VU AMSTERDAM

Reversed Revenue Management

How to use Revenue Management for customer advantage

David Los

23-10-2012

Research Paper Business Analytics

Supervisors: Alwin Haensel and Ger Koole

Faculteit der Exacte Wetenschappen
Vrije Universiteit
De Boelelaan 1081
1081 HV Amsterdam
Nederland

Foreword

This paper is part of the course *Research Paper Business Analytics* from the master program of Business Analytics. The goal of the research is to learn about the different steps when doing an individual research. The research must have practical value and should cover the three main aspects of Business Analytics: informatics, mathematics, and business.

In this paper, I will discuss the effect of Revenue Management on airline ticket prices and how customers can profit from that. I will investigate whether it is possible to advice customers to wait or to buy a ticket immediately based on my own data investigation. Moreover, I will discuss what is known about this subject so far.

I would like to thank my supervisor Alwin Haensel for helping me in the research.

Abstract

It is common belief that airline ticket prices solely rise as time progresses. In practice however, it can be observed that also price *drops* occur. In this research a model is developed to assist the customer in determining when to buy a ticket. This research is partly based on a paper called "*To Buy or Not to Buy: Mining Airfare Data to Minimize Ticket Purchase Price*".

Airline companies are trying to maximize their profit using a strategy called Revenue Management. Revenue Management is a pricing strategy that uses the assumption that the market can be segmented. The prices are separated through the use of different customer groups, of which each group corresponds to a different type of contract. The goal when using Revenue Management is to allocate the total capacity into each of these smaller groups, such that the expected profit is maximized.

The reason that a price drop can occur is because Revenue Management is based on a forecast of the demand. Forecasts are per definition almost never accurate. Moreover, the forecasts that are used in the airline industry are made on a daily basis, usually during the night. The most accurate and sophisticated forecasting techniques need a long time to compute the calculations. This time exceeds the limited time of the airline companies. Therefore, simpler and more rough forecasts are made in the airline industry. When it is observed that the forecast for demand is not achieved, price drops may occur.

In this research a k-Nearest Neighbor algorithm is used to train a model on data of a 106 flights. This model is trained on the explained variable that describes whether a price drop will occur within four days. The explanatory variables that were used to increase the accuracy of the model are all based on simple flight information. This information is for instance the current ticket price, the name of the airline, the destination of the flight, and a history of seven days of ticket prices. Furthermore, statistics based on the seven day history of those prices were also used.

When applying the final model to several combinations of training and test sets, inconsistent results were realized. Remarkably, the paper that was the basis of this research does not describe on what data their model is tested. To be able to compare their results with the results from this research, the model was also applied purely to the training set. The result of this was better than testing on our previously defined test sets. An average amount of $20,37 was saved per customer when applying the model to the training set. Although on average a profit is realized, high variation occurs in using the model. This means that further research is necessary before the model is practicably usable.

Contents

Introduction

The concept of Revenue Management can be hard to understand for those with a limited amount of mathematical knowledge. Many customers still believe that all companies manually decide what to do with their prices. In practice, we can see that airline companies have fully automated processes that determine the optimal price for every possible ticket on a daily basis. This does not always have to be a logic price or it might not be intuitively clear why they changed it. Moreover, many consumers still believe that from the initial date that one can purchase a ticket the prices only rises and that the price will reach a maximum on the departure date. This might be indeed the optimal price curve for airline companies, but in practice the price will also drop every now and then. It is my goal to help customers in their quest of buying a cheap airline ticket by developing a model that advices to the customer.

This research will cover the question whether it is possible to advice customers when to buy an airline ticket in order to save money. To investigate this, some relevant questions will be answered:

1. How does Revenue Management work?
2. What is known about the subject?
3. Is it possible to build a model to advice the customer whether to wait or to buy?

The first question covers Revenue Management, the concept of determining the price of each ticket. I will explain the basics of Revenue Management and how we can use this to help us with predicting the price of airline tickets. After that, I will discuss a paper that already investigated this on a small number of flights in America. Last, I will explain how I tried to build a model that can advice the customer.

Revenue Management

Companies that want to increase their profit usually try to decrease their costs. Revenue Management tries to increase the profit by increasing the revenue while maintaining the same amount of costs. With Revenue Management, companies try to exploit that different people have different needs and wishes. Typically, the companies try to make several products out of one product by changing the conditions, and price the multiple products differently. In order for Revenue Management to be applicable to a business, a few assumptions have to be met:

1. You have a finite amount of capacity
2. The market can be segmented.
3. There is a time limit to sell the resources (the resources sold are perishable).

Revenue Management is applied in many domains, of which one is aviation: "The airline industry is one of the most sophisticated in its use of dynamic pricing strategies in an attempt to maximize its revenue"[1]. Other domains where Revenue Management is used are car rental, hospitality, advertising (on television and online), parking and retailing.

It is interesting that everything about the products that an airline company sells, being the tickets, is fixed in advance. The seats within economy or business class are all the same and will not change. Still, it could be possible that someone in seat 23F bought a ticket for $150 and someone in seat 23E bought a ticket for $400. While dynamic pricing can be very advantageous for the company, in some sectors customers can consider this as unfair. For instance if two people buy the same book, but one pays double the price of the other. In the airline sector this seems to be tolerated more, because it is widely known that ticket prices fluctuate often. The price of a book is considered to be more constant.

Revenue Management can be split into two parts: quantity based and price based[2]. Quantity based Revenue Management has prices of products set in advance and focuses on deciding how much of each product is to be offered. On the other hand, price based Revenue Management uses the variation of the price of a product to manage demand.

In most literature, we can see that airlines typically fall into the first category. We can see this clearly from one of the flights from the data, a roundtrip from Amsterdam to Zurich with flight carrier Swiss International, see figure 1. If we take a closer look at other flight carriers, we see another price process. For example the roundtrip from Amsterdam to Madrid with flight carrier KLM, see figure 2. More different prices are involved here, which can have several explanations. Possibly, KLM has more different classes than Swiss International and prices each of these classes differently. Another option is that the classes from KLM do not have fixed prices, but can be subject to change.

