Global Journal of Information Technology: Emerging Technologies

Volume 7, Issue 3, (2017) 128-148

www.gjit.eu

Airline revenue management via data mining

Cuneyt Bahadir, Information Technologies MSc Program, Bahcesehir University, Istanbul 34349, Turkey.
Adem Karahoca*, Software Engineering Department, Bahcesehir University, Istanbul 34349, Turkey.

Suggested Citation:
Bahadir, C. & Karahoca, A. (2016). Airline revenue management via data mining. *Global Journal on Technology*. *7*(3), 128-148.

Received June 02, 2017; revised October 01, 2017; accepted November 09, 2017.
Selection and peer review under responsibility of Prof. Dr. Dogan Ibrahim, Near East University, North Cyprus.

Abstract

Revenue maximisation has been of paramount interest in the airline industry during the past few decades, and numerous studies have been reported, aiming at robust analyses. Principal analysis techniques in most of these studies include computational-based prediction algorithms that are used for a given dataset. In this study, airline specific data, which consists of cabin class passenger data, cabin class supplied capacity data, distance of flights, season, year–month data and revenue data, are analysed using various prediction algorithms. Consistencies and accuracies of different algorithms are compared and reported.

Keywords: Airline industry, airline revenue data, prediction algorithms, Weka, Bayesian network, sequential minimal optimisation, support vector machines, multilayer perceptron, radial basis function network.

* ADDRESS FOR CORRESPONDENCE: **Adem Karahoca,** Software Engineering Department, Bahcesehir University, Istanbul 34349, Turkey.
E-mail address: adem.karahoca@eng.bau.edu.tr / Tel.: +90-212-381-0560

1. Introduction

An airline is a company that provides air transport services for travelling passengers and freight. Airlines lease or own their aircraft with which to supply these services and may form partnerships or alliances with other airlines for mutual benefit. Generally, airline companies are recognised with an air operating certificate or license issued by a governmental aviation body.

The airline industry was first commenced with Zeppelin in Germany in the early 1900s and continued with aircraft. At the beginning of the industry, one aircraft could only transport few people. At that time, it was a big gain to achieve the transportation of one person in a relatively short time through a relatively long distance. However, as time passed, people's needs changed, and accordingly their expectations. People wanted to reach far destinations in a short time. As the technology evolved, aircraft got bigger, faster and with the ability to fly in high altitudes. Hence, the airline industry began to bloom. This also affected many industries such as material (composite) technologies, jet engine technologies, etc.

During the second quarter of the 20th century, the airline industry had improved in the USA. With the advantage of being a continent country, in the USA, domestic flights were the thrust force for the industry.

According to a research done by Airbus Company in 2012, the centre of gravity of the airline industry had crossed the Atlantic Ocean towards the end of the 2000s. In 2010, it reached the middle part of the Mediterranean Sea. Due to the economic increase in the Far East, it is continuing to move towards the east, and it is expected to reach Cyprus by the year 2030. This is also a big chance for our country, Turkey, to increase its pie in the global airline industry.

The airline industry is a highly competitive industry. According to the market survey published by Boeing Company in June 2013, new technologies and intense competitions continue to push the airline yields downward. The yield was around 16 cent per seat-mile in the 1970s, whereas it was almost 7 cent per seat-mile in 2010.

On the other hand, in a research done by Airbus Company in 2012, they used the International Civil Aviation Organisation's data in their research, showing that the airline industry has been increasing steadily since 1970. This indicates that the industry doubles every 15 years, and grows by 5% each year. Also, it will keep growing at 5.1% between 2011 and 2021, and 4.4% between 2021 and 2031.

However, the same research depicts that although there are good growing rates in the industry, it is really hard to make profit due to increasing costs and high competition.

The industry has two main big costs: first, aviation fuel whose price is directly dependent on global oil prices and second, aircraft-stuff costs, which is also dependent on the global market. From this perspective, the only manageable parameter for profit is revenue.

Due to high competition and basically not manageable cost parameters, revenue management had been the most important topic in the airline industry for many years. In order to maximise the revenue, reservation class methodology had been introduced many years ago. The basic aim is to optimise the itinerary fare according to demand. Moreover, origin/destination (O/D) model has been come out due to complex itineraries. This model is aimed to fetch passengers' true origin and true destination for the defined airport or city. For instance, on a flight from Atatürk International Airport in İstanbul to Frankfurt International Airport, there might be a number of passengers with different origins and destinations. One might be travelling from İstanbul to Frankfurt, while another may be connecting in Frankfurt to fly to Berlin. Another passenger may go to Hamburg from Tel Aviv through İstanbul and Frankfurt International Airports, respectively. In this simple example, one single flight İstanbul-Frankfurt serves the demand for at least three different origins/destinations: İstanbul-Frankfurt; İstanbul-Berlin and Tel Aviv-Hamburg. If we measure only the number of passengers who are travelling from İstanbul to Frankfurt, then this will not reflect the number of passengers who have İstanbul origin and Frankfurt destination. Because of the other O/D passengers who use İstanbul to

Frankfurt flight as a connecting flight have not been taken into account. In contrast, in an O/D model, the true travel volume is estimated based on the passengers' complete journey by the given two cities. The airline industry can understand the current need and predict the future demand with this approach.

In this study, airline specific data, which consist of cabin class passenger data, cabin class supplied capacity data, distance of flights, season, year–month data and revenue data, are tested in terms of prediction algorithms. It is expected to determine whether these algorithms are convenient or not.

In the airline industry, there are two common types of business models, which are also called carrier models: 1) flag carrier and 2) low cost carrier.

In the flag carrier model, the crucial point is the hub phenomenon. A hub is the area that the airline collects its passengers from and distributes them through the hub.

On the other hand, in the low cost carrier model, there is no hub centre and the carrier sets its business plan to transport its passengers directly from the origin to the destination.

In this study, independent of the business model, it is assumed that aircraft execute one-way direct flight. Therefore, there is no need to analyse all the segments in a possible itinerary.

Moreover, in the airline business, the product is mainly the seat that the airline provides. Product differentiation can be occurred up on seat (cabin) class which provides different service types in the aircraft during the flight. Commonly, there are four cabin classes, which are first class, business class, comfort class and economy class. In this study's dataset, there are only two cabin classes, business and economy classes. This viewpoint is the main constraint of the study.

