THE REPUBLIC OF TURKEY
BAHCESEHIR UNIVERSITY

AIRLINE REVENUE MANAGEMENT VIA DATA MINING

Master's Thesis

CÜNEYT BAHADIR

ISTANBUL, 2015

THE REPUBLIC OF TURKEY
BAHCESEHIR UNIVERSITY

GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES
INFORMATION TECHNOLOGIES

AIRLINE REVENUE MANAGEMENT VIA DATA MINING

Master's Thesis

CÜNEYT BAHADIR

Supervisor: PROF. DR. ADEM KARAHOCA

İSTANBUL, 2015

THE REPUBLIC OF TURKEY
BAHCESEHIR UNIVERSITY

GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES
INFORMATION TECHNOLOGIES

Name of the thesis: Airline Revenue Management Via Data Mining
Name/Last Name of the Student: Cüneyt Bahadır
Date of the Defense of Thesis:

The thesis has been approved by the Graduate School of Natural and Applied Sciences.

Assoc. Prof. Dr. Nafiz ARICA
Graduate School Director
Signature

I certify that this thesis meets all the requirements as a thesis for the degree of Master of Arts.

Prof. Dr. Adem KARAHOCA
Program Coordinator
Signature

This is to certify that we have read this thesis and we find it fully adequate in scope, quality and content, as a thesis for the degree of Master of Arts.

Examining Comittee Members	Signature
Thesis Supervisor Prof. Dr. Adem KARAHOCA	----------
Member Prof. Dr. İbrahim PINAR	----------
Member Asst. Prof. Dilek KARAHOCA	----------

ACKNOWLEDGMENT

I would like to express my sincere gratitude both personally and professionally, to my thesis supervisor, Prof. Dr. Adem Karahoca, for his endless support in preparation of this study. His knowledge, patience and understanding made possible the successful completion of the thesis.

I would like to thank Bahçeşehir University and the Graduate School of Natural and Applied Sciences for providing me a beautiful academic environment.

Also, I am indebted to my parents for their self-sacrifice and patience that kept me altogether throughout my education lifetime. Most importantly, I would like to thank my wife for her endless support.

İstanbul, 2016 Cüneyt BAHADIR

ÖZET

VERİ MADENCİLİĞİYLE HAVAYOLU GELİR YÖNETİMİ

Cüneyt Bahadır

Bilgi Teknolojileri

Tez Danışmanı: Prof. Dr. Adem Karahoca

Aralık 2015, 56 Sayfa

Gelir maksimizasyonu havayolu endüstrisinde uzun yıllardır büyük bir ilgi ile araştırılan ve üzerinde de çokça çalışma yapılan bir konudur. Araştırma konuları genelde tanımlı bir veri kümesine bilgisayar tabanlı tahmin algoritmalarının uygulanmasıyla gerçekleşir. Bu tezde de havayolu endüstrine ait kabin bazlı yolcu sayısı, kabin bazlı arz edilen koltuk sayısı, uçuş mesafesi, sezon bilgisi, ay yıl bilgisi ve gelir muhtelif tahmin algoritmalarıyla analiz edilmiştir. Doğruluk ve tutarlılıkları mukayese edilerek raporlanmıştır.

Bölüm 2'de endüstrideki gelir yönetim modelleri ve kullanılmış tahmin algoritmaları hakkında özet literatür araştırması verilmiştir.

Bölüm 3'te algoritmalarda kullanılan veri kümesi tanıtılmıştır. Bununla birlikte, Bayesian Network, Sequential Minimal Optimization, Support Vector Machines, Multilayer Perceptron ve Radial Basis Function Network algoritmaları müzakere edilmiştir.

Bölüm 4'te tahmin algoritmalarının çıktıları analiz edilmiştir.

Son olarak, veri kümesi için seçilmiş tahmin algoritmalar mukayese edilmiş ve çıktılar üzerinde değerlendirme yapılmıştır.

Anahtar Kelimeler: Havayolu Endüstrisi, Hayayolu Gelir Datası, Tahmin Algoritmaları, Weka, Bayesian Network

ABSTRACT

AIRLINE REVENUE MANAGEMENT VIA DATA MINING

Cüneyt Bahadır

Information Technologies

Thesis Supervisor: Prof. Dr. Adem Karahoca

December 2015, 56 Pages

Revenue maximization has been of a paramount interest to Airline Industry during the last few decades and numerous studies have been reported aiming robust analyses. Principle analysis techniques in most of these studies include computational-based prediction algorithms that are used for a given data set. In this thesis, 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, is analyzed with various prediction algorithms. Consistencies and accuracies of different algorithms are compared and reported.

In Section 2, a brief literature review is given on airline revenue management models and on prediction algorithms that are used in airline industry via Weka.

In Section 3, the data set that is use in the algorithms is described. Also, the predictions algorithms, Bayesian Network, Sequential Minimal Optimization, Support Vector Machines, Multilayer Perceptron, and Radial Basis Function Network is discussed.

In Section 4, the outcomes of prediction algorithms are analyzed.

Lastly, selected prediction algorithms for the data set are compared and a conclusion on resulting outcome is given in Section 5.

Keywords: Airline Industry, Airline Revenue Data, Prediction Algorithms, Weka, Bayesian Network

CONTENTS

TABLES

FIGURES

ABBREVIATIONS/SYMBOLS

ArzC	:	Business cabin class seat amount
ArzY	:	Economy cabin class seat amount
BayesNet	:	Bayesian Network
FP Rate	:	False Positive Rate
ICAO	:	International Civil Aviation Organization
MLP	:	Multilayer Perceptron
PaxC	:	Passenger amount in the business class
PaxY	:	Passenger amount in the economy class
RBFNetwork	:	Radial Basis Function Network
ROC	:	Receiver Operating Characteristic
SMO	:	Sequential Minimal Optimization
SVM	:	Support Vector Machines
TP Rate	:	True Positive Rate

1. INTRODUCTION

An airline is a company that provides air transport services for traveling 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 recognized with an air operating certificate or license issued by a governmental aviation body.

Airline industry was first commenced with zeppelin in Germany in early 1900's and continued with aircrafts. 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 the time goes people's needs have been changed, and accordingly their expectations. People wanted to reach far destinations in a short time. As the technology evolves, aircraft getting bigger, faster, and ability to flow in high altitude is increased. And so on, airline industry has been bloomed. It also affected many industry such as material (composite) technologies, jet engine technologies, etc.

During the second quarter of 20^{th} century, airline industry had been improved in the USA. 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 center of gravity of airline industry had crossed the Atlantic Ocean towards the end of 2000's. In 2010, it reached middle part of Mediterranean Sea. Due to the increasing in economic in Far East, it is continuing to move towards east, and it is expected to reach Cyprus area in the year of 2030. This is also a big chance for our country, Turkey, to increase its pie from the global airline industry.

Airline industry is a highly competitive industry. According to market outlook which was 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 1970's, whereas it was almost 7 cent per seat-mile in 2010.

