

RESEARCH PAPER

An Automated Machine Learning Based Decision Support System to Predict Hotel Booking Cancellations

Nuno Antonio¹, Ana de Almeida² and Luis Nunes³¹ ISCTE-IUL, Instituto de Telecomunicações, PT² ISCTE-IUL, CISUC, ISTAR-IUL, PT³ ISCTE-IUL, Instituto de Telecomunicações, ISTAR-IUL, PTCorresponding author: Nuno Antonio (nuno_miguel_antonio@iscte-iul.pt)

Booking cancellations negatively contribute to the production of accurate forecasts, which comprise a critical tool in the hospitality industry. Research has shown that with today's computational power and advanced machine learning algorithms it is possible to build models to predict bookings cancellation likelihood. However, the effectiveness of these models has never been evaluated in a real environment. To fill this gap and investigate how these models can be implemented in a decision support system and its impact on demand-management decisions, a prototype was built and deployed in two hotels. The prototype, based on an automated machine learning system designed to learn continuously, led to two important research contributions. First, the development of a training method and weighting mechanism designed to capture changes in cancellations patterns over time and learn from previous days' predictions hits and errors. Second, the creation of a new measure – Minimum Frequency – to measure the precision of predictions over time. From a business standpoint, the prototype demonstrated its effectiveness, with results exceeding 84% in accuracy, 82% in precision, and 88% in Area Under the Curve (AUC). The system allowed hotels to predict their net demand and thus making better decisions about which bookings to accept and reject, what prices to make, and how many rooms to oversell. The systematic prediction of bookings with high probability of being canceled allowed hotels to reduce cancellations by 37 percentage points by acting to avoid their cancellation.

Keywords: A/B testing; data science; decision support systems; machine learning; predictive analytics; revenue management

1 Introduction

Revenue management – “the application of information systems and pricing strategies to allocate the right capacity to the right customer at the right price at the right time” (Kimes and Wirtz, 2003, p. 125) – was originally developed in 1966 in the aviation industry (Chiang et al, 2007) and has been gradually implemented in other service industries, such as hotels, rental cars and golf courses (Chiang et al, 2007; Kimes and Wirtz, 2003). The hospitality industry adapted the former definition into: “making the right room available for the right guest and the right price at the right time via the right distribution channel” (Mehrotra and Ruttley, 2006, p. 2). To comply, hotels accept bookings in advance. A booking symbolizes a contract between the customer and the hotel, which gives the customer the right to use the service in the future at a settled price. Usually, an option to cancel the contract prior to the service provision is included. However, the option to cancel a service prior to its provision places all risk on hotels, which have to guarantee rooms to customers who honor their bookings but also bear the cost of vacant rooms when a booking is cancelled or a customer does not show (Talluri and Van Ryzin, 2005). Consequently, cancellations have a significant impact on demand-management decisions in the context of revenue management.

Although accurate forecasts are a critical tool in terms of revenue management performance, forecasts are evidently affected by cancellations (Talluri and Van Ryzin, 2005). Cancellations can represent 20% of the total bookings received by hotels (Morales and Wang, 2010). This value increases to 60% in the case of airport/roadside hotels (Liu, 2004). In an attempt to balance losses, hotels resort to the implementation of overbooking strategies and restrictive cancellation policies (Mehrotra and Ruttley, 2006; Smith et al, 2015; Talluri and Van Ryzin, 2005). However, these demand-management decisions can negatively impact on the hotel's revenue and social reputation. Overbooking can cause a hotel to deny service provision to a customer, which indisposes the customer toward the hotel and generates costs for relocation of the customer to an alternative hotel (Noone and Lee, 2011). This relocation may also introduce the customer to a hotel that he/she may enjoy and cause a loss of future reservations from the customer (Mehrotra and Ruttley, 2006). Restrictive cancellation policies, particularly non-refundable policies and policies with 48 hours advance cancellation deadlines (Smith et al, 2015) decrease not only revenue due to the application of significant price discounts but also the number of bookings (Chen et al, 2011; Smith et al, 2015).

Several studies address topics related to the methods employed to moderate the consequences of cancellations in revenue and inventory allocation, cancellation policies and overbooking strategies (Ivanov, 2014; Hayes and Miller, 2011; Talluri and Van Ryzin, 2005). The majority of published studies focus on the airline industry, which differs from the hospitality industry in a considerable number of characteristics (Freisleben and Gleichmann, 1993; Garrow and Ferguson, 2008; Hueglin and Vannotti, 2001; Lemke, 2010; Subramanian et al, 1999; Yoon et al, 2012). But, the number of studies related to the hospitality industry has been increasing, demonstrating the importance of the topic for this industry (Caicedo-Torres and Payares, 2016; Weatherford and Kimes, 2003; Zvi Schwartz et al, 2016). The majority of those studies employ traditional statistics methodologies, and only a few take advantage of machine learning methodologies and techniques (Caicedo-Torres and Payares, 2016; Freisleben and Gleichmann, 1993; Hueglin and Vannotti, 2001). The same comment applies to research on demand forecasting to predict cancellations where, despite the existence of several studies on the subject, only four studies are specific to the hospitality industry (Antonio et al, 2017b; Huang et al, 2013; Liu, 2004; Morales and Wang, 2010). Moreover, only three studies use hotel specific data (Property Management Systems – PMS data) (Antonio et al, 2017b, 2017a; Liu, 2004). The remaining studies use Passenger Name Record (PNR) data, which are an airline industry standard created by the International Air Transport Association (International Civil Aviation Organization, 2010).

Most of the studies in booking cancellations prediction consider it as a regression problem. Only a few of the previously published studies approach the subject as a classification problem (Antonio et al, 2017a) and focus on global cancellation rate forecasting instead of each booking's cancellation probability (Antonio et al, 2017b). In fact, Morales and Wang (Morales and Wang, 2010, p. 556) stated that "it is hard to imagine that one can predict whether a booking will be canceled or not with high accuracy". However, António et al. have demonstrated that the likelihood of booking cancellations can be predicted with high accuracy (Antonio et al, 2017a, 2017b). The sum of all bookings predicted as likely to cancel can be deduced from the demand to calculate the hotel net demand, i.e., the demand excluding bookings that will likely cancel. Equipped with an accurate demand value, a hotel's revenue manager can make sounder and informed demand-management decisions and improve overbooking strategies and cancellation policies.

Business analytics relies on the continuous advances of systems to support decision making (Holsapple et al, 2014). Predictive analytics is the better known of the three aspects that comprise the business analytics orientation dimension. The other two being descriptive and prescriptive analytics. In the context of quantitative empirical modeling, predictive analytics is defined as "the building and assessment of a model aimed at making empirical predictions" (Shmueli and Koppius, 2011). The model should comprise two components: empirical predictive models developed to predict new/future observations and methods for evaluating the predictive power of these models. Predictive analytics has an important role in theory building, theory testing, and relevance assessment. Scientific research in predictive analytics can assume different roles: generation of new theory, development of evaluation measurements, competing theories comparison, improvement of existing models, assessment of relevance and of predictability. (Shmueli and Koppius, 2011) conducted a literature survey to investigate the extent to which predictive analytics was integrated into research empirical Information Systems and concluded that only 7 of the 52 papers with predictive claims did in fact employ predictive analytics. This shortage of studies on the subject is also recognized by (Holsapple et al, 2014) and by (Domingos, 2012). The may be explained by the difficulty to overcome obstacles for its operationalization (Abbott, 2014) at several levels: management (a shift in resources allocation and mentalities is required), data (existence of quality data on the subject), modeling (model complexity issues), deployment (integration and practicality issues).

Theoretically, building models to predict cancellation of hotel bookings is possible (Antonio et al, 2017a, 2017b). Despite the increasing importance of predictive analytics and machine learning in business applications (Domingos, 2012; Holsapple et al, 2014; Shmueli and Koppius, 2011), no scientific documented examples are available to address this particular problem from an empirical perspective, that is, predictive analytics research. This study describes the development and implementation of a prototype of what (Demirkan and Delen, 2013) named as service-oriented decision support system. In this case, a cloud based system to be used as a service (Analytics-as-a-Service). This system could be used by itself or as a component of a Revenue Management System (RMS). The prototype is based on an automated machine learning model, designed to learn continuously from new PMS data and from previous predictions hits and errors. To assess the prototype performance in real production environments, the prototype was made available for two hotels with different characteristics. The deployment incorporated active hotel actions to prevent cancellations of bookings predicted with high probability, which has also been evaluated. Hopefully, this study will foster further research on the application of cloud-based service-oriented decision support systems to different business areas.

2 Data and Methods

Economic theories such as rationing, free entry, price discrimination, and monopoly pricing provide insights that are essential to revenue management. Certain economic fundamentals and assumptions serve as the basis of revenue management in the hospitality industry, namely, product perishability, limited capacity, high fixed and low variable costs, unequal demand over time, plausibility of forecasting demand, plausibility of segmenting demand, and different price elasticities of market segments (Ivanov, 2014; Talluri and Van Ryzin, 2005). Revenue management practice often diverges from classical economic theory in important aspects (Talluri and Van Ryzin, 2005). For example, the application of price elasticity demand theories in the hospitality industry is more theoretical than it is practical. For instance, customers can always change to a different hotel if the price increases or stay at a hotel even when the price of other hotel bookings decreases due to brand loyalty (Ivanov, 2014). This gap between theory and application renders the empirical evaluation of a machine learning model to predict hotel booking's cancellations an undeniable problem that should be addressed in the context of Design Science Research (DSR). DSR requires the development of an artifact, in particular, a prototype, which fulfils the two requirements of DSR: relevance—by addressing a real business need—and rigor—by applying the proper body of knowledge in the artifact development (Cleven et al, 2009; Hevner et al, 2004). In this case, this body of knowledge is encompassed by data science fields: computer science (machine learning, databases, and data visualization), statistics and domain knowledge (Flath and Stein, 2018).

