INDIAN INSTITUTE OF INFORMATION TECHNOLOGY
ALLAHABAD

7^{TH} SEMESTER PROJECT

Web Traffic Data Exploration and Time Series Analysis

Submitted To:
Dr. Pavan Chakraborty
IIIT Allahabad

Submitted By:
Srikar Anand (IIT2017504)
Akshay Gupta (IIT2017505)

CERTIFICATE FROM SUPERVISOR

We hereby declare that the work presented in this mid semester project report of B.Tech (IT) 7^{th} Semester entitled "Web Traffic Data Exploration and Time Series Analysis" , submitted by us at Indian Institute of Information Technology, Allahabad, is an authenticated record of our original work carried out from August 2020 to December 2020 under the guidance of Dr. Pavan Chakraborty.

Due acknowledgements have been made in the text to all other material used. The project was done in full compliance with the requirements and constraints of the prescribed curriculum.

Date : 03-12-2020
Place : IIIT Allahabad

Supervisor:
Dr Pavan Chakraborty

Contents

1 Abstract

This report titled "Web Traffic Data Exploration and Time Series Analysis " describes the 7^{th} semester project in which web traffic prediction is done through careful data analysis involving various data exploration methods and then time series analysis is done to predict the future of web traffic for the given Wikipedia articles. In the recent times web traffic analysis has been aggressively being used by web site owners to predict the traffic on web pages and correspondingly scale up or down the infrastructure depending upon the traffic in order to provide efficient services as well as reduce cost maintenance.
***Keywords*:** Web Traffic, Time Series, Data Analysis,Wikipedia Articles, ARIMA , LSTM ,CNN .

2 Introduction

Time series analysis is of utmost importance in the modern world where every industry wants to enhance its revenue by predicting the future of their product and adjusting their production , consumption and maintain their market . Time series analysis of data not only helps in forecasting but also helps in understanding the past of subject in consideration statistically . Sequential or temporal observations emerge in many key real-world problems, ranging from biological data, financial markets, weather forecasting, network congestion to audio and video processing [**1**]. The field of time series encapsulates many different problems, ranging from analysis and inference to classification and forecast.
Web traffic analysis and forecast is an active area of research and with different methods involved to regularly monitor the user behavior , business pattern , detect anomaly , sell advertisements based on user pattern , scale infrastructure and make business decisions based on that [**2**]. From various machine and deep learning models to genetic algorithm to different sophisticated mechanism have already been developed to explore this area .

3 Motivation

Forecasting and determining web traffic enables to plan efficient resource allocation which in turn results in effective analysis of upcoming growth and revenues for web site owner which includes :- 1.determining an effective strategy for load balancing of web pages residing in the cloud , 2. forecasting future trends based on historical data , 3. detect anomaly and security concerns, 4. sell advertising products or make business decisions.

4 Problem Definition

This project is aimed to provide efficient system where user can see the the short/long term trend of any wikipedia article. For this purpose various data methods of data analysis is to be done with Big Data at disposal . To study and predict the result , time series analysis is to be done on those datasets. Time Series analysis can be done through different time series models . In this project , we used Arima , LSTM and CNN models to predict and forecast the timeseries of wikiepedia articles.

5 Literature Review

5.1 Forecasting Web Page Views : Methods and Observations

This paper titled as "Forecasting Web Page Views : Methods and Observations" was published in Journal of Machine Learning Research ,2008 by Jia Li, Visiting Scientist , Google Labs and Andrew W. Moore , Engineering Director , Google Labs.[3]
In this research paper , main focus is given to extract trends and season pattern which have great impact on web page view prediction. Time Series is decomposed into trend , season and noise and can be represented as:
$X_t = L_t + I_t + N_t$
For the short-term prediction , HW procedure and SSM procedure has been used , since in web pages there exist seasonality at multiple time scales . They used different methods to find remedy of sudden massive noise which could have detrimental impact on forecasting . For the long-term prediction , ESSF algorithm is used to extract the global trend and scale the long term season effect after removal of short term seasonal impact from HW .

- **Holt-Winters**
  This is a probabilistic approximation method in which the current state is an exponentially weighted running average of recent season-adjusted observations. The forecasting of h time units ahead is given by this following equation:
  $\hat{x}_{t+h} = L_t + hT_t + I_{t-d+h mod d}$,where $x_1, x_2, ...x_t$ is given time series and $\hat{x_t}$ is predicted at time t.

