Cancellation Predictor for Revenue Management

applied in the hospitality industry

Vrije Universiteit Amsterdam
Business Analytics - Research Paper

Supervisor:

prof.dr. G.M. Koole
ger.koole@vu.nl

D. S. Hopman
daniel.hopman@vu.nl

Student:

R. van Leeuwen
rik.van.leeuwen@student.vu.nl

13 februari 2018

Samenvatting Cancellations of reservations influence the revenue significantly, a cancellation model is key to counter this problem. A Passenger Name Record (PNR) approach is used to create a classification model which is applied on a dataset of an international hospitality company, with the aim to increase revenue. The goal is to create a model which can be implemented, not only, in the hospitality industry but also in the airline industry or the car rental industry for example. This paper, part of the Master's program of Business Analytics at Vrije Universiteit Amsterdam, describes the steps from raw data until implementation. Four models will be applied: Naïve Bayes, logistic regression, decision tree and random forest. In terms of accuracy, precision and F-score, random forest performs the best. The features refundable (y/n), lead time, channel are important ones, according to the different models.

Keywords: revenue management, cancellation model, Passenger Name Record (PNR), machine learning, hospitality

Inhoudsopgave

1 Introduction

Selling the *right* room to the *right* customer at the *right* moment for the *right* price is the challenge in the hospitality industry. With revenue management (RM) strategies, hotels attempt to optimize their revenue with, for example, dynamic pricing and allocation (Talluri and van Ryzin [1]). The classical way of RM in the hospitality industry is selling a fixed number (the capacity) of rooms, which are perishable at a fixed deadline (the booking horizon). Based on historical reservations, market information, guest information and more available information, hotels choose optimal controls in the form of dynamic pricing and capacity allocation to maximize their revenue. These controls are the price setting and availabilities for various room types. RM is mainly associated with the airline industry and hospitality industry, but it is also applied in the car rental industry and the financial sector.

In today's world, hotels offer refundable and non-refundable rates to guests. Recently, there's an increased interest in refundable rates, where guests still have to possibility to cancel (last-minute). Whereas guests value the flexibility, hotels are dealing with the risk of empty rooms, and thus loss of revenue, which is a problem for the industry. As such, RM systems can be further improved by taking cancellations into account. High cancellation rates can lead to consequent loss of revenue due to empty rooms. With last-minute cancellations and "no-shows", the capacity allocation is no longer optimal because hotels do not succeed in attracting guests on such short notice.

There are roughly two ways of taking cancellations into account: first Rajopadhye [2] introduced a new method of giving the RM system "net demand". Net demand is defined as the number of bookings minus the number of cancellations. A disadvantage of this method is that a part of the reality is neglected, which may result in adding uncertainty to the model. The second method is incorporating a cancellation predictor into the RM system. This approach can give insights in the behavior of cancellations.

This paper describes an approach how an industry, such as hospitality, can overcome this problem by creating a cancellation predictor. This system allows controlled overbookings and relies on subsequent cancellations to keep the remaining number of bookings at the check-in date at, or just below, the capacity of the hotel. In order to have a successful overbooking policy, a prediction model for cancellations is key, where the focus of this paper is aimed at.

The reasons of cancellation may happen for multiple reasons: illness, bad weather or a rescheduled meeting. Additionally, Chen and Xie (2013) and Chen, Schwartz, and Vargas [3] rightfully mention that cancellations occur due to deal-seeking customers via Online Traval Agencies, such as Booking.com and Expedia. This particular group of guests value the cancellation policies more than others. But it is fair to state that the hospitality sector does allow overbookings to cut down the losses of cancellations. According to Rothstein [4], industries allow overbookings since 1985. Even though there is a risk of having more arriving guests than the capacity of the hotel.

To ensure optimal capacity allocation, the risks of overbooking have to be specified and calculated. It may occur that there are not enough cancellations on the check-in date, and thus overbooking is a problem. In absence of "last-minute"cancellations or "no-shows", a hotel may have out-of-order rooms that can be used in such emergency. As a second solution, a hotel could check if customers can go to another property of the same brand in the surrounding area. Third, it could be that an employee occupies a room in the hotel. Then the employee could go to another property to ensure the guests stay at the hotel where they originally made the reservation. If all aforementioned options are unavailable, then the hotel manager needs to make a reservation at its competitor. Because a customer expects a certain quality level, the competitor should comply with this level. Additionally, the hotel manager can compensate for the inconvenience by offering vouchers, discounts at bars or a night for free to the customer. Because of the additional costs of overbooking and potential reputation damage, it is important to keep the risk of overbooking low.

The remaining part of the paper is structured as follows. In Section 2, several approaches and their corresponding cancellation models are discussed. An overview of the available data and an explanatory analysis are presented in Section 3. In Section 4, the data preprocessing is described. This is an important step for every machine learning algorithm: creating the input and selecting available data. The methodology is described in Section 5. Section 6 is dedicated to the implementation of the model. In Section 7, the results are presented. Finally, Section 8 includes the conclusion and discussion.

2 Literature Review

Cancellation models are applied in the hospitality industry as well as in the airline industry, but are complex due to the various features and models. Benefits can be achieved, in terms of revenue or competition, when a cancellation model is added to a revenue management (RM) system. The system underestimates the demand when the cancellations are overestimated, which leads to miscalculated capacity allocation or too low price setting. The consequence of underestimating may result in being fully occupied too many days before the check-in date, which leads to a decrease in revenue. There are roughly two different approaches for cancellation models: first, forecasting of cancellation rates and second, classify each reservation individually, so-called Passenger Name Record (PNR) approach.

Accurately forecasting the arrivals is one of the key inputs of a successful RM system. Pölt [5] stated that, on a general setting, a reduction of 20 percent forecast error (demand, capacity and price forecasting) can translate into 1 percent of incremental increase of revenue by a RM system. As with arrivals, the forecasting of cancellation rate is also a technique that is applied in the hospitality industry and airline industry. Morales & Wang [6] describes such a technique with the application of data mining techniques. They concluded that a combination of multiple models is necessary in order to deal with time-dependency (booking horizon).