2.1 System design

The system has several different objectives: the automation of the modeling tasks; the deliverance of information for the hotel to act upon; and to register information that enables the assessment of the performance of the booking's cancellation prediction in a real production environment. To fulfill the systems' objectives and the requirements of a service-oriented decision support system (Demirkan and Delen, 2013), the system was designed based on the following specifications:

- For modeling:
 - The system trains daily with a dataset of all reservations on-the-books, enabling it to learn with changes in bookings and changes of patterns that occur over time.
 - Each day, the system builds a new model and automatically executes hyper-tuning of parameters. Performance is compared with the performance results of the previous seven days to obtain a decision that enables the model parameters to be replaced with new parameters or continued use of the previous parameters.
 - The predictions and performance results of the preceding days are stored in a database for evaluation, and where applicable, reused as model elaboration features.
 - The system trains by incorporating the incorrect predictions of previous days as penalizations and the correct predictions of previous days as rewards, with costs being class-dependent (false positives have higher costs than those of other miss-classifications).
 - 50% of the new bookings should be marked as the "control group", indicating that details of these particular bookings would never be shown to hotels and enables A/B testing.
 - Global demand and net demand for future dates are calculated based on existing bookings and model prediction results.

- Usability:
 - A web-based platform with a visualization component should be accessible by hotel staff and researchers anywhere and anytime.
 - Hotels should have a login per staff user to access the application.
 - Every action executed by hotel staff should be logged.
 - Global totals, totals per room type of demand, and net demand are displayed in a planning screen.
 - Details of bookings that were identified as likely to cancel (and not part of the “control group”) for the current date or previous days should be available for consultation.
 - Booking attributes that may lead to the identification of customers should be displayed or recorded by the system (to enable research purpose usage).
 - The system should report the actions made toward bookings that were identified as likely to cancel to prevent their cancellation.
 - The system must provide the visualization of the model performance results daily.
 - The system must provide the analysis of model predictions and effective performance results without disclosing the results of the A/B testing.

2.2 Hotel participation, data understanding and data description

Convincing hotels to participate in the project was a challenge for two reasons. First, hotels were required to share their data with the researchers. Second, hotels were required to commit resources to the project, particularly human resources. Hotels' staff was required to use the prototype on a daily basis and incorporate the prototype predictions in their demand-management decisions. Hotels' staff was also required to analyze the bookings predicted as likely to cancel and decide which to contact to try to prevent a cancellation.

A Portuguese hotel chain (that required anonymity) agreed to participate in the project and provided consent to access the PMS data of two of their hotels. One is a resort hotel (H1), and the other is a city hotel (H2); both have more than 200 rooms and are classified as four star hotels. Data was available from July 2015 to August 2017. Because H2 engaged in a soft-opening process until the end of August 2015, only data from September 2015 onwards was considered for the modeling of H2.

Figure 1 presents the cancellation ratios,¹ which oscillate between 25.7% in 2015 and 30.8% in 2017 for H1 and exceed 40% for H2. Note that these values, especially the values for H2, substantially exceed the value indicated by Morales and Wang (20% cancellations) (Morales and Wang, 2010). Excluding the previously referred initial period of July and August 2015 for H2, an analysis of the expected arrivals per month shows a growing tendency for cancellation (**Figure 2**). Within this timespan, the monthly cancellation ratio exceeded 35% in June, July, and August 2017. **Figure 2** also suggests that seasonality has an important role in the cancellations behaviors of H1: in the off-peak season months (November and January) cancellations

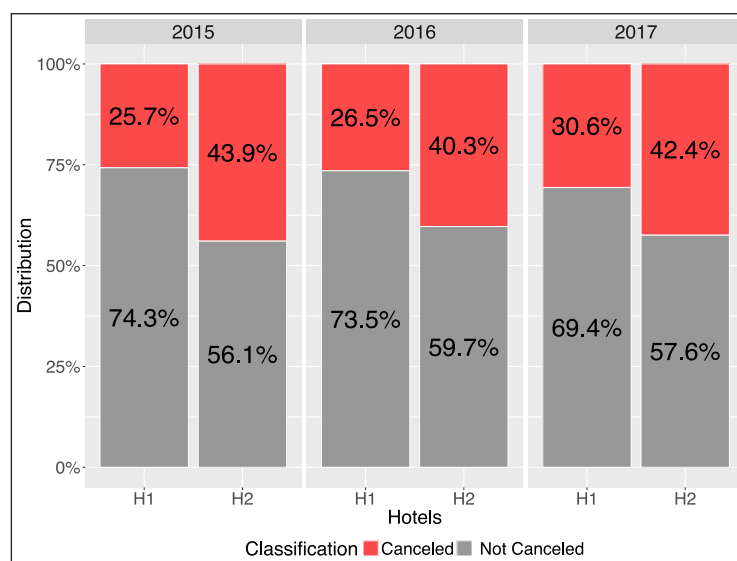


Figure 1: Cancellation ratio per year.

¹ In the scope of this work, “no-shows”, i.e., customers who have made a reservation but failed to cancel it or did not check-in on the expected date of arrival are treated as cancellations.



Figure 2: Cancellation ratio per month.

decrease to a minimum but increase to a maximum from June to September. Since the cancellation ratio of hotel H2 is nearly 50%, class imbalance is not a problem for this hotel. In contrast, H1 shows a considerably lower ratio due to class asymmetry.

The two hotels' datasets and their description can be found and download from an open data paper (Antonio et al, 2018).

2.3 Machine learning model

The CRISP-DM methodology (Chapman et al, 2000) was employed to build the system's machine learning models. Although models that were previously developed served as the starting point (Antonio et al, 2017a), multiple adjustments required employing the different CRISP-DM phases until the final prototype models were obtained. Models that employ PMS data produce better results than models that employ PNR data (Antonio et al, 2017a). Nevertheless, the deployment in a real production environment revealed a tendency for models to overfit data: the models did not generalize well for unknown bookings, that is, bookings for a date in the future not included in the model development process, which is a common issue in machine learning models (Domingos, 2012). Further analysis revealed that two issues that had considerable influence on the performance in the production environment: data leakage and "dataset shift", i.e., "where the joint distribution of inputs and outputs differs between training and test stage" (Quiñero-Candela et al, 2009, p. xi). Distribution shift main reasons were: a) based on the booking status outcome (canceled or not canceled) and due to the speed at which the hospitality business changes, the stratified dataset splitting strategy for the creation of the training and testing datasets did not guarantee a comparable distribution among both the training datasets and the testing datasets; b) the rapid growth of the tourism industry in recent years and the increasing annual demand causes a rapid increase in the prices (Average Daily Rate (ADR)) and *LeadTime*,² which contribute to differences in the distributions of inputs and outputs over time. In addition, this fast pace of operations causes the continued arrival of new players (Online Travel Agencies (OTAs) and the disappearance of other players, namely, "traditional" travel agencies and travel operators. These constant transformations contribute to a change in the representative weight of these entities in the hotel operation, which influences the distribution of certain features, such as *ADR*, *LeadTime*, *Agency* or *Company*, over time. Consequently, two major changes occurred: in dataset construction and dataset splitting, and in feature selection and engineering, which are detailed in the following sub-sections.

Another important change in the construction of these models is the use of the highly effective machine learning gradient tree boosting algorithm XGBoost (Chen and Guestrin, 2016) to build the classification models to predict each booking's cancellation outcome. XGBoost is a decision tree-based

² *LeadTime* is the hospitality term for the number of days prior to the arrival date for a booking.

ensemble algorithm that is recognized as one of the most effective and fast algorithms among classification (and regression) algorithms. The effectiveness of XGBoost, particularly in terms of controlling overfitting, is achieved by a set of parameters that enable fine-tuning of the model's complexity, including parameters to add randomness to make training more robust to noise. These parameters include the definition of the subsample of observations to use in each decision tree and the subsample of features to use per decision tree and per tree level. For the estimation of model parameters, including the learning rate and boosting, a combination of two well-known techniques—grid-search and random-search—was employed (Bergstra et al, 2011). The values for the parameters were selected from the model presenting the better error rate, from a total of 100 iterations of ten-fold cross-validations, over a maximum ensemble of 200 trees. In cross-validation, the parameter “early stop” was set to 8, indicating that training was stopped after eight rounds of training set error improvements without a correspondent improvement of the test set error to avoid overfitting. For each iteration, parameters were randomly selected according to limits that were previously established during manual optimization experiments. The list of parameters and source code to select its values and the established limits is provided in **Table 1**. Each of the parameters' impact on the model estimation is detailed in the XGBoost documentation (Chen and Guestrin, 2016).

2.3.1 Data splitting and data construction

Considering the existing “dataset shift” problem and that the selection of the data-splitting method should depend on the characteristics of data, such as size and structure (Kuhn and Johnson, 2013), a method borrowed from time series techniques was employed to create the training and testing datasets: convenience splitting (Reitermanová, 2010). This data-splitting method enables the capture of “non-stationary temporal data”: data that “changes behavior with time and therefore should be reflected in the modeling data and sampling strategies” (Abbott, 2014, p. 197). Convenience splitting involves the division of the dataset in discrete “time” blocks. In this case, the dataset was divided into blocks of “month/year” of bookings' arrival dates. From each block, 75% of bookings were assigned to the training dataset, and the remaining 25% of bookings were assigned to the testing dataset.

Data for hotel forecasting has two dimensions: the first dimension is related to booking creation, and the second dimension is related to the period of stay (Weatherford and Kimes, 2003). Regardless of cancellation policies, a booking can be canceled anytime between the date of its creation and the expected date of arrival. Consequently, at any moment in time, bookings with three types of status coexist in a hotel PMS database (**Figure 3**): (A) Effective – bookings with an arrival date that is prior to or equal to the current date, for which customers already checked-out or are checked-in; (B) Canceled – bookings with an arrival date set for any moment in time (past or future) but which were already canceled; (C) Unknown – bookings with an arrival date that is equal to or later than the current date and that have not been canceled prior to the current date but can be canceled in between the current date and the date of arrival.

