y-BIS 2019 Conference Book: Recent Advances in Data Science and Business Analytics

Mimar Sinan Fine Arts University
Department of Statistics,
Fındıklı Campus
http://ybis2019.msgsu.edu.tr

September, 25 - 28, 2019
Istanbul, TURKEY

MİMAR SİNAN FINE ARTS UNIVERSITY

Statistics Business Industry

isi

TÜRKİYE CUMHURİYET MERKEZ BANKASI

tam faktoring

Proceeding Book of the

y-BIS Conference 2019:

Recent Advances in Data Science and Business Analytics

y-BIS 2019
YOUNG BUSINESS AND INDUSTRIAL STATISTICIANS WORKSHOP
ON RECENT ADVANCES IN DATA SCIENCE AND BUSINESS ANALYTICS
SEPTEMBER 25-28, 2019
isi
Statistics
Business
Industry
SPONSORED BY ISBIS
MIMAR SINAN FINE ARTS UNIVERSITY
ISTANBUL, TURKEY

September 25 – 28, 2019
Mimar Sinan Fine Arts University / Fındıklı Campus

Edited by

Ozan Kocadagli
Ali Erkoc
Bilge Baser
Nihan Acar Denizli
Tahir Ekin
LOC of y-BIS 2019

Web design by
Ali Zafer Dalar

Cover photo by
Aylin Yaman Kocadagli

Cover design by
Ali Mertcan Kose
Damla Ilter

Published by

Mimar Sinan Fine Arts University Publications : 884

ISBN:

978-605-5005-95-5

Serial Number:

eMSGSÜ-FEF-İST-019/09-Kat1

Contents

Part I. Introduction

Part II. Scientific Program

Part III. Keynote Lectures

Part IV. Short Courses

Part V. Invited Sessions

Part VI. Contributed Papers (Abstract)

Part VII. Contributed Papers (Full)

Part VIII. Poster (Abstract)

Part IX. Poster (Full)

Part X. List of Participants

Part XI. Sponsors and Supporting Institutions

Part I

Preface

Welcome to y-BIS 2019 Conference: Recent Advances in Data Science and Business Analytics.

On the behalf of the Local Organizing Committee we are pleased to welcome you to y-BIS 2019 Conference: ISBIS Young Business and Industrial Statisticians Workshop on Recent Advances in Data Science and Business Analytics, sponsored by ISBIS (International Society for Business and Industrial Statistics) and Mimar Sinan Fine Arts University. This is the fourth conference arranged by ISBIS/y-BIS where the second one was organized in Istanbul before, 2013 Joint Meeting of Young Business and Industrial Statisticians Meeting.

The purpose of y-BIS 2019 is to bring together young statisticians and professionals working in Academia and in Industry. The conference will offer opportunities to meet each other, to share scientific and professional experiences, and to promote new collaborations and international cooperation. This conference will cover many researches in the academia and business world such as finance, medicine, insurance, energy, etc.

The program covers 10 Keynote Speakers, 8 workshops with 11 speakers, 3 invited and 16 contributed parallel sessions with 70 speakers and one poster session. We would like to thank all the speakers and, in particular the Keynote Speakers, Workshop and Invited Paper Session organizers who helped greatly to improve the scientific program of the conference.

The end of y-BIS 2019, all the presented studies have been published as a full-paper or abstract in the conference book with ISBN under the refereeing procedure and editorial policy of the conference. In addition, it is expected that, after refereeing process the selected papers will be directed to the five special issues of the journals: Applied Stochastic Models in Business and Industry, Istanbul Business Research Journal of Ambient Intelligence and Humanized Computing, Journal of Computational and Applied Mathematics and Turkish Journal of Forecasting.

The program also includes social events which will allow nearly 200 participants to know each other and to get experience of Turkish culture and history in addition to the taste of the Turkish cuisine and hospitality.

The organizers would like to thank to all the institutions that have provided financial support to make this organization possible. Many thanks to Faculty of Sciences and Letters of Mimar Sinan Fine Arts University, ISBIS-International Society for Business and Industrial Statistics, The Central Bank of Turkey and Tam Faktoring. Lastly, I really appreciate Local Organizing and Scientific Program Committees for their efforts performing on y-BIS 2019.

I am looking forward to seeing you in the next scientific events of ISI/ISBIS.

On the behalf of the Local Organizing Committee,

Ozan Kocadagli

(General Chair of y-BIS 2019)

Dear colleagues,

We are excited for your participation in the 2019 y-BIS (Young Business and Industrial Statisticians) Conference on Recent Advances in Data Science and Business Analytics.

y-BIS the Young Statisticians' group in the International Society for Business and Industrial Statistics (ISBIS), was formed in 2008. The purpose of y-BIS is to bring together young researchers and professionals working on business, financial and industrial statistics, to help support their career development.

ISBIS is an association of the International Statistical Institute (ISI) that is dedicated to the promotion of business and industrial statistics worldwide. ISBIS promotes applications, research, and best current practices in business and industrial statistics, facilitates technology transfer, and fosters communications among members and practitioners worldwide. Please visit http://www.isbis-isi.org/index.html for more information.

y-BIS has organized conferences previously in Lisbon (2012), Istanbul (2013) and Hamedan (2017). We are excited to get back to Istanbul for the fourth y-BIS Conference.

We would like to extend our sincere thanks to the organizing committee of y-BIS 2019. They have done a great job putting together a great scientific and social program. The scientific program includes a great mix of keynote speakers, short courses and sessions. We are sure that this will turn out to be a great conference.

Tahir Ekin (2017-2019 y-BIS Chair) and Luca Frigau (2019-2021 y-BIS Chair)

Committees

Ozan Kocadagli (Mimar Sinan Fine Arts University, Istanbul, Turkey)

(General Chair of y-BIS 2019)

The Local Organizing Committee

- Ali Erkoc (Co-chair, Mimar Sinan Fine Arts University, Istanbul, Turkey)
- Bilge Baser (Co-chair, Mimar Sinan Fine Arts University, Istanbul, Turkey)
- Nihan Acar (Co-Chair, Mimar Sinan Fine Arts University, Istanbul, Turkey)
- Ali Zafer Dalar (Giresun University, Giresun, Turkey)
- Berk Kucukaltan (Trakya University, Edirne, Turkey)
- Busenur Kizilaslan (Marmara University, Istanbul, Turkey)
- Coskun Parim (Yildiz Technical University, Istanbul, Turkey)
- Damla Ilter (Mimar Sinan Fine Arts University, Istanbul, Turkey)
- Ezgi Özer (Istanbul Okan University, Istanbul,Turkey)
- Metin Yangin (Mimar Sinan Fine Arts University, Istanbul, Turkey)
- Neslihan Gokmen (Istanbul Technical University, Istanbul, Turkey)
- Selin Saridas (Mimar Sinan Fine Arts University, Istanbul,Turkey)
- Turgut Ozaltindis (Mimar Sinan Fine Arts University, Istanbul, Turkey)
- Zeynep Atli (Mimar Sinan Fine Arts University, Istanbul, Turkey)
- Zeynep Bal (Mimar Sinan Fine Arts University, Istanbul, Turkey)

