Team: Detectives
Members: Bandaru Kundana Sri, Venkatesan S.

Datathon Challenge:

Team Name:	Detectives
Project Title:	**Context-Aware Deep Learning for Proactive Network Anomaly Detection and Congestion Forecasting**
Focus:	Inference Model (Anomaly Detection) and Predictive Model (Congestion Forecasting) with External Data Integration

Abstract:

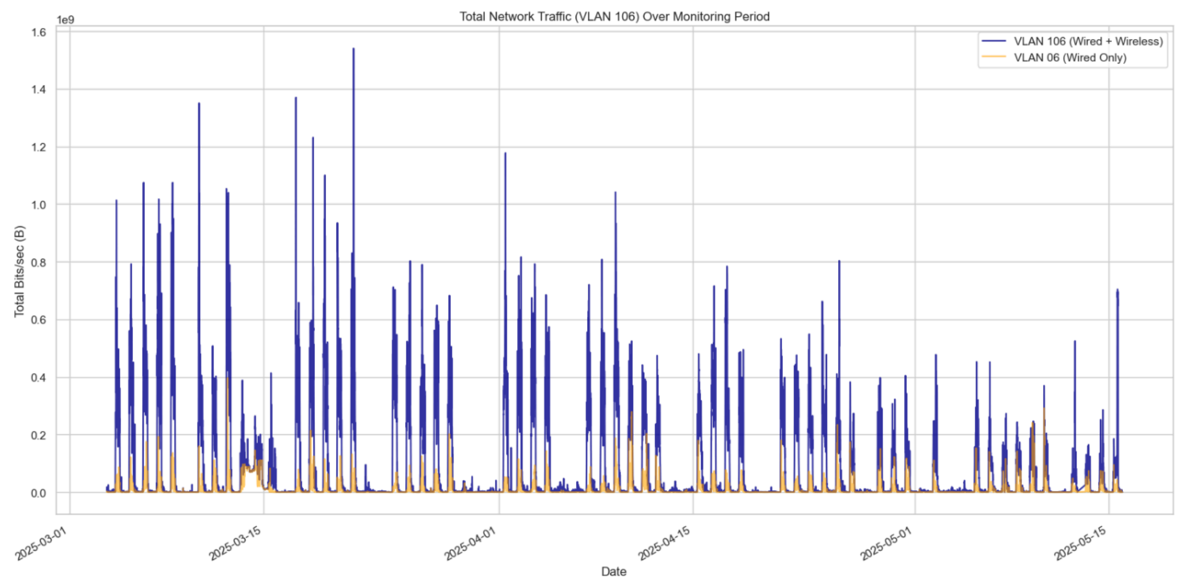
This project leverages a **data-centric** approach and **time-series network data** for **dual-output deep learning**. **STL decomposition** isolated **Trend**, **Seasonality**, and the **Residual anomaly signal**. Key features engineered include **Packet Ratio** and **VLAN-Specific Signatures**.

The **Sequence-to-Sequence LSTM** serves as the core. The **Predictive Model** forecasts **network congestion** (high **bits/sec** and **packets/sec**) using a **Congestion Severity Index (CSI)** for **preemptive SDN resource allocation**. The **Inference Model** uses **Isolation Forest** on residuals for **anomaly classification** (**DDoS** vs. **Massive Download**).

External Dataset Integration utilizes the **Academic Schedule** to validate the **correlation hypothesis** and refine forecasts during **Finals week** and **Break** periods.For **Extendibility** we used SDN-Campus dataset as external dataset representing application level activity (ie, how much traffic occurred in the network).Deliverables include a **robust Anomaly Alert System** and a **Congestion Forecast Dashboard**.

EDA performed and Analysis:

1. Total Network Traffic (VLAN 106 vs. VLAN 06)



- **Observation:** VLAN 106 (Wired + Wireless) shows dramatically higher peak traffic than VLAN 06 (Wired Only), with peaks reaching up to **1.6×109 bits/sec (≈1.6 Gbps)**. VLAN 06 peaks are generally much lower.

- **Analysis:**
  - **VLAN 106 is the Congestion Target**
  - **Strong Weekly Seasonality:** The highest traffic peaks occur predictably every 7 days (or 5 days for the academic week). Notice the **troughs (weekends)** are clearly visible, the weekends dates are

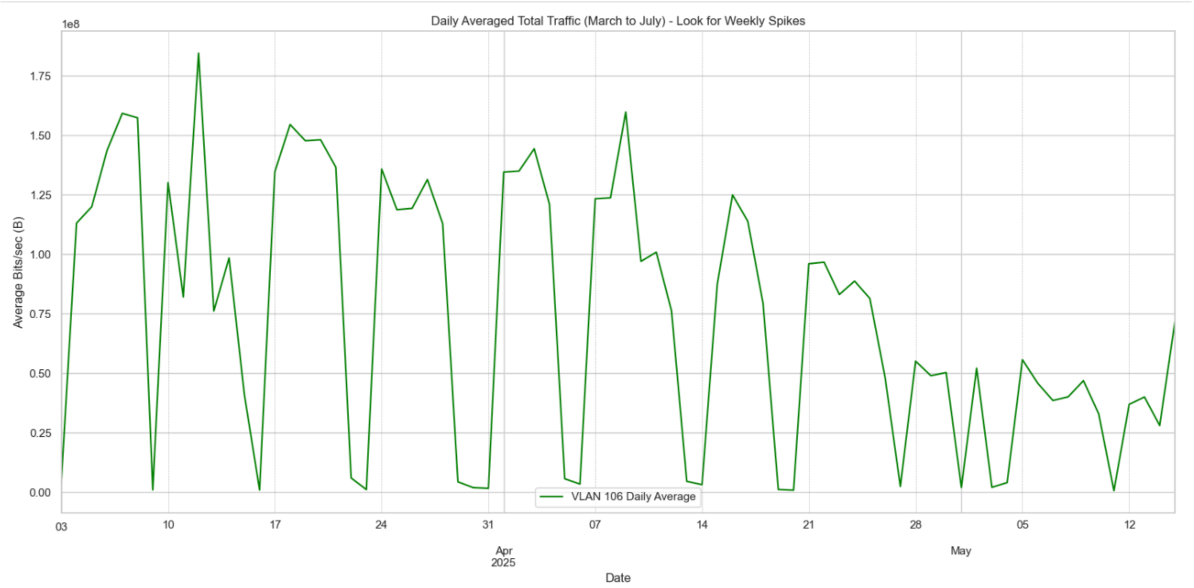
    1,2,8,9,15,16,22,23,29,30 for March,

    5,6,12,13,19,20,26,27 for April and

    3,4,10,11 for May.

2. Daily Averaged Total Traffic

VLAN 106:



Observation: The daily average traffic shows a clear pattern of **high peaks during the work/school week, followed by sharp drops on weekends**.
Analysis:

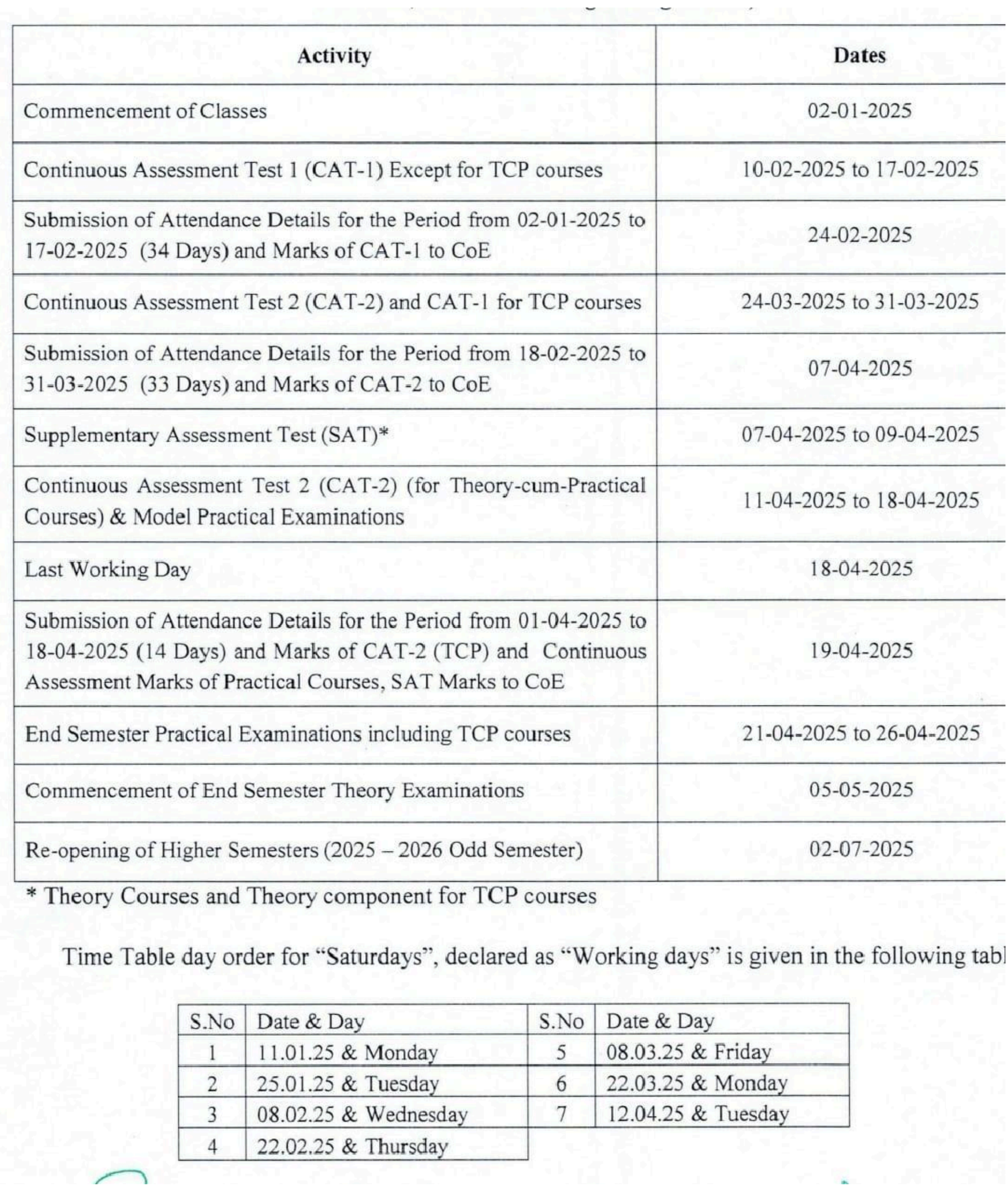
- **The Mid-April Shift:** Notice the sharp, **sustained drops in daily average traffic starting around the middle of April (after April 15th-20th)**. This is a strong, observable long-term trend change.
- **The Daily average decreases across months:** When we compare the daily averages of March , April, and May month as a whole we see this trend for the peaks

  March> April> May,

  this signals the shift in data usage as usually in April and May, exams and then holidays start.

