Passenger Name Record Data Mining Based Cancellation Forecasting for Revenue Management

Dolores Romero Morales * and Jingbo Wang †

April 16, 2008

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 forecasting models proposed in the literature, which mainly address the no-show case. Using a real-world dataset, we examine the performance of the existing models and propose new promising ones based on Support Vector Machines. We also illustrate how the set of relevant variables to describe cancellation behavior is very different in different stages of the booking horizon, which confirms the dynamic aspect of this problem.

Keywords: revenue management, cancellation forecasting, PNR data mining, two-class probability estimation, support vector machines

*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 [27]. The classical revenue management scenario is that in which a service provider sells a fixed number (the capacity) of identical 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" [24] 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" [27], which is the actual capacity plus a buffer representing the number of bookings that are expected to be canceled. In both cases, accuracy of cancellation rates is crucial to the revenue management system.

Airlines and hotels routinely practice overbooking [26], 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, over-

booking is only needed in the very late part of the booking horizon. This has led to think that forecasting cancellation is only necessary close to the delivery of the service [4]. However, cancellation forecasting is not only required to determine the levels of overbooking.

Another equally important purpose is to contribute to the estimation of net demand [4, 24]. During the booking horizon, and as mentioned in [13], the revenue management system constantly needs two parts of estimate of the net demand: the net demand to come and 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, and around 20% of them are eventually canceled. In average, the number of on-going bookings reaches 40% (respectively 80%) of the number of total show-ups at 10 (respectively 3) days prior to the time of service. However, about 16% (respectively 11%) of them are canceled at a later point of time, which means that net demand is just 33% (respectively 71%). As a result, errors in estimating these two cancellation proportions could have a big impact on the estimation of total net demand. The proportion of cancellation is even higher in airlines. A recent study points out that "cancellation proportions of 30% or more are not uncommon today", see [13].

Needless to say that the bias caused by errors in cancellation 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 will make the system underestimate net demand and therefore set the price too low (in order to attract more demand). Similarly, overestimating cancellation will make a system using allocation control open a cheaper fare too early or for too long. As far as the authors are aware, there are not citations on revenue increase and cancellation forecasting. On a more general setting, Pölt [21] 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 [6] 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 forecasting. Motivated by large Passenger Name Record (PNR) databases, several recent attempts have been made to use PNR information to improve the accuracy of no-show forecasting with the help of Data Mining [10] tools, and have shown promising results [9, 12, 17, 18]. A recent study modeled cancellation as a survival process and used logistic regression to forecast the hazard rate [13]. In this study, we discuss PNR data mining based forecasting for cancellations happening at any time during the booking horizon. Our contribution is two-fold. First, 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, concluding that multiple models should be built to capture such complex dynamics. Second, and by noticing that the task of PNR cancellation forecasting can be seen as a two-class probability estimation problem [10], we study the performance of state-of-the-art methods in this field, three of which are applied to cancellation forecasting for the first time. Our numerical results on the real-world dataset suggest that Support Vector Machines (SVM) [5] is a very promising method to forecasting cancellations.

The rest of the paper is structured as follows. In Section 2 we describe in detail the cancellation forecasting problem and the existing models in the literature. In Section 3, we introduce the real-world dataset that inspired this work. Using this dataset, we illustrate how the set of relevant variables to describe cancellation behavior is very different in different stages of the booking horizon. In Section 4, we discuss state-of-the-art two-class probability estimation techniques including SVM. In Section 5, we present our numerical results which confirm that SVM is a promising technique. Conclusions and some lines for future research are discussed in Section 6.

2 Problem formulation and literature review

In this section, we formulate the cancellation forecasting problem and describe the different models proposed in the literature. Before that we introduce some notation that will be used throughout the paper.

2.1 Definitions and basics

2.1.1 The ultimate goal

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 proportion of cancellation among all on-going bookings at time $T-t$. Forecasting models for $Q(t)$ are built based on historical booking records.

2.1.2 The information at hand

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

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

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

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 $T_s - t$.

2.1.3 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, while 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".

2.2 Existing forecasting models

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 [9, 12, 17, 18].

2.2.1 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 [17], 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. In addition, some ad-hoc rules based on the idea of giving historical booking records with similar seasonal information greater weight might also be applied. For example, to forecast proportion of cancellation for a flight on a Monday in July, historical booking records for a flight on Monday or in July will be given greater weights.