The International Scientific Program Committee (Referees, Editorial Review Board)

- Alev Bakir (Turkey)
- Ali Shojaie (USA)
- Ali Erkoc (Turkey)
- Arzu Baygul (Turkey)
- Ayfer Ezgi Yilmaz (Turkey)
- Aytac Atac (Germany)
- Babak Zafari (USA)
- Bahadir elmas (Turkey)
- Balaji Raman (India)
- Benay Uzer (Turkey)
- Bilge Baser (Turkey)
- Caterina Liberati (Italy)
- Deniz Inan (Turkey)
- Elif Coker (Turkey)
- Elif Ozge Ozdamar (Turkey)
- Emilie Devijver (France)
- Emre Dunder (Turkey)
- Emre Nadar (Turkey)
- Erkan Sirin (Turkey)
- Esra Akdeniz (Turkey)
- Esra Pamukcu (Turkey)
- Ettore Lanzarone (Italy)
- Francesca Ieva (Italy)
- Fulya Gokalp Yavuz (Turkey)
- Gregor Kastner (Austria)
- Han- Ming Wu Hank (Taiwan)
- Jeff Goldsmith (USA)
- Jitka Hrabakova (Czechia)
- Kathrin Plankensteiner (Austria)
- Kristine Lurz (Germany)
- Laura Lotero- Velez (Colombia)
- Laura Trinchera (France)
- Luca Frigau (Italy)
- Marie Perrot-Dockes (France)
- Meral Yay (Turkey)
- Miguel Angel Ortiz Barrios (Colombia)
- Mustafa Murat Arat (USA)
- Naciye Tuba Yilmaz Soydan (Turkey)
- Nina Senitschnig (Austria)
- Nihan Acar Denizli (Turkey)
- Nuriye Sancar (Cyprus)
- Oguz Akbilgic (USA)
- Olawale Awe (Nigeria)
- Ozan Kocadagli (Turkey)
- Ozlem Deniz Basar (Turkey)
- Ozlem Turksen (Turkey)
- Paulo Canas Rodrigues (Brazil)
- Pedro Delicado (Spain)
- Rahim Mahmoudvand (Iran)
- Ridvan Keskin (Turkey)
- Seda Tolun Tayali (Turkey)
- Shima Mohebbi (USA)
- Tahir Ekin (USA)
- Tevfik Aktekin (USA)
- Tuba Yilmaz Gozbasi (Turkey)
- Ufuk Beyaztas (Turkey)
- Ufuk Yolcu (Turkey)
- Yasmin Said (USA)

Scientific Advisory Committee (Referees, Editorial Review Board)

Part II

Scientific Program

TIME	25th SEPTEMBER 2019 WEDNESDAY	
08:00-09:00	REGISTRATION	
09:00-09:30	OPENING CEREMONY	
09:30-10:30	Owl Hall *(Session Chair: Gulay BASARIR)* **KEYNOTE SPEAKER 1 : Aytul ERCIL** (Vispera, Sabancı University) *The irreparable Rise of Artificial Intelligence*	
10:30-11:00	COFFEE BREAK (POSTER SESSIONS)	
11:00-12:00	Owl Hall *(Session Chair: Semra ERPOLAT TASABAT)* **KEYNOTE SPEAKER 2: Umut Satır GURBUZ** (IBM) *Preparing to Exist in the Age of Artificial Intelligence*	
12:00-13:30	LUNCH	
13:30-15:00	WORKSHOP 1	**Big Data: Introduction to Hadoop big data ecosystem** Erkan SIRIN - ***Room 201*** ***Session Chair: Ali ERKOC*** **Retail Analytics with Dynamic Linear Models Using R** Balaji RAMAN - ***Room 202*** ***Session Chair: Nihan ACAR DENIZLI*** **Visualization with QlikView (How to make dashboards)** Rahim MAHMOUDVAND - ***Room 203*** ***Session Chair: Bilge BASER***
15:00-15:30	COFFEE BREAK	
15:30-17:00	WORKSHOP 2	**Introduction to Apache Spark, Data analysis and Machine Learning with Apache Spark** Erkan SIRIN- ***Room 201*** ***Session Chair: Ufuk BEYAZTAS*** **Introduction to DLM and Kalman filter, Setting up DLM in R using packages astsa, dlm and INLA, Real-life applications** Balaji RAMAN - ***Room 202*** ***Session Chair: Fatih KIZILASLAN*** **Fraud Analystics** Tahir EKIN - ***Room 203*** ***Session Chair: Arzu BAYGUL***
17:00-18:00	Owl Hall *(Session Chair: Eylem DENIZ)* **KEYNOTE SPEAKER 3: Erkal BIYIKLIOGLU** (Tam Factoring) *Financial Risk and Data Analysis*	
18:30	WELCOME RECEPTION	
TIME	**26th SEPTEMBER THURSDAY**	
09:00-10:00	Owl Hall *(Session Chair: Gulay ILONA TELSIZ)* **KEYNOTE SPEAKER 4 : Gerhard Wilhelm WEBER** (Poznan University of Technology, Adviser to EURO Conferences) *RMARS under Cross-Polytope Uncertainty - Prediction of Natural Gas Consumption*	
10:00-10:30	COFFEE BREAK	
10:00-10:30	POSTER SESSION	**The Examination of Real Estate Prices in Istanbul by Using Hybrid Hierarchical K-Means Clustering** Ilkay TUG, Betul KAN KILINC **Statistical properties and modeling of stable-like word count time series in nation-wide language data** Hayafumi WATANABE
10:30-12:00	INVITED PAPER SESSION 1	**CLASSIFICATION BASED ALGORITHMS: METHODS AND APPLICATION - *Room 201 (Session Chair: Luca FRIGAU)*** **Classification-based Approach for Validating Image Segmentation Algorithms** Luca FRIGAU, Francesco MOLA **Portfolio composition strategy through a P-Spline based clustering approach** Carmela IORIO, Giuseppe PANDOLFO **Network-based Semisupervised Clustering** Giulia CONTU, Claudio CONVERSANO, Luca FRIGAU
10:30-12:00	INVITED PAPER SESSION 2	**TUPRAS SESSION (Data Mining and Big Data Analytics in Refinery Processes) - *Room 202 (Session Chair: Cagla ODABASI)*** **Fault Detection and Diagnosis Methodology in Refineries: A Data-Driven Approach** Cagla ODABASI **Big Data Solutions in Refineries with Heat Exchangers** Ocan SAHIN, Cagla ODABASI
10:30-12:00	WORKSHOP 3	**Real world applications/cases of transportation analytics-optimization with a potential demo** Tuba YILMAZ GOZBASI (Optiyol, Ozyegin University), Ozan GOZBASI (Optiyol, Bosphorus University) - ***Room 203*** ***Session Chair: Bahadır ELMAS***
12:00-13:00	LUNCH	
13:00-13:50	Owl Hall *(Session Chair: Baris ASIKGIL)* **KEYNOTE SPEAKER 5 : Barıs SURUCU** (METU) *Data Analytics and Machine Learning: Real Life Applications in Various Field*	
14:00 - 14:50	Owl Hall *(Session Chair: Caterina LIBERATTI)* **KEYNOTE SPEAKER 6 : HAMPARSUM BOZDOGAN (University of Tennessee)** *Robust Bayesian Relevance Vector Machines in Regression and Supervised Classification using Information Complexity and Genetic Algorithm (Application in Early Detection of Heart Attack Classification Problem)*	
15:00-16:40	**STATISTICS THEORY I - *Room 201 (Session Chair: Mahmude Revan OZKALE)*** **A Robust Method for Estimation of Models with Random Effects** Beste Hamiye BEYAZTAS **A Percentile Bootstrap Based Method on Dependent Data: Harrell Davis Quantile Estimator vs NO Quantile Estimator** Gozde NAVRUZ, A. Firat OZDEMIR **Evaluating New Optimization Methods for Two Parameter Ridge Estimator via Genetic Algorithm** Erkut TEKELI, Selahattin KACIRANLAR, Nimet OZBAY **Stochastic Linear Restrictions in Generalized Linear Models** Mahmude Revan OZKALE **The GO estimator: A New Generalization of Lasso** Murat GENC, Mahmude Revan OZKALE	
15:00-16:40	**BUSINESS/FINANCE I - *Room 202 (Session Chair: Ayca CAKMAK PEHLIVANLI)*** **An Approach for Considering Claim Amount and Varying** Atefeh MORADI,Maryam SHARAFI, Rahim MAHMOUDVAND **Churn Analysis for Factoring: An Application in Turkish Factoring Sector** Enis GUMUSTAS, Huseyin BUDAK **Opportunities in Location Based Customer Analytics** Murat OZTURKMEN	

