

A Project Report on
**Creating a Data Driven Organization
(Level 2 Organization)**

Submitted in partial fulfillment of the requirements for
the award of the degree of

Master of Science in Data Science and Big Data Analytics
by

Suraj Shrimant Mane

41907

Under the Guidance of
Prof. Esmita Gupta



Department of Information Technology

B. K. Birla College of Arts, Science and Commerce (Autonomous), Kalyan
B. K. Birla College Road, Near RTO, Kalyan

UNIVERSITY OF MUMBAI

Academic Year 2021-2022

Acknowledgement

This Project Report entitled “*Creating a Data Driven Organization (Level 2 Organization)*” Submitted by “*Suraj Shrimant Mane*”(41907) is approved for the partial fulfillment of the requirement for the award of the degree of *Master of Science* from *University of Mumbai*.

(Name)
Co-Guide

(Name)
Guide

Prof. Esmita Gupta
Head, Department of Information Technology

Place: B. K. Birla College (Autonomous), Kalyan
Date:

CERTIFICATE

This is to certify that the project entitled “***Creating a Data Driven Organization (Level 2 Organization)***” submitted by “***Suraj Shrimant Mane***”(41907) for the partial fulfillment of the requirement for award of a degree ***Master of Science***, to the University of Mumbai, is a bonafide work carried out during academic year 2021-2022.

(Name)
Co-Guide

(Name)
Guide

Prof. Esmita Gupta
Head Department of IT

Dr. Avinash Patil
Principal

External Examiner(s)

1.

2.

Place: B. K. Birla College (Autonomous), Kalyan
Date:

Declaration

I, the undersigned Mr. Suraj Shrimant Mane declare that the work embodied in this project work hereby, ***“Creating a Data Driven Organization (Level 2 Organization)”***, forms my own contribution to the research work carried out under the guidance of Ms. Harshada Topale is a result of my own research work and has not been previously submitted to any other University for any degree to this or any other University.

I have adhered to all the principles of academic honesty and integrity and have not misinterpreted or fabricated or falsified any idea/data/fact/source in my submission.

I would also like to mention that I have signed the NDA (Non-Disclosure Agreement) with Cloud Counselage Pvt. Ltd., the agreement terms as I cannot reveal the code and logic related to the project, I understand that any violations of the above will be cause for disciplinary action by the company and can also evoke penal action against me.

(Signature)

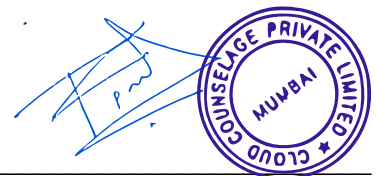
(Suraj Shrimant Mane., 41907)

Date:

Approval Sheet

This Project “Creating a Data Driven Organization (Level 2 Organization)” by Mr. Tushar Topale (Founder & CVO, Cloud Counselage Pvt. Ltd.) is approved to use as a final year master’s in Data Science and BDA. This Project was built and managed under supervision of Ms. Harshada Topale (Director & Managing Partner, Cloud Counselage Pvt. Ltd.).

This Project is also approved to showcase as a final year project by Mrs. Esmita Gupta Vice Principal of B.K. Birla College Kalyan (W).



Mr. Tushar Topale

(Founder & CVO)



Mrs. Harshada Topale

(Director & Managing Partner)

Mrs. Esmita Gupta

(Vice Principal, B. K. Birla College)

Abstract

Modern businesses rely on efficient management of their data assets. Data management and analytics have become decisive success factors for enterprises and companies during the last decades. However, the different aspects of transforming an organization into a data-driven one leave a lot of room for scientific inquiry. There are various aspects to consider in the transition to a data-driven organization. The paper investigates key elements and corresponding requirements of data-driven organizations to foster consensual definition in the research field. The work is grounded in the theory of organizational design and presents deductively generated results collected in a literature review. Key contributions of the paper are a shared understanding and a proposition for key elements of data-driven organizations.

Keywords

Data-driven organization, structured literature review, data capture, data management, data analytics.

Index

Sr. No.	No	Index	Page No.
Chapter 1: Introduction	1.1	Introduction	1
	1.2	Problem Statement	1
Chapter 2: Literature Review	2.1	Literature Review	2
Chapter 3: Building DDO	3.1	Building Data Driven Organization	5
	3.2	Creating Self Service Culture	5
	3.3	Build a Data Driven Culture	5
	3.4	Organizational Structure that Supports a Self-Service Culture	6
	3.5	Fostering a Culture of Data-Driven Decision-Making	7
Chapter 4: Planning	4.1	Level of Company	8
	4.2	Identify Company Level using Questionnaire	9
	4.3	Detail Report	13
	4.4	Level 1 to Level 2	14
	4.5	Data Collection	16
	4.6	Tpot Machine Learning	18
Chapter 5: Methodology	5.1	Proposed Model	19
	5.2	Proposed Methodology	20
	5.3	Requirements	23
Chapter 6: Design a Model	6.1	Flow Charts	24
Chapter 7: Experiment Result and Performance Analysis	7.1	Experimental Result	25
Chapter 8: Conclusion and Future Scope	8.1	Conclusion and Future Scope	26
Chapter 9: Reference	9.1	Reference	27

List of Figures

1.1	Organizational Hub-and-Spoke Model.....	6
1.2	Overview of the Maturity Model for Data and Analytics.....	8
1.3	Tpot Machine Learning Pipeline.....	18
1.4	Cleaning Flow Chart.....	24
1.5	Predicting Flow Chart.....	24

Chapter 1

1.1 Introduction

The continuous development and penetration of digital technologies lead to disruptive changes in the economy and society (European Commission 2020). These digital technologies enable novel ways to leveraging data for optimizing business processes and finding new configurations for innovative data driven business models. Data differs from typical resources as they enable reproducible services at almost zero marginal cost and offer a plethora of utilization options independent from specific devices.

Organizations need to integrate analytics technologies (e.g., predictive or prescriptive analytics) to leverage the potential of data to uncover previously undiscovered business insights and opportunities. Usually, papers employ the term “data-driven” to illustrate the role of data as a central resource. Throughout our research, we found several publications investigating the term data-driven. The oldest paper we found was from 1983, exploring data-driven automation in manufacturing industries and recommended handling data smarter to achieve substantial cost savings. Nowadays, we know that the term data-driven is relevant for manufacturing and all industries and society in general. Forecasts predict that the amount of data stored worldwide will increase fivefold from 33 Zettabytes in 2018 to 175 Zettabytes in 2025. Subsequently, considering the economic value of data is an issue that cannot be ignored. Organizations must embrace this disruptive change process and integrate data into existing and new processes. Since there is still no adequate description for this type of organization, we are trying to identify key elements and derive conceptual requirements. Further, we describe those organizations as data-driven organizations (DDO).

In Today’s date many companies are trying to become data-driven organizations, so I am working on this DDO Project to make Cloud Counselage as data-driven organization.

1.2 Problem Statement

1. How to Create a Self-Service Culture
2. Identify Company on which Level for DDO
3. Once it identifies Company’s Level, what are Solution and Strategy that we Provide a Company
4. Which Model is useful for Company’s Data Cleaning and Prediction.

Chapter 2

Literature Review

Gartner Survey Shows Organizations Are Slow to Advance in Data and Analytics. A worldwide survey* of 196 organizations by Gartner, Inc. showed that 91 percent of organizations have not yet reached a "transformational" level of maturity in data and analytics, despite this area being a number one investment priority for CIOs in recent years.

"Most organizations should be doing better with data and analytics, given the potential benefits," said Nick Heudecker, research vice president at Gartner. "Organizations at transformational levels of maturity enjoy increased agility, better integration with partners and suppliers, and easier use of advanced predictive and prescriptive forms of analytics. This all translates to competitive advantage and differentiation."

"It's easy to get carried away with new technologies such as machine learning and artificial intelligence," added Mr. Heudecker. "But traditional forms of analytics and business intelligence remain a crucial part of how organizations are run today, and this is unlikely to change in the near future."

Data maturity is a measurement that demonstrates the level at which a company makes the most out of their data. To achieve a high level of data maturity, data must be firmly embedded throughout the business and fully integrated into all decision-making and activities.

The 'maturity' part of the phrase 'data maturity' is directed at a company. If a company is mature with their data, it means that they utilise their data effectively to ensure that they are getting the best out of it, and they do this in a responsible way that maximises security for the people or business whose data they are handling.

The types of data can be specific to an individual (including personal details such as age, address, contact information) or metadata that explains how long visitors visit a web application, how they interact with it (what they click on etc.) and how likely they are to return. These are just some examples of numerous groups of data that can be essential to businesses.

Tree-based Pipeline Optimization Tool, or TPOT for short, is a Python library for automated machine learning. TPOT uses a tree-based structure to represent a model pipeline for a predictive modeling problem, including data preparation and modeling algorithms and model hyperparameters.

An evolutionary algorithm called the Tree-based Pipeline Optimization Tool (TPOT) that automatically designs and optimizes machine learning pipelines.

An optimization procedure is then performed to find a tree structure that performs best for a given dataset. Specifically, a genetic programming algorithm, designed to perform a stochastic global optimization on programs represented as trees.

TPOT uses a version of genetic programming to automatically design and optimize a series of data transformations and machine learning models that attempt to maximize the classification accuracy for a given Supervised learning data set.

For machine learning models and tools require you to prepare data before it can be fit to a particular ML model. One hot encoding is a process of converting categorical data variables so they can be provided to machine learning algorithms to improve predictions. One hot encoding is a crucial part of feature engineering for machine learning.

One hot encoding is one method of converting data to prepare it for an algorithm and get a better prediction. With one-hot, we convert each categorical value into a new categorical column and assign a binary value of 1 or 0 to those columns. Each integer value is represented as a binary vector. All the values are zero, and the index is marked with a 1.

