Forecasting Cancellation Rates for Services Booking Revenue Management Using Data Mining

Dolores Romero Morales * and Jingbo Wang †

June 7, 2009

Abstract

Revenue management (RM) enhances the revenues of a company by means of demand-management decisions. An RM system must take into account the possibility that a booking may be canceled, or that a booked customer may fail to show up at the time of service (no-show). We review the Passenger Name Record data mining based cancellation rate forecasting models proposed in the literature, which mainly address the no-show case. Using a real-world dataset, we illustrate how the set of relevant variables to describe cancellation behavior is very different in different stages of the booking horizon, which not only confirms the dynamic aspect of this problem, but will also help revenue managers better understand the drivers of cancellation. Finally, we examine the performance of the state-of-the-art data mining methods when applied to Passenger Name Record based cancellation rate forecasting.

Keywords: revenue management, cancellation rate forecasting, PNR data mining, two-class probability estimation, time-dependency

*Saïd Business School, University of Oxford, Park End Street, Oxford OX1 1HP, United Kingdom; email: dolores.romero-morales@sbs.ox.ac.uk

†Saïd Business School, University of Oxford, Park End Street, Oxford OX1 1HP, United Kingdom; email: jingbo.wang@sbs.ox.ac.uk

1 Introduction

Revenue Management aims at enhancing the revenues of a company by means of demand-management decisions such as dynamic pricing and capacity allocation (Talluri and van Ryzin 2004). The classical revenue management scenario is that in which a service provider sells a fixed number (the capacity) of perishable service products (air seats, hotel rooms, etc.) through a booking process which ends at a fixed deadline (the booking horizon). A revenue management system collects and stores booking records and market information and uses them to forecast the demand and learn customer behaviors. Then, during the booking horizon, it chooses optimal controls based on these input in order to maximize the revenue. The controls are in the form of dynamic pricing and capacity allocation, which are the prices and availabilities of various fares.

A revenue management system must take into account the possibility that a booking may be canceled, or that a booked customer may fail to show up at the time of service (no-show), which is a special case of cancellation that happens at the time of service. A way to take cancellations into account is to work with "net demand" (Rajopadhye et al. 2001) instead of demand. Here net demand is defined as the number of demand requests minus the number of cancellations. Alternatively, we may still work with demand but use a "virtual capacity" (Talluri and van Ryzin 2004), which is the actual capacity plus a buffer representing the number of bookings that are expected to be canceled. In both cases, accurate cancellation rates are crucial to the revenue management system.

Airlines and hotels routinely practice overbooking (Rothstein 1985), which is accepting more bookings than actual capacity based on estimated number of cancellations, or in other words accepting bookings up to the virtual capacity. Obviously, overbooking is only required after the number of on-going bookings, i.e. the bookings that have not been canceled yet among the existing ones, is close to the actual capacity. With the usual demand-management tools, such as dynamic pricing or capacity allocation control, the revenue management system will raise the price or close cheaper fares if demand is high, such that the capacity will never be sold out very early. As a result, overbooking is only needed in the very late part of the booking horizon. This has led people to think that

cancellation rate forecasting is only necessary close to the delivery of the service (Chatterjee 2001).

However, cancellation rate forecasting is not required merely to determine the levels of overbooking. Another equally important purpose is to contribute to the estimation of net demand (Chatterjee 2001, Rajopadhye et al. 2001). During the booking horizon, and as mentioned in Iliescu et al. (2008), the revenue management system constantly needs two parts of estimate of the net demand: (1) the net demand to come and (2) the net demand among on-going bookings. Forecasting the latter one is simply a cancellation problem. To give an idea of the difference between demand and net demand, from a major hotel chain in the UK, we have obtained a database of bookings with nearly 240,000 entries made between 2004 and 2006, where around 20% of them are eventually canceled. On average, the number of on-going bookings reaches 80% of the number of total show-ups at 3 days prior to the time of service; however about 11% of them are canceled at a later point of time, which means that net demand is just 71%. Similarly, the number of on-going bookings reaches 40% of the number of total show-ups at 10 days prior to the time of service, while about 16% of them are canceled at a later point of time such that the net demand is just 33%. As a result, errors in estimating these two cancellation rates could have a big impact on the estimation of total net demand. The rate of cancellation is even higher in airlines. A recent study points out that "cancellation rates of 30% or more are not uncommon today", see Iliescu et al. (2008).

Needless to say that the bias caused by errors in cancellation rate forecasting could affect revenue outcome because demand-management decisions, either in the form of pricing or allocation control, are always made based on estimated net demand. For example, in a revenue management system using dynamic pricing, overestimating cancellation rate will make the system underestimate net demand and therefore set the price too low in order to attract more demand. As far as the authors are aware, there are no direct citations on the revenue effect of an increase in the accuracy of cancellation rate forecasting. On a more general setting, Pölt (1998) estimates that a 20% reduction of forecast error (including demand, capacity and price forecasting) can translate into a 1% incremental increase in revenue generated from the revenue management system. Davenport (2006) cites the

example of the Marriott hotel chain which, through revenue management techniques, has been able to increase its actual revenue from 83% to 91% of the optimal revenue it could have obtained.

Because of the irreversible revenue losses at the time of service, research and practice to date have focused on no-show rate forecasting. Motivated by large Passenger Name Record (PNR[1]) databases, several recent attempts have been made to use PNR information to improve the accuracy of no-show rate forecasting with the help of Data Mining (Hastie et al. 2001) tools, and have shown promising results (Gorin et al. 2006, Hueglin and Vannotti 2001, Lawrence et al. 2003, Neuling et al. 2003). A recent study modeled cancellation as a survival process and used logistic regression to forecast the hazard rate (Iliescu et al. 2008). In this study, we discuss PNR data mining based forecasting for cancellations happening at any time during the booking horizon. Our main contribution is that we address the modeling of the behavior of customers in different stages of the booking horizon. We illustrate by means of a real-world dataset, how the set of relevant variables is very different in different stages of the booking horizon. Knowledge on these dynamics can help a revenue manager better understand the drivers of cancellation. On the other hand, such complex dynamics also require careful treatment in building forecasting models. We propose that multiple models can be built for different stages of the booking horizon to eliminate the time-dependency effect. In addition, and by noticing that the task of PNR cancellation rate forecasting can be seen as a two-class probability estimation problem (Hastie et al. 2001), we study the performance of state-of-the-art methods in this field. Our numerical results on the hotel dataset suggest that these methods are indeed promising.

The rest of the paper is structured as follows. In Section 2 we introduce some necessary notation and describe in detail the cancellation rate forecasting problem. In Section 3 we review existing models for cancellation rate forecasting and state-of-the-art methods in two-class probability estimation. In Section 4, we first introduce the real-world dataset that inspired this work. Using this dataset, we then illustrate how the set of relevant variables to describe cancellation behavior is very different in different stages of the booking horizon,

[1]Please note that "Passenger Name Record" simply means a booking record in the database of a computer reservation system. It is a standard term in many service industries including airline, hotel, rental car, railway, tourism and so on, although the word "passenger" might not be accurate in some cases.

and discuss how to model such time-dependency. In Section 5, we present our numerical results on applying the methods described in Section 3 to cancellation rate forecasting. Conclusions and some lines for future research are discussed in Section 6.

