RC22732 (W0302-266) February 28, 2003
Computer Science

IBM Research Report

Passenger-Based Predictive Modeling of Airline No-show Rates

Richard D. Lawrence, Se June Hong
IBM Research Division
Thomas J. Watson Research Center
P.O. Box 218
Yorktown Heights, NY 10598

Jacques Cherrier
Air Canada
P.O. Box 9000
Dorval, Quebec H4Y 1C2

Research Division
Almaden - Austin - Beijing - Delhi - Haifa - India - T. J. Watson - Tokyo - Zurich

Passenger-Based Predictive Modeling of Airline No-show Rates

Richard D. Lawrence
IBM T. J. Watson Research Ctr
P. O. Box 218
Yorktown Heights, NY 10598
ricklawr@us.ibm.com

Se June Hong
IBM T. J. Watson Research Ctr
P. O. Box 218
Yorktown Heights, NY 10598
sjhong@us.ibm.com

Jacques Cherrier
Air Canada
P.O. Box 9000
Dorval, Quebec H4Y 1C2
jacques.cherrier@aircanada.ca

ABSTRACT

Airlines routinely overbook flights based on the expectation that some fraction of booked passengers will not show for each flight. Accurate forecasts of the expected number of no-shows for each flight can increase airline revenue by reducing the number of spoiled seats (empty seats that might otherwise have been sold) and the number of involuntary denied boardings at the departure gate. Conventional no-show forecasting methods typically average the no-show rates of historically similar flights, without the use of passenger-specific information.

We develop two classes of models to predict cabin-level no-show rates using specific information on the individual passengers booked on each flight. The first of these models computes the no-show probability for each passenger, using both the cabin-level historical forecast and the extracted passenger features as explanatory variables. This *passenger-level* model is implemented using three different predictive methods: a C4.5 decision-tree, a segmented Naive Bayes algorithm, and a new aggregation method for an ensemble of probabilistic models. The second *cabin-level* model is formulated using the desired cabin-level no-show rate as the response variable. Inputs to this model include the predicted cabin-level no-show rates derived from the various passenger-level models, as well as simple statistics of the features of the cabin passenger population. The cabin-level model is implemented using either linear regression, or as a direct probability model with explicit incorporation of the cabin-level no-show rates derived from the passenger-level model outputs.

The new passenger-based models are compared to a conventional historical model, using train and evaluation data sets taken from over 1 million passenger name records. Standard metrics such as lift curves and mean-square cabin-level errors establish the improved accuracy of the passenger-based models over the historical model. All models are also evaluated using a simple revenue model, and it is shown that the cabin-level passenger-based model can produce between 0.4% and 3.2% revenue gain over the conventional model, depending on the revenue-model parameters.

SIGKDD '03 August 24-27, 2003, Washington, DC, USA.

Categories and Subject Descriptors

H.2.8 [**Information Systems**]: Database Applications—*Data mining*

General Terms

Data mining

Keywords

Airline overbooking, no-show forecasting, predictive modeling, classification, probabilistic estimation, model aggregation

1. INTRODUCTION

The practice of optimizing revenue by controlling the availability and pricing of airline seats is commonly referred to as revenue management[7]. Sophisticated revenue management systems are in use at all major airlines today, and are widely viewed as a critical component of an airline's overall logistics framework. Rather than offering identical seats at a common fare, revenue management systems introduce multiple booking classes differentiated by the offered fare as well as other possible restrictions such as cancellation options or overnight-stay requirements. The number of seats allocated to each booking class is determined by the estimated demand for each class. Sales of tickets in each class are controlled in an attempt to maximize revenue. For example, it is desirable to reserve seats in high-fare classes for last-minute travelers willing to pay higher fares, while limiting the number of seats sold in lower-fare classes earlier in the booking process. Revenue management establishes booking policies to determine whether to accept or reject a booking in a specific booking class, given the current number of bookings and expected additional demand prior to departure.

Booking policies must also take into account the possibility that a booking may be cancelled prior to departure, or that a booked passenger may fail to show up at the time of flight departure. Such "no-shows" will result in lost revenue if the flight departs with empty seats that might otherwise have been sold. For this reason, airlines will typically accept bookings in excess of the cabin capacity, based on

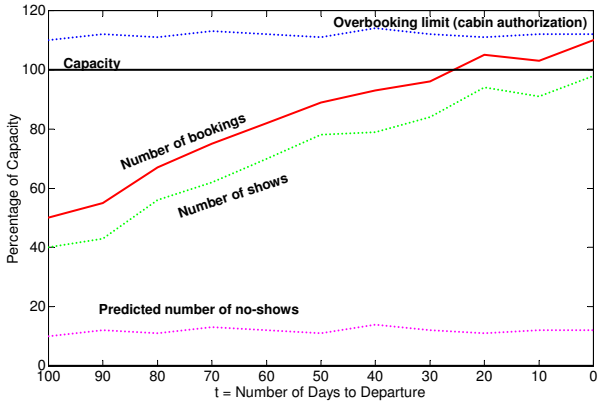

Figure 1: The booking process.

estimates of the number of no-shows. Accurate forecasts of the expected number of no-shows for each flight are essential inputs to the determination of the overbooking level: under-prediction of no-shows leads to loss of potential revenue from empty (or spoiled) seats, while over-prediction can produce a significant cost penalty associated with denied boardings at the departure gate. Optimal booking policies seek the maximum revenue as a tradeoff between the revenue due to additional sales offet by the cost of any denied boardings that may result. In general, airlines introduce contraints to strictly control the number of involuntary denied boardings.

Conventional forecasting methods (see [4]) predict cabin-level no-show rates via statistical methods applied within groups of similar historical flights. The input data for these historical models is limited to past no-show rates at the fare-class level, and does not include information specific to individual passengers. In this paper, we use detailed passenger information to develop improved models to predict cabin-level no-show rates, and compare the results of several passenger-based models with a simple baseline historical method. In the following section, we describe the no-show forecasting problem in more detail.