On the other hand, a research done by Airbus Company in 2012, also they used International Civil Aviation Organization's (ICAO) data in their research, shows that, airline industry has been increasing steadily since 1970. It indicates that the industry doubles every fifteen years, and grow 5 percent for each years. Also, it keeps growing 5.1 percent between 2011 and 2021, 4.4 percent between 2021 and 2031.

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

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

Due to high competition and basically not manageable cost parameter, revenue management had been the most important subject in airline industry for many years. In order to maximize the revenue, reservation class methodology had been introduced in many years ago. The basic aim is to optimize the itinerary fare according to demand. Moreover, origin/destination (O/D) model has been come out due to complex itineraries. This model is aiming to fetch passengers' true origin and true destination for defined airport or city. For instance, on a flight from Atatürk International Airport (AHL) in İstanbul to Frankfurt International Airport there might be number of passengers with different origins and destinations. One might be travelling from İstanbul to Frankfurt, another may be connecting in Frankfurt to fly to Berlin. And, another one going 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, 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 by the given two cities are estimated based on the passengers' complete journey. Airline industry can understand the current need and predict the future demand with this approach.

In this thesis, 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, will be tested in terms of prediction algorithms. It is expected to determine whether these algorithms are convenient or not.

In airline industry, there are two common types of business models which are also called carrier models:

1. Flag carrier,
2. Low cost carrier.

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

On the other hand, in the low cost carrier model, there is no hub center and the carrier sets its business plan to transport its passengers directly from the origin to the destination. In this thesis, independent of business model, it is assumed that aircrafts execute one way direct flight. Therefore, there is no need to analyze all segments in a possible itinerary.

Moreover, in airline business, the product is mainly the seat that 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 thesis' data set there are only two cabin classes which are business and economy classes.

In Section 2, we will give a brief literature review on airline revenue management models and prediction algorithms in Weka.

In Section 3, we will describe the data set that we will use in the algorithms. Also, the predictions algorithms will be discussed.
In Section 4, the outcomes of prediction algorithms will be stated. Then, we will analyze these outcomes.

Lastly, we will compare the given prediction algorithms for our data set and we will conclude our finding in Section 5.

2. LITERATURE REVIEW

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

2.1 AIRLINE REVENUE MANAGEMENT MODELS

Every mercantile establishment is set up base 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". Airline Industry is one of the industries that most of studies have been done during the last decades in order to analyze revenue maximization. In airline industry short term costs are mostly fixed and variable cost per passenger are small. Thus, it is enough to research on booking policies which maximize revenues.

Pak (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 airline industry among the other industries.

According to Morales and Wang (2010) revenue management increases the revenue of 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 et al. (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, the unit revenue, and the 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 in the future. This situation forms a big part of revenue management problem of 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 is studied by Belobaba in 1987.

The development of revenue management system has progressed from simple leg control through segment control, and finally to origin-destination control. Thus, many extensions of marginal seat revenue approach have been investigated by researchers. See Dun-leavy and Phillips, 2009 for details.

According to Mcgill and Van Ryzin (1999), in most situations it is enough to seek booking policies that maximize revenues in order to maximize 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 in 1958. They give description of statistical models on passenger booking, cancellation, and no-show behavior. The authors compare Poisson, Negative Bionamial, and Gamma models of 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 (1987) analyzes regression experiments with airline data. Sa states that the performance of revenue management system can be improved by using regression techniques with comparing time series analysis or historical averages. Also,

the effects of promotional sear sales on forecasting and revenue management is discussed by Botimer in 1997.

2.2 PREDICTION ALGORITHMS IN WEKA

Data mining is the ability to fetch information from very large scaled data. Nowadays data is getting bigger and bigger. Therefore, in order to analyze huge data and to catch up with a meaningful outcomes, some necessary tools are needed. Machine learning tools are the most commonly used method in order to process huge data and figure out results in data mining processes. Weka is the one of the popular machine learning software that widely used in academic world. It was developed by researchers from Waikato University, New Zeland. There are many pre-defined methods and classifiers in Weka such as BayesNet, RBFNetwork, SMO, SVM, etc. 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 ailrine 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. study data-mining techniques to identify similar days in the National Airspace System in terms of the cause and location of historically implemented ground delay programs. 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 programs.

Gallo and Kepto 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 square 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. use the AutoBayes method to generate customized 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 will describe our data set and also we will describe the methods that we used to test our data set.

A data set is made of a set of data items. It basically composed of two dimensional spreadsheet or database table. Weka implements data set by constructing instances which consist of many attributes.

3.1 DATA SET

The data set used in this thesis is composed of airline market information which shows some old information about a specific market through three years and 36 months.

The data set has 2596 instances and 8 attributes, which also means 2596 rows, 8 columns. 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 that the instance occurs. There are 36 distinct YearMonth datum, 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 that the instance occurs. There are 4 distinct seasons, which are "Winter, Spring, Summer, Fall". Data

type is “nominal”. There are 643 Winter, 643 Spring, 653 Summer, 657 Fall items in the data set.

iii. **Km:** Km value shows the distance of a flight from its origin to the destination. There are 71 distinct Km values, which also indicates distinct markets. It can be assumed that each distinct Km show 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 6448 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 1020 distinct values in the data set. The minimum value is 36 seats and the maximum value is 7916 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 2421 distinct values in the data set. The minimum value is 180 seats and the maximum value is 58376 seats.

vi. **PaxC:** This attribute shows the passenger amount that flown in the business class in a defined year-month period for a market. There are 1230 distinct values in the data set. The minimum value is 5 seats and the maximum value is 5407 seats.

vii. **PaxY:** This attribute shows the passenger amount that flown in the economy class in a defined year-month period for a market. There are 2456 distinct values in the data set. The minimum value is 95 seats and the maximum value is 49585 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 2596 distinct

values in the data set. The minimum value is 18731 and the maximum value is 15151561.

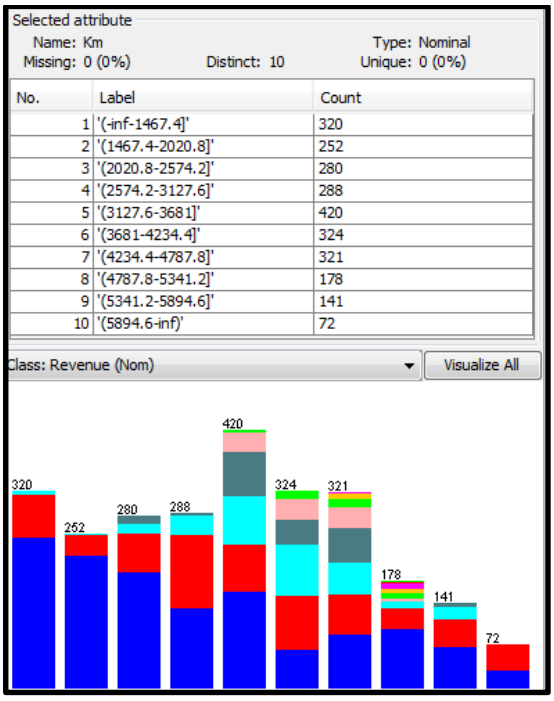
3.1.1 Discretization

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

After discretization is completed, it is seen that all attributes form into 10 bins. It is also seen that how many instances are in each bin. The colors in each column depicts the revenue interval value for corresponding bin in Figures 3.1 to 3.6.

