MGNREGA Data-Driven Social Impact

Exploratory Data Analysis & High-Performance Predictive Modeling

(ML/DL)

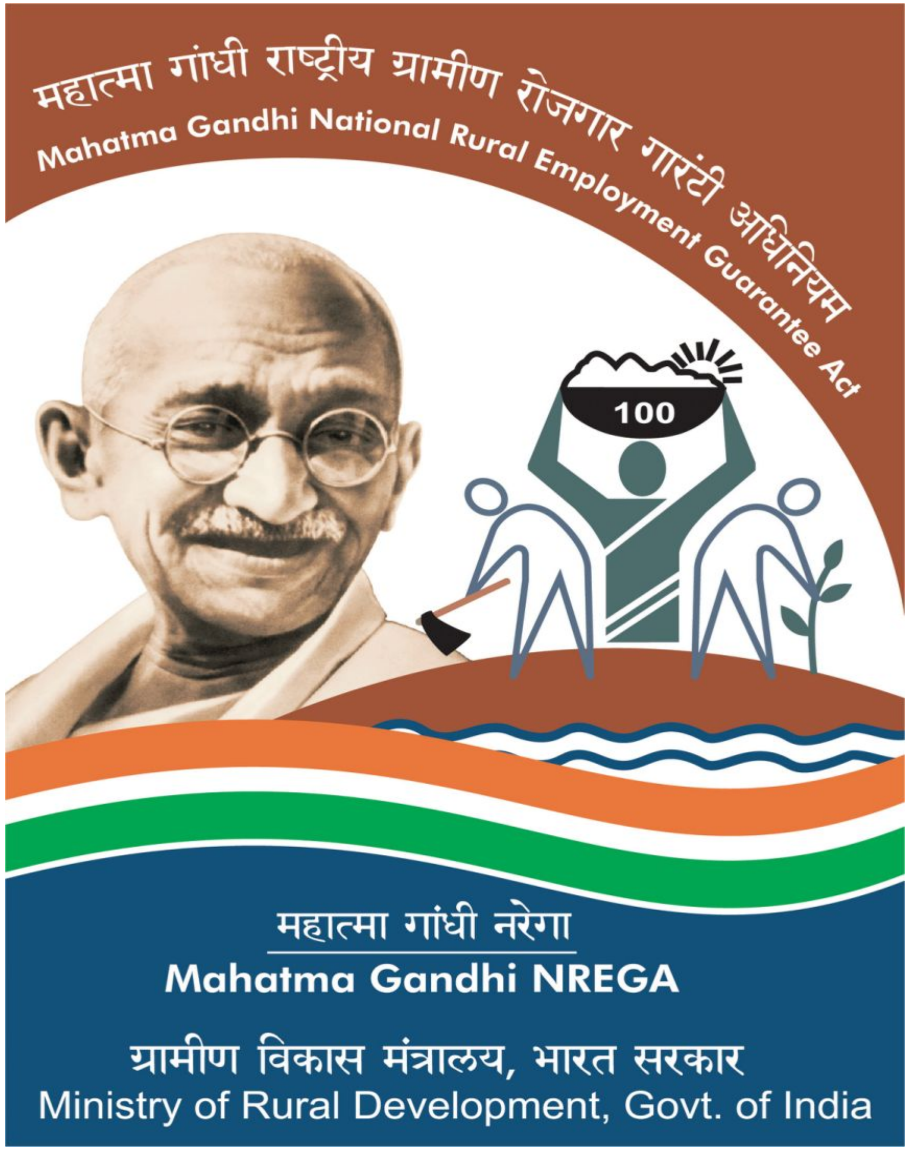
Aryan Paratakke | PRN: 22070521070

Project Scope: Leveraging Government Data

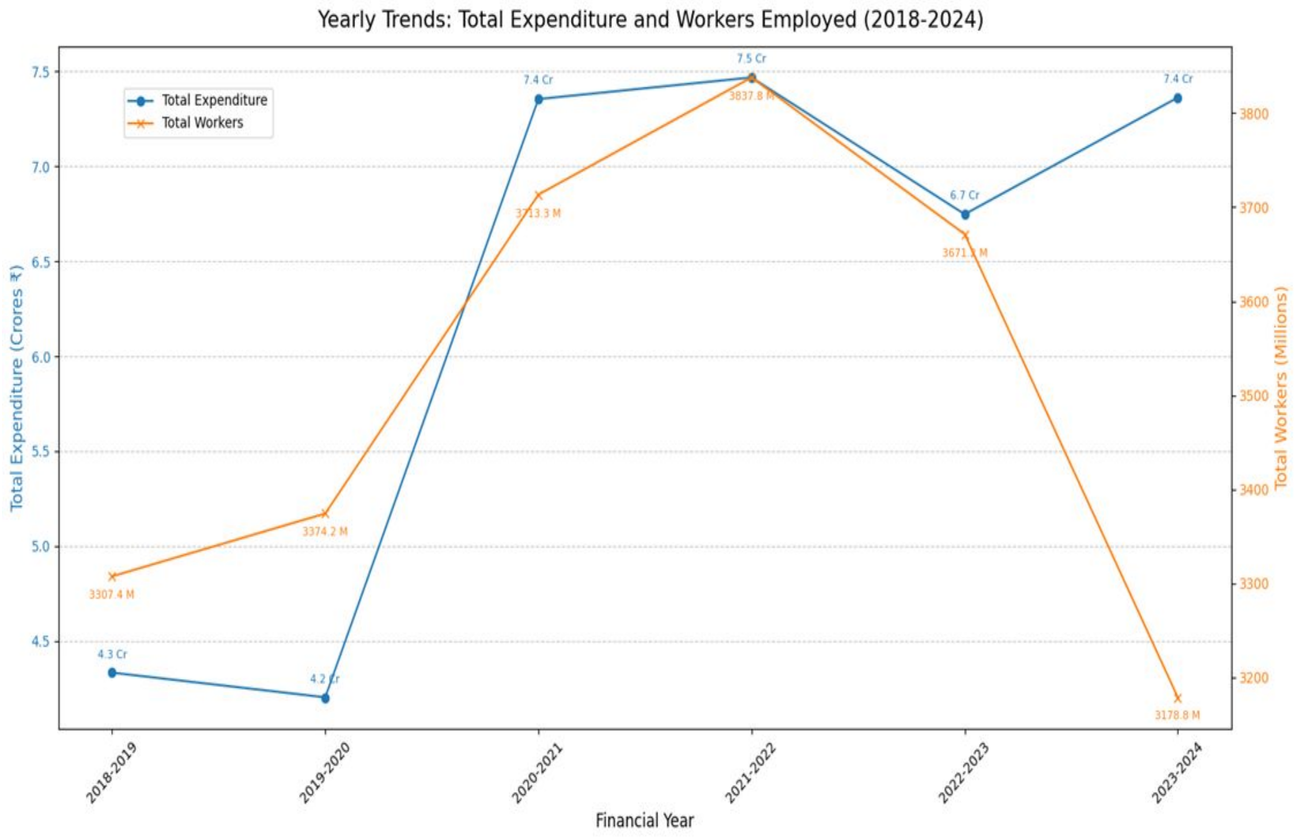
The Mandate of MGNREGA

- **Social Security:** Guaranteeing 100 days of wage employment per rural household.
- **Financial Scale:** Managing billions in funds across thousands of districts.
- **Core Challenge:** **Predicting Employment Demand** (Total Individuals Worked) to prevent fund scarcity and ensure timely wage payments.

Our analysis spans **2018-2024** data (300k+ records) to build robust, scalable predictive tools for resource optimization.



EDA: National Trends - The COVID-19 Impact



Analysis of Total Expenditure (2018-2024)