- **State Space Model**
  The underlying principle is guided by Markov Model where states are characterized by Markov process and is a linear combination of states added with Gaussian noise.
  When web page view showed long range seasonality and trends both the above model were far from predicting the outcome . In such case trend was further

decomposed to yearly season y_t, global linear trend u_t and volatility part n_t such that the new equation would be:
$x_t = u_t + y_t + n_t + I_t + N_t;$ where $L_t = u_t + y_t + n_t$.

- **Elastic Smooth Season Filtering**
  ESSF is used for long term prediction of web page views keeping the track of long term trend , and seasonality pattern as in the modified equation . Smoothness regularization and scalable yearly season improved prediction accuracy in comparison to earlier used models.

5.2 Web Traffic Prediction of Wikipedia pages

This paper titled as "Web Traffic Prediction of Wikipedia Pages" was published in IEEE conference on Big Data in 2018 by Navyasree Petluri and Eyhab Al-Masri , School of Engineering and Technology, University of Washington.[2]
This research paper used RNN seq2seq model to predict the web page views.The prediction model was built on : (1) number of hits , (2) features that were extracted from URLs , (3) day of week - analyses the weekly seasonality information ,(4) year to year auto correlation, (5) page popularity and (6) lagged page views . These were the features used by RNN seq2seq model based on Encoder/Decoder Architecture to predict the outcome. Rolling median was used as measure of median to avoid sensitivity based on sudden spike in the data. Feature window is used instead of entire time series length to capture the trend which is later used to derive and forecast weekly , monthly or yearly trend.

5.3 WaveNet: A Generative Model For Raw Audio

This paper titile as "WaveNet : A Generative Model For Raw Audio" was pubslished which introduces a new deep neural network for generating audio waveforms [4]. This research paper used causal convolutions to make sure that the model cannot violate the ordering in which the data was modelled. For any given time-series , the prediction p($x_{t+1}|x_t, x_{t-1}, x_{t-2}...x_1$) emitted by time-step t cannot depend on any future time-steps e.g. x_{t+1}, x_{t+2} etc , Only the past can influence the time-series prediction.
While training , the conditional predictions can be made in parallel because of all the time-steps of ground truth with the given time-series is known. Generating with the model , all the predicted values at any time is fed back into the model to predict the next sample. In this paper , dilated convolutional network was used to further increase the receptive field by the order of magnitude , this way computational cost was also minimised. A dilated convolutional network is one where the filter is applied over an larger area greater than its length by skipping the input values by

certain step . Since , this model does not have recurrent connections , it is typically faster to train than that of RNN for long sequences. One of the problem associated with causal convolutional network is that they require many layers or large filters to increase the receptive field. WaveNets produces samples that outperformed the then TTS systems in subjective naturalness. It also showed promising results when applied to music , audio modelling and speech recognition.

6 Data Analysis and Experimentation

The dataset used here has been collected from "Web Traffic Time Series Forecasting", released by Google **[1]**. The dataset contains web traffic data of several wikipedia pages from 1st July, 2015 to 31st December, 2016. Each entry in the dataset contains the url link to the wikipedia page, and number of views for each day. The pages belong to 7 different languages (English, Japanese, German, French, Chinese, Russian, and Spanish) and the media pages. This variation in the languages may cause problems to analyze the data, hence the dataset has been divided into individual dataframe for each language and studied separately.

6.1 Autocorrelation and Partial-autocorrelation

In statistics, correlation is a statistical relationship, whether causal or not, between two variables or data. In other words, correlation shows us the strength of relationship between two variables. In time series analysis, the current value of the series is actually related to its value at a certain time in the past. ACF and PACF (Autocorrelation and Partial-autocorrelation) help to identify this in a time series. ACF and PACF plots are frequently used in time-series analysis and forecasting. These graphs summarize the strength of the time-series with the previous time steps. The correlation between two series x and y can be determined by using the formula,

$$\rho_{x,y} = \frac{\Sigma(x - x_m)(y - y_m)}{\sqrt{\Sigma(x - x_m)^2 \Sigma(y - y_m)^2}}$$

In Autocorrelation, the correlation has to be calculated between the series and itself a few time steps back. Hence, it is given by,

$$\gamma_{x,h} = \frac{\Sigma(x_t - x_m)(x_{t-h} - x_m)}{\sqrt{\Sigma(x_t - x_m)^2 \Sigma(x_{t-h} - x_m)^2}}$$

where x, y are series and h is the lag.[4]

