**Mitigating Hotel Revenue Management System Risk Using Anomaly Detection in Short Booking Windows**

Pranoy Bhattacharya, Pratyusha Gajavalli, Surya Sathiyamurthy, Rishabh Sharma, Evan Tiffany, Matthew A. Lanham

Purdue University, Department of Management, 403 W. State Street, West Lafayette, IN 47907

sharm451@purdue.edu; pgajaval@purdue.edu; ssathiya@purdue.edu; bhatta56@purdue.edu; etiffany@purdue.edu; [lanhamm@purdue.edu](mailto:lanhamm@purdue.edu)

**Abstract**

We investigate forecasting and anomaly detection approaches to support improved pricing decisions for economy hotels. The motivation for this research is that in the economy hotel industry, the most frequent hotel booking time often occurs within 48 hours prior to arrival. This two-day booking window, often due to last-minute stops when traveling, makes planning more challenging compared to non-economy hotel chains. From the hotel’s perspective, they want to satisfy demand at a price that achieves the greatest revenue. It has been shown that traditional revenue management systems that first predict demand, then optimize price, perform well for non-economy hotels but have room for improvement in the short booking window situation. We show that the ability to identify and adjust price anomalies over time provides a revenue risk management layer that can help the business offer more competitive pricing and increase daily revenues. To achieve this, we utilized PCA and Clustering-based anomaly detection methods to signal anomalies within the 48-hour window using historical data. This optimal forecast is then integrated with a deviation-based anomaly detection layer to provide an alert system to the hotel for unexpected boosts and declines in demand which help adjust prices accordingly and mitigate risk.

**Keywords:** Anomaly detection**,** Alert system**,** Short-term demand forecasting, Price Optimization, Revenue Maximization

**Introduction**

The hospitality industry faces the unique challenge of implementing a pricing strategy within a dynamic system of demand: with cancelations showing at random and anomalies in demand due to externalities like changes in weather and special events. These externalities can hinder a business from profiting off this additional demand in the location. Moreover, the fluctuation in demand creates difficult labor usage prediction. The COVID-19 pandemic has exacerbated this fluctuation in demand, with cancelations of trips and events leading to an increase in cancelations in bookings as well. As a means of overcoming these two issues with a single solution, anomaly detection is being researched within this industry to promote a better forecasting of high traffic times and better prepare each residence for these abnormalities.

An overarching issue with demand forecasting in this industry is the effect the labor shortage has on the productivity of the bookings. As described in the Forbes article (Adams, 2021), the labor shortage has brought costly consequences to the industry. The article describes the issue as two-fold: individuals seeing the decrease in the purchasing power of the dollar in combination with workers feeling disheartened by the treatment they receive and feeling undervalued within the organization. With this, the industry has witnessed a decrease in labor, specifically at the lowest levels of the company. While the article focuses on “investing in its professionals and setting a new standard for wages at every position” (Adams, 2021), an additional route to investigate is a better understanding of demand forecasting leading to increased accuracy in predictions and better pre-detection of anomalies in the demand. This sentiment is echoed by the Wall Street Journal (2022), which describes the deep results of the pandemic on the hospitality industry: “People are going back to hotels. But with supply chain shortages holding up goods and workers quitting, the industry is having to figure out new ways to be hospitable.” Properly forecasting demand allows managers to properly staff, giving the few employees enough support to feel sustained, helping in keeping current staffing while continuing to hire new staff.

Pricing strategies within the hospitality industry have become much more dependent on data (Deloitte, 2017) as this can help competitively price bookings and increase revenue. “By automating data collection and using analytics to track customer preferences, hospitality companies can help improve staff interactions with guests at every touchpoint.” (Deloitte, 2017). By automating many processes and increasing data on traffic, companies within this industry can not only understand their guests better but provide them with a better experience during their stay. This understanding of the importance of data within the industry provides a small bridge toward utilizing historical data to better predict demand and therefore optimize the pricing of rooms within a chain. Using this data will help companies within this industry better forecast demand and detect anomalies before their occurrence to offer better service to guests and properly price their bookings based on these predictions.

The necessity for data in the reinvigoration of the hospitality industry is reiterated by Deloitte UK (Abdelkodos, P., & Cheng, J., 2021). “Getting forecasting right assists with financial and operational decision-making and can mean leveraging a slim window of opportunity or missing out entirely.” Businesses within the industry need to intertwine data into their recovery process to optimize profitability in a time where every dollar counts. Utilizing forecasting to assess revenue management in terms of pricing and operational cost analysis will promote the comeback of the industry.

A literature review outlining the best practices in forecasting and anomaly detection of time series data will follow this section. This will be followed by the methodology. Section four will outline the various methods utilized, comparing the results of these practices. Section five will follow with the performance of these models. Finally, Section six will conclude the paper, providing future suggestions and insight from the process, as well as the accomplishments of the research.

**Literature Review**

In assessing literature in the scholarly realm, the focus was on primarily predicting anomalies in an environment with volatile demand. Additionally, it was vital to find industry research as well, addressing the commonalities between hotel industry demands. Overall, five papers were analyzed providing a thorough analysis preparing different aspects of our project.

Unsupervised Anomaly Detection in Time Series Using LSTM-Based Autoencoders (Provotar, Linder, Veres, 2019) focuses on explaining the process and advantage of using autoencoder-based methods for anomaly detection. This paper also discusses how to choose an algorithm based on the nature of the data provided. It highlights differences among classical machine learning methods, time series methods, and neural networks methods catering to different use cases. It describes the detailed architecture and algorithm of One class SVM mode1, Loess decomposition model, and autoencoder-based LSTM model and compares their performance on an artificial signal dataset as well as a rare sound events dataset. The study proved that the autoencoder model worked best on time series data with no clear period. This paper helped us understand the importance of knowing the nature of data available and the type of anomaly required to be detected while choosing the algorithm, but the algorithms discussed in this paper were tested on data that is different from the demand data of the hotel industry.

