Data Mining Applications in Social Lending and Anchorage Planning

A thesis submitted to the
Graduate School of Natural and Applied Sciences

by

Milad MALEKIPIRBAZARI

in partial fulfillment for the
degree of Master of Science

in

Industrial and Systems Engineering

This is to certify that we have read this thesis and that in our opinion it is fully adequate, in scope and quality, as a thesis for the degree of Master of Science in Industrial and Systems Engineering.

APPROVED BY:

Assoc. Prof. Vural Aksakallı .....
(Thesis Advisor)

Asst. Prof. Ahmet Bulut .....

Asst. Prof. Ali Fuat Alkaya .....

This is to confirm that this thesis complies with all the standards set by the Graduate School of Natural and Applied Sciences of İstanbul Şehir University:

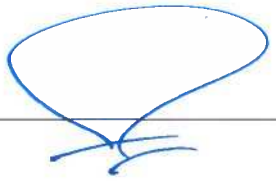
DATE OF APPROVAL: 18 August 2015

SEAL/SIGNATURE:

Declaration of Authorship

I, Milad MALEKIPIRBAZARI, declare that this thesis titled, 'Data Mining Applications in Social Lending and Anchorage Planning' and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed: 

Date: 18/8/2015

Data Mining Applications in Social Lending and Anchorage Planning

Milad MALEKIPIRBAZARI

Abstract

In today's data-driven world, various industries resort to data mining on a regular basis for competitive advantage and sustained growth. In this thesis, we consider employment of data mining techniques in two application domains: social lending and anchorage planning.

With the advance of electronic commerce and social platforms, social lending (also known as peer-to-peer lending) has emerged as a viable platform where lenders and borrowers can do business without the help of institutional intermediaries such as banks. Social lending has gained significant momentum recently, with some platforms reaching multi-billion dollar loan circulation in a short amount of time. On the other hand, sustainability and possible widespread adoption of such platforms depend heavily on reliable risk attribution to individual borrowers. For this purpose, we propose a random forest (RF) based classification method for predicting borrower status. Our results on data from the popular social lending platform Lending Club (LC) indicate the RF-based method outperforms the FICO credit scores as well as LC grades in identification of good borrowers.

The second data mining application domain we consider pertains to maritime transportation. In particular, we first provide a comprehensive statistical analysis on a new anchorage data set gathered for nine recent consecutive years in Istanbul anchorages. We introduce a data mining framework with the aim of identifying a good estimate for anchorage duration for a given vessel. Our goal is to develop an understanding of key factors relevant to vessel anchorage and devise an effective methodology for predicting anchorage duration, which is critical for efficient anchorage planning. In addition, along with a temporal analysis of vessel type traffic, we forecast vessel type traffic for the next three years using the statistical ARIMA model. Our results suggest an overall decrease in berthing vessels, yet a pronounced increase in LPG barges. This finding is rather significant as this type of vessel is more prone to accidents and any such accident would pose a great danger to the Strait.

Keywords: Data mining, social lending, anchorage planning, random forests, decision tree, ARIMA model

Sosyal Kredilendirme ve Demirleme Planlamasında Veri Madenciliği Uygulamaları

Milad MALEKIPIRBAZARI

Öz

Günümüzün veri odaklı dünyasında çeşitli endüstriler rekabet üstünlüğü sağlamak ve devamlı gelişim için düzenli bir şekilde veri madenciliğine başvurmaktadır. Bu tezde, iki uygulama alanında veri madenciliği tekniklerinin kullanılması ele alınmaktadır: sosyal kredilendirme ve demirleme planlama.

Elektronik ticaret ve sosyal platformların gelişmesi ile sosyal kredilendirme, kredi verenlerin ve kredi kullanıcılarının bankalar gibi kurumsal aracıların yardımı olmadan iş yapabildiği gerçekçi bir platform olarak ortaya çıkmıştır. Sosyal kredilendirme kısa sürede milyarlarca dolarlık kredi sirkülasyonu sağlayan bazı platformlarla birlikte son zamanlarda ciddi bir ivme kazanmıştır. Öte yandan, bu tür platformların sürdürülebilirliği ve yaygın bir şekilde kullanılması bireysel kredi kullanıcılarının riskinin doğru tahmin edilmesine bağlıdır. Bu amaçla, kredi kullanıcılarının risk durumunu tahmin etmek için rasgele ormanlar (RO) tabanlı bir sınıflandırma yöntemi öneriyoruz. Popüler sosyal kredilendirme platformu Lending Club (LC) verileri üzerindeki çalışmalarımız, RO tabanlı yöntemin iyi kredi kullanıcılarının tanımlanmasında LC sonuçlarının yanı sıra FICO kredi puanlarından da daha sağlıklı tahmin verdiğini göstermektedir.

İkinci veri madenciliği uygulama alanı olarak deniz taşımacılığı ele alınmaktadır. Özellikle, İstanbul'da son dokuz yılda toplanan demirleme bilgilerini içeren yeni bir veri seti üzerine kapsamlı bir analiz sunulmaktadır. Belirli bir gemi için demirleme süresinin tahmin edilebilmesi amacıyla bir veri madenciliği yapısı sağlanmaktadır. Amacımız gemi demirlemesi ile ilgili önemli faktörlerden bir anlam çıkarmak ve verimli demirleme planlaması için önemli olan demirleme süresi tahmini için etkili bir yöntem geliştirmektir. Buna ek olarak, gemi tipi trafiği zamansal analizi ile birlikte istatistiksel ARIMA modeli kullanılarak önümüzdeki üç yıl için gemi tipi trafiği tahmin edilmiştir. Sonuçlarımız, gemi demirleme sayısında genel bir düşüş, fakat LPG taşıyıcılarında belirgin bir artış olduğunu göstermektedir. Bu sonuçlar oldukça önemlidir çünkü bu tür gemiler kazalara daha yatkındır ve herhangi bir kaza anında Boğaz için büyük tehlike teşkil etmektedir.

Anahtar Sözcükler: Veri madenciliği, sosyal kredilendirme, demirleme planlama, rastgele ormanlar, karar ağacı, ARIMA modeli

Acknowledgments

My deepest gratitude is to my adviser, Assoc. Prof. Vural Aksakallı, for his continued support during my graduate studies at Istanbul Sehir University. I have been fortunate to have an adviser who persuaded me to explore on my own and guided me in every step when I made a mistake. His guidance helped me immensely during my research as well as writing of this thesis. I hope that one day I could become an adviser to my students as good as Dr. Vural has been to me. I could not have imagined having a better adviser and mentor.

Special thanks to my committee members, Asst. Prof. Ahmet Bulut and Asst. Prof. Ali Fuat Alkaya, for their kindness and helpful insights and suggestions.

The research in Chapter 2 was published in the Journal of Expert Systems with Applications under the title *Risk Assessment in Social Lending via Random Forests*, Volume 42, Issue 10, 15 June 2015, pp. 4621–4631. The study in Chapter 4 was presented by me in IEEE 5th International Conference on Industrial Engineering and Operations Management in Dubai, under the title *A Temporal Analysis of Ship Types in Istanbul Strait Anchorages*.

Finally, I wish to thank my parents and my brother for their support and encouragement throughout my studies.

Contents

List of Figures

List of Tables

Abbreviations

AICc	**A**kaike **I**nformation **C**riterion with **C**orrection
ARIMA	**A**utoregressive **I**ntegrated **M**oving **A**verage
AUC	**A**rea **U**nder the **C**urve
BIC	**B**aysian **I**nformation **C**riterion
CI	**C**onfidence **I**nterval
CV	**C**ross **V**alidation
d	**d**epth
DC	**D**angerous **C**argo
DGCS	**D**irectorate **G**eneral of **C**oastal **S**afety
DTI	**D**ebth to **I**ncome
FP	**F**aulse **P**ositive
GT	**G**ross **T**onnage
k-NN	**k**-**N**earest **N**eighbors
L	**L**ength
LC	**L**ending **C**lub
LPG	**L**iquified **P**etroluem **G**as
LR	**L**ogistic **R**egression
LTSS	**L**ocal **T**raffic **S**eparation **S**chemes
P2P	**P**eer to **P**eer
PPP	**P**recise **P**oint **P**ositioning
RF	**R**andom **F**orest
RMSE	**R**oot **M**ean **S**quare **E**rror
ROC	**R**eceiving **O**perating **C**haracteristic
SVM	**S**upport **V**ector **M**achine
TP	**T**rue **P**ositive

Chapter 1

Introduction

1.1 Overview

Data mining can briefly be defined as extracting meaning from data. In today's data-driven world, various industries employ data mining techniques on a regular basis for their mission-critical processes in order to gain competitive advantage and help business grow. Below is a partial list of areas where data mining is widely used:

- Financial data analysis
- Retail industry
- Electronic commerce
- Telecommunications
- Biological data analysis
- Logistics and supply chain management

There is a wide variety of textbooks on data mining and machine learning that can be recommended for readers new to the field. Presented below is a short review of four such popular textbooks ordered by their level of technicality.

- **Data Mining: Practical Machine Learning Tools and Techniques** (Witten and Frank [1]): This book contains a thorough review of introductory machine

learning concepts along with practical guidelines on utilization of popular machine learning tools and techniques in real-world data mining situations. The book discusses a number of basic data mining topics including preparing inputs, interpreting outputs, evaluating results, and algorithms.

- **Introduction to Data Mining** (Tan et al. [2]): This book presents fundamental data mining concepts and algorithms where each topic is organized into two chapters: basic concepts that provide necessary background for understanding the topic followed by more advanced concepts and algorithms.

- **Introduction to Machine Learning** (Alpaydin [3]): This book presents a comprehensive review of the subject; covering a broad range of topics not typically included in introductory machine learning texts. Subjects include Bayesian decision theory; parametric, semi-parametric, and non-parametric methods; multivariate analysis; hidden Markov models; reinforcement learning; kernel machines; graphical models; Bayesian estimation; and statistical testing.

- **Pattern Recognition and Machine Learning** (Bishop [4]): This is one of the prominent books on pattern recognition including a Bayesian viewpoint. With a fairly technical presentation, the book discusses approximate inference algorithms that allow fast approximate answers in situations where exact answers are not feasible. The book also presents in detail graphical models in machine learning.

The next three chapters of this thesis present applications of data mining in two particular application areas: social lending and anchorage planning. These chapters are independent from each other and therefore their literature reviews and recaps are incorporated within themselves.

1.2 Organization of the Thesis

The rest of this thesis is organized as follows:

Chapter 2 presents applications of data mining in social lending. With the advance of electronic commerce and social platforms, social lending (also known as peer-to-peer lending) has emerged as a viable platform where lenders and borrowers can do business

without the help of institutional intermediaries such as banks. Social lending has gained significant momentum recently, with some platforms reaching multi-billion dollar loan circulation in a short amount of time. On the other hand, sustainability and possible widespread adoption of such platforms depend heavily on reliable risk attribution to individual borrowers. For this purpose, we propose a random forest (RF) based classification method for predicting borrower status. Our results on data from the popular social lending platform Lending Club (LC) indicate the RF-based method outperforms the FICO credit scores as well as LC grades in identification of good borrowers.

Chapter 3 provides a comprehensive statistical analysis on a new anchorage data set gathered between 2006 and 2014 in Istanbul anchorages. In this chapter, we introduce a data mining framework with the aim of identifying a good estimate for anchorage duration for a given vessel. Our goal is to develop an understanding of key factors relevant to vessel anchorage and devise an effective methodology for predicting anchorage duration, which is critical for efficient anchorage planning.

Chapter 4 presents a temporal analysis of vessel type traffic between 2006 and 2014 in Istanbul anchorages using the statistical ARIMA model. Our results suggest an overall decrease of berthing vessels, yet a pronounced increase in LPG barges. This finding is rather significant as this type of vessel is more prone to accidents and any such accident would pose a great danger to the Strait.

Chapter 5 presents our detailed summary and conclusions for the previous three chapters and proposes several directions for future research.

Chapter 2

Risk Assessment in Social Lending via Random Forests

2.1 Introduction

Social lending, also known as peer-to-peer (P2P) lending, is emerging as an alternative to banks where individual members lend and borrow money using an online trading platform without the help of official financial intermediaries such as banks. The attractive feature of doing business on a peered platform is the higher potential of mutual profitability. Borrowers can obtain loans at lower interest rates and lenders can loan money at better rates than what they can get from a bank. In particular, via social lending, lenders can find a multitude of potential borrowers and choose among them the ones they wish to lend. Since the ultimate savers are predominantly consumers, and consumers are the individuals who are actually lending in the social lending model, there is no need to increase the liquidity of the loans by securitizing them. Since social lending is powered by the Internet, it would not take much effort to connect small communities such as towns, religious, or ethnic groups for the purpose of intra-community lending and borrowing.

The popular social lending platforms currently in use today are the U.S.-based Prosper[1] and Lending Club Corp.[2], UK-based Zopa Ltd.[3], and Germany-based Smava GmbH[4].

[1] http://www.prosper.com
[2] http://www.lendingclub.com
[3] http://www.zopa.com
[4] http://www.smava.de

All of these social platforms rely on the credit scores provided by a cooperating credit reporting agency; Experian, TransUnion LLC, Equifax Inc., and Schufa Holding AG respectively. The popularity of these platforms is growing as recently indicated by Lending Club (LC) which has reached 6.2 billion USD in total loans by January 2015 and transformed into a 8.5 billion USD publicly-traded company, becoming the world's largest social lending platform [5].

Our standpoint in this work, which is consistent with other studies in this context, is that even though social lenders can base their investment strategy on the traditional financial credit scores provided by external agencies, available data suggest that social lending tends to have different dynamics when compared to traditional lending. For instance, the distribution of lenders' bids on social loan listings when indexed by time follows a power law [6], an indication of a herding behavior. Assume that lenders and loan listings are denoted by nodes and an edge between them denotes that the lender is interested in the corresponding listing. Since the distribution of bids indicates a bias towards highly connected nodes, this in reality means that once a loan listing has a hundred or more lenders bid on it, then that specific listing is more likely to attract more and more lenders. This in turn makes the corresponding listing more likely to get funded in the end due to high lender interest.

A comprehensive analysis of LC loan data by Emekter et al. [7] reveals two key findings:

1. There exists a selection bias in the sense that high-income borrowers with the highest FICO credit scores[5] do not borrow from LC. In particular, top one third of the consumers with respect to FICO scores do not create any loan listings on LC.

2. Higher interest rates charged on the higher risk borrowers are not worth the risk. Specifically, higher rates charged for the borrowers with low LC's own credit grade are not high enough to overcome the greater default risk that the lenders take.

The above two findings imply that, from a profitability point of view, identifying the "good borrowers", i.e., those who will pay back their loan in full within due time, is of great importance for investors participating in social lending. Profitability of social investors, on the other hand, is a critical component in continued interest in social lending

[5]FICO, a publicly traded corporation, produces scoring models that are most commonly used and distributed by TransUnion, Equifax, and Experian.

as well as overall sustainability of the social lending market. In this regard, subsequent to a risk and return efficiency analysis, Emekter et al. [7] suggest that "the lenders would be better off to lend only to the safest borrowers with the highest LC grades". Despite this suggestion, we show in this work that even borrowers with the highest FICO scores or LC grades are not necessarily good borrowers, which in turn indicate that traditional financial score metrics are not well-equipped to capture the non-conventional dynamics prevalent in social lending.

In order to improve identification of good borrowers within the context of social lending, this chapter proposes and presents comparisons of different machine learning methods including random forests (RFs), support vector machines (SVM), logistic regression (LR), and k-nearest neighbor (k-NN) classifiers. Our computational results on LC data between January 2012 and September 2014 for a total loan amount of about one billion USD indicate that random forests outperform the other classification methods and stand as a scalable and powerful approach for predicting borrower status. In fact, an empirical comparison reveals that RFs significantly outperform both FICO scores and LC grades in identification of the best borrowers in terms of low default probability.

The rest of this manuscript is organized as follows: Section 2.2 provides a brief literature review on social lending. In Section 2.3, we introduce basics of social lending, describe the financial features used in prediction, and provide an exploratory data analysis. Section 2.4 presents the classifiers and Section 2.5 provides a comprehensive experimental comparison. Our summary and conclusions are presented in Section 2.6.

2.2 Related Work

An application of machine learning principles in social lending is the use of Gaussian mixture models on the Prosper data set containing loan transactions between November 2005 and December 2008 [8]. An interesting finding therein is that if an individual with a high-risk FICO score belongs to a trusted social community, then this individual's social membership can still help secure a loan. Thus, even though a high-risk credit score usually means lack of access to traditional bank-mediated financial markets, a positive social feature can outweigh a highly negative financial feature in socially mediated markets.

Complex behavioral dynamics further complicate the social lending process. For example, the simple auction mechanism used in some social lending platforms can lead to unpredictable payments for the borrower. An incentive compatible mechanism might be more suitable to eliminate this inefficiency where lenders report their true interest rate and do not change their rate dynamically [9]. Otherwise, such inefficiencies enable users with adversarial interests to use the lending platform as an arbitrage opportunity: borrow at 10% and then loan at 20% [10].

The notion of groups was introduced into social lending with Prosper. Users of this platform can form groups around an affinity that all members share such as a certain topical interest, geographic location, a peer group, or simply around the reasons to borrow. Groups have leaders that act as mediators of loan activity. This mediation can be in the form of pre-evaluating group members, endorsing potential borrowers, inverting and diffusing the risk of a particular group member default among all group members, or encouraging all members to proactively screen new members and apply peer pressure. Empirical studies show that when a group leader in a lending platform mediates the group actively, the risk factor drops considerably. In addition, if a group leader recommends a loan listing put together by one of the group members, this endorsement increases the chance of the loan being issued and also decreases the final interest rate [11].

There exist several studies proposing a set of guidelines in order to make purely rational investment decisions in social lending. In one such study on Prosper loan data that includes loan transactions between November 2005 and March 2007, irrespective of the financial credit rating categories[6], three simple rules help decrease the risk of a default [12]. These investment rules are as follows:

1. Invest only in borrowers without any delinquent accounts.
2. Invest only in borrowers that satisfy Rule 1 and that have a debt-to-income (DTI) ratio less than 20%.
3. Invest in borrowers that satisfy Rule 2 and that have no credit inquiry reports during the last 6 months.