For model improvement, all “C” bookings were removed from the dataset construction. As expected, for future dates, only canceled bookings (“B”) are considered in the dataset apart from a small period after the current data. Although a severe imbalance is introduced in the dataset (with respect to future dates), the benefits of this change outweighed the losses since it reduces the risk of leakage and training with incorrect data.

Table 1: Models' estimation parameters selection source code.

| Parameter | R source code |
|------------------|---------------------|
| colsample_bytree | runif(1, 0.4, 0.8) |
| eta | runif(1, 0.01, 0.3) |
| gamma | runif(1, 0, 0.2) |
| lambda | runif(1, 0, 0.5) |
| max_delta_step | sample(1:5, 1) |
| max_depth | sample(2:4, 1) |
| min_child_weight | sample(1:5, 1) |

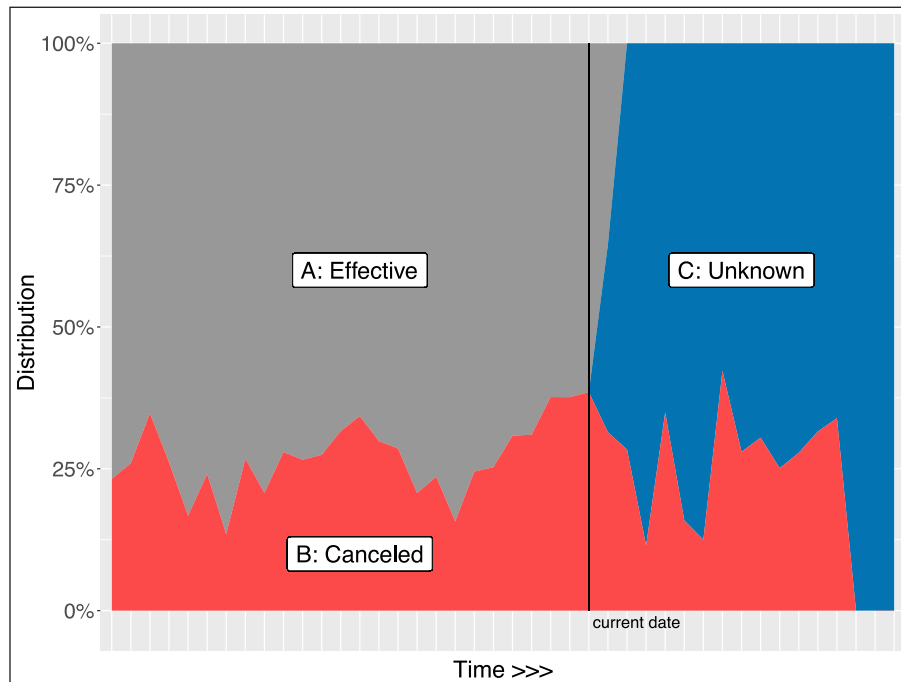


Figure 3: H1 bookings status at a moment in time.

2.3.2 Feature selection and feature engineering

Feature selection and engineering tasks require not only technical knowledge but also intuition, creativity, and domain knowledge (Domingos, 2012; Flath and Stein, 2018). Feature selection and engineering are amongst the most important factors for the success of machine learning projects. This importance can be confirmed by the role that the transformations here implemented had in circumventing the problem of initial poorly predictive training results: due to the removal of “C” bookings, models began to predict most future arrivals as “likely to cancel”.

Five major transformations of the datasets were performed by using feature selection and feature engineering transformations that required hundreds of iterations to train models, results’ evaluations, and building new transformations.

In the first step, features that did not contribute to model improvement or introduced noise were removed. The identification of these features was made with the XGBoost feature importance metric. The removed features included *AssignedRoomType*, *RequiredCarParkingSpaces* and *ReservedRoomType*.

In the second step, *Country* was also removed from the modeling datasets because it introduced leakage in the model. The leakage was due to the default filling of Portugal as country of origin in the bookings, information that was only confirmed and corrected at check-in.

Although some authors consider that demand patterns substantially differ by day of the week (Haensel and Koole, 2011), this case was observed in terms of cancellations. While experiments with a feature representing the arrival day of the week—*ArrivalDateWeekDay*—did not improve the performance of any of the hotel models, the splitting of the total number of nights of stay improved the performance of both hotel models. That feature was split into a feature representing the number of stayed weekend nights (*StaysInWeekendNights*), and another feature representing the number of stayed weeknights (*StaysInWeekNights*).

Even though seasonality is a phenomenon of recognized importance in the tourism industry (Song and Li, 2008), this is not the case here since only already canceled bookings (“B” bookings) for future dates are included in the datasets. Thus, all features associated with time, that is, features that can capture seasonality, had to be removed from the modeling datasets. As previously explained, to prevent that models predict that the vast majority of bookings would be cancelled, required the removal of features representing the arrival date of: the month (*ArrivalDateDayOfMonth*), the month number (*ArrivalDateMonth*), yearly week number (*ArrivalDateWeekNumber*) and year (*ArrivalDateYear*).

The features *LeadTime* and *ADR* were reengineered into multidimensional features *LiveTime* and *ADRThirdQuartileDeviation*, respectively. These features were created from multiple variables and contributed to improve model’s accuracy due to the information gain obtained with those associations (Abbott,

2014; Kuhn and Johnson, 2013). *LiveTime* was differentiated from *LeadTime* by capturing another information attribute: the number of days prior to arrival at which the booking was canceled. *ADRThirdQuartileDeviation* was differentiated from *ADR* by capturing the *ADR* distribution and amplitude. Bookings with an expensive price (compared with similar bookings for the same period, room type, and distribution channel) tend have a higher rate of cancellation. However, the *ADR* does not provide information about its position in relation to similar bookings. To capture this positioning of a booking price against similar bookings, several modeling iterations were necessary to uncover an engineered feature that would incorporate price, on a normalized scale, for any period of the year. This new feature—feature *ADRThirdQuartileDeviation*—is calculated by the formula (1).

$$ADRThirdQuartileDeviation = \frac{ADR}{ADR \text{ of 3rd quartile of distribution channel, per room type, per week/year}}. \quad (1)$$

Considering that high cardinality can cause slow model training and overfitting (Abbott, 2014), the last transformation involved treating the high degree of cardinality of some of the categorical features, such as *Agency* and *Company*. These features were re-encoded into two additional features using the R “vtreat” package (Mount and Zumel, 2017). This served to mitigate the effects of high cardinality and attenuated the overfitting effects of features containing categorical levels that were sparsely employed, such as some of the levels in the features *DistributionChannel*, *Meal* and *MarketSegment*.

These transformations produced a modeling dataset with a set of features that considerably differed from the modeling dataset employed in previous studies (Antonio et al, 2017a, 2017b). A complete list of these features and their descriptions is listed in Appendix A.

2.4 System architecture and modeling

To comply with the previously mentioned prototype requirements and specifications and to render the system technically reliable and capable of adequate performance, the system was built on top of the Microsoft Azure cloud platform, taking advantage of several open-source components and technologies available as services in this platform (Figure 4): one HDInsight Linux based, Hadoop and Spark cluster with R Server. This component enabled Hadoop/Spark-based big data processing, enabled R to be used in the Spark context and took advantage of XGBoost (Chen and Guestrin, 2016) performance efficacy by utilizing the cluster capabilities to distribute the processing among the different machines; one SQL database to process and store logs for all operations. This component also stored all prediction results with actions of the users; One web server. This component published the visualization layer in the form of a dynamic website, built in C# and asp.net. In this website, users can consult demand, predictions, and report the actions made for bookings identified as likely to cancel.

Since each hotel had a unique PMS database located in servers at the hotels' premises, a fully automated Extract, Transform and Load (ETL) process was created in each of the hotels for a daily extraction of all bookings from the hotels PMS', transformation of the data into a CSV dataset file, and loading into the Hadoop cluster.

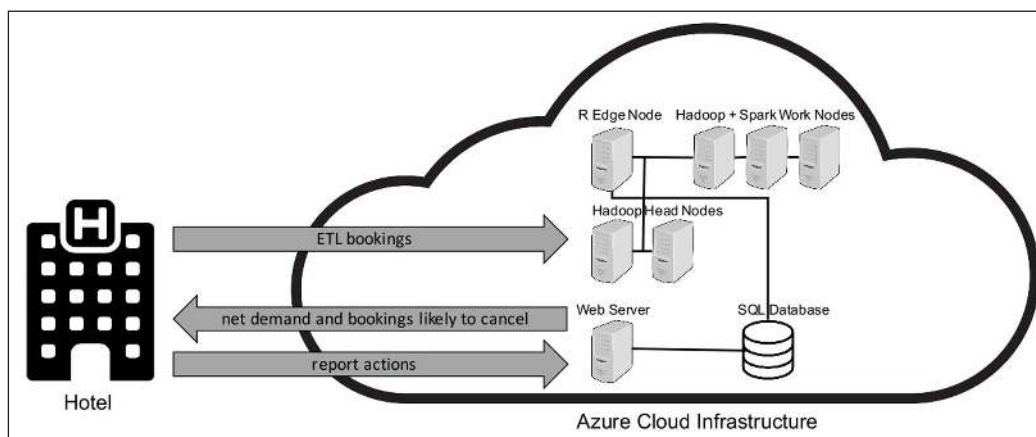


Figure 4: System architecture diagram.

