Forecasting mobile network traffic
Assignment for a PhD position at IMDEA Networks Institute
Networks Data Science group

Darix SAMANI SIEWE (darix.siewe@aims.ac.rw)

Abstract

In this work, we present a regression model to forecast the Internet traffic activity in the city of Milan using a Big Data Challenge Data that was collected over two months from November to December 2013. To achieve this goal, we initially performed a basic data exploration to get insight into the data set, which reveals that Seasonal Auto-Regressive Integrating Moving Average (SARAMIX) can suit better for the prediction. The evaluation process of this model showed an accuracy performance of 80%.

Keywords: Time Series, Telecommunication, Internet Traffic, Milan city.

Introduction

Time-series analysis and forecasting are crucial for predicting future trends, behaviors, and behaviors based on historical data. There are many models for time series forecasting from classical models such as statistical models to machine learning and deep learning models. Choosing the best model that must adapt to our time series requires some investigation in data analysis to identify patterns and assess the assumption (stationary, seasonality, etc.) of each model. Our implementation follows the pipeline in Figure 1. This work is organized into two tasks: the first task (Section 1) is focused on the exploratory data analysis of our time series dataset, and the second task (Section 2) is about training and building the model.

Figure 1: Workflow of the work.

1 Task 1: Exploratory Data Analysis

In this section, we will focus on data analysis of our dataset in order to extract relevant information.

1.1 Description of Hardware and Software

The following table gives a brief description of the hardware and software we used during this work.

Table 1: Brief description of Hardware and software

Information	Value
Processor	Intel® Core™ i7-10700 CPU @ 2.90GHz × 16
Graphic	Mesa Intel® UHD Graphics 630 (CML GT2)
Disk Capacity	512.1 GB
Operating System	Linux, Debian 11
RAM	8 RAMs
Programming Language	Python3.9.2

1.2 Preparation and Processing of our dataset: data cleaning

The purpose of this section is to explain the preparation and pre-processing of the dataset. The dataset used in this work is accessible via the website [3, 2]. The size of our dataset is approximately 20GB and the number of features

of our dataset is 7. Loading this dataset using a standard process was a bit challenging due to the limited capacity of our hardware. Therefore, we speed up the loading time using DASK [1], which implements a parallel process. Afterward, we extract only the relevant features such as time interval, square id, and internet traffic activity, which are all numerical values. The data cleaning is performed on these selected features, replacing the NA missing value value by zero. Finally, we convert all timestamps into appropriate data formats and aggregate data hourly. The code of this data cleaning and pre-processing step is in the notebook name **pre-processing data.ipynb** file.

1.3 Distribution of our dataset

In this section of our work, we were asked to plot the density probability function of the internet traffic in the city, using over 10000 computer samples that each represent the total two-month traffic in one geographical area. Since Internet traffic activity is continuous data, we need to convert it into numbers of bins, there are many methods to do that, and one of the most popular methods is freedman diaconis because it handles well the outlier in the data and captures more information.

$$bins\_numbers = 2 * (IQR)/n^{\frac{1}{3}}$$

where $IQR = IQ(0.75) - IQ(0.25)$ is the intervale interquantile and n the number of sample in our dataset.

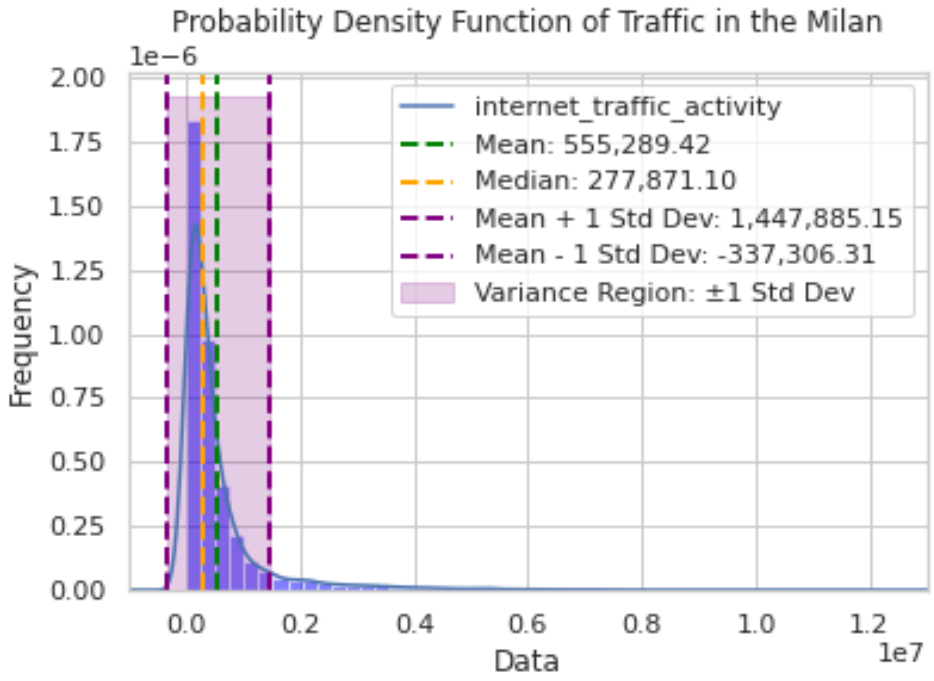

Figure 2: Distribution of internet traffic activity in one geographical area in the city Milan.

The following table gives the summary statistic of this distribution :

Table 2: Summary statistic of this distribution.

Mean	Median	Standard Deviation	Variance	Skewness	Kurtosis
555,289.42	277,871.10	892,595.73	796,727,141,100.52	4.26	25.50

Figure 2 shows us that our data is much more distributed between my mean and the median, we can see that there is almost no outlier in our data because most of the data is distributed in the variance region.

