Linear Regression Analysis of Network Traffic Data

Applied Machine Learning & Data Analytics

Dataset: Internet Firewall Logs
Professor: Dr. Samir Rustamov

Team Members:
Vusal Shirinbayli
Gabil Gurbanov

Dataset

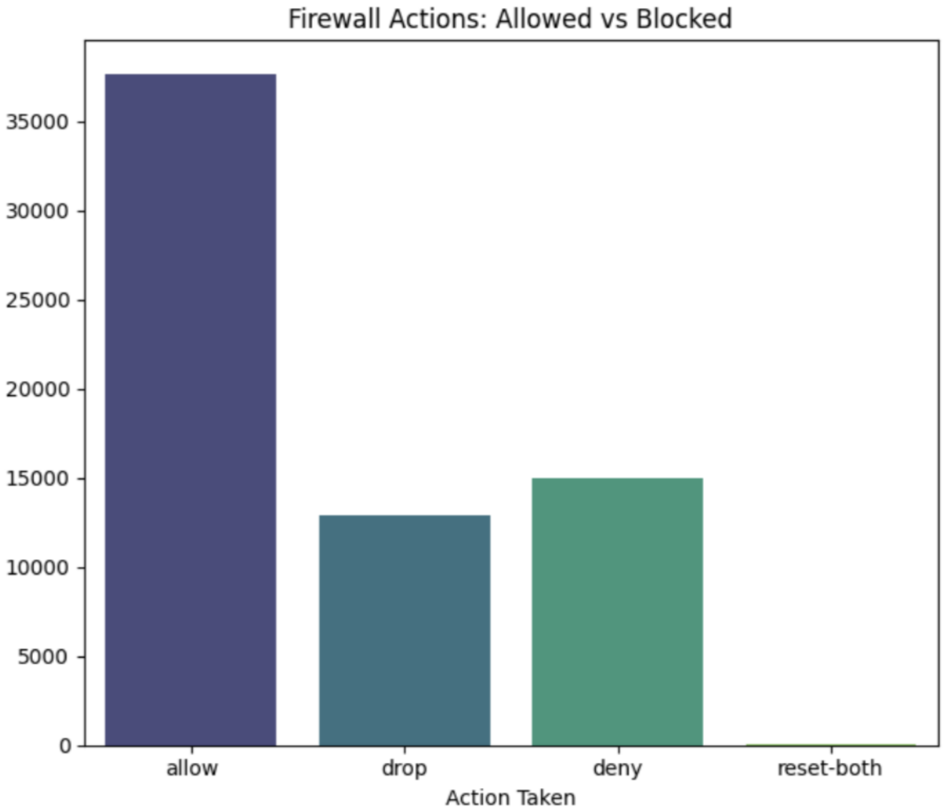
Source: Kaggle Firewall Dataset

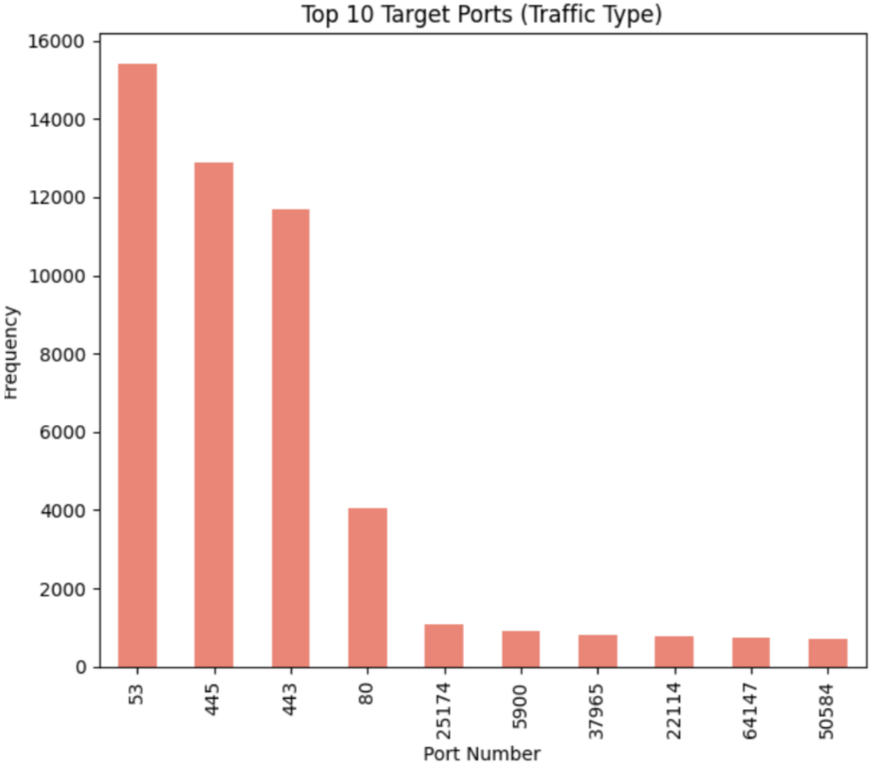
Observations: 65,532

Features: 12

Target: Bytes

