IDC ANALYZE THE FUTURE

IDC SOLUTION SPOTLIGHT

Sponsored by: DataRobot

Today's businesses leverage time series data to extract greater insights about future performance measures. Automated machine learning tools offer important advantages over traditional analytics tools such as their simplicity, agility and the interpretability of their results.

Accelerating Time Series Analysis with Automated Machine Learning

January 2019

Written by: Dr. Chris Marshall, Associate Vice President, Big Data and Analytics, Cognitive Computing, IDC Asia/Pacific; and Jessie Cai, Senior Research Manager, Big Data and Analytics, Cognitive Computing, IDC Asia/Pacific

Introduction

Over the past few years, the number of data pipelines feeding into analytical data stores have substantially increased to support an ever broader set of business needs. This shift significantly changes the type of analytics required, from descriptive aggregated analytics about past performance to a greater focus on current and future performance using more fine-grained metrics. The shift also pressures organizations to invest in technologies to manage and analyze time series data. Such data reflect many aspects of business processes, customer behaviors and asset performance, with analytics used to rapidly identify deviations from the norm that negatively impact performance or represent new opportunities. As traditional statistical tools struggle to leverage multivariate inputs, uncover actionable insights and make an impact on actual business operations, this IDC Solution Spotlight examines how automated machine learning tools can augment the analysis, modeling and prediction of time series data to deliver easily understood and actionable insights for businesses in a simple and agile fashion.

The Nature of Time Series and Its Applications

Every aspect of the world, and the natural and human systems within it, changes with time. Time series can be formally defined as a series of data points tagged by regularly spaced time stamps. Time series analysis refers to the use of statistical or machine learning methods to analyze time series data – one series or many, by extracting meaningful patterns in output variables (such as trends, seasonality and special events) and their co-relation with input variables enabling predictions about how changes in input variables affect output variables.

Applications of time series analysis abound, and many of them forecast future demand for better operational planning. This allows for the prediction of the future state of Chicago housing starts next year, the demand for a Bangkok hotel room next week, the footfall on a London street in the next hour, the workload on a local router over the next five minutes, and even the clicks and conversions of a shopping portal in the next 30 seconds. Apart from making forecasts, time series analysis can also provide insights for complex systems, for example, to identify contributing factors for equipment downtime in a manufacturing site or to detect anomalous signals from IT system logs. Adding additional contextual metadata or related time series can augment the original analysis, enabling what-if questions like the impact of an upcoming storm on a wind farm's electricity generation or the effect of advertising on sales.

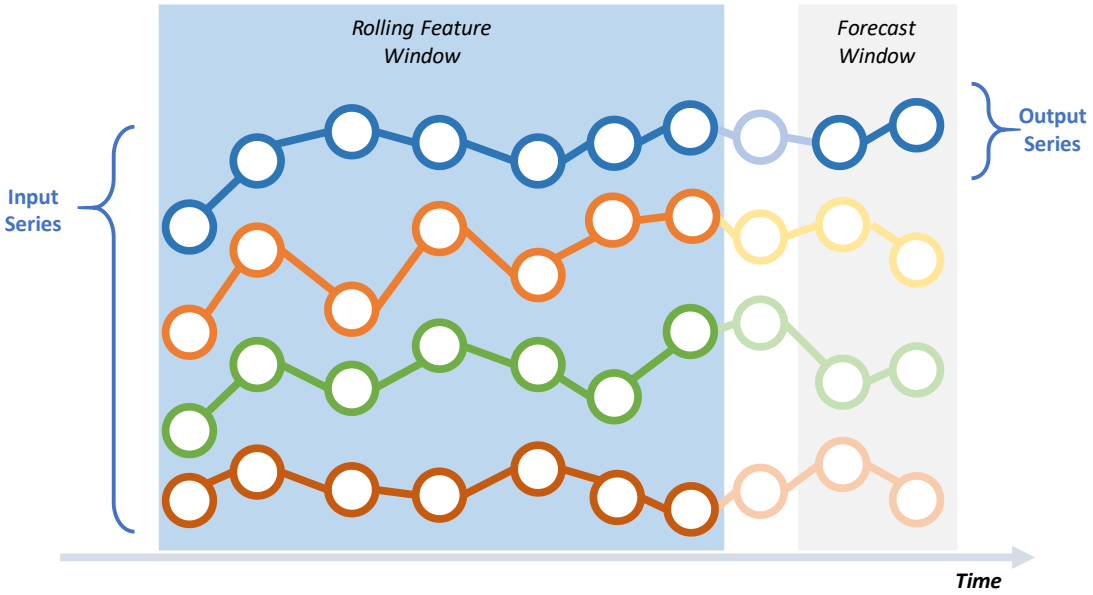
The assumption here is that there is an internal structure embedded in the data which can be at least partially accounted for by the change in time or extraneous variables that are themselves changing. Variables may be independent such as season, weather, public holiday, planned events, work schedules or more complex interdependent variables such as macroeconomics, energy supply and stock market position.