“Even the most accurate and effective models don’t stay active indefinitely” (Abbott, 2014, p. 498). To overcome this vulnerability and to enable the system to continuously learn from new data, the system was designed to incorporate the “Champion-challenger” approach (Abbott, 2014, p. 508). Rather than waiting for a decrease in model performance to build a new model, a challenger model is built on a daily basis and its performance compared with the performance of the current model. The model with superior results will be selected. This fully automated daily cycle, which is illustrated in the diagram of **Figure 5**, is composed of eight steps:

1. ETL PMS data to cluster: at a predefined time, an SQL jobs extracts all bookings from the PMS database, transforms data to the format required by the modeling component and loads the data to the Hadoop cluster via a Windows Powershell script.
2. Data preparation: this step includes the selection of data, definition of the training and testing datasets, removal of the unused features (Section 2.3), data cleaning, construction of engineered features, and calculation of a weight per booking/observation (as next explained).
3. Build “challenger model”: using the training dataset, a ten-fold cross-validation mixed grid/random-search is executed to hyper-tune model parameters. The model is trained with the selected hyper-tuned parameters.
4. Build “champion model”: train a model with the parameters employed on the previous day.
5. Assess models’ performance: both models are fed with the testing set and both Accuracy and AUC metrics values are compared. When the “challenger” model outperforms the “champion” model for the last seven days’ average and on at least four of these days in both metrics, the “challenger” is selected to be the model. Otherwise, the use of the “champion” model will continue.
6. Apply the selected model to expected arrivals: this step involves the application of the selected model to all future arrivals (“C”-type bookings”) and predict their outcome.
7. Evaluate results: for both models, calculation of classic machine learning performance metrics (Accuracy, AUC, Precision, F1Score, Sensitivity and Specificity), regarding both the training datasets and the testing datasets. Calculate the ratio of predicted bookings as likely to cancel for future arrivals (“C” type bookings).
8. Record results in database: all performance metrics and all predictions of the current day are recorded in the database to enable further analysis and enable the use of previous predictions in the creation of the weighting mechanism.

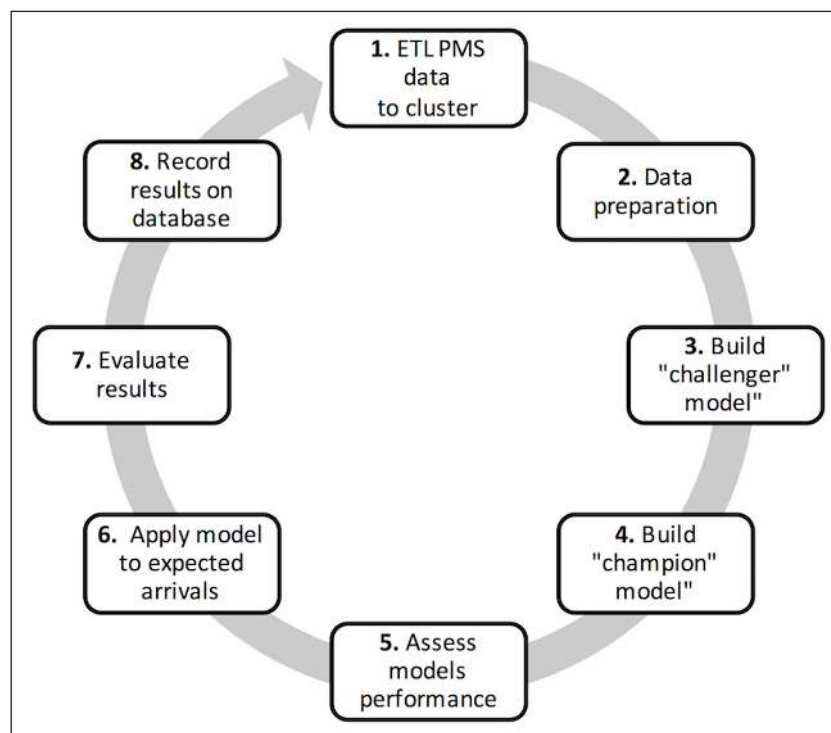


Figure 5: Daily automation cycle diagram.

Note that, since cancellation patterns change over time and because the system was required to learn continuously, a weighting mechanism was created to attribute higher importance to recent bookings and to incorporate a cost-sensitive learning by example weighting based on previous predictions hits and errors (Abe et al, 2004). In fact, hotel bookings are dynamic, i.e. over time there is a change in bookings' attributes (e.g. arrival date, length of stay, number of persons, etc.). On the other hand, time to arrival influences cancellations: a booking can be predicted as "likely to cancel" one of the days, but as "not likely to cancel" on the next day. Measuring the precision of previous predictions on unstable observations required the development of a new measure, Minimum Frequency (MF):

$$MF = \frac{\sum_{i=1}^n \hat{y}_i}{n} \tag{2}$$

MF is calculated by Formula (2), where n is the number of days since the booking has arrived to the hotel and has been processed by the predictive system and \hat{y}_i is the prediction classification for each day i it was processed. The prediction is binary: 0 for classified as "not likely to cancel" or 1 when classified as "likely to cancel".

As illustrated in **Figure 6**, the weighting mechanism is comprised of two components. The "time component" calculates the base weight according to the booking antiquity. Then, the "previous predictions component" uses the booking outcome status and the MF measure to assign a penalization to every false negative and false positive observations on the dataset, or a bonus to true positive predictions. The MF threshold to classify if prediction was correct was set to 0.5.

2.5 Development and deployment

The main component of this system prototype—the modeling component—was written in R (R Core Team, 2016) and continuously run in the R Edge node of the HDInsight cluster. Every day, at a predefined hour, this component executed the daily automation cycle described in the previous section. This modeling component and its visualization component were deployed in April 2017. After a set of tests, adaptations, and optimizations, the system was made available to hoteliers on the 1st of May 2017. However, it was not until the end of May that hotels started to utilize the prototype in a systematic manner. Initially, the evaluation period was defined to run from June to September of 2017. However, due to hotel human resources constraints, this period had to be shortened and completed at the end of August 2017.

An initial kickoff meeting was held in April to provide training to hotel users (revenue management team) about the visualization component of the system. The training explained how users should report actions to prevent the predicted cancellation of bookings, consult logs and analyze modeling performance results. The training also discussed how to visualize a planning for future dates and how to identify bookings that were predicted as likely to cancel. The main screen of the prototype visualization component (planning for

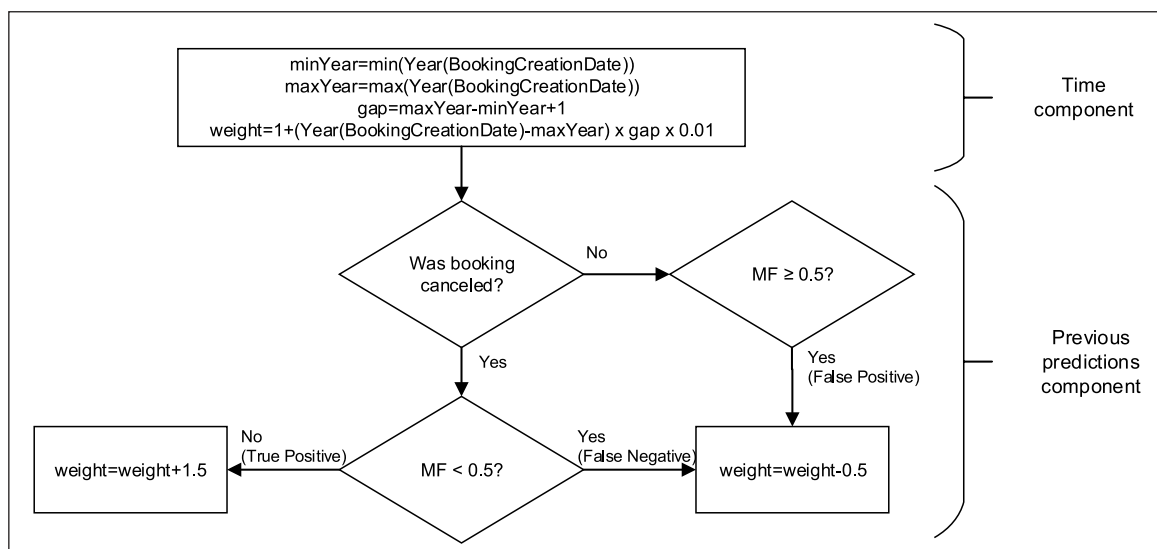


Figure 6: Observations weighting mechanism diagram.

future dates screen, **Figure 7**) enables users to visualize the demand for each room type (smaller font) and the net demand (larger font) for current and future dates one year in advance. The net demand is calculated by deducing the total number of bookings that were predicted to be cancelled. The planning also exhibited the daily totals of demand, occupation ratios, and pickup (difference in the total bookings between a date, which is the previous day by default, and the day of the visualization). A button on each of the day lines enables users to check the PMS identification (*Folio number*) of the bookings that were identified as likely to cancel and additional information, including booking attributes such as arrival date, nights, departure date, number of persons, ADR, total room revenue and frequency, which was the number of days that the booking was identified as likely to cancel in relation to the total number of days that the booking was processed by the system (**Figure 7**). For A/B testing, note that 50% of the bookings were used as a control group (“A” group) and the remaining 50% of the bookings were used as the verification group (“B” group). Users could only view the details of bookings in the “B” group that were predicted as likely to cancel.

A click on the *Folio number* enables users to report to the researchers the actions that were taken to avoid a booking cancellation, including how the action was executed and what was offered to (or asked of) the customer.

2.5.1 Operation

To try to avoid cancellation of bookings that were identified as likely to cancel, the hotel revenue management team had *carte blanche* from the hotel chain board to offer any type of services or discounts they deemed suitable according to the booking potential revenue loss. These discounts included breakfast discounts to customers who have booked room-only rates, free room-type upgrades or discounts on room-type upgrades, free meals or discounts on meal packages, and discounts on other services such as car parking, SPA treatments, and free tickets for local attractions.