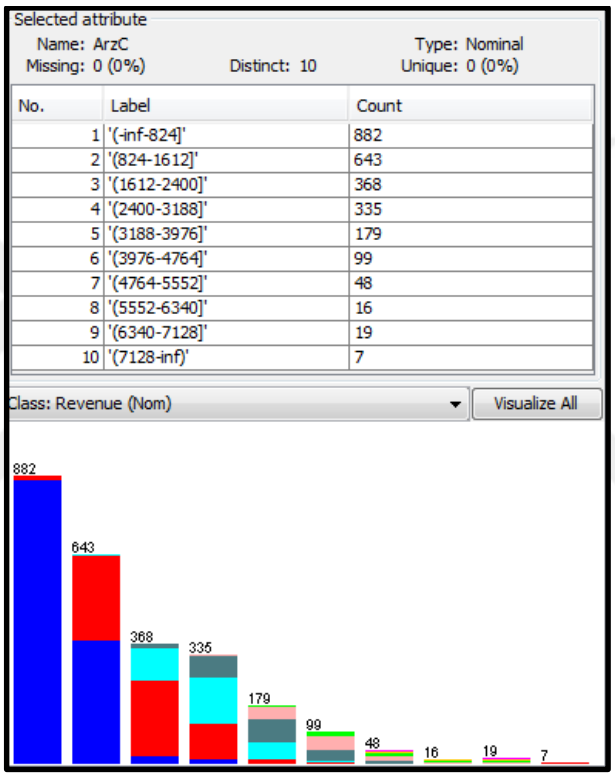
In Figure 3.1 discretization results for Km has been given. This attribute depicts the distances between origin and destination of a flight.
Except bin ten, the number of instances in each bin range between 141 and 420. The fifth bin has the maximum number of instances which is 420. In the figure each bin has the ten different colors. These colors show the ratio of each classes.

Figure 3.1: Discretization result for Km



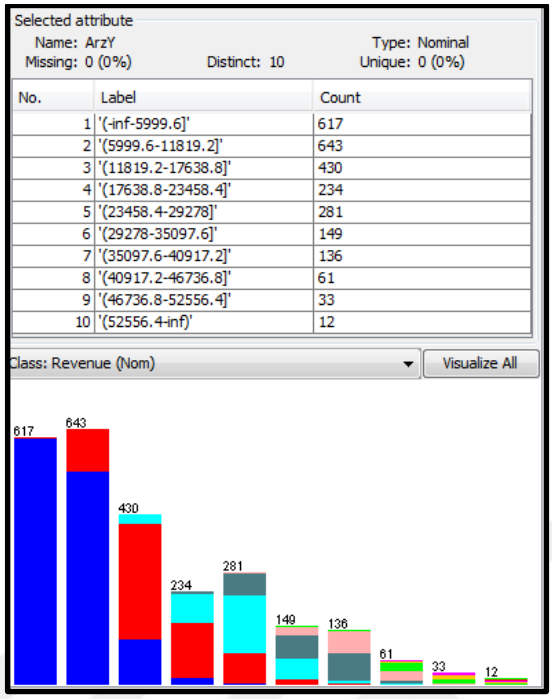
Discretization results for ArzC is given in Figures in 3.2. The number of instances mostly decrease as bin label increases. The number of instances in the first six bin dominate the total number of instances. Each bin shows the number of available business class seat capacity in flights for specific market in a month. For instances, second bin shows there are 643 flights whose total available business seat capacities are between 824 and 1612.

Figure 3.2: Discretization result for ArzC



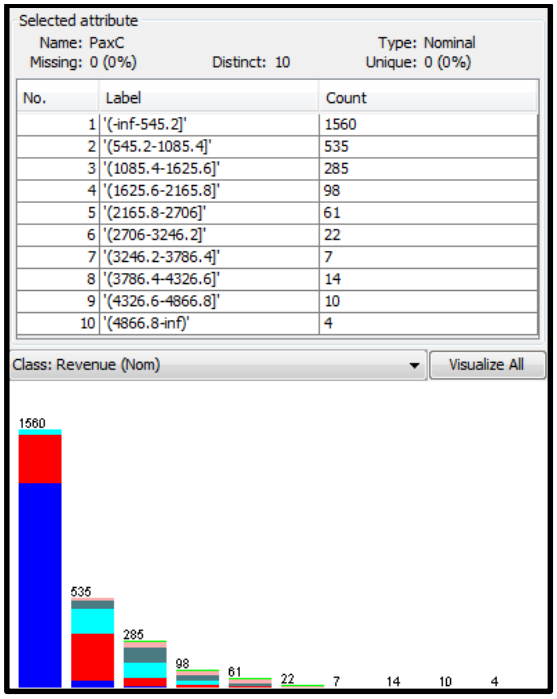
Discretization results for ArzY is given in Figures in 3.3. Similar to ArzC, the number of instances mostly decrease as bin label increases. Each bin shows the number of available economy class seat capacity in flights for specific market in a month. If you consider the fourth bin in the figure 3.3, there are 234 flights whose total available business seat capacities are between 17638 and 23458. Someone may easily compare the ratio of number of classes in each bin just by comparing the colors in each bin.

Figure 3.3: Discretization result for ArzY



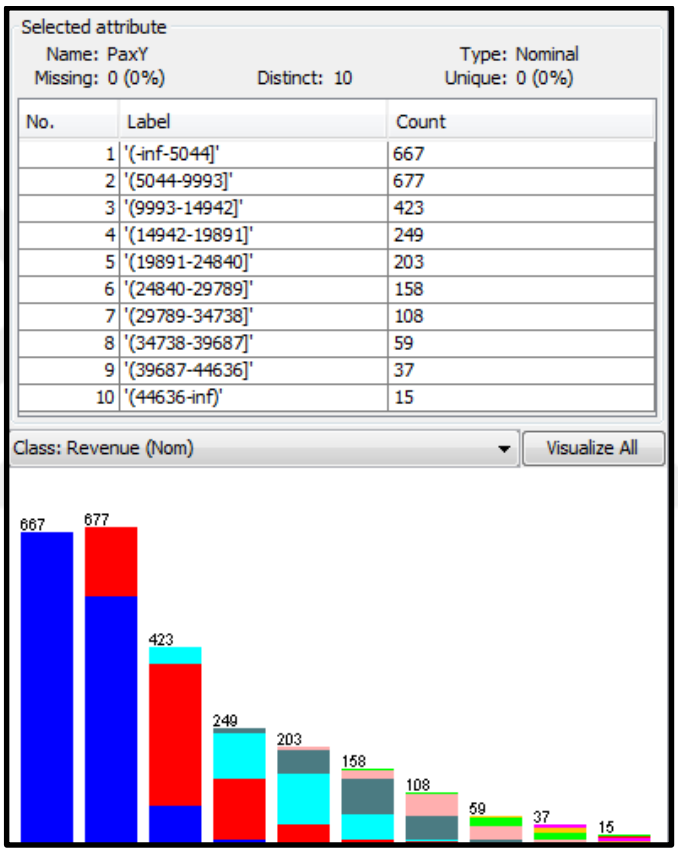
Discretization results for PaxC is tabulated in Figures 3.4. For PaxC the number of instances in bin 1-3 dominate the total number of instances. This means that the number of business passengers up to 1625 are most likely to be seen in a month for a specific flight market.

