Forecasting Optic Fiber Deployment for enhanced Revenue Planning

A Final Report for the BDM Capstone Project

Submitted by

Name: Archit Handa

Roll Number: 22f2000744

IITM Online BS Degree Program,

Indian Institute of Technology, Madras, Chennai

Tamil Nadu, India, 600036

Contents

1 Executive Summary and Title 3

2 Detailed Explanation of Analysis Process/Method 3

2.1 Data Collection Process 3

2.2 Data Preprocessing 4

2.3 Data Aggregation for Overall Daily Overview 4

2.4 Engineering Features - Work Duration and Daily Productivity 5

2.5 Modeling 6

2.6 Regression Analysis 6

2.7 Time Series Data Analysis 7

2.7.1 Seasonal Decomposition 7

2.7.2 Augmented Dickey-Fuller Test for Stationarity 11

2.7.3 Autocorrelation Plot for T&D Percentage 12

2.7.4 SARIMAX Model 13

3 Results and Findings 13

3.1 MLR OLS Model 13

3.2 Regularized MLR OLS Model 15

3.3 SARIMAX Model 15

3.4 Modeling Results 16

3.5 Ensemble Model 17

3.6 Idle Capacity Analysis and Results 17

4 Interpretation of Results and Recommendations 19

4.1 Recommendation 1: Forecasting Model and Data Improvement Strategies 19

4.2 Recommendation 2: Over-Utilization of Equipment 19

Declaration Statement

I am working on a Project titled “Forecasting Optic Fiber Deployment for enhanced Revenue Planning”. I extend my appreciation to **Sterlite Technologies Limited (STL)**, for providing the necessary resources that enabled me to conduct my project.

I hereby assert that the data presented and assessed in this project report is genuine and precise to the utmost extent of my knowledge and capabilities. The data has been gathered from primary sources and carefully analyzed to assure its reliability.

Additionally, I affirm that all procedures employed for the purpose of data collection and analysis have been duly explained in this report. The outcomes and inferences derived from the data are an accurate depiction of the findings acquired through analytical procedures.

I am dedicated to adhering to the principles of academic honesty and integrity, and I am receptive to any additional examination or validation of the data contained in this project report.

I understand that the execution of this project is intended for individual completion and is not to be undertaken collectively. I thus affirm that I am not engaged in any form of collaboration with other individuals, and that all the work undertaken has been solely conducted by me. In the event that plagiarism is detected in the report at any stage of the project's completion, I am fully aware and prepared to accept disciplinary measures imposed by the relevant authority.

I understand that all recommendations made in this project report are within the context of the academic project taken up towards course fulfillment in the BS Degree Program offered by IIT Madras. The institution does not endorse any of the claims or comments.

Signature of Candidate:

Name: Archit Handa

Date: 14th April, 2024

1 Executive Summary and Title

This BDM Capstone Project aims to model and streamline the 'trenching and ducting' process for laying out optic fiber cables at Sterlite Technologies Limited (STL). STL currently operates in the B2B segment and provides its clients with telecommunication infrastructure solutions. 'Trenching and ducting' is a crucial step in the capacity planning process for the firm and in turn serves a major role in the overall revenue planning for the business vertical. The major issue that STL currently faces is inability to predict with high accuracy how much optic fiber they will be able to lay in a month, and thus they are unable to effectively plan and execute their future revenue projections. Inefficiency planning leads to delays in turn around time (TAT) and resource wastage, increasing costs and indirectly affecting profitability.

To tackle this problem, in this final report, the process of data collection and preprocessing will be touched upon. Following this, Regression Analysis and Time Series Analysis will be performed using dedicated python packages. Next, the analysis will be used to train and test Forecasting Models that can aid the organization to locate, understand and, potentially, bypass factors or variables that are acting as hindrances. The report will also talk about another problem that STL is facing but did not realize (Idle Capacity and Machine Over-Utilization). The report will conclude by summarizing the results and findings, and provide recommendations for STL to inculcate in their business operations that can make the capacity planning process more streamlined and help eliminate potential bottlenecks or problems.

2 Detailed Explanation of Analysis Process/Method

2.1 Data Collection Process

As previously mentioned in greater detail in the Midterm Report, the entire process of collecting data was relatively well streamlined. Sterlite had already deployed an in-house open-source tool, 'ForceField', in which the field engineer would input details regarding the operations that were performed during the working day. These details include the **type of activity** being performed, **type of machine** being employed for the activity as well as the **count of machines** being used for the same, the **start and end time** of using the machine, details about the span they are working at (**State, CMP,** and **Span** names), the **Scope** of the span (the targeted length of optic fiber to be laid; mathematically, the sum of targets for all active spans being worked upon) along with how much **Trenching and Ducting output** they achieved that particular day. Besides this, the tool also had the ability to note down if the machine was actively **working** on the day or was laying **idle** (either not in-use or undergoing maintenance due to breakdown).

After the field engineer fills in the details, the information is sent back to a dedicated server set up by STL from which data is retrieved as an MS Excel file for further analysis.

2.2 Data Preprocessing

Once the data is requested from the API, it is not usable for model building in its current state as it is extremely raw. Thus, multiple preprocessing steps are required to be performed using Excel and python libraries like `pandas`, `numpy` and `scikit-learn`. Firstly, the 'Trenching and Ducting' (T&D) data is filtered by choosing the required activity and then all unusable features are dropped from the table. These features include details pertaining to the field engineer herself/himself as well as output details for other activities, like 'Manhole Installation' and 'DIT', which (as per Mr. Singh) do not contribute significantly to the overall capacity planning process as much as T&D does.

Post this, there were missing values in the dataset that needed to be imputed. The imputation was performed depending on what the feature tried to convey as well as what conditions do other variables portray. For example, for some entries the 'Machine Status' (machine is working or idle/breakdown) was missing; however, there was a T&D output value, signifying work had been performed. Hence, the 'Machine Status' was filled in as Working.

Another instance for imputation, was when the 'Machine Status' was Idle/Breakdown, then other attributes such as start and end times along with the output were not captured. Since the data was in chronological order, for the datetime values, missing values were filled by interpolating the other values, and for the output, the missing value was replaced with 0.

It is worth mentioning that FieldForce has captured data since October 2021; however, the initial months' data was extremely crude and noisy. Upon discussing about the issue with Mr. Singh and Mr. Kumar, it was concluded that this had been the case because the staff was not well-versed with the tool. Several workshops and vocational programs had to be organized to educate them how to operate the tool: what details are required to be filled and under which field. Thus, the data for the first three months is not used for analysis and model building, i.e. data post 1st January, 2022 is considered.

2.3 Data Aggregation for Overall Daily Overview

Even though the data is cleaned now, it is still not usable for the purpose of forecasting. As Mr. Singh had requested, he wished to have a model where the certain inputs like machine type and count are fed in and he can receive a forecasted prediction for total T&D output for the next month. This requires the data to capture a daily snapshot of the work performed in the entire country.

On the other hand, the current data was a daily snapshot of each individual span being worked upon and that too for each separate machine type. Though such data provides granularity for further analysis, for the given request of Mr. Singh, such minute details will act as noise causing the model to overfit. Thus, daily data for the individual spans and machine types had to be aggregated using advanced pandas operations and methods. The aggregation was performed for each machine type as well as overall to

further check performance of each machine type individually. This is discussed further in the following section.

2.4 Engineering Features - Work Duration and Daily Productivity

While the dataset is now ready for a simple baseline modeling, yet for in-depth analysis regarding the performance measures of the machine types being used, two new features had to be engineered: **Work Duration** and **Daily Productivity**.

Work Duration was calculated using the formula below:

$$Work\ Duration\ =\ Machine\ End\ Time\ -\ Machine\ Start\ Time$$

On the other hand, Daily Productivity, as discussed in the midterm report, was calculated as:

$$Daily\ Productivity\ =\ \frac{Daily\ Output\ (in\ meters)}{Machine\ Count}$$

These measures will be used ahead to check how the machines tend to perform throughout the year.