2 Problem definition

In this section, we first introduce some notation that will be used throughout the paper, and then formulate the cancellation rate forecasting problem.

2.1 The cancellation rate

Suppose there is a repeating service (for example a daily flight) managed by a revenue management system, where the time we deliver the service characterizes the service instance. Let us first look at one single future instance of the service scheduled at time T, for instance, the flight on next Monday. Let t denote the time-to-service, i.e., the remaining time before the service instance will take place. At t units of time from delivering this service instance, we have a collection of "on-going" bookings, say $\mathcal{O}(t)$. These bookings have been made before time point $T-t$ and have not been canceled before time point $T-t$. In order to estimate the net demand among $\mathcal{O}(t)$, we need to forecast how many bookings in $\mathcal{O}(t)$ will be eventually canceled. Let us use $Q(t)$ to denote the forecasted rate of cancellation among all on-going bookings at time $T-t$. The goal of this study is to develop models for forecasting $Q(t)$ for any time-to-service t.

2.2 The information at hand

Forecasting models for $Q(t)$ are built based on historical booking records. A booking record, commonly known as Passenger Name Record (PNR), of the service can be represented as $R = (T_s, t_b, X, \ell, t_c)$. The entry T_s is equal to the time of service (which as said before defines the service instance); t_b is the time of booking (in units of time before T_s); X is the vector of detailed information of the individual booking stored in the system such as

purpose (business or leisure) and channel. Finally, ℓ is the label of cancellation status where $\ell = 1$ indicates a booking which was eventually canceled and $\ell = 0$ is a show-up; while t_c is the time of cancellation (also in units of time before T_s), where a show-up has $t_c = -1$. A booking record for a service instance which has not been delivered yet may not be complete. In particular, for any booking in $\mathcal{O}(t)$, both ℓ and t_c will have empty entries.

A historical booking record of the service is characterized by T_s being a time point in the past ($T_s < T - t$), and therefore is a complete booking record, in which the entries ℓ and t_c are known. Let $\mathcal{H}$ denote the set of historical bookings at hand. When building a forecasting model for $Q(t)$, we can only consider the subset $\mathcal{H}(t) \subset \mathcal{H}$ defined as the set of historical bookings satisfying the following two conditions[2]

$$t_b > t \tag{1}$$

$$t_c \leq t. \tag{2}$$

Inequalities (1) and (2) tell us that the booking was an on-going booking for the service instance scheduled at T_s, at time point $T_s - t$.

2.3 Modeling and forecasting

Before we discuss forecasting models for $Q(t)$, we would like to stress the distinction between modeling and final forecast. First, forecasting models for $Q(t)$ are built well in advance and only updated from time to time, whereas the final output $Q(t)$ is calculated in real time during the booking horizon as time goes. Second, although at any point of time during the booking horizon $Q(t)$ is just a single number, the models must be able to calculate $Q(t)$ for any value of t, since t will change as time develops. In other words, we are building models that can produce a complete "cancellation curve".

[2]Here we assume t_b and t_c of any historical booking record are always known, which is true in most modern revenue management systems. In the case where either t_b or t_c is missing, which is commonly referred to as "censored data", $\mathcal{H}(t)$ cannot be constructed and methods that can handle censored data, such as survival analysis (see e.g. Iliescu et al. (2008)), should be used.

3 Related work

A number of models have been proposed for $Q(t)$ in the literature, which can be classified into two main categories: seasonal average models and PNR data mining models. Although the former has been very popular in practice, nowadays it is acknowledged that the latter leads to superior accuracy (Gorin et al. 2006, Hueglin and Vannotti 2001, Lawrence et al. 2003, Neuling et al. 2003). The PNR data mining approach models cancellation rate forecasting as a two-class probability estimation problem. In this section, we review existing seasonal and data mining models and state-of-the-art methods in two-class probability estimation that will be applied to cancellation rate forecasting in this study.

3.1 Existing forecasting models

3.1.1 Seasonal average models

Let us use S to denote the subvector of X regarding seasonal information, such as time of day, day of week, month of year and weather, on the service instance, which we recall is defined by the time of service T_s. Weighted averaging models based on S (also called "seasonal average") were very popular in the early days of revenue management (Lawrence et al. 2003), partly because S was often the only information available in the revenue management systems at that time. Exponential smoothing has been widely used to derive the weights. In the simplest case, weights only depend on the time of service T_s, and are set as a geometric progression increasing from earliest to latest service instance.

3.1.2 PNR data mining models

As revenue management systems develop, R starts to include information regarding each individual booking, i.e. proper PNR information, such as purpose (business or leisure), channel and many others. As a result, there has been an increasing tendency to move towards PNR data mining (Hastie et al. 2001) based models. Different from seasonal average models, which obtain $Q(t)$ directly, PNR data mining models are two-stage in nature. A PNR data mining model forecasts the probability of cancellation of each booking

in $\mathcal{O}(t)$ and then calculates $Q(t)$ as the average of these probabilities, i.e.,

$$Q(t) = \frac{1}{\text{card}(\mathcal{O}(t))} \sum_{R \in \mathcal{O}(t)} Q(t, R),$$

where $Q(t, R)$ denotes the forecasted probability of cancellation of an on-going booking R and $\text{card}(\mathcal{O}(t))$ is the cardinality of the set $\mathcal{O}(t)$. In this case, forecasting $Q(t, R)$ can be seen as a two-class probability estimation problem, where the two classes are "canceled" and "not canceled".

A natural alternative to the two-class probability estimation approach is to classify each booking into the two classes "canceled" and "not canceled", and then calculate $Q(t)$ as the number of bookings classified as cancellation divided by the total number of on-going bookings. However, if classification is directly used to calculate $Q(t)$, the accuracy is very poor. Firstly, this is because the associated classification problem is very difficult, which is also intuitive: it is hard to imagine that one can predict whether a booking will be canceled or not with high accuracy simply by looking at PNR information. Secondly, using classification directly means there are only two possible values for $Q(t, R)$: 0% and 100%, which are clearly too rough. Fortunately, in the revenue management context, the classification or even probability of cancellation of an individual booking is not important. What the system needs is just $Q(t)$, the aggregated cancellation rate among all on-going bookings, and for this purpose, the two-class probability estimation approach works well.

Several PNR data mining models have been proposed in the literature, however most of them focus on the no-show case ($t = 0$). Freisleben and Gleichmann (1993) and Wu and Lin (1999) both trained neural networks, but their PNR data contained just seasonal information. Gorin et al. (2006) used weighted averaging with weight determined by three PNR variables using ad-hoc rules. Hueglin and Vannotti (2001) used a simple decision tree (Ripley 2008) with merely 15 nodes, as well as logistic regression to build predictive models; Lawrence et al. (2003) used ProbE (Apte et al. 2002), a hybrid of decision tree and Naive Bayes, C4.5 decision tree (Quinlan 1993) and the Adjusted Probability Model (Hong et al. 2002), an extension of Naive Bayes. Neuling et al. (2003) also used C4.5 decision tree. They all claimed significant improvement (5% to 15% reduced error) on accuracy over

seasonal average models.

