current set of features, and the regression metric is checked again. This gives the first set of features that can be removed with cross-validation. The next step is implementing a permutation importance algorithm to rank the features. This helps in removing the ones with negative or small importance.

Table 2 The feature set for modeling

Category	Sub-category	Frequency
Reservations	Rooms Unsold group rooms Cancellations Pickup Moving average	Daily
Market segmentation	Group Retail Discount etc.	N/A
Day of the week	Monday Tuesday etc.	N/A
Month of the year	Jan Feb March etc.	N/A
Day of the week seasonality	$\cos \frac{2D\pi}{7}$ $\sin \frac{2D\pi}{7}$	N/A
Month of the year seasonality	$\cos \frac{2M\pi}{12}$ $\sin \frac{2M\pi}{12}$	N/A
Lead time	DBA DBA bucket	Daily
Demand	High Medium Low	Daily
Average daily rate	Market segment rate Market rate Moving average Pickup	Daily
Market pricing	Market rate Moving average Plus-minus	Daily
Events	Yes No	Daily
Flights	Flight arrivals Flight departures	Daily

Unconstrained and constrained demand

In the lodging industry, unconstrained and constrained demand are significant pieces of information for a hotel. Unconstrained demand in a market is the total demand interested in lodging in a particular market and is irrespective of the capacity in that market. In practicality, the capacity cannot be exceeded which makes it necessary to constrain the total demand to the capacity of the property. The market forecaster forecasts the unconstrained demand by the market segment regardless of the supply of rooms in the market. In addition, the forecasts are price agnostic and do not include day rooms or oversold rooms. The forecasted unconstrained demand is the demand based on the arrival time and the current lead time. Thereafter, the forecasted unconstrained demand is constrained by the constraining equation given below to compute the demand based on the supply of the market. Thereafter, the demand is aggregated by segment which is total unconstrained and constrained forecasts.

d_{uncons} = model output

$$D_{uncons} = \sum d_{uncons}$$

$$k = \frac{1}{D_{uncons}},$$

$$d_{cons} = kd_{uncons}$$

$$D_{cons} = \sum d_{cons}$$

where d_{uncons} represents the forecasted unconstrained segment demand. D_{uncons} represents the aggregate of d_{uncons} . The variable k from represents the constraint constant which is the inverse of D_{uncons} . The variable d_{cons} represents the segmented constrained demand which is the product of the constraint constant and unconstrained segment demand. D_{cons} is the total constrained demand and the aggregation of the d_{cons} .

Modeling

The goal of the market forecasting framework is to reduce latency, scale both horizontally (markets), and vertically (dynamic feature selection). The arrival process (in-flow of demand) of each lead date for each market segment is stochastic in nature. Therefore, the distribution of bookings over the lead time of each date for each segment is considered in the training model. Moreover, data streams change over time leading to the change of underlying (static) model relationships and parameters in the data between the features and the demand variable. For example, there are perpetual changes of event dates, conference locations, and traveler preferences are always evolving. Accordingly, such changes need to be taken into account to update the model parameters

and maintain high performance. To address this issue, sliding window X (training data) is drifted by a single increment to generate training data on receiving the latest data stream. This helps in preserving dynamic training instances positioned within the window and discarding the past instances that minimize the generalization error.

Using these defined time windows, the machine learning model is trained by minimizing the loss function such as mean absolute percentage error (MAPE). This loss function is further segregated by lead-time buckets (0–3, 4–7, 8–14, 15–30, 31–45, 46–60, 61–90, 91–365 days).

Results and discussions

In this section, the results of the various machine learning algorithms are compared and analyzed to measure the performance of each forecast model. Each model was trained on two years of historical segmented and aggregate market reservations and actual data, starting from March 1, 2017, to February 28, 2019. For model evaluation, a test set with segmented and aggregate market reservations and actual occupancy from March 1, 2019, to September 30, 2019, was used. Each algorithm's performance is evaluated by using MAPE based on one day, monthly, and aggregate error reports. The MAPE of each model is based at the aggregate level, which is the total constrained demand, D_{cons} .

As new information pours in daily, the market forecast model runs, and a new set of predictions are made. Consequently, measuring the accuracy of a forecast on one particular day is not an accurate approach, but rather, the whole set of forecasts must be considered for the arrival date.