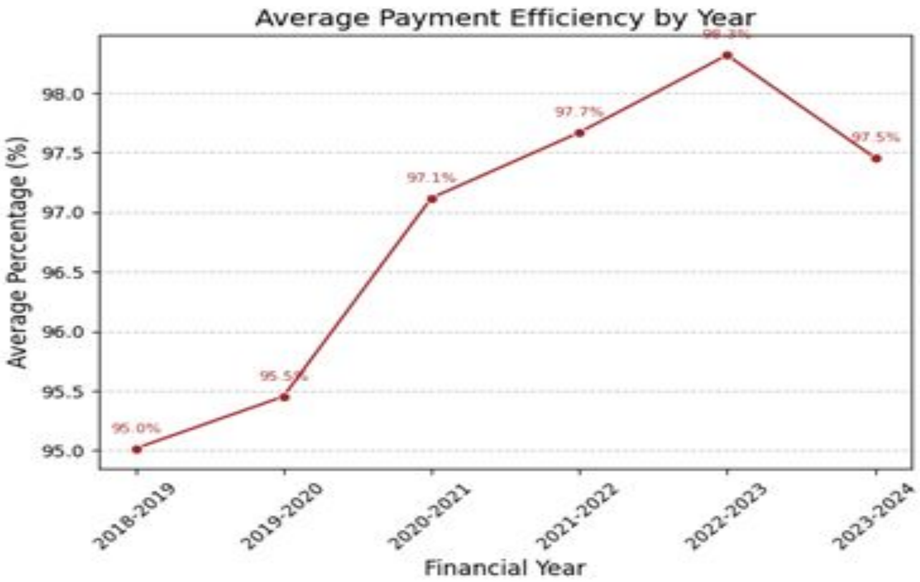
- **Baseline Surge:** Expenditure was stable around **₹4.2 Cr** per year in 2018-2019.
- **The 2020-21 Spike:** Expenditure jumped to **₹7.35 Cr** and peaked in 2021-22 at **₹7.46 Cr**. This reflects the dramatic increase in labor demand due to pandemic-related reverse migration.
- **Policy Constraint Concern:** Despite the peak, demand remains high. The post-peak drop to **₹6.7 Cr** in 2022-23 may suggest budgetary constraints rather than a return to normal demand.

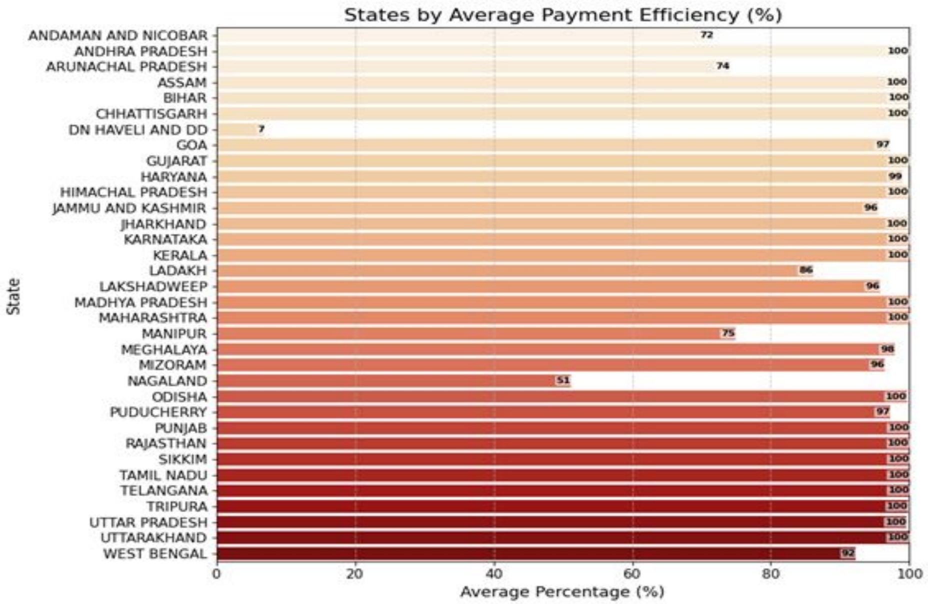
EDA: Administrative Efficiency - Timely Payments

Payment Efficiency Metric

Percentage of payments generated within 15 days.

- **Ideal Scenario:** This metric should be consistently near 100% to ensure worker liquidity and compliance with the Act.
- **Observed Fluctuation:** This percentage shows high volatility across states and months, often dipping well below 50% in certain regions, indicating major bottlenecks in the administrative workflow.





