RESEARCH ARTICLE   OPEN ACCESS

'Essay on Practical Implementation Problems in Data Analytics Project Execution

Samhitha Challa,
Independent Researcher and Data Scientist

ABSTRACT
Businesses today heavily rely on data analytics for decision making. Managers at all levels need to back their decision with data and analysis. Short technology evolution cycles leave no time for managers to learn and become proficient, given theconstantly updating required skills list. Hence they rely on data scientist and business analysts to carry out data analytics projects. Even the analysts face many practical implementation issues during project execution. This paper attempts to discuss few practical implementation problems in data analytics projects.Because data analytics requires thorough understanding of math, business and technology – limitations to interpreting analytics reportsneed to be understood and accommodated for, before managers make decisions based on them. This paper draws from personal experiences in the industry.
Aim: To discuss practical implementation problems in data analytics and big data projects. To caution the managers and practitioners, of data science, on the biases in reporting.

Date of Submission: 16-10-2017   Date of acceptance: 27-10-2017

I. INTRODUCTION

Unprecedented rate of development in distributed computing and storage capabilities, madelarge scale data processing a reality. Many companies are switching to data driven decision making from the conventional, experience-based decision making process. This means for every decision made in a company, data analytics reports are the reason why managers are looking at that problem, why they are planning to implement a particular recommendation and also how they estimate impact of that change. Thereby making, designing an analytics study from data acquisition to insight generation and interpreting reports, the most crucial aspect of a manager’s role. And by extension, every manager must either be a business-savvy data analyst or someone who is trained to clearly understand research methodology [1] and analytics study design along with other managerial skills, to fulfill his/her role as a manager and decision maker.

Practical bottlenecks in making data driven decisions are the reality of today and this paper attempts to discuss few of them. The idea of racing past the competitors in adoption of data analytics,has contributed to growth of data science and misinterpretation equally.Awareness about practical implementation issues in data analytics projects, can help managers and data analyst plan ahead, in setting up an analytics department and plan to avoid hiccups in executing an analytics project.

Data Analytics is using data to find patterns (descriptive analytics), algorithms to design/developoptimized outcomes (prescriptive analytics) and forecast required metrics (predictive analytics).Data Analytics is a combination of statistics, business and technology. Every analytics project goes through the following phases:

1. Hypothesis Design
2. Data Acquisition
3. Data Wrangling and Data Exploration
4. Statistical Study to prove/ disprove a hypothesis
5. Observation Listing / Insight Generation / Recommendation
6. Report Generation
7. Decision Making

Biases in each stage of this project get carried on to the next, distorting findings and observations, which are inputs to decision making. So it is of utmost importance to identify such biases and avoid major problems which may arise due to inefficient decision making based on such input.

II. DISCUSSION ON PRACTICAL IMPLEMENTATION PROBLEMS IN EACH PHASE OF AN ANALYTICS PROJECT

Hypothesis Design:

Hypothesis Design in most cases is equivalent to identifying an operational or marketing or sales related problem for which cause or process-blockneeds to be identified, and/or prediction needs to be made to estimate the future losses/gains.In most cases hypothesis design revolve around the decisions and criteria on which managers are incentivized.

In general, hypothesis must be clear, specific, falsifiable, logically consistent and testable. Since most of the analytics studies in companies are done to improve performance or identify the cause of decline in performance it must be designed by keeping in mind the goals of the company. Because most of the studies are initiated by managers, it is only natural to see the analytics study objectives to positively line up with criteria on which managers are incentivized. Therefore such studies may be beneficial to the manager per say but, these studies may not give a complete solution to advancing company's goals. [2]

To understand this more clearly let us consider the case of call center operations manager in a product based company. For a call center operations manager, incentives are generally designed around expense reduction in a product based company because, the sole purpose of call center operation is to improve customer satisfactionduring product usage, so for a product company call center operation is an expenditure. The manager designs analytics study and hypothesis around expense reduction to achieve his expense saving targets. The decisions made based on such study may not necessarily improve customer satisfaction – the actual aim of the product support centerof that company.

The analytics study goals and hypothesis design must be done with overall company's goals in mind.