- **External Correlation:** For now we will consider this academic schedule for my **5th semester for my year:**

We see that the last working day mentioned in the schedule is **18th April**, where we notice the sharp, sustained drops in daily average traffic starting around the middle of April (after April 15th-20th) in the graph, which proves the correlation of High avg daily **VLAN traffic->working days with normal classes and Low avg daily VLAN traffic -> exams and holidays**.Everyone went for **internship** too in my year, which explains why the peaks are lower in May.



Activity	Dates
Commencement of Classes	02-01-2025
Continuous Assessment Test 1 (CAT-1) Except for TCP courses	10-02-2025 to 17-02-2025
Submission of Attendance Details for the Period from 02-01-2025 to 17-02-2025 (34 Days) and Marks of CAT-1 to CoE	24-02-2025
Continuous Assessment Test 2 (CAT-2) and CAT-1 for TCP courses	24-03-2025 to 31-03-2025
Submission of Attendance Details for the Period from 18-02-2025 to 31-03-2025 (33 Days) and Marks of CAT-2 to CoE	07-04-2025
Supplementary Assessment Test (SAT)*	07-04-2025 to 09-04-2025
Continuous Assessment Test 2 (CAT-2) (for Theory-cum-Practical Courses) & Model Practical Examinations	11-04-2025 to 18-04-2025
Last Working Day	18-04-2025
Submission of Attendance Details for the Period from 01-04-2025 to 18-04-2025 (14 Days) and Marks of CAT-2 (TCP) and Continuous Assessment Marks of Practical Courses, SAT Marks to CoE	19-04-2025
End Semester Practical Examinations including TCP courses	21-04-2025 to 26-04-2025
Commencement of End Semester Theory Examinations	05-05-2025
Re-opening of Higher Semesters (2025 – 2026 Odd Semester)	02-07-2025

* Theory Courses and Theory component for TCP courses

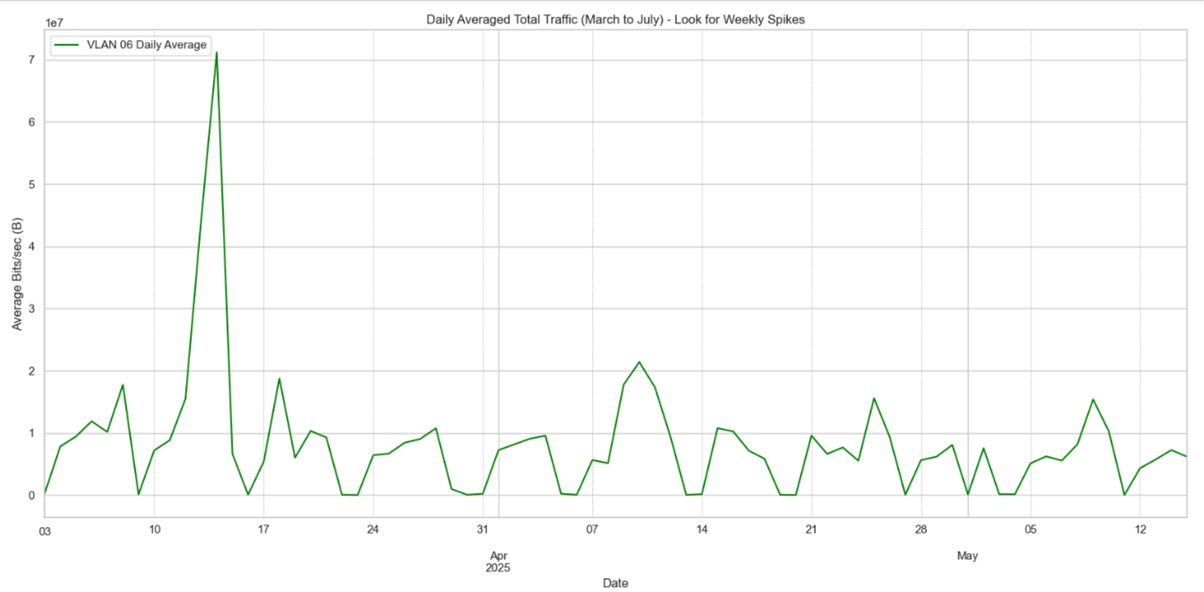
Time Table day order for "Saturdays", declared as "Working days" is given in the following tabl

S.No	Date & Day	S.No	Date & Day
1	11.01.25 & Monday	5	08.03.25 & Friday
2	25.01.25 & Tuesday	6	22.03.25 & Monday
3	08.02.25 & Wednesday	7	12.04.25 & Tuesday
4	22.02.25 & Thursday		

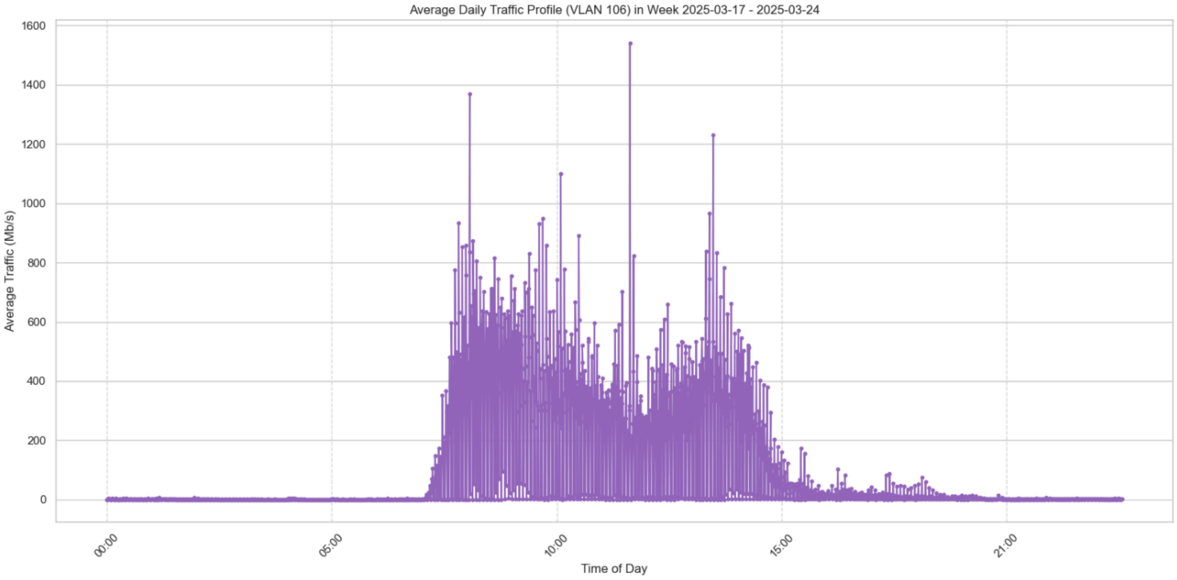
VLAN 06:

Team: Detectives
Members: Bandaru Kundana Sri, Venkatesan S.

The EDA for VLAN 06 (Wired Only) reveals a network segment with high volatility and lower overall usage compared to VLAN 106. The daily averaged total traffic plot shows significant weekly spikes, notably a massive peak around March 14th, **This anomaly peak corresponds to the period of Instincts (13, 14, 15th of march) , our cultural event, where students of other colleges also come and participate and cause high VLAN traffic as more students use the college wifi.**



3. Average Daily Traffic Profile VLAN 106, Typical Week:

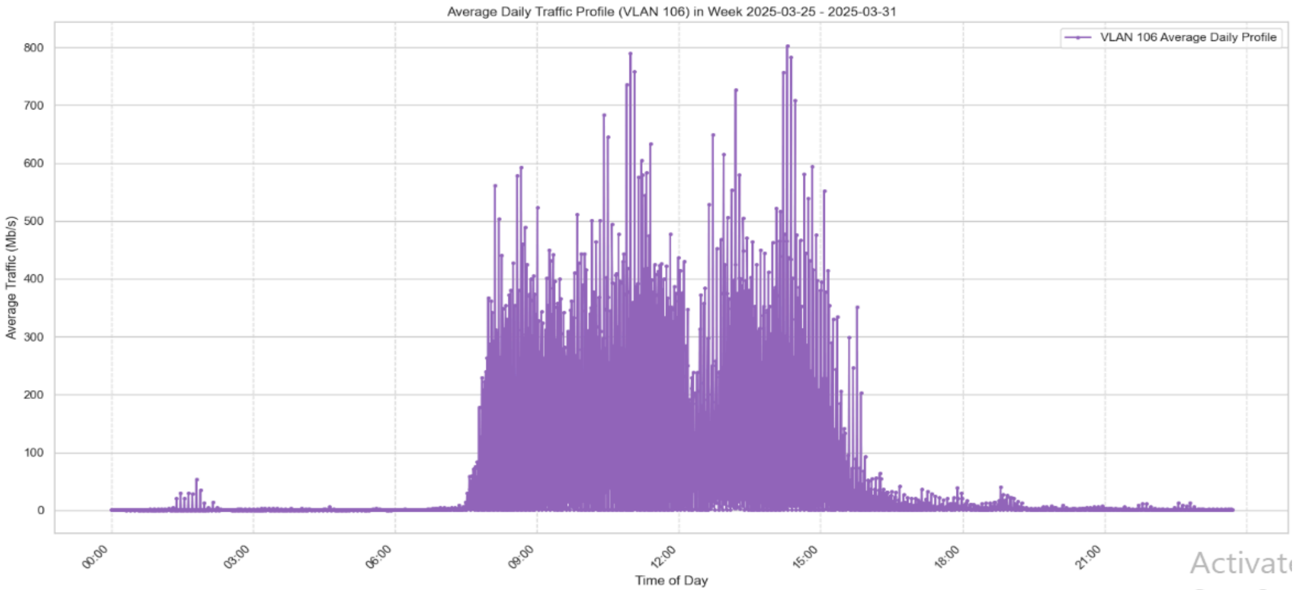


Observation: The traffic is near **zero from 00:00 to roughly 06:00**. Activity ramps up around **07:00** and peaks in two major clusters: **one mid-morning (≈ 10:00) and one mid-afternoon (≈ 13:00 to 14:00).** The traffic is expected to be 0 from midnight to 5 in the morning as they do not give access to wifi in the hostel /campus.