Time Series Problems Are Hard

Time series problems are hard because they attempt to uncover a potentially evolving inner structure within historical data and extrapolate into the future. The following core characteristics of time series problems are depicted in Figure 1:

- **Regularly spaced time intervals** are an important feature of any time series. When arranged differently, by day or by week or by month, for example, the same algorithm could yield very different models and predictive powers.
- One or more time series are used as **inputs**, captured over a **rolling feature window** which defines the time range of feature discovery.
- The **forecast window** of time series is usually an extrapolation of the targeted **output series**, in which the model developed from the input series is used to predict the future state of the target – for instance, first quarter sales in the coming year or clicks in the next 30 minutes following a web promotion.

FIGURE 1: ***Key Characteristics of Time Series Problems***



Source: IDC, 2019

The time series modeling process seeks to discover how the change in the input series results in the change in the output series. The process is complex and iterative. It starts by determining which input series to use, and may then involve preparing the data by splitting, cleaning and segmenting the data. This is followed by feature extraction, model building and backtesting until an acceptable result based on some predefined accuracy criteria is achieved. Next comes model interpretation and evaluation, and finally, deployment into applications – the frequently underemphasized step

in which the model is stabilized, scaled and embedded in local business operations to produce useful outputs and interventions. Periodically, the entire time series modeling process needs to be repeated as new data arrives.

Traditional Approaches to Time Series Analysis

Time series analysis is not new. Classical statistical techniques for econometric analysis of trends, cycles and randomness have been staples for decades. Since the dawn of computing, techniques such as ARIMA (autoregressive, integrated, moving average) and their many variants (VARIMA for vectors, GARCH for time varying volatility, etc.) have been widely used to solve econometric, business and operational problems.

These are typically parametric, often univariate, models that make strong assumptions about random variables' distributions and the model's stationarity over time – these models are highly structured, easily explained, require limited data and generally produce adequate approximations to sampled data sets. At the same time, these classical methods are known for their limitations as follows:

- **Dependence on statistical assumptions.** The validity of assumptions including linearity, normality and stationarity is critical for classical statistical analyses to apply in practical setups. They cannot be relaxed for the model derived to make sense, thus requiring more iterations for cleaning, resampling and model checking.
- **Poor adaptability to multivariate analysis.** Multivariate problems involve multiple input time series, and in solving them, classical statistical methods often suffer reduced prediction power and accuracy. This greatly limits the applicability of time series analysis in complex interdependent systems in the real world.
- **Poor predictive power for special or extreme events** for which, by definition, limited historical data is available.

Higher Demand for Time Series Analysis

Just in time inventories and the rise of e-commerce have forced businesses to up their game in analyzing and supplying customer demand. Furthermore, digital initiatives such as the Internet of Things (IoT) and digital transformation projects have dramatically increased the availability and richness of time series data about demand and operations. This trend is set to continue and has forced companies to enhance their time series specific capabilities across a variety of industries. Below are some examples:

- **Retail.** Since 2013, Walmart has started sharing on-shelf availability data with their suppliers for better replenishment execution. Suppliers have to use the data to predict the next shelving time window and deliver the replenishment goods in full quantities in the required time window for at least 85% of the time, or a fine of 3% of the cost of goods will be imposed.
- **Transportation.** Uber relies on time series forecasts to predict supply and demand in fine-grained spatial-temporal analysis to direct drivers to high demand areas before they arise. This is essential to their business model – connecting drivers and passengers in a timely fashion through their platform service.
- **Energy.** One of Europe's largest integrated electric power companies federated 80+ equipment sensor readings and maintenance logs to predict sealing fluid loss and vibration-induced failures, as well as other anomalies. Their application of time series analytics provided users and operators with up to three weeks' advance notice of impending failures with high confidence and no false alarms.
- **Manufacturing.** Haier adopted an IT operations orchestration tool that analyzes IT system logs as time series. This reduces the time taken to identify issues and investigate root causes from days to hours to even minutes in some cases.

The Promises and Challenges of Machine Learning

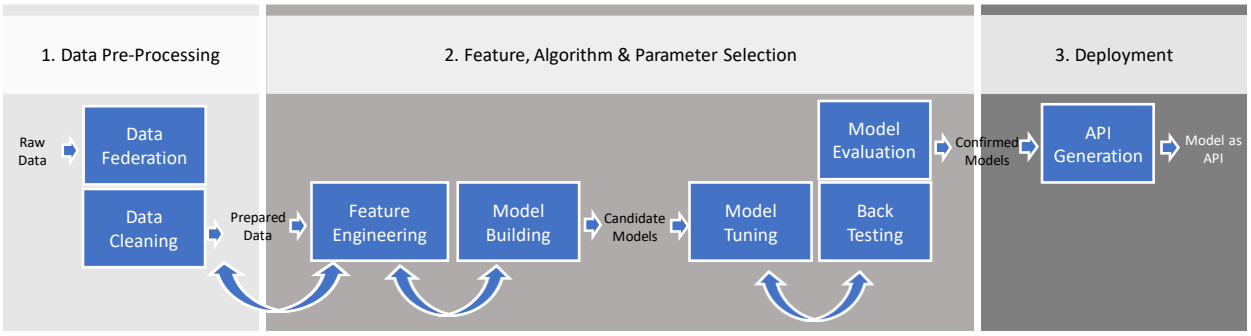
The more complex real-world time series problems outlined above have huge data sets with many potential features, captured over long time periods. Most, if not all, of these problems are multivariate in nature, for which classical techniques can be adapted but only with dramatically reduced accuracy and predictive power. Into this breach, steps machine learning techniques such as ridge regressors, boosted trees and neural networks.

Not only do machine learning approaches sidestep the typical linearity, distributional and stationarity assumptions of classical statistical modeling, they can also produce usable models, with higher adaptability to multivariate analysis and less risk of overfitting. Typically, machine learning approaches are well suited for a huge number of feature-rich data records – increasingly the case in our instrumented 3rd Platform world. Consider one such application – IT operations, where IDC predicts that by 2022, three quarters of operations will be supplanted by machine learning based analytics and automation, resulting in over 25% OPEX savings.

However, machine learning also presents a few challenges:

- **Lack of skills.** Perhaps the biggest single obstacle to implementing machine learning algorithms to time series is a human one. Data scientists and machine learning specialists remain a scarce resource in many organizations. In APEJ for example, only 23.7% of organizations have data scientists, and among these data scientists, only 20.5% of them have an extensive educational background in computer science and machine learning.
- **Complex process.** Organizations face difficulty in the multiple steps of the complex and iterative machine learning model building process which involves disparate stages of data preparation, feature engineering, model building, model evaluation and deployment. Figure 2 describes the typical flow, keeping in mind that iterations are required not only between steps but also within steps.

FIGURE 2: ***Machine Learning Model Development Process***



Source: IDC, 2019

- **Many disjointed tools.** Different software tools cover each stage of the complex machine learning model development process. Unlike traditional software development, where developers look to one tool to cover multiple stages, machine learning developers often use different tools and algorithms to determine their effect on accuracy and performance. Inevitably this leads to significant customization and tuning that can be difficult to systematize and replicate.

