Topic: Traffic Forecasting

Course: Network Measurement and Data Analysis Laboratory

Students: Bahram Hedayati and Mahsa Delaram

Matricola: 10870276 - 10847175

School of Industrial and Information Engineering

Master of Telecommunication

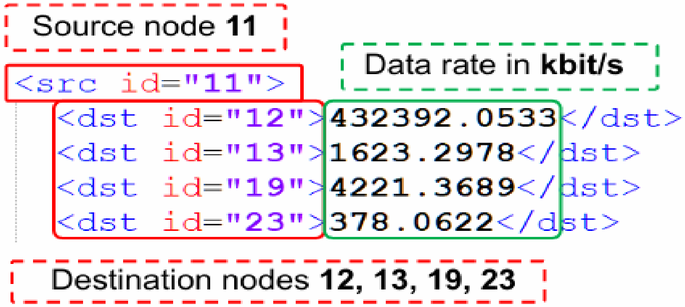
2023/24

Project Specification

Project #7-8 – Traffic forecasting Dataset

- GÉANT is the research network that carries traffic between universities and research institutions in Europe
- GÉANT is composed of 23 routers connected with 38 links
- GÉANT uses SONET technology to multiplex traffic with different bitrates into one optical signal
- Channel with the smallest bitrate that can be created in SONET is 50 Mbit/s

- Each file in the dataset describes total traffic in kbit/s between pairs of routers [1, 2]

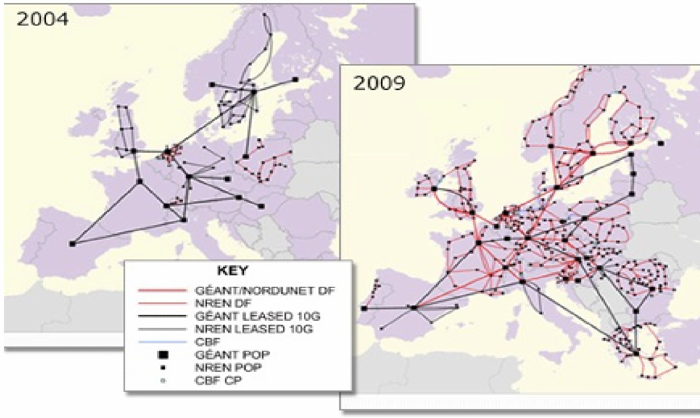


Source node **11**

Data rate in **kbit/s**

```
<src id="11">
  <dst id="12">432392.0533</dst>
  <dst id="13">1623.2978</dst>
  <dst id="19">4221.3689</dst>
  <dst id="23">378.0622</dst>
```

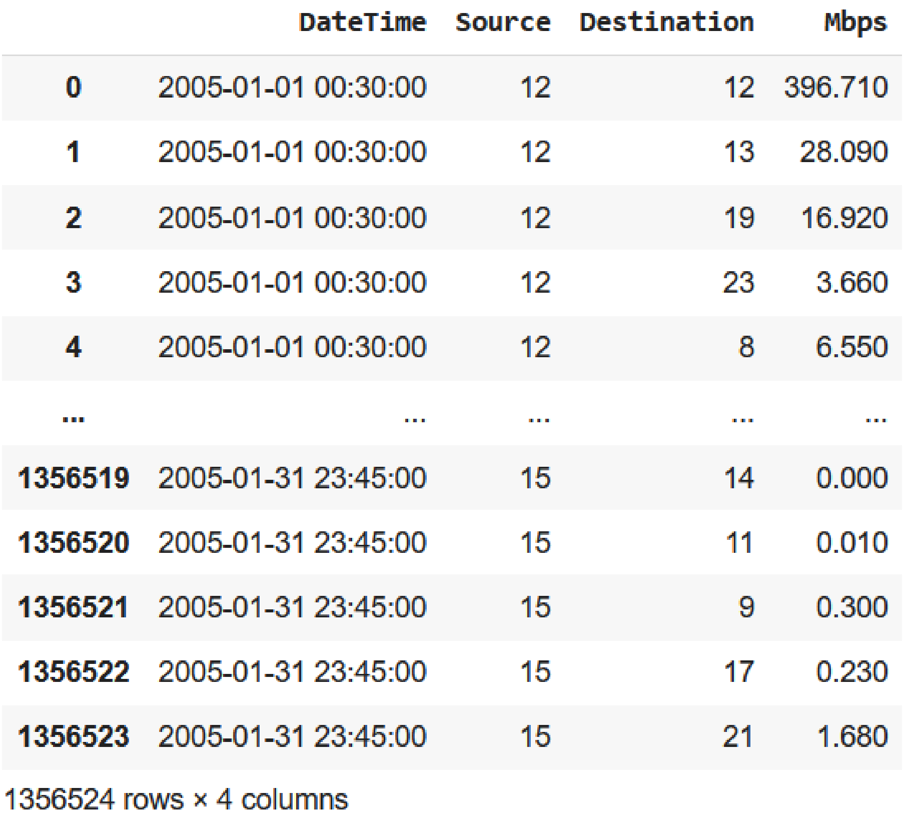
Destination nodes **12, 13, 19, 23**



- Dataset includes 2941 files: traffic at 15 min intervals for 1 month

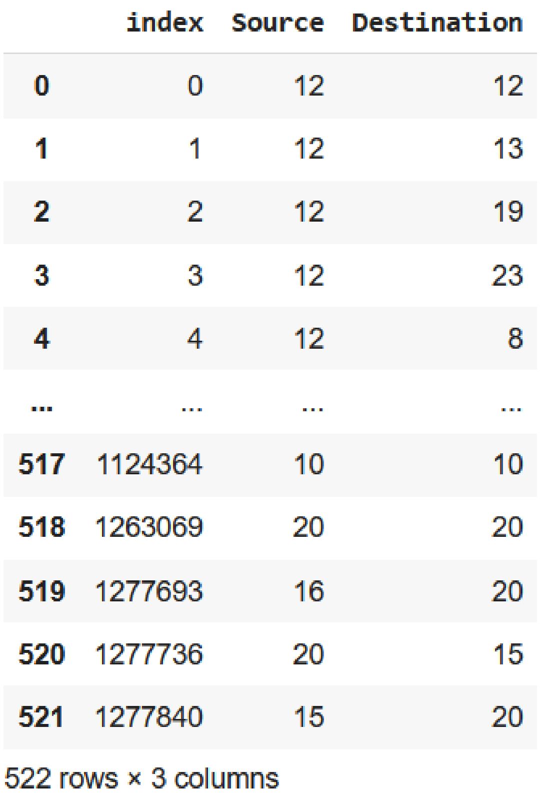
[1] https://totem.info.ucl.ac.be/dataset.html
[2] https://dl.acm.org/doi/10.1145/1111322.1111341

Dataset Construction in Pandas



	DateTime	Source	Destination	Mbps
0	2005-01-01 00:30:00	12	12	396.710
1	2005-01-01 00:30:00	12	13	28.090
2	2005-01-01 00:30:00	12	19	16.920
3	2005-01-01 00:30:00	12	23	3.660
4	2005-01-01 00:30:00	12	8	6.550
...	...	...	...	...
1356519	2005-01-31 23:45:00	15	14	0.000
1356520	2005-01-31 23:45:00	15	11	0.010
1356521	2005-01-31 23:45:00	15	9	0.300
1356522	2005-01-31 23:45:00	15	17	0.230
1356523	2005-01-31 23:45:00	15	21	1.680

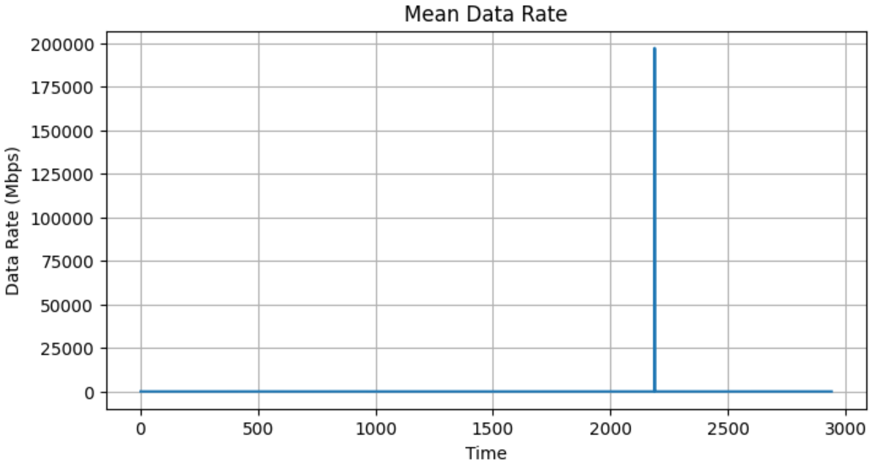
1356524 rows × 4 columns

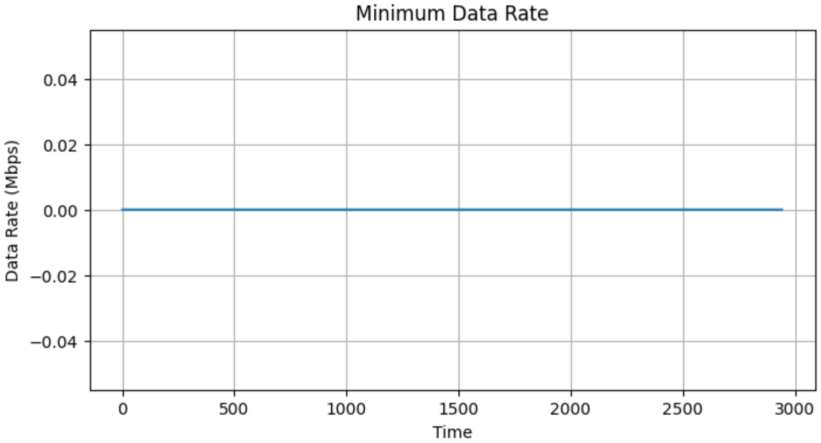


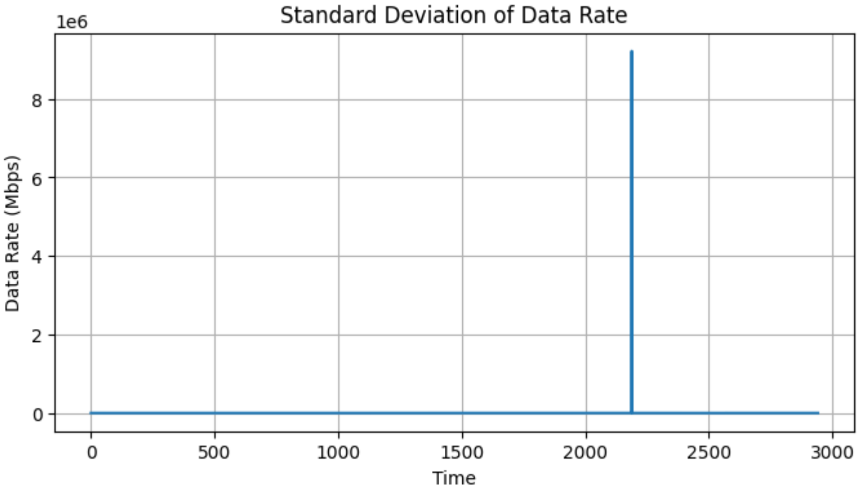
	index	Source	Destination
0	0	12	12
1	1	12	13
2	2	12	19
3	3	12	23
4	4	12	8
...	...	...	...
517	1124364	10	10
518	1263069	20	20
519	1277693	16	20
520	1277736	20	15
521	1277840	15	20