One hot encoding is useful for data that has no relationship to each other. Machine learning algorithms treat the order of numbers as an attribute of significance. In other words, they will read a higher number as better or more important than a lower number. While this is helpful for some ordinal situations, some input data does not have any ranking for category values, and this can lead to issues with predictions and poor performance. That's when one hot encoding saves the day.

One hot encoding makes our training data more useful and expressive, and it can be rescaled easily. By using numeric values, we more easily determine a probability for our values. In particular, one hot encoding is used for our output values, since it provides more nuanced predictions than single labels.

Extremely Randomized Trees Classifier (Extra Trees Classifier) is a type of ensemble learning technique which aggregates the results of multiple de-correlated decision trees collected in a "forest" to output its classification result. In concept, it is very similar to a Random Forest Classifier and only differs from it in the manner of construction of the decision trees in the forest. Each Decision Tree in the Extra Trees Forest is constructed from the original training sample. Then, at each test node, each tree is provided with a random sample of k features from the feature-set from which each decision tree must select the best feature to split the data based on some mathematical criteria (typically the Gini Index). This random sample of features leads to the creation of multiple de-correlated decision trees.

ExtraTreesClassifier is an ensemble learning method fundamentally based on decision trees. ExtraTreesClassifier, like RandomForest, randomizes certain decisions and subsets of data to minimize over-learning from the data and overfitting.

Extra Trees has Low Variance

Extra Trees is like Random Forest, in that it builds multiple trees and splits nodes using random subsets of features, but with two key differences: it does not bootstrap observations (meaning it samples without replacement), and nodes are split on random splits, not best splits.

So, in summary, ExtraTrees:

- builds multiple trees with bootstrap = False by default, which means it samples without replacement
- nodes are split based on random splits among a random subset of the features selected at every node

In Extra Trees, randomness doesn't come from bootstrapping of data, but rather comes from the random splits of all observations.

Predictive modeling is a statistical technique using machine learning and data mining to predict and forecast likely future outcomes with the aid of historical and existing data. It works by

analyzing current and historical data and projecting what it learns on a model generated to forecast likely outcomes. Predictive modeling can be used to predict just about anything, from TV ratings and a customer's next purchase to credit risks and corporate earnings.

A predictive model is not fixed; it is validated or revised regularly to incorporate changes in the underlying data. In other words, it's not a one-and-done prediction. Predictive models make assumptions based on what has happened in the past and what is happening now. If incoming, new data shows changes in what is happening now, the impact on the likely future outcome must be recalculated, too. For example, a software company could model historical sales data against marketing expenditures across multiple regions to create a model for future revenue based on the impact of the marketing spend.

Chapter 3

3.1 Building a Data-Driven Organization

In our opinion, a data-driven organization should possess three things:

1. A culture in which everyone buys into the idea of using data to make business decisions
2. An organizational structure that supports a data-driven culture
3. Technology that supports a data-driven culture and makes data self-service

3.2 Creating a Self-Service Culture

The most important and arguably the most difficult aspect of transitioning to a data-driven organization that practices DataOps is the cultural shift required to move to a data mindset. This shift entails identifying and building a cultural framework that enables all the people involved in a data initiative from the producers of the data, to the people who build the models, to the people who analyze it, to the employees who use it in their jobs to collaborate on making data the heart of organizational decision-making

3.3 Build a Data Driven Culture

1. Hire data visionaries

We need people who see the “big picture” and understand all the ways that employees can use data to improve the business. Although this certainly includes analyzing marketing, sales, and customer data, it doesn’t end there. Data-driven decisions can help with internal operations, such as making customer service and support more efficient, and cutting costs from inventory

2. Organize your data into a single data store accessible to everyone

All of the data in the universe won’t help if that data is inaccessible to the people who need it to make business decisions. A data-driven company consolidates its data while keeping it continuously up to date so that employees have access to the most accurate information at any given point in time. This means eliminating data silos and effectively democratizing data access. There are, of course, always data security and compliance issues, but making data available to everyone is an important feature of a self-service data culture. Always allow employees to see the data that affects their work. They need to see this not only at a granular level, but also in a holistic way that helps them to understand the bigger picture. Doing this will make your employees more informed, skilled, and enthusiastic about using data to improve the business.

3. Empower all employees

All employees should feel comfortable taking initiative when it comes to suggesting ways that data can be used. This kind of mentality goes well beyond just using data, of course. If you build a company where all employees feel free to give opinions as long as they are backed up by data even if those opinions contradict senior executives’ assumptions, you are building an organization where the best ideas will naturally gravitate to the top and keep you competitive in even the fastest-moving markets.

4. Invest in the right self-service data tools

Our data, even if readily accessible, won’t help your business much if most of employees can’t

understand it or don't apply it to business problems. We can solve this problem by investing in the right data tools. we should pick tools based on your goals, but as a starting point, tools should make it easy for employees to access, share, and analyze data. we might want tools that can be directly embedded into the business tools that already use; for example, Excel and Tableau. And make sure to invest in training for these tools.

5. Hold employees accountable

Technology will take us only so far. We also need to put incentives in place to encourage employees to use the technology and tools. We also should have a way to measure and grade progress toward a self-service data culture. This means holding employees accountable for their actions and progress when they effectively use data to drive business decisions. Only when company reward employees for actions based on data will you achieve true cultural transformation.

3.4 Organizational Structure that Supports a Self-Service Culture

In most successful data-driven organizations there is a central data team that publishes data and manages the infrastructure used to publish that data. In others, there might be multiple data teams embedded in different departments, each catering to the needs of that department. Ironically, the latter model is typically less successful in creating a data-driven culture, even though data teams are there in each department. The reason is simple: such an organization creates low connectivity between the different departments and ends up creating data silos. A strong, functional, central data team is therefore extremely important in creating connectivity between the different departments of an organization. They usually publish the most important datasets, making sure that there is a single source of truth that underpins the analyses.

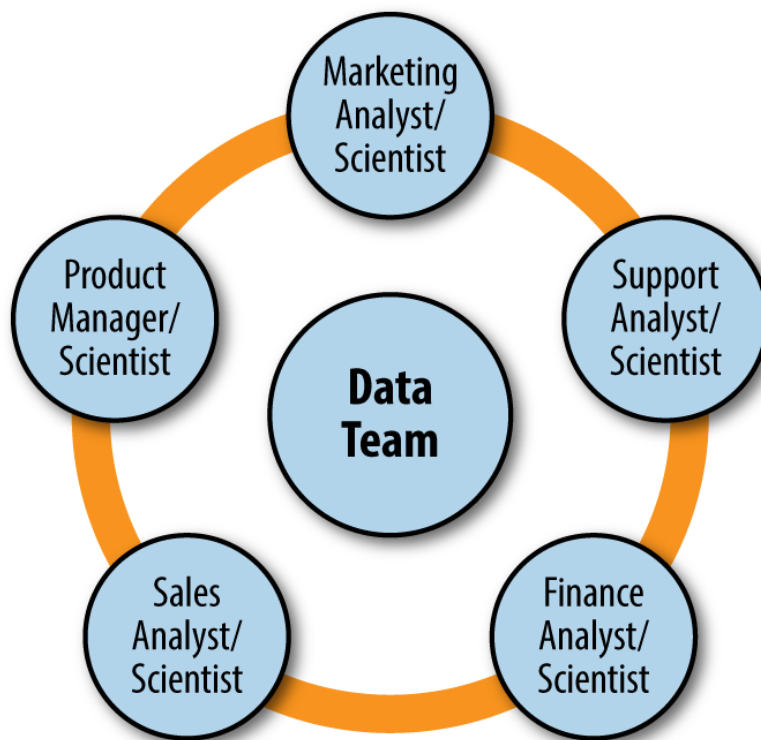


Figure 1.1 Organizational hub-and-spoke model (source: Qubole)

3.5 Fostering a Culture of Data-Driven Decision-Making

To succeed at becoming a data-driven organization, employees should always use data to start, continue, or conclude every single business decision, no matter how major or minor.

What drives companies to have a successful data-driven culture? It's important to understand that it's not necessarily about the data itself. That's secondary. The technology itself comes in third. Data-driven decision-making is first and foremost about the organization.

Regardless of whether you have acknowledged it, your business already has a culture of decision-making. That culture might not be geared toward a data-driven approach. All too many companies subscribe to the "HIPPO" (highest-paid person in the office) method of decision-making, whereby the senior person in the meeting gets to make the final choice. Needless to say, this HIPPO can be wrong. But unless you have the data as well as the permission coming from the very top of the organization to argue back, that decision stands.

Chapter 4

Planning

4.1 Level of Company

A worldwide survey* of 196 organizations by Gartner, Inc. showed that 91 percent of organizations have not yet reached a "transformational" level of maturity in data and analytics, despite this area being a number one investment priority for CIOs in recent years.

"Most organizations should be doing better with data and analytics, given the potential benefits," said Nick Heudecker, research vice president at Gartner. "Organizations at transformational levels of maturity enjoy increased agility, better integration with partners and suppliers, and easier use of advanced predictive and prescriptive forms of analytics. This all translates to competitive advantage and differentiation."