[6]Prosper grades its individual platform users into credit grade buckets in the increasing risk order as AA, A, B, C, D, E, and HR (high-risk) depending purely on credit scores assigned by Experian.

In studies conducted on social communities, herding (denser clustering following a power law regime) effects usually prevail [13–15]. Empirical studies show that the tendency of an individual to join a given community is effected by the number of friends in this community and the inter-connectedness of this individual's friends within the community. Such behavioral bias also exist in investment decisions of lenders at Prosper. The loan data between 2006 and 2008 show that previous lender decisions effected subsequent lender decisions and lender decisions were not made purely rationally [16]. For the interested reader, there exist other real-world networks (such as airports and power grid transmission lines) and other social networks (such as DBLP and LiveJournal) that also exhibit a herding behavior [17, 18].

The closest study to ours is the work of Emekter et al. [7] where the authors analyze LC data between May 2007 and June 2012 and present a logistic regression (LR) model for predicting default probability of a borrower. Their model includes FICO scores as well as LC grades in default prediction. In contrast, our study uses all the available financial features other than the FICO scores and LC grades in order to assess the relevance and prediction power of these two metrics in social lending. Nonetheless, we show in Section 2.4 that one can get much better prediction accuracy using random forests compared to LR even with the exact same features used in building this LR model.

2.3 Overview and Data Analysis

2.3.1 Social Lending Overview

The LC social lending platform works as follows:

- Eligible borrowers with respect to LC's selection criteria create a loan listing.
- LC determines an interest rate for the loan based on the borrower's LC credit grade.
- Lenders have access to borrowers' financial information such as FICO score, LC grade, debt to income ratio, home ownership status, and number of delinquent accounts for their evaluation of potential borrowers. A loan listing includes other

details such as the reason to borrow and relevant demographic information for lenders to review.

- Lenders assume the risk of defaults. Thus, a common lending strategy is to spread out a certain amount of money across a large number of borrowers, thereby reducing the probability of a loss.
- For a loan to be granted, the borrower needs to find enough lenders to cover the entire loan amount.
- A loan listing either expires without enough interest or it gets fully funded before the expiration date.
- LC collects a certain service fee from lenders for any payments they receive from the borrowers.

2.3.2 Explanatory Data Analysis

The data set used in our study was retrieved from LC website LendingClub.com for the period between January 2012 and September 2014. The data contain approximately 350K borrower records with all the credit information normally used by credit agencies to assign FICO scores. Since many of these records are loans that have not been issued or reached maturity yet, and thus do not contain information about the empirical creditworthiness of the borrowers, we filtered the data to only include loans with a status of fully-paid or defaulted. There were about 68K such loans, which correspond to a total loan amount of about one billion USD.

A critical first step in any machine learning application is selection of features[7] with good predictive power of the response variable. LC data contain a total of 35 financial and other borrower-specific features for each loan record. Upon a careful analysis, we selected 23 of these features to be used in predictive modeling. Next, we generated a total of 15 features to be used in our models subsequent to certain data manipulations such as taking ratios of the original attributes, using logarithms to compress exponentially-distributed data, converting dates to time lengths, and eliminating outliers. Data pre-processing and manipulation tasks were conducted using the open-source statistical software R whereas

[7] In this manuscript, the terms feature, attribute, and predictor are used synonymously.

the machine learning tasks were carried out with the open-source machine learning software WEKA. Features used in our predictive models are described below.

- **Loan Status (Response Variable)**: Binary variable indicating whether the borrower defaulted on the loan or fully paid off the loan. For convenience, the defaulted loans shall be referred to as "bad" and fully paid ones as "good" loans.
- **Annual Income**: The annual income information provided by the borrower during registration. Data manipulation: the natural logarithm function.
- **Credit Age**: Date of the earliest credit line opened by borrower, converted to months. Data manipulation: the natural logarithm function.
- **Delinquencies**: The number of delinquencies in the last two years for the borrower. Data manipulation: right-censor ≥ 2. That is, values greater than 2 were set to 2.
- **Employment Length**: Employment length in years. Possible values are integers between 0 and 10 with 0 meaning less than one year and 10 meaning ten or more years.
- **Home Ownership**: The home ownership status information provided by the borrower during registration. The possible values are: Rent, Own, and Mortgage.
- **Inquiries**: The number of credit inquiries in the last six months on the borrower. Data manipulation: right-censor ≥ 3.
- **Loan Amount**: Dollar amount of the loan. This amount cannot exceed $35,000.
- **Loan Purpose**: A category provided by the borrower for the loan request. Possible values are: Debt Consolidation, Home Improvement, Credit Card, Moving, Small Business, Car, Major Purchase, Vacation, Medical, Renewable Energy, House, Wedding, and Other.
- **Open Accounts**: The number of open credit lines on the borrower's credit file.
- **Total Accounts**: The total number of credit lines currently on the borrower's credit file.
- **Term**: The number of monthly payments on the loan. Values can be either 36 Months or 60 Months.

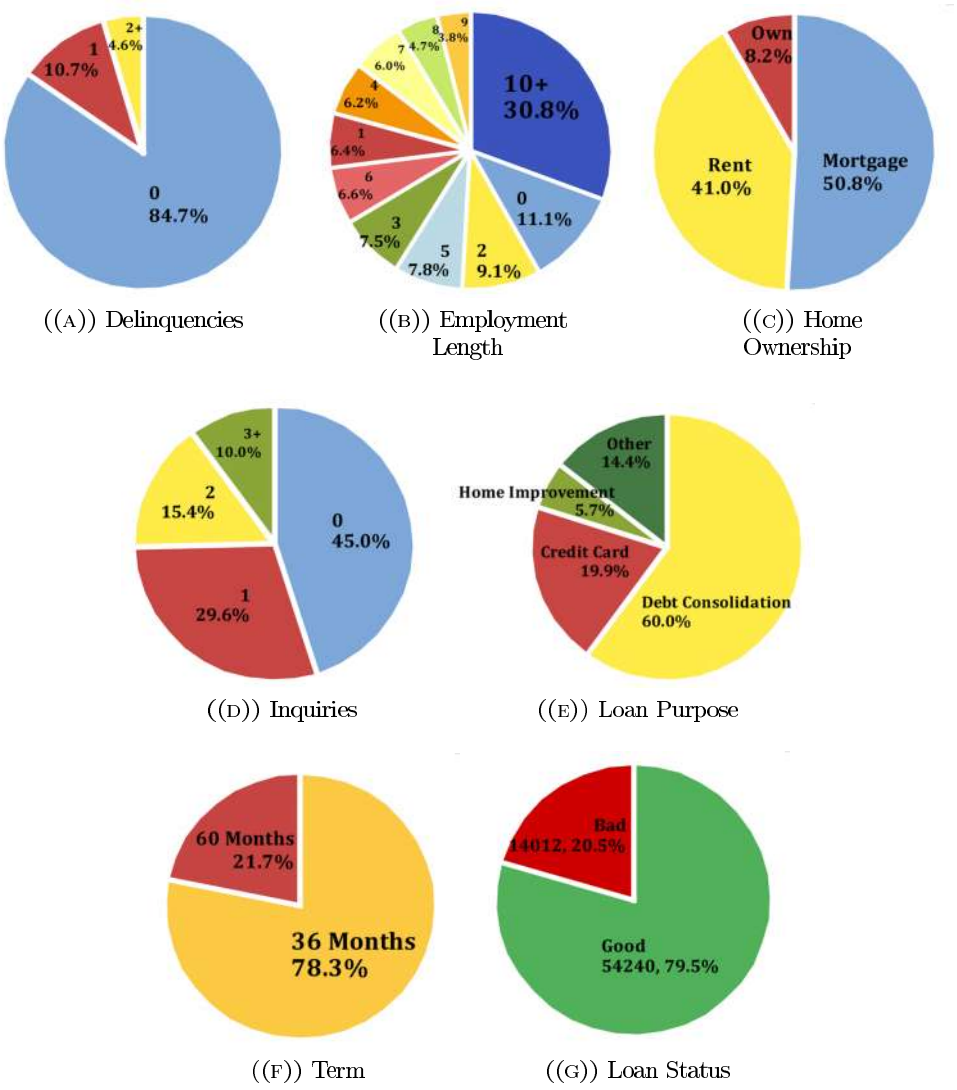
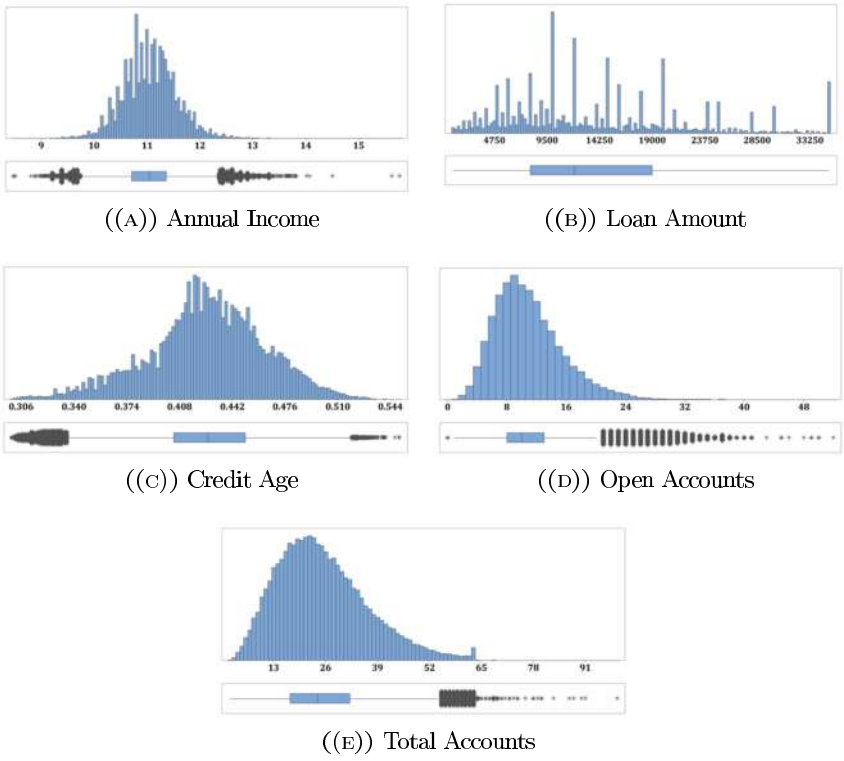


FIGURE 2.1: Pie charts of Delinquencies, Employment Length, Home Ownership, Inquiries, Loan Purpose, Term and Loan Status.

Pie charts of some of the features above are shown in Figure 2.1 whereas histograms and box plots of the other attributes are displayed in Figure 2.2. Next, we describe the features that are simple ratios of other features. These ratios represent a meta-normalization of certain borrower characteristics that otherwise would not be easily captured. These ratio features are explained below and displayed in Figure 2.3.

- **DTI (Debt to Income Ratio)**: Ratio of the borrower's total monthly debt payments to the borrower's monthly income.



((A)) Annual Income

((B)) Loan Amount

((C)) Credit Age

((D)) Open Accounts

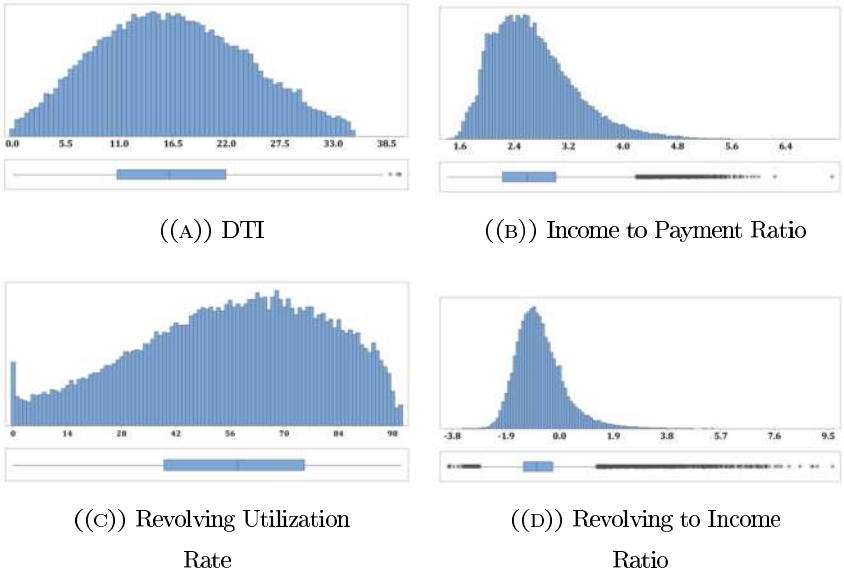
((E)) Total Accounts

FIGURE 2.2: Histograms and box plots of (log) Annual Income, Loan Amount, (log) Credit Age, Open Accounts, and Total Accounts.

- **Income to Payment Ratio**: Ratio of the loan's monthly payments to monthly income. Data manipulation: the natural logarithm function. This is a non-standard financial feature that we introduce in this chapter. The idea behind this feature is that a monthly payment of $500 may be inconsequential for someone who makes $10,000 per month, but it would be a life-changer for someone who makes $1,000 per month. Considered separately, it might be difficult for a machine learning algorithm to capture the essence of this attribute.

- **Revolving Utilization Rate**: The amount of credit the borrower is using relative to all available revolving credit (i.e., credits that do not have a fixed number of payments such as credit cards).

- **Revolving to Income Ratio**: Ratio of revolving credit balance to the borrower's monthly income. Data manipulation: the natural logarithm function. This is another non-standard financial feature that we introduce in this chapter.



((A)) DTI

((B)) Income to Payment Ratio

((C)) Revolving Utilization Rate

((D)) Revolving to Income Ratio

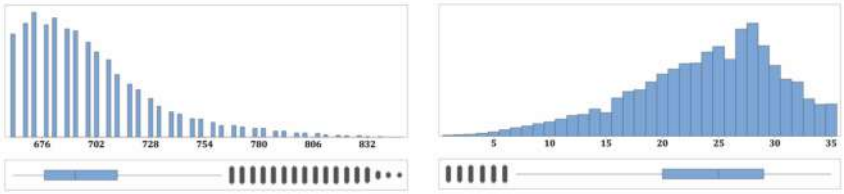
FIGURE 2.3: Histograms and box plots of DTI, Income to Payment Ratio, Revolving Utilization Rate, and Revolving to Income Ratio.

Two additional attributes that are present in the LC data are FICO scores and LC grades, which are described below. Our primary purpose in this chapter is to assess relevance and prediction power of these two attributes regarding borrower risk attribution and therefore they are excluded from our predictive models.

- **FICO Score**: FICO scores are the standard credit scores that are used in majority of the lending decisions in the US. They are calculated from various financial attributes derived from the borrowers' credit records. In the LC data, FICO scores are reported as two numbers in the form of FICO low and FICO high. We average these two numbers and call it the FICO score. Histograms and box plots of the FICO scores are shown in Figure 2.4(a).

- **LC Grade**: LC assigns a grade from A to G to each loan using a proprietary algorithm based on the loan characteristics and risk assessment of the borrower. These seven grades are further divided into 5 levels resulting in loan grades of A1 through G5 with A1 corresponding to the safest loan and G5 corresponding to the most risky loan. LC assigns interest rates to each loan ranging from 7% to 25% such that the interest rate monotonically increases as the loan grade decreases. For convenience, we converted these grades to numbers between 1 and 35 where

35 corresponds to A1 (highest grade) and 1 corresponds to G5 (lowest grade). Histograms and box plots of the LC grades are depicted in Figure 2.4(b).



((A)) FICO Score ((B)) LC Grade

FIGURE 2.4: Histograms and box plots of FICO scores and LC grades.

2.3.3 Statistical Significance of Features

Out of the 15 features we consider, 12 of them are numeric. These attributes were standardized to have zero mean and unit standard deviation before any model building process. The remaining 3 attributes are nominal: Home Ownership (3 levels), Loan Purpose (13 levels), and Term (2 levels). Prior to building the first three machine learning models; namely, k-NN, LR, and SVM; the nominal attributes were binarized, in which case we were left with a total of $12 + 3 + 13 + 2 = 30$ numeric attributes. On the other hand, in the case of RFs, the nominal attributes were untouched.

For a better understanding of the data and the attributes, we computed Information Gain and correlation for numeric and binarized nominal attributes with respect to the Loan Status. These two statistics are displayed in Table 2.1 for the top 15 attributes. The two statistics result in similar rankings with LC grade being the highest followed by Income to Payment Ratio, Annual Income, FICO score, and DTI. Nonetheless, even the LC grade, which solely determines the interest rate for a particular loan, has a correlation less than 0.02 and an Information Gain less than 0.03. We observe in general that, when considered separately, the available attributes are quite weakly correlated with the Loan Status.

TABLE 2.1: Information Gain and correlation with respect to Loan Status.