2. NO-SHOW FORECASTING AND OVERBOOKING

Figure 1 shows a simplified view of the booking process, where the capacity can refer to either the cabin capacity or the allocation of seats to each booking class. Bookings are accepted well in advance of the flight departure date, and no-show forecasting models provide continuously updated predictions of the number of no-shows in the cabin. These forecasts are used to set the overbooking limit for the cabin or booking class; in the simplest case, the overbooking limit is taken as the capacity plus the estimated number of no-shows. Bookings are accepted up to this limit, and hence the total bookings may exceed the capacity. Figure 1 shows an ideal situation where overbooking, based on a correct estimate for the number of no-shows, has resulted in 100% capacity utilization with no denied boardings.

Conventional forecasting methods predict the number of no-shows using time-series methods such as taking the seasonally-weighted moving average of no-shows for previous instances of the same flight leg.[1] Since they do not depend on the specific characteristics of the passengers booked at the time, such forecasts can be provided early in the booking process, with continuous refinement as more recent historical data become available. Passenger-based no-show models compute the no-show probability for each booked passenger, and hence are incapable of producing a complete forecast early in the booking process when relatively few bookings have been made. As discussed in [4], the no-show forecast at any time prior to departure can be taken as a weighted average of the historical and passenger-based forecasts, with increasing weights assigned to the passenger-based forecast as more bookings are realized.

Recently reported passenger-based predictive models employ explanatory variables extracted from databases containing specific information on each passenger and the passenger's itinerary. Hueglin *et. al.* [4, 2] have applied classification trees and logistic regression to the problems of predicting both no-shows and cancellations at the passenger level throughout the booking phase. Kalka and Weber [5] at Lufthansa have used induction trees to compute passenger-level no-show probabilities, and compared their accuracy with conventional, historical-based methods. Continental Airlines [9] describes a decision-tree model based on a relatively small number of input passenger records and features. Selby [12] discusses an application of radial basis functions to passenger-based forecasting, but does not report specific results.

These methods have in common the objective of computing the probability that each individual passenger will not show for the booked flight. The passenger-based probabilities are summed to produce the desired no-show rates at the booking-class or cabin-level. In this paper, we develop two different passenger-based no-show predictive models. The first follows earlier work (but with different explanatory features and predictive methodologies) to compute the no-show probability for each passenger. Unlike previous work cited above, we use the standard historical forecast as input to this *passenger-level* classification problem. Using the results of this passenger-level model, we develop a second *cabin-level* model to predict directly the required cabin-level no-show rate. This latter approach has not been explored previously.

In the following four sections, we describe the available input data, the predictive models, the features used as explanatory variables in the models, and the relative performance of these models. Section 7 develops a simple revenue model in order to provide estimates of the overall revenue impact attributable to the use of these new passenger-level no-show forecasting models.

3. DATA

Information about individual passengers is available in the form of Passenger Name Records (PNRs), which are typically transferred to a PNR database from an airline's flight reservation system. A new PNR is generated whenever a customer makes a flight reservation. In general, a single

[1]A flight leg refers to a section of a flight involving a single takeoff and landing, with no boarding or deplaning of passengers at any intermediate stops. We will often use "flight" in place of flight leg.

Table 1: Summary of input data

	Complete data	Train set	Evaluation set
Number of PNRs	1,258,169	883,269	374,900
Number of flights	15,019	10,931	4,088
First departure date	5/1/2002	5/1/2002	7/11/2002
Last departure date	7/31/2002	7/10/2002	7/31/2002
Mean no-show rate	0.0995	0.0963	0.1071

PNR represents an itinerary that typically will contain multiple flights (segments), and may include more than one passenger flying the same itinerary. In this paper, we will use PNR to refer to a single passenger flying a single flight. The PNR data includes, for each passenger, specifics of all flights in the itinerary, the booking class, and passenger-specific information such as frequent-flier membership, ticketing status, and the agent or channel through which the booking originated. Each PNR is tagged with a label indicating whether the passenger was a no-show for the specified flight.

In order to faciliate comparison with Air Canada's revenue management system, we extracted only PNRs for coach (Y) cabin on single-leg flights departing from Toronto. Table 1 summarizes the characteristics of the resulting filtered PNRs used in our analysis. Approximately 1.26 million PNRs are included, covering a total of 15,019 flights with departure dates spanning a 3 month period. Train and evaluation sets are determined via a split on departure date: the training data spans 71 days of departures, with the subsequent evaluation period covering 21 days. Note that the mean no-show rate in the evaluation data is somewhat lower than the during the test phase.

The data used for our analysis represents a snapshot of the PNR database taken immediately prior to flight departure (t = 0 in Figure 1). For this reason, our no-show forecasts are also computed immediately prior to departure. This implies no limitation on the methodology, however, since our models could be applied to a PNR database snapshot taken at any time t prior to departure (see Figure 1), and the resulting passenger-based forecast combined with a conventional historical forecast using some weighting scheme similar to [4], as mentioned in Section 2.

4. MODEL DEVELOPMENT

In this section, we begin with a description of a conventional historical-based statistical model, and then describe two different passenger-based no-show forecast models. The final subsection discusses the specific predictive techniques used for the passenger-based models.

4.1 Historical Model

Historical models predict the number of no-shows on an upcoming flight by computing the mean no-show rate over a group of similar historical flights:

$$\rho^m_{hist}(b) = \frac{\sum_{m' \in \mathcal{N}_m} \mathrm{NS}^{m'}(b)}{\sum_{m' \in \mathcal{N}_m} \mathrm{B}^{m'}(b)} \ , \tag{1}$$

where

m denotes a flight characterized by a unique flight number and departure date,

$\rho^m_{hist}(b)$ is the predicted no-show fraction in booking class b for flight m,

$\mathrm{NS}^m(b)$ is the number of no-shows in booking class b on flight m,

$\mathrm{B}^m(b)$ is the number of passengers booked in class b on flight m, and

$\mathcal{N}_m$ denotes the set of similar historical flights for flight m.

In practice, $\mathcal{N}_m$ might include earlier flights with the same origin and destination as flight m, departing on the same day of the week as flight m. We will discuss specific implementations of the historical model in Section 6.

