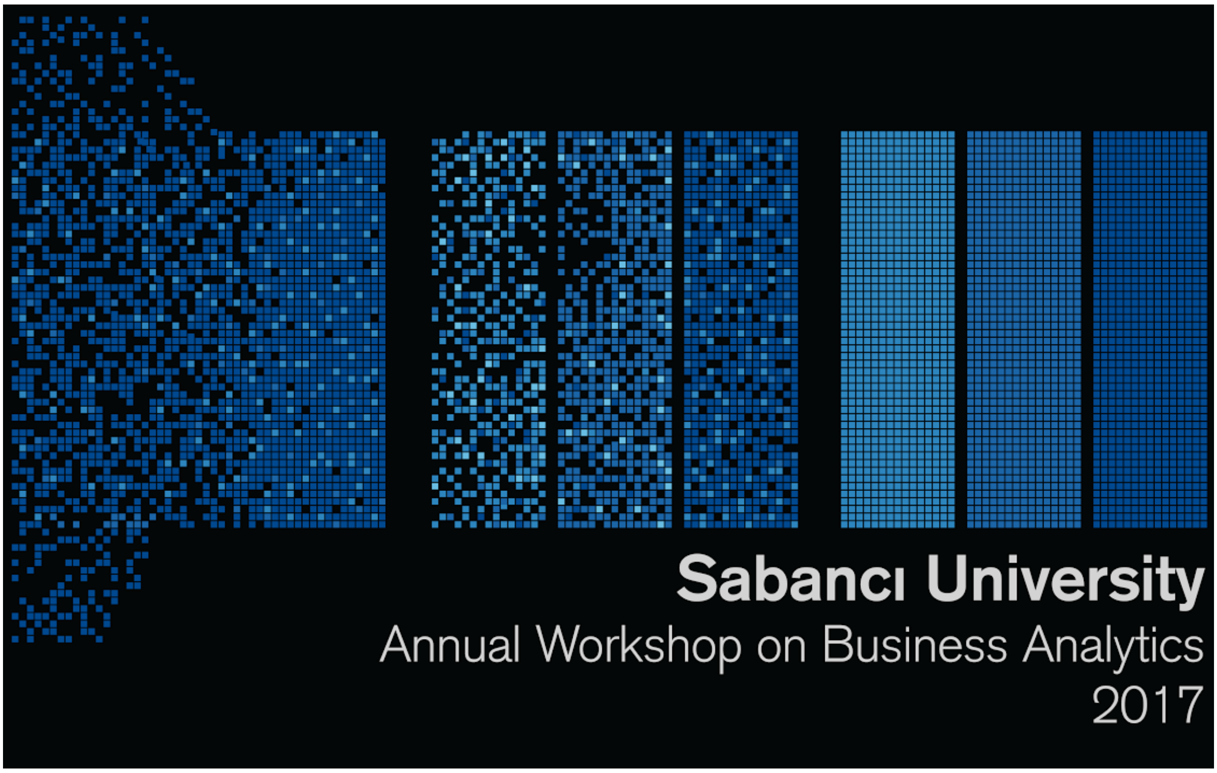

Sabancı University

Annual Workshop on Business Analytics 2017

October 13, 2017

Program

09:00- 09:10:	Registration
09:10- 09:15:	Opening remarks
09:15- 10:00:	Keynote speech Gül Ege, Senior Director, SAS Advanced Analytics R&D
10:00 – 11:15:	Session #1: Marketing Analytics “Power of Analytics: Two Applications in Apparel Retail”, Mustafa Gökçe Baydoğan, Boğaziçi University “Who Benefits from Brand Exits? Why?”, Barış Depecik, Bilkent University “Using Data Analytics for Customer Churn Prediction: A Case Study for an Online Retailer”, Birol Yüceoğlu, Migros
11:15 – 11:30:	Coffee break
11:30 – 11:50:	Software tutorial #1: BonAir
11:50 – 13:05:	Session #2: New Methods “Are the labels informative enough? -Semi–Supervised Probabilistic Distance Clustering and the Uncertainty of Classification”, Cem İyigün, Middle East Technical University “Variable Selection in Regression using Maximal Correlation and Distance Correlation”, Deniz Yenigün, Bilgi University “Independent Agent Segmentation for the Insurance Sector”, Ece Egemen, Sabancı University “Insurance Fraud Detection: A Network Modeling and Behavioral Analytics Approach”, Gökhan Göktürk, Sabancı University
13:05 – 14:00:	Lunch break
14:00 – 14:40:	Software tutorial #2: SAS
14:40 – 15:10:	Coffee break and poster session
15:10 – 16:25:	Session #3: Healthcare Applications in Analytics “A Decision Support System for Contract Negotiation Processes of the Health Care Insurance Companies”, Kemal Kılıç, Sabancı University “Identifying key miRNA-mRNA regulatory programs in cancer using Bayesian reduced-rank regression”, Mehmet Gönen, Koç University “Next-Day Stochastic Operating Room Scheduling with Limited Historical Data”, Enis Kayış, Özyeğin University
16:25 – 17:00:	Software tutorial #3: TIMI
17:00 – 18:00:	Closing cocktail

Abstracts

Keynote Speech

"Business Value of Advanced Analytics and Current Trends", Dr. Gül Ege, Senior Director, Advanced Analytics R&D, SAS Institute Inc, Cary, NC

Leading with analytics creates a competitive edge across all business domains. As the scale, nature, and speed of data rapidly increase, the need for innovations in analytical algorithms and technology becomes critical. The expectation is that data-driven business decisions will need to be revisited more rapidly, near real time, and sometimes real time.

Often, solutions to the most challenging problems require the invention of new techniques and algorithms across multiple analytical disciplines such as forecasting, estimation, predictive modeling, data mining and optimization.

This presentation will provide examples from multiple business domains, discuss the challenges of new data sources, the requirements for performance in streaming data, and highlight the importance of analytical talent to realize maximum business value from the potential created by technology and computational innovations.

Session #1: Marketing Analytics

"Power of Analytics: Two Applications in Apparel Retail", Mustafa Gökçe Baydoğan, Boğaziçi University

