BUSINESS DATA MANAGEMENT:
CAPSTONE PROJECT

Forecasting Optic Fiber Deployment for enhanced Revenue Planning

Archit Handa (22f2000744)

Business Overview

Sterlite Technologies Limited (STL) is a global **optical and digital solutions** company, which aims to deliver **end-to-end data network solutions** to its clients. STL currently operates in the **B2B** segment.

Project Focus:

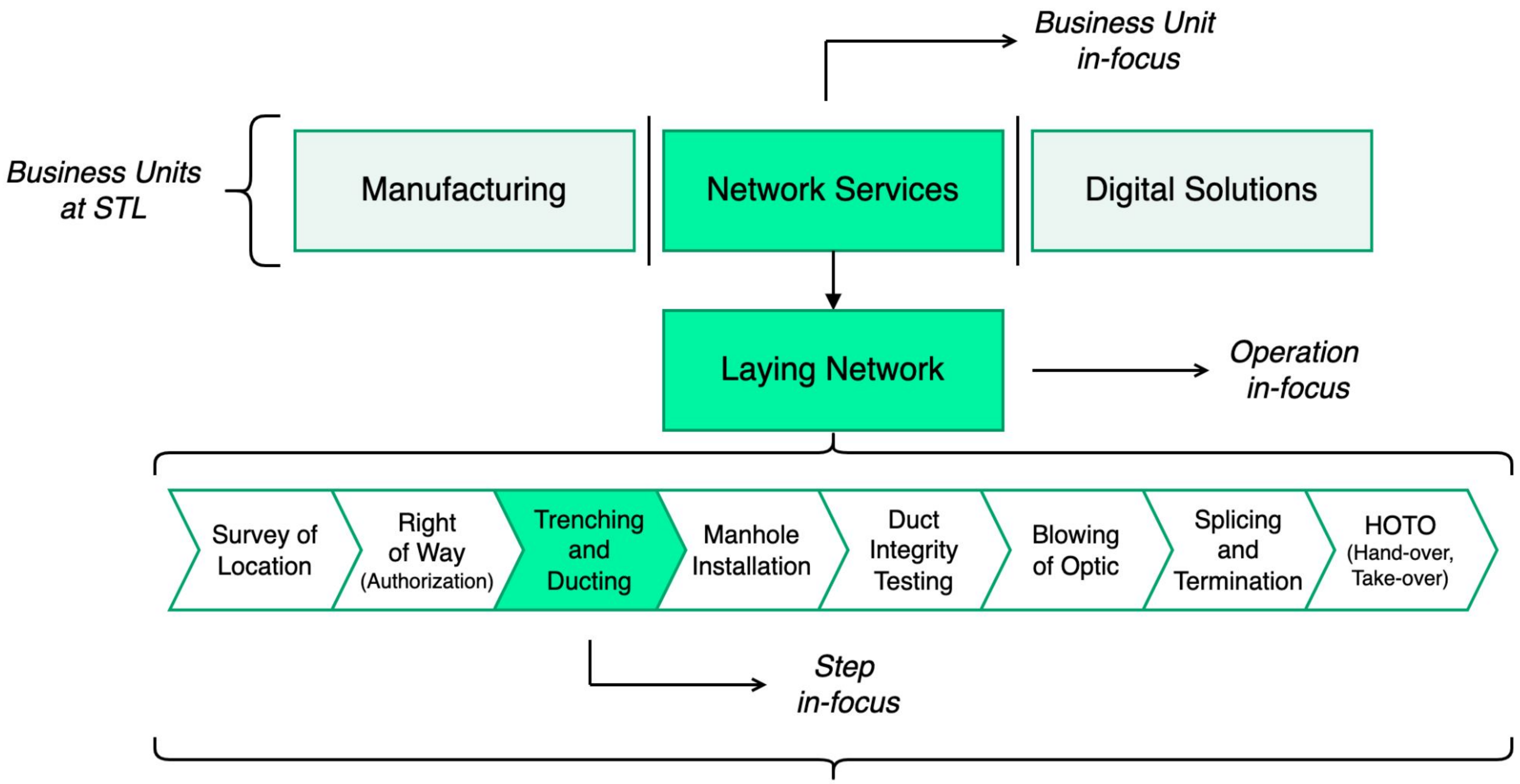
Trenching and Ducting (T&D), a crucial step in laying out optic fiber, thus in **Capacity** and overall **Revenue Planning** for the Organisation

Business Problem:

Inability to estimate optic fiber layout output for a month leading to **deviations from Revenue Projections** which causes **delays in Turn Around Time (TAT)** and **Resource Wastage**.

Solution:

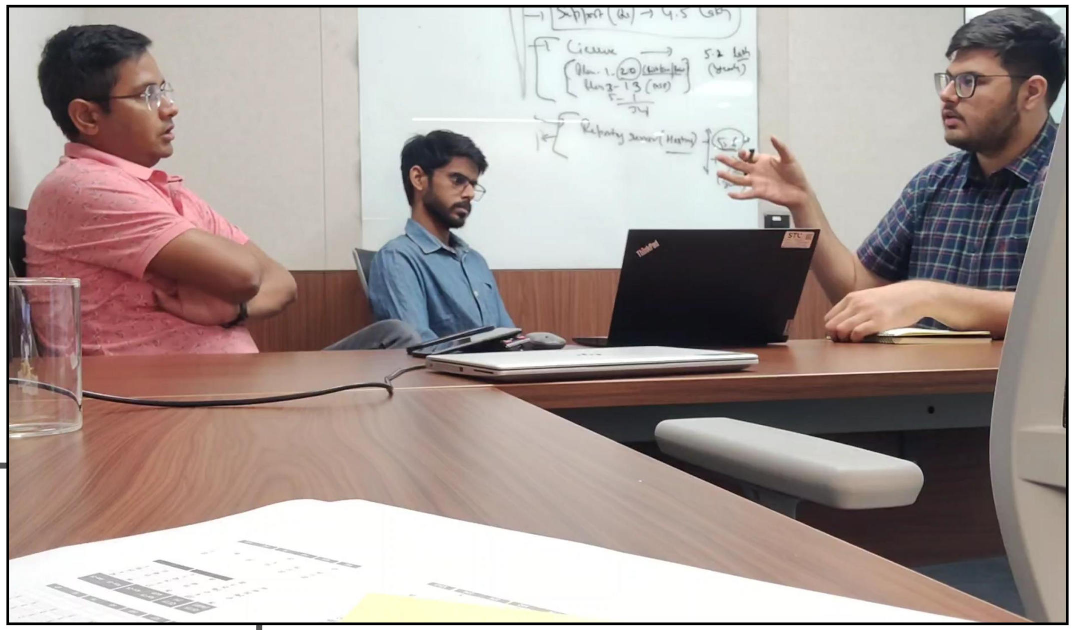
Machine Learning Model that integrates concepts of **Linear Regression** and **Time Series Analysis**.



Entire Process of Laying out an Optic Fibre Network

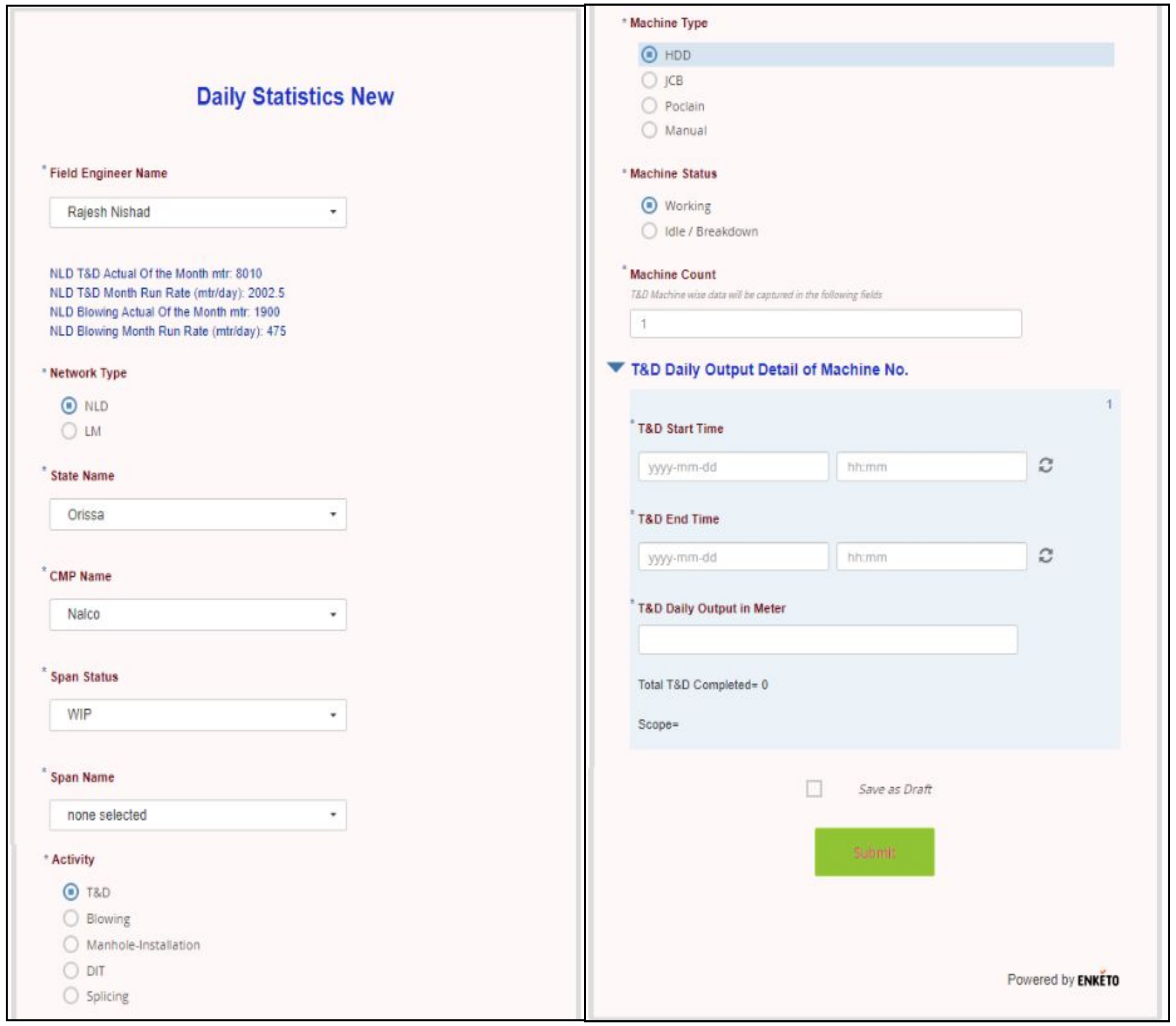
People of Contact:

- **Mr. Pankaj Singh**
  PMO Lead at STL
- **Mr. Shailendra Kumar**
  PMO Analytics Engineer at STL



Data Collection

- Data collected using STL’s in-house tool: **FieldForce**
- Daily details input by Field Engineer
- Data collected from **October 2021 till date**



Daily Statistics New

* Field Engineer Name
Rajesh Nishad

NLD T&D Actual Of the Month mtr: 8010
NLD T&D Month Run Rate (mtr/day): 2002.5
NLD Blowing Actual Of the Month mtr: 1900
NLD Blowing Month Run Rate (mtr/day): 475

* Network Type
NLD
LM

* State Name
Orissa

* CMP Name
Nalco

* Span Status
WIP

* Span Name
none selected

* Activity
T&D
Blowing
Manhole-Installation
DIT
Splicing

* Machine Type
HDD
JCB
Poclain
Manual

* Machine Status
Working
Idle / Breakdown

* Machine Count
T&D Machine wise data will be captured in the following fields
1

T&D Daily Output Detail of Machine No.
1

* T&D Start Time
yyyy-mm-dd hh:mm

* T&D End Time
yyyy-mm-dd hh:mm

* T&D Daily Output in Meter

Total T&D Completed= 0
Scope=

Save as Draft
Submit

Powered by ENKETO

Data Preprocessing

Step 1: Imputation

- **Unusable fields were dropped** such as Field Engineer Details
- **Missing values** were filled according to **patterns exhibited**
- **First 3-month data** was **extremely noisy** due to **field engineers’ unawareness** regarding the tool → **Dropped for further analysis**

Step 2: Aggregation

- Current state captures **span-wise data**; however, **planning** is to be performed for **all spans in the country as a whole**
- **Daily Statistics** allow **granularity**; however, from modeling perspective, can **cause the estimator to overfit**

Step 3: Feature Engineering

- Work Duration = Machine End Time – Machine Start Time
- Daily Productivity = Daily Output (in meter) / Machine Count

_device	FE_Name	State_Name	CMP_Name	Span_Name	Activity	Machine_Type	Machine_Status	Machine_Count	Trenching_And_Ducting_Start_Time	Trenching_And_Ducting_End_Time	Trenching_And_Ducting_Daily_Output_in_Meter	Scope_kms
FieldForce:	[redacted]	Meghalaya	Shillong	AS-NE Border-1 to NE-AS Bo	Trenching-and-Ducting	HDD	Working	2	2023-09-08 3:	2023-09-08 12	610	23.0
FieldForce:	[redacted]	Meghalaya	Shillong	AS-NE Border-1 to NE-AS Bo	Trenching-and-Ducting	HDD	Working	2	2023-09-08 3:	2023-09-08 12	610	23.0
FieldForce:	[redacted]	MP	Ujjain	Moman Badodiya-Bhaisoda	Trenching-and-Ducting	JCB	Idle-Breakdown	1				
FieldForce:	[redacted]	MP	Ujjain	Nalkheda (NP)-Kayra	Trenching-and-Ducting	JCB	Idle-Breakdown	1				
FieldForce:	[redacted]	MP	Khargone	Mohana-Bagarda	Trenching-and-Ducting	JCB	Working	1	2023-09-06 5:	2023-09-06 12	300	18.5
FieldForce:	[redacted]	MP	Khargone	Mohana-Bagarda	Trenching-and-Ducting	JCB	Working	1	2023-09-07 4:	2023-09-07 12	200	18.5
FieldForce:	[redacted]	Orissa	Bhawanipatna	Titlagarh-Sirol	Trenching-and-Ducting	HDD	Working	1	2023-09-08 3:	2023-09-08 11	310	21.01
FieldForce:	[redacted]	MP	Chhindwara	Dongariya-Garra (CT)	Trenching-and-Ducting	HDD	Working	1	2023-09-08 5:	2023-09-08 12	70	11.36
FieldForce:	[redacted]	MP	Chhindwara	Patan-Lakhnadon	Trenching-and-Ducting	Poclain	Idle-Breakdown	1				
FieldForce:	[redacted]	Orissa	Rayagada	Sorispadar-Parajasuku	Trenching-and-Ducting	HDD	Working	1	2023-09-08 5:	2023-09-08 12	160	
STL:bgVlvM	[redacted]	Orissa	Rayagada	Balimela-Balimela	Trenching-and-Ducting	Manual	Working		2023-09-08 2:	2023-09-08 0:1	180	0.0
FieldForce:	[redacted]	MP	Ratlam	Rawti-Ratlam	Trenching-and-Ducting	JCB	Working	1	2023-09-08 3:	2023-09-08 11	100	27.0
FieldForce:	[redacted]	MP	Shahdol	Sheori Chandas-Purga	Trenching-and-Ducting	HDD	Working	1	2023-09-08 4:	2023-09-08 13	300	49.74
FieldForce:	[redacted]	Orissa	Bhawanipatna	Telenpali-Badbanjipali	Trenching-and-Ducting	HDD	Working	1	2023-09-07 3:	2023-09-07 13	310	22.12
FieldForce:	[redacted]	Orissa	Bhawanipatna	Telenpali-Badbanjipali	Trenching-and-Ducting	HDD	Working	1	2023-09-08 2:	2023-09-08 13	240	22.12
FieldForce:	[redacted]	Orissa	Keonjhar	Jalahari-Jaroli	Trenching-and-Ducting	JCB	Idle-Breakdown	1				

