Data Wrangling to Data Science in Revenue Management: A Dynamic Pricing Model for the Vacation Rental Industry

Changxuan Liu, Nelson Chou, Yachu Liu, Jiangtao Xie, Matthew A. Lanham
Purdue University, Department of Management, 403 W. State Street, West Lafayette, IN 47907
chou71@purdue.edu, liu1498@purdue.edu, liu2371@purdue.edu, xie310@purdue.edu, lanhamm@purdue.edu

ABSTRACT

We develop a dynamic pricing model for the vacation rental industry that fuses public and proprietary data to optimally price a vacation rental, while maximizing the firm's key performance index (KPI). The motivation of this research is the increased interest in expanding data science and analytics capabilities within the revenue management business area. Bigger picture, it is believed that the future of revenue management will begin to influence other industries where perishability of revenue exists, apart from the traditional service industry. In collaboration with a world-wide vacation rental company we fuse the firm's proprietary vacation rental data with other rentals on the market using customized web crawlers that capture the listed price and features of those rentals. Our model ingests this data to make pricing recommendations that optimize the firm's KPI. Our pricing model uses price, distribution channels, unit features, lead time, stay length, and other features, and provides a propensity of unit purchase given a posted price, and the current market competition. Lastly, we can recommend the price that maximize the overall KPI. This not only lets the organization plan better, but also allows the firm greater flexibility to improve and price more aggressively in geographical areas where it is expanding.

Keywords: Revenue management, Pricing, Predictive Analytics, Web Scraping, Optimization

1. Introduction

It is a critical time for business to rethink their approach for managing profit in the digital era. Digital transformation has been widely recognized as one of the most disruptive business developments. A study by MIT Sloan and Deloitte found there were digital transformations undergoing in over 70% of the companies surveyed (Briggs, Henry et al. 2019). Among the business models created by digital transformation, the 'As a Service' (XaaS) model was one of the best-performing and fastest-developing (Newman 2017). Many industries are joining the service model by adding knowledge, subscriptions, solutions, and expertise into the product line. Examples include UPS Solutions for logistics industry and GE Software for manufacturing industry. While companies enjoy the increasing revenue from selling services, lessons from the traditional service sector on the perishability of service products should not be overlooked. The objective of our study was to propose a data-driven approach to maximize the expected revenue from perishable service products based in the context of the vacation rental business, and ultimately generalize the application of revenue management to businesses launching 'XaaS' digital transformation in other industries.

Perishability of service products leads to sunk revenue once a sales opportunity was unsuccessful. For example, the revenue associated with an unsold flight seat or an unoccupied hotel room is lost permanently once the plane took off or the day passed. Driven by the significant fixed cost and fixed capacity, airlines and hotels use mass market segmentation and revenue management to ensure the correct price was charged to the correct customer at the correct time (Walker and Walker, 2004). Revenue management involves controlling price and distribution channels to achieve better KPIs. Occupancy rate and revenue per available room (RevPAR) are the most commonly used. In the context of the lodging industry, the KPIs can be calculated using Equations 1 and 2. Similar formulas can be applicable for other service providers.

$$\text{Equation 1: } Occupancy = \frac{roomnights\ sold}{total\ roomnights\ available}$$

$$\text{Equation 2: } RevPAR = \frac{total\ rooms}{total\ roomnights\ available}$$

The concept of revenue management can be applied to a broader range of industries under four conditions: fixed capacity, predictable demand, perishable product, and possibility of customer segmentation. The business problem in this study is to find the revenue management strategy that achieves the best outcome, where best is defined by the greatest key performance index (KPI). We framed the business problem into the analytical problem by developing a descriptive model of sales (effect) caused by price and different offerings. This assumed cause-and-effect relationship was solved as a predictive model to estimate the effect the inputs had on the outcome (sales). Using the predictive model parameter estimates, we then formulated an optimization/decision model to maximize the KPI calculated based on the prediction results using the predictive model inputs that the firm would have control over (e.g. price) as decision variables to be solved for.

Using a market-leading hospitality company in the vacation rental business, as a prototype, we aimed to use analytics to recommend pricing decisions with the goal of motivating the evolution of revenue management into other industries developing perishable service products during digital transformation. Our methodology incorporated both internal data from the partner company and public competitor data scraped from the Internet. Using historical booking records and inventory information we tested multiple supervised-learning algorithms to understand the effects of our predictive model on demand. Given a potential booking with information on time, location, service offering, and external competition, the model predicted whether the potential booking would be booked or not. Our optimization model was designed to take the prediction results to find the recommended price that would achieve the highest expected revenue over all bookings. For example, if a booking was offered at a price of $250 and the predictive model predicted the probability of being booked to be 0.47, then as we change the price lower the probability will increase. The optimization model helps us find the optimal price for this unit as well as all others while considering several business and market constraints.

The remaining sections of this paper are organized as follows. First, the existing literature on revenue management practices are summarized in the literature review. Second, we describe the characteristics of the data provided by the partner company and scrapped from the Internet. Third, we explain the design of our analytical solution that establishes the descriptive model into a predictive model on future bookings, and how we use those parameter estimates in our overall optimization model. Our model assumptions are clearly highlighted in that section. Fourth, the predictive models investigated are evaluated to identify the best performing model, and our pricing recommendations from the optimization were compared with the current strategies to assess the effectiveness of our proposed solution. Lastly, we reflect on the model, data, and available information used in our study and their limitations and suggest directions for future research in this area.

2. Literature Review

The research on dynamic pricing for the hotel industry has been scarce. Many research papers cover the pricing strategy for traditional hotels but papers covering dynamic pricing for vacation rentals are limited. Traditional hotels accommodate both short-term and long-term stay customers while vacation rentals focus more on the long-term stay customer (e.g. seven days or more) and customers usually book at least one month in advance. Though there are fundamental behavior differences between traditional hotel customers and vacation rental customers, the pricing strategy for traditional hotels still provides some insights for setting the right pricing for vacation rentals. Erdem and Jiang (2016) analyzed over 70 hotel Revenue Management (RM) related research articles and provided an overview of the research progress and challenges of Revenue Management. Buckhiester (2011) states that the key components of RM include capacity alignment, competitive benchmarking, strategic pricing, demand forecasting, business mix, and distribution channel management. Kimes (2011) concluded in her study that "*the future was going to be much more strategic in nature and will be more strongly driven by technology in which function space will be the new frontier*". However, many companies still have not embraced the emerging technologies required to develop more strategic approaches.