We describe two cases of data-driven and optimization-based analytics applications in apparel retailing. i) Attribute-based demand estimation and Price Elasticity and Demand Estimation for Markdown Price Optimization: In fashion retailing, a key input in markdown price optimization is the estimation of price elasticity and forecasting of demand for products with a short demand history. We observe empirically that products with distinctively different product characteristics or inventory distribution patterns yield significantly different reactions to price changes. Retailers have historically relied on managers' experience to differentiate between products with different characteristics. We develop a methodology to identify key product attributes that explain the variation in historical item-response to price changes. We then embed an attribute-based price elasticity estimation technique in item-level demand forecasting. Finally, we demonstrate the benefits of an online learning layer in demand forecasting. We find that this methods result in significantly more accurate price elasticity estimation and demand forecasts, which in turn lead to more profitable markdown decisions. We describe the implementation results with a fashion retailer achieving 300 basis points higher gross margin. ii) Inventory allocation of regular and new products: We describe the development of a forecasting and allocation system for managing the flow of inventory of products to stores in an apparel retailer. We develop a methodology to estimate demand for products with no history or little demand data and feed that into a stochastic inventory allocation model. Live controlled experiments at a fashion retailer demonstrate a 4-5% reduction in lost sales and a parallel increase in revenues.

"Who Benefits from Brand Exits? Why?", Barış Depecik, Bilkent University

The increasing frequency of brand exits raises two questions pertinent for both manufacturers and retailers: When a brand disappears from the market, (1) what brands are better positioned to benefit from the exit? and (2) what marketing efforts do influence the realignment of sales after the exit? To answer these questions, we develop a dynamic brand sales response model. The model allows for examination of the long-term effects of a brand exit on sales and identification of the drivers of excess demand redistribution following the exit. We apply the model to 232 brand exit events in five repeat-purchase categories and examine the effects of these exits on 1802 incumbent brands' sales. We find that (1) frequent sales promotions in the post-exit period bring bigger permanent gains from brand exits, (2) incumbent brands which have product portfolios similar to those of the deleted brands gain disproportionally more after exits, (3) brand-specific characteristics explain variations in sales response, yet most variables have complex net effects, and these effects vary by category. Our results generate insights that will help manufacturers and retailers better manage their product portfolios and marketing investments to maximize gains from brand exits.