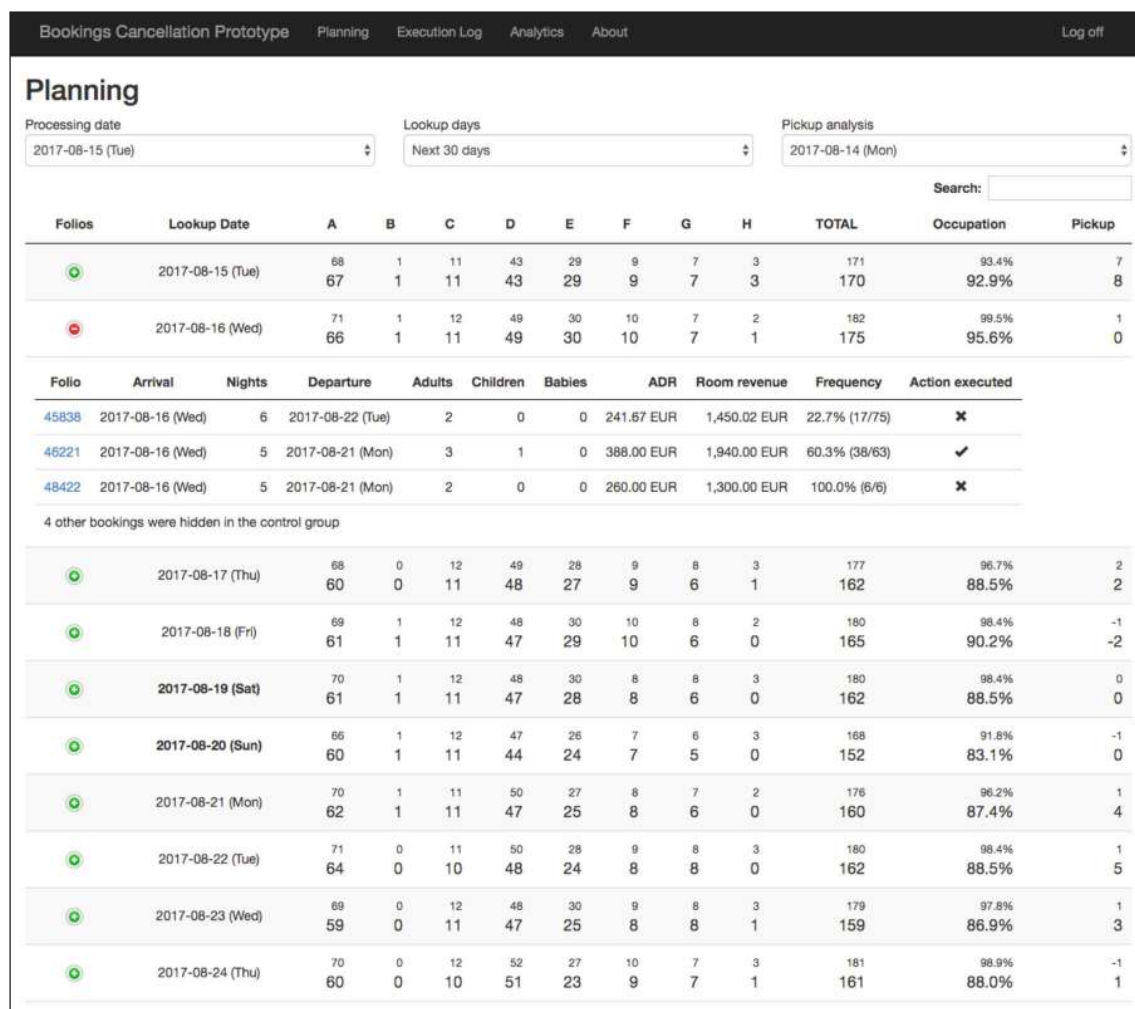


Figure 7: Prototype's main screen – Planning.

Initial contacts with customers revealed that this type of approach was very demanding in terms of human resources costs and financially costly. Customers started to request additional discounts (e.g., when offered a 20% discount on breakfast, customers would ask for free car parking), which can generate a higher costs/less margin and be a time-consuming task. Therefore, the hotel revenue management team rapidly decided to change the policy of these contacts and, with the authors' agreement, decided to inquire about technicalities, such as the type of bed preferred by customers, the expected hour of arrival to ensure that rooms could be prepared in a timely manner, children's ages (for the size of beds/cots), car license plate (to accelerate the check-in process), or credit card details, when the customers were not present or the data were not validated. The hotel staff also made themselves available to clarify any questions that customers may have regarding their stay, the hotel, or the region prior to their arrival. This information enables the hotels to provide a better and customized service to customers, also enhancing the quality of service.

The system identified a high number of predicted cancellations. Since the hotels did not have sufficient resources to contact all customers, hoteliers defined selection criteria for which bookings were to be contacted:

1. Arrival date should be three days in advance of the current date, at a minimum;
2. The booking should be made at a reasonable price or yield high room revenue;
3. The customer had to be directly contactable (e.g., extranet contacts or direct emails). Note that this criterion excluded any customers who were travelling with traditional travel agencies or other partners that did not disclose direct contact with their customers (e.g., Hotelbeds).
4. The customer's nationality and language were identifiable, and the latter had to be dominated by some of the hotel staff. Therefore, the hotels only contacted customers who spoke Portuguese, Spanish, German, English, or French.
5. Only bookings classified as likely to cancel at least 50% of the time the booking had been processed by the model (frequency) should be chosen. However, this criterion was not mandatory: if sufficient resources are available, lower frequencies with a high revenue would be contacted.

Most the contacts were made via personalized direct emails or via their original booking platform (e.g., Booking.com extranet or Expedia.com extranet). Using templates for each language, texts were always personalized for each customer.

3 Results

3.1 Quantitative results

The proposed approach shows that the capacity of the system to continuously learn with the daily incorporation of new bookings—both with changes to existing bookings and with the outcome of previous predictions—and the ability to automatically build a new model every day produced a system that achieved satisfactory quantitative results.

The chosen “Champion-challenger” strategy showed that the system required a relatively short time to stabilize. In the case of H1, the system commuted to the challenger model only twice within the first two weeks of deployment. Similarly, for H2, the system changed four times in the first four weeks of deployment. Since this time, the champion model has been consistent. This stability does not imply that the model will not change again but implies that the system only changes after proven performance. This finding can be explained by the criteria specifications for the challenger model to be selected, requiring the challenger model to demonstrate a superior performance compared with the performance of the champion model. These criteria ensure that a challenger model that performed very well on a particular day is not promptly selected.

From the perspective of classic machine learning performance metrics, since models were built and assessed daily, the results cannot be presented for the entire period. Because daily results were very similar, only the performance metrics for the last day are presented in **Table 2**. As expected, these results are slightly inferior to those reported by the authors in the previous theoretical study (Antonio et al, 2017a). Current models are less prone to overfitting, more robust, and do not exhibit problems of over-classification for future arrivals. On August 31, 2017, the percentage of future arrivals that were identified as likely to cancel was 18.6% for H1 and 26.4% for H2, which is consistent with the hotels' cancellations rates (as displayed in **Figures 1** and **2**). Similarly, the differences among hotels' cancellation rates are also present in the models' performance metrics, which consistently present superior values for H2.

A/B testing also presented stimulating results. For arrivals expected between June 2017 and August 2017 (excluding bookings canceled prior to the model deployment – April 2017), the number of bookings on which hotels acted to avoid cancellations was rather low (5.4% for H1 and 4.8% for H2), the percentage of canceled bookings in group “A” (the group that was not included) is 0.6% higher than the results for group “B” (**Table 3**). This finding translates into a relative decrease in group “B” cancellations of 2.5% for H1 and a relative decrease in group “B” cancellations of 2.0% for H2. Note that these differences are not sufficient to consider the results as statistically significant. The Cohen’s h size effect (Cohen, 1988), i.e., the difference in the cancellation rate, would have to exceed 7.9% for H1 and exceed 5.5% for H2 (at a significance level of 0.05, using a power of test of 0.80). The Chi-square test of independence also shows that this difference is not statistically significant for any of the hotels: for H1, we obtain $\chi^2(1) = 0.144$ and $p = 0.705$; for H2, we obtain $\chi^2(1) = 0.234$, $p = 0.629$.

Assessing the system by the MF ratio confirms the system’s predictions precision. As depicted in **Figure 8**, a MF decrease is followed by a decrease in the cancellation ratio. The cancellation ratio for bookings that were predicted as likely to cancel every time they were processed (MF = 100%) was 50.1% for H1 and 57.4% for H2. These values decrease to 39.8% for H1 and 38.4% for H2 with bookings that were predicted as likely to cancel at least 50% of the times that they were processed (MF ≥ 50%). These values contrast the total cancellation ratio (MF ≥ 0%) of 24.3% for H1 and 25.2% for H2.

Note that this cancellation ratio can be higher if hotels had not contacted some of the bookings to avoid cancellation. Considering the low number of bookings acted on to prevent cancellations in relation to the total number of bookings that were predicted as likely to cancel (**Table 3**), these actions had a significant impact on avoiding cancellations. The analysis of the “B” groups, the groups of bookings to which the hotels had access to the details of bookings predicted as likely to cancel, shows a substantial difference in terms of the cancellation rates between the bookings were no actions were made and bookings were actions were made (**Table 4**). For all “B” group bookings with MF ≥ 0%, this difference is 13.8 percentage points for H1, which translates to a relative decrease in cancellations of 56%. For H2, this difference is greater, with a value of 18.1 percentage points, translating to a relative decrease in cancellations of 70%. A Chi-square test of independence confirms that this difference is statistically significant for both hotels: H1: $\chi^2(1) = 9.978$, $p = 0.002$; H2: $\chi^2(1) = 31.873$, $p < 0.001$. For “B” group bookings predicted as likely to cancel in at least half of the days that they were processed (MF ≥ 50%), the differences are substantial. The differences in the cancellation ratio are 37.1 percentage points for H1 and 37.8 percentage points for H2, which corresponds to relative decreases in cancellations of 82% for H1 and 83% for H2. A Chi-square test of independence confirms that this difference is statistically significant for both hotels: H1: $\chi^2(1) = 33.609$, $p < 0.001$; H2: $\chi^2(1) = 58.373$, $p < 0.001$.