Time Series Anomaly Detection: Detection of Anomalous Drops with Limited Features and Sparse Examples in Noisy Periodic Data (Shipmon, Gurevitch, Piselli, Edwards, 2017) focuses on detecting sustained anomalies in periodic, but noisy, traffic patterns of Google. As labeled data for anomalies was not provided, the study performs regression and predicts the expected value first, then creates anomaly detection rules that compare actual values with the predicted values. To predict expected values, they explored a TFLearn Deep Neural Network (DNN) Regressor model with Relu6 activation function and Adam optimizer, a Recurring Neural Network (RNN) model with Exponential Linear Units (ELU) activation and Adam optimizer, Long Short-Term Memory (LSTM) model with ELU activation and Adam optimizer and Fourier series model with Adagrad optimizer. To detect anomalies, they explored the Accumulator method with fixed threshold and variance-based algorithms and Gaussian Tail Probability methods. It is observed that the hybrid model, a mixture of both anomaly detection models, applied over predictions from Fourier and RNN models is most effective with the fewest false positives while also providing the lowest MSE loss on the validation dataset. This study helped us understand the potential methodology to be adopted for anomaly detection in time series, but it is restricted to detecting sustained anomalies rather than short or momentary ones. Since short anomalies are also equally trivial to us as we are using the number of bookings per day data, further exploration is required.

Anomaly Detection for Univariate Time-Series Data (Chakravarty, Dabiri, Nayyar, Vishwasrao, 2015) describes three patterns of anomaly detection: threshold, change in pattern, and change in frequency. The researchers implemented different algorithms to detect each of these anomaly types. The point anomaly algorithm utilizes standard deviations from a mean value to extract any abnormalities in the data. This approach, therefore, assumes normality of the data but works well in finding short-term abnormalities. In this method, alpha must be set to provide a standard for the degree of abnormality. A change in pattern can utilize statistical methods to notice changes in means to bring these anomalies to the forefront. This method sets a window to find an average with k points in each window. K will be determined before the process to notice these abnormal mean periods. The final abnormality is a change in frequency within the data. To detect this type of anomaly, the researchers utilized an F1 score to determine these anomalies. The detection rates for each were 0.68, 0.68, and 0.51 respectively in terms of accuracy and have provided insight into beneficial methods of detecting different types of anomalies.

Demand forecasting within this industry takes form in different ways, with no fully correct method. Demand Forecasting Model Using Hotel Clustering Findings for Hospitality Industry (Kaya, Yilmaz, Yaslan, Öğüdücü, Çıngı, 2022) demonstrate the value in clustering prior to forecasting demand, an approach to be considered. As clusters may not be structurally defined within this industry or assigned randomly, the initial clustering assignment helps provide some indication into naturally forming clusters in terms of the data. This, therefore, allows optimal binning in the forecasting process to provide better-forecasted results within each cluster. In testing this process, the team utilized K-means clustering, hotel embedding, and an Attention LSTM model, comparing this against a base model as well as a plethora of ensemble and deep-learning models. The Attention LSTM model outperformed the others, providing a lower mean absolute error.

Anomaly Detection: Predicting hotel booking cancellations (Timamopoulos, 2020) outlines types of anomalies, detection of these anomalies, and classification methods. Beginning, the paper emphasizes three anomaly types: point anomalies, contextual anomalies, and collective anomalies. Point anomalies cover short periods, for example, a contextual example of a concert that increases demand for a few days. A contextual anomaly consists of an abnormality in data-dependent on the time. Say a hotel can expect an average of 20 bookings a night during the summer season, but only 10 a night during typical school months, seeing 20 bookings during the slower periods would be an anomaly at that time. Finally, collective anomalies represent a large group of anomalous data, say booking rates at the start of Covid-19. In detecting these different types of anomalies, the study utilized Logistic Regression, Naïve Bayes, KNN, SVM, Decision Tree, Random Forest, Gradient Boosting, and XG Boost. The study concluded that a fine-tuned decision tree helped predict specifically booking cancellations in this case but aided in the anomaly detection process for this concept.

|  |  |  |  |
| --- | --- | --- | --- |
| Authors | Title | Reasons for study | Conclusions |
| Provotar, Linder, Veres | Unsupervised Anomaly Detection in Time Series Using LSTM-Based Autoencoders | Proving the usefulness of autoencoder LSTM models. | Autoencoder LSTM models outperformed the SVM and Loess models. This type of model works well with data without defined periods. |
| Shipmon, Gurevitch, Piselli, Edwards | Time Series Anomaly Detection: Detection of Anomalous Drops with Limited Features and Sparse Examples in Noisy Periodic Data | Inquiring about the model choice in anomaly detection over long periods. | The hybrid method, utilizing accumulation and Gaussian tail probabilities provided the best anomaly detection for long-term anomalies. |
| Chakravarty, Dabiri, Nayyar, Vishwasrao | Anomaly Detection for Univariate Time-Series Data | Finding the optimal detection methods for anomalies of different types. | Utilizing normal distribution ranges, the K-means process, and F1 scores were found to be successful for three different types of anomalies. |
| Kaya, Yilmaz, Yaslan, Öğüdücü, Çıngı | Demand Forecasting Model Using Hotel Clustering Findings for Hospitality Industry | Quantifying the benefits of clustering in demand forecasting for the hotel industry. | Clustering in combination with the LSTM model provided the best results. Clustering helped the model find patterns in the data that the hotel chain was unable to find alone. |
| Timamopoulos | Anomaly Detection: Predicting hotel booking cancellations | Finding the optimal algorithm for determining anomalies in booking cancellations. | A fine-tuned decision tree provided the best results in this study for this type of anomaly. |

