CA1 Mini Project Report: Exploratory Data Analysis of MGNREGA Data

Name: Aryan Paratakke

PRN: 22070521070

Batch: 2022-26

Semester: 7th

College: Symbiosis Institute of Technology (SIT), Nagpur

Subject: Machine Learning (CA1)

Instructor: Dr. Piyush Chauhan

1. Introduction: Leveraging Machine Learning for Social Impact

This report documents the Exploratory Data Analysis (EDA) performed on the MGNREGA dataset as part of the Machine Learning CA1 assessment. The primary objective is to analyze the implementation and impact of the Mahatma Gandhi National Rural Employment Guarantee Act (MGNREGA) across various states and districts in India. The analysis aims to uncover trends, patterns, and insights related to employment generation, financial expenditure, and administrative efficiency under the scheme.

The chosen dataset is a powerful tool for this project, as it provides real-world data from a government source. By applying data cleaning and machine learning techniques, we can move beyond simple observation to derive actionable insights that could help improve the scheme's social impact.

2. Dataset Overview

2.1 Source and Purpose

Source: The data was obtained from the official Government of India's Open Government Data (OGD) platform, specifically the API endpoint for "District-wise MGNREGA Data at a Glance" (data.gov.in).

Purpose: The dataset contains detailed, monthly metrics on the MGNREGA scheme's implementation across different districts and states of India. It is ideal for analyzing regional disparities, seasonal trends, and the effectiveness of the program.

2.2 Initial Data State

The dataset, as it was initially loaded, contained 302,753 records and 36 columns. An initial inspection showed a mixture of data types, with many numerical fields improperly stored as object (string) types due to inconsistencies in the source data. A summary of the initial state is provided below to illustrate the starting point of our data cleaning process.

Summary of Initial DataFrame:

--- DataFrame Info after initial load ---

RangeIndex: 302753 entries, 0 to 302752 Data columns (total 36 columns): No Column Non-Null Count Dtype

0 fin_year 302752 non-null object 1 month 302752 non-null object 2 state_code 302752 non-null float64 3 State 302752 non-null object 4 district_code 302752 non-null float64 5 District 302752 non-null object 6 Approved_Labour_Budget 302752 non-null float64

7 Average_Wage_rate_per_day_per_person 302752 non-null float64 8 Average_days_of_employment_provided_per_Household 302752 non-null float64 9 Differently_abled_persons_worked 302752 non-null float64 10 Material_and_skilled_Wages 302752 non-null float64 11 Number_of_Completed_Works 302752 non-null float64 12 Number_of_GPs_with_NIL_exp 302752 non-null float64 13 Number_of_Ongoing_Works 302752 non-null float64 14

Persondays_of_Central_Liability_so_far 302752 non-null float64 15 SC_persondays 302752 non-null float64 16 SC_workers_against_active_workers 302752 non-null float64 17 ST_persondays 302752 non-null float64

18 ST_workers_against_active_workers 302752 non-null float64 19 Total_Adm_Expenditure 302752 non-null float64 20 Total_Exp 302752 non-null float64 21 Total_Households_Worked 302752 non-null float64 22 Total_Individuals_Worked 302752 non-null float64 23 Total_No_of_Active_Job_Cards 302752 non-null float64 24 Total_No_of_Active_Workers 302752 non-null float64 25 Total_No_of_HHs_completed_100_Days_of_Wage_Employment 302752 non-null float64 26 Total_No_of_JobCards_issued 302752 non-null float64 27 Total_No_of_Workers 302752 non-null float64 28 Total_No_of_Works_Takenup 302752 non-null float64 29 Wages 302752 non-null float64 30 Women_Persondays 302752 non-null float64 31 percent_of_Category_B_Works 302752 non-null float64 32 percent_of_Expenditure_on_Agriculture_Allied_Works 302752 non-null float64 33 percent_of_NRM_Expenditure 302752 non-null float64 34 percentage_payments_gererated_within_15_days 302752 non-null float64 35 Remarks 1 non-null object dtypes: float64(31), object(5) memory usage: 83.2+ MB