The Figure 1 below shows a snapshot of the data before and after preprocessing and feature engineering. In the first table, the 'FE_Name' (Field Engineer Name) column has been redacted for privacy reasons.

Figure 1: Snapshots of Data (Above) Data as retrieved from the FieldForce API (unsuitable for modeling) (Below) Data post processing and feature engineering (usable for modeling)

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

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

The data now is ready for modeling and in-depth analysis.

2.5 Modeling

Please note that for all models that are created in this project, the data for the last 2 months (February and March 2024) has been kept for testing; the models are trained on the rest of the data (data from January 2022 to January 2024).

2.6 Regression Analysis

Since the problem at hand involves predicting the T&D Output (a continuous numerical variable), ***Regression Analysis*** seems the most appropriate statistical modeling method to begin with. Using the `statsmodels` api for Python, an `OLS` (Ordinary Least Squares) was fitted as a MLR (Multiple Linear Regression) model.

Since the ultimate aim is forecasting, the inputs to the model can only be variables that Mr. Singh has control over. These included the count for each machine type, overall target scope, and the month for which prediction needs to be made.

The MLR OLS model was trained and predictions were made on the test dataset as shown in the Figure 2 below.

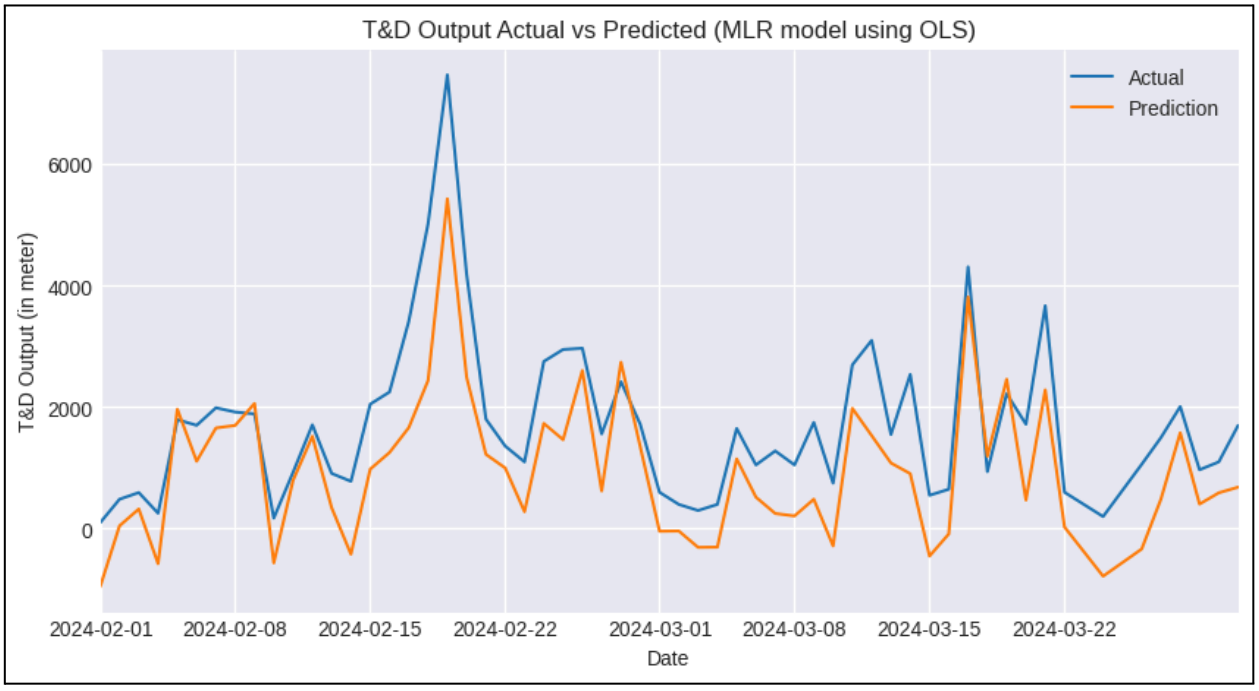


Figure 2: Actual vs Predicted T&D Output values using MLR OLS for Test Months

Results from this model are discussed in the 'Results and Findings' section ahead. It is also mentioned there that this model was not reliable for predicting, and thus, arose a need to regularize the same.

Upon applying L2-regularization on the MLR OLS model with a regularization strength of $\alpha = 10$, the best MLR OLS model was achieved.

The prediction results for the same are shown in the Figure 3 below.

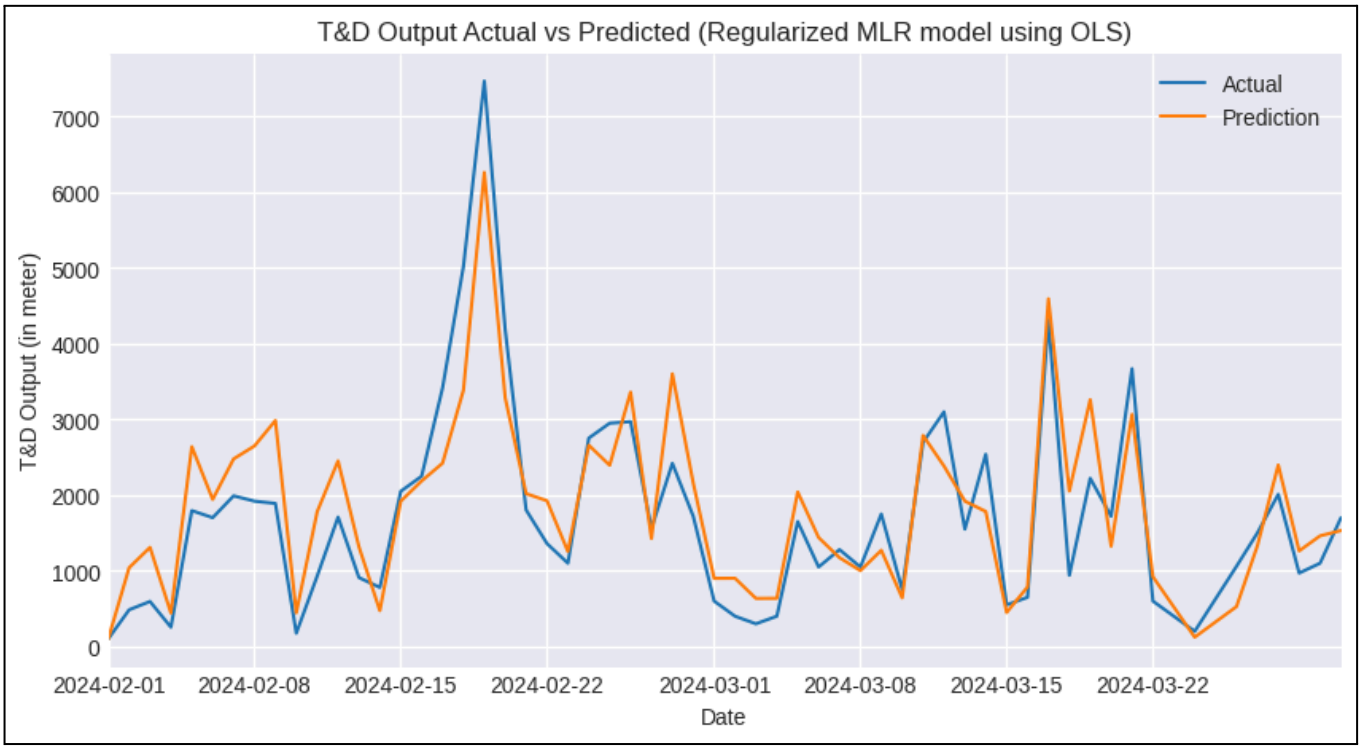


Figure 3: Actual vs Predicted T&D Output values using Regularized MLR OLS

While the predictions do look similar for both the models, the regularized model was much more accurate than the previous one. Nevertheless, it still had a minor imperfection: it could not account for the effect of seasonality during prediction. This is further talked about in the 'Results and Findings' section ahead.