In Section 2, we give a brief literature review of airline revenue management models and prediction algorithms in Weka. In Section 3, we describe the dataset that we will use in the algorithms. Also, the predictions algorithms are discussed. In Section 4, the outcomes of prediction algorithms are stated. Then, we analyse the outcomes. Lastly, we compare the given prediction algorithms for our dataset and we conclude our findings in Section 5.

2. Literature Review

In this section, we give a brief literature review of airline revenue management models and prediction algorithms in Weka.

2.1. Airline Revenue Management Models

Every mercantile establishment is set up based on one main objective, which is 'profit'. Therefore, each establishment looks for how to make profit and increase its profit ratio. There is basically one approach to increase profit, which is 'increase revenue, decrease cost'. The airline industry is one of the industries on which most studies have been done during the past decades, in order to analyze revenue maximization. In the airline industry short-term costs are mostly fixed and the variable cost per passenger is small. Thus, it is enough to research on booking policies which maximize revenues.

Pak and Piersma (2002) stated that revenue management has become an important discipline because of the improvements on decision support systems and computer science. He also stated that revenue management has become more important in the airline industry compared to other industries.

According to Morales and Wang (2010), revenue management increases the revenue of a company as long as demand management is achieved. In other words, a revenue management system should consider the possible cancel bookings and no-show values.

Cao, Ding, He & Zhang (2010) described that airlines' over-booking rights have effect on the profit in terms of maximization. If airlines achieve to foresee no-show quantity, they can easily reduce the

number of involuntary denied boarding and the number of spoiled seats, which result in increase in revenue.

Doganis (2006) stated that the profitability of an airline depends on the relation of three variables, which are the unit cost, unit revenue and load factors achieved. Hence, he claims that costs, fares and load factors must be adjusted to produce more profit. However; this process is dynamic and complex, and so, it is difficult due to pricing instability in the airline industry.

For a current booking request, it might be fulfilled at the current price or it might be held in anticipation of a higher one in the future. This situation forms a large part of the revenue management problem in the airline industry. In the case of single leg and single product, a solution for this problem can be found by using the expected marginal seat revenue approach, which was studied by Belobaba in 1987.

The development of the revenue management system has progressed from simple leg control through segment control, and finally to origin–destination control. Thus, many extensions of the marginal seat revenue approach have been investigated by researchers (Dunleavy & Phillips, 2009).

According to Mcgill and Van Ryzin (1999), in most situations it is enough to seek booking policies that maximise revenues in order to maximise profit for the objective in revenue management.

In mercantile establishments, forecasting is an important part of the planning process. Especially, it is much more crucial in airline revenue management. The reason for this is the booking limits determine airline profits and this has a direct effect on forecasts.

There are several models for demand distributions in the literature. One of the first of these models is given by Beckmann and Bobkowski (1958). They give a description of the statistical models on passenger booking, cancellation and no-show behavior. The authors compare Poisson, negative binomial and gamma models of the total passenger arrivals, and they state evidences of reasonable fit for the gamma distribution to airline data.

In the book of Taneja (1978) traditional regression techniques are described for aggregate airline forecasting. Later, Sa (1998) analysed regression experiments with airline data. Sa (1998) states that the performance of revenue management system can be improved using regression techniques by comparing time series analysis or historical averages. Also, the effect of promotional sear sales on forecasting and revenue management was discussed by Botimer (1997).

2.2. Prediction Algorithms in Weka

Data mining is the ability to fetch information from very large-scale data. Nowadays, data are getting bigger and bigger. Therefore, in order to analyse huge data and to catch up with meaningful outcomes, some necessary tools are needed. Machine learning tools are the most commonly used method to process huge data and figure out results in data mining processes. Weka is the one of the popular machine learning software that is widely used in the academic world. It was developed by researchers from Waikato University. There are many pre-defined methods and classifiers in Weka such as BayesNet, radial basis function network (RBFNetwork), sequential minimal optimisation (SMO) and support vector machine (SVM). These methods can be easily applied to given data and so many researcher use Weka to perform data analysis in many different topics. Next we will give some of these studies applied to the airline industry.

Mack *et al.* (2011) describe a tree augmented naive Bayesian classifier that forms the basis for systematically extending aircraft diagnosis reference models using flight data from systems operating with and without faults.

Mukherjee *et al.* (2014) studied data-mining techniques to identify similar days in the National Airspace System in terms of the cause and location of historically implemented ground delay programmes. They study a modified k-means clustering algorithm which was applied to all days from

2010 through 2012. They identified 45 national-level daily clusters that represent unique combinations of historically implemented ground delay programmes.

Gallo and Kepto (2014) examined the relationship between expected meteorological conditions as specified by TAF reports and actual ground conditions as specified by hourly METAR reports for Chicago-Midway (MDW) and Seattle-Tacoma (SEA) airports for the period September–December 2011. Chi-squared analyses indicated that although the relationship between TAF and METAR at each airport was statistically significant, the corresponding Kappa agreement coefficients showed that this relationship was nearly twice as strong at MDW as at SEA.

Schumann *et al.,* (2011) use the AutoBayes method to generate customised data analysis algorithms that process large sets of aircraft radar track data in order to estimate parameters and uncertainties.

3. Data and Methods

In this section, we describe our dataset and the methods used to test our dataset. A dataset covers a set of data items and is basically composed of a two-dimensional spreadsheet or database table. Weka implements datasets by constructing instances that consist of many attributes.

3.1. Dataset

The dataset used in this study is composed of airline market information which shows some historical information about a specific market through 3 years and 36 months.

The dataset has 2,596 instances and 8 attributes, which also means 2,596 rows and 8 columns.