Key Predictors: packets, elapsed_time_sec, nat_source_port, action

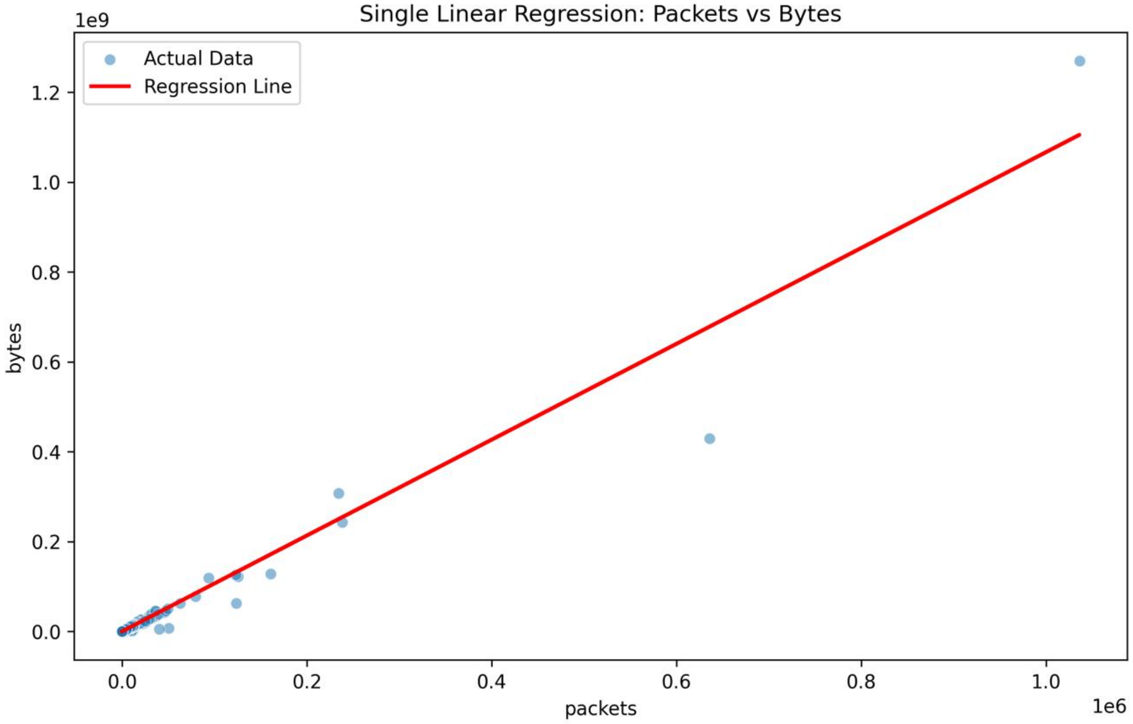

Firewall Actions: Allowed vs Blocked
35000
30000
25000
20000
15000
10000
5000
0
allow
drop
deny
reset-both
Action Taken


Top 10 Target Ports (Traffic Type)
Frequency
16000
14000
12000
10000
8000
6000
4000
2000
0
53
445
443
80
25174
5900
37965
22114
64147
50584
Port Number