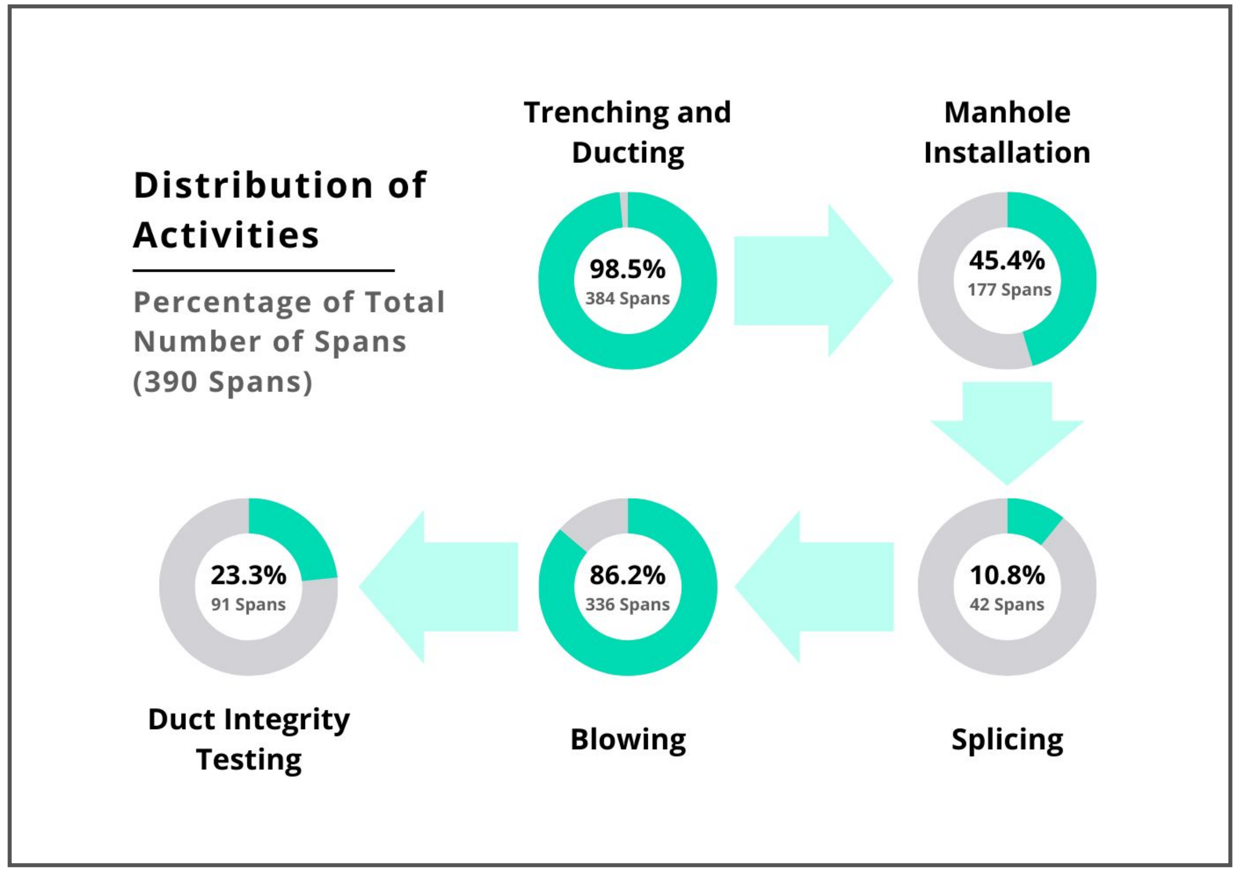
Raw Data before Processing

Data Tools
MS Excel, Pandas, Numpy, Scikit-Learn

Cleaned Data after Processing

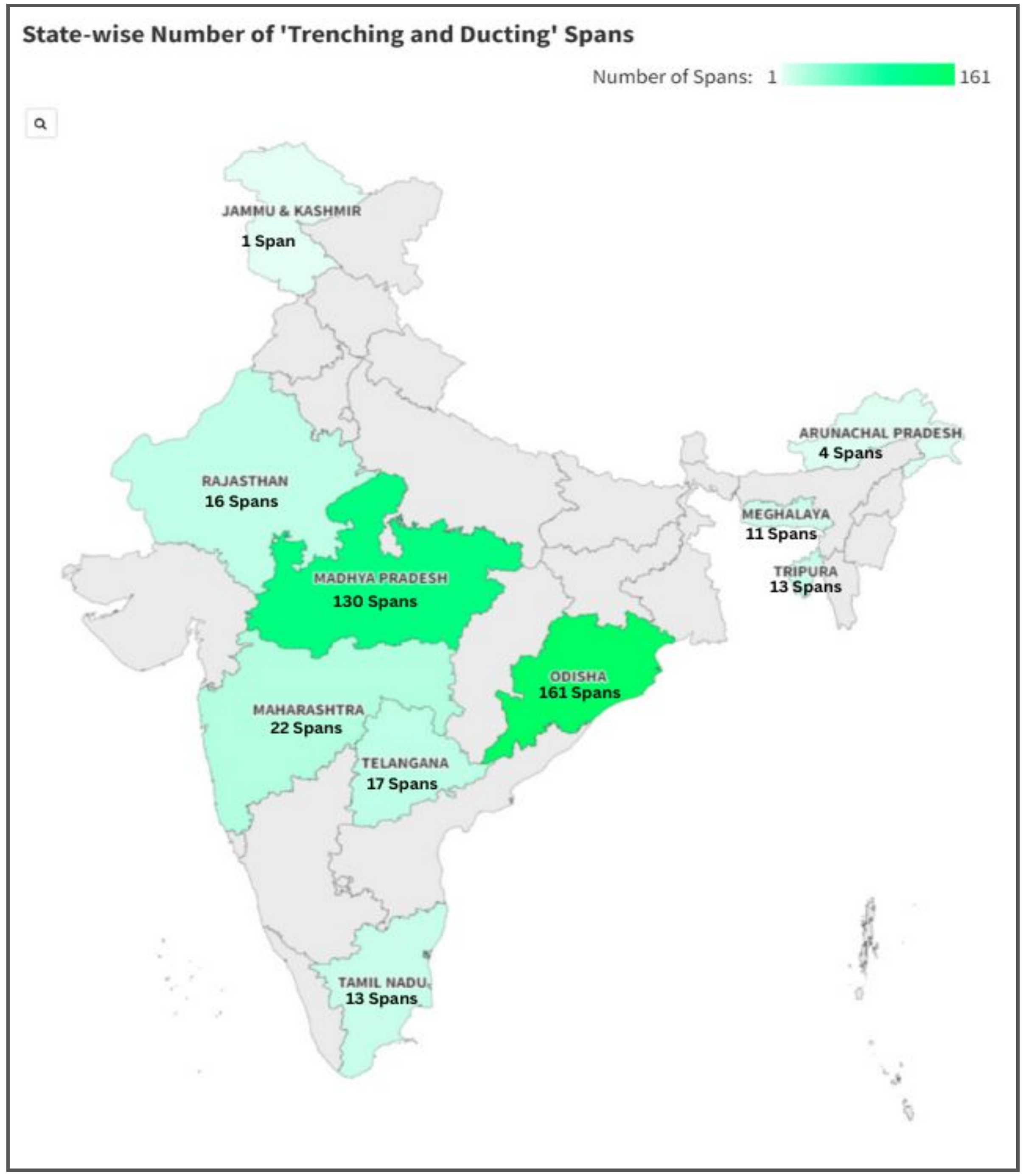
Date	Scope	JCB_Count	JCB_Hours	JCB_Output	JCB_Productivity	HDD_Count	HDD_Hours	HDD_Output	HDD_Productivity	Poclain_Count	Poclain_Hours	Poclain_Output	Poclain_Productivity	Total_Output	Total_Hours	Total_Count	Total_Productivity
2022-01-01	410070.33	5	36.38	1980	396.00	5	22.52	650	130.00	2	2.02	0	0.00	2630	60.92	12	219.17
2022-01-02	406919.00	5	37.20	1700	340.00	7	48.43	1500	214.29	3	20.25	110	36.67	3310	105.88	15	220.67
2022-01-03	400914.00	8	57.80	3050	381.25	7	35.28	975	139.29	2	15.03	60	30.00	4085	108.12	17	240.29
2022-01-04	394909.00	6	35.27	1000	166.67	9	58.72	1812	201.33	3	8.77	100	33.33	2912	102.75	18	161.78
2022-01-05	388904.00	8	49.90	1540	192.50	8	49.07	1595	199.38	3	24.92	450	150.00	3585	123.88	19	188.68
2022-01-06	382899.00	6	45.73	1290	215.00	10	59.65	1712	171.20	3	16.85	0	0.00	3002	122.23	19	158.00
2022-01-07	376894.00	6	44.62	1420	236.67	8	54.00	1625	203.13	2	2.80	0	0.00	3045	101.42	16	190.31
2022-01-08	370889.00	5	37.63	2300	460.00	9	53.25	1694	188.22	2	7.95	150	75.00	4144	98.83	16	259.00
2022-01-09	364884.00	5	37.53	1750	350.00	11	68.88	2400	218.18	1	10.00	0	0.00	4150	116.42	17	244.12
2022-01-10	358879.00	8	64.77	2225	278.13	8	51.83	2205	275.63	2	15.13	200	100.00	4630	131.73	18	257.22
2022-01-11	353752.33	7	58.97	2490	355.71	8	49.25	2245	280.63	2	23.02	100	50.00	4835	131.23	17	284.41
2022-01-12	348625.67	6	40.78	1640	273.33	10	60.63	3176	317.60	2	20.78	200	100.00	5016	122.20	18	278.67
2022-01-13	343499.00	5	46.92	1530	306.00	7	35.98	2111	301.57	3	27.27	300	100.00	3941	110.17	15	262.73
2022-01-14	338372.33	6	49.32	1640	273.33	6	32.32	2107	351.17	2	12.00	0	0.00	3747	93.63	14	267.64
2022-01-15	333245.67	6	50.23	1330	221.67	6	34.02	2019	336.50	3	25.33	470	156.67	3819	109.58	15	254.60
2022-01-16	328119.00	5	47.65	1660	332.00	6	34.07	1980	330.00	2	19.05	0	0.00	3640	100.77	13	280.00
2022-01-17	322882.00	4	36.53	1400	350.00	8	39.82	2750	343.75	3	22.08	200	66.67	4350	98.43	15	290.00
2022-01-18	316694.33	5	44.45	850	170.00	10	64.93	3749	374.90	3	30.83	100	33.33	4699	140.22	18	261.06
2022-01-19	310506.67	4	35.58	1300	325.00	7	59.67	2364	337.71	2	8.73	150	75.00	3814	103.98	13	293.38
2022-01-20	306896.67	7	61.18	2071	295.86	7	39.08	2252	321.71	2	17.25	0	0.00	4323	117.52	16	270.19
2022-01-21	303286.67	5	46.73	1593	318.60	5	43.00	2640	528.00	2	18.50	0	0.00	4233	108.23	12	352.75
2022-01-22	300938.33	7	62.23	1545	220.71	7	44.45	2485	355.00	2	17.50	0	0.00	4030	124.18	16	251.88

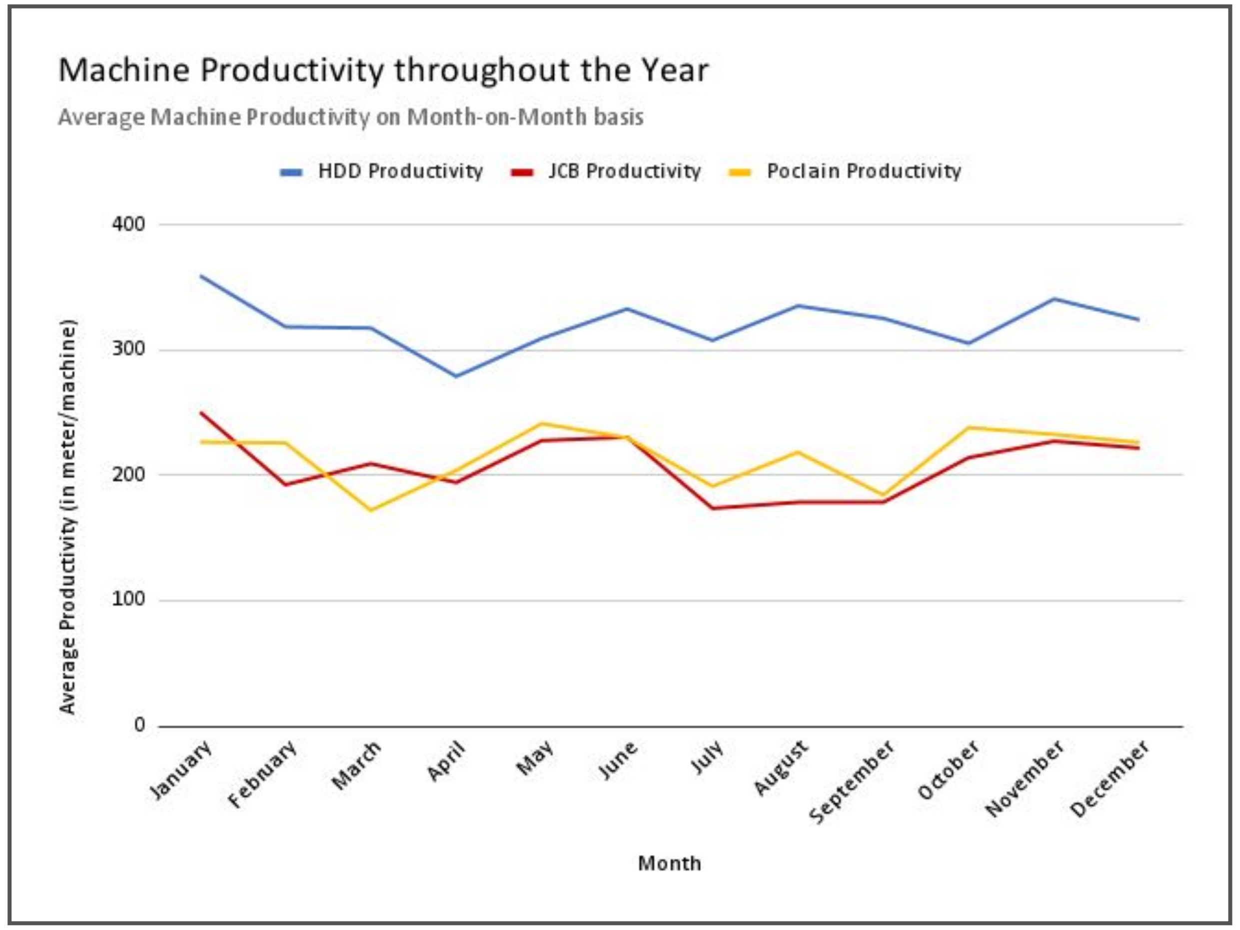
Descriptive Statistics



- Manhole Installation, Splicing, and DIT are quick to execute
- T&D and Blowing go hand-in-hand with a 10-day lag
- Significant Weight of T&D in Overall Capacity Planning