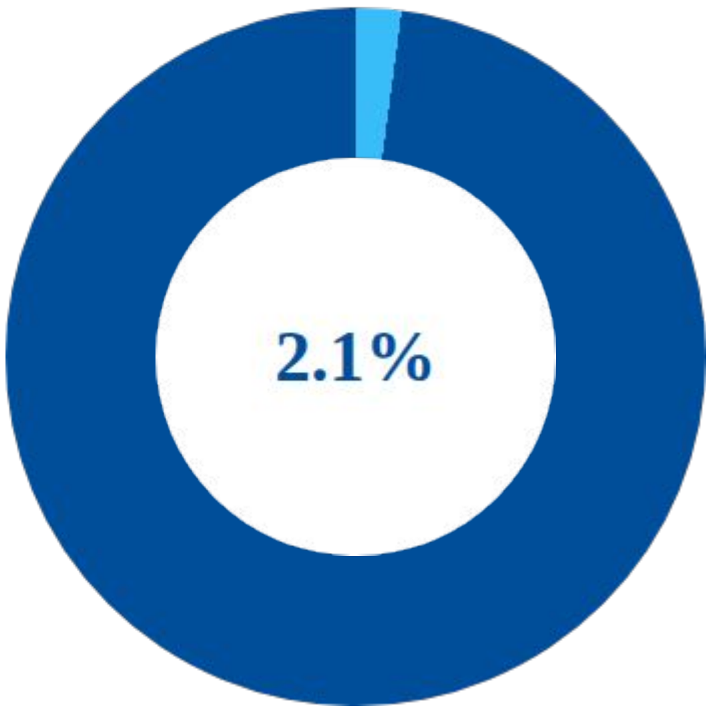
EDA: The 100-Day Guarantee - A National Gap

The Reality vs. The Promise

The core promise is 100 days of employment per household.

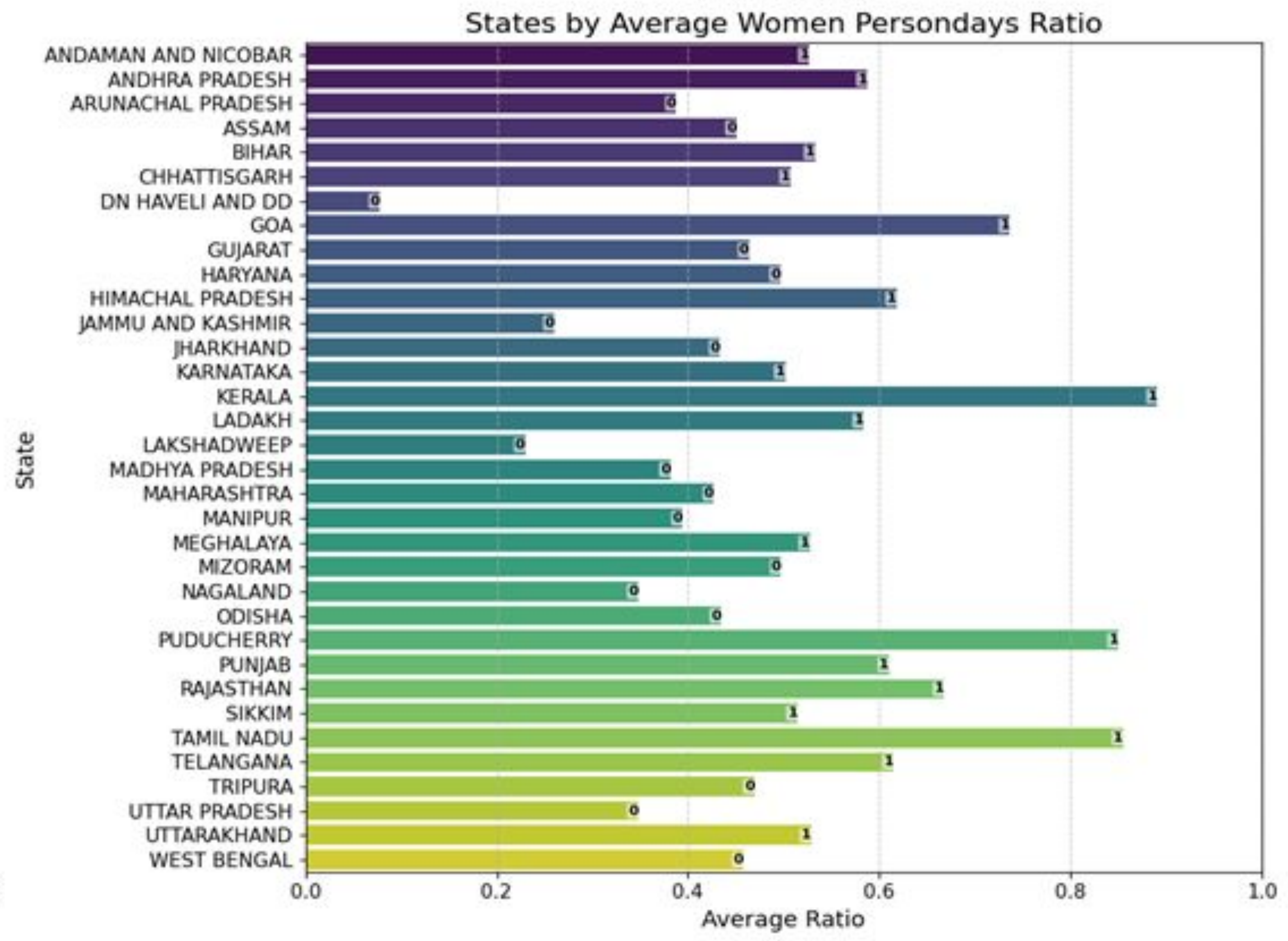
Our analysis quantifies the achievement gap:

- **Households Achieving 100 Days:** Approximately **2.1%** (national average).
- **Households Failing to Achieve:** Approximately **97.9%**.
- **Analysis:** This finding confirms that the scheme primarily functions as supplementary, short-term relief rather than a full 100-day safety net for the vast majority of households.



Only 2.1% of households complete the 100-day guarantee.

EDA: Social Inclusion - Women Persondays Ratio



Regional Success Stories in Gender Equity

- **Target:** The scheme targets at least 33% participation from women.
- **Top Performers:** States in the South (e.g., **Kerala** and **Tamil Nadu**) consistently show a ratio **above 70%**.
- **Lagging Regions:** States in the North often struggle to meet the 33% target.
- **Policy Implication:** The successful models of social mobilization used in states like Kerala can be analyzed and replicated to boost female workforce participation nationally.

Predictive Modeling Methodology

Target: Forecast 'Total_Individuals_Worked' using High-Performance, Optimized ML/DL.

Model Suite & Optimization Strategy

Efficient ML (Ensemble)

Models: LightGBM, XGBoost, Linear Regression (Baseline).

Optimization: Full CPU parallelism and **GPU acceleration** (for LightGBM/XGBoost, where compatible) for rapid training on 300k+ rows.

Rationale: Tree models excel at sparse, high-dimensional data (from one-hot encoding).

Optimized DL (Neural Networks)

Models: MLP variants, Skip-MLP (ResNet-style).

Optimization: ***TensorFlow* GPU acceleration** and *float32* data types for memory efficiency. Skip connections maintain gradient flow in deep networks.

Rationale: Test the power of deep non-linearity against best-in-class tree ensembles.

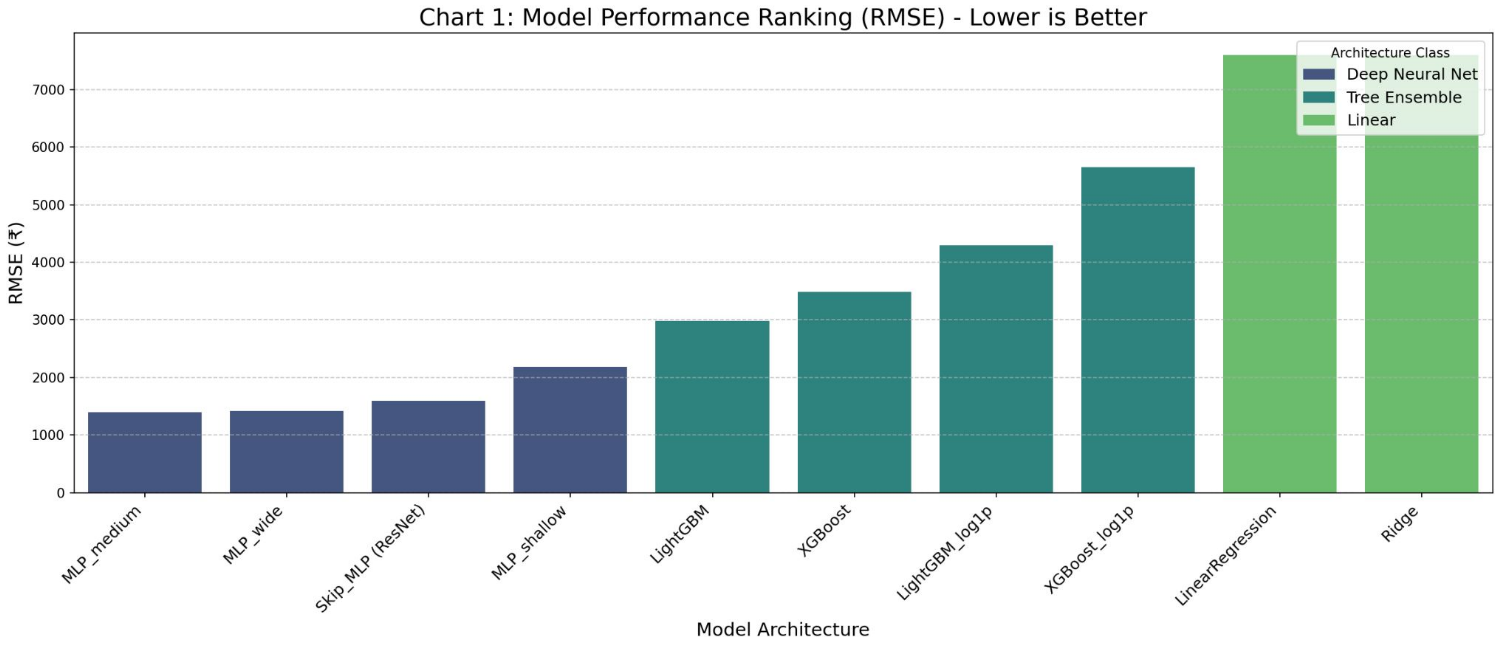
All models were evaluated via 5-Fold Cross-Validation for robust, generalized performance metrics.