This association between bookings for which customers were contacted and bookings for which customers were not contacted can be measured to compare bookings for which customers were not contacted and effectively canceled against those for which customers were contacted. For bookings that were predicted

Table 2: Performance metrics on the 31st of August 2017.

| Hotel | Dataset | Accuracy | Precision | F1Score | AUC | Sensitivity | Specificity |
|-------|---------|----------|-----------|---------|--------|-------------|-------------|
| H1 | Train | 0.8646 | 0.8484 | 0.7410 | 0.9227 | 0.6577 | 0.9510 |
| | Test | 0.8486 | 0.8205 | 0.7016 | 0.8864 | 0.6128 | 0.9452 |
| H2 | Train | 0.8701 | 0.8849 | 0.8460 | 0.9438 | 0.8103 | 0.9171 |
| | Test | 0.8563 | 0.8731 | 0.8274 | 0.9276 | 0.7862 | 0.9110 |

Table 3: A/B testing effective cancellation summary.

| Hotel | Group | Canceled | Not canceled | Total | % Canceled | Actions | % Actions |
|-------|-------|----------|--------------|-------|------------|---------|-----------|
| H1 | A | 486 | 1,489 | 1,975 | 24.6% | N/A | N/A |
| | B | 483 | 1,526 | 2,009 | 24.0% | 109 | 5.4% |
| H2 | A | 1,043 | 3,060 | 4,103 | 25.4% | N/A | N/A |
| | B | 1,025 | 3,086 | 4,111 | 24.9% | 196 | 4.8% |

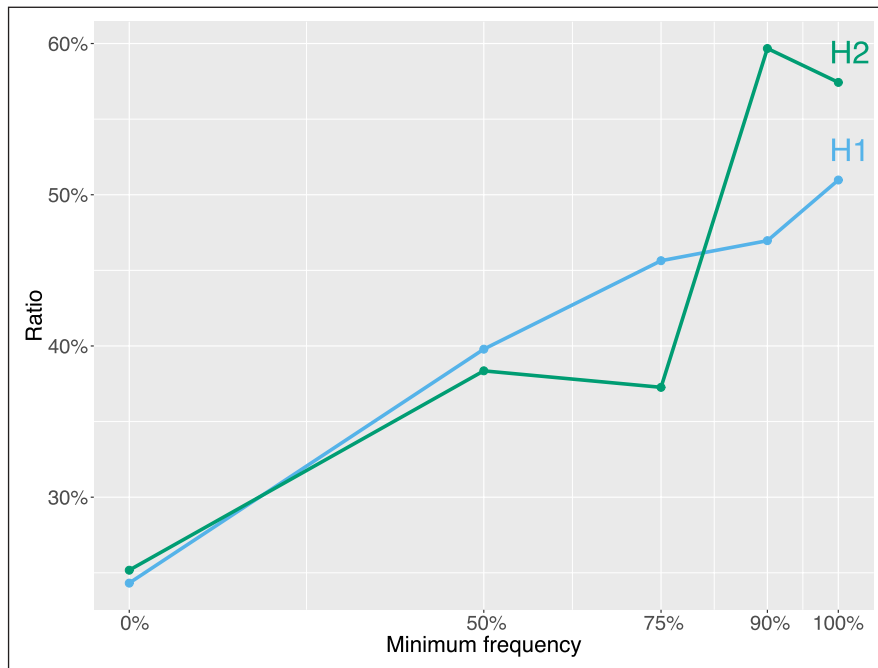


Figure 8: Cancellation ratio by minimum frequency.

Note: MF threshold levels were selected based on the users' criteria to select the bookings to contact. The majority of time, users only selected bookings with a MF equal to or greater than 50%.

Table 4: "B" group cancellation results summary.

| Hotel | Action | MF ≥ 0% (all bookings) | | | MF ≥ 50% | | |
|-------|--------|------------------------|--------------|------------|----------|--------------|------------|
| | | Canceled | Not canceled | % Canceled | Canceled | Not canceled | % Canceled |
| H1 | No | 471 | 1,429 | 24.8% | 125 | 153 | 45.0% |
| | Yes | 12 | 97 | 11.0% | 6 | 70 | 7.9% |
| H2 | No | 1,010 | 2,905 | 25.8% | 269 | 325 | 45.3% |
| | Yes | 15 | 181 | 7.7% | 9 | 111 | 7.5% |

as likely to cancel with an MF ≥ 50%, not contacting the guest entails a cancellation enhancer factor at a magnitude of 9.3 for H1, and a magnitude of 10.0 for H2, with 95% CIs [4.20, 24.83] and [5.26, 21.74], respectively. The lower cancellation rate of all bookings contacted by hotels, independent of their prediction as likely to cancel (MF ≥ 0%), indicates that contacting customers of bookings may reduce the number of cancellations. Because contacting all customers requires resources that are unavailable most of the time, these results highlight the importance of having a booking cancellation prediction model to identify in which bookings invest the limited available resources.

From a financial perspective, despite the low number of contacted customers of bookings, the analysis of the results emphasizes the impact to prevent cancellation of bookings that are identified as likely to cancel. Considering the proportion of bookings where actions to prevent cancellations were taken and did not effectively cancel in relation to those with no actions taken, the room revenue that has not been lost to cancellations is € 16,680.97¹ for H1 and € 22,144.77 in H2. For both hotels, the actions taken prevented a total revenue loss of € 38,825.75. This amount corresponds to a monthly average of € 12,941.91 of room revenue that is not lost to cancellations during the three months of the system's deployment. Some of this value would not have been lost even if cancellations occurred since hotels would eventually re-sell some of the rooms' nights. Cancellations increase uncertainty and prevent hotels' revenue management teams to increase prices, confirming the positive impact on the hotel business performance of contacting customers of bookings that are identified as likely to cancel.

Another interesting aspect is the fact that some customers who were contacted replied on the same day or the following day with an effective cancellation. This finding may not be negative since hotels can immediately reserve the canceled rooms for other customers.

3.2 Qualitative results

From the periodic interviews with the hotel chain revenue management team and the project final interview, four important considerations were highlighted.

First, users suggested that the system should be fully integrated with the PMS or should be able to display each booking's complete details. Users indicated that this requirement can expedite the time required to identify the details of each booking that was predicted as likely to cancel. This situation also limits the total number of customers of that they managed to contact about their bookings.

Second, hotels recognized that they seldom took advantage of the "net demand" as an indicator in their demand-management decisions and acknowledged their resistance to change instead of a lack of confidence in the system as the main reason. In situations in which the hotel was overbooked or situations that required decisions for short term dates, they considered the system "net demand" measure to decide whether to open or close sales at certain time. As an example, the H2 team mentioned that at approximately 06:00 PM, the hotel was fully booked for the night, they decided to accept two walk-ins because the system identified that four of the bookings remaining to check-in were identified by the system as likely to cancel. Half of these four bookings canceled.

Third, hotel users recognize that the system may have a positive impact on the hotel's social reputation because most customers who were contacted engaged in conversation with the hotel staff, showed appreciation for the hotel concerns and thanked them.

Last, all users positively answered when asked if they would continue to use the system if it was made available as a permanent tool.

4 Discussion and Conclusion

This study contributes to reduce the paucity of studies in predictive analytics and demonstrates how Analytics-as-a-Service decision support systems can be built and deployed.

From a scientific standpoint, this study discusses several of the roles of predictive analytics in scientific research, including the development of new indicators for assessment of performance. In fact, one of the major contributions of this study is the development of the new measurement—MF—for evaluating the performance of binary classification problems when observation characteristics are unstable or when the outcome of the prediction is affected by time. Other of the important contributions of this study is the development of a weighting system allowing for training the model to enhance the importance of more recent observations and simultaneously learn from previous predictions. Additionally, this study also demonstrates how data-splitting method selection and domain knowledge in feature engineering are of paramount importance in machine learning modeling and the influence for the improvement of prediction models.

The development and deployment of the models in a prototype tested in real-world conditions enabled the assessment of the system's relevance and predictability, other relevant roles of predictive analytics research. Although the benefits of the application of machine learning in business information systems are advocated by several authors, so far, only a few studies demonstrate those benefits in applied research, this study being one of them.

Another distinctive point in this study is the use of open-source tools such as Linux, R, and Hadoop to build a cloud-based service-oriented decision support system. The system's performance and results prove the adequacy and usefulness of these tools for the problem of booking cancellation prediction. The Linux Hadoop/Spark cluster running R Server enabled the modeling process to be distributed through different cluster machines, taking advantage of the available computational power and the powerful XGboost tree boosting machine learning method. The results validated the value of the system architecture design for running an automated machine learning system that daily incorporates new data and utilizes previous prediction errors and hits for continuous improvement.

From a business standpoint, this study also presented significant results. First, the study showed that the final results of the different hotels were similar: Accuracy greater than 0.84, Precision greater than 0.82, and AUC greater than 0.88. Second, the bookings cancellation ratio in predicted as likely to cancel bookings attained 39.8% for H1 and 38.4% for H2 in at least half of the days' processed ($MF \geq 50\%$) and exceed the cancellation ratio of all bookings ($MF \geq 0\%$): 24.3% for H1 and 25.2% for H2. These results stress the satisfactory level of precision of the models. Third, despite the difficulties associated with contacting customers prior to their arrival (including the costs associated with the contact), the identification of possible cancellations enables hotels to take action for preventing effective cancellation at a limited cost. The decrease in the number of actual cancellations on bookings where customers were contacted, a total in excess of 37 percentage points, corresponds to a relative cancellation decrease of 82% for H1 and 83% for H2. These

findings indicate that the actions taken for preventing cancellations in identified as cancellable bookings amounted in a total revenue in the order of approximately € 39,000.00. Although all future bookings identified as likely to cancel cannot be contacted, the results indicate that an increase in the number of contacted customers may prevent additional cancellations and revenue loss.