1.4 Time series of network traffic during the first two weeks at three areas

The purpose of this section is to plot the time series of internet traffic activity of three areas in the city of Milan during the two months: the area with the highest total traffic, the area with Square ID 4159, and the area with Square ID 4556 during the first two weeks. We first need to find the square ID with the highest traffic internet

activity. To achieve that, we need to compute the total traffic activity in each area and sort by decreasing order. The following table gives us the five areas with the highest total traffic activity :

Table 3: Five highest areas with total internet traffic activity.

square id	total internet traffic activity
5161	1.274006e+07
5059	1.117085e+07
5259	1.048578e+07
5061	9.584334e+06
5258	8.707440e+06

Table 3 shows the areas with total internet traffic is the area with square id: 5161.

The following plot is the time series plot of three areas in the city of Milan of their internet traffic activity during the first two weeks

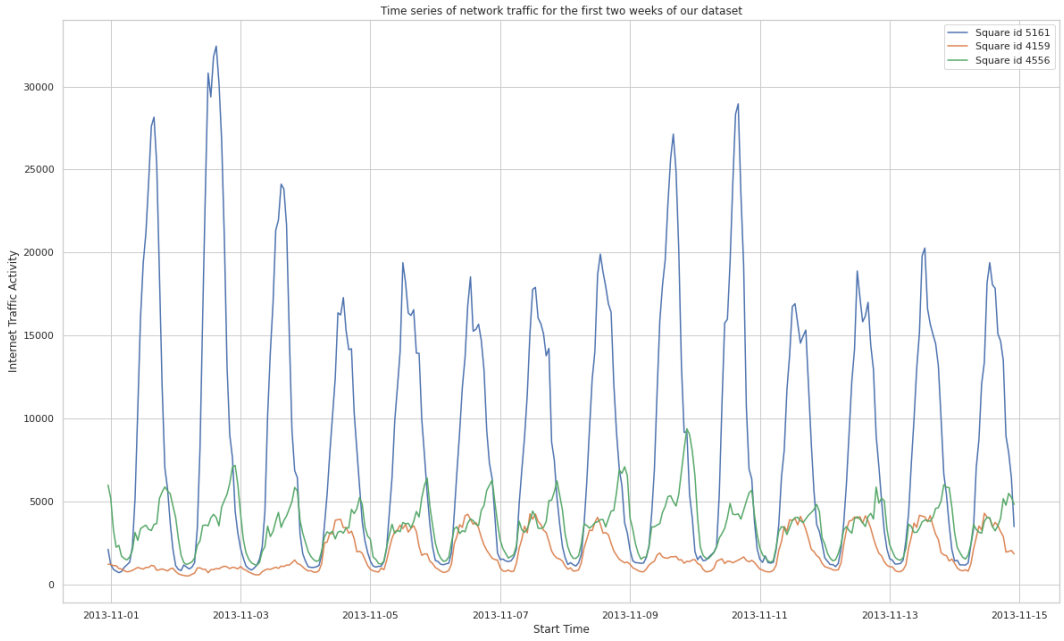

Figure 3: Time series plot of the internet activity of areas: 5161, 4159, 4556 during first two weeks.

From Figure 3, we can see many differences between these three areas. The interpretation is as follows:

- square ID 5161: In the area with the highest total activity traffic, we observed that the internet traffic activity has the same oscillation(period) but with a different peak(different amplitude).
- square ID 4159: This area has two variations: the traffic internet is almost constant(line) from 10-11-2013 to 04-11-2013 and also from 09-11-2013 to 13-11-2013 and has the same oscillation(sinusoidal) the other time.
- square ID 4556: In this area, the internet traffic has the same oscillation(period) over the time but with different amplitude(peak).

It is important to note that, the interpretation above is just the graphical interpretation of our time series, furthermore, we need a statistical model to check some of the assumptions.

1.5 Further analysis

In this section, we performed a deeper analysis to find other useful information that will help us better understand the behavior of our time series over time. In order to do that, we analyze the behaviors of our time series by hourly and by a week of day.

The following figure shows us the boxplot of internet traffic in the area with the highest internet traffic by day of the week.

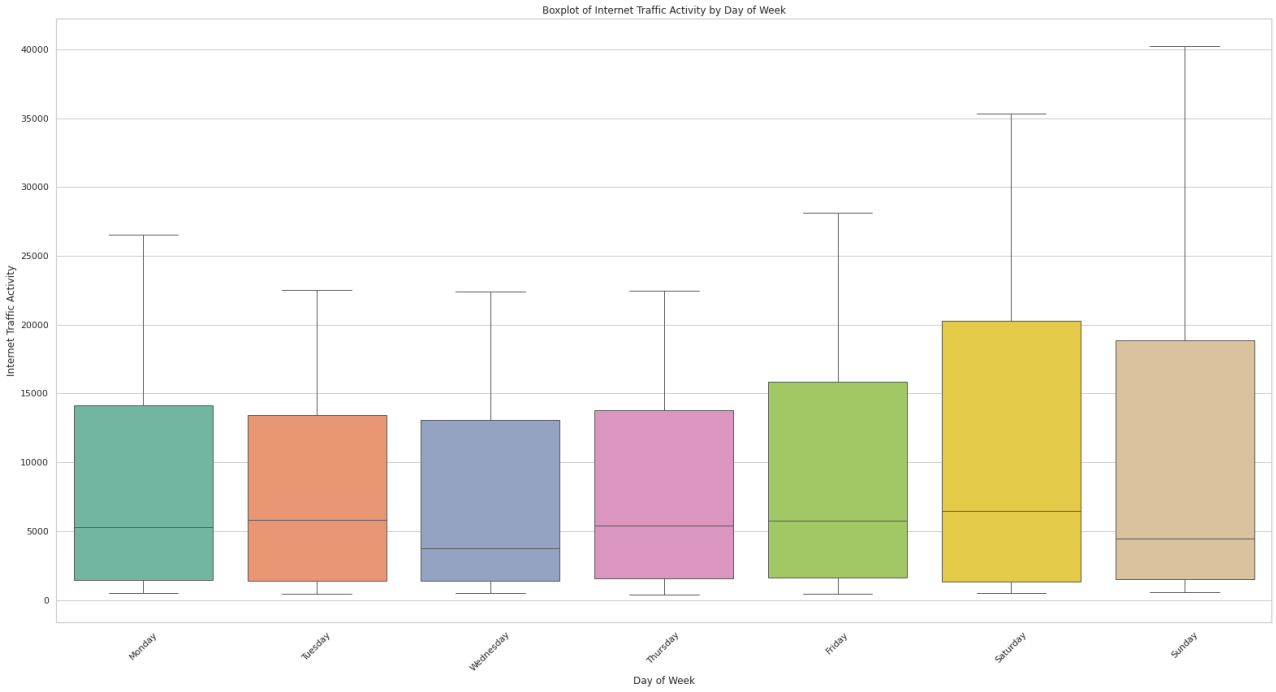

Figure 4: Boxplot of internet traffic of area 5161 by day of week.