TIME	26th SEPTEMBER THURSDAY
15:00-16:40	**TIME SERIES/MODELING - *Room 203*** ***(Session Chair: Rahim MAHMOUDVAND)*** **Conditional Autoregressive Model Approach to Generalized Linear Spatial Models by CARBayes** Leyla BAKACAK KARABENLI, Serpil AKTAS ALTUNAY **Highlighting a Mathematical Property of Sample ACF for Time Series Analysis** Rahim MAHMOUDVAND **A New Approach to Econometric Modelling of Monthly Total Air Passengers: A Case Study for Ataturk Airport** Resit CELIK, Hasan Aykut KARABOGA, İbrahim DEMIR **Feature Selection Approaches for Machine Learning Classifiers on Yearly Credit Scoring Data** Damla ILTER, Ozan KOCADAGLI, Nalini RAVISHANKER
15:00-16:40	**FUNCTIONAL DATA ANALYSIS - *Owl Hall (Session Chair: Gulhayat GOLBASI SIMSEK)*** **Investigation of the Electricity Consumption of Provinces of Turkey using Functional Principal Components Analysis** Sumeyye INAL, Gulhayat GOLBASI SIMSEK **On function-on-function regression: Partial least squares approach** Ufuk BEYAZTAS, Han Lin SHANG **Wavelet Regression for Noisy Data** Gokce Nur TASAGIL, Eylem DENIZ **A Functional Data Framework to Analyse the Effect of Quinoa Consumption on Blood Glucose Levels** Nihan ACAR DENIZLI, Pedro DELICADO, Belchin KOSTOV, Diana A. DIAZ RIZOLLO, Antoni SISO, Ramon GOMIS
16:40-17:00	COFFEE BREAK
17:00-18:40	**BUSINESS/FINANCE II - *Room 201 (Session Chair: Ipek DEVECI KOCAKOC)*** **Predicting Business Survival From Their Websites** Desamparados BLAZQUEZ, Lisa CROSATO, Josep DOMENECH, Caterina LIBEATI **Fast Fault Finding Methods in Smart Manufacturing Lines with Augmented Reality Applications** Adem KAYAR, Fatih OZTURK, Ozkan KAYACAN **A Customer Segmentation Model Proposal for Hospitals: LRFM-V** Ipek DEVECI KOCAKOC, Pınar OZKAN
17:00-18:40	**APPLIED STATISTICS I - *Room 202*** ***(Session Chair: Gul INAN)*** **The Effect of Weights on Multi-rater Weighted Kappa Coefficients** Ayfer Ezgi YILMAZ **Probabilistic Structural Equation Modeling Approach to Investigate the Relationships Between Passenger Perceived Value, Image, Trust, Satisfaction and Loyalty** Tugay KARADAG, Gulhayat GOLBASI SIMSEK **Two Structural Equation Modelling Approaches for Cloud Use in Software Development** Erhan PISIRIR, Cuneyt SEVGI, Oumout CHOUSEINOGLOU, Erkan UCAR **Joint Modeling the Frequency and Duration of Physical Activity from a Lifestyle Intervention Trial** Gul INAN, Juned SIDDIQUE
17:00-18:40	**BIOSTATISTICS / BIOINFORMATICS - *Room 203*** ***(Session Chair: Candan GURSES)*** **Analyzing the Competition of HIV-1 Phenotypes with a Quantum Computation Perspective** Bilge BASER **HIV-1 Prostease Cleavage Site Prediction with Generating Dataset Using a New Encoding Scheme Based on Physicochemical Properties** Metin YANGIN, Ayca CAKMAK PEHLIVANLI, Bilge BASER **Time-Frequency Analysis of EEG Signals: Visual Identification of Epileptic Patterns** Ezgi OZER, Arnaldo Guimaraes BATISTA, Ozan KOCADAGLI
TIME	**27th SEPTEMBER FRIDAY**
09:00-10:00	**Owl Hall** ***(Session Chair: Tahir EKIN)*** **KEYNOTE SPEAKER 7 : Sotiris BERSIMIS (University of Piraeus)** *Using Data Analytics for Fraud Detection in Health Care: Applications and Some Results*
10:00-10:30	COFFEE BREAK
10:30-12:00	INVITED PAPER SESSION 3 **RECENT ADVANCES ON FUNCTIONAL DATA ANALYSIS - *Room 201 (Session Chair: Nihan ACAR DENIZLI)*** **From Multivariate to Functional Classification** José Luis TORRECILLA **Functional Linear Model for Monitoring and Prediction of Profiles** Alessia PINI **Depth-Based Functional Time Series Forecasting** Antonio ELIAS, Raúl JIMENEZ
10:30-12:00	WORKSHOP 4 **Innovation in Germany (with potential emphasis on internet of things, Supply Chain Analytics)** Aytac ATAC (Supply Chain Wizard) - ***Room 202*** ***Session Chair: Ozge CAGCAG YOLCU***
10:30-12:00	**APPLIED STATISTICS II - *Room 203*** ***(Session Chair: Elif COKER)*** **Serial Mediation Model of Leader Member Interaction in Work Values and Job Satisfaction** Meral YAY, Mine AFACAN FINDIKLI, Ali Mertcan KOSE **Finding the Determinants of National Problem Perceptions of Turkish Citizens** Ipek DEVECI KARAKOC, Ozlem KIREN GURLER **Analysis of Data Comparing the use of Different Social Media for Scientific Research Across Different Countries of the World** Fatima HARIS **Analysis of the Science Scores of Turkish Students in PISA 2015 via Multilevel Models** Elif COKER **A StarCraft 2 Player Skill Modeling** Natasa A. CIROVIC, Zoran Z. CIROVIC
12:00-13:30	LUNCH
13:30-14:30	**Owl Hall** ***(Session Chair: Birsen EYGI ERDOGAN)*** **KEYNOTE SPEAKER 8 : Bahar KINAY ERGUNEY** (CISCO) *Big Data and IoE*