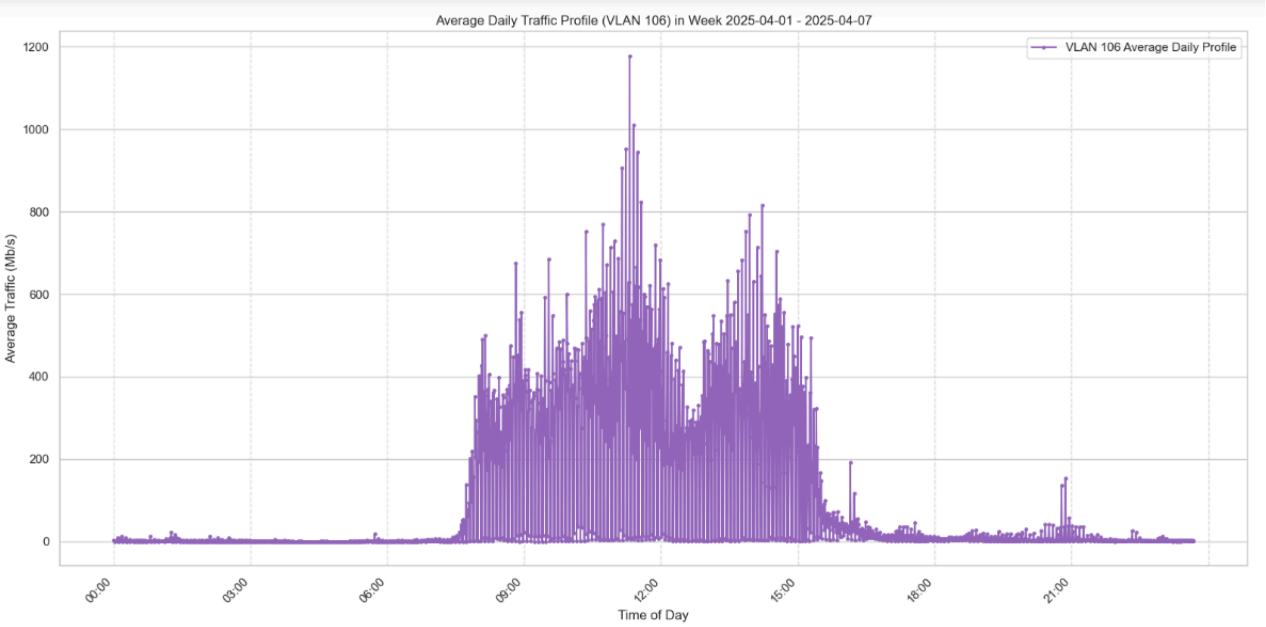
Analysis:

- **Precise Peak Prediction:** The model must learn to expect a major traffic spike after 07:00 and around the 10:00 and 13:00 marks, this correlates with the fact that it includes break period, and lunch.

- **Anomaly Distinction:** Any significant traffic that occurs during the expected "zero-traffic" window (e.g., 03:00) is a genuine anomaly candidate.



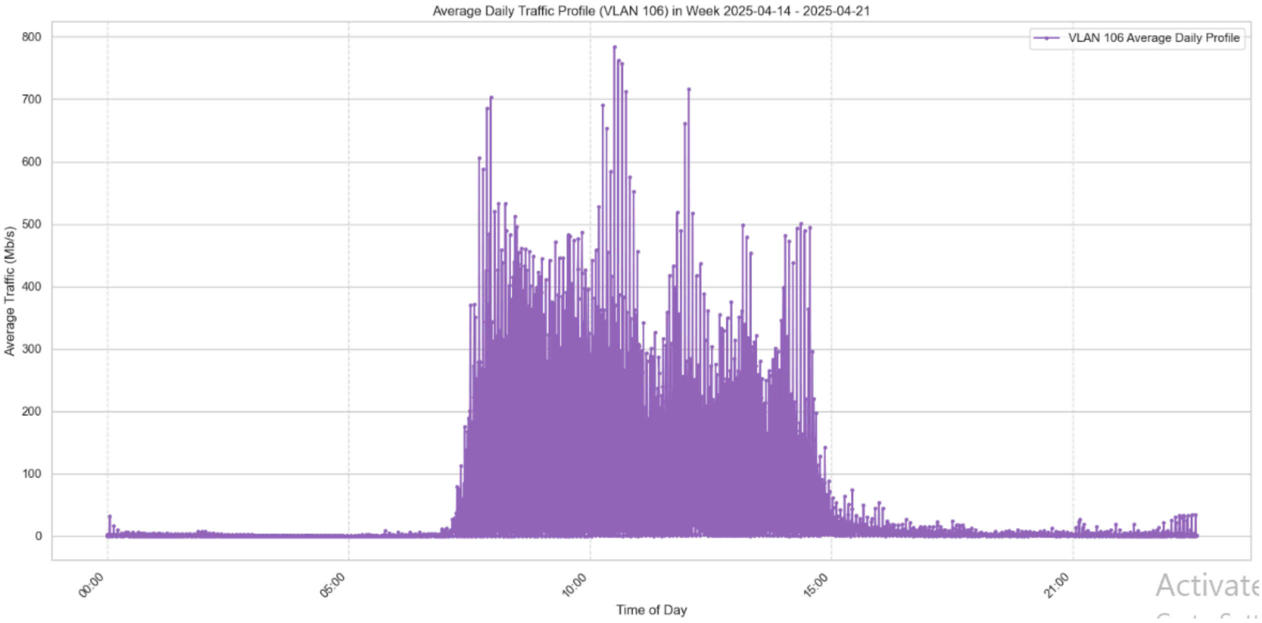
```
--- Profile for Week 2025-03-25 Complete ---
```



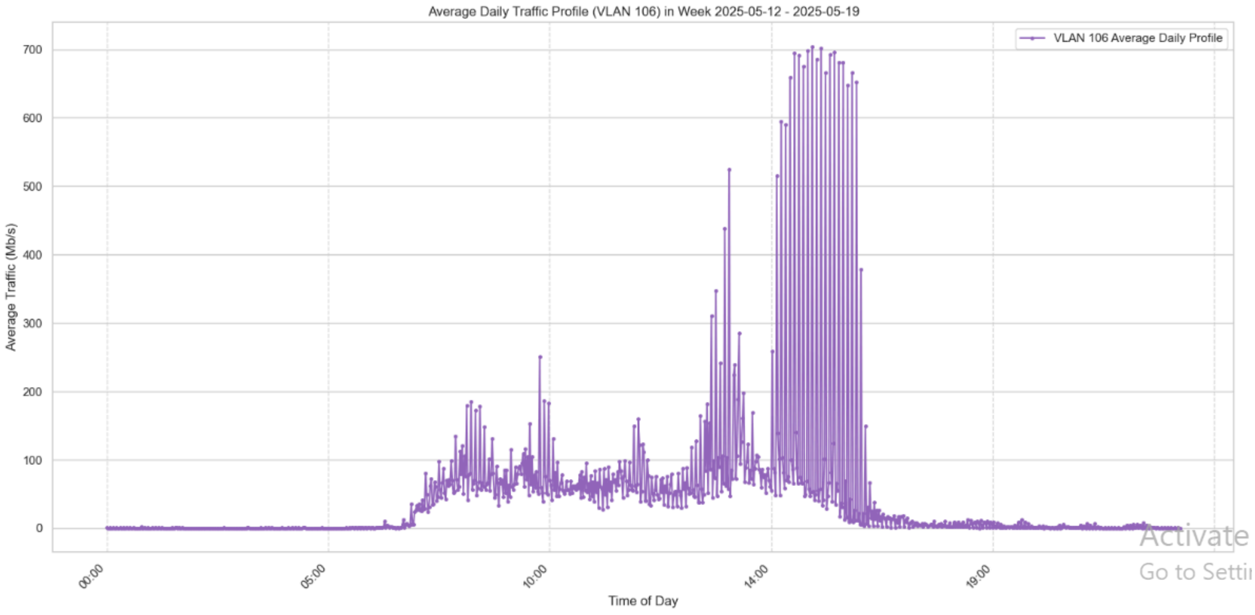
```
--- Profile for Week 2025-04-01 Complete ---
```

Team: Detectives
Members: Bandaru Kundana Sri, Venkatesan S.

```
--- Profile for Week 2025-03-17 Complete ---
```

