Introduction To
Big Data and Analytics: Pathways to Maturity
“The Original Big Data and Analytics Minitrack”

Stephen H. Kaisler, D.Sc.
SHK & Associates
Laurel, D 20723
Skaisler1@comcast.net
(301) 498-4244

Frank J. Armour, Ph.D.
Kogod School of Business
American University
Washington, DC 20016
fjarmour@gmail.com
(202) 251-3554

J. Alberto Espinosa, Ph.D.
Kogod School of Business
American University
Washington, DC 20016
alberto@american.edu
(202) 885-1958

The Big Data and Analytics minitrack of the Decision Analytics, Mobile Services, and Service Science has selected six papers to constitute this minitrack. This year the majority of papers focused on techniques for improving analytical approaches.

Our first paper, “Easy and Efficient Hyperparameter Optimization (HPO) to Address Some Artificial Intelligence “ilities””, by Trevor Bihl, Joe Schoenbeck, Daniel Steeneck, and Jeremy Jordan, addresses improving the selection of parameters for AI applications that yield robust results. Program parameters are often learned experimentally and experientally. Applying the same algorithm or set of rules to different domains or problems may yield less than satisfactory results. Automating parameter optimization can lead to faster deployment of AI applications that meet the desired levels of the “ilities”, such as reliability, repeatability, explainability, and usability, among others, that are demanded of production systems.

The authors note that there “are no hard and fast rules” for hyperparameter selection. Moreover, hyperparameter selection depend on the data itself. Thus, they are part of the “art of algorithm design”. They note that there several approaches to HPO, but each requires some advanced knowledge of mathematic and algorithms as well as a deep understanding of the domain at the level of a subject matter expert.

The authors have developed a framework which embeds CRISP-DM (CRoss-Industry Standard Process for Data Mining) and show how this technique facilitates the use of general methods for HPO. Finally, they provide a short taxonomy of AI HPO methods.

Our second paper, “Understanding Customer Preferences Using Image Classification – A Case Study of an Online Travel Community”, by Ines Brusch, is an innovative application of standard classification methods – SVM and CNN -to image data. The author believes that the plethora of travel images posted online on social media can be used to identify user travel preferences. Drawing on previous research, the author believes that the content of images can be analyzed to identify common locales. From this data, user preferences for locales can be extracted and used by travel companies to improve recommendations to customers for their next journey.

Her analysis involves a two-part process. First, for a given travel portal, user-supplied images are categorized using image analysis methods. In the second step, data about real users and their images was captured. These images were automatically classified into categories such as food and beverage, mountain panoramas, and cityscapes. The images were then segmented using cluster analysis. The segments were then compared with the holiday styles provided by the user. The end result was that at least one travel style of the user could be correctly identified in ~93% of the cases.

This paper shows that combining user –supplied preference data about travel locations and associated activities coupled with detailed analysis of images can be used to build a profile individual users that can be used by recommender systems to provide information to users about future destinations. The author notes that a richer set of image analysis methods can yield information that can be combined to form a richer profile. And, extracting data about locations and activities from Internet sources can provide the basis for better recommendations.

Our third paper, “Model Interpretation and Explainability”, by Dan Dolk and Dan Kridel, addresses the problem of how AI systems explain how they have reached the answers they did. This area has only recently received significant attention from the research community about how to capture the analysis process in order to generate cogent, coherent explanations for the user. The note that AI/ML programs are becoming sophisticated enough that they

URI: https://hdl.handle.net/10125/63856
978-0-9981331-3-3

may "soon outstrip human ability to understand and manage their results". While we are skeptical of this observation, it is essential to begin improving the explanatory capability of AI/ML programs in order to explain to decision makers what the programs have derived in the way of results.

The authors examine several analytical methods and several explanatory techniques. They applied several standard models to a dataset of 20,000 observations. They then used the explainability techniques (SKLEARN, GAM, SHAP and LIME) the predictions of each of the analytical methods and the output of the models.