There has been research that has focused more the on various statistical methods that could maximize revenue. For example, Guo, Ling et al. (2013) focused on using an online reservation system to set optimal pricing based on market segmentation (e.g. clustering), with the goal of discriminating different characteristics of customers based on price. The proposed pricing model based on different market segments could then be applied to different seasons with different capacities. The model assumed a fixed demand function and fixed amount of hotel capacities. For linear demand, the optimal pricing for n segments can be obtained based on the condition that total capacities are larger than (c ≥ $a_0 - b$) or smaller than the demand (c < $a_0 - b$) under each segment. For a non-linear demand curve, a recursive dynamic programing algorithm was developed. The algorithm calculates the number of segmentations for the non-linear function and after the K segmentation is defined, the profit of each Kth segment can be derived based on the known solution at stage $k - 1$. Maximizing the profit, the optimal price is obtained for each K segment. The model results yielded different room rates for peak and slack seasons for different lead times. The limitation of their model is that it considered customers have homogenous price sensitivity and it did not consider customers with different stay lengths and cancellations prior to arrival dates.

Ye, Qian et al. (2018) published a study on dynamic pricing at Airbnb using their online home sharing platform. They focused on using the daily prices in a given city to provide pricing suggestions for their owner portal. The challenge they faced is that demand estimation is extremely difficult to predict due to time-varying (seasonality and lead time) and list-varying (different housing types and positive review variation) effects. The initial approach was to build a gradient boosting machine (GBM) model to predict booking probability for each market. They found this difficult to do accurately. Their alternative approach was to construct a pricing suggestion matrix that factors in calendar price set by the host, guest booking probability, and market demand signals. The offered price was further adjusted by considering suggested price vs actual booked price to come up with a dynamic pricing strategy. They conclude that this pricing strategy has proven to perform better than a direct max-revenue pricing strategy.

Ling, Guo et al. (2012) provides a methodology for setting an optimal price for a long-term hotel stay. First, they categorize hotel rooms into short-term and long-term stays. The short-term stay customer per day is estimated using a Poisson distribution. The probability that long-term stay customer accepts a room rate follows a non-increasing convex function on the basis that the lower the room rate, the higher the demand and vice versa. In the hotel business, the availability of long-term stay rooms is affected by the short-stay rooms, given that long-term stay customers are charged with a lower rate per night. The authors found that customers with long-term stay do not always enjoy the best rate from hotels, because the rates are affected by the occupancy rate of each hotel. Thus, hotels should try to improve overall occupancy rate so that higher rates can be charged on long-term stay customers to maximize the overall revenue.

In order to maximize total revenue, Aziz, Saleh et al. (2011) proposed a new optimization/decision pricing model. The objective function is the summation over all nights of the multiplication of the price of a certain night and the occupied room for that night. The only restriction is the number of reservations cannot exceed the total capacity at a particular night. Their approach extended a classical deterministic model, which maximized the summation over nights for a price associated with a price class, the stay length, and the optimal allocation to a stay of type, by providing a dynamic pricing model that incorporates price elasticity of demand. By partitioning the total capacity of the hotel into segments and examining the models, the evidence showed the dynamic pricing model could contribute more to increasing total revenue. A generic outline of the model is shown below:

$$\begin{array}{ll} \text{Maximize} & \text{Subject to} \\ \displaystyle\sum_{l=1}^{Max\ l} P_l O_l & \begin{array}{l} O_l \leqslant C_l \quad \forall l \\ P_l \leqslant 0 \quad \forall l \end{array} \end{array}$$

Tse and Tung Poon (2012) found that most of studies assume an algebraic relationship among the rate and demand. However, in some situations the demand function is a superior forecasting model than the traditional method, as it might generate unrealistic solutions. Therefore, they might consider a different function: $\frac{1}{2}\left(\left(1-\frac{1}{\beta}\right)\right)r_0+\gamma$, to achieve maximized revenue, where β is the regression coefficient, r_0 is the initial equitable room rate and γ is the variable cost per room sold. They claim this method is more reliable both in theory and in a realistic situation.

Our study provides a dynamic pricing model that assumes a descriptive (cause-and-effect) relationship exists. We estimate the effect of demand (booking the unit) based on price, unit, and market characteristics in a analogous fashion to Airbnb's study. However, we try to optimize the offered price using an optimization/decision model using ideas from (Aziz, Saleh et al. 2011). Our model formulation in additional to crawling and incorporating this external competitor data to examine the potential improvements makes is what makes our study novel.

3. Data

3.1 Internal Data

To predict whether a unit will be sold on the particular day, variables related to the unit booking (e.g. booked or not, daily rent, arrival date, stay length, lead time, number of guests) and room features (location, room capacity, room cluster) were included in the model. This data was collected during the arrival periods corresponding to 2018 – 2022.

Internal data consists of two categories, booking data and inventory data. Booking data had 200,000 observations and captured 2018-2022 unit-level historical and future bookings statuses, arrival dates, departure dates, booking dates, total rent amount, capacity information, distribution channel, and customer demographic information. Inventory data had 3.5 million observations for the same time period and consists of unit-level daily rent, availability, status, and room features (amenities). This data was broken into three proprietaries geographical regions: A, B and C as shown in Table 1. Region A's unit features are comprised of 400+ columns essentially measuring three aspects: unit parking features, unit pool/hot-tub/sauna indicators, and unit proximity; Region C's unit features comprised of 174 columns measuring mainly six aspects: unit parking, unit pool/hot-tub/sauna indicators, unit proximity, unit services available, unit entertainment, and unit bathroom features. Region B's unit features comprised 112 columns measuring eight aspects: unit parking features, pool/hot-tub/sauna indicators, unit proximity, unit services available, unit entertainment, unit exterior features, unit extra sleeping area, and unit kitchen features. The data used in this analysis were for the year of 2018.