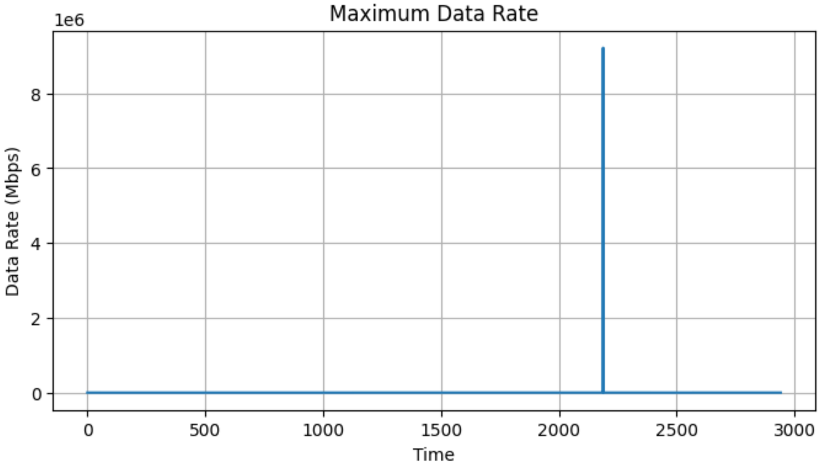
522 rows × 3 columns

Raw Data

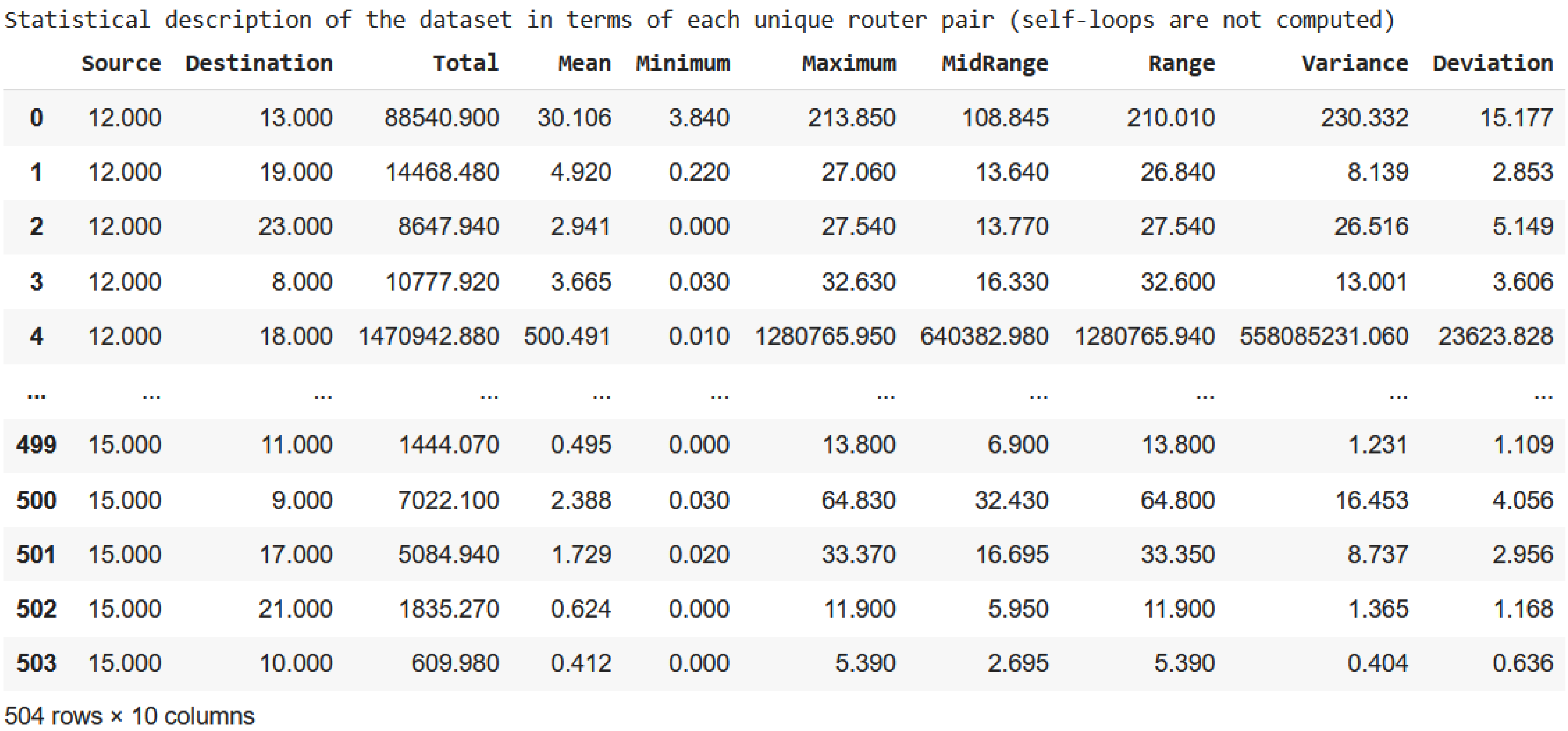

Mean Data Rate
Data Rate (Mbps)
200000
175000
150000
125000
100000
75000
50000
25000
0
0
500
1000
1500
2000
2500
3000
Time


Minimum Data Rate
Data Rate (Mbps)
0.04
0.02
0.00
−0.02
−0.04
0
500
1000
1500
2000
2500
3000
Time


Standard Deviation of Data Rate
1e6
Data Rate (Mbps)
8
6
4
2
0
0
500
1000
1500
2000
2500
3000
Time


Maximum Data Rate
1e6
Data Rate (Mbps)
8
6
4
2
0
0
500
1000
1500
2000
2500
3000
Time

Statistical Description of Raw Data



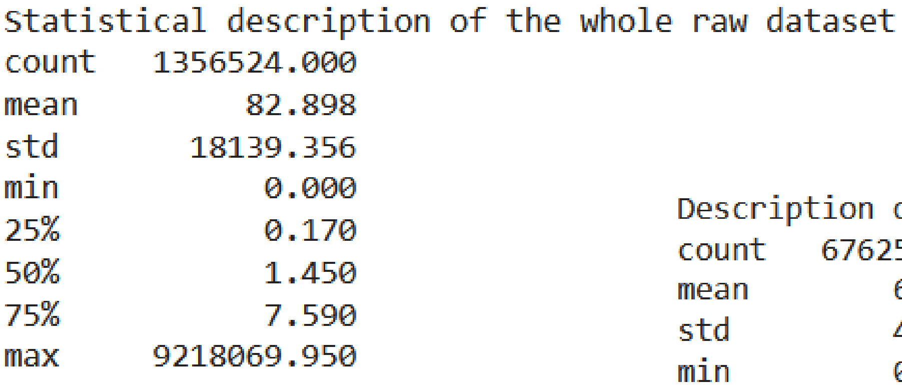
Statistical description of the dataset in terms of each unique router pair (self-loops are not computed)

	Source	Destination	Total	Mean	Minimum	Maximum	MidRange	Range	Variance	Deviation
0	12.000	13.000	88540.900	30.106	3.840	213.850	108.845	210.010	230.332	15.177
1	12.000	19.000	14468.480	4.920	0.220	27.060	13.640	26.840	8.139	2.853
2	12.000	23.000	8647.940	2.941	0.000	27.540	13.770	27.540	26.516	5.149
3	12.000	8.000	10777.920	3.665	0.030	32.630	16.330	32.600	13.001	3.606
4	12.000	18.000	1470942.880	500.491	0.010	1280765.950	640382.980	1280765.940	558085231.060	23623.828
...	...	...	...	...	...	...	...	...	...	...
499	15.000	11.000	1444.070	0.495	0.000	13.800	6.900	13.800	1.231	1.109
500	15.000	9.000	7022.100	2.388	0.030	64.830	32.430	64.800	16.453	4.056
501	15.000	17.000	5084.940	1.729	0.020	33.370	16.695	33.350	8.737	2.956
502	15.000	21.000	1835.270	0.624	0.000	11.900	5.950	11.900	1.365	1.168
503	15.000	10.000	609.980	0.412	0.000	5.390	2.695	5.390	0.404	0.636

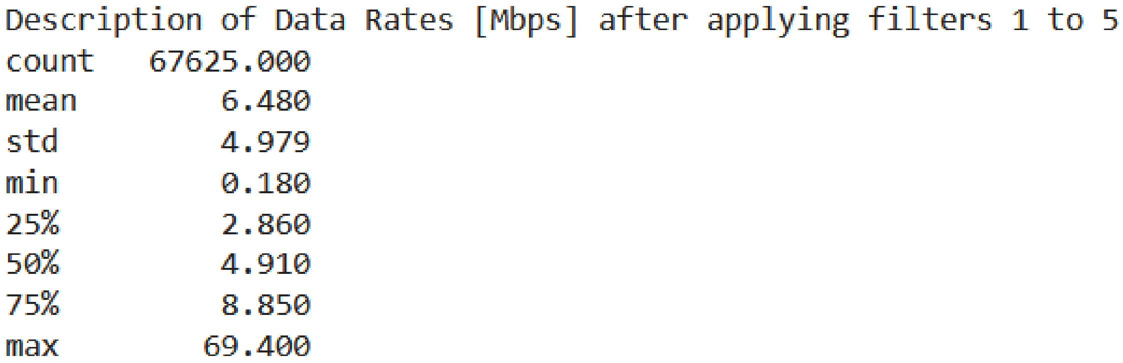
504 rows × 10 columns

Subset selection of source-destination pairs

1. Source-Destination pairs with the same source and destination (self-loops)
2. Source-Destination pairs that have some data rate higher than 99% percentile
3. Source-Destination pairs that have some data rate lower than 25% percentile
4. Source-Destination pairs that did not have any data rate during at least one day
5. Source-Destination pairs that have a standard deviation higher than 5 over their data rates.