This study highlights how a service-oriented decision support system, based on an automated machine learning model, designed in accordance to DSR to address an unsolved problem in a unique and innovative manner, can be constructed and implemented. The DSR approach demonstrated the importance of instantiations in terms of information technology research. It was the construction of the prototype that uncovered the limitations of the previously developed models and led to the design of new solutions to overcome those limitations. The measurable impact of the system on business performance highlights the benefits for revenue management in service-based industries (such as hospitality, airlines, rent-a-car, cruise ships, among other) of using Analytics-as-a-Service decision support systems to take advantage of the available data and technology to improve decision making.

4.1 Limitations and future studies

As expected, this study presents some limitations that are an incentive for further research. Although XGBoost produces a performance metric that enables modelers to comprehend the features that are employed in the models and the degree of importance of the features in a model's construction, the study of its importance and impact on business operation was beyond the scope of this study. Future research can explore the predictive power of features not only to better understand cancellation drivers but also to use this knowledge to improve cancellation policies.

Although the dataset for H1 presented a class imbalance this issue was not addressed. However, future research can address this issue to improve results.

Another limitation of this study was the difficulty of collecting the number of customers who responded to the hotels' contact. This could have been interesting for measuring the effective reach of the customers contacted. However, due to the multiplicity of channels that a customer can use to book a hotel and the many different persons/departments who can handle the contact, registering this process was impossible. The hotels' revenue management team estimates this number to be very low, probably less than 10%.

Two additional limitations, which are imposed by research requirements, contributed to the low number of contacted bookings. The first limitation was the fact that the system was designed to include A/B testing and did not allow hotel users to obtain the details of bookings in the "A" group. The second limitation was the time invested in the selection of the bookings to contact and the time required to obtain the contacts of these bookings, because it required the consultation of booking details in the hotels' PMS. In a real production system, the inexistence of these limitations would enable all bookings to be selected, which allows users to check booking details directly in the system and hotels to contact a larger number of customers within the same amount of time.

Approximately two years of data were available for training but it did not include features that can explicitly capture seasonality. The hospitality industry, especially in resort hotels, is an industry where seasonality has an important influence on business. The use of data in a wider timespan with the inclusion of time/season specific features has the potential to enable the development of models with other performances and capabilities. These models can also benefit from the introduction of features from other data sources related to factors that affect hotel customers' booking/cancellation decisions, such as competitors' prices, competitors' social reputation, weather, and events.

The latter proposed system can generate new features for improving the model performance. Since bookings that were acted on are canceled less frequently than bookings in which no action was taken, a feature with the indication if and what category of action was taken would probably improve model performance. Additionally, recording the actions made in each booking to avoid cancellation (e.g., offering a room upgrade or asking about the bed type preference) has a potential use for another machine learning model capable of recommending the actions that should be executed in the bookings that are predicted as likely to cancel. This finding can prompt the development of a fully automated system. A system that not only can predict a bookings cancellation outcome but also can select which customers to contact, make initial contact, and engage in a discussion with the customer via a chat bot, only requiring human intervention in the aspects of the discussion where the system is not prepared to answer.

Finally, booking cancellation prediction is just one example of the type of revenue management problems that can employ service-oriented support systems to help decision making. Future research should explore the development and implementation of systems for predicting overall demand, customer lifetime value, social reputation ratings, service delays or slow responses to customers' requests, among others.

Appendix A – Features description

Table A.1: Features description.

| Feature | Type | Description |
|----------------------------------|------|---|
| <i>ADRThirdQuartileDeviation</i> | N, E | Ratio calculated by the division of the booking ADR by the third quartile value, of all bookings of the same distribution channel, same reserved room type, for the same expected week/year of arrival. |
| <i>Adults</i> | N, I | Number of adults |
| <i>Agent</i> | C, I | ID of agency (if booked via an agency) |
| <i>Babies</i> | N, I | Number of babies |
| <i>BookingChanges</i> | N, E | Heuristic created by summing the number of booking changes (amendments) prior to arrival that can indicate cancellation intentions (arrival or departure dates, number of persons, type of meal, ADR, or reserved room type) |
| <i>Children</i> | N, I | Number of children |
| <i>Company</i> | C, I | ID of company/corporation (if an account was associated with it) |
| <i>CustomerType</i> | C, E | Type of customer (group, contract, transient, or transient-party); this last category is a heuristic built when the booking is transient but is fully or partially paid in conjunction with other bookings (e.g., small groups, such as families who require more than one room) |
| <i>DaysInWaitingList</i> | N, I | Number of days the booking was on a waiting list prior to confirming the availability and being confirmed as a booking |
| <i>DepositType</i> | C, E | Since hotels had different cancellation and deposit policies, a heuristic was developed to define the deposit type (nonrefundable, refundable, no deposit): payment made in full before the arrival date was considered to be a “nonrefundable” deposit, and a partial payment before arrival was considered to be a “refundable” deposit; otherwise, it was considered to be “no deposit” |
| <i>DistributionChannel</i> | C, I | Distribution channel used to make the booking |
| <i>IsRepeatedGuest</i> | C, E | Binary value that indicates if the booking holder, at the time of booking creation, was a repeat guest at the hotel (0: no; 1: yes); created by comparing the time of booking with the guest profile creation record |
| <i>LiveTime</i> | N, E | Number of days from booking creation according to the booking status: for “A” type bookings, it was calculated as the number of days between booking creation and arrival; for “B” bookings, the elapsed number of days between the date of booking creation and the cancellation date was employed; for “C” bookings, the elapsed number of days between the date of creation and the processing date (current date) |
| <i>MarketSegment</i> | C, I | Market segment to which the booking was classified as |
| <i>Meal</i> | C, I | ID of meal requested by the guest |
| <i>PreviousCancellationRatio</i> | N, E | Ratio created by the division of the guest’s number of previous cancellations by the guest’s previous number of bookings at the hotel |
| <i>StaysInWeekendNights</i> | N, E | From the total length of stay, the number of weekend nights (Saturday and Sunday) |
| <i>StaysInWeekNights</i> | N, E | From the total length of stay, the number of weekday nights (Monday through Friday) |
| <i>TotalOfSpecialRequests</i> | N, E | Number of special requests (e.g., fruit basket and sea view) |
| <i>WasInWaitingList</i> | C, I | Binary value that indicates if the booking was entered on a waiting list or directly entered as a booking (0: normal booking; 1: waiting list); |

Type legend: C- Categorical, E- Engineered, I-Input, N- Numerical.

Acknowledgements

The authors thank the board of the hotel chain for their support of this study. Without their support, this project would not be possible. The authors also thank the hotel chain revenue management team. Without their openness to embrace the system predictions in their demand-management decisions and their effort to act on bookings that are identified as likely to cancel, this study would not exist.

This study was partly supported by a Microsoft Azure Data Science Award granted to the first author.

Competing Interests

The authors have no competing interests to declare.

References

- Abbott, D.** 2014. *Applied predictive analytics: Principles and techniques for the professional data analyst*. Indianapolis, IN, USA: Wiley.
- Abe, N, Zadrozny, B and Langford, J.** 2004. An iterative method for multi-class cost-sensitive learning. In: *Proceedings of the Tenth ACN SIGKDD International Conference on Knowledge Discovery and Data Mining*. Presented at the tenth ACN SIGKDD International Conference on Knowledge Discovery and Data Mining. Seattle, WA. pp. 3–11. DOI: <https://doi.org/10.1145/1014052.1014056>
- Antonio, N, de Almeida, A and Nunes, L.** 2017a. Predicting hotel booking cancellation to decrease uncertainty and increase revenue. *Tourism & Management Studies*, 13(2): 25–39. DOI: <https://doi.org/10.18089/tms.2017.13203>
- Antonio, N, de Almeida, A and Nunes, L.** 2017b. Using data science to predict hotel booking cancellations. In: Vasant, P and M, K (eds.), *Handbook of Research on Holistic Optimization Techniques in the Hospitality, Tourism, and Travel Industry*. Hershey, PA, USA: Business Science Reference. pp. 141–167. DOI: <https://doi.org/10.4018/978-1-5225-1054-3.ch006>
- Antonio, N, de Almeida, A and Nunes, L.** 2018. Hotel booking demand datasets. *Data in Brief*. DOI: <https://doi.org/10.1016/j.dib.2018.11.126>
- Bergstra, JS, Bardenet, R, Bengio, Y and Kégl, B.** 2011. Algorithms for hyper-parameter optimization. In: *Advances in Neural Information Processing Systems 2011*. Presented at the Neural Information Processing Systems Conference. Granada, Spain. p. 9.
- Caicedo-Torres, W and Payares, F.** 2016. A machine learning model for occupancy rates and demand forecasting in the hospitality industry. In: *Advances in Artificial Intelligence – IBERAMIA 2016*. Presented at the Ibero-American Conference on Artificial Intelligence. Cham: Springer. pp. 201–211. DOI: https://doi.org/10.1007/978-3-319-47955-2_17
- Chapman, P, Clinton, J, Kerber, R, Khabaza, T, Reinartz, T, Shearer, C and Wirth, R.** 2000. CRISP-DM 1.0: Step-by-step data mining guide. *The Modeling Agency*. Available at <https://the-modeling-agency.com/crisp-dm.pdf> [Last accessed 10 September 2015].
- Chen, C-C, Schwartz, Z and Vargas, P.** 2011. The search for the best deal: How hotel cancellation policies affect the search and booking decisions of deal-seeking customers. *International Journal of Hospitality Management*, 30(1): 129–135. DOI: <https://doi.org/10.1016/j.ijhm.2010.03.010>
- Chen, T and Guestrin, C.** 2016. Xgboost: A scalable tree boosting system. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM. pp. 785–794. Available at <http://dl.acm.org/citation.cfm?id=2939785> [Last accessed 22 January 2017]. DOI: <https://doi.org/10.1145/2939672.2939785>
- Chiang, W-C, Chen, JC and Xu, X.** 2007. An overview of research on revenue management: current issues and future research. *International Journal of Revenue Management*, 1(1): 97–128. DOI: <https://doi.org/10.1504/IJRM.2007.011196>
- Cleven, A, Gubler, P and Hüner, KM.** 2009. Design alternatives for the evaluation of Design Science research artifacts. In: *Proceedings of the 4th International Conference on Design Science Research in Information Systems and Technology*. New York, NY, USA: ACM. pp. 19:1–19:8. DOI: <https://doi.org/10.1145/1555619.1555645>
- Cohen, J.** 1988. *Statistical power analysis for the behavioral sciences*. 2nd ed. Hillsdale, N.J.: L. Erlbaum Associates.