14:30-15:00	COFFEE BREAK
15:00-16:40	**CLUSTERING / CLASSIFICATION** - ***Room 201*** ***(Session Chair: Nurdan COLAKOGLU)*** **Hierarchically Built Trees with Probability of Placing Clusters** Nebahat BOZKUS, Stuart BARBER **Comparison of Internal Validity Indices According to Distance Measurements in Clustering Analysis** Derya ALKIN, Aydin KARAKOCA, Ibrahim DEMIR **Comparison of Multi-class Classification Algorithms on Early Diagnosis of Heart Diseases** Nurdan COLAKOGLU, Berke AKKAYA **An Application of XGBOOST on Diabetes Dataset** Gulcin YANGIN, Elif Ozge OZDAMAR **How Does Resampling Affect the Classification Performance of Support Vector Machines on Imbalanced Churn Data** Serra CELIK , Seda TOLUN

TIME	27th SEPTEMBER FRIDAY
15:00-16:40	**STATISTICS THEORY II** - ***Room 202 (Session Chair: Berk KUCUKALTAN)*** **Stress-Strength Reliability Estimation of Series System with Cold Standby Redundancy at System and Component Levels** Gulce CURAN, Fatih KIZILASLAN **Statistical Inference of Consecutive k-out-of-n System in Stress-Strength Setup Based on Two Parameter Proportional Hazard Rate Family** Duygu DEMIRAY, Fatih KIZILASLAN **Approximation of Continuous Random Variables for the Evaluation of the Reliability Parameter of Complex Stress-Strength Models** Alessandro BARBIERO **Chaos Control in Chaotic Dynamical Systems Via Auto-tuning Hamilton Energy Feedback** Atike Reza AHRABI, Hamid Reza KOBRAVI
15:00-16:40	**WORKSHOP 5** **Medical Analytics / Bioinformatics (Clinical Research Statistics, Medical Informatics, Validation Studies with** Arzu BAYGUL, Cagdas AKTAN and Neslihan GOKMEN - ***Room 203*** ***Session Chair: Esra AKDENIZ***
16:40-17:10	COFFEE BREAK
17:10-18:50	**REGRESSION/FUZZY MODELING** - ***Room 201 (Session Chair: Ozlem TURKSEN)*** **A Seemingly Unrelated Regression Modeling for Extraction Process in Green Chemistry** Ozlem TURKSEN, Serhan TUNCEL, Nilufer VURAL **The Effect of WoE Transformation on Credit Scoring By Using Logistic Regression** Zeynep BAL, M. Aydin ERAR **Bivariate Intuitionistic Fuzzy Time Series Prediction Model** Ozge CAGCAG YOLCU, Erol EGRIOGLU, Eren BAS, Ufuk YOLCU **Nonlinear Neural Network for Cardinality Constraint Portfolio Optimization Problem: Sector-wise analysis of ISE-all Shares** Ilgım YAMAN, Turkay ERBAY DALKILIC **Statistical and Fuzzy Modeling of Extraction Process in Green Chemistry** Nilufer VURAL, Ozlem TURKSEN
17:10-18:50	**BUSINESS/FINANCE III** - ***Room 202*** ***(Session Chair: Esra N. KILCI)*** **Bitcoin Cash: Returns Distributions and Dissimilarity Analysis** Muhammad SHERAZ, Vasile PREDA, Silvia DEDU **Do Confidence Indicators Have Impact on Macro-financial Indicators? Analysis on the Financial Services and Real Sector Confidence Indexes:** Esra N. KILCI **Granger-Causality-Based Portfolio Selection In The Moroccan Stock Market** Abdelhamid Hamidi ALAOUI
17:10-18:50	**STATISTICS THEORY III** - ***Room 203*** ***(Session Chair: Nursel KOYUNCU)*** **Multivariate Skew Laplace Normal Distribution: Properties and Applications** Fatma Zehra DOGRU, Olcay ARSLAN **Fitting lognormal distribution to actuarial data** Mahdi MAHDIZADEH, Ehsan ZAMANZADE **Gamma and Inverse Gaussian Distributions in Fitting Parametric Shared Frailty Models with Missing Data** Nursel KOYUNCU, Marthin PIUS, Nihal ATA TUTKUN
20:00	GALA DINNER
TIME	**28th SEPTEMBER SATURDAY**
08:45-09:30	**Owl Hall** ***(Session Chair: Elif Ozge OZDAMAR)*** **KEYNOTE SPEAKER 9 : Selim DELILOGLU (Data Analyst at Telecommunication Sector)** *Fundamental Skills for Data Science & Business Analytics*
09:30-10:15	**Owl Hall** ***(Session Chair: Meral YAY)*** **KEYNOTE SPEAKER 10 : Ayse OZMEN** *Multi-objective Sparse Regression Models for short- and long-term Natural Gas Demand Prediction*
10:15-10:30	COFFEE BREAK
10:30-12:00	**WORKSHOP 6** **Hands-on Introduction Course in R(Acquiring data from different sources on command line and R, data pre-processing, a map package to map one of the up-to-date data (potentially with 2019 Turkish local election data), SQL in R** Fulya GOKALP (METU) - **Owl Hall** **Session Chair: Deniz INAN**
10:30-12:00	**BUSINESS/FINANCE IV** - ***Room 201*** ***(Session Chair: Mujgan TEZ)*** **Risk-based Fraud Analysis for Bank Loans With Autonomous Machine Learning** Yunus Emre GUNDOGMUS, Mert NUHUZ, Mujgan TEZ **Bivariate Credibility Premiums Distinguishing Between Two Claims Types in Third Party Liability Insurance** Pervin BAYLAN, Serdar KURT, Neslihan DEMIREL, Jeffrey S. PAI **Methods for Optimum Establishment of Government-imposed global budget caps -Monitoring Pharmaceutical Expenditures using SPM** Nika ELISAVET **Prediction of Claim Probability in the Presence of Excess Zeros** Aslihan SENTURK ACAR
10:30-12:00	**OUTLIER DETECTION** - ***Room 202*** ***(Session Chair: Erkan SIRIN)*** **Detection and Handling Outliers in Longitudinal Data: Can Wavelets Decomposition Be a Solution?** Marwa BENGHOUL, Berna YAZICI, Ahmet SEZER **Outlier Detection on Big Data** Erkan SIRIN, Hacer KARACAN **Identification of Vehicle Warranty Data and Anomaly Detection by Means of Machine Learning Methods** Halil Ibrahim CELENLI, Esin OZKAN
10:30-12:00	**OPTIMIZATION/DECISION MAKING** - ***Room 203*** ***(Session Chair: Semra ERPOLAT TASABAT)*** **Alternative Subway Project Selection with TOPSIS Method Using Different Weighting Techniques** Nihan YUCEL, Semra ERPOLAT TASABAT **Recycle Project with RFM Analysis** Esra AKCA, Semra ERPOLAT TASABAT **Inferences About Development Levels of Countries with Data Envelopment Analysis** Semra ERPOLAT TASABAT