The attributes are as follows:

i. YearMonth
ii. Season
iii. Km (Distance)
iv. ArzC
v. ArzY
vi. PaxC
vii. PaxY
viii. Revenue

i. YearMonth: This attribute shows the year and month information in which the instance occurs. There are 36 distinct YearMonth data, which starts with (min value) year 2011 and month 01 and ends with 2013–12. Data type is 'date'. In Weka, date type is used in 'YYYY-MM-DD' format. Here, only year and month are used, therefore the format is 'YYYY-MM'.
ii. Season: This attribute shows the season information in which the instance occurs. There are 4 distinct seasons, 'Winter, Spring, Summer, Fall'. Data type is 'nominal'. There are 643 Winter, 643 Spring, 653 Summer, 657 Fall items in the dataset.
iii. Km: Km value shows the distance of a flight from its origin to the destination. There are 71 distinct Km values, which also indicate distinct markets. It can be assumed that each distinct Km shows a market from hub X to destination Y_n, where n is $1 \leq n \leq 71$. The minimum value of this attribute is 914 km and the maximum value is 6,448 km.
iv. ArzC: This attribute shows supplied business cabin class seat amount in a defined year–month period for a market. This value is calculated and declared before the flight. There are 1,020 distinct values in the dataset. The minimum value is 36 seats and the maximum value is 7,916 seats.
v. ArzY: This attribute shows the supplied economy cabin class seat amount in a defined year–month period for a market. This value is calculated and declared before the flight. There are 2,421 distinct values in the dataset. The minimum value is 180 seats and the maximum value is 58,376 seats.

vi. PaxC: This attribute shows the passenger amount that is flown in the business class in a defined year–month period for a market. There are 1,230 distinct values in the dataset. The minimum value is five seats and the maximum value is 5,407 seats.
vii. PaxY: This attribute shows the passenger amount that is flown in the economy class in a defined year–month period for a market. There are 2,456 distinct values in the dataset. The minimum value is 95 seats and the maximum value is 49,585 seats.
viii. Revenue: This attribute shows the revenue amount that is gained from flown flights in a defined year–month period for a market. There are 2,596 distinct values in the dataset. The minimum value is 18,731 and the maximum value is 15,151,561.

3.1.1. Discretisation

As many attributes in the dataset are composed of numeric values, it is hard to handle them while classifying. Therefore, an instance filter that discretises a range of numeric attributes into nominal attributes will be used in the dataset. Once discretisation is completed, the data form will be changed to nominal.

After discretisation is complete, it is seen that all the attributes form into 10 bins. It is also seen how many instances are in each bin. The colours in each column depict the revenue interval value for the corresponding bin in Figures 1–6.

In Figure 1, discretisation results for Km are given. This attribute depicts the distances between the origin and destination of a flight.

Except for bin ten, the number of instances in each bin ranges between 141 and 420. The fifth bin has the maximum number of instances, which is 420. In the figure, each bin has ten different colours. These colours show the ratio of each class.

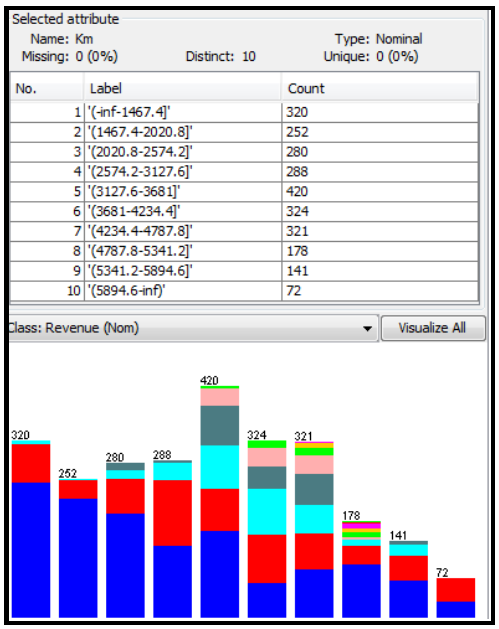

Figure 1. Discretisation results for Km

The discretisation results for ArzC are given in Figure 2. The number of instances mostly decreases as the bin label increases. The number of instances in the first six bins dominates the total number of instances. Each bin shows the number of available business class seat capacity in flights for specific markets in a month. For instance, the second bin shows that there are 643 flights whose total available business seat capacities are between 824 and 1,612.

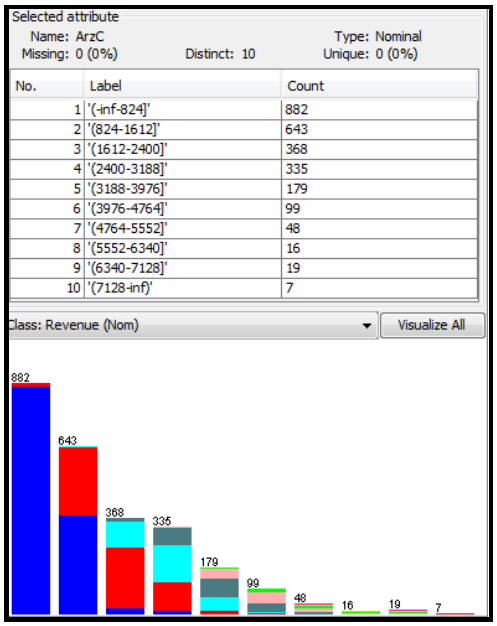

Figure 2. Discretisation results for ArzC

Discretisation results for ArzY are given in Figures 3. Similar to ArzC, the number of instances mostly decreases as the bin label increases. Each bin shows the number of available economy class seat capacity in flights for the specific market in a month. If you consider the fourth bin in Figure 3, there are 234 flights whose total available business seat capacities are between 17,638 and 23,458. The ratio of the number of classes in each bin is easily comparable by just comparing the colours in each bin.

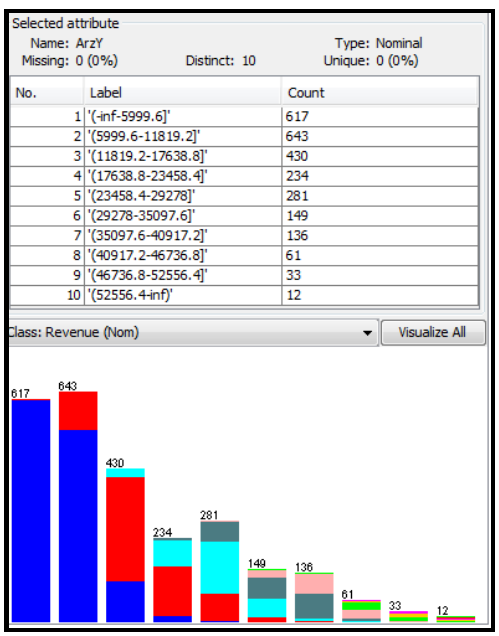

Figure 3. Discretisation results for ArzY

The discretisation results are tabulated in Figure 4 for PaxC. For PaxC, the number of instances in bins 1–3 dominates the total number of instances. This means that the number of business passengers up to 1,625 is most likely to be seen in a month for a specific flight market.