In figure 4, we see that all day of the week, we have almost the same average traffic (5000) and the traffic is higher during the weekend.

The following figure is about the internet traffic of 10 top areas (10 areas with the highest internet traffic activity) hourly.

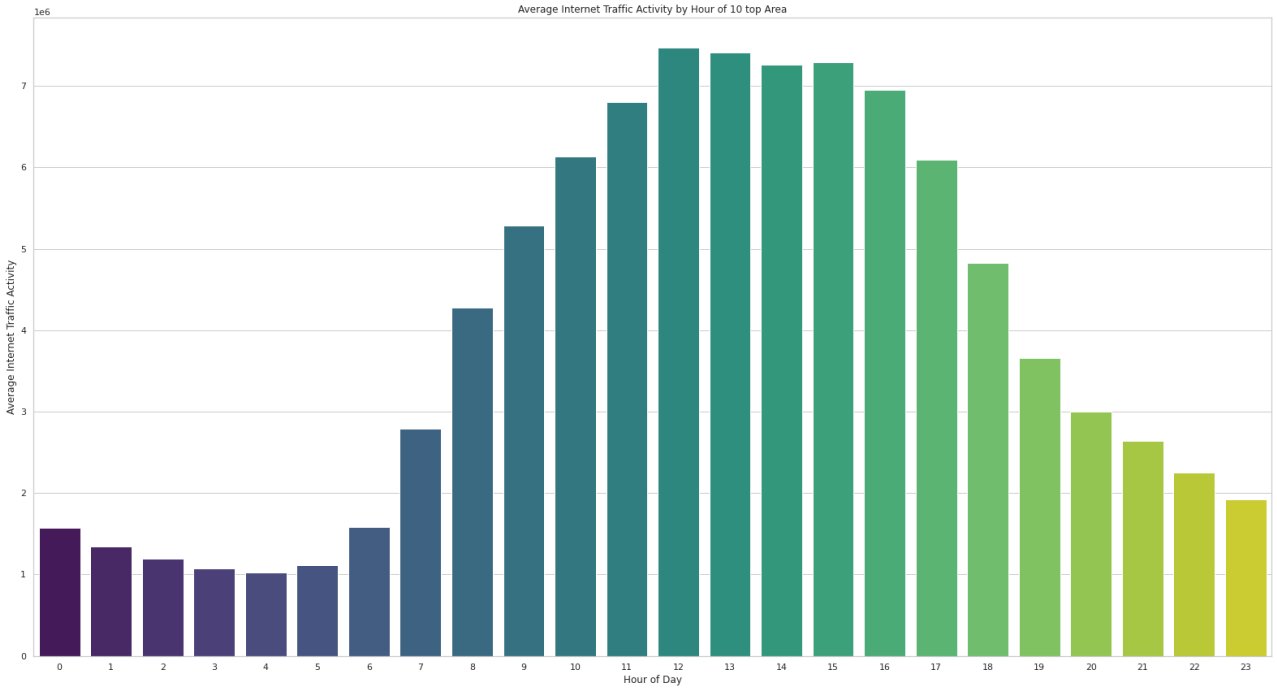

Figure 5: Average internet traffic activity in 10 top areas by hours.

In Figure 5, the behaviors met our expectations because the internet traffic is higher between 7 and 21 hours which is normal.

2 Task 2: Training and building time series Model

This section aims to build a time series model to predict internet traffic activity in three areas of interest. Before delving into the building model, we need to investigate a good choice model that meets our situation well; some models work only with the assumption. In this section, our methodology is to check some assumptions of the most popular time series models. After checking these assumptions, the next step is to find a good parameter for this model that we can use as the most popular statistical method and finally build the model for three areas of interest. The primary purpose of this section is the build a time series model using the data until 15 December to predict the future internet traffic from 16 December to 22 December and compare predicted internet traffic and actual traffic.

2.1 ARIMA model and his variant

In this section, we will plan to use the ARIMA model which is the most popular model for forecasting time series across the world. Before depth into the training model, we need to explain some mathematical explanation of this model and its variants.

2.1.1 ARIMA model explanation

ARIMA is the most popular model for analyzing and forecasting time series. ARIMA stands for Autoregressive integrated moving average essentially creates a linear equation that describes and forecasts your time series data. This equation is generated through three separate parts which can be described as:

- AR Model : Auto-regression: equation terms created based on past data points. Models attempt to predict future values based on past values. AR models require the time series to be stationary. We denote by $x_a(t)$ the traffic observed at the area a during the time interval t.

$$x_a(t) = b_0 + b_1 x_a(t-1) + b_2 x_a(t-2) + \ldots + b_p x_a(t-p) + \epsilon_a(t)$$

The equation above is such as a linear regression model, the parameter of this model can be estimated using a method such as least square (which the goal is to minimize the error terms $SSE = \sum \epsilon_t^2$) or Maximum Likelihood. Important to note that the error $\epsilon_a(t)$ follows the normal distribution with parameter $(0, \sigma^2)$

The AR model however takes an order, p, which will dictate how many prior time steps to use in the regression.