Figure 3.4: Discretization result for PaxC



Discretization results for PaxY is tabulated in Figures 3.5. For PaxY approximately 85 percent of the total number of instances are occurred in the first five bin. This means that the number of business passengers up to 24840 are most likely to be seen in a month for a specific flight market. In the first bin, almost all instances classified in the first revenue class. In the second and third bins, we see that there are two and three revenue class respectively.

Figure 3.5: Discretization result for PaxY

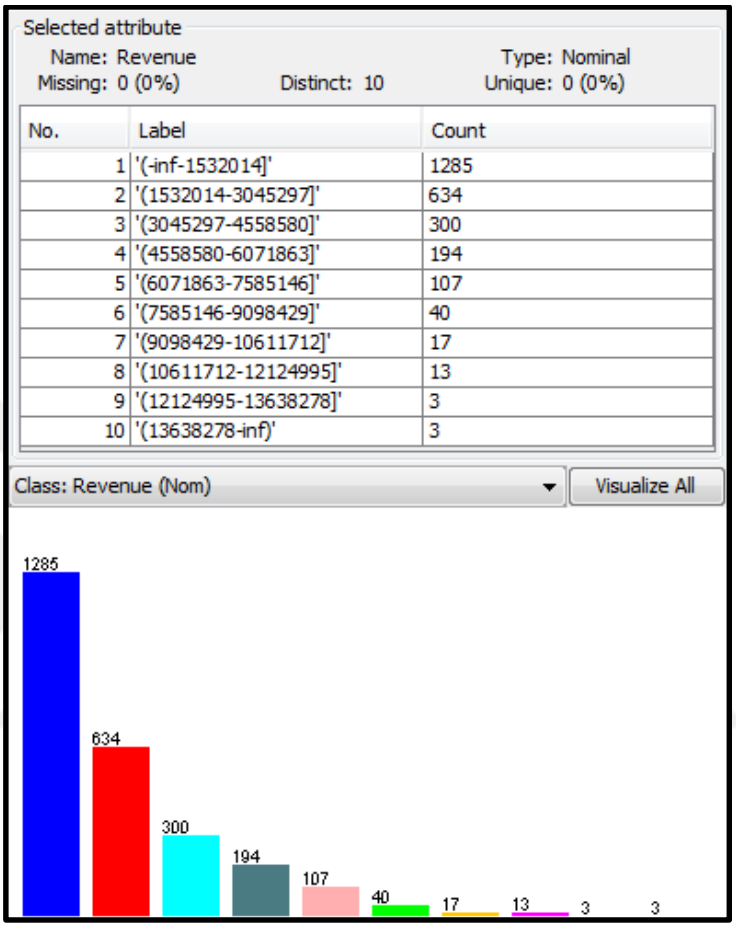


Discretization results for revenue is tabulated in Figures 3.6. The attribute revenue is our class column. In other words, classification will be executed according to revenue. Having completed discretization, we have ten bins those shows the revenue intervals.

Classes are in this attribute are not exact value. It consists of revenue intervals which of them has 1.500.000 incremental pitch. The first class has the most number of instances which is 1285. For the total number of instances perspective, the first five bin dominate

the total number of instances. In the figure each bin has different color. Each color symbolizes the different class value.

Figure 3.6: Discretization result for Revenue



3.2 METHODS

Machine learning is a subfield of computer science and it is developed from the studies pattern 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 program instruction directly. They build models from the given input data to obtain prediction or decision.

Computational statistics is a discipline which study the design of algorithms for statistical method implementation on computers. Also it has a strong connection to mathematical

optimization .Machine learning is closely related to computational statistics. Machine learning is applied in different computing tasks in which designing and programming algorithm is not feasible. The main application of machine learnings are spam filtering, recognition algorithms, and search engines. Even though machine learning and data mining seem to be as same field, however, machine learning mostly focusing on data analysis.

In this subsection, we will briefly describe the methods that we are going to test our data set with.

3.2.1 BayesNet

The properties of Bayesian Networks (Bayesnet) are first summarized by Pearl (1988) and Neapolitan (1989). Also they established Bayesnet as a field of study. Pearl (1988) emphasized especially 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 represent events as conditional probabilities, which involves random variables. Bayesnet can compute the values of a subset of variables by using given values of another subset of variables.

Bayesnet has several advantages for data analysis. We will mention two of them here. First, Bayesian networks can handle incomplete data sets without any extra computations or calculations. If there are two variables which are highly anti-correlated, then most of the prediction algorithms need the inputs for every possible cases. However, Bayesnet works fine for this type 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 will apply Bayesnet to our data set and then interpret the results in Section 4.1.

3.2.2 SMO

The Sequential Minimal Optimization (SMO) algorithm is proposed by Platt (1998) for training support vector classifier. Training a support vector machine requires the solution of a very large quadratic programming optimization 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 Sequential Minimal Optimization (SMO) are closely related to applications of support vector machine, which are given in the next subsection.

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

The main advantage of 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 data sets, SMO can be more than a thousand times faster than the chunking algorithms (Platt, 1998).

We will apply SMO to our data set and then interpret the results in Section 4.2.

3.2.3 SVM

Support Vector Machine (SVM) was first introduced in 1992 by Boser, Guyon and Vapnik. The current standard version was proposed by Cortes and Vapnik (1995). They

proposed this algorithm especially for two-group classification problems. Previously SVM implemented for the restricted case where the training data without errors. They extended the implementation of SVM for non-separable training data.

Application of Support Vector Machine includes text and hypertext categorization, classification of images, and classification of proteins in medical sciences. SVM becomes popular because it is especially very successful in hand written digit recognition. SVM is mostly regarded as important example kernel methods which is one of the key area 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 it predicts the category that these new examples belong be assuring of the gap.

One of the main advantages of Support Vector Machine is that it has good performance even with a large number of inputs. On the hands 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 will be applied to our data set and then interpreted.

3.2.4 MLP

Multilayer Perceptron (MLP) is first proposed by Rosenblatt in 1961. He simplified artificial neural networks problem by considering a particular a type of neural network which is called perceptron. For the 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.

Multilayer Perceptron (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 hand written character recognition. For more 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 connected to the next layer. MLP uses supervised learning technic which is called backpropogation for training the network. More 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 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 will apply MLP to our data set and then interpret the results in Section 4.4.

3.2.5 RBFNETWORK

A Radial Basis Function Network (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 generalization in terms of interpolation between known data points, and also they 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.

