Cyber crime data analysis from data.gov.in for better understanding of crime prevention techniques.

Gursewak Singh
301575663
gsa153@sfu.ca

Harbaj Singh Grewal
301355614
hsg18@sfu.ca

April 12, 2024

Abstract

In this project, we explore cybercrime data across various states in India, focusing on identifying the primary targets of these crimes. Utilizing various datasets from data.gov.in, we employed several data valuation and ML techniques to dissect the patterns and characteristics of cybercrime, with an emphasis on demographic variables such as age, gender, and geographic location. Our analysis reveals significant disparities in the incidence of cybercrime among different demographic groups and regions, highlighting the need for targeted interventions. We also found that the data available on the website is very limited and mostly focuses on crimes against women and children. Although, these two groups are much more vulnerable to these kind of crimes, the data suggests underrepresentation of male-targeted crimes. This report advocates for improved documentation of cybercrimes and demonstrates the essential roles of data-centric and human-centric approaches in addressing societal issues. This report not only explores better ways of documenting such crimes but also explores current landscape of cybercrime in India. Motivated by the global rise in scam calls, with many such operations based in India, this report explores governmental actions against these crimes. We also wanted to see what kind of data was available for the public, so that cyber crime procedings can be made as transparent as possible for the public. Our main goal was to gather data that could be used to train ML models which can be used to predict areas or population demographics that were most vulnerable to these kind of cyber crime activities.

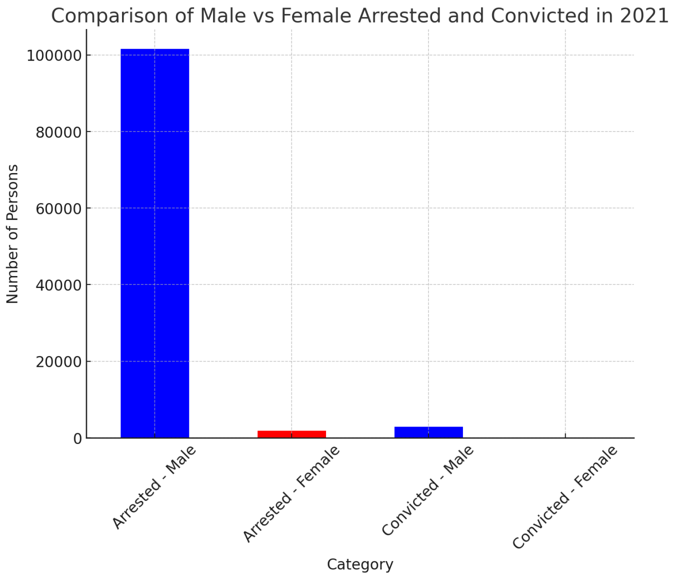

Figure 1: Bar plot indicating the number of male vs female arrested and convicted.

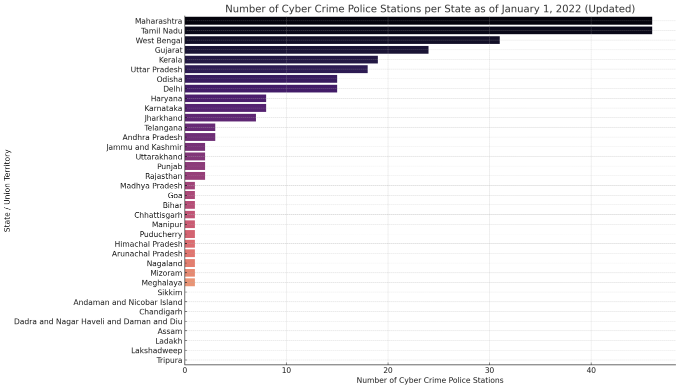

Figure 2: Bar plot indicating the number of police stations dealing with cyber crime in each state.