--- First 4 rows of the DataFrame ---

fin_year	month	state_code	State	district_code	District	Approved_Labour_Budget	Average_Wage_rate_per_day_per_person	Average_days_of_employment_provided_per_Household	Differently_abled_persons_worked	Material_and_skilled_Wages	Number_of_Completed_Works	Number_of_GPs_with_NIL_exp	Number_of_Ongoing_Works	Persondays_of_Central_Liability_so_far	SC_persondays	SC_workers_against_active_workers	ST_persondays	ST_workers_against_active_workers	T e
2019-2020	June	35	UTTARAKHAND	3506	RUDRA PRAYAG	323294	181.7353366	27	8	6.2480323	375	8	4047	193779	35244	14078	182	14	
2019-2020	June	35	UTTARAKHAND	3508	NAINITAL	252505	177.6638997	29	21	18.6859487	925	64	4835	307676	59778	16595	859	372	
2019-2020	June	35	UTTARAKHAND	3512	BAGESHWAR	241752	171.565915	25	57	0	501	74	2027	176189	42992	19052	961	395	
2019-2020	June	37	LADAKH	3707	LEH (LADAKH)	0	85814.19344	10	0	122.118293	297	2	1031	701	0	0	701	34337	

[4 rows x 36 columns]

Key Observations from Initial Load: * The dataset contains `302,753` entries and `36` columns. * Many columns, such as `fin_year`, `month`, `State`, `District`, and `Remarks`, are of `object` (string) data type. This indicates potential inconsistencies in data entry, such as mixed month name abbreviations or leading/trailing whitespace. * `Remarks` is a highly sparse column with only 1 non-null value, suggesting it is not a useful feature for analysis and should be handled appropriately. * Key numerical codes (`state_code`, `district_code`) are `float64` and contain a small number of `NaN` values. This requires conversion to a proper integer type for clean categorical representation. * The `Average_Wage_rate_per_day_per_person` and `percentage_payments_gererated_within_15_days` columns, while initially loaded as `float64`, contained extreme, physically and mathematically impossible outliers that needed specific treatment.

2.3 Column Details (After Cleaning and Feature Engineering)

Below is a summary of the key columns and their final state after our comprehensive data cleaning and feature engineering process. These data types and metrics form the basis of all subsequent analysis.

Column Name	Final Data Type	Descriptio & Significanc
`fin_year`	`object`	The financia year.
`month`	`object`	The calenda month.
`state_code`	`int64`	Numerical code for the state.
`State`	`object`	Name of th state.
`district_code`	`int64`	Numerical code for the district.
`District`	`object`	Name of th district.
`Approved_Labour_Budget`	`float64`	The total approved budget.

Column Name	Final Data Type	Descriptio & Significan
Average_Wage_rate_per_day_per_person	float64	Average da wage rate. **(Critical social indicator)**
Average_days_of_employment_provided_per_Household	float64	The averag number of employmen days provid per household.
Differently_abled_persons_worked	float64	Number of differently-abled perso who availed work.
Material_and_skilled_Wages	float64	Expenditure on material and skilled wages.
Number_of_Completed_Works	float64	Total numb of works completed.
Number_of_GPs_with_NIL_exp	float64	Number of Gram Panchayats with zero expenditure
Number_of_Ongoing_Works	float64	

Column Name	Final Data Type	Descriptio & Significanc
		Number of ongoing works.
Persondays_of_Central_Liability_so_far	float64	Total personday generated under centr liability.
SC_persondays	float64	Persondays generated f Scheduled Castes.
SC_workers_against_active_workers	float64	Ratio of SC workers to active workers.
ST_persondays	float64	Persondays generated f Scheduled Tribes.
ST_workers_against_active_workers	float64	Ratio of ST workers to active workers.
Total_Adm_Expenditure	float64	Total administrat expenditure
Total_Exp	float64	Total overal expenditure

Column Name	Final Data Type	Descriptio & Significanc
`Total_Households_Worked`	`float64`	Total households that workec
`Total_Individuals_Worked`	`float64`	Total individuals who workec
`Total_No_of_Active_Job_Cards`	`float64`	Number of active job cards.
`Total_No_of_Active_Workers`	`float64`	Number of active workers.
`Total_No_of_HHs_completed_100_Days_of_Wage_Employment`	`float64`	Households that completed 100 days of wage employmen **(Crucial social impact metric)**
`Total_No_of_JobCards_issued`	`float64`	Total job cards issue
`Total_No_of_Workers`	`float64`	Total worke registered.
`Total_No_of_Works_Takenup`	`float64`	Total works taken up.

