Data Science Accelerator: Empowering citizen data scientists

Organizations are rapidly adopting advanced analytics to enable data-driven business decisions. As a result, the demand for data science experts is growing. However, there is a huge gap between the demand and supply of data science talent. McKinsey confirms that 12% annual growth in demand for data scientists outpaces the supply of 7% data science graduates per year, which would lead to a shortfall of some 250,000 data scientists by 2024[1].

Hence, there is a need to fill the current skills gap by enabling citizen data scientists in the organization. Citizen data scientists, though not experts in the field of data science, can leverage analytics and statistics while delving into the data to derive additional insights. Citizen data scientists complement expert data scientists by bringing in their own expertise and unique skills to the process.

Technology is a key enabler of these non-specialists (citizen data scientists).

Data Science Accelerator empowers the citizen data scientists to extend their reach to data and analytics via approaches that automate data science stages for insight discovery.

This paper explores an approach to accelerate data science by providing automated predictions on dataset via automation workbench and reusable modules.

Data science acceleration platform

Data Science Accelerator (DSA) powers automated predictions on dataset via automation workbench to empower citizen data scientists and reuse modules that make data scientists more productive. It leverages open-source technologies in a cloud environment to create a user friendly workflow.

Figure 1 highlights the key features essential to a data science acceleration platform.

Data transformation	Data exploration	Model development	Model validation
Read data from different sources	Histogram	Hypothesis testing	Model validation scores in test data
Data summary	Pie chart	Generalized linear regression	Model summary
Data sampler	Box plot	Logistic regression	Concordance & discordance
Selection of rows/columns	Distribution	Time series model - ARMA,ARIMA, classical, exponential	AIC/BIC - Measures of relative quality
Missing value imputation	Scatter map/ Scatter plot	K-Means (Clustering)	R-square & Adjusted R-square - Measures of accuracy/fitness
Outliers detection/ imputation	Heat map	Random forest classification	P Value - Measure of statistical significance
Edit domain/rename values & features	Percentile distribution	Random forest regression	F/T/Chi square Statistics - Model statistic
Feature constructor	Measures of central tendency	Scoring of model	Predictions (on new data)
			Confusion matrix - Classification
			ROC analysis - Measure for diagnostic ability for binary classifier
			Lift Curve - Measure of performance of model

Figure 1: Key features of data science acceleration platform

Accelerate your data science journey with automated predictions

Figure 2 showcases the workflow of DSA across different data science phases that transform data to information, and to insights.

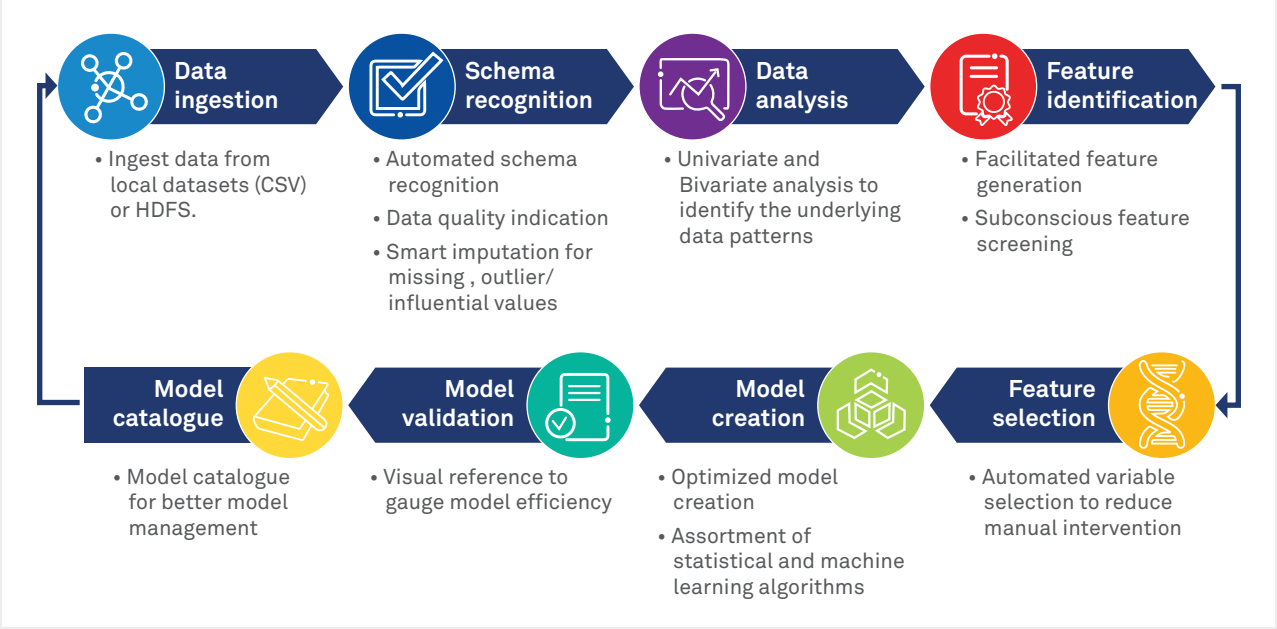

Figure 2: Workflow of data science accelerator

Features and functionalities

The key features and functionalities of DSA across different data science phases are:

1. Data analysis: Provides a quick summary of the dataset and highlights the underlying data patterns and relationships.

a. Univariate analysis: DSA has additional features viz. Automatic schema recognition for smart detection of data type and ID variable, which enables creation of right univariate distribution based on the variate type detected.

b. Bivariate analysis: Distinctive techniques such as Weight of Evidence (WoE) at decile level, chi-squared test, cross tabulation and tukey test are used to explain association or disassociation between a pair of target and explanatory variable at pre-defined significance level.