2.2.2 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 [10] 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 card($\mathcal{O}(t)$) is the cardinality of the set $\mathcal{O}(t)$. This means that forecasts of seasonal average models can be provided very early and for any value of t, even before the booking horizon starts, as long as the seasonal information on the service instance S is known. However, 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 $T - t$. Such difference relies on the fact that S is defined at the service level while R also contains information at the booking level.

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 [8] and Wu and Lin [28] both trained neural networks, but their PNR data contained just seasonal information. Gorin et *al.* [9] used weighted averaging with weight determined by three PNR variables using ad-hoc rules. Hueglin and Vannotti [12] used a simple decision tree [25] with merely 15 nodes, as well as logistic regression to build predictive models; Lawrence et *al.* [17] used ProbE [1], a hybrid of decision tree and Naive Bayes, C4.5 decision tree [23] and the Adjusted Probability Model [11], an extension of Naive Bayes. Neuling et *al.* [18] also used C4.5 decision tree. They all claimed significant improvement (5% to 15% reduced error) on accuracy over seasonal average models.

3 PNR data mining based cancellation forecasting

PNR data mining models are clearly the future direction, as they use the most information as well as state-of-the-art techniques, and have shown promising results in no-show forecasting. However, extending PNR data mining based forecasting from no-show to cancellation at any time is not a straightforward task. As discussed in Section 2.1.3, the extended model should be able to produce a complete "cancellation curve", i.e. calculating $Q(t, R)$ for any t, and not just for $t = 0$ (the no-show case). Here the question of how to model t properly arises. In this section, we will identify a solution based on our analysis of the dynamics in a real-world PNR dataset.

3.1 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 proportion of cancellation 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 proportions 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 proportions are very big compared with their mean values – which shows how random the process of cancellation is and therefore the challenge we face.

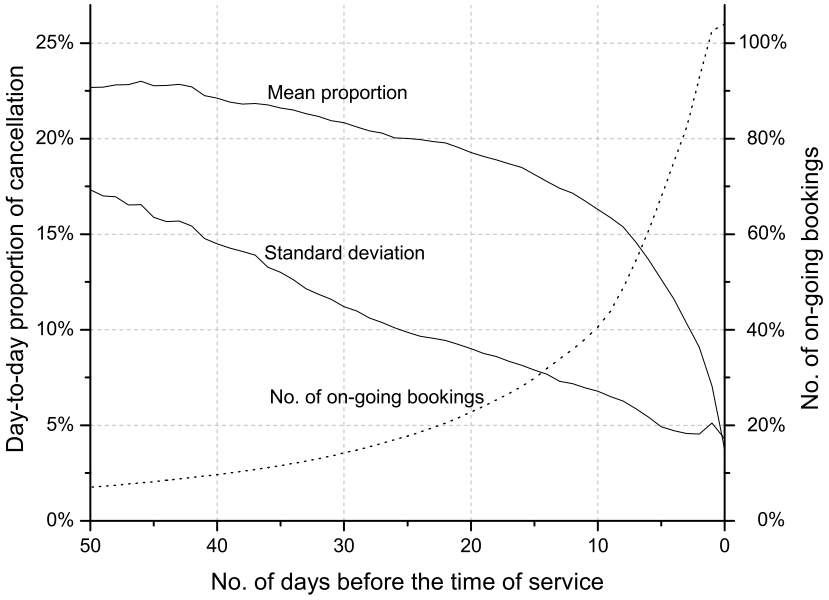

Figure 1: Proportion of cancellation across the booking horizon

We may recall that a PNR entry is given by $R = (T_s, t_b, X, \ell, t_c)$, see Section 2.1.2. In our hotel dataset X contains a total of 13 variables, which together with t_b ("time-booking"), are used as explanatory variables to forecast cancellations. Details of

the 14 explanatory variables can be found in Table 1. Simple preprocessing has been done. For example, variables such as "company" and "agent" originally had hundreds of possible values but many of them were actually very rare, so we grouped infrequent values together. Besides, continuous variables such as "timebooking", "length" and "price" have been discretized.