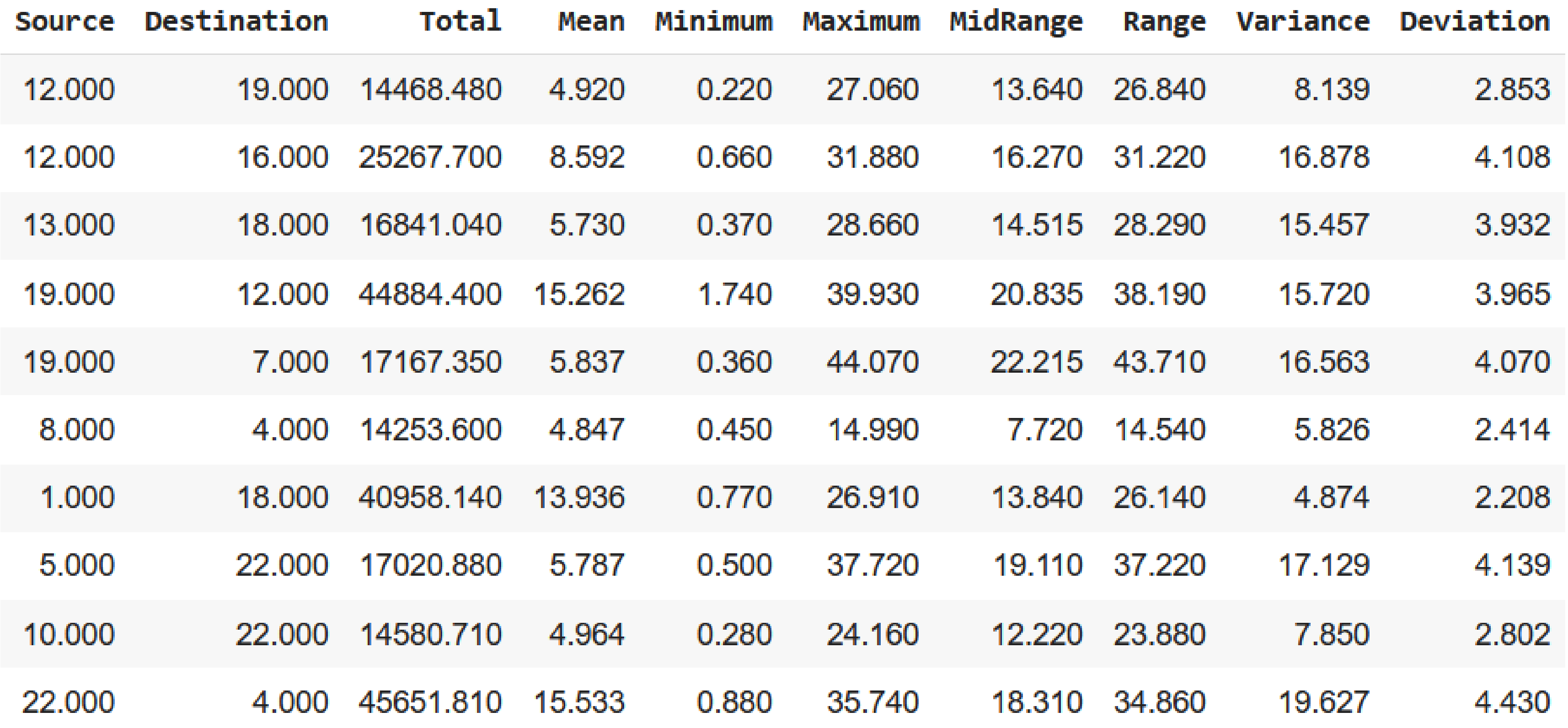


```
Statistical description of the whole raw dataset
count   1356524.000
mean         82.898
std       18139.356
min           0.000
25%           0.170
50%           1.450
75%           7.590
max     9218069.950
```



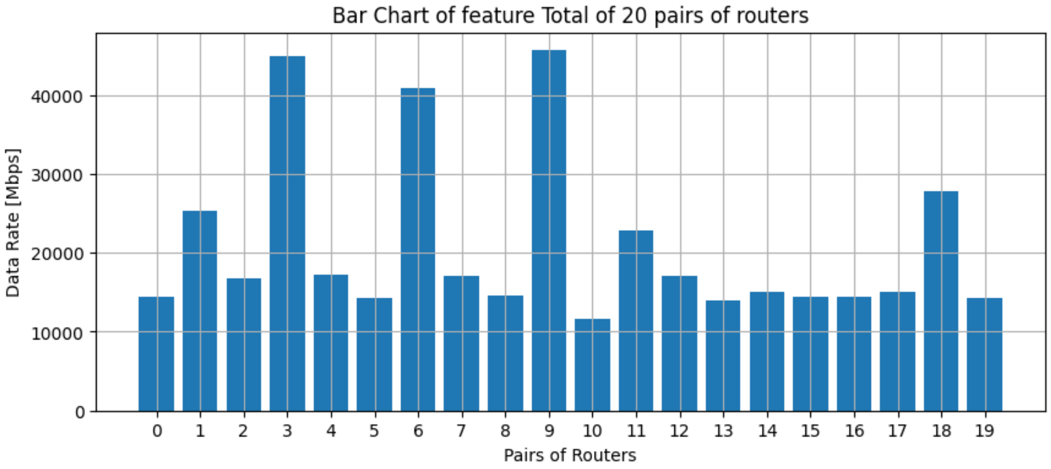
```
Description of Data Rates [Mbps] after applying filters 1 to 5
count   67625.000
mean        6.480
std         4.979
min         0.180
25%         2.860
50%         4.910
75%         8.850
max        69.400
```

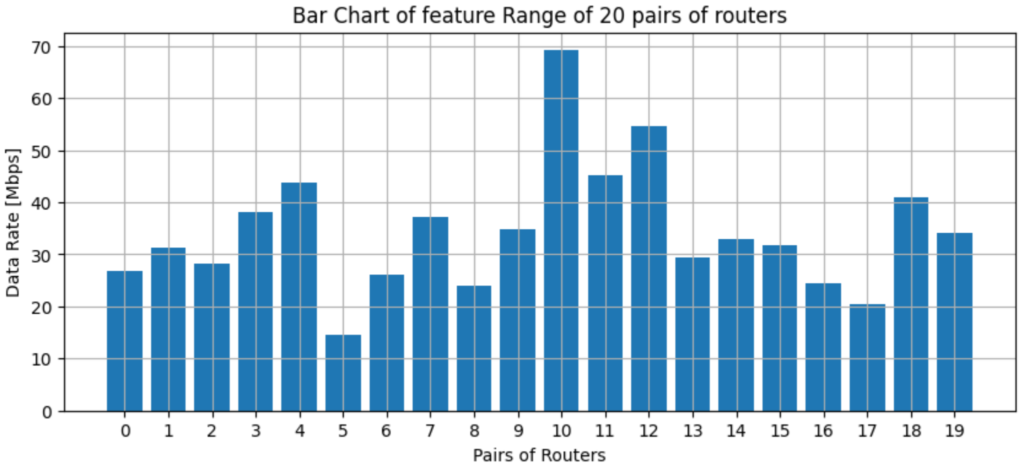
Statistical Description of selected subset



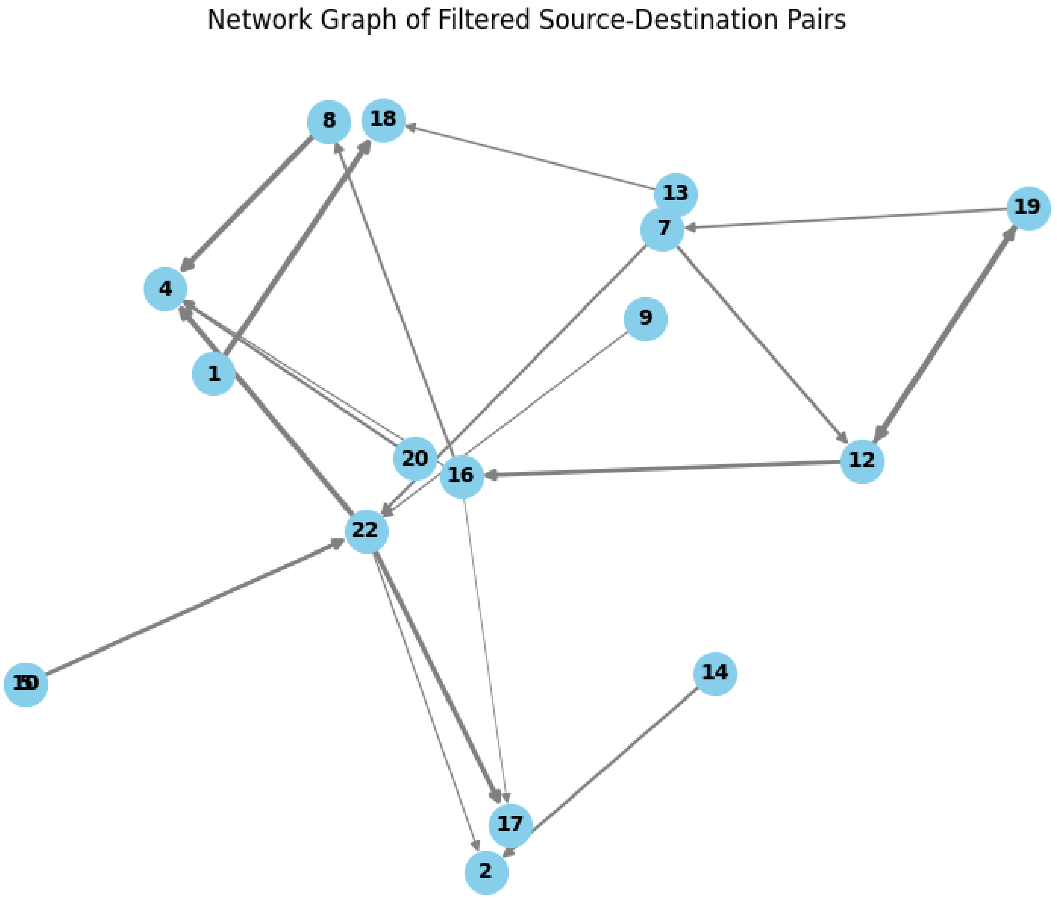
Source	Destination	Total	Mean	Minimum	Maximum	MidRange	Range	Variance	Deviation
12.000	19.000	14468.480	4.920	0.220	27.060	13.640	26.840	8.139	2.853
12.000	16.000	25267.700	8.592	0.660	31.880	16.270	31.220	16.878	4.108
13.000	18.000	16841.040	5.730	0.370	28.660	14.515	28.290	15.457	3.932
19.000	12.000	44884.400	15.262	1.740	39.930	20.835	38.190	15.720	3.965
19.000	7.000	17167.350	5.837	0.360	44.070	22.215	43.710	16.563	4.070
8.000	4.000	14253.600	4.847	0.450	14.990	7.720	14.540	5.826	2.414
1.000	18.000	40958.140	13.936	0.770	26.910	13.840	26.140	4.874	2.208
5.000	22.000	17020.880	5.787	0.500	37.720	19.110	37.220	17.129	4.139
10.000	22.000	14580.710	4.964	0.280	24.160	12.220	23.880	7.850	2.802
22.000	4.000	45651.810	15.533	0.880	35.740	18.310	34.860	19.627	4.430