2. Feature engineering: DSA provides a user-friendly environment to:

a. Drop column(s) via single click button

b. Impute missing values

c. Cap outliers based on position and influential value in the range scale

d. Provide data transformation options such as split variables, standardize or normalize data, flag creation

e. Auto-discard variables basis Strength of Association (SOA) and Variance Inflation Factor (VIF)

Thus, it reduces process time for transmutation, screening and ensembling data features, leading to significant effort and cycle time reduction in model development stage.

3. Model creation: Allows user to build a model as per requirements using the provided options:

a. Test-control sampling - User creates development and validation data from the modelling data using user-defined algorithm by choosing sampling method and train ratio

b. Expression split - Helps create hold-out samples based on chosen partition option

4. Model selection: Recommends the most suitable model for solving the problem viz. classification or regression for the chosen target variable. This feature provides guided assistance throughout the journey of model development. Further, model performance is validated via confusion matrix, Mean Squared Error (MSE), concordance value, etc.

5. User control options: Provides the following additional options to scrutinize the model generated:

a. Download scored data

b. Download transformed train and test data

c. Score model with new data

d. Download validation charts

e. Rebuild and save model

6. Model catalogue: Stores and enlists all the models created by the user. This provides a quick glimpse of the various models created using diverse set of variables for the same dataset, thereby enabling user to compare performance of different models via accuracy, precision, R squared value, F1 score etc.

Thus, DSA not only helps improve productivity of data scientists by reducing process time through automation of repetitive tasks, but also empowers citizen data scientists via code-free analytics. It provides guided assistance through a series of statistically robust steps with minimal supervision from specialists.

The future roadmap

Unlike other automated data science tools, which are essentially “black boxes”, wherein most of the intermediate techniques are hidden and beyond the control of the analyst, DSA provides transparency and control to experienced data scientists to understand under-the-hood rationale at every analytical process and override recommendations, if any. Further, based on the data and the semantics, it auto-suggests the most suitable model for solving the problem viz. classification or regression for the chosen target variable. This feature is pivotal for citizen data scientists, as there is a guided assistance provided by the platform. In addition, it has an in-built ability to scale and process large data volumes through distributed execution on Spark clusters which enables easy implementation.

DSA capabilities can be further enhanced by:

- Extending the algorithmic footprint in supervised and unsupervised learning through addition of advanced modelling techniques such as XGBoost, CART etc. under model selection
- Providing seamless integration with deployment environment
- Implementing model management framework that helps monitor model performance in operational environment and identifies suitable timing for updating the models

Reference

[1]https://mck.co/2KKhr3s

About the authors

Bavya Venkateswaran
Consultant, Wipro Limited.

Bavya is currently responsible for solution design and implementation of advanced analytics and artificial intelligence-based use cases on Data Discovery Platform (DDP), Wipro's proprietary Insights-as-a-Service offering. She has been instrumental in delivering turnkey solutions in Energy & Utilities, Human Resources, and Healthcare domains.

Harish Kumar Chauhan
Project Lead, Wipro Limited.

Harish leads Engineering division of Platform and Solutions in the Analytics group at Wipro. He is responsible for creating, developing and implementing various analytics solutions. He has experience in designing big data architecture in cloud environment and ideating for automation in the industry.

Dipojjwal Ghosh
Principal Consultant, Wipro Limited.

Dipojjwal is currently involved in the development of analytical apps for Consumer and Utility domains on DDP, Wipro's proprietary Insights-as-a-Service offering. He has around 10 years of research and analytical experience in Manufacturing, Energy, Natural Resources, and Retail domain.

Wipro Limited
Doddakannelli, Sarjapur Road,
Bangalore-560 035,
India

Tel: +91 (80) 2844 0011
Fax: +91 (80) 2844 0256
wipro.com

Wipro Limited (NYSE: WIT, BSE: 507685, NSE: WIPRO) is a leading global information technology, consulting and business process services company. We harness the power of cognitive computing, hyper-automation, robotics, cloud, analytics and emerging technologies to help our clients adapt to the digital world and make them successful. A company recognized globally for its comprehensive portfolio of services, strong commitment to sustainability and good corporate citizenship, we have over 175,000 dedicated employees serving clients across six continents. Together, we discover ideas and connect the dots to build a better and a bold new future.

For more information, please write to us at **info@wipro.com**

IND/B&T/FEB-JUL2019