- **Managing experiments.** Machine learning algorithms have many configurable parameters, and it is a lot of work to track which set of parameters has been tuned in each experiment in order to converge to an optimal model. Many organizations struggle with the agile development methodologies needed to make this effective.
- **Deploying machine learning models.** As well as the model's scalability to handle large volumes, another major consideration when deploying machine learning models in production is interoperability across different internal platforms, and the wide range of deployment tools and environments it needs to run in (e.g., REST serving, batch inference or mobile apps).

These challenges are more specific to time series data:

- **Data preparation.** Data preparation for time series can be tricky and should incorporate not only common issues such as missing values, standardization, wrong entries, outliers, etc., but also the right sizing of time intervals. Different time intervals could include or exclude features that are critical for model building.
- **Testing and evaluation traps.** Evaluation of time series models is also difficult. A time series model of high accuracy, according to one or a few metrics, might not have sufficient predictive power. In general, evaluation criteria have to be carefully selected for the specific use case.
- **Model interpretability.** Model interpretability is required more for time series data than other data types because time series analyses are often tied with critical business decisions. Users can rarely make such decisions without a clear understanding of the model's underlying assumptions. This poses greater challenge for machine learning methods, which are often associated with a lack of transparency.

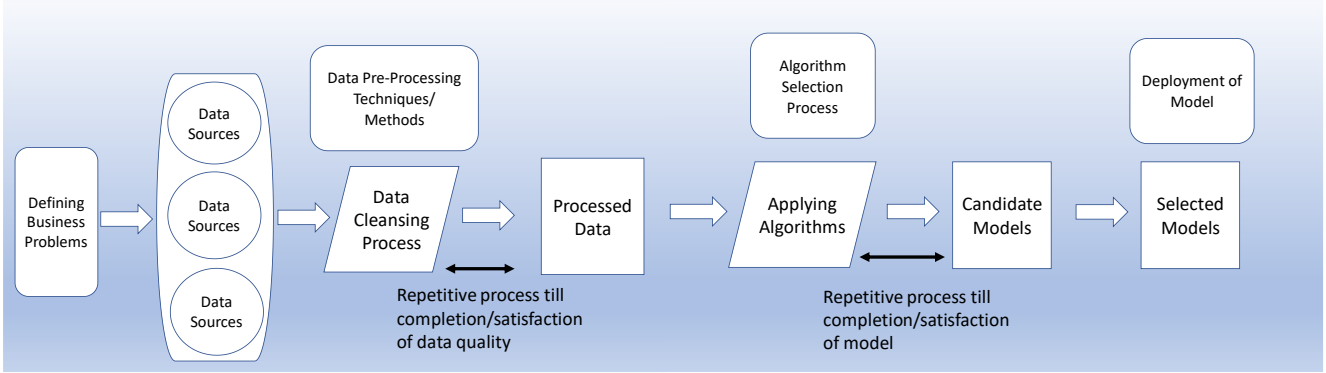
The Rise of Automated Machine Learning

According to a recent IDC survey, more than 46% of companies in Western Europe stated that they face difficulty in multiple steps of the machine learning model building process, including data preparation, feature engineering, model building and model evaluation. No surprise therefore, that vendors have started to alleviate the complexity of the machine learning model development process by streamlining or automating the model building process.

Most vendors characterize "automated machine learning" functionality as providing separate tools and preconfigured models and libraries to support the different tasks within the end-to-end machine learning model development process. A few vendors are adopting a new approach to automated machine learning, automatically searching through the space of potential models and, for a given data set, choosing the best model based on some predefined criteria (see Figure 3).

These approaches are not mutually exclusive; automated machine learning models can be switched to manual mode to allow stepped execution. Both approaches promise significant increases in the productivity of data scientists as they develop complex machine learning models. However, for most of the vendors, automated machine learning functionality, especially for time series data, remains limited.

FIGURE 3: ***Machine Learning Model Development Process***



Source: IDC, 2019

DataRobot's Automated Machine Learning Platform for Time Series

DataRobot, a leading machine learning platform vendor based in Boston, U.S.A, has recently incorporated time series handling capability to their automated machine learning platform. The company is among the very first to release an automated machine learning product, in as early as 2012. The product provides turn-key end-to-end automation to search through the modeling space for customers to produce "optimal" target models.