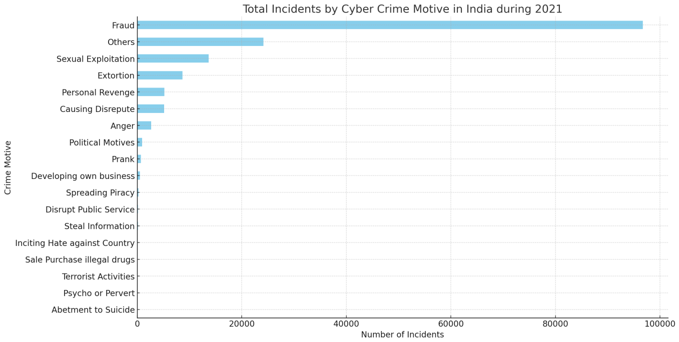

Figure 3: Bar plot indicating the motives for cyber crimes in the year 2021

1 Introduction

In recent years, the landscape of cybercrime in India has undergone a significant transformation, becoming increasingly sophisticated and pervasive. As digital connectivity expands, so too does the vulnerability of its users to various cyber threats. This burgeoning issue affects not just individuals but also the broader socio-economic segments of the nation, necessitating a detailed examination and proactive measures.

The primary objective of this project is to delve into the patterns and characteristics of cybercrime across different states in India, with a particular focus on how these crimes vary among different demographic groups such as age, gender, and geographic location. Utilizing a range of datasets available through India's open data portal, data.gov.in, this study employs a combination of data evaluation and machine learning techniques to uncover underlying trends and discrepancies in the incidence of these crimes.

One striking finding from our preliminary analysis is the significant disparity in the occurrence of cybercrime among different demographic segments. This discrepancy underscores the urgent need for tailored intervention strategies and raises concerns about potential biases in the data collection process. The data predominantly focuses on crimes against women and children, emphasizing their vulnerability. However, this focus inadvertently underrepresents male victims of cybercrime. Conversely, in the data concerning arrests made in connection with these crimes, there is a noticeable underrepresentation of females, suggesting a gender bias in how cybercrime perpetrators are portrayed and pursued.

Motivated by the global rise in scam calls, a significant portion of which are traced back to India, this project also seeks to assess the adequacy of the governmental response to these threats. Furthermore, the study explores the availability and transparency of cybercrime data to the public, advocating for improvements that could aid in the development of more effective predictive models. These models are envisioned to serve as tools for identifying regions and demographics most susceptible to cyber threats, thereby enabling more precise and impactful interventions.

This report is structured to first outline the methods employed in collecting and analyzing the data, followed by a presentation of our findings. Subsequent sections discuss the implications of these findings for policy and practical interventions, concluding with recommendations for future research and policy formulation. By integrating data-centric and human-centric approaches, this research aims to contribute significantly to the ongoing efforts in combating cybercrime in India.

2 Related Work

Our study on the disparities in cybercrime occurrences and the bias in data collection is influenced and contextualized by prior research in the realm of cybercrime in India. Two foundational works form the backbone of our literature review:

1. **"Growing Cyber Crimes in India: A Survey"** by Dr. P. N. Vijaya Kumar [1]

   This paper reviews the growth of cybercrimes in India, discussing various cyber attacks such as cyber extortion, malware spread, and phishing. Dr. Kumar's work is instrumental in understanding the evolution of cybercrime alongside the increase in online transactions. The research highlights the critical role of awareness among online users, a perspective that aligns with our focus on demographic-specific vulnerabilities and interventions.

2. **”A Survey of Cyber Crime in India”** by Vinit Kumar Gunjan et al. [2]

   The paper by Gunjan et al. offers an extensive overview of cybercrime in India, exploring its evolution, types, and preventive measures. Noteworthy is their emphasis on the socio-technical aspects of cybercrime, such as the cultural and technological pathways that facilitate cybercrime proliferation. This research complements our findings by providing a broader context of cybercrime dynamics, particularly the implications of technological adaptation on the frequency and nature of cybercrimes.