- Data available from various soil and weather landscapes
- Positive from a modeling viewpoint

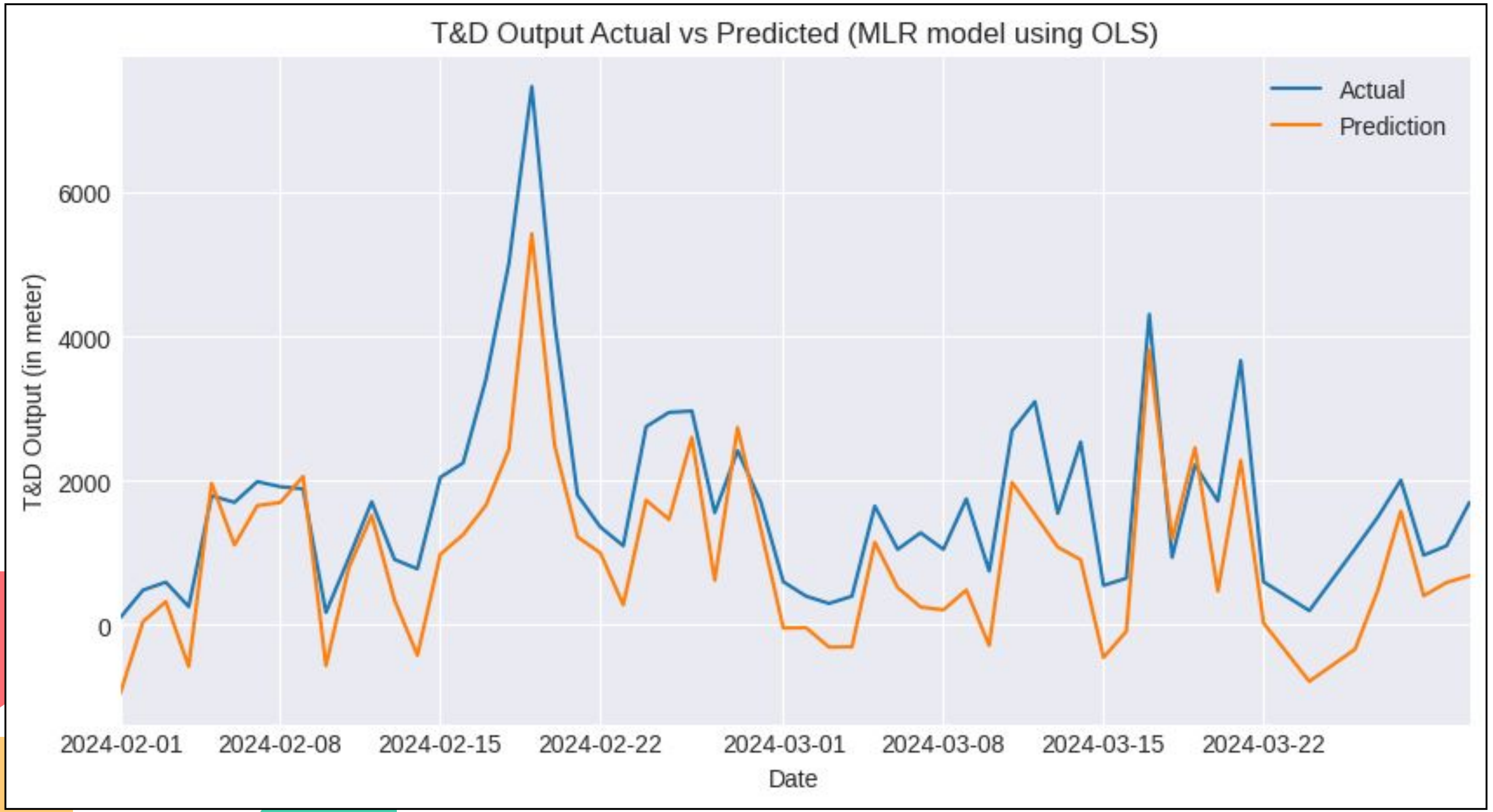




- HDD has highest productivity
- Fluctuations hint at certain cyclical patterns that should be decoded

Data Analysis and Modelling

Linear Regression: MLR OLS




```
                            OLS Regression Results
==============================================================================
Dep. Variable:           Total_Output   R-squared:                       0.962
Model:                            OLS   Adj. R-squared:                  0.962
Method:                 Least Squares   F-statistic:                     3828.
Date:                Sun, 14 Apr 2024   Prob (F-statistic):               0.00
Time:                        15:26:01   Log-Likelihood:                -6794.1
No. Observations:                 760   AIC:                         1.360e+04
Df Residuals:                     754   BIC:                         1.363e+04
Df Model:                           5
Covariance Type:            nonrobust
=================================================================================
                    coef    std err          t      P>|t|      [0.025      0.975]
---------------------------------------------------------------------------------
const          -927.7405    171.363     -5.414      0.000   -1264.145    -591.336
JCB_Count       271.9102      8.427     32.267      0.000     255.367     288.453
HDD_Count       361.1449      6.803     53.082      0.000     347.789     374.501
Poclain_Count   228.4512     27.589      8.280      0.000     174.290     282.612
Scope_km         -0.3843      0.081     -4.771      0.000      -0.542      -0.226
Month             9.1140     19.805      0.460      0.646     -29.766      47.994
==============================================================================
Omnibus:                      148.154   Durbin-Watson:                   0.745
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              479.465
Skew:                           0.924   Prob(JB):                    7.68e-105
Kurtosis:                       6.424   Cond. No.                     6.66e+03
==============================================================================
```

Metrics

R2 Score shows **Overfitting**:
0.962 (Train) | 0.478 (Test)

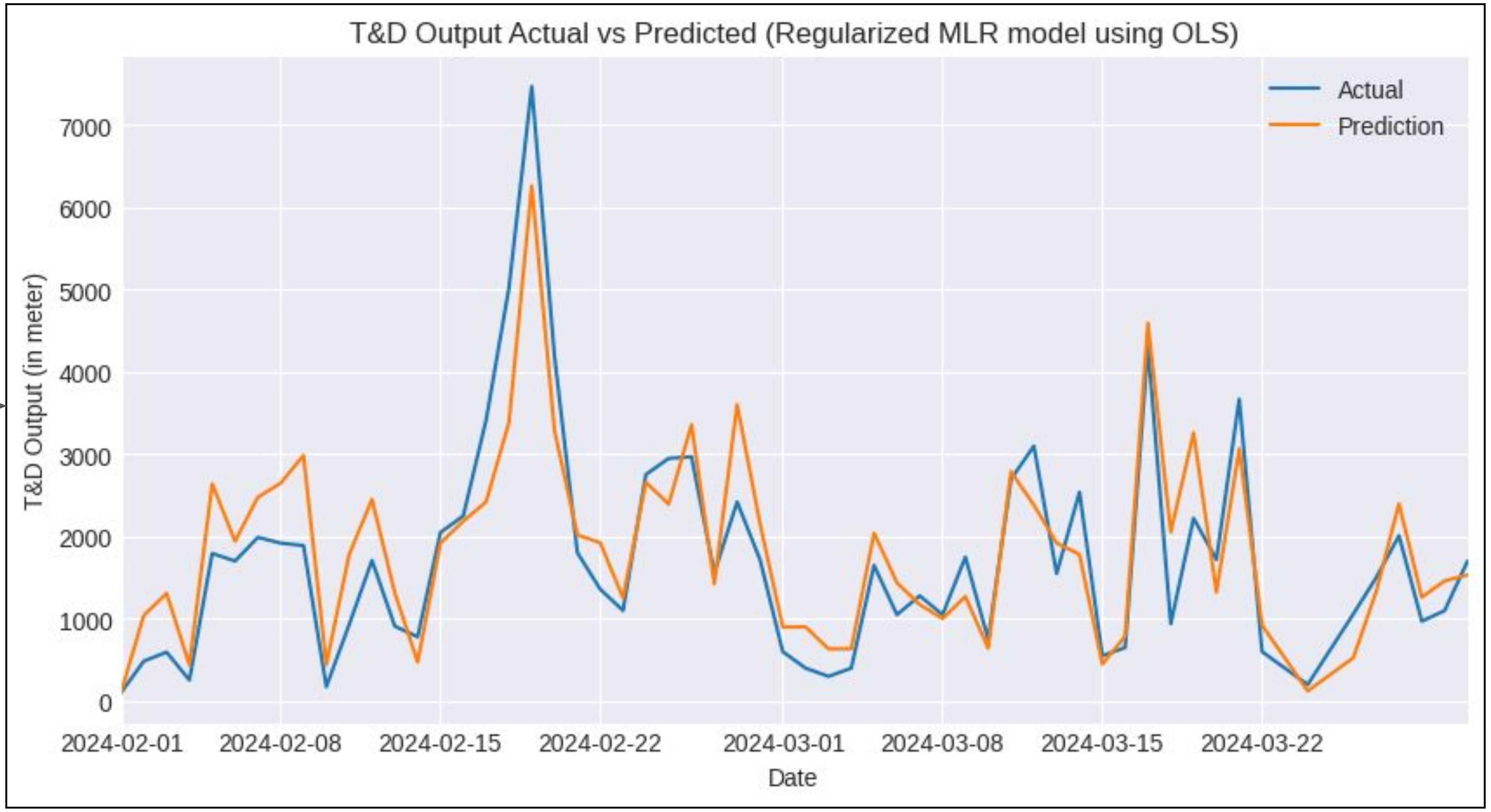
Percentage Difference:
44.53%

- Accurately predicts the Spikes and Dips
- Does not consider the Month variable, hence the **Seasonality Effect**