	Bookings data			Inventory data
Region	Booked	Canceled	Grand total	Grand total
A	138,643	30,377	169,020	2,691,958
B	24,475	7,149	31,624	538,448

C	16,714	2,343	19,057	234,048
Grand total	179,832	39,869	219,701	2,464,454

Table 1: Summary for Bookings and Inventory

Tables 2 and 3 below provide a data dictionary for the booking and inventory.

Variable	Type	Description
Unit ID	Numeric	ID of the Unit
Unit Name	Categorical	Name of the Unit
Booking Status	Categorical	Status of the booking (V or X)
Stay Type	Categorical	Type of the stay (e.g. Paid Guest, UnPaid Guest, Owner etc.)
Book Date	Date	Date when the Booking was confirmed
Arrival Date	Date	Date of Arrival
Departure Date	Date	Date of Departure
Cancel Date	Date	Date if the booking was cancelled, else NULL
Update Date	Date	Date if the record was modified in the system
Num Adults	Numeric	Guest Information
Num Children	Numeric	Guest Information
Num Infants	Numeric	Guest Information
Num Pets	Numeric	Guest Information
Rent	Numeric	Total Rent paid for the entire booking
Currency	Categorical	Currency
Unit Status	Categorical	Status of the Unit
Bed	Numeric	The number of beds in the unit
Bath	Numeric	The number of baths in the unit
Occupancy	Numeric	The number of people that can be accommodated
Unit Type	Categorical	The type of Unit for e.g. Home, Condo etc.
Unit View	Categorical	Featured View on Website
Country	Categorical	Location
State	Categorical	Location
City	Categorical	Location
Latitude	Numeric	Geographical Latitude
Longitude	Numeric	Geographical Longitude
Property Name	Categorical	Name of the Property in the system
Building Name	Categorical	Name of the Building where the Unit is located
Unit View	Categorical	Featured View on Website
Building City	Categorical	City where the Building is located
Others	Categorical	Other columns including room features, cancellation policy, etc.

Table 2: Selected Data dictionary for booking data

Variable	Type	Description
Unit ID	Categorical	UnitId in the System
Unit Name	Categorical	Name for the Unit
Date	Date	Inventory Stay Date
Stay Year	Numeric	Year of the Stay Date
Stay Week	Numeric	Week of the Stay Date
Unit Status	Categorical	Current Status of the Unit
Bed	Numeric	The number of beds in the unit
Bath	Numeric	The number of baths in the unit
Occupancy	Numeric	The number of people that can be accommodated
Unit Type	Categorical	The type of Unit, e.g. Home, Condo etc.
Unit View	Categorical	Featured View from the Website
Country	Categorical	Country where the unit is located

State	Categorical	State where the unit is located
City	Categorical	City where the unit is located
Latitude	Numeric	Geographical Latitude
Longitude	Numeric	Geographical Longitude
Property Name	Categorical	Name of the Property in the system
Building Name	Categorical	Name of the Building where the Unit is located
Unit View	Categorical	Featured View from the Website
Building City	Categorical	City where the Building is located
Supply	Numeric	Total Supply except Suspended Units (Built from Status Current Field)
Capacity	Numeric	Total Supply Except Suspended and Blocked Units (Built from Status Current Field)
Rate	Numeric	Current Price
Others	Categorical	Other columns including room features, cancellation policy, etc.

Table 3: Selected Data dictionary for inventory data

3.2 External Data

External data consisted of information harvested from Airbnb, a national vacation housing website, and primary regional vacation rental websites. We developed web scrapers for the top two competitors for each region and for Airbnb as the national competitors. Information collected covered the 2019 full calendar year and includes website name, unit name, city, arrival and departure date, daily/weekly rents, unit capacity, and relevant fees.

4. Methodology

The primary problem focus in our study is to recommend pricing plans to achieve a maximum KPI. RevPAR, as defined previously, is the most popular KPI in this industry. RevPAR equals the occupancy rate multiplied by the price. For this study, we used expected revenue, a similar concept to RevPAR as the target variable to maximize, defined in Equation 3:

Equation 3: $Expected\ Revenue = Prob(a\ unit\ sold) \times\ Price$

Expected revenue represents the RevPAR on the unit level. In order to align the business problem of maximizing the total expected revenue into an analytical problem, we broke the problem down into two sub problems.

1) **A predictive model**: Predict whether a booking with a unit on a day will be successful given the requirement of the customer (lead time, planned length of stay, etc.), price, and the choice of distribution channel;
2) **An optimization model**: Calculation of KPI (RevPAR) using the predicted booking results and search for the best price (our decision variable) as the proposed recommendation.

Figure 1 describes our methodological design, followed by step-by-step explanations on the modeling process, with assumptions made.