Given $\rho^m_{hist}(b)$, the predicted number of no-shows NS^m_{hist} in the cabin (over all booking classes) is

$$\mathrm{NS}^m_{hist} = \sum_{b=1}^{N_b} \rho^m_{hist}(b)\mathrm{B}^m(b) \ , \tag{2}$$

where N_b is the number of booking classes. The cabin-level no-show rate for flight m is

$$\rho^m_{hist} = \frac{\mathrm{NS}^m_{hist}}{\mathrm{B}^m} \ , \tag{3}$$

where B^m is the total number of bookings in the cabin.

4.2 Passenger-level Model

We begin by defining a set of features (explanatory variables) characterizing each passenger flying a single flight in a PNR. Upper-case letters will be used to denote these features, with the values of the features given in lower-case. Let $X_i, i = 1, \ldots, I$ denote I features associated with each passenger. Combining all features yields the feature vector

$$\mathbf{X} = [X_1, \ldots, X_I] \ .$$

Each passenger, $n = 1, \ldots, N$, booked on flight m is represented by the vector of feature values

$$\mathbf{x}^m_n = [x^m_{n,1}, \ldots, x^m_{n,i}, \ldots, x^m_{n,I}] \ ,$$

and a class label C with values $\mathbf{c}^m_n$ denoting either NOSHOW (NS) or SHOW.

In addition to these PNR-based features, we use the cabin-level no-show rate ρ^m_{hist}, predicted using our historical model, as an additional model input. It is understood that a passenger n on flight m inherits ρ^m_{hist} for that flight. The passenger-level predictive model is then stated as follows: given a set of class labels c^m_n, a set of feature vectors $\mathbf{x}^m_n$, and a cabin-level historical prediction ρ^m_{hist}, predict the output class of passenger n on flight m:

$$P(C = c^m_n | \rho^m_{hist}, \mathbf{X} = \mathbf{x}^m_n) \ .$$

We are specifically interested in the no-show probability, $c^m_n = \mathrm{NS}$, and write this probability in the simplified form

$$P(\mathrm{NS}|\rho^m_{hist}, \mathbf{x}^m_n) \ . \tag{4}$$

The number of no-shows in the cabin is estimated as

$$\mathrm{NS}^m_{pax} = \sum_{n \in m} P(\mathrm{NS}|\rho^m_{hist}, \mathbf{x}^m_n) \ . \tag{5}$$

Here, the summation is taken over the passengers n booked on flight m, and "pax" refers to a "passenger-level" model because we are computing the no-show probability for each

individual passenger. The passenger-level no-show rate ρ^m_{pax} for the cabin is defined as in Equation (3), *i.e.*

$$\rho^m_{pax} = \frac{\mathrm{NS}^m_{pax}}{\mathrm{B}^m} \ . \tag{6}$$

4.3 Cabin-level Model

Since it is the cabin-level no-show rate that is ultimately required, it is of interest to formulate a predictive model using ρ^m_{actual} directly as the response, rather than computing the no-show probabilities for each passenger in the cabin, and summing them to obtain the cabin rate. We refer to this as a "cabin-level" model since we predicting the no-show rate for the cabin. Of course, we continue to use passenger-specific information in developing such a cabin-level model. An analogous approach can also be used to predict no-show rates at the fare-class level.

Explanatory features for the cabin-level model are obtained as simple statistics of the in-cabin passenger population. If $x^m_{n,i}$ is a real (continuous) feature, the analogous cabin-level feature is obtained simply by taking its mean (and possibly higher-order moments) over all passengers n on flight m, *e.g.*

$$\hat{x}^m_i \equiv \frac{1}{N_m} \sum_{n \in m} x^m_{n,i} \ , \tag{7}$$

where N_m is the number of passengers booked on flight m. If $x^m_{n,i}$ is a nominal feature with J discrete values, we first perform the usual $1 \rightarrow$ J expansion to create J new binary features, and then take the means (and possibly higher-order moments) of each new feature as in Equation (7). For example, if the passenger-level feature is a binary variable (YES/NO) indicating frequent-flier membership, the resulting two (mean) cabin-level features will be the fractions of the in-cabin passengers with YES and NO labels, respectively. In general, the J cabin-level features generated from a J-value nominal passenger-level feature will sum to one. Performing these operations on all passenger-based features results in a cabin-level feature vector $\hat{\mathbf{x}}^m$ for flight m. The length $\hat{I}$ of $\hat{\mathbf{x}}^m$ (*i.e.* the number of cabin-level features) will depend upon the number of discrete values for each of the passenger-level nominal features, as well as number of moments retained; in our models, $\hat{I}$ is typically several times I.

Given the expanded feature vectors, the cabin-level no-show rate can be modeled as either a regression problem,

$$\rho^m_{cabin} = \hat{f}(\rho^m_{pax}, \hat{\mathbf{x}}^m) \ , \quad 0 \le \rho^m_{cabin} \le 1 \ , \tag{8}$$

or directly as a cabin-level probabilistic model,

$$\rho^m_{cabin} = P(\mathrm{NS}|\rho^m_{pax}, \hat{\mathbf{x}}^m) \ . \tag{9}$$

Note that in either formulation, the cabin-level no-show rate ρ^m_{pax} predicted by the passenger-level model is used as an input feature. Since Equation (4) can be implemented using any of a number of probability-estimation methods, equations (8) and (9) can be extended to include as inputs the values of ρ^m_{pax} estimated by multiple passenger-level approximation techniques, *e.g.*

$$\rho^m_{cabin} = P(\mathrm{NS}|\rho^m_{pax_1}, \ldots, \rho^m_{pax_L}, \hat{\mathbf{x}}^m) \ . \tag{10}$$

The specific implementations are discussed in the following subsection.

4.4 Predictive Models

The passenger-level model given by Equation (4) can be implemented using any classification method capable of generating the normalized in-class probabilities required to evaluate Equation (5). Obvious candidates include the conventional Naive Bayes [8] and decision-tree algorithms such as C4.5 [11]. In addition to C4.5, we have implemented the passenger-level model using ProbE [1] and APMR [3]. Brief discriptions of these methods are provided here.