12:00-12:30	CLOSING
14:00-17:30	SOCIAL EVENT (Bosphorus Boat Tour)

How Does Resampling Affect the Classification Performance of Support Vector Machines on Imbalanced Churn Data?

Serra Çelik
Informatics Dept.,
Istanbul University
erra.celik@istanbul.edu.tr

Seda Tolun Tayalı
Quantitative Methods Dept.,
School of Business, Istanbul University
stolun@istanbul.edu.tr

Abstract

Churn prediction is an important task for companies. Determining a customer as a probable churner beforehand and keeping her as a result of customer relationship management efforts directly increases the profit of a business. However, churn datasets are imbalanced by nature, which negatively affects the classification performance of algorithms such as the popular Support Vector Machines (SVM). This study handles the classification of imbalanced churn data by applying resampling techniques; Random Under-Sampling, Clustering Based Under-Sampling, Random Over-Sampling, and Synthetic Minority Oversampling Technique as a preprocessing step for to construct a more balanced dataset and investigates their effects on the classification performance of SVM with different kernel functions. The results show that the classification performance of Support Vector Machines improves when resampling is implemented to an imbalanced churn data, especially with Radial Basis Function, and with 5x2 cross validation.

Keywords: under-sampling; over-sampling; churn; binary classification

1. Introduction

In today's competitive business world, building and maintaining successful relationships with customers is an inevitable necessity to survive in a market. Within the context of customer relationship management (CRM), there are several tasks [1] such as; customer profiling, sentiment analysis, churn prediction, and direct marketing and many of these problems are handled by data analytics. The concentration of these problems is understanding the customers and their behavioral patterns.

Customer churn, also known as attrition [2], is an important and expensive problem for businesses. It is explained by the likelihood of customers terminating doing business with a company. The literature has known for a fact that gaining new customers is much more expensive than retaining current ones. Also, within the CRM context, the existing customers are more prone to being in communication with the company and spend more than the new ones. Reichheld and Shefter [3] state from companies' perspective that "*Increasing customer retention rates by 5% increases profits by 25% to 95%*".

In the telecommunications sector, an average customer's monthly spending varies between 20$ and 80$ depending on the country [4]. Customers switching between operators, which is by definition customer churn, is quite common in the telecommunication sector [5].

The annual churn rate in the sector is around 30% in average and acquiring new customers is at least 5 times more expensive than keeping the existing ones [6]. Therefore, losing high number of customers results in high losses for the telecom companies because of lost acquisitions as well as of certain CRM efforts such as reducing the prices to keep the highly potential churners somehow in the company portfolio. This makes churn analysis and consequently the reduction of churn rate a crucial goal for telecom companies.

Customer classification is a good way for realizing churn analysis. Classification is a supervised learning task, where the input data consist of the values each example (customer) takes with respect to the attributes included in the model and the target attribute takes a categorical value refering to the class that the customer belongs to. However, the class imbalance inherited in churn analysis [7-8] as in other CRM datasets -such as fraud detection, response modeling, and credit evaluation- is a reason turning customer classification into a challenging task.

While examining the relation between the data set characteristics and the classification performance, Kwon and Sim [9] mention data imbalance as one of the characteristics that has an effect on the performance of classification algorithms. Imbalance data either distorts the performace of classification or causes overfitting and gives high accuracies.

The literature agrees on two main approaches when dealing with imbalance datasets [10];

1- Data-level approach: Preprocessing steps to balance the classes. Resampling and feature selection techniques are the main data-level approaches.
2- Algorithm-level approach: Modifications of traditional classifiers developed especially for learning from imbalanced datasets such as; one-class learning, cost-sensitive learning, ensemble methods, and recently hybrid approaches.

The literature on churn prediction in the telecommunications sector is more focused on the latter approach, yet there are a few dealing with the effects of the former.