Five machine learning regression algorithms and a single time series model are used for building the market forecast model: SARIMAX, LightGBM, Catboost, Random Forest, XGBoost, and Stacked (Stacking). SARIMAX is a Seasonal Autoregressive Integrated Moving Average model that supports exogenous variables. SARIMAX is used as the baseline model to provide the required metric of comparison. Three machine learning algorithms are leveraged to build the Stacked Regressor, viz. the LightGBM and Catboost Regressors were used as the inputs to the XGBoost Regressor. The MAPE was calculated by comparing the one-year forecast based on the date of the forecast and the respective actual occupancy. For example, a forecast made on Jan 1st, 2019 will have predictions of the occupancy for Jan 1st, 2019 to Dec 31, 2019. Subsequently, the actual occupancy for each date is compared to the set of predictions and the MAPE is computed. In the following table, a forecast was made as of March 1, 2019 for the 365 days. Subsequently, the actual occupancy of the market was used to determine the MAPE. In Table 3, a summary of each algorithm's error can be seen.

After calculating the MAPE for each algorithm, the top three performing machine learning models were the LightGBM, Random Forest, and Stacked models. Each of them shows an improvement of at least 2% from the baseline time series model, SARIMAX. The top three performing models are evaluated further in this section.

In a market-level forecast context, to better understand the performance of the models, the error is tracked yearly, monthly as well as at daily levels, especially, for event days. The error uses the aggregate actual and constrained forecasted demand, D_{cons} . For these three models, one-day error reports were generated for specific event days in the New York City market. The one-day error report can be $L = j - i + 1$

$$MAPE(D, i, j) = \frac{100}{j} \sum_{t=i}^j \left| \frac{A(D) - F_t(D)}{A(D)} \right|$$

$$Curve = [(dba_{0,3}, MAPE(Date, 0.3)), \dots, (dba_{91,365}, MAPE(Date, 91, 365))]$$

formulated by the grouping the forecast for a particular date in days before arrival (DBA) buckets and calculating the MAPE for each of these DBA buckets. The selected days were Easter, Memorial Day, Independence Day, and Labor Day. Figures 10, 11, 12, and 13 provide the daily accuracy tracker for these days, respectively. The daily accuracy tracker measures the total accuracy of the model. The predictions are tracked at each lead time and then compared against the actual occupancy of the market. Thereafter, the MAPE would be calculated by grouping the lead time into buckets (DBA Bucket: 0–3, 4–7, 8–14, 15–30, 31–45, 46–60, 61–90, 91–365).

Table 3 Model comparison

Model	MAPE (%)
SARIMAX	5.20
Random forest	3.77
CatBoost	4.20
LightGBM	3.66
Stacked	3.77
XGBoost	3.95

D can be specified as a singular date or a range of dates. A and F are the actual and forecasted values, respectively. The variables i and j represent the DBA interval, where $j > i$. L is the span of the DBA bucket. The *curve* is the list of ordered pairs representing the accuracy curve. This provides a view of how the forecast performed as the lead time progressed to the arrival date. For example, to track the accuracy of May 27, 2019, the forecast at each DBA (lead time) leading up to the arrival date would be compared with the actual occupancy (say, 70%). The forecasts from Stacked in the DBA (0–3) bucket are 73.93%, 74.75%, and 75.78%. Using the equations above the MAPE would be 6.73%.

Fig. 10 1-Day accuracy report for Easter (04/21/19) in the New York City market

Fig. 11 1-Day accuracy report for Memorial Day (05/27/19) in the New York City market

Fig. 12 1-Day accuracy report for Independence Day (07/04/19) in the New York City market

Table 4 presents the MAPE and percentage of Over/Under-forecasted dates. The MAPE is calculated by comparing the actual occupancy with all of the forecasts by lead time. The percentage of Over/Under-forecasted dates are determined by summing the negative and positive errors by the lead time for each model and stay date. From Table 4, it can be observed that the three models under-forecasted for a majority of lead dates in three out of the four event dates. Moreover, the stacked model does not perform well in comparison with the other boosting models in terms of model performance. On the other hand, the performance of

the other two models is comparable with one outperforming the other in some dates and vice versa.