Level 1 Basic	Level 2 Opportunistic	Level 3 Systematic	Level 4 Differentiating	Level 5 Transformational
<ul style="list-style-type: none"> Data is not exploited, it is used D&A is managed in silos People argue about whose data is correct 	<ul style="list-style-type: none"> IT attempts to formalize information availability requirements Progress is hampered by culture; inconsistent incentives 	<ul style="list-style-type: none"> Different content types are still treated differently Strategy and vision formed (five pages) 	<ul style="list-style-type: none"> Executives champion and communicate best practices 	<ul style="list-style-type: none"> D&A is central to business strategy
<ul style="list-style-type: none"> Analysis is ad hoc Spreadsheet and information firefighting Transactional 	<ul style="list-style-type: none"> Organizational barriers and lack of leadership Strategy is over 100 pages; not business-relevant Data quality and insight efforts, but still in silos 	<ul style="list-style-type: none"> Agile emerges Exogenous data sources are readily integrated Business executives become D&A champions 	<ul style="list-style-type: none"> Business-led/ driven, with CDO D&A is an indispensable fuel for performance and innovation, and linked across programs Program mgmt. mentality for ongoing synergy Link to outcome and data used for ROI 	<ul style="list-style-type: none"> Data value influences investments Strategy and execution aligned and continually improved Outside-in perspective CDO sits on board

D&A = data and analytics; ROI = return on investment

© 2017 Gartner, Inc.

Figure 1.2 Overview of the Maturity Model for Data and Analytics

The model identifies four types of analytics:

- Descriptive Analytics
- Diagnostic Analytics
- Predictive Analytics
- Prescriptive Analytics

4.2 Identify Company Level using Questionnaire

This is the Questionnaire that we asked companies and know about their company level on their answers.

Strategy:

Data and Analytics Maturity Assessment

STRATEGY

- What level of sponsorship does data & analytics have in your Organisation?
 1. We do not have a sponsor
 2. We have a champion within the team but don't have a senior sponsor willing to invest
 3. We have a senior sponsor (not C-Level) who is currently working to get C Level engagement
 4. We have a C-Level sponsor who doesn't yet have data and analytics objective(s) and/or a willingness to invest
 5. We have an engaged C-Level sponsor who has data & analytics objectives and a willingness to invest
- How well does your data & analytics strategy support your business strategy and associated goals?
 1. We don't have a data and analytics strategy
 2. Strategy in place for some functions but very siloed in nature
 3. Strategy in place for majority of functions but with a lack of understanding of business priorities of different functions
 4. Consistent view across business but more around operational reporting than supporting longer term strategic decision-making
 5. Data and analytics are used to drive operational excellence, enhance customer experience, guide product innovation and improve risk management

- Do you have a data and analytics road map?
 1. We have no roadmap in place
 2. We have an incomplete roadmap
 3. We have an existing roadmap but it's not regularly updated
 4. We have a completed roadmap which is shared internally
 5. We plan regularly with reviews and updates

- Do you track the benefits of your analytics projects?
 1. We do not track benefits
 2. Benefits are tracked within a few functions, but without consideration of the impact on adjacent departments performance
 3. Benefits are tracked within the majority of functions, but without consideration of the impact on the adjacent function's performance
 4. Benefits tracked within functions for majority of use cases / projects and some consideration given to impact on adjacent functions
 5. Analytics benefits measured and tracked for each use case / project and communicated across the business

Process:

PROCESS

- How do you filter, prioritise and fund incoming analytics requests?
 1. No demand management in place
 2. Requests reviewed sporadically but no consistent processes in place to prioritise requests
 3. Some processes in place to manage analytics requests but generally completed in an ad-hoc fashion
 4. Processes in place to prioritise incoming requests but only infrequently
 5. Processes in place across the organisation to filter, prioritise and fund incoming analytics requests on a frequent and regular basis
- How do you maintain and remove data and dashboards?
 1. No reporting lifecycle management in place
 2. Only ad-hoc reviews in place
 3. Some processes in place at a functional level, but only in silos across the Organisation
 4. Processes in place in most functions
 5. Processes and automation in place across the entire Organisation, in an integrated fashion

- How are business processes documented and communicated?
 1. Business processes are not documented
 2. Documented for some functions but not all, with inconsistencies in documentation and often out of date
 3. Documented for most functions but some are out of date
 4. Documented for all functions, generally up to date
 5. Fully documented across the whole organisation in a consistent manner and fully up to date
- How is the usage of analytics driven into your Organisation?
 1. No measures in place to ensure adoption of analytics solutions
 2. Brief instructions emailed out with link to dashboard
 3. Detailed but static user guides distributed via email with link to dashboards. Ad-hoc tracking of adoption
 4. Detailed user guides delivered via meetings and accessible online. Adoption tracked for majority of dashboards but typically not against clearly defined success criteria
 5. Detailed user guides delivered via meetings and accessible online, supported by on-demand videos / Gifs. Adoption measured for all dashboards with clearly articulated success criteria

Data:

DATA

➤ What types of data are you able to generate value from?

1. Mainly simple, structured data. Data is stored in files like Excel and csv. There may be some transactional systems but the data is only available within those systems
2. Mainly using structured data sources required for regular reporting. These come from files or system extracts. Ad-hoc data files are also used in departments
3. Structured data is regularly used as the main type of data across the organisation. Unstructured and semi-structured data is also being used in an ad-hoc manner, and there are general guidelines and best practices for dealing with these data types
4. All data types can be used, whilst structured data is used widely and consistently. Semi- and unstructured data is being used more widely in silos across the Organisation
5. All value is extracted from all available data, regardless of type, and used in the most appropriate way as part of everyday business processes.

➤ How readily available and governed is the organisation's data for analytics in all its forms and can people find and access it?

1. People don't typically know what data is available and gather and create their own sets of data, usually manually.
2. Restricted access is given to regular data sources. Users often extract this data and add other data for their individual purposes. Data in files is not well controlled
3. Structured business data is available for use to all users that need it. New data that proves to be useful is made generally available. Users also make use of semi and unstructured data

➤ How consistently is the data refined, processed and governed to the correct level required for the specific analytics?

1. There is no official governance of the data. Any data cleansing is dependent on individual need and knowledge
2. Access to regular data sources is governed. Data outside of these systems is not. Data correction may happen within the data systems or by individuals that extract data from these systems for reporting
3. Data extracted from systems into models are cleansed and governed. Other sources use are still often ungoverned and reliant on individual knowledge. Standardisation of reference data is occurring, but not enterprise-wide
4. All regular data access is governed and audited. This data has passed through standard cleansing processes. Governance and cleansing processes are mature and well managed
5. All regulated data has the same level of data quality applied. Access to data is well governed and auditable. All data has the appropriate level of processing, cleansing and auditing applied

4. There is a well-understood availability of all structured data.

External sources can be made available. Semi- and unstructured data is also available

5. Data is fully available via regulated channels as appropriate for the business. All data models and definitions are understood and are used correctly across the business. Data for experiments is also available appropriately.

➤ Is your data defined, standardized and consistently modelled for analytics purposes?

1. No defined models, individuals build models based on specific needs. Data definitions are dependent on individual knowledge
2. Data from regular systems have an understood structure within those systems. Departments also create their own KPIs and calculations
3. Data models are being built to enable cross-system reporting but often stored in data silos. Standardised KPI definitions are defined but sometimes at department level only
4. All structured business data is well understood and modelled for analytics. Standardised data definitions and calculations exist.
5. All regulated data is well defined and modelled. New data has a well understood process for being made available in the same quality way

Platforms:

PLATFORMS

➤ How do you store and manage all of your data assets securely?

1. Data stored in files, folders and transactional sources systems. Data security does not often exist
2. Structured data stored in silos or files across departments. Unstructured or semi-structured data is not well managed. Data security implementation is inconsistent
3. Data storage is able to store most of the different types of data in the organisation. There is a central location for data assets but still some silos. Data security is a central requirement
4. Well-established data storage strategy to cover the majority of data, including data lake, data warehouse, streaming and advanced analytics requirements. Well-established security and data access policies in place
5. The platform is in place to effectively store and manage all the data assets of the organisation in a secure and trusted manner

➤ How do you perform cleansing, shaping and modelling of your data in line with your data strategy?

1. Ad-hoc editing of data, often in Excel or directly in data files
2. Excel is still mainly used, with some macros helping. Moderate use of BI tools with data prep capability are now being used to enhance this process. Departments may also be building their own siloed data models
3. Using analyst-led data prep tools, often to build dashboard-specific data models. Still done at department-level, rather than a central team. Some centralisation occurring
4. Enterprise-wide data preparation capability fully integrated with enterprise data platforms. Well-established methodology-driven

➤ What tools are available to support your users to do the reporting, analysis and data interactions required to deliver on your organisation's data & analytics strategy?

1. Individuals use various tools to access data, usually Excel as the dominant tool. Source system data analysis is restricted to canned reports on those systems
2. There are departmental-based tools in use, especially for BI reporting and analytics. Excel is still prevalent and dominant. Some specialists use database tools. Access to semi- and unstructured data stores is uncommon and, on an ad-hoc basis
3. There is established use of analyst-based analysis and visualisation desktop software (eg Alteryx, Tableau, Qlik, PowerBI) - Any advanced analysis is done using the predictive capabilities in those tools. Some advanced analysis also done on data sandboxes
4. There are well-established, enterprise-wide tools for data analytics, ad-hoc data prep, data science as well as data feeds to downstream processes
5. Leveraging all the tools available synergistically. Self-service analytics are deployed to the organisation via online analytics tools. Everyone in the business is a Citizen Data Scientist

modelling approaches. Ad-hoc requirements are serviced with an agreed set of tools

5. Using all integrated tools fully and automatically. Well defined data models and methodology governed and created by the right teams. Adding new data sources is understood and well-practiced

➤ How well do you manage master data?

1. No sources of this information available. It's based on local knowledge
2. There are some departmental-based data dictionaries but they are inconsistent. Master data is also departmental-based but some core organisational master data lists exist
3. There are some published data dictionaries and generally a centralised approach to providing this information. It's mostly done via document sharing rather than specialist tools. Centralising of core data is occurring
4. Central data dictionary exists to provide information about data assets and the core data is also centrally managed and provided. The approaches are well understood by users
5. Data metadata is aligned and maintained, search is easy, governance is tight, data properly tagged and defined. All managed by a central data team

Analysis:

ANALYSIS

➤ What type of analysis does your team/ organisation complete?

