Harnessing Government Data: Analytical Perspectives on Karnataka and India

Internship Report
submitted in Partial Fulfillment for the Award of
Master of Technology

by

Manas Gupta
(IMT2019050)

to

INTERNATIONAL INSTITUTE OF INFORMATION TECHNOLOGY BANGALORE
JUNE 2024

CERTIFICATE

This is to certify that the internship report titled, **'Harnessing Government Data: Analytical Perspectives on Karnataka and India'** submitted by **'Manas Gupta' (IMT2019050)** is a bona fide work carried out under my supervision at 'Web Science Lab' from 10-01-24 to 28–06-24, in partial fulfilment of the Master of Technology course of International Institute of Information Technology Bangalore.

His performance during the internship was satisfactory,

(Signature of the Supervisor)
Supervisor Name
Address of the Company
(with seal)

Date: 22 June 2024
Place: Bengaluru, Karnataka

Undertaking by the Student

I, Manas Gupta, hereby declare that the report of the internship program titled, “Harnessing Government Data: Analytical Perspectives on Karnataka and India” is prepared by me. I also confirm that, the report is only prepared for my academic requirement and not for any other purposes.

I also confirm that, the submitted softcopy has been reviewed and approved for submission by my supervisor.

(IMT2019050)

Date: 22 June 2024
Place: Bengaluru, Karnataka

ACKNOWLEDGEMENT

I would like to express my sincere gratitude to the International Institute of Information Technology Bangalore (IIIT Bangalore) for providing me with the opportunity to undertake this internship. The support and resources provided by the institution have been invaluable to my learning and development.

I am deeply indebted to Prof. Badrinath R and Prof. Srinath Srinivasa for their exceptional guidance and mentorship throughout this internship. Their insights, advice, and encouragement have been instrumental in the successful completion of this project.

I would also like to extend my heartfelt thanks to my colleagues at the Web Science Lab, particularly Abraham GK, for their unwavering support and assistance. Their collaboration and help have greatly contributed to the progress and success of my work.

CONTENTS

ABSTRACT

During my internship at the Web Science Lab (WSL) at IIIT Bangalore, I concentrated on the analysis of government data from Karnataka and India, aiming to derive insights and build predictive models to inform policy decisions. Initially, I explored the Karnataka Data Lake website to understand the breadth of available data and identify datasets suitable for time series analysis. This comprehensive review involved assessing various datasets for their relevance, completeness, and analytical potential. Through this process, I identified several key datasets that were viable for further exploration and analysis.

To analyze the selected data, I employed a variety of advanced analytical techniques, including Multi-Linear Regression, ARIMA (AutoRegressive Integrated Moving Average), and Facebook's Prophet model. Each of these models brought unique strengths to the analysis. Multi-Linear Regression was utilized to understand the relationships between multiple predictors and a single outcome variable, providing insights into how different factors interacted. ARIMA was chosen for its effectiveness in modeling time series data, especially where temporal dependencies were present. Facebook's Prophet model was particularly useful for handling seasonal data and managing missing values, ensuring robustness in the predictive analysis. By applying these models, I was able to perform thorough analyses and uncover significant patterns and trends within the datasets.

In addition to analyzing the initial datasets, I sought out additional government data to enhance the scope of our analysis. This involved attending data meetups and thoroughly scouring various government websites to find relevant datasets. I successfully compiled 20 years' worth of data related to central government schemes. This extensive data acquisition process included thorough data processing and filtering to clean and prepare the data for analysis, ensuring it was suitable for building reliable predictive models.

The culmination of my efforts was the development of a predictive model aimed at forecasting key metrics such as Maternal Mortality Rate (MMR) and Infant Mortality Rate (IMR) based on a set of independent variables. The model utilized the processed and filtered data to generate reliable predictions, which could be instrumental in informing and improving government policies and programs. These predictive insights provided actionable data that could help shape effective interventions and strategies in public health and other critical areas.

Overall, my internship experience at WSL was deeply enriching, offering hands-on experience in data analysis, model building, and data processing within the context of government datasets. This opportunity not only honed my technical skills but also enhanced my understanding of the practical applications of data science in the public sector. The work I undertook has the potential to contribute significantly to data-driven decision-making processes, ultimately aiding in the development of more effective and targeted government policies and programs.