- MA Model: an MA model is a linear regression of the current value of the series against previously observed white noise error terms. Similarly to AR models, MA models also take an order term, q, which will dictate how many prior errors will be considered. The equation of this model can be expressed as:

$$x_a(t) = c_0 + c_1\epsilon_a(t-1) + c_2\epsilon_a(t-2) + \ldots + c_q\epsilon_a(t-q) + \epsilon_a(t)$$

the the previous model explanation is also a linear regression problem and the parameter of this model can be estimated using methods like least square and maximum Likelihood. q is the number of past error terms.

- I (integration or differencing): accounting for overall "trend" in the data. The main purpose of the act of differentiation is to make a time series stationary. There are many methods to make time-series stationary we will discuss them in the next session. This part of the model accounts for general trends that occur throughout the time series data. d is number of differencing

The mathematical expression of ARIMA model can be summary in the following equation:

$$x_a(t) = b + \sum_{k=1}^{p} b_k x_a(t-k) + \sum_{i=1}^{q} c_i x \epsilon_a(t-q) + \epsilon_a(t)$$

where b, b_k and b_i, p, p and q are the parameters of our model that the regression parameter can be estimated using the least square, maximum Likelihood, and the order p, q, and d can be determined using a statistical method with hyperparameter tuning.

2.1.2 Variation of ARIMA model

There are many variations of ARIMA, most the popular variation is SARIMA(Seasonal ARIMA) and SARIMAX(Seasonal ARIMA with exogenous variation) etc...

- SARIMA: This model is very similar to the ARIMA model, except that there is an additional set of autoregressive and moving average components.The additional lags are offset by the frequency of seasonality (ex. 12 — monthly, 24 — hourly)

  The equation of SARIMA can be expressed as:

$$x_a(t) = b + \sum_{k=1}^{p} b_k x_a(t-k) + \sum_{i=1}^{q} c_i x \epsilon_a(t-q) + \sum_{k=1}^{P} \phi_k x_a(t-sn) + \sum_{i=1}^{Q} \tau_i x \epsilon_a(t-sn) + \epsilon_a(t)$$

  This models allow for differencing data by seasonal frequency, yet also by non-seasonal differencing.

- SARIMAX: this model takes into account exogenous variables, or in other words, uses external data in our forecast. Some real-world examples of exogenous variables include gold price, oil price, outdoor temperature, and exchange rate.

  It is interesting to think that all exogenous factors are still technically indirectly modeled in the historical model forecast. That being said, if we include external data, the model will respond much quicker to its effect than if we rely on the influence of lagging terms.

$$x_a(t) = b + \sum_{k=1}^{p} b_k x_a(t-k) + \sum_{i=1}^{q} c_i x \epsilon_a(t-q) + \sum_{j=1}^{r} \beta_j x_j(t) + \sum_{k=1}^{P} \phi_k x_a(t-sn) + \sum_{i=1}^{Q} \tau_i x \epsilon_a(t-sn) + \epsilon_a(t)$$

2.2 Stationary

Before delving into the model training, we need to check all assumptions of the model. Most popular time series models work with the assumption of stationary over time. Stationary means constant statistical properties over time, such as means and variance, and autocorrelation (correlation with the lagged value). Two main components can make time series non-stationary are trends and seasonality. Trends are about the direction of the time series over time. If the time series changes direction many times (increase or decrease), that means over time series has a trend. Seasonality is when the data have almost the same pattern over a repeated time range, such as hourly, daily, etc. There are many methods to test if our time series is stationary; one of the popular methods is the Dickey-Fuller Test.

2.2.1 Dickey-Fuller Test

The main principle behind this statistical test is to test if our time series is stationary. The null hypothesis is that our time series is not stationary, and the alternative hypothesis is our time series is stationary. The results comprise a statistic test and critical values for different confidence levels. If the statistic test is less than the critical value, we can reject the null hypothesis and say that the series is stationary.

2.2.2 Test stationary for area 5161

Figure 6 shows the mean-variance and internet traffic variance for area 5161 over time. In this figure, we can see small variances of mean and variance over time.

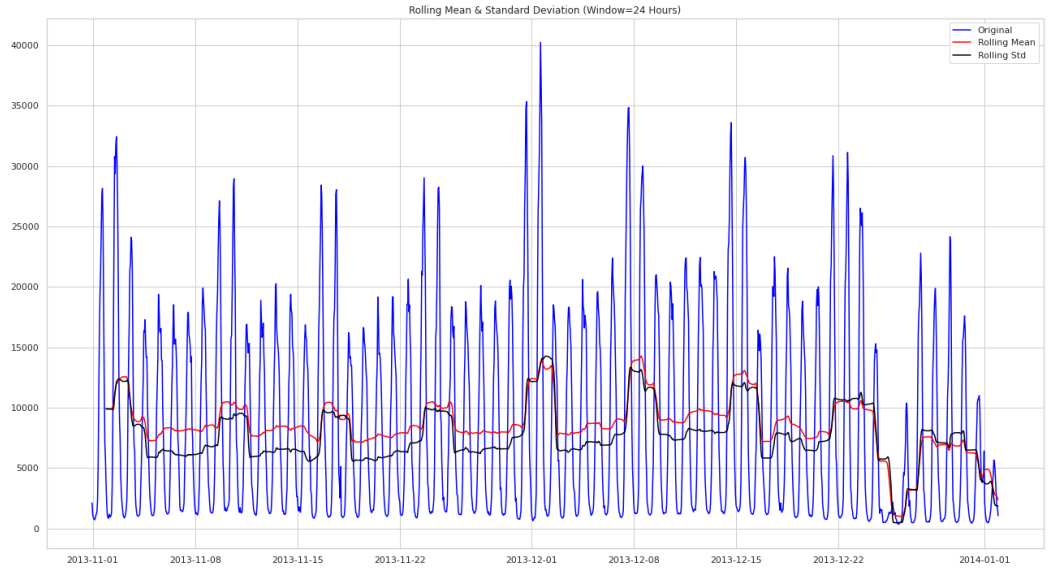

Figure 6: Rolling mean and Rolling standard deviation of time series in area 5161 over the time

The following table gives the statistical summary of the Dickey-Test:

Table 4: Summary test statistic dickey-Test for area 5161.