1. Descriptive Analytics is widespread - answering the what has happened questions. This is typically transactional and operational
2. Diagnostic analytics is widespread - performing analysis to understand why things happened the way they did
3. Predictive analytics is widespread - using historic data to understand what will happen in the future
4. Prescriptive Analytics is widespread - using analytics to generate foresight to establish what should be done in the future
5. Self-Learning Analytics (AI and machine learning) is widespread - using analytics to establish what we do not yet know

➤ How automated is your analysis?

1. Little or no automation in place. All analytics and reporting tasks are management manually
2. Automation is used sporadically and is restricted to specific uses, using legacy tools (e.g. VBA in Excel)
3. Core reports have been automated using modern BI platforms but this is not widely available to all analysts
4. Business function reports have been automated and all analysts have access to a modern BI platform that supports automation
5. Data science projects and machine learning is present. AI is being used to automate processes in some cases

➤ How is your analysis shared?

1. Static reports and information are shared manually (e.g., via email or printouts)
2. Interactive reports are shared manually (e.g., via email). Multiple versions of the same analysis exist
3. Analysis is shared via secure file sharing methods, however, duplicate content still exists due to a lack of governance
4. Secure, scalable, browser-based modern analytics platforms are leveraged and accessible by the majority of the organisation
5. Secure, scalable, browser-based modern analytics platforms are leveraged and accessible by the everyone in the organisation, and self-service analytics is widespread

➤ How do people access your analysis? Is it secure?

1. Security is not really considered and analysis is fully accessible to those who have access to it (e.g., hiding tabs in Excel)
2. Simple, single-level security measures are applied manually (e.g., Password protected Excel models)
3. Security is centralised per application, and controlled at the user level
4. Security is centralised across applications, controlled at the user level, and is also scalable
5. Highest levels of data encryption and security features are deployed (e.g., biometrics), including previously mentioned considerations.

Culture & Skills:

CULTURE & SKILLS

➤ How developed and structured is your data & analytics community?

1. There is no tangible data and analytics community present
2. An informal community exists via analytics focused internal meetings to discuss and share analysis. Official roles are not defined
3. Formal communication channels (e.g., Teams), and resources (e.g., intranet, newsletters and discussion forums), are used to promote data and analytics across the organisation
4. Regular activities to accelerate and reinforce data and analytics take place and community leadership roles are defined
5. Structured community activities take place, driven by defined community roles and there is widespread awareness across the organisation

➤ What is the extent, range and level of technical and soft skills in your team/ organisation?

1. Analytics skills are basic or non-existent and are limited to legacy platforms (e.g., simple Excel functions)
2. People are proficient with legacy analytics platforms (e.g., Advanced Excel users) and some are using self-service BI analytics platforms
3. Analysts are skilled in modern BI platforms to automate or accelerate analytics processes
4. Analysts are able to take ownership and automate analytics workflows using online analytics platforms (e.g., Tableau Online or Alteryx Server)

5. Analysts are highly proficient across the team and across multiple platforms, including data science and AI

➤ Do you have a structured approach to developing the relevant skills in your people?

1. No structured learning program is provided and all learning is done "on the job" or is self-applied
2. There is a limited budget for training courses and conferences. Learning materials are accessed in an ad-hoc manner
3. Training programmes are held in-house on a fixed schedule, and delivered across multiple modern BI platforms
4. Regular analytics learning program is a prerequisite for analysts to enter the full-time role
5. Regular analytics learning program is a prerequisite for analysts and on-demand training is always available

➤ Do you have standards in place that drive best practices?

1. There are no agreed or published standards
2. Documentation exists around analytics processes and solutions and the need for standards is recognised
3. Thorough guidelines and documentation exists about internal and external best practices and measures are taken to enforce these
4. Processes enforce documentation while solutions are being developed. Alignment to standards is enforced
5. Advanced processes are in place to enforce documentation and application of best practices (e.g. through Artificial Intelligence)

I am implemented this and I asked these questions to Cloud Counselage PVT. LTD and they answer it.

Based on answers I found that its level 1 company for DDO. So, we need to Improve this Company Level 1 to Level 2.

4.3 Detail Report:

This is the Detail Report about Level 1 and Level 2.

Why We Choose Only Level 1 and Improved to Level 2?

Because our Company (CC) follow all the level 1 Characteristics that's why it's Level 1 and We need to Improve this to Level 2, So we focus on Level 1 and Level 2. This project is mainly Focus on only Level 1 and Level 2.

In this report we can see there are Characteristics, what is Goal, What Activities we need to do to move to next level, what are the task into next level, what are the Challenges are we faced while doing this, Some Strategies and lastly Success Factor.

Level	Characteristics	Goal (of level)	Activities to move to next level	Tasks to move into next level	Challenges	Strategies	Success factor
Level 1: Basic (Aware)	Business & IT Leaders start understanding and acknowledge the importance of information and Enterprise Information Management (EIM) > Data is not exploited, it is used > Data & Analytics is managed in Silos > People argue about whose data is correct > analysis is ad-hoc > Spreadsheet and information firefighting > Transactional	Develop a basic information management system to, a. formalise information capturing, b. formalise information availability c. formalise reporting	1. Get Stakeholder Buy-in 2. Identify the Strategic position vis-a-vis Data (Refer resp. tools tab) 3. Develop a Data Architecture 4. Develop a Data Office 5. Develop a Data Governance Framework 6. Draft policies	1. Start building SOPs to follow a certain activity 2. Start maintaining data on cloud or on CRMS 3. Integrate the existing data 4. Plan strategies on a monthly basis and monitor them 5. Start understanding the kind of data that is required or extremely important for a particular task (Device a mechanism for data capture) 6. Evaluate the business strategy 7. Teaching various analytical softwares to the team members (ITIL certifications) 8. Implement some security levels (who as access to which type of data) 9. Conduct team building activities 10. Build/ design a data architecture	Lot of data is scattered. Time consuming to get all data at one place	Invest a lot of time in pre-work instead of directly taking up the task. Plan strategies on a monthly basis. Make fact based decisions	Maintenance of data and it's capacity to help make some business driven decisions.
Level 2: Opportunistic (Reactive)	Sharing of information takes place between the teams. Adherence to information management system is low > it attempts formalize information availability requirements > Progress is hampered by culture; inconsistent incentives > Organizational barriers and lack of leadership > strategy is over 100 pages; not business relevant > data quality and insight efforts but still in silos	Relook at the strategy and vision, Normalise the information system (i.e. ensure it is adopted and accepted by teams), ensuring business executives become data and analytics champions	1. Identify the Strategic position vis-a-vis Data (Refer resp. tools tab) 2. Adjust the architecture & process based on the Strategic position 3. Develop Policies to create a data driven culture and inculcate Data-driven mindset 4. Training & Development of Business Executives	1. Start strategising for smaller periods - 15 days or 1 month. Evaluate and improve the strategies. 2. Conducting montly company events to share insights about the data/ company's performance 3. Focus is more on normalizing the data that has been received 4. Doing data quality checks every week 5. Implement automations for various tasks that do not need manual intervention 6. Teaching various analytical softwares to the team members (ITIL certifications) 7. Have a centralised database maintained and do a quality check around that data. (mysql) 8. Have Bi-yearly compensation cycles 9. Conduct leadership workshops and structured tech workshops 10. SLAs should be defined	The tasks might be difficult to follow in comparision with the original way of working.	Picking up data that is only useful and required and using it to make better strategies.	More things accomplished in lesser time.

4.4 Level 1 to Level 2

This are the Level 2 Characteristics with their Solutions

1. It attempts formalize information availability requirements

Some of the Tips for Successful Requirements Gathering:

- Establish Project Goals and Objectives Early.
- Document Every Requirements Elicitation Activity.
- Be Transparent with Requirements Documentation.
- Talk To the Right Stakeholders and Users.
- Don't Make Assumptions About Requirements.
- Confirm.
- Practice Active Listening

2. Progress is hampered by culture; inconsistent incentives

Change Your Organizational Culture:

- Define desired values and behaviors.
- Align culture with strategy and processes.
- Connect culture and accountability.
- Have visible proponents.
- Define the non-negotiables.
- Align your culture with your brand.
- Measure your efforts.
- Don't rush it

Improve Workplace Culture:

- Build strong employee relationships.
- Connect people to a purpose.
- Encourage frequent employee recognition.
- Create positive employee experiences.
- Open up transparency and communication.
- Give teams the autonomy they seek.
- Schedule regular and meaningful one-to-ones.

3. Organizational Barriers and lack of leadership

This are the organizational Barriers

- Environmental barriers
- Interpersonal communication barriers
- Cultural barriers
- Decision-making barriers
- Insecurity within teams
- Remote working barriers

Lack of Leadership

- Have a Clear Vision
- Show You're Passionate
- Walk the Walk
- Make Concrete Plans
- Remember That it's Not About You
- Stay Positive
- Improve Your Communication Skills
- Admit Your Weaknesses
- Keep on Learning
- Think Critically
- Handle Conflicts with Grace
- Learn How to Delegate
- Encourage Creativity and Contributions
- Give Rewards and Recognition
- Discover Your Leadership Style

4. Strategy is over 100 pages but not business relevant

A business strategy, in simple terms, is a documented plan on how an organisation is setting out to achieve their goals. A business strategy contains a number of key principles that outlines how a company will go about attaining these goals. For example, it will explain, how to deal with your competitors, look at the needs and expectations of customers, and will examine the long-term growth and sustainability of their organisation.

5. Data quality and insight efforts but still in silos

Improve Data Quality

- Improve data collection.
- Improve data organization.
- Cleanse data regularly.
- Normalize your data.
- Integrate data across departments.
- Segment data for analysis

How to break down data silos in 4 steps:

The solutions to silos are technological and organizational. Centralizing data for analysis has become much faster and easier in the cloud.