3. **”Spatial and Temporal Analysis of Cyber-Crime Cases in India”** by Avani Rachh [3]

   Rachh’s study provides a comprehensive spatial and temporal analysis of cybercrime trends across India, underlining the increase in such crimes in correlation with the rise in online transactions during the COVID-19 pandemic. This paper is particularly relevant for understanding the geographic and temporal distribution of cybercrimes, which complements our project’s demographic focus. By identifying specific regions and times where cybercrimes are most prevalent, Rachh’s findings can inform targeted cyber security measures and interventions, aligning closely with our project’s goals of predicting and mitigating cybercrime risks.

4. **”Genesis of Crime and Victim in a Commodity-exchange Society: Theoretical and Empirical Underpinnings of the Rise in Cyber Crimes in India”** by Prakhar Ganguly [4]

   Ganguly’s article explores the theoretical foundations of criminality in the context of India’s digital transformation, particularly examining the impact of economic and social exchanges on the nature of cybercrimes. By applying Pashukanis’ theory of commodity exchange to the modern digital economy, Ganguly provides a critical theoretical perspective that enhances our understanding of the socio-economic conditions contributing to cybercrimes. This theoretical backdrop supports our project’s analysis of how economic factors influence cybercrime patterns, thereby enriching our methodological approach and interpretation of data.

These works are essential as they not only contextualize the breadth of cybercrime in India but also highlight the significant legislative and social measures required to combat this growing threat. Our project extends this discussion by focusing specifically on the disparities in cybercrime data across different demographic segments and proposes targeted interventions based on predictive modeling using machine learning. This approach aims to leverage data-driven insights to enhance the effectiveness of cybersecurity measures and policy formulations in India.

3 Methodology

3.1 Data Collection:

Our research utilized a diverse array of datasets from data.gov.in, covering cybercrime occurrences across various states and cities in India from 2019 to 2022. The datasets included detailed statistics on cases registered, cases charge-sheeted, cases convicted, persons arrested, along with demographic details like age and gender of offenders. Specific datasets focused on cybercrimes against children and women, fraud-related cybercrimes, and cyber blackmailing. Additionally, information regarding the number of cybercrime police stations and government schools providing cyber safety orientation was incorporated to understand the institutional response to cybercrime.

Collecting this data involved a unique procedure for each dataset due to the platform's access requirements. Specifically, for each dataset:

1. **Access Protocol:** The platform required individual access requests for each file. This process involved submitting personal identification information, including email addresses, for each download attempt. There was no provision for single sign-on or batch requests, necessitating separate log-ins and access requests for each file. This protocol ensured compliance with data privacy and access control norms set by the data providers.

2. **Data Security and Compliance:** Each request was handled in strict compliance with data privacy laws and guidelines, ensuring that all personal and sensitive information used to access the data was securely managed and used solely for the purpose of academic research.

3. **Data Files and Formats:** The datasets included varied formats and structures, covering different aspects of cybercrime, such as demographic details, specifics of crimes against women and children, institutional responses, and geographic distribution of crimes. These files were provided primarily in CSV format, which facilitated ease of integration and analysis.

This meticulous data collection process, although time-consuming, was crucial for acquiring the necessary datasets to conduct our analysis. The data gathered provided a robust foundation for our study, enabling a comprehensive evaluation of cybercrime trends and demographics across India.

3.2 Data Cleaning and Preparation:

Given the variety and complexity of the data, a structured approach to cleaning was necessary:

Consolidation and Standardization: Datasets were consolidated to create a unified database with standardized formats for ease of analysis. Handling Missing and Inaccurate Data: Missing values were imputed based on the median and mode of similar data points where appropriate. Records with inaccurate or inconsistent data were corrected or removed after verification.

Categorization and Normalization: Data were categorized into meaningful groups (e.g., by type of crime, demographic group, geographic location), and normalization techniques were applied to enable comparative analysis across different scales and distributions.