How the DataRobot Platform Works

The platform is designed to allow users with limited data science or software development background to build and deploy machine learning models. The interface is easy to grasp. Zero coding is required, and users just need to bring along the data and specify the prediction target. Much of the rest of the modeling process including data partitioning, feature engineering, model training, testing and evaluation, etc., can be done automatically on clicking the "Start" button. A list of models is generated, scored and ranked in a leaderboard, each with a summary of model performance for users to drill down. Various charts are available to help users to visualize impacts of extracted features and connect them with domain knowledge. Users can choose to further finetune the shortlisted models. Once a model is chosen, the tool can generate an API for users to deploy to a target system.

The design of the DataRobot platform seeks to reduce the complexity of the machine learning model building process, delivering high quality models, and relieving data analytics professionals from coding and model tuning tasks that are repetitive in nature. At the same time, it is possible to switch to manual mode for trained data scientists to make adjustments. The platform is based on various open sourced machine learning libraries and frameworks including H2O, Scikit-Learn, R, Tensorflow, Spark and XGBoost, which are themselves further curated by a team of highly ranked data scientists. Within DataRobot is a workflow visualization component showing which algorithm is being applied at which step and in which iteration, helping ensure good model governance.

Specially for Time Series Data

The platform accommodates both classical and machine learning methods for large time series data sets. Depending on the actual data set, the two schools of methods can be applied in a hybrid manner. The platform does data exploration and time interval selection automatically and differently based on different algorithms used in different iterations. Systematic backtesting is designed in to avoid common pitfalls in model evaluation. As of November 2018, DataRobot Time Series is the only commercially released product that can automate the machine learning modeling process for time series data. Customers are already seeing improved predictive power over time. Moreover, the

DataRobot platform is equipped with features such as autoscaling to manage compute while tackling difficult time series problems.

Automated Machine Learning Case Study

Steward Health Care is a very large for-profit private hospital network in the United States. Like other hospitals, operational inefficiencies can pose challenges, which are further exacerbated by large variation in demand. The availability of a large amount of historical time series data, though, gave hospital executives confidence that there are opportunities to stay proactive of upcoming volume. These go beyond simple projections of patient volumes to incorporate in-depth analyses of contributing factors, and better predictive models. Faced by a lack of skills internally, the team used DataRobot Time Series to identify, build and test new machine learning models. A total of 1,536 models were built and incorporated into the hospital's own proprietary labor management dashboard that supports all 38 hospitals in the network, after verifying the screened factors such as weekdays, school vacations, sports events, and even moon phases. According to the embedded accuracy tracker of the dashboard, some models are predicting day-specific volumes with 95% accuracy. As Erin Sullivan, Executive Director of Information Systems and Software Development at Steward Health Care, put it, "A lot of people use the buzzwords like predictive analytics or machine learning, but we're actually doing it. We have a product that's out on the hospital floors now, and we see it work." Indeed, the hospital is well on the way toward achieving their cost-cutting goals:

- US$2 million in annual savings from the reduction in registered nurse-hours for eight of the 38 hospitals in Steward's network.
- Another US$10 million in savings per year from the reduced length of patient stay.

Opportunities Ahead

In time, more organizations will come to realize the value of automated machine learning, which offers greater accessibility, productivity and explainability for AI and advanced analytics technologies. There are several directions the DataRobot platform can take going forward to further help organizations harness data science capabilities:

- **Collaboration features** involving multiple team members in the modeling process. The DataRobot platform should be accessible to various business roles, with collaborative tools to accelerate the forming of a cohesive internal ecosystem centered around data. Through this, data literacy can be promoted and model interpretability can be reinforced by domain expertise.
- **Pre-built components for reuse** including components of previously deployed models that can be potentially reused by others facing similar problems. Documentation on the use cases, algorithms, data types, etc., should be in place to facilitate searches, assemblies and deployments for new users.
- **Community to promote machine learning innovation**. Having a user community and encouraging the sharing of best practices, problem solving tips and business results is very important. Often, innovation is less of a technology issue, but more a problem of mindset. Having an active user community can make changes happen faster and at a larger scale.
- **Partner for computation acceleration** to offer more options for computation resource optimization and acceleration. This becomes more important as bigger data require deeper learning, and customers will appreciate the help to run the model selection process in the most cost-efficient manner.