Due to this flaw, arose a need to perform ***Time Series Analysis***.

2.7 Time Series Data Analysis

2.7.1 Seasonal Decomposition

As aforementioned, the regularized MLR OLS model seems quite accurate; however, it still does not account for the seasonality effect on the T&D Output. As per Mr. Singh, since the entire process of 'Trenching and Ducting' is physically-intensive in terms of both machine and labor, conditions like weather and festivities can play a major impact on how productive a month is.

As previously touched upon in the midterm report, monsoon weather and hurricanes (especially for Orissa and other coastal states) can make conditions inoperable for man and machines. Additionally, the month of October tends to be less productive since labor tends to take off to celebrate festivals like Diwali. This is evident through the graph in the Figure 4 below, which showcases productivity as both actual meters of output as well as percentage of scope.

Periods from April to June tend to be most productive, followed by a sharp decline with the onset of rains in July. October is least productive, and a decline during the harsh winters of January.

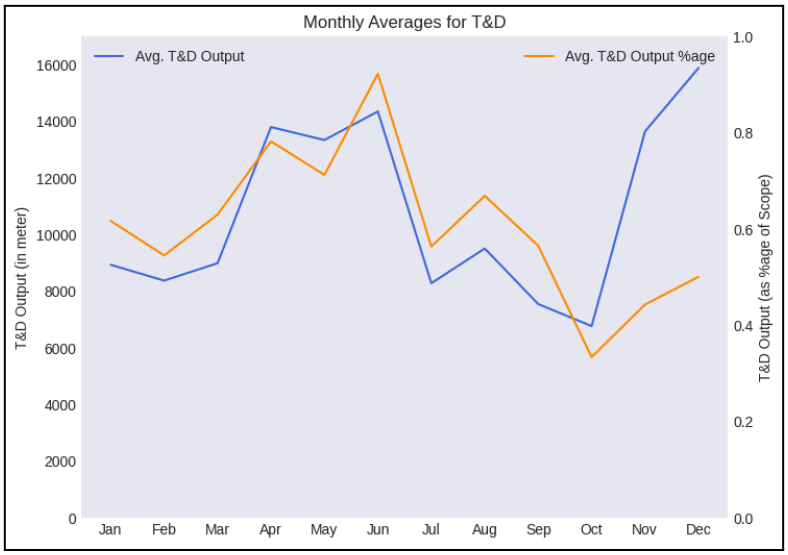


Figure 4: Monthly Average of T&D output (Meter and Percentage of Scope)

This serves as evidence that certain trends due to time, which are not considered by the Regression Models, need to be included in the forecasting model. Thus, justifying the need for Time Series Analysis.

Firstly, the original series is plotted and decomposed into the trend, seasonality, and residuals (or noise) components using `seasonal_decompose` function from the `StatsModels` python package. The results are visible in the Figure 5 below.

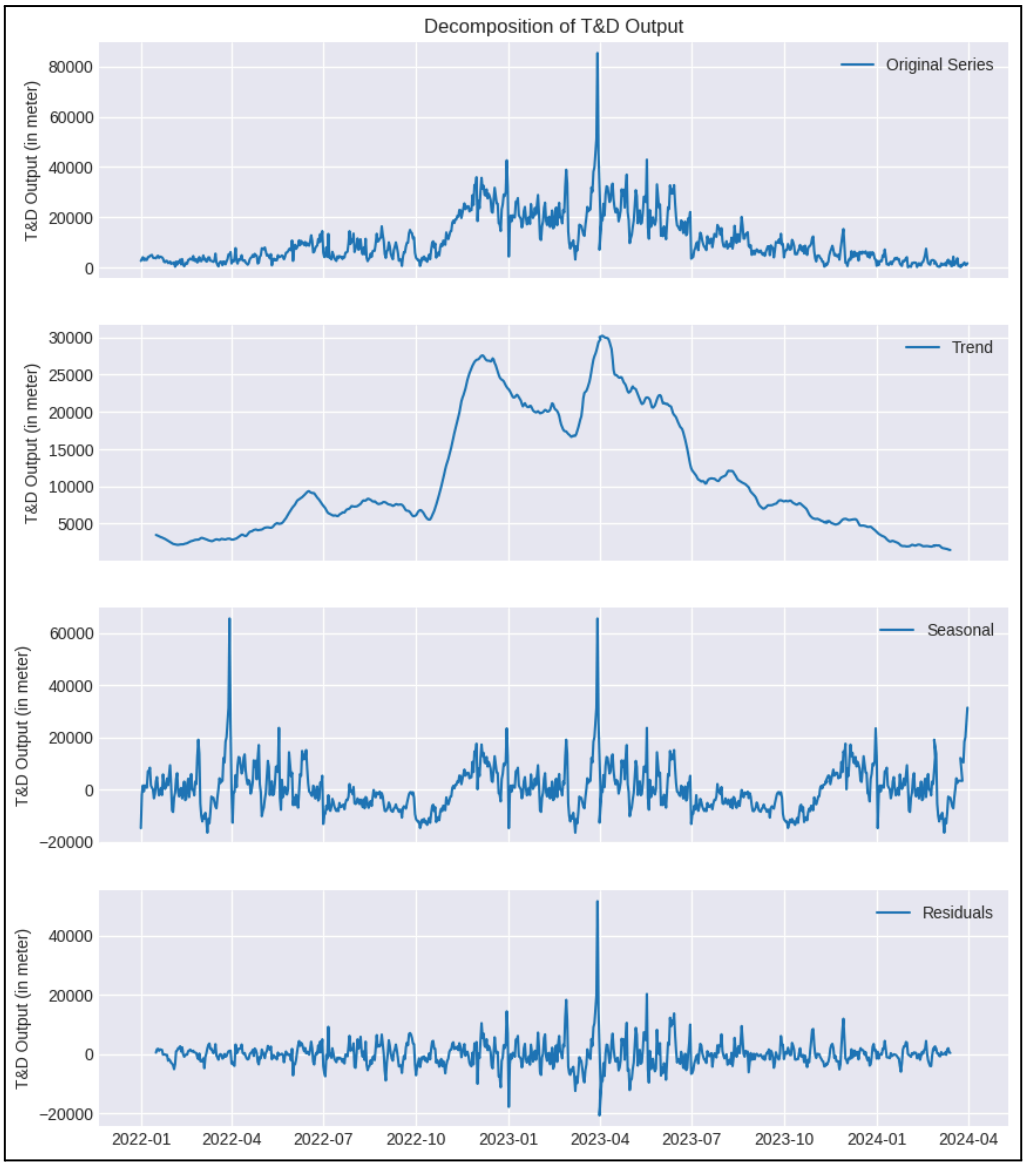


Figure 5: Decomposition of the Actual T&D Output
(Top) Actual Series; (Second from Top) Overall Trend of the Series;
(Second from Bottom) Seasonality Effect considering a 365-day period
(Bottom) Remaining Error/Residuals/Noise in the Series

The results from the graph are as follow:

- **Seasonality Results**: From the first look, the subplot for seasonality is not the exact same as the seasonality effect shown in Figure 5 above; however, it must be noted that Figure # plot was a monthly average. Moreover, both the plots do capture the cyclical nature of seasonality with peaks in April, overall high productivity till June, followed by a fall till October (lowest), followed by an increase in November and December, and finally a slight downfall from January to March. This is better visible from the Moving Average plotted considering a 30-day period (for monthly average) in the Figure 6 below.