The development in the airline industry is that airlines are not obliged to publish their prices anymore in advance of taking bookings[3]. In the past, fare tariff books and print media were used to advertise. Furthermore, it was more difficult to manage different prices for administrative reasons.

[1] *To Buy or Not to Buy: Mining Airfare Data to Minimize Ticket Purchase Price*, Etzioni et al, 2003.
[2] *Application of Revenue Management to the Manufacturing Industry*, Blumenthal et al, 2009.
[3] *The theory and practice of Revenue Management,* Tulluri and Van Ryzin, 2004, p 515-524.

With the increasing use of internet it became easier for airlines to publish their new prices. As a result, the gap between quantity based and price based RM is getting narrower within the airline industry. Especially the low-cost airlines are using Revenue Management that is closer to price based Revenue Management than to quantity based Revenue Management.

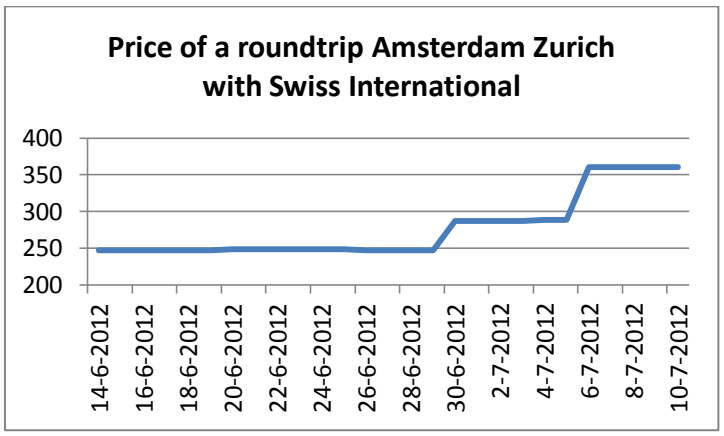

Figure 1

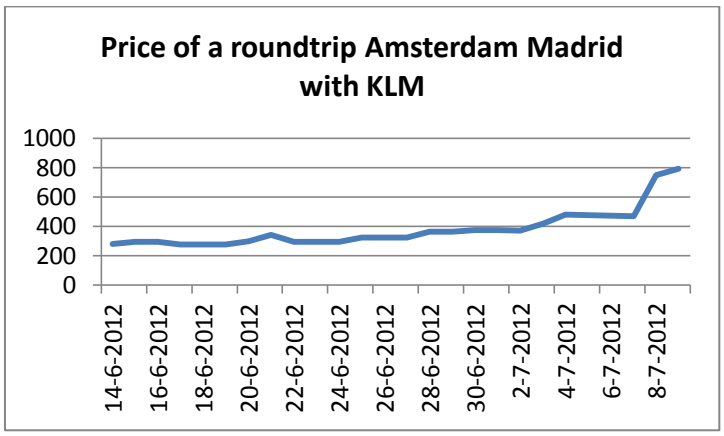

Figure 2

Demand

The demand for a particular good or service depends on a variety of factors. "Key influences include the tastes of consumers, the levels of consumer income, the price and quality of the product in question and the prices of other goods, especially goods that are close substitutes."[4]

Some of these influences are easy to adjust, such as the price. Because one of the assumptions was that the market can be segmented, it is possible to have different customers pay different prices for the same product.

If all other factors remain constant, demand can be expressed as a function of one variable: the price. Let p be the price of a product and d(p) the demand corresponding to that price. It is intuitively clear that the function d(p) is a decreasing function of the price p. An important assumption in many

[4] *Air Travel Demand Elasticities: Concepts, Issues and Measurement*, Gillen et al, 2003.

demand models is that the demand for each product is an independent stochastic process, thus not influenced by the availability for different classes.[5]

When using Revenue Management, it is interesting to measure the difference in demand related to the difference in the price. This is calculated using the *price elasticity*. The price elasticity can be defined as "the percentage change in quantity demanded in response to a one percent change in price"[6]. The concept of Revenue Management uses the price elasticity to generate a variety of prices in order to maximize revenue. Using the demand function, we can define the price elasticity as

$$E(p) \equiv \frac{p}{d}\frac{\partial d}{\partial p} = \frac{\partial \ln(d)}{\partial \ln(p)}.$$

If an increase in price causes a relative bigger decrease in demand, the demand is called *"elastic"*; this occurs when $-\infty < E(p) < -1$. On the other hand, if a price rise causes a relative smaller decline in the demand, the demand is called *"inelastic";* this is the case when $-1 < E(p) < 0$.

Research[3] shows that in the airline sector the price elasticity behaves different for six distinct markets. The demand is less elastic for long hauls than it is for short hauls. The demand for leisure travel is more elastic than for business travel. Furthermore, international travel is less sensitive to changes in price than domestic travel.

The demand function can be used to calculate the revenue as a function of price. Revenue is normally defined as the amount of products sold times the price of each product. Using our functions of demand we can write that as

$$r(p) \equiv p * d(p).$$

In Revenue Management it is our goal to find the maximum possible revenue, using the right combination of demand and price of a product. Thus we are looking for the maximum of r(p) or

$$\frac{\partial}{\partial p} r(p) = 0.$$

$\frac{\partial}{\partial p} r(p)$ is also called the *marginal revenue,* denoted by J(p). Therefore, the expression for J(p) is

$$J(p) \equiv \frac{\partial}{\partial p}(p * d(p)) \; = \mathrm{d}(p) + p\, d'(p).$$

Some common demand functions are linear demand, exponential demand, constant-elasticity demand and logit demand.

Forecast

Airplane ticket prices are generated based on many factors -e.g. USD currency fluctuations, oil price, airport fees, destination, supply and demand- of which the demand forecast is essential to dynamically update the prices. A change in the price of an airplane ticket is often the result of a

[5] *What Variables Affect The Demand For Airline Travel*, Gissimee Doe
[6] http://en.wikipedia.org/wiki/Price_elasticity_of_demand, reference to Png, Ivan (1999) p57.

change in capacity or a change in demand. For instance, if the forecast tells you that five business men will arrive on the last day you would probably want to leave five seats free in the airplane. But if it turns out that not even one of them actually comes you have five spare seats. These seats could have been sold easily to economy passengers. Making a good demand-forecast is therefore essential for airlines.