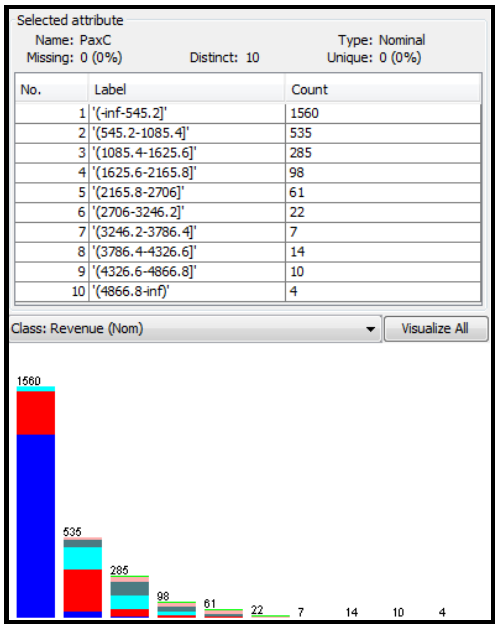

Figure 4. Discretisation results for PaxC

The discretisation results for PaxY are tabulated in Figure 5. For PaxY, approximately 85 % of the total number of instances occurs in the first five bins. This means that the number of business passengers up to 24,840 is most likely to be seen in a month for a specific flight market. In the first bin, almost all instances were classified in the first revenue class. In the second and third bins, there are two and three revenue classes, respectively.

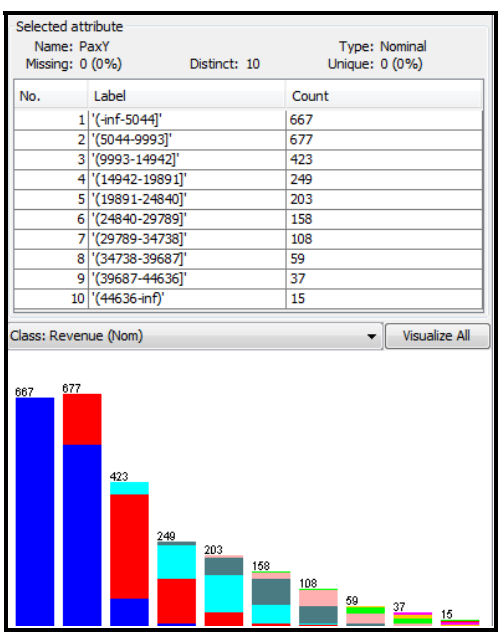

Figure 5. Discretisation results for PaxY

The discretisation results are listed for revenue in Figure 6. The attribute revenue is the class column. In other words, classification is executed according to the revenue. Having completed discretisation, we have ten bins which show the revenue intervals.

Classes in this attribute are not of exact value. They consist of revenue intervals which have 1,500,000 incremental pitch. The first class has the most number of instances, which is 1,285. For the total number of instances perspective, the first five bins dominate the total number of instances. In the figure, each bin has a different colour. Each colour symbolises the different class value.

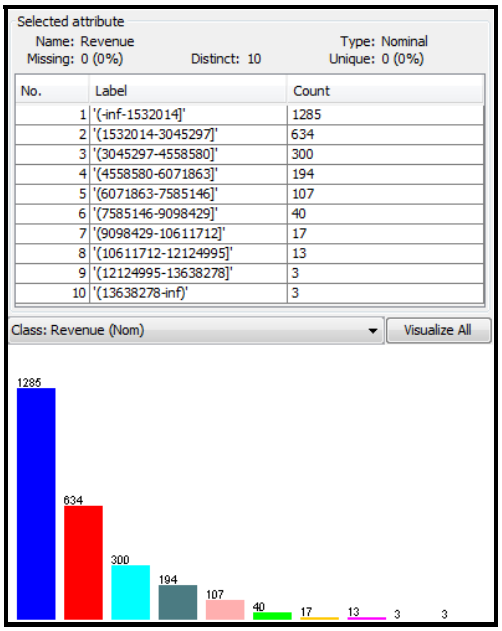

Figure 6. Discretisation results for revenue

3.2. Methods

Machine learning is a sub-field of computer science and it is developed from studies pattern of recognition and computational learning theory in artificial intelligence. Machine learning deals with the structure of algorithms that convert data into information and make predictions from them. These algorithms do not apply static programme instructions directly. They build models from the given input data to obtain a prediction or decision.

Computational statistics is a discipline that studies the design of algorithms for statistical method implementation on computers. Also, it has a strong connection to mathematical optimisation. Machine learning is closely related to computational statistics. Machine learning is applied in different computing tasks in which the designing and programming algorithm is not feasible. The main application of machine learnings is spam filtering, recognition algorithms and search engines. Even though machine learning and data mining seem to be in the same field, machine learning mostly focuses on data analysis.

In this section, we briefly describe the methods that we test our dataset with.

3.2.1. BayesNet

The properties of Bayesian networks (Bayesnet) were first summarised by Pearl (1988) and Neapolitan (1989). They also established Bayesnet as a field of study. Pearl (1988) especially emphasised the following two points: the often subjective nature of the input information and the reliance on Bayes' conditioning as the basis for updating information.

Bayesian networks are powerful tools to represent and inference the knowledge and the reasoning mechanism. Bayesnet is applied in many different fields such as bioinformatics, medicine, engineering and risk analysis. It also has many applications in financial and marketing informatics.

Bayesnet represents events as conditional probabilities, which involves random variables. Bayesnet can compute the values of a subset of variables by using the given values of another subset of variables.

Bayesnet has several advantages for data analysis. We mention two of them here. First, Bayesian networks can handle incomplete datasets without any extra computations or calculations. If there are two variables that are highly anti-correlated, then most of the prediction algorithms need the inputs for every possible case. However, Bayesnet works fine for these types of datasets as well.

Second, Bayesnet allows researchers to learn more about the relationships between variables. This feature helps to understand the problem easily and to make better prediction with the current data.

We apply Bayesnet to our dataset and interpret the results in Section 4.1.

3.2.2. SMO

The SMO algorithm was proposed by Platt (1988) for training a support vector classifier. Training an SVM requires the solution of a very large quadratic programming optimisation problem. Platt's algorithm breaks this large quadratic problem into a series of smallest possible quadratic problems, and then these small problems are solved analytically.

The applications of SMO are closely related to applications of SVM, which are given in the following subsection.

In Weka, the implementation of SMO implementation globally replaces all missing values. It also transforms nominal attributes into binary attributes, and it normalises all attributes.