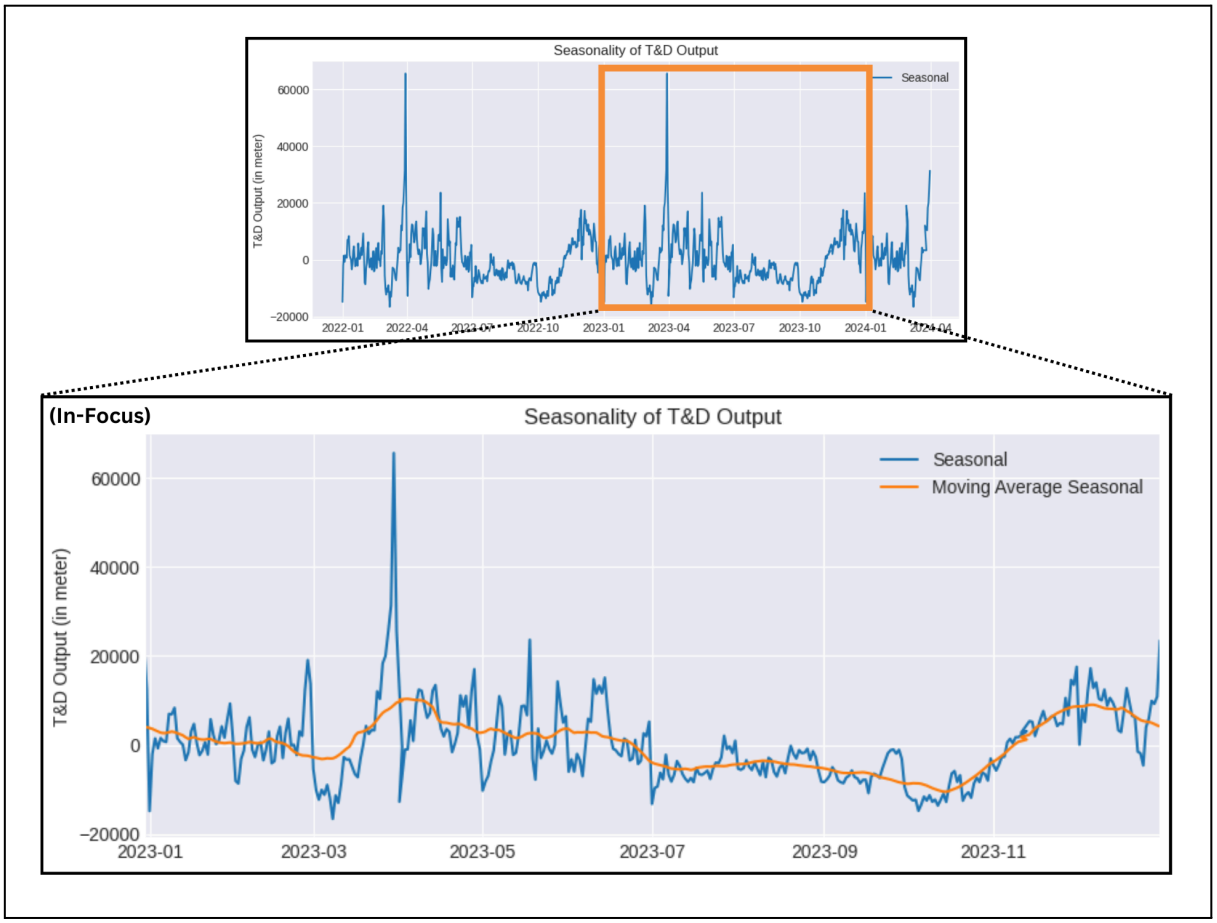


Figure 6: (Top) Seasonal Decomposition of the Actual T&D Output from Figure 5; (In-focus) Seasonality for a Year, repeats itself in a Year with its 30-day Moving Average

Do note, the orange box in the top plot shows that the seasonal effect repeats each year and thus findings are found using any year (2023 in this case) as a reference, as shown in the bottom plot.

- **Trend Results**: In the Figure 5 above, one can see the upward shock in the overall trend from October 2022. The plot then plateaus till June 2023 and finally tapers down. Upon enquiring Mr. Singh

regarding this unusual behavior, he highlighted that STL had signed a deal for multiple spans in Madhya Pradesh and Orissa in October 2022 that increased their overall target from roughly 2.000 km to over 5,900 km. This justifies the upward shock. For the tapering down section post-June 2023, he pointed out that only a few spans have been signed since then to prevent any ***backorders*** as STL is trying to finish its pending commitments.

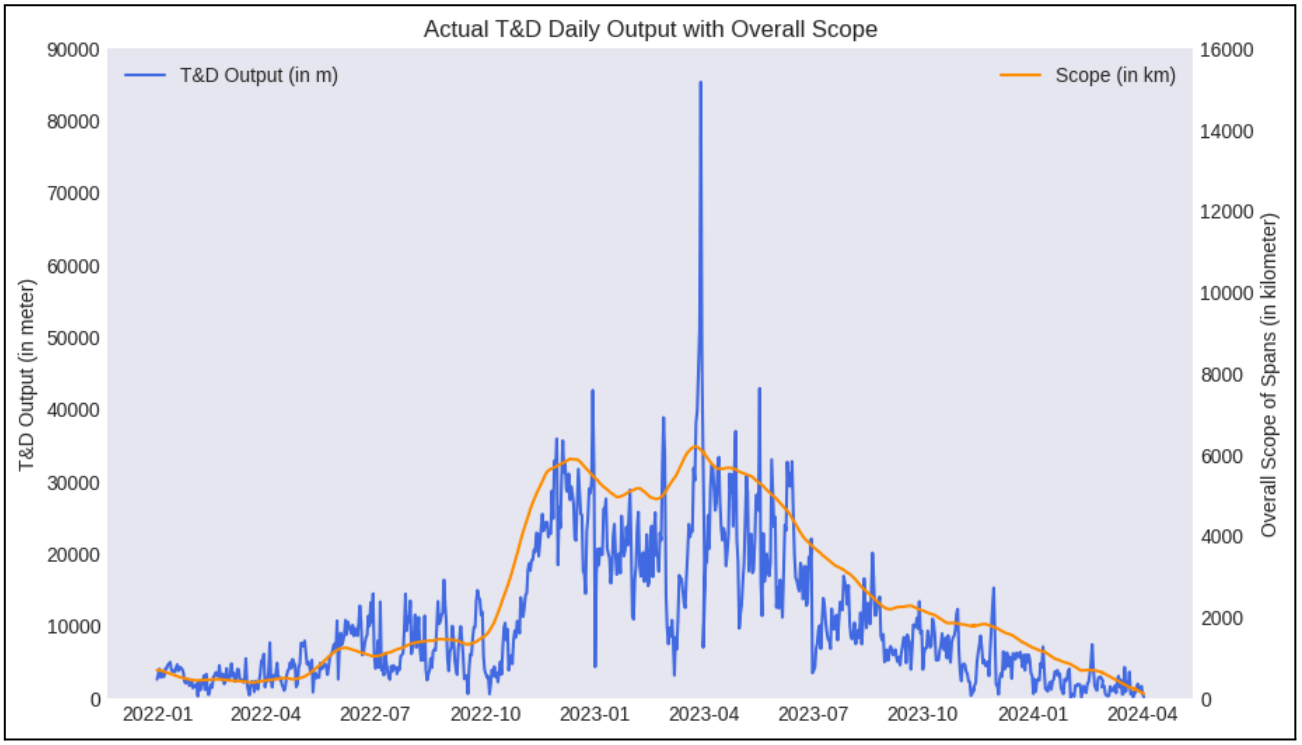


Figure 7: Graph of Actual T&D Output with the Overall Scope

Given the impact of orders on the Actual Output, a graph was plotted for both actual T&D output and the Scope of active spans as seen in Figure 7 above. Overall Scope essentially monitors the cumulative target for all active spans and thus, is the best feature to monitor the impact of customer orders. The graph shows a strong correlation between T&D output and Scope. The correlation coefficient was also calculated between the two features using the formula:

$$Correlation\ Coefficient\ (\rho) = \frac{\Sigma(x-\bar{x})(y-\bar{y})}{\sqrt{(\Sigma(x-\bar{x})^2)(\Sigma(y-\bar{y})^2)}}$$

where x and y correspond to T&D Output and Scope respectively. The value for the coefficient turned out to be 0.838 (83.8%) suggesting that the two features are highly correlated. Therefore, Scope can provide us with better forecasting ability.