**Data**

The data used in the study belongs to a hotel company and is divided into three datasets. The first dataset focuses on property information and has property details such as property status, brand, and location. The second dataset consists of room level details for each property such as room type, bed count, and facilities available. This data combined with the property information will make up the inventory. The last dataset contains annual reservations from the year 2015 through 2021 and consists of booking features like the number of tenants, date of stay, booked rate, booking number, and loyalty membership status. An additional dataset containing rate structure was shared to show the differences between rate plans.

Preprocessing and exploratory data analysis utilized each of these datasets to identify the most relevant features for the model. Key features used are PropertyCode, LocationType, RevenueManger, Manager\_email\_id, PropertyClassType, CreationDateTime, StayDate, and Nights.

*Table 1:Property Information Data*

|  |  |  |
| --- | --- | --- |
| **Variable** | **Data Type** | **Description** |
| PropertyCode | varchar | Unique property ID for hotel |
| PropertyDescription | nvarchar | Property Name |
| IsActive | bit | 0 = property is not active in system 1 = property is active in system |
| PropertyStatus | varchar | Current Status of property |
| Brand | varchar | Associated sub-brand |
| Portfolio | nvarchar | Ownership Group |
| Street1 | nvarchar | Street Address of hotel |
| City | nvarchar | City of hotel |
| StateProvince | nvarchar | State of Hotel |
| PostalCode | nvarchar | Postal Code of hotel |
| CountryCode | varchar | Country Code of hotel |
| Country | varchar | Country of hotel |
| Latitude | decimal | Latitude of hotel |
| Longitude | decimal | Longitude of hotel |
| PhoneNumber | varchar | PhoneNumber of hotel |
| BuildYear | int | Year hotel was built |
| LastRenovationYear | int | Year of the last renovation at the hotel |
| OpenDate | date | Original open date of the hotel within the brand |
| RevenueManager | varchar | Currently assigned RevenueManager |
| RegionName | varchar | Currently assigned RegionName |
| RvpoName | varchar | Currently Assigned RvpoName of Ops |
| RegionDescription | varchar | Currently Assigned RegionDescription |
| DivisionName | varchar | Currently Assigned DivisionName |
| SvpName | varchar | Currently Assigned SVP or Ops |
| CityRegionName | varchar | Associated STR Market - Geographic area typically made up of a Metropolitan Statistical Area |
| SubmarketName | varchar | Associated STR Sub Market - Geographic area that is a subset of a market |
| CountryRegionName | varchar | Metropolitan Statistical Area - MSA is a geographic entity defined by the Office of Management and Budget (OMB) for use by federal statistical agencies in the collection |
| LocationType | varchar | Property classification is driven by physical location regardless of amenities or services offered. Location types include:  Location Segments and Types include: Urban: Densely populated location in a large metropolitan area.  Airport: Hotel near an airport that primarily serves demand from airport traffic. Distance may vary. Interstate/Motorway: Property near major highway, motorway, or other major roads with the primary source of business via passerby travel.  Hotels located in suburban areas have the suburban classification. Resort: Property located in a resort area or market where a significant source of business is derived from leisure/destination travel |
| TimezoneName | varchar | TimezoneName of hotel |
| TimezoneUtcOffset | int | The difference in hours between UTC and local time |
| RoomCount | int | Count of daily inventory currently available to sell at the hotel.  Includes out of order rooms |
| CurrencyCode | varchar | CurrencyCode of hotel |
| FirstStayDate | date | FirstStayDate recorded for hotel |
| LastStayDate | date | LastStayDate recorded for hotel |
| LastAuditDate | datetime2 | LastAuditDate recorded for hotel |
| PropertyClassType | nvarchar | The distinction between a property being an extended say hotel or not |
| Tenure | nvarchar | Categorized times (in months) since the property opened in a brand |
| DualLocation1 | varchar | Linked PropertyCode if Dual Location |
| IsDualBrand | varchar | FALSE = property is not a dual-branded location TRUE = Property is a dual-branded location |

Table 2: Summary of Room Category

|  |  |  |
| --- | --- | --- |
| **Variable** | **Data Type** | **Description** |
| property | varchar | Unique property ID for hotel |
| RoomTypeCode | varchar | Room type code |
| RoomTypeDescription | varchar | Room type description |
| BedCount | int | No. of beds in the room |
| IsSmoking | bit | 0 = Non-Smoking 1 = Smoking |
| IsAccessible | bit | 0 = No 1 = Yes |
| RoomType | varchar | Description of room type |
| BedType | varchar | Description of bed type |
| room | int | No. of rooms present for a particular room type |
| active | bit | 0 = No 1 = yes |
| handicap | bit | 0 = handicap access, not present 1 = handicap access present |
| floor | int | floor number |
| lastupdated | datetype | data last updated |