"Using Data Analytics for Customer Churn Prediction: A Case Study for an Online Retailer", Birol Yüceoğlu, Migros

We present a new decision support system for predicting customer churn for a leading online fast-moving consumer goods retailer. The proposed system is based on the recent data analytics and machine learning tools. The main steps of our development are the data cleaning, the feature extraction and the prediction with supervised learning methods. The prediction step requires a careful definition of a churning customer in the retail sector for it is not obvious to identify the agents in such a non-contractual setting. After discussing different scenarios for customer attrition, we conduct an empirical study by using the past data of the company. We present the proposed decision support system as a promising application of data analytics to predict churn in online retail. We discuss the results of our tests and features affecting customer churn.

Session #2: New Methods

"Are the labels informative enough? Semi–Supervised Probabilistic Distance Clustering and the Uncertainty of Classification", Cem İyigün, Middle East Technical University

In this study we first discuss unsupervised and semi-supervised clustering and then focus on the latter one. Semi–supervised clustering is an attempt to reconcile clustering (unsupervised learning) and classification (supervised learning, using prior information on the data.) These two modes of data analysis are combined in a parameterized model. The results (cluster centers, classification rule) depend on the parameter θ, an insensitivity to θ indicates that the prior information is in agreement with the intrinsic cluster structure, and is otherwise redundant. This explains why some data sets in the literature give good results for all reasonable classification methods. The uncertainty of classification is represented here by the geometric mean of the membership probabilities, shown to be an entropic distance related to the Kullback–Leibler divergence.

"Variable Selection in Regression using Maximal Correlation and Distance Correlation", Deniz Yenigün, Bilgi University

In most of the regression problems the first task is to select the most influential predictors explaining the response, and removing the others from the model. These problems are usually referred to as the variable selection problems in the statistical literature. Numerous methods have been proposed in this field, most of which address linear models. In this study we propose two variable selection criteria for regression based on two powerful dependence measures, maximal correlation and distance correlation. We focus on these two measures since they fully or partially satisfy the Renyi postulates for dependence measures, and thus they are able to detect nonlinear dependence structures. Therefore, our methods are considered to be appropriate in linear as well as nonlinear regression models. Both methods are easy to implement and they perform well. We illustrate the performances of the proposed methods via simulations, and compare them with two benchmark methods, stepwise AIC and lasso. In several cases with linear dependence all four methods turned out to be comparable. In the presence of nonlinear or uncorrelated dependencies, we observed that our proposed methods may be favorable. An application of the proposed methods to a real financial data set is also provided.

"Independent Agent Segmentation for the Insurance Sector", Ece Egemen, Sabancı University

As one of the leading players in a fragmented insurance sector in Turkey, Aksigorta aims to better model its network of independent agents and the behavioral groups that reflect agent channel dynamics in the industry. In this study, we use Machine Learning approaches to create agent segmentation models in three dimensions: *efficiency segmentation* based on a proposed efficiency model by insurance product group; *response segmentation* that models an agent response index as the peak time differences of production with respect to total Aksigorta production; and governance segmentation based on a beta index calculated by comparing the volume productions of the agents and the total Aksigorta volume productions. In our analysis, we use internal company data including insurance premium, commission and volume of sales for each product group for each agent in 2014 and 2015. We also use external data such as real estate sales, foreign import/export, volumetric and quantitative information on credit and deposit types for the same period, as well as point of interest (POI) data around each agent location. We calculate the agent segment level (e.g. strong, medium, weak) under each dimension and together with the sub-segmentations, we produce a total of 36 clusters for 1967 active independent agents. Implementation of the results by Aksigorta is in progress, as the company plans to use the new agent segments in a number of strategic and operational moves.