However, a single model could give additional information about the dynamics of features in the model. Petraru [7] concludes that cancellation rate forecasting in combination with a overbooking policy can increase the revenue of airlines. A company should be careful with such a policy and not be too aggressive with overbookings, otherwise the benefits would be offset by the costs of compensating denied guests. Cancellation forecasting could increase revenue gains by 0.12% even when no overbooking is applied. Revenue managers or RM systems can change the price, knowing the forecasted cancellation rate, in order to attract more demand. When the number of overbookings is increased, the revenue gain could end up between 1.15% and 3.13%.

Classifying cancellations based on PNR is popular in the airline industry according to Petraru [7]. With such a structure, data could easily be transfered between other airlines when a passenger has multiple flights, with different airlines, to reach their destination. It is a record based data structure that includes information about the reservation and the guest. Examples of features that are reservation related are reservation date, check-in date, price and via which channel the reservation has been made. Guest related features are name, surname, date of birth, gender and more. This approach takes the cancellations on a more disaggregated level into account than forecasting. There is no defined structure of a PNR system, so a company determines the features that they want to save.

In the hospitality industry, a few papers are written about cancellation models with such a PNR approach. Antnio, Almeida and Nunes [8] used, among others, a decision tree with a result of 98.6% accuracy. With the use of classification, a revenue manager gains insights about the dynamics of a cancellation. And so, revenue managers can anticipate better in case of high cancellation rates.

3 Data

3.1 Available Data

For this article, data of an international hospitality company is available, containing reservation data and guest data. The guest data is connected to the reservation data via an unique key. In this section, the available features are described and basic statistics are presented.

There are 1,277,844 reservations registered from 2009-01-01 until 2017-01-01. Over these years, several properties opened their doors. Therefore, not every property has the same number of reservations, also due to the different capacity per property. The property names are converted to numbers due to confidentiality, the properties are in this paper named 1 till 7. The cancellation rate of all reservations is 16.3%, but this rate varies over different properties, with the lowest rate of 11,5% for property 1 and the highest rate of 28.3% for property 5. An overview of the statistics about the number of records and the cancellation rate per property can be found in Table 1.

Property	**Records**	**Cancellation Rate**
1	374,941	11.5%
2	146,391	15.4%
3	219,221	12,3%
4	148,652	20.9%
5	78,525	28.3%
6	76,880	17.6%
7	233,202	20.1%
Overall	1,277,844	16.3%

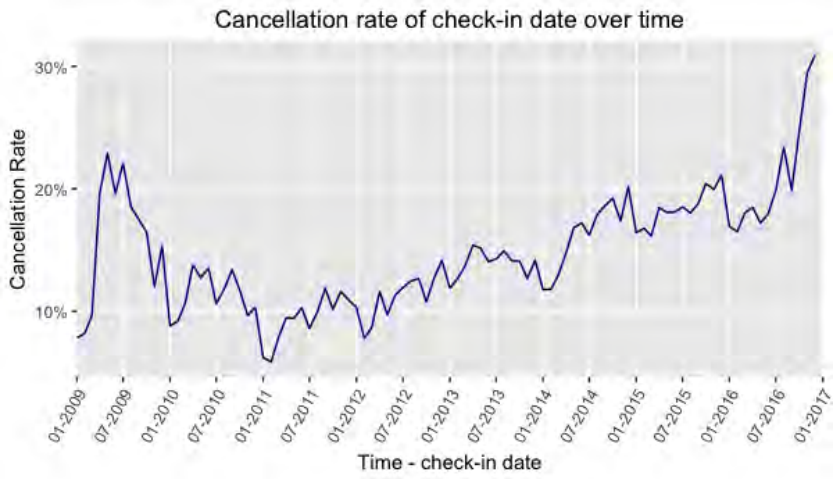
Tabel 1. The initial number of records and cancellation rate per property

The reservation data and guest data are merged into a single row in order to obtain a single Passenger Name Record (PNR) setup instead of two separate tables. The reservation-, arrival- and departure date are part of the initial features. Other features are hotel code, room rate (rate per night) and channel. A maximum of two guests can stay in a single room, therefore the number of guests is a feature. Every reservation is made with a rate plan, which mainly indicates if a reservation is refundable or whether a guest has booked a breakfast. There is a column that indicates this rate plan. It is possible to make a reservation as a group, which is indicated by a code. Such a group reservation is usually offered a discounted rate. This code is unique for every group reservation. There is a possibility that companies can make a reservation, if this is the case the company name can be found as one of the features. These corporate reservations often benefit from a deal with the company, for instance a fixed rate or a discount. Specific information is included in the rate codes of those reservations. The guest information is largely unknown. For example, data availability for age, gender,

and city is only 0.01%, 4.3%, and 4.4%, respectively. For this reason only the email address is taken into account in order to calculate repeat percentages. The last feature is the status of the reservation that indicates if there was a check-out, cancellation or no-show.

3.2 Exploratory data analysis

From the initial data, valuable insights are gained and presented below by summarizing the main characteristics. The cancellation rate over time of the check-in date is presented in Figure 1. There is an upwards trend over the last five years. The deal seekers, as mentioned in the introduction, are possibly one of the causes of this increase. Another factor could be the Online Travel Agencies, who make it more attractive to place a reservation via their system. They are adding extras to the reservation, for example a flexible cancellation option. During the first quarter of each year, the rate is the lowest except in 2009. Revenue managers at this hospitality company confirm that January is a difficult period attracting demand, which result in a lower cancellation rate. This results in a cancellation rate which is lower than other months. The rate in the last quarter of 2016 is around 10% higher than the same quarters previous years.