Table 5 outlines the comparison of the best and worst forecast errors (absolute) for each model. The worst and best forecast dates are determined by searching for the highest and lowest error for all stay dates grouped by lead time. While LightGBM and Stacked models are able to give the best performance at 65 days before arrival (DBA), Random Forest achieves the same performance at only 17 days before arrival (DBA). Furthermore, the worst forecast performance for the Stacked model is at 3 months before

Fig. 13 1-Day accuracy report for Labor Day (09/02/19) in the New York City market

Table 4 Event day error

Stay Date	Model	MAPE (%)	% of dates Over-forecasted	% of dates Under-forecasted
April 21, 2019	LightGBM	5.43	7.69	92.31
	Random forest	3.46	7.69	92.31
	Stacked	7.52	0	100
May 27, 2019	LightGBM	5.52	23.86	76.14
	Random forest	5.27	7.95	92.05
	Stacked	8.78	0	100
July 4, 2019	LightGBM	1.81	60.32	39.68
	Random forest	2.79	95.24	4.76
	Stacked	1.72	75.40	24.60
September 2, 2019	LightGBM	2.59	37.63	62.37
	Random forest	3.30	9.14	90.86
	Stacked	3.54	16.13	83.87

Table 5 Comparison of best and worst forecast error

Model type	Error category	Stay date	Absolute difference (%)	DBA
LightGBM	Best	8/16/2019	0	65
	Worst	7/1/2019	13.72	122
Random forest	Best	9/2/2019	0	17
	Worst	9/1/2019	14.07	184
Stacked	Best	8/16/2019	0	65
	Worst	5/27/2019	14.13	87

arrival, while LightGBM and Random Forest forecasts the worst 4 and 6 months before arrival, respectively.

For a more holistic view of model performance, the monthly accuracy tracker provides a perspective of how the models performed. The monthly MAPE is calculated by comparing the actual occupancy of the entire month with the forecasts by lead time.

Table 6 presents these results from March to September. The Stacked model has the highest under-forecasting percentage for all months while the LightGBM and Random Forest models have a mix of over and under-forecasting.

Table 6 Monthly error comparison

Month	Model	MAPE (%)	% of dates Over-forecasted	% of dates Under-forecasted
March	LightGBM	1.97	78.83	21.17
	Random forest	2.16	85.89	14.11
	Stacked	3.69	13.10	86.90
April	LightGBM	1.99	52.33	47.67
	Random forest	2.13	68.10	31.90
	Stacked	2.99	17.28	82.72
May	LightGBM	2.34	52.03	47.97
	Random Forest	2.56	48.60	51.40
	Stacked	2.81	20.23	79.77
June	LightGBM	1.90	34.20	65.80
	Random forest	2.12	33.09	66.91
	Stacked	2.24	14.45	85.55
July	LightGBM	2.90	37.94	62.06
	Random Forest	2.96	42.57	57.43
	Stacked	3.05	14.70	85.30
August	LightGBM	2.65	43.41	56.59
	Random forest	2.50	45.43	54.57
	Stacked	2.45	31.55	68.45
September	LightGBM	3.22	63.51	36.49
	Random forest	3.49	50.66	49.34
	Stacked	3.05	46.67	53.33

Table 7 Aggregate model comparison

Model	MAPE (%)
LightGBM	2.65
Random forest	2.76
Stacked	2.78

Moreover, the LightGBM model has the lowest MAPE for five out of the seven months ranging from 1.90 to 3.22%.

At last, model performance is presented in an aggregate view. Table 7 presents a model comparison of the aggregate MAPE for all lead dates leading to stay dates.

Fig. 14 Forecast as of Jan 1, 2020

To summarize the results and discussions above, the LightGBM tends to perform better when compared to other models. This indicates that the LightGBM-based market forecast model can be useful to a hotel revenue manager and it instills a satisfactory level of trust.

Forecast Interpretation

Market forecast architecture (Fig. 14) provides valuable insights for 365 days into the future that offers visibility into market performance. This information is crucial for hoteliers to anticipate dynamic changes in the market and formulate strategies to tackle those conditions. The forecast measures provide an understanding of the market booking data and pace to compare against their hotel booking movements. They would be able to have knowledge of the market intelligence on future dates for each stay date and accordingly formulate a strategy to complement and/or outpace the market. With booking lead times statistics, hoteliers would be able to make better decisions around the availability of rooms.

All these performance measures are needed both at an aggregate level and also at a segmented level to facilitate prescriptive actions at a particular segment level. Moreover, they would also have the knowledge of market pricing categorized by hotel rating and, thus, strategize hotel pricing movements. Finally, the algorithm also provides information on events' so that the revenue managers can take early action. Therefore, with the forecast market demand for each stay date in the future, hoteliers would be well equipped to keep prices up to date.