Radial Basis Function Networks 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 modeling non-linear data and can be trained in one stage. It also learn the given application quickly. Another advantage of RBFNetwork is that it is useful in solving problems where the input data are corrupted with additive noise.

We will apply RBFNetwork to our data set and then interpret the results in Section 4.1.

4. FINDINGS

In this section, we will give the outcomes of the algorithms for our data set. The data set used in this thesis is composed of airline market information which shows some old information about a specific market through three years and 36 months. The data set has 2596 instances and 8 attributes, which also means 2596 rows, 8 columns. Finding predictive relationship in the data set is our main objective. In order to do that, we basically classified our data set according to "revenue", by using classification method, or in other words machine learning algorithms. Because our data set 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 which composed of 1.500.000 incremental pitch instead of exact value classes. At the end of classification process, all instances were classified according to their revenue interval results.

Weka is a software in data mining area that has been used by many 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, Linux, etc. The outcomes will be given in this section are obtained by applying the mentioned algorithms in Section 3 to our data set in Weka.

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

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

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

In order to evaluate result, it is the crucial point to have different test data and training data. However, if the data is limited, whole data set 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 limited data scenario. The first one is "percentage split" and the second one is "cross validation".

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

In cross validation option weka divided our data set into ten pieces. Ten is a variable which can be set by user. We took nine of them to use in training and last pieces for testing And then we took another nine pieces for training and the remaining one piece for testing. Weka did this cycle for 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 BayesNet method which are obtained via Weka is as follows;

Table 4.1: Summary of BayesNet Outcome

Correctly Classified Instances	1974	76.0401%
Incorrectly Classified Instances	622	23.9599%

Kappa statistic	0.6472	
Mean absolute error	0.0519	
Root mean squared error	0.1898	
Relative absolute error	38.395%	
Root relative squared error	73.075%	
Total Number of Instances	2596	

1974 instances of the total 2596 instances are classified correctly. This corresponds to 76 percent 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 of zero, it means the lack of any relation. If it approaches to one for 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 mean absolute error can be formulated as MAE = $\sum(|f(xi) - yi|)/N$ 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 = $\sum(f(xi) - yi)^2/N$. 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 root mean squared error (RMSE) is then evaluated by RMSE = $\sqrt{MSE}$.

In MSE and RMSE, the formulas include a quadratic function and hence large errors are high weighted. Thus, RMSE is used for especially for the models where large errors are undesired.

However, for the following four lines in table 4.1 which depicts errors are not particularly useful in a classification task, because they are measures used to assess performance when the task is numeric prediction.

Table 4.2: 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)'
Avg.	0.76	0.068	0.766	0.76	0.763	0.955	

TP (True Positive) Rate is the rate of correctly classified instances as a given class. As seen from the above table, TP rates are high for the classes where the revenue is relatively low.

FP (False Positive) 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 very similar to the values that correspond to TP rate and FP rate.

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 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.

ROC (receiver operating characteristic) 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 y axis is the TP rate. The area under a ROC curve can be maximum 1. If the ROC area is close to 1 which means that classifier method is working successfully. The results of the method is worthless if the ROC area is close to 0.5. For instance; for the second class (the revenue is between 1532014-3045297), the ROC area is relatively smaller than the seventh class (the revenue is between 9098429-10611712) because the FP rate of the seventh class is significantly less than the FP rate of the second class even the TP rate of the second class is greater than the TP rate of the seventh class. However, the ROC areas of all classes are greater than 0.9 which means that the method is working good for all classes.

The values for confusion matrix of the BayesNet are tabulated in Table 4.3.

Table 4.3: BayesNet Confusion Matrix

a	b	c	d	e	f	g	h	i	j	classified as
1175	105	5	0	0	0	0	0	0	0	a = '(-inf-1532014]'

92	426	101	7	7	1	0	0	0	0	b = '(1532014-3045297]'
0	66	183	47	3	1	0	0	0	0	c = '(3045297-4558580]'
0	1	57	107	25	4	0	0	0	0	d = '(4558580-6071863]'
0	0	3	30	58	11	5	0	0	0	e = '(6071863-7585146]'
0	0	0	1	15	13	8	3	0	0	f = '(7585146-9098429]'
0	0	0	0	1	5	6	5	0	0	g = '(9098429-10611712]'
0	0	0	0	0	2	5	5	1	0	h = '(10611712-12124995]'
0	0	0	0	0	0	0	2	1	0	i = '(12124995-13638278]'
0	0	0	0	0	0	0	3	0	0	j = '(13638278-inf)'

A confusion matrix visualizes the performance of the algorithm. Each column of confusion matrix represents the number of instances in 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 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

The summary of the results of SMO method which are obtained via Weka is as follows;

Table 4.4: Summary of SMO Outcome

Correctly Classified Instances	2217	85.4006%
Incorrectly Classified Instances	379	14.5994%
Kappa statistic	0.7831	
Mean absolute error	0.1608	
Root mean squared error	0.2735	
Relative absolute error	119.044%	
Root relative squared error	105.332%	
Total Number of Instances	2596	

2217 instances of the total 2596 instances are classified correctly. This corresponds to 85 percent of the total instances. This shows that a great amount of total instances are classified correctly.

The corresponding Kappa statistic is 0.78.

Table 4.5: SMO Detailed Accuracy by Class

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.955	0.05	0.95	0.955	0.952	0.97	'(-inf-1532014]'
	0.838	0.052	0.838	0.838	0.838	0.911	'(1532014-3045297]'
	0.75	0.037	0.726	0.75	0.738	0.945	'(3045297-4558580]'
	0.67	0.021	0.718	0.67	0.693	0.969	'(4558580-6071863]'
	0.71	0.013	0.697	0.71	0.704	0.976	'(6071863-7585146]'
	0.45	0.007	0.5	0.45	0.474	0.973	'(7585146-9098429]'
	0.412	0.003	0.467	0.412	0.437	0.983	'(9098429-10611712]'
	0.077	0.003	0.1	0.077	0.087	0.993	'(10611712-12124995]'

	0.667	0.001	0.5	0.667	0.571	0.999	'(12124995-13638278]'
	0	0.002	0	0	0	0.997	'(13638278-inf)'
Avg.	0.854	0.044	0.853	0.854	0.853	0.953	

TP (True Positive) Rate is the rate of correctly classified instances as a given class. As seen from the above table, TP rates get the highest value for the classes where the revenue is relatively low. Even this looks like similar to the values in Bayesnet, it is different because as the revenue increases, TP lowers until a point which is in the 4^{th} class, and then it increases in the next class. Through the following next three class it decreases and again increase. Finally, it gets its lowest value for TP rate in the 8^{th} class.

FP (False Positive) 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 it moves 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 (receiver operating characteristic) area is 0.911 which shows that the values are very close the maximum value of 1. Thus, this classifier method work successfully for our data set.