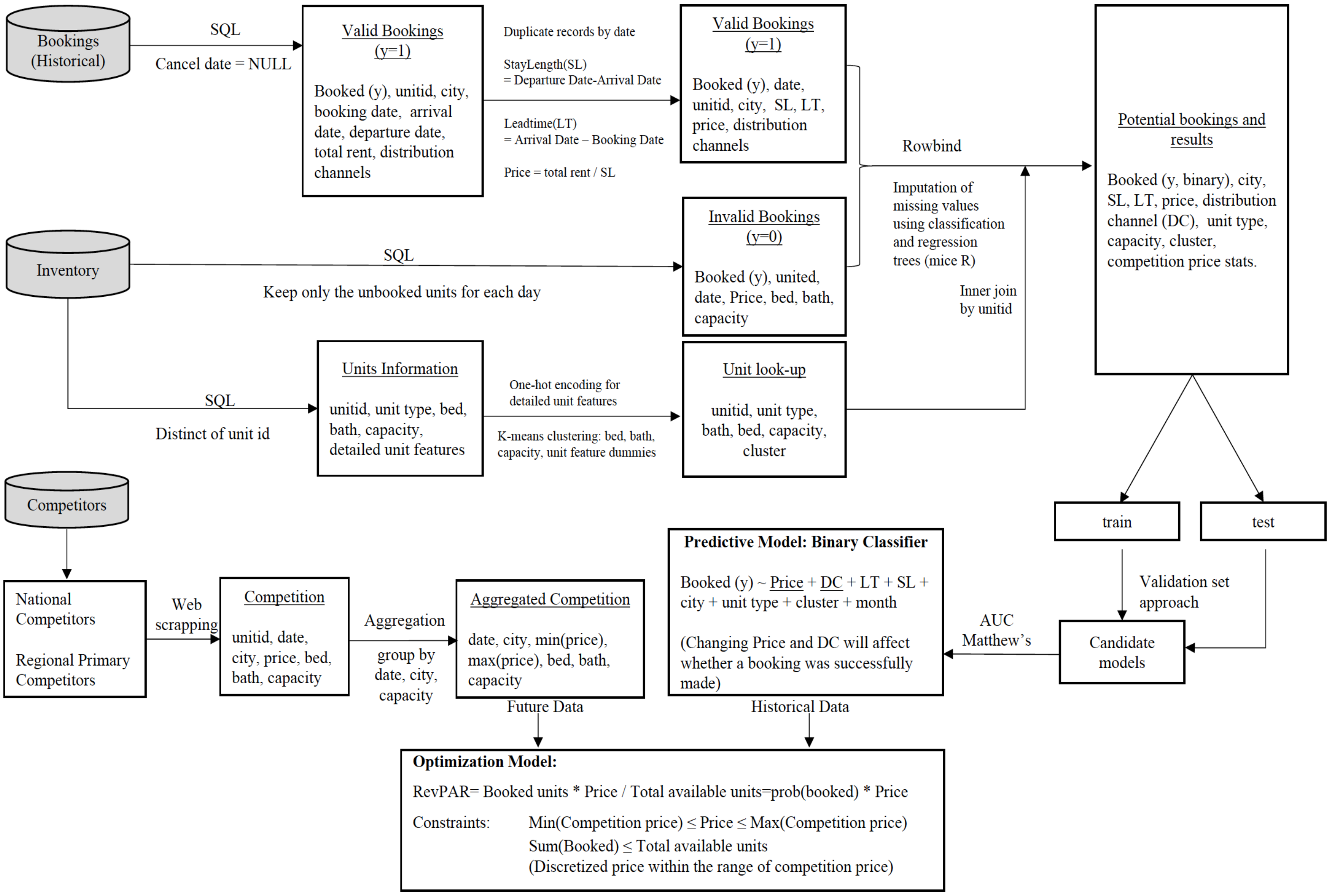
Bookings (Historical)
SQL
Cancel date = NULL
Valid Bookings (y=1)
Booked (y), unitid, city, booking date, arrival date, departure date, total rent, distribution channels
Duplicate records by date
StayLength(SL) = Departure Date-Arrival Date
Leadtime(LT) = Arrival Date – Booking Date
Price = total rent / SL
Valid Bookings (y=1)
Booked (y), date, unitid, city, SL, LT, price, distribution channels
Rowbind
Imputation of missing values using classification and regression trees (mice R)
Inner join by unitid
Potential bookings and results
Booked (y, binary), city, SL, LT, price, distribution channel (DC), unit type, capacity, cluster, competition price stats.
Inventory
SQL
Keep only the unbooked units for each day
Invalid Bookings (y=0)
Booked (y), united, date, Price, bed, bath, capacity
SQL
Distinct of unit id
Units Information
unitid, unit type, bed, bath, capacity, detailed unit features
One-hot encoding for detailed unit features
K-means clustering: bed, bath, capacity, unit feature dummies
Unit look-up
unitid, unit type, bath, bed, capacity, cluster
train
test
Validation set approach
Candidate models
AUC Matthew's
Competitors
National Competitors
Regional Primary Competitors
Web scrapping
Competition
unitid, date, city, price, bed, bath, capacity
Aggregation
group by date, city, capacity
Aggregated Competition
date, city, min(price), max(price), bed, bath, capacity
Predictive Model: Binary Classifier
Booked (y) ~ Price + DC + LT + SL + city + unit type + cluster + month
(Changing Price and DC will affect whether a booking was successfully made)
Future Data
Historical Data
Optimization Model:
RevPAR= Booked units * Price / Total available units=prob(booked) * Price
Constraints: Min(Competition price) ≤ Price ≤ Max(Competition price)
Sum(Booked) ≤ Total available units
(Discretized price within the range of competition price)

Figure 1: Methodology

4.1 Internal Data cleaning

Booking data: Booking records were stored with information of the deal, the travel plan, and features of the booked unit. We decomposed the data to the granularity of daily level. The rent column kept the total rent of the entire stay period. Daily rent was obtained by dividing rent by length of stay. Bookings cancelled were removed.

Inventory data: Unit information on a daily level with features and unit status including whether the unit was blocked, booked, or available. Features for all units were used to form categories for available units. Only the unbooked records were treated as unsuccessful deals and combined with the booking records.