- **Residuals Results**: As seen in the Figure 5 above, the residuals also tend to spike during shocks due to Scope. Yet overall, residuals exhibit ideal properties of noise: near-0 mean and consistent variance (fluctuating near 0) and no discernable pattern (if a pattern existed, it would have meant that the decomposition was not proper and certain aspects had been overlooked, which is not the case here).

2.7.2 Augmented Dickey-Fuller Test for Stationarity

For time series data, models tend to perform better when the data is stationary. Stationarity implies that the statistical features such as mean and variance do not vary over time. However, considering the T&D output as it is does not meet this requirement, quite evident from the shocks due to Scope. Thus, to better incorporate the effect of Scope and make the underlying data stationary, a T&D output as a percentage of Scope is used instead for modeling. The new target is computed as follow:

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

The new target trend is shown in the Figure 8 below.

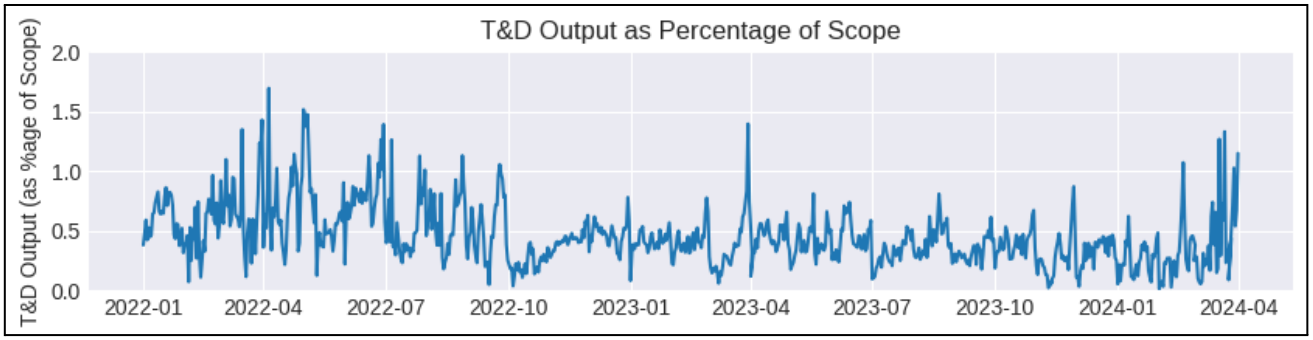


Figure 8: Graph of T&D Output as Percentage of Scope

Graphically, T&D output as percentage of scope (T&D percentage) seems more consistent and no unusual shocks are present like in T&D output. Yet, it is worth mentioning that post-October 2022 the 'noisy' behavior of the T&D percentage does reduce and tends to smoothen out. This is further discussed in sections ahead.

To statistically test stationarity, an **Augmented Dickey-Fuller Test** was run on T&D percentage using the `adfuller()` function in the `StatsModels` python package.

The test is performed with the following null and alternative hypotheses:

- **Null Hypothesis (H_0)**: A ***Unit Root*** (a characteristic of time series data) is present in the series which implies non-stationarity.
- **Alternative Hypothesis (H_A)**: A Unit Root does not exist which implies stationarity.

The value for the test statistic computed was -5.596 with an extremely small P-value of 8.296×10^{-5}. The critical value at 1% Significance Level was -4.386. Since the p-value is lesser than even 0.01 (α = 1%), the null hypothesis can be rejected and the alternative can be adopted, stating that stationarity exists in the T&D percentage.

2.7.3 Autocorrelation Plot for T&D Percentage

Next step is to understand which time series model (ARIMAX or its seasonal variant SARIMAX) would be ideal for training and forecasting. To aid in this decision-making process, an **Auto-Correlation Function** (ACF) Plot for T&D Percentage is plotted, as seen in Figure 9 below. An ACF plot highlights how much a day's T&D percentage is related to a previous day's percentage.

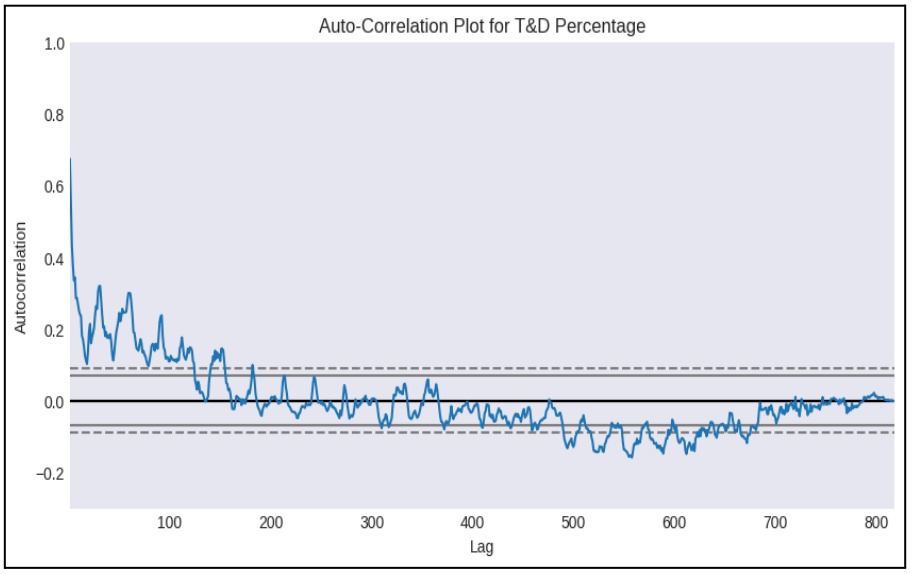


Figure 9: ACF Plot for T&D Percentage

The plot does exhibit cyclical behavior suggesting that a **SARIMAX** (**S**easonal **A**uto-**R**egressive **I**ntegrated **M**oving-**A**verage with e**X**ogenous variables [exogenous for inputs like machine counts]) model would be better suited for modeling.

To determine its parameters, further **Partial Auto-Correlation Function** (PACF) and ACF plots are graphed for a more focused lag value of 40 than close to 800. A PACF plot draws correlation between the time series data and the lag (similar to ACF) but ignores the effect of intermediate lags.