Some airplane tickets allow customers to cancel or change the flight. Therefore, more revenue can be earned if cancellation can be predicted. Furthermore, sometimes customers don't go on a flight even if they did not cancel it; this is called a *no-show.* A no-show might occur when someone get refused at the airport, when someone does not get at the flight on time or for a variety of other reasons. Both no-show and cancellation-rate forecasts can be used to estimate the amount of extra seats that can be sold on an airplane; this is called *overbooking.*

But more types of forecasts are required. In quantity based Revenue Management, the behavior of customer arrivals during a booking period can be necessary. This type of forecasting is called *booking-curve* or *booking-profile* forecasting.

Forecasts are automated within the airline sector, because of the large amount of forecasts that has to be made. Since some airlines like KLM-Air France change their prices up to once a day, a new forecast has to be calculated every day. In the case of KLM-Air France this happens during the night. A good forecast is essential, but the best forecasting methods that uses the newest techniques are often impossible because of the time-constraint that it has to be finished over one night.

Many different forecasting models are used in Revenue Management, but all of them focus on speed, simplicity, and robustness. Many models and methods that are used in practice are from the ad-hoc forecasting class. Ad-hoc forecasting methods have t he properties that they are intuitive, perform well in practice, and easy to program. The reason that they are called ad-hoc is that "they are implemented without respect to a properly defined statistical model".[7]

Ad-hoc forecast methods work best when the structure of the data can be translated into three components: the level or average of the data, a trend, and seasonality. The strategy that is used with ad-hoc forecasting methods is to *smooth* the data so that noise is less present and then estimate the level, trend and seasonality components in the data.

A simple model from the ad-hoc class is the moving average method. If we define y_t as an observed value at time t and x_t as the forecast value at time t we could compute the forecast for the fourth period by:

$$x_4 = \frac{y_1 + y_2 + y_3}{3}.$$

The idea behind the moving average method is that the most recent observations have a bigger contribution towards the forecast than older data. That is why x_t is usually not calculated by taking all previous observations but only the last N. The forecast for period $t + 1$ is therefore given by:

[7] *Forecasting, Structural Time Series Models and the Kalman Filter,* By Andrew C. Harvey.

$$x_{t+1} = \frac{y_t + y_{t-1} + \cdots + y_{t-N+1}}{N}.$$

The moving average method gives an equal weight to each of the last observations. Sometimes it can be necessary to assign a different weight to observations, usually the recent observations are weighted higher. In these cases another forecasting method is usually used: exponential smoothing. Exponential smoothing is one of the most popular forecasting methods in Revenue Management. This algorithm has property that it is simple, robust and usually achieves a good accuracy. Exponential smoothing can take all the three component estimates into account: the estimate of the average for a period, the estimate of the trend for a period and the estimate of the seasonality factor for a period. The most simple form of exponential smoothing is called *simple exponential smoothing*. This method works by using one parameter α, where $0 < \alpha < 1$, which is called the *smoothing factor*. If we recall the notation, we can recursively define a forecast for period t+1 by:

$$x_{t+1} = \alpha y_t + (1-\alpha)x_t\,.$$

If we would expand the formula by substituting lower x_t with the formula above, we would get:

$$\begin{aligned} x_{t+1} &= \alpha y_t + (1-\alpha)x_t \\ &= \alpha y_t + (1-\alpha)(\alpha y_{t-1} + (1-\alpha)x_{t-1}) \\ &= \alpha y_t + \alpha(1-\alpha)x_{t-1} + (1-\alpha)^2 x_{t-1} \\ &\vdots \\ &= \alpha \sum_{i=0}^{\infty} (1-\alpha)^j x_{t-j}. \end{aligned}$$

Now we can see that exponential smoothing is the weighted average of the previous observations with exponentially decreasing weights.

When α is large, the recent observations will be more important than when α is small. Therefore, a small value of α will cause the forecast to be more smooth, and a high value of α will cause the forecast to be more susceptible to recent changes.

Airline Revenue Management

In the airline industry Revenue Management works on a very specific way. This section will explain the usual procedures.
The tickets are separated into multiple categories, called *booking classes*. For example, the economy class usually has eight or more booking classes. Each of these classes is sold under a different price. The cheaper classes will be sold first and the more expensive classes later.
Under normal conditions, if we simplify it by using only four classes, it will look something like this.

Booking Class	Price
S	€ 100
K	€ 150
B	€ 200
Y	€ 250

In this example we say that the total amount of seats in the plane, denoted by C, is 40. It is now the job of the airline to allocate the total capacity to the smaller booking classes. The amount of capacity that is assigned to each booking class depends on the estimated demand for those classes. If we go back to the example, let us say that the airline estimates the following total demand D per class:

Booking Class	Expected Demand
S	40
K	15
B	10
Y	5

The way that this information is put into an optimization problem is by use of the formulation of a Deterministic Linear Programming (DLP) model. This model uses only the *expected* demand E[D] in its calculations and does not need any information about the distribution of D. Because of time constraints, often approximation algorithms are used to solve this model and thereby compute the optimal amount of space for each booking class.

Let us define $f \in \mathbb{R}^n$ as the vector that represents the price of a ticket in each of the n booking classes. Let the vector $x \in \mathbb{R}^n$ be the decision variable, that represents the amount of space that should be reserved for each of the n booking classes. Furthermore, the matrix A is the identity matrix, with ones on the main diagonal and zeros elsewhere. Now, we can define the DLP model as follows:

Max $f^T x$

s.t. $Ax \leq C$

$0 \leq x \leq E[D]$

If we go back to our example, let x_1 be amount of space that should be reserved for booking class S, x_2 the amount of space for booking class K, etc. Now we get the following model:

Max $100x_1 + 150x_2 + 200x_3 + 250x_4$ (1)

$s.t.$ $x_1 + x_2 + x_3 + x_4 \leq 40$ (2)

$0 \leq x_1 \leq 40$ (3)

$0 \leq x_2 \leq 15$ (4)

$0 \leq x_3 \leq 10$ (5)

$0 \leq x_4 \leq 5$ (6)

When we solve this model, we find an optimal solution of €6500, where $x_1 = 10, x_2 = 15, x_3 = 10, x_4 = 5$.