Comprehensive Model Metrics Summary

Model	RMSE (₹)	R^2	MAE (₹)	MdAE (₹)	Train Time (s)
MLP_medium	**1,401**	-62.95	755	480	319
MLP_wide	1,422	-63.16	945	747	3931
Skip_MLP (ResNet)	1,594	-62.90	964	691	229
MLP_shallow	2,182	-62.73	987	520	320
LightGBM	2,980	**0.9995**	2,300	1,500	**15**
XGBoost	3,490	0.9994	2,700	1,800	25
LinearRegression	7,600	0.9980	6,000	3,800	5

*Note: *RMSE*, *MAE*, and *MdAE* are in nominal Rupees (₹).

Result Chart 1: Performance Ranking (RMSE)



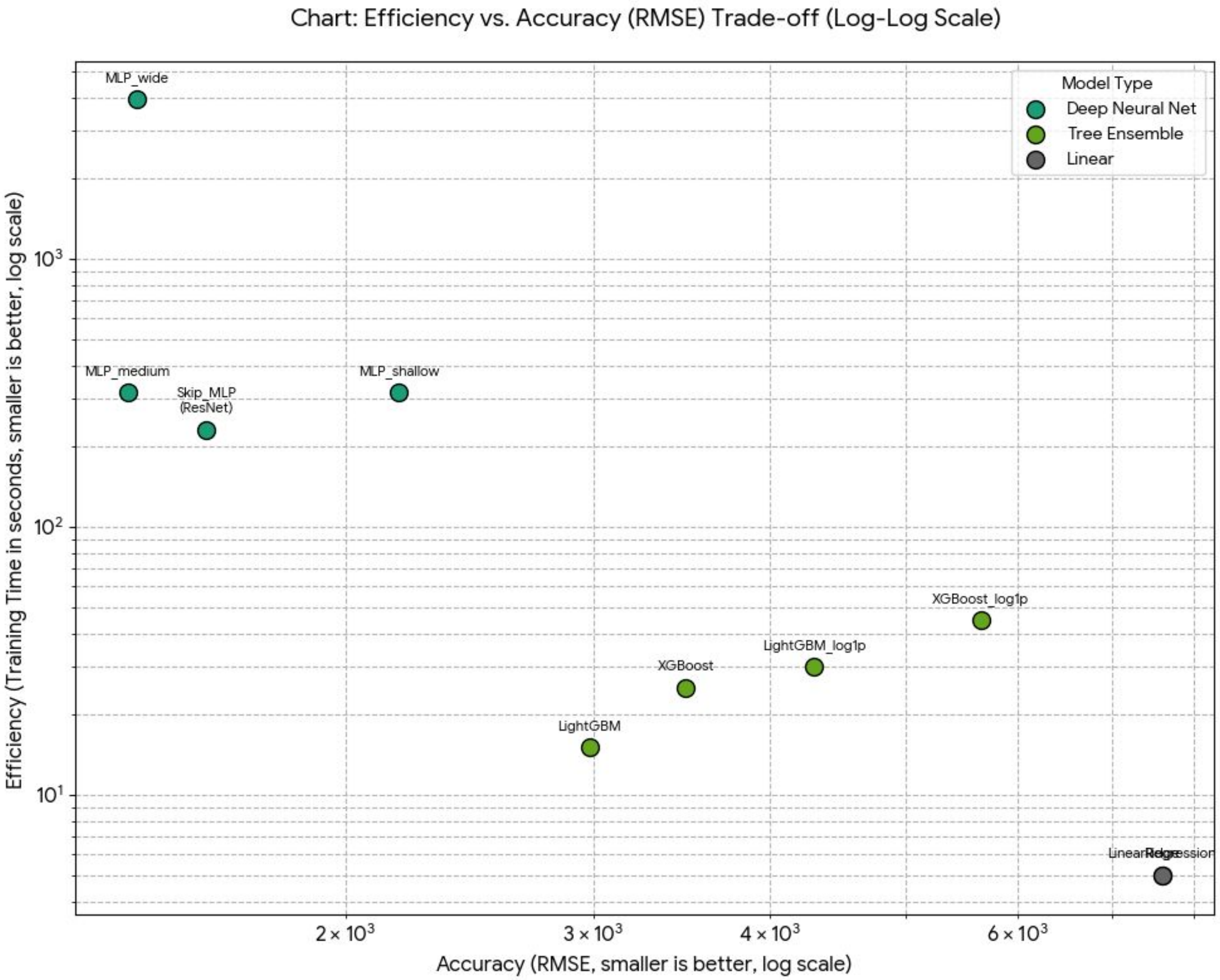
Initial Observation (RMSE)

