SAVITRIBAI PHULE PUNE UNIVERSITY, PUNE

A
MINI PROJECT REPORT
ON
"GINA CASE STUDY"

SUBMITTED TO THE SAVITRIBAI PHULE PUNE UNIVERSITY, PUNE
IN THE FULFILLMENT OF THE REQUIREMENT
OF

Data Science and Big Data Analytics Laboratory
Third Year Computer Engineering
Academic Year 2023-24

BY

Name of Students:	Roll No.:
Sanskruti Kabadi	**3101065**

Under the Guidance of
Mr. S. D. Dighe

Sinhgad Institutes

DEPARTMENT OF COMPUTER ENGINEERING
STES'S SINHGAD INSTITUTE OF TECHNOLOGY AND SCIENCE
NARHE, PUNE – 411041

Sinhgad Institutes

DEPARTMENT OF COMPUTER ENGINEERING
STES'S SINHGAD INSTITUTE OF TECHNOLOGY AND SCIENCE
NARHE, PUNE – 411041

CERTIFICATE

This is to certify that,

Name of Students:	Roll No.:
Sanskruti Kabadi	**3101065**

studying in TE Computer Engineering Course SEM-VI has successfully completed their DSBDA Lab Mini-Project work titled **GINA CASE STUDY** at Sinhgad Institute of Technology and Science, Narhe in the fulfillment of the Bachelor's Degree in Computer Engineering in **Savitribai Phule Pune University**, during the academic year 2023- 2024.

Mr. S. D. Dighe	**Dr. G. S. Navale**	**Dr. S. D. Markande**
Guide	Head of Department	Principal

SINHGAD INSTITUTE OF TECHNOLOGY AND SCIENCE, NARHE, PUNE

Place : Pune

Date :

ACKNOWLEDGEMENT

I take this opportunity to acknowledge each and every one who contributed towards our work. We express our sincere gratitude towards guide **Mr. S. D. Dighe**, Assistant Professor at Sinhgad Institute of Technology and Science, Narhe, Pune for her valuable inputs, guidance and support throughout the course.

I wish to express our thanks to **Dr. G. S. Navale**, Head of Computer Engineering Department, Sinhgad Institute of Technology and Science, Narhe for giving us all the help and important suggestions all over the Work.

I thank all the teaching staff members, for their indispensable support and priceless suggestions. We also thank our friends and family for their help in collecting data, without their help DSBDA Lab Mini Project report have not been completed. At the end our special thanks to **Dr. S. D. Markande**, Principal Sinhgad Institute of Technology and Science, Narhe for providing ambience in the college, which motivate us to work.

Name of student Signature

Sanskruti Kabadi

CONTENT

INTRODUCTION

GINA: Global Innovation Network and Analysis:

- GINA stands for Global Innovation Network and Analysis
- The GINA case study exemplifies how a team utilized the Data Analytics Lifecycle to analyze innovation data at EMC.
- GINA, an acronym for Global Innovation Network and Analysis, is a team of senior technologists situated in centers of excellence around the world.
- Motivated by the desire to foster global idea sharing and knowledge dissemination among geographically dispersed members, the GINA team envisioned their approach as a means to achieve this.

There are Three key goals mention below:

1. **Store Formal and Informal Data:** This includes capturing both documented and undocumented information.
2. **Track Research from Global Technologists:** The repository would serve as a central hub for tracking research activities conducted by technologists across the globe.
3. **Mine the Data for Patterns and Insights:** By analyzing the data, the team aimed to uncover valuable patterns and insights that could be leveraged to enhance their operations and strategic direction.

Overview of Data Analytics Lifecycle:

1. Discovery:
- Understand the business problem and goals.
- Identify stakeholders and data sources.
- Develop initial hypotheses.

2. Data Preparation:
- Clean, transform, and organize data for analysis.
- This is often the most time-consuming step.

3. Model Planning:
- Choose the right analytical techniques based on the data and objectives.
- Research existing solutions for similar problems.

4. Model Building:
- Develop and test models using training and testing data.
- Evaluate model performance and refine as needed.

5. Communicate Results:
- Share key findings and insights with stakeholders.
- Ensure results are statistically sound and well-documented.

6. Operationalize:
- Deploy the model into production for real-world use.
- Monitor and maintain the model's accuracy over time.

Problem Statement: Write a case study on Global Innovation Network and Analysis (GINA). Components of analytic plan are 1. Discovery business problem framed, 2. Data, 3. Model planning analytic technique 4. Results and Key findings

PHASE 1: DISCOVERY

Team Members and Roles

- **Business User & Project Sponsor:** Vice President from the Office of the CTO (Chief Technology Officer)
- **Business Analyst (BA):** IT Representative
- **Data Engineer & DBA:** Dedicated personnel

Data Inventory

The data resided in two primary categories:

1. **Years of Idea Submissions:** Data from internal innovation contests spanning multiple years.
2. **Unstructured Data:** Minutes and notes capturing innovation and research activities from around the world.