The authors in [11] use both random under- and over-sampling prior to applying several classifying. The study in [12] internally applies four rules generation algorithms based on the rough set theory (RST) with cross validation using six over-sampling techniques on four publicly available dataset. Among the combinations of experiments the ones integrated with over-sampling show better performance as in [13] and [14] that use over-sampling, under-sampling and SMOTE with random forests in the algorithmic level. On the other hand, Verbeke et al [8] examine the effect of over-sampling on the performance of a customer churn prediction model for a telecom dataset and conclude that the dataset structure and the classification technique can change the results completely.

The authors in [15] propose using a genetic programming based approach with Adaboost algorithm and compare results of KNN and random forest for an imbalanced dataset with a 7.3% churner ratio. The study in [16] apply decision trees with Renyi and Tsallis entropies on a dataset that consists 1.96% of the samples as churners.

Ensemble classifers are popular methods in algorithm-level approach and random forest technique is oftenly preferred [17-18]. The study in [19] try to understand the perfomance of negative correlation learning (NCL) ensembles and a multilayer perceptron trained ensembles. Although the focus of the study is on these two techniques, we can easily see that support vector machines (SVM) taken as one of the reference classifiers gives the highest accuracy.

The findings of the study in [20] show that trying to have equally distributed classes is not necessary and an imbalance ratio of 1:3 (minority class: majority class) is a good option for sampling methods. The authors state that SVM can gain benefits from resampling especially through cost sensitive ones.

There are studies that propose hybrid approaches [21, 22] as well as studies combining feature selection methods either with sampling methods [7, 23] or with ensemble algorithms [24]. The study in [25] handles a huge dataset in the telecom sector. They first applied random forest for feature selection. Then, they proposed an under-sampling approach through clustering and one-sided sampling predetermining the value of k and the imbalance ratio and finally applied decision trees.

The binary classification of imbalanced datasets is a hot topic of data analytics in recent years and churn analysis is one of the application areas. Particularly churn in the telecommunications sector is a specific problem that needs a focused attention since the cost of misclassification is already known. However, the literature does not provide with mature and effective techniques but rather invests in newly developed algorithms [26]. We can safely conclude from the existing research in the field of customer churn prediction that there is not a single model that could give the highest accuracy in all of the cases. Instead, the performance of every algorithm will differ according to the characteristics of the data.

This study examines mainly the effect of a data-level approach on SVM for a binary classification task. The aim is to find the answer to the question “how does resampling affect the classification performance of support vector machines on imbalanced churn data?” for the telecom domain. Section 2 describes the dataset and the methodology followed. Section 3 explains the framework that includes the setup and the evaluation metrics. Section 4 provides the findings and elaborates on them and in the last section, the study concludes with possible research improvements to this study.

2. Modeling Churn with Class Imbalance

This section summarizes the methodology followed in this study.

2.1. Data Set Description

The telecom “churn" dataset is from the UCI Machine Learning Repository [27]. The original dataset has 3333 observations and 21 features in total. Three features -“State”, “Area.Code”, and “Phone”- are manually eliminated prior to the analyses. The imbalanced ratio is 6:1 [majority class (2850 obs.): minority class (483 obs.)].

2.2. Handling Binary Class Imbalance

The study uses resampling, a data-level approach, to handle the class imbalance problem. For this, the preferred resampling methods are;

Under-sampling techniques

- Random Under-Sampling (RUS): Samples, equal to the number of the minority class or multiples of it, are randomly drawn from the majority class.
- Clustering Based Under-Sampling (CLUSBUS): The training set is divided into groups via clustering techniques. The number of samples to be selected from each cluster belonging to the majority class is calculated and combined with the minority class units. Thus, a new training set is constructed.

Over-sampling techniques

- Random Over-Sampling (ROS): Samples are generated for the minority class so that equal number of majority class or multiples of it is achieved.
- Synthetic Minority Over-Sampling Technique (SMOTE): Samples are generated for the minority class based on the

k-nearest-neighbor method. The parameter k is used to determine the number of samples to be generated for a minority sample.

2.3. Support Vector Machines (SVM) for Classification

Support vector machines is a powerful machine learning method based on structural risk minimization [28], which is proven to show good performance especially for binary classification tasks. Although this performance is not as good for imbalanced classes, motivated by the findings in [20] this study investigates the effect of resampling methods on SVM with different kernels applied.

The customer churn dataset is a binary classification problem, where the customers are coded either as a churner (1) or a non-churner (0). The SVM models with linear, polynomial, sigmoid, and radial basis function kernels are trained once resampling methods are applied and the dataset is partitioned as training and test datasets accordingly.

3. Experimental Framework

The experimental framework of this study is briefly explained in this section. The analyses are implemented in R software. For the SVM parameter optimization, we use grid search with a small size. However, a further fine tuning is not applied in order to prevent the drastic altering of decision boundaries and hence to keep the generalization capability of the models that can resist to minor changes in data.

3.1. Experimental Setup

The study tracks the following setup path:
Step1: The training and test sets are defined in the way that they preserve the specified imbalanced ratio.
Step2: SVM with different kernel functions are trained on the training sets constructed in the previous step.
Step3: Parameters of SVM models are optimized through grid search.
Step4: The final models are selected and used for the evaluation on test sets.

The training and the test sets are split based on 5-fold cross-validation (CV). 5x2 CV method is additionally realized for random under-sampling and random over-sampling.

For the CLUSBUS sampling, Partitioning Around Medoids (PAM) is preferred as the clustering algorithm because of the mixed structure of the dataset features.

3.2. Evaluation Metrics

Evaluation metrics are important when comparing different experimental results. This study uses “Balanced Accuracy”, “Sensitivity”, and “Lift” measures, which are known to eliminate the drawbacks of traditional metrics for imbalanced classes and expose the correctly classification of the churner class, which is the goal in churn prediction tasks. Table 1 provides the basis for the computation of the evaluation metrics:

Table 1. Confusion matrix

	Classified as	
Actual	Churner	Non-churner
Churner	True Positive (TP)	False Negative (FN)
Non-churner	False Positive (FP)	True Negative (TN)

4. Results and Discussion

To answer our research question, the study includes the analyses conducted with and without resampling techniques. The tables 2-5 show the evaluation metric values of SVM results with respect to different divisions of training and test sets, imbalanced ratios (IR), and kernel functions. Table 7 shows the best results of each resampling method containing the confusion matrix values, whereas Table 6 provides the base comparison results achieved with SVM without any resampling applied.