Test Statistic	-3.900012
p-value	0.002035
Lags Used	24.000000
Number of Observations Used	1463.000000
Critical Value (1%)	-3.434828
Critical Value (5%)	-2.863518
Critical Value (10%)	-2.567823

In table 4, we can see that the $p - value = 0.002035 < 0.05$, which means we reject our null hypothesis: the time series is not stationary. Hence, the time series of area 5161 is stationary. We can use the ARIMA model because all assumptions are met for the area with the highest internet traffic activity

2.2.3 Test stationary for area 4159

Figure 7 shows the mean and internet traffic variance for area 4159 over time. In this figure, we can see that there are small variances of mean and variance over time. The Rolling Mean and Rolling std are almost constant over time.

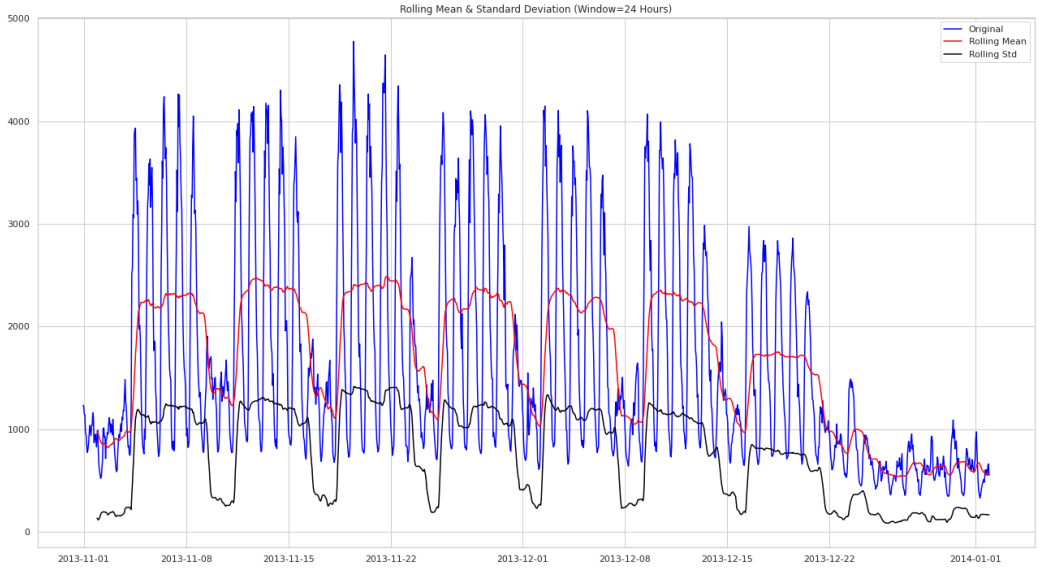

Figure 7: Rolling mean and Rolling standard deviation of time series in area 4159 over the time

The following table gives us the summary statistics for Dickey-Test:

Table 5: summary test statistic dickey-Test for area 4159

Test Statistic	-3.670918
p-value	0.004541
Lags Used	24.000000
Number of Observations Used	1463.000000
Critical Value (1%)	-3.434828
Critical Value (5%)	-2.863518
Critical Value (10%)	-2.567823

In table 5, we can see that the $p - value = 0.004541 < 0.05$, which means we reject our null hypothesis: the time series is stationary. Hence, the time series of area 5161 is stationary.

2.2.4 Test stationary for area 4556

The plot below shows that the Rolling Mean and Rolling std are almost constant over time, and the p-value 0.000012 is less than 0.05, which means we reject the hypothesis H_0 (which is our time series is non-stationary). Hence, our with series is stationary. We can use ARIMA model because all assumptions are met for the area with square_id: 4556.

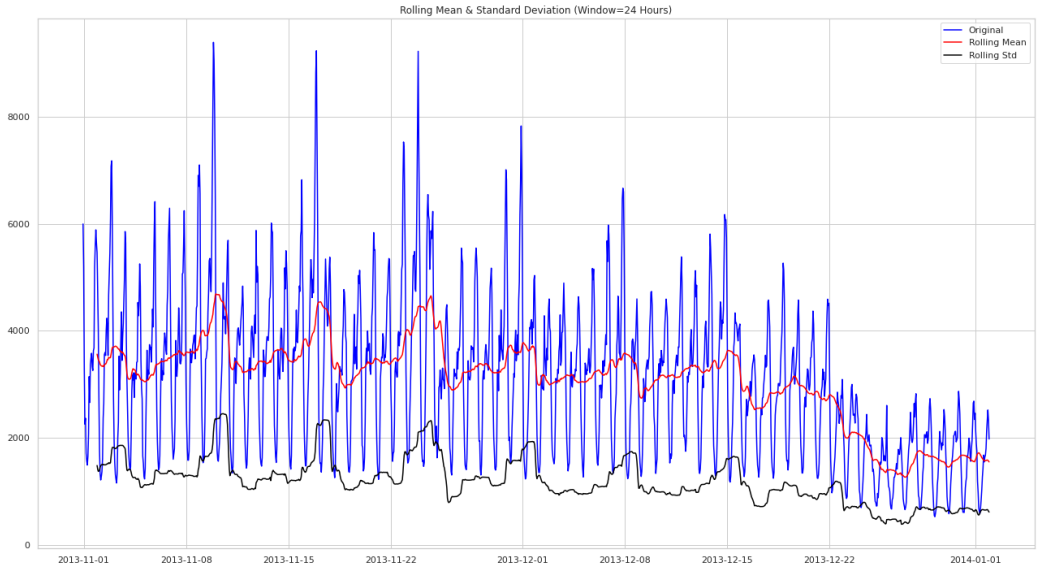

Figure 8: Rolling mean and Rolling standard deviation of time series in area 4556 over the time

The following table gives us the summary statistics for Dickey-Test:

Table 6: summary test statistic dickey-Test for area 4159

Test Statistic	-1.775915
p-value	0.392483
Lags Used	24.000000
Number of Observations Used	1463.000000
Critical Value (1%)	-3.434828
Critical Value (5%)	-2.863518
Critical Value (10%)	-2.567823

In table 6, we can see that the $p - value = 0.392483 > 0.05$ means we don't reject our null hypothesis, which is that the time series is not stationary. Hence, the time series of area 5161 is not stationary. For this case, we need to make this time series stationary. There are many methods to make time series stationary.