ProbE: IBM ProbE (for probabilistic estimation) [1] is a scalable data mining engine particularly well-suited for implementing segmentation-based modeling techniques in which the input data records are partitioned into segments, and separate predictive models are developed for each segment. ProbE is able to perform the segmentation and predictive modeling within each segment simultaneously, thereby optimizing the segmentation as to maximize the overall predictive accuracy and thus to produce better models. ProbE uses a top-down tree-based algorithm for constructing segmentations, along with a collection of other algorithms to construct the models within each segment. For the passenger-level model, we used the stepwise Naive Baye algorithm, and hence our ProbE implementation can be viewed as a segmented, feature-selected, Naive Bayes predictive model.

Following [13], we calibrated the ProbE-computed probabilities against the actual densities observed in the train data. The calibration curve was generated using a window-based smoothing technique, and then used to recalibrate the probabilities computed for each record in the test data.

APMR: The Adjusted Probability Model (APM)[3] is a new approach to aggregating the outputs of multiple probabilistic models. Given L model outputs, the l-th model output, $P_l(C = c|\mathbf{X})$, contributes to the APM ensemble output as follows:

$$P(C = c|X) \propto P(C = c) \prod_{l=1}^{L} \left[\frac{P_l(C = c|\mathbf{X})}{P(C = c)} \right]^{\alpha_l} \ , \tag{11}$$

where the $P_l(C = c|\mathbf{X})$ is the probability of observing class label c, given features $\mathbf{X}$ as in Section 4.2. The α_l values are the importance weight for the l-th model fitted for the minimum loss à la logistic regression. When the α values are set to 1, and the individual models are just single-feature models, i.e., $P_l(C = c|\mathbf{X}) = P(C = c|\mathbf{X}_l)$, the APM becomes identical to the Naive Bayes model as shown in [3]. Unlike the Naive Bayes model, APM is not adversely affected by correlated features. The regularized version of APM, called APMR, is used in our modeling.

We now consider the specific application of APMR to the passenger-level and cabin-level models defined above. To illustrate, we consider first the calculation of $P(\mathrm{NS}|\mathbf{x}^m_n)$ in Equation (4), where we neglect ρ^m_{hist} for the moment. The APM estimate is

$$P(\mathrm{NS}|\mathbf{x}^m_n) \propto P(\mathrm{NS}) \prod_{i=1}^{I} \left[\frac{P(\mathrm{NS}|x^m_{n,i})}{P(\mathrm{NS})} \right]^{\alpha_i} \ , \tag{12}$$

where $P(\mathrm{NS})$ is the prior probability of observing a no-show. If outputs $\rho_1, ..., \rho_L$ from L other probabilistic models are available, Equation (11) can be written to include these estimates explicitly,

$$P(\mathrm{NS}|\rho_1, ..., \rho_L, \mathbf{x}^m_n) \propto P(\mathrm{NS}) \prod_{l=l}^{L} \left[\frac{\rho_l}{P(\mathrm{NS})} \right]^{\alpha_l} \prod_{i=1}^{I} \left[\frac{P(\mathrm{NS}|x^m_{n,i})}{P(\mathrm{NS})} \right]^{\alpha_i} . \tag{13}$$

Table 2: Explanatory variables

Name	Description	Type	Number of Values	Information Gain
No-show label	Target label (Y == no-show)	nominal	2	NA
Passenger_Ticketed	Ticket number issued? (Y/N)	nominal	2	0.0563
Frequent_Flier	Air Canada Frequent Flier? (Y/N)	nominal	2	0.0162
Arrival_Airport	Identifies unique flight leg	nominal	79	0.0112
Flight_Leg_Group	Flight-legs grouped by no-show rate	nominal	5	0.0092
PNR_Originator	Booking originator (*e.g.* AA)	nominal	5	0.0087
Hist_Noshow_Rate	Cabin no-show rate from historical model	real	(5 bins)	0.0079
Booking_Class	Booking class	nominal	11	0.0044
Booking_Class_By_Noshow	Booking class by no-show rate	nominal	3	0.0040
Segment_distance	Segment distance	real	(5 bins)	0.0034
Gender	Derived from title (including unknown)	nominal	3	0.0031
Booking_Class_By_Yield	Booking class by yield	nominal	3	0.0018
Binned_Days_Since_PNR_Creation	Binned days since PNR was created	nominal	5	0.0014
Binned_Departure_Hour	Binned flight departure hour	nominal	4	0.0012
Connecting_Flight	Connecting flight? (Y/N)	nominal	2	0.0007
Binned_Num_Passengers	Number of passengers in this PNR	nominal	3	0.0006
Special_Meal	Ordered special-meal? (Y/N)	nominal	2	0.0005
Binned_Churn	Binned churn values	nominal	3	0.0004
Binned_Num_Segments	Number of flight segments in this PNR	nominal	3	0.0003
Departure_Day	Departure day of week	nominal	7	0.0002

The passenger-level model [Equation (4)] is realized using ρ^m_{hist} as a model input in Equation (13). The cabin-level model [Equation (9)] is obtained using Equation (13) with $\rho_l, l = 1, 2, 3$, taken as the outputs of the passenger-level models implemented using C4.5, ProbE, and APMR, respectively.

Conventional linear regression [6] is used to solve the regression form of the cabin-level model given by Equation (8).

5. FEATURE EXTRACTION

Table 2 summarizes the features extracted for each PNR, sorted by the information gain computed for each feature. Information gain [8, 11] is a popular metric for measuring the contribution of a feature to determination of a class label. It is important to note that information gain measures the contribution of the feature in isolation, and it is possible for a feature with relatively low information gain to improve the predictive model via interaction with other features.

For purposes of computing information gain, the real features were discretized into 5 equal-population bins. The discretized values were also used as input to the APMR and C4.5 passenger models, while ProbE accepts the continuous inputs directly, with discretization handled internally.

Table 2 shows that whether a passenger is ticketed, and membership in a frequent flier program have the highest information gain. Although not explicitly shown in this table, ticketed passengers, as well as frequent fliers, are significantly more likely to show for a flight. Other important features include the flight-leg destination and the PNR originator (*i.e.* the channel through which the booking was made). The no-show rate predicted for the entire cabin using the conventional historical model [Equation 3] is also an important feature at the passenger level. Whether or not a special meal was ordered, by itself, does not appear to contribute much. The Binned_Churn feature captures the number of times the passenger has made itinerary changes after the PNR was created. We expected that passengers making frequent itinerary changes would be more likely to not show, but this premise is not supported by the information gain.