Hypothesis

The analysis would be categorized into two key areas:

1. **Descriptive Analytics:** This initial phase aimed to uncover insights into current innovation trends, fostering further creativity, collaboration, and asset generation.
2. **Predictive Analytics:** By analyzing past data, the team intended to develop forecasts to inform executive management on future investment opportunities.

PHASE 2: DATA PREPARATION

Establishing an Analytical Sandbox

The initial step involved setting up a dedicated analytical sandbox. This secure environment would serve as a storage and experimentation space for the data, facilitating analysis without compromising the integrity of the original dataset.

Data Conditioning and Normalization

During the data exploration process, the team identified the need for data conditioning and normalization. This included tasks like:

- **Correcting inconsistencies:** Addressing issues such as misspelled names, extra spaces, and formatting variations.
- **Identifying missing datasets:** Determining if critical datasets were absent and exploring strategies to mitigate potential gaps.
-

Data Quality Assessment

The team recognized the importance of data quality for subsequent analysis. They acknowledged that poor quality data could negatively impact the reliability and accuracy of the extracted insights. Therefore, they established data quality thresholds to determine the level of cleaning and normalization required for the specific project goals.

PHASE 3: MODEL PLANNING

Milestone Identification:

- Defining clear milestones within the innovation process.
- These milestones would represent critical stages that ideas must traverse to reach successful outcomes.

Idea Journey Mapping:

- Tracing the movement of ideas through each identified milestone, highlighting the paths they take towards achieving the established goals.
- This mapping would involve analyzing both ideas that ultimately succeed and those that ultimately fail.

Comparative Analysis:

- Comparing and contrasting the journeys of successful and unsuccessful ideas. This would involve analyzing factors like:
  - Time taken to reach each milestone.
  - Resources utilized at each stage.
  - Decision-making points that influence outcomes.

Selection of Analytical Techniques:

- Determining appropriate analytical methods to compare times and outcomes for the identified ideas.
- The choice of methods could range from basic statistical tests (t-tests) to more complex machine learning algorithms (classification algorithms) depending on the data characteristics and complexity of the analysis.

Components of Analytic Plan	GINA Case Study
Discovery Business Problem Framed	Tracking global knowledge growth, ensuring effective knowledge transfer, and quickly converting it into corporate assets. Executing on these three elements should accelerate innovation.
Initial Hypotheses	An increase in geographic knowledge transfer improves the speed of idea delivery.
Data	Five years of innovation idea submissions and history; six months of textual notes from global innovation and research activities
Model Planning Analytic Technique	Social network analysis, social graphs, clustering, and regression analysis
Result and Key Findings	1. Identified hidden, high-value innovators and found ways to share their knowledge 2. Informed investment decisions in university research projects 3. Created tools to help submitters improve ideas with idea recommender systems

Fig: Analytic Planning from the EMC GINA Project

PHASE 4: MODEL BUILDING

Natural Language Processing (NLP) Techniques:
A data scientist utilized NLP techniques to analyze the textual descriptions of innovation roadmap ideas. This could involve tasks such as:

- Extracting key concepts and themes from the descriptions.
- Identifying sentiment and potential challenges expressed within the text.
- Categorizing ideas based on thematic similarities.

Social Network Analysis (SNA):
The team employed Social Network Analysis (SNA) using R and RStudio. This involved:

- Building social graphs that represent the connections and interactions between researchers and their innovation projects.
- Visualizing these networks to identify key players, collaboration patterns, and potential knowledge silos.

PHASE 5: COMMUNICATE RESULTS

- **Identification of Boundary Spanners and Hidden Innovators:** Boundary spanners are individuals who bridge different teams or departments, facilitating knowledge flow. Hidden innovators are talented individuals whose contributions may not have been readily apparent through traditional methods. By analyzing the data, GINA was able to identify these valuable assets within the organization.

- **Enhanced Knowledge Sharing:** The project significantly promoted knowledge sharing related to innovation and researchers. This included sharing across various disciplines within the company, as well as fostering connections with external researchers.

- **Intellectual Property and Research Opportunities:** GINA's insights enabled EMC to cultivate additional intellectual property leads based on research topics identified through the analysis. It also opened doors to forging collaborative relationships with universities for joint academic research, particularly in the fields of data science and big data.

- **Discovery of Hidden Innovator Cluster:** The study successfully identified a cluster of "hidden innovators" located in the Cork, Ireland office. This discovery, which might have been missed through traditional methods, highlighted the value of social network analysis in uncovering valuable contributors.