Data Acquisition

Every company captures data in a unique way and there are processes set up for data capture of transactions, customer details, etc. Data capturing set-up in companies is designed to log every customer and every transaction, which means the entire data set collected is the population and not the sample. So sampling bias is never the case. But the quality of data captured is the question of concern. The data captured must be accurate, timely, complete and consistent, failing which will result in poor and ineffective analysis and decision making input. [3]

A problem even with the best data capturing infrastructure is that, there is only a set of defined data attributes available. Most analytic studies are performed on data that has been captured with the available attributes. Sometimes to falsify a hypothesis certain data attribute - which does not exist in the data capture mechanism – could prove vital. In that case, one cannot create a new data capture mechanism to just go ahead with the study and even if they did they wouldn't be able to generate that data attributes for the previously collected data points.

The general work around in this case is to create a metric which may help answer hypothesis falsification.Depending on the metric, instead of a measurable attribute, adds bias to the data – depending on how the metric was designed.

Another serious concern while creating a data capture mechanism is that including all the data attributes is not very storage efficient idea. Limitations on how many days' worth data can be achieved also determines availability of data for analysis.

Data Wrangling and Data Exploration:

Data Wrangling is manipulation of data to treat missing values, eliminating outliers, combine tables, etc. in order to create data sets containing all relevant attributes. Data Exploration means finding relations between variables in the hypothesis – positive or negative or no correlation, comparing distribution of variables across time, etc.

One problem in data acquisition is that, data is present in multiple tables. When the labeling is not consistent across tables or when each of the table does not have a unique identifier for a single data point across all tables combining data becomes tedious and sometimes reduces the quality of data on which the analysis is run.

In datasets where missing value data points account for a small percentage of all the data points, missing value data points are removed from the data set. Only when missing value data points are more than 15% of the data, missing value treatment is put in place. In many cases missing value treatment is replacingthe missing value data points with mean value of the population. This creates a bias. What if the missing values were outliers or extreme data points? And this bias that we created in the process cannot be quantified.

There are specific statistical methods to deal with missing values. Depending on the type of missing values – missing completely at random, missing at random, missing not at random – missing data treatment techniques can be used. Few techniques that analysts should prefer to adopt are list-wise deletion, pair-wise deletion, single imputation, maximum likelihood and multiple imputation. [4][5]

Statistical Study for Hypothesis Testing:

Descriptive analytics does not leave much scope for bias when the data sets are small and manageable. In case of structured big data, only way to understand the distribution is by creating qualitative category wise statistical summaries and looking at the processed numbers because data size creates constraints for visualization. The problem with viewing data from category wise aggregation level is that sometimes the categories accommodate too many subgroups in which each has a different statistical reality than the whole. Hence any conclusion drawn from data aggregated at any qualitative category level may be biased unless every sub category level data points exhibit the same properties as the category level distribution.

Prescriptive analytics is affected by the above mentioned problem and one other one. Prescriptive analytics involves finding cause of a problem and recommending a solution – generally causal inferences. Causal Inference studies are generally prone to confirmation bias. Confirmation bias is using data to fit the alternative hypothesis instead of attempting to disprove null hypothesis.

Predictive analytics have both the above mentioned bias and one other. To forecast/predict the future, we need a model/algorithm and data concerning the features and attributes of the variable that needs to be predicted. A model/algorithm for prediction is chosen based on distribution and properties of feature and attribute data set. When theseavailable data distributions match the assumptions for one predictive model, that model is used for forecast/ prediction. Another confirmation bias happens to occur when the analyst/study designer believes a particular model must be used for prediction, data points are chosen to confirm the use of that model, instead of investigating which model to use.

Observation Listing / Insight Generation / Recommendation:

These are the outputs of a statistical study. General tendency among practitioners is to exclude hypothesis that were proved false - this leaves out valuable information which may be useful for decision making.Many times the manager who asks for the analytics study, is the reporting manager of the team presenting the study. So the team may want to be cautious of adding certain observations which show negative effect of a past decision. A compounding effect of such misreporting may prove harmful to the operations of a firm.

The situation is further complicated in the case of an outsourced team running an analytics project for a company. The objectives alignment is different for each of the parties. The manager who outsourced the study, looks to improve the company condition, whereas the company doing the analytics study is fixed on pipelining projects from that manager – which means the analytics team is pressured into providing observations which are aligned with manager's incentives rather than company goals.

This selective observation listing will generate biased reports. Open communications are a must for transparent reporting.