All 19 features shown in Table 2 were used as input to the passenger-level models, since, as noted above, information gain neglects possible interaction with other features. The same features, with the exception of the flight destination, were provided as initial input to the cabin-level models, along with the output probabilities from the C4.5, ProbE, and APMR passenger-level models. Due to the expansions discussed in Section 4.3, a total of 65 features were generated for the cabin-level models.

6. RESULTS

In this section, we compare results computed using the historical, passenger-level, and cabin-level models described in Section 4. As shown in Table 1, the models were built using approximately 880,000 PNRs booked on 10,931 flights, and evaluated against 374,900 PNRs booked on 4088 flights. All results shown here are for the evaluation set of PNRs and flights.

Revenue management systems such as PROS [10] employ proprietary history-based models to predict the required no-show rates. Since we did not have access to such a production-level model, we developed a historical model that appears to be reasonably representative of standard historical models, with the possible exception that it uses a shorter historical period since we had access to only 3 months of flight data. Given that we use our historical model as a baseline for comparisons with our passenger-based models, we explored a range of parameters in order to generate the lowest errors against the evaluation set. The resulting best model for our data was computed via Equation (1), evaluated with $\mathcal{N}_m$ taken over identical flight legs (*i.e.* same origin and destination), departing on the same day of the week in the 45 days prior to the target flight m.

Note that unlike the passenger-based models, which are

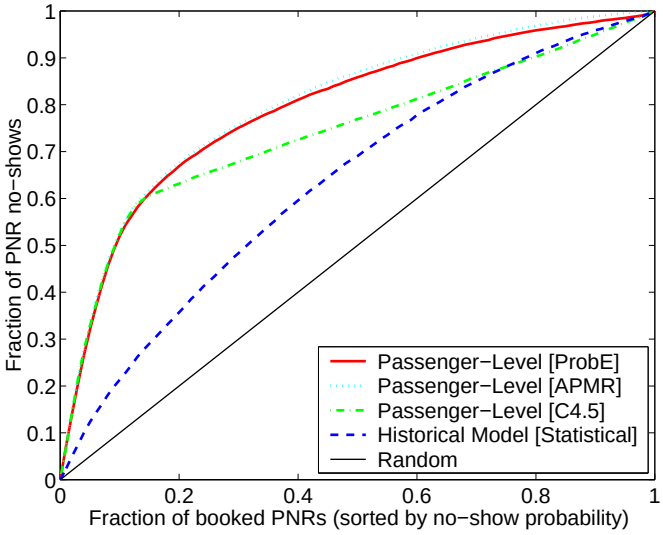

Figure 2: Lift curves for 374,900 evaluation PNRs.

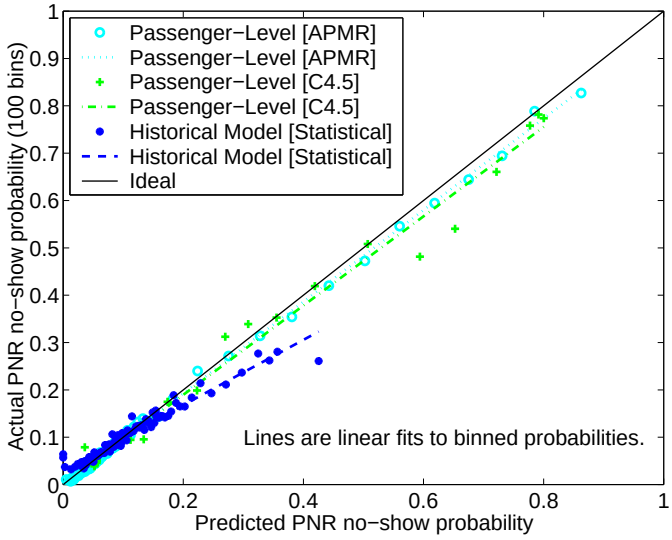

Figure 3: Actual versus predicted passenger-level no-show probabilities.

built using only data from the train set, the historical model uses a moving time-series incorporating data from the most recently available flights. The historical model can also be viewed as a segmented model in which a separate, trivial model (taking the mean of past instances) is constructed in each of many segments formed by selecting identical flights departing on the same day of the week. In other words, the historical model can be viewed as an extremely bushy decision tree generated by hand.

Separate APMR passenger-level models were built in each of 5 distinct segments formed from splits over the important features Passenger_Ticketed, Frequent_Flier, and PNR_Originator shown in Table 2. These segments were determined via limited experimention to produce the minimum loss over the training set. As noted in Section 4.4, ProbE automatically generates a segmented model. The segmentation tree of the ProbE model and the upper nodes of the C4.5 tree generated splits based on these three features.

6.1 Passenger-level No-show Probabilities

Figure 2 shows a conventional lift curve computed using the three different implementations (Section 4.4) of the passenger-level model. Each point on the lift curve shows the fraction of actual no-shows observed in a sample of PNRs selected in order of decreasing no-show probability. The diagonal line shows the baseline case in which it is assumed that the probabilities are drawn from a random distribution. In the case of the historical model, the no-show probability for each passenger is taken as the no-show rate for the passenger's booking class computed using Equation (1). All three implementations of the passenger-level model identify approximately 52% of the actual no-shows in the first 10% of the sorted PNRs. The C4.5 result degrades because many of the lower-probability PNRs are assigned to the same leaf node, producing identical probabilities that yield essentially random orderings in this region. The ProbE and APMR results are very similar, with APMR producing a slightly better lift than ProbE. The lift curve for the historical method is much poorer than any of the passenger-level models, identifying only 21% of the no-shows in the first 10% of the sorted evaluation PNRs. The dramatic difference between these lift curves demonstrates the significant improvement by using PNR-based features in the model.