Table 3: Annual Reservations Data

|  |  |  |
| --- | --- | --- |
| **Variable** | **Data Type** | **Description** |
| Property | varchar | Unique property ID for hotel |
| AccountUid | varchar | Unique reservation ID |
| Adult | tinyint | number of adults – an additional charge may apply |
| Child1 | tinyint | number of infants - no additional charges |
| Child2 | tinyint | number of infancts - no additional charges |
| HasPet | bit | 0 = no pets ; 1 = has pets. |
| CreationDateTime | datetime2 | UTC date and time of which entry was created |
| StayDate | date | Date of reservation and date of business |
| ArrivalDate | date | Date of arrival for reservation |
| DepartureDate | date | Date of departure for reservation |
| CancelNumber | varchar | Unique cancel ID |
| CancelDateTime | datetime2 | UTC date and time of which reservation was canceled |
| CheckInDateTime | datetime2 | UTC date and time of which guest checked in to the hotel |
| CheckOutDateTime | datetime2 | UTC date and time of which guest checked out of the hotel |
| LastUpdateDateTime | datetime2 | UTC date and time of which entry was last updated |
| Status | nvarchar | Current status of reservation   Reserved - reservation made In Houseguest is currently checked in  Checked Out - guest has checked out Canceled - guest has canceled the reservation Wait List - guest is on the waitlist - typically due to limited inventory No Show - guest did not check-in or cancel the reservation. May or may not generate revenue based on hotel policies. |
| IsLoyaltyMember | bit | 0 = no Loyalty Member ID attached to the reservation ; 1 = Loyalty Member ID attached to the reservation |
| IsGroup | bit | 0 = no Group ID attached to reservation ; 1 = Group ID attached to reservation |
| GroupUid | varchar | unique group id |
| Source | nvarchar | Booking distribution channel. |
| SourceCategory1 | varchar | Level 1 of Source Category Hierarchy Source categories were created for reservations starting in 2019. Reservations prior to 2019 may have NULL level 1 source category. See Source if the value is NULL |
| SourceCategory2 | varchar | Level 2 of Source Category Hierarchy |
| SourceCategory3 | varchar | Level 3 of Source Category Hierarchy |
| Company | nvarchar | Name of Company for reservation |
| IsVolumePlus | bit | 0 = no Volumn Plus ID attached to reservation ; 1 = Volume Plus ID attached to reservation  Volume Plus is Corporate Negotiated accounts |
| IsTravelAgency | bit | 0 = no Travel Agency ID attached to reservation ; 1 = Travel Agency ID attached to reservation |
| Rateplan | varchar | The booked rate plan for StayDate. Rateplan can change over duration of a reservation |
| RateplanTier | varchar | Associated tier for rateplan: BASE - BASE RATES DY1 - DYNAMIC YIELD LEVEL 1  DY2 - DYNAMIC YIELD LEVEL 2 GROUP -GROUP TIER HOT - HIGH OCCUPANCY TIER NEGOTIATED - NEGOTIATED TIER |
| Roomtype | varchar | Room Desciption Code. Not unique to the hotel |
| Room | varchar | Room ID. Unique to hotel |
| BookedRate | decimal | Booked rate for StayDate |
| Nights | int | Count of StayDates |

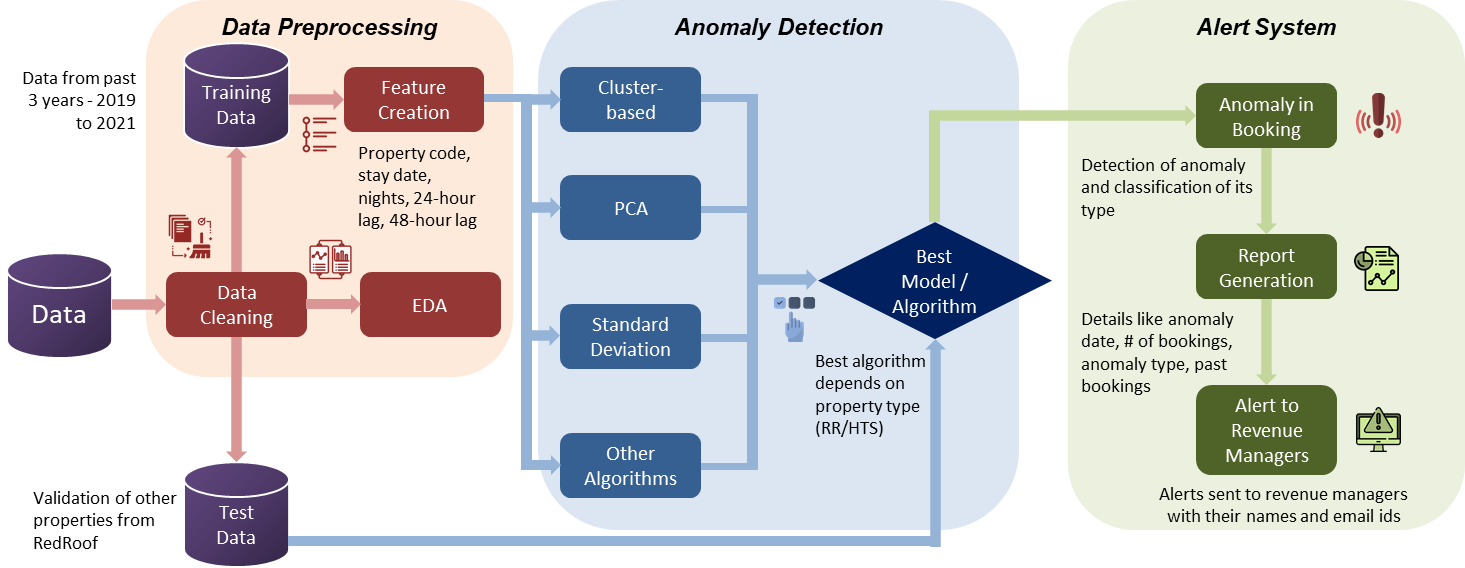
**Methodology**

The methodology used in our project is divided into various steps from Data Cleaning and partitioning to Email Notification based Alert systems. The flow of the methodology used is sequential and at the final step, the team of revenue managers across each hotel would receive an email alert based on the outlier detected in the booking pattern along with the details like booking date, and the number of bookings. The core main parts of the methodology are the implementation of

* Outlier Detection Model
* Email - Alert System

The graphical representation of the methodology workflow can be found below in Figure 1.