Rank	Attribute	Information Gain	Correlation
1	LC Grade	0.0294	0.198
2	Income to Payment Ratio	0.0191	0.143
3	Annual Income	0.0126	0.126
4	FICO Score	0.0124	0.125
5	DTI	0.0102	-0.119
6	Revolving Line Utilized	0.0092	-0.111
7	Term	0.0074	-0.104
8	Revolving to Income Ratio	0.0071	0.079
9	Total Accounts	0.0042	0.073
10	Home Ownership: Mortgage	0.0041	0.075
11	Home Ownership: Rent	0.0037	-0.072
12	Loan Amount	0.0026	-0.044
13	Loan Purpose: Small Business	0.0019	-0.054
14	Inquiries	0.0018	0.002
15	Credit Age	0.0017	0.044

2.4 Methodology

2.4.1 The Mathematical Model

In line with the notation in James et al. [19], we model the borrower status prediction problem as follows: Let the binary response variable be denoted by Y such that

$$Y = \begin{cases} -1, & \text{if Bad Borrower} \\ +1, & \text{if Good Borrower.} \end{cases}$$

Our mathematical model is given as

$$Y = f(\mathbf{X}) + \epsilon$$

where $\mathbf{X} = (X_1, \ldots, X_p)$ is the feature vector, p is the number of features, and ϵ is the irreducible error term in the model that captures measurement errors and other noise in the data. Note that even though the function f is assumed to exist, it is almost never known in practice. In addition, even if f is known, there would still be prediction errors

due to the fact at each $\mathbf{X} = \mathbf{x}$, there is a distribution of possible Y values in general. Thus, the function f is defined as

$$f(\mathbf{x}) := E(Y|\mathbf{X} = \mathbf{x})$$

where the right-hand-side (RHS) is the expected value of Y for a particular realization $\mathbf{x}$ of the feature vector $\mathbf{X}$. Rather than assuming a binary value, $f(\mathbf{x})$ specifies the probability of a borrower being a good borrower, which can be expressed as

$$f(\mathbf{x}) = E(Y|\mathbf{X} = \mathbf{x}) = Pr(Y = 1|\mathbf{X} = \mathbf{x}) \tag{2.1}$$

An important property of $f(\mathbf{x})$ is that it is the function that minimizes $E[(Y-g(\mathbf{X}))^2|\mathbf{X} = \mathbf{x}]$ over all functions g at all points $\mathbf{X} = \mathbf{x}$. That is,

$$f(x) = \arg\min_g E[(Y - g(\mathbf{X}))^2|\mathbf{X} = \mathbf{x}]$$

Our goal in this chapter is to find a good estimate of $f(\mathbf{x})$, which we denote by $\hat{f}(\mathbf{x})$. For any $\hat{f}(\mathbf{x})$, it holds that

$$E[(Y - \hat{f}(\mathbf{X}))^2|\mathbf{X} = \mathbf{x}] = [f(\mathbf{x}) - \hat{f}(\mathbf{x})]^2 + Var(\epsilon)$$

where the first RHS term denotes the reducible error and the second RHS term denotes the irreducible error that is inherent in Y. Thus, the learning problem at hand is to find a good estimate $\hat{f}^*(\mathbf{x})$ that minimizes the reducible error, i.e., $[f(\mathbf{x} - \hat{f}(\mathbf{x})]^2$. We consider a total of four different classifiers for this task, which are described below.

2.4.2 k-Nearest Neighbors (k-NN)

The algorithm of k-Nearest Neighbors (k-NN) is a very simple, yet popular and powerful non-parametric method widely used in classification. The inputs to the algorithm are the k nearest training instances to the test instance with respect to a certain distance function, typically with a small k between 1 and 10. The classification is based on a majority vote of these k nearest neighbors. It might sometimes be helpful to weight the neighbors' contributions such that the closer neighbors have more effect on the final decision that the farther ones [20]. In our setting, the estimate $\hat{f}_{k-\text{NN}}(\mathbf{x})$ of k-NN can

be expressed as

$$\hat{f}_{k-\mathrm{NN}}(\mathbf{x}) := \mathrm{Majority}(Y|\mathbf{X} \in \mathcal{N}_k(\mathbf{x}))$$

where *Majority* denotes the majority vote function and $\mathcal{N}_k$ denotes the k closest neighbors of $\mathbf{x}$ with respect to the Euclidean distance in the $p-$dimensional space. In our implementation, we take $k = 1$ for simplicity and convenience.

2.4.3 Logistic Regression (LR)

A (simple) linear regression estimator $\hat{f}_{\mathrm{L}}$ is defined as a linear combination of the individual attributes in the form

$$\hat{f}_{\mathrm{L}}(\mathbf{x}) := \sum_{i=0}^{p} \beta_i x_i = \boldsymbol{\beta}'\mathbf{x} \ (x_0 = 1)$$

where the estimates $\hat{\beta}_i$ of β_i are computed via a least-squares fit for $i = 0, \ldots, p$ using the observations in the training data. A more appropriate technique for binary classification is the logistic regression estimator below:

$$\hat{f}_{\mathrm{LR}}(\mathbf{x}) := \frac{e^{\boldsymbol{\beta}'\mathbf{x}}}{1 + e^{\boldsymbol{\beta}'\mathbf{x}}}$$

In LR, the estimates $\hat{\beta}_i$ are calculated using the maximum likelihood method [21, 22]. A useful property of LR is that it guarantees an output value between 0 and 1 that can be interpreted as class-conditional probabilities in classification problems.

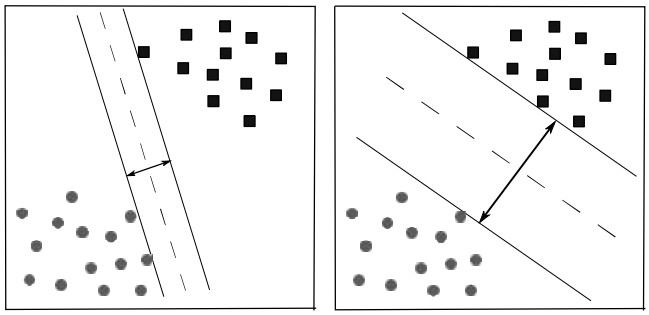
2.4.4 Support Vector Machines (SVM)

For a data set with a binary response variable, Support Vector Machines separate the data into two regions (one for each class) in the $p-$dimensional feature space via a hyperplane with the maximum margin width between instances of the two classes [23]. This process is illustrated in Figure 2.5 for two different cases in two-dimensional space. Maximizing the margin width decreases the complexity of the model and the overall risk of errors. When the data is not separable by a hyperplane, which is usually the case in practice, a soft margin is used. In this situation, a positive slack is added to the instances on the wrong side of the margin. This slack increases proportional to how

far the corresponding instance is from the margin. The goal is to minimize the sum of these slacks while maximizing the width of the margin. A regularization parameter C governs the relative cost of each objective in the optimization process, which is taken as 1.0 in our implementation. The SVM optimization problem can be stated as a quadratic programming problem in the dual form as

$$\max_{\alpha_l}\left(\sum_{l=1}^{n}\alpha_l + \frac{1}{2}\sum_{l,r=1}^{n} y_l y_r \alpha_l \alpha_r K(x_l, x_r)\right)$$

subject to the constraints $0 \leq \alpha_l \leq C$ for training instances $l = 1$ to n and, $\sum_{l=1}^{n} y_l \alpha_l = 0$ where α_l is the Lagrange multiplier associated with the instance l. In this chapter, we employ the quadratic polynomial kernel $K(x_l, x_r) = (1 + x_l.x_r)^2$. We chose SVMs as a candidate classifier as they haven been used successfully in similar applications before [24–26].



((A)) Small margin ((B)) Large margin

FIGURE 2.5: Separating hyperplanes with small and large margins in SVM.

2.4.5 Random Forests (RF)

Decision tree learning is a popular classification method that grows a tree-based structure with class-conditional probabilities at the end of the tree branches. The decision tree starts with a root node and gradually builds sub-trees with internal nodes that are connected by emanating branches and ends with terminal nodes called leaves. Each internal node corresponds to a test of a feature (e.g., a borrower owns a house or otherwise) and branches represent a binary partition of the test attribute. The process of building a decision tree is a divide-and-conquer approach in the sense that the root node corresponds to the entire training data and each node split corresponds to a partitioning of the

available data at that node based on the test condition for the associated feature. There are two critical issues in decision tree learning: (1) how to choose the split attribute at each internal node and (2) how many levels to have at each tree branch, i.e., when to stop splitting. Within the context of random forests, which are collections of decision trees, splitting is done with respect to *Gini Index*, which is described below, and number of levels in each tree branch is controlled by an algorithm parameter d [27].

The Gini Index at an internal tree node is calculated as follows: For a candidate (nominal) split attribute X_i, denote possible levels as $L_1, \ldots, L_J$. Gini Index for this attribute is calculated as

$$G(X_i) := \sum_{j=1}^{J} Pr(X_i = L_j)(1 - Pr(X_i = L_j)) = 1 - \sum_{j=1}^{J} Pr(X_i = L_j)^2.$$

Once Gini Indices are computed for each candidate split attribute, the split is done on the attribute that has the highest Gini Index.

Decision tree classifiers have several attractive properties: they are easy to interpret, they can handle both numerical and nominal data, and they are easy to build. Nonetheless, decision trees are not always competitive with other classification techniques. Thus, in order to improve the accuracy of trees, one sometimes need to employ ensemble methods such as boosting (iterative learning from misclassified instances) and bagging (building multiple trees and combining the results). Random Forests (RFs), which can be seen as an enhanced bagging technique, is a powerful method for constructing a forest of random decision trees. RFs de-correlate the decision trees in the forest via randomization of split attributes that leads to an improvement over traditionally bagged trees and reduces the variance when averaged over the trees [27]. RFs also build multiple decision trees on bootstrapped training samples. However, while building these trees, the candidate split attributes in each tree are chosen by a random selection of m attributes from the full set of p attributes, as illustrated in Figure 2.6. The split is allowed to use only one of these m attributes and a fresh selection of m attributes is made at each split. In each tree, splitting is continued until the tree reaches a depth of d. In our implementation of RFs, we chose a forest size of 80 with $m = 5$ and $d = 25$. Fine-tuning of these parameters is described in Section 2.4.8.

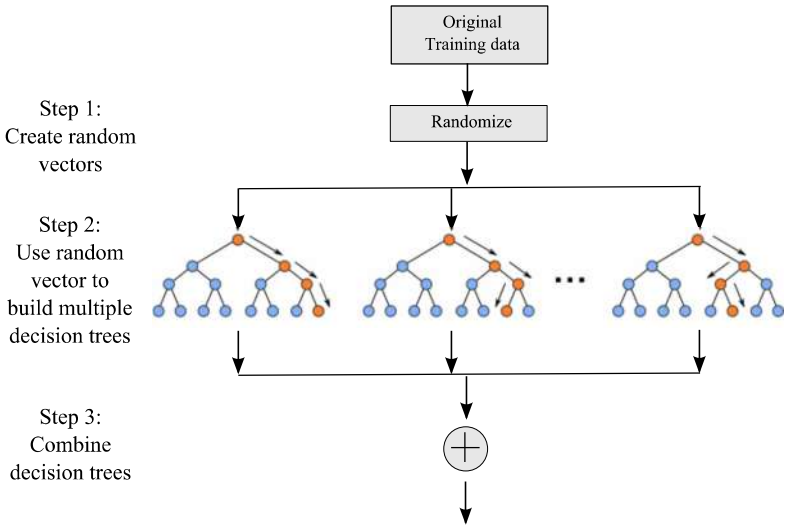


FIGURE 2.6: Demonstration of the Random Forest methodology.

2.4.6 Cost Sensitive Analysis

Since accepting a bad borrower carries much greater risk than rejecting a good borrower, we use a weighted cost matrix in our models to increase the cost of misclassification associated with bad borrowers. As suggested in Schebesch and Stecking [28], we performed all our experiments in this chapter with a 5-to-1 cost ratio such that misclassification of a bad borrower (as a good borrower) is 5 times more costly than misclassification of a good borrower (as a bad borrower). For this purpose, we used the *CostSensitiveClassifier* meta-function within WEKA.

2.4.7 Model Evaluation and Assessment

For evaluation of the candidate classifiers, we employed the popular 5-fold cross-validation (CV) that has been shown to provide a good trade-off between model over-fitting and under-fitting in general [29]. Thus, the data set of 68K loans was first partitioned into five equally-sized slices. For each one of the 5 CV folds, one slice was set aside for testing and the remaining 4 slices were used for training the classifiers. This process is illustrated in Figure 2.7. CV specifically enables us to test the performance of a model on every instance in the available data set without having used it in the training phase.

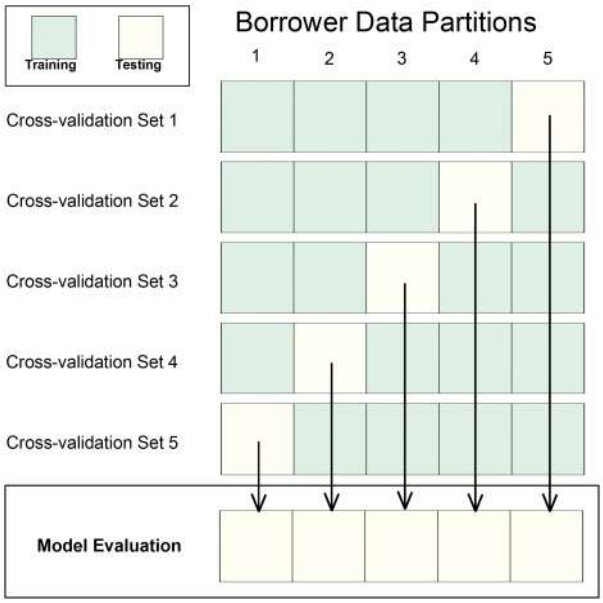


FIGURE 2.7: Demonstration of 5-fold cross validation.

Regarding performance assessment of the classifiers, we took an average over the 5 folds in CV of the following 7 performance metrics: (1) overall classification accuracy rate on the test slice in the fold, (2) the usual area under the Receiving Operating Characteristic (ROC) curve (AUC) for the good class in the test slice, (3) root mean square error (RMSE), (4)-(7) the confusion matrix: true positive (TP) and false positive (FP) rates for the good and bad classes respectively in the test slice.

2.4.8 RF Parameter Fine-Tuning

This section investigates the effects of forest size, the number of split features m, and the maximum tree depth d on the performance of RFs in our problem so that we can identify the optimal values for these parameters. For simplicity and convenience, this fine-tuning process was conducted in two steps: (1) determining the optimal forest size and, (2) determining the optimal m and d values. First, by using the "auto" option in WEKA for m and d, we constructed forests with sizes ranging from 1 to 500 trees in increments of 10 trees. The comparison results are shown in Figure 2.8 where we observe that building trees beyond 80 did not result in considerable additional performance, yet increased the run time considerably. Thus, we settled on a forest size of 80 as a reasonable trade-off between execution time and classification performance.

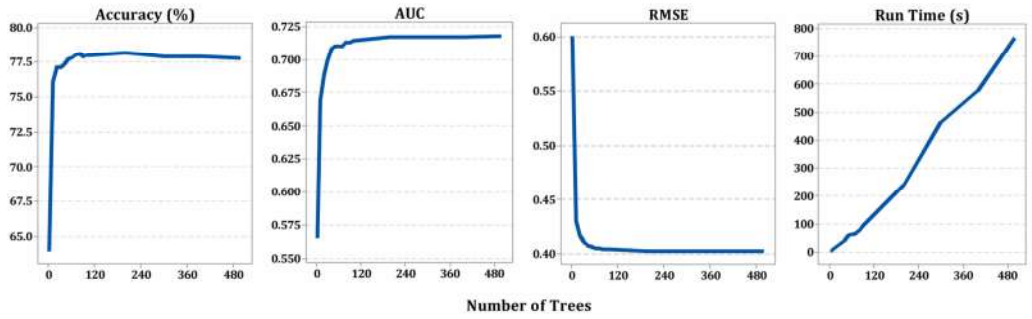
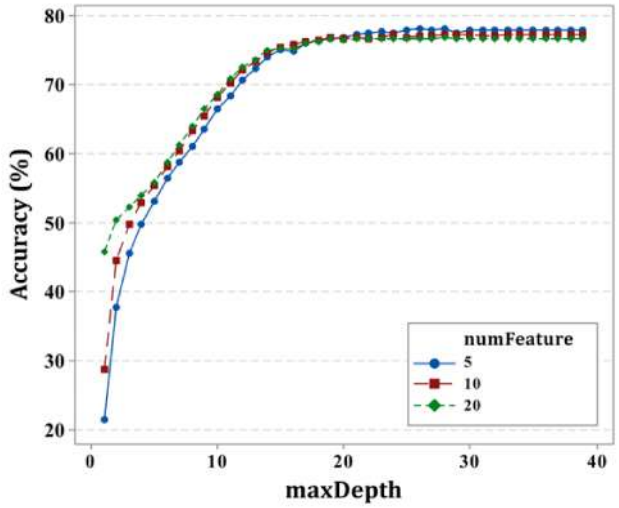
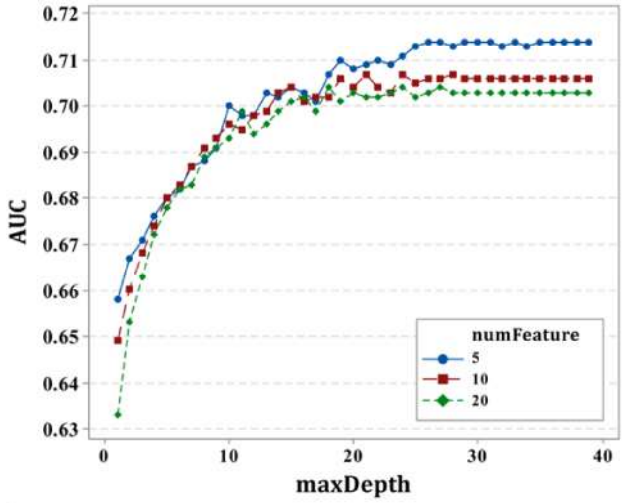


FIGURE 2.8: RF performance with respect to the forest size. Run times shown correspond to the total time for the 5-fold CV process for each size.

The suggested value for the number of split features m is $\lceil \log_2 (p) \rceil$ [27], which in our case corresponds to 5. In order to identify a good value for m, we considered $m = 5, 10, 20$ and for each m value, we experimented with d values ranging from 1 to 40. For each m, d combination considered, we used a forest size of 80. Experimental results shown in Figure 2.9 indicate that (1) 5 split features does indeed exhibit the best performance and (2) accuracy rates and AUCs practically flatten out after a tree depth of 25 regardless of the m value. Thus, in our implementation, we settled on a tree size of 80 with $m = 5$ and $d = 25$.



((A)) Accuracy



((B)) AUC

FIGURE 2.9: RF performance with respect to number of split features and maximum depth.

2.5 Experimental Results

2.5.1 Comparison of the Classifiers

This section presents comparisons of the classifiers on the LC data as evaluated via 5-fold CV. The results are shown in Table 2.2 and Figure 2.10 respectively. Overall, we observe that RFs have the highest accuracy rate at 78.0% and the highest AUC at 0.71 with the

lowest RMSE at 0.42. Whereas k-NN has the second highest accuracy rate at 70.1%, LR has the second highest AUC at 0.68. Based on these results, we declare RFs to be the best overall classifier compared to the other three alternatives.

TABLE 2.2: Performance comparison of the classifiers. AUC denotes area under the ROC curve for the good class, RMSE the root mean square error, TP true positive, and FP false positive respectively.