The main advantage of the SMO algorithm is that the amount of memory required is linear in the training set size. This allows SMO to handle very large training sets. Another advantage is that matrix computations are avoided. On real-world sparse datasets, SMO can be more than a thousand times faster than the chunking algorithms (Platt, 1988).

We apply SMO to our dataset and interpret the results in Section 4.2.

3.2.3. Support Vector Machine

SVM was first introduced by Boser *et al.* (1992). The current standard version was proposed by Cortes and Vapnik (1995). They proposed this algorithm especially for two-group classification problems. Previously, SVM was implemented for the restricted case where the training data is without errors. They extended the implementation of SVM for non-separable training data.

Application of SVM includes text and hypertext categorisation, classification of images, and classification of proteins in medical sciences. SVM becomes popular because it is especially very successful in handwritten digit recognition. SVM is mostly regarded as an important example for kernel methods, which are one of the key areas in machine learning. Nowadays, SVM is regarded as one of the first choice for classification problems.

An SVM model is a representation of examples as points in space. The aim is that the examples of the separate categories are divided by a clear gap, which is as wide as possible. Then, it maps new examples into the same space and predicts the category that these new examples belong to, assuring of the gap.

One of the main advantages of SVM is that it has good performance even with a large number of inputs. On the other hand, SVM has some limitations on the speed of running time and size in both training and test data. Also, SVM has a complex algorithmic structure, and it requires an extensive memory capacity.

In Section 4.3, SVM is applied to our dataset and then interpreted.

3.2.4. Multilayer Perceptron

Multilayer perceptron (MLP) was first proposed by Rosenblatt (1961). He simplified artificial neural networks problem by considering a particular type of neural network called perceptron. For perceptrons, the neurons are distributed in layers with feed-forward connection. They also discovered the perceptron learning rule with its corresponding convergence theorem, which could be used for training of perceptrons.

MLP is a popular machine learning method, especially for its application in speech recognition and image recognition. The real-world applications include data compression, financial prediction, speech and handwritten character recognition. For details, see Wasserman and Schwartz (1988). An MLP is a model that maps the input data onto a set of appropriate outputs. An MLP contains multiple layers of nodes where each layer is connected to the next layer. MLP uses the supervised learning technique, which is called back propagation for training the network. Details can be found in Rosenblatt (1961).

An MLP model is well suited to problems that people are good at solving, but for which computers are not. The main advantages of the MLP model are adaptive learning and self-organization, which is creating its own representation of the information it receives during learning time and real-time operation.

We apply MLP to our dataset and interpret the results in Section 4.4.

3.2.5. Radial Basis Function Networks

A RBFNetwork is a model that uses radial basis functions. It was first formulated by Broomhead and Lowe in 1988. They discussed the relationship between learning in adaptive layer networks and the fitting of data with high-dimensional surfaces. From this, they obtained a generalisation in terms of interpolation between known data points, and they also obtained a rational approach to the theory of such networks.

RBFNetworks can be used to solve a set of common problems. These problems include function approximation, times series prediction and system control.

RBFNetworks have three layers in most cases. These are an input layer, a hidden layer and a linear output layer. The output of the network is a scalar function of the input data.

RBFNetworks are good at modelling non-linear data and can be trained in one stage. It also learns the given application quickly. Another advantage of the RBFNetwork is that it is useful for solving problems where the input data are corrupted with additive noise.

We apply RBFNetwork to our dataset and interpret the results in Section 4.1.

4. Findings

In this section, we give the outcomes of the algorithms for our dataset. The dataset used in this study is composed of airline market information, which gives some old information about a specific market through 3 years and 36 months. The dataset has 2,596 instances and 8 attributes, which also means 2,596 rows and 8 columns. Finding the predictive relationship in the dataset is our main objective. In order to do that, we basically classified our dataset according to 'revenue', by using the classification method, or in other words, machine learning algorithms. Because our dataset is composed of numeric values, first we convert them into nominal values by discretizing in order to achieve classification. At the end of discretization process, we got ten revenue intervals composed of 1,500,000 incremental pitches instead of exact value classes. At the end of classification process, all instances were classified according to their revenue interval results.

Weka is software in the data mining area that has been used by researchers in different fields. It basically uses machine learning algorithms in order to make predictive analyses. It was developed in Java platform and is able to be used in general operating systems such as Windows, Mac OS and Linux. The outcomes given in this section are obtained by applying the mentioned algorithms in Section 3 to our dataset in Weka.

For each algorithm, we give the summary of the outcome, detailed accuracy by class and confusion matrix via Weka.

As mentioned above, finding the predictive relationship in the dataset is our main objective. The dataset which is used find that predictive relationship is called a training set. All used machine learning algorithms in this study used the same training data in order to compare the abilities of the algorithms. Meanwhile, test data is the dataset that is used to measure the predictive abilities of algorithms.

In this study, we feed machine learning algorithms with training data and then they produce classifiers. After that, we test each classifier with the test data and get evaluation results.

In order to evaluate the result, it is a crucial point to have different test data and training data. However, if the data are limited, the whole dataset can be divided into two parts. It would be better if the number of instances in training data is much larger than the number of instances in test data.

In Weka, there are two options for the limited data scenario: 'percentage split' and 'cross-validation'.

In the percentage split option you may split your data as the test and the training data by defining a certain percentage rate such as 90% for training and the remaining 10% for test data. Then you can run the algorithm. Weka chooses random 90% and 10% parts from data and executes the algorithm. However, if you repeat it again in the same ratio, you get exactly the same result, because Weka ran the same random logic for the same ratios in order to guarantee the repeatability.

In the cross-validation option, Weka divided our dataset into ten pieces. Ten is a variable that can be set by the user. We took nine of them to use in training and the last pieces for testing, and then we took another nine pieces for training and the remaining one piece for testing. Weka did this cycle ten times, using a different segment for testing each time. At the end, Weka averages the results.

4.1. BayesNet

The summary of the results of the BayesNet method obtained via Weka is as follows:

Table 1. Summary of BayesNet classifier

Correctly classified instances	1974	76.04%
Incorrectly classified instances	622	23.96%
Kappa statistic	0.65	
Mean absolute error	0.05	
Root mean squared error	0.19	
Relative absolute error	38.4%	
Root relative squared error	73.1%	
Total number of instances	2596	

1,974 of the total of 2,596 instances are classified correctly. This corresponds to 76% of the total instances. The corresponding Kappa statistic is 0.65.

Correctly classified instances show the accuracy of the model.