PNR data mining models make use of the individual booking information of $R \in \mathcal{O}(t)$, but this collection of bookings is only available at time $T-t$. However, forecasts from seasonal average models can be provided before time $T-t$, and for any value of t, as long as the seasonal information on the service instance S is known. When the focus is not only on cancellation rates among on-going bookings but also among bookings to come, PNR and seasonal average models have a complementary role. In this case, weighted average of PNR and seasonal average forecasts can be used, as suggested in Lawrence et al. (2003) and Neuling et al. (2003). The idea is: PNR forecast is used for on-going bookings, seasonal average forecast is used for bookings to come, whose number is roughly estimated (usually capacity minus card($\mathcal{O}(t)$)), and then the weighted average of the two forecasts is the overall forecast.

3.2 Two-class probability estimation methods

The literature on multi-class probability estimation (Caruana and Niculescu-Mizil 2006) is rich. In this study, state-of-the-art methods from two popular categories, tree based methods and kernel based methods, will be applied to cancellation rate forecasting. Most of these methods have roots in classification, which is one of the main tasks in data mining.

Provost and Domingos (2003) showed that traditional decision trees aiming at improving classification accuracy usually do not give good class probability estimates, mostly because the pruning and collapsing procedures lead to leaf nodes with large number of observations and then make the probability estimates too rough, i.e., all observations on the same leaf node share the same probability estimate. This coincides with the poor performance of C4.5 decision tree when forecasting no-show rate (Lawrence et al. 2003). Provost and Domingos (2003) then developed the first probability estimation tree method: C4.4 tree (C4.4), which divides tree nodes based on information gain criteria and uses no pruning or collapsing. C4.4 was confirmed to outperform C4.5 in terms of probability estimation in Fierens et al. (2005). Another recent probability estimation tree algorithm is the Minimum Squared Expected Error tree (MSEE) proposed in Nielsen (2004). The splitting criterion

proposed here is to minimize the squared expected error of class probability estimates, which means the method is specifically designed for probability estimation but not classification. Besides, it introduces a top down smoothing scheme using information at upper nodes to adjust class probability estimates at leaves, which also appears to be quite helpful for probability estimation.

While C4.4 and MSEE build a single tree, it has long been observed that the performance of a single tree can usually be improved by growing an ensemble of trees (Ho 1998). Random Forest (RF) (Breiman 2001) combines several successful techniques in building tree ensembles such as bagging (Breiman 1996) and random split selection (Dietterich 2000) to construct a collection of decision trees. Let Ω denote the population under consideration. Random forest is a combination of tree classifiers in which each tree is built using a different bootstrap sample of Ω with size $|\Omega|$; only a small randomly chosen set of explanatory variables are considered for splitting each node; and each tree is fully grown and not pruned. The class probability can be obtained by letting these tree classifiers vote for the class membership of the object under consideration and then counting their votes. Recent studies suggest that random forest is among the most competitive probability estimation method (Caruana and Niculescu-Mizil 2004, 2006).

Recently, kernel based methods, notably Support Vector Machine (SVM) (Platt 2000) and Kernel Logistic Regression (KLR) (Wahba 1999, Zhu and Hastie 2005), have been developed for class probability estimation and obtained promising results. These methods make use of kernel functions which map input data points to a higher dimensional space. As a result, a linear method in the new space becomes non-linear in the original space and is therefore able to model non-linear relationships between dependent and independent variables.

The best known element of kernel methods is probably SVM, which was originally proposed as a classification method (Cristianini and Shawe-Taylor 2000). For two classes, it tries to find a separating hyperplane for the two classes in the higher dimensional space, i.e. where each object is correctly classified, which maximizes the margin. To account for non-separable data and to avoid overfitting, the soft approach was proposed, in which some objects may be misclassified. Consider a training set I of vectors (x^i, y^i), where $x^i \in \mathbb{R}^N$ is the vector of explanatory variables and $y^i \in \{-1, +1\}$ is the class label. Let

ϕ be the mapping function (the kernel). The hyperplane is found by solving the following optimization problem:

$$\text{minimize } \frac{1}{2}\omega^\top\omega + C\sum_{i\in I}\xi_i \tag{3}$$

subject to

$$\begin{aligned} y^i(\omega^\top\phi(x^i)+b) &\geq 1-\xi_i \quad i\in I & (4)\\ \xi_i &\geq 0 \qquad i\in I & (5)\\ \omega, b &\qquad \text{free}, & (6)\end{aligned}$$

where $\xi = (\xi_i)$ is the vector of slack variables and C is a penalty parameter which trades off misclassification and margin. The dual of this problem can be written as a convex quadratic problem, which can be solved efficiently by standard techniques, see Cristianini and Shawe-Taylor (2000). Once the optimal solution (ω^*, b^*, ξ^*) is found, new objects are classified using the decision function $f(x) = \text{sign}((\omega^*)^\top\phi(x) + b^*)$. Platt (2000) and Wu et al. (2004) use logistic regression to link a data point's distance to the hyperplane defined by ω^* and b^* with its class probability and thus making it possible to use SVM for probability estimation.

A more direct approach than the above logistic regression based calibration is to replace Formula (4) by:

$$\log(1+e^{-y^i(\omega^\top\phi(x^i)+b)}) \geq 1-\xi_i \quad i\in I. \tag{7}$$

Then by applying the duality principle, the optimization problem becomes maximizing the log-likelihood function associated with the probabilistic model:

$$\text{Prob}(y|x) = \frac{1}{1+e^{-y^i(\omega^\top\phi(x^i)+b)}}. \tag{8}$$

This approach is called Kernel Logistic Regression (Wahba 1999, Zhu and Hastie 2005), in which a data point's distance to the hyperplane defined by ω^* and b^* is used to approximate its log-odds of class probability.

4 Cancellation rate forecasting and time dependency

Extending PNR data mining based forecasting from no-show to cancellation at any time is not a straightforward task. As discussed in Section 2.3, the extended model should be able to produce a complete "cancellation curve", i.e. calculating $Q(t, R)$ for any time-to-service t, and not just for $t = 0$ (the no-show case). In this section, we will illustrate how the cancellation behavior of customers is dependent on the time-to-service t based on our analysis of a real-world PNR dataset. Knowledge on these dynamics will not only help in building more accurate forecasting models, but also in better understanding the drivers of cancellation.

4.1 The real-world PNR dataset

We have collected the complete reservation record of a hotel in a major hotel chain in the UK for 974 days of services between 2004 and 2006, which contains nearly 240,000 booking records. Throughout the rest of the paper, and without loss of generality, time will be measured in days. We calculate the actual cancellation rate at $t = 0, 1, \ldots, 50$ for each of the 974 service instances, and Figure 1 shows the mean value and standard deviation of these cancellation rates across the 974 service instances. The other curve in Figure 1 shows the growth of the number of on-going bookings relative to the total number of show-ups in the same period. As we can see, the standard deviations of cancellation rates are very big compared with their mean values – which shows how random the process of cancellation is and therefore the challenge we face.

