Data Science Project Life Cycle

Suresh Bommireddy
PMP, India

Abstract— the setting up a successful Data Science capability is not so easy. It contains its own challenges. To overcome some of those challenges during the execution and implementation, we need to focus on key outcomes for each phase of data science project life cycle, which I have highlighted below.

Key words: Data Science, Data Science Project Life Cycle

I. INTRODUCTION

Data Science is becoming more important day by day as we are drowning in data, but we are still starving for knowledge! Continued business value generation from data is critical to success, however embedding the insights generated into decision making and instigating a cultural shift to a data driven outcomes, would be the ultimate game-changer. From insight to action and to value creation can be achieved by following key outcomes of Data Science project life cycle.

II. PHASES OF DATA SCIENCE PROJECT LIFE CYCLE

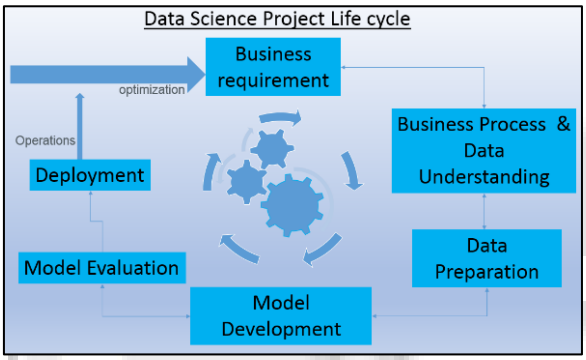

Fig. 1: DSPLC

A. *Business Requirement Phase*

1) Key Outcomes

- Motivation factor(s) to start the project with focus on end objective
- Overview about the business
- Document high-level expectation from key stakeholder

2) Examples

- Needs to get clear expectation from key stakeholders
- Process improvement to increases the productivity
- Eliminating wastage in the process (yield or reduction in time)
- Derive hidden insight which can be used for forecasting/ prediction

3) Reviewer(s)

- Business experts and key stakeholders

B. *Business process and data understanding*

1) Key Outcomes

- Step by step details of process
- Understanding to process and variables associated to that process
- Operational data mapping with variables and sources of bias, missing data, data quality issues.

2) Examples

- Document end to end business process (transactional /Operational) steps (Batch Processing / Semi-batch Processing/ Continuous Processing)
- Document details of data variables description, type of data and excepted range and values (time series data/ batch data/ transactional data etc.)

3) Reviewer(s)

- Domain Experts and Key stakeholders

C. *Data Preparation*

1) Key Outcomes

a) Exploratory Data Analysis

- Gain maximum insight into the data set and its underlying structure
- Uncover a parsimonious model, one which explains the data with a minimum number of predictor variables
- Check assumptions associated with any model fitting or hypothesis test
- Create a list of outliers or other anomalies
- Find parameter estimates and their associated confidence intervals or margins of error
- Identify the most influential variables
- Data cleaning including clear strategy defined for missing data
- Check and clean-up for missing data and other mistakes

2) Example

- Missing value treatment
- Outlier detection and treatment
- Influential points detection and treatment
- Bias correction
- Data Transformation (Normalization, Standardization, Box-Cox, Log)

3) Reviewer(s)

- Statistics data mining experts, domain experts and data scientists

D. *Model Development*

1) Key Outcomes

- Identify best support model based on the business case (Predictive and/or descriptive)
- Data selection method to be adapted based on bias, precision and accuracy to be select for training data vs. testing data (N-Fold, K-fold cross validation, Random forests, K-NN, Bootstrap etc.) to train the model
- Outcome results should be documented in details to justify the best model

2) Reviewer(s)

- Statistics data mining experts and data scientists

E. *Model Evaluation*

1) Key Outcomes

- Identify the best support model based on data
- Classification Models - (accuracy, precision, recall, F-score, AUC (area under the ROC curve), average log loss, training log loss etc.)

- Regression models - (mean absolute error (MAE), root mean squared error (RMSE), relative absolute error (RAE), relative squared error (RSE), mean zero one error (MZOE), coefficient of determination(R2) etc.)
- Clustering Models - (Sweep clustering, maximal distance to cluster centre, average distance to cluster centre etc.)
- Project document evaluation by 3-fold reviewer's method

2) Reviewer(s)

- Statistics data mining experts, Domain experts and data scientists

F. Deployment

1) Key Outcomes

- Models are deployed to a production or production-like environment for final user acceptance
- Should be operationally acceptable
- Need to be scalable for optimization
- Maintain version control as well as need to tag to respective domain

2) Reviewer(s)

- Domain experts, Data Scientists, Operational leaders and business Key stakeholder

III. DATA SCIENCE PROJECT TEAM SKILL SET REQUIREMENT

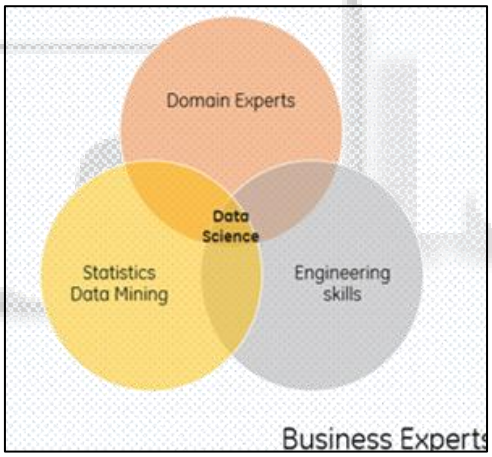

Fig. 2: Data Science Project Team Skill Set Requirement

- Data Science which is a fusion of different technologies like Statistics, Machine Learning and "modern "Database approaches.
- Domain expert(s) experts with special knowledge or skills an area
- Statistics data mining statistical thinking process based on data in constructing statistical models for decision making under uncertainties.
- Engineer skillset Design, construct, install, test and maintain highly scalable data management systems etc.

REFERENCES

[1] Life Cycle of Data Science Projects Posted by Vincent Granville on October 10, 2016 http://www.datasciencecentral.com/profiles/blogs/life-cycle-of-data-science-projects
[2] "The Data Science Process, Rediscovered" By Matthew Mayo, KDnuggets Published on Mar, 2016 http://www.kdnuggets.com/2016/03/data-science-process-rediscovered.html
[3] "Three Essential Components of a Successful Data Science Team" By Jack Danielson. Published on Aug, 2015 http://www.kdnuggets.com/2015/08/3-components-successful-data-science-team.html
[4] "Data Science Team Roles" By Michael Walker published Oct, 2016 http://www.datascienceassn.org/content/data-science-team-roles