Rank	**Classifier**	**Accuracy**	**AUC**	**RMSE**	**TP Rate**		**FP Rate**	
					Good	Bad	Good	Bad
1	Random Forest	78.0%	0.71	0.42	0.88	0.31	0.69	0.13
2	Nearest Neighbor	70.1%	0.53	0.55	0.82	0.25	0.74	0.18
3	Support Vector Machine	63.3%	0.62	0.68	0.47	0.78	0.22	0.53
4	Logistic Regression	54.5%	0.68	0.51	0.49	0.77	0.23	0.51

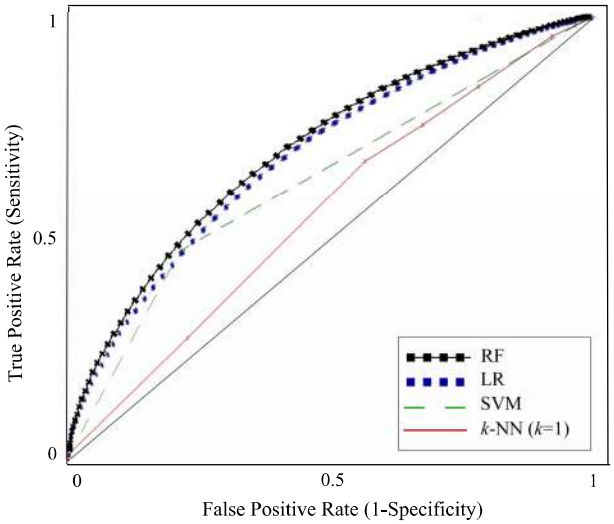


FIGURE 2.10: Comparison of ROC curves for different classifiers.

2.5.2 Comparison Against FICO Scores and LC Grades

We now compare the performance of RFs against FICO scores and LC grades. Our methodology in this comparison is to compute thresholds for all three metrics that result in identical acceptance rates and then compare the ratio of the number of defaults to the number of loans corresponding to that acceptance rate. For instance, if we were to accept only borrowers with a FICO score above 750, we would be accepting about 8% of borrowers. This specific subset of accepted borrowers has a default rate of 8.2%.

Then, we can perform a threshold sweep over the RF scores to find the threshold that corresponds to exactly 8% acceptance rate. The RF classifier has only 3.1% of the borrowers defaulting in this subset. In this way, we can declare RF scores to be superior to FICO scores at predicting creditworthiness in the top 8% acceptance rate.

Comparison results for acceptance rates ranging from 0.1% to 20% for all three metrics are shown in Figure 2.11. Within the 3% acceptance rate, we see that RFs do not misclassify any bad borrowers as good borrowers. In particular, we observe that up to the 10% acceptance rate, RFs outperform the other two metrics in identification of good borrowers. However, this superior performance of RFs for the top 10% of the borrowers comes at the cost of misclassifying good borrowers (as bad borrowers) outside of this acceptance rate. Specifically, beyond the 10% acceptance mark, even though RFs are still superior to FICO scores, LC grades show better performance.

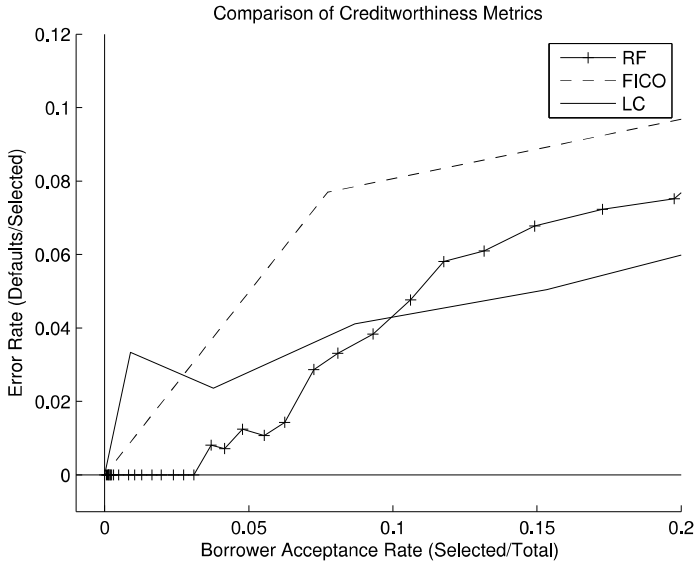


FIGURE 2.11: Comparison of creditworthiness metrics in terms of error rate vs. borrower acceptance rate.

2.5.3 Comparison Against Existing Methods

This section compares our RF methodology against the current state-of-the-art in LC loan status prediction in the literature, which is the logistic regression model in Emekter et al. [7]. This model advocates utilization of only four attributes: FICO score, LC score,

DTI, and Revolving Utilization Rate. As before, we trained the 4-feature LR with a 5-to-1 cost ratio for fairness. Table 2.3 shows that all-attribute RF (except FICO scores and LC grades) significantly outperforms the 4-feature LR of Emekter et al. [7] with an accuracy of 78.0% vs. 51.0%. In particular, the 4-feature LR's misclassification of bad borrowers (as good) is 56% whereas this ratio is only 13% for all-feature RF.

We also considered the question if one would be better served by using RFs with the above 4 features as opposed to LR. This comparison is also given in Table 2.3. It can be seen that 4-feature RF achieves an overall accuracy of 69.8% compared to 51.0% accuracy of 4-feature LR. In addition, the 4-feature RF's misclassification rate for bad borrowers is 20% compared to 56% rate of the 4-feature LR. Thus, we conclude that even with the same 4 features considered in Emekter et al. [7], RFs exhibit superior performance compared to LR.

TABLE 2.3: Comparison of RF and LR with the features FICO score, LC grade, DTI, and Revolving Utilization Rate against RF with all the feature (except FICO scores and LC grades). The first row is taken directly from Table 2.2.

Classifier	**Accuracy**	**AUC**	**RMSE**	**TP Rate**		**FP Rate**	
				Good	Bad	Good	Bad
(All-Feature) Random Forest	78.0%	0.71	0.42	0.88	0.31	0.69	0.13
(4-Feature) Random Forest	69.8%	0.61	0.45	0.79	0.32	0.68	0.20
(4-Feature) Logistic Regression	51.0%	0.65	0.52	0.45	0.76	0.24	0.56

2.6 Chapter Recap

In order to compute the risk score of an individual, financial features such as past financial history, existence of delinquent accounts, debt to income ratio (DTI) and various other financial features are used. In this chapter, we present an RF-based methodology for identification of good borrowers in social lending using the world's largest social lending platform LendingClub.com's publicly available historical records. We introduce non-standard financial features in order to increase reliability of the computed risk scores, and we propose and present a comparison of the machine learning methods RF, SVM, LR, and k-NN for identifying good borrowers in social lending. Our computational results indicate that RFs outperform the other classifiers as well as the FICO scores and LC grades in predicting good customers.

Chapter 3

A Statistical Analysis of Istanbul Strait Anchorage Traffic Between 2006 and 2014

3.1 Introduction

The Strait of Istanbul is the only sea route between Mediterranean, Aegean and the Black Seas and it is one of the busiest waterways in the world. The Strait divides the City of Istanbul into European and Asian parts and makes the city a significant logistics node in the entire region [30]. Commercial vessels have the legal right to pass freely through the two straits of the Marmara Sea, namely, Istanbul and Dardanelles Straits, and drop anchor at their north and south sides during peacetime [31]. However, lack of alternative routes and the bottleneck shape of these two straits, along with global shipping development and limited anchorage capacity of Turkish anchorages, result in heavy traffic and occasional accidents. These complications for commercial ships in these straits necessitate consistent and advanced traffic control and anchorage management. And the fact that we don't have any data of the vessels' departures make it even harder to manage the anchorages. Therefore, a comprehensive analysis of traffic flow of vessels and clarifying the patterns of anchorages seems necessary. The results would create a better understanding of traffic and anchorage behavior with the aim of developing and shifting the policies into an optimized state. In this regard, and to fill this critical void,

implementing machine learning and data mining frameworks on recorded data regarding all the vessels anchored in the Turkish anchorages, one can provide a useful description of the variables effecting a vessel's anchorage.

In 1970s, the field of Marine traffic engineering was introduced by Toyoda and Fuji, with the aim of improving marine traffic regulations and better performance in navigation facilities which was followed by a number of academicians for several years [33–41]. Although after the 1990s relevant studies reduced significantly, recently optimizing anchorage utilization and anchorage planning especially during peak periods has been the center of focus and debate. In this regard, Bijwaard and Knapp [42] by exploiting duration analysis and generating ship life cycles aimed to improve assessments of ship lives with the purpose of reducing the possibility of incidents. The capacity of multiple anchorages was evaluated by Huang et al. [43] using a simulation-based model, and some methods regarding improvement of space utilization were proposed accordingly. In 2013, concerning optimal navigation of ships while obstacles are in the way, a graph theoretical resolution was presented and applied on an ice navigation case study [44]. And Silveira et al. [45] analyzed the risk of collision near the ports of Portugal by developing an algorithm using the available data. Moreover, there are some researches focusing on the improvement of marine traffic management specifically in Istanbul Strait using several strategies like offering a mathematical formulation of the current scheduling [46], proposing a specific navigation safety support model [47], suggesting Local Traffic Separation Schemes (LTSS) [48], evaluating the performance of an online Precise Point Positioning (PPP) service for positioning in Halic Bay [49], and using generic fuzzy analytic hierarchy methods for risk evaluation in Istanbul Strait [50]. However, in order to have an effective and optimal strategy for anchorage planning and achieving the best maritime traffic model, knowing the patterns of anchorage main factors for different category of vessels is a dominant factor. An aspect which all these investigations lack mostly because of insufficient data and restricted utilization of data mining approaches.

Although no investigation has yet been published concentrating on the analysis of Istanbul Strait anchorages as well as anchorage duration, there are few researches executed machine learning and data mining algorithms in order to analyze the traffic data and extract the information required to manage the marine traffic flow. Among these, two different classification techniques were applied on a data of ship arrivals at a port for a period of one week in order to predict their future locations [51]. Using clustering

and statistics, a data mining platform was presented in order to have a proper prediction of marine traffic flow Tang and Shao [52]. Tsou [53] employed association rule discovery method for the data gathered for the sea area of Keelung Harbor regarding navigation conditions. And a clustering algorithm along with three neighborhood models were employed in order to detect vessel traffic areas in the Shanghai Strait Oo et al. [54].

This manuscript is mainly concerned with providing useful information regarding vessel anchorage behaviors especially in Istanbul Strait Anchorages. This investigation is based on available data consisted of several anchorage related attributes. Applying statistical analysis on this data, which mainly consists of anchorage reason, type, length, flags and date of anchorage recorded for the last nine years (from 2006 to 2014), we describe and illustrate the behavior and relations of these variables. Afterwards, an association analysis discussed relation between zone and reason of anchorage in Istanbul Strait Anchorages. Finally, by applying two famous data mining approaches, linear regression and decision trees, we investigate the predictability of gross tonnage and anchorage duration separately.

3.2 Statistical Analysis

In order to have a good management and planning, one should have a general understanding of the historical data. In this regard, we preform a comprehensive statistical analysis for each attribute as well as all possible relations between them. The following section describes the nature and structure of the ship anchorage data in Istanbul Strait.

3.2.1 Resource Data

The available data recorded by Turkish Directorate General of Coastal Safety (DGCS) includes historical records of 13 attributes related to the anchorage of vessels from 2006 to 2014 in the anchorages of Istanbul. There are 443339 observations in the data set with both categorical data such as ship type, flag and anchorage reason, as well as numeric attributes such as anchorage duration, length and gross tonnage of ships.

3.2.2 Structure of the Data

As mentioned, the data contains information related to ships anchored in a specific anchorage of Istanbul Strait. In order to have a better understanding of this data, some details of these variables are reported in Table 3.1, and discussed in more detail in this section.

TABLE 3.1: Description of the parameters.

Attribute	Type	Number of possible values/ Range
Reason	Nominal	5
Zone	Nominal	3
Year	Ordinal	9
Month	Ordinal	12
Ship Type	Nominal	73
Flag	Nominal	126
Arrival Country	Nominal	165
Departure Country	Nominal	200
Arrival Port	Nominal	1778
Departure Port	Nominal	1397
Length (m)	Numeric	14.9 - 328.56
Gross Tonnage	Numeric	40 - 129325
Duration (h)	Numeric	0.2 - 2651.88

3.2.2.1 Reason, Zone, Month and Year of Anchorage

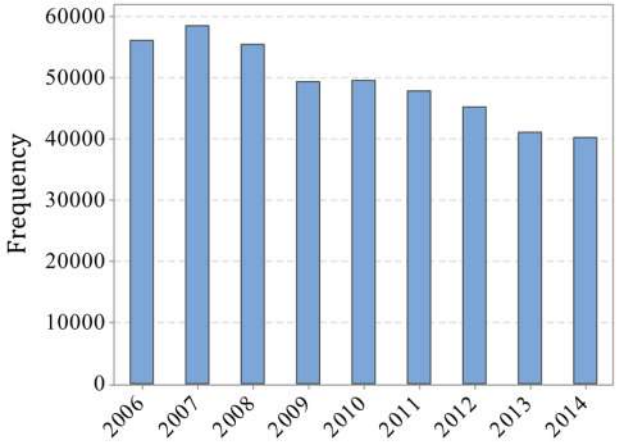
Different types of vessels anchor in Istanbul anchorages for different reasons, such as waiting for instructions from owners or authorities (planning), bunkering and supply, port operations, and rough weather conditions. These anchorages are divided into three major areas according to their geographical position: Southern (locally known as Ahırkapı or Guney), Northern (Kuzey) and Eastern (Kartal). In Tables 3.2 and 3.3, and Figure 3.1, detailed information about classes and frequencies of four attributes of reason, zone, month and year recorded through nine consecutive years, from 2006 to 2014, is given. From Table 3.2, it can be inferred that more than 90% of anchorages were because of planning and supply, and Table 3.3 shows the ratio of anchorages in the zones of Southern, Eastern and Northern are approximately 35:1:14. Moreover, the number of anchorages were slightly decreasing through these years (Figure 3.1(a)), which may relates to the economic and policy changes of Turkey and other countries like Russia. According to the chart of anchorage month (Figure 3.1(b)), there is no significant drop in any month, which states no different policies for different times of a year.

TABLE 3.2: Anchorage distribution by anchorage reason

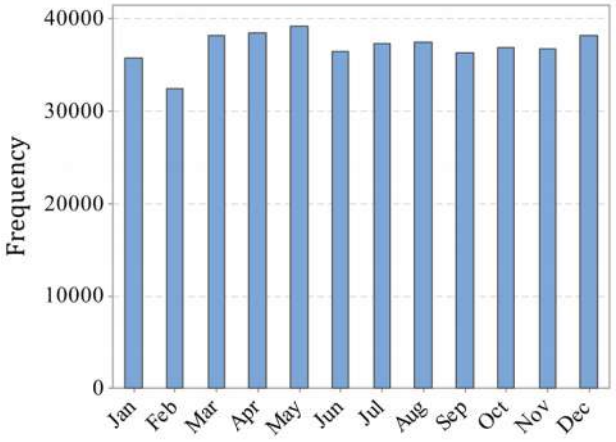
Reason of Anchorage	Number of Anchorages	Percent	Duration Amount (h)	Percent
Planning	258432	58.29	3493539.688	57.76
Port	13181	2.97	167780.32	2.77
Supply	156176	35.23	1856045.70	30.69
Weather	10919	2.46	350497.785	5.80
Other	4631	1.04	180220.61	2.98
Total	443339	100.00	6048084.10	100.00

TABLE 3.3: Anchorage distribution by anchorage zone

Zone of Anchorage	Number of Anchorages	Percent	Duration Amount (h)	Percent
Eastern	7790	1.76	124328.471	2.06
Northern	120937	27.28	1621102.650	26.80
Southern	314612	70.96	4302652.98	71.14
Total	443339	100.00	6048084.10	100.00

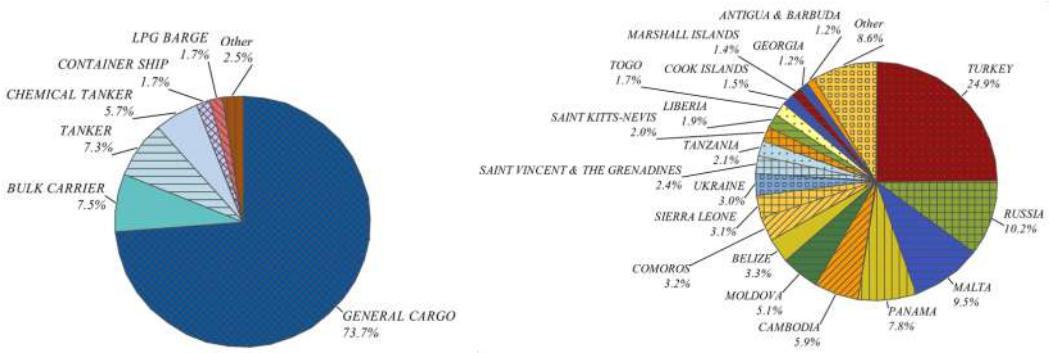


((A)) Year of Anchorage



((B)) Month of Anchorage

FIGURE 3.1: Charts of year and month



((A)) Ship Type ((B)) Vessel Flag

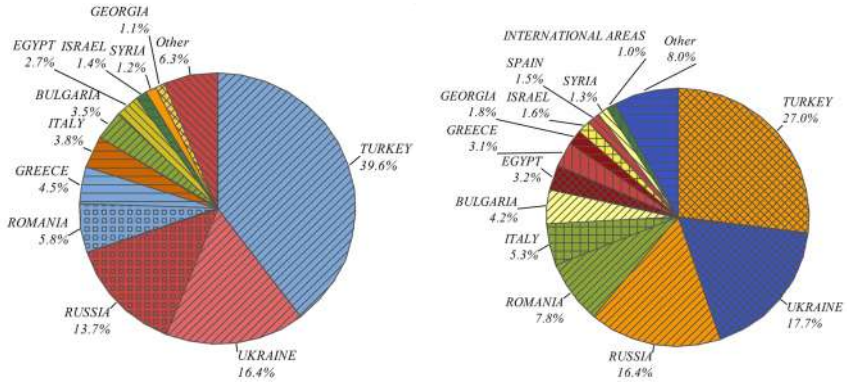
FIGURE 3.2: Pie charts of ship type and flag according to the frequency of their categories.