Name	**No. of Values**	**Description**	**Info Gain Ratio**
Timebooking	16	No. of days to service when the booking is created	0.327%
Refundable	2	Is the booking refundable?	0.266%
Company	34	Is the booking from a company? If so, which company?	0.149%
Ratecode	59	Rate code of the booking	0.147%
Market	44	Market sector the booking is from	0.089%
Agent	44	Is the booking from an agent? If so, which agent?	0.086%
Channel	5	Channel (internet, phone, etc.) used to make the booking	0.084%
System	25	Reservation system used to make the booking	0.065%
Length	8	No. of nights of stay	0.062%
Roomtype	21	Room type	0.042%
Price	6	Price of the booking	0.035%
Month	12	Month of year of the service	0.008%
Day	7	Day of week of the service	0.003%
Group	2	Is the booking part of a group?	0.003%

Table 1: Explanatory variables for building PNR based models

Not all these 14 variables are of equal importance for cancellation forecasting. To illustrate this, we use the fact that the task of PNR cancellation forecasting can be seen as a two-class probability estimation problem. In this context, the *information gain ratio* [23] is a commonly used measure of a discrete 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 and the negative class. Let p be the proportion of

objects in the positive class within Ω. Consider a discrete explanatory variable x, with m possible values. Let Ω_i be the subset of Ω defined by the i-th value 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.

Table 1 shows the information gain ratios of the 14 explanatory variables for the no-show forecasting problem ($\Omega = \mathcal{H}(0)$). Explanatory variables "timebooking", "refundable", "company" and "ratecode" have the highest information gain ratios, which means they are most likely to affect the probability of no-show of a booking. On the other hand, variables "month", "day" and "group" have extremely low information gain ratios, which means they are almost irrelevant to the probability of no-show. However, we should note that these ratios are only valid for no-show probability $(t = 0)$, and might change significantly at a different time point. Therefore, it is still necessary to explore other time points, which will be done in the next section.

3.2 PNR data mining – how to get a practical solution?

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. In the following we will argue that the dynamics of the

cancellation process as well as the impact of variable t make the first approach more plausible.

To illustrate the dynamics of the cancellation process, we have calculated the information gain ratios of the 14 explanatory variables in Table 1 for $t = 1, 2, 4, 8, 16$ and 32. These results can be found in Figure 2 together with the ratios at the time of service ($t = 0$), already given in Table 1. Different from Table 1, Figure 2 does not use absolute value of information gain ratio but relative value, i.e., a variable's information gain ratio over the total information gain ratio among all the 14 variables. Besides, we have dropped out variables "price", "length", "month" and "day", which have the lowest average information gain ratios at the seven time points. As we can see, the relative importance of variables changes dramatically in different stages of the booking horizon. For example, "timebooking" is the most important variable at the time of service, but is almost irrelevant when $t \geq 4$; "group" is very important at $t = 2$ and $t = 4$ but is less important in earlier and later stages.

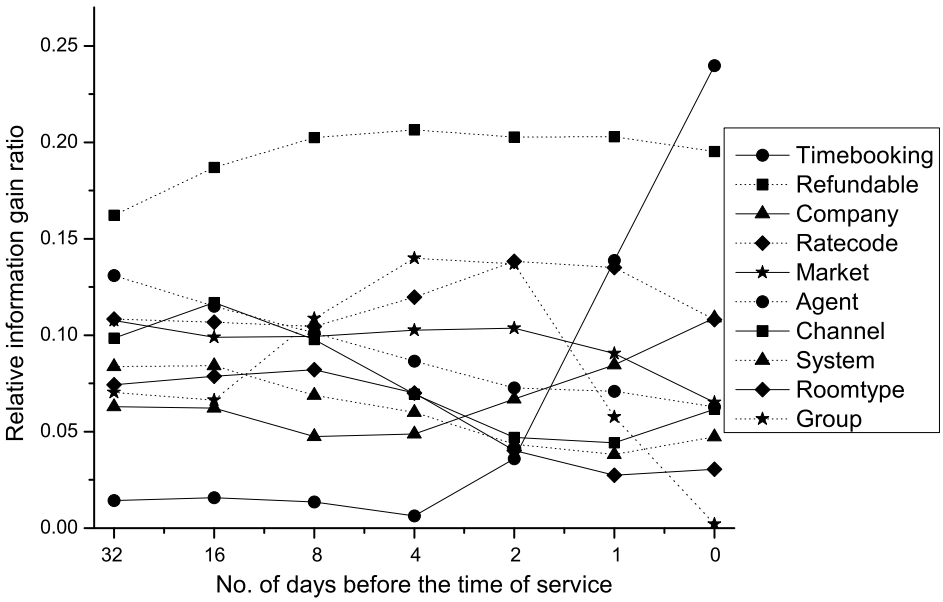

Figure 2: Relative information gain ratio for seven time points

Thus, in general, an explanatory variable's impact on cancellation probability can