- Change management
- Develop a way to centralize data
- Integrate data
- Establish governed self-service access.

4.5 Data Collection


This data collection is for Prediction and this data is collected by events.

This is the one of the events with their form. Firstly, need to register for the event and buy a free ticket and then fill all the details.

CLOUD COUNSELAGE PVT. LTD.

Data Visualization using Power BI

12th March 2022 4:00 PM



AUG 22

Copy of Data Visualization using Power BI

by Cloud Counselage

513 followers [Follow](#)

Free

[Register](#)

Data Visualization using Power BI

About this event

In our constant endeavor to help and support students, freshers, and young professionals aspiring for an IT career, we are arranging a webinar on **12th March 2022** that would help the individuals gain some insights on the topic "Data Visualization using Power BI". We are pleased to invite

Date and time

Mon, August 22, 2022
4:00 PM – 5:00 PM IST

Copy of Data Visualization using Power BI

Mon, Aug 22, 2022 4:00 PM - 5:00 PM IST

CCPL - Data Visualization using Power BI Tkt

Free

Sales end on Aug 22, 2022

1


Powered by **eventbrite**

English (US)

[Register](#)

Data Visualization using Power BI

12th March 2022 4:00 PM



Order summary

1 x CCPL - Data Visualization using Power BI Tkt	\$0.00
Total	\$0.00

Email address *

Cell phone *

College Name *

How did you come to know about this event? *

- ☐ Youtube
- ☐ Facebook
- ☐ Instagram
- ☐ LinkedIn
- ☐ Whatsapp
- ☐ Others

☐ I accept the [Eventbrite Terms of Service](#), [Community Guidelines](#), and [Privacy Policy](#). (Required)

Eventbrite isn't responsible for the health and safety of this event. Please follow the organizer's safety policies as well as local laws and restrictions.

Powered by [eventbrite](#)

Register

4.6 Tpot Machine Learning

Automated Machine Learning (AutoML) refers to techniques for automatically discovering well-performing models for predictive modeling tasks with very little user involvement.

TPOT is an open-source library for performing AutoML in Python. It makes use of the popular Scikit-Learn machine learning library for data transforms and machine learning algorithms and uses a Genetic Programming stochastic global search procedure to efficiently discover a top-performing model pipeline for a given dataset.

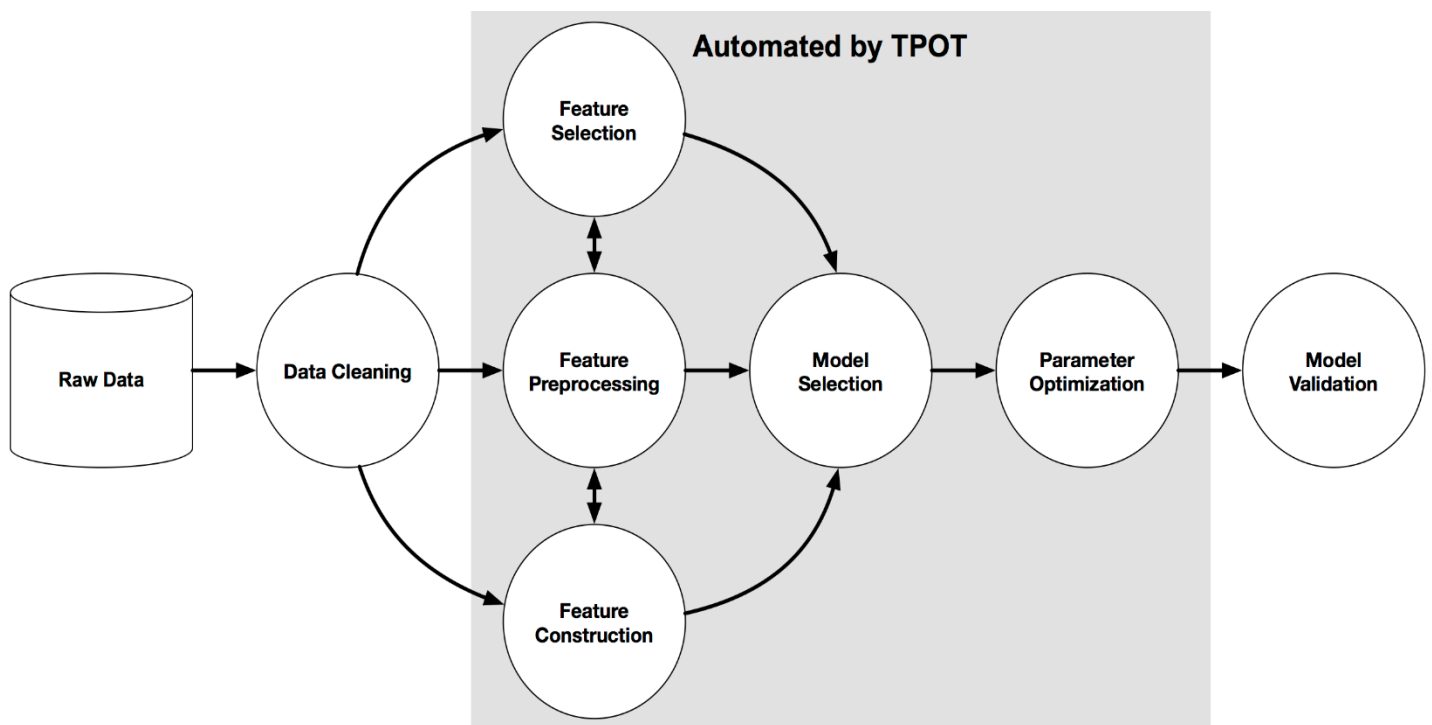


Figure 1.3 Tpot Machine Learning Pipeline

Chapter 5

Methodology

5.1 Proposed Model

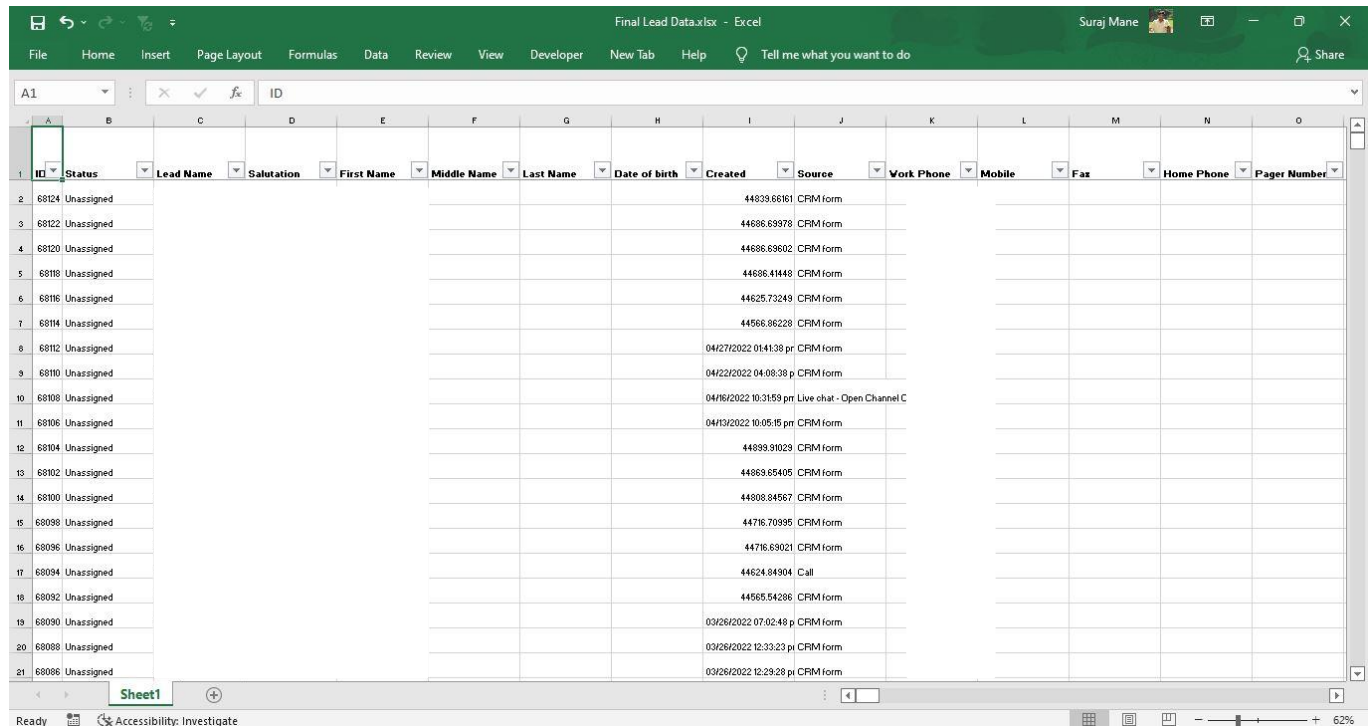
The Goal of Proposed DDO is to make a company data driven. In a data-driven approach, decisions are made based on data instead of intuition. That's because a data-driven strategy uses facts and hard information rather than gut instinct. Using a data-driven approach makes it easier to be objective about decisions.

This Problem is solved by making a model for Data Cleaning and Prediction. Firstly, we collect the data from CRM forms now we have a raw data and we need to clean that dataset using machine learning and make a data clean for next step. Now, we create the model for data cleaning, for data cleaning process we view the full dataset and check the null values and null columns. If in our data null values are present then the data shows the wrong information and it misleads, so we drop all the blank columns and columns that have more null values and in our data some of the columns that have fewer null values so we fill all those null values with blank. For this data there are many similar columns that information is very useful so I merged all of those columns and make a data proper way. In this dataset I need to make 1 column called Year of graduation using calculations, so we do some math operations and create that column and we replace all the blanks of numerical column to 0. Now our Data is Clean and also our model is ready for next work.