In practice, the solution of this problem is not often used. Rather, the optimal dual variables are used. Dual variables are the variables that are used when trying to solve the dual variant of a Linear Program. The dual problem provides an upper bound to the primal (i.e. the initial) problem. The solution to the dual problem is called the shadow prices and will tell us the marginal worth of one additional unit of any of the resources (in this case, the capacity) from the primal problem.

Let us define the optimal value of y_1 to be the shadow price of constraint (2), y_2 the shadow price of constraint (3), etc. The dual problem of the DLP model can be defined as follows:

Min $C^T y_1 + E[D]^T y_{2,\ldots,n}$

s.t. $y_1 + A^T y_{2,\ldots,n} \geq f$

$y \geq 0$

In this formulation, the difference is made between the first decision variable from the vector y and the other decision variables. This is because we are especially interested in the value of y_1 because this tells us how much one place of capacity in the airplane is worth. In other words, y_1 is a representation of the bidprice for a ticket.

If we take the dual of our example, we get:

Min $40y_1 + 40y_2 + 15y_3 + 10y_4 + 5y_5$

s.t. $y_1 + y_2 \geq 100$

$y_1 + y_3 \geq 150$

$y_1 + y_3 \geq 200$

$y_1 + y_5 \geq 250$

$y_i \geq 0$ i=1,…,5

When we solve this dual model, we find an optimal solution of €6500, where $y_1 = 100, y_2, = 0, y_3 = 50, y_4 = 100, y_5 = 150$.

This information can now be used to compute the price of a ticket over time. This model can be adjusted continually over time to calculate the ticket prices when the capacity or the expected demand changes.

When we assume that the customers of the cheaper classes arrive before the customers of the more expensive classes and that the customers arrive approximately, the course of the price will look like in figure 3.

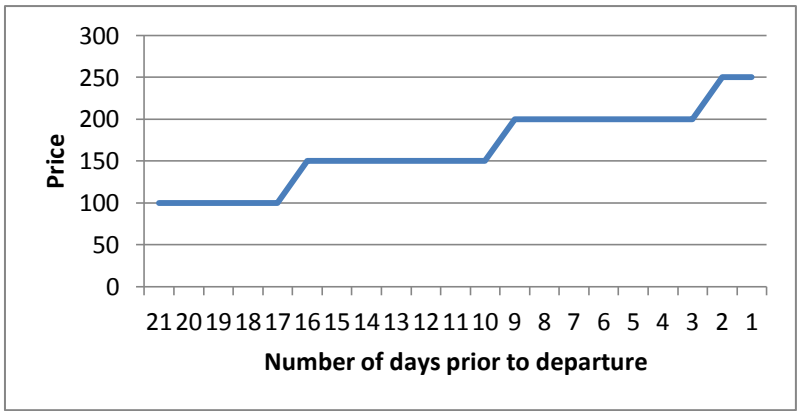

Figure 3

Some airlines also change the prices of their tickets because external factors, such as USD currency fluctuations and a change in the price of oil, etc. We could add some noise to simulate the change in these factors. If we do so, the price of a ticket would look like in figure 4.

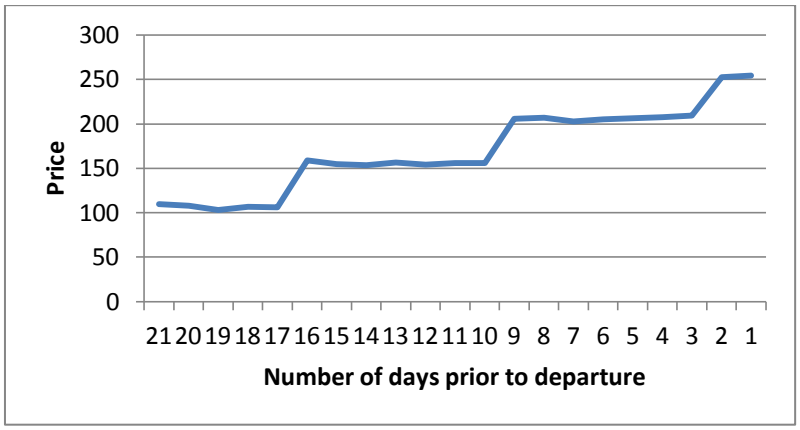

Figure 4

It is hard to predict the small price drops of this kind. But, this is not the only kind of price drop that we are looking for in this research. It is also possible that a lower booking class is opened once more in a later period of time. The reason for this is that the arrivals do *not* arrive approximately uniform. Airlines use the remaining capacity and a daily updated demand forecast to make sure that the flight will end up full. It is a waste to fly with seats unsold, because the costs will remain roughly the same independent of the amount of people that are on the flight. This means that every seat sold is extra

revenue and thus extra profit. Bid prices are used to keep adjusting prices in a way that the airplane will end up just full at the moment it will leave. If it is full before that, you miss the very expensive tickets from people that decide last-minute that they have to go on that flight (for instance because of an important business meeting). If it is not full, you miss easy revenue.

Airlines have absolute freedom to do what they want with the classes. Because of this, it is possible to open en close booking classes with ease. This is possible because the classes correspond with a type of contract that has certain constraints, but the seats and the actual flight will remain the same. Some types of constraints can be: the possibility of cancellation, a minimum stay between the outbound and return flight, or the possibility of changing the flight to another day or time. It does not matter what kind ticket you bought with any combination of these constraints, the moment you enter the airplane everything will be the same.

Paper discussion

Multiple researches have been done on this subject. One of them is the basis of a company called Farecast, that has been bought by Microsoft's Bing/travel. In this chapter, the paper from that research will be discussed.

The paper is called *To Buy or Not to Buy: Mining Airfare Data to Minimize Ticket Purchase Price*[8]. Three questions are answered:

1. What is the behavior of airline ticket prices over time?
2. What data mining methods are able to detect patterns in price data?
3. Can Web price tracking coupled with data mining save consumers money in practice?