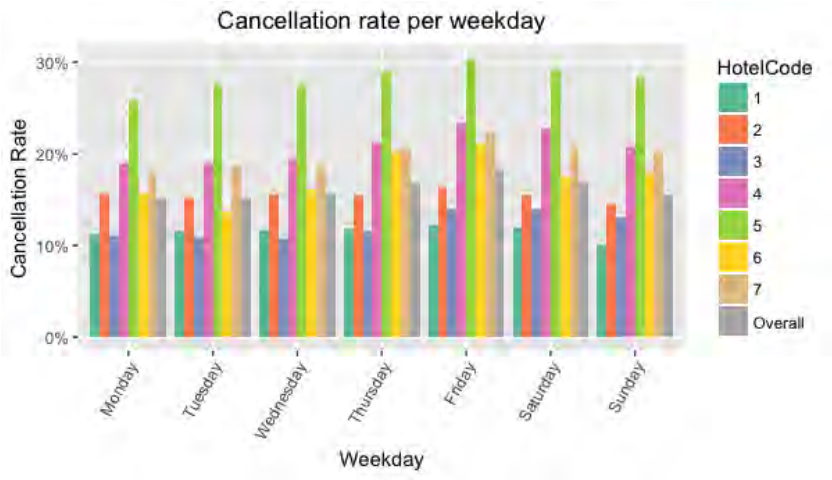
Figuur 1. Cancellation rate of check-in date over time

Because in Figure 1 the cancellation differers per quarter, a table has been made of the cancellation rate per month of the check-in date, see Table 3. In this overview, the overall cancellation rate is also given. The differences between properties are significant, for example in December where property 5 has a cancellation rate of 37.2% and property 1 is 12.3%.

Property	Jan	Feb	Mar	Apr	May	June	July	Aug	Sept	Oct	Nov	Dec
1	10.6	10.1	10.8	13.0	12.2	11.6	10.3	11.3	12.5	11.6	12.4	12.3
2	14.5	14.3	13.9	14.9	14.5	16.4	13.7	14.3	16.8	16.3	18.7	16.1
3	10.7	11.1	12.2	11.3	10.8	11.7	12.1	14.2	14.2	12.9	12.8	14.0
4	17.9	18.7	19.7	20.3	21.6	22.2	22.5	23.0	20.4	20.8	19.2	24.1
5	30.2	26.0	27.7	24.3	25.3	24.1	26.5	27.3	29.1	31.7	30.9	37.2
6	12.4	14.6	15.4	17.0	16.8	16.3	21.6	16.1	18.8	26.8	16.3	18.7
7	19.2	20.5	18.3	20.3	17.4	21.8	21.7	22.2	21.6	21.6	18.3	20.0
Overall	14.3	14.2	15.1	16.7	16.3	16.3	16.1	16.8	17.1	17.5	16.6	18.4

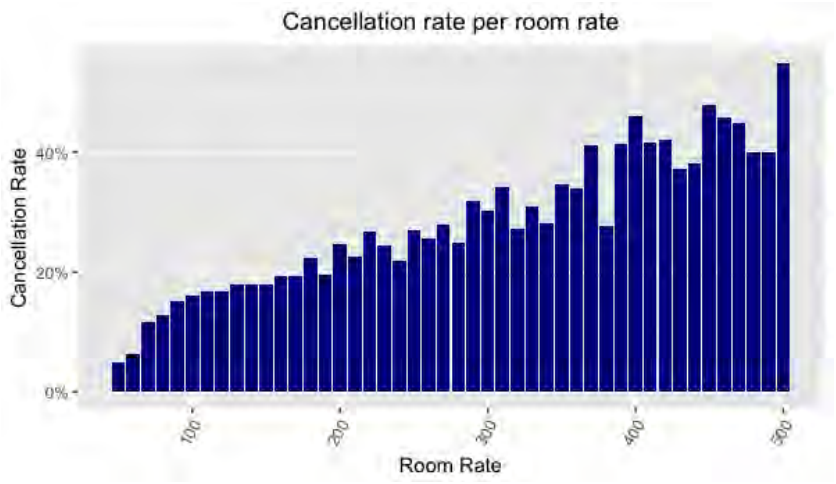
Tabel 2. Cancellation rate per month per property and overall

Another way to look at the check-in dates is per weekday. Figure 2 shows the cancellation rate per weekday of the check-in date for each property. The overall rate per weekday is also presented in Figure 2. Again, property 5 has the highest cancellation rate for each single weekday. The differences per weekday are in some cases over 15%, at Friday for example between property 5 and property 3. The overall cancellation rate per weekday differs at the most 1.1%, so that is not a large difference. But for individual properties, this difference is around 5%.



Figuur 2. Cancellation rate per weekday

Figure 3 contains a visualization of the cancellation rate per room rate aggregated into bins of 10 euro. A minimum and maximum price per hotel are defined, but due to manual changes the room rate could differ from that interval. These observations, which are 1.2% of the total, are not shown in the visualization. The rate of cancellation increases when the price per night increases.



Figuur 3. Cancellation rate per room rate

There are four channels through which reservations can come in, these are BOOKING, DIR, HCM and WEB. The first channel is from Online Travel Agencies (OTA) such as Booking.com and Expedia, the second channel is direct walk-ins, the third is for companies and fourth is via their own website and other online services. The overall distribution of channels is respectively 43.6%, 6.5%, 10.8% and 39.1%. The corresponding cancellation rate is 21.8 %, 12.8%, 12.4% and 8.1%. There are differences per channel, but these numbers increase when the cancellation rate per channel per hotel is computed. In Table 3 an overview is given of these statistics. For each property individually, the cancellation rate of OTA's is the highest. The highest cancellation rate of all is property 5 when reservation come through this channel, which is also the property with the highest overall cancellation rate.

Property	**OTA**	**DIR**	**HCM**	**WEB**
1	14.4	6.2	11.8	9.4
2	17.0	5.4	12.9	13.5
3	18.8	9.2	9.9	9.6
4	31.1	9.5	15.6	14.6
5	36.4	10.9	14.0	23.6
6	20.9	8.5	12.3	16.4
7	27.9	8.7	16.3	14.8
Overall	21.8	12.8	12.4	8.1

Tabel 3. Cancellation rate per hotel per channel

4 Processing

4.1 Feature Engineering

The creation of new features can boost the performance of a model. Howbert [9] states that well-conceived new features can sometimes capture the important information in a dataset much more effective than the original features. This can be achieved by creating a new feature or map exiting features to new space for example. From the initial features, new features are created.

A date by itself cannot be used in classification problem. However, dates can be transformed to categories, such as weekday. As shown in Section 3, the month and weekday of the check-in date can contain differences between properties which can be valuable for a model. The weekday and month is derived from reservation date, check-in date and check-out date. Resulting in six new features, which are named as follows ReservationMonth, ReservationWeekday, ArrivalMonth, ArrivalWeekday, DepartureMonth and DepartureWeekday.