In Autocorrelation, the relationship between the series and the lagged time series is also affected by the observations of time steps in-between. But, in Partial-Autocorrelation, the effect of such intervening observations is removed. The partial autocorrelation at lag k is the correlation that results after removing the effect of any correlations due to the terms at shorter lags. With this intuition, the formula for the PACF can be given as

$\Phi_{x,h} = \rho(x_t - P(x_t|x_{t+1}, .., x_{t+h-1}), x_{t+h} - P(x_t|x_{t+1}, .., x_{t+h-1}))$
$P(w, z) = \Sigma_{w,z}\Sigma_{z,z}^{-1}z$
$\Sigma_{w,z}$ is the Covariance Matrix between w and z

ACF and PACF give the different plots when applied to various languages, of which English and Spanish are shown below

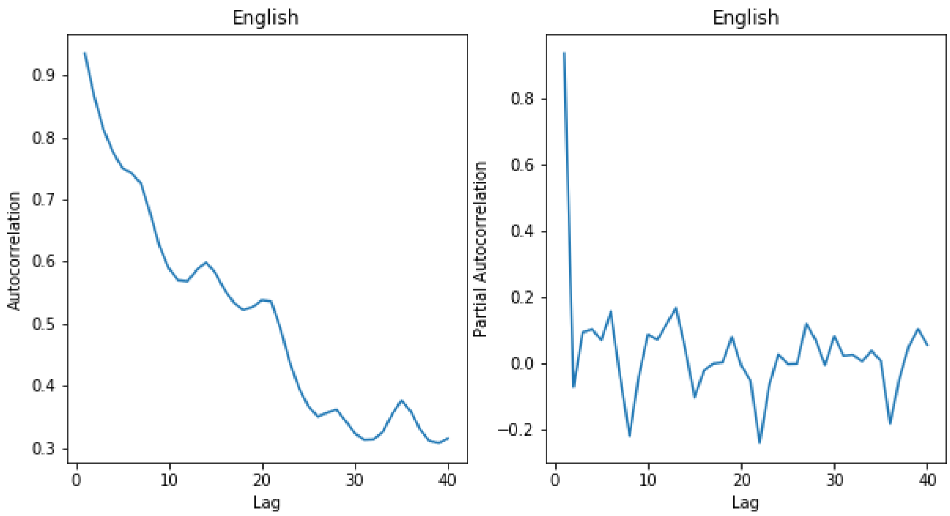

Figure 6.1.1 : ACF and PACF applied on English data

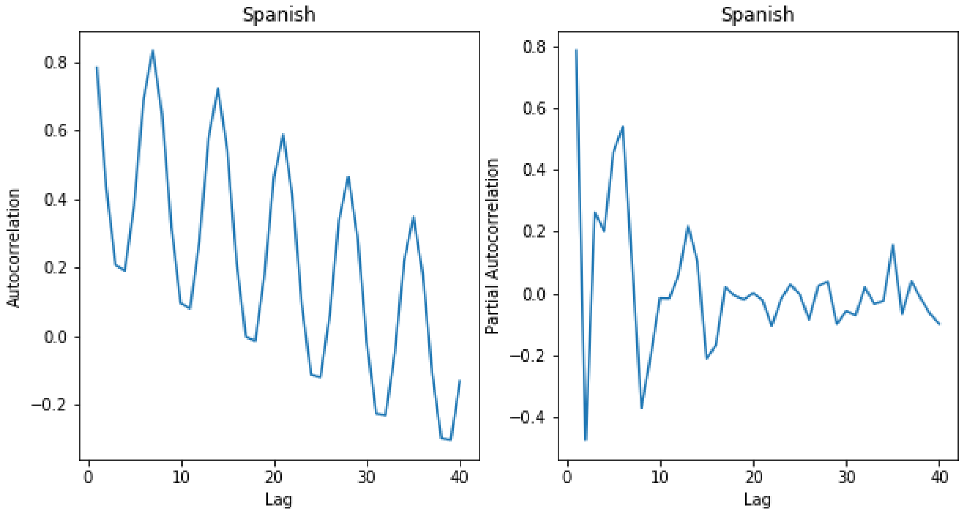

Figure 6.1.2 : ACF and PACF applied on Spanish data

ACF and PACF plots can be used to estimate the hyperparameters that will be used to train ARIMA model for time series forecasting. The peaks where the ACF and PACF plots oscillate with respect to the lag can be considered near optimal parameters for ARIMA model. Based on the above plots, it can be concluded that there won't be any lag required for the English data. For Spanish, there is a trend of 7 days per peak. Hence a parameter of 7 can be used for Spanish data. Similarly, it can be estimated for all other languages as well.