2.2.5 Seasonality and trend for time series

This section focuses on studying seasonality and trends for each of the three areas. The seasonality helps us determine the repeating short-term cycle in the series, and the three factors help determine if the time series value increases or decreases over time.

Figure 9 shows that the seasonality and trend using the decomposition method of area 5161:

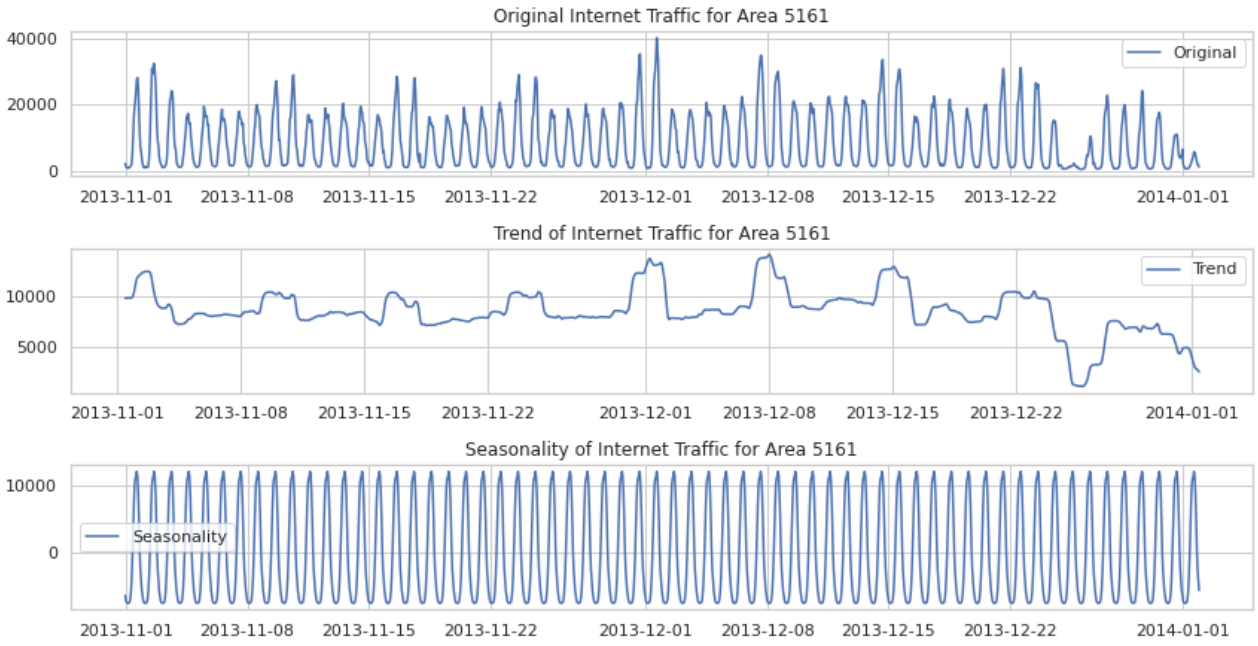

Figure 9: Seasonality and trend for area 5161.

For this area, we can see that the time series has seasonality because the plot of our seasonality has almost the same pattern as the sinusoidal pattern.

The following figure shows us the seasonality and trend using the decomposition method of area 4556:

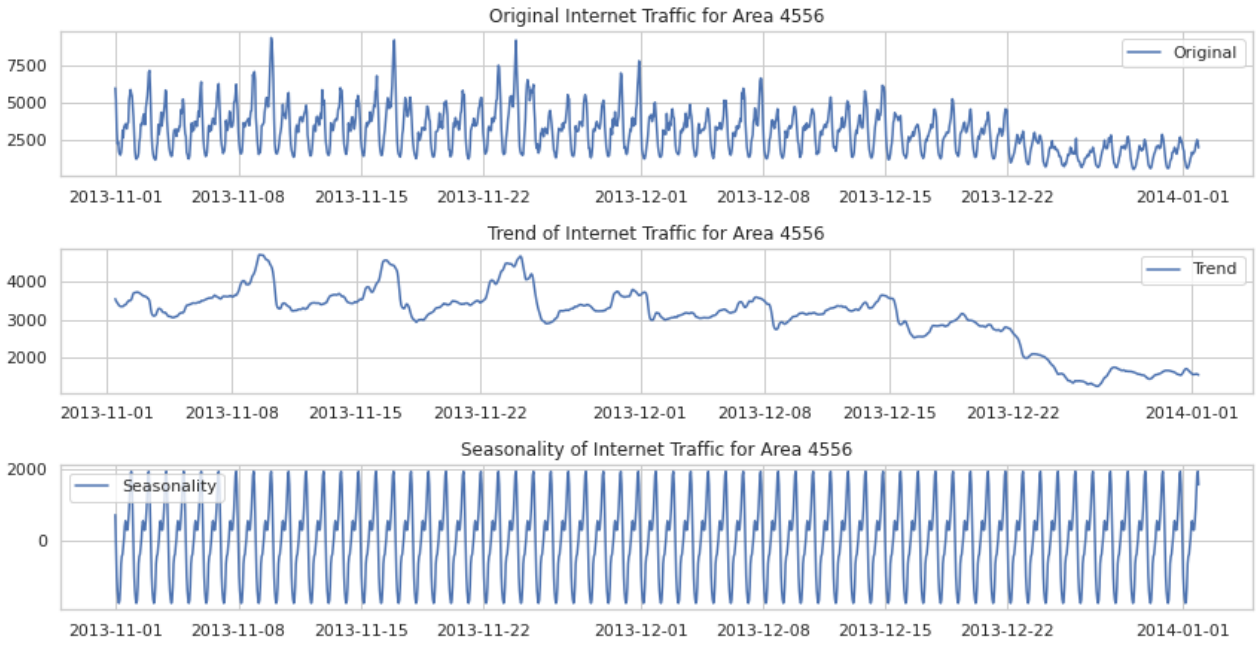

Figure 10: Seasonality and trend for area 4556.