Unit features: We one-hot-encoded all the categorical features. To avoid introducing a large number of dummy variables, we performed k-means clustering on the unit features in each of the three different regions separately. The Silhouette plots are as shown below in Figure 2 to 4:

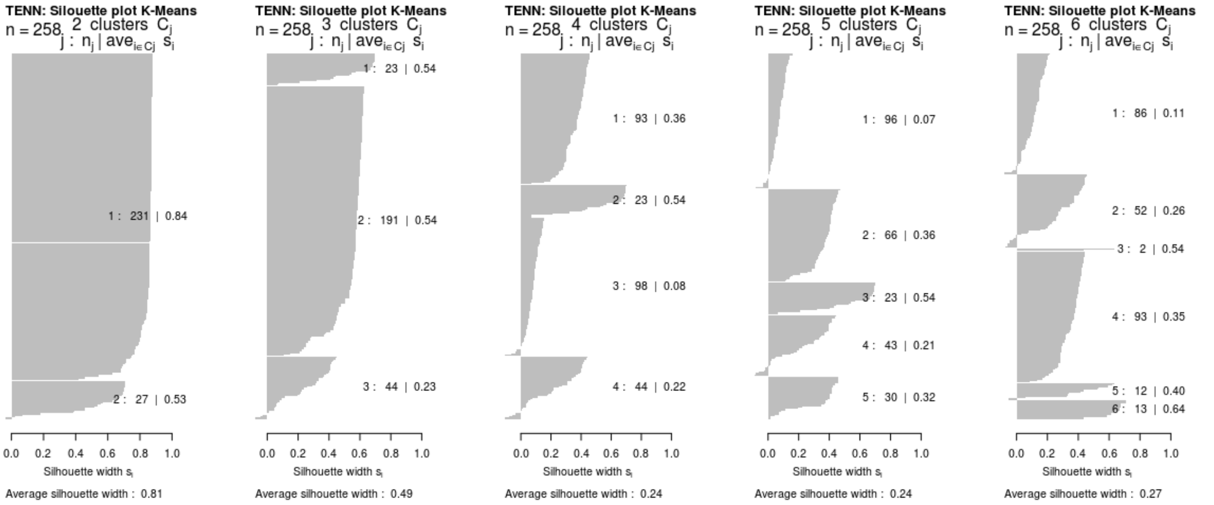

Figure 2: Silhouette plot in Region A

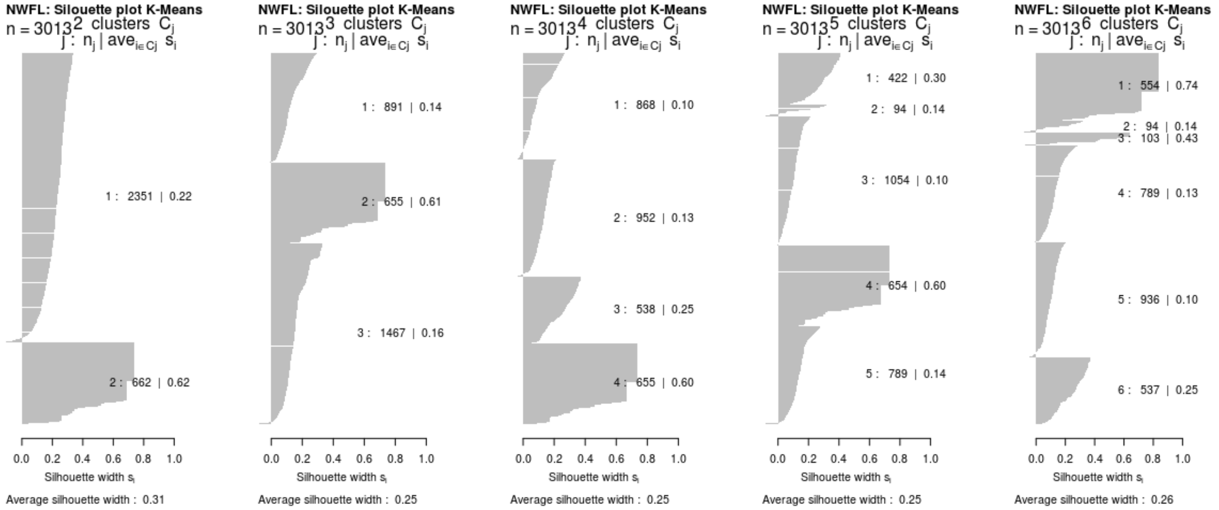

Figure 3: Silhouette plot in Region B

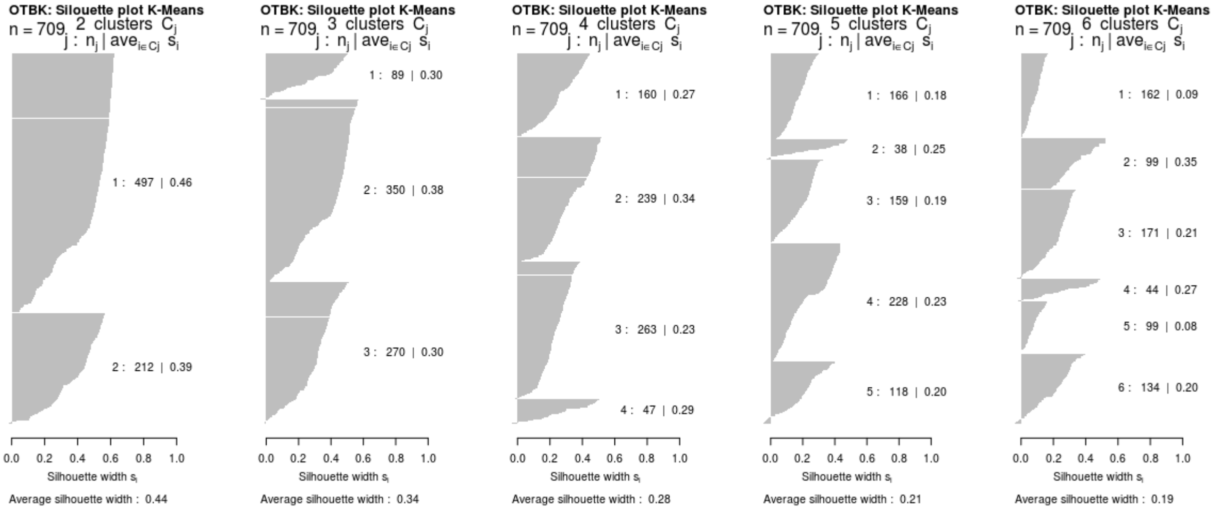

Figure 4: Silhouette plot in Region C

The silhouette plots indicated that in all regions, two clusters yielded the best clustering result. Therefore, all units from the same region were categorized into two clusters. The cluster of a unit was expected to capture the class of the unit, separating full-service luxury units from economical units.