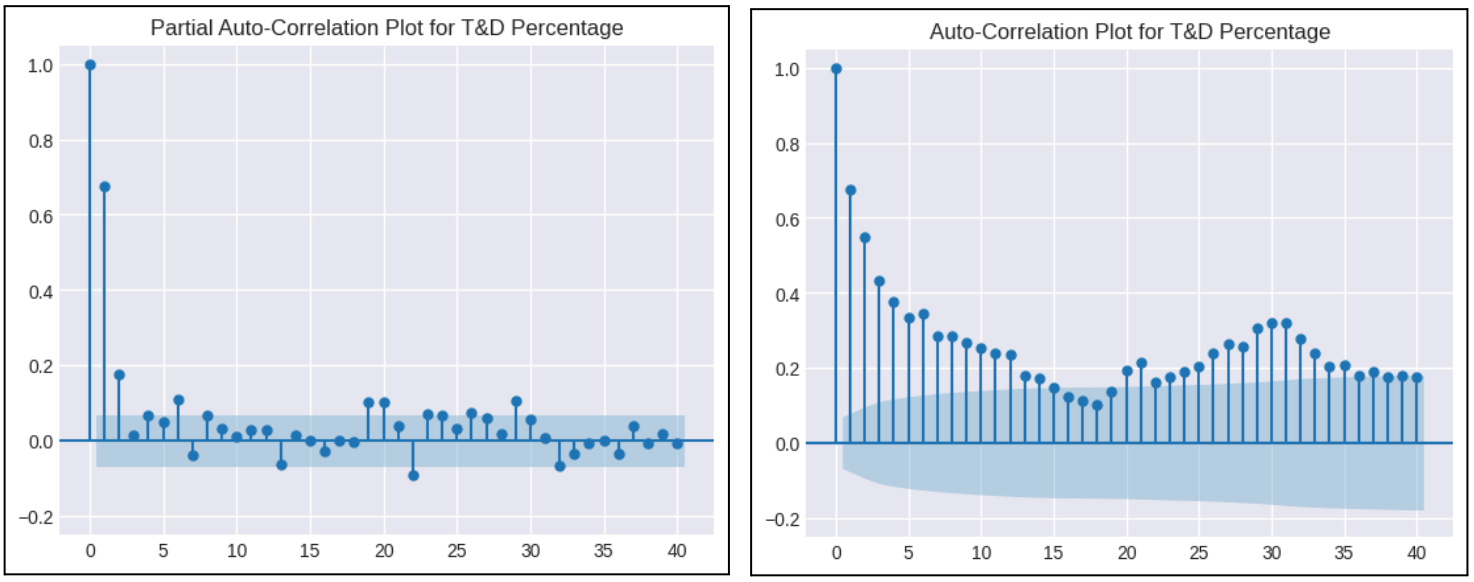


Figure 10: PACF and ACF Plots for T&D Percentage (Close-up view for 40 Lags; Blue Shaded Region shows the Significance Limits)

From the PACF, the first 3 lags are out of the significance limit (shown by the blue shaded region) and thus a good upper-limit approximation for the Auto-Regressive parameter '*p*', which determines the

memory of the model (how far to look for making predictions). From the ACF, the first 15 lags are significant and serve as a good upper-limit estimate for the Moving-Average parameter '*q*', which defines how many errors of the past can help in predicting for the future. For the Integration parameter '*d*', since the data already exhibits stationarity, a maximum value of 1 will work appropriately.

2.7.4 SARIMAX Model

Now that the upper limits for the parameters are known, modeling can begin. As mentioned earlier, the SARIMAX model is being trained. To determine the best model parameters, `auto_arima` object from the `Pyramid-Arima` python library is used.

Upon training the `auto_arima`, the best model turned out to be a `SARIMAX(3,1,1)x(1,0,[],30)` model[1]. The predictions made using the same are displayed in Figure 11 below. The exogenous variables provided alongside the T&D percentage values included machine counts for each type.

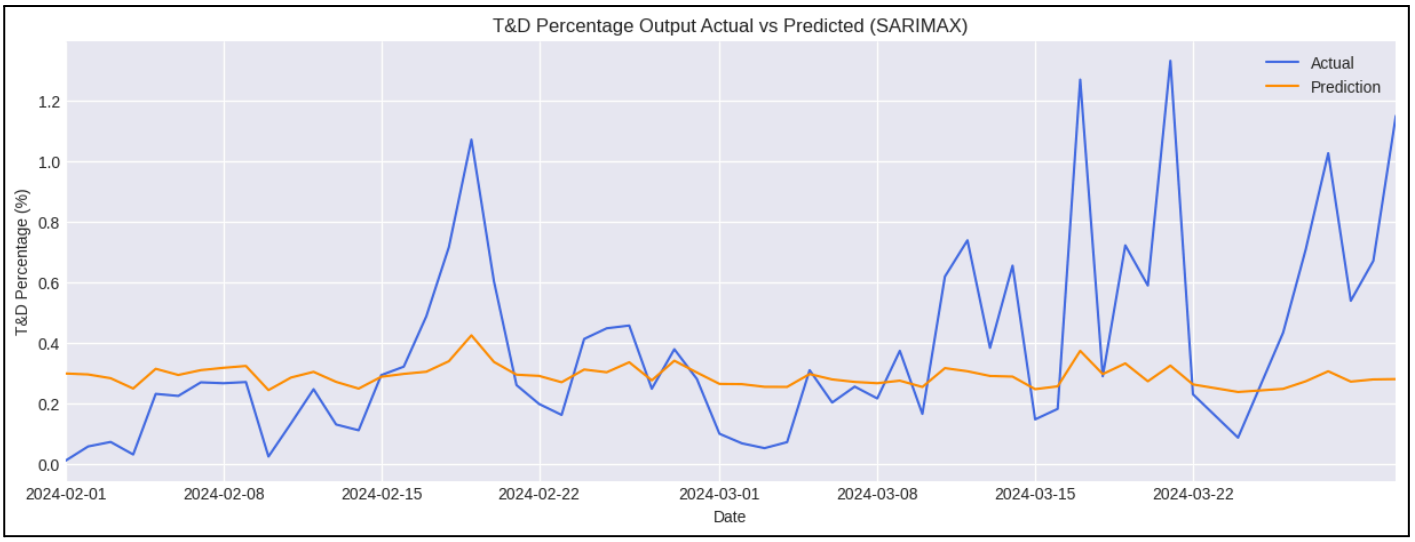


Figure 11: Actual vs Predicted T&D Percentage values using SARIMAX

Even though the model does not vary so much as the MLR OLS models do, the prediction does float around the mean of the test data.

Once the T&D Percentage is predicted, the T&D Output can be computed using the following formula:

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

3 Results and Findings

3.1 MLR OLS Model

Figure 12 below presents the Regression Results for the unregularized MLR OLS Model. The red box highlights the F-value and the corresponding P-value (Prob (F-Statistic)). Since the P-value is near 0, it

[1] `SARIMAX(3,1,1)x(1,0,[],30)` are the parameters for AR, I, MA and Seasonal AR, I, MA and Period respectively.

signifies that the regression model is **significant** even for a significance level of 1%. Moreover, five of the six regressors have a p-value close to 0 (as shown in the blue box) portraying that those five variables (**Machine Counts** and **Scope** in km along with the **intercept** term) are highly significant to make a prediction. Unfortunately, for **Month** the p-value is 0.646 which is much greater than a 10% significance level (0.1). This in turn means that the regression model does not consider the Month variable, and in turn, the seasonality aspect during prediction (this aspect has been previously considered during time series analysis).

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

Figure 12: Regression Results for MLR OLS

As shown previously in Figure 2 above and again as a snapshot in Figure 13 below for quick reference, predictions were made for the months of February and March 2024.

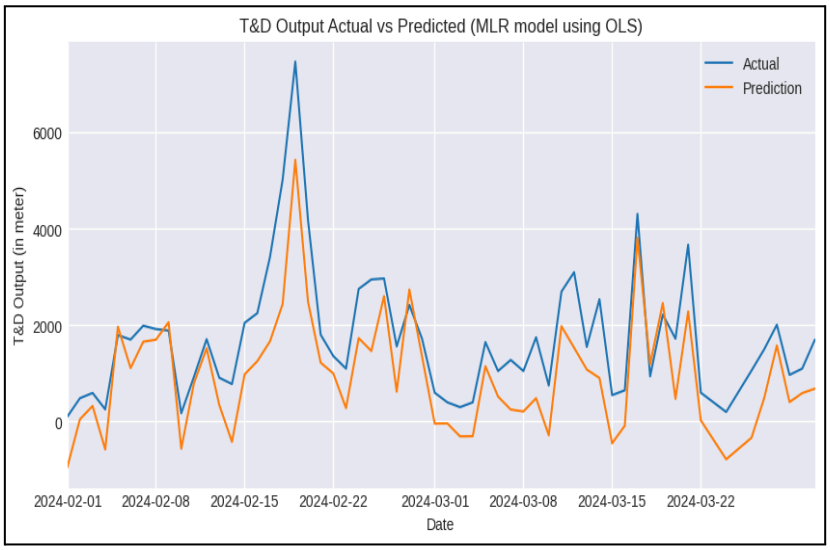


Figure 13: Actual vs Predicted T&D Output values using MLR OLS for Test Months

The figure shows that the model is able to predict the spikes and drops in the plots accurately; however, the model does predict negative values for T&D output, which are not reasonable. Moreover, the percentage difference between the actual and predicted output turned out to be 44.53% using the formula:

$$\%\ Difference = \left| \frac{\Sigma\ Actual\ Daily\ Output - \Sigma\ Predicted\ Daily\ Output}{\Sigma\ Actual\ Daily\ Output} \times 100\% \right|$$

3.2 Regularized MLR OLS Model

Moreover, from a statistical point of view, the greenbox in the Figure 12 above highlights the R2-Score (scoring metric for Regression Analysis) being too high (0.962) while the value computed for the test set turned out to be just 0.478. This shows that the model is overfitting. Thus, arose the need to further refine the model by performing **regularization**. The prediction results for the regularized model are shown in Figure 14 again for quick reference.