Single Predictor Model

Model: Bytes = β0 + β1·Packets

Packets chosen due to strong intuitive relationship



Single Regression Results

Intercept ≈ -12,585

Slope ≈ 1,066

Each packet adds ~1066 bytes

p-value ≈ 0.000

T statistic ≈ 1109.02

$R^2 \approx 0.949$

Residual Sum of Squares (RSS) ≈ 1.046×10^{17}

Residual Standard Error (RSE) ≈ 1 263 656 bytes ≈ 1.2 Megabytes

Correlation r ≈ 0.974

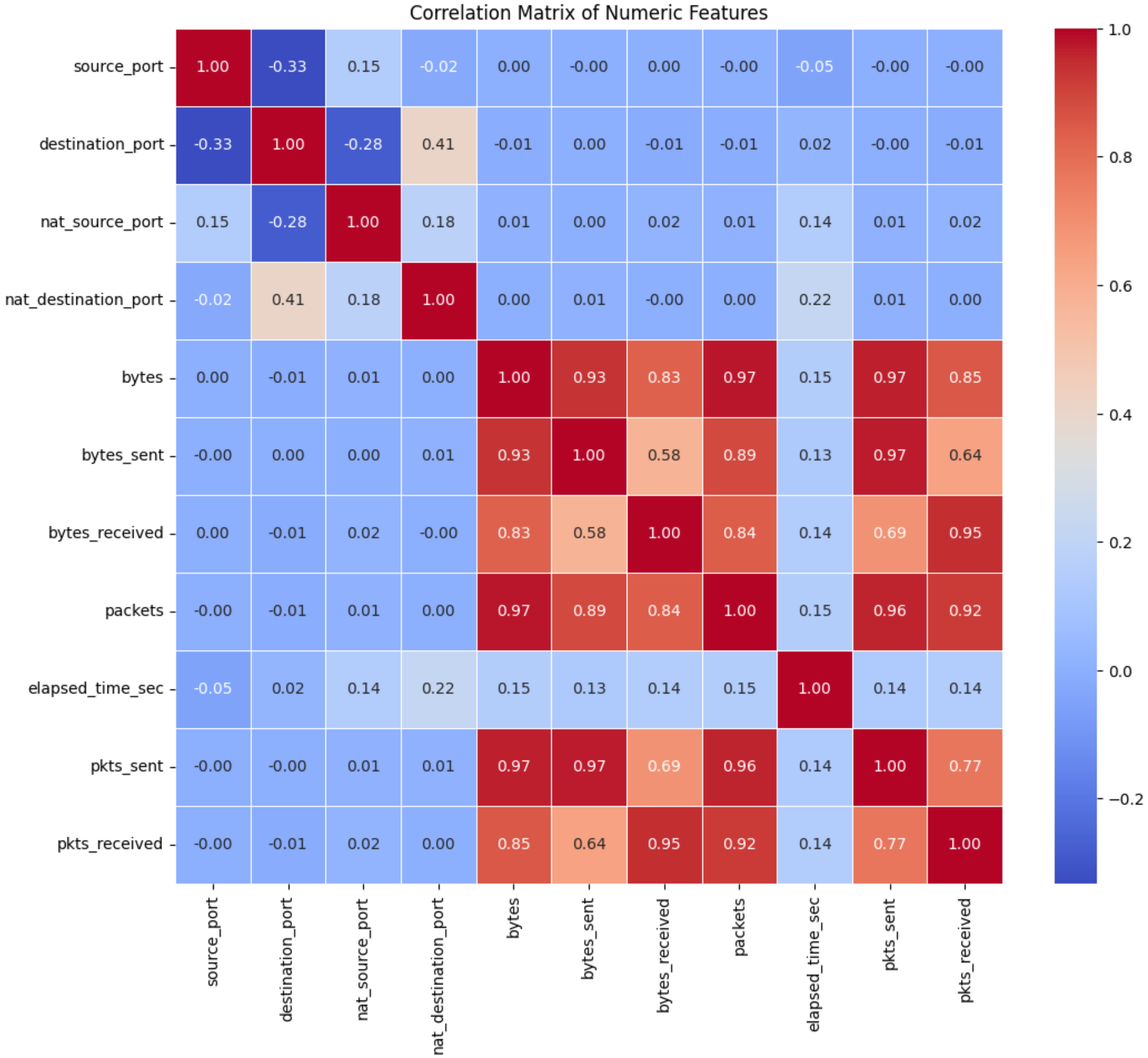
Multiple Regression Model

Predictors added:

- Elapsed time (sec)
- NAT source port

Metric	Single predictor	Multiple predictors	Conclusion
R-squared	0.9494	0.9495	Negligible improvement
P-values	< 0.05	All < 0.05	All variables are useful