As an illustration, when the market demand is high (major conferences, concerts, festivals, etc.), the hotel would want to increase the prices, put stay controls, and also at the same time adjust their distribution strategy to facilitate maximum revenue during those nights. Therefore, in the New York City market, on high-demand nights like United Nations General Assembly Debate (September 22, 2020) (Fig. 15) or Black Friday (November 27, 2020) (Fig. 16), travelers usually book their trips beforehand (i.e., at much higher days before arrival) compared to average market demand nights and hoteliers would want to set their pricing strategy early enough to capture the demand.

On the other hand, during low market demand, the hotel would want to switch over its focus on increasing its occupancy through price and promotions. For example, low demand nights like Mother's Day (10th May 2020) (Fig. 17), in order to drive occupancy, the hotelier would want to have a different pricing strategy.

Conclusion

Unlike previous works, this research focused on having the knowledge of the market, market segments, and what affects consumers' patterns of travel by lead time. To account for accuracy, scalability, and performance, the most fitting way to maximize the market forecasting accuracy is to leverage machine learning methods for forecasting. This research determined the most appropriate machine learning algorithm to forecast market demand (both aggregated and segmented) and dealt with the seasonality and events of a market. The LightGBM model provided the best error rate with

Fig. 15 (Pre-COVID-19) UNGA 2020 forecast

Fig. 16 (Pre-COVID-19) Thanksgiving 2020 forecast

Fig. 17 (pre-COVID-19) Mother's Day 2020 forecast

a MAPE of 2.65% at an aggregated market-level compared to other models and traditional time series model (SARIMAX). Accordingly, a forecasting framework is engineered which reduces latency, can scale both horizontally and vertically, can handle model decay, and supports real-time data processing.

However, the research is subject to a few limitations like the availability of data and, more importantly, data integrity. Proper care has been taken in data pre-processing, and the results are reliable and accurate that the machine learning models perform very well in comparison to traditional time

series models, like SARIMAX. However, there might be a few instances where the market might not possess enough data for machine learning models to be deployed. Under such cases, the concept of “progressive enhancement and graceful degradation” is proposed. Depending on the data availability, an applicable number of features are created, and a machine learning model is deployed. For instance, if only market pricing records are available, but no market reservation records, then demand is forecasted based on market pricing (recall, the concept of price demand elasticity).

Future research should also include a study of a combination of machine learning and statistical modeling to check what, if any, additional forecast accuracy can be achieved and also make it more scalable. Moreover, a study of the amount of historical data necessary to train the machine learning forecast model would be helpful. Secondly, a market is also affected by unexpected weather events like hurricanes, a pandemic like COVID-19, or cancelation of a big event like Mobile World Congress, 2020, Facebook Developer Conference, 2020. Therefore, future research may apply the concept of “now-casting” to deal with such uncertainties and adjust the forecasted demand pattern.