The number of days between dates can be a new feature, for example the length of stay and the lead time. The length of stay is defined as the number of nights a guest is planned to stay. The lead time is defined as the number of days between reservation date and arrival date. Another feature that was created is called NumBusinessNights, which is the number of business night. A night of the week is defined as a business night if the night is between Monday and Thursday.

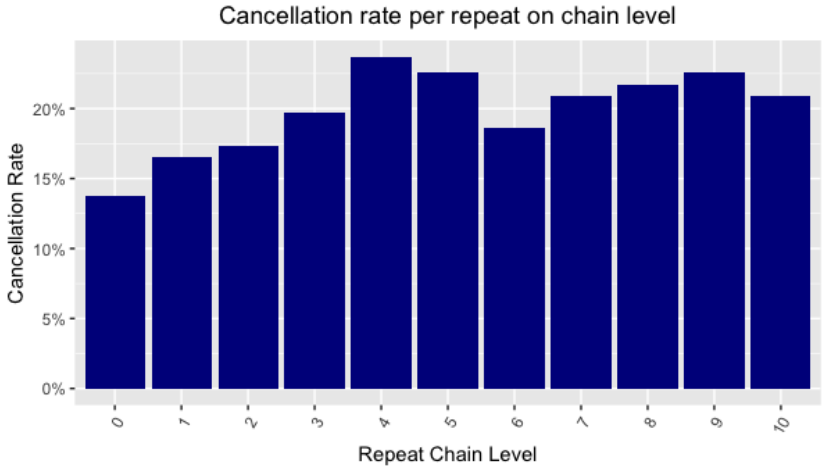
An algorithm does not gain information of different group codes or company codes. Therefore, these numbers are translated to binary columns, because group codes are unique and company codes are unique per company. Both column indicate whenever the reservation is part of a group reservation or a company reservation.

The feature RateCode has 353 unique values, 65.4% of these unique values occur less than 100 times. These codes change over time and many of these are not used for several years, for example with an opening of a property in 2011. However, each of them defines if the reservation is refundable. For this reason, RateCode is changed to a binary column where 1 is refundable and 0 is non-refundable reservation.

Unfortunately, the quality of demographic information of guests is poor. However, every email address has a domain, for example .nl for email addresses from the Netherlands or .uk for email addresses from Great Britain. Each reservation has a feature CountryCode which contains the country of the email domain. There are 194 unique CountryCodes matched to all the email address. There are some domains which does not belong to a country, such as .com or .net. These domains are labeled otherwise, respectively Commercial or Network. Table 9, which can be found in the Appendix, contains the top 20 frequency of the column CountryCode, which covers 90.3% of the records. The goals is to see if a machine learning technique finds dependency between cancellation and CountryCode.

The final added features contain information about repeats and cancellations, which are based on email address. One of the repeat features is based on chain

level and the other on hotel level. The same as for the cancellation features. In total there are four new features added, named RepeatChain, RepeatHotel, CancellationsChain, CancellationsHotel. These features will be set to 0 for a new customer, because now a model is still able to classify the new instance. Ideally, the outcome for a hospitality company is that the higher number of repeats, the cancellation rate lowers. With such an outcome companies can measure the loyalty of their guests. Tepeci [10] states that a loyalty program is more profitable than other marketing activities, such as price cuts or promotional programs. In Figure 4, the cancellation rate the repeat number 1 to 10 on chain level is visualized, the rate fluctuates between 16% and 23 %. The number of repeats has been set to a maximum of 10, due to readability of the graph.



Figuur 4. Cancellation rate per repeat on chain level

4.2 Selection

The overall quality of the data is high, but some features require a selection. A subset selection reduces the dimensionality of the data or feature without creating new features. According to Howbert [9], a selection can be beneficial if redundant, irrelevant or noisy features are removed. All in order to speed up the learning process of the model, enhance generalization and alleviate the curse of dimensionality.

The first step in the selection process is deleting test reservations, which are made to test the system. The RateCode tells whenever a record was a test, there are 1308 records removed from the dataset.

A revenue manager should know in advance how many reservations are likely to cancel otherwise no actions can be executed to counter empty rooms. There

are some records that have a negative lead time. This implies that first a guest stayed in the hotel and after or during the stay placed a reservation. Various minor technical reasons or situations could cause this, for example a reservation is extended with a few nights. Therefore, the records with a lead time smaller than 0 are excluded from the dataset, since it is 35,293 and 3.3% of these records is canceled.

The room rate is bounded by the minimum and maximum price, which is different for each property. These boundaries reduces 1.2% of the total amount of reservation. These deviations are caused by manual changes of employees of the company. There is no log of the original price, unfortunately the rate could not be restored. The number of records is unknown which are manually changed and are still in between the bounds.

Because this paper uses the Passenger Name Record (PNR) approach, the records without a email address are removed from the dataset since the repeat statistics are based on email address. In total there are 96,451 records removed, which is 7.5% of the total amount.

Besides removing records, a limit can be set to the maximum number of unique values of a feature. The feature CountryCode needs such a limit since there are 194 different domains labeled. Some of the values only occur one, therefore a limit will be set to the maximum of 20. The rest of the 9.7% will be set at a new category 'Other'. This is also applied on the Channel feature, which is discussed in Section 3. As with the CountryCode, the maximum number of repeats are set to a maximum of 10.

From the creation of features, records can be detected that contains a error. For example with the creation of the LeadTime feature. In some cases, the lead time was below zero. Which is not possible, because it implies that a guest first stayed in a hotel and later on a reservation was made. These reservations are removed from the data set.

In the end, 17.5% records are removed from the dataset. The number of remaining records per property can be found in Table 10, in the Appendix with the corresponding reduction per property. There are still over a million reservation in the data set available where the algorithms can be trained and tested on. All of the feature engineering and selection has been done with the use of SQL Server. Table 4 shows an overview of the features, types, possible values and a description. This is the set of features that will be used as input for each of the machine learning algorithms.