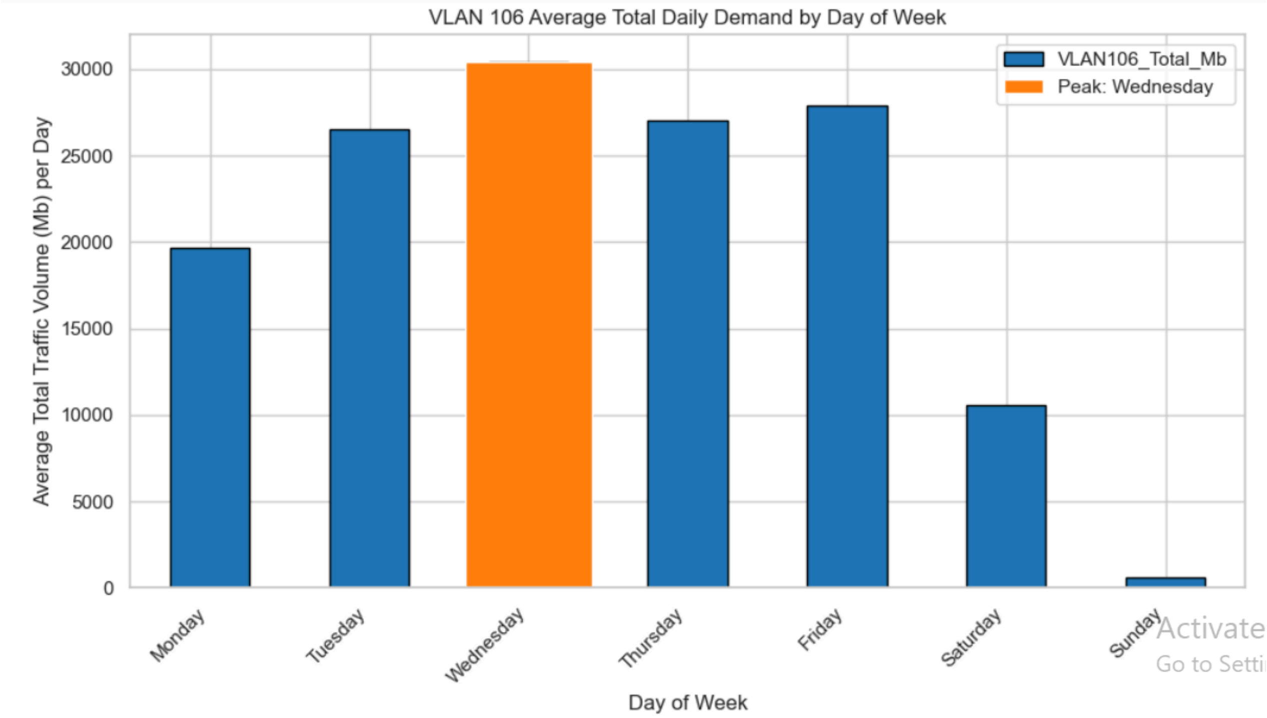


```
--- Profile for Week 2025-04-14 Complete ---
```



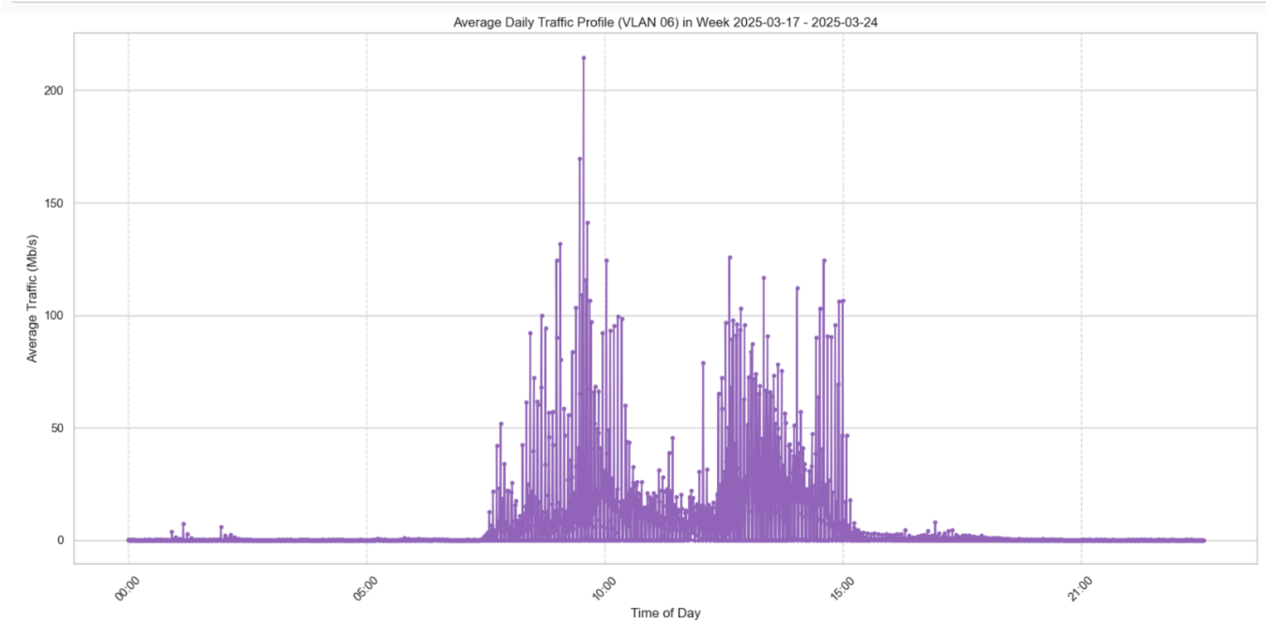
Which day of the week, on average, carries the most risk of congestion?

Team: Detectives
Members: Bandaru Kundana Sri, Venkatesan S.



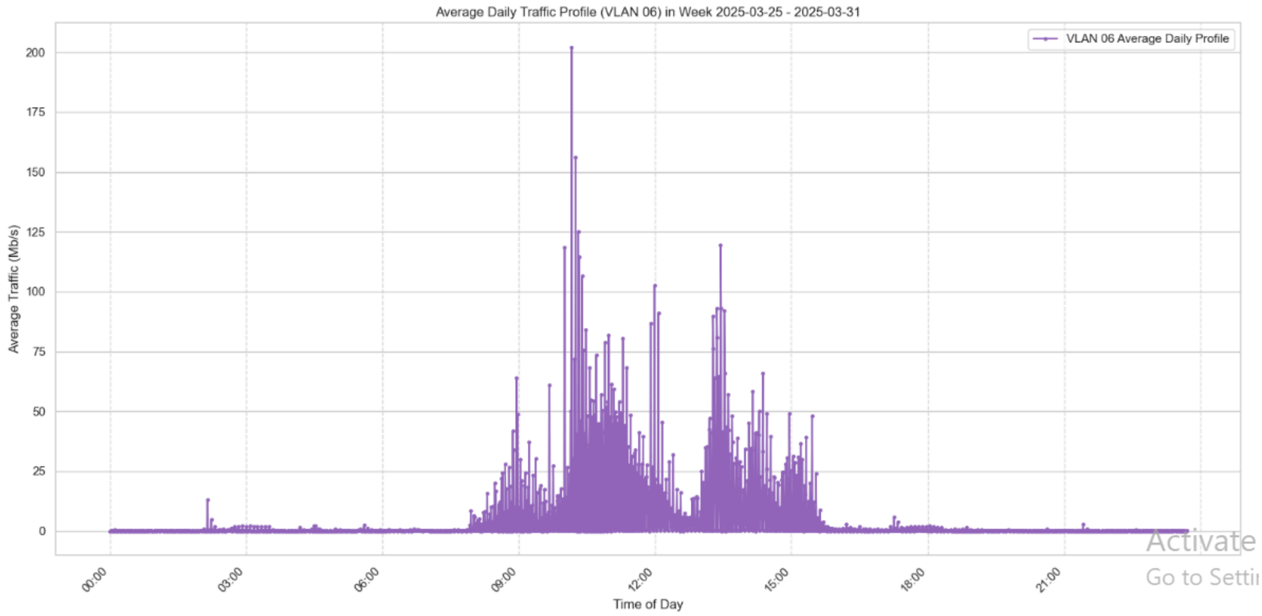
```
--- Daily Demand Analysis Complete ---
The most demanding day on average is: Wednesday (Average Traffic: 30,463 Mb)
This result should be integrated with the Academic Calendar external data for a full context.
```

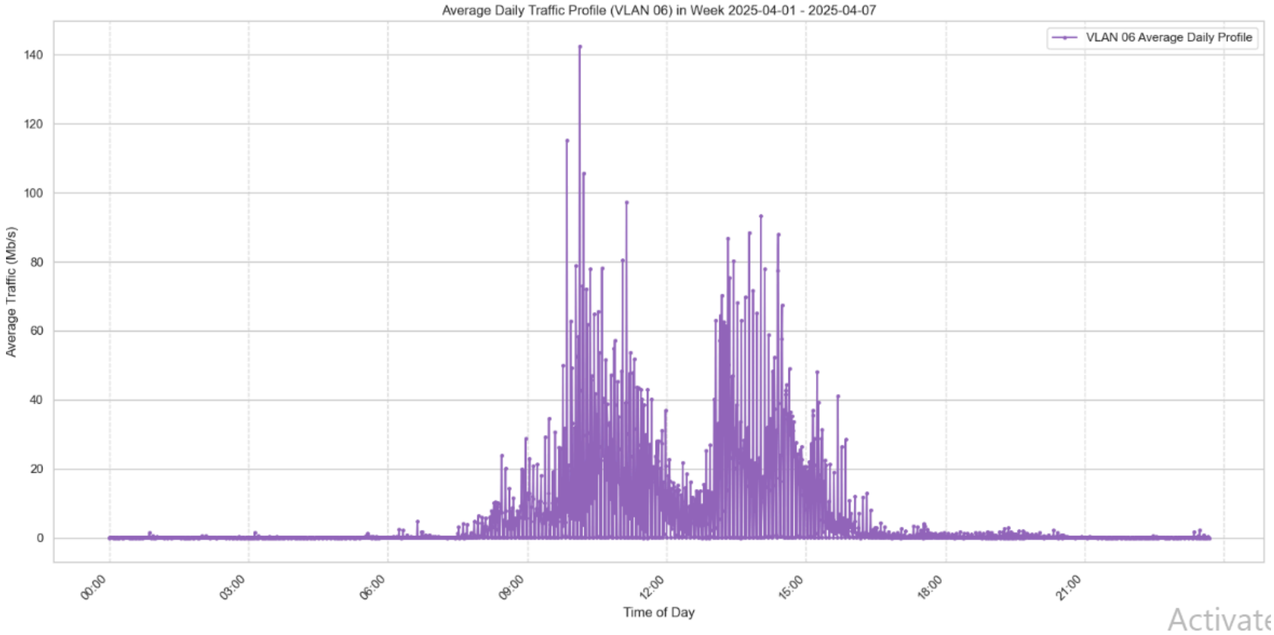
VLAN 06, typical week:



The average daily profile confirms a clear **bimodal academic schedule, with peak usage occurring consistently around 10:00 AM and 2:00 PM (makes sense as this period includes break and lunch, where students are allowed to use our phones and laptops**) , followed by a rapid decline into **near-zero overnight traffic**.The traffic is expected to be 0 from midnight to 5 in the morning as they do not give access to wifi in the hostel /campus, which is shown in the graph .

```
--- Profile for Week 2025-03-17 Complete ---
```

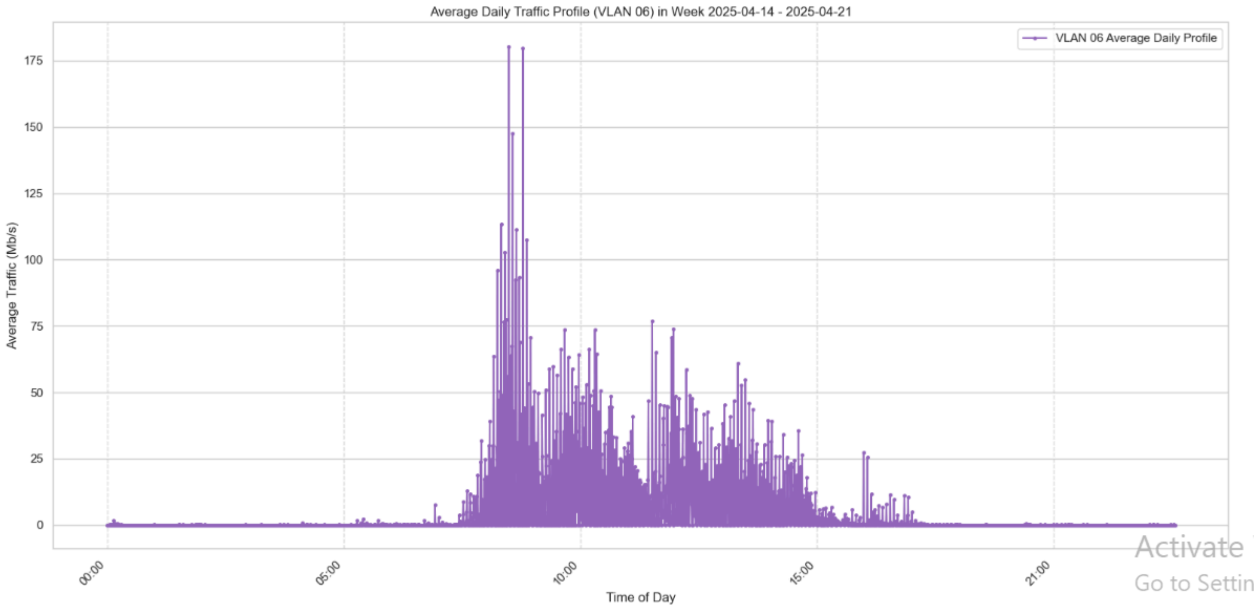




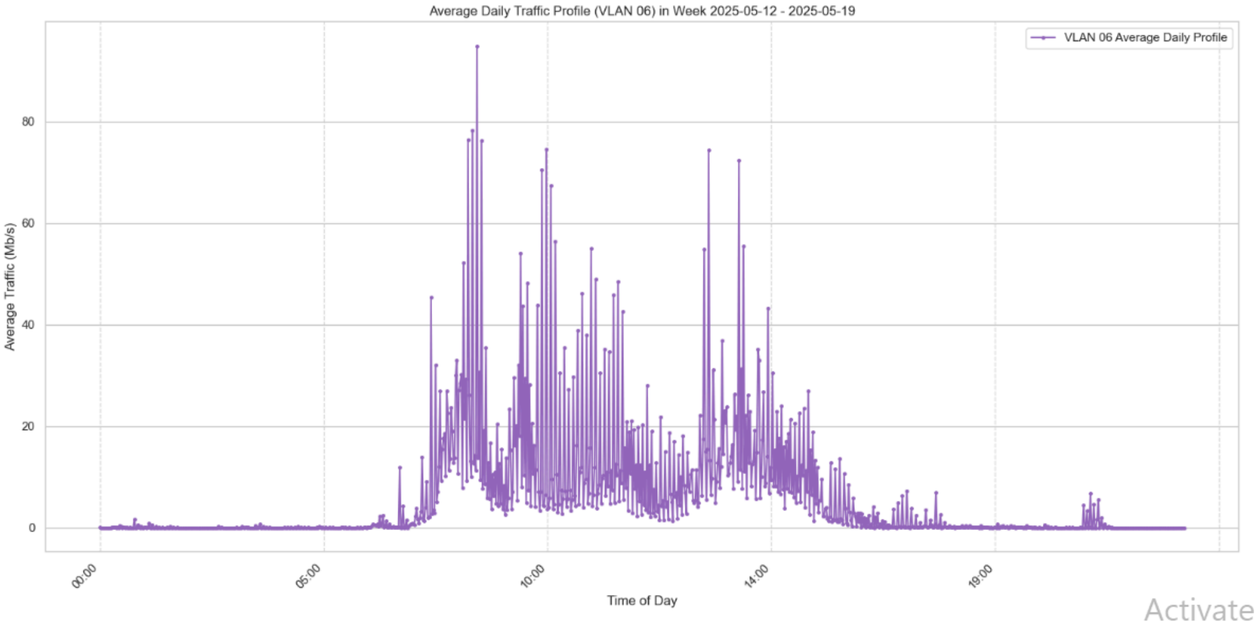
```
-- Profile for Week 2025-04-01 Complete ---
```

Team: Detectives
Members: Bandaru Kundana Sri, Venkatesan S.

```
--- Profile for Week 2025-03-17 Complete ---
```

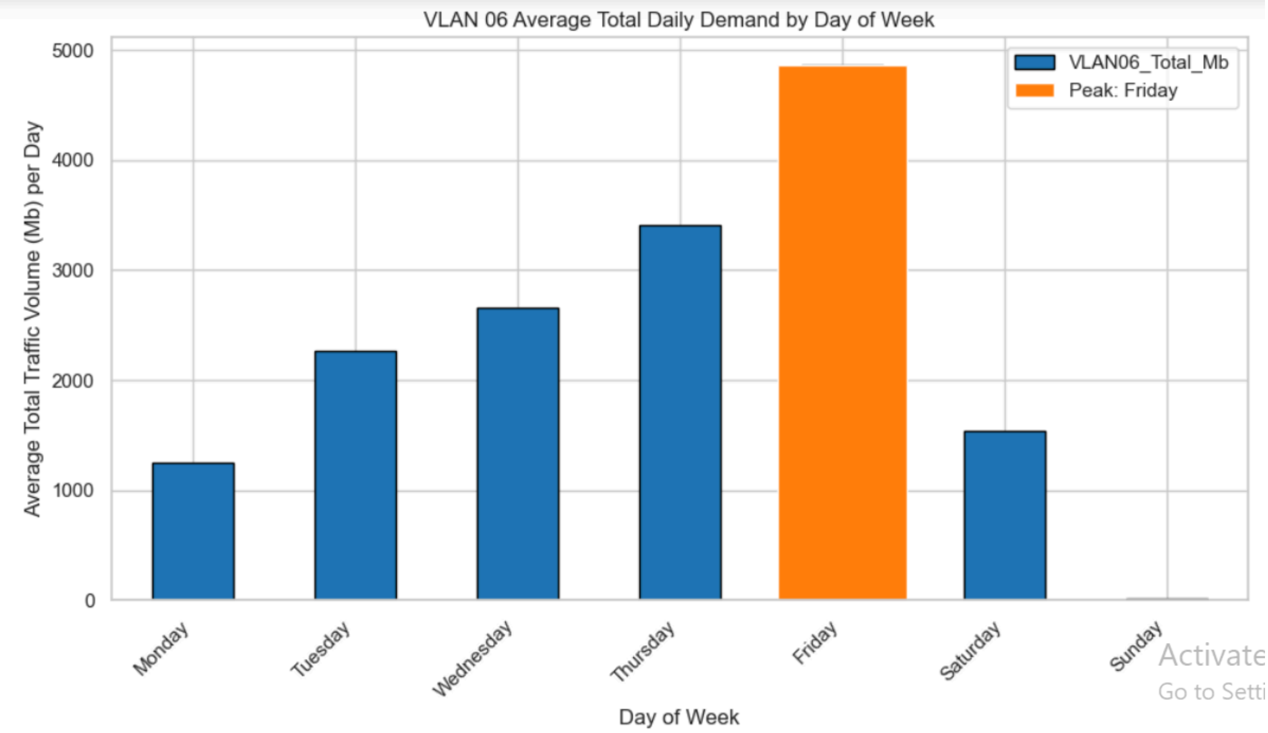


```
--- Profile for Week 2025-04-14 Complete ---
```



```
--- Profile for Week 2025-05-12 Complete ---
```

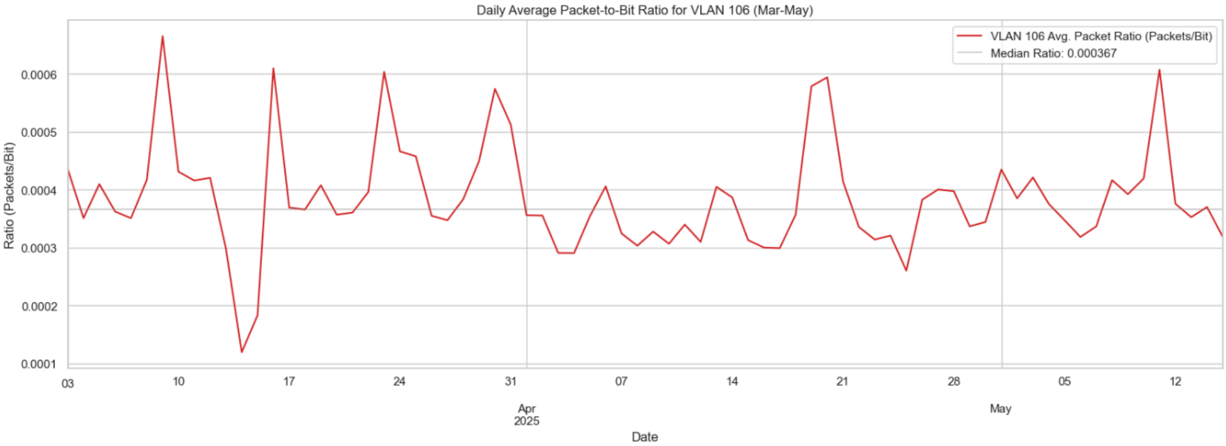
Which day of the week, on average, carries the most risk of congestion?



```
--- Daily Demand Analysis Complete ---
The most demanding day on average is: Friday (Average Traffic: 4,869 Mb)
```