be very different at different time points. 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. However, modeling these interactions will bring additional computational burden, making this approach less attractive.

Other data mining methods, such as decision trees, can build a single model by including t as an additional explanatory variable. In general, this new explanatory variable is very influential compared to the rest. In our real-world dataset, the information gain ratio of variable t (with seven discrete values $0, 1, 2, 4, 8, 16$ and 32) is 0.1427, while the total information gain ratio of the 14 variables in Table 1 is only 0.0248, less than 18% of that of t alone. This is not surprising since, in general, the probability of cancellation decreases significantly as t decreases, see Figure 1. As a result, the single decision tree would always first split on t, becoming a collection of sub-trees, one for each value of t. Single-model and multiple-model approaches are then equivalent. Summarizing the discussions above, we believe the multiple-model approach is a plausible choice. In all subsequent experiments, we will only show results for the multiple-model approach.

4 Data mining methods for PNR based forecasting

In this section, we will discuss the data mining methods that will be tested in Section 5. The task of PNR cancellation forecasting can be modeled as a two-class probability estimation problem [10], where the two classes are "canceled" and "not canceled". Therefore, we will include five state-of-the-art methods in two-class probability estimation, three of which are applied to the cancellation forecasting problem for the first time, and two weighted data mining methods, which can be seen as a combination of

PNR data mining and seasonal average.

4.1 Standard data mining methods

There is a rich literature on multi-class probability estimation [3], with roots in classification – one of the key tasks of data mining. In this study, we will compare the performance of state-of-the-art methods from three popular categories: decision tree based methods, Naive Bayes based methods and support vector machine (SVM) based methods. Besides, the classical two-class probability estimation method, logistic regression (LR) will also be included in the comparison.

Provost and Domingos [22] showed that traditional decision trees aiming at improving classification accuracy usually do not give good class probability estimates, which coincides with the poor performance of C4.5 decision tree when forecasting no-shows [17]. Provost and Domingos [22] then developed the first real 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 [7]. Another probability estimation tree is the Minimum Squared Expected Error tree (MSEE) [19]. 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. To the best of our knowledge, both C4.4 and MSEE will be applied to the cancellation forecasting problem for the first time.

Naive Bayes is a classical classification method which has been widely used to derive class probability estimates [10]. It is also the origin of several more recent methods such as Naive Bayes Tree [16] and Adjusted Probability Model (APM) [11]. APM was applied to the no-show forecasting problem in [17] and actually turned out to be the best method reported in their experiments. So we also include APM into our

comparison. Basically, APM adaptively adjusts the weights of different explanatory variables in order to maximize the accuracy of estimated class probabilities.

SVM was originally proposed as a classification method [5]. For two classes, it maps input data points to a higher dimensional space and tries to find a separating hyperplane for the two classes, i.e. where each object is correctly classified, which maximizes the margin. 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 hyperplane is found by solving the following optimization problem:

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

subject to

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

where ξ_i, $i \in I$, are the slack variables and C is a penalty parameter which tradeoffs 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 [5]. Once the optimal solution (ω^*, b^*) is found, new objects are classified using the decision function

$$f(x) = \text{sign}((\omega^*)^\top \phi(x) + b^*).$$

Platt [20] and Wu et *al.* [29] use logistic regression to link a data point's distance to the hyperplane $((\omega^*)^\top \phi(x) + b^*)$ with its class probability and thus making it possible to use SVM for probability estimation. Recent studies [2, 3] have shown that SVM is among the most competitive methods in terms of probability estimation. To the best of our knowledge, SVM has not been applied to the cancellation forecasting problem before.