6.2 Fast Fourier Transform

Fourier Transform shows various frequencies' amplitudes present in the data. Especially when the data looks periodic, the strongest frequencies will be at peaks, which identifies the trend in the data. Since the Fourier Transform gives complex values for amplitudes, so the magnitude of the amplitude would be plotted. Below are the plots for English and Spanish data, when applied Fast Fourier Transform.

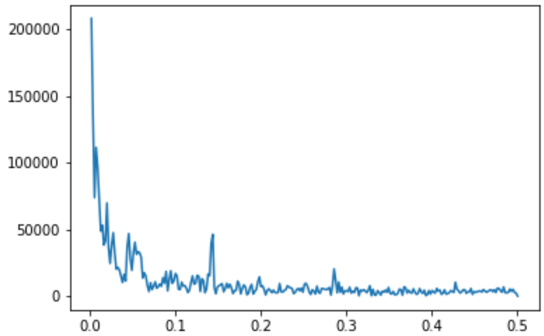

Figure 6.2.1 : FFT applied on English data

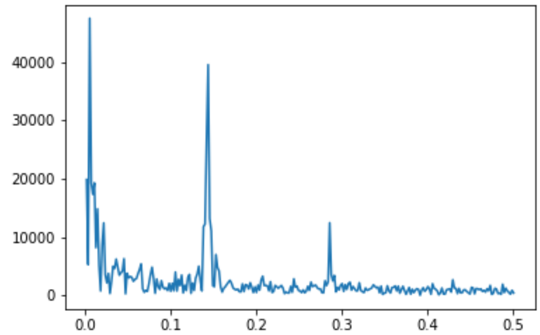

Figure 6.2.2 : FFT applied on Spanish data

From the plots, it can be seen that for Spanish, there is a huge peak at around 0.15 and a small peak at 0.29, which means a time period of 7 days and 3.5 days respectively. For English the peaks are present similar to Spanish, only smaller in the relative amplitude. Since Spanish is more periodic and regular, its relative peak amplitude is higher than that of English. So it can be deduced that the trends mostly are weekly and once per 3 days. This is as expected because the browsing habits might differ weekdays to weekends which results in the peaks. So, it is known now that page views are not at all smooth. There is some regular variation from day to day, but there are also large effects that can happen quite suddenly.

7 Proposed Methodology

7.1 Data Analysis

As seen in the previous section, various types of functions and filters have been applied on the data to study it. And the following conclusions have been achieved:

- Though the data isn't fully stable, its periodic and predictable mostly.
- Applying ACF and PACF on the data showed that data is stable enough across different languages, and it follows a trend. The hyperparameters for different languages are figured out from the graphs.
- Fourier Transform of the data showed the strongest frequencies present in the data. This gave a clue on the stability of the data as well. Multiple peaks imply less stability. Most of the data is not too smooth, so appropriate model must be used for prediction.
- Decomposing the data proved the nature of the data is approximately linear in terms of seasonality, but non-linear in the trend. So the models like ARIMA and LSTM can be applied

7.2 Modelling

Using various Data exploration and analysis techniques , parameters for ARIMA was decided , also using the keras framework on top of tensorflow 2 deep learning models were constructed using LSTM and CNN to make predictions.

7.2.1 Auto-Regressive Integrated Moving Average (ARIMA)

The Auto-Regressive moving average model (ARMA) model is useful for studying the uni-variate time series. ARMA model as the name suggests is the combination of Auto-Regression (AR) and Moving-Averages(MA). The mathematical equation for regression consists of the present output as the linear function of the past output with the error term included [**7**]. But in practice the data shows the non stationary behaviour and so we need to use the concept of generalised ARMA model ,also referred to as ARIMA model (Auto-Regressive Integrated Moving Averages),this covers the case of non stationary data as well. ARIMA is a triple (p,d,q) meaning that the values p,d,and q needs to be defined. These values can be taken from the results of analysis section, where the repetitive patterns of the data has been shown. ARIMA model would be a good fit with our data, having said that the parameters for the model have already been determined.
The mathematical equation for regression consists of the present output as the linear function of the past output with the error term included. Mathematically the AR(p) model can be expressed as :