We create another model for prediction, for the data we already done many events in our company and once event complete google form is shared to everyone. Because of this we have raw data in various excel files, this all excel files have same columns so we merge all of this excel file into one master data, now our data is ready but still this data is raw so we need to clean first. Our Prediction is for which students is come for our next event and whom we need to send invite for next event. Firstly, we import the various libraries like NumPy, pandas, matplotlib, seaborn, sklearn, tpot classifier and accuracy scores etc. now we start to clean the data with very basic things like drop the null columns and fill the empty cells and merging the columns etc. now data is clean and ready for prediction but this data is categorical so we can't do prediction for categorical column so we need to make this columns to numerical for this we use one hot encoder and make our data into categorical form. We split the data into the train test split in 70-30 manner and use tpot for best possible algorithm to give us best predicted model with high accuracy and we apply this model to our data and makes our prediction and check the accuracy score.

5.2 Propose Methodology:

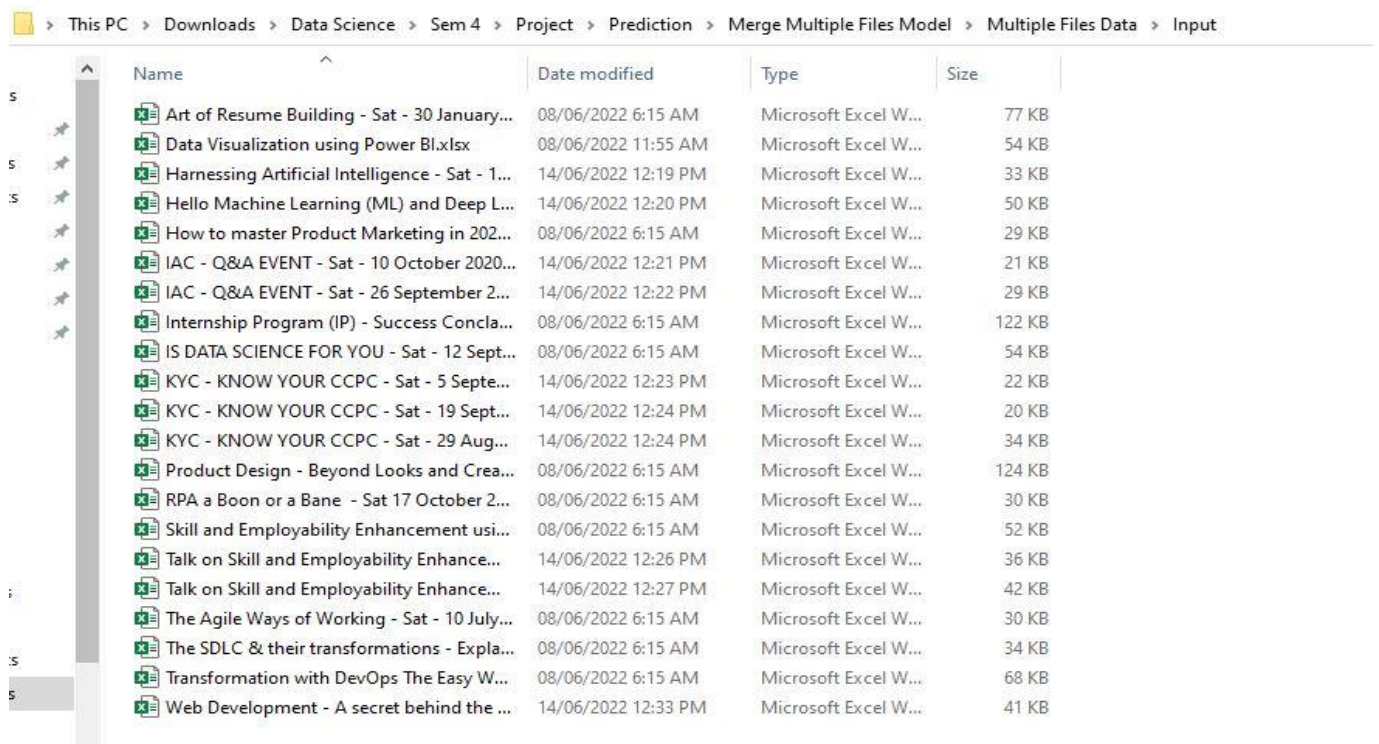
I. **Cleaning Model:** Our Dataset is huge, we have 26951 rows and 274 columns



ID	Status	Lead Name	Salutation	First Name	Middle Name	Last Name	Date of birth	Created	Source	Work Phone	Mobile	Fax	Home Phone	Pager Number
68124	Unassigned							44839.66161	CRM form					
68122	Unassigned							44686.63978	CRM form					
68120	Unassigned							44686.69602	CRM form					
68118	Unassigned							44686.41448	CRM form					
68116	Unassigned							44625.73249	CRM form					
68114	Unassigned							44566.86228	CRM form					
68112	Unassigned							04/27/2022 01:41:38 pm	CRM form					
68110	Unassigned							04/22/2022 04:08:38 p	CRM form					
68108	Unassigned							04/16/2022 10:31:59 pm	Live chat - Open Channel C					
68106	Unassigned							04/13/2022 10:05:16 pm	CRM form					
68104	Unassigned							44699.31023	CRM form					
68102	Unassigned							44689.85405	CRM form					
68100	Unassigned							44808.84567	CRM form					
68098	Unassigned							44716.70995	CRM form					
68096	Unassigned							44716.69021	CRM form					
68094	Unassigned							44624.84904	Call					
68092	Unassigned							44565.54286	CRM form					
68090	Unassigned							03/26/2022 07:02:49 p	CRM form					
68088	Unassigned							03/26/2022 12:33:23 pm	CRM form					
68086	Unassigned							03/26/2022 12:29:28 pm	CRM form					

II. **Prediction Model:** for prediction model data is in various excel files and We merge all this excel file into 1 Master Data.

Raw Data:



Name	Date modified	Type	Size
Art of Resume Building - Sat - 30 January...	08/06/2022 6:15 AM	Microsoft Excel W...	77 KB
Data Visualization using Power BI.xlsx	08/06/2022 11:55 AM	Microsoft Excel W...	54 KB
Harnessing Artificial Intelligence - Sat - 1...	14/06/2022 12:19 PM	Microsoft Excel W...	33 KB
Hello Machine Learning (ML) and Deep L...	14/06/2022 12:20 PM	Microsoft Excel W...	50 KB
How to master Product Marketing in 202...	08/06/2022 6:15 AM	Microsoft Excel W...	29 KB
IAC - Q&A EVENT - Sat - 10 October 2020...	14/06/2022 12:21 PM	Microsoft Excel W...	21 KB
IAC - Q&A EVENT - Sat - 26 September 2...	14/06/2022 12:22 PM	Microsoft Excel W...	29 KB
Internship Program (IP) - Success Concla...	08/06/2022 6:15 AM	Microsoft Excel W...	122 KB
IS DATA SCIENCE FOR YOU - Sat - 12 Sept...	08/06/2022 6:15 AM	Microsoft Excel W...	54 KB
KYC - KNOW YOUR CCPC - Sat - 5 Septe...	14/06/2022 12:23 PM	Microsoft Excel W...	22 KB
KYC - KNOW YOUR CCPC - Sat - 19 Sept...	14/06/2022 12:24 PM	Microsoft Excel W...	20 KB
KYC - KNOW YOUR CCPC - Sat - 29 Aug...	14/06/2022 12:24 PM	Microsoft Excel W...	34 KB
Product Design - Beyond Looks and Crea...	08/06/2022 6:15 AM	Microsoft Excel W...	124 KB
RPA a Boon or a Bane - Sat 17 October 2...	08/06/2022 6:15 AM	Microsoft Excel W...	30 KB
Skill and Employability Enhancement usi...	08/06/2022 6:15 AM	Microsoft Excel W...	52 KB
Talk on Skill and Employability Enhance...	14/06/2022 12:26 PM	Microsoft Excel W...	36 KB
Talk on Skill and Employability Enhance...	14/06/2022 12:27 PM	Microsoft Excel W...	42 KB
The Agile Ways of Working - Sat - 10 July...	08/06/2022 6:15 AM	Microsoft Excel W...	30 KB
The SDLC & their transformations - Expla...	08/06/2022 6:15 AM	Microsoft Excel W...	34 KB
Transformation with DevOps The Easy W...	08/06/2022 6:15 AM	Microsoft Excel W...	68 KB
Web Development - A secret behind the ...	14/06/2022 12:33 PM	Microsoft Excel W...	41 KB

Merging Multiple Excel sheet into one Master Data:

Code:

Importing required Libraries

```
In [ ]: import pandas as pd
import os
```

Path Location

```
In [ ]: input_loc = "C:/Users/suraj/Downloads/Data Science/Sem 2/Machine Learning/CC Model/Predictive Analysis/Input/"
output_loc = "C:/Users/suraj/Downloads/Data Science/Sem 2/Machine Learning/CC Model/Predictive Analysis/Output/"
```

```
In [ ]: fileList = os.listdir(input_loc)
fileList
```

Merging and Exporting Dataset

```
In [ ]: finalDf = pd.DataFrame()

for files in fileList:
    if files.endswith(".xlsx"):
        df = pd.read_excel(input_loc+files)
        finalDf = finalDf.append(df)

finalDf.to_excel(output_loc+"Master_Data.xlsx",index=False)
```

Exported Master Data Excel File:

This PC > Downloads > Data Science > Sem 4 > Project > Prediction > Merge Multiple Files Model > Multiple Files Data > Output				
	Name	Date modified	Type	Size
5	Master_Data.xlsx	22/06/2022 8:45 PM	Microsoft Excel W...	664 KB
5				
5				