3.2.2.2 Ship Type and Flag

Commercial vessels due to their purpose and size have various categories, which in this region, general cargo, bulk carrier, tanker, chemical tanker, container ship and LPG barge are the most common ship types. Moreover, each vessel is distinguished with a flag, demonstrating its country of registration, and the vessels are required to follow the rules of its flag state. According to our data base for the last nine years, the flags of Turkey, Russia, Malta and Panama constituted more than fifty percent of the flag states of the whole anchored vessels. In Figure 3.2, more information is displayed about these two variables, and the ratio of their categories those with more than one percentage frequency is illustrated.

3.2.2.3 Departure and Arrival Port and Country

The other important information recorded in our data base is the specific port and country a vessel departed from and planned to arrive at, which considering the unique location of the Strait of Istanbul, could reveal a considerable portion of sea transportation in the region. As demonstrated in Figure 3.3, most transports are associated with the countries of Turkey, Ukraine, Russia and Romania.



((A)) Departure Country ((B)) Arrival Country

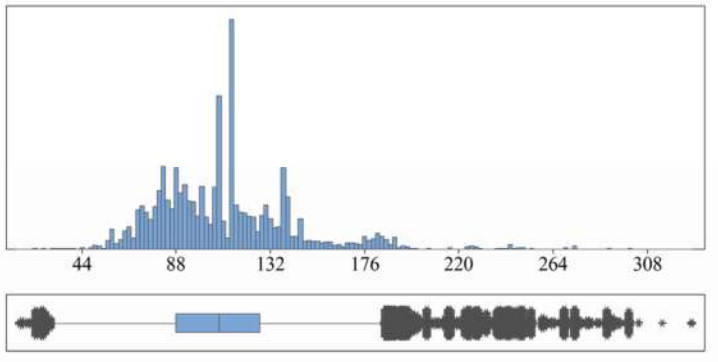
FIGURE 3.3: Pie charts of departure and arrival country with the frequency of their categories.

3.2.2.4 Length and Gross Tonnage

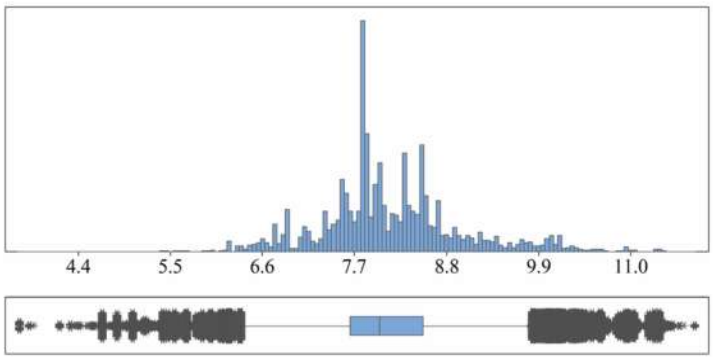
Cargo vessel size is mainly associated with two features of ship length and tonnage. The former is the distance between the first and the last part of a vessel, and the latter refers to the cargo volume of a ship, mostly expressed as Gross Tonnage (GT) which is a linear function of the tonnage. These measures are generally used to evaluate the cost of shipping and berth as well as the safety requirements. So, based on our data, and considering the importance of these two attributes, a detailed statistics regarding ship length and gross tonnage is reported in Figure 3.4. According to this figure, although the ranges are large, their common values seems restricted.

3.2.2.5 Anchorage Duration

As discussed before, the time a vessel intends to stay in an anchorage is a critical parameter for planning and scheduling the anchorage specifics. Fortunately, we have access to this valuable measure through our data base and information about anchorage duration of a total number of about 444 thousand cases is demonstrated in Figure 3.5. Based on this data, commercial vessels anchor in the anchorages of Istanbul Strait with the average duration of around half a day.



((A)) Histogram and boxplot for Length



((B)) Histogram and boxplot for Gross Tonnage (log)

FIGURE 3.4: Summary reports for length and gross tonnage of the anchored vessels

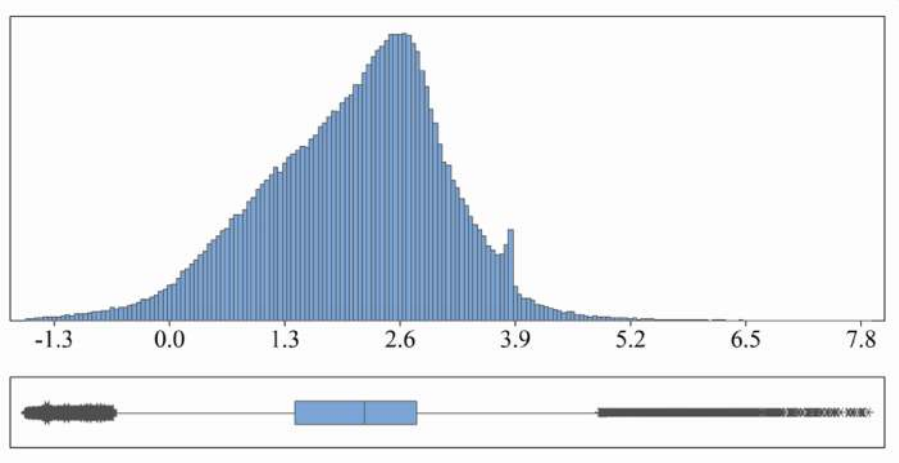


FIGURE 3.5: Histogram and boxplot for anchorage duration of the vessels (log)

3.2.3 Group Comparisons

It wouldn't be realistic to assume that there is no connection and dependency between the mentioned parameters concerning ship anchorage. In this regard, using cluster and line plots in this section and association analysis in section 3.3, we aim to analyze and investigate possible relations between different attributesin more detail.

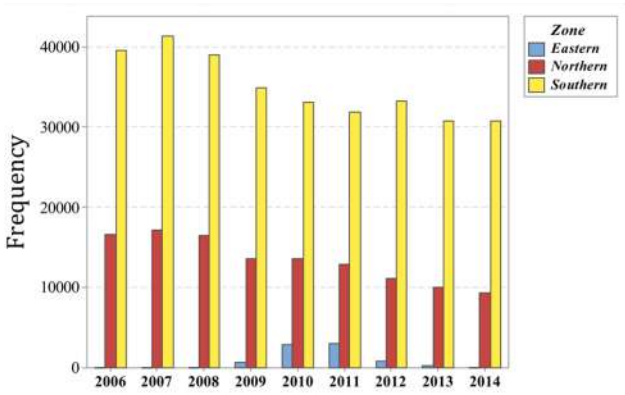
As illustrated in Figure 3.6, some noticeable patterns between parameters of reason, zone, year and month of anchorage exist. According to the chart of 3.6(a), the constant decrease of anchorages exists in both Northern and Southern zones, while most anchorages in the zone of Eastern occurred in 2010 and 2011, and the service reduced afterwards. Moreover, our data do not contain information about anchorages in Eastern anchorage zone in 2014. Due to the chart of 3.6(b), Northern zone is mostly associated with anchorage reason of planning, while Southern zone experience all types of anchorages with different reasons. From the plot of 3.6(c), it can be inferred that, as expected, the anchorages as a consequence of rough weather condition were significantly less from April to August.

Furthermore, there are some associations between departure and arrival countries and zone of anchorage; Data suggests that most anchorages in Northern zone were for vessels departed from some Eastern European countries like Russia, Ukraine and Romania, and almost all anchorages in Northern zone were related to vessels departed from Turkey.

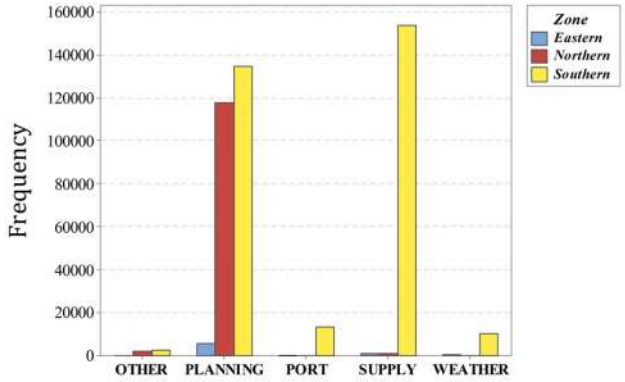
In anchorage planning, vessel size as well as ship type are the dominant parameters in the optimal utilization of berth locations. In this regard, Figure 3.7 depicts boxplot of vessel length for different ship types, where clearly there is a connection between Length and type of a vessel.

We display the relation between vessel length, and reason and zone of anchorages in Figure 3.8. While the anchorage reason of rough weather condition is related to the vessels with shorter lengths, larger vessels anchored mostly because of supplying. Also, Eastern zone has been operating since 2009 in order to decrease the heavy traffic in Southern zone, and it seems Eastern zone has been considered for anchorage of short ships; That's probably the reason of higher length mean in Southern zone after 2008 compared with Northern zone.

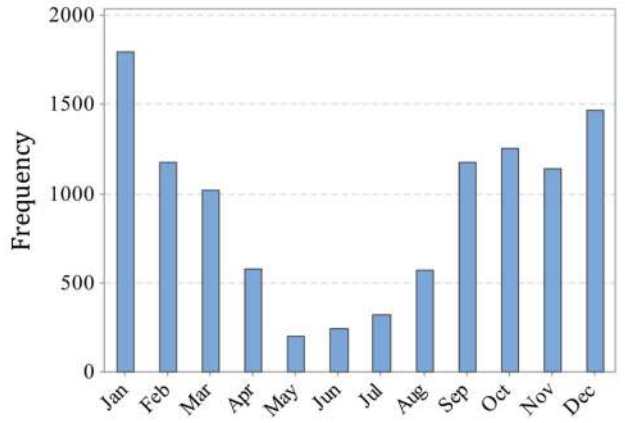
We perform same analysis on gross tonnage of vessels, however obtaining same trends as ship length made it unnecessary to present them. Similar results for length and gross



((A)) Year and Zone



((B)) Reason and Zone



((C)) Weather and Month

FIGURE 3.6: Plots of reason, zone, year and month of anchorage

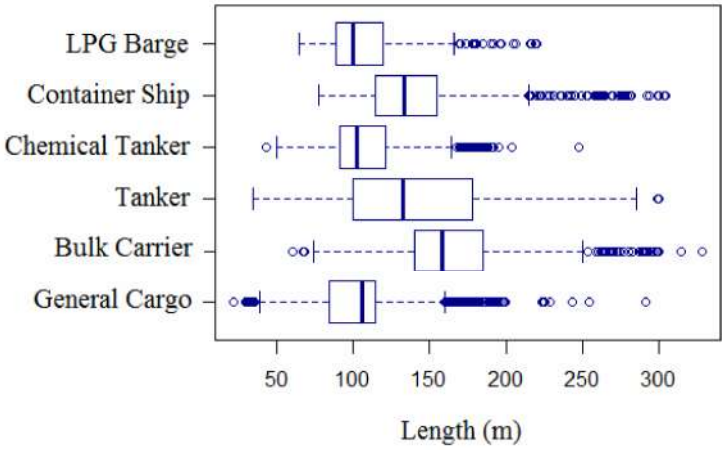
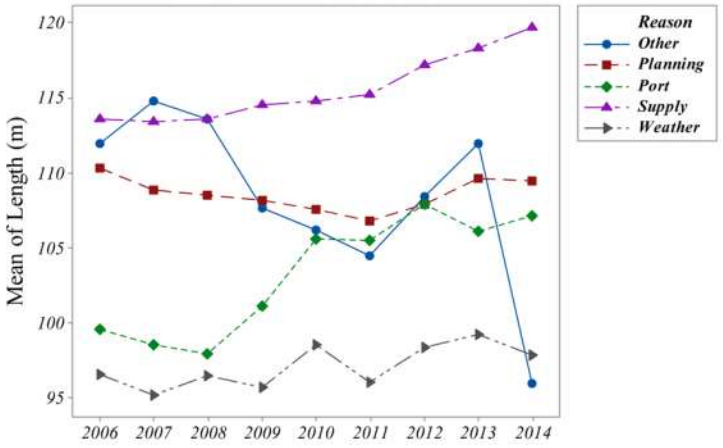
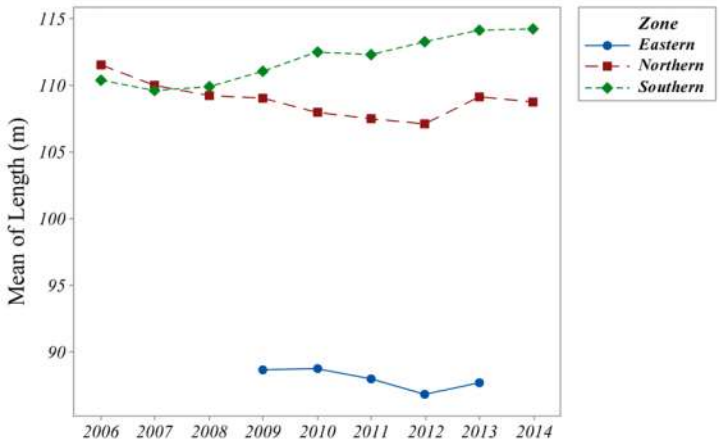


FIGURE 3.7: Boxplot of vessel length for different ship types



((A)) Length vs. year and reason



((B)) Length vs. year and zone

FIGURE 3.8: Line plots of ship length according to reason, zone and year.

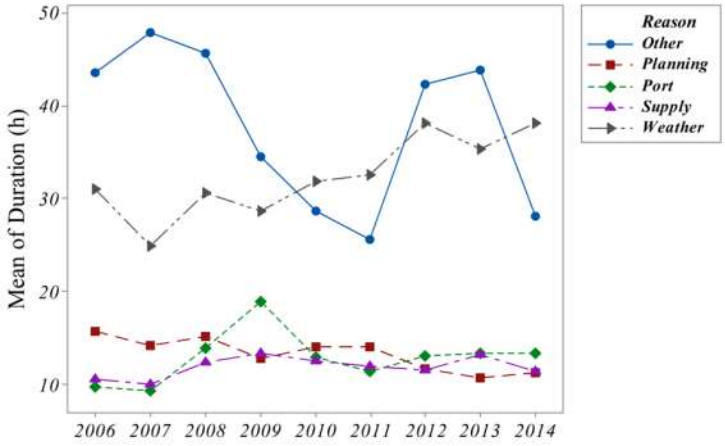
tonnage necessitates more investigation regarding the possible relation between these two representatives of vessel size. In this regard, linear regression analysis is performed to clarify their association.

Mean anchorage duration of vessels regarding the attributes of reason, zone and year is shown in Figure 3.9. Although reasons of planning, port and supply have a close and stable duration mean through the time, the anchorages because of rough weather condition and other causes have irregular trends. Actually, the variations of duration mean for rough weather condition can be explained by atmosphere changes through these nine years, however irregular alterations of other causes cannot be clarified, due to the fact that the exact reasons weren't recorded specifically. Unlike Southern zone, Northern and Eastern zones do not have a constant rate of mean duration, which along with the unknown causes of anchorages can make the duration prediction even more complex.

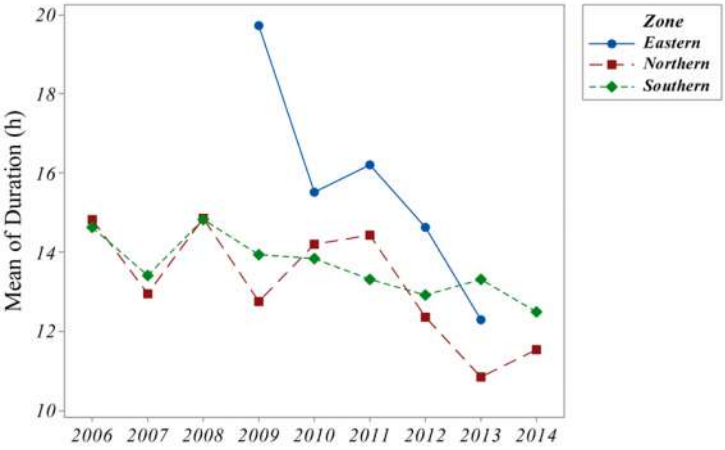
3.3 Association Analysis

Association analysis is one method for generating helpful knowledge and mining variables associations for better understanding of data features as well as constructing predictive models for regular incidents. The other well-known term of this investigation is Market Basket Analysis, as it originally was developed to analyze behavior of customers in markets in order to discover the products customers purchase at the same time, and develop better sale strategies accordingly.

In this chapter, we perform association analysis of reason and zone of ship anchorage to discover the relationship between the cause of berth and the region they anchor. Thorough this, an improved ship navigation system and traffic management can be reached. Each class of these two variables is called an itemset, where the expression of $X \rightarrow Y$ implies the association rule between them. The relative frequency of X and Y together is their support (counts of transactions containing X and Y over the total counts of transactions), and certainty measure of the acquired rule is termed confidence (counts of transactions containing X and Y over the counts of transactions containing X). As both terms of support and confidence should be high enough in order to have



((A)) Duration vs. year and reason



((B)) Duration vs. year and zone

FIGURE 3.9: Line plots of anchorage duration according to reason, zone and year.

representative rules with high accuracy, usually a minimum threshold would be assigned for them to disqualify unimportant rules.

3.4 Prediction of Anchorage Duration

Data mining is the process of extracting valuable information from a set of data. This computational process contains discovering previously unknown patterns from large amount of data using mathematical, statistical and computer science procedures, and transforming it into a comprehensible structure. Besides using a specific data mining software like R, Weka and Rapid Miner, there are numerous data mining techniques, each of which has their own advantages and disadvantages and should be selected according to the

nature of parameters and constraints exist in the data set. The most regular techniques are artificial neural networks, decision trees, Naïve Bayes and Nearest-neighbor. Each of these methods analyzes the set with a different tactic.

As mentioned before, one of our goals here is to provide an approach for estimating the duration of ship anchorage by presenting the factors responsible for the variation of this factor. This section briefly reviews the concepts and techniques of machine learning and data mining used in this chapter in order to analyze the data.