For this area, we can see that the time series has seasonality because the plot of our seasonality has almost the same pattern and sinusoidal pattern. Still, at the end of the year, the trend decreases rapidly.

The following figure shows us the seasonality and trend using the decomposition method of area 4159:

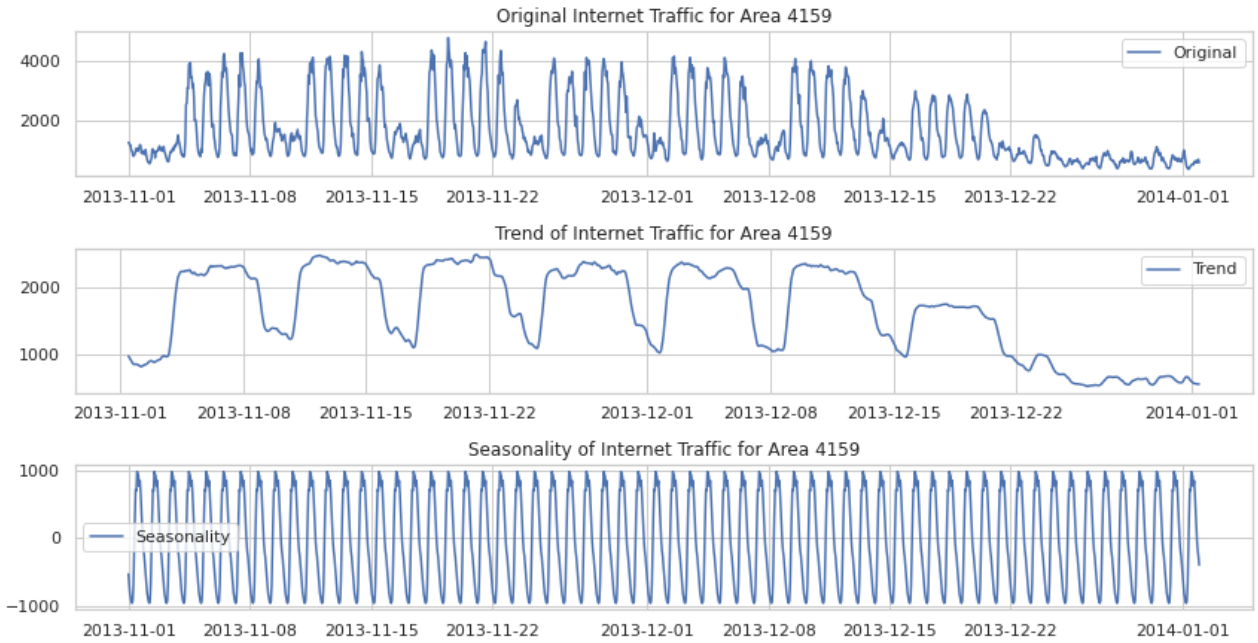

Figure 11: Seasonality and trend for area 4159.

For this area, we can see that the time series has seasonality because the plot of our seasonality has almost the same pattern and sinusoidal pattern, but at the end of the year, the trend decreases slowly.

2.3 ACF and PACF method

Once we have checked the stationary each of the time series in the area of interest, the next step is the find the range value of p, p, and d using the method Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF):

- Autocorrelation Function (ACF): computes the autocorrelation for a given time series. Autocorrelation is the correlation between observations of a time series separated by k time steps.
- Partial Autocorrelation Function (PACF): This measures the correlation between the time series with a lagged version of itself but after eliminating the variations already explained by the intervening comparisons. Compute the strength of the relationship while also accounting for any intermediate lags.

Both ACF and PACF produce plots for up to any arbitrary amount of lags, which easily visualizes the autocorrelational strength that each lag has on a given observation.

2.4 Hyperparamater tuning

Another method to find a suitable model is hyperparameter tuning, which helps us find a good method for each area. The main idea behind hyperparameter turning for time series is to find the best model that minimizes AIC. Akaike's Information Criterion (AIC), which was useful in selecting predictors for regression, is also useful for determining the order of an ARIMA model. It can be written as

$$AIC = -2\log(L) + 2(p + q + k + 1)$$

where L is the likelihood of the data, $k = 1$ if $c \neq 0$ and $k = 0$ if $c = 0$. Note that, the last term in parentheses is the number of parameters in the model (including σ^2, the variance of the residuals).

The Bayesian Information Criterion can be written as:

$$BIC = AIC + [log(T) - 2](p + q + k + 1)$$

Good models are obtained by minimizing the AIC, AICc, or BIC. Our preference is to use the AICc. To find the best model for each of the three areas, we used the autoarima package.

By using this method, we found that all three areas have the same parameter.

2.5 Model selection and prediction

We have seen above that our time series has many proprieties like stationary, seasonality, and trend. We have checked with appropriate statistical methods. We should use the model that captures all these properties. While ARIMA and SARIMA are good models, SARIMAX is better because it adapts to this time series. SARIMAX is an extension of SARIMA, as we have explained above, but it added additional variables (the 'exogenous' part). SARIMAX stands for Seasonal AutoRegressive Integrated Moving Average with eXogenous regressors.