The kappa statistic measures the agreement of prediction with the true class. For instance; 1.0 signifies complete agreement. In other words, the kappa statistic is an indicator of correlation coefficient. If it gets the value 0, it means the lack of any relation. If it approaches to 1, there is a very strong statistical relation between the class label and attributes of instances.

The mean absolute error (MAE) calculates the average absolute value of the error in a set of instances. So it omits the sign of the errors. For continuous variables, it may be considered as the corresponding accuracy. The MAE can be formulated as MAE = $\sum\left(\left| f\left(xi\right) - yi\right|\right) / N$, where $y = f(x)$ is built by a regression algorithm to predict a numeric instance in a model and N represents the number of instances. Thus, MAE is the average over the absolute values of the differences between actual values and predicted values. Here, MAE is evaluated with a linear function and hence all differences are equally weighted.

Similar to MAE, the root mean squared error (RMSE) calculates the average squares of the error in a set of instances. The corresponding formula can be given in the following form: RMSE = $\sum\left(f\left(xi\right) - yi\right)^2 / N$. This formula can be interpreted as first squaring the differences between actual values and predicted values, and then taking the average over the sample. The RMSE is then evaluated by RMSE = $\sqrt{\mathrm{MSE}}$. In MSE and RMSE, the formulas include a quadratic function and hence large errors are high weighted. Thus, RMSE is used for especially the models where large errors are undesired.

However, the four lines in Table 1 depict that errors are not particularly useful in a classification task, because they are measures used to assess performance when the task is numeric prediction.

Table 2. BayesNet detailed accuracy by class

TP rate	FP rate	Precision	Recall	F-measure	ROC area	Class
0.914	0.07	0.927	0.914	0.921	0.979	'(-inf-1532014]'
0.672	0.088	0.712	0.672	0.692	0.913	'(1532014-3045297]'
0.61	0.072	0.524	0.61	0.564	0.929	'(3045297-4558580]'
0.552	0.035	0.557	0.552	0.554	0.956	'(4558580-6071863]'
0.542	0.02	0.532	0.542	0.537	0.971	'(6071863-7585146]'
0.325	0.009	0.351	0.325	0.338	0.976	'(7585146-9098429]'
0.353	0.007	0.25	0.353	0.293	0.992	'(9098429-10611712]'
0.385	0.005	0.278	0.385	0.323	0.993	'(10611712-12124995]'
0.333	0	0.5	0.333	0.4	0.999	'(12124995-13638278]'
0	0	0	0	0	0.99	'(13638278-inf)'
0.76	0.068	0.766	0.76	0.763	0.955	←Weighted avg.

True positive (TP) rate is the rate of correctly classified instances as a given class. As given in Table 2, TP rates are high for the classes where the revenue is relatively low. False positive (FP) Rate is the rate of incorrectly classified instances as a given class. Similar to the TP rates, FP rates are low for the classes where the revenue relatively high.

Precision is the proportion of instances that are correctly classified in a class divided by the total instances classified as that class. The values in the precision column behave similar to the values that correspond to the TP and FP rates.

Recall is the proportion of instances that are correctly classified in a class divided by the total instances in that class. This is equivalent to the TP rate.

F-measure is a combined measure for precision and recall and it can be calculated as 2 * Precision * Recall / (Precision + Recall). It can be also expressed as the harmonic mean of recall and precision.

The receiver operating characteristic (ROC) curve illustrates the performance of a classifier method. The ROC curve is plotted on a graph where the *x*-axis is the FP rate and the *y*-axis is the TP rate. The area under a ROC curve can be a maximum of 1. If the ROC area is close to 1, it means that the classifier method is working successfully. The results of the method are worthless if the ROC area is close to 0.5. For instance, for the second class (the revenue is between 1,532,014 and 304,5297), the ROC area is relatively smaller than the seventh class (the revenue is between 9098429 and 10611712) because the FP rate of the seventh class is significantly lesser than the FP rate of the second class, even though the TP rate of the second class is greater than the TP rate of the seventh class. However, the ROC areas of all the classes are greater than 0.9, which means that the method is working well for all the classes.

A confusion matrix visualizes the performance of the algorithm. Each column of confusion matrix represents the number of instances in the predicted class, and each row represents the number of instances in an actual class. The sum of the values in the main diagonal of a confusion matrix gives the number of correctly classified instances. The sum of the remaining values gives the number of incorrectly classified instances. TP rate, FP rate, and precision can be evaluated by the confusion matrix with the following formulas.

TP rate can be evaluated for class *x* with the following formula:

The value in cell (*x*, *x*)/the sum of row *x*.

FP rate can be evaluated for class *x* with the following formula:

(The sum of column *x* – the value in cell (*x*, *x*))/(the number of instances – the sum of row *x*).

Precision can be evaluated for class *x* with the following formula:

The value in cell (*x*, *x*)/the sum of column *x*.

4.2. SMO

TP rate is the rate of correctly classified instances as a given class. TP rates get the highest value for the classes where the revenue is relatively low. Even this looks similar to the values in Bayesnet; it is different because as the revenue increases, TP lowers until a point that it is in the fourth class, and then it increases in the next class. Through the following next three classes it decreases and again increases. Finally, it gets its lowest value for TP rate in the eighth class.

FP rate is the rate of incorrectly classified instances as a given class. Similar to the TP rates, FP rates get their lowest value in the first class and then move in a sinusoidal path like the values in TP rates.

The largest precision value is 0.95, which is for the first class and the smallest non-zero value is 0.1.

The smallest value for ROC area is 0.911, which shows that the values are very close to the maximum value of 1. Thus, this classifier method works successfully for our dataset.

4.3. SVM

2,005 instances of the total of 2,596 instances are classified correctly. This corresponds to 77% of the total instances. The corresponding kappa statistic is 0.66.

TP rates get the highest values for the classes where the revenue is relatively low. However, it dramatically lowers while the class number increases. The TP rate reaches 0 value in the seventh class.

Correspondingly, the FP rate has non-zero values for the first five classes, and it has a value of 0 for the last five classes.

The highest ROC value occurs in the first class, which is 0.933. Then it dramatically falls, after the first class it never reaches a 0.9 value. Instead, it converges to a value of 0.5 in the sixth class. Thus, this classifier method does not work successfully for our dataset.

4.4. MLP

2,176 instances of the total of 2,596 instances are classified correctly. This corresponds to 84% of the total instances. The corresponding kappa statistic is 0.76.