*Figure 1: Methodological approach*



1. **Data Cleaning**

The total number of bookings per day was calculated using annual reservations data (Table 3) aggregated on a day-by-day basis. It was necessary as each booking contained multiple observations, and forecasting using this data would be erroneous as the anticipated booking would not represent the overall booking for that day. We use the following formula to calculate the 48-hour booking per day from the data:

**CreationDateTime – min(StayDate, ArrivalDate) < 2**

1. **Data Partitioning**

After cleaning the data, the training data from a few specific properties were used to train different anomaly detection models. Different outlier/anomaly detection algorithms seemed to work well for different types of properties based on their location, demand, etc. To avoid getting skewed results from time series based on the recent COVID-19 pandemic, we handpicked a few properties at random and trained the anomaly detection model. Data selected for testing or validating the anomaly detection model were from other properties that were not a part of the testing data set.

1. **Feature selection for different types of models**

Two different features were considered for respective training and validation datasets. These features are the number of bookings and the dates for these bookings. These were tabulated in such a way that each row, from the timeframe of interest, had the date and the total number of bookings for that particular date. This was needed in order to create a time series plot for training the anomaly detection models. The X-axis consisted of the timeline while the Y-axis consisted of the number of bookings. This time series plot helps the model in determining which particular day was an anomaly in terms of that day having significantly higher bookings or lower bookings.

1. **Anomaly Detection Model Comparison**

To determine the best-performing model, we utilized an accurate measurement of the true positive rate between our 12 proposed models. This approach was done for both types of hotels within the given dataset to test any differences in model types by hotel type.

1. **Email based Alert System**

After the anomaly is detected in a 48-hour window or other windows of interest, a report is generated that has the following details

* A snapshot of the time-series with markers indicating the anomalies
* The dates of the anomalies detected in the time window of interest
* The number of bookings associated with those dates
* Priority of the anomaly – high, medium, low
* Details of the property like name and location
* An in-depth detail of all the bookings in the past 45 days

After this report is generated, it is emailed to the people of interest like the manager of the hotel, the revenue manager of a group of hotels, etc. After receiving these reports, the concerned people can take necessary action in order to increase or decrease the price of that particular hotel property depending on the type of anomaly being detected. This would help the hotels maximize their revenue as they can make of the high demand and increase the pricing of their hotel rooms accordingly.

**Models**

In all, we tested the use of twelve different models to select the optimal model in anomaly detection. Our data utilized two different types of hotels; therefore, we tested each method on each type of hotel to form the conclusions. The models included an isolation forest, angle-based outlier detection, KNN, clustering-based outlier detection, connectivity-based outlier detections, histogram-based outlier detection, local outlier factor, one-class SVM, and minimum covariance determinant, principal component analysis, subspace outlier detection, and stochastic outlier detection.

Isolation forests utilize binary trees to detect abnormalities within the data through similarity clustering. The clusters allowed for easy use for different types of bookings and provide a good strategy for grouping and detecting abnormalities. Some abnormalities, however, were found within these clusters, showing a downside in the anomaly detection.

Angle-based outlier detection utilizes angles between points in a plane to detect those points that are determined to be outliers. The process assesses the degree of angles between three points, with the vertex being the point of the question. The degree helps identify dissimilar points. This model works well with high dimensionality; however, the provided data was not high in dimension. This led to less success in this analysis.

KNN modeling is an unsupervised approach to anomaly detection. With KNN, analysis is able to be done without labeling a point anomalous or not. It provides a way to detect anomalies without having a training set, allowing us to perform the model with all the data. A con is that performance cannot fully be measured and there’s no set model.

Clustering-based outlier detection is similar to isolation forests. This method provides dense clusters and classifies all points outside of these clusters as anomalous. This method is easy to visualize and interpret. With the data provided, however, an anomaly may fit into one of the clusters simply dependent on the seasonality and be easily missed.

Connectivity-based outlier detection determines a degree of an outlier for each point and uses that degree as a metric to classify outliers. The method requires labels and provides a set model for future use and allows the company to determine the degree they determine to be an outlier. This method can also easily miss some seasonality-dependent outliers.

Histogram-based outlier detection is a typically used anomaly detection. This finds histogram-based off features and finds anomalies within each feature-based histogram. It can provide an easy and quick method of finding anomalies. With the data provided, this approach doesn’t detect anomalies very well, specifically those anomalies at the lower end.

The local outlier factor is another unsupervised method that utilizes an anomaly score through a density deviation of surrounding points. Similar to the connectivity-based approach, the company can set a threshold for anomalies to provide variability in the degree of the anomaly, permitting the company to adjust its approach based on the intensity of the anomaly. This can be unstable with anomaly predictions given the data we’re using.

Another unsupervised anomaly detection method is one-class SVM. This method provides a robust analysis, understanding that there could be much fewer anomalous points than non-anomalous points. Additionally, the non-anomalous points are not required to have commonalities. It provides a novel method of determining outliers, but the lack of cohesion between non-anomalous points affects the performance of the model.