"Insurance Fraud Detection: A Network Modeling and Behavioral Analytics Approach", Gökhan Göktürk, Sabancı University

Insurance fraud arises in a variety of contexts and sectors such as home, auto and health insurance. Automobile claim fraud added $7+bn in excess payments to auto-injury claims paid in the U.S. in 2012 where insurers have identified potential fraud in 7.4% of auto claims. In Turkey, the Insurance Information and Monitoring Center (IIMC), a nationwide not-for-profit agency, reports fewer insurance claims as fraudulent compared to the U.S. and E.U. benchmarks, which suggests that more fraud cases can potentially be found. Avoiding undue costs from such cases would clearly help the industry and customers who will pay less premiums in the long run. In this paper, we analyze a unique setting where IIMC oversees all insurance contracts and claims in Turkey, and is in a unique central position to detect fraud. While studies have attempted insurance fraud detection using data mining and machine learning, none to our knowledge has exploited the availability of such a comprehensive dataset. We consider such a dataset from an auto accident insurance fraud viewpoint and propose a two-stage analytical approach to identify fraud cases: 1) develop a network-based model to link various parties involved, and use it to identify (behavioral) predictor variables; 2) use a gradient boosting algorithm to classify accident loss claims as fraudulent or not. Our approach trains to 97% success rate with the known cases of fraud, and identifies thousands of new cases for the IIMC to investigate.

Session #3: Health Care Applications in Analytics

“**A Decision Support System for Contract Negotiation Processes of the Health Care Insurance Companies**”, Kemal Kılıç, Sabancı University

As the main players in health insurance, the healthcare insurance companies (HCICs) have a crucial role in healthcare ecosystem. Basically, the HCICs collect the premiums of the policies from the insurees, and pay for the services delivered by the Healthcare Providers (HCPs) to their customers. In order to keep the sustainability, HCICs have to determine the optimal premiums for the policies of the insurees and the optimal prices for the healthcare services in the service contracts signed with HCPs. The service contracts are mostly renewed annually after the peer to peer negotiations with HCPs. But since the HCPs are more informed regarding to the healthcare services provided to the patients, there is an information asymmetry between HCISs and HCPs during the service contract negotiations in the favor of HCPs. This creates a disadvantageous situation for the HCISs during the negotiation processes. In this study, we developed a decision support system (DSS) framework for the HCISs which shortlists the services that a discount should be considered in the negotiations. Note that it is neither feasible (due to the time restrictions) nor pleasant to request a discount on each medical service provided by the HCPs during the negotiations. Therefore, only a few of the items which would both yield substantial return (in terms of cost reduction) to the HCISs and at the same time are acceptable by the HCPs should be brought to the negotiation table. As a result the problem is treated as a multi criteria decision making problem and infamous AHP is utilized as part of the model developed for the HCISs. As part of the research, novel conceptualizations of the acceptability score and the return of the discount are made and real life historical data is extensively used in order to determine their values for each possible (HCP, cost item, discount percentage) triplets. As part of the developed framework, the options that are located at the Pareto Efficient Frontier are shortlisted for the HCISs to be considered during their negotiation process.

“Identifying key miRNA-mRNA regulatory programs in cancer using Bayesian reduced-rank regression”, Mehmet Gönen, Koç University

MicroRNAs (miRNAs) are a class of noncoding RNAs that regulate messenger RNAs (mRNAs) either by degradation or by translational repression. miRNA alterations have a major effect in the initiation and progression of human cancer. That is why it is quite important to identify cancer-specific miRNA-mRNA regulatory programs. In this study, we found regulatory programs of 31 different cancer types using a novel Bayesian reduced-rank regression model on matched miRNA and mRNA profiles of around 8,000 primary tumors.

“Next-Day Stochastic Operating Room Scheduling with Limited Historical Data”,
Enis Kayış, Özyeğin University

Effective operating room (OR) scheduling when surgery durations are uncertain requires accurate surgery duration estimates. True surgery duration distributions are not known in practice and hence estimates are used as a proxy. Unfortunately, these estimates are developed using limited past data. In this talk, we present parametric and nonparametric models for next day OR scheduling, quantify and analyze the effect of limited past data on generating OR schedules. We find that the number of data points may not be practically available in most settings to generate near-optimal schedules.