Report Generation:

Most analytics reports are narratives with data proofs. A lot of emphasis is given to narratives in generation of reports, as narratives are the easiest way to explain and showcase findings. [6]

In most cases the entirety of the study produces facts which are independent. Using business context these facts are stitched together to generate a report which seems to create a causal explanation for events that happen to occur independently. Such explanations sound very logical and obvious but have no study to back them sometimes.

Here is a simple example to understand this.Let us look at the following two facts.

Fact 1: Sales of pencils fell by 5% during time back to school season.

Fact 2: Sales of pens increased by 6% during back to school season.

If the report states these two facts as is, it is accurate from statistical study stand point. If the report states that customers prefer purchase of pens over pencils during the back-to-school season and inventory must be accordingly stacked with greater number of pens in the next back-to-school season it might be wrong. To write this statement a study on datasets with previous back-to-school season sales records must be consulted to see – that pattern of purchase in the past seasons, availability of pencils during the last back-to-school season, etc. in each season of purchase, to conduct a causal inference study.

Every statement in an analytics report must be written with careful consideration and explanation as to why that variable, data or graph was included in the analysis. [7]

Decision Making:

Managers are most likely to implement or act on recommendations that are in line with their thinking and previous decisions. This is because managers and analysts alike, have a frame of reference built unconsciously through which they view the facts before creating a narrative. [8]

In many reports the limitation of the study are generally not stated. It is very rare to see a forecast report stating limitations as follows - with "e" error rate, calculated considering "m"method, the forecast will be significant to consider only if data attributes have the same distributions as in the past "x" months without any major change in business environment.

Reports must also make sure to mention the time duration until which the recommendations remain valid – because of the constant changes in customer and company environment the recommendation will also hold validity until the assumptions in the study hold valid. This is a serious concern for decision makers as implementing certain recommendations may take longer than one can plan for. The decision to use a recommendation heavily depends on this.

III. CONCLUSION

Despite the widespread use of analytics in the industry there are many aspects of implementation and execution that are plagued with problems that need practitioners' attention. This paper attempts to highlight those practical implementation problems that the author frequently observed. This work serves as a starting point towards conscious research, analytic study design and execution with a goal to improve effective decision making.

REFERENCES

[1] Wixom, Barbara; Ariyachandra, Thilini; Douglas, David; Goul, Michael; Gupta, Babita; Iyer, Lakshmi; Kulkarni, Uday; Mooney, John G.; Phillips-Wren, Gloria; and Turetken, Ozgur (2014), The Current State of Business Intelligence in Academia: The Arrival of Big Data, *Communications of the Association for Information Systems*: Vol. 34 , Article 1.

[2] Barton, D., Court, D., 2012. Making advanced analytics work for you. Harvard Business Review 90 79-83

[3] Benjamin T.Hazena, Christopher A.Booneb, Jeremy D.Ezellc, L. AllisonJones-Farmerc, (2014),Data quality for data science, predictive analytics, and big data in supply chain management: An introduction to the problem and suggestions for research and applications, *International Journal of Production Economics,* Volume 154, Pages 72-80

[4] Daniel A. Newman, Missing Data : Five Practical Guide Lines (2014), *Organizational Research Methods,* Volume: 17 issue: 4, Pages 372-411

[5] Yiran Dong, Chao-Ying Joanne Peng (2013), Principled missing data methods for researchers, SpringerPlus, https://doi.org/10.1186/2193-1801-2-222, Online ISSN: 2193-1801

[6] Nassim Nicholas Taleb, The Narrative Fallacy, Nassim Nicholas Taleb (Ed.), *The Black Swan: The Impact of the Highly Improbable*, 2,(London: Penguin Books - 2010),*62-84*

[7] Bodily, Robert; Verbert, Katrien, Trends and issues in student-facing learning analytics reporting systems research (2017), *Proceedings of the Seventh International Learning Analytics & Knowledge Conference pages:309-318LAK edition:17 location:Vancouver, British Columbia, Canada date:March 13 - 17, 2017*

[8] Rajeev Sharma, Sunil Mithas, AtreyiKankanhalli, Transforming decision-making processes: a research agenda for understanding the impact of business analytics on organisations (2014), *European Journal of Information Systems,* Vol. 23, Issue 4, Pages 433-441

International Journal of Engineering Research and Applications (IJERA) is **UGC approved** Journal with Sl. No. 4525, Journal no. 47088. Indexed in Cross Ref, Index Copernicus (ICV 80.82), NASA, Ads, Researcher Id Thomson Reuters, DOAJ.