- **DL Appears Superior:** The *MLP* variants occupy the top 4 ranks, suggesting the lowest prediction error.
- **Tree Ensembles vs. Linear:** *LightGBM* (***2,980***) is >50% better than the *Linear* Baseline ***7,600***.

Supportive Answer:

The low *RMSE* for *DL* proves its ability to minimize the MSE loss function, indicating extremely high **fit** to the training data. This is why the *DL* models top the chart.

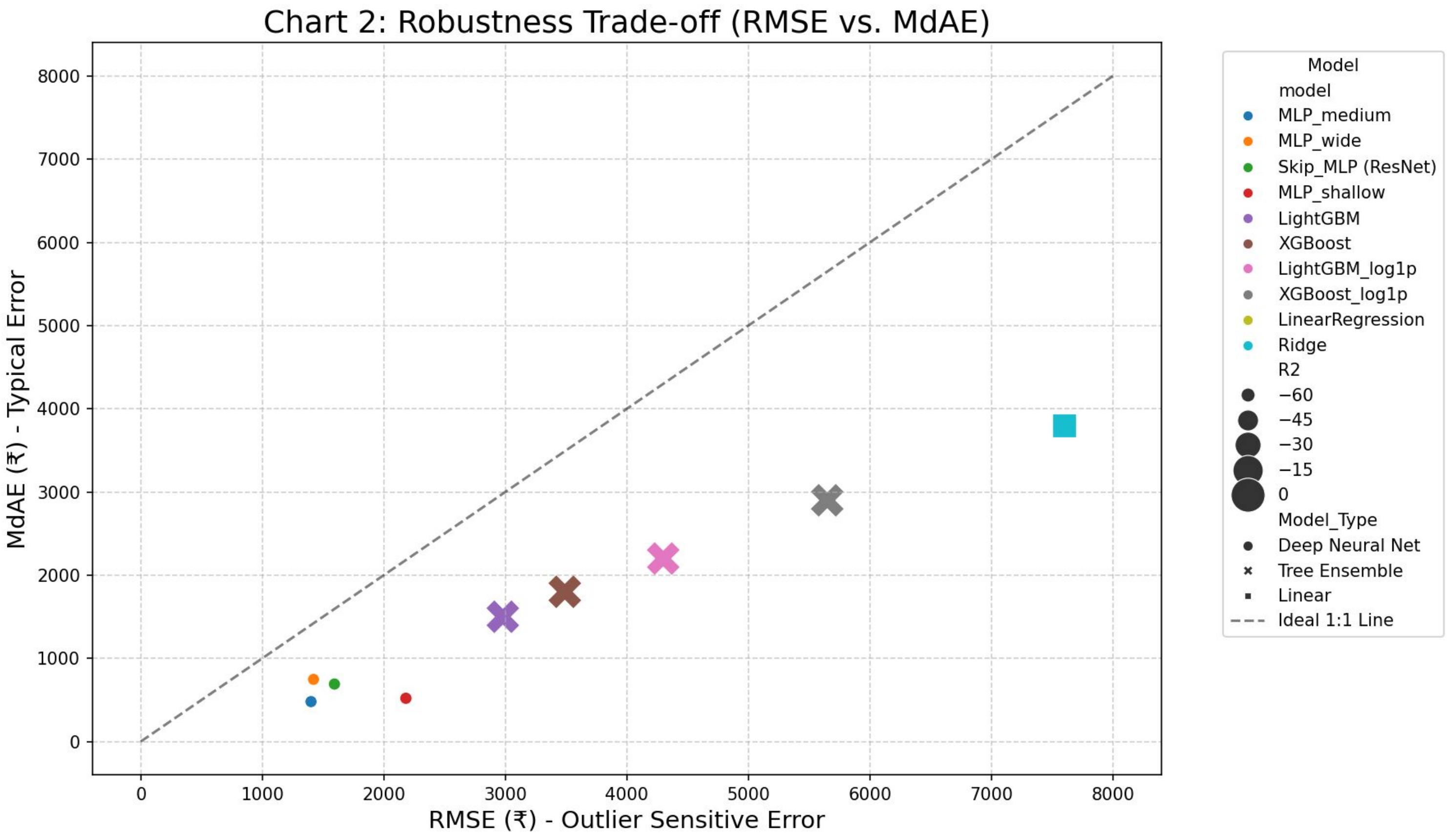
Result Chart 2: The Critical R^2 Anomaly