We may recall that a PNR entry is given by $R = (T_s, t_b, X, \ell, t_c)$, see Section 2.2. In our hotel dataset X contains a total of 13 variables, which together with t_b (`timebooking`), are used as explanatory variables to forecast cancellation rates. Among the 14 variables, 11 are nominal, 2 (`length` and `timebooking`) take non-negative integer values and 1 (`price`) takes non-negative real number values. Simple preprocessing has been done. Nominal variables `company`, `ratecode`, `market`, `agent` and `system` originally each had hundreds of possible values but many of them were actually very rare, so we grouped all infrequent values (with less than 500 observations in the whole dataset, or equivalently with frequency

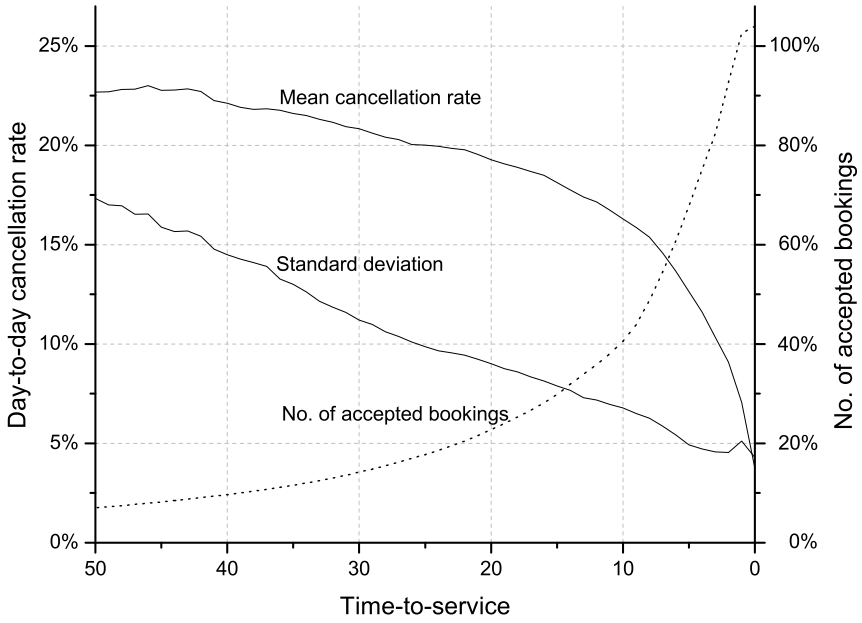

Figure 1: Cancellation rates across the booking horizon

less than 0.21%) together. We found that using the numerical values of the two integer variables `length` and `timebooking` directly often implies very strong assumptions and will seriously distort the probability estimates given by certain methods. For example, when `timebooking` is used as a numerical variable in a logistic regression model, it implies that `timebooking` has linear relationship with the log-odds of cancellation probability, which is too rigid and has caused logistic regression to give substantially suboptimal results in our comparison. Therefore, we decided to use the two integer variables as nominal variables and the same grouping for infrequent values (with less than 500 observations in the whole dataset) was applied. Details of the 14 explanatory variables (after preprocessing) can be found in Table 1.

4.2 Time-dependency of the relative importance of variables

In addition to building accurate PNR data mining based forecasting models, another important function of the explanatory variables is to help the revenue manager understand customer behavior and in particular the drivers of cancellation. Therefore, it is necessary to examine the relevance of the explanatory variables to cancellation rate forecasting, and such examination should be carried out for different stages of the booking horizon. In this

Name	Type	Possible Values	Description
Agent	Nominal	44 categories (after grouping)	Which agent is the booking from?
Channel	Nominal	5 categories	Channel used to make the booking
Company	Nominal	34 categories (after grouping)	Company of the customer
Day	Nominal	7 categories	Day of week of the service
Group	Nominal	2 categories	Is the booking part of a group?
Length	Integer	9 categories (after grouping)	No. of nights of stay
Market	Nominal	44 categories (after grouping)	Market sector the booking is from
Month	Nominal	12 categories	Month of year of the service
Price	Continuous	Non-negative real numbers	Price of the booking
Ratecode	Nominal	59 categories (after grouping)	Rate code of the booking
Refundable	Nominal	2 categories	Is the booking refundable?
Roomtype	Nominal	21 categories	Room type
System	Nominal	25 categories (after grouping)	Reservation system used
Timebooking	Integer	49 categories (after grouping)	t_b (measured in number of days)

Table 1: Explanatory variables for building PNR based models

section we illustrate that the relative importance of these variables is different at different stages of the booking horizon (different time-to-service values t). In other words, the drivers of cancellation change from one stage of the booking horizon to another. To illustrate this time-dependency of variable importance, we use the fact that the task of PNR cancellation rate forecasting can be seen as a two-class probability estimation problem. In this context, we have tried two widely used methods for measuring variable importance.

The *information gain ratio* (Quinlan 1993) measures a nominal explanatory variable's relevance to a multi-class probability estimation problem. For two classes, the information gain ratio is defined as follows. Suppose Ω is the population under consideration, which is divided into the positive class (canceled) and the negative class (not canceled). Let p be the proportion of objects in the positive class within Ω. Consider a nominal explanatory variable x, with m possible categories. Let Ω_i be the subset of Ω defined by the i-th category of x, and p_i the proportion of objects in the positive class within Ω_i $(i = 1, \ldots, m)$. Then, the information gain ratio of the explanatory variable x is defined as

$$GainRatio(x) = \frac{Info(p) - \sum_{i=1}^{m} \frac{|\Omega_i|}{|\Omega|} Info(p_i)}{-\sum_{i=1}^{m} \frac{|\Omega_i|}{|\Omega|} \log_2 (\frac{|\Omega_i|}{|\Omega|})},$$

where

$$Info(p) = -p \log_2(p) - (1-p) \log_2(1-p).$$

Generally, the larger the information gain ratio, the more relevant the explanatory variable

is to the probability estimation problem.

To complement the information gain ratio, which is a univariate metric, we propose to use the built-in variable importance assessment function of Random Forest (Breiman 2001), which is a multivariate metric. To assess the importance of a variable x, it works as follows. For every tree in the forest, since it is constructed using a different bootstrap sample of Ω with size $|\Omega|$, about one-third of the objects in Ω are not used in its construction and are called its "out-of-bag" data. Use the tree to classify its out-of-bag data and count the percentage of correct classification. Now randomly permute the values of variable x in the out-of-bag data and use the tree to classify them. Subtract the percentage of correct classification in the variable-x-permuted out-of-bag data from the percentage of correct classification in the original out-of-bag data. The average of this number over all trees in the forest is the importance score for variable x. By assuming this number is independently and normally distributed across trees, standard error and significance level of the score can be computed. Intuitively, this importance score has the interpretation of "the drop of classification accuracy caused by disabling variable x from the random forest", which is quite similar to ideas in the Wrapper approach (Kohavi and John 1997) and Recursive Feature Elimination method (Guyon et al. 2002) in the feature selection (Liu and Motoda 2007) literature, i.e., using the change in objective function such as classification accuracy or log-likelihood when one variable is removed as a measure of variable importance.

Table 4 shows the information gain ratios and random forest importance scores of the 14 explanatory variables at seven time points $t = 0$ and $t = 2^i$, $i = 0, 1, \ldots, 5$ (with $\Omega = \mathcal{H}(t)$). Please note that the information gain ratio is defined for nominal variables only and therefore cannot be calculated for continuous variable `price`. All random forest importance scores are statistically significant at 1% level. It is clear that the summation of the ratios/scores of the 14 variables (the last row of Table 4) are quite different at different time points, especially for information gain ratio. Therefore, it is better to use relative ratio/score rather than absolute ones when examining the importance of variables at different time points. The relative ratio/score of a variable at a certain time point can be calculated by dividing its absolute ratio/score by the total ratio/score of all the variables at that time point.