References

- Amit, Y., S. Bauer, R. Kohavi, L. Breiman, T. K. Ho, J. Kleinberg, Y. Freund, P.L. Bartlett, and J.C. Lee. 1997. Random Forests. Machine Learning. Kluwer Academic Publishers, January 1, 1997. <https://link.springer.com/article/10.1023/A:1010933404324>.
- Bar-On, Raphael Raymond. 1999. The Measurement of Seasonality and Its Economic Impacts. *Tourism Economics* 5 (4): 43758. <https://doi.org/10.1177/135481669900500409>.
- Blackmar, Frank W. 1912. *Economics*. New York: Macmillan Co.
- Breiman, Leo. 1996. Bagging Predictors. *Machine Learning* 24 (2): 12340. <https://doi.org/10.1007/bf00058655>.
- Breiman, Leo. 1996. Stacked Regressions. *Machine Learning* 24 (1): 4964. <https://doi.org/10.1007/bf00117832>.
- Butler, Richard. 1998. Seasonality in Tourism: Issues and Implications. *The Tourist Review* 53 (3): 1824. <https://doi.org/10.1108/eb058278>.
- Chen, Tianqi, and Carlos Guestrin. XGBoost. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD 16, 2016. <https://doi.org/https://doi.org/10.1145/2939672.2939785>.
- Cooper, W. L., T. Homem-de-Mello, and A. J. Kleywegt. 2006. Models of the spiral-down effect in revenue management. *Operations Research* 54 (5): 968–987. <https://doi.org/10.1287/opre.1060.0304>.
- Demico, Fred J., Yan Lin, Ling Liu, Ldia Rejt, Srikanth Beldona, and Diccon Bancroft. 2006. The Effect of Holidays on Hotel Daily Revenue. *Journal of Hospitality & Tourism Research* 30 (1): 11733. <https://doi.org/10.1177/1096348005284489>.
- Dorogush, Anna Veronika, Vasily Ershov, and Gulin Andrey. 2018. CatBoost: Gradient Boosting with Categorical Features Support. arXiv.org, October 24, 2018. <https://arxiv.org/abs/1810.11363v1>.
- Friedman, Jerome H. 2001. Greedy Function Approximation: A Gradient Boosting Machine. *The Annals of Statistics* 29 (5): 11891232. <https://doi.org/10.1214/aos/1013203451>.
- General. Seasonally Adjusting Vehicle Miles Traveled—Bureau of Transportation Statistics. Accessed February 27, 2020. <http://www.bts.gov/explore-topics-and-geography/topics/seasonally-adjusting-vehicle-miles-traveled>.
- Global FocusNorth America. How the Sharing Economy Is Transforming the Short-Term Rental Industry. Knowledge@Wharton. Accessed February 27, 2020. <https://knowledge.wharton.upenn.edu/article/short-term-rentals-the-transformation-in-real-estate-and-travel-set-to-check-in/>.
- Ivanov, Stanislav. 2014. *Hotel Revenue Management: From Theory to Practice*. Zangador Ltd.
- Ke, Guolin, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. Accessed February 27, 2020. <https://papers.nips.cc/paper/6907-lightgbm-a-highly-efficient-gradient-boosting-decision-tree.pdf>.
- Kimes, S.E. 1999. Group forecasting accuracy in hotels. *Journal of the Operational Research Society* 50 (11): 1104–1110. <https://doi.org/10.1057/palgrave.jors.2600770>.
- Lundberg, Scott, and Su-In Lee. 2017. A Unified Approach to Interpreting Model Predictions. arXiv.org, November 25, 2017. <https://arxiv.org/abs/1705.07874>.
- Oxford Economics Global Travel Market Study. TripAdvisor Insights, October 25, 2018. <http://www.tripadvisor.com/TripAdvisorInsights/w2841>.
- Rajopadhye, Mihir, Mounir Ben Ghalia, Paul P. Wang, Timothy Baker, and Craig V. Eister. 2001. Forecasting Uncertain Hotel Room Demand. *Information Sciences* 132 (1–4): 111. [https://doi.org/10.1016/s0020-0255\(00\)00082-7](https://doi.org/10.1016/s0020-0255(00)00082-7).
- Sigala, Marianna. 2015. From Demand Elasticity to Market Plasticity: A Market Approach for Developing Revenue Management Strategies in Tourism. *Journal of Travel & Tourism Marketing* 32 (7): 81234. <https://doi.org/10.1080/10548408.2015.1063801>.
- Vinod, Ben. 2004. Unlocking the Value of Revenue Management in the Hotel Industry. *Journal of Revenue and Pricing Management* 3 (2): 17890. <https://doi.org/10.1057/palgrave.rpm.5170105>.
- Weatherford, Larry R., and Sheryl E. Kimes. 2003. A Comparison of Forecasting Methods for Hotel Revenue Management. *International Journal of Forecasting* 19 (3): 40115. [https://doi.org/10.1016/s0169-2070\(02\)00011-0](https://doi.org/10.1016/s0169-2070(02)00011-0).
- Zakhary, Athanasios, Amir F. Atiya, Hisham El-Shishiny, and Neamat E. Gayar. 2009. Forecasting Hotel Arrivals and Occupancy Using Monte Carlo Simulation. *Journal of Revenue and Pricing Management* 10 (4): 34466. <https://doi.org/10.1057/rpm.2009.42>.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Rimo Das serves as a Data Scientist at LodgIQ, where he is working on dynamic pricing, demand forecasting, and uncertainty estimation. He earned an MS in Business Analytics from University of California, Irvine, and an MBA from Symbiosis International University with a concentration in marketing.

Harshinder Chadha serves as a Data Scientist at LodgIQ, where he is working on demand forecasting and pricing optimization. He earned a BS in Mechanical Engineering from San Jose State University.

Somnath Banerjee serves as the CTO of LodgIQ, an AI/ML-based revenue management platform. He started his career at Texas Instruments Research and later worked at Paul Allen's Interval Research. He founded and operated a multinational technology company for 8 years and led its successful acquisition. He is also very passionate about cryptocurrencies, algorithmic trading, and digital art. He earned a Bachelor of Technology in Electronics and Communication Engineering from IIT, Kharagpur and an MS in Computer Science from Southern Methodist University.