Variable	Type	Possible Values	Description
ReservationDateWeekday	Nominal	7	Weekday of reservation date
ReservationDateMonth	Nominal	12	Month of reservation date
LeadTime	Integer	365	Days between reservation date and check-in date
CheckInDateWeekday	Nominal	7	Weekday of check-in date
CheckInDateMonth	Nominal	12	Month of check-in date
CheckOutDateWeekday	Nominal	7	Weekday of check-out date
CheckOutDateMonth	Nominal	12	Month of check-out date
HotelCode	Nominal	7	Hotel of reservation
LOS	Integer	40	Length of stay
NumBusinessNights	Integer	31	Number of nights between Monday and Thursday
Refundable	Boolean	2	Is the reservation refundable?
PartOfGroup	Boolean	2	Part of a group booking?
PartOfCompany	Boolean	2	Company booking?
RoomRate	Integer	56	Price per night
RoomRevenue	Integer	304	Total price of room
ChannelCode	Nominal	4 (after grouping)	Channel used to make the reservation
NumberGuests	Integer	2	Number of guests
CountryCode	Nominal	20 (after grouping)	Country of email domain
RepeatChain	Integer	10 (after grouping)	Number of repeats of guest
RepeatHotel	Integer	10 (after grouping)	Number of repeats of guest at hotel
CancellationsChain	Integer	10 (after grouping)	Number of cancellations of guest
CancellationsHotel	Integer	10 (after grouping)	Number of cancellations of guest at hotel
ReservationStatus	Boolean	2	Is the booking canceled?

Tabel 4. Overview of the all features

5 Methodology

The input and output of the model is defined in the previous sections, therefore it is a supervised classification problem. Supervised learning is the machine learning task of mapping a function to labeled training data, according to Mohri et al. [11]. This way of classification will analyses the training data and apply the mapped function to new data examples. The ideal outcome is that the mapped function will accurately classify unseen data points. Generally, the input is transformed into a vector of features. Attention should be paid to the number of input features due to the curse of dimensionality, but there should be enough information to accurately classify the new data point. According to Bellman [12], the curse of dimensionally occurs when data in high-dimensional spaces get analyzed.

5.1 Naïve Bayes

The Naïve Bayes model is a probabilistic classifier based on Bayes' theorem. This theorem estimates the probability of an event based on prior knowledge that are associated to that event such as features. Equation (1) describes Bayes' theorem with the assumptions that A and B are events and $P(B) \neq 0$.

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \tag{1}$$

To break down Equation (1), $P(A)$ and $P(B)$ are the observation of A and B without regard of each other. The conditional probability is $P(B|A)$, the probability of event B given A. And, $P(A|B)$ is the probability of event A given B. In a Naïve Bayes model, A is defined as the possible outcome classes, for this problem it is '1' or '0', and B is defined as a vector of the n features that are known.

One of the advantages of a Naïve Bayes classifier is the ability to converge quickly and it is easy to implement, which makes it possible to be trained more often than other algorithms. While this model is typically seen as a simple one, it may still perform as good as, perhaps even better, than a more complex model. The simplicity may be seen as an advantage: easy to understand or a disadvantage: the algorithm may be too simplistic.

5.2 Logistic Regression

Logistic regression is a type of regression model where the dependent variable is categorical, which is in this paper a binary variable '1' and '0' that is representing if the reservation is canceled yes/no. The model was developed by Cox in 1958 [13]. The model estimates the probability of a binary response. These type of models can handle features that are continuous and/or categorical.

The model uses a logistic function to estimate these probabilities. The input is any real input t. The function is defined as Equation (2).

$$\sigma(t) = \frac{1}{1 + e^{-t}} \tag{2}$$

In order to improve these models, the logarithm of numerical features are added. From the Table 4, there are 10 features which are integer. So, there are ten new features added for this model.

The advantage of logistic regression is that the outcome is a probability, therefore the classification the border for '1' or '0' can be shifted to an optimum. Typically, the variance of such a model is low, which makes it more robust to noise in the dataset. The trade off is, in general, a higher bias, which is missing relevant relations between features.

5.3 Decision Tree

A decision tree is an algorithm that builds a flow-chart like illustration that shows a possible outcome of a decision. In order to build a tree, the algorithm finds variables that separate the data into two groups the best. The algorithm prefers splits such that it classifies the maximum number of observation correctly. This step is repeated until the stopping criteria have been met, which are a minimum number of observations per category before attempting to split the tree and a split must decrease the overall lack of fit by a factor. The latter is also known as the cost complexity factor. A decision tree is know for a low bias but a high variance.

An advantage of this algorithm is that the graphical representation makes it intuitive to understand, which is a plus for revenue managers because they typically do not have experience in machine learning algorithms. A disadvantage could be over fitting, because a tree can easily grow. This downside can be countered by the stopping criteria or by applying a random forest, which is explained in the next subsection.

5.4 Random Forest

Random forest is an ensemble approach that operates by composing a aggregation of decision trees. Ho [16] came in 1995 with a first idea of random decision forests but Breiman [17] extended that idea in 2001 as what now is known as the random forest algorithm. The principle behind ensemble methods is combining 'weak' learners in order to form together a 'strong' learner. A sample is taken from the data, and the tree will be built with that sample. Every time a new sample is taken, and so a new tree is built. The number of trees is a setting of the algorithm. When a new input is presented to the random forest, it will be classified by each of the trees. The average of the outcome will be taken and so the new input will be classified.

One of the advantages is a fast runtime and it counters overfitting. But the more trees there are set to train the model, the longer it takes to predict new data. Generally, in practical situations it is fast enough, but the run-time performance should be taken into consideration.

5.5 Approach

Cross-validation is applied to evaluate the performance of the models, to be more specific k-fold cross-validation, described by Hastie et al. [14] . Smola and Vishwanathan [15] showed that k-fold cross-validation can be computationally expense. But it overcomes the threat of overfitting of a model, for this research $k = 10$ is taken. The idea behind this technique is splitting k times the dataset into a training set and test set. These sets are respectively 80% of the total records of a property and 20% of the total records of a property. Each model is trained on the same training set and tested on the same test set. The accuracy of each outcome will be computed. Eventually, the models are hold against a benchmark to rank the models. The benchmark is defined as the accuracy if the are no predicted cancellations. The benchmark per property is derived from the CancellationRate column of Table 1. Next to the accuracy, the precision and recall is computed. The precision is defined as the percentage of correctly classified instances decided by the total number of test instances. Recall is also known as the sensitivity, which is the number of true positives divided by all correctly classified test instances. A measurement that contains the precision and recall is the F-score, which is a score about the test's accuracy. For completeness, the confusion matrix is also shown in Section 7. From the confusion matrix, the above three measurements are computed.