1.
An interesting discovery that they make is that you can divide the airlines into two different categories. The major airlines and the price-fighters. Pricing policies are different for the two categories but seem to be consistent for airlines within the same category. The prices for tickets at the major airlines are high and will often change. The fares at the price-fighter category are lower and more balanced.
2.
Five different data mining methods were examined: Rule learning (Ripper rule learning system), Q-learning, Time Series, stacked generalization and some hand crafted rules. The best performing method was that of stacked generalization. In this method they combined the results of Ripper, Q-learning, and time series.
3.
To investigate whether one of their methods could be used in practice, simulation was used. They simulated arriving passengers and let their algorithms decide whether they should buy a ticket immediately or buy a ticket later. This process is repeated until the passenger buys a ticket or when there is no capacity left in economy class. To punish that last option, the algorithm will then buy a business class ticket. In this way, the passenger is always able to go on that flight.

Results
The best performing algorithm was the stacked generalization method, which they call HAMLET. The HAMLET algorithm achieved 61,8% of the total possible amount of savings. On average, if every passenger would follow the suggestions from HAMLET 4,4% would be saved.

Discussion

The results from this research are remarkable, if we take into account the lack of information about the number of seats available and only 41 days of data prior to departure. This research is similar to the research described in the paper on some aspects. The notable differences are:

- The research was done for flights in America, while this research focuses on flights mainly from Europe.
- They have 21 days of data from the period prior to departure, whereas in this research there are 28 days of data prior to departure.
- They got two routes (from Los Angeles to Boston and from Seattle to Washington DC), in this research 22 routes are used, all departing from Amsterdam.
- They have 6 airlines and record the lowest fare. In this paper 17 airlines are used and all of the prices are stored, not only the minimum. The initial idea was to start with only KLM data,

[8] Can be downloaded for free at http://knight.cis.temple.edu/~yates//papers/hamlet-kdd03.pdf

but when it turned out that the amount of data in the dataset from KLM was not sufficient, data from the other carriers was also used.

- They used stacked generalization, with Ripper, Q-learning, and time series. This research also uses a combined method, with a combination of time series forecast and k-NN.
- In the dataset used in this research, planes did not seem to get full more than 12 hours prior to departure. For this reason the business class punishment that was used in the paper is not implemented in this research.

A remarkable point is that the paper does not describe on what data the model is tested. It is not clear how the results are achieved. It might be possible that they did not use the combination of a trainings- and test set. Therefore, in this research, the model was also tested on the *training set* to be able to make better comparisons between the two researches.

Examining the data

Data preprocessing

The price of a flight is varying not only in time but also per destination. For instance the average price of a roundtrip ticket from Amsterdam to Zurich is €365 whereas the average price of a roundtrip ticket from Amsterdam to Istanbul is €559. This difference makes it troublesome to compare ticket prices from different destinations. For this reason the prices were firstly normalized. This normalization happened by dividing every price by the average over all price observations for that particular flight.

The normalized values can be used to compute statistics over all flights. In figure 5 there is a figure about the average normalized price and in figure 6 there is a figure about the standard deviation of the normalized values.

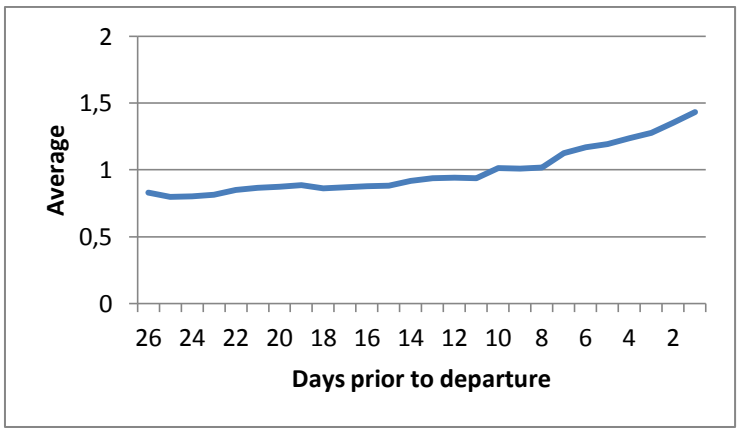

Figure 5

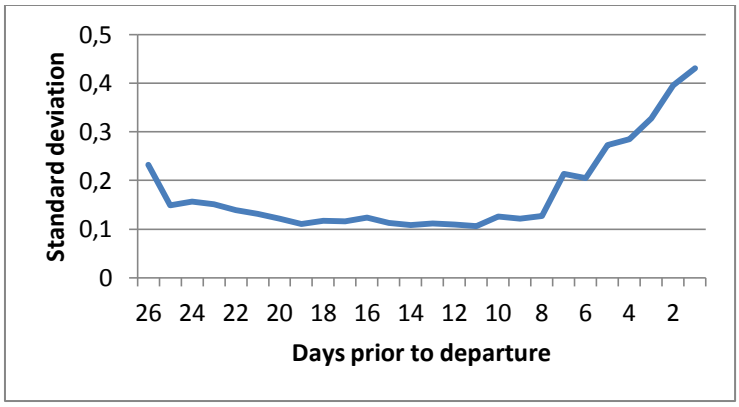

Figure 6

In the graph about the average, an upward trend is clearly visible. The graph about the standard deviation tells us that at the days before departure, many different scenarios are possible. The high standard deviation at the end represents that some tickets rise in price because the flight is almost full, and others will go to low very low prices because the flight is expected not leave with capacity left. While training the model, it seemed hard to get a high accuracy on the last 6 days, probably because of the large differences. Another issue is that if the model is wrong and it tells the customer

to wait and the price will rise, the losses are much larger at the end than compared with the days before. Of course, this also works the other way around; but we can see from the graph with the normalized average that the price is more likely to rise.

With this information the model was modified to not take the last few days before departure into account. It was found that if the model would always output a 'buy immediately' at the last days, more profit could be made. Ignoring the last *six* days was found locally optimal, which could be explained when looking at the rise of both standard deviation and average in the two graphs.

Training

To explain how the model was built and improved, an example will be used. In table 1 there is a list of fictional prices and the corresponding amount of days prior to departure. In figure 7, a visual representation of those prices are given.