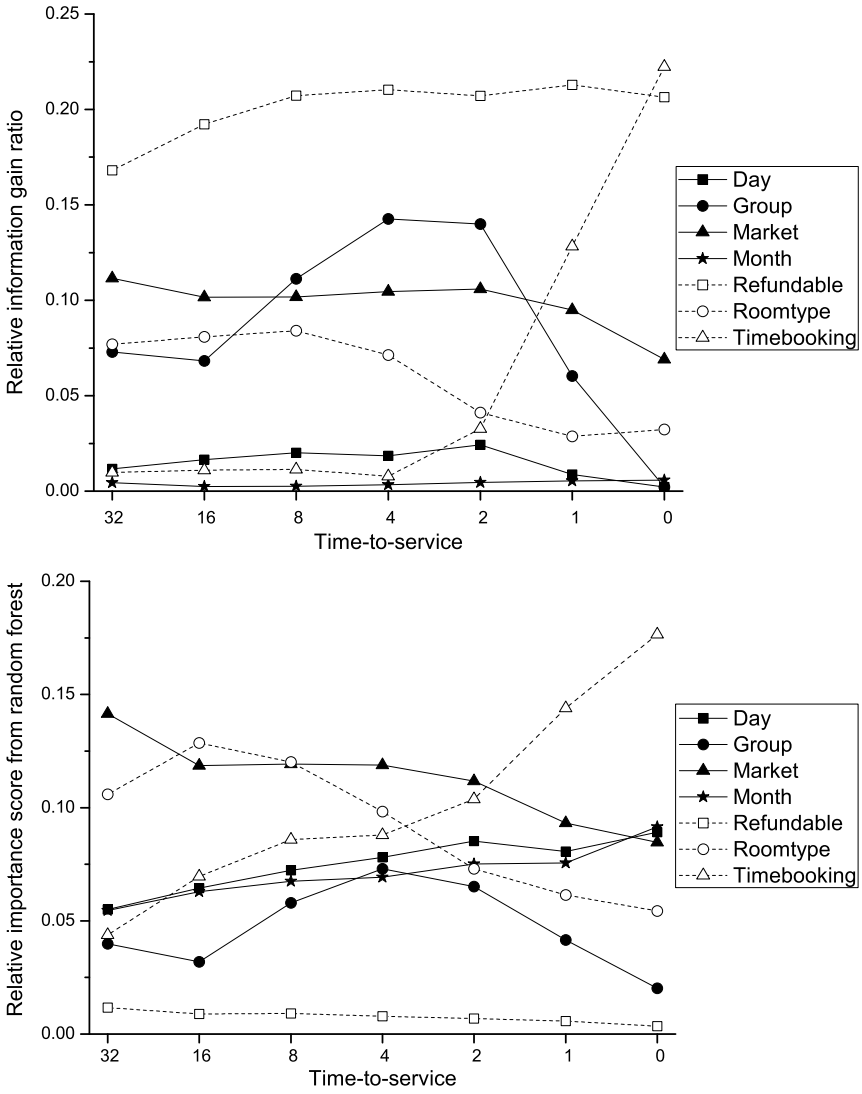

Figure 2: Relative information gain ratio and random forest importance score

Figure 2 shows the relative ratios/scores of seven variables presenting the most interesting patterns. The first observation is that the two different metrics agree with each other to some extent but could also give very different results. For example, variable `refundable` is consistently high on information gain ratio but low on random forest importance score. This suggests that variable `refundable` is very informative individually, however most of the information it brings has already been explained by other variables. Variables `month` and `day` are just the opposite, being less important individually and more important when working together with other variables.

Despite the discrepancies between the two metrics, some interesting patterns are identified by both of them. The relative importance of variable `timebooking` increases sharply

when it gets close to the time of service and is in fact the most important variable at the time of service ($t = 0$) under both metrics. The relative importance of variable `group` is higher at $t = 2$ and $t = 4$ than in earlier and later stages. In particular, both metrics suggest that `group` is among the least relevant variables when $t = 0$. Variables `market` and `roomtype` are generally less important in later stages of the booking horizon than in earlier stages. These observations suggest that the relative importance of variables changes significantly in different stages of the booking horizon, i.e., the relative importance of variables is highly time-dependent. This time-dependency of variable importance has two important implications. First, it becomes compulsory to take the time-dependency of variable importance into account when building forecasting models. Second, it offers revenue managers a chance to better understand the cancellation process by raising questions such as why some variables are more important in certain stages but not the others like in the case of `group`. With the help of the improved understanding, not only forecasting accuracy will be improved, better cancellation policies will also become possible.

4.3 Modeling the time-dependency

When dealing not only with no-show but cancellations at any time, the most crucial issue is how to model the variable time-to-service t properly. The two options seem to be quite natural. First, the forecasting for different values of t can be carried out separately, i.e., a data mining model has to be trained for each t. Alternatively, a single predictive model is built to forecast the probability of cancellation for any value of t. At a first glance, the second approach seems more attractive as it uses the same model for any value of t. However, as the information gain ratios and random forest importance scores have suggested, an explanatory variable's impact on cancellation probability is highly dependent on t. If the single-model approach were to be used, some methods, such as logistic regression, would need to model the interactions between each explanatory variable and t. For example, instead of using the variable `group`, three variables "group with $t < 2$", "group with $2 \leq t \leq 4$" and "group with $t > 4$" can be used, each taking positive value only if the booking is being considered during the corresponding range of time-to-service and also part of a group. Summarizing the discussions above, we believe the multiple-model approach

is a plausible choice. Although multiple models will need to be built, it takes the time-dependency into account in a simple an effective way by isolating different stages of the booking horizon from each other. For this simplicity, it will be used in all our experiments in the next section.

5 Experimental results on forecasting accuracy

In this section we use the hotel dataset introduced in Section 4 to test the methods discussed in Section 3.2. First, we discuss the choice of time points at which predictive models are evaluated in Section 5.1. Implementation and tuning details of the methods being tested are given in Section 5.2. We introduce the performance metric for evaluating forecasting methods in Section 5.3. Finally the numerical results are presented in Section 5.4.

5.1 Discrete choice for time-to-service

In Section 4.2, we used a discrete choice for time-to-service t when calculating the information gain ratios and random forest importance scores. In this section we are interested in the performance of different methods in forecasting the cancellation curve. Below we argue that the discrete choice $t = 0$ and $t = 2^i$, $i = 0, 1, \ldots, 5$, in Section 4.2 is still adequate. Clearly, the further from the time of service, the more dispersed the time-to-service values need to be chosen. The reason for this is that, when t_1 is close to t_2 and both are far enough from the time of service, the sets of historical bookings $\mathcal{H}(t_1)$ and $\mathcal{H}(t_2)$ (respectively on-going bookings $\mathcal{O}(t_1)$ and $\mathcal{O}(t_2)$) are very similar to each other and so are the corresponding predictive models and their accuracy. The seven time-to-service values we have chosen roughly mark the time when the number of on-going bookings reaches 105% (show-ups plus no-shows), 100%, 95%, 75%, 50%, 25% and 10% of the number of show-ups, see Figure 1. They are relatively far from each other in terms of number of on-going bookings and represent different stages of the booking horizon from early to very late. The performance at the seven time points will be a good indicator of a predictive model's ability to capture the dynamics of the cancellation process.