In the author's sense, given the analytical techniques were making binary decisions regarding loan applications, they determined which features had the greatest impact given the expected predictions. They rightly note that complex explanatory techniques are in nascent stage. Their major contribution is to identify a discontinuity between static and dynamic explainability models. Further, they note that the explainability techniques themselves need to be explained to end users as well.

Our fourth paper, entitled "A New Metric for Lumpy and Intermittent Demand Forecasts: Stock-keeping-oriented Prediction Error Costs", by Dominik Martin, Phillipp Spitzer, and Niklas Kuhl, presents a new metric for assessing the accuracy of predictive results from a model. As the authors note, "there is no overall best performance metric" which can be applied to any forecasting problem. In particular, traditional measures fail spectacularly when dealing with intermittent demands.

Internittent demand is often characterized by lumpy intervals often having large fluctuations in the actual demand. As a result, algorithms expecting a relatively smooth event sequences and/or time series often lead to misleading results. The authors propose a new metric which measures the cost of the difference between actual and predicted values. A perfect prediction should yield a metric value of zero. The greater the deviation the greater the cost. The proposed metric - *Stock-keeping-oriented Prediction Error Costs (SPEC)*- calculates an error term for each time step of the forecast. SPEC calculates errors in both magnitude and time.

Using both simulated and real data, the authors evaluated the performance of SPEC. They demonstrated that SPEC generates both reliable and valid results compared to other metrics. As a result, given historical data and forecast data, SPEC can assess how good the predicted data are given valid historical data. By tweaking the alpha parameters, SPEC can forecast an overall cost out to he forecast horizon. This provides organizations with one mechanism for attempting to tune demand given a set of resources.

Our fifth paper, entitled "Exploring Critical Success Factors in Agile Analytics Projects", by Mikhail Tsoy and D. Sandy Staples, examines critical success factors in agile analytics projects. This area has not been extensively studied. The authors add ten new factors to the success factors proposed by Chow and Cao [1]. Their success factors were proposed over ten years ago when agile analytics was relatively immature. Now that agile analytics has entered an early mature phase, revisiting success factors can shed new light on how to evaluate analytics.

Through a literature survey, the authors identified additional factors to be considered from other projects because, as they noted, no papers directly addressed analytics projects. The organized the combined set of success factors into twelve attributes. The authors selected two projects for study – one reasonably successful and one not very successful. Through interviews, they gathered data for analysis. The bulk of the paper presents their analysis of the two projects according to these attributes and their findings.

The successful project had many of these attributes at a very strong level, while the unsuccessful project did not have many of these attributes. The two projects served to demonstrate that the revised attributes incorporating the success factors are essential to a successful analytics project. These results provide guidance to project managers undertaking agile analytics project about aspects of the project to focus on to help in successful execution.

One paper was withdrawn. The co-chairs encourages the authors to resubmit for HICSS-54 in 2021.

The co-chairs believe that several of these papers yield innovative results that, further developed and applied to larger data sets, will provide the basis for tools to assist organizations in managing their business operations. We note that the transition from research to viable tools that can be used on a periodic basis for assessing business operations often a difficult one and make take considerable time from when the research is first reported to the availability of viable tools. We continue to encourage this type of research as well as case studies and practical applications in order to further the methods, tools, and techniques available to organizational managers.

The co-chairs thank all authors who submitted papers to the HICS-53 Big Data and Analytics minitrack. And we thank all those authors who participated in the reviewing process to select the six papers represented by these two sessions. For authors whose papers were not selected, please review the comments and consider submitting a revised and/or enhanced paper to HICSS-54.

Respectfully,

Steve Kaisler
Frank Armour
Alberto Espinosa
Big Data and Analytics Minitrack Co-Chairs

References

[1] Chow, T., and D.-B. Cao, "A survey study of critical success factors in agile software projects", *Journal of Systems and Software 81*(6), 2008, pp. 961–971