Minimum covariance determinant anomaly detection forms a multidimensional Gaussian distribution as a means to classify anomalies. The multidimensional aspect provides a robust determination of anomalous points. This method reduces the effect that outliers have on the dataset to better determine which points are outliers. This can overpredict outliers, though.

Principal component analysis edits a coordinate system to a set of normal values and determines outliers on the new plane. An anomaly score is determined as a metric away from the normal score and the anomalies are then determined. The process provides a preventative measure for sporadic anomalies.

Subspace outlier detection utilizes subspace clustering to find anomalous data. The process first finds underlying subspaces within the data then clusters within each subspace, and finally clustering all the clusters together. Similar to angle-based outlier detection, this method works well with high-dimensionality data but has some short-comings with the data provided due to the small dimensions.

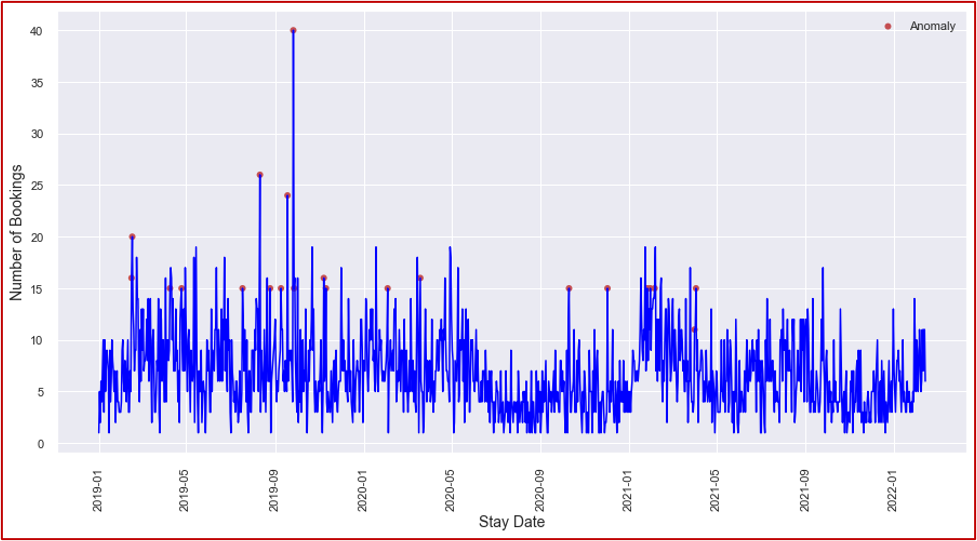
Stochastic outlier detection is the final model tested. It utilizes model fitting to find low probability data points in the data. This provides another unsupervised approach to outlier detection that performs very well in most datasets.

Testing on this range of models allowed us to determine the benefits of each model within the scope of our dataset specifically, allowing us to find optimal anomaly detection with datasets on booking numbers. We hoped to deliver a variety from the beginning of the process to easily narrow our focus when utilizing the model on the chain hotels wholly.

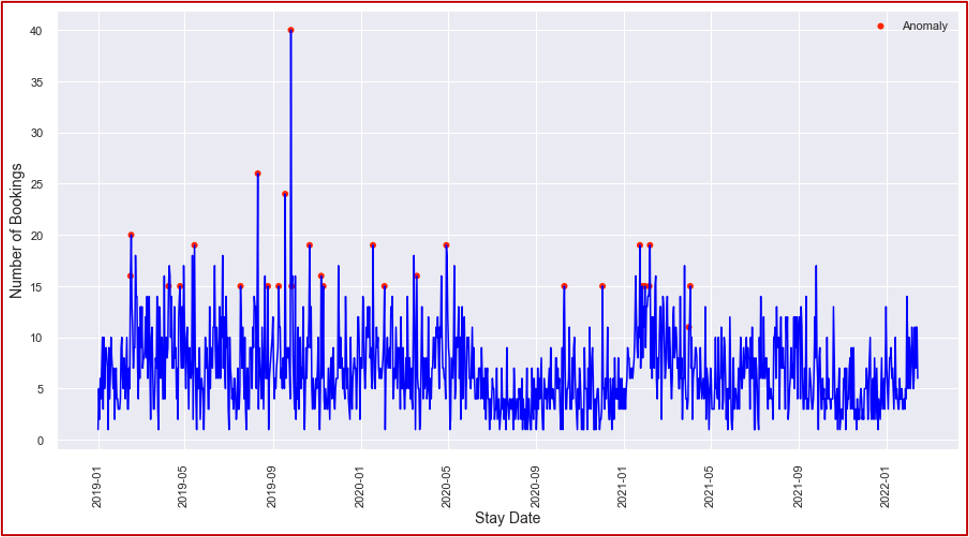
**Results**

In our model selection, we stratified the process, allowing for optimal results to be found for the different property types. Due to different operational strategies, the results and optimal models differed slightly for the different property types (there were 2 primary types of property based on the different business needs, one was used for shorter business stays with fewer amenities whereas the other was used for longer stays and had more amenities). Beginning with the shorter-stay category, we found success initially through the utilization of PCA. The strategy grasped the high demand anomalies very well but noticeably captured a number of false positives as well. To overcome this, as the model would alert management an unnecessary number of times, creating distrust for the system, we added a secondary level that omits anomalies found within 0-1 standard deviations from the mean and included any missed anomaly over three standard deviations above the mean.