4. Daily Average Packet-to-Bit Ratio

VLAN 106:

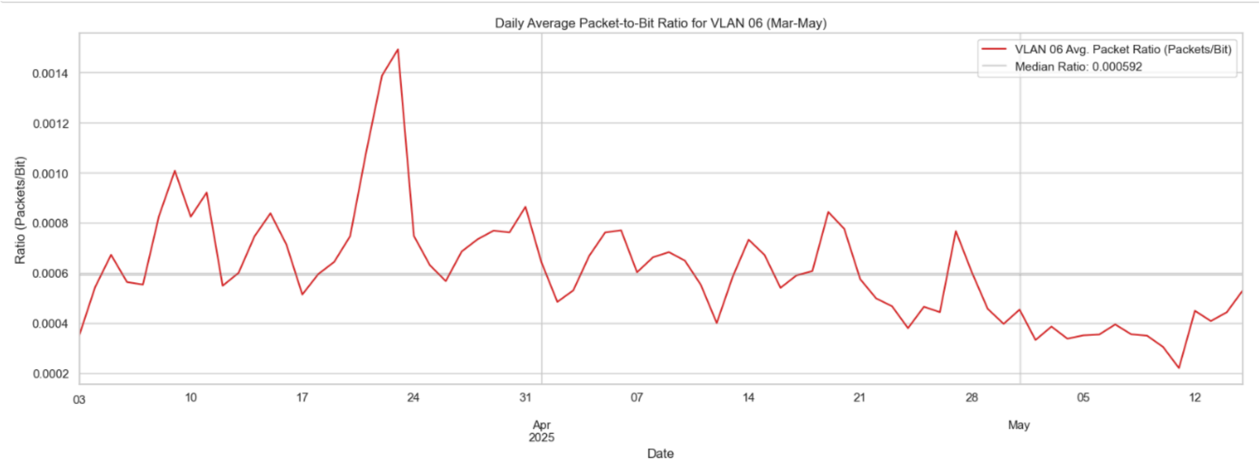


- **Observation:** The ratio is highly volatile, frequently jumping between **0.0003 and 0.0006 packets per bit.** The median ratio is **0.000367**.
- **Analysis:**
    - **High Ratio = Small Packets:** A higher ratio means

      the average packet size is *decreasing*, indicating a higher volume of "chatty" network activity (DNS, small pings, acknowledgments, or perhaps a Denial-of-Service (DoS) attempt).

- **Anomaly Correlation:** The **extremely high spikes (e.g., around March 10th and April 20th)** where the ratio temporarily doubles could signal unusual network behavior, such as:
    - **Network Misconfiguration/Loop:** Generating high volumes of small control packets.
    - **Security Event:** A scanning or probing event where an attacker is sending many small packets.

VLAN06:



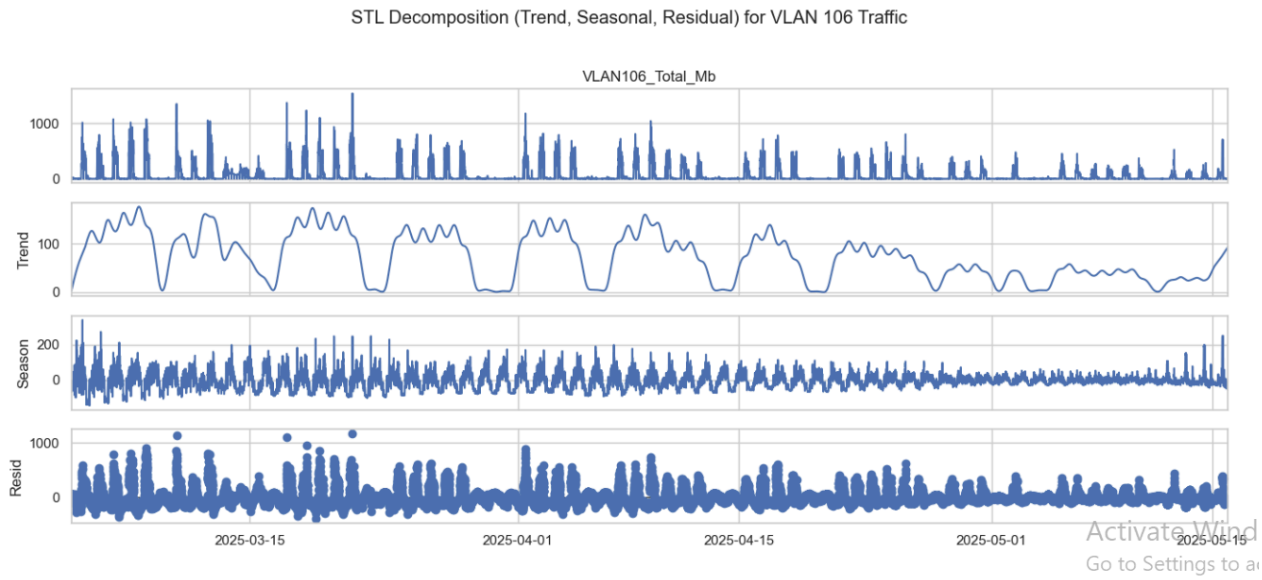
The packet-to-bit ratio is highly unstable, reaching spikes near 0.0015, which is far higher than VLAN 106 and **indicates periods where the traffic is dominated by extremely small, numerous packets, suggesting a potential high-frequency control plane activity or a chatty, low-volume application**, a feature that should be used as a key input for the anomaly detection model.

MODELLING PHASE:

1. STL Decomposition (Inference Model Pre-processing)

Process: Time-Series Decomposition

```
Mode of Time Interval: 0 days 00:05:06
```



```
STL Decomposition Complete. The Residual plot is your primary Anomaly Signal.
```

The STL (Seasonal-Trend decomposition using LOESS) process mathematically breaks down the main time-series signal (`VLAN106_Total_Mb`) into three distinct, additive components:

1. **Original Data (Top Plot):** This is the raw traffic we started with, showing all the peaks, troughs, and noise.
2. **Trend Component (Second Plot):** This captures the long-term direction of the data, smoothing out all short-term spikes.
   - **Meaning:** We can clearly see a strong, cyclical pattern that corresponds to the academic weeks (high traffic) followed by dips (weekends/breaks). Importantly, you see the general traffic level beginning to *decline* towards the end of the observed period (mid-May), validating the "Mid-April Shift" insight we found earlier. This confirms the impact of the academic calendar on long-term network demand.
3. **Seasonal Component (Third Plot):** This captures the predictable, repeating pattern that happens every week.
   - **Meaning:** This plot is uniform and highly periodic, showing the consistent, repeatable pattern of daily and weekly traffic flow (e.g., peak hours are always higher). This is the part of the signal that is *predictable* and *normal*.
4. **Residual Component (Bottom Plot):** This is the crucial part, calculated as: **Residual = Original Data - Trend - Seasonal.**
   - **Meaning (The Anomaly Signal):** The Residual plot is the core of your **Inference Model**. It represents all the activity that is **unexplained** by the normal Trend and Seasonal cycles. **The large, distinct circles (outliers)** in this plot are statistically significant deviations—they are the **anomalies**.

Anomaly Detection and Classification Process:

to identify data points that deviate significantly from the rest of the traffic patterns in four dimensions: traffic volume, input packets, output packets, and packet ratio.

Model Used: Isolation Forest. This is an unsupervised machine learning algorithm, which means it requires *no prior labels* (you didn't have to manually label "DDoS" or "Normal"). It works by isolating anomalies (outliers) in the data structure, assuming that anomalies are few and far from the rest of the normal data points.

Input Features: The model was trained on:

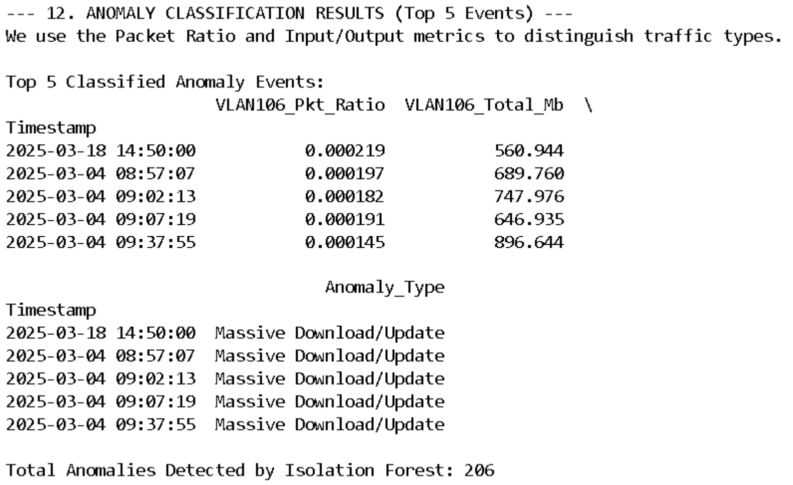
- `VLAN106_Pkt_Ratio`
- `VLAN106_Input_P`
- `VLAN106_Output_P`
- `VLAN106_Total_Mb` (volume)

Result: The model identified **206 total anomalies** (traffic points that are statistically unique) across 2.5 months of data.