4.2 Data Imputation

The **mice** package in R was used to impute the missing values of unbooked units. Information such as distribution channels, rent price, lead time and length of stay were unavailable for unsuccessful bookings coming from their inventory database. Combining the successful and unsuccessful bookings resulted in a data set that allowed us to predict the probability that a unit would be sold or not. However, imputing missing values we had make sure we were creating reasonable values. All rows with a target variable $Y = 0$ contained missing values and all other rows with $Y = 1$ were complete. As a result, the imputation assumed that (1) the listing price equals inventory rate, and (2) the probability distribution of imputed variables for unsuccessful bookings followed the same distribution post-imputation of those variables for successful bookings. Also, we assumed that pricing would always influence a booking outcome as vacation rentals are price elastic, while in reality some units would remain unsold under any price during slack seasons. After performing model-based imputation, the successful bookings and unoccupied inventories were joined.

4.3 Incorporation of competitor data

We extracted statistical information per unit on a specific week for all competitors and matched that info to a proprietary unit. We aggregated the competitors' data by city, date, and capacity and combined it with the internal data set. Summary statistics of the matched competitors, such as number of competitors, average price, standard deviation of price, and median price were captured. The competition conditions were joined with the main dataset on unit and date.

4.4 Preprocessing

Clustering results were added to the main dataset which incorporates both internal offerings, result offerings, unit information, and the price statistics from external competitors. The continuous variables were preprocessed using a z-score standardization to avoid bias from different scales when the predictive model learns. Categorical variables were one hot encoded to create dummy variables and at least one variable encoding was removed per set to avoid having perfect linear combinations of categorical features. Week of year for each row was extracted from the date field and one-hot encoded to capture the seasonality of the market.

4.5 Train and evaluate predictive models

We trained and compared four different supervised algorithms, namely a logistic regression (Logit), a random forest (RF), a feed-forward artificial neural network (NNet), and extreme gradient boosting tree (XGBTree) to predict the probability of unit being sold on a particular week. We partitioned the data so that 70% of the data set was used to train the models and the remaining 30% was used to evaluate the generalizable performance of the models.

4.6 Optimization model

After obtaining a predictive model that best predicts the probability of a unit being sold at a given price, we tested different discrete price sets to get the maximized expected revenue. Meaning the predicted probability of the unit being sold multiplied by a possible price, to obtained expected revenue for a unit. The possible price was constrained by a competitors' maximum and minimum price. The price generating the maximum expectation of revenue was selected as the recommended price for the unit during a week. The optimization was done unit by unit. Hence, our current formulation does not capture the internal competition among units under the same network. The chance of being sold should be constrained by market seasonality and a network's market share so that the optimization could tell the revenue to sacrifice and that to utilize.

Optimization function:

$$max(\sum_{t=1}^{T}\sum_{n=1}^{N}[\,\mathrm{E}(revenue_{n,t})\,])$$

$$= max(\sum_{t=1}^{T}\sum_{n=1}^{N}[\,XGB.prob(price_{n,t}) \, * \, price_{n,t}\,])$$

Constrained by:

$$price_{n,t}/10 \; \in \; N^{*}$$

$$price_{n,t} \in [\min(compPrice_{n,t}),\ \max(compPrice_{n,t})]$$

$$\sum_{n=1}^{N}[\,XGB.prob(price_{n,t})] \;\le \frac{Demand_{t} * MktShr_{wynd,t}}{N}$$

where

T: the number of weeks within the optimization time range, T=53 in this study
N: the number of units to be optimized
$Price_{n,t}$: The price for unit n during week t

5. Result

For the predictive model, we used the area under curve (AUC) as our statistical performance measure to compare models. Comparisons are displayed in Figure 5. From all three regions, the XGBTree model had the best performance among all three regions was not overfitting. The Logistic regression model provided us the ability to interpret and draw insights about different features. However, in order to proceed to the optimization phrase, we chose the XGBTree since it has the best predictive performance.

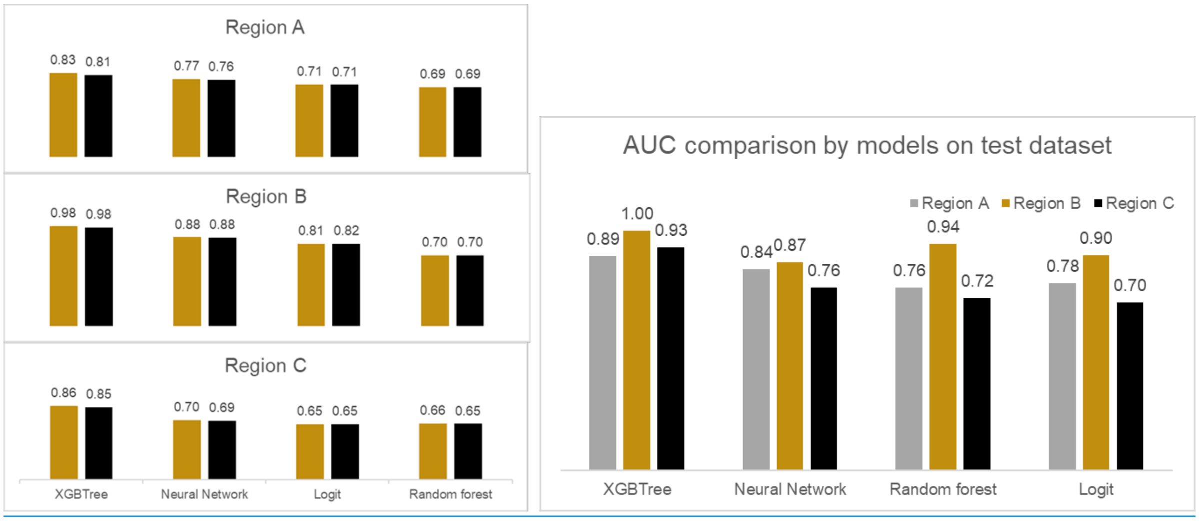

Figure 5: Model result comparison

In addition to performance, price was one of the features used in the tree. Tree models, unlike parametric models such as logistic regression may not use (or split) on the all the input features. In our case our predictive model is used to optimize the revenue for a given price, thus if price was not split on in the tree, when the price is changed it will not impact the predictive probability, nor the expected revenue. Luckily, price was an important feature in this model.