*Figure 2: Model of anomalies using PCA only*



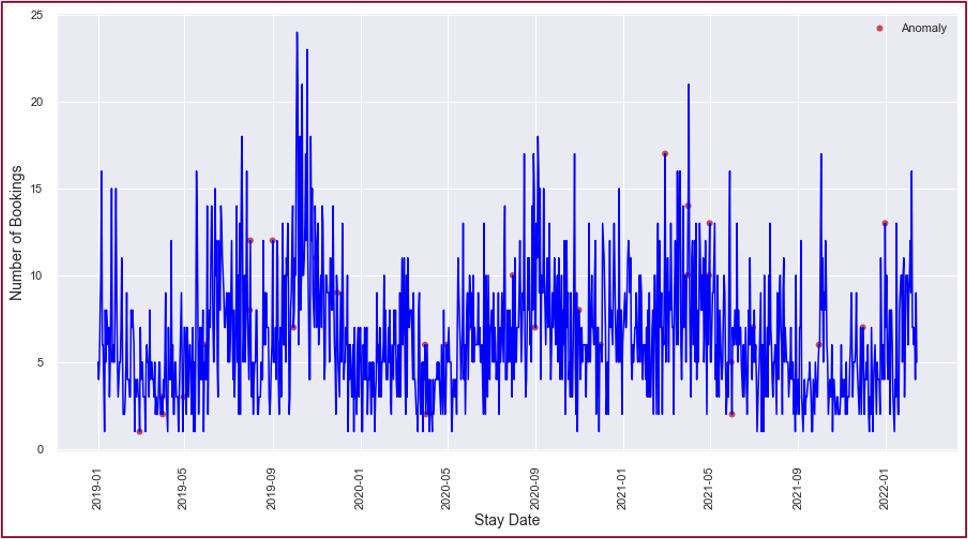
*Figure 3: Model of anomalies using PCA and standard deviations*



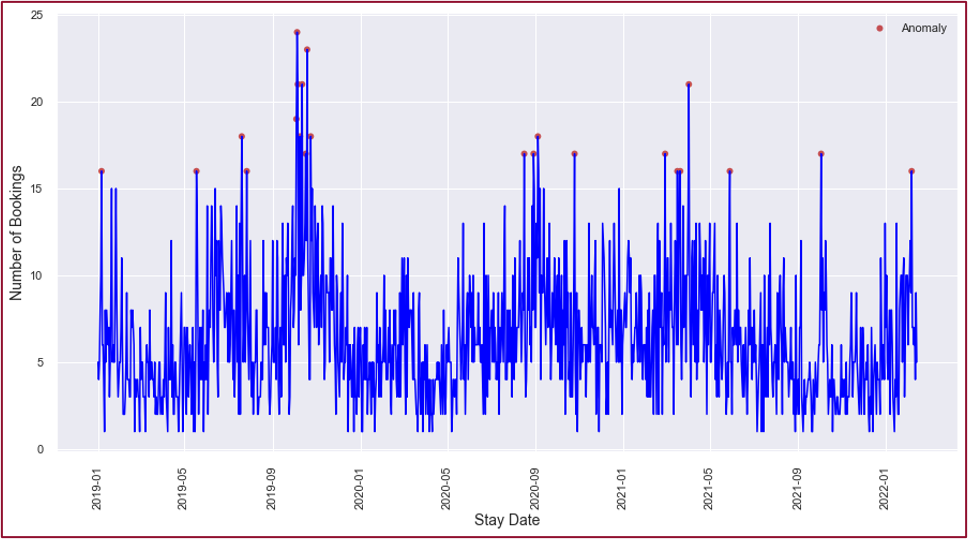
The figures above show the anomalies detected using both methods. In Figure 3, compared with Figure 2, you can see the added benefit of the standard deviation layer in finding anomalies and avoiding over-alerting the chain with false positives.

In assessing the property type used for longer stays, the same initial models were tested and compared. Results were split when comparing the performance of the models, so a combination of clustering-based anomaly detection and PCA was implemented in order to provide the best results. Moreover, the results retained a lot of false positives while missing some obvious anomalies. With this, the standard deviation requirements used for the shorter-stay modeling were implemented again within the longer-stay properties.

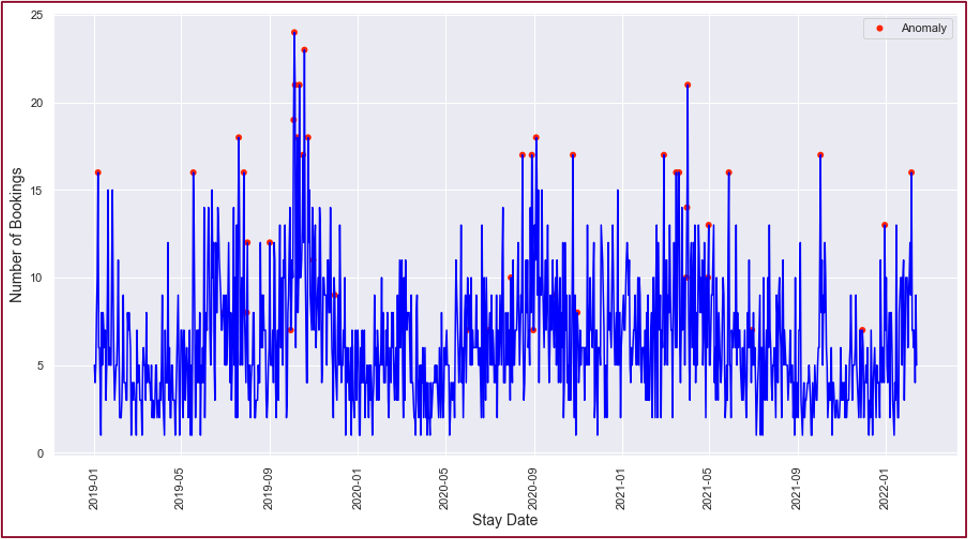
*Figure 4: PCA anomaly detection*



*Figure 5: Clustering-based anomaly detection*



*Figure 6: Clustering-based and PCA anomaly detection with standard deviation layer*



The above figures show the performance of the two base layers as well as the detection provided in the complete model. As seen in comparisons, the final composite model provides detection of the high anomalous data well and limits the number of false positives predicted in the model, allowing for appropriate detection of high demand without alerting numerous times for falsely predicted anomalous data.