The results show that SVM applied with RBF kernel suits better for a churn dataset. RBF generally performs better than SVM with other kernel functions for all the runs in resampling methods.

The random under-sampling method works best with the imbalanced ratios of 3:1 and 4:1 (Table 2 and Table 5). Both RUS and ROS give fairly good results with 5x2 CV applied together with RBF kernel. Yet, the best results obtained with RUS seems better than the ones with ROS (Table 2 and Table 3). The under-sampling methods, RUS and CLUSBUS, give the highest accuracies when the majority class is reduced to three and four times of the minority class. The performance results of ROS are superior than SMOTE's.

Table 2. SVM Results with Random Under-Sampling

		5-fold CV				**5x2 fold CV**			
IR	**RUS**	**Sig.**	**Lin.**	**RBF**	**Poly.**	**Sig.**	**Lin.**	**RBF**	**Poly.**
1:1	*Sensitivity*	0.39	0.41	0.54	0.50	0.36	0.38	0.47	0.46
	Bal.Acc.	0.67	0.68	0.75	0.72	0.65	0.67	0.72	0.71
	Lift	2.38	2.48	3.29	3.06	2.35	2.49	3.12	3.27
2:1	*Sensitivity*	0.62	0.53	0.54	0.66	0.42	0.46	0.44	0.61
	Bal.Acc.	0.75	0.72	0.74	0.79	0.68	0.70	0.70	0.78
	Lift	3.77	3.23	3.28	4.06	3.04	3.32	3.14	4.30
3:1	*Sensitivity*	0.44	0.44	***0.88***	0.74	0.28	0.48	***0.82***	0.65
	Bal.Acc.	0.65	0.65	***0.90***	0.83	0.62	0.69	***0.87***	0.79
	Lift	2.66	2.66	***5.39***	4.51	3.73	3.37	***5.32***	4.59
4:1	*Sensitivity*	0.45	0.44	0.76	*0.80*	0.48	0.50	***0.81***	0.76
	Bal.Acc.	0.66	0.65	0.84	*0.86*	0.68	0.69	***0.87***	0.84
	Lift	2.77	2.66	4.62	*4.88*	3.51	3.63	***5.90***	5.41

Table 3. SVM Results with Random Over-Sampling

			5-fold CV				5x2 fold CV		
IR	**ROS**	**Sig.**	**Lin.**	**RBF**	**Poly.**	**Sig.**	**Lin.**	**RBF**	**Poly.**
1:1	*Sensitivity*	0.42	0.43	0.66	0.61	0.41	0.41	*0.70*	0.56
	Bal.Acc.	0.68	0.69	0.79	0.76	0.67	0.67	*0.82*	0.73
	Lift	2.54	2.62	4.06	3.72	2.58	2.56	*5.08*	3.49
2:1	*Sensitivity*	0.53	0.48	0.70	0.61	0.53	0.50	*0.73*	0.66
	Bal.Acc.	0.72	0.70	0.80	0.76	0.71	0.70	*0.83*	0.80
	Lift	3.21	2.97	4.26	3.75	3.29	3.11	*5.29*	4.76
3:1	*Sensitivity*	0.57	0.55	0.80	0.72	0.60	0.54	***0.76***	0.64
	Bal.Acc.	0.72	0.72	0.85	0.82	0.74	0.70	***0.84***	0.79
	Lift	3.48	3.35	4.91	4.43	3.98	3.33	***5.50***	4.99

Table 4. SVM Results with SMOTE 5 fold CV

			5 fold CV		
IR	**SMOTE**	**Sig.**	**Lin.**	**RBF**	**Poly.**
1:1	*Sensitivity*	0.39	0.40	0.46	0.46
	Bal.Acc.	0.67	0.68	0.69	0.70
	Lift	2.40	2.47	2.80	2.84
1.5:1	*Sensitivity*	0.44	0.44	0.57	0.50
	Bal.Acc.	0.65	0.65	0.74	0.71
	Lift	2.66	2.66	3.49	3.08
2:1	*Sensitivity*	0.44	0.44	**0.62**	0.51
	Bal.Acc.	0.65	0.65	**0.77**	0.72
	Lift	2.66	2.66	**3.78**	3.13

The analysis done without any resampling uses the imbalanced ratio of the dataset as it is (6:1). Looking at the confusion matrix values (Table 6), it is obvious that SVM with sigmoid and linear kernel perform extremely poor in terms of detecting the churner class. This is also the reason why some metrics turn out to be non-available. On the other hand, the RBF results and give high sensitivity ratios, referring to the TP rates. An important error to be minimized for churn analysis is the FNs, referring to the actual churners who are predicted as non-churners. Polynomial kernel result seems to be more successful in terms of FN rate.

Comparing the results in Table 6 with the results in Table 7 that summarizes the best performances with resampling methods, we can say that the performance of SVM is improved when resampling is applied. The results in Table 7 are sorted based on the lift value. Overall, the values show that the resampling results improve the classification performance of SVM on imbalanced churn data the most when used with radial basis function (RBF) and with 5x2 CV.

Table 5. SVM Results with CLUSBUS

			5 fold CV		
IR	**CLUSBUS**	**Sig.**	**Lin.**	**RBF**	**Poly.**
1:1	*Sensitivity*	0.39	0.41	0.53	0.49
	Bal.Acc.	0.67	0.68	0.75	0.72
	Lift	2.38	2.48	3.21	3.02
2:1	*Sensitivity*	0.47	0.50	*0.75*	*0.68*
	Bal.Acc.	0.69	0.71	*0.85*	*0.81*
	Lift	2.86	3.06	*4.56*	*4.17*
3:1	*Sensitivity*	0.43	0.46	*0.73*	*0.72*
	Bal.Acc.	0.65	0.67	*0.83*	*0.83*
	Lift	2.64	2.82	*4.45*	*4.43*
4:1	*Sensitivity*	0.47	0.51	***0.86***	0.76
	Bal.Acc.	0.67	0.69	***0.89***	0.84
	Lift	2.87	3.12	***5.26***	4.62

Table 6. SVM Results without Using Resampling

		IR	TN	FP	FN	TP	Sensitivity	Bal. Accuracy	Lift
5-fold CV	**Sigmoid**	6:1	558	109	0	0	NA	NA	NA
	Linear	6:1	558	109	0	0	NA	NA	NA
	RBF	6:1	547	46	11	63	0.85	0.89	5.21
	Polynomial	6:1	549	48	9	61	0.87	0.90	5.33