The statistics used to measure the goodness of our model are measured with mean absolute error (MAE) or mean absolute percentage error (MAPE). For each of the models below, we summarize these metrics in the table below. This metric is calculated with the expression:

$$MAE = \frac{1}{n}\sum_{i=0}^{n} |\hat{y}_i - y_i|$$

$$MAPE = \frac{1}{n}\sum_{t=0}^{n} |\frac{\hat{y}_i - y_i}{\hat{y}_i}|$$

Using hyperparameter tuning, we found the same value for the tree area. This parameter is split into two parts order and seasonal order:

- order(1,0,1): In this case, $p = 1$, $d = 0$ and $q = 1$
- seasonal order (1,0,2,24) In this case $P = 1$, $D = 0$, $Q = 2$ and $s = 24$.

The following figure shows us the predicted and the actual value for area 5161:

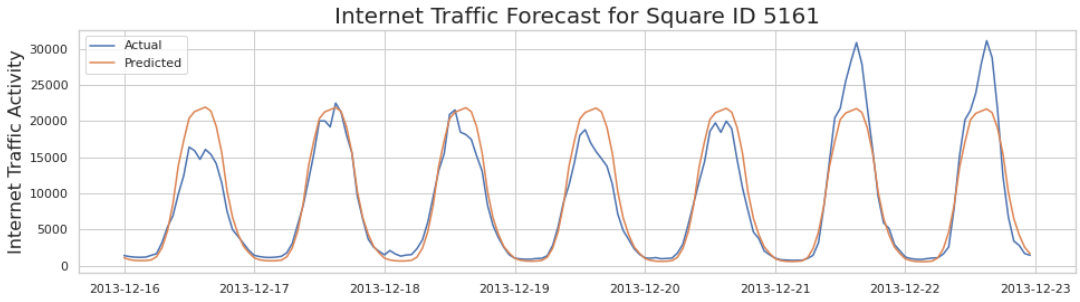

Figure 12: Predicted and actual value for the area 5161

We can see that the predicted value is almost the same as the actual value except on the last day.

The following table gives a statistical summary of the model.

Table 7: Summary accuracy for the area 5161.

MAE	MAPE	ACI
1683.7131649751661	0.23923806809059073%	18802.049568567316

The following figure shows the predicted and the actual value for area 4159:

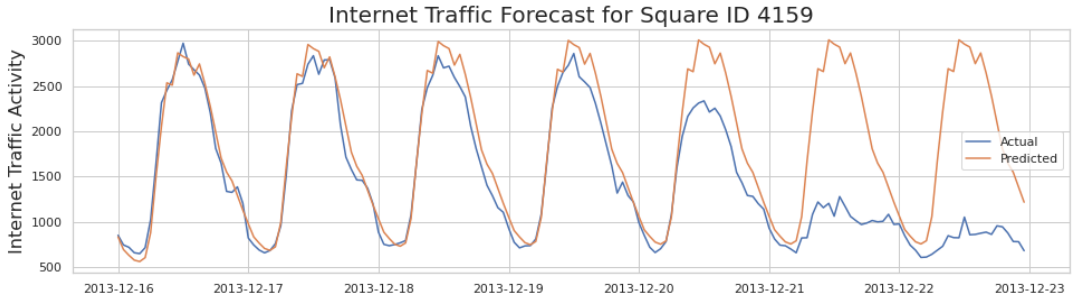

Figure 13: Predicted and actual value for the area 4159.

The predicted value is not good after 19 December 2013. The following table gives a statistical summary of the model.

Table 8: Summary accuracy for the area 4159.

MAE	MAPE	ACI
397.370893441464	0.37304417433436626%	14817.98615582478

We can see that the MAPE is higher than in the previous area. The reason is that after 12 December, the model didn't predict a good value.

The following figure shows us the predicted and the actual value for area 4556:

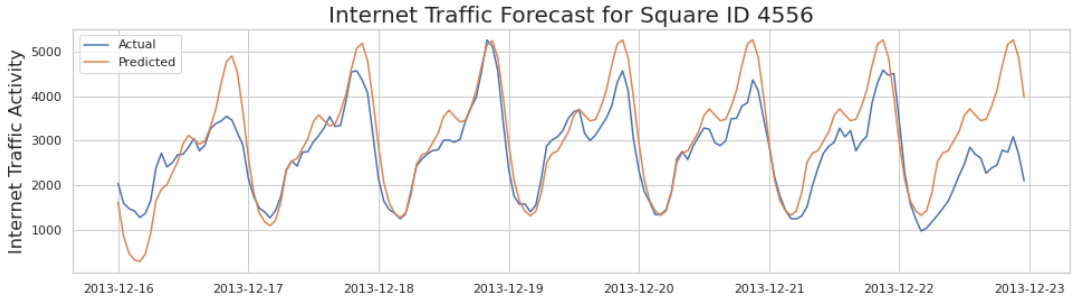

Figure 14: Predicted and actual value for the area 4159.

we observed that the predicted value is not good after 19 December 2013. The following table gives us the statistical summary for the model.

Table 9: Summary accuracy for the area 4556.

MAE	MAPE	ACI
502.3057079524565	0.20404037225955823%	16232.97169068984

The accuracy for this area using our model is good because our model has almost 80% of good predictions.

2.6 Recommendation

At the end of this study, we have seen that it is important to understand our dataset before fitting the dataset to a model. This understanding is carried out throughout the assessment of the statistical assumptions (e.g., hypothesis) and statistical properties over time. While fitting the dataset, we noticed that features with big numbers can increase the processing time. As part of the recommendations, we can highlight that standardizing operations on those numerical values (e.g., our main feature, internet traffic activity) can help to reduce the processing time. Another suggestion to improve the fitting algorithm is to improve the condition number of the covariance matrix (condition number: 5.96e+27).

References

[1] *Dask — Dask documentation*. URL: https://docs.dask.org/en/stable/.

[2] Telecom Italia. *Milano Grid*. Version V1. 2015. DOI: 10.7910/DVN/QJWLFU. URL: https://doi.org/10.7910/DVN/QJWLFU.

[3] Telecom Italia. *Telecommunications - SMS, Call, Internet - MI*. Version V1. 2015. DOI: 10.7910/DVN/EGZHFV. URL: https://doi.org/10.7910/DVN/EGZHFV.