PHASE 6: OPERATIONALIZE

Key Findings:
The GINA project yielded several key findings that will inform future efforts:

- **Data Availability:** The project highlighted the need for a more robust data collection strategy moving forward.
- **Data Sensitivity:** The analysis revealed the presence of some potentially sensitive data within the innovation ecosystem. A data governance framework needs to be established to ensure proper handling of such information.
- **Business Intelligence (BI) Improvement:** The project identified a need for a parallel initiative focused on improving basic BI activities. This could involve tasks such as data standardization and ensuring data quality across various departments.

Model Re-evaluation:
The team acknowledged the importance of continuously monitoring and re-evaluating the model after deployment. Establishing a mechanism for ongoing assessment will allow the model to adapt and evolve over time, ensuring its continued relevance and effectiveness in supporting innovation efforts.

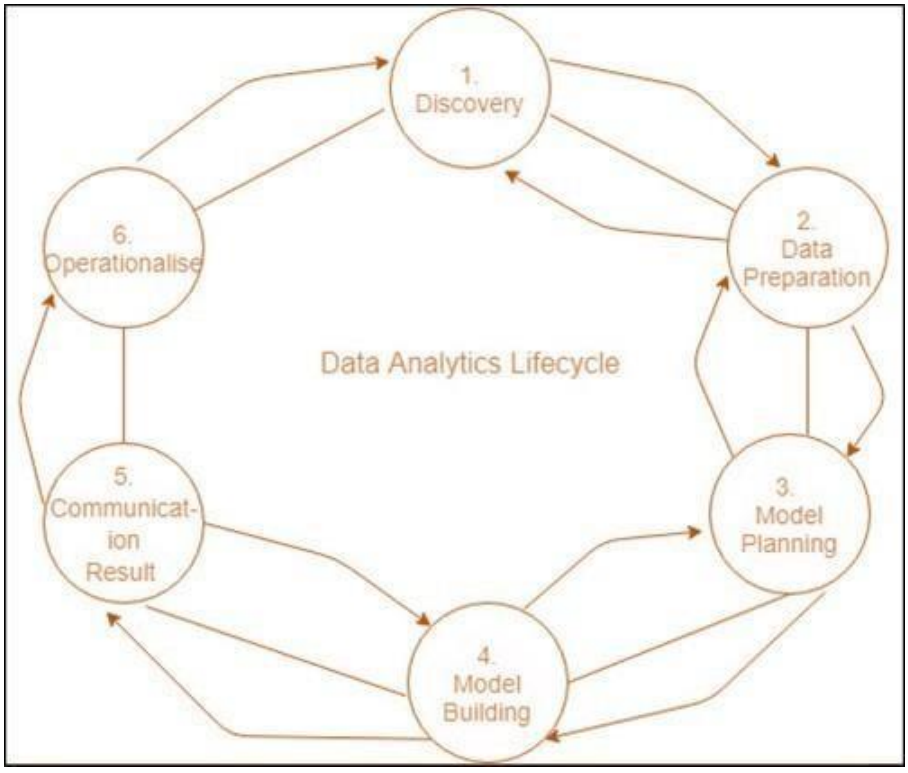

Fig: Data Analytics Lifecycle

SUMMARY AND CONCLUSION

Summary:

- The Data Analytics Lifecycle is an organized approach for managing and executing analytic projects.
- It consists of six distinct phases, providing a structured framework for project management.
- The preparation phase (phases 1 and 2) typically consumes the majority of the project timeline.
- There are seven essential roles required for a data science team to effectively execute projects within this lifecycle.
- The course focuses on the quantitative disciplines that underpin data analytics, including: Mathematics, Statistics, Machine Learning
- In addition, the course will provide an overview of Big Data analytics, including exploring key algorithms used for handling large and complex datasets.

Conclusion:

GINA (Global Innovation Network and Analysis) is a tool that provides insights and analysis to support innovation and technology development. It enables companies and organizations to stay up-to-date with the latest industry trends and technology advancements, allowing them to make informed decisions and stay competitive in their respective fields. The platform offers a range of features, including technology scouting, competitor analysis, and IP portfolio management, and can be customized to suit the needs of individual companies. Overall, GINA is a valuable resource for companies looking to innovate and stay ahead of the curve in their industries.

REFERENCES

1) Data-science-and-big-data-analy-nieizv_boaok
2) https://www.slideshare.net/SurakshaSanghavi/case-study-107609829 5.2
3) https://bhavanakhivsara.files.wordpress.com/2018/06/datascience-and-big-data analysis book.pdf
4) slidetodoc.com/data-analytics-lifecycle-data-analytics-lifecycle-data-science/
5) www.slideserve.com/avian/gina-a-network-of-geo-spatial-data-and-activities
6) bhavanakhivsara.files.wordpress.com/2018/07/data-analytics-unit-i-1.pptx
7) slideplayer.com/slide/17674162/
8) https://www.javatpoint.com/life-cycle-phases-of-data-analytics
9) https://onlinelibrary.wiley.com/doi/abs/10.1002/9781119183686.ch2