3.4.1 Data Transformation

In order to have a nominal structure of the overall data, we transform three numeric parameters of length, gross tonnage and duration into several bins using proper discretization methods. For our independent attributes of length and gross tonnage, we perform discretization with equal entropy which resulted in 20 and 17 classes respectively. Our response variable, anchorage duration, was divided into five intervals with equal frequency, presented in Table 3.4.

TABLE 3.4: Discretized duration with five intervals.

Representation	Description	Range (h)
VS	Very Short Duration	0 - 3.5
S	Short Duration	3.5 - 7.0
M	Medium Duration	7.0 - 11.5
L	Long Duration	11.5 - 18.5
VL	Very Long Duration	18.5 - ∞

3.4.1.1 Dimension Reduction

Having large number of attributes in the data set not only will take too much time and hard drive memory, but also it might reduce the accuracy of our prediction and make it difficult for us to interpret the model. In this regard, in the current section, finding the most important and effective parameters and removing the rest from the evaluation without sacrificing any important information is described and justified.

Various attribute ranking and attribute selection methods have been proposed in the data mining literature, with the aim of discarding unrelated or redundant parameters from a given data set, like Information Gain, Gain Ratio, symmetrical uncertainty, one-R

and Chi-square test. So as to investigate the effect of each attribute on anchorage time, in this chapter, we perform attribute evaluator of Information Gain for feature evaluation and attribute ranking.

In order to identify redundancy of some attributes, chi-square test was executed on each pair of nominal variables, as well as correlation test for numeric ones. Applying these approaches, those pairs with high correlation and very little p-values are considered as attributes with high association, one of which can be removed from our training set without losing any valuable information.

3.4.2 Prediction Models

As discussed earlier, numerous data mining approaches are available to generate a suitable model for an existing data set. In this chapter, according to the nature of the data, i.e. categorical parameters and large amount of observations, we select three famous classifiers of Decision Tree, Nearest Neighbor and Naïve Bayes, and after finding the best model for each method, we compared their performances accordingly. The following sections contain brief explanation of each of these machine learning techniques.

3.4.2.1 Decision Tree

Decision tree learning is one of the common and visual techniques of data mining which employs the predictive model of decision tree by mapping an item's observations to decisions about objective value of the item. A decision tree is a tree where each interior node is labeled with an input parameter, and the exiting lines from the node are marked with possible classes of the other features, leading to several leaves representing a class label of the objective variable.

Decision tree classifier has several advantages compared to other machine learning approaches; it is easy to comprehend and interpret, capable of handling both numerical and categorical data, able of performing proper classification for large datasets in reasonable time, and requires little data preprocessing. However, having too many leaves might generate complex trees with lower accuracy, termed overfitting. Therefore, in order to avoid this problem, some mechanisms known as pruning are employed to get rid of those

problematic leaves, which decrease the complexity of the ultimate classifier as well as improve the predictive accuracy.

3.4.2.2 Nearest Neighbour

The approach of k-Nearest Neighbors (k-NN) is a non-parametric technique for classification and regression applications. The k-NN algorithm receives the k nearest training instances in the space and give the result based on a majority vote of these k neighbors. Typically, k is small and mostly between 1 and 10, also it may be helpful to consider some weights for the neighbors' contributions such that the closer neighbors have more impact on the final decision.

3.4.2.3 Naïve Bayes

In data mining, Naïve Bayes classifier is one of the probabilistic classifiers in regard to employing Bayes' theorem with the assumption of strong independency between the variables. In other words, it considers the significance of a specific feature being isolated from the occurrence or absence of any other attribute, regarding the class variable. In spite of this oversimplified assumption, this classifier operates reasonably well in many complex cases, and it simply requires a small training data for the classification.

3.4.3 Performance Criteria for Model Evaluation

Model evaluation is an essential step in the procedure of developing the final model. Using this performance evaluation, we can compare the possible established models, and find the best one with highest accuracy, as well as assessing the acceptability of a finalized model. There are several approaches for the evaluation, such as finding re-substitution error, and Hold-out method. In the former, the assessment is performing on the training data used for the model generation, and the latter is using a test set different from the training set, in order to evaluate the performance of the model.

3.5 Results and Discussion

This section utilizes the approaches defined previously, and consists the results of linear regression, association analysis, and evaluation of different methods for duration prediction.

3.5.1 Regression analysis Results

Similar results for length (L) and gross tonnage (GT) can be explained by Figure 3.10, where scatter plot of these two variables are illustrated, and as expected, these two representatives of ship size have a considerable association with each other. Also considering a two degree polynomial regression, which is displayed in the figure with red line, we achieved an equation with the accuracy of 87.58% to demonstrate their association (Equation 3.1)

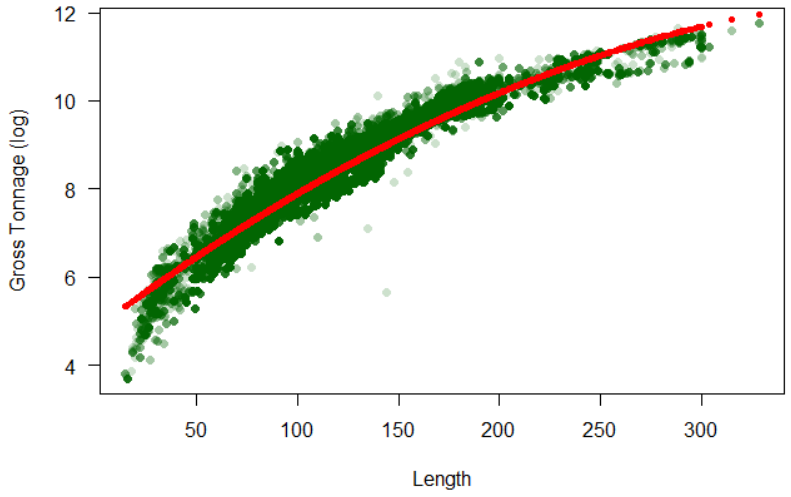


FIGURE 3.10: Scatter plot of gross tonnage versus ship length

$$log(GT) = 4.83 + 0.0345 \times L - 0.000039 \times L^2 \quad (R^2 = 0.8758) \tag{3.1}$$

3.5.2 Association Analysis Results

Regarding the figure illustrating the relation between reason and zone (Figure 3.6), we can infer their strong association with each other; most anchoring in Northern and Eastern zone is because of Planning, and anchoring in Southern zone includes mostly Planning and Supply. To quantify these observations, we perform an Apriori association analysis (minsup threshold: 0.25 or minconf threshold: 0.75) on these two attributes, and the best rules found are reported in Table 3.5.

TABLE 3.5: Association rules for variables of reason and zone.

	Association Rule	Support	Confidence
1	{Port} → {Southern}	0.03	1
2	{Supply} → {Southern}	0.34	0.98
3	{Northern} → {Planning}	0.27	0.97
4	{Weather} → {Southern}	0.02	0.91
5	{Eastern} → {Planning}	0.01	0.75
6	{Planning} → {Southern}	0.30	0.52
7	{Southern} → {Supply}	0.34	0.49
8	{Planning} → {Northern}	0.27	0.46
9	{Southern} → {Planning}	0.30	0.43

3.5.3 Classification Results

3.5.3.1 Data Preprocessing Results

With the aim of comparing the dependency of anchorage duration with other attributes, we employed the evaluator of Information Gain, and the result of its attribute ranking is presented in Table 3.6. From this Table, we can infer that the parameter of reason has a very strong relation with the duration of anchorage over other attributes, which even alone can be considered as the sole estimator of duration extent. Furthermore, attributes of flag and ship type are among the lowest values revealing their small association with our dependent variable, which suggests removing them from our training set.

With the aim of reducing the number of attributes, association inspection is performed through correlation test and Chi-square analysis. The former resulted in high correlation of 0.931 for two attributes of length and gross tonnage, and the latter revealed that reason and zone have p-value less than 0.005, expressing their high association. These results were expected according to the results of linear regression and association analysis explained in previous sections. According to the higher Information Gain measure of

TABLE 3.6: Attributes ranked by Information Gain

Rank	Attribute	Information Gain
1	Reason	0.153
2	Zone	0.039
3	Arrival Port	0.031
4	Departure Port	0.031
5	Month	0.015
6	Arrival Country	0.012
7	Departure Country	0.011
8	Gross Tonnage	0.008
9	Year	0.008
10	Length	0.008
11	Flag	0.008
12	Ship Type	0.003

gross tonnage than length, we keep gross tonnage as a better representative of these two categories. Knowing the importance of reason parameter in the prediction, excluding the attribute of zone from our training set seems reasonable as well.

Moreover, having attributes with large number of classes, like departure and arrival ports in our data, might make the final model less effective and less accurate. The typical solution is to combine their classes and generate higher level grouping. In fact, in this case, there exists two other parameters with higher grouping named departure and arrival countries, therefore these two can be a good representation for departure and arrival ports.

Consequently, these analyses lead to six important features, presented in Table 3.7.

TABLE 3.7: Variables determined as significant.

Rank	Variable	Explanation
1	Reason	Cause of anchorage
2	Month	Month the ship anchored
3	Arrival Country	Country the vessel arrived at
4	Departure Country	Country the vessel departed from
5	Gross Tonnage	Representation of cargo volume and ship size
6	Year	Year the ship anchored

3.5.3.2 Evaluation of predictive models

We evaluate the effect of pruning in Decision Tree and number of neighbors in Nearest Neighbor using two accuracy evaluators, namely re-substitution and hold-out method. Afterwards, in order to have a better insight of our method performances, we compare the prediction accuracy of these three classifiers accordingly.

Re-substitution and Hold-out accuracies of Decision Tree classifier is compared with two other methods of Nearest Neighbor, and Naïve Bayes displayed in Table 3.8. It can be inferred by increasing the number of neighbors in k-NN, re-substitution accuracy decreases because the nearer points have more similar classes, however Hold-out accuracy improves effectively, which can be the result of reducing the effect of noises in the classification procedure. According to these results, for re-substitution method pruning the data will decrease the accuracy excessively, and the reason is missing some data because of the pruning. However, concerning the Hold-out method pruning increases the accuracy due to error reduction behavior of pruning procedure. And more importantly, it is shown that among these five, Decision Tree has the best performance, and is recommended as the best classifier for both approaches.

TABLE 3.8: Prediction accuracies

Method	Re-substitution accuracy	Hold-out method accuracy
Nearest Neighbor (IB1)	0.525	0.274
Nearest Neighbor (IB5)	0.475	0.350
Naïve Bayes	0.345	0.353
Decision Tree (unprunned)	0.750	0.310
Decision Tree (prunned)	0.480	0.380

3.6 Chapter Recap

In this chapter, we present a comprehensive statistical analysis of a new data set on Istanbul Strait Anchorages. We provide descriptive statistics on attributes including flag, arrival and departure country, length and tonnage of the vessel as well as zone, reason, month, and year of the anchorage. Next we explore statistical relationship between these attributes. Using linear regression, we find high correlation of length and gross tonnage of vessels, and through an association analysis we show a meaningful relation between reason and zone of anchorage. Regarding the prediction of anchorage duration, we compare different classifiers and we infer that decision tree is the best method. We also show that "reason of anchorage" is the dominant attribute among the whole variables for duration prediction of vessels in these anchorages.

Chapter 4

A Temporal Analysis of Vessel Type Traffic in Istanbul Strait Anchorages

4.1 Introduction

Business and transportation in today's world are going through repeated improvement and reorganization requiring dynamic management and planning. The fact that more than 90% of the world's trade is seaborne underlines the importance of sea transportation management in general. On the other hand, the Istanbul Strait, one of the busiest waterways in the world and the only sea route between the Mediterranean, Aegean, and the Black Sea, is a logistic node in the region that necessitates constant attention [30]. In this regard, previous studies concentrated on the improvement of marine traffic management in the Istanbul Strait via a number of methodologies including the following: proposing a mathematical formulation for maritime scheduling [46], suggesting a particular navigation safety support model [47], offering local traffic separation schemes [48], performance evaluation of the service of an online precise point positioning (PPP) system with the aim of positioning in Halic Bay [49], and applying the method of generic fuzzy analytic hierarchy for evaluating maritime risks [50]. On the other hand, a maritime accident analysis in the Southern Anchorage Area can be found in Aydogdu et al. [55].

Tens of thousands of vessels pass through the Istanbul Strait every year and a large portion of them berth in the Strait's anchorage areas for a period of time for various reasons such as bunkering and supply, port operation, lay-up, waiting for Strait passage,

and waiting due to bad weather conditions (in this work, berthing and anchoring are used interchangeably). Turkish Maritime Authorities do not allow the vessels to anchor arbitrarily as there exist different zones for different types of vessels. For instance, Figure 4.1 shows the specific zones for different types of vessels in the Southern Anchorage Area where vessels are shown as squares. This zoning separates the vessels with dangerous cargo, long stays, and harbor approach categories from each other, mainly because each type require different management and strait passage planning. Therefore, a good estimate of future vessel type traffic is a key parameter for managing the anchorage areas in an effective manner. In this regard, original research contributions of this manuscript are as follows: (1) We present a statistical analysis of vessel types in the most recent eight years (between 2006 and 2013) using a new historical dataset not available in the literature before. (2) We provide an application of the statistical forecasting methodology of Autoregressive Integrated Moving Average (ARIMA) [56] to predict vessel type traffic over the next three years in the Strait anchorages. We remark a similar forecasting methodology was performed on monthly converted traffic data for ports in Korea with the aim of estimating future traffic volume by using arrival vessel data per tonnage [57]. Our goal in this exercise is to assess whether a change in anchorage area zones is necessary in the near future for better anchorage zoning as well as more efficient marine traffic management in the Istanbul Strait.

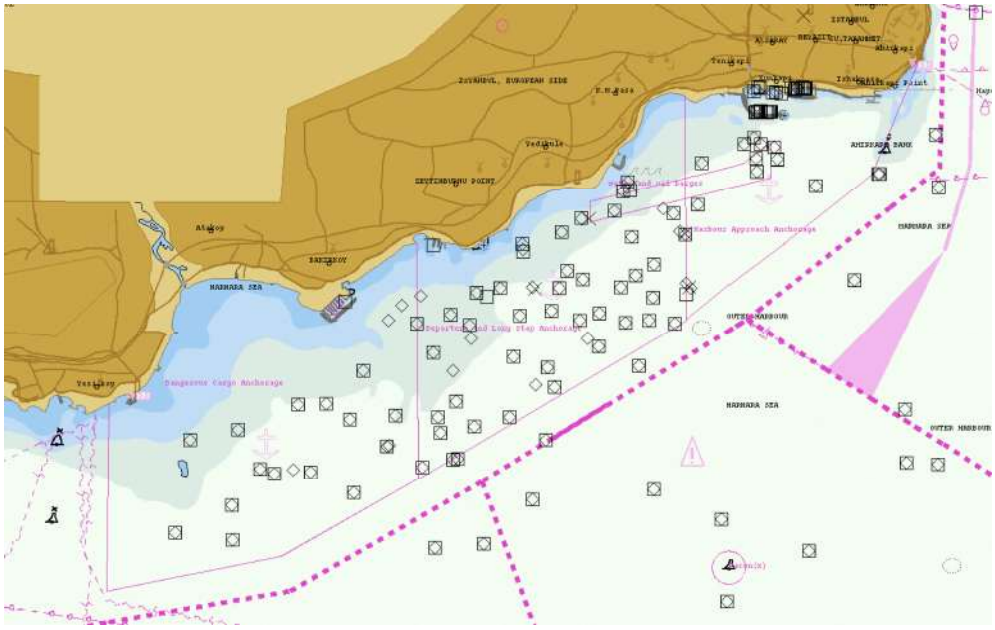


FIGURE 4.1: Anchorage zones for different vessel types in the Southern Anchorage Area

4.2 Methodology

The historical data that was made available to us by the Turkish Directorate General of Coastal Safety contains historical information on a number of attributes associated with each vessel anchored in the Istanbul Strait anchorage areas from 2006 to 2013 with close to half million vessel records. Using the commercial Minitab Software, we performed a temporal analysis of this data, including the frequency of different vessel types, association of vessel types and length, and a time series analysis of vessel types.

We use the powerful ARIMA model and the popular open-source statistical programming language R to predict the number of vessels for each vessel type anchoring in Istanbul berth regions for the upcoming three years. The ARIMA model is specified as ARIMA (p, d, q) where parameters of p, d, and q are non-negative integers referring the order of autoregressive, non-seasonal differences, and lagged forecast errors respectively. For a given time series, we employ the classic AICc (Akaike Information Criterion with Correction for Finite Values) metric to determine the optimal values of the p, d, q parameters (all else equal, the model with the smallest AICc value is chosen). We use the auto.arima function in R for this optimization task. AICc is a measure based on the likelihood function allowing us to compare the relative performance of competing statistical models for a given dataset. We also report the BIC (Bayesian Information Criterion) metric for each forecast.

4.3 Results and Discussion

4.3.1 Results of Statistical Analysis

In the Istanbul Strait, there are three anchorage areas: Southern (locally known as Ahirkapi or Guney), Northern (Kuzey) and Eastern (Kartal). The vessel types we considered and their dangerous cargo status are as follows:

1. General Cargo (non-dangerous)
2. Bulk Carrier (non-dangerous)
3. Container Ship (non-dangerous)

4. (Product and Crude Oil) Tanker (dangerous)

5. Chemical Tanker (dangerous)

6. LPG Barge (dangerous)

Other vessel types respectively constitute less than 1% of the anchorage traffic and therefore they are not considered in further detail. Table 4.1 shows the combined distribution of vessel types across 2006 and 2013 and Figure 4.2 demonstrates a yearly combined percentage breakdown of the same data. It can be inferred from this figure the total number of anchored vessels was moderately decreasing throughout these years, which could be related to the global financial crisis in 2008 as well as policy and/or economic changes in Turkey and other countries using the Strait for maritime transportation.