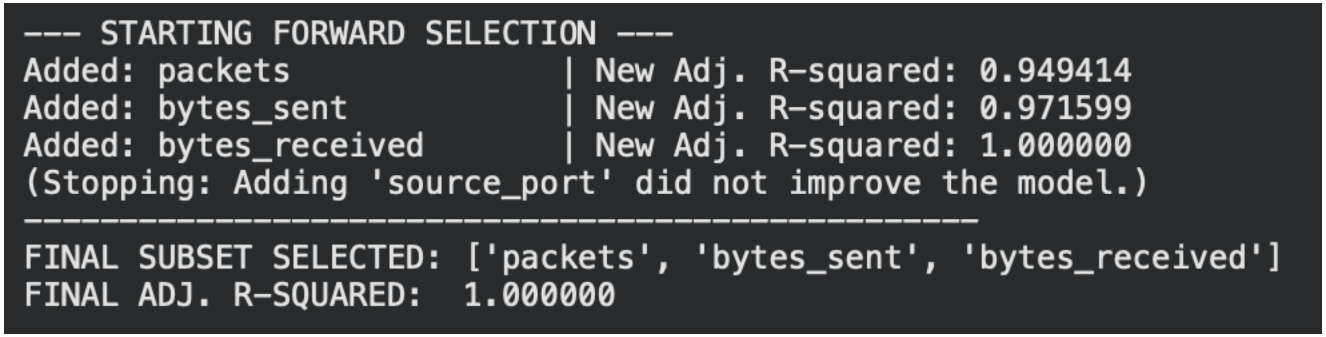
Correlation Matrix



Correlation Matrix of Numeric Features

	source_port	destination_port	nat_source_port	nat_destination_port	bytes	bytes_sent	bytes_received	packets	elapsed_time_sec	pkts_sent	pkts_received
source_port	1.00	-0.33	0.15	-0.02	0.00	-0.00	0.00	-0.00	-0.05	-0.00	-0.00
destination_port	-0.33	1.00	-0.28	0.41	-0.01	0.00	-0.01	-0.01	0.02	-0.00	-0.01
nat_source_port	0.15	-0.28	1.00	0.18	0.01	0.00	0.02	0.01	0.14	0.01	0.02
nat_destination_port	-0.02	0.41	0.18	1.00	0.00	0.01	-0.00	0.00	0.22	0.01	0.00
bytes	0.00	-0.01	0.01	0.00	1.00	0.93	0.83	0.97	0.15	0.97	0.85
bytes_sent	-0.00	0.00	0.00	0.01	0.93	1.00	0.58	0.89	0.13	0.97	0.64
bytes_received	0.00	-0.01	0.02	-0.00	0.83	0.58	1.00	0.84	0.14	0.69	0.95
packets	-0.00	-0.01	0.01	0.00	0.97	0.89	0.84	1.00	0.15	0.96	0.92
elapsed_time_sec	-0.05	0.02	0.14	0.22	0.15	0.13	0.14	0.15	1.00	0.14	0.14
pkts_sent	-0.00	-0.00	0.01	0.01	0.97	0.97	0.69	0.96	0.14	1.00	0.77
pkts_received	-0.00	-0.01	0.02	0.00	0.85	0.64	0.95	0.92	0.14	0.77	1.00

Forward Selection



```
--- STARTING FORWARD SELECTION ---
Added: packets              | New Adj. R-squared: 0.949414
Added: bytes_sent           | New Adj. R-squared: 0.971599
Added: bytes_received       | New Adj. R-squared: 1.000000
(Stopping: Adding 'source_port' did not improve the model.)
--------------------------------------------------
FINAL SUBSET SELECTED: ['packets', 'bytes_sent', 'bytes_received']
FINAL ADJ. R-SQUARED:  1.000000
```

Prediction & Interaction Analysis

Packets=50, Time=10s, NAT Port=5000

Predicted Bytes ≈ 43,656

95% confidence interval [32,084, 55,228]

Packets × Action interaction

We tested if 'Action' (Allow vs Deny) affects the slope.

Result: Significant Interaction ($p < 0.05$).

'Allowed' packets carry more data (payload) than 'Denied' packets (headers only).