Models will be created for each property individually since the differences between models are significant for several features, for example the cancellation rate per month differs over 15%. As an initial model, a Naïve Bayes classifier has been used to confirm this initial thought. The benchmark, after the cleaning and selection, is 83.7%. A general model has been made and the accuracy is 80.4%. There is a difference of 3.3% between the benchmark and the accuracy. But when a model is created for each hotel, the difference, for each hotel, is smaller than 3.3%. Further results are presented in Section 8.

By applying machine learning algorithms to each property, the dynamics between features per property can be highlighted. These dynamics will help the revenue managers to understand whenever the cancellations act in similar way for different properties. The importance tells which features are used in the model and which are not. Each feature will get a score and the five highest ranked scores per property per model are presented in tables in Section 8. This importance can be extracted for logistic regression, decision tree and random forest. It is impossible to make a ranking between the features of the Naïve Bayes model. Because only the a-priori probabilities could be extracted from each model. That is, how frequently each level of class occurs in the training dataset. But this does not indicate the importance of the features of the Naïve Bayes model.

6 Results

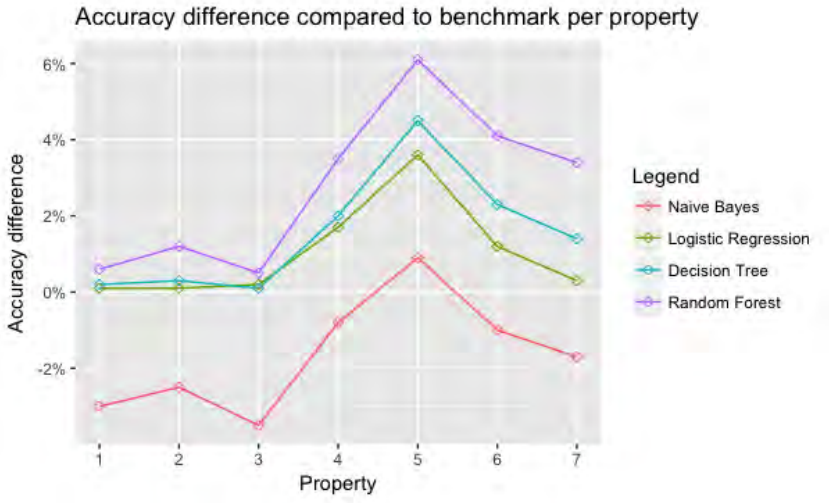
In this section the results of each model and the importance of features will be discussed in order to gain extra insights about the dynamics within a machine learning algorithm. The confusion matrix, accuracy, precision, recall and F-score per model for each hotel is presented in Table 5. For each hotel, random forest has the highest accuracy and is also the only on that is above the benchmark with property 3. After the random forest, the logistic regression and decision tree are competitive to each other, at six of the seven properties decision tree is better than logistic regression. At an accuracy point of view, the Naïve Bayes model does cross the benchmark only once. In Figure 5, a graphical overview is given of the accuracy compared to the benchmark.

Prop.	Algorithm	TP	FP	FN	TN	Benchmark	Accuracy	Precision	Recall	F-score
1	Naïve Bayes	48340	5768	2664	951	0.884	0.854	0.893	0.948	0.920
	Logistic regression	50696	6532	37	77	0.884	0.885	0.886	0.999	0.939
	Decision tree	50789	6353	250	331	0.884	0.886	0.889	0.995	0.939
	Random forest	50129	5652	627	933	0.884	0.890	0.899	0.988	0.941
2	Naïve Bayes	21043	3323	1421	785	0.846	0.821	0.864	0.937	0.899
	Logistic regression	22202	3930	100	157	0.846	0.847	0.850	0.996	0.917
	Decision tree	22196	3752	267	356	0.846	0.849	0.855	0.988	0.917
	Random forest	21745	3161	583	900	0.846	0.858	0.873	0.974	0.921
3	Naïve Bayes	30042	3779	2014	720	0.877	0.842	0.888	0.937	0.912
	Logistic regression	31577	4297	56	85	0.877	0.879	0.880	0.998	0.936
	Decision tree	31704	4108	355	390	0.877	0.878	0.885	0.989	0.934
	Random forest	30868	3504	729	914	0.877	0.882	0.898	0.977	0.936
4	Naïve Bayes	18542	3795	1772	1585	0.791	0.783	0.830	0.913	0.869
	Logistic regression	19427	4397	432	902	0.791	0.808	0.815	0.978	0.889
	Decision tree	19301	3848	1008	1537	0.791	0.811	0.834	0.950	0.888
	Random forest	18637	3154	1223	2144	0.791	0.826	0.855	0.938	0.895
5	Naïve Bayes	8701	2377	1546	1705	0.717	0.726	0.785	0.849	0.816
	Logistic regression	9175	2700	754	1332	0.717	0.753	0.773	0.924	0.842
	Decision tree	9001	2149	1264	1914	0.717	0.762	0.807	0.877	0.841
	Random forest	8715	1873	1225	2148	0.717	0.778	0.823	0.877	0.849
6	Naïve Bayes	10730	1801	798	644	0.824	0.814	0.856	0.931	0.892
	Logistic regression	10890	2012	191	364	0.824	0.836	0.844	0.983	0.908
	Decision tree	11417	1821	354	640	0.824	0.847	0.862	0.970	0.913
	Random forest	11157	1411	404	967	0.824	0.865	0.883	0.964	0.922
7	Naïve Bayes	26612	5719	2108	1451	0.799	0.782	0.823	0.927	0.872
	Logistic regression	25866	6181	361	631	0.799	0.802	0.807	0.986	0.888
	Decision tree	27071	5093	1615	2110	0.799	0.813	0.842	0.944	0.890
	Random forest	24665	3981	1549	2844	0.799	0.833	0.861	0.941	0.899

Tabel 5. Overview of performance of the algorithms per property

In terms of precision, the random forest was also the best model out of the four. Decision tree is placed the second, except for property 3 with a difference of 0.01 in comparison with logistic regression. Third ranked is logistic regression and fourth is the Naïve Bayes model. The F-score of random forest, which is a combination between precision and recall, is the highest ranked at all properties, except for property 3 there is the measure equal to the logistic regression model. Furthermore, the results are in-line with the findings of Fernández-Delgado, Cernadas, Barro and Amorim [18] in 2014. They tested 179 classifiers from 17 families and concluded that classifiers from random forest family performs generally the best.