Data Analysis and Modelling

Linear Regression: Regularized MLR OLS with $\alpha = 10$

Regularization used to tackle problem of **Overfitting**



Metrics

R2 Score shows **Good Fit**: 0.958 (Train) | 0.807 (Test)

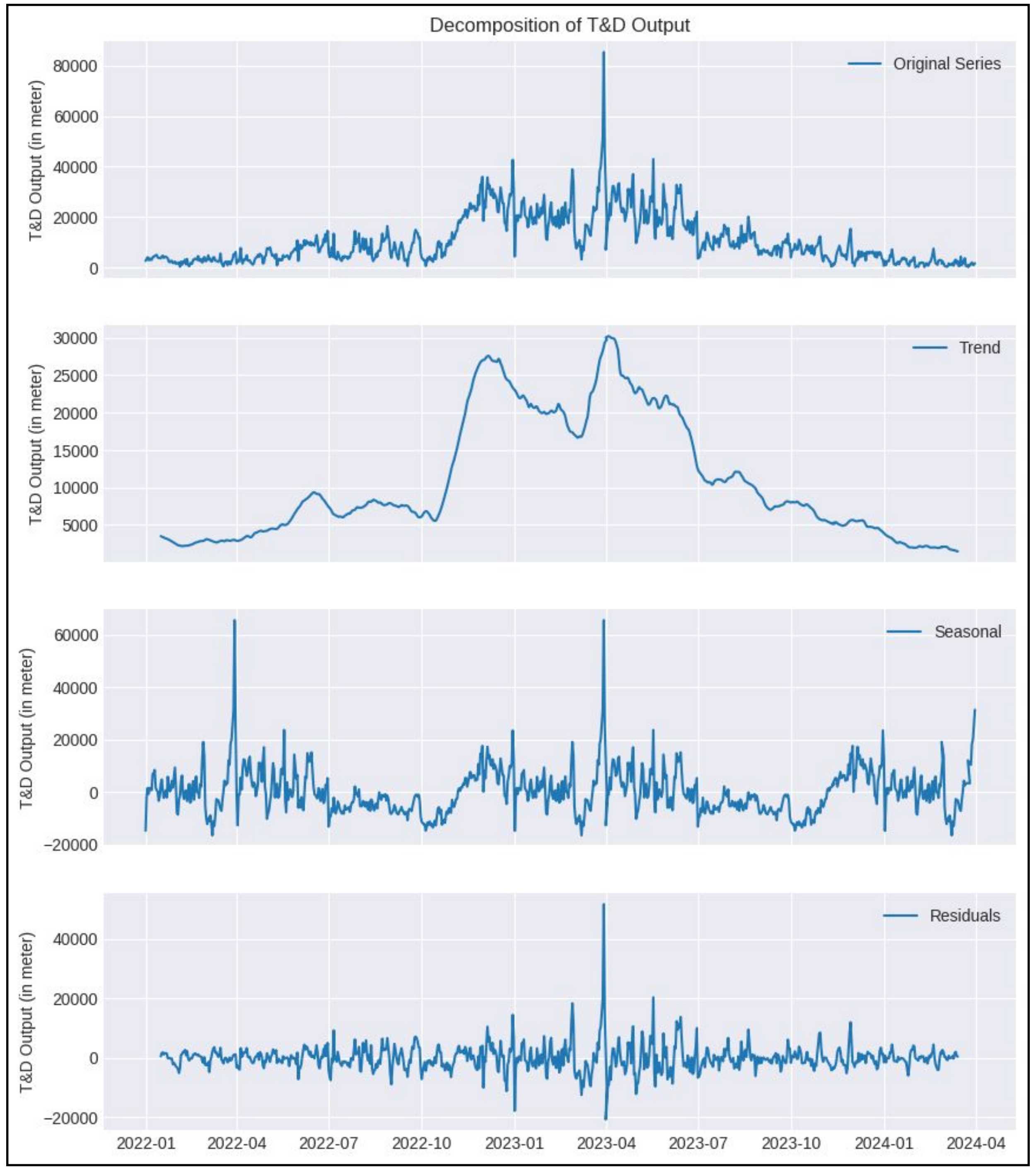
Percentage Difference: 6.3%

- Statistically, much better than MLR OLS
- Yet still does not consider the **Seasonality Effect**

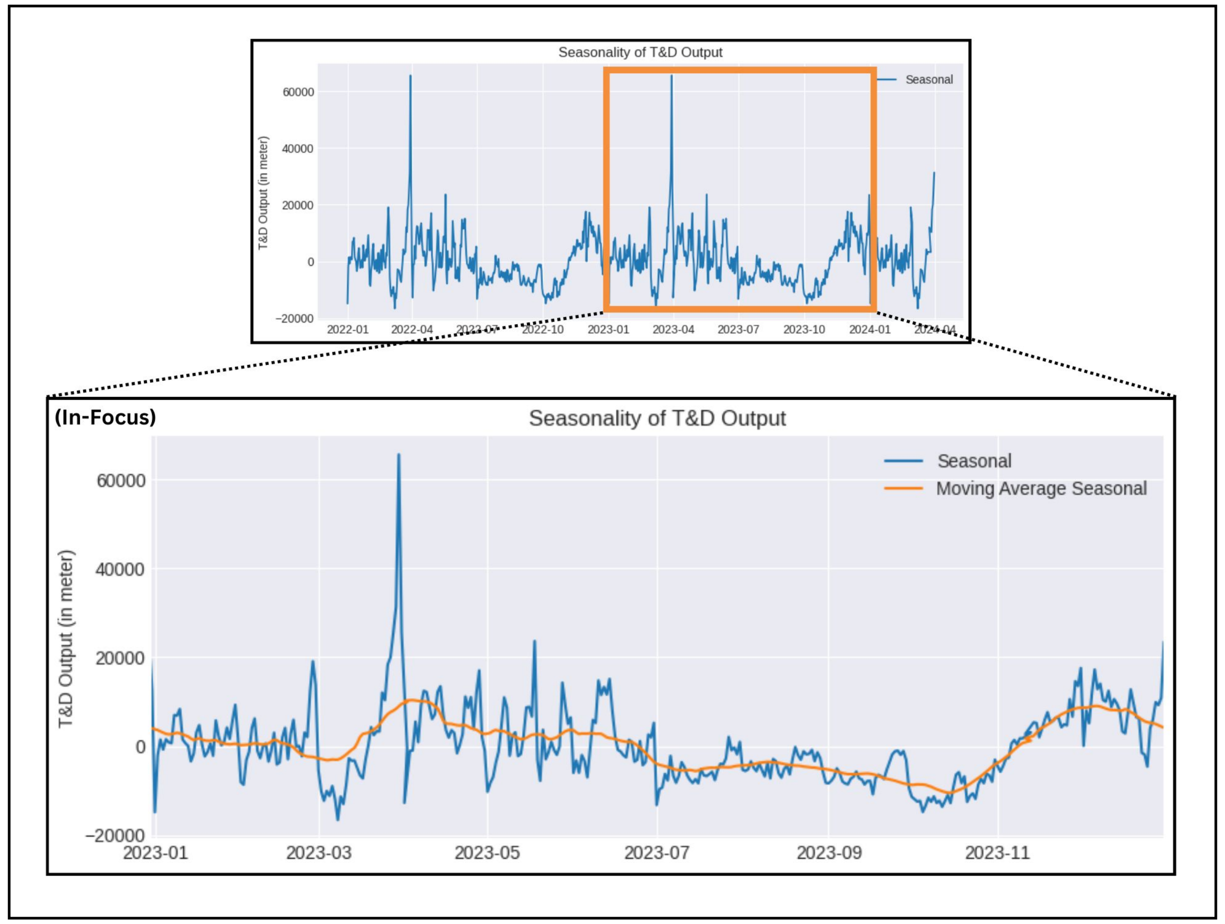
Data Analysis and Modelling

Time Series Analysis

Seasonal Decomposition



Seasonality (In-focus)