5.2 Forecasting methods and the tuning

In addition to the methods discussed in Section 3.2, we also include logistic regression (LR), the classical two-class probability estimation method, and simple average (AVG), which is simply the average cancellation rate among all training data, and seasonal average (SAVG) into our comparison for benchmarking purpose. For the sake of completeness, a list of abbreviation and short description of these methods can be found in Table 2.

Abbreviation	**Description**
AVG	Average cancellation rate among all training data
SAVG	Seasonally averaged cancellation rate among all training data
LR	Logistic Regression
C4.4	C4.4 Probability Estimation Tree (Provost and Domingos 2003)
MSEE	Minimum Squared Expected Error Tree (Nielsen 2004)
RF	Random Forest (Breiman 2001)
SVM	Support Vector Machine (Cristianini and Shawe-Taylor 2000)
KLR	Kernel Logistic Regression (Wahba 1999)

Table 2: List of methods being tested in Table 3

The SVM implementation we use is the `SVMlight` software package (Joachims 1999). The KLR implementation we use is the `myKLR` software package based on the algorithm of Keerthi et al. (2005). We have tried SVM and KLR with linear, polynomial, sigmoid and Radial Basis Function (RBF) kernels. For simplicity, we will only show results with RBF kernel in our comparison, as RBF kernel dominates the other three in most of our experiments. For SVM and KLR, parameter selection is necessary. For the penalty parameter C, we have tested three values, namely 0.1, 1 and 10; for parameter γ of the RBF kernel, we have tested three values, namely 0.01, 0.1 and 1. In all the experiments, we use two-fold crossvalidation (Kohavi 1995) to find the best parameter. Here we only perform a two-fold crossvalidation because SVM and KLR training is very time consuming. We find that most of the times the parameter selected by this procedure works well. The RF implementation we use is the original Fortran program (available at `http://www.stat.berkeley.edu/~breiman/RandomForests/`) with default settings and the size of the forest is 500. As suggested in Breiman (2001), we have chosen the parameter controlling the number of variables considered for splitting each node as the one with the smallest out-of-bag error estimate in the training dataset.

For SAVG, we set the weight of booking records for training to increase exponentially with

a smoothing factor of 0.002 from the oldest to the latest booking records. In addition, SAVG also gives double weight to booking records for a service on the same weekday or in the same month (could be doubled twice) as the future service to be forecasted for.

5.3 Performance metric

Despite the potential asymmetrical costs associated with over and under-estimating $Q(t)$, cost-based evaluation for accuracy in forecasting $Q(t)$ is unnecessary under the revenue management context. When $t > 0$, both over and under-estimating $Q(t)$ will cause the bias in estimation of net demand, which then might cause wrong demand management decisions, and finally affect the revenue potential that could have been captured. In this case, the two types of errors are equally damaging, and there is no need for discriminating among them by means of different costs. In the case of no-shows ($t = 0$), there are indeed different costs associated with the two types of errors, i.e., the cost of denied service is higher than the cost of unused capacity. However, in revenue management systems, the imbalanced costs are modeled in the modules to which the cancellation rates are fed. For instance, when setting the overbooking levels the imbalances are taken into account and a conservative approach is normally used to avoid too many displaced customers.

For a given testing service instance and time-of-service t, use $\mathcal{O}(t)$ to denote the set of on-going bookings. We will measure a model's performance in forecasting $Q(t)$ by its *absolute error* defined as:

$$\text{err}^{\text{abs}}(t) = |\text{card}(\mathcal{O}(t)) \times Q(t) - \sum_{R \in \mathcal{O}(t)} \ell| \tag{9}$$

where $\text{card}(\mathcal{O}(t)) \times Q(t)$ and $\sum_{R \in \mathcal{O}(t)} \ell$ are the forecasted and actual number of cancellations in on-going bookings $\mathcal{O}(t)$ respectively. This is also a more appropriate measure than the absolute difference between $Q(t)$ and actual rate of cancellation $\sum_{R \in \mathcal{O}(t)} \ell / \text{card}(\mathcal{O}(t))$, because it gives more weight to service instances with a large number of bookings.

5.4 The numerical results

Single-run training and testing using all the data at once seem to be a common practice of evaluating performance of predictive models in previous studies on no-show rate forecasting (Hueglin and Vannotti 2001, Lawrence et al. 2003, Neuling et al. 2003). However, any statistical experiment with only one run may be unreliable. Therefore, we have designed a randomized experiment. Instead of adopting the popular practice in machine learning of randomly splitting the whole dataset into training and testing sets (Devijver and Kittler 1982), it is more appropriate to use the older part of the data for training and the newer part for testing, since PNR data are naturally ordered by time. The randomized experiment consists of 20 runs. In each run, we randomly choose one fourth of the data of the first 700 service instances (with replacement) to train predictive models and randomly choose 25 service instances out of the last 274 (with replacement) to test the accuracy.

In each of the 20 runs and for each time point t, we calculate the total number of on-going bookings in the 25 service instances n_b and total number of cancellations among these bookings n_c. For each forecasting method, we report the total absolute errors (TAE) over the 25 service instances, where the absolute error for each service instance is given by (9). Table 3 shows the mean and standard deviation of n_b, n_c and the TAE over the 20 runs at the seven time points. To summarize the goodness of a method at the seven time points, in each of the 20 runs, we also calculate SUM, which is the summation of the TAEs at the seven time points. The mean and standard deviation of SUM in the 20 runs is also shown in Table 3. Finally, we divide the mean SUM of each method by the mean SUM of AVG, which reflects its relative goodness against AVG, and the results are shown in the last row of Table 3.

Several conclusions can be drawn from Table 3. First, seasonal average (SAVG) is actually slightly worse than simple average (AVG). This indicates that the connection between seasonal information and cancellation behavior is not very strong. At $t = 0$, the mean TAEs of SAVG and AVG are actually larger than n_c, i.e., the average absolute error is larger than the forecast itself. This is caused by high day-to-day variability of cancellation rate at $t = 0$. In Figure 1, the standard deviation is almost as large as the mean cancellation rate. Here, simply using the forecast $Q(t) = 0$ would beat SAVG and AVG.