The lift curve measures the relative ranking of the computed probabilities, but the accuracy of the probabilities is perhaps more important here given the need to compute cabin-level no-show rates via Equation (5). Figure 3 shows the binned actual no-show density as a function of the predicted no-show probability for each PNR in the evaluation set. Each curve is computed by sorting the probabilities in increasing order, binning the result in 100 equal-population bins, and then calculating the density of actual no-shows in each bin. Note that the passenger-level APMR model produces a very strong correlation with the actual density, and displays relatively little bias as shown by the close alignment with the ideal curve.

6.2 Cabin-level Error Analysis

In this sub-section, we summarize the accuracy of the various models measured by the errors in the predicted no-show rates for the coach (Y) cabin. The root-mean-square (RMS) error in the Y-cabin no-show rate is given by

$$\epsilon_{model}(\rho) \equiv \sqrt{\frac{1}{N_f}\sum_{m=1}^{N_f}[\rho^m_{model} - \rho^m_{actual}]^2} \ ,$$

where N_f is the number of flights in the evaluation set. The error in the predicted number of no-shows is given by

$$\epsilon_{model}(\mathrm{NS}) \equiv \sqrt{\frac{1}{N_f}\sum_{m=1}^{N_f}[\mathrm{NS}^m_{model} - \mathrm{NS}^m_{actual}]^2} \ .$$

Taking the standard historical forecast as a baseline, we define the improvement $\delta_{model}(\mathrm{NS})$ in the no-show count for each model relative to the historical model as

$$\delta_{model}(\mathrm{NS}) \equiv \epsilon_{hist}(\mathrm{NS}) - \epsilon_{model}(\mathrm{NS}) \ .$$

Table 3 summarizes the accuracy of different models using the cabin-level error metrics defined above. The Prior

Table 3: Summary of cabin-level errors for all evaluation flights

Model	$\epsilon_{model}(\rho)$	ϵ_{model}(NS)	δ_{model}(NS)
Prior Probability	0.07422	7.98	-1.98
Historical [Stat]	0.06466	6.00	0.00
Passenger-Level [C4.5]	0.05657	5.50	0.50
Passenger-Level [ProbE]	0.05445	5.24	0.75
Passenger-Level [APMR]	0.05358	4.80	1.20
Cabin-Level [Regression]	0.05387	4.99	1.01
Cabin-Level [APMR]	0.05315	4.80	1.20
4088 evaluation flights: $\overline{\rho}_{actual} = 0.1071$ $\overline{\text{NS}}_{actual} = 9.82$			

Probability results refer to the simplest possible model in which the cabin-level no-show rate for each evaluation flight is taken simply as the mean no-show rate over the training set. Not surprisingly, the historical model out-performs this model noticeably, measured by the errors in both the no-show rate and the no-show count in the cabin.

The C4.5, ProbE, and APMR passenger-level models demonstrate progressively better performance relative to the historical model, with the APMR passenger-level model producing an improvement of 1.20 seat over the historical model. Although ProbE and APMR produced similar lift curves (Figure 2), the APMR errors are significantly smaller than ProbE. The cabin-level errors in in Table 3 reflects the accuracy of the computed probabilities, and one reason could be that APMR minimizes a loss function that directly penalizes incorrect estimation of the class probabilities.[2] We also note that the APMR performance is much closer to ProbE when a single APMR model is built over all the train data, without using the 5-way segmentation discussed above. Comparing performance of the APMR passenger-level and cabin-level models, we observe that the cabin-level model produces a more accurate no-show rate, but the improvements in the no-show counts are identically 1.20 seat. The regression formulation of the cabin-level model does surprisngly well, significantly outperforming two of the three passenger-level models.

Heuglin *et. al.* [4] report results using a slightly different improvement metric,

$$\hat{\delta}_{model} \equiv \frac{1}{N_f} \sum_{m=1}^{N_f} \frac{|\text{NS}^m_{hist} - \text{NS}^m_{actual}| - |\text{NS}^m_{model} - \text{NS}^m_{actual}|}{S^m} ,$$

where S^m is the number of shows for flight m. Their Figure 4 plots this improvement during the booking phase, and shows an improvement of approximately 1.2% immediately prior to departure. Using this metric, the APMR passenger-level and cabin-level models both produce improvements of 1.1%. Note that these values depend immediately on the choice of historical model used in the comparison. Also, only 42 evalation flights are used in [4], whereas our results are for 4,088 flights. Direct comparison with the other methods mentioned in Section 2 is not possible due to imprecise or missing error metrics.

While the errors shown in Table 3 provide a useful summary of the accuracy of the methods measured by conventional data-mining metrics, they do not immediately convey the impact on revenue (reflecting additional seats sold versus possible denied boardings) due to the improved predictions. This revenue impact is explored in the following section.

[2]Recall from Section 4.4 that the ProbE passenger-level probabilities are calibrated following [13]; this recalibration does improve the ProbE accuracy.

7. REVENUE ANALYSIS

As noted in the Introduction, more accurate no-show forecasting models can increase overall revenue by selling seats that otherwise would fly empty, and by avoiding the cost of denied boardings at the departure gate. However, evaluating this revenue impact using past flights is not straightforward: even if an improved model were to free up an additional seat for sale, we have no way of knowing whether there was demand for this seat at departure time. For this reason, we first develop a specific scenario assuming parameterized levels of additional demand during the booking process, and then evaluate each of the predictive models in terms of the revenue generated under this scenario.

7.1 Revenue Scenario

For each coach cabin m in our database, we know B^m, the number of final-booked passengers, NS^m, the number of no-shows among the final booked pasengers, and C^m, the cabin capacity. The number of shows is $S^m = B^m - \text{NS}^m$. The evaluation scenario is summarized as follows.

1. The total demand D^m for seats is assumed to exceed the number of final bookings *i.e.*

$$D^m = f_{demand} B^m , \qquad f_{demand} > 1 , \qquad (14)$$

where f_{demand} is the multiplicative demand factor.

2. The no-show predictions for each model are used to generate an estimate for the expanded cabin $\tilde{C}^m_{model}$, an upper limit on the total number of bookings that can be accepted (see Figure 1). The most straightforward approach is to expand the cabin by the predicted number of no-shows,

$$\tilde{C}^m_{model} = C^m + \text{NS}^m_{model} . \qquad (15)$$

It is possible to extend this expression to include a model-dependent optimization parameter determined by maximizing total revenue. However, we retain Equation (15) in order to focus more immediately on the impact of the accuracy of the no-show forecast, without the additional complication of separately optimizing each model.