4.2 Weighted data mining methods

Some existing PNR data mining models [12, 17, 18] have included seasonal information in S as explanatory variables. However, they still neglect an important issue when dealing with historical data. Since customer behaviors change gradually over time, older booking records may not be as representative as newer ones. Therefore, it is reasonable to give more weight to recent bookings in order to better capture the current behavior of customers. Based on this idea, we have developed weighted PNR data mining methods in which weights increase exponentially from the oldest to the latest booking records. We have chosen MSEE and SVM as the base methods, since they have shown to be the most promising ones, see Section 5.3. The two weighted methods will be denoted by WMSEE and WSVM.

This essentially makes both WMSEE and WSVM a combination of PNR data mining and seasonal average. The same idea of training SVM with weighted data has been used by Yang et *al.* [30] in the context of outliers and SVMs. We should note that WSVM would require a slightly different formulation from the one proposed above, in which the second term in the objective function will be replaced by $C \sum_{i \in I} s_i \xi_i$, where s_i is the weight assigned to object $i \in I$.

5 Experimental results on forecasting accuracy

In this section we use the hotel dataset introduced in Section 3 to test the methods discussed in Section 4. We also include simple average (AVG), which is simply the proportion of cancellations 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
LR	Logistic Regression
C4.4	C4.4 Probability Estimation Tree [22]
MSEE	Minimum Squared Expected Error Tree [19]
APM	Adjusted Probability Model [11], a variant of Naive Bayes
SVM	Support Vector Machine [5]
WMSEE	Seasonally weighted Minimum Squared Expected Error Tree [19]
WSVM	Seasonally weighted Support Vector Machine [5]
AVG	Average proportion of cancellation among all training data
SAVG	Seasonally averaged proportion of cancellation among all training data

Table 2: List of methods being tested in Table 3

5.1 Discrete choice for time-to-service

In practice, predictive models are normally built at a limited number of so-called *checkpoints* for t. Moreover, the further from the time of service, the more dispersed the checkpoints should be from each other. 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. In our experiment, predictive models for $Q(t)$ are built at seven selected checkpoints $t = 0, 1, 2, 4, 8, 16$ and 32.

The seven checkpoints we have chosen roughly mark the time points when the number of on-going bookings reaches 105% (time of service before no-show), 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 checkpoints will be a good indicator of a predictive model's ability to capture the dynamics of the cancellation process.

5.2 The tuning

The SVM implementation we use is the `SVMlight` software package [14]. We have tried SVMs with linear, polynomial, sigmoid and Radial Basis Function (RBF) kernels. For simplicity, we will only show results of SVM with RBF kernel in our comparison, as RBF kernel dominates the other three in most of our experiments. For SVM, parameter selection is necessary. In all the experiments, we use two-fold cross validation [15] to find the best parameter. Here we only perform a two-fold cross validation because SVM training is very time consuming. We find that most of the times the parameter selected by this procedure works very well. Another interesting observation is that different parameters are selected at the seven checkpoints, which further confirms that the cancellation behaviors of customers are different at different stages of the booking horizon and multiple models are more appropriate.

For WSVM, WMSEE and 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 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 forecasting [12, 17, 18]. However, any statistical experiment with only one run may be unreliable. Therefore, we have designed a randomized experiment. Instead of adopting the popular practice of randomly splitting the whole dataset into training and testing sets, we believe 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 to train predictive models and

randomly choose 25 service instances out of the last 274 to test the accuracy, having in total 500 service instances in the testing sample.

For each of the 500 testing service instances, we use the following formula to measure a model's accuracy at checkpoint t:

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

where $\mathrm{card}(\mathcal{O}(t)) \times Q(t)$ and $\sum_{R \in \mathcal{O}(t)} \ell$ are the forecasted and actual number of cancellations respectively. This is a more appropriate measure than the absolute difference between $Q(t)$ and actual proportion of cancellation $\sum_{R \in \mathcal{O}(t)} \ell / \mathrm{card}(\mathcal{O}(t))$, because it gives more weight to service instances with a large number of bookings. Table 3 shows the total absolute errors of the 500 days for each checkpoint t, while n_b is the total number of on-going bookings t days prior to the delivery of each service instance and n_c is the total number of cancellations among these bookings.