1. INTRODUCTION

1.1 About the Company

The Web Science Lab (WSL) at IIIT Bangalore is dedicated to developing models that extract semantics and analyze the impact of the web on various aspects of human life. The lab's research extends to applying web technologies across diverse fields such as education. Moreover, WSL undertakes projects focused on understanding and leveraging both open data and open-ended data.

Typical research endeavors at WSL include:

- **Extracting Semantic Relationships:** Analyzing text and social media data to uncover semantic associations.
- **Socio-Cognitive Web Models:** Developing computational models to study socio-cognitive phenomena on the web, such as the formation of collective opinions and the rise of celebrities.
- **Semantic Data Integration:** Integrating formal web data semantically.

1.2 Your role, responsibilities and contribution

During my internship at the Web Science Lab (WSL) at IIIT Bangalore, I served as a Data Analyst. My role focused on data acquisition, analysis, and predictive modeling of government datasets from Karnataka and India.

Role: As a Data Analyst, my primary role was to extract, analyze, and interpret government data to derive insights and build models to support informed decision-making and policy formulation.

Responsibilities:

1. **Data Exploration:**
    - Conducted an in-depth review of the Karnataka Data Lake website to understand the breadth of available datasets.
    - Identified datasets suitable for time series analysis, assessing their relevance, completeness, and potential for further exploration.
2. **Algorithm Selection and Implementation:**
    - Evaluated and selected appropriate algorithms for data analysis, including Multi-Linear Regression, ARIMA (AutoRegressive Integrated Moving Average), and Facebook's Prophet model.
    - Applied these analytical techniques to the selected datasets, leveraging the strengths of each model:
        - Multi-Linear Regression: Used to understand relationships between multiple predictors and an outcome variable.
        - ARIMA: Employed for modeling time series data, especially with temporal dependencies.
        - Facebook's Prophet: Utilized for handling seasonal data and managing missing values.
3. **Data Acquisition:**
    - Sought additional government data by attending data meetups and conducting thorough searches on various government websites.

- Successfully compiled 20 years' worth of data related to central government schemes, including PMAY, PMGAY, Jal Jeevan Mission, JSY, and JSSY.

4. **Data Processing and Cleaning:**
    - Performed extensive data processing and filtering to clean and prepare the acquired datasets for analysis.
    - Addressed issues such as missing values, data inconsistencies, and outliers to ensure high-quality, reliable data for modeling.
5. **Predictive Modeling:**
    - Developed predictive models to forecast key metrics, such as Maternal Mortality Rate (MMR) and Infant Mortality Rate (IMR), using independent variables related to various central government schemes.
    - Trained and validated these models to ensure they provided accurate and actionable predictions.
6. **Reporting and Documentation:**
    - Documented the methodologies, findings, and insights derived from the data analysis process.
    - Prepared detailed reports and presentations to effectively communicate the results to the research team and other stakeholders.

2. INTERNSHIP DETAILS

2.1 Tasks and responsibilities assigned

During my internship at the Web Science Lab (WSL) at IIIT Bangalore, I undertook several tasks and responsibilities aimed at analyzing and modeling government data to derive insights and predictive models. My work primarily focused on understanding the factors influencing Maternal Mortality Rate (MMR) using extensive datasets.

Initial Analysis and Data Compilation:

Objective: To analyze the Karnataka Data Lake and other relevant sources to identify key datasets for time series analysis related to MMR.

Responsibilities:

- Extracted district-wise data from the "Karnataka at a Glance" reports, covering various health indicators for four years.
- Key indicators included:
    - Number of First Referral Units
    - Number of 108 Ambulance Vehicles and Patients Benefited
    - Number of 24/7 Working Hospitals and Deliveries Occurred
    - Beneficiaries of Janani Suraksha Yojana
    - Beneficiaries Receiving Madilu Kit
    - Beneficiaries under the Supplementary Nutrition Programme
- Normalized these indicators relative to the population of each district.

Outcomes:

- Compiled a structured dataset suitable for analysis.

- Conducted a time series analysis but found the limited data insufficient for building a reliable predictive model.
- Linear regression analysis revealed that the number of patients benefited by ambulances negatively correlated with maternal mortality, while other variables showed no significant relationships.

Search for a Larger Dataset:

Objective: To find a more extensive dataset that could improve the robustness of the predictive models.