The confusion matrix of the SMO algorithm as is shown in the Table 4.6 which is given below

Table 4.6: SMO Confusion Matrix

a	b	c	d	e	f	g	h	i	j	classified as
1227	58	0	0	0	0	0	0	0	0	a = '(-inf-1532014]'
65	531	38	0	0	0	0	0	0	0	b = '(1532014-3045297]'

0	45	225	30	0	0	0	0	0	0	c = '(3045297-4558580]'
0	0	46	130	18	0	0	0	0	0	d = '(4558580-6071863]'
0	0	0	21	76	10	0	0	0	0	e = '(6071863-7585146]'
0	0	1	0	14	18	4	2	1	0	f = '(7585146-9098429]'
0	0	0	0	0	6	7	3	0	1	g = '(9098429-10611712]'
0	0	0	0	1	2	4	1	1	4	h = '(10611712-12124995]'
0	0	0	0	0	0	0	1	2	0	i = '(12124995-13638278]'
0	0	0	0	0	0	0	3	0	0	j = '(13638278-inf)'

4.3 SVM

The summary of the results of SVM method which are obtained via Weka is as follows;

Table 4.7: Summary of SVM Outcome

Correctly Classified Instances	2005	77.2342%
Incorrectly Classified Instances	591	22.7658%
Kappa statistic	0.6585	
Mean absolute error	0.0455	
Root mean squared error	0.2134	
Relative absolute error	33.7093%	
Root relative squared error	82.1669%	
Total Number of Instances	2596	

2005 instances of the total 2596 intances are classified correctly. This corresponds to 77 percent of the total instances. The corresponding Kappa statistic is 0.66.

Table 4.8: SVM Detailed Accuracy by Class

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.94	0.074	0.926	0.94	0.933	0.933	'(-inf-1532014]'
	0.763	0.1	0.712	0.763	0.737	0.832	'(1532014-3045297]'
	0.563	0.051	0.593	0.563	0.578	0.756	'(3045297-4558580]'
	0.557	0.045	0.5	0.557	0.527	0.756	'(4558580-6071863]'
	0.327	0.029	0.324	0.327	0.326	0.649	'(6071863-7585146]'
	0.025	0	0.5	0.025	0.048	0.512	'(7585146-9098429]'
	0	0	0	0	0	0.5	'(9098429-10611712]'
	0	0	0	0	0	0.5	'(10611712-12124995]'
	0	0	0	0	0	0.5	'(12124995-13638278]'
	0	0	0	0	0	0.5	'(13638278-inf)'
Avg.	0.772	0.071	0.759	0.772	0.762	0.85	

As seen from the above table, TP rates get the highest values for the classes where the revenue is relatively low. However, it dramatically lower while the class number increases. TP rate reaches zero value in the 7^{th} class.

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

The highest ROC value occurs in the first class which is 0.933. And then it dramatically falls down, after the first class it never reaches a 0.9 value. Rather than, it converges to a value 0.5 in the 6^{th} class. Thus, this classifier method does not work successfully for our data set.

The values for confusion matrix of the SVM are tabulated in Table 4.9.

Table 4.9: SVM Confusion Matrix

a	b	c	d	e	f	g	h	i	j	classified as
1208	77	0	0	0	0	0	0	0	0	a = '(-inf-1532014]'
97	484	42	7	4	0	0	0	0	0	b = '(1532014-3045297]'
0	102	169	27	2	0	0	0	0	0	c = '(3045297-4558580]'
0	8	71	108	7	0	0	0	0	0	d = '(4558580-6071863]'
0	0	3	69	35	0	0	0	0	0	e = '(6071863-7585146]'
0	0	0	5	34	1	0	0	0	0	f = '(7585146-9098429]'
0	2	0	0	15	0	0	0	0	0	g = '(9098429-10611712]'
0	2	0	0	10	1	0	0	0	0	h = '(10611712-12124995]'
0	3	0	0	0	0	0	0	0	0	i = '(12124995-13638278]'
0	2	0	0	1	0	0	0	0	0	j = '(13638278-inf)'

4.4 MLP

The summary of the results of MLP method which are obtained via Weka is as follows;

Table 4.10: Summary of MLP Outcome

Correctly Classified Instances	2176	83.8213%
Incorrectly Classified Instances	420	16.1787%
Kappa statistic	0.76	
Mean absolute error	0.0351	
Root mean squared error	0.1681	
Relative absolute error	26.0183%	
Root relative squared error	64.7219%	
Total Number of Instances	2596	

2176 instances of the total 2596 instances are classified correctly. This corresponds to 84 percent of the total instances. The corresponding Kappa statistic is 0.76.

Table 4.11: MLP Detailed Accuracy by Class

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.949	0.053	0.946	0.949	0.948	0.983	'(-inf-1532014]'
	0.808	0.058	0.819	0.808	0.813	0.95	'(1532014-3045297]'
	0.693	0.037	0.707	0.693	0.7	0.944	'(3045297-4558580]'
	0.727	0.03	0.662	0.727	0.693	0.966	'(4558580-6071863]'
	0.626	0.016	0.632	0.626	0.629	0.962	'(6071863-7585146]'
	0.35	0.007	0.438	0.35	0.389	0.956	'(7585146-9098429]'
	0.471	0.003	0.471	0.471	0.471	0.953	'(9098429-10611712]'
	0.385	0.004	0.333	0.385	0.357	0.974	'(10611712-12124995]'
	0.333	0.001	0.25	0.333	0.286	0.861	'(12124995-13638278]'
	0	0	0	0	0	0.414	'(13638278-inf)'
Avg.	0.838	0.048	0.837	0.838	0.838	0.967	

As seen from the above table, TP rates get the highest value for the classes where 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 (receiver operating characteristic) area is 0.967. Even though this value is good, the last two values are below 0.9. Except the last two classes, results show meaningful ROC value which are bigger than 0.95.

The values for confusion matrix of the MLP are tabulated in Table 4.12.

Table 4.12: MLP Confusion Matrix

a	b	c	d	e	f	g	h	i	j	classified as
1220	64	0	1	0	0	0	0	0	0	a = '(-inf-1532014]'
69	512	50	2	1	0	0	0	0	0	b = '(1532014-3045297]'
0	44	208	44	3	0	0	0	1	0	c = '(3045297-4558580]'
1	2	29	141	20	1	0	0	0	0	d = '(4558580-6071863]'
0	3	3	23	67	11	0	0	0	0	e = '(6071863-7585146]'
0	0	2	1	14	14	6	3	0	0	f = '(7585146-9098429]'
0	0	2	1	1	3	8	2	0	0	g = '(9098429-10611712]'
0	0	0	0	0	3	3	5	2	0	h = '(10611712-12124995]'
0	0	0	0	0	0	0	2	1	0	i = '(12124995-13638278]'
0	0	0	0	0	0	0	3	0	0	j = '(13638278-inf)'

4.5 RBFNETWORK

The summary of the results of RBFNetwork method which are obtained via Weka is as follows;

Table 4.13: Summary of RBFNetwork Outcome

Correctly Classified Instances	2037	78.4669%
Incorrectly Classified Instances	559	21.5331%
Kappa statistic	0.6801	
Mean absolute error	0.0541	
Root mean squared error	0.1747	
Relative absolute error	40.0622%	
Root relative squared error	67.2598%	
Total Number of Instances	2596	

2037 instances of the total 2596 instances are classified correctly. This corresponds to 78 percent of the total instances. The corresponding Kappa statistic is 0.68.