5.1 Region C

For Region C, the overall aggregated optimized result suggests 18% revenue increase in 2018, which is 8.7 million. The optimization captured more revenue from week 5 to week 27. The result also shows the comparison of the sold and unsold ratio before and after optimization. Before the optimization, 25% units are sold while setting the price using the optimization model, suggests 87% of the units would be sold. The average daily rates (ADR) for a sold unit are generally lower than those of unsold units. Thus, the model seems to suggest that by lowering the price, more units could be sold, and the overall revenue could be improved because many unsold units fall under the high price range bucket. However, one issue is that the optimization model does not consider different demands based on seasonality, therefore, the optimization result seems too optimistic.

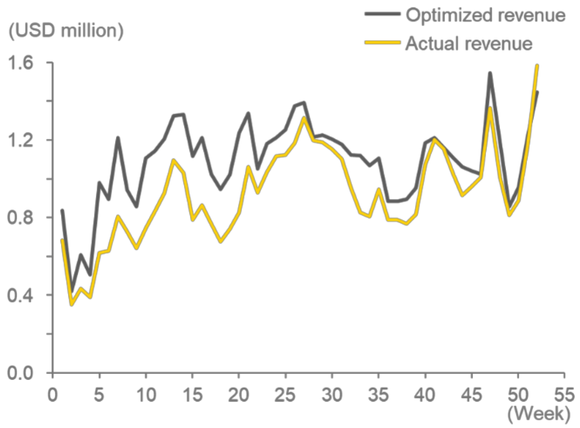

Figure 6: optimization result of region C

5.2 Region B

Region B generally has a higher price range than the other regions. Our optimization result was not promising. Using the model suggests an overall 21% decrease in annual revenue, which is 14.4 million, and fails to capture the peak season. One potential reason is that the predictive model trained for this region does not capture price sensitivity as well as other regions. For example, during a peak season, units could be sold at mid-to-high range prices, but during a slack season, lower prices do not guarantee bookings.

The sold and unsold ratio is similar with the result of Region C. In terms of ADR, the sold units are much lower than those of unsold units. The model may suggest that the high ADR causes units to be unsold and thus suggests a much lower price, leading to overall revenue decline in a year. Therefore, if we can identify consumer browsing behavior, e.g. capture numbers of page views for each unit, we could narrow down to the units with a reasonable number of views and only build model using those units so that the model does not get biased by the high selling price.

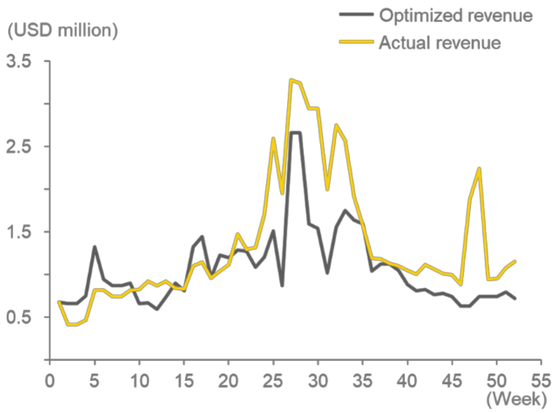

Figure 7: optimization result for region B

5.3 Region A

The result for Region A realized a 32% decrease in revenue, which is 140 million, versus actual revenue. This region has a special peak pattern during the middle of the year which is different from the other two regions. The peak season highly affects the price suggestion. The sold and unsold unit result after optimization for Region A is different from the results of the other regions. This region's competition price range varies from a low-price range to very high price range. This situation together with the unique peak season pattern, potentially affected the pricing suggestion for, leading to abnormal rate suggestions and a lower annual revenue. A better competition pricing comparison in the same segment and treatment of outliers are required for a region like this to further improve the optimization result.

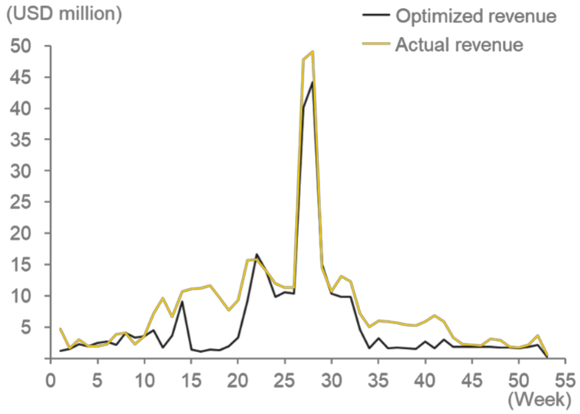

Figure 8: optimization result for region A

6. Study limitations

The optimization model is currently too optimistic for one region and too pessimistic for the other two regions due to assumptions and data limitations. In order to make our solution more robust, the following information could be collected and tested.

- Competitor data

  Currently the competition data is not comprehensive enough and still needs further consistent scraping. The more data that is collected, the more accurate the model accuracy can be. Since the current internal data is from 2018 and external data is from 2019 and beyond, one of the model assumptions assumes competition prices in 2018 and 2019 have similar patterns. This assumption can be eliminated once more 2019 competition data is collected and the data can be combined with 2019 internal data to build a model with data of the same timeline. In addition, if a competitor's unit feature information is collected and processed, units that are more homogenous can be compared. Finally, more competition data on a weekly level could also reveal pricing patterns better.