Amount of days prior to departure	Price
8	50
7	56
6	56
5	60
4	70
3	50
2	50
1	60
0	80

Table 1

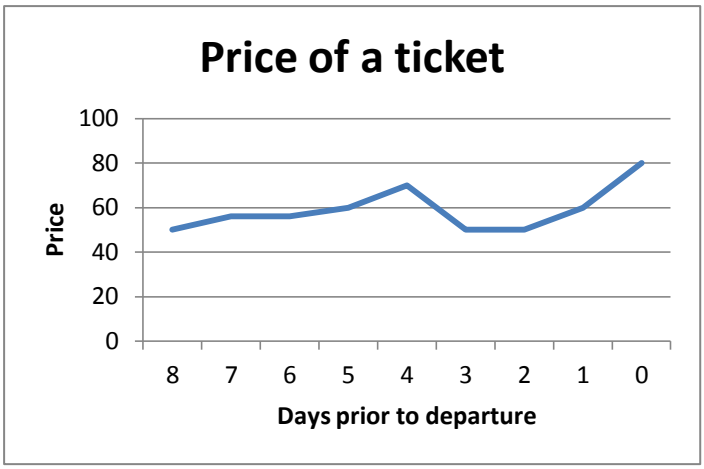

Figure 7

If an arrival occurs between 8 and 4 days prior to departure, it is optimal to wait because of the lower price at 3 days prior to departure. After 3 days prior to departure, the price is only rising. Thus, if an arrival occurs in that interval, it is optimal to buy a ticket immediately.

For training purposes, the model is presented with one line of information and a label value. The information on each line are values of predefined parameters, referred to as *explanatory variables*. The label value is a zero/one variable that contains a zero if buying is the optimal advice, and contains a one if waiting is the optimal advice. The label value is referred to as the *explained variable.* If we go back to our example, some lines of trainingsdata could be:

Price	Airline	Destination	Amount of days prior to departure	Wait = 1, Buy = 0
50	KLM	Berlin	8	1
60	KLM	Berlin	1	0

Table 2

Wait or buy – the explained variable

In this paper, a model is discussed that predicts whether to wait or to buy a ticket immediately. It is therefore important that a definition of waiting or buying is given and how the optimal value is determined when training a model.

The optimal time of buying would be when the ticket price is at a minimum. It is not possible to guide every user of the model towards this minimum because some customers arrive after the moment where the price hits a minimum. A better way to define the optimal buying time is the minimum price after the arrival of a potential customer. Although this has some practical issues. For instance, if the advice is to keep waiting until just before departure, not all customers might want to wait that long, taking the small risk that the flight will end up full before a ticket is purchased. Or if the model

tells the customer to wait, while seeing the price rise every day, one might become impatient. In this research, it is decided to limit the amount of days in which the price should get lower than the price that is observed on arrival.

Therefore, the explained variable was set to a predicted decrease in price within four days. But also small variations were used while testing and training. For instance, there has been experimented with a predicted price drop within seven days. Furthermore, an explained variable was used that only took significant price drops into account. Only if a price would be predicted to drop more than five percent, the explained variable would be classified as a 'wait'. Also other percentages were used to test for results. In the end, the best results came with a four day wait with no constraint on the amount of the price drop.

Explanatory variables

One part of the research was to find explanatory variables that could increase the accuracy when predicting whether to buy immediately or not.

the following explanatory variables were used in the final model:

The number of days prior to departure
Description: This variable represents the number of days that remain until the flight leaves. The underlying assumption is that a trend exists that is not dependent on the date and time of the year but only relative to the moment of departure.
Representation: integer value.

5 different variables all using an 7-day period average

A weighted average and a normal average was used over the past seven days for several explanatory variables. The weighted average used the following weights.

Amount of days before current	7	6	5	4	3	2	1
Weight	$^1/_{28}$	$^2/_{28}$	$^3/_{28}$	$^4/_{28}$	$^5/_{28}$	$^6/_{28}$	$^7/_{28}$

Table 3

1. **Direction of weighted average**
   Description: The first variable that uses the weighted average is the *direction* of the weighted average. This variable describes if the weighted average is higher or lower than the weighted average of the day before.
   Representation: a binomial value; it is equal to zero if it is lower than the weighted average of the day before, it is equal to one if it is higher.
2. **Current divided by unweighted**
   Description: The current price divided by the unweighted seven day average. Representation: real number.
3. **Excluded current divided by unweighted**

Description: Nearly the same as number two. The difference to number two is that the current price is not involved in the average. This means that the seven day average is from $t-1$ until $t-8$, whereas at number two the seven day average is from t until $t-7$.
Representation: real number.

4. **Comparison excluded current**
   Description: This variable is comparing the ratio from the variable at number three to the average of all ratios measured from this constant before.
   Representation: real number.
5. **Time series forecast.**
   Description: Uses a weighted average forecasting method.
   Representation: The variable is binomial, a zero indicating that the price is predicted to drop and a one indicating that the price is predicted to rise.

Max_1 and Max_2
Description: Both of these variables tell if the current price is higher or lower than the maximum of the previous seven days. Max_1 is represented as the ratio between the current price and the maximum of the previous seven days. The underlying assumption here is that the model can use the extra information from *how much* the price is higher or lower than the maximum of the previous seven days. Max_1 uses a binomial variable that only describes *if* the current price is higher or lower than the maximum of the previous seven days.
Representation: This variable is used as a ratio and as a binomial variable.

$Momentum_1$ and $Momentum_2$
Description: Defined as the weighted 7 day average divided by the normal 7 day average. If the weighted 7 day average is higher than the normal 7 day average that means that the current prices (with more weight) are higher than the older prices. For the same reasons as Max_1 and Max_2, the momentum variable has two different representations.
Representation: This variable is used as a ratio and as a binomial variable.

Stdev
Description: This variable represents the standard deviation of all the prices from that flight since the beginning of the data for that flight is available.
Representation: real number.

Destination
Description: The destination of the roundtrip.
Representation: categorical.

Airline
Description: The airline that is used for the flight.
Representation: categorical.