TABLE 4.1: Number of vessels anchored across different types

Vessel Type	**Year**							
	2006	***2007***	***2008***	***2009***	***2010***	***2011***	***2012***	***2013***
General Cargo	40272	43424	41159	36871	37246	35701	33008	29764
Bulk Carrier	4113	3660	3666	3693	3496	3483	3968	3519
Tanker	4846	4859	4206	3189	3345	3294	2887	2621
Chemical Tanker	3337	3283	3158	3234	2758	2607	2686	2255
Container Vessel	1044	1056	1049	650	854	860	681	667
LPG Barge	675	627	592	623	766	832	1055	1232

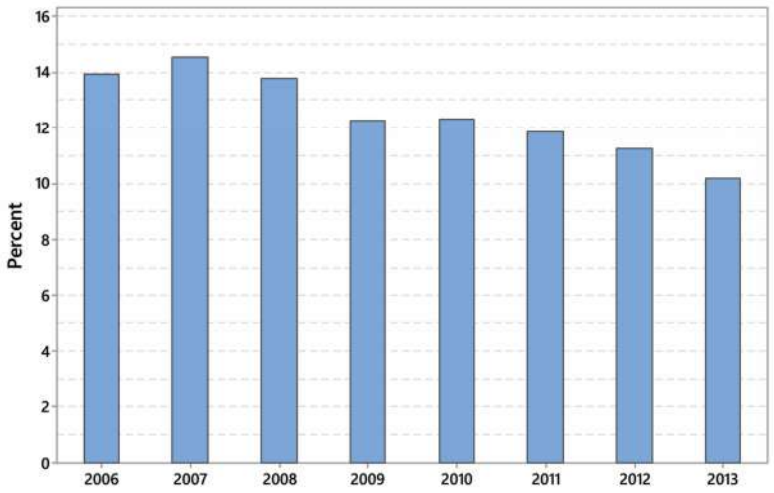


FIGURE 4.2: Percentage of vessels anchored in the Istanbul Strait

Figure 4.3 shows vessel type ratios with more than 1% frequency. We observe that vessels classified as general cargo constitute almost three quarters of the anchorage traffic.

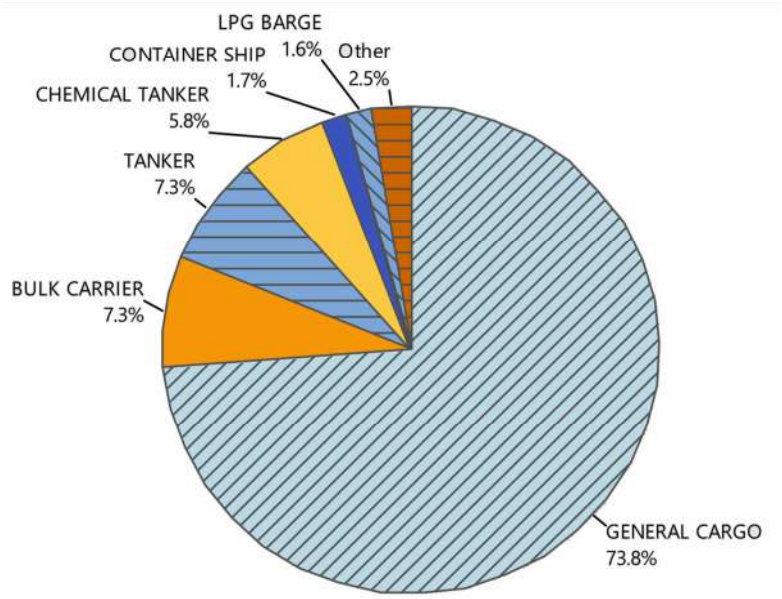


FIGURE 4.3: Pie chart of vessel type traffic breakdown between 2006 and 2013

In anchorage planning, besides vessel type, the size of the vessels play an important role, as it is the dominant parameter in the optimal utilization of berth locations [58]. In this regard, Figure 4.4 depicts 95% vessel length confidence interval (CI) plots for different vessel types. These intervals, which were calculated by their respective standard deviations, indicates vessel lengths and vessel types are in a good association with each other, which not only implies the importance of vessel type analysis, but also justifies zoning practices in the anchorage areas.

A yearly breakdown of relative traffic for each vessel type is shown in Figure 4.5. The figure suggests there is a somewhat steady decrease in general cargo vessels and tanker ships in general. Yet, bulk carrier traffic seems to be relatively stable. On the other hand, LPG barge type vessels show a steady and relatively significant increase in these eight consecutive years.

4.3.2 Times Series Forecast

Subsequent to a performance comparison against several competing forecasting models, the ARIMA model was chosen in order to forecast vessel type traffic in the three anchorage areas over the next three years. Table 4.2 presents the yearly forecasts for each vessel type along with optimal model parameters, AICc, and BIC metrics in addition

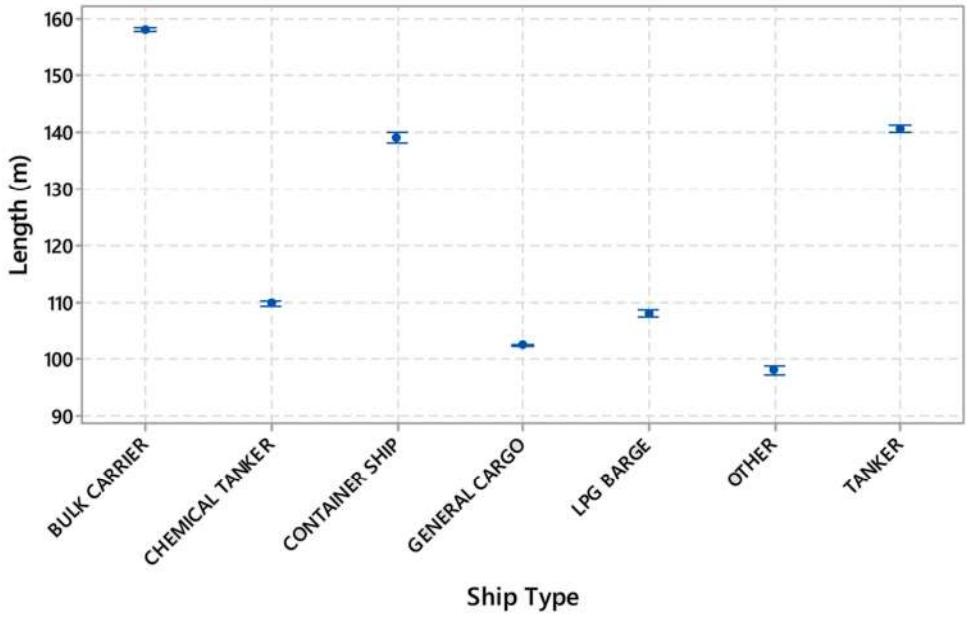


FIGURE 4.4: 95% vessel length confidence intervals for different vessel types

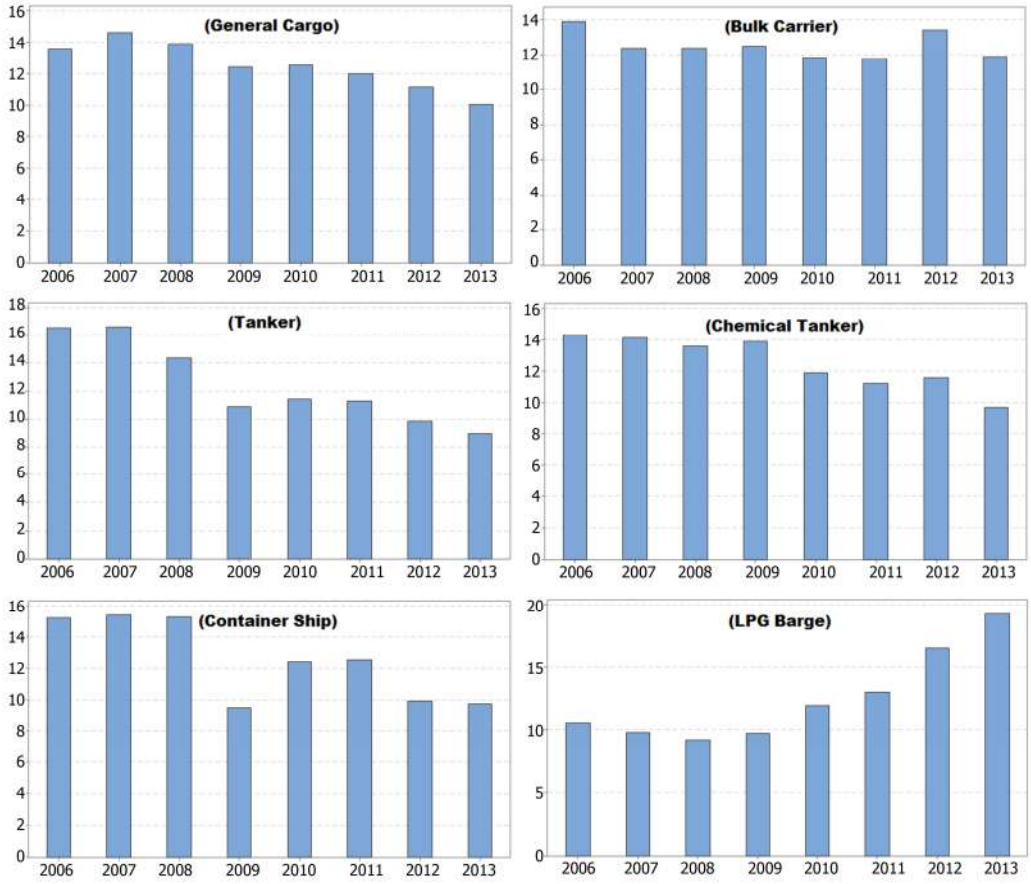


FIGURE 4.5: Percentage of vessels anchored in the Istanbul Strait for each vessel type

to 80% and 95% prediction intervals. Graphical illustrations of these predictions are displayed in Figure 4.6. A visual inspection of this figure suggests that except bulk carriers that do not have a clear trend, other forecasts appear to be reasonable. The three anchorage areas require different planning and management strategies as they differ by environmental and physical characteristics of their respective geographical locations. For this reason, individual area forecasts were performed and their illustrations are displayed in Figure 4.7 and Figure 4.8 for Southern and Northern Areas respectively. Our traffic analysis revealed when compared against the Southern and Northern Areas, the Eastern Area has much smaller traffic with no clear trends. Therefore, we do not report individual traffic forecasts for the Eastern Anchorage Area.

TABLE 4.2: Yearly forecast of vessel traffic and model statistics

		2014			2015			2016		
Vessel Type	**Metrics**	***Forecast***	***Prediction Interval***		***Forecast***	***Prediction Interval***		***Forecast***	***Prediction Interval***	
			80%	***95%***		***80%***	***95%***		***80%***	***95%***
General Cargo	ARIMA(0,1,0) AICc=117.07 BIC=113.96	28263	[25038, 31487]	[23332, 33194]	26762	[22202, 31322]	[19788, 33736]	25261	[19676, 30845]	[16719, 33802]
Bulk Carrier	ARIMA(0,0,0) AICc=114.97 BIC=112.72	3699	[3425, 3974]	[3280, 4119]	3700	[3425, 3974]	[3280, 4119]	3700	[3425, 3974]	[3280, 4119]
Tanker	ARIMA(0,1,0) AICc=95.35 BIC=92.24	2303	[1776, 2831]	[1496, 3110]	1985	[1239, 2731]	[844, 3126]	1667	[754, 2581]	[270, 3065]
Chemical Tanker	ARIMA(0,1,0) AICc=87.99 BIC=84.88	2100	[1815, 2386]	[1663, 2538]	1946	[1542, 2350]	[1328, 2564]	1791	[1296, 2286]	[1034, 2548]
Container Vessel	ARIMA(0,1,0) AICc=85.97 BIC=82.86	613	[372, 855]	[244, 982]	559	[218, 901]	[37, 1082]	505	[87, 924]	[-134, 1145]
LPG Barge	ARIMA(0,2,0) AICc=74.17 BIC=72.96	1409	[1292, 1292]	[1231, 1587]	1586	[1325, 1847]	[1187, 1985]	1763	[1326, 2200]	[1095, 2431]

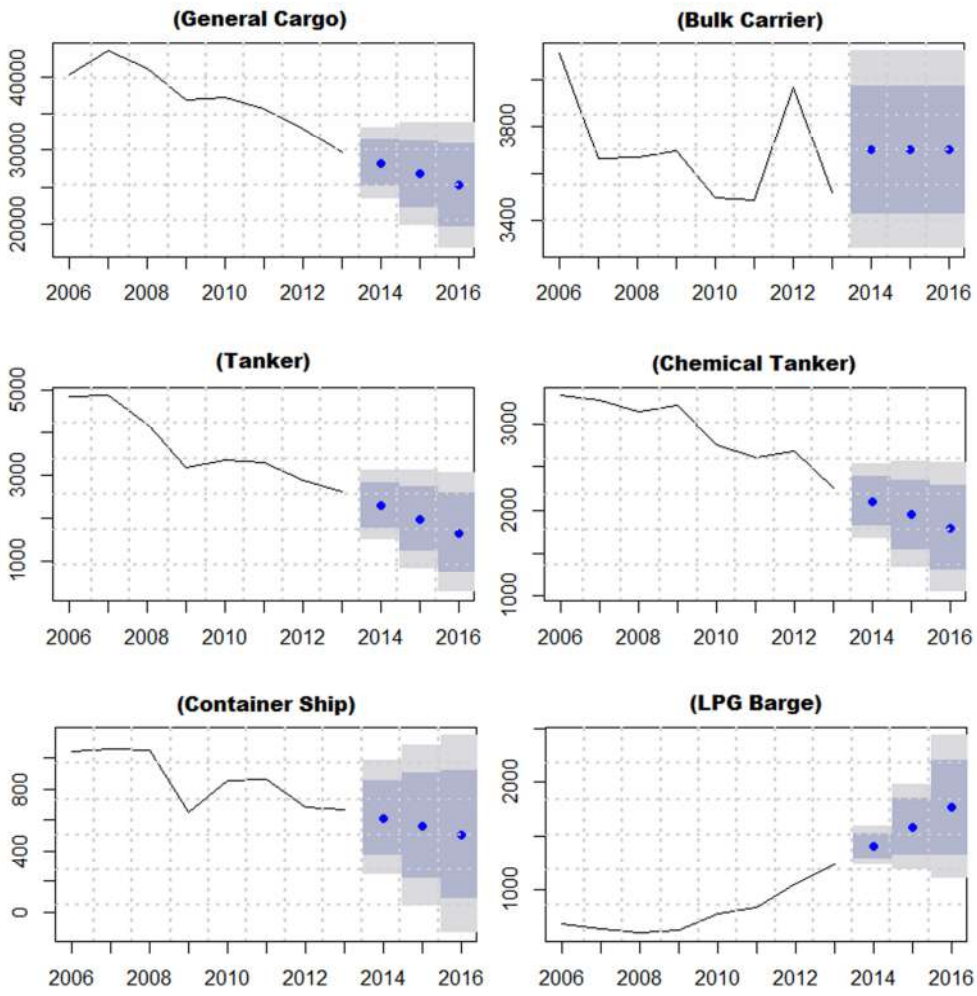


FIGURE 4.6: Three-year forecasts for vessel traffic in the three Strait anchorage areas combined

4.4 Chapter Recap

In this chapter, we present a temporal analysis of vessel type traffic inside the anchorage areas in the Istanbul Strait employing a new historical data set for the years between 2006 and 2013. This analysis consists of exploring the frequency of different vessel types, association of vessel types and length, and a time series analysis of vessel types. Furthermore, we forecast the number of vessels for each vessel type anchoring in Istanbul berth regions for the upcoming three years using the statistical ARIMA model. Our goal with this exercise is to provide a short-term outlook for assisting appropriate strategic decisions regarding anchorage area planning and management in the Istanbul Strait. Our

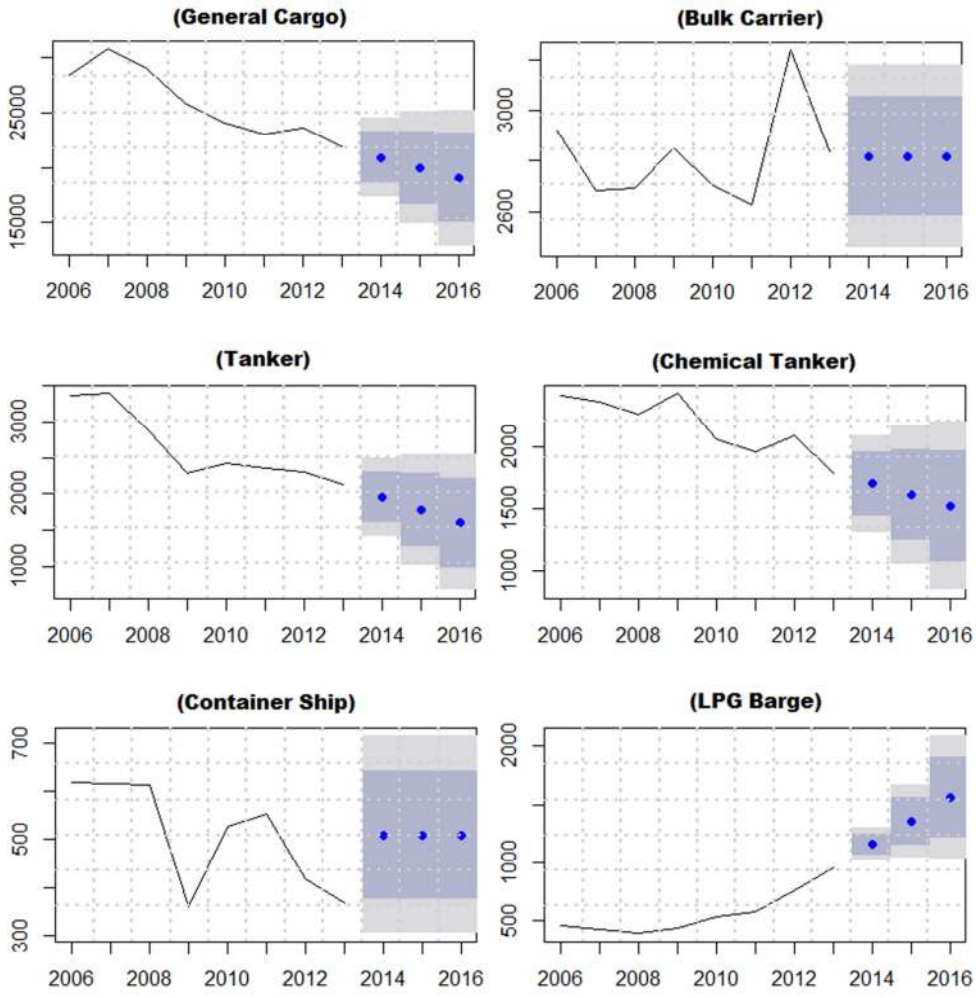


FIGURE 4.7: Three-year forecasts for vessel traffic in the Southern Anchorage Area

results suggest an overall decrease of berthing vessels, yet a pronounced increase in LPG barges.