Polynomial Regression

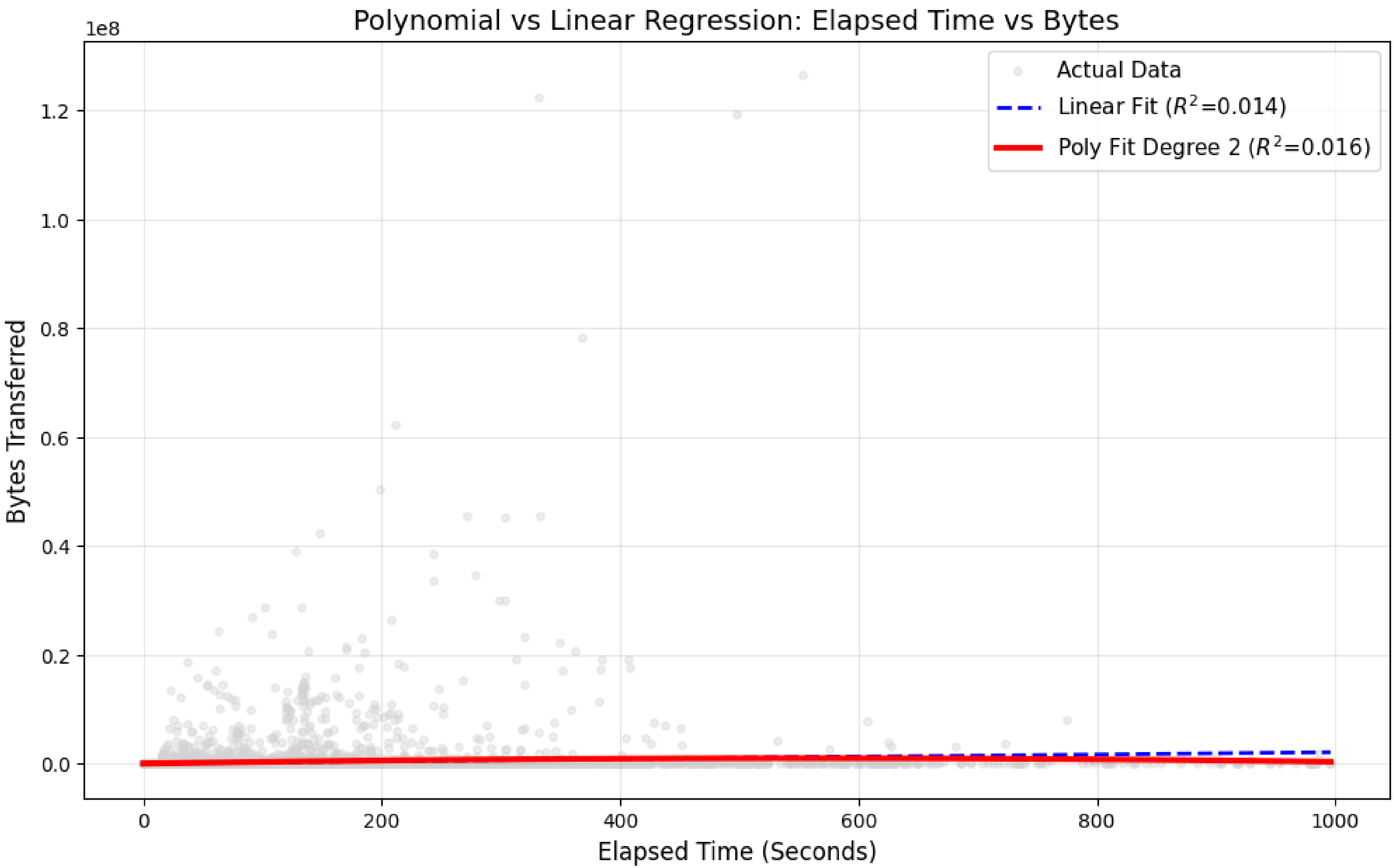
Hypothesis: Does traffic grow non-linearly with time?

Model Comparison (Time vs Bytes):

1. Linear Fit: $R^2 = 0.0136$

2. Polynomial (Deg 2): $R^2 = 0.0161$

Polynomial fits slightly better, but Time is a weak predictor compared to Packets.


Polynomial vs Linear Regression: Elapsed Time vs Bytes
1e8
1.2
1.0
0.8
0.6
0.4
0.2
0.0
Bytes Transferred
0
200
400
600
800
1000
Elapsed Time (Seconds)
Actual Data
Linear Fit (R^2=0.014)
Poly Fit Degree 2 (R^2=0.016)

Conclusion

Best Model: Single Predictor (Packets) is sufficient ($R^2 \approx 0.95$).

Adding Time and Port increased complexity but added negligible accuracy (< 0.001 improvement).

The volume of data transferred is almost entirely determined by the number of packets sent, regardless of connection duration.

Thank you for your attention