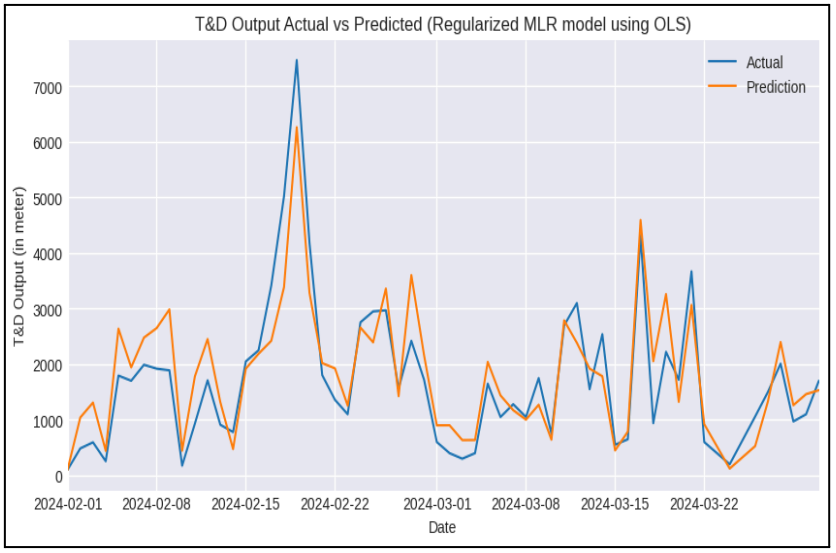


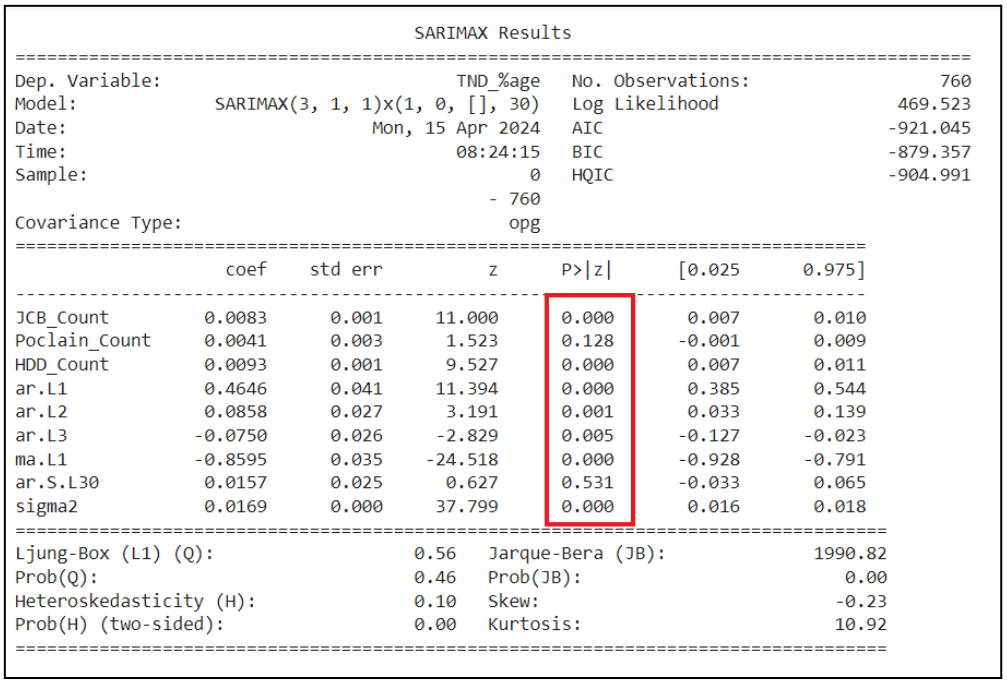
Figure 14: Actual vs Predicted T&D Output values using Regularized MLR OLS

This model fares much better than the previous one. It does not predict negative values as output and also the R2-scores for the train and test sets are much more comparable (0.958 vs 0.807 respectively). The sum of the actual T&D outputs for the 2 months turned out to be 1,00,582.00 m (100.58 km), while the predicted sum was 1,06,923.40 m (106.92 km). Thus, the percentage difference between actual and predicted outputs has decreased from 44.53% to a significantly lesser 6.3%.

3.3 SARIMAX Model

Figure 15 below shows the Results Summary from the SARIMAX Model. As evident from the p-values highlighted in the red box, except 2 features, all other features are significant for prediction (since p-value is less than 1% significance level).

Moreover, the predicted T&D Output computed using formula mentioned earlier turned out to be 92,095.06 m (92.10 km), which has a percentage difference of 8.44% (slightly greater than the Regularized MLR OLS).



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	HQIC	-904.991
	- 760		
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

Figure 15: Summary Results for SARIMAX

3.4 Modeling Results

The Regularized MLR OLS and SARIMAX Models perform comparable to each other. While SARIMAX considers the seasonality aspect, the Regularized MLR OLS does obtain a better accuracy. However, it is also worth noting the nature of predictions these models make.

The SARIMAX predictions tend to ***under-estimate*** the actual values. From a business perspective, such a situation can have severe detrimental impacts:

- Underestimating the output might cause STL to not be able to fulfill contracts on-time and severe delays in the final HOTO (Hand-Over Take-Over). This may cause customer dissatisfaction, hampering the reputation that STL has earned over the years.
- A situation might arise where more machines are needed to fulfill the customer contracts. In order to do so, STL might need to place a machine order on an immediate basis. Mr. Singh further pointed out that these short-notice contracts usually come at a premium pricing, which will act as an extra cost for STL eating their profit margins
- Overall, underestimation is a situation to avoid since it can potentially disrupt STL's plans and budgets. Furthermore, inefficient allocation of resources, including equipment, machine, and manpower, are likely scenarios that can further add onto STL's costs.

On the other hand, the Regularized MLR OLS predictions tend to ***over-estimate*** the actual values. Such a situation can again have adverse effects for the business:

- Overestimating the output can also cause misallocation of resources; however, this time, more men, machines, and raw materials will be contracted than necessary. Thus, it will add onto the cost for STL.

- Moreover, there is a high chance for many machines to not be used, **Under-utilization of Resources**, which may lead to **Idle Capacity**, and hence, reduced productivity.

- STL's capital could also get **tied-up** in inventory (excess raw material bought) for T&D. Furthermore, this may result in wastage if these resources cannot be used later. Also, STL would have to bear the **store-keeping costs** for these additional resources.

- Lastly, overestimation can create a **stressful** environment for the employees, which can **bring their morale down**, especially if they are asked to **work overtime** to achieve unrealistic targets.

Thus, both situations that arise from SARIMAX (Underestimation) and Regularized MLR OLS (Overestimation) are not suitable from a business perspective. However, interestingly, since one is underestimating while the other is overestimating, their average is likely to be a better estimate and mitigate the ill-effects of both the situations.

3.5 Ensemble Model

Thus, the final model devised for the forecasting is an ensemble created from both SARIMAX and Regularized MLR OLS. Thus, the final prediction for the test case is 99,509.23 m (99.51 km) with a percentage difference of 1.07%. This indicates that the ensemble is best for prediction.

Mr. Singh and Mr. Kumar were presented with these results and they were quite pleased with the model. In fact, Mr. Kumar has made a prediction using the ensemble for the current month of April 2024 and will compare it with the actual results he obtains at the end of the month.