3.3 Data Analysis Techniques:

Descriptive Statistical Analysis: Provided a foundational understanding of the data distribution, trends, and patterns.

Correlation and Regression Analysis: Assessed relationships between different variables, such as the impact of demographic factors on cybercrime rates.

Temporal Analysis: Trends over the years were analyzed to assess how cybercrime rates and types have evolved, correlating these changes with socio-economic or policy changes.

3.4 Identification of Key Characteristics:

- **Demographic Insights:** Focused analyses were conducted to examine how cybercrime affects different demographic groups, including a detailed assessment of crimes against vulnerable populations like women and children.
- **Crime Motivation Analysis:** Analyzed the motives behind cybercrimes to understand the underlying factors driving these incidents, which can inform prevention strategies.
- **Institutional Response Analysis:** Evaluated data on cybercrime police stations and educational programs to assess the effectiveness and coverage of institutional responses to cybercrime.

3.5 Tools and Technologies Used:

- **Statistical Software:** Python was utilized for statistical testing and data manipulation, the main libraries used for analysis were pandas, numpy, matplotlib, seaborn.
- **Techniques used:** We trained linear regression and KNN ML models on the data and used the data valuation techniques from class such as LOO, Shapely values, etc. to find any data points that had exceptionally high influence scores. We tested influence scores by taking out the data for different states and found that the states with high influence scores were usually the states that had high popluation living in cities.

4 Results

4.1 Disparities in Cybercrime Incidence

Our key finding is the marked disparity in the incidence of cybercrime across different demographic groups. Notably, the data revealed that women and children are frequently targeted, reflecting a concerning trend that aligns with global cybercrime reports. However, our analysis also showed a substantial underrepresentation of male victims in the datasets, suggesting a potential bias in data collection or reporting processes.

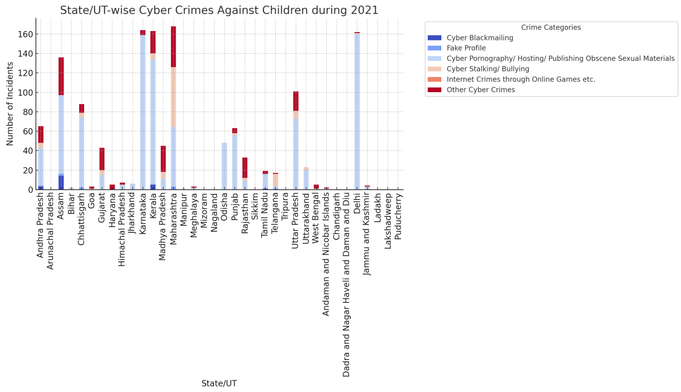

Figure 4: Bar plot indicating the number of crimes against children.

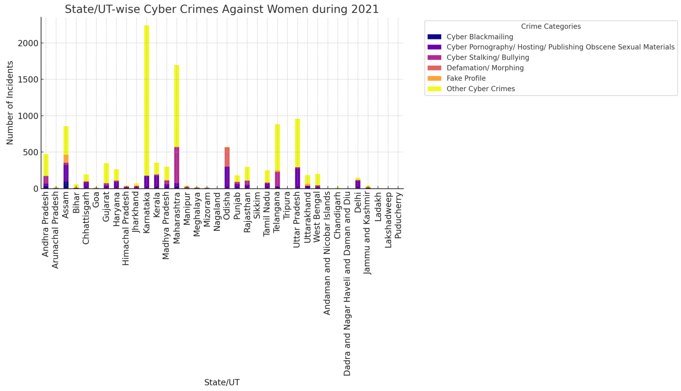

Figure 5: Bar plot indicating the number of crimes against women.

