2017 2nd International Conference on Telecommunication and Networks (TEL-NET 2017)

Challenges of Automated Machine Learning on Causal Impact Analytics for Policy Evaluation

Yuh-Jong Hu and Shu-Wei Huang
Dept. of Computer Science, NCCU Taipei, Taiwan
{yjong.hu, wei.90211} at gmail.com

***Abstract*—Automated machine learning (AutoML) refers to the full aspects of automation machine learning without human in the analytics loop. The main goals of big data analytics are to determine correlation, prediction, and cause-effect among high-dimensional data features. Until now, AutoML systems were primarily proposed for classification and regression, and lacked causal impact analytics. In this study, we address the possible challenges of extending AutoML on causal impact analytics for a policy evaluation. A simplified causal inference model has been implemented on a generic AutoML system Spark ML pipeline for a scenario of policy evaluation for (inter-)national stock market impacts analytics based on GDELT big datasets.**

***Index Terms*—Big data analytics, automated machine learning (AutoML), causal impact analytics, counterfactuals, Bayesian Structural Time Series (BSTS), high-dimensional machine learning, policy evaluation.**

I. Introduction

Data science has been a research field for several decades. Harvard Business Review calls data science the sexiest job in the 21st century [1]. In fact, this hot new field promises to revolutionize various industries, from business to government, health care to academia. A high volume and dimension of big data are constantly generated and collected from different sources throughout the Web and mobile devices. Big data with diverse formats are stored in so-called Datalakes in the cloud computing platforms for high availability and performance.

High-dimensional data values are extracted for a statistical learning model to derive correlation between feature (or predictor) and response variables through regression or classification techniques [2]. Moreover, we can predict possible future outcomes and even answer causal impact queries about the effects of interventions for a policy evaluation [3].

Big data analytics, an emerging related field of data science, is an interdisciplinary research area that includes machine (or statistical) learning, cloud computing, and other related application domains. The primary challenge of big data analytics is to devise a new foundation to manage high-dimensional datasets to effectively balance between bias and variance in an optimal learning model construction process.

There are several research challenges on high-dimensional data analytics from a machine learning perspective:

First, we are facing intensive computation and storage requirements in order to satisfy the resources' need of data extraction, transformation, load (ETL) processing, and model construction. The large size and variety of datasets are beyond the ability of commonly used hardware and software tools to capture, manage, and process information adequately. Therefore, a standalone computer cannot meet this high computing and storage demand. We need high performance and flexible cloud computing platforms to tackle this problem.

Second, we can construct a complex analytic model with a low bias to fit a training dataset. However, this will increase its variance when this model is applied to the unknown response of a testing dataset. Therefore, we should balance the bias-variance while searching for an optimal model from an enormous amount of features combinations.

Third, in addition to balancing between statistical accuracy and computational efficiency, we need to address the problems of data heterogeneity, noise accumulation, spurious correlations, and incidental endogeneity in the high-dimensional datasets analytics [4]. These issues make traditional statistical methods invalid, especially regarding the feature selection of a model construction for high-dimensional datasets.

Fourth, the goals of big data analytics can be classified into three levels: (1) discover features correlations, (2) predict response feature outcomes, and (3) perform causal inference on counterfactuals. Counterfactuals can be defined as individual treatment effects for precision medicine or average treatment effects for a specific policy evaluation.

While one can discover a model's feature correlations its response values may be not able to be accurately predicted. Furthermore, while one can predict a model's response values, it might not be possible to explain which intervention (or causal) factors are effective in the causal inference [5].

Compared with standard statistical analysis, such as regression and classification, the purpose of causal inference goes one step further: its aim is to infer not only beliefs or probabilities under static conditions, but also the dynamics of beliefs under changing conditions. In fact, correlation does not imply causation [3] [5].

Fifth, AutoML intends to take the machine learning expert out of a data analytics processing loop. The big data analytics pipeline includes a consecutive series of stages: preprocess raw data, decide appropriate algorithms for modeling, select appropriate features, and optimize a model's hyperparameters to justify an optimal model with a low bias and variance combination, and finally, predict new data through a statistical inference with the optimal model. An AutoML system faces a challenge to automatically perform these processes without human intervention.