Column Name	Final Data Type	Descriptio & Significanc
Wages	float64	Total wages paid.
Women_Persondays	float64	Persondays generated b women.
percent_of_Category_B_Works	float64	Percentage Category B works.
percent_of_Expenditure_on_Agriculture_Allied_Works	float64	Percentage expenditure on agricultu and allied works.
percent_of_NRM_Expenditure	float64	Percentage expenditure on Natural Resource Managemer works.
percentage_payments_gererated_within_15_days	float64	Percentage payments generated within 15 days. **(Key efficiency indicator)**
Remarks	object	Free text remarks

Column Name	Final Data Type	Description & Significance
		(mostly nul and droppe
Women_Persondays_Ratio	float64	**Engineere Feature:** Proportion o total persondays generated b women.
SC_Persondays_Ratio	float64	**Engineere Feature:** Proportion o total persondays generated b SC individuals.
ST_Persondays_Ratio	float64	**Engineere Feature:** Proportion o total persondays generated b ST individuals.
100_Days_HH_Ratio	float64	**Engineere Feature:** Ratio of households completing 100 days to

Column Name	Final Data Type	Descriptio & Significanc
		total households worked.

dtypes: float64(31), object(5)

memory usage: 83.2+ MB

3. Data Cleaning & Preprocessing (ETL)

The initial dataset, as sourced from the government API, was in a raw state that required a robust ETL (Extract, Transform, Load) pipeline to ensure data integrity and suitability for analysis. This process was critical for addressing inconsistencies, outliers, and preparing the data for meaningful insights.

3.1 Robust Data Type Conversion and Outlier Handling

The raw data presented several challenges, including columns with incorrect data types and the presence of erroneous outliers. A systematic approach was implemented to correct these issues:

- **Initial Data Types**: The initial data contained columns with a mix of data types. Specifically, a large number of numerical columns were incorrectly loaded as object (string) types. This required a programmatic approach to convert them.

- **Extreme Outliers**: Key financial and performance metrics showed physically or mathematically impossible values. For example, the Average_Wage_rate_per_day_per_person column contained outliers in the tens of millions, and percentage_payments_gererated_within_15_days had values far exceeding 100%.

Handling Strategy:

1. Systematic Numerical Conversion: All numerical columns were explicitly converted to float64 using pd.to_numeric(errors='coerce') to handle any non-numeric entries gracefully by converting them to NaN.

2. Imputation of NaN and inf Values: np.inf values resulting from division by zero were converted to NaN. All NaN values were then filled with 0, based on the assumption that for these metrics, a missing value represents zero activity.

3. Targeted Outlier Treatment:

    - `Average_Wage_rate_per_day_per_person`: Values were capped at a plausible upper limit (₹5,000). Zero values (where an average wage is illogical) were replaced with the median of the valid wage distribution. This corrected the extreme outliers while preserving the integrity of the data.
    - `percentage_payments_gererated_within_15_days`: This metric was clipped to a valid range of [0, 100] to ensure mathematical correctness.

3.2 Temporal Data Alignment and Sorting

Accurate time-series analysis requires data to be correctly aligned with the financial year. The Indian financial year, running from April to March, necessitated a custom sorting approach.

1. Financial Year Month Ordering: A custom ordered categorical data type was created for the month column, explicitly defining the sequence from 'April' to 'March'. This ensured all monthly trend visualizations would be chronologically accurate.

2. Multi-level Sorting Hierarchy: The entire DataFrame was sorted according to a strict hierarchy:

    - fin_year (ascending)
    - month (using the custom financial year order, ascending)
    - state_code (ascending)
    - district_code (ascending)

This sorting process ensured that all subsequent analyses, from yearly trends to geospatial comparisons, were based on a perfectly ordered dataset.