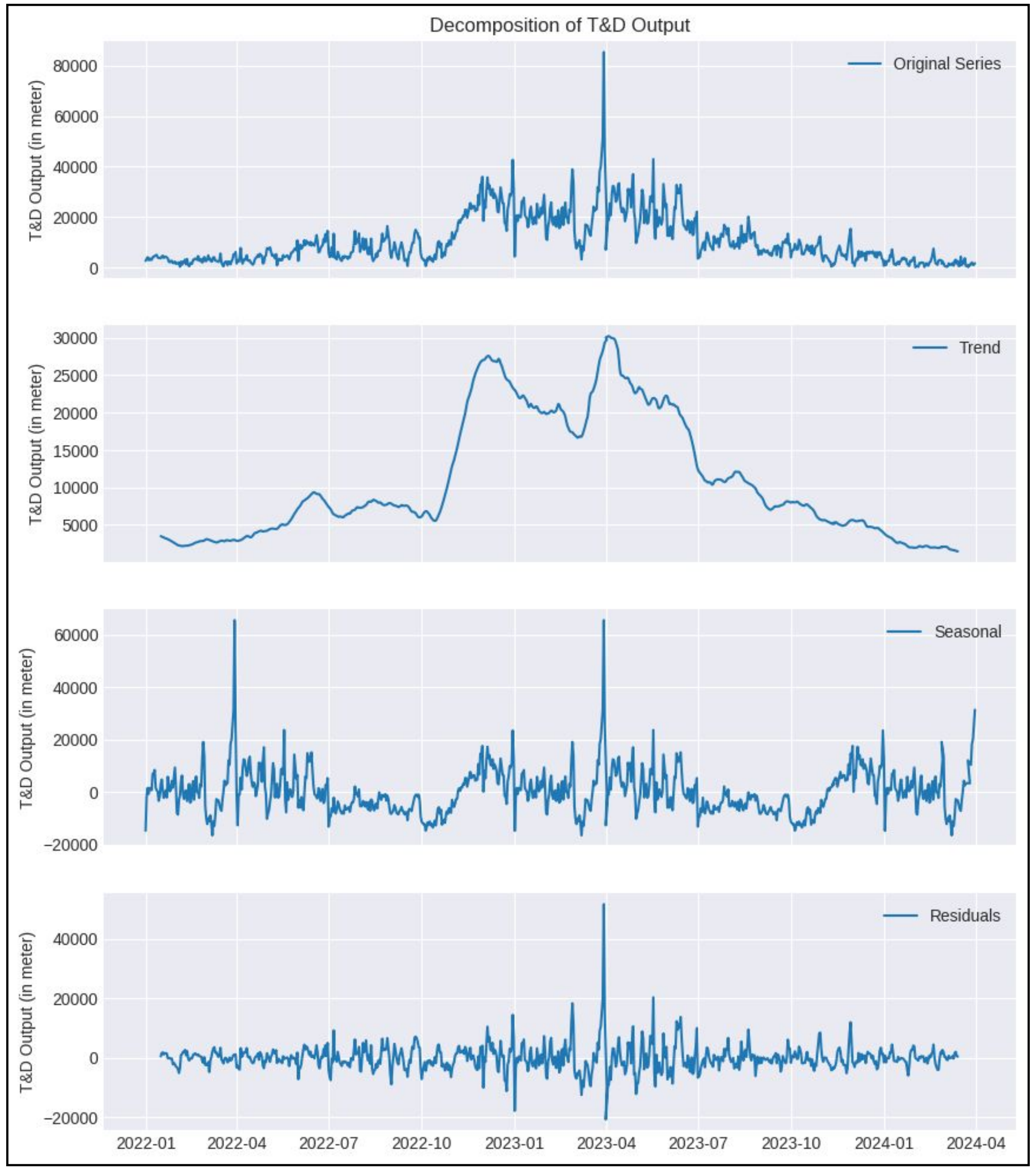
30-day Moving Average shows:

- Peak in April
- Overall High Productivity till June
- Fall till October (Monsoon + Festive)
- Increase till December
- Slight downfall till March

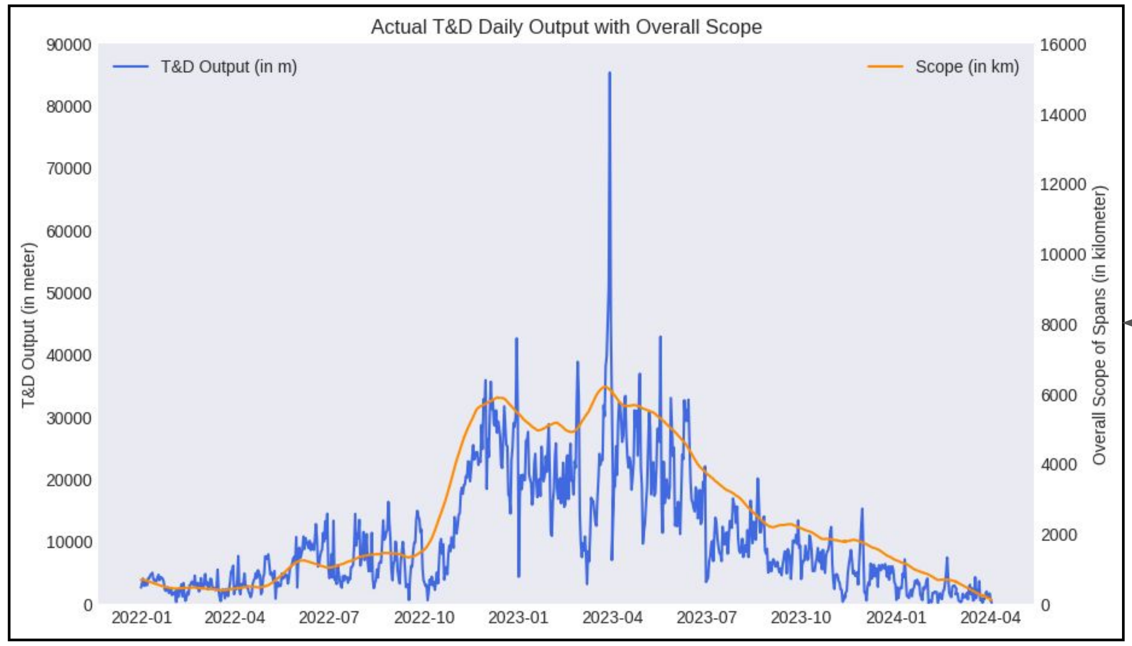
Data Analysis and Modelling

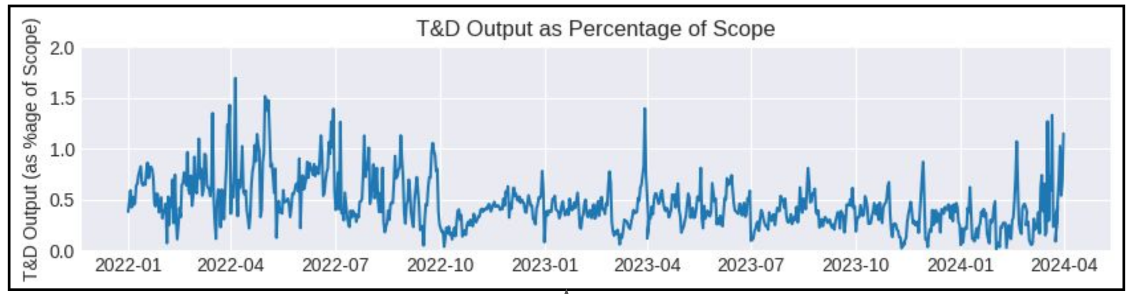
Time Series Analysis

Seasonal Decomposition



Trend (In-focus)



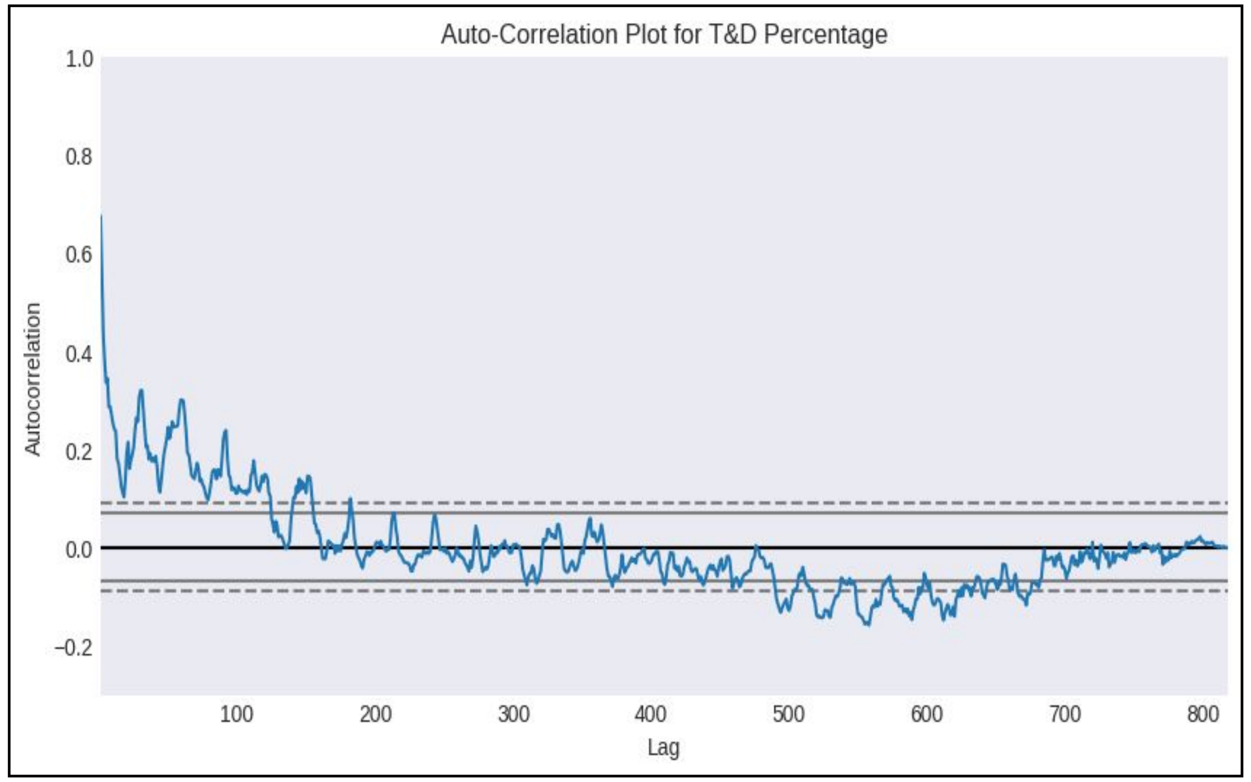


- Upward Shock in October 2022; multiple deals signed in MP and Orissa
- Taper-down post-June 2023; STL wanted to prevent **backorders** and finish pending commitments
- **Strong Correlation** of T&D with Scope: 83.8%
- Use T&D as Percentage of Scope (**T&D Percentage**) as the new target variable

- T&D Percentage exhibits **stationarity**; positive for time-series forecasting
- Mathematically backed by the **Augmented Dickey-Fuller Test**: Test-statistic = -5.6; p-value = 8.296×10^{-5}; supports Stationarity