Moreover, a big data analytics pipeline also consists of several feedback loops in order to improve the quantity of a model's analytics. In a high-dimensional data analytics scenario, we face several research issues while building an AutoML system. For example, how do we select a feasible machine learning algorithm, which is suitable for a large scale of features with possibly sparse and missing values in some features? Once a machine learning algorithm is decided, how do we optimize a learning model by tweaking its hyperparameters? In addition, which stages of the data analytics pipeline can be possibly automated, and which stages cannot?

In this study, we first exploit various AutoML techniques to answer above proposed research issues. Our final ambitious goal is to address the challenges of automated causal inference for a high-dimensional dataset with known (or unknown) causal model in the cloud computing platform. Currently, this study aims to enhance current AutoMLs, which only tackles the problem of causal impact analytics with a known causal model for a specific intervention factor. In contrast with existing machine learning algorithms, which only focus on correlation and prediction, the Bayesian Structural Time Series (BSTS) model, a type of structural equation model, is more appropriate to infer causal impacts for a policy evaluation [6].

We intend to construct an AutoML pipeline with BSTS to perform causal inference on high-dimensional datasets with a scenario of policy evaluation. The simplified AutoML system for causal inference has been implemented to track and discover the (inter-)national stock market impacts by using the GDELT [1] collective news media big datasets.

A. Research Issues and Contributions

This paper addresses the following major *research issues*:

1) How do we construct an AutoML system through optimization of a machine learning algorithm selection with its hyperparameters?
2) How can we apply causal inference on high-dimensional datasets, where numerous samples might have missing or sparse values with possibly imbalanced response values?
3) What are the major challenges to extend current AutoML systems for causal model discovery and inference?
4) How can a real policy evaluation scenario be implemented to justify the feasibility of using a simplified causal impact analytics technique to achieve the objective of AutoML system in the cloud computing platform?

This paper makes the following three contributions: First, we have established an extended AutoML system with causal impact analytics services in the cloud computing environment. This extends the current AutoML systems, which only focus on classification and regression problems without providing cause-effect analytics. Second, we point out the major challenges and obstacles on establishing a full set of AutoML systems for known (and unknown) causal model discovery and inference. Third, we have implemented a real policy evaluation scenario on (inter-)national stock market impacts analytics by using an extended AutoML system.

This paper is organized as follows. In Section I, we give an introduction. Then, we provide a brief background knowledge in Section II. In Section III, we address related work. In Section IV, we investigate current AutoML system techniques and argue why they are not robust enough to handle the causal impact analytics problem. In Section V, we point out the possible challenges to fully extend current AutoML systems regarding automated causal model discovery and inference. We propose an extended AutoML pipeline system with lightweight causal impact analytics capabilities. A real policy evaluation scenario for inferencing (inter-)national stock market impacts is demonstrated in Section VI. Finally, in Section VII, we conclude this paper and point out possible future work.

II. Background

AutoML can be formalized as a *Combined Algorithm Selection and Hyperparameter optimization (CASH)* problem used by `Auto-WEKA` [7]. The CASH problem is to find the joint machine learning algorithm and hyperparameter setting that minimizes the loss function over the training and validation datasets. The `Auto-Sklearn` system, based on the popular machine learning framework Scikit-learn [8], outperforms previous Auto-ML systems, such as `Auto-WEKA 2.0` [9] and `Hyperopt-Sklearn` [10] at the ChaLearn AutoML Challenge 2015 [11].

AutoML Challenge ChaLearn is a competition event to develop state-of-the-art in fully automatic machine learning without any human intervention on a wide range of problems. However, this event mainly offered challengers the opportunity to solve classification problems, fewer opportunities to address regression problems were offered [12] [13]. Automated causal inference challenge was not even addressed in this event except the Causality Workbench challenge[2].

We are aware of the getting importance of solving causal inference problems for counterfactuals analytics. Therefore, we are exploiting this research issue and proposing a preliminary automated machine learning pipeline for causal impact analytics. We further point out the possible research challenges of AutoML on big data analytics from a causal inference perspective.