TP rates get the highest value for the classes in which the revenue is relatively low. Similar to the TP rates, FP rates are relatively high as the revenue values are relatively low.

The largest precision value is 0.95, which is for the first class and the smallest non-zero value is 0.25. The precision value decreases gradually as the revenue increases.

The weighted average value for ROC area is 0.97. Even though this value is good, the last two values are below 0.9. Except for the last two classes, the results show meaningful ROC values that are bigger than 0.95.

4.5. RBFNetwork

2,037 instances of the total 2,596 instances are classified correctly. This corresponds to 78% of the total instances. The corresponding kappa statistic is 0.68.

TP rates are high for the classes where the revenue is relatively low. The incorrectly classified instances, FP rates, are similar to the TP rates. FP rates are low for the classes where the revenue is relatively high.

The values in the precision column behave similar to the values that correspond to the TP and FP rates.

Even if the weighted average value for ROC area seems good because of the value 0.958, the results are not as good as their average values. For the first six classes, the ROC area is greater than 0.9. On the other hand, the remaining areas are smaller.

5. Discussion

In the airline industry, profit is the most important issue for the sustainability of a company. Although this issue looks the same as in other industries, there is one difference, which is that the costs in the airline industry are not as manageable as in other industries. Therefore, revenue management becomes the most crucial point for the airline industry in order to make profit. From this point of view, for many years a number of research works and surveys have been done and are still continuing.

In this section, we discuss our findings pertaining to the outcomes of the algorithms stated in Section 4.

In this study, machine learning algorithms were used to classify the airline revenue-related data. There can be many attributes that affect revenue. In this study, only the available seat values in business and economy classes, the number of passenger in both these classes, distance of flight, year–month values and season values have been considered as attributes.

In the research, by using machine learning algorithms, a reasonable classification method was tried to be achieved in order to evaluate the revenue interval of attributes. We use the term ‘interval’, because our dataset is composed of numeric values. We obtained nominal values after applying the discretisation method. Our class value is not a specific unique value. Instead, our class values are composed of revenue intervals. It has ten intervals, starting from 0. Each interval has a 1,500,000 incremental pitch.

In order to evaluate the classification method, we need training data and test data. Instead of preparing two separate datasets due to limited instances, we used the cross-validation option in Weka. In this option, Weka divided the dataset into ten pieces. Then, it took nine of them to use as

training data and the last piece for testing. Weka did the same cycle for every different nine pieces and the remaining one piece. At the end it returns their averages.

We check the accuracy of the classification model by sorting out the number of correctly classified instances.

When the classification algorithms were considered, SMO seems to be the best ability to classify the given instance. It is the most accurate model. It achieved to classify 85.4% of instances correctly, which means 2,217 instances over 2,596. MLP is the second best for classification ability. It reached the value of 83.8%. The value of 2,176 over 2,596 is quite close to SMO. The remaining are RBFNetwork, SVM and BayesNet, whose correctly classified instances rates are 78.5%, 77.2% and 76%, respectively. The classification abilities of all the algorithms are higher than 75%, which can be considered as good.

The kappa statistic measures the agreement of prediction with the true class. For instance; 1.0 signifies complete agreement. In other words, the kappa statistic is an indicator of the correlation coefficient. If it gets the value of 0, it means the lack of any relation. If it approaches to 1, it indicates very strong statistical relation between the class label and attributes of instances. Corresponding to values of the correctly classified instances, the values for kappa statistic are in the same pattern; SMO has the highest value whereas BayesNet has the lowest value. The algorithm SMO has the most statistical relation. All the information pertaining to classification can be seen in in Tables 3 and 4.

Table 3. Classifying outcomes of algorithms

	BayesNet	SMO	SVM	MLP	RBFNetwork
Correctly classified instances	1974	2217	2005	2176	2037
Incorrectly classified instances	622	379	591	420	559
Kappa statistic	0.6472	0.7831	0.6585	0.76	0.6801
Mean absolute error	0.0519	0.1608	0.0455	0.0351	0.0541
Root mean squared error	0.190	0.274	0.213	0.168	0.175
Relative absolute error	38.40%	119.0%	33.71%	26.02%	40.06%
Root relative squared error	73.08%	105.3%	82.17%	64.72%	67.26%
Total number of instances	2596	2596	2596	2596	2596

Table 4. Classifying percentage outcomes of algorithms

RATES	BayesNet	SMO	SVM	MLP	RBFNetwork
Correctly classified instances	76.0%	85.4%	77.2%	83.8%	78.5%
Incorrectly classified instances	24.0%	14.6%	22.8%	16.2%	21.5%

Table 5. Weighted averages of detailed accuracy outcomes

Weighted avg.	TP rate	FP rate	Precision	Recall	F-measure	ROC area
BayesNet	0.76	0.068	0.766	0.76	0.763	0.955
SMO	0.854	0.044	0.853	0.854	0.853	0.953
SVM	0.772	0.071	0.759	0.772	0.762	0.85
MLP	0.838	0.048	0.837	0.838	0.838	0.967
RBFNetwork	0.785	0.068	0.784	0.785	0.784	0.958

In Table 5, weighted averages of detailed outcomes are tabulated. The best TP rate value is obtained from SMO algorithms. BayesNet gives the worst value for TP rate, which is 0.76.

The best value for ROC area is 0.967 and it is obtained from MLP. 0.85 is the smallest value for ROC area and its algorithm is SVM. Except for the value of SVM, all the remaining ROC values are greater than 0.95.

Considering the confusion matrices in the previous section, a large number of instances occur in the first five classes corresponding to the revenue attribute. Therefore, we will analyse the detailed accuracies of each algorithm for the first five classes.

For the first class, even though the highest TP rate value seems to belong to SMO, the value for SVM is very close to SMO. All the TP rate values are greater than 0.90 in the first class. Looking at the ROC area values, all of them are higher than 0.97, which is good. MLP and RBFNetwork have the same values, which is 0.983. The detailed accuracy outcomes for the first class are tabulated in Table 6.