**Conclusions**

With the implementation of anomaly detection as an addition to demand forecasting, the economy hotel industry is able to substantially improve its revenue while simultaneously providing better preparation to meet the true demand on its daily booking rates. We found a layered model that helped encounter the many types of anomalous data present within this industry. We found through research on anomaly detection that results will vary dependent on the type of data. Within the economy hospitality industry, our optimal method was the combination model for the 48-hour window. From this strategy, a true positive rate of 83% on average across the sample subset was produced with a false positive rate of 1%. These statistics were found through visual and historical input from the originators of the data.

*Table 4: Results from anomaly detection*

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Property Count** | **Total Data Points** | **Actual Positive** | **Actual Negative**  **(Total - Actual + ve)** | **Detected Positives** | **True Positive (TP)** | **False Positive (FP)** |
| 11 | 10,522 | 175 | 10,347 | 238 | 146 | 92 |

Table 4 displays the performance of the two models to provide an assessment of the model. The model overall provides thorough attainment of true positives while limiting the number of false positives so as to not over-alert the company, therefore, decreasing the benefit and reliability to the company.

*Table 5: Revenue Analysis*

|  |  |  |  |
| --- | --- | --- | --- |
| **Alert Type** | **Number of Bookings** | **Actual Revenue** | **Increased Revenue** |
| **High** | 4615 | 100% | 50% |
| **Medium** | 1288 | 100% | 35% |
| **Low** | 69 | 100% | 20% |
| **Total Number of Bookings in Anomaly Days** | 5972 | 597200 | 277210 |

Table 5, above, outlines the revenue optimization setup utilized to show the monetary benefit of anomaly detection within the company. Assuming the increased revenue laid out in the table above, a YTD revenue increase of 2.6% to 3.5% was projected. It can also lead to increased revenue of 34% to 46% on the particular days these anomalies are detected. The missed revenue opportunity for the company is an area that would provide a significant competitive advantage over competitors within the economy hotel industry.

Future studies may look into the optimal anomaly detection method for a larger window, that may aid in the prediction of unknown planned events or simply unexpectedly high travel times. With this, additional functionality providing more autonomous feedback, such as including dynamic pricing strategies and staffing suggestions would provide higher functionality to the company and allow for optimized revenue attainment. Anomaly detection, wholly, can benefit the readiness of an economy hotel chain in pricing strategy and staffing, aiding in providing the best guest experience while also generating the highest possible profits for the company.

**References**

1. Abdelkodos, P., & Cheng, J. (2021, May 5). *Forecasting for success in the hospitality industry against a backdrop of uncertainty*. Deloitte United Kingdom. Retrieved April 1, 2022, from <https://www2.deloitte.com/uk/en/pages/consumer-business/articles/forecasting-for-success-in-the-hospitality-industry-against-a-backdrop-of-uncertainty.html>
2. Adams, C. (2021, October 4). *Council post: Why the hospitality industry must reinvent itself as labor shortage still looms*. Forbes. Retrieved April 1, 2022, from <https://www.forbes.com/sites/forbesbusinesscouncil/2021/10/04/why-the-hospitality-industry-must-reinvent-itself-as-labor-shortage-still-looms/?sh=5a09603317b6>
3. Chakravarty, Dabiri, Nayyar, Vishwasrao (2016). Anomaly Detection for Univariate Time-Series Data. <https://courses.cs.vt.edu/cs5824/Fall15/project_reports/nayyar_vishwasrao_chakravarty_dabiri.pdf>
4. Deloitte. (2017, May 30). *Maximizing revenue in Hospitality*. The Wall Street Journal. Retrieved April 1, 2022, from <https://deloitte.wsj.com/articles/maximizing-revenue-in-hospitality-1496116942>
5. Kaya, Yılmaz, Y., Yaslan, Y., Öğüdücü, Ş. G., & Çıngı, F. (2022). Demand forecasting model using hotel clustering findings for hospitality industry. Information Processing & Management, 59(1), 102816–. <https://doi.org/10.1016/j.ipm.2021.102816>
6. Provotar, Linder, Y. M., & Veres, M. M. (2019). Unsupervised Anomaly Detection in Time Series Using LSTM-Based Autoencoders. 2019 IEEE International Conference on Advanced Trends in Information Theory (ATIT), 513–517. <https://doi.org/10.1109/ATIT49449.2019.9030505>
7. Shipmon, Gurevitch, J. M., Piselli, P. M., & Edwards, S. T. (2017). Time Series Anomaly Detection; Detection of anomalous drops with limited features and sparse examples in noisy highly periodic data.
8. Timamopoulos, Christos. (2020). Anomaly Detection: Predicting hotel booking cancellations. [Timamopoulos\_Anomaly\_Detection\_thesis.pdf (ihu.edu.gr)](https://repository.ihu.edu.gr/xmlui/bitstream/handle/11544/29631/Timamopoulos_Anomaly_Detection_thesis.pdf?sequence=1#:~:text=cancellation%20prediction%20as%20a%20supervised%20anomaly%20detection%20concept%2C,to%20forecast%20booking%20cancellations%20with%20overall%20accuracy%2099%25.)