t		n_b	n_c	Total Absolute Errors (TAE) AVG	SAVG	LR	C4.4	MSEE	RF	SVM	KLR
32	Mean	613.7	110.7	55.4	55.1	48.8	49.1	51.0	43.3	49.8	47.3
	Std Dev	68.4	15.7	6.0	5.6	7.6	5.8	5.7	6.1	6.6	5.7
16	Mean	1353.3	222.7	87.7	88.2	83.1	76.9	77.1	72.6	78.8	74.9
	Std Dev	78.3	24.8	9.5	11.1	10.4	9.9	8.6	11.1	8.2	9.1
8	Mean	2356.1	328.9	118.4	118.4	100.6	100.7	98.7	99.0	98.4	100.8
	Std Dev	137.1	24.5	18.3	19.0	17.7	18.1	15.9	14.8	17.0	14.7
4	Mean	3669.9	377.9	127.6	127.1	116.9	116.2	114.2	110.9	115.0	115.3
	Std Dev	270.8	35.9	19.5	17.2	16.9	21.1	18.2	17.6	18.0	17.7
2	Mean	4685.9	397.8	149.0	149.9	129.6	124.6	119.5	129.8	117.4	134.9
	Std Dev	306.2	47.5	29.6	30.8	26.8	28.9	25.5	30.3	29.5	29.5
1	Mean	5214.6	363.7	218.1	219.1	163.2	164.1	158.5	160.6	156.1	158.0
	Std Dev	473.0	64.3	50.9	52.1	34.0	34.2	34.3	37.3	39.7	32.3
0	Mean	5169.5	166.2	178.1	179.8	112.9	107.9	106.9	107.1	104.6	102.6
	Std Dev	294.4	46.7	27.4	24.9	15.4	20.6	14.4	23.1	19.8	15.4
SUM	Mean	23062.8	1967.8	934.3	937.6	755.2	739.5	725.7	723.3	719.9	733.9
	Std Dev	729.3	112.0	69.4	68.1	50.8	39.4	42.6	58.6	53.1	44.9
SUM mean relative to that of AVG				100.0	100.4	80.8	79.2	77.7	77.4	77.1	78.5

Table 3: Forecasting performance of the eight methods in 20 randomized runs

On the other hand, PNR data mining models are much more accurate than SAVG and AVG, on average reducing the error by more than 20% at the seven time points. This is in accordance with existing literature, and demonstrates that PNR data mining based cancellation rate forecasting is indeed promising. Among the six data mining based methods, SVM gives the smallest mean of SUM but the margin from the other methods is small. The two single tree based methods C4.4 and MSEE are the best in terms of standard deviation of SUM, although their mean of SUM is not the best. On the other hand, LR is the least accurate in terms of mean of SUM, reducing the error of AVG by 19.2% while the other five methods reduce 22% (2.8% more) on average.

In order to enable an easy comparison of the performance of the six data mining based methods, we plot the relative mean TAE of each method against that of SVM at the seven time points in Figure 3. An interesting observation from Figure 3 is that the best performing method (in terms of mean TAE) is different at different time points. For example, KLR is the best at $t = 0$, SVM is the best at $t = 1$, 2 and 8, and RF is the best at $t = 4$, 16 and 32. Besides, we have also observed that different parameters are often selected for SVM, KLR and RF at the seven time points. These behaviors are possibly caused by the time-dependency of the cancellation rate forecasting problem, i.e., the two-class probability estimation problems associated with different stages of the booking horizon have different

characteristics and therefore lead to different parameters and relative performance.

6 Conclusions and future direction

Forecasting for cancellations happening at any time during the booking horizon is a complex problem. As shown in Section 4.2, in different stages of the booking horizon (i.e. for different values of the variable time-to-service t), the set of factors influencing the probability that a booking is canceled is very different. Knowledge on these dynamics can help a revenue manager better understand the drivers of cancellation. On the other hand, such complex dynamics also require forecasting models to take the time-dependency into account. We propose that multiple models can be built for different stages of the booking horizon (for different values of t). This idea of building multiple models could also be applied to other time-dependent forecasting problems in demography, econometrics and transportation (Iliescu et al. 2008). Alternatively, a single model taking into account the time-dependency could also be built, which would probably give even more insights on the dynamics of the cancellation process. Although the complexity of designing such models would make it out of the scope of this study, we believe it is a promising direction for future

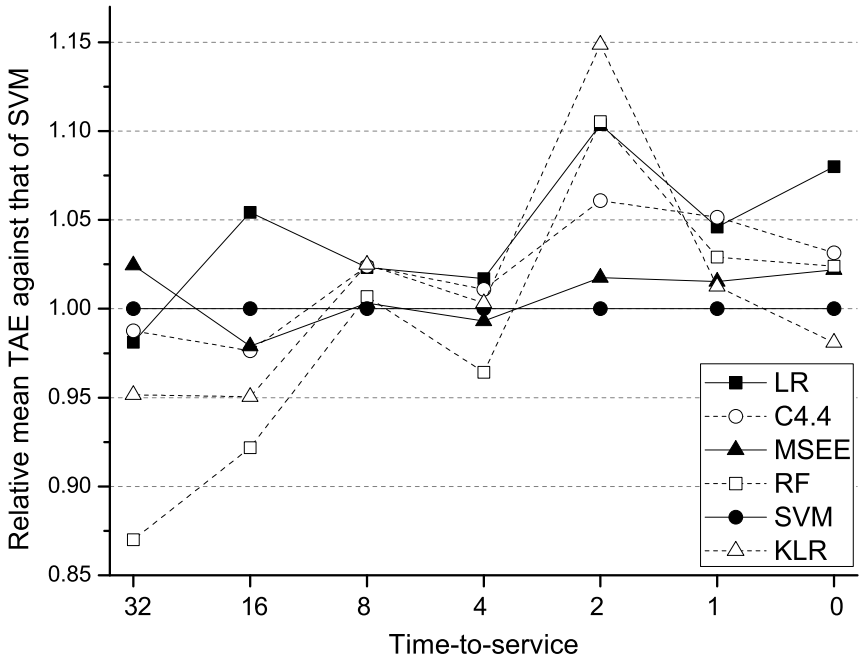

Figure 3: Relative performance of the six data mining based methods

research.

In this study, we apply state-of-the-art data mining based two-class probability estimation methods to cancellation rate forecasting and the results are promising. In addition to obtaining accurate cancellation rate forecasts, understanding how PNR variables influence the probability of cancellation, or in other words the drivers of cancellation, is an important concern of revenue managers. However this is a challenging task, since the time-dependency of the relative importance of variables means that the analysis of drivers of cancellation must be coordinated across multiple time points along the booking horizon. We will leave the PNR data mining based analysis of drivers of cancellation as a future research topic.

References

C. Apte, R. Natarajan, E. P. D. Pednault, and F. Tipu. A probabilistic estimation framework for predictive modeling analytics. *IBM Systems Journal*, 41(3):438–448, 2002.

L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

L. Breiman. Bagging predictors. *Machine Learning*, 26(2):123–140, 1996.

R. Caruana and A. Niculescu-Mizil. Data mining in metric space: an empirical analysis of supervised learning performance criteria. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 69–78, 2004.

R. Caruana and A. Niculescu-Mizil. An empirical comparison of supervised learning algorithms. In *Proceedings of the Twenty-Third International Conference on Machine Learning*, pages 161–168, 2006.

H. Chatterjee. Forecasting for cancellations. Presentation at *AGIFORS 2001 Reservations and Yield Management Conference*, Bangkok, Thailand, 2001.

N. Cristianini and J. Shawe-Taylor. *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods*. Cambridge University Press, Cambridge, UK, 2000.

T. H. Davenport. Competing on analytics. *Harvard Business Review*, 84(1):98–107, 2006.

P. A. Devijver and J. Kittler. *Pattern Recognition: A Statistical Approach*. Prentice-Hall, London, 1982.

T. G. Dietterich. An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization. *Machine Learning*, 40(2):139–157, 2000.

D. Fierens, J. Ramon, H. Blockeel, and M. Bruynooghe. A comparison of approaches for learning probability trees. In *Proceedings of the Sixteenth European Conference on Machine Learning*, pages 556–563, 2005.