t	n_b	n_c	LR	C4.4	MSEE	APM	SVM	WMSEE	WSVM	AVG	SAVG
0	103390	3323	2287	2154	2116	2169	2056	2077	1984	3562	3595
1	104291	7273	3294	3406	3316	3331	3125	3340	3122	4362	4382
2	93718	7955	2573	2504	2396	2428	2382	2365	2334	2980	2999
4	73397	7558	2323	2290	2277	2352	2271	2253	2260	2551	2541
8	47121	6578	2014	2020	1962	2045	1928	1917	1938	2369	2369
16	27066	4454	1660	1542	1541	1641	1605	1559	1600	1755	1763
32	12273	2214	985	983	1020	998	974	1021	985	1108	1103
Total absolute error			15137	14900	14628	14963	14342	14532	14223	18686	18752
(Relative to WSVM)			106.4	104.8	102.8	105.2	100.8	102.2	100	131.4	131.8

Table 3: Total absolute errors across 500 service instances in randomized experiment

Several conclusions can be drawn from Table 3. First, PNR data mining models are much more accurate than seasonal average (SAVG), which is just marginally better than simple average (AVG). This indicates that the connection between seasonal information and cancellation behavior is not very strong, which also agrees with the low information gain ratios of variables "day" and "month", see Table 1. Second, WSVM and SVM give the best overall accuracy, in average reducing about 30% of the error of seasonal average (SAVG). If we look at the relative error to WSVM

in Table 3, WMSEE, the best non-SVM method, gives 2.2% more error; APM and LR, the two methods that have been used for no-show forecasting in the literature, gives 5.2% and 6.4% more errors respectively. Such results suggest that applying SVM to the large-scale cancellation forecasting problem in Revenue Management is promising. Third, the two weighted data mining methods WSVM and WMSEE are slightly (around 0.5% to 1%) more accurate than the original SVM and MSEE respectively. However we should note that SVM and MSEE are already the best among the five data mining methods, so such improvement is not trivial and proves the effectiveness of combining PNR data mining and seasonal average. Fourth, we can see a rough trend that more errors are reduced when it is closer to the time of service, for every data mining based method. For example, in Table 3, for $t = 32$ the seven data mining based methods (both original and weighted) in average only reduce 9.8% of the error from seasonal average (SAVG); this ratio is 19.1% for $t = 2$ and 41.0% for $t = 0$. This is good news for practitioners since the closer to the time of service, the more important cancellation forecasting is.

6 Conclusions and future direction

Forecasting for cancellations happening at any time during the booking horizon is a complex problem. As shown in Section 3.2, 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. Therefore, it is necessary to build different models 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 [13].

In this study, we build predictive models based on Support Vector Machines, which outperform models based on several other state-of-the-art data mining methods. We also try to combine PNR data mining and seasonal average and the results are promising. However, there is still a scope for fundamental research in this area, since the

data mining methods proposed for two-class probability estimation problems in the literature have been tested mostly on moderate-size datasets, meaning these methods may underperform in large-size datasets, as is the case for revenue management datasets. Also due to the magnitude of revenue management datasets, computational efficiency becomes a major challenge. For example, the SVM models we build for our hotel dataset usually take hours to train. Considering that airline datasets could be even larger, there is a need to make competitive methods such as SVM more efficient.

References

[1] 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.

[2] 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.

[3] 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.

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

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

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

[7] 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.

[8] 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.

[9] 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.

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

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] R. Kohavi. Scaling up the accuracy of Naive-Bayes classifiers: A decision-tree hybrid. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, pages 202–207, 1996.

[17] 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.

[18] 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.

[19] 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.

[20] J. C. Platt. Probabilistic outputs for support vector machines and comparison to regularized likelihood methods. In A. J. Smola, P. Bartlett, B. Schölkopf, and D. Schuurmans, editors, *Advances in Large Margin Classifiers*, Cambridge, MA, 2000. MIT Press.

[21] 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.

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

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

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

[25] B. D. Ripley. *Pattern Recognition and Neural Networks*. Cambridge University Press, Cambridge, UK, 1996.

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

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

[28] 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.

[29] 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.

[30] X. Yang, Q. Song, and A. Cao. Weighted support vector machine for data classification. In *Proceedings of IJCNN'05 International Joint Conference on Neural Networks*, pages 859–864, 2005.