III. Related Work

`Scikit-learn` is a Python-based open source package for numerous machine learning algorithms to enforce

[1]GDELT - Global Data on Events, Location and Tone (http://www.gdeltproject.org/)

[2]See http://www.causality.inf.ethz.ch/

the basic function of AutoML [8]. It only aims to solve medium-scale supervised and unsupervised learning problems. `Auto-WEKA` is another AutoML approach based on a widely used open source WEKA package [7]. `Auto-WEKA 2.0` is an upgraded version of `Auto-WEKA` with regression capabilities and parallel execution of model search optimization [9]. `Auto-Sklearn` was the winner of the 2015 AutoML Challenge ChaLearn event, as it extends the capabilities of `Auto-WEKA` and `Scikit-learn` systems.

`Auto-Sklearn` initially applies meta-learning to find good instantiations of machine learning frameworks across datasets. It utilizes the Bayesian optimizer to find a best-performing hyperparameter setting and retain all the candidate models it automatically trains to construct an ensemble of these models [11].

The CCD Causal Software suite, including Tetrad, Py Causal, and R-Causal[3], offers easy-to-use software for causal discovery from large biomedical datasets by applying Bayesian and constraint based algorithms. However, it still needs further study to determine how to integrate automatic causal discovery into AutoML systems.

Until now, most of the AutoML systems lacked two important services: One does not yet fully deal with the regression problem, and the other only focuses on small- to medium-sized datasets. In this study, we intend to investigate these two issues and exploit possible solutions. Moreover, we further explore the causal impact issue of big data analytics to see what are the possible challenges for the AutoML system.

IV. AutoML for Classification and Regression

AutoML refers to the full aspects of automation machine learning without human intervention. The stages of an AutoML system processing includes dataset preprocessing, learning algorithms selection for model construction, hyperparameter and parameter optimization, and final results interpretation, etc. The core of AutoML system is the optimal machine learning model construction.

This is a "bi-level optimization program", in which a lower objective is to train the parameters of a specific model, and an upper objective is to tweak the hyperparameters from numerous models [12]. Both are optimized simultaneously. Before enforcing this bi-level optimization program, we need to first select an appropriate learning algorithm fitted to the proposed dataset.

Why is an AutoML system so hard to build especially for big data analytics? First, we need to address automatic ETL data preprocessing while facing data heterogeneity, sparsity, missing values, and noise accumulation issues. In addition, we need to automatically transform various data representation that fits the processing for the automatic learning algorithm.

Second, numerous machine learning algorithms are available for selection on an optimal model construction. Which one is the best selected algorithm for a particular dataset? Can the AutoML system itself decides automatically without human help? In fact, meta-learning techniques have been proposed to automatically match datasets and learning algorithms [14]. A domain expert of the data scientist team can learn about the performance of machine learning algorithms across datasets from past experiences. Meta-learning mimics this strategy by using meta-features to describe which algorithms are effective on what datasets.

Third, once a specific machine learning algorithm is selected, next challenge will be a search of optimal hyperparameter and parameter combinations. If the objective loss function is a convex (or concave) and not a closed form, we proceed to search for an optimal iteration direction through a gradient descent (or ascent) optimization technique. Here, we must avoid overfitting and underfitting by balancing its bias-variance combination in a model construction.

Datasets are first first divided into two parts for a specific model's training and testing evaluation. Training datasets are further subdivided as training and validation datasets in offline model search phase. The k-fold Cross-Validation (CV) process averages the k-validation measures from k-validation outputs to decide an optimal combination of hyperparameter and parameter for a particular machine learning algorithm [15]. In general, we need to compare among various candidate learning algorithms proposed by meta-learning and decide which is the final optimal model for online testing.

A. ETL Data Preprocessing

When a data analyst performs data preprocessing in the ETL stage, (s)he faces several challenges: scaling, normalization, missing value imputation, variable coding for categorical variables, and continuous variable discretization, etc. Furthermore, if number of features are far more than number of data instances in a dataset, the feature preprocessors should automatically provide a capability to reduce the dimensionality of feature space through the Principal Component Analysis (PCA) technique. In addition, data preprocessors should also automatically provide rescaling, re-sampling, one-hot encoding, missing value handling, and skewed data balancing services before a machine learning modeling stage is performed.

B. Meta-learning for Model Selection

Meta-learning techniques learn about the performance of various machine learning algorithms across datasets to quickly suggest possible machine learning algorithms for model selection. However, they only provide coarse, not fine-grained, information on the performance of the machine learning framework. Therefore, we need hyperparameter tuning algorithms, such as the Bayesian optimization technique. In general, meta-learning for learning model selection complements the Bayesian optimization for hyperparameter tuning.