B. Freisleben and G. Gleichmann. Controlling airline seat allocations with neural networks. In *Proceedings of the Twenty-Sixth Hawaii International Conference on System Sciences*, pages 635–642, 1993.

T. Gorin, W. G. Brunger, and M. M. White. No-show forecasting: A blended cost-based, PNR-adjusted approach. *Journal of Revenue and Pricing Management*, 5(3):188–206, 2006.

I. Guyon, J. Weston, S. Barnhill, and V. Vapnik. Gene selection for cancer classification using support vector machines. *Machine Learning*, 46(1–3):389–422, 2002.

T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference and Prediction*. Springer-Verlag, New York, NY, 2001.

T. K. Ho. The random subspace method for constructing decision forests. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 8(20):832–844, 1998.

S. J. Hong, J. R. M. Hosking, and R. Natarajan. Ensemble modeling through multiplicative adjustment of class probability. In *Proceedings of the Second IEEE International Conference on Data Mining*, pages 621–624, 2002.

C. Hueglin and F. Vannotti. 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*, pages 438–442, 2001.

D. C. Iliescu, L. A. Garrow, and R. A. Parker. A hazard model of US airline passengers' refund and exchange behavior. *Transportation Research Part B*, 42(3):229–242, 2008.

T. Joachims. Making large–scale SVM learning practical. In B. Schölkopf, C. J. C. Burges, and A. J. Smola, editors, *Advances in Kernel Methods — Support Vector Learning*, pages 169–184, Cambridge, MA, 1999. MIT Press.

S. S. Keerthi, K. Duan, S. K. Shevade, and A. Poo. A fast dual algorithm for kernel logistic regression. *Machine Learning*, 61(1–3):151–165, 2005.

R. Kohavi. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence*, pages 1137–1145, 1995.

R. Kohavi and G. John. Wrappers for feature subset selection. *Artificial Intelligence*, 97(1–2): 273–324, 1997.

R. D. Lawrence, S. J. Hong, and J. Cherrier. Passenger-based predictive modeling of airline no-show rates. In *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 397–406, 2003.

H. Liu and H. Motoda. *Computational Methods of Feature Selection*. Chapman & Hall/CRC, New York, NY, 2007.

R. Neuling, S. Riedel, and K. U. Kalka. New approaches to origin and destination and no-show forecasting: Excavating the passenger name records treasure. *Journal of Revenue and Pricing Management*, 3(1):62–72, 2003.

R. Nielsen. MOB-ESP and other improvements in probability estimation. In *Proceedings of the Twentieth Conference in Uncertainty in Artificial Intelligence*, pages 418–425, 2004.

J. C. Platt. Probabilistic outputs for support vector machines and comparison to regularized likelihood methods. In Alexander J. Smola, Peter Bartlett, Bernhard Schölkopf, and Dale Schuurmans, editors, *Advances in Large Margin Classifiers*, pages 61–74, Cambridge, MA, 2000. MIT Press.

S. Pölt. Forecasting is difficult – especially if it refers to the future. Presentation at *AGIFORS 1998 Reservations and Yield Management Study Group Annual Meeting*, Melbourne, Australia, 1998.

F. Provost and P. Domingos. Tree induction for probability based ranking. *Machine Learning*, 52(3):199–215, 2003.

J. R. Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann, San Mateo, CA, 1993.

M. Rajopadhye, M. B. Ghalia, and P. P. Wang. Forecasting uncertain hotel room demand. *Information Sciences*, 132(1):1–11, 2001.

B. D. Ripley. *Pattern Recognition and Neural Networks*. Cambridge University Press, Cambridge, UK, 2008.

M. Rothstein. OR and the airline overbooking problem. *Operations Research*, 33(2):237–248, 1985.

K. T. Talluri and G. J. van Ryzin. *The Theory and Practice of Revenue Management*. Kluwer Academic Publishers, Dordrecht, The Netherlands, 2004.

G. Wahba. Support vector machines, reproducing kernel Hilbert spaces, and randomized GACV. In B. Schölkopf, C. J. C. Burges, and A. J. Smola, editors, *Advances in Kernel Methods — Support Vector Learning*, pages 69–88, Cambridge, MA, 1999. MIT Press.

K. T. Wu and F. C. Lin. Forecasting airline seat show rates with neural networks. In *Proceedings of IJCNN'99 International Joint Conference on Neural Networks*, pages 3974–3977, 1999.

T. F. Wu, C. J. Lin, and R. C. Weng. Probability estimates for multi-class classification by pairwise coupling. *Journal of Machine Learning Research*, 5:975–1005, 2004.

J. Zhu and T. Hastie. Kernel logistic regression and the import vector machine. *Journal of Computational and Graphical Statistics*, 14(1):185–205, 2005.

	Information gain ratio (in ‰)							Random forest importance score (in %)						
Time-to-service	32	16	8	4	2	1	0	32	16	8	4	2	1	0
Agent	11.43	7.81	4.75	2.41	1.88	1.69	0.86	4.39	5.30	4.58	3.89	3.46	3.06	3.18
Channel	8.60	7.95	4.59	1.93	1.22	1.05	0.84	3.11	3.84	3.25	2.14	2.21	2.08	2.21
Company	5.50	4.23	2.23	1.36	1.73	2.02	1.49	1.81	2.57	2.71	2.66	2.93	2.92	3.68
Day	0.99	1.09	0.92	0.51	0.62	0.20	0.03	2.73	4.30	4.70	4.43	4.87	4.50	5.43
Group	6.15	4.51	5.10	3.89	3.54	1.37	0.03	1.98	2.13	3.76	4.14	3.72	2.32	1.23
Length	3.59	1.89	1.46	1.35	1.73	1.57	0.63	2.89	3.29	3.12	3.03	3.14	3.37	2.59
Market	9.40	6.72	4.66	2.85	2.68	2.16	0.89	7.01	7.92	7.75	6.74	6.39	5.19	5.16
Month	0.37	0.16	0.12	0.09	0.12	0.12	0.08	2.70	4.20	4.38	3.93	4.29	4.22	5.58
Price	-	-	-	-	-	-	-	4.76	6.75	6.69	5.96	6.29	6.23	6.99
Ratecode	9.46	7.25	4.91	3.33	3.58	3.22	1.47	4.43	5.70	4.88	4.73	5.41	5.87	6.03
Refundable	14.16	12.71	9.50	5.74	5.25	4.83	2.66	0.58	0.59	0.59	0.45	0.39	0.32	0.21
Roomtype	6.49	5.34	3.85	1.94	1.04	0.65	0.42	5.25	8.59	7.81	5.57	4.17	3.43	3.31
System	7.31	5.72	3.23	1.67	1.13	0.91	0.65	5.74	6.97	5.16	4.07	3.94	4.22	4.58
Timebooking	0.83	0.73	0.52	0.21	0.83	2.91	2.87	2.17	4.65	5.58	4.99	5.93	8.03	10.76
TOTAL	84.26	66.10	45.84	27.27	25.33	22.69	12.90	49.55	66.79	64.95	56.72	57.15	55.75	60.94

Table 4: Information gain ratios and random forest importance scores of the 14 explanatory variables at seven time points