$$y_t = c + \sum_{i=1}^{p} \phi_i y_{t-i} + \epsilon_t$$

where ϕ_i are the estimates and ϵ_i is the error value corresponding to the time 't'. The estimates are found using Regression Algorithms either using Least Squares

or Gradient Descent ,the value of error is a probabilistic function (can be assumed Gaussian).The term 'c' is also included in the estimate as this parameter also needs to be calculated.
The moving average MA(q) assumes the past error values as explanatory variables.so ,instead of dependence upon the previous instances of variables , these are dependent upon the error values,Hence new prediction MU(q) will be given as:

$$y_t = \mu + \sum_{i=1}^{q} \theta_i \epsilon_{t-i} + \epsilon_t$$

Here θ are the model parameters, and μ is the mean values. So combining these will give the ARMA model. The problem with this model is that it can be used only in the case of univariate ,time-stationary models.But in practice the data shows the non stationary behaviour and so we need to use the concept of generalised ARMA model ,also referred to as ARIMA model (Auto-Regressive Integrated Moving Averages),this covers the case of non stationary data as well. ARIMA is a triple (p,d,q) meaning that the values p,d,and q needs to be defined. The mathematical formulation of the Arima(p,d,q) is given as:

$$(1 - \sum_{i=1}^{p} \phi_i L^i)(1 - L)^d y_t = (1 + \sum_{j=1}^{q} \theta_j L^j)\epsilon_t$$

Here the parameters in the triple p,d, and q are non negative integers, Generally the value of d is considered to be 1 as this covers most of the cases, d=0 gives back our natural ARMA model. In our experiment we considered the value of d as 1.

7.2.2 Long Short Term Memory (LSTM)

Long Short Term Memory (LSTM) are the special kind of Recurrent Neural Networks (RNN). LSTM have the capability to learn the long term dependencies. As such, the sequence of observations must be transformed into multiple examples from which the LSTM can learn [**6**]. From the results of data analysis, we can say that LSTM model can be used directly to predict the data, without any requirement of data-transformation. LSTM would have good efficiency, especially in this case where data is highly periodic and depends on the past data most of the times.
The common architecture of LSTM unit consists of the cell and the three gates which are input gate, output gate, and forget gate. The job of cell is to remember the values over the arbitrary intervals of the time while the job of the gates is to regulate and control the flow of data and information from in and out of the cell.The cell keeps the track of the dependencies between the elements of the imput sequence. The input gate's work is to regulate the flow of new values in the cell.Forget gate controls that for how much unit of time value should be present in the cell. Output's gate job is to check that for how much extent the value present in the cell will contribute to the computation of the output activation of the LSTM unit.The activation functions used for the gates of LSTM are usually sigmoid activation function

and the hyperbolic tangent function.

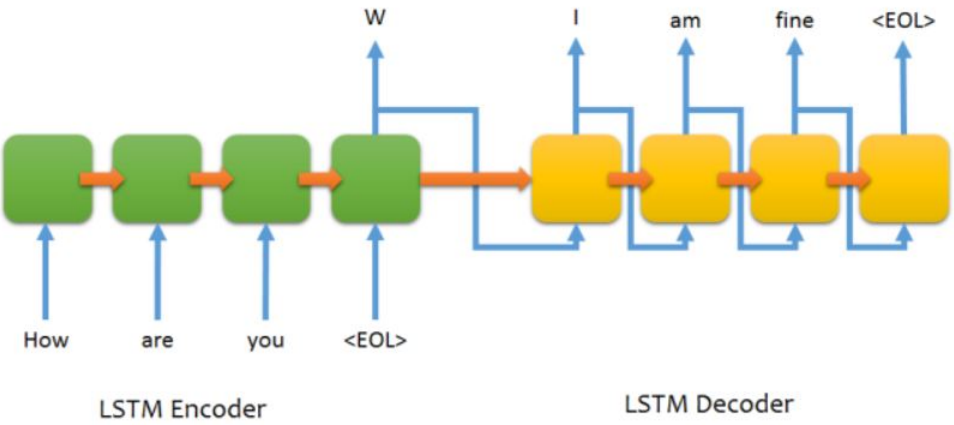

Figure 7.2.2.1 : LSTM
Source: Adapted from [09]