Table 7. Best SVM Results with Resampling Methods

		IR	TN	FP	FN	TP	Sensitivity	Bal. Accuracy	Lift
5x2 fold CV	**RUS+RBF**	4:1	1412	115	27	113	0.81	0.87	5.90
5x2 fold CV	**ROS+RBF**	3:1	1399	112	37	119	0.76	0.84	5.50
5-fold CV	**RUS+RBF**	3:1	549	43	9	66	0.88	0.90	5.39
5-fold CV	**CLUSBUS+RBF**	4:1	548	48	10	61	0.86	0.89	5.26

5. Conclusion

The purpose of this study is to investigate the effects of resampling techniques on Support Vector Machines for imbalanced customer churn data. We can conclude that resampling techniques, especially random under-sampling improves the classification performance of SVM. Also, support vector machines with RBF yields a better performance than the other kernel functions. The study handles the imbalanced dataset classification from a data-level approach. As we now see how these resampling techniques affect the performance of SVM for a telecom customer churn dataset, other under-resampling techniques could also be investigated. A further improvement to this study would be to formulize the problem as a multi-classification task and see whether resampling improves the classification performance in such setting.

References

[1] Krishna, G. and Ravi V. (2016). Evolutionary computing applied to customer relationship management: A survey. *Engineering Applications of Artificial Intelligence*, 56 (November), 30-59.

[2] Singh, H. and Samalia H.V. (2014). A business intelligence perspective for churn management. *Procedia – Social and Behavioral Sciences*, 109 (January), 51-56.

[3] Reichheld, F. and Schefter P. (2000). E-loyalty: Your secret weapon on the web. *Harvard Business Review*, 78 (July-August), 105-113.

[4] Mattison, R. (2005). The telco churn management handbook, null edition, Lulu.com, XiT Press, Oakwood Hills, Illinois, USA.

[5] Zhang, Y., et al. (2011). Behavior-based telecommunication churn prediction with neural network approach. *Proceedings – 2011 International Symposium on Computer Science and Society, ISCCS 2011*, 307-310.

[6] Lu, J. (2002). Predicting customer churn in the telecommunications industry- An application of survival analysis modeling using SAS. *SAS User Group International (SUG127) Online Proceedings*, 114-127.

[7] Idris, A., Riswan M. and Khan A. (2012). Churn prediction in telecom using Random Forest and PSO based data balancing in combination with various feature selection strategies. *Computers and Electrical Engineering*, 38 (6), 1808-1819.

[8] Verbeke, W., et al. (2012). New insights into churn prediction in the telecommunication sector: A profit driven data mining approach. *European Journal of Operational Research*, 218 (1), 211-229.

[9] Kwon, O. and Sim J.M. (2013). Effects of data set features on the performances of classification algorithms. *Expert Systems with Applications*, *40* (5), 1847-1857.

[10] Ali, A., Shamsuddin S.M. and Ralescu A.L. (2015). Classification with class imbalance problem: A review. *International Journal Advances in Soft Computing and its Applications*, 7 (3), 176-204.

[11] Qureshi, S.A., et al. (2013). Telecommunication subscribers' churn prediction model using machine learning. *Eighth International Conference on Digital Information Management (ICDIM 2013)*. IEEE, 131-136.

[12] Amin, A., et al. (2016). Comparing oversampling techniques to handle the class imbalance problem: A customer churn prediction case study. *IEEE Access*, (4), 7940-7957.

[13] Gui, C. (2017). Analysis of imbalanced data set problem: The case of churn prediction for telecommunication. *Artificial Intelligence Research*, 6 (2), 93-99.

[14] Hanif, A. and Azhar N. (2017). Resolving class imbalance and feature selection in customer churn dataset. In: *2017 International Conference on Frontiers of Information Technology (FIT)*. IEEE, 82-86.

[15] Idris, A., Khan A. and Lee Y.S. (2012). Genetic programming and adaboosting based churn prediction for telecom. In: *2012 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*. IEEE, 1328-1332.

[16] Gajowniczek, K., Ząbkowski T. and Orłowski, A. (2015). Comparison of decision trees with Rényi and Tsallis entropy applied for imbalanced churn dataset. In: *2015 Federated Conference on Computer Science and Information Systems (FedCSIS), IEEE,* 39-44.

[17] De Bock, K.W. and Van Den Poel, D. (2011). An empirical evaluation of rotation-based ensemble classifiers for customer churn prediction. *Expert Systems with Applications,* 38 (10), 12293-12301.

[18] Xie, Y. et al. (2009). Customer churn prediction using improved balanced random forests. *Expert Systems with Applications*, 36 (3), 5445-5449.

[19] Rodan, A. et al. (2015). Negative correlation learning for customer churn prediction: A comparison study. *The Scientific World Journal*, 1-7. http://dx.doi.org/10.1155/2015/473283

[20] Zhu, B. et al. (2018). Benchmarking sampling techniques for imbalance learning in churn prediction. *Journal of the Operational Research Society*, 69 (1), 49-65.

[21] Idris, A. and Khan A. (2016). Churn prediction system for telecom using filter–wrapper and ensemble classification. *The Computer Journal*, 60 (3), 410-430.

[22] Ahmed, A.A. and Maheswari, D. (2017). Churn prediction on huge telecom data using hybrid firefly based classification. *Egyptian Informatics Journal*, 18 (3), 215-220.

[23] Kim, Y. (2006). Toward a successful CRM: variable selection, sampling, and ensemble. *Decision Support Systems*, 41 (2), 542-553.

[24] Idris, A., Khan, A. and Lee, Y.S. (2013). Intelligent churn prediction in telecom: employing mRMR feature selection and RotBoost based ensemble classification. *Applied intelligence*, *39* (3), 659-672.

[25] Li, H. et al. (2016). Supervised massive data analysis for telecommunication customer churn prediction. In *2016 IEEE International Conferences on Big Data and Cloud Computing (BDCloud), Social Computing and Networking (SocialCom), Sustainable Computing and Communications (SustainCom) (BDCloud-SocialCom SustainCom), IEEE,* 163-169.

[26] Haixiang, G. et al. (2017). Learning from class-imbalanced data: Review of methods and applications. *Expert Systems with Applications*, *73*, 220-239.

[27] Blake, C.L. and Merz, C.J. (1998). Churn Data Set, UCI Repository of Machine Learning Databases, http://www.ics.uci.edu/∼mlearn/MLRepository.html, University of California, Department of Information and Computer Science, Irvine, CA.

[28] Vapnik, V.N. (1999). An overview of statistical learning theory. *IEEE transactions on neural networks, 10* (5), 988-999.