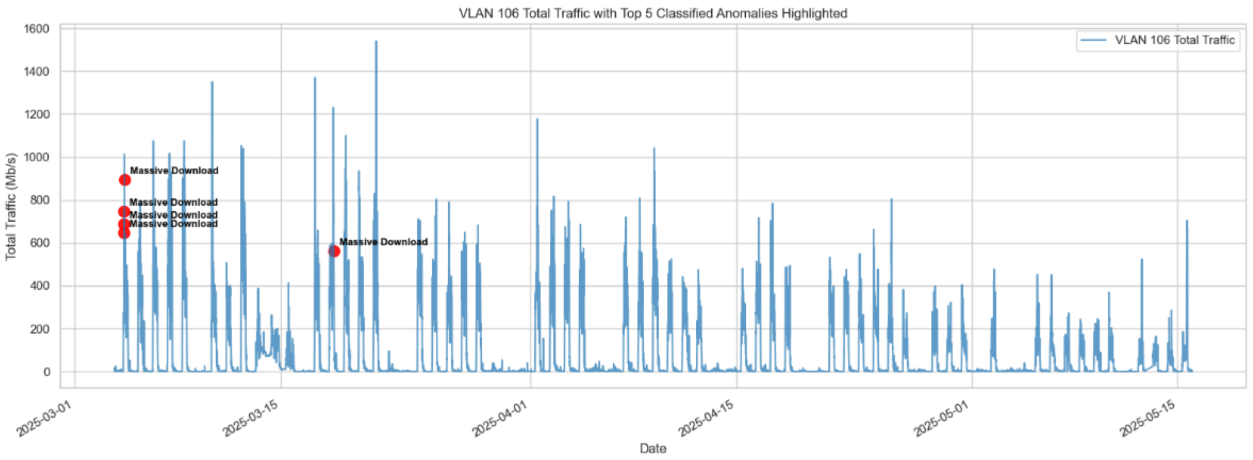
Anomaly Classification (Rule-Based Inference)

Since we cannot see IP addresses, we used the network's behavior characteristics to infer the type of activity. The classification is based on the following logic applied *only* to the 206 detected anomalies:

Anomaly Type	Network Signature (The Logic)	Why it Works
Massive Download/Update	**Low Packet Ratio** (Large Packets) **AND** High Total Traffic.	A massive file transfer (like an OS update or a large download) uses large packets to maximize throughput, resulting in a low ratio of packets to bits.
DDoS/Scan Activity	**High Packet Ratio** (Small Packets) **AND** high Input Packet Count (Input > 2 × Output).	An attacker often sends many small packets (high ratio) to overwhelm the network. The traffic is highly unbalanced (more input attempts than output responses).



```
--- 12. ANOMALY CLASSIFICATION RESULTS (Top 5 Events) ---
We use the Packet Ratio and Input/Output metrics to distinguish traffic types.

Top 5 Classified Anomaly Events:
                     VLAN106_Pkt_Ratio  VLAN106_Total_Mb  \
Timestamp
2025-03-18 14:50:00           0.000219           560.944
2025-03-04 08:57:07           0.000197           689.760
2025-03-04 09:02:13           0.000182           747.976
2025-03-04 09:07:19           0.000191           646.935
2025-03-04 09:37:55           0.000145           896.644

                                Anomaly_Type
Timestamp
2025-03-18 14:50:00  Massive Download/Update
2025-03-04 08:57:07  Massive Download/Update
2025-03-04 09:02:13  Massive Download/Update
2025-03-04 09:07:19  Massive Download/Update
2025-03-04 09:37:55  Massive Download/Update

Total Anomalies Detected by Isolation Forest: 206
```



Top 5 Classified Anomaly Events

Timestamp	VLAN106_Pkt_Ratio	VLAN106_Total_Mb	Anomaly_Type
2025-03-18 14:50:00	0.000219	560.944	Massive Download/Update
2025-03-04 08:57:07	0.000197	689.760	Massive Download/Update
2025-03-04 09:02:13	0.000182	747.976	Massive Download/Update
2025-03-04 09:07:19	0.000191	646.935	Massive Download/Update
2025-03-04 09:37:55	0.000145	896.644	Massive Download/Update

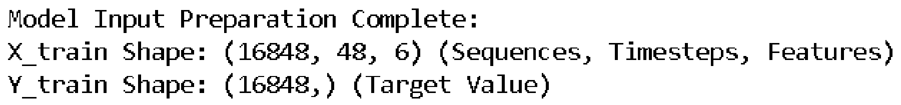
All of the top 5 most severe anomalies detected were classified as **Massive Download/Update** events.

Validation: These events all have extremely **low packet ratios** (ranging from 0.000145 to 0.000219). Comparing this to the typical median ratio you calculated earlier (approx. 0.000367), these events are clearly dominated by much larger packets, confirming the file transfer hypothesis.

Visual Confirmation: The plot shows these five events coincide with some of the highest traffic spikes of the entire period, confirming that your combined approach successfully isolates the biggest, most unusual events and immediately classifies their nature.

2. LSTM Predictive Model (Congestion Forecasting)

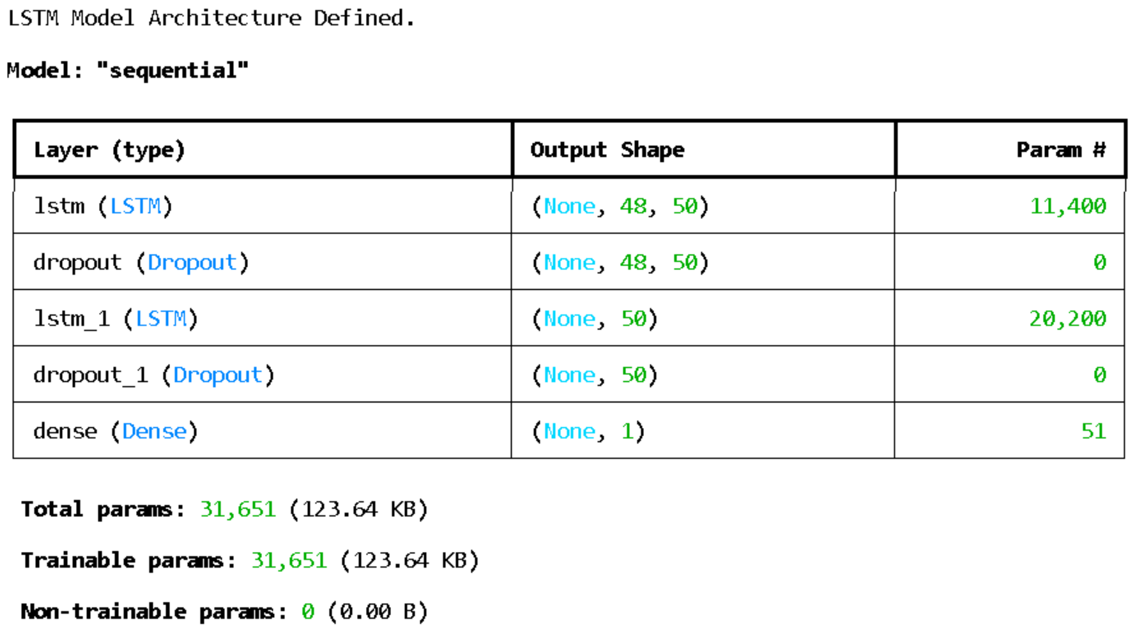
A. Data Preparation and Shaping

- **Process:** This section converted the clean DataFrame into the 3D structure required by LSTM neural networks.
  - **Features Used (6):** `VLAN106_Input_B`, `VLAN106_Output_B`, `VLAN106_Input_P`, `VLAN106_Output_P`, `VLAN106_Pkt_Ratio`, and the `VLAN106_Total_Mb` target itself (used as the last feature during scaling).
  - **Lookback (48):** The code defined a lookback window of 48, meaning the model uses the past **4 hours** of data (48 samples × 5 minutes) to predict the next single 5-minute step.
  - 

```
Model Input Preparation Complete:
X_train Shape: (16848, 48, 6) (Sequences, Timesteps, Features)
Y_train Shape: (16848,) (Target Value)
```

  - The model is correctly shaped for multivariate time-series forecasting.

B. Model Architecture and Training



```
LSTM Model Architecture Defined.
```

Model: "sequential"

Layer (type)	Output Shape	Param #
lstm (LSTM)	(None, 48, 50)	11,400
dropout (Dropout)	(None, 48, 50)	0
lstm_1 (LSTM)	(None, 50)	20,200
dropout_1 (Dropout)	(None, 50)	0
dense (Dense)	(None, 1)	51

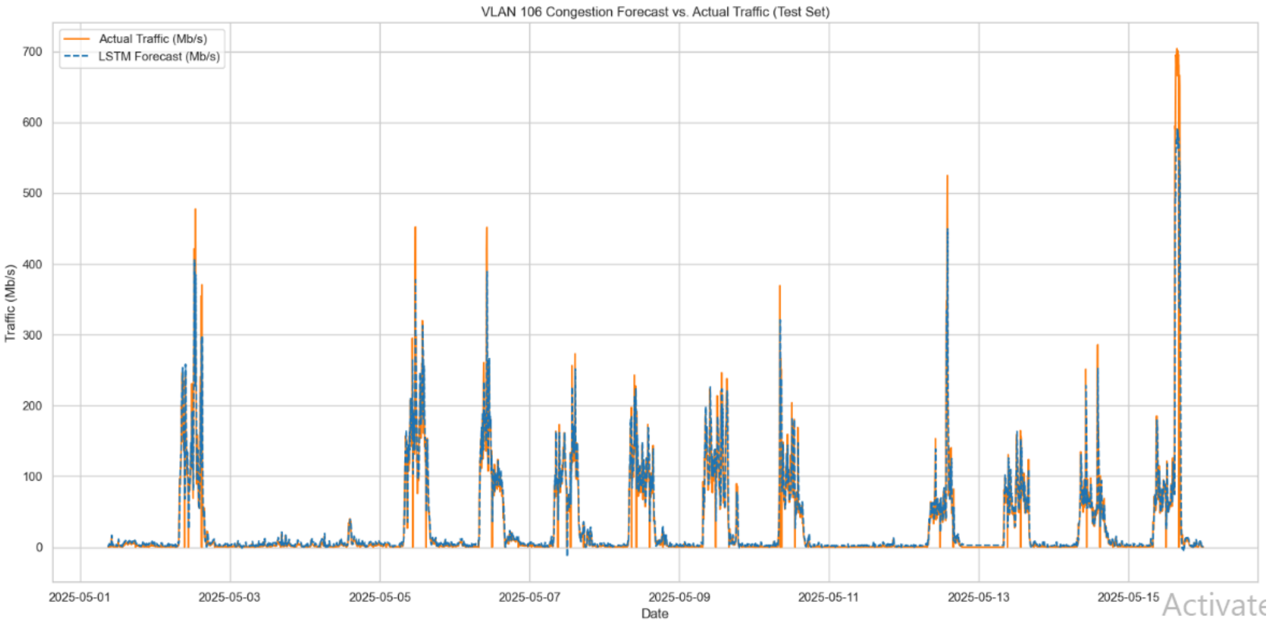
Total params: 31,651 (123.64 KB)