- Customer browsing behavior and demographic information

  Customer browsing behavior and demographic information are currently missing. For example, inventory data does not capture number of page visits by customers so that frequent viewed units cannot be identified. If additional customer browsing behavior information can be captured such as distribution channel information, page browsing time length, and customer demographic information, this might help improve the model.

- Economic conditions

  Demand elasticity for the market is unknown. To understand how price influences the unit sold it is important to better understand the demand for vacation rentals in general.

Other limitations from the model selection perspective:

- For the predictive model, the most accurate model was chosen as an intermediate input to the price optimization model. However, to gain more interpretable insights from the predictive phase, an alternative model such as logistic regression or random forest might have been more ideal candidates.
- For the optimization model, the optimization is on a unit and day level, which is 365 days for each unit. However, in practice, optimization might be based on the same unit type and week level. This would reduce the problem size. The current formulation might lead to two potential problems, one is too sensitive, which leads to more inaccuracy, and the other is computationally expensive. In the future, further work on fine tuning the constraints such as only computing one recommendation price for same unit type in one week can potentially improve the optimization result.

To address the above limitations, a road map has been developed for future work, as shown in Figure 9. As of now, only internal historical data has been analyzed. The next step requires a more comprehensive competition dataset, such as room features. This information along with customer browsing behavior information have potential to further improve the predictive model.

Another area is to identify more business constraints for the optimization model. Currently, the price constraint is subject to the minimum and maximum of competitor prices. In the future, only competitors that are in the same segment should be compared. Moreover, units that have predicted probabilities greater or equal to 0.50 are considered booked according to our current formulation. This booking threshold could be further adjusted. Finally, one price for same unit types under the same week is another constraint to be explored. Other constraints such as demand elasticity, overbooking, and revenue higher than the previous year could also be taken into consideration.

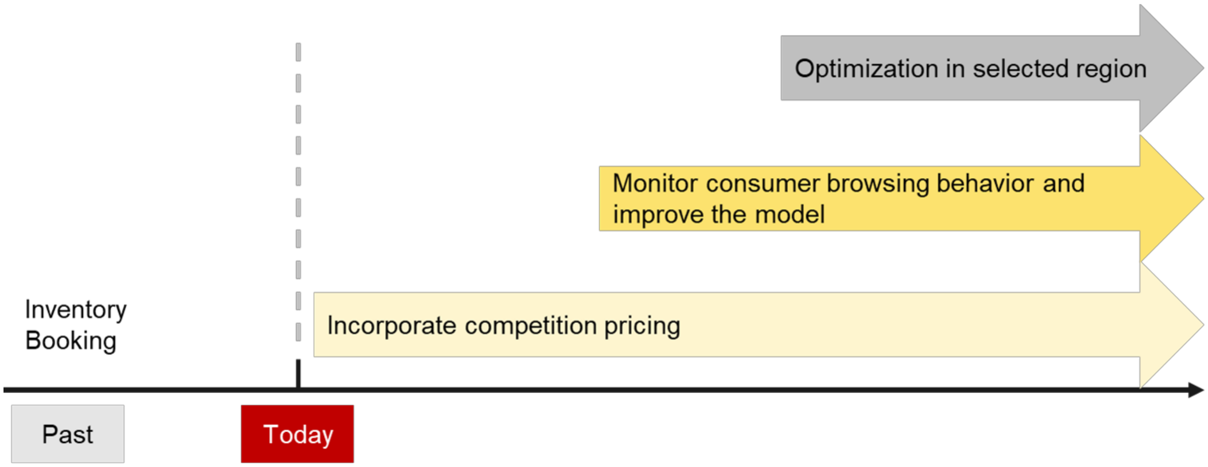
Optimization in selected region
Monitor consumer browsing behavior and improve the model
Inventory Booking
Incorporate competition pricing
Past
Today

Figure 9: Plan for future deployment

7. Conclusion

In this study, the main business objective is to scrape external data and merge it with internal price/demand data with the goal of providing insights to influence pricing decisions. The business problem was transformed into three analytical approaches. First, set up an external competition data collection mechanism. Second, incorporate external information with internal information to build and evaluate a predictive model that predicts the probability whether a unit will be booked/sold or not. Finally, leverage the predictive model as the input to an optimization model. We demonstrate that the web crawler, the predictive model, and the optimization model can be used separately or used as a streamlined solution. Further research and testing would likely improve the applicability of this being used for decision-making.

References

Aziz, H. A., et al. (2011). "Dynamic room pricing model for hotel revenue management systems." Egyptian Informatics Journal **12**(3): 177-183.

Briggs, B., et al. (2019) Beyond the digital frontier: Mapping your future Digital transformation demystified.

Erdem, M. and L. Jiang (2016). "An overview of hotel revenue management research and emerging key patterns in the third millennium." Journal of Hospitality and Tourism Technology **7**(3): 300-312.

Guo, X., et al. (2013). "Optimal pricing strategy based on market segmentation for service products using online reservation systems: An application to hotel rooms." International Journal of Hospitality Management **35**: 274-281.

Kimes, S. E. (2011). The future of hotel revenue management. Journal of Revenue and Pricing Management, 10(1), 62-72.

Ling, L., et al. (2012). "Optimal pricing strategy of hotel for long-term stay." International Journal of Services Technology and Management **17**(1): 72-86.

Newman, D. (2017) Why The 'As-A-Service' Model Works So Well For Digital Transformation. Forbes.com. Retreived from https://www.forbes.com/sites/danielnewman/2017/06/27/why-the-as-a-service-model-works-so-well-for-digital-transformation/#359743b16490

Tse, T. S. and Y. Tung Poon (2012). "Revenue management: resolving a revenue optimization paradox." International journal of contemporary hospitality management **24**(4): 507-521.

Walker, J. R. and J. T. Walker (2004). Introduction to hospitality management, Prentice Hall Upper Saddle River, NJ.

Ye, P., et al. (2018). Customized Regression Model for Airbnb Dynamic Pricing. Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, ACM.