- Demirkan, H** and **Delen, D.** 2013. Leveraging the capabilities of service-oriented decision support systems: Putting analytics and big data in cloud. *Decision Support Systems*, 55(1): 412–421. DOI: <https://doi.org/10.1016/j.dss.2012.05.048>
- Domingos, P.** 2012. A few useful things to know about machine learning. *Communications of the ACM*, 55(10): 78–87. DOI: <https://doi.org/10.1145/2347736.2347755>
- Flath, CM** and **Stein, N.** 2018. Towards a data science toolbox for industrial analytics applications. *Computers in Industry*, 94: 16–25. DOI: <https://doi.org/10.1016/j.compind.2017.09.003>
- Freisleben, B** and **Gleichmann, G.** 1993. Controlling airline seat allocations with neural networks. In: *Proceeding of the Twenty-Sixth Hawaii International Conference on System Sciences, 1993*. Presented at the Proceeding of the Twenty-Sixth Hawaii International Conference on System Sciences, 4: 635–642.
- Garrow, L** and **Ferguson, M.** 2008. Revenue management and the analytics explosion: Perspectives from industry experts. *Journal of Revenue and Pricing Management*, 7(2): 219–229. DOI: <https://doi.org/10.1057/rpm.2008.3>
- Haensel, A** and **Koole, G.** 2011. Booking horizon forecasting with dynamic updating: A case study of hotel reservation data. *International Journal of Forecasting*, 27(3): 942–960. DOI: <https://doi.org/10.1016/j.ijforecast.2010.10.004>
- Hayes, DK** and **Miller, AA.** 2011. *Revenue management for the hospitality industry*. Hoboken, NJ, USA: John Wiley & Sons, Inc.
- Hevner, AR, March, ST, Park, J** and **Ram, S.** 2004. Design science in information systems research. *MIS Quarterly*, 28(1): 75–105. DOI: <https://doi.org/10.2307/25148625>
- Holsapple, C, Lee-Post, A** and **Pakath, R.** 2014. A unified foundation for business analytics. *Decision Support Systems*, 64: 130–141. DOI: <https://doi.org/10.1016/j.dss.2014.05.013>
- Huang, H-C, Chang, AY** and **Ho, C-C.** 2013. Using artificial neural networks to establish a customer-cancellation prediction model. *Przeglad Elektrotechniczny*, 89(1b): 178–180.
- Hueglin, C** and **Vannotti, F.** 2001. Data mining techniques to improve forecast accuracy in airline business. In: *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Presented at the 7th ACM SIGKDD international conference on knowledge discovery and data mining. San Francisco, CA, USA: ACM. pp. 438–442. Available at <http://dl.acm.org/citation.cfm?id=502578> [Last accessed 14 October 2015]. DOI: <https://doi.org/10.1145/502512.502578>
- International Civil Aviation Organization.** 2010. Guidelines on Passenger Name Record (PNR) data. Available at https://www.iata.org/iata/passenger-data-toolkit/assets/doc_library/04-pnr/New%20Doc%209944%201st%20Edition%20PNR.pdf [Last accessed 17 February 2016].
- Ivanov, S.** 2014. *Hotel revenue management: From theory to practice*. Varna, Bulgaria: Zangador. Available at [Last accessed 7 June 2015].
- Kimes, SE** and **Wirtz, J.** 2003. Has revenue management become acceptable? Findings from an International study on the perceived fairness of rate fences. *Journal of Service Research*, 6(2): 125–135. DOI: <https://doi.org/10.1177/1094670503257038>
- Kuhn, M** and **Johnson, K.** 2013. *Applied predictive modeling*. New York, NY: Springer New York. Available at [Last accessed 7 June 2015].
- Lemke, C.** 2010. Combinations of time series forecasts: When and why are they beneficial? Bournemouth University. Available at http://dec.bournemouth.ac.uk/staff/bgabrys/publications/C_Lemke_PhD_thesis.pdf [Last accessed 5 October 2015].
- Liu, PH.** 2004. Hotel demand/cancellation analysis and estimation of unconstrained demand using statistical methods. In: Yeoman, I and McMahon-Beattie, U (eds.), *Revenue Management and Pricing: Case Studies and Applications*. Cengage Learning EMEA. pp. 91–108.
- Mehrotra, R** and **Ruttley, J.** 2006. *Revenue management (second ed.)*. Washington, DC, USA: American Hotel & Lodging Association (AHLA).
- Morales, DR** and **Wang, J.** 2010. Forecasting cancellation rates for services booking revenue management using data mining. *European Journal of Operational Research*, 202(2): 554–562. DOI: <https://doi.org/10.1016/j.ejor.2009.06.006>
- Mount, J** and **Zumel, N.** 2017. *vtreat: A statistically sound "data.frame" processor/conditioner*. Available at <https://cran.r-project.org/web/packages/vtreat/index.html> [Last accessed 30 June 2017].
- Noone, BM** and **Lee, CH.** 2011. Hotel overbooking: The effect of overcompensation on customers' reactions to denied service. *Journal of Hospitality & Tourism Research*, 35(3): 334–357. DOI: <https://doi.org/10.1177/1096348010382238>

- Quiñonero-Candela, J, Sugiyama, M, Schwaighofer, A and Lawrence, ND.** (eds.) 2009. *Dataset shift in machine learning*. Cambridge, Massachusetts: MIT Press. DOI: <https://doi.org/10.7551/mitpress/9780262170055.001.0001>
- R Core Team.** 2016. *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Available at <https://www.R-project.org/>.
- Reitermanová, Z.** 2010. Data splitting. In: *WDS'10 Proceeding of Contributing Papers. Presented at the 19th Annual Conference of Doctoral Students, Charles University, Faculty of Mathematics and Physics*. Praha: Matfyzpress. pp. 31–36.
- Schwartz, Z, Uysal, M, Webb, T and Altin, M.** 2016. Hotel daily occupancy forecasting with competitive sets: a recursive algorithm. *International Journal of Contemporary Hospitality Management*, 28(2): 267–285. DOI: <https://doi.org/10.1108/IJCHM-10-2014-0507>
- Shmueli, G and Koppius, OR.** 2011. Predictive analytics in information systems research. *Mis Quarterly*, 35(3): 553–572. DOI: <https://doi.org/10.2307/23042796>
- Smith, SJ, Parsa, HG, Bujisic, M and van der Rest, J-P.** 2015. Hotel cancellation policies, distributive and procedural fairness, and consumer patronage: A study of the lodging industry. *Journal of Travel & Tourism Marketing*, 32(7): 886–906. DOI: <https://doi.org/10.1080/10548408.2015.1063864>
- Song, H and Li, G.** 2008. Tourism demand modelling and forecasting—A review of recent research. *Tourism Management*, 29(2): 203–220. DOI: <https://doi.org/10.1016/j.tourman.2007.07.016>
- Subramanian, J, Stidham, S, Jr and Lautenbacher, CJ.** 1999. Airline yield management with overbooking, cancellations, and no-shows. *Transportation Science*, 33(2): 147–167. DOI: <https://doi.org/10.1287/trsc.33.2.147>
- Talluri, KT and Van Ryzin, G.** 2005. *The theory and practice of revenue management*. New York, NY: Springer. DOI: <https://doi.org/10.1007/b139000>
- Weatherford, LR and Kimes, SE.** 2003. A comparison of forecasting methods for hotel revenue management. *International Journal of Forecasting*, 19(3): 401–415. DOI: [https://doi.org/10.1016/S0169-2070\(02\)00011-0](https://doi.org/10.1016/S0169-2070(02)00011-0)
- Yoon, MG, Lee, HY and Song, YS.** 2012. Linear approximation approach for a stochastic seat allocation problem with cancellation & refund policy in airlines. *Journal of Air Transport Management*, 23: 41–46. DOI: <https://doi.org/10.1016/j.jairtraman.2012.01.013>

How to cite this article: Antonio, N, de Almeida, A and Nunes, L. 2019. An Automated Machine Learning Based Decision Support System to Predict Hotel Booking Cancellations. *Data Science Journal*, 18: 32, pp.1–20. DOI: <https://doi.org/10.5334/dsj-2019-032>

Submitted: 05 December 2018 **Accepted:** 19 June 2019 **Published:** 08 July 2019

Copyright: © 2019 The Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International License (CC-BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited. See <http://creativecommons.org/licenses/by/4.0/>.

 *Data Science Journal* is a peer-reviewed open access journal published by Ubiquity Press.

OPEN ACCESS 