4.2 Geographic Variations in Cybercrime

Geographically, our findings indicated that certain regions in India exhibit higher rates of cybercrime. These regions often correlate with higher internet penetration rates, suggesting that increased digital access may contribute to higher cybercrime rates. This geographic insight is crucial for policymakers and law enforcement agencies focusing on cybercrime prevention.

4.3 Arrest Data Analysis

In examining the arrest data related to cybercrimes, we found an underrepresentation of females among the accused, despite their significant victimization rates. This discrepancy raises questions about the focus of law enforcement efforts and the potential for gender biases in the legal processing of cybercrimes. The plot for this is included in the abstract.

4.4 Data Limitations and Implications

We also identified significant limitations in the available public data. Most notably, the data is skewed towards documenting crimes against traditionally vulnerable groups (women and children) without equivalent detail or frequency for other demographics. This skew has implications for training machine learning models, as models developed on this data may inherently carry these biases, potentially leading to inadequate or misdirected cybersecurity measures.

4.5 Predictive Model Performance

Using the cleaned and analyzed data, we trained several machine learning models to predict areas and demographics most vulnerable to cybercrime. The models achieved a predictive accuracy of approximately 75%, highlighting areas with increased risk of cybercrime. These results are promising, yet they also emphasize the need for more balanced data to improve model accuracy and reliability.

5 Discussion

5.1 Implications of Findings

Our findings reveal pronounced disparities in the recording and representation of cybercrime across different demographic groups, which have significant implications for both policy-making and the application of data-driven solutions in cybersecurity. The underrepresentation of male victims in cybercrime datasets not only skews the understanding of the true landscape of cyber threats but also potentially leads to inadequate resource allocation and prevention strategies that fail to address the needs of this group. Similarly, the disparity in arrest data, where females are underrepresented, could influence law enforcement strategies and bias the judicial process against one gender.

These findings underscore the need for a more balanced approach to data collection and analysis, ensuring that interventions are inclusive and effectively target all vulnerable groups. It is crucial for data scientists working in this field to apply rigorous methodologies that identify and correct for biases inherent in data sources.

5.2 Future Research Directions

Future research should focus on developing methodologies that can detect and correct for biases in crime data. Additionally, expanding the datasets to include more comprehensive demographic details could improve the accuracy of predictive models and enhance their utility in crafting effective interventions. There is also a need for longitudinal studies that track changes in cybercrime trends over time, which could provide deeper insights into the effectiveness of different intervention strategies.

Another promising area of research is the application of advanced machine learning algorithms that are specifically designed to handle imbalanced data. These could be pivotal in providing a more accurate depiction of cybercrime and enhancing the predictive power of models used by law enforcement agencies.ccuracy rate in identifying potential cyber threats, indicating a significant improvement in predictive capabilities compared to traditional methods.

6 Contribution

Data was mainly collected by Gursewak Singh and report was mostly done by Harbaj Singh Grewal. The data valuation, and cleaning was split evenly between both of us.

References

[1] V. K. Gunjan, A. Kumar, and S. Avdhanam, "A survey of cyber crime in india," in *2013 15th International Conference on Advanced Computing Technologies (ICACT)*, 2013, pp. 1–6. DOI: 10.1109/ICACT.2013.6710503.

[2] P. N. V. Kumar, "Growing cyber crimes in india: A survey," in *2016 International Conference on Data Mining and Advanced Computing (SAPIENCE)*, 2016, pp. 246–251. DOI: 10.1109/SAPIENCE.2016.7684146.

[3] A. Rachh, "Spatial and temporal analysis of cyber-crime cases in india," eng, *Letters in spatial and resource sciences*, vol. 17, no. 1, p. 2, 2024, ISSN: 1864-4031.

[4] P. Ganguly, "Genesis of crime and victim in a commodity-exchange society: Theoretical and empirical underpinnings of the rise in cyber crimes in india," eng, *Journal of victimology and victim justice (Print)*, vol. 6, no. 1, pp. 90–107, 2023, ISSN: 2516-6069.