Master Excel Data:

Master_Data.xlsx - Excel																										
Suraj Mane																										
File Home Insert Page Layout Formulas Data Review View Developer New Tab Help Tell me what you want to do																										
AB2 X f Students																										
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
	Order #	Order Date	First Name	Last Name	Email	Desat	Price	Ticket Type	Attend	Gross	Order Type	Netre	total	Fee	Intre	Paym	Attend St	Adi	Adi	ome	ome	ome	ome	ome	ome	ome
1	1585206831	2021-01-20 04:46:30				1		Art of Resume Building Tkt	2.2E+03		Free Order	USD	0	0	0	0	Attending									
2	1585206831	2021-01-20 04:50:36				1		Art of Resume Building Tkt	2.2E+03		Free Order	USD	0	0	0	0	Attending									
3	1585212527	2021-01-20 05:00:19				1		Art of Resume Building Tkt	2.2E+03		Free Order	USD	0	0	0	0	Attending									
4	1585315573	2021-01-20 09:01:33				1		Art of Resume Building Tkt	2.2E+03		Free Order	USD	0	0	0	0	Attending									
5	1585791017	2021-01-20 11:24:01				1		Art of Resume Building Tkt	2.2E+03		Free Order	USD	0	0	0	0	Attending									
6	1586410659	2021-01-21 06:43:24				1		Art of Resume Building Tkt	2.2E+03		Free Order	USD	0	0	0	0	Attending									
7	1586410571	2021-01-21 06:52:01				1		Art of Resume Building Tkt	2.2E+03		Free Order	USD	0	0	0	0	Attending									
8	1586414035	2021-01-21 06:53:32				1		Art of Resume Building Tkt	2.2E+03		Free Order	USD	0	0	0	0	Attending									
9	1586427027	2021-01-21 07:37:14				1		Art of Resume Building Tkt	2.2E+03		Free Order	USD	0	0	0	0	Attending									
10	1586429615	2021-01-21 07:41:53				1		Art of Resume Building Tkt	2.2E+03		Free Order	USD	0	0	0	0	Attending									
11	1586434355	2021-01-21 07:55:44				1		Art of Resume Building Tkt	2.2E+03		Free Order	USD	0	0	0	0	Attending									
12	1586435235	2021-01-21 07:56:28				1		Art of Resume Building Tkt	2.2E+03		Free Order	USD	0	0	0	0	Attending									
13	1586453437	2021-01-21 08:36:51				1		Art of Resume Building Tkt	2.2E+03		Free Order	USD	0	0	0	0	Attending									
14	1586500527	2021-01-21 09:53:03				1		Art of Resume Building Tkt	2.2E+03		Free Order	USD	0	0	0	0	Attending									
15	1586416229	2021-01-21 10:53:16				1		Art of Resume Building Tkt	2.2E+03		Free Order	USD	0	0	0	0	Attending									
16	1586657805	2021-01-21 12:28:22				1		Art of Resume Building Tkt	2.2E+03		Free Order	USD	0	0	0	0	Attending									
17	1586655343	2021-01-21 16:11:32				1		Art of Resume Building Tkt	2.2E+03		Free Order	USD	0	0	0	0	Attending									
18	1586626687	2021-01-21 17:02:30				1		Art of Resume Building Tkt	2.2E+03		Free Order	USD	0	0	0	0	Attending									
19	1587566783	2021-01-22 04:14:42				1		Art of Resume Building Tkt	2.2E+03		Free Order	USD	0	0	0	0	Attending									
20	1587578219	2021-01-22 04:41:12				1		Art of Resume Building Tkt	2.2E+03		Free Order	USD	0	0	0	0	Attending									
21	1587578855	2021-01-22 04:42:33				1		Art of Resume Building Tkt	2.2E+03		Free Order	USD	0	0	0	0	Attending									
22	1587580385	2021-01-22 04:43:43				1		Art of Resume Building Tkt	2.2E+03		Free Order	USD	0	0	0	0	Attending									
23	1587581635	2021-01-22 04:46:52				1		Art of Resume Building Tkt	2.2E+03		Free Order	USD	0	0	0	0	Attending									
24	1587582579	2021-01-22 04:48:40				1		Art of Resume Building Tkt	2.2E+03		Free Order	USD	0	0	0	0	Attending									
25	1587583103	2021-01-22 04:48:52				1		Art of Resume Building Tkt	2.2E+03		Free Order	USD	0	0	0	0	Attending									
26	1587585277	2021-01-22 04:55:06				1		Art of Resume Building Tkt	2.2E+03		Free Order	USD	0	0	0	0	Attending									
27	1587586133	2021-01-22 04:58:40				1		Art of Resume Building Tkt	2.2E+03		Free Order	USD	0	0	0	0	Attending									
28	1587587077	2021-01-22 05:04:15				1		Art of Resume Building Tkt	2.2E+03		Free Order	USD	0	0	0	0	Attending									
29	1587583263	2021-01-22 05:16:36				1		Art of Resume Building Tkt	2.2E+03		Free Order	USD	0	0	0	0	Attending									
30	1587588211	2021-01-22 05:26:32				1		Art of Resume Building Tkt	2.2E+03		Free Order	USD	0	0	0	0	Attending									
31	1587584401	2021-01-22 05:26:56				1		Art of Resume Building Tkt	2.2E+03		Free Order	USD	0	0	0	0	Attending									
32	1587624395	2021-01-22 06:36:38				1		Art of Resume Building Tkt	2.2E+03		Free Order	USD	0	0	0	0	Attending									
33	1587630019	2021-01-22 06:43:06				1		Art of Resume Building Tkt	2.2E+03		Free Order	USD	0	0	0	0	Attending									
34	1587642605	2021-01-22 07:10:27				1		Art of Resume Building Tkt	2.2E+03		Free Order	USD	0	0	0	0	Attending									
35	1587645319	2021-01-22 07:24:23				1		Art of Resume Building Tkt	2.2E+03		Free Order	USD	0	0	0	0	Attending									
36	1587650219	2021-01-22 07:35:02				1		Art of Resume Building Tkt	2.2E+03		Free Order	USD	0	0	0	0	Attending									
37	1587655333	2021-01-22 07:45:39				1		Art of Resume Building Tkt	2.2E+03		Free Order	USD	0	0	0	0	Attending									
38	1587662075	2021-01-22 08:00:43				1		Art of Resume Building Tkt	2.2E+03		Free Order	USD	0	0	0	0	Attending									
39	1587682807	2021-01-22 08:36:13				1		Art of Resume Building Tkt	2.2E+03		Free Order	USD	0	0	0	0	Attending									
40	1587686607	2021-01-22 08:58:44				1		Art of Resume Building Tkt	2.2E+03		Free Order	USD	0	0	0	0	Attending									
41	1587708555	2021-01-22 09:15:40				1		Art of Resume Building Tkt	2.2E+03		Free Order	USD	0	0	0	0	Attending									
42	1587758511	2021-01-22 10:28:15				1		Art of Resume Building Tkt	2.2E+03		Free Order	USD	0	0	0	0	Attending									
43	1587807421	2021-01-22 11:41:05				1		Art of Resume Building Tkt	2.2E+03		Free Order	USD	0	0	0	0	Attending									
44	1587854551	2021-01-22 11:16:46				1		Art of Resume Building Tkt	2.2E+03		Free Order	USD	0	0	0	0	Attending									
45	1588191197	2021-01-22 11:47:15				1		Art of Resume Building Tkt	2.2E+03		Free Order	USD	0	0	0	0	Attending									
46	1588226735	2021-01-22 18:19:36				1		Art of Resume Building Tkt	2.2E+03		Free Order	USD	0	0	0	0	Attending									
47	1588497751	2021-01-23 01:52:51				1		Art of Resume Building Tkt	2.2E+03		Free Order	USD	0	0	0	0	Attending									
48	1588612363	2021-01-23 03:48:26				1		Art of Resume Building Tkt	2.2E+03		Free Order	USD	0	0	0	0	Attending									
49	1588674723	2021-01-23 03:52:39				1		Art of Resume Building Tkt	2.2E+03		Free Order	USD	0	0	0	0	Attending									
50	1588674723	2021-01-23 03:52:39				1		Art of Resume Building Tkt	2.2E+03		Free Order	USD	0	0	0	0	Attending									
51	1588674723	2021-01-23 03:52:39				1		Art of Resume Building Tkt	2.2E+03		Free Order	USD	0	0	0	0	Attending									

5.3 Requirements

Python 3:

Python is a popular programming language. It was created by Guido van Rossum, and released in 1991. It is widely used for web development, data science, machine learning and AI applications.

Google Colab Notebook and Jupyter notebook:

This notebook is used to do a machine learning coding in interactive development environment. These notebooks are easy to use and easy to share.

Tpot:

Tree-based Pipeline Optimization Tool, It is a Python library for automated machine learning. TPOT uses a tree-based structure to represent a model pipeline for a predictive modeling problem, including data preparation and modeling algorithms and model hyperparameters

One-hot Encoding:

In this strategy, each category value is converted into a new column and assigned a 1 or 0 (notation for true/false) value to the column. Let's consider the previous example of bridge type and safety levels with one-hot encoding.

Extra-trees classifier:

This class implements a meta estimator that fits a number of randomized decision trees (a.k.a. extra-trees) on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting.

Pandas:

pandas is a fast, powerful, flexible and easy to use open source data analysis and manipulation tool, built on top of the Python programming language

NumPy:

NumPy is the fundamental package for scientific computing in Python. It is a Python library that provides a multidimensional array object, various derived objects and an assortment of routines for fast operations on arrays, including mathematical, logical, shape manipulation, sorting, selecting, I/O, discrete Fourier transforms, basic linear algebra, basic statistical operations, random simulation and much more.