Testing

The model needs a warm-up period of seven days. In this research we assume that if a customer arrives, the price history of that flight-only the seven days before the arrival- is available. The model will give an advice on the time of arrival. If the model tells the customer to wait, it means that the model expects that a price drop will occur within four days. On the next day, the model expects an input of the new price of that day; after that, the advice is updated. This is repeated until the buy-advice occurs. Then, the procedure will stop and the model does not need to be updated anymore. Two scenarios are possible:

1) The current price is less than the initial price and the model is successful, the customer saved money.
2) The current price is equal to or more than the initial price and the model is not successful, meaning that price drops were not present or could not be exploited.

Simulation

To measure the performance of the model, the decision was made to take every possible scenario into account and compute the total amount of savings and losses. An important assumption for this to be realistic is that the chance of an arrival is equally likely on every day. For instance, if it is possible to book a ticket from 25 days before the flight leaves, 25 distinct passenger arrivals will be simulated.

Results

The model was first trained on a large part of the data, called the *training set*. The training set was chosen to be as big as possible, with still a reasonable amount left for testing. With a 106 flights total in our possession, initially 97 flights were used for training. These flights were used to tweak the parameters of our model. For more information about this dataset, see Appendix A under “Trial dataset”.

The program that was used to train and test the model is called Rapidminer from Rapid-I. The model was tested for results with a decision tree and a k-Nearest Neighbor algorithm. Eventually, using the previously mentioned dataset, the k-Nearest Neighbor algorithm was found to be the best for the model. K was set to 4 neighbors. This 4-Nearest Neighbor is used for the results in the rest of this paper.

The output from Rapidminer with information about the model. A ‘0’ represents the ‘buy-advice’ and a ‘1’ represents the ‘wait-advice’.

	true 0	true 1	class precision
predicted 0	1146	184	86.17%
predicted 1	198	637	76.29%
class recall	85.27%	77.59%	

Table 4

accuracy: 82.36% +/- 3.14% (mikro: 82.36%)

This model was applied to the test set consisting of nine flights using Rapidminer. The results from this action were put into excel to be able to combine the advices with the prices of the tickets. The combination of the price and ticket were processed using Visual Basic for Applications to a simple *.txt file. The resulting file was used in a Java-written simulation program.

Output from Java file that does the simulation:

Flight number	Total savings/losses
1	$ 3,00
2	$ -49,98
3	$ -223,71
4	$ 145,00
5	$ 239,00
6	$ 65,71
7	$ 9,54
8	$ -25,00
9	$ -30,40
Total	$ 133,16

The results from the initial simulation look promising. Although the deviation seems to be high. To be sure that the profit at the end of the nine flights was not only luck, a more sophisticated combination between training and testing has been made. For more information about these datasets, see the three training- and test sets in Appendix A.

Rapidminer outcomes:

On test set1

accuracy: 69.86% +/- 4.81% (mikro: 69.86%)

	true 0	true 1	class precision
pred. 0	598	215	73.55%
pred. 1	201	366	64.55%
class recall	74.84%	62.99%	

Table 5

On test set2

accuracy: 69.06% +/- 4.15% (mikro: 69.06%)

	true 0	true 1	class precision
pred. 0	561	205	73.24%
pred. 1	222	392	63.84%
class recall	71.65%	65.66%	

Table 6

On test set3

accuracy: 69.70% +/- 4.14% (mikro: 69.70%)

	true 0	true 1	class precision
pred. 0	568	200	73.96%
pred. 1	209	373	64.09%
class recall	73.10%	65.10%	

Table 7

Simulation outcomes

	Test set 1	Test set 2	Test set 3
Total savings/losses	$ -15.722,99	$ -4.575,58	$ -6.680,77
Percentage of profit	46%	27%	38%
Total passengers simulated	139	135	133
Average profit/loss per person	$ -113,12	$ -33,89	$ -50,23

Table 8

We can see in table 8 that the model does not perform well on the test sets. The model made a loss in every one of the test sets. To be able to compare this research with the research that is described in the section “Paper discussion”, the model was also used on the training set. The results from that are in table 9.

	Training set 1	Training set 2	Training set 3
Total savings/losses	$ 12.662,03	$ 11.935,83	$ 11.935,83
Percentage of profit	74%	70%	70%
Total passengers simulated	562	616	616
Average profit/loss per person	$ 22,53	$ 19,38	$ 19,38

Table 9

Discussion

As we can see in the paper discussed in the section “Paper discussion”, equivalent researches have been done with success. However, in this research it was not managed to make a model that would make a profit in a high percentage of the cases. Reasons for this are mainly the time limit of this research and the limited dataset that was available. The results of this research suggest that a model can be built that helps the customers in their quest of buying a cheap airline ticket.

If this subject will be further investigated, the author suggests the use of a more extended dataset of flights. It would be a good idea to first try a couple of routes only and build a model on only that part, before extending it to a broader use. The data used in this research was for only one arrival date. For a coming research the author would suggest using a different number of arrival dates combined with data of multiple flights per arrival date.

Instead of fixing the departure date and varying the purchase date, also equivalent models can be used to gain more insight in the process of fixing the purchase date and varying the time of purchase. For more information about this subject, see the chapter “Other part of Revenue Management“.

If this type of research will be more successful, airlines might respond to it by changing their pricing strategies. We can already see subtle changes in the strategy that is most likely because of these types of research. For more information about this subject, see the chapter “Response of airlines”.

Other part of Revenue Management

In this research we focus on fixing the flight, and varying the day of purchase. Another option is to fix the time of purchase and vary the departure date. In practice, this is already done by customers that want to go on a holiday, but do not mind going a few days earlier or later. For an individual it can be hard to find a cheap ticket because the airlines will focus on selling you an expensive one. Tools are already available to help the customer to make the process more insightful. For example Bing / Travel is already capable of making the graphs as in figure 8.