C. Optimal Hyperparameter and Parameter Tuning

By using a surrogate function to approximate the true objective blackbox function of a learning model, several Sequential Model-Based Global Optimization (SMBO) techniques have been proposed for a hyperparameter optimization. They

[3]Center for Causal Discovery (CCD): http://www.ccd.pitt.edu

include Bayesian optimization [16], Sequential Model-based Algorithm Configuration (SMAC), and Tree-Structured Parzen Estimator (TPE) [10] [11]. Among them, Bayesian optimization is a popular technique for tweaking hyperparameters. It first fits a probabilistic model to capture the relationship between hyperparameter settings and the model's performance. Then, it evaluates the hyperparameter settings and updates with the results through iteration.

D. Optimal Model Testing and Selection

In ML framework, data processors and feature preprocessors automatically enforce the ETL preprocessing services for a classifier (or regressor). The ultimate goal of current AutoML systems is to automatically achieve data analytic modeling for classification and regression. The important stages of the AutoML system aim to solve the CASH problem through preparing datasets and algorithms, executing meta-learning for a machine learning framework, and searching for an optimal classification (or regression) framework for future testing. Meta-learning initializes the Bayesian optimizer and automated ensemble construction from configurations evaluated during optimization.

V. AUTOML FOR CAUSAL IMPACT ANALYTICS

A. Three Approaches of Causal Impact Analytics

In contrast to machine learning analytics through ordinary statistical inference, causal impact analytics can be achieved by causal inference with three approaches: potential outcomes for counterfactuals, non-parametric structural equations, and graphical models in path diagram. Even though these approaches are created from different research fields, in fact they are related to each other [3] [17]. Machine learning analytics for estimating features correlation and prediction are based on observational data only. However, two types of situations are possible in causal inference models: an observational world, an experimental world. Causal inference problems are best worked out by experimental data.

However, randomized experiments can be unethical, infeasible, time consuming, or expensive. Besides, observational data are abundant and easy to obtain. Therefore, causal effects of counterfactuals can often be estimated only from observation data by imposing causal assumptions. Causal structures can be represented as graphs, where random variables are indicated as nodes and causal relationships are indicated as edges.

If the causal structure is known, estimating the counterfactuals and impacts of intervention of a causal graph are much easier than the unknown causal structure cases. The impacts of intervention factors can be computed by using covariate adjustments, inverse probability weighting, or instrumental variables techniques. Furthermore, we can learn unknown causal structures from observational data by using causal model search methods, such as constraint-based or scored-based algorithms [18].

B. Challenges of Causal Impact Analytics on AutoML

The goal of the automatic causal model search is to discover any possible true causal graph of a population from observation samples with the joint probability distribution over the population with background knowledge, such as parametric assumptions, time order, etc. We propose several challenges of applying AutoML for causal inference as follows based on recent causal discovery studies [19]:

1) ETL preprocessing of observational datasets used for machine learning models might not be useful for causal model searches and expected counterfactuals predictions. Moreover, automatic unknown causal model searching is different from automatic machine learning modeling regarding features selection to discover their correlation and the outcomes prediction.
2) Although three causal model representations and inferences are interoperable, how to smoothly integrate three causal model techniques into current AutoML systems is still unknown.
3) Although we have several automated causal model discovery algorithms, there are significant impediments regarding the improvement of these algorithms. The performance of these algorithms is difficult to evaluate if the manipulated and unmanipulated datasets are not available to compare to each other.
4) Although an unmanipulated causal graph can be manipulated to a fixed variable, we still must apply several assumptions when computing the causal impact probability density, including causal Markov assumption, causal sufficiency, and causal faithfulness.
5) The outputs of causal models search are Markov equivalence classes. When the vertexes number of possible causal graphs increases, such as high-dimensional datasets, the number of directed acyclic graphs (DAGs) increase exponentially. This makes it difficult to search for an optimal causal model to be compatible with the probability density of sample observational datasets, given the causal Markov and faithful assumptions.

VI. A SIMPLIFIED CAUSAL INFERENCE MODEL

In this study, we use a simplified type of non-parametric structural equation model, i.e., BSTS, to find the cause-effect of an intervention factor in an extended version of AutoML system with Spark ML pipeline (see Figure 1).

Before an intervention factor was introduced, we first trained a treatment group's time series pattern from a control group viewpoint, given a set of observational data from a fixed time period. Then, in the validation phase we applied a cross-validation technique on a treatment group time series pattern, given a set of observational data for a consecutive fixed time period. This is a statistical inference on the training and validation phases from machine learning viewpoints.

After an intervention was introduced, in an observational context, we predicted what would happen for a treatment group regarding its unobserved counterfactuals, given a set

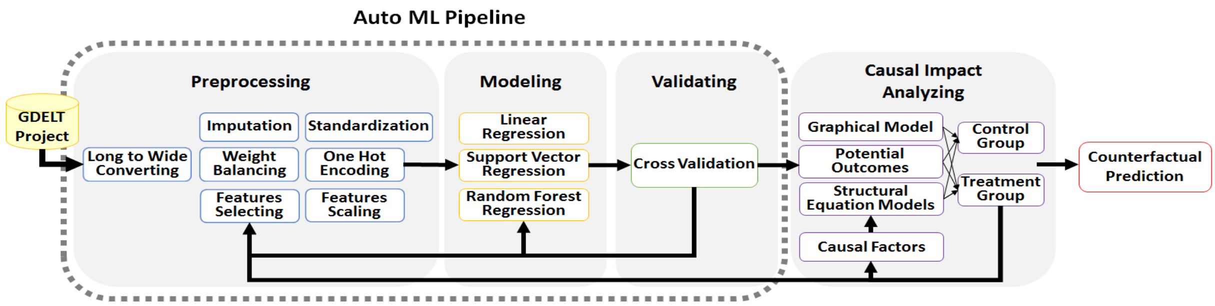

Fig. 1. A multiple stages of AutoML pipeline system for cause impact services, where an AutoML pipeline system offers causal impact analytics with three approaches, i.e., graphical model, potential outcomes, and structural equation models. In the pre-intervention phase, we first learn a treatment group's time series pattern from a control group. Then, in the post-intervention phase, unobserved predicted counterfactuals of a treatment group are computed based on the pre-intervention phase control's group learning model. A specific causal factor is given to derive average treatment effects by differencing between a treatment group's observed outcomes and its unobserved predicted counterfactuals.

of observational data in the post-intervention phase. This will change the corresponding structural equation with a intervention factor inside and leave the other structural equations unchanged. Finally, we compared a treatment group's real observed data with unobserved predicted counterfactuals from a control group's learning model and computed the average treatment effects for a specific intervention factor.

Current AutoML system stages should be revised and extended to cope with the causal inferences in observational and experimental data analytics. For example, we first train a control group on observational data to mimic a treatment group before an intervention factor was introduced. Once an intervention factor is empowered, we can compute unobserved predicted counterfactuals and compute the differences between a treatment group's counterfactuals and observed data.

Time series analysis by state space methods from econometrics perspectives have been well-established for intervention factors identification and predicted counterfactual analysis [20]. Graphical models from computer science perspectives have been proposed to answer three types of causal queries about [3] (1) the effects of potential interventions, (2) probabilities of counterfactuals and (3) direct and indirect effects.

Currently, we only leverage an AutoML system's analytics capabilities to solve time series regression problem without human involved in the pre-intervention phase. Use fixed time interval of Random Forest algorithm for training and validation of the big datasets. After that, we still manually apply a type of structural equation model, e.g., BSTS, to predict the expected counterfactuals and derive average treatment effects.

A. Bayesian Structural Time Series (BSTS)

A classic time series model with *constant* level, linear time trend, and regressor components. The 'local linear trend model' is a stochastic generalization of a classic time series model, where a level and trend can vary through time. Structural time series can be shown as a local linear trend model with a pair of equations: observation equation and state equation, where features (or covariates) provided with static or dynamic coefficients for state and regressor variables [6].

BSTS, a specific type of structural equation models, is used to infer the temporary evaluation on the impact of covariates, incorporating empirical priors on the parameters in a Bayesian approach, and coping with multiple sources of trend and seasonal variations [21]. BSTS uses a Markov Chain Monte Carol (MCMC) algorithm for posterior inference of regression simulated outcomes and hyperparameter θ when a well-known maximum likelihood distribution function to represent its structural equation model is unknown.

BSTS combined with the Google Trend and Correlate outputs are used for short-term economic prediction, i.e., Nowcasting [22]. Moreover, BSTS can be applied to learn the pre-intervention structural time series model and derive the unobserved predicted counterfactuals in the post-intervention period to yield a probability density over the temporal evolution of causal impacts.

B. A Real Policy Evaluation Scenario

GDELT system has been used to collect and monitor the world's news media in 15-minutes real time intervals of every day since 2014/02. GDELT is automatically encoded by the TABARI system as taxonomy of CAMEO event data types, and they are currently publicly available via Google's BigQuery. In this study, a real policy evaluation scenario is applied to the GDELT big dataset to evaluate the sentiment escalation (or de-escalation) tones between countries. This GDELT dataset after preprocessing can learn and track the S&P 500 stock market index through sentiment escalation (or de-escalation) tones between countries.

The original GDELT dataset only has 58 features, which are categorical types with 20,000-280,000 records available each day. These categorical data types are first transferred into numeric values. The ETL data preprocessing stage includes long-to-wide feature converting, weight balancing, and one-hot encoding. After the ETL data preprocessing stage, we have transformed the original high-tall type dataset into a low-fat one and created with a high-dimensional dataset in 35,582 features, including 35,548 sentiment escalation (or de-escalation) tones between countries, CAMEO's 20 event classification, Goldstein Scale, and AvgTone. The `feature_selection`

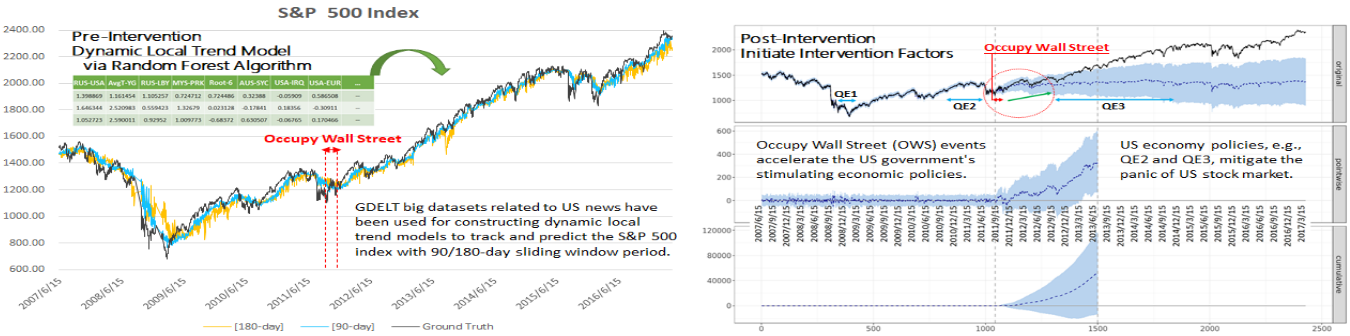

Fig. 2. The GDELT news media datasets have been used for tracking and predicting the S&P 500 stock indexes through Random Forest algorithm from 2007/06/13 to 2017/04/21 (see left-hand side). Among them, the news media continuously report on the Occupy Wall Street (OWS) movement along with the US government economic stimulating policies had explicit impacts on the stock indexes after 2011/11 (see right-hand side).

package from Scikit-learn provides the `f_regression` library to filter out the most important 300 features to track the S&P 500 stock indexes within the entire tracking period.

In the post-intervention phase, we first identify a possible intervention factor of a policy to evaluate its impacts to a stock market index. The causal impacts are based on the stock index differences between real observed outcomes and unobserved predicted counterfactuals derived from the dynamic Random Forest algorithm. A policy evaluation refers to the concepts of news media reports, such as Occupy Wall Street events along with the US government economic stimulating policies on reflecting the escalation (or de-escalation) tones, which might have influences on the stock market (see Figure 2).

VII. Conclusion and Future Work

AutoML system's standard processing pipeline for classification has been explicitly demonstrated. Current AutoML systems only provide correlation and prediction analytics services, Thus, they lack causal inference capabilities on intervention factor discoveries and counterfactuals prediction. We explicitly address the potential challenges of empowering causal impact analytics in the current AutoML systems.

A real policy evaluation scenario has been implemented by using the GDELT news media datasets along with the government economic policy to learn and predict the S&P 500 stock market index variation with a possible intervention factor. This has been demonstrated as the news media report on the Occupy Wall Street movement events along with the government economic stimulating policies for their impacts on the stock indexes increasing.

Future work will be exploiting on how to seamlessly integrate the well-known automatic causal structures discovery and inference techniques for high-dimensional datasets into the AutoML systems in the cloud computing platform.

References

[1] H. T. Davenport and D. J. Patil, "Data scientist: The sexiest job of the 21st century," *Harvard Business Review*, Oct. 2012.
[2] B. Efron and T. Hastie, *Computer Age: Statistical Inference - Algorithms, Evidence, and Data Science*. Cambridge Univeristy Press, 2017.
[3] J. Pearl, *Causality: Models, Reasoning, and Inference*, 2nd ed. Cambridge University Press, 2009.
[4] J. Fan, F. Han, and H. Liu, "Challenges of big data analysis," *National Science Review*, vol. 1, pp. 293–314, 2014.
[5] G. Shmueli, "To explain or to predict?" *Statistical Science*, vol. 25, no. 3, pp. 289–310, 2010.
[6] H. K. Brodersen *et al.*, "Inferring causal impact using bayesian structural time-series models," *The Annals of Applied Statistics*, vol. 9, pp. 247–274, 2015.
[7] C. Thornton *et al.*, "Auto-WEKA: Combined selection and hyperparameter optimization of classification algorithms," in *KDD'13*. ACM, 2013, pp. 847–855.
[8] F. Pedregosa *et al.*, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.
[9] L. Kotthoff *et al.*, "Auto-WEKA 2.0: Automatic model selection and hyperparameter optimization in WEKA," *Journal of Machine Learning Research*, vol. 17, pp. 1–5, 2016.
[10] B. Komer, J. Bergstra, and C. Eliasmith, "Hyperopt-Sklearn: Automatic hyperparameter configuration for Scikit-learn," in *Proc. of the 13th Python in Science Conf. (SCIPY 2014)*, 2014.
[11] M. Feurer *et al.*, "Efficient and robust automated machine learning," in *Proceedings of the 28th International Conference on Neural Information Processing Systems*, ser. NIPS'15. MIT Press, 2015, pp. 2755–2763.
[12] I. Guyon *et al.*, "Design of the 2015 ChaLearn AutoML challenge," in *2015 International Joint Conference on Neural Networks (IJCNN)*. IEEE, July 2015.
[13] ——, "A brief review of the ChaLearn AutoML challenge: Anytime any-dataset learning without human intervention," in *ICML 2016 AutoML Workshop*, 2016.
[14] P. Brazdil *et al.*, *Metalearning: Applications to Data Mining*. Springer, 2009.
[15] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd Edition*. Springer, 2013.
[16] A. Klein *et al.*, "Fast bayesian optimization of machine learning hyperparameters on large datasets," in *Proc. of the 20th Int. Conf. on Artificial Intelligence and Statistics (AISTATS)*, 2017.
[17] T. Richardson. Drawing causal inference from big data, sackler colloquia. National Academy of Science. [Online]. Available: http://www.pnas.org/content/113/27/7308.full
[18] H. M. Maathuis and P. Nandy, "A review of some recent advances in causal inference," in *Handbook of Big Data*, P. Bűhlmann *et al.*, Eds. Chapman and Hall/CRC Press, 2016.
[19] P. Spirtes and K. Zhang, "Causal discovery and inference: Concepts and recent methodological advances," *Applied Informatics*, vol. 3, no. 3, 2016.
[20] C. A. Harvey, *Forcasting, Structural Time Series Models and The Kalman Filter*. Cambridge Univeristy Press, 1990.
[21] R. H. Varian, "Big data: New tricks for econometrics," *Journal of Economic Perspectives*, vol. 28, no. 2, pp. 3–28, 2014.
[22] L. S. Scott, "Predicting the present with Bayesiam structural time series," *Int. J. Mathematical Modeling and Numerical Optimization*, vol. 5, no. 1-2, pp. 4–23, 2014.