3.6 Idle Capacity Analysis and Results

Besides developing a forecasting model, an interesting topic that arose from one of the discussions was about **Idle Capacity**, when machines are not active (Machine Status is 'Idle/Breakdown').

The ownership of the machines deployed is either STL's purchased company machines or machines hired on contractual basis. According to Mr. Singh, about 80% of the machines are contracted. Upon being asked why these two channels are opted, Mr. Singh answered that there are 3 main reasons:

- Firstly, contracting machines is significantly cheaper than purchasing them.

- However, contractors do sometimes ask for surge prices; and thus, STL has purchased a small fleet of such machines.

- Most Interestingly, contractors are not available at each span location. For instance, for certain spans that lie in remote locations of Arunachal Pradesh and J&K, such heavy-duty machines cannot be procured easily. These are the sites where STL has to deploy its own machines.

Based on insights gained from Mr. Singh and Idle/Breakdown machine status data, the following Figure 16 was created. It shows that Contract-Heavy states do tend to fall under the trap of under-utilization.

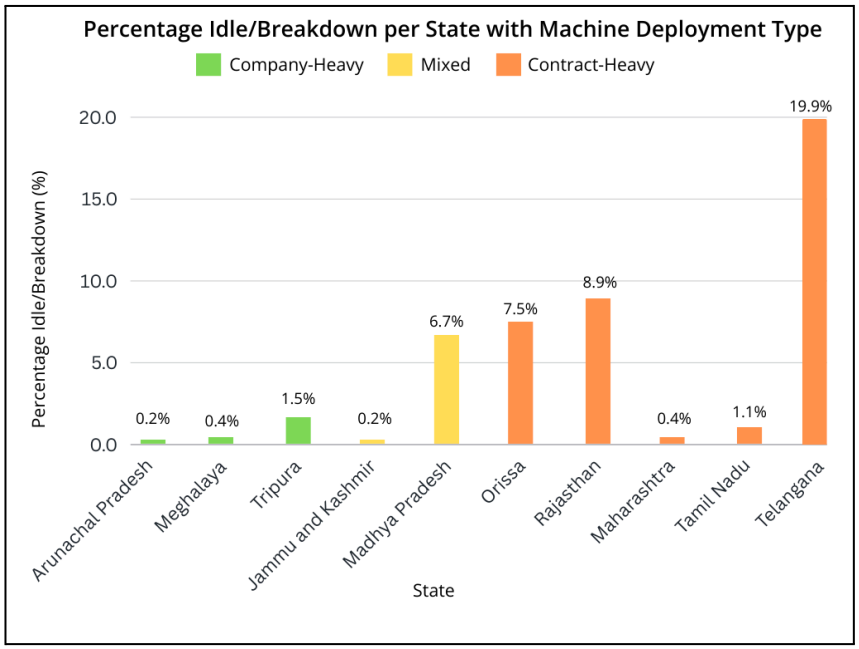


Figure 16: Percentage Idle/Breakdown per State given the Machine Deployment Type

Interestingly, from the actual trend of T&D Output as in Figure 5 above, one can notice the extremely noisy behavior of the plots. This suggests that a period of high output is followed by a period of low output especially when the scope increased in October 2022. A possible reason for this could be that machines were overused as employees wanted to achieve targets quickly; however, due to overuse, machines had to be taken for repair more frequently and thus, are noted as being idle. This is further showcased by the increase in Percentage Idle/Breakdown after October 2022 as shown beside. The Percentage Idle/Breakdown has roughly doubled after Scope increase in October 2022.

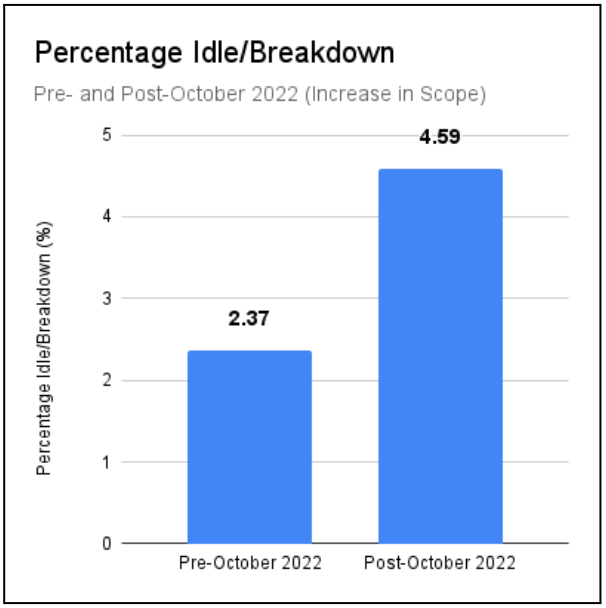


Figure 17: Percentage Idle/Breakdown Before and After Scope Increase in October 2022

4 Interpretation of Results and Recommendations

From the results discussed above, the following recommendations can be suggested to STL:

4.1 Recommendation 1: Forecasting Model and Data Improvement Strategies

The final ensemble model delivers high accuracy with a percentage difference of just (1.07%). Also, as mentioned above, Mr. Kumar is testing the model's real-world accuracy for the month of April 2024. The Ensemble model seems robust as it incorporates the properties of both regression and time series analysis. It is able to capture the impact of different variables such as machine count and type along with the trends that have been portrayed in the historical data. It is also able to handle shocks such as those caused by an increase in Scope.

An added advantage of the model is that it predicts daily output and then cumulates to provide a monthly output. This means that the model can be tweaked to forecast for different time periods, such as weekly, bi-weekly or even quarterly outputs.

Nevertheless, like all Machine Learning and Statistical Models, this model too needs to be retrained as and when new data arrives. The more data it churns, the more accurate will the results be. Moreover, to track seasonality better, more factors can be tracked, such as 'Rainfall' and its effect on the 'Soil Conditions' such as moisture content and rock content. Incorporating fields corresponding to such variables within their ForceField tool can help bring about a greater understanding of the local systems.

Moreover, as brought about in an earlier discussion, the locality of the span also plays a crucial role not only in terms of local soil conditions but also the local authorities, such as the Forest Department and Local Markets, from which approval must be taken before any work is performed. Such information regarding spans can be noted into the system during the 'Survey' step.

Therefore, employing the model for Capacity Planning but also remembering to retrain it with newer data and better data (more factors incorporated) can be extremely useful for STL to plan and deploy the machines and labor at their spans accordingly.

4.2 Recommendation 2: Over-Utilization of Equipment

As mentioned in the 'Idle Capacity Analysis and Results' section above, it is evident that machines and hence, labor, are being overworked, as suggested by the increase in Idle Percentage. This is perhaps because employees are trying to achieve targets that have been overestimated from the currently-used manual Capacity Planning strategy.

By using the model to forecast targets, STL will be able to mitigate this issue significantly. However, STL should also deploy certain strategies to prevent such a problem from other aspects.

- STL can conduct more vocational and educational workshops to help the employees understand under what circumstances the machine should be stopped to prevent breakdown.
- To enforce this, STL can further incentivize appropriate usage of machines by recognizing and rewarding employees.
- STL can instruct the Field Engineers to monitor the machine usage. More sophisticated sensors to detect operating hours and machine temperatures can also be used.
- Currently, there is no fixed work schedule that has been established by STL. STL can enforce clear operating guidelines and shifts that instructs employees to adhere to maximum operating hours and load capacities. While shifts are dedicated for timely maintenance, these can be further streamlined.

In conclusion, by employing the model developed as a part of this BDM Capstone Project as well as implementing the suggestions stated in the recommendations, STL can perform Capacity Planning, and hence, generate Revenue Projections in a more timely manner and that too which can yield better forecasting results. It can, thus, allocate its resources more efficiently and in turn maximize profits. It can also mitigate the problem of machine over-utilization by taking preventive measures for the same.