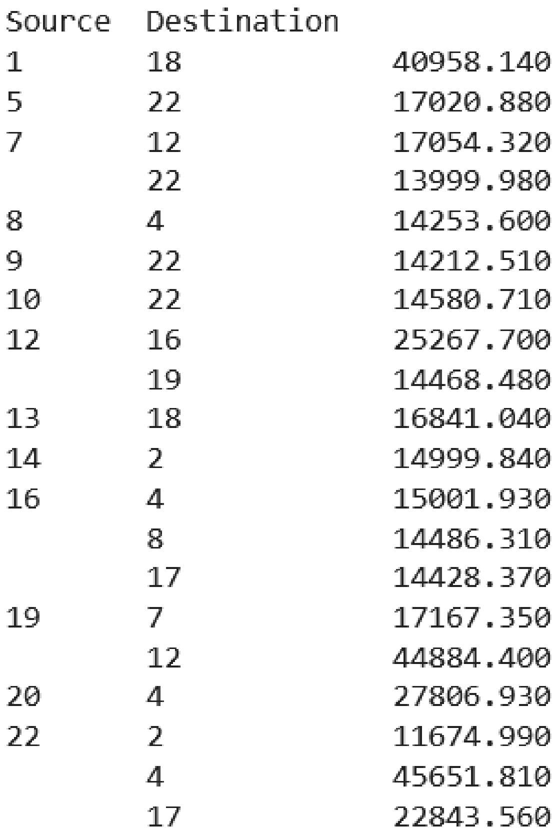
Total and Range features of data rate





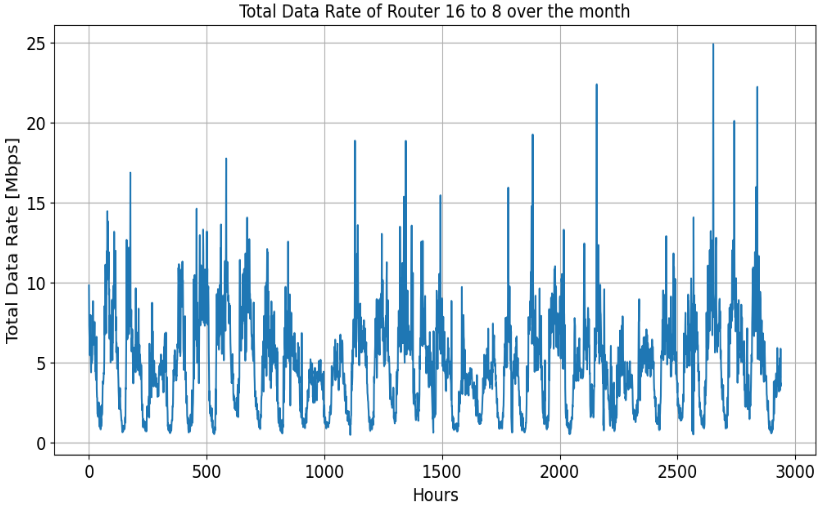
Selected subset of source-destination pairs



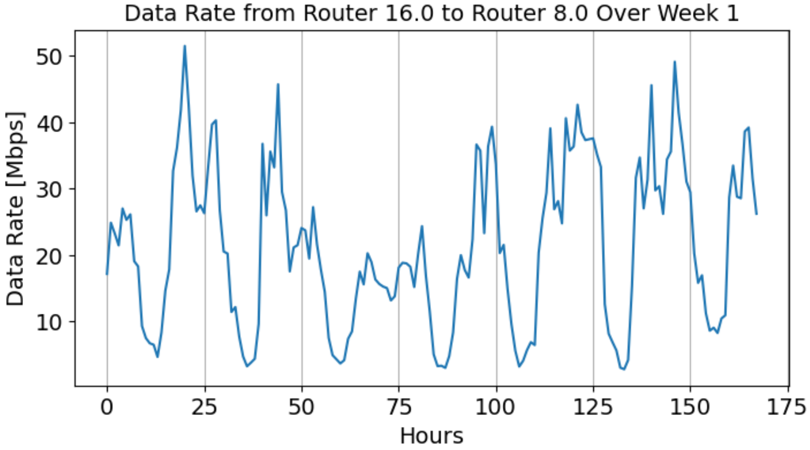


```
Source  Destination
1       18             40958.140
5       22             17020.880
7       12             17054.320
        22             13999.980
8       4              14253.600
9       22             14212.510
10      22             14580.710
12      16             25267.700
        19             14468.480
13      18             16841.040
14      2              14999.840
16      4              15001.930
        8              14486.310
        17             14428.370
19      7              17167.350
        12             44884.400
20      4              27806.930
22      2              11674.990
        4              45651.810
        17             22843.560
```

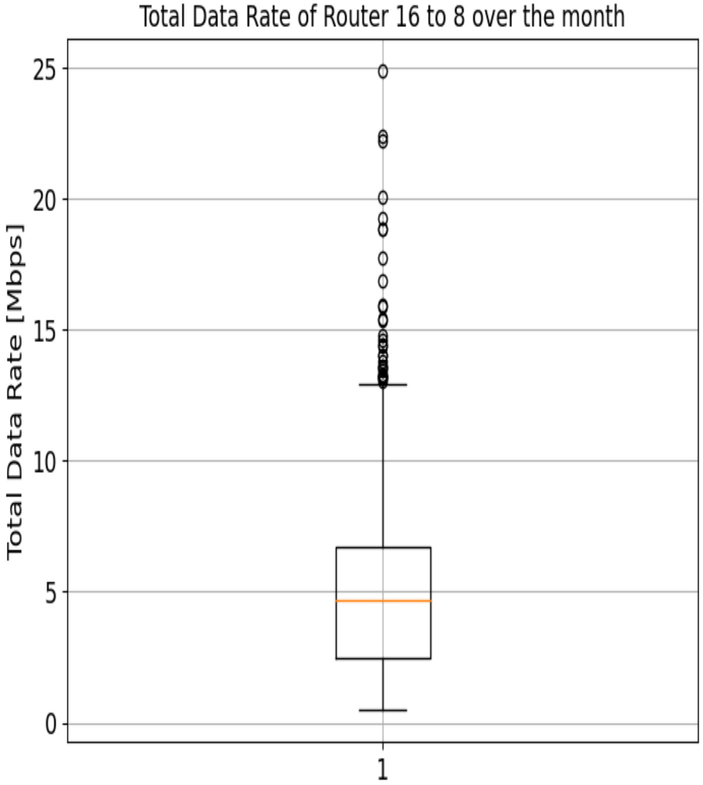
Data Rate of Router 16 to 8 over the whole month



Total Data Rate of Router 16 to 8 over the month

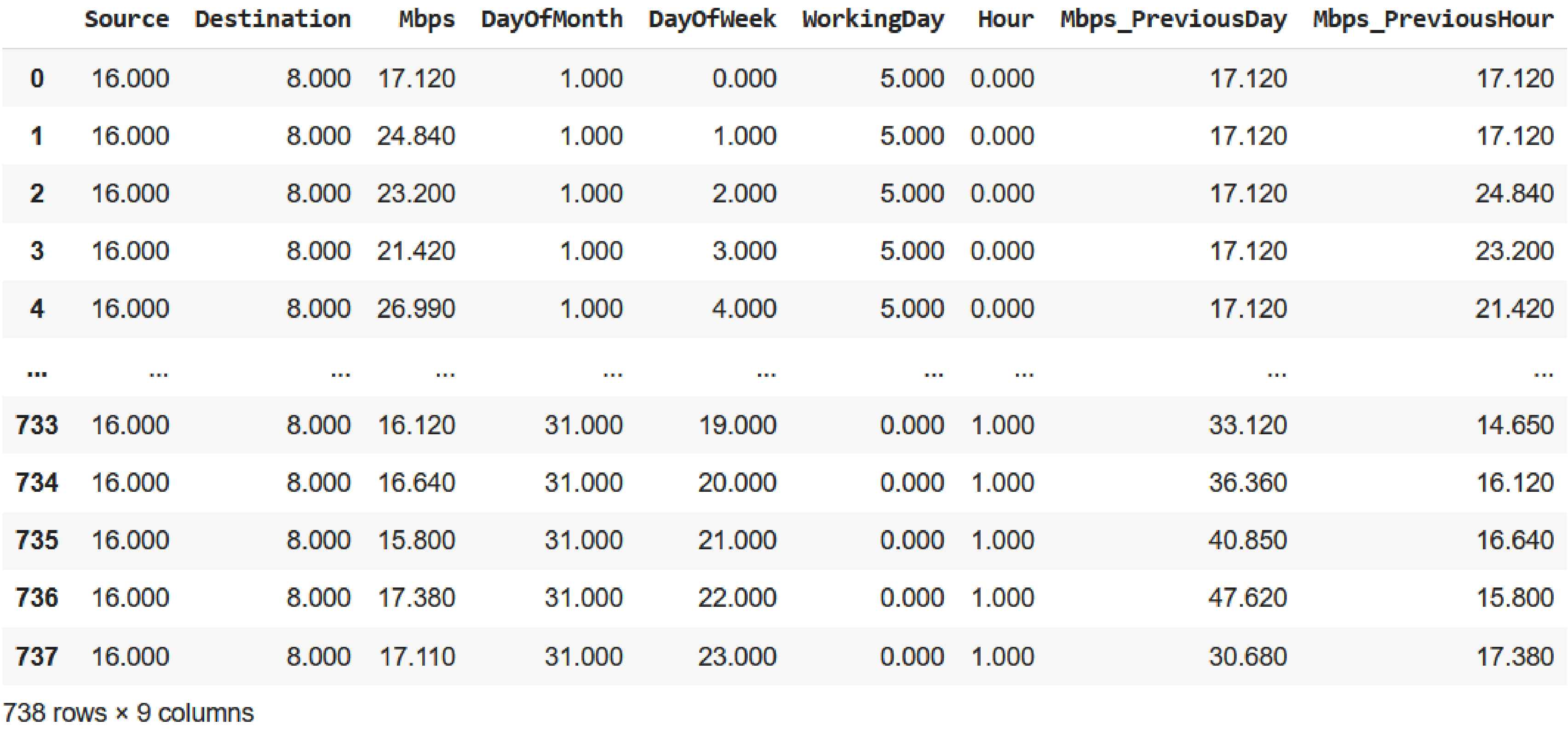


Data Rate from Router 16.0 to Router 8.0 Over Week 1



Total Data Rate of Router 16 to 8 over the month

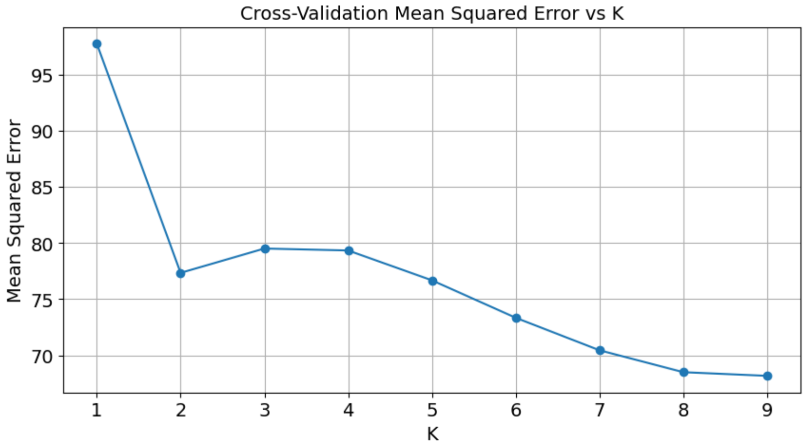
Feature Extraction



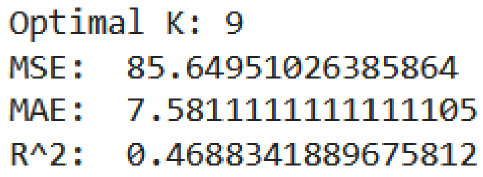
	Source	Destination	Mbps	DayOfMonth	DayOfWeek	WorkingDay	Hour	Mbps_PreviousDay	Mbps_PreviousHour
0	16.000	8.000	17.120	1.000	0.000	5.000	0.000	17.120	17.120
1	16.000	8.000	24.840	1.000	1.000	5.000	0.000	17.120	17.120
2	16.000	8.000	23.200	1.000	2.000	5.000	0.000	17.120	24.840
3	16.000	8.000	21.420	1.000	3.000	5.000	0.000	17.120	23.200
4	16.000	8.000	26.990	1.000	4.000	5.000	0.000	17.120	21.420
...	...	...	...	...	...	...	...	...	...
733	16.000	8.000	16.120	31.000	19.000	0.000	1.000	33.120	14.650
734	16.000	8.000	16.640	31.000	20.000	0.000	1.000	36.360	16.120
735	16.000	8.000	15.800	31.000	21.000	0.000	1.000	40.850	16.640
736	16.000	8.000	17.380	31.000	22.000	0.000	1.000	47.620	15.800
737	16.000	8.000	17.110	31.000	23.000	0.000	1.000	30.680	17.380