The False Negatives are fore some properties higher in comparison with the True Negatives regarding the Naïve Bayes algorithm. Which implies that such a algorithm has great difficulties with classifying the instances correctly. For all the other algorithms the False Negatives are never higher than the True Negatives. However, the number of False Negatives are relatively high with respect to the True Negatives.



Figuur 5. Accuracy difference compared to the benchmark per property

Next, the dynamics of each algorithm will be discussed in order to gain extra insights of the differences between the models, except for the Naïve Bayes. Starting with the model that has the highest accuracy of the four, random forest. In Table 6 the five highest ranked features per property are presented. The feature Refundable is the most important one, which is expected because this feature tells if it is even possible to cancel the reservation. The LeadTime is for five of the seven properties the second most important one. So the number of days before a reservation was made, plays a crucial role in cancellation behavior.

Prop.	1	2	3	4	5
1	Refundable	LeadTime	ChannelCode	Reservation-DateWeekday	NumberGuests
2	Refundable	LeadTime	ChannelCode	Reservation-DateWeekday	RoomRevenue
3	Refundable	ChannelCode	LeadTime	RoomRevenue	RoomRate
4	Refundable	LeadTime	ChannelCode	RoomRate	Reservation-DateWeekday
5	Refundable	ChannelCode	LeadTime	RoomRate	CheckIn-DateWeekday
6	Refundable	LeadTime	Reservation-DateWeekday	ChannelCode	RoomRevenue
7	Refundable	LeadTime	ChannelCode	Reservation-DateWeekday	Reservation-DateMonth

Tabel 6. Ranked feature importance for each property with Random Forest

In Table 7 the five highest ranked features per property are presented of the logistic regression model. The feature Refundable is always at the top, which is logical because when a booking in non-refundable the guest cannot cancel their reservation. ChannelCode and LeadTimeLog are quite competitive to each other, these are the second and third most important features. The other important features are LeadTime, NumberOfGuests, ReservationDateMonth and CheckInDateWeekday. In comparison with the random forest, the highest three features, Refundable, LeadTime and ChannelCode, are the same, except for the log taken of a numerical feature.

Prop.	1	2	3	4	5
1	Refundable	LeadTimeLog	ChannelCode	NumberGuests	PartOfGroup
2	Refundable	LeadTimeLog	ChannelCode	LeadTime	ReservationDateMonth
3	Refundable	LeadTimeLog	ChannelCode	PartOfGroup	ReservationDateMonth
4	Refundable	ChannelCode	LeadTimeLog	PartOfGroup	CheckInDateWeekday
5	Refundable	ChannelCode	LeadTimeLog	LeadTime	ReservationDateMonth
6	Refundable	LeadTimeLog	ChannelCode	PartOfGroup	ReservationDateMonth
7	Refundable	LeadTimeLog	ChannelCode	PartOfGroup	NumberGuests

Tabel 7. Ranked feature importance for each property with Logistic Regression

The importance for the decision trees are extracted from the models, these results can be found in Table 8. The most important feature is not at every property Refundable, which is a different in comparison with random forest and logistic regression. The feature Refundable is ranked as first, second and even third. At six out of seven properties, the feature LeadTime is the most important one. From the previous two importance tables, Table 6 and Table 7,

this features is ranked second or third. The RoomRevenue and RoomRate do play a more important role for a single decision tree than at the random forest model.

Prop.	**1**	**2**	**3**	**4**	**5**
1	LeadTime	Refundable	RoomRevenue	ChannelCode	LOS
2	LeadTime	LOS	Refundable	RoomRevenue	NumBusinessNights
3	LeadTime	ChannelCode	Refundable	RoomRate	NumberGuests
4	Refundable	RoomRevenue	LOS	LeadTime	NumBusinessNights
5	LeadTime	Refundable	ChannelCode	Roomrevenue	NumberGuests
6	LeadTime	Refundable	CheckOutMonth	CheckInMonth	RoomRate
7	LeadTime	NumberGuests	Refundable	ChannelCode	PartOfGroup

Tabel 8. Ranked feature importance for each property with decision tree

7 Implementation

A machine learning algorithm may improve the overall objective of a department or could result in a competitive advantage. This goal will be achieved if a successful implementation leads to new insights or a more effective recourse usage. However, a machine learning algorithm has to engage employees in order to make use of it, for example by creating a sense of urgency. This sense of urgency can be created by explaining the problem of cancellations and the possible improvement in terms of revenue. Another way of engaging employees is by explaining what the idea of a model is and the performance of it. Therefore, a basic understanding about the model is preferable, implying that a white-box method is more desirable. There is the possibility to explain and interpret black-box method by decomposing the feature contribution. However, the idea of a black-box method can be explainable for employees that do not have experience in the field of machine learning. Take a random forest for example, in the basis it is a black-box method because the algorithm ensembles x number of trees. Yet, the concept is explainable, the model combines different decision trees and a decision tree is an intuitive algorithm. Revenue managers will understand the algorithm better if a decomposition of the features is given, which may be in line with their experience or not.

The number of expected cancellation should be communicated with the revenue management (RM) system, either automatically or manually. Extra rooms should be sold in order to anticipate on the potential loss in terms of revenue. With an automated approach, an algorithm or heuristic is of the essence which decides when there should be extra rooms available. This is key in order to counter the problem of cancellations. The number of predicted cancellations should not be added instantly to the number of available rooms in the RM system. In such a situation, when all the rooms are instantly added, there is a possibility that the revenue can be even lower instead of doing nothing. The demand should increase to gain revenue of those extra available rooms, and a way to do so is lowering the price. A price drop can result in a negative influence with respect to the revenue or the brand. For example, selling a room for a rate of 500 is better than selling five rooms for 99. It may happen that the current price is lower than the price of guests who made a reservation way in advance. These guests do not get 'rewarded' to be early-birds, which should be the case according to this international hospitality company. A consequence can be talking negatively about the brand.