Responsibilities:

- Conducted an exhaustive search across multiple government websites, including MOSIP, RBI, CAG, Digital SANSAD, RSDEBATE, data.gov.in, etc.
- Compiled a 20-year dataset encompassing various independent variables, such as:
    - PMGSY (Pradhan Mantri Gram Sadak Yojana) for roads
    - PMAY (Pradhan Mantri Awas Yojana) for housing
    - Jal Jeevan Mission and National Rural Drinking Water Programme (NRDWP) for water supply
    - JSY (Janani Suraksha Yojana) and JSSY (Janani Shishu Suraksha Karyakram) for maternal health

Deliverables:

- Comprehensive datasets covering a span of 20 years.
- Detailed reports on data acquisition, normalization, and processing.
- Time series and regression analysis reports highlighting key insights and challenges.
- Predictive models and their performance analysis, providing actionable insights despite the data limitations.

Through these tasks and responsibilities, I gained extensive hands-on experience in data analysis, model building, and the practical challenges of working with government datasets. My work contributed to the broader objectives of the WSL by enhancing the lab's data repository and providing a foundation for future research aimed at improving public health policies.

2.2 Technologies/Methodologies used

During my internship at the Web Science Lab (WSL) at IIIT Bangalore, I utilized a range of technologies and methodologies to successfully accomplish the assigned tasks related to data analysis and predictive modeling. These tools and approaches enabled me to manage, process, and analyze extensive datasets effectively.

Technologies and Tools:

1. **Python:**
    - **Libraries:** Pandas, NumPy, Scikit-learn, Statsmodels, Fbprophet, Matplotlib, Seaborn
    - **Usage:** Python was the primary programming language used for data manipulation, analysis, and visualization. Libraries such as Pandas and NumPy facilitated efficient data handling and preprocessing. Scikit-learn was used for implementing regression models, Statsmodels for ARIMA models, and Fbprophet for time series forecasting. Visualization libraries Matplotlib and Seaborn helped in creating insightful graphs and plots.
2. **Excel:**
    - **Usage:** Microsoft Excel was used for initial data inspection, cleaning, and basic analysis. It was particularly useful for handling smaller datasets and performing preliminary checks before importing data into Python.
3. **Git:**
    - **Usage:** Git was used for version control, enabling collaborative work and efficient management of code changes and updates.
4. **Google Sheets:**
    - **Usage:** Google Sheets facilitated collaborative data entry and sharing among team members, allowing real-time updates and easy accessibility.

Methodologies:

1. **Data Collection and Preprocessing:**
    - **Data Sources:** Karnataka Data Lake, government websites (MOSIP, RBI, CAG, Digital SANSAD, RSDEBATE, data.gov.in), and reports ("Karnataka at a Glance").

- **Techniques:** Extracted data from PDFs and websites, cleaned and normalized datasets, handled missing values, and ensured data consistency across different sources.

2. **Exploratory Data Analysis (EDA):**
    - **Techniques:** Performed EDA to understand the distribution and relationships within the data. Used statistical summaries, visualizations (histograms, box plots, scatter plots), and correlation matrices to identify patterns and anomalies.
3. **Time Series Analysis:**
    - **Models Used:** ARIMA, Facebook's Prophet
    - **Techniques:** Conducted time series analysis to explore trends, seasonality, and temporal dependencies. ARIMA models were used for their effectiveness in handling time-dependent data, while Facebook's Prophet was employed for its robustness in dealing with seasonality and missing values.
4. **Regression Analysis:**
    - **Models Used:** Multi-Linear Regression
    - **Techniques:** Utilized regression models to examine relationships between multiple independent variables and the dependent variable (MMR). This included combining datasets over multiple years and performing statistical tests to validate the models.
5. **Predictive Modeling:**
    - **Techniques:** Developed and validated predictive models to forecast key metrics such as Maternal Mortality Rate (MMR) using various government schemes as independent variables. Iteratively refined models based on performance metrics and validation results.
6. **Reporting and Visualization:**
    - **Tools:** Matplotlib, Seaborn, Excel
    - **Techniques:** Created comprehensive reports and visualizations to communicate findings effectively. Used plots and charts to highlight key insights, trends, and correlations.