738 rows × 9 columns

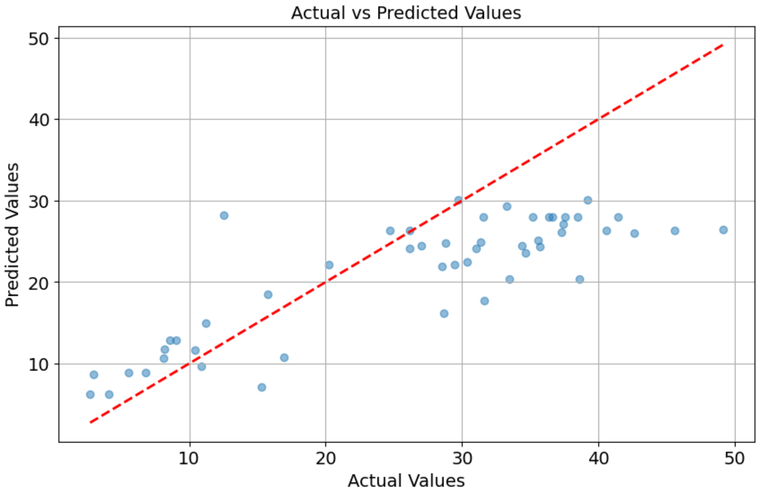
K-Neighbor Regressor



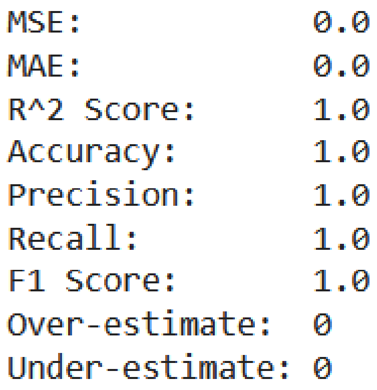
Performance Metrics for traffic forecasting



```
Optimal K: 9
MSE:  85.64951026385864
MAE:  7.5811111111111105
R^2:  0.4688341889675812
```

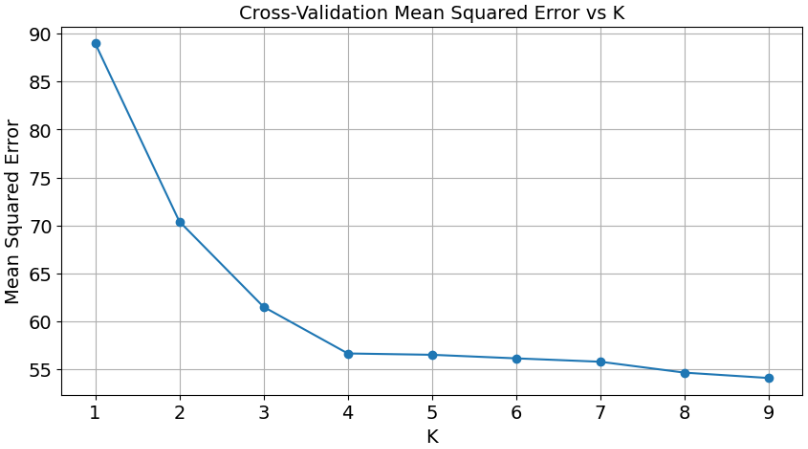


Performance Metrics for prediction of 50 Mbps Channel's count

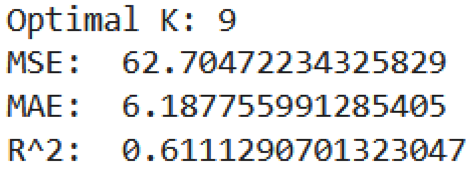


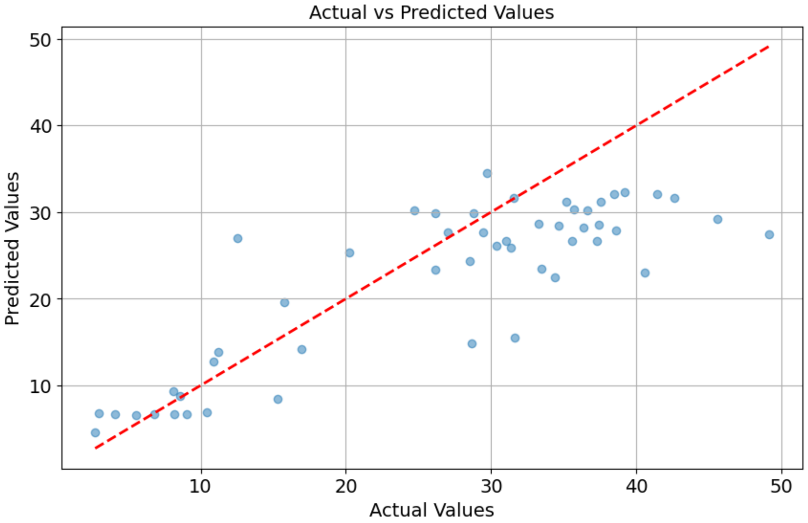
```
MSE:            0.0
MAE:            0.0
R^2 Score:      1.0
Accuracy:       1.0
Precision:      1.0
Recall:         1.0
F1 Score:       1.0
Over-estimate:  0
Under-estimate: 0
```

K-Neighbor Regressor (Without normalization)

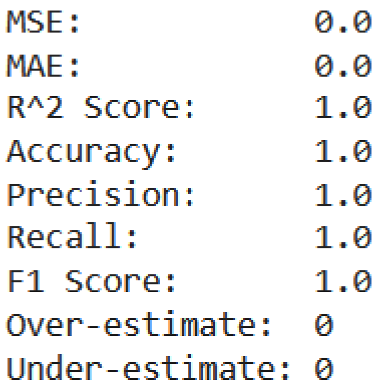


Performance Metrics for traffic forecasting

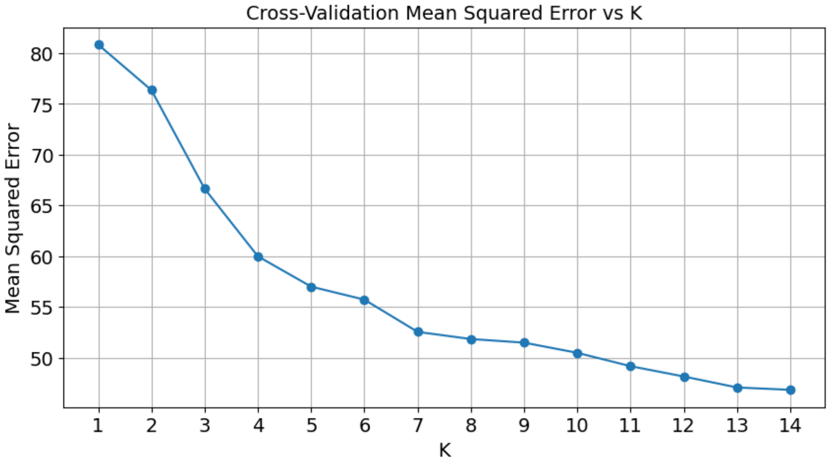

```
Optimal K: 9
MSE:  62.70472234325829
MAE:  6.187755991285405
R^2:  0.6111290701323047
```



Performance Metrics for Channel forecasting

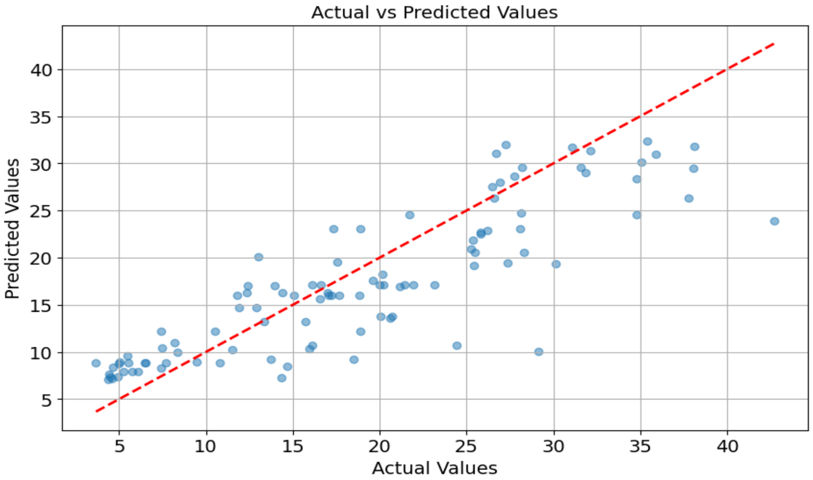

```
MSE:            0.0
MAE:            0.0
R^2 Score:      1.0
Accuracy:       1.0
Precision:      1.0
Recall:         1.0
F1 Score:       1.0
Over-estimate:  0
Under-estimate: 0
```

K-Neighbor Regressor (Larger Dataset)



Performance Metrics for traffic forecasting

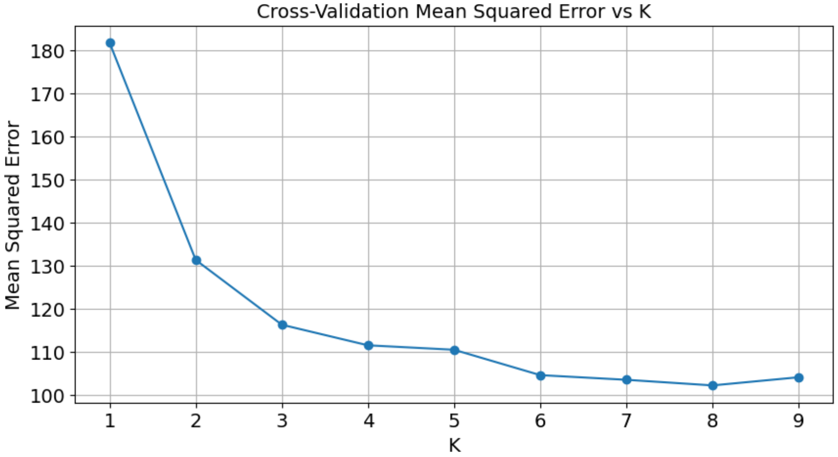
```
Optimal K: 14
MSE:  27.806879460497065
MAE:  4.081909476661952
R^2:  0.7079396008415211
```



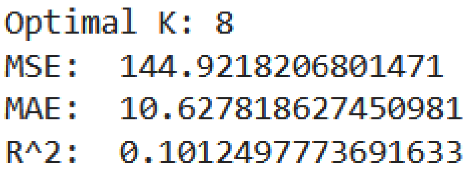
Performance Metrics for Channel forecasting

```
MSE:           0.0
MAE:           0.0
R^2 Score:     1.0
Accuracy:      1.0
Precision:     1.0
Recall:        1.0
F1 Score:      1.0
Over-estimate: 0
Under-estimate: 0
```