Another way to make use of a cancellation model is to manually change the allocation or availability of rooms in the RM system. These changes should be made by revenue managers. This approach can be implemented faster than the automatic approach. A dashboard can be created which indicates which reservations are likely to cancel for each single day in the future. The revenue manager can look at which days there is a high cancellation rate, and so increase the number of available rooms for those days. However, the speed and complexity of nowadays revenue management system is such that the advantages of using models are not clear if there is a manual input setting added.

8 Conclusion and discussion

The data analysis in Section 3 provided insights in the dataset, which can be used for revenue managers to have a deeper understanding in cancellation behavior. Since 2010, the overall cancellation rate is slightly increasing. One of the causes could be deal-seekers, who are pointed out in the introduction. On an aggregated seasonal level, the mutual differences are quantified. These findings help in the understanding of seasonal influences. The focus of Section 3 was also highlighting the differences between properties. Revenue managers know that those are present, but now these are partly quantified. Next the the seasonal influence and the differences per property, the cancellation rate slightly increase when the room rate increases. This may have a negative influence on the revenue when the hotel is almost fully occupied and the room rate has been set too high. This situation may lead to extra cancellations.

From the results, presented in the Section 6, the conclusion can be made that random forest is the algorithm that performs the best. Due to the highest accuracy for each property and also in terms of precision, it beats the other models. The dynamics within this algorithm are consistent, which is measured with the importance of the features. As expected, the feature extracted from RateCode, the boolean Refundable, is the most important feature in each model. Also, the feature LeadTime, which is extracted from the difference between reservation date and check-in date, turned out to be a relevant feature as well. These kind of findings, confirm that feature engineering and selection are one of the important steps in creating a useful model.

As mentioned in the results, the False Negatives are relatively high with respect to the True Negatives. This implies that a negatives result of an instance is not confidence to be an actual negative result, especially for the Naïve Bayes algorithm. To counter this problem extensive research can be done in finding new features or adding new columns to the reservations. For example, the price of the room at the moment of canceling in comparison with the price at the moment of booking. Besides, there were not that many features present in the original dataset that contained information about the guest itself, such as point of sale or gender. This could have a positive influence on the classification process. Next to feature engineering, other models can be tested which are more black-box, for instance a neural network could be applied.

The approach that is suggested in this paper can be applied in different industries such as the airline industry and car rental industry, as well as the models that are applied. Inevitably, issues will be faced regarding the data selection and feature engineering because the data structure is in a different format, however the same type of steps can be taken.

Revenue managers would benefit from the implementation of this algorithm into a dashboard, because these predictions should create more control in certain situations. The ability to see which reservations are likely to cancel in a overbooking situation for example. Besides their experience, they can act with more knowledge in crucial situations with the goal of generating extra revenue.

Another point of discussion is the influence of competitors. Since the raise of Online Travel Agencies, such as Booking.com or Expedia, revenue managers should keep an eye on their direct competitors. There is no information taken into account about this external factor. There is no literature about the influence of competitors in hospitality, regarding dependence between cancellation and prices or ranking. It could be innovative research if it is proven that there is a form of dependence between cancellations and competitors.

9 Appendix

Position	CountryCode	Records
1	Commercial	645787
2	United Kingdom	72751
3	Netherlands	61533
4	Germany	32921
5	France	30203
6	Network	20063
7	Italy	13434
8	Belgium	13287
9	Europe	9041
10	Russia	7764
11	Switzerland	6407
12	Organization	5600
13	Japan	5167
14	Canada	4924
15	Education	4141
16	Australia	4099
17	Brazil	3944
18	Spain	3805
19	Norway	3540
20	Austria	3361

Tabel 9. Top 20 frequency of CountryCodes derived from email address

Property	Records	Reduction
1	288,610	23.0%
2	132,852	9.2%
3	182,762	16.6%
4	128,467	13.6%
5	71,637	8.8%
6	69,857	9.1%
7	179,393	23.1%
Overall	1,053,649	17.5%

Tabel 10. Final number of records and reduction percentage after selection

Referenties

1. K. T. Talluri, G. J. van Ryzin: The Theory and Practice of Revenue Management, Kluwer Academic Publishers, 2004.
2. M. Rajopadhye, M. B. Ghalia, P. P. Wang: Forecasting uncertain hotel room demand. Information Sciences, 2001.
3. C. Chen, Z. Schwartz, P. Vargas: 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, 2011.
4. M. Rothstein: OR and airline overbooking problem. Operations Research, 1985.
5. S. Pölt: Forecasting is difficult - especially if it refers to the future. Reservations and Yield Management Study Group Annual Meeting Proceedings, 1998.
6. D. Morales, J. Wang: Forecasting Cancellation Rates for Services Booking Revenue Management Using Data Mining, European Journal of Operational Research, 2009.
7. O. Petraru: Airline passenger cancellations : modeling, forecasting and impacts on revenue management, Massachusetts Institute of Technology, 2016.
8. N. Antnio, A. Almeida, L. Nunes: Predicting hotel booking cancellations to decrease uncertainty and increase revenue, Tourism & Management Studies, 2017.
9. J. Howbert: Introduction to Machine Learning, University of Washington Bothell, 2012.
10. M. Tepeci: Increasing brand loyalty in the hospitality industry, School of Hotel, Restaurant, and Recreation Management, 1999.
11. M. Mohri, A. Rostamizadeh, A. Talwalkar: Foundations of Machine Learning, MIT Press, 2012.
12. R. Bellman: Adaptive Control Processes: A Guided Tour, Princeton Legacy Library, 1961.
13. D. Cox: The Regression Analysis of Binary Sequences, Journal of the Royal Statistical Society, 1958.
14. T. Hastie, R. Tibshirani, J. Friedman: The elements of statistical learning, Springer series in statistics, 2001.
15. A. Smola, S. V. N. Vishwanathan, J. Friedman: Introduction to machine learning, Cambridge University Press, 2010.
16. T. Ho: Random Decision Forests, Proceedings of the 3rd International Conference on Document Analysis and Recognition, 1995.
17. L. Breiman: Random Forests, Machine Learning, 2001.
18. M. Fernández-Delgado, E. Cernadas, S. Barro, D. Amorim: Do we need hundreds of classifiers to solve real world classification problems?, The Journal of Machine Learning Research, 2014.