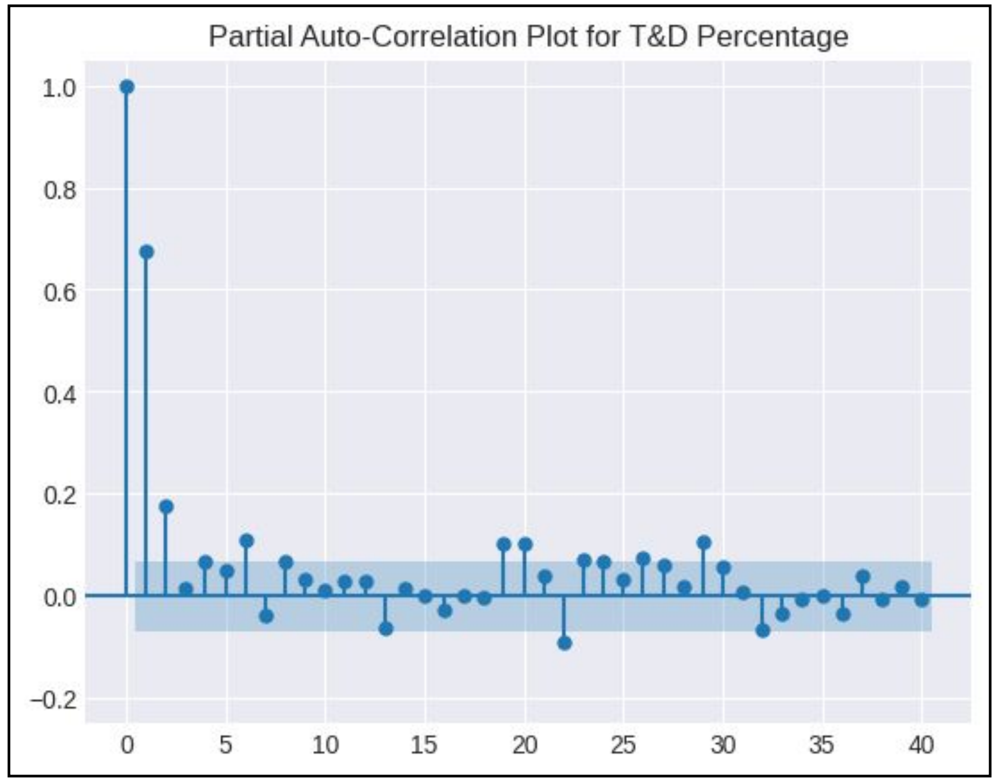
Data Analysis and Modelling
Time Series Analysis

Auto-Correlation Function (ACF) Plot



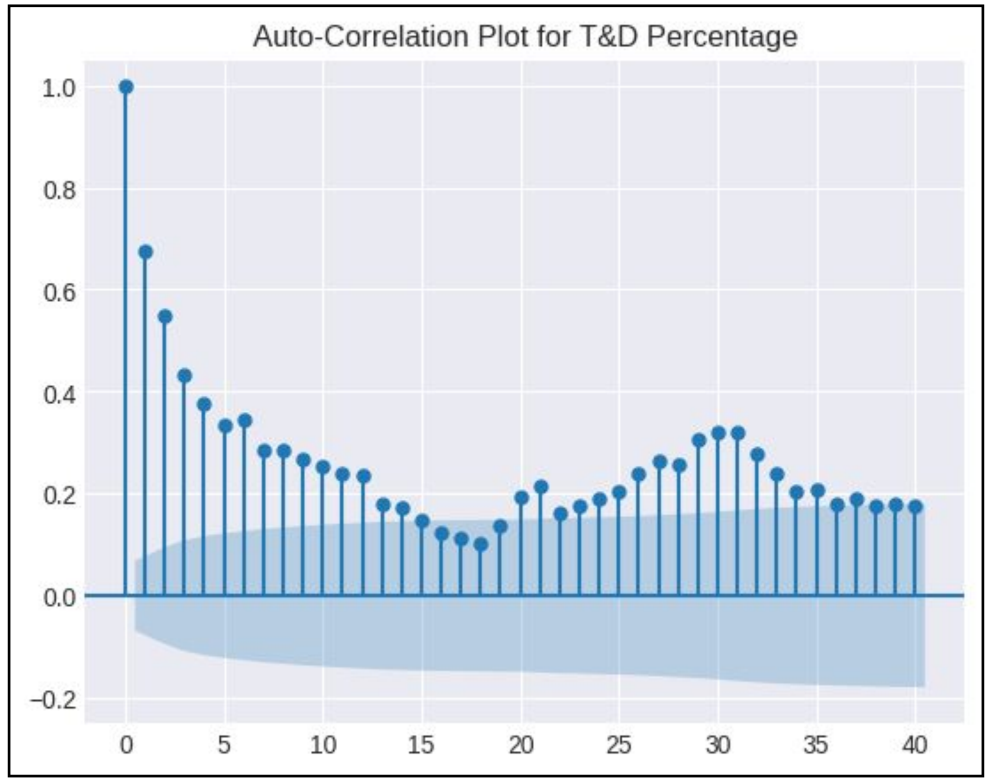
Cyclical nature → **SARIMAX** over ARIMAX

Partial Auto-Correlation Function (PACF) Plot - 40 Lags



First **3** lags are out of the significance limit → Upper Limit for **Auto Regressive (AR) parameter 'p'**

ACF Plot - 40 Lags

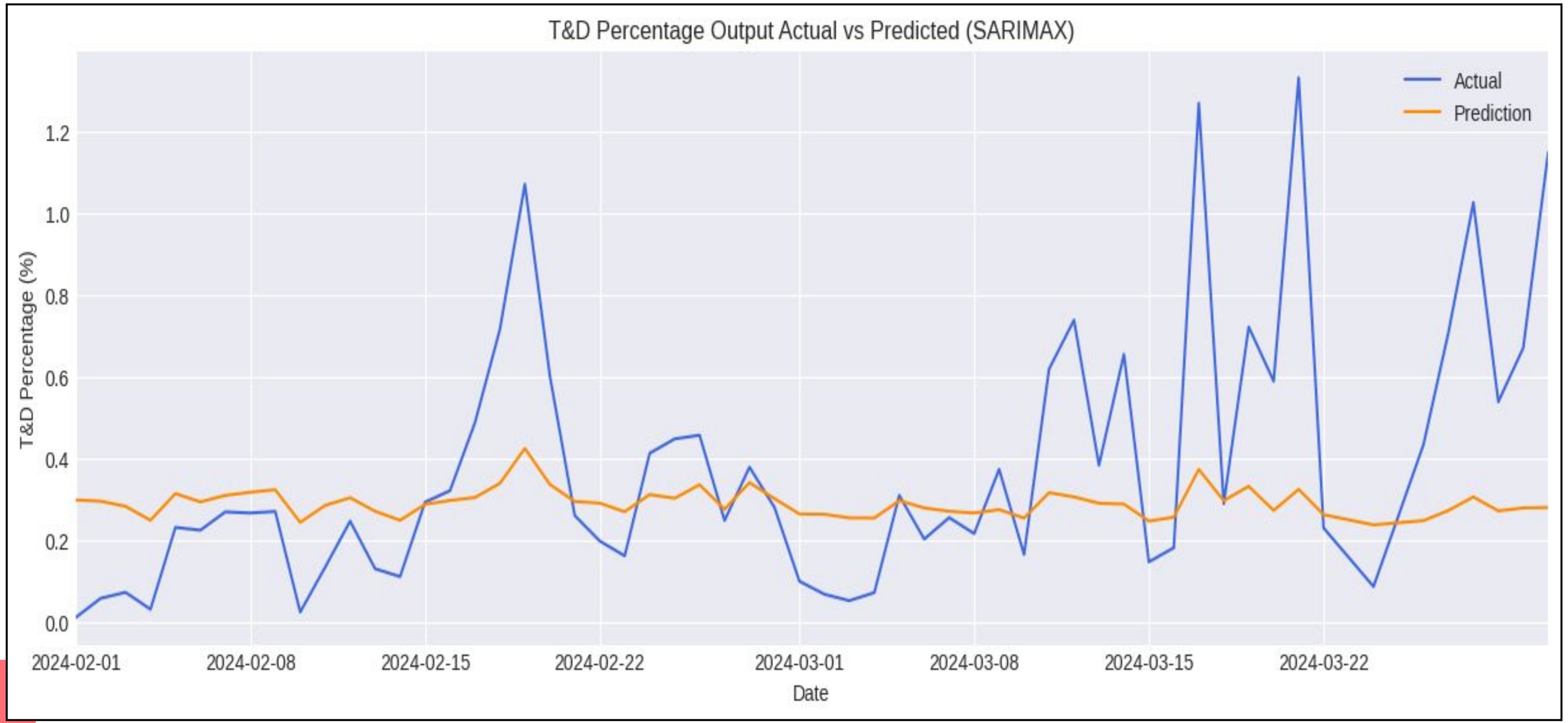


First **15** lags are significant → Upper Limit for **Moving Average (MA) parameter 'q'**

Training an Pyramid-ARIMA Auto-ARIMA model with upper limits → **SARIMAX(3,1,1)x(1,0,[],30)**

Data Analysis and Modelling

Time Series Analysis: SARIMAX



SARIMAX Results

Dep. Variable:	TND_%age	No. Observations:	760
Model:	SARIMAX(3, 1, 1)x(1, 0, [], 30)	Log Likelihood	469.523
Date:	Mon, 15 Apr 2024	AIC	-921.045
Time:	08:24:15	BIC	-879.357
Sample:	0 - 760	HQIC	-904.991
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
JCB_Count	0.0083	0.001	11.000	0.000	0.007	0.010
Poclain_Count	0.0041	0.003	1.523	0.128	-0.001	0.009
HDD_Count	0.0093	0.001	9.527	0.000	0.007	0.011
ar.L1	0.4646	0.041	11.394	0.000	0.385	0.544
ar.L2	0.0858	0.027	3.191	0.001	0.033	0.139
ar.L3	-0.0750	0.026	-2.829	0.005	-0.127	-0.023
ma.L1	-0.8595	0.035	-24.518	0.000	-0.928	-0.791
ar.S.L30	0.0157	0.025	0.627	0.531	-0.033	0.065
sigma2	0.0169	0.000	37.799	0.000	0.016	0.018