3. Given the total demand, incremental bookings ΔB^m_{model} are accepted up to the minimum of the postulated demand and the expanded cabin, *i.e.*

$$\Delta B^m_{model} = max(0, min(D^m, \tilde{C}^m_{model}) - B^m) \ .$$

All incremental bookings are assumed to show, so the final number of shows is

$$S^m_{model} = S^m + \Delta B^m_{model} \ . \tag{16}$$

4. The number of denied boardings produced by each model is

$$\mathrm{DB}^m_{model} = max(0, S^m_{model} - C^m) \ .$$

5. The revenue for flight m is

$$R^m_{model} = \alpha S^m_{model} - \beta \mathrm{DB}^m_{model} \ ,$$

where α is the mean revenue per seat sold, and β is the mean cost of a denied boarding. Without loss of generality, we set $\alpha \equiv 1$. The total revenue is simply the sum over all evaluation flights,

$$R_{model} = \sum_{m=1}^{N_f} R^m_{model} \ . \tag{17}$$

We compare each model with a baseline model in which no overbooking is permitted,

$$\tilde{C}^m_{noOB} \equiv C^m \ ,$$

and define the normalized revenue for each model relative to this baseline:

$$\bar{R}_{model} \equiv \frac{R_{model} - R_{noOB}}{R_{noOB}} \ . \tag{18}$$

The revenue improvement for each model relative to the historical method is

$$\delta \bar{R}_{model} \equiv \bar{R}_{model} - \bar{R}_{hist} \ . \tag{19}$$

7.2 Revenue-Model Results

Table 4 summarizes the percent revenue improvement for all models as a function of increasing levels of demand above the actual booking as in Equation (14). The revenue improvement of the APMR cabin-level model ranges from 0.41% with 10% excess demand, to 1.21% with 40% excess demand. These results are for a relatively high denied boarding cost ($\beta = 4$), but this is not unreasonable given the strong penalty most airlines attach to involuntary denied boardings. It is interesting to note that while the accuracy of the passenger-level and cabin-level APMR models are very similar in Table 3, the cabin-level model produces larger revenue improvements. Table 3 captures only the mean errors, while the revenue improvement incorporates the impact of the asymmetric costs associated with positive and negative no-show prediction errors, leading to denied boardings and spoiled seats, respectively.

It is important to note that many of the evaluation flights are significantly under-booked ($B << C$). Hence, even with the assumed additional demand, an improved no-show forecast will not produce additional revenue for these flights because the number of additional seats that can be sold is bounded by the demand [see Equation (3)], and there is no chance for denied boardings since such flights will not be overbooked. For this reason, Figure 4 examines the revenue for the best passenger-based model, the APMR cabin-level model, as a function of a threshold on the bookings as a fraction of capacity. The z-axis is the revenue improvement defined in Equation (19), and the x-axis is the denied boarding cost. The y-axis sets a threshold based on the ratio of booked passengers (B) to cabin capacity (C); for example, the plane generated by $min(B/C) \equiv 0$ includes all (4,088) evaluation flights, while $min(B/C) \equiv 1$ includes only the subset (633) of flights which are overbooked in the actual data. Considering only the subset of evaluation flights that are actually overbooked, revenue improvements (for $\beta = 4$) range from 1.68% to 3.09% as the excess demand increases from 10% to 40%. These improvements are larger than those in Table 4 because we are considering the subset of evaluation flights where accuracy of the no-show forecast is critical in order to avoid denied boardings and spoiled seats.

The range of revenue improvements computed under this scenario can be significant because additional revenue from improved overbooking is essentially pure profit, and profit margins in the airline industry have decreased significantly over the past two years.

8. DISCUSSION

We have shown that that models incorporating specific information on individual passengers can produce more accurate predictions of no-show rates than conventional, historical-based, statistical methods. Performance of the different methods developed here was measured using three different metrics: conventional lift curves, mean-square errors in the cabin-level no-show rates, and revenue gain evaluated for a scenario postulating various levels of incremental demand during the booking process. These metrics measure different model characteristics, and it is interesting to note that they provide different insights into the various methods. The lift curves for the three passenger-level models shown in Figure 2 are quite similar (for the first 10% of records), and yet the cabin-level errors in Table 3 show noticeable differences in accuracy for the passenger-level models. As noted in Section 6.2, the improved performance of the APMR passenger-level model appears to be due to the specific loss function minimized in the APMR algorithm. Table 3 also suggests that the accuracy of the APMR passenger-level and cabin-level errors are comparable. However, measured by a revenue metric (Table 4) that heavily penalizes denied boardings (which result when a model over-predicts the number of no-shows), the cabin-level model out-performs the passenger-level model. From this study, we conclude that APMR is particularly suitable for the no-show forecasting problem.

An interesting aspect of the model development here is the combined use of models generated at different levels of data granularity. Given features for the passengers composing each cabin, the natural approach to computing behavior (*e.g.* no-show rates) at the cabin level is via combination of the predicted behaviors (*e.g.* no-show probabilities) for each passenger. Given multiple implementations of the passenger-level model, and statistics of the passenger features for each cabin, we define a coarser, cabin-level model with the response variable taken as the desired cabin-level no-show rate. Results given here, along with results of other models built against subsets of our data, suggest that this latter approach can generate more accurate predictions of the aggregate behavior. There are many other data-mining applications where models can be built based on individual

Table 4: Summary of percent revenue improvement for all evaluation flights

Model	$f_{demand} = 1.1$	$f_{demand} = 1.2$	$f_{demand} = 1.3$	$f_{demand} = 1.4$
Prior Probability	-0.27	-0.34	-0.29	-0.31
Historical [Statistical]	**0.00**	**0.00**	**0.00**	**0.00**
Passenger-Level [C4.5]	0.09	0.20	0.27	0.42
Passenger-Level [ProbE]	0.15	0.33	0.42	0.54
Passenger-Level [APMR]	0.38	0.73	0.93	1.10
Cabin-Level [Regression]	0.30	0.57	0.73	0.87
Cabin-Level [APMR]	0.41	0.78	1.03	1.21
4,088 evaluation flights $\beta \equiv 4$				