Table 4.14: RBFNetwork Detailed Accuracy by Class

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.93	0.074	0.925	0.93	0.927	0.983	'(-inf-1532014]'
	0.727	0.088	0.727	0.727	0.727	0.928	'(1532014-3045297]'
	0.603	0.051	0.605	0.603	0.604	0.941	'(3045297-4558580]'
	0.588	0.034	0.582	0.588	0.585	0.96	'(4558580-6071863]'
	0.542	0.019	0.552	0.542	0.547	0.971	'(6071863-7585146]'
	0.45	0.006	0.529	0.45	0.486	0.908	'(7585146-9098429]'
	0.235	0.003	0.308	0.235	0.267	0.804	'(9098429-10611712]'
	0.385	0.004	0.313	0.385	0.345	0.773	'(10611712-12124995]'
	0.333	0.002	0.2	0.333	0.25	0.826	'(12124995-13638278]'
	0	0.001	0	0	0	0.111	'(13638278-inf)'
Avg.	0.785	0.068	0.784	0.785	0.784	0.958	

As seen from the above table, 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 relatively high.

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

Even if the weighted average value for ROC area seems good because of the value 0.958, results are not as good as their average values. For the first 6 class ROC area is greater than 0.9. On the other hand, the remaining areas are smaller.

The values for confusion matrix of the RBFNetwork are tabulated in Table 4.15.

Table 4.15: RBFNetwork Confusion Matrix

a	b	c	d	e	f	g	h	i	j	classified as
1195	87	3	0	0	0	0	0	0	0	a = '(-inf-1532014]'
97	461	66	5	4	0	0	0	1	0	b = '(1532014-3045297]'
0	79	181	38	1	0	0	1	0	0	c = '(3045297-4558580]'
0	5	46	114	28	1	0	0	0	0	d = '(4558580-6071863]'
0	2	3	37	58	6	1	0	0	0	e = '(6071863-7585146]'
0	0	0	2	12	18	5	1	1	1	f = '(7585146-9098429]'
0	0	0	0	2	7	4	4	0	0	g = '(9098429-10611712]'
0	0	0	0	0	2	3	5	2	1	h = '(10611712-12124995]'
0	0	0	0	0	0	0	2	1	0	i = '(12124995-13638278]'
0	0	0	0	0	0	0	3	0	0	j = '(13638278-inf)'

5. DISCUSSION

In airline industry profit is the most important issue for the sustainability of a company. Although this issue looks like same with other industries, there is one difference which is that; cost in airline industry is not as manageable as in other industries. Therefore revenue management become the most crucial point for airline industry in order to make profit. From this point of view, for many years lots of researches and surveys have been done and still continuing.

In this section, we will 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 those affect revenue. In this study only the available seat values in both 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 the machine learning algorithms, a reasonable classification method were tried to achieved in order to evaluate revenue interval of attributes. We use the term “interval”, because our data set is composed of numeric value. We got nominal values after applying the discretization method to them. Our class value is not a specific unique value. Instead, our class values are composed of revenue intervals. It has ten intervals, starting from zero. Each interval has 1.500.000 incremental pitch.

In order to evaluate the classification method, we need training data and test data. Instead of preparing two separate data set due to limited instances, we used cross validation option in Weka. In this option Weka divided the data set into ten pieces. Then took nine of them to use as training data and the last piece for testing. Weka did the same cycle every different nine pieces and the remaining one piece. At the end return averages of them.

We will check the accuracy of 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 the 85.4 percent of instances correctly which means 2217 instances over 2596. MLP is the second best for classification ability. It reached the value of 83.8 percent The value 2176 over 2596 is quiet close to SMO. The remaining are RBFNetwork, SVM and BayesNet; whose correctly classified instances rates are 78.5 percent, 77.2 percent, 76 percent, respectively. The classification abilities of all algorithms are higher than 75 percent, which can be considered 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 correlation coefficient. If it gets the value of zero, it means the lack of any relation. If it approaches to one for very strong statistical relation between the class label and attributes of instances. Corresponding to values correctly classified instances, the values for kappa statistic are in 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 detail in the tables 4.16 and 4.17.

Table 4.16: 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 4.17: 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 4.18: 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 the Table 4.18 above, 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 the value of SVM, all the remaining ROC values are bigger than 0.95.

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

For the first class, even though the highest TP rate value seems to be belonged to SMO, the value for SVM is very close to SMO. All the TP rate values are bigger than 0.90 in the first class. Looking the ROC area values, all of them are higher than 0.97 which is

good. MLP and RBFNetwork have the same value which is 0.983. The detailed accuracy outcomes for the first class are tabulated in Table 4.19 in below.

Table 4.19: 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 seems 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 ROC area of MLP algorithm is the best, which is 0.95. Except the value 0.832 of SVM algorithms, All remaining ROC values are higher than 0.90. The detailed accuracy outcomes for the second class are tabulated in Table 4.20 in below.

Table 4.20: Detailed Accuracy Outcomes for the Second Class

'(1532014-3045297]'	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area
BayesNet	0.672	0.088	0.712	0.672	0.692	0.913
SMO	0.838	0.052	0.838	0.838	0.838	0.911
SVM	0.763	0.1	0.712	0.763	0.737	0.832
MLP	0.808	0.058	0.819	0.808	0.813	0.95
RBFNetwork	0.727	0.088	0.727	0.727	0.727	0.928

For the third class, TP rate values are getting smaller relatively. However, the value for SMO still seems good, which is 0.75. The worst is the 0.563 which belongs to SVM.

Except the value 0.756 of SVM algorithms, All remaining ROC values are higher than 0.92. The detailed accuracy outcomes for the third class are tabulated in Table 4.21 in below.

Table 4.21: Detailed Accuracy Outcomes for the Third Class

'(3045297-4558580]'	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area
BayesNet	0.61	0.072	0.524	0.61	0.564	0.929
SMO	0.75	0.037	0.726	0.75	0.738	0.945
SVM	0.563	0.051	0.593	0.563	0.578	0.756
MLP	0.693	0.037	0.707	0.693	0.7	0.944
RBFNetwork	0.603	0.051	0.605	0.603	0.604	0.941

For the fourth class, the values for TP rate are relatively low in each algorithm. Only the value for MLP is greater than 0.7. Although TP rate values are small, due to the small FP rate values, the values for ROC area seem good. The biggest value is 0.969 which belongs to SMO. Beyond that, all the ROC values are bigger than 0.95 except the value for SVM which is 0.756. The detailed accuracy outcomes for the foruth class are tabulated in Table 4.22 in below.