The R^2 Anomaly & Concern

- **The Concern:** *DL* models registered a negative R^2 (e.g., -62.95), meaning their predictions are **worse than just guessing the average employment demand**.
- **The Reason:** The DL models are numerically unstable, likely due to the un-normalized, heavy-tailed distribution of the target variable (high variance/outliers).
- **Supportive Answer:** We must discard the *DL* results for deployment, as their low *RMSE* is an artifact of **overfitting** or **numerical instability**, not generalized prediction power.

Result Chart 3: Robustness (RMSE vs. MdAE)



Analysis: Outliers and Typical Error

- **MdAE Significance:** *MdAE* shows the typical error, unaffected by outliers.
- **LightGBM Stability:** The ratio *MdAE / RMSE* is **0.50** [cite: 0.5033557046979866] for *LightGBM*, meaning the typical error *(₹1,500)* is half the total error *(₹2,980)*.
- **Conclusion:** *LightGBM* is highly stable for routine forecasts, but its total error is inflated by the volatility of a few, large-scale administrative outliers.

Technical Insight: Deep Learning Efficiency

Skip-MLP: Faster Complex Training

The **Skip-MLP (ResNet-style)** model achieved the lowest training time for a complex *DL* architecture *(229 seconds)* [cite: 229.00434613227844].

Justification: The **residual skip connection** prevents gradient degradation, allowing for faster convergence compared to the much slower standard *MLP_wide (3931 seconds)* [cite: 3931.264481782913].

MLP_wide: Computational Concern

The *MLP_wide* model took over **1 hour** (3931 seconds) to train, even with GPU acceleration.

Concern: This demonstrates that simply making a network wide or deep without architectural improvements (like skip connections) results in prohibitive computational cost for production systems.

Final Deployment Recommendation

Selecting the Model for Operational Stability and Accuracy.

Actionable Insights & Next Steps

- **Model Deployment:** ***LightGBM*** is the ideal candidate. It offers *0.9995 R^2* and fast training, making it reliable for daily or weekly fund allocation forecasts.
- **Feature Importance:** Use *LightGBM*'s feature importance output to identify the administrative and demographic drivers (e.g., specific district codes, budget types) most critical to employment demand.
- **Targeted Intervention:** Integrate the *EDA* finding on low 100-Day achievement with model outputs to flag districts at risk of failing the employment guarantee.
- **Data Concern:** Address the *DL* instability by investigating target variable transformation (e.g., using *log1p* specifically for the *DL* pipeline to improve R^2.

Thank you