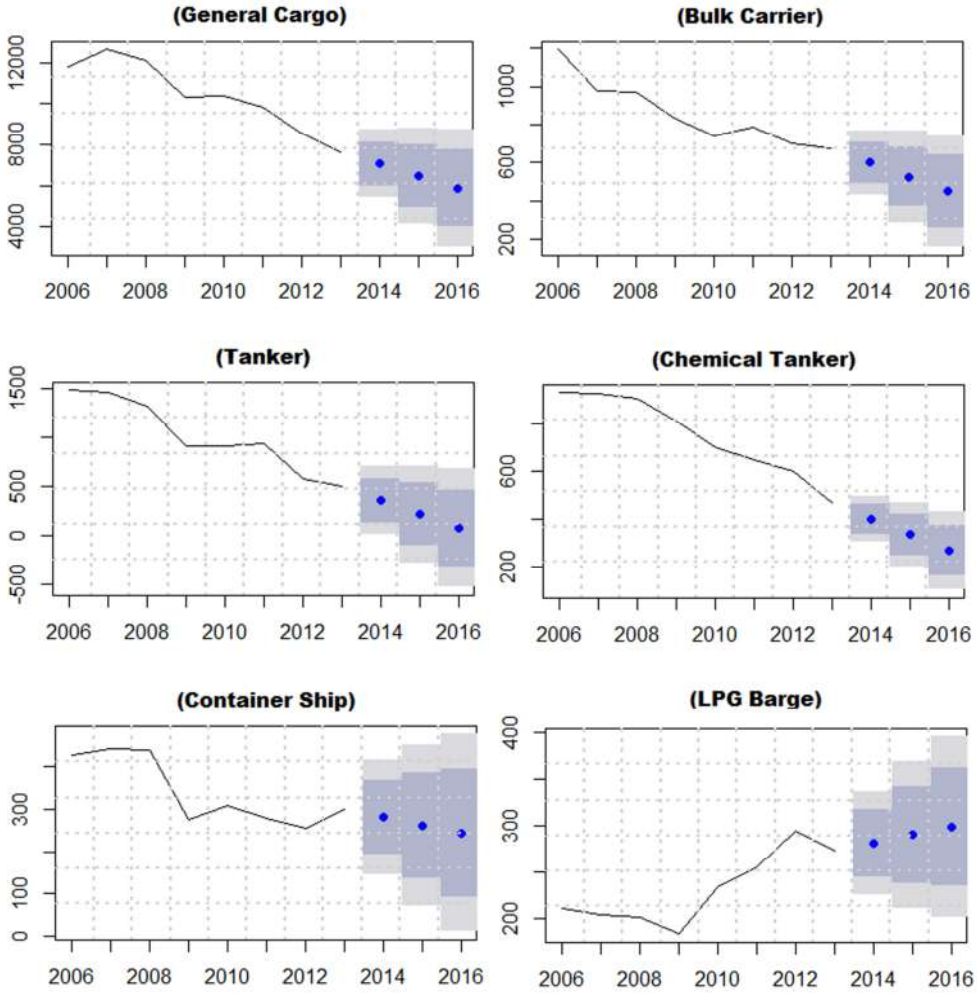


FIGURE 4.8: Three-year forecasts for vessel traffic in the Northern Anchorage Area

Chapter 5

Summary, Conclusions, and Directions for Future Research

This thesis considers deployment of data mining techniques in two application domains: social lending and anchorage planning. In this chapter, we provide summary and conclusions of our work on a chapter-by-chapter basis, which is followed by several directions for future research.

5.1 Summary and Conclusions

Chapter 2 presents a random forest (RF) based methodology for identification of good borrowers in social lending using data from the world's largest social lending platform. We introduce non-standard financial features in order to increase reliability of the computed risk scores, and we propose and present a comparison of the machine learning methods RF, SVM, LR, and k-NN for identifying good borrowers in social lending. Our computational results indicate that RFs outperform the other classifiers as well as the FICO scores and LC grades in predicting good customers. A limitation of our approach is that while RFs are quite powerful in predicting good borrower status, this comes at the cost of misclassifying some of the good borrowers as bad who are not the best (i.e., those not on top when ranked based on RF scores). In particular, RFs are superior in identifying best of the best borrowers, but they exhibit a decrease in relative performance beyond the 10% acceptance rate mark. Nonetheless, as mentioned earlier, it has

been shown that high risk borrowers are not worth the higher returns in general from a risk/ expected return trade-off point of view. Consequently, RFs stand as a more logical choice for potential LC lenders for making investment decisions. In particular, a lender can simply choose a borrower, say, in the top 3% of the population with respect to RF scores, and be confident that even though the returns might not be as high, there is practically no risk of a default.

Chapter 3 presents a comprehensive analysis on a new data set provided by Turkish Directorate General of Coastal Safety regarding Istanbul Strait anchorages for the years between 2006 and 2013. This analysis covers presenting frequency and histograms of several attributes including flag, arrival and departure country, length and tonnage of the vessel as well as zone, reason, month and year of the anchorage. And statistical relationship between these variables is explored accordingly. Employing linear regression technique, this analysis indicates high correlation of length and gross tonnage of vessels, and through an association analysis, we comprehend a meaningful relation between reason and zone of anchorage (p-value < 0.005). Regarding the prediction of anchorage duration, we compared different classifiers, and we infer that decision tree is the best method in order to estimate the duration, where using important attributes, classification accuracy of 75 percent and just by using key variables hold-out accuracy of 38 percent is achieved.

Chapter 4 presents a temporal analysis of vessel type traffic inside the anchorage areas in the Istanbul Strait using the historical data set for the years between 2006 and 2013. This analysis includes exploring the frequency of different vessel types, association of vessel types and length, and a time series analysis of vessel types. In addition, we forecast the number of vessels for each vessel type anchoring in Istanbul berth regions for the upcoming three years using the statistical ARIMA model. Our goal with this exercise is to provide a short-term outlook for assisting appropriate strategic decisions regarding anchorage area planning and management in the Istanbul Strait. Our results suggest an overall decrease of berthing vessels, yet a pronounced increase in LPG barges. Maritime accident statistics indicate collision and contact type accidents tend to be higher for vessels smaller than 10,000 gross tons [55]. On the other hand, most LPG barges are in this small vessel category. In addition, LPG is considered to be a very dangerous cargo. Thus, our finding that a sharp increase in the number of LPG barges is expected over the

next several years implies a need for revising current anchorage zoning and management practices.

5.2 Directions for Future Research

We now discuss several directions for future research.

In chapter 2, we do not consider any adjustments to interest rates for a trade-off analysis between default risk and higher returns. Hypothetically, should LC allow for the loan interest rates to be determined in a free market via an auctioning mechanism, lenders would likely be less hard-pressed to loan only to best of the best borrowers as the much higher rates might outweigh relatively high default risks. We believe there is a considerable amount of room in analysis of dynamic interest rates in the LC platform and optimal balancing of risk and return trade-offs based on lenders' risk preferences, which constitute a rather important direction for future research.

There exist a vast amount of data on borrower characteristics on social networks such as Facebook and Twitter. We believe that there is a significant potential in mining of social media data for more accurate borrower risk attribution and thereby integration of social media into social lending.

One other direction for future research in Chapter 2 is further fine-tuning of the SVM parameters (including trying different kernel functions) for loan status prediction. This fine-tuning process is a rather time-consuming task as it requires a careful search in multi-dimensional parameter space. In our implementation, we conducted only a limited amount of experiments for SVM fine-tuning. It is plausible that a carefully tuned SVM can show similar performance to RFs for classification.

Our results suggest the increase in LPG barges is likely to be offset by a decrease in both tanker types (chemical and product/crude oil), which are the three types of vessels required to anchor in designated Dangerous Cargo (DC) Zones inside the anchorages. Therefore, it is not immediately clear whether an urgent change in DC zoning is required. However, given an overall decrease in non-DC vessel types (i.e., general cargo, bulk carrier, and container ships), it appears an enlargement of current anchorage areas is probably not necessary at this point as any potential increase in DC zones might perhaps

be achieved by shrinking non-DC zones. That being the case, future trends in the Istanbul Strait anchorages need to be investigated in conjunction with global and regional trade, as well as general trends in maritime transportation. Thus, we would like to raise caution that our analysis in this study should merely be seen as a very first step in such a future trends assessment.

We also believe even more comprehensive studies need to be conducted on how current anchorage planning strategies need to be modified as a response to these future trends in order to provide more efficient, more effective, and safer vessel traffic services in the Strait anchorages.

Bibliography

[1] I. H. Witten and E. Frank. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2005.

[2] P. N. Tan, M. Stainbach, and V. Kumar. *Introduction to data mining*. Pearson, 2006.

[3] E. Alpaydin. *Introduction to machine learning*. MIT Press, 2014.

[4] C. M. Bishop. *Pattern recognition and machine learning*. Springer, 2006.

[5] LendingClub.com. Accessed January 27th, 2015. [Online]: http://www.lendingclub.com/public/about-us.action.

[6] G. J. Rodgers and D. Zheng. A herding model with preferential attachment and fragmentation. *Physica A*, 308:375–380, 2002.

[7] R. Emekter, Y. Tu, B. Jirasakuldech, and M. Lud. Evaluating credit risk and loan performance in online peer-to-peer (p2p) lending. *Applied Economics*, 47:54–70, 2015.

[8] S. H. Lopez. Social interactions in p2p lending. In *3rd Workshop on Social Network Mining and Analysis*, pages 1–8, 2009.

[9] N. Chen, A. Ghosh, and N. Lambert. Social lending. In *10th ACM Conference on Electronic Commerce*, pages 335–344, 2009.

[10] A. Steelman. Bypassing banks. In *Region Focus*, pages 37–40, 2006.

[11] S. C. Berger and F. Gleisner. Emergence of financial intermediaries in electronic markets: The case of online p2p lending. *Official Open Access Journal of VHB*, 2 (1):39–65, 2009.

[12] M. Klafft. Online peer-to-peer lending: A lender's perspective. In *International Conference on E-learning*, pages 371–375, 2008.

[13] R. Gao and J. Feng. An overview study on P2P lending. *International Business and Management*, 8(2):14–18, 2014.

[14] E. Lee and B. Lee. Herding behavior in online P2P lending: An empirical investigation. *Electronic Commerce Research and Applications*, 11(5):495–503, 2012.

[15] H. Yum, B. Lee, and M. Chae. From the wisdom of crowds to my own judgment in microfinance through online peer-to-peer lending platforms. *Electronic Commerce Research and Applications*, 11(5):469–483, 2012.

[16] D. Shen, C. Krumme, and A. Lippman. Follow the profit or the herd? exploring social effects in peer-to-peer lending. In *International Conference on Social Computing*, 2010.

[17] L. A. N. Amaral, A. Scala, M. Barthelemy, and H. E. Stanley. Classes of small-world networks. *PNAS*, 97(21):11149–11152, 2000.

[18] L. Backstrom, D. Huttenlocher, J. Kleinberg, and X. Lan. Group formation in large social networks: Membership, growth, and evolution. In *12th ACM International Conference on Knowledge Discovery and Data Mining*, pages 44–54, 2006.

[19] G. James, D. Witten, T. Hastie, and R. Tibshirani. *An Introduction to Statistical Learning*. Springer, NY, 2013.

[20] D. Aha and D. Kibler. Instance-based learning algorithms. *Machine Learning*, 6: 37–66, 1991.

[21] N. Landwehr, M. Hall, and E. Frank. Logistic model trees. *Machine Learning*, 59: 161–205, 2005.

[22] M. Sumner, E. Frank, and M. Hall. Speeding up logistic model tree induction. In *9th European Conference on Principles and Practice of Knowledge Discovery in Databases*, 2005.

[23] C. Cortes and V. Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, 1995.

[24] L. Zhou, K. K. Lai, and L. Yu. Least squares support vector machines ensemble models for credit scoring. *Expert Systems with Applications*, 37(1):127–133, 2010.

[25] L. Han, L. Han, and H. Zhao. Orthogonal support vector machine for credit scoring. *Engineering Applications of Artificial Intelligence*, 26(2):848–862, 2013.

[26] K. B. Schebesch and R. Stecking. Support vector machines for classifying and describing credit applicants: detecting typical and critical regions. *Journal of the Operational Research Society*, 56:1082–1088, 2005.

[27] L. Breiman. Random forests. *Machine Learning*, 45:5–32, 2001.

[28] K. B. Schebesch and R. Stecking. Support vector machines for credit scoring: Extension to non standard cases. In *Innovations in Classification, Data Science, and Information Systems*, Studies in Classification, Data Analysis, and Knowledge Organization, pages 498–505. Springer Berlin Heidelberg, 2005.

[29] C. L. Huang, M. C. Chen, and C. J. Wang. Credit scoring with a data mining approach based on support vector machines. *Expert Systems with Applications*, 33: 847–856, 2007.

[30] D. Ozdemir. Strategic choice for Istanbul: A domestic or international orientation for logistics? *Cities*, 27:154–163, 2010.

[31] J. Morgan. The Turkish straits: Christos L. Rozakis and Petros N. Stagos Martinus Nijhoff publishers, Dordrecht, 1987. *Marine Policy*, 13:173–174, 1989.

[32] S. Toyoda and Y. Fuji. Marine traffic engineering. *The Journal of Navigation*, 24: 24–34, 1971.

[33] A. Yamaguchi and S. Sakaki. Traffic surveys in Japan. *The Journal of Navigation*, 24:521–534, 1971.

[34] J. Draper and C. Bennett. Modelling encounter rates in marine traffic flows with particular application to the Dover Strait. *The Journal of Navigation*, 25:381–382, 1972.

[35] Y. Fujii. Two centuries of navigation: Development of marine traffic engineering in Japan. *The Journal of Navigation*, 30:86–93, 1977.

[36] K. Hara. A method for estimating the voyage distribution of marine traffic. *The Journal of Navigation*, 30:386–393, 1977.

[37] M. D. Ciletti. Traffic models for use in vessel traffic systems. *The Journal of Navigation*, 31:104–116, 1978.

[38] E. M. Goodwin. Marine encounter rates. *The Journal of Navigation*, 31:357–369, 1978.

[39] M. R. Bradshaw and K. D. Jones. Information systems in ports. *The Journal of Navigation*, 33:370–378, 1980.

[40] B. A. Colley, R. G. Curtis, and C. T. Stockel. A marine traffic flow and collision avoidance computer simulation. *The Journal of Navigation*, 37:232–250, 1984.

[41] T. Degre. The management of marine traffic, a survey of current and possible future measures. *The Journal of Navigation*, 48:53–69, 1995.

[42] G. E. Bijwaard and S. Knapp. Analysis of ship life cycles - the impact of economic cycles and ship inspections. *Marine Policy*, 33:350–369, 2009.

[43] S. Y. Huang, W. J. Hsu, and Y. He. Assessing capacity and improving utilization of anchorages. *Transportation Research Part E: Logistics and Transportation Review*, 47:216–227, 2011.

[44] I. Ari, V. Aksakalli, V. Aydogdu, and S. Kum. Optimal ship navigation with safety distance and realistic turn constraints. *European Journal of Operational Research*, 229:707–717, 2013.

[45] P. A. M. Silveira, A. P. Teixeira, and C. G. Soares. Use of AIS data to characterise marine traffic patterns and ship collision risk off the Coast of Portugal. *The Journal of Navigation*, 66:879–898, 2013.

[46] O. S. Uluscu, B. Ozbas, T. Altiok, I. Or, and T. Yilmaz. Transit vessel scheduling in the Strait of Istanbul. *The Journal of Navigation*, 62:59–77, 2009.

[47] M. A. Yazici and E. N. Otay. A navigation safety support model for the Strait of Istanbul. *The Journal of Navigation*, 62:609–630, 2009.

[48] Y. V. Aydogdu, C. Yurtoren, J. S. Park, and Y. S. Park. A study on local traffic management to improve marine traffic safety in the Istanbul Strait. *The Journal of Navigation*, 65:99–112, 2012.

[49] R. M. Alkan and T. Ocalan. Usability of the GPS Precise Point Positioning Technique in marine applications. *The Journal of Navigation*, 66:579–588, 2013.

[50] Y. V. Aydogdu. A comparison of maritime risk perception and accident statistics in the Istanbul Straight. *The Journal of Navigation*, 67:129–144, 2014.

[51] R. Lagerweij. Learning a model of ship movements. Master's thesis, University of Amsterdam, 2009.

[52] C. Tang and Z. Shao. Data mining platform based on AIS data. In *International Conference on Transportation Engineering*, pages 4465–4470, Chengdu, July 2009.

[53] M. C. Tsou. Discovering knowledge from AIS database for application in VTS. *The Journal of Navigation*, 63:449–469, 2010.

[54] K. M. S. Oo, C. Shi, H. Qinyou, and A. Weintrit. Clustering analysis and identification of marine traffic congested zones at Wusongkou, Shanghai. *Zeszyty Naukowe Akademii Morskiej W Gdyni*, 67:101–113, 2010.

[55] Y. V. Aydogdu, S. Kum, C. Yurtoren, E. Pitirlioglu, and T. Sanal. Marine accident analysis at Ahirkapi anchorage area in southern entrance of the Istanbul Strait. In *Asia Navigation Conference, Kobe, Japan*, 2012.

[56] G. E. P. Box, G. M. Jenkins, and G. C. Reinsel. *Time series analysis forecasting and control*. Wiley, 2008.

[57] S. R. Yoo, J. S. Jeong, J. Y. Jong, and G. K. Park. Forecast of marine traffic volume using time series model. In *International Conference on Fuzzy Theory and Its Application, Taipei, Taiwan*, 2013.

[58] D. Oz, V. Aksakalli, A. F. Alkaya, and V. Aydogdu. An anchorage planning strategy with safety and utilization considerations. *Computers Operations Research*, 62:12–22, 2015.