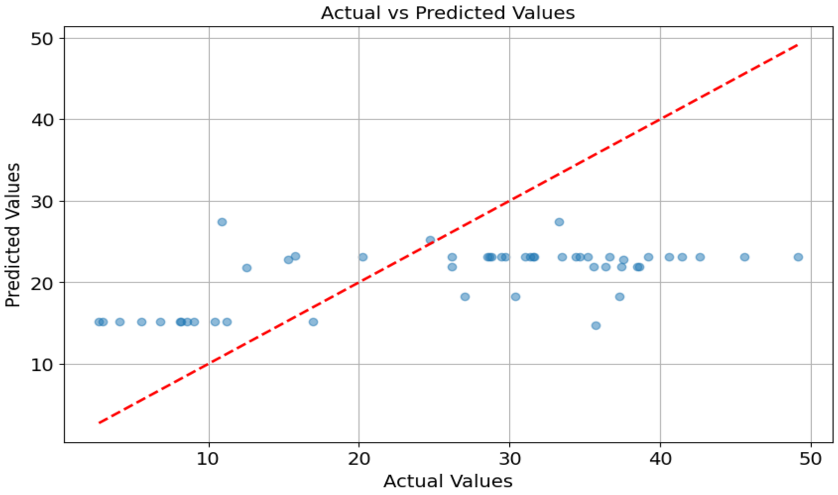
K-Neighbor Regressor
(Removing 'DayOfWeek', 'Hour', 'Mbps_PreviousHour’ features)



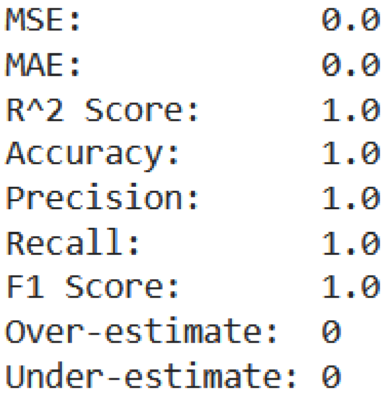
Performance Metrics for traffic forecasting



```
Optimal K: 8
MSE:   144.9218206801471
MAE:   10.627818627450981
R^2:   0.1012497773691633
```

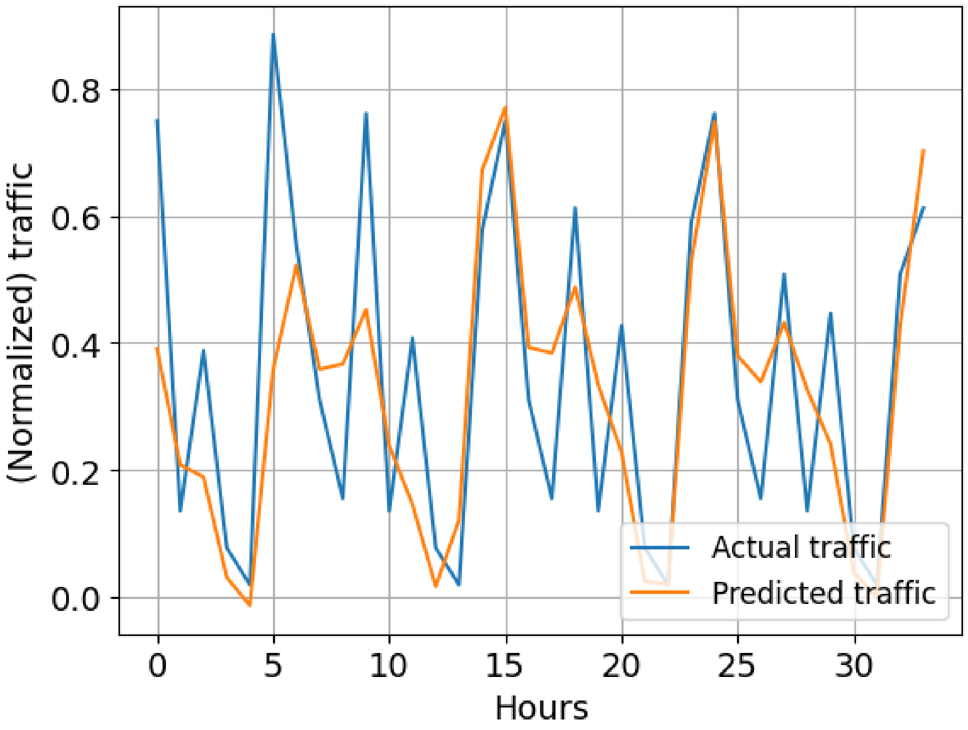


Performance Metrics for prediction of 50 Mbps Channel’s count

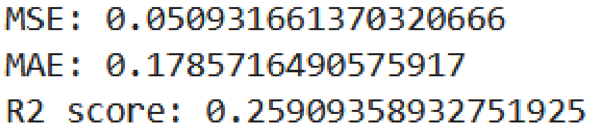


```
MSE:             0.0
MAE:             0.0
R^2 Score:       1.0
Accuracy:        1.0
Precision:       1.0
Recall:          1.0
F1 Score:        1.0
Over-estimate:   0
Under-estimate:  0
```

LSTM (before tuning of hyperparameters)

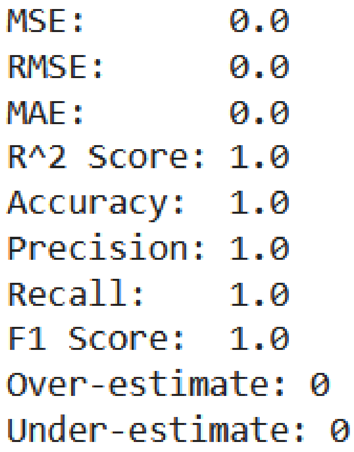


Performance Metrics for traffic forecasting



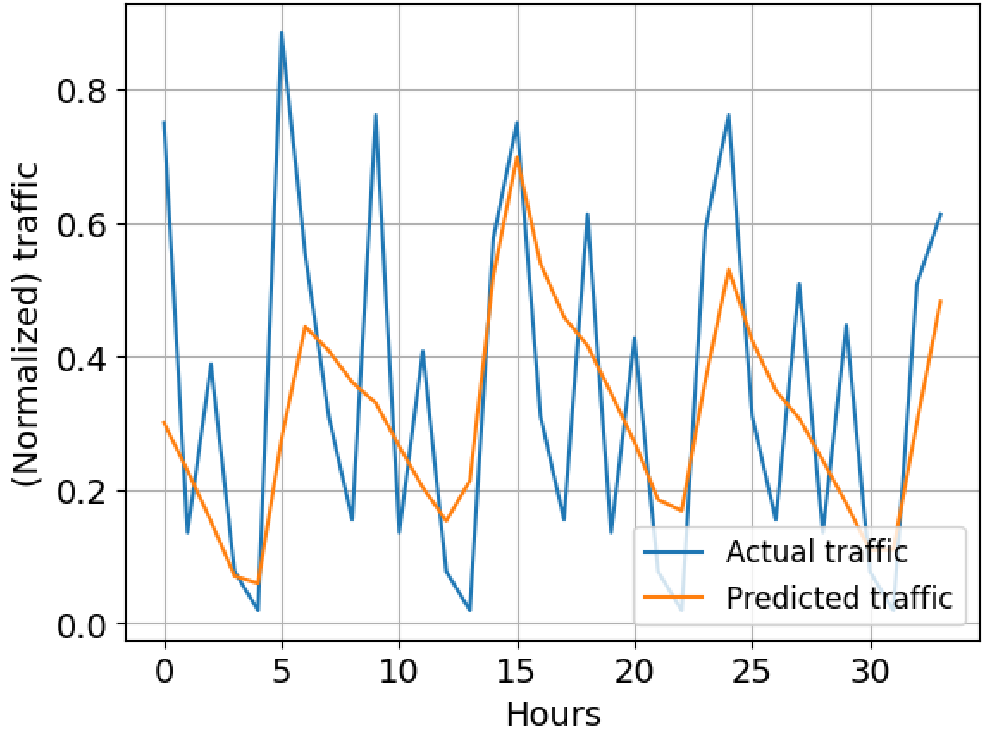
```
MSE: 0.05093166137032066
MAE: 0.1785716490575917
R2 score: 0.25909358932751925
```

Performance Metrics for Channel forecasting

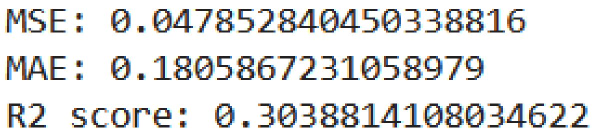


```
MSE:        0.0
RMSE:       0.0
MAE:        0.0
R^2 Score:  1.0
Accuracy:   1.0
Precision:  1.0
Recall:     1.0
F1 Score:   1.0
Over-estimate: 0
Under-estimate: 0
```

LSTM (after tuning of hyperparameters)

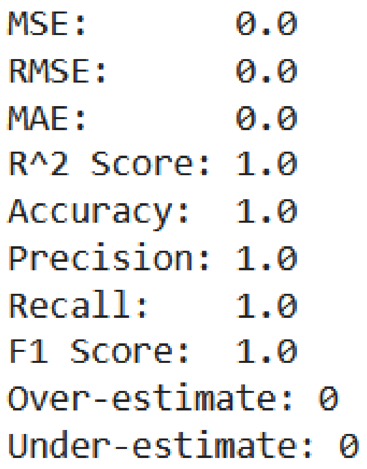


Performance Metrics for traffic forecasting



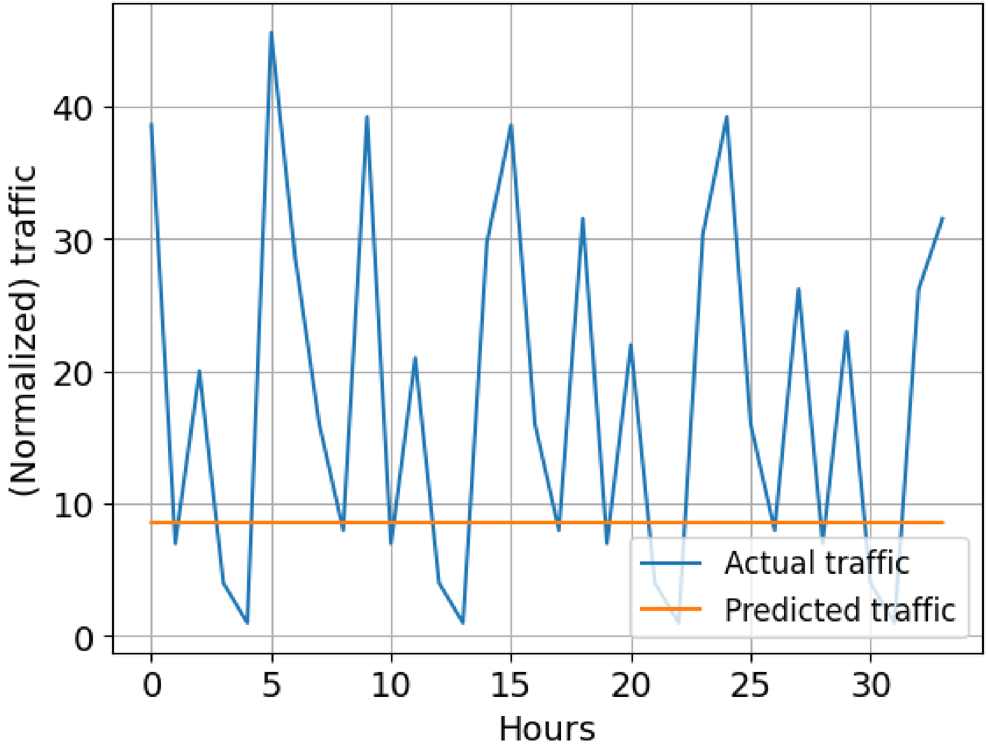
```
MSE: 0.047852840450338816
MAE: 0.1805867231058979
R2 score: 0.3038814108034622
```