Seaborn:

Seaborn is a library for making statistical graphics in Python. It builds on top of matplotlib and integrates closely with pandas' data structures. Seaborn helps you explore and understand your data.

Matplotlib:

Matplotlib is a comprehensive library for creating static, animated, and interactive visualizations in Python. Matplotlib makes easy things easy and hard things possible.

Sklearn:

Scikit-learn (Sklearn) is the most useful and robust library for machine learning in Python. It provides a selection of efficient tools for machine learning and statistical modeling including classification, regression, clustering and dimensionality reduction via a consistent interface in Python.

Chapter 6

Design a Model:

6.1 Flow Charts:

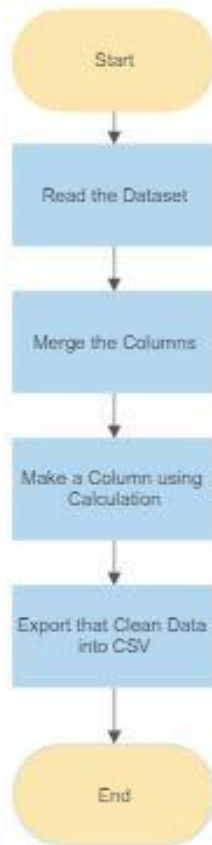


Figure 1.4 Cleaning Flow Chart

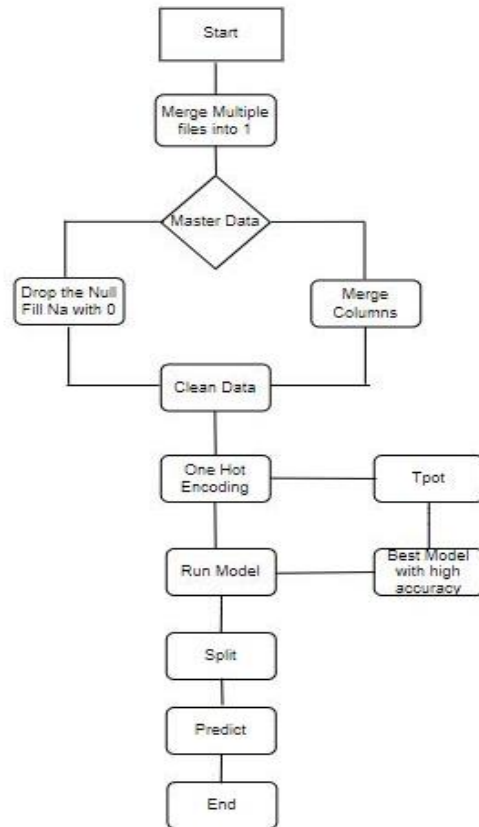


Figure 1.5 Prediction Flow Chart

Chapter 7

7.1 Experimental Result:

Confusion matrix

```
[ ] from sklearn.metrics import confusion_matrix
    tab_etc = confusion_matrix(df_predict, Y_test)
```

tab_etc

```
array([[1569,  2],
       [  0,  0]])
```

```
[ ] df_scores_comp = pd.DataFrame({'Actual':Y_test, 'Predicted':pred_etc})
    df_scores_comp
```

	Actual	Predicted
207	0	0
3345	0	0
72	0	0
5134	0	0
4568	0	0
2949	0	0
681	0	0
521	0	0
4374	0	0
4415	0	0
2718	0	0
2973	0	0

```
[ ] tab_etc.diagonal().sum() / tab_etc.sum()
```

```
0.9987269255251432
```

```
[ ] # Some important features
    feature_importance = pd.DataFrame({'Features': X_train.columns, 'Score': etc.feature_importances_})
    feature_importance.head(20)
```

	Features	Score
0	"IS DATA SCIENCE FOR YOU?" Ticket	0.005385
1	Art of Resume Building Tkt	0.013818
2	CCPL - Artificial Intelligence Webinar Tkt	0.002328
3	CCPL - Data Visualization using Power BI Tkt	0.582971
4	CCPL - Product Design & Full Stack Tkt	0.034978
5	CCPL - Product Marketing Tkt	0.004572
6	CCPL - The Agile Ways of Working Tkt	0.001795
7	CCPL - The SDLC & their transformations -Exp Tkt	0.002820
8	Hello ML and DL Tkt	0.005588
9	IAC - Q&A EVENT Tkt 10 OCT	0.000424
10	IAC - Q&A EVENT Tkt 26 SEP	0.001977
11	Internship Program(IP) Success Conclave	0.031438
12	KYC - Know Your CCPC 19 SEP	0.000402
13	KYC - Know Your CCPC 29 AUG	0.002134
14	KYC - Know Your CCPC 5 SEP	0.000274
15	RPA: A Boon or A Bane - Ticket	0.001397
16	Skill and Employability Enhancement using CCPC Tkt	0.004525
17	Talk on Skill and Employability Enhancement Ticket 11 JULY	0.001149
18	Talk on Skill and Employability Enhancement Tkt 1 AUG	0.005018
19	Transformation with DevOps: The Easy Way Tkt	0.018805

Model Accuracy = 0.9987269255251432

Prediction : Invite all the students that attend at least 2 times in any event

Result:

Model's Accuracy is 99% and its predict that all of the students who have at least 2 times present in events send invite all of them. So, we send invitation to all of the students for any event in next month

Chapter 8

Conclusion and Future Scope:

As the technology are blooming so we have our data cleaning and prediction model to help to make a company DDO. For data cleaning model we do some basic data cleaning things and some advanced thing like merging the two columns and remove the other text that we don't want in our columns, also creating a column using some logic and calculation. This help to company because Cloud Counselage PVT.LTD has same data format for everything so this help to clean every data also for prediction model. For prediction model first we merge the many excel files into 1 master data and then we clean this all the data and start to predict the which students are come for next event and whom we send the invitation for next event. So we using one hot encoder for categorical column and now we use tpot for giving us a best model for prediction and it gives extra tree classifier with best accuracy and our prediction is to invite all of the students who have at least 2 times present in events send invite all of them. So, we send invitation to all of the students for any event in next month and it is helpful for company for next events and work.

Chapter 9

Reference

- Anwar, M., Khan, S. Z., and Shah, S. Z. A. 2018. "Big Data Capabilities and Firm's Performance: A Mediating Role of Competitive Advantage," *Journal of Information & Knowledge Management* (17:04), p. 1850045 (doi: 10.1142/S0219649218500454).
- Ashrafi, A., Zare Ravasan, A., Trkman, P., and Afshari, S. 2019. "The role of business analytics capabilities in bolstering firms' agility and performance," *International Journal of Information Management* (47), pp. 1-15 (doi: 10.1016/j.ijinfomgt.2018.12.005).
- Berndtsson, M., Forsberg, D., Stein, D., and Svahn, T. 2018. "Becoming a Data-Driven Organisation," *ECIS 2018 Proceedings*, pp. 1-9.
- Chae, B., Olson, D., and Sheu, C. 2014. "The impact of supply chain analytics on operational performance: a resource-based view," *International Journal of Production Research* (52:16), pp. 4695-4710 (doi: 10.1080/00207543.2013.861616).
- Cheah, S., and Wang, S. 2017. "Big data-driven business model innovation by traditional industries in the Chinese economy," *Journal of Chinese Economic and Foreign Trade Studies* (10:3), pp. 229-251 (doi: 10.1108/JCEFTS-05-2017-0013).
- Davenport, T. 2006. "Competing on Analytics," *Harvard Business Review*, pp. 1-10.
- Devaraf, S., and Kohli, R. 2003. "Performance Impacts of Information Technology: Is Actual Usage the Missing Link?" *Management Science*.
- Dubey, R., Gunasekaran, A., and Childe, S. J. 2019. "Big data analytics capability in supply chain agility," *Management Decision* (8), pp. 2092-2112 (doi: 10.1108/MD-01-2018-0119).
- European Commission 2020. A European strategy for data, Brussels.
- Falletta, S. 2014. "Organizational Diagnostic Models - A Review and Synthesis," *Organizational Intelligence Institute*.
- Galbraith, J. R. 2016. "THE STAR MODEL™,"
- Grant, R. M. 1996. "Toward a knowledge-based theory of the firm," *Strategic Management Journal*.
- Guggenberger, T., Möller, F., Boualouch, K., and Otto, B. 2020. "Towards a Unifying Understanding of Digital Business Models," *24th Pacific Asia Conference on Information Systems*, pp. 1-14.
- Hartmann, P., Zaki, M., and Feldmann, Niels, Neely, A. 2016. "Capturing value from big data: A taxonomy of data-driven business models used by start-up firms," *International journal of operations & production management*, pp. 1-19.
- Henderson, D., Earley, S., Sebastian-Coleman, L., Sykora, E., and Smith, E. 2017. DAMA-DMBOK: Data

management body of knowledge, Basking Ridge, New Jersey: Technics Publications.

Hernaus, T., Aleksic, A., and Klindzic, M. 2013. "Organizing for Competitiveness – Structural and Process Characteristics of Organizational Design," *Contemporary Economics* (7:4), pp. 25-40 (doi: 10.5709/ce.1897-9254.122).

Jurgen, R. K. 1983. "Data-driven automation," *IEEE Spectrum* (20:5), pp. 33-35 (doi: 10.1109/MSPEC.1983.6369900).

Kates, A., and Galbraith, J. R. 2007. *Designing your organization: Using the star model to solve 5 critical design challenges*, San Francisco: Jossey-Bass.

Kearny, C., Gerber, A., and van der Merwe, A. 2016. "Data-driven enterprise architecture and the TOGAF ADM phases," in 2016 IEEE International Conference on Systems, Man, and Cybernetics (SMC), Budapest, Hungary, IEEE, pp. 4603-4608.

Klotzer, C., and Pflaum, A. 2015. "Cyber-physical.