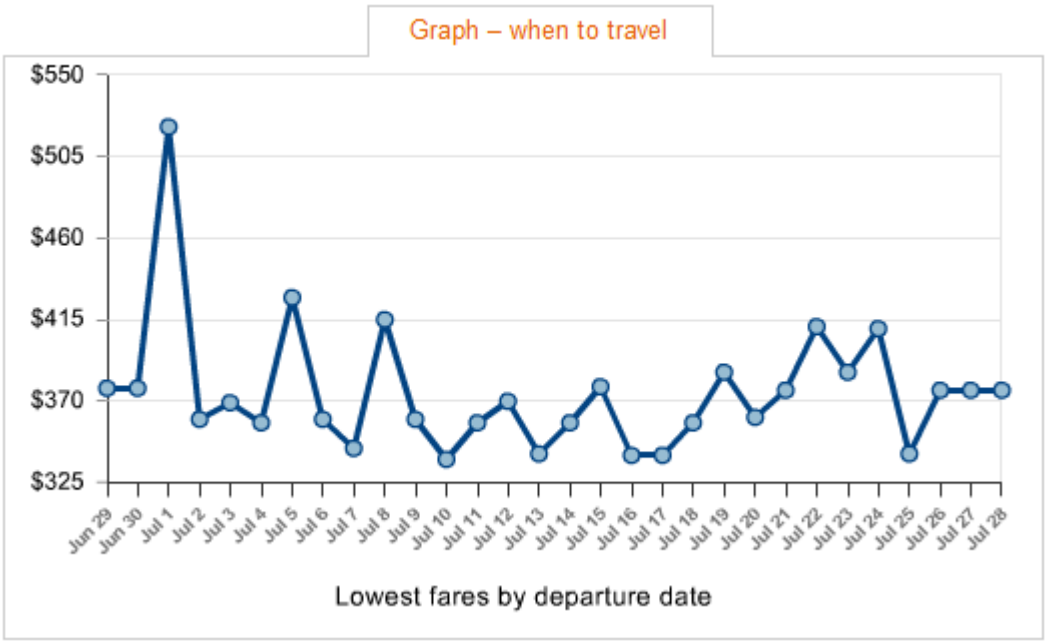

Figure 8

This subject was not covered in this paper because this type of research focuses more on seasonality. Seasonality is a topic that was not able to be covered because of the lack of yearly data. A couple years of data for the same departure date is necessary to train a model that predicts such values.

Response of airlines

Imagine that from now on, everybody would use a price predictor to decide whether to buy an airline ticket immediately or to wait. This would have a major effect on the airlines. Not only the revenue would decrease, furthermore the demand for tickets would be more clustered. If demand is clustered it is harder to determine the right prices. The information when buying early is essential for airlines when using Revenue Management. You do not know anything about people willing to buy a ticket but waiting for the right moment. You would prefer knowing that they want to buy a ticket sooner.

Airlines will have to find a solution to this 'problem'. One solution could be that an airline uses the information from a site like Bing/travel as extra input to price the tickets. For example, if Bing/travel advices the customer to wait, you could decide to lower the price, making it more attractive to buy now. Or if Bing advices the customer to buy now, you could also increase the price, hoping that one would still buy the ticket because Bing tells them to.

It is visible that airlines are already reacting towards the changes on this field. Some airlines offer *refund policies*. These policies allow customers to get a refund if the price of a ticket drops below the price that they paid for that ticket. In this way, airlines want to encourage people to buy early, not having to be afraid of a sudden price drop.

Moreover, one airline –the Dutch KLM- started with a "Time to think"-option[9]. When you buy such an option, you pay a certain price (around €10 to €15) for an option that will allow you to buy the ticket at any day in a period for up to 14 days. When you buy the ticket you can buy it for the same price as is stated in the option with the same fare conditions. If the price is lower than the price that is stated in the option, you can chose not to exercise it and buy the ticket without the option.

[9] http://www.klm.com/travel/it_en/plan_and_book/booking/booking_options/tttoptionita.htm.

Appendix A

Total dataset

Destination	Number of flights
Zurich	2
London	4
Oslo	8
Los Angeles	2
Bombay	1
Johannesburg	1
New York	9
Paris	8
Rio de Janeiro	1
Paramaribo	1
Athens	1
Madrid	6
Toronto	5
Helsinki	7
Prague	9
Warsaw	6
Vienna	6
Istanbul	6
Berlin	7
Dublin	6
Rome	8
Hong Kong	2
Sum	**106**

Trial dataset

Destination	Number of flights in training set	Number of flights in test set
Zurich	1	1
London	4	
Oslo	7	1
Los Angeles	2	
Bombay	1	
Johannesburg	1	
New York	8	1
Paris	7	1
Rio de Janeiro	1	
Paramaribo	1	
Athens	1	
Madrid	5	1
Toronto	5	
Helsinki	6	1
Prague	9	
Warsaw	5	1
Vienna	6	
Istanbul	6	
Berlin	6	1
Dublin	6	
Rome	7	1
Hong Kong	2	
Sum	**97**	**9**

Training and testingset 1

Destination	Number of flights in training set	Number of flights in test set
Zurich	1	1
London	2	2
Oslo	4	4
Los Angeles	0	1
Bombay	0	1
Johannesburg	0	1
New York	4	5
Paris	8	
Rio de Janeiro	1	
Paramaribo	1	
Athens	1	
Madrid	4	2
Toronto	5	
Helsinki	7	
Prague	9	
Warsaw	4	2
Vienna	6	
Istanbul	6	
Berlin	7	
Dublin	6	
Rome	8	
Hong Kong	2	
Sum	**86**	**19**

Training and testingset 2

Destination	Number of flights in training set	Number of flights in test set
Zurich	2	
London	4	
Oslo	5	3
Los Angeles	2	
Bombay	1	
Johannesburg	1	
New York	9	
Paris	4	4
Rio de Janeiro	0	1
Paramaribo	0	1
Athens	0	1
Madrid	4	2
Toronto	3	2
Helsinki	4	3
Prague	9	
Warsaw	6	
Vienna	6	
Istanbul	6	
Berlin	7	
Dublin	6	
Rome	6	2
Hong Kong	2	
Sum	87	19

Training and testingset 3

Destination	Number of flights in training set	Number of flights in test set
Zurich	2	
London	4	
Oslo	8	
Los Angeles	2	
Bombay	1	
Johannesburg	1	
New York	9	
Paris	8	
Rio de Janeiro	1	
Paramaribo	1	
Athens	1	
Madrid	6	
Toronto	5	
Helsinki	7	
Prague	5	4
Warsaw	4	2
Vienna	4	2
Istanbul	4	2
Berlin	3	4
Dublin	3	3
Rome	4	4
Hong Kong	1	
Sum	84	21