Performance Metrics for Channel forecasting

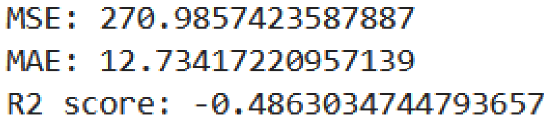


```
MSE:        0.0
RMSE:       0.0
MAE:        0.0
R^2 Score:  1.0
Accuracy:   1.0
Precision:  1.0
Recall:     1.0
F1 Score:   1.0
Over-estimate: 0
Under-estimate: 0
```

LSTM (before tuning of hyperparameters) (Without Normalization)

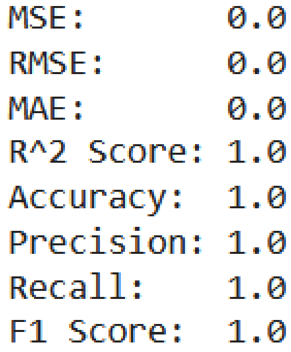


Performance Metrics for traffic forecasting



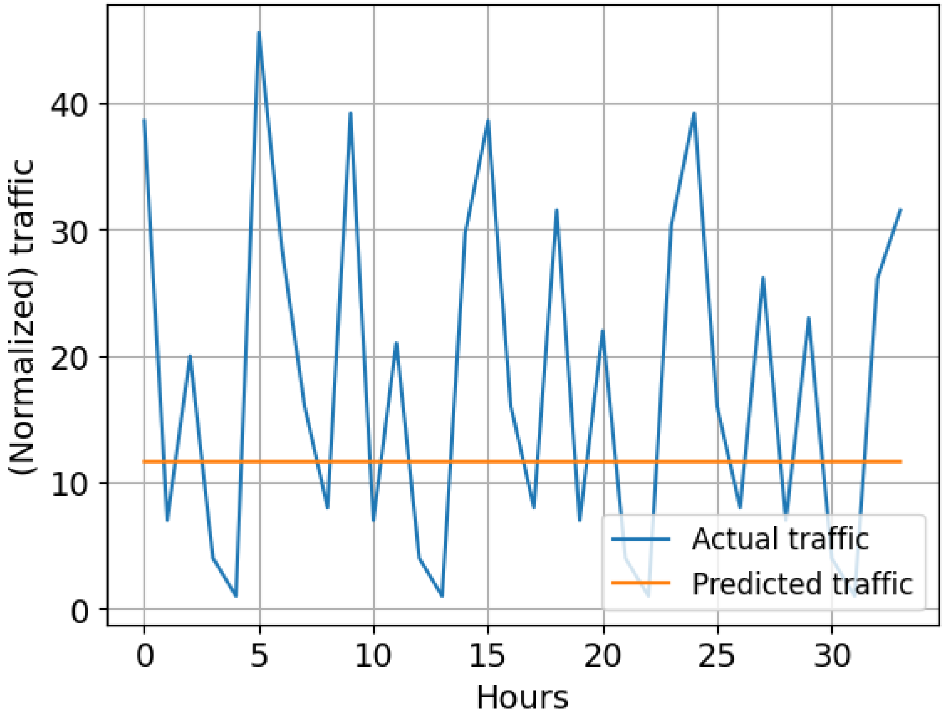
```
MSE: 270.9857423587887
MAE: 12.73417220957139
R2 score: -0.4863034744793657
```

Performance Metrics for Channel forecasting

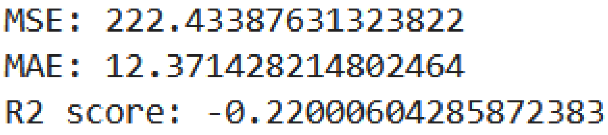


```
MSE:        0.0
RMSE:       0.0
MAE:        0.0
R^2 Score:  1.0
Accuracy:   1.0
Precision:  1.0
Recall:     1.0
F1 Score:   1.0
```

LSTM (after tuning of hyperparameters) (Without Normalization)

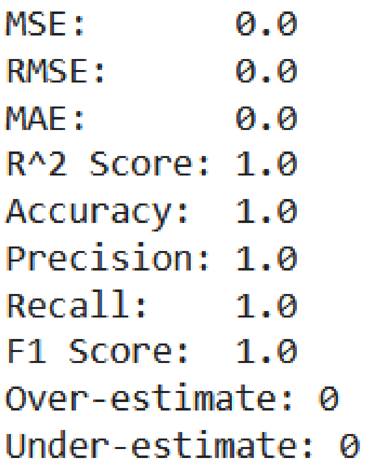


Performance Metrics for traffic forecasting



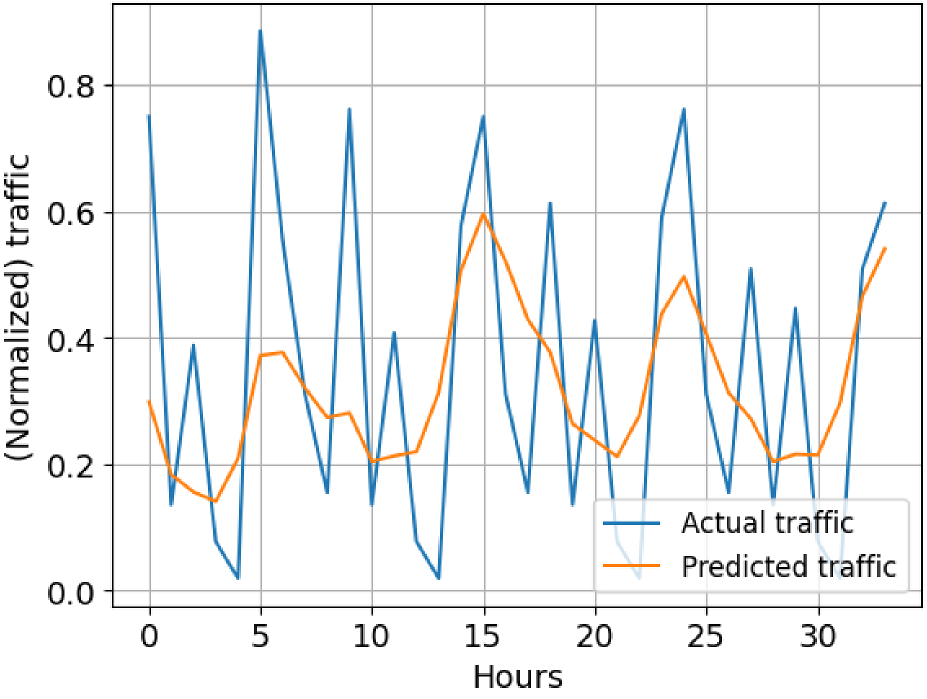
```
MSE: 222.43387631323822
MAE: 12.371428214802464
R2 score: -0.22000604285872383
```

Performance Metrics for Channel forecasting

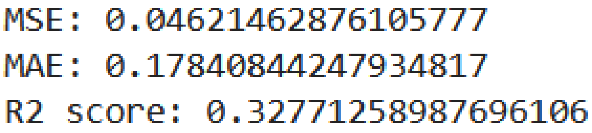


```
MSE:        0.0
RMSE:       0.0
MAE:        0.0
R^2 Score:  1.0
Accuracy:   1.0
Precision:  1.0
Recall:     1.0
F1 Score:   1.0
Over-estimate: 0
Under-estimate: 0
```

LSTM (before tuning of hyperparameters)
(Removing 'DayOfWeek', 'Hour', 'Mbps_PreviousHour' features)

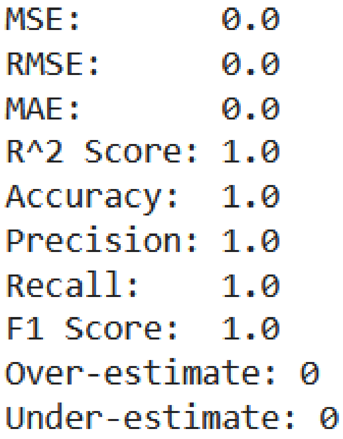


Performance Metrics for traffic forecasting



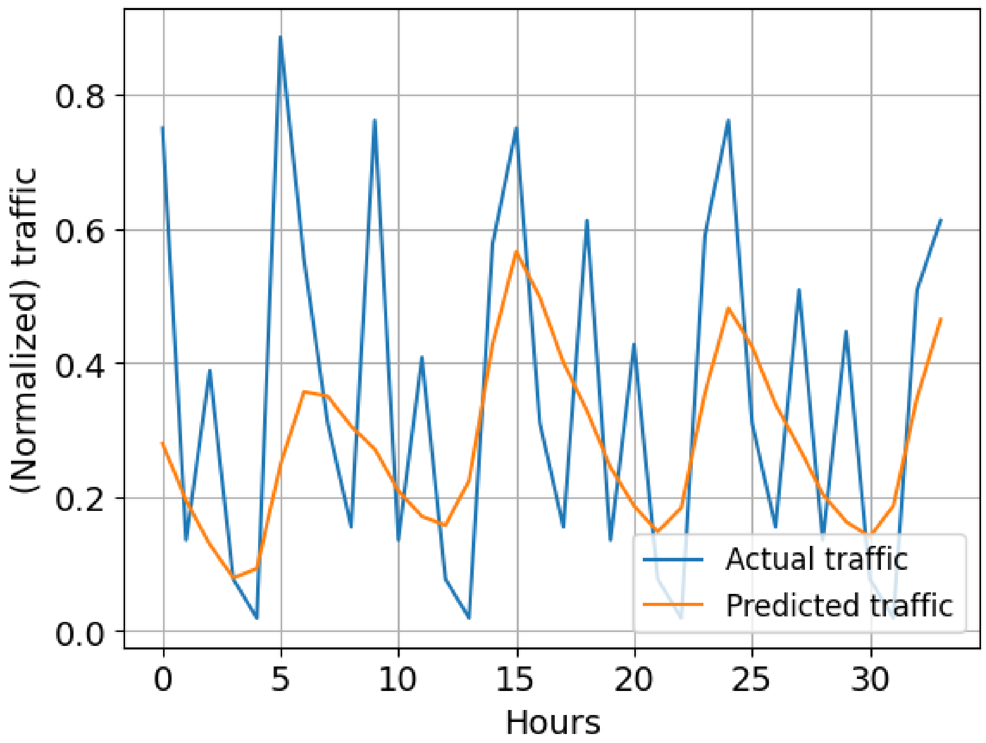
```
MSE: 0.04621462876105777
MAE: 0.17840844247934817
R2 score: 0.32771258987696106
```