Sample of Month Mapping to Sort Key:

This table demonstrates the successful mapping of month strings to their numerical financial year order, which enabled correct sorting.

--- Sample of Month Mapping to Sort Key --- | month | month_sort_key | | :-------- | :------------- | | April | 0 | | May | 1 | | June | 2 | | July | 3 | | Aug | 4 | | August | 4 | | Sep | 5 | | September | 5 | | Oct | 6 | | October | 6 | | Nov | 7 | | November | 7 | | Dec | 8 | | December | 8 | | Jan | 9 | | January | 9 | | Feb | 10 | | February | 10 | | Mar | 11 | | March | 11 |

DataFrame Sorted Successfully (Head showing multi-level sort):

--- DataFrame sorted successfully. --- fin_year month state_code State district_code District 0 2018-2019 April 1 ANDAMAN AND NICOBAR 101 SOUTH ANDAMAN 1 2018-2019 April 1 ANDAMAN AND NICOBAR 102 NICOBARS 2 2018-2019 April 1 ANDAMAN AND NICOBAR 103 NORTH ANDAMAN 3 2018-2019 April 2 ANDHRA PRADESH 201 SRIKAKULAM 4 2018-2019 April 2 ANDHRA PRADESH 202 VIZIANAGARAM 5 2018-2019 April 2 ANDHRA PRADESH 203 VISAKHAPATNAM 6 2018-2019 April 2 ANDHRA PRADESH 204 EAST GODAVARI 7 2018-2019 April 2 ANDHRA PRADESH 205 WEST GODAVARI 8 2018-2019 April 2 ANDHRA PRADESH 206 KRISHNA 9 2018-2019 April 2 ANDHRA PRADESH 207 GUNTUR 10 2018-2019 April 2 ANDHRA PRADESH 208 PRAKASAM 11 2018-2019 April 2 ANDHRA PRADESH 209 NELLORE 12 2018-2019 April 2 ANDHRA PRADESH 210 CHITTOOR 13 2018-2019 April 2 ANDHRA PRADESH 211 ANANTAPUR 14 2018-2019 April 2 ANDHRA PRADESH 212 KADAPA 15 2018-2019 April 2 ANDHRA PRADESH 213 KURNOOL 16 2018-2019 April 3 ASSAM 301 BARPETA 17 2018-2019 April 3 ASSAM 302 CACHAR 18 2018-2019 April 3 ASSAM 303 DHEMAJI 19 2018-2019 April 3 ASSAM 304 DIBRUGARH

3.3 Anomaly Investigation and Filtering

A critical discovery during the ETL phase was a significant anomaly in the data for the 2024-2025 financial year. This year showed disproportionately high metrics, which upon investigation, was found to be due to a change in the data's reporting granularity, not a genuine surge in activity. The 2025-2026 financial year was also excluded as the data was incomplete.

To ensure our analysis was based on genuinely comparable historical trends, a professional decision was made to filter out both the '2024-2025' and '2025-2026' financial years.

DataFrame Shape After Filtering Anomalous Years:

--- Filtering Data for Consistent Historical Analysis (Excluding 2024-2025 and 2025-2026) --- DataFrame shape after filtering ['2024-2025', '2025-2026']: (50892, 36)

Financial Years Remaining in Data:

Financial years remaining in data: ['2018-2019' '2019-2020' '2020-2021' '2021-2022' '2022-2023' '2023-2024'] Categories (6, object): ['2018-2019', '2019-2020', '2020-2021', '2021-2022', '2022-2023', '2023-2024']

3.4 Feature Engineering

To gain deeper insights, several ratio-based features were meticulously engineered. These normalized metrics are vital for fair comparisons across states of different sizes and over time.

Women_Persondays_Ratio: Proportion of total employment generated by women.

SC_Persondays_Ratio / ST_Persondays_Ratio: Proportion of employment for marginalized groups.

100_Days_HH_Ratio: Ratio of households completing 100 days of work to total households that worked, a key indicator of scheme effectiveness.