By leveraging these technologies and methodologies, I was able to perform thorough data analysis, develop predictive models, and provide actionable insights despite the challenges posed by data limitations. This integrated approach ensured that the tasks were completed efficiently and effectively, contributing to the broader objectives of the Web Science Lab.

2.3 Challenges and the Solutions

1. Insufficient and Inconsistent Data Availability:

Challenge: The data available on the Karnataka Data Lake was not suitable for time series analysis as the healthcare data was only available for certain years.

Solution:

- **Alternative Data Source:** I discovered an alternative dataset called "Karnataka at a Glance" through a colleague in the Web Science Lab. This dataset had healthcare data available for the years 2017-2022.
- **Data Extraction and Processing:** The data was in PDF format, with some sections in Kannada. I extracted the data from the PDFs into Excel and processed it to make it usable for analysis.

2. Data in Non-Digital and Non-English Format:

Challenge: Some of the data in the "Karnataka at a Glance" reports was in Kannada, making it difficult to understand and process.

Solution:

- **Language Translation:** Used translation tools and assistance from native Kannada speakers to translate the data into English.
- **Digital Conversion:** Converted the data from PDF to Excel using OCR (Optical Character Recognition) software, ensuring that the data was accurately captured and digitized.

3. Clarifying Roles and Responsibilities in Collaborative Tasks:

Challenge: Some tasks required collaboration with other teams, and initially, the roles and responsibilities of each team member were unclear.

Solution:

- **Engaging with Teams:** Increased engagement with other teams to better understand their roles and expertise.
- **Delegating Tasks:** As I became more familiar with team members' strengths, I delegated tasks more effectively, which improved overall efficiency and task completion.
- **Regular Communication:** Established regular communication channels to ensure everyone was on the same page, facilitating smoother collaboration.

3. SKILLS DEVELOPED

3.1 Details of the Technical Skills (TS)

During my internship at the Web Science Lab (WSL) at IIIT Bangalore, I developed robust technical skills essential for data analysis and modeling. I gained expertise in data extraction, cleaning, and normalization, handling diverse datasets extracted from PDFs and translating Kannada data into English for analysis. Utilizing advanced statistical methods like Multi-Linear Regression and time series models such as ARIMA and Prophet, I analyzed healthcare data to uncover insights despite challenges like data sparsity and policy changes. I also mastered data integration using SQL and Python, applying machine learning techniques to predict health indicators and creating clear visualizations with tools like Matplotlib and Seaborn to effectively communicate findings.

3.2 Details of the Soft Skills (SS)

In terms of soft skills, I enhanced problem-solving capabilities by finding alternative data sources and implementing solutions for incomplete datasets. Effective communication was crucial as I documented and presented findings, ensuring clarity for the research team and stakeholders. Time management skills were pivotal in handling multiple tasks efficiently, while collaborative teamwork allowed me to learn from and contribute to interdisciplinary projects. Additionally, my ability to work independently and take initiative enabled me to navigate challenges autonomously, demonstrating proactive problem-solving and self-directed project management throughout the internship.

4. KEY LEARNING FROM THE INTERNSHIP

During my internship at the Web Science Lab (WSL) at IIIT Bangalore, I gained valuable insights into practical data science and research. One of the key takeaways was mastering the intricacies of data analysis, from extracting and normalizing diverse datasets to applying statistical models like Multi-Linear Regression, ARIMA, and Prophet for forecasting. I learned firsthand the importance of meticulous data preprocessing and its impact on analysis outcomes, honing my technical skills in Python, SQL, and data manipulation techniques.

Navigating challenges such as data scarcity and policy-driven data variations taught me adaptive problem-solving skills. I developed strategies to find alternative data sources, handle missing data through imputation, and analyze temporal effects of policy changes. Collaborating within interdisciplinary teams enhanced my ability to communicate complex findings effectively and appreciate diverse viewpoints, fostering an environment of innovation and shared learning.

This internship underscored the dynamic nature of data science, emphasizing the need for continuous learning and adaptation. Staying updated with evolving methodologies and technologies became essential, reinforcing my proactive approach to research and problem-solving. Overall, this experience deepened my passion for leveraging data to inform policy decisions and improve public health initiatives, motivating me to pursue further studies in data science with a focus on societal impact.