Ljung-Box (L1) (Q):	0.56	Jarque-Bera (JB):	1990.82
Prob(Q):	0.46	Prob(JB):	0.00
Heteroskedasticity (H):	0.10	Skew:	-0.23
Prob(H) (two-sided):	0.00	Kurtosis:	10.92

$$T\&D\ Output\ (in\ meter) = Predicted\ T\&D\ \%age \times Scope\ (in\ kilometer) \times \frac{1{,}000\ m/km}{100\%}$$

Metrics

Percentage Difference:
8.44%

- **Exogenous Variable:** Machine Count
- No peaks but floats around the mean
- **Not as accurate** as Regularised MLR OLS
- Yet considers the **Seasonality Effect**

Results and Findings

Model Comparison

- Regularised MLR OLS and SARIMAX perform well
- MLR model obtains better accuracy
- SARIMAX model considers the seasonality aspect
- MLR model tends to **Over-estimate**
- SARIMAX model tends to **Under-estimate**

Over-Estimation:

- Misallocation of Resources
- Recruit more than necessary
- Added Costs
  - Store-keeping; Extra salaries
- Capital **tied-up** in inventory
- **Idle Capacity** due to **Over Utilization** of Resources
- **Stressful** Working Environment

Under-Estimation:

- Inefficient allocation of Resources
- Recruit less men and machines
- Inability to meet deadlines
- **Delays** in final HOTO
- **Customer Dissatisfaction**
- Added Costs
  - **Premium Pricing** at short-notice

Results and Findings

Model Comparison

Ensemble Model

- Combine the results of Regularised MLR OLS and SARIMAX
- Reduces the Over and Under-Estimation
- **Percentage Difference:** 1.07%
- Mr. Singh and Mr. Kumar were pleased with the results
- STL is testing the model for final deployment

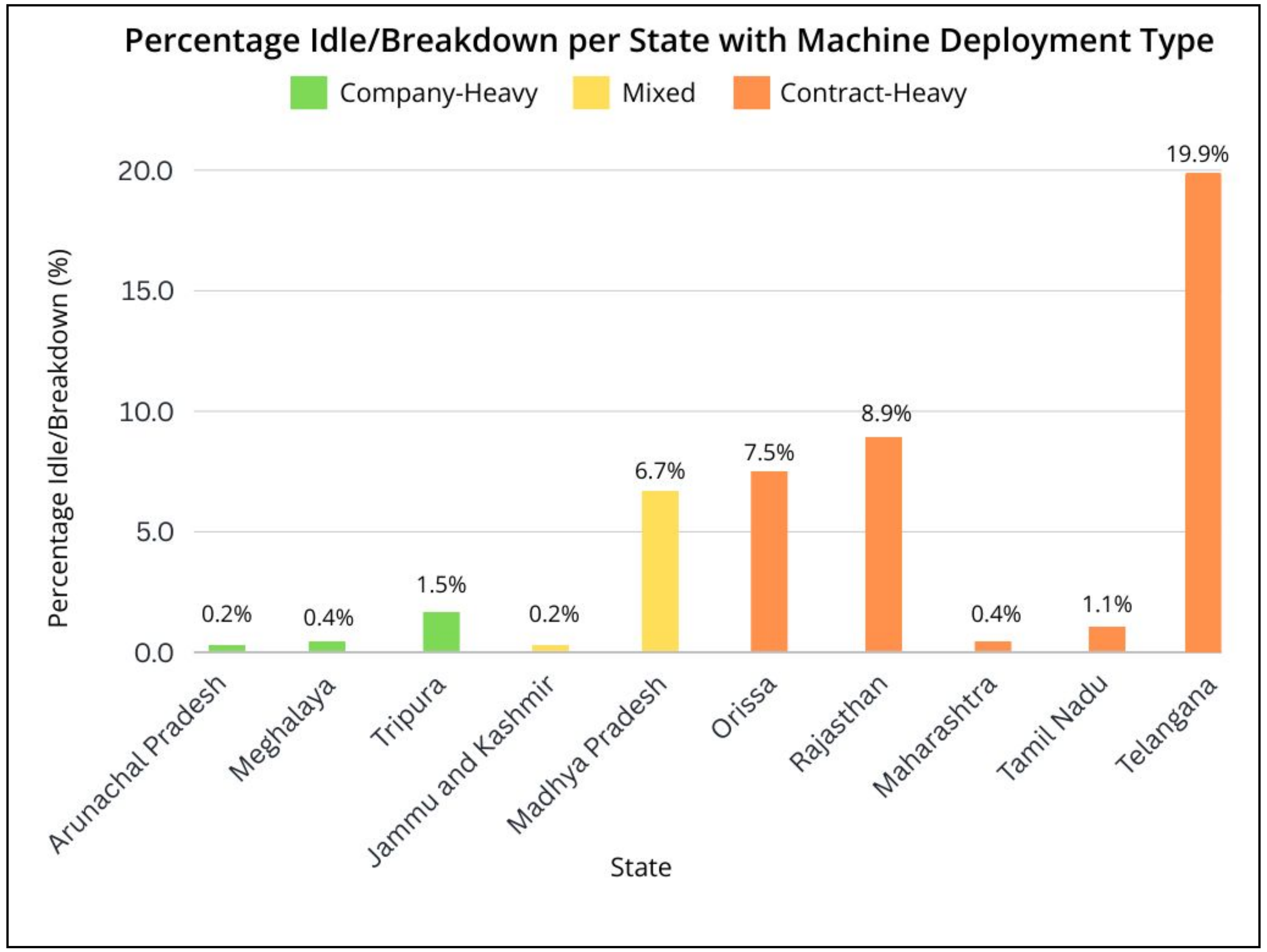
Results and Findings

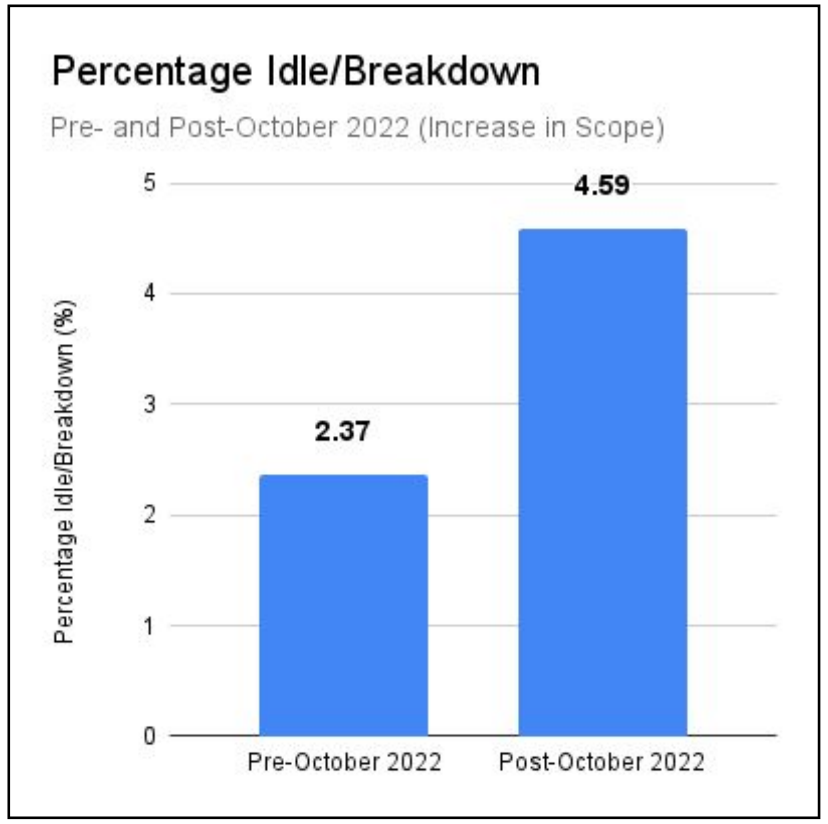
Model Comparison

Ensemble Model

Idle Capacity Analysis and Results

- 80% of machines contracted; 20% company-owned
- **Contract-heavy states** tend to fall under the trap of **over-utilization** due to **machine burnout**
- Periods of High output **immediately followed** by periods of Low output → *Erratic T&D Trendlines*
  - Suggests overuse of labour and capital
  - **2x** increase in Percentage Idle/Breakdown after signing the MP and Orissa Deals in October 2022





Recommendations

1 Forecasting Model and Data Improvement Strategies

- Ensemble Model delivers **high accuracy**
- **Robust**: Properties of both Linear Regression and Time Series Analysis
- Model can be tweaked for **weekly, monthly or quarterly** prediction
- Model is easy to **incrementally retrain** for more data
- More **localised data** (for e.g., soil conditions, climatic conditions, local authorities) could be captured

2 Prevent Over-Utilization of Equipment

- Current manual capacity planning overestimates the target
- Conduct more **vocational training** regarding machine breaks and failure
- **Monitor machine usage** either manually (field engineers) or automatically (sensors)
- Establish a **fixed work schedule**
- Enforce **maximum operating hours** (about 10 hrs/day)
- Dedicate shifts for timely **equipment maintenance**

Thank You!

Links:

- Access Related Resources (Project Reports, Project Poster, Recorded Discussions):
  - Google Drive: https://drive.google.com/drive/folders/1XElSqxikWTIhyDnVc45CHc4o4lp_Pued?usp=sharing
  - GitHub: https://github.com/Archit-Handa/Optic-Fiber-Deployment-Prediction-Model

- Contact Me:
  - Archit Handa | UG Student - BS in Data Science and Applications, IIT Madras
  - Email: archit20handa@gmail.com | LinkedIn: linkedin.com/in/archit-handa | GitHub: github.com/Archit-Handa
  - My Resume/CV: https://drive.google.com/file/d/1qNNr9vAACO7qvYVpD8zHfLU_D_TMBUEu/view?usp=sharing