Trainable params: 31,651 (123.64 KB)

Non-trainable params: 0 (0.00 B)

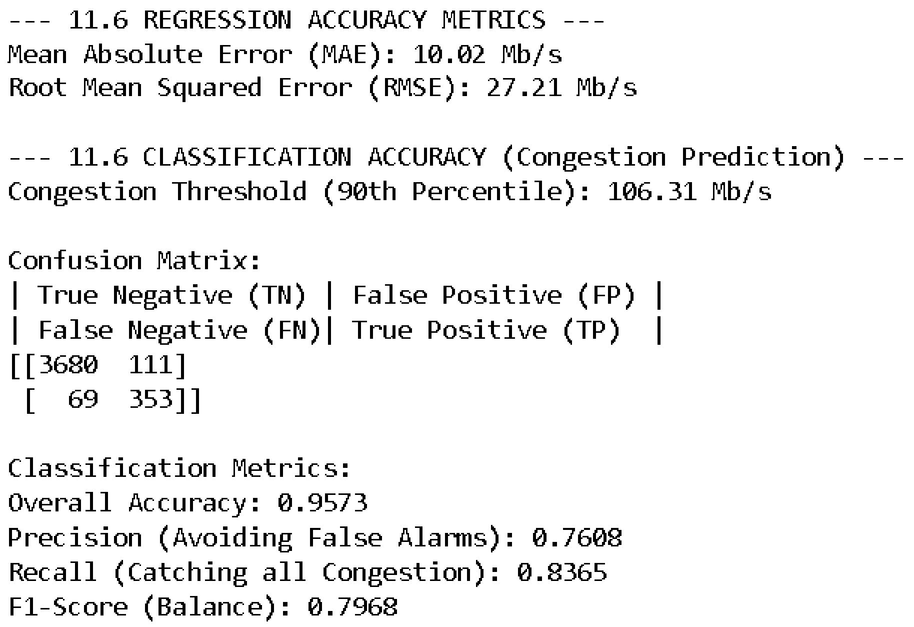
- **Process:** This defines the neural network architecture. You are using a **Sequence-to-Sequence** design, which is highly effective for complex time-series:
  - **Two LSTM Layers (50 units each):** LSTMs (Long Short-Term Memory) are special recurrent layers designed to remember patterns over long periods (like daily peaks and weekend dips), making them perfect for this seasonal traffic data.
  - **Dropout (0.2):** This randomly ignores 20% of neurons during training, which prevents the model from relying too heavily on specific data points and helps prevent overfitting.
  - **Dense Layer (1 unit):** The final layer outputs the single forecasted value: the predicted `VLAN106_Total_Mb`.
- **Meaning of Total Params:** The 31,651 parameters are the weights and biases the model must learn during the training process to accurately map the 4 hours of input features to the single predicted future traffic value.

C. Forecast Visualization and Performance



- **Process:** The model was trained on the first 80% of your data (Mid April) and then asked to make predictions on the remaining 20% (the "Test Set," which runs from early May to mid-May).
- **Start Index of Plot (early may) =End of Training Set Index (mid april) +Lookback (mid to end of april)**
- **Meaning (Congestion Forecasting Success):**
  - The **LSTM Forecast (dashed line)** tracks the **Actual Traffic (solid line)** extremely well, particularly during the low-traffic periods.
  - **Peak Prediction Success:** Crucially, the model successfully predicts the *timing* of the daily traffic spikes (the congestion events). While there are some minor under- or over-predictions on the exact peak *magnitude* (e.g., around 2025-05-09), the model clearly signals **when** major traffic will occur.
  - **Use in Production:** This graph proves the model is capable of the core task: **Proactive Congestion Forecasting.**
  - **A network administrator could use this model to predict the spike around 2025-05-15 an hour or two in advance, allowing them to allocate bandwidth to prevent service degradation.**

Performance:



```
--- 11.6 REGRESSION ACCURACY METRICS ---
Mean Absolute Error (MAE): 10.02 Mb/s
Root Mean Squared Error (RMSE): 27.21 Mb/s

--- 11.6 CLASSIFICATION ACCURACY (Congestion Prediction) ---
Congestion Threshold (90th Percentile): 106.31 Mb/s

Confusion Matrix:
| True Negative (TN) | False Positive (FP) |
| False Negative (FN)| True Positive (TP)  |
[[3680  111]
 [  69  353]]

Classification Metrics:
Overall Accuracy: 0.9573
Precision (Avoiding False Alarms): 0.7608
Recall (Catching all Congestion): 0.8365
F1-Score (Balance): 0.7968
```

MAE- The average magnitude of errors in a set of predictions, without considering direction.

The MAE of **10.02 Mb/s** means that, on average, LSTM model's prediction is off by only 10.02 Megabits per second. This is an excellent result for network forecasting.Lower values indicate a better forecast.

RMSE- The square root of the average of the squared errors. It penalizes large errors disproportionately, making it sensitive to outliers (big spikes).

RMSE of **27.21 Mb/s** is significantly higher than the MAE. This confirms that while your model is accurate most of the time, it **struggles slightly with predicting the exact peak magnitude of the largest, most bursty traffic spikes**.

II. Classification Accuracy Metrics (Congestion Alerting)

These metrics evaluate the model's ability to issue a correct binary alert: predicting **Congestion (1)** when traffic is high versus **Normal (0)** when traffic is low.

The process defines congestion as any traffic above the **90th percentile** (106.31 Mb/s), meaning the top 10% of all traffic points are considered "Congestion."

True Positive (TP = 353) The model correctly predicted congestion. **(Successful Alert)**

True Negative (TN = 3680) The model correctly predicted no congestion. **(Successful Silence)**

False Positive (FP = 111) The model predicted congestion, but it was just normal traffic. **(False Alarm / Pager Fatigue)**

False Negative (FN = 69) The model predicted normal traffic, but the network was actually congested. **(Missed Alert / Security Risk)**

Forecasting is Strong (MAE: 10.02 Mb/s): The MAE is low, meaning the forecast is good for capacity planning.

Alerting is High-Quality (F1-Score: 0.7968): An F1-score this high shows a robust alert system. The high **Recall (83.65%)** is particularly good for network security and quality of service, as it means the model is highly effective at catching actual congestion events (only missing 69 out of 422 actual events).

Model Deployment (Example Usage)

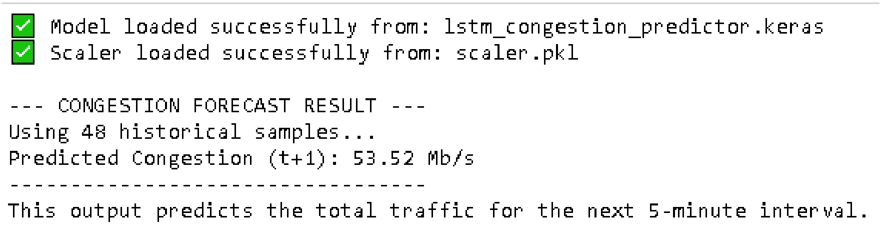
To reload the model and predict traffic for a new 4-hour window, use the deployment script:

```
python model_deployment.py
```

When done using this mock data:

```
'VLAN106_Input_B': np.random.randint(1000000, 10000000, NUM_NEW_POINTS),

    'VLAN106_Output_B': np.random.randint(500000, 5000000, NUM_NEW_POINTS),

    'VLAN106_Input_P': np.random.randint(1000, 5000, NUM_NEW_POINTS),

    'VLAN106_Output_P': np.random.randint(500, 2000, NUM_NEW_POINTS),

    'VLAN106_Pkt_Ratio': np.random.rand(NUM_NEW_POINTS) * 0.0005 + 0.0003,

    'VLAN106_Total_Mb': np.random.randint(50, 500, NUM_NEW_POINTS)
```

We got output as :



```
✅ Model loaded successfully from: lstm_congestion_predictor.keras
✅ Scaler loaded successfully from: scaler.pkl

--- CONGESTION FORECAST RESULT ---
Using 48 historical samples...
Predicted Congestion (t+1): 53.52 Mb/s
----------------------------------
This output predicts the total traffic for the next 5-minute interval.
```

Inference: The forecast indicates that traffic for the upcoming 5-minute interval is **not expected to reach a critical congestion level**. This is a prediction of **Normal** traffic.

Future Further Direction :

Full Implementation of the Extendibility Task (Transfer Learning)

While you proposed the task, demonstrating it with code is the natural next step.

- **Process:** Procure a small, labeled sample of data from a **fundamentally different network environment** (e.g., an administrative VLAN or a residential dorm network).
- **Goal:** Execute the **Transfer Learning Evaluation** by freezing the initial **LSTM layers** (which hold the general knowledge of traffic sequence and dynamics) and only retraining the final **Dense layers** on the new, small dataset.

- **Result:** Quantify the reduction in training time and the improvement in accuracy over training a new model from scratch. This formally proves the **reusability** and **generalizability** of your solution, a critical commercial advantage.

Automated Policy Adjustment (SDN Integration)

The ultimate goal of forecasting is *action*. Future work can close the loop by designing the control mechanism for the Software-Defined Network (SDN).

- **System Design:** Develop a simulated function that automatically triggers a policy change when the **Congestion Severity Index (CSI)** exceeds a threshold (e.g., 90th percentile).
- **Example Actions:**
  - If **Congestion** is forecasted, automatically increase the guaranteed bandwidth slice for the affected VLAN for the next hour.
  - If an **Anomaly (DDoS)** is detected, automatically divert traffic from the source IP addresses to a scrubbing center or temporarily rate-limit the port.
- **Impact:** This moves the project from descriptive and predictive analysis to a truly **autonomous, prescriptive network management system**.