In time-series analysis , the LSTM model can be thought as a encoder-decoder framework , where an arbitrary long input sequence to an arbitrary long output sequence is fed and encoded in an intermediate state. This encoded state can be treated as the representation of entire history sequence of input sequence that provides context to the decoder to produce an output sequence.

```
Model: "functional_1"
__________________________________________________________________________________________________
Layer (type)                    Output Shape         Param #     Connected to
==================================================================================================
input_3 (InputLayer)            [(None, None, 1)]    0
__________________________________________________________________________________________________
input_4 (InputLayer)            [(None, None, 1)]    0
__________________________________________________________________________________________________
lstm_2 (LSTM)                   [(None, 50), (None,  10400       input_3[0][0]
__________________________________________________________________________________________________
lstm_3 (LSTM)                   [(None, None, 50), ( 10400       input_4[0][0]
                                                                 lstm_2[0][1]
                                                                 lstm_2[0][2]
__________________________________________________________________________________________________
dense_1 (Dense)                 (None, None, 1)      51          lstm_3[0][0]
==================================================================================================
Total params: 20,851
Trainable params: 20,851
Non-trainable params: 0
__________________________________________________________________________________________________
```

Figure 7.2.2.2 : LSTM summary of the used model

7.2.3 Convolutional Neural Network (CNN)

Standard tasks for CNNs are generally image recognition or text classification etc. These kind of neural networks are very powerful to learn spatial invariant patterns. In the time-series analysis , In this project , CNN inspired by $WaveNet$[4] architecture were modified to learn temporal pattern. With the help of dilated causal convolution layer which helps in maintaining the temporal order and long-term dependencies without any explosion in complexity of the architecture , it is made possible. In a traditional 1 dimensional CNN, a mask or filter is applied over the input to get the convolution of input which is then fed as input to the hidden layers. But in case of temporal pattern , extra care has to be taken such that future values don't influence the past. In other words, the temporal order must be maintained which is done by maintaining the causal structure . Causal structure provide the needed temporal flow but long-term dependencies should also be handled , these are done through dilation of convolution layer as shown in graph below.

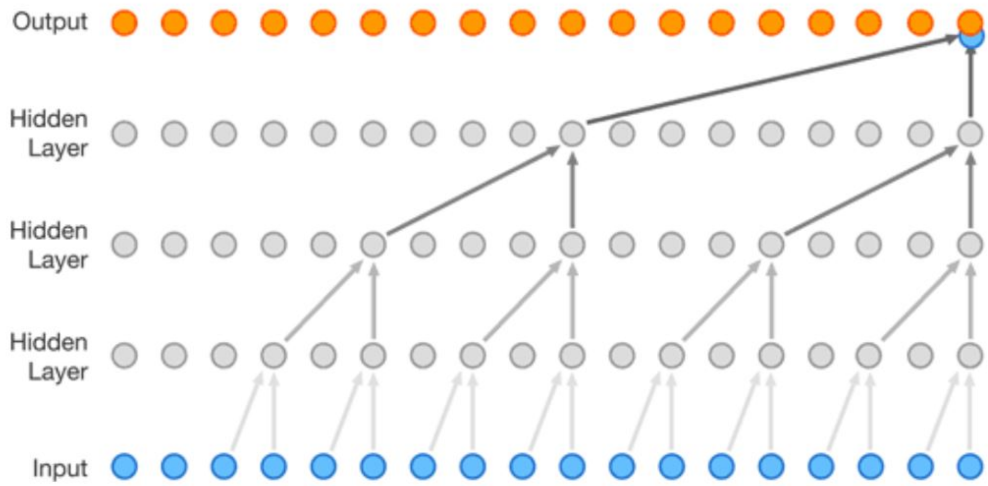

Figure 7.2.3 : A dilated CNN
Source: Adapted from [4]

The dilated convolutions allows to increase the receptive area exponentially where filters are not applied in sequential manner to the input but instead use a skip with a set dilation rate. Increasing dilation rate in an exponential manner, it was possible to capture history 256 days of history with 9 dilated convolution layer. In this project , a stack of 8-dilated convolution layer followed by 2 dense layer were used which is then used to learn pattern in time-series. Use of Activated Gating and Skip Connections improves the accuracy of the model.

8 Result

8.1 ARIMA