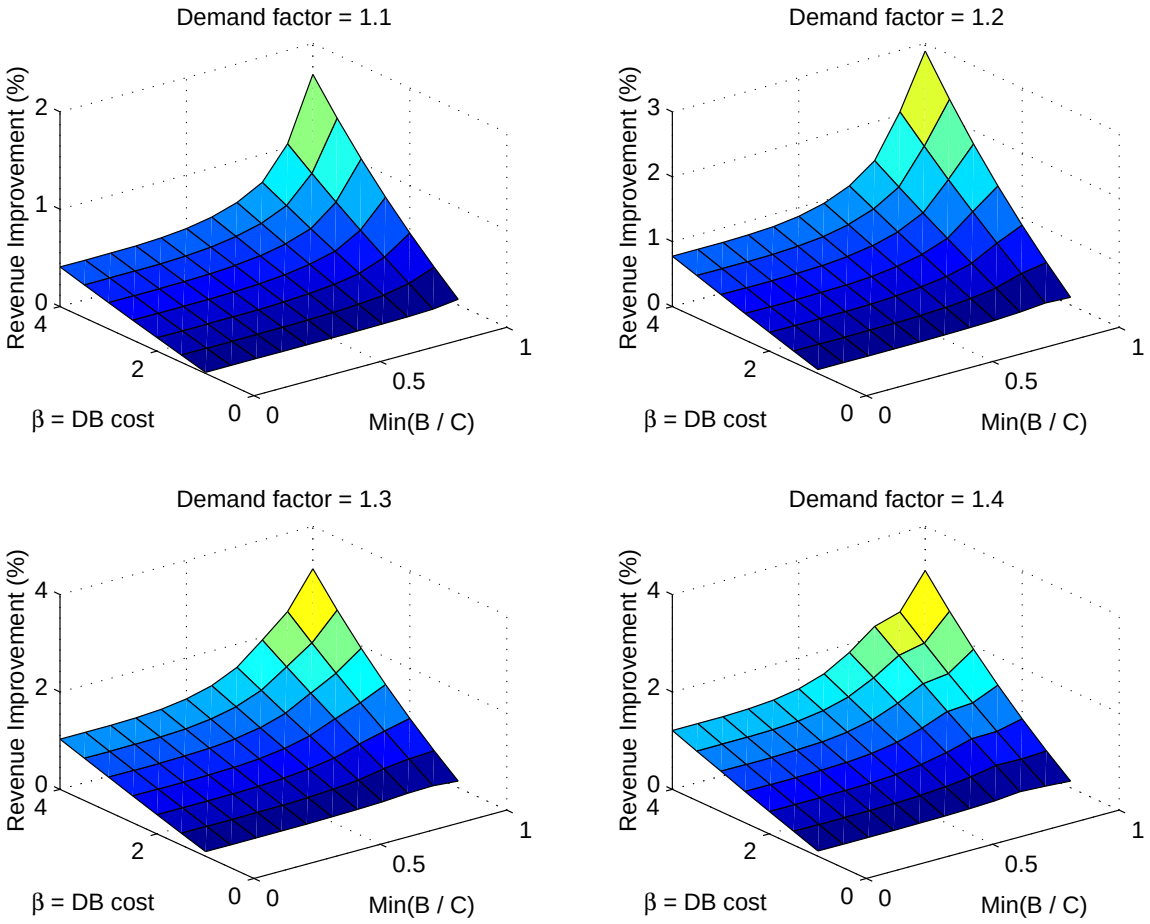

Figure 4: Revenue improvement for the APMR cabin-level model

data, yet the desired prediction is some domain-specific aggregate behavior. We will further study the applicability of our two-stage model generation for these applications.

9. REFERENCES

[1] C. Apte, *et.al.* A probabilistic estimation framework for predictive modeling analytics. *IBM Systems Journal*, 41(3), 438–448, August, 2002. http://www.research.ibm.com/journal/sj/413/apte.html

[2] H. Feyen and C. Hueglin. Data mining techniques to improve forecast accuracy in airline business. *Presentation at 2001 AGIFORS meeting*, Bangkok, Thailand, http://www.agifors.org/studygrp/revmgm/2001/FeyenHuglin.ppt

[3] S.J. Hong, J. Hosking, and R. Natarajan. Ensemble modeling through multiplicative adjustment of class probability. *Proc 2002 IEEE International Conference on Data Mining*, 621–624, 2002.

[4] C. Hueglin and F. Vannotti. Data mining techniques to improve forecast accuracy in airline business. *Proc The Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 438–442, 2001.

[5] H-U. Kalka and K. Weber. PNR-based no-show forecast. *Presentation at 2000 AGIFORS meeting*, New York, NY, 2000. http://www.agifors.org/studygrp/revmgm/2000/agifors2000_lufthansa_systems_berlinweber.ppt

[6] *MATLAB: The Language of Technical Computing.* The Mathworks, Inc. http://www.mathworks.com, 2003.

[7] J. McGill and G. Van Ryzin. Revenue management: research overview and prospects. *Transportation Science*, 33(2), 233–256, 1999.

[8] T. Mitchell. *Machine Learning.* McGraw-Hill, Boston, Massachusetts, 1997.

[9] H. Pastor. What exactly is data mining?. *Presentation at 2000 AGIFORS meeting*, New York, NY, 2000. http://www.agifors.org/studygrp/revmgm/2000/judypastordatamining.ppt

[10] *The PROS 5 Revenue Management System.* http://www.prosrm.com, 2003.

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

[12] D. Selby. Materialisation forecasting: a data mining perspective. *Operations Research in Space and Air*, 393–406, Kluwer Academic Publishers, Boston, 2003.

[13] B. Zadrozny and C. Elkan. Obtaining calibrated probability estimates from decision trees and naive Bayesian classifiers. *Proc Eighth International Conference on Machine Learning (ICML'01)*, 609–616, 2001.