Table 6. Detailed accuracy outcomes for the first class

'(-inf-1532014]'	TP rate	FP rate	Precision	Recall	F-measure	ROC area
BayesNet	0.914	0.07	0.927	0.914	0.921	0.979
SMO	0.955	0.05	0.95	0.955	0.952	0.97
SVM	0.94	0.074	0.926	0.94	0.933	0.933
MLP	0.949	0.053	0.946	0.949	0.948	0.983
RBFNetwork	0.93	0.074	0.925	0.93	0.927	0.983

For the second class, TP rate values for SMO and MLP seem better than the others. It seems that classification has been done better by these two algorithms in the second class. SVM has the lowest FP rate value, which is 0.1. The value for the ROC area of the MLP algorithm is the best, which is 0.95. Except for the value 0.832 of the SVM algorithms, all the remaining ROC values are higher than 0.90. The detailed accuracy outcomes for the second class are tabulated in Table 6.

In this study, some data mining methods were used in order to make revenue prediction.

The dataset used in this study is composed of airline market information, which shows some old information about a specific market through 3 years and 36 months. The data mining research was conducted for 2,596 instances, each of which is compounds of eight attributes. They are YearMonth, Season, Km, ArzC, ArzY, PaxC, PaxY and Revenue. The Weka program was run for this dataset by selecting Bayesian Network (BayesNet), SMO, SVM, MLP and RBFNetwork classification models, respectively, and when the obtained outcomes are considered, it is seen that the SMO classification algorithm is the one with the highest accuracy. MLP has the second best accuracy. However, if we look at the weighted average point of view, due to the lowest values in FP rates, MLP has slightly greater ROC values than SMO. It can be said that for our dataset the SMO and MLP algorithms are the most convenient algorithms.

6. Conclusion

In this study some data mining methods were applied over the dataset that belongs to the airline industry. This research aimed to predict revenue by using classification methods. There can be many attributes that have an effect on revenue. The dataset used in this study is only composed of YearMonth, Season, Km, ArzC, ArzY, PaxC, PaxY and Revenue attributes. Because the dataset's attributes are numeric values, we first convert them into nominal values by applying discretisation.

Due to the limited amount of data, cross-fold method was used in order to produce input data. The input data has two subsets: training data and test data. Using the cross-fold method, Weka divided the whole dataset into ten pieces and then used nine of them as training and the remaining part as test

data. Having completed this cycle ten times for every different nine pieces and the last one piece, it produced output.

Our output is composed of revenue classes which have a 1,500,000 incremental pitch. Also, we can see how many instances were correctly classified in which classes.

The Weka program was run for this dataset by selecting BayesNet, SMO, SVM, MLP and RBFNetwork classification models, respectively. The comprehensive comparison of the outcomes of classification models was presented in Section 5. SMO classification algorithm is the one with the highest accuracy. It achieved to classify the 85.4% of instances correctly, which means 2,217 instances over 2,596. MLP has the second best accuracy with the rate of 83.8%. The remaining are RBFNetwork, SVM and BayesNet; whose correctly classified instances rates are 78.5%, 77.2% and 76%, respectively.

References

Beckmann, M. J., & Bobkowski, F. (1958). Airline demand: An analysis of some frequency distributions. *Naval Res. Logistics*, *5*(1), 43–51.

Boser, B. E., Guyon, I. M., & Vapnik, V. N. (1992). A training algorithm for optimal margin classifiers. *In Proceedings of the fifth annual workshop on Computational learning theory* (pp. 144-152).

Botimer,T.C. (1997). *Select ideas on forecasting with sales relative to bucketing and 'seasonality.'* Houston, TX: Continental Airlines, Inc.

Broomhead, D. S., & Lowe, D. (1988). Multivariable functional interpolation and adaptive networks. *Complex Syst., 2,* 321–355.

Cao, R. Z., Ding, W., He, X. Y., & Zhang, H. (2010, July). Data mining techniques to improve no-show forecasting. In *Service Operations and Logistics and Informatics (SOLI), 2010 IEEE International Conference on* (pp. 40-45).

Cortes,C., & Vapnik, V. (1995). Support-vector networks. *Mach. Learn.*, *20*(3), 273-280.

Doganis, R. (2006). *The airline business*. UK: London.

Dunleavy, H., & Phillips, G. (2009). The future of airline revenue management. *J. Revenue Pricing Manag.*, *8*(4), 388–395

Gallo, M. A., & Kepto, M. (2014). The relationship between 2011 METAR and TAF data at Chicago-Midway and Seattle-Tacoma Airports. *Collegiate Aviation Rev.*, *32*(1), 18-26.

Grabbe, S., Sridhar, B., & Mukherjee, A. (2014). Clustering days and hours with similar airport traffic and weather conditions. *Journal of Aerospace Information Systems, 11*(11), 751-763.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA data mining software: An update. *ACM SIGKDD Explorations Newsletter, 11*(1), 10-18.

Mack, D. L., Biswas, G., Koutsoukos, X. D., Mylaraswamy, D., & Hadden, G. (2011). Deriving bayesian classifiers from flight data to enhance aircraft diagnosis models. In *Annual Conference of the Prognostics and Health Management Society.*

McGill, J. I, & Van Ryzin, G. J. (1999). Revenue management: Research overview and prospects. *Transportation Sci*., *33*(2), 233–256.

Morales, D. R., & Wang, J. (2010). Forecasting cancellation rates for services booking revenue management using data mining. *Eur. J. Oper. Res*., *202*(2), 554–562.

Neapolitan, R.E. (1989). *Probabilistic reasoning in expert systems: Theory and algorithms*. New York, NY: Wiley.

Pak, K., & Piersma, N. (2002). Overview of OR techniques for airline revenue management. *Statistica Neerlandica, 56*(4), 479-495.

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems*. San Francisco, CA: Morgan Kaufmann.

Platt, J. (1998). Fast training of support vector machines using sequential minimal optimization. In *Advances in kernel methods – Support vector learning*. Cambridge, MA: MIT Press.

Rosenblatt, F. (1961). *Principles of neurodynamics: Perceptrons and the theory of brain mechanisms*. Washington, DC: Spartan Books.

Sa, J. (1998). *Reservations forecasting in airline yield management* (Unpublished Master Thesis). Flight Transportation Laboratory, Massachusetts Institute of Technology, Cambridge, MA.

Taneja, N. K. (1978). *Airline traffic forecasting: Aregression analysis approach*. Lexington, MA: Lexington Books.

Theodore, C. B., & Belobaba, B. B. (1999). Airline pricing and fare product differentiation: A new theoretical framework. *J. Oper. Res. Society*, *50*(11), 1085–1097.

Wasserman, P. D., & Schwartz, T. (1998). Neural networks. II. What are they and why is everybody so interested in them now. *IEEE Expert*., *3*(1), 10–15.