Table 4.22: Detailed Accuracy Outcomes for the Fourth Class

'(4558580-6071863]'	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area
BayesNet	0.552	0.035	0.557	0.552	0.554	0.956
SMO	0.67	0.021	0.718	0.67	0.693	0.969
SVM	0.557	0.045	0.5	0.557	0.527	0.756
MLP	0.727	0.03	0.662	0.727	0.693	0.966
RBFNetwork	0.588	0.034	0.582	0.588	0.585	0.96

For the fifth class, the TP rate value and FP rate value of SVM algorithm is smallest and biggest respectively. Therefore it has the worst ROC area value which is 0.649. SMO still has relatively a good TP rate in the fifth class. Due to this reason, it has the highest ROC

area value which is 0.976. All remaining ROC values are higher than 0.96. The detailed accuracy outcomes for the fifth class are tabulated in Table 4.23 in below.

Table 4.23: Detailed Accuracy Outcomes for the Fifth Class

'(6071863-7585146]'	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area
BayesNet	0.542	0.02	0.532	0.542	0.537	0.971
SMO	0.71	0.013	0.697	0.71	0.704	0.976
SVM	0.327	0.029	0.324	0.327	0.326	0.649
MLP	0.626	0.016	0.632	0.626	0.629	0.962
RBFNetwork	0.542	0.019	0.552	0.542	0.547	0.971

In this study some data mining methods were used in order to make revenue prediction. The data set used in this thesis is composed of airline market information which shows some old information about a specific market through three years and 36 months. The data mining research was conducted for 2596 instances each of which compounds of 8 attributes. They are YearMonth, Season, Km, ArzC, ArzY, PaxC, PaxY and Revenue. Weka program was run for this data set by selecting Bayesian Network (BayesNet), Sequential Minimal Optimization (SMO), Support Vector Machine (SVM), Multi Layer Perceptron (MLP), Radial Basis Function Network (RBFNetwork) classification models respectively, and when the obtained outcomes are considered, it is seen that SMO classification algorithm is the one with highest accuracy. MLP has the second best accuracy. However, if we look at for the weighted average point of view, due to lowest values in FP rates, MLP has slightly bigger ROC value than SMO.

It can be said that for our data set Sequential Minimal Optimization algorithm and Multi Layer Perceptron algorithm are the most convenient algorithms.

6. CONCLUSION

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

Due to the limited amount of data, cross fold method was used in order to produce input data. The Input data has two sub set. The first is the training data and the second is test data. By using cross fold method, Weka divided whole data set into ten pieces. 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 basically composed of revenue classes which have 1.500.000 incremental pitch. Also we can see how many instances were correctly classified in which classes.

Weka program was run for this data set by selecting Bayesian Network (BayesNet), Sequential Minimal Optimization (SMO), Support Vector Machine (SVM), Multi Layer Perceptron (MLP), Radial Basis Function Network (RBFNetwork) classification models respectively. The comprehensive comparison of the outcomes of classification models was presented in the section 5. SMO classification algorithm is the one with highest accuracy. It achieved to classify the 85.4 percent of instances correctly which means 2217 instances over 2596. MLP has the second best accuracy with the rate of 83.8 percent. The remaining are RBFNetwork, SVM and BayesNet; whose correctly classified instances rates are 78.5 percent, 77.2 percent, 76 percent, respectively.

REFERENCES

Books

Doganis, R. The Airline Business 2nd ed. London 2006

Neapolitan, R. E. 1989. Probabilistic reasoning in expert systems: theory and algorithms. Wiley.

Rosenblatt, F., 1961. Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms. Spartan Books, Washington DC

Taneja, N. K., 1978. Airline Traffic Forecasting: A Regression Analysis Approach. Lexington Books.

Periodicals

Airbus Global Market Forecast 2012-2031, 2012, Airbus Inc.

Beckmann, M.J. and Bobkowski, F., 1958. Airline Demand: An Analysis of Some Frequency Distributions. *Naval Res. Logistics Q. 5*, pp. 43–51.

Boeing Current Market Outlook 2012-2032, 2012, Boeing Inc.

Boser, B. E.; Guyon, I. M.; Vapnik, V. N., 1992. A training algorithm for optimal margin classifiers. *Proceedings of the fifth annual workshop on Computational learning theory - COLT '92*. pp. 144.

Botimer, T. C. 1997. Select Ideas on Forecasting with Sales Relative to Bucketing and 'Seasonality,' Unpublished Company Report, Continental Airlines, Inc.

Broomhead, D. S.; Lowe, D., 1988. Radial basis functions, multi-variable functional interpolation and adaptive networks. *Technical report*. RSRE. 4148.

Broomhead, D. S.; Lowe, D. 1988. Multivariable functional interpolation and adaptive networks. *Complex Systems 2*. pp. 321–355.

Cao, R. Z., et al. "Data mining techniques to improve no-show forecasting." *Service Operations and Logistics and Informatics (SOLI). 2010 IEEE International Conference on*. IEEE, 2010.

Cortes, C.; Vapnik, V., 1995. "Support-vector networks". *Machine Learning 20 (3)*, pp. 273.

Dunleavy, H.; Phillips, G., 2009. The future of airline revenue management. *Journal of Revenue and Pricing Management*

Gallo, M. A., and Kepto, M 2014. The Relationship Between 2011 METAR and TAF Data at Chicago-Midway and Seattle-Tacoma Airports. *Collegiate Aviation Review 32.1* pp. 18.

Grabbe, S., Sridhar, B., and Mukherjee, A., 2014. Clustering Days and Hours with Similar Airport Traffic and Weather Conditions. *Journal of Aerospace Information Systems 11.11* pp. 751-763.

Hall, M., et al. 2009. The WEKA data mining software: an update. ACM SIGKDD explorations newsletter 11.1. pp. 10-18.

Mack, D., et al. 2011. Deriving Bayesian Classifiers from Flight Data to Enhance Aircraft Diagnosis Models. *Annual Conference of the Prognostics and Health Management Society.*

McGill, J. I, Van Ryzin G. J., 1999. Revenue Management: Research Overview and Prospects

Morales, D. R., and Wang, J., 2010. Forecasting cancellation rates for services booking revenue management using data mining. *European Journal of Operational Research 202.2.* pp. 554-562.

Pak, K. and Piersma, N., 2002. Overview of operational research techniques for airline revenue management. Vol 56, pp. 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. *Advances in Kernel Methods - Support Vector Learning*.

Sa, J., 1998. Reservations Forecasting in Airline Yield Management. *Master's thesis, Flight Transportation Laboratory, Massachussetts Institute of Technology, Cambridge, MA.*

Schumann, J., Cate, K, and Lee, A., 2011. Analysis of air traffic track data with the autobayes synthesis system. *Logic-Based Program Synthesis and Transformation*. Springer Berlin Heidelberg, pp. 21-36.

Theodore, C. B. and Belobaba, P. P., 1999. Airline pricing and fare product differentiation: A new theoretical framework. *The Journal of the Operational Research Society*

Wasserman, P. D., Schwartz, T., 1998. Neural networks. II. *What are they and why is everybody so interested in them now*, vol. 3, issue 1, pp. 10-15