Performance Metrics for Channel forecasting

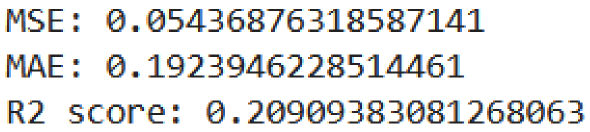


```
MSE:        0.0
RMSE:       0.0
MAE:        0.0
R^2 Score:  1.0
Accuracy:   1.0
Precision:  1.0
Recall:     1.0
F1 Score:   1.0
Over-estimate: 0
Under-estimate: 0
```

LSTM (after tuning of hyperparameters) (Removing 'DayOfWeek', 'Hour', 'Mbps_PreviousHour’ features)

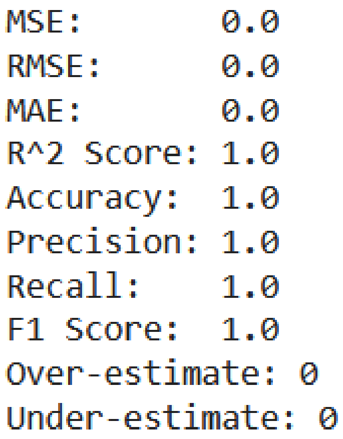


Performance Metrics for traffic forecasting



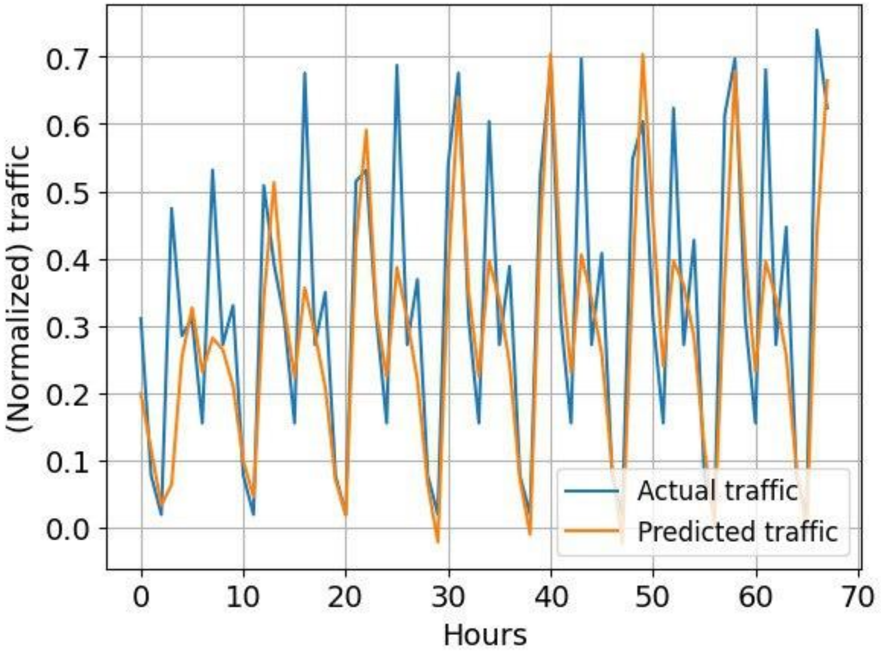
```
MSE: 0.05436876318587141
MAE: 0.1923946228514461
R2 score: 0.20909383081268063
```

Performance Metrics for Channel forecasting

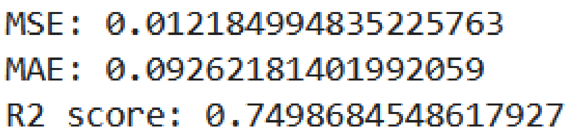


```
MSE:        0.0
RMSE:       0.0
MAE:        0.0
R^2 Score:  1.0
Accuracy:   1.0
Precision:  1.0
Recall:     1.0
F1 Score:   1.0
Over-estimate: 0
Under-estimate: 0
```

LSTM (before tuning of hyperparameters) (Larger Dataset)

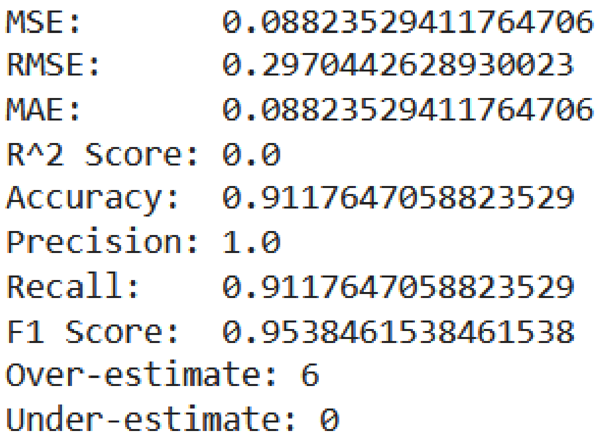


Performance Metrics for traffic forecasting



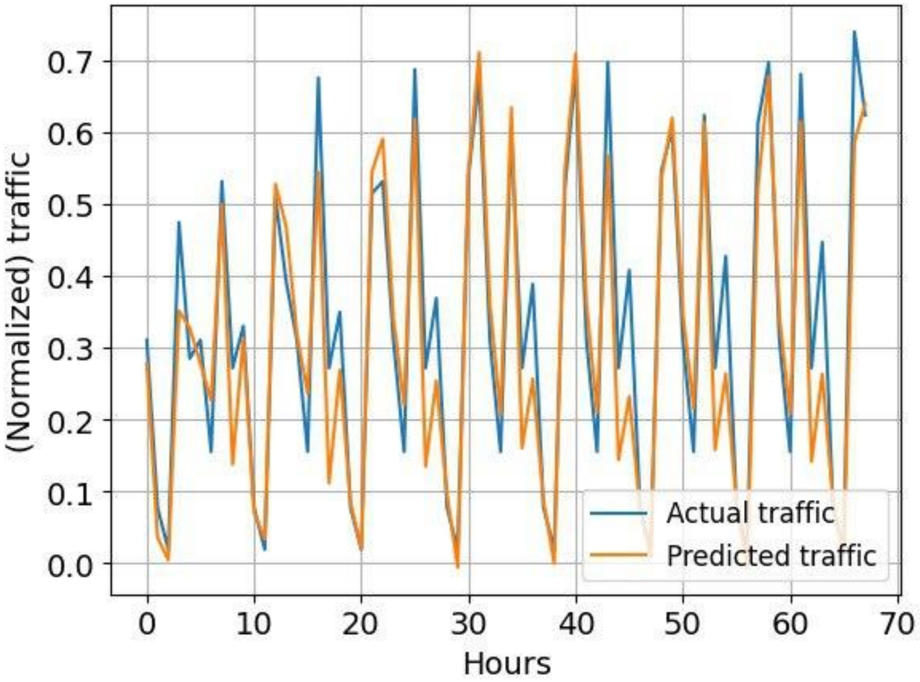
```
MSE: 0.012184994835225763
MAE: 0.09262181401992059
R2 score: 0.7498684548617927
```

Performance Metrics for Channel forecasting

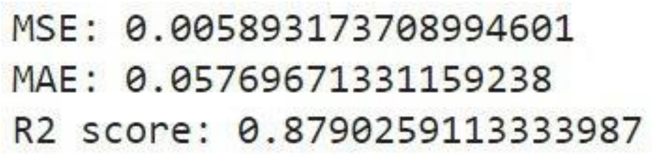


```
MSE:        0.08823529411764706
RMSE:       0.29704426289300235
MAE:        0.08823529411764706
R^2 Score: 0.0
Accuracy:  0.9117647058823529
Precision: 1.0
Recall:    0.9117647058823529
F1 Score:  0.9538461538461538
Over-estimate: 6
Under-estimate: 0
```

LSTM (after tuning of hyperparameters) (Larger Dataset)

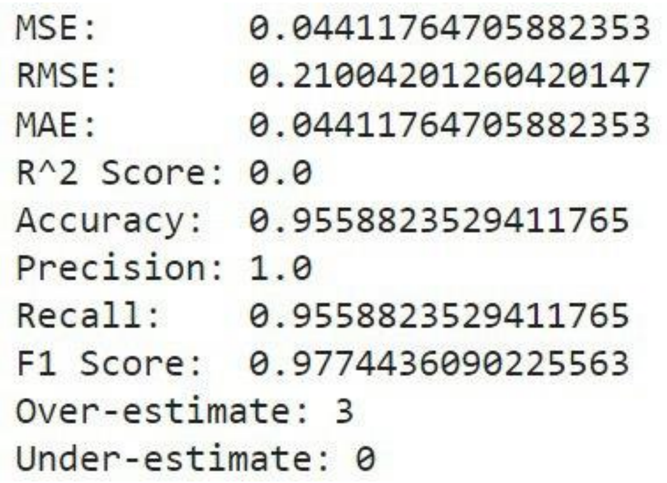


Performance Metrics for traffic forecasting



```
MSE: 0.005893173708994601
MAE: 0.05769671331159238
R2 score: 0.8790259113333987
```

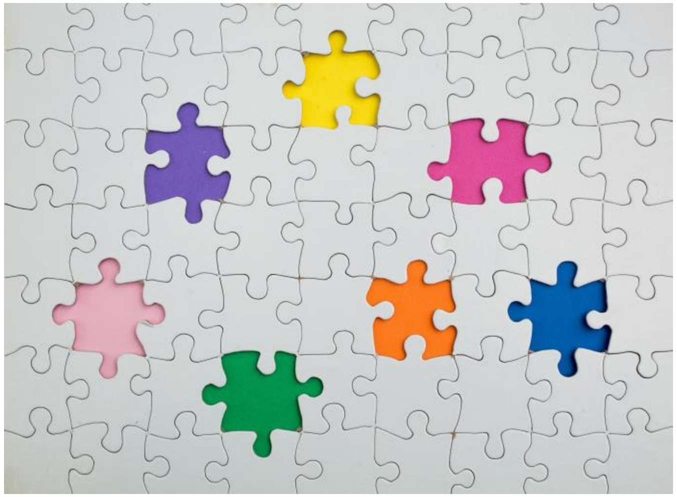
Performance Metrics for Channel forecasting



```
MSE:        0.04411764705882353
RMSE:       0.21004201260420147
MAE:        0.04411764705882353
R^2 Score: 0.0
Accuracy:   0.9558823529411765
Precision: 1.0
Recall:     0.9558823529411765
F1 Score:   0.9774436090225563
Over-estimate: 3
Under-estimate: 0
```

KNR and LSTM vs. Missing values

Missing values in the dataset significantly impact on the performance of the KNR and LSTM algorithms as they significantly encountered error when they had received a dataset with some missing values randomly distributed through the dataset and different features.



Conclusion

- At all test scenarios in terms of traffic prediction, KNR worked worse than LSTM as it reaches considerable higher values of MSE, MAE, and R2 Score.
- When the number of features reduces KNR acts worse, while LSTM works as well as before.
- Unnormalized features negatively impact on the performance of LSTM, while KNR is not too sensitive about it.
- Hyper-parameter tuning significantly impact on the improvement of algorithm when the size of the dataset is large.

Conclusion

- Both KNR and LSTM reach the same result in terms of channel prediction as it is an easier task compared to the traffic prediction. However, LSTM has more computational complexity. So, the more inputs, the much longer time we have to wait for running LSTM.
- The least value of MSE obtained for running of LSTM on a larger dataset.
- We can conclude than KNR can be better than LSTM when the size of dataset is large and we are not too sensitive about the exact values which are going to be predicted.

Thank
You