5. CONCLUSION

My internship at the Web Science Lab (WSL) at IIIT Bangalore has been instrumental in shaping both my professional trajectory and personal growth. Working with real-world government datasets, I honed critical skills in data analysis, statistical modeling, and data management. This hands-on experience provided me with a deep understanding of data preprocessing, application of advanced analytical techniques such as Multi-Linear Regression, ARIMA, and Prophet models, and the nuances of interpreting results for actionable insights. These skills are foundational for my career aspirations in data science, particularly in driving evidence-based decision-making and policy formulation.

Collaborating within a multidisciplinary team at WSL expanded my communication and teamwork abilities. Engaging with diverse perspectives and contributing to collective goals underscored the importance of effective collaboration in achieving impactful outcomes. Personally, this internship reinforced my passion for using data-driven approaches to tackle societal challenges, particularly in improving healthcare and informing public policy. It deepened my appreciation for continuous learning and adaptability in the dynamic field of data science, motivating me to stay informed about emerging technologies and methodologies.

In conclusion, my internship at WSL has not only equipped me with technical skills and practical experience but also instilled a strong sense of purpose in leveraging data for societal good. It has prepared me to contribute meaningfully to the evolving landscape of data science, aspiring to drive positive change through rigorous analysis and innovative solutions. This transformative experience has laid a solid foundation for my future academic pursuits and career endeavors, inspiring me to pursue opportunities where I can apply my skills to make a tangible impact in the realm of public health and policy.

GLOSSARY

1) PMGSY - Pradhan Mantri Gram Sadak Yojana
2) JSY - Janani Suraksha Yojana
3) PMAY - Pradhan Matri Awas Yojana
4) JSSY - Janani Shishu Suraksha Karyakaram
5) ARIMA - Autoregressive Integrated Moving Average

BIBLIOGRAPHY

1) Ministry of Statistics and Programme Implementation. (2022). *Statistical Profile of Youth in India 2022*. Retrieved from https://www.mospi.gov.in/sites/default/files/publication_reports/Youth_in_India_2022/Statistical_Profile.pdf

2) West Bengal Health Department. (2006). *Infant Mortality Rate 2004-2006*. Retrieved from https://www.wbhealth.gov.in/other_files/2006/14_17.html

3) Press Information Bureau, Government of India. (2013). *Infant Mortality Rate 2009-2011*. Retrieved from https://pib.gov.in/newsite/PrintRelease.aspx?relid=98399

4) Ministry of Health and Family Welfare, Government of India. (2020). *State/UT-wise Details of Infant Mortality Rate (IMR) Sample Registration System (SRS) during 2020*. Retrieved from https://data.gov.in/resource/stateut-wise-details-infant-mortality-rate-sample-registration-system-srs-during-2020

5) Reserve Bank of India. (2019). *Maternal Mortality Rate (MMR) in Karnataka (2001-2019)*. Retrieved from https://www.rbi.org.in/scripts/PublicationsView.aspx?id=22076

6) National Health Mission, Karnataka. (2017). *Health Profile of Karnataka 2017-18*. Retrieved from https://nhm.karnataka.gov.in/Demography/Karnataka%20Health%20Profile%202017-18.pdf

7) Sansad Adarsh Gram Yojana. (2020). *Janani Suraksha Yojana (JSY) Data*. Retrieved from https://sansad.in/getFile/loksabhaquestions/annex/1711/AU296.pdf?source=pqals

8) Ministry of Rural Development, Government of India. (2021). *Pradhan Mantri Gram Sadak Yojana (PMGSY) Data*. Retrieved from https://omms.nic.in/

9) Ministry of Housing and Urban Affairs, Government of India. (2019). *Pradhan Mantri Awas Yojana (PMAY) Data*. Retrieved from https://rsdebate.nic.in/bitstream/123456789/591812/1/IQ_223_02082011_U301_p235_p241.pdf

10) Ministry of Health and Family Welfare, Government of India. (2019). *Janani Shishu Suraksha Karyakaram (JSSK) Data*. Retrieved from https://sansad.in/getFile/loksabhaquestions/annex/173/AU3327.pdf?source=pqals

11) Ministry of Women and Child Development, Government of India. (2022). *Pradhan Mantri Matru Vandana Yojana (PMMVY) Data*. Retrieved from https://sansad.in/getFile/loksabhaquestions/annex/1712/AU2734.pdf?source=pqals