Essential Guidance

With a substantially larger number of data pipelines feeding analytics to support a broader set of business goals, businesses need to have a greater focus on future performance on different levels of granularity – for example, at a product, store or equipment level. The shift has placed greater pressure on organizations to improve their capabilities to manage time series data that reflect various aspects of business processes, customer behaviors and asset performance to rapidly identify deviations from the norm that negatively impact performance or represent new opportunities.

Analysts across many industries, working on time series applications, are faced with more time varying attributes, more data types, more historical data and more frequently updated data. However, traditional analytics techniques fail to perform at such a scale and are hamstrung by their restrictive assumptions. Powerful machine learning tools can capture richer features for the linkages between changing time series variables – a richness that can significantly improve the ability to forecast future time series values.

In the past, machine learning approaches have been hamstrung by the lack of trained data analytics professionals able to develop and optimize models. The advent of fully automated machine learning tools such as DataRobot Time Series promises to greatly improve the accessibility of machine learning technology for real world problems, drastically sidestep limitations, and achieve more accurate and robust future predictions. For businesses, this translates to deeper understanding and better control of a myriad of operations using time series data. For data scientists, such productivity tools will support effective talent retention and continuous skills democratization.

For this to happen, data analytics professionals and business users must address time series modeling holistically, from managing the data pipelines to accelerating model deployment. This implies that analysts must:

- Collect business problems that develop and evolve over time. Time series data is often undervalued and underused within an organization. Be it to predict a future value based on historical data, or to understand a system's behavior in response to changing conditions, many business problems fit well in the time series category.
- Establish a data management strategy to effectively identify, acquire, access and federate high-quality time series data assets from both production and development environments. Furthermore, data governance policies for these data pipelines need to be in place and aligned with well-defined use cases to fuel continuous model generation and update.
- Have automated and containerized API deployment capability to ensure the mobility and scalability of machine learning models in heterogenous systems. Data driven intelligence is most effective when it is embedded into operation workflows to augment or automate decision making in real time, wherever they sit and whenever they are required.

MESSAGE FROM THE SPONSOR

DataRobot helps enterprises embrace artificial intelligence (AI). Its automated machine learning platform harnesses hundreds of cutting-edge open source algorithms to discover the best machine learning models for every situation, empowering users of all skill levels to consistently make smarter, faster business decisions. The DataRobot platform automates, trains, and evaluates models in parallel, delivering AI applications at scale. DataRobot captures the knowledge, experience, and best practices of the world's leading data scientists, delivering high levels of automation and ease-of-use for machine learning initiatives. For more information, visit datarobot.com.

About the analysts:

Dr. Chris Marshall, *Associate Vice President, Big Data and Analytics, Cognitive Computing, IDC Asia/Pacific*

Dr. Chris Marshall is Associate Vice President for IDC Asia/Pacific, responsible for the Analytics, Big Data and Artificial Intelligence practice. Dr. Marshall's core research coverage includes the development of Data Analytics and Machine Learning competencies and their implications – the threats and opportunities facing organizations as they seek to augment and automate their knowledge-based work.

Jessie Cai, *Senior Research Manager, Big Data and Analytics, Cognitive Computing, IDC Asia/Pacific*

Jessie Cai is an analyst in the Big Data & Analytics Practice for IDC Asia/Pacific, serving as Senior Research Manager for Cognitive Computing/Artificial Intelligence. Cognitive systems and an army of underpinning technologies are the natural next development stage of Big Data. The materialization of its business values will further drive digital transformation across all industries.

IDC Custom Solutions

IDC Corporate USA

5 Speen Street
Framingham, MA 01701, USA

T 508.872.8200

F 508.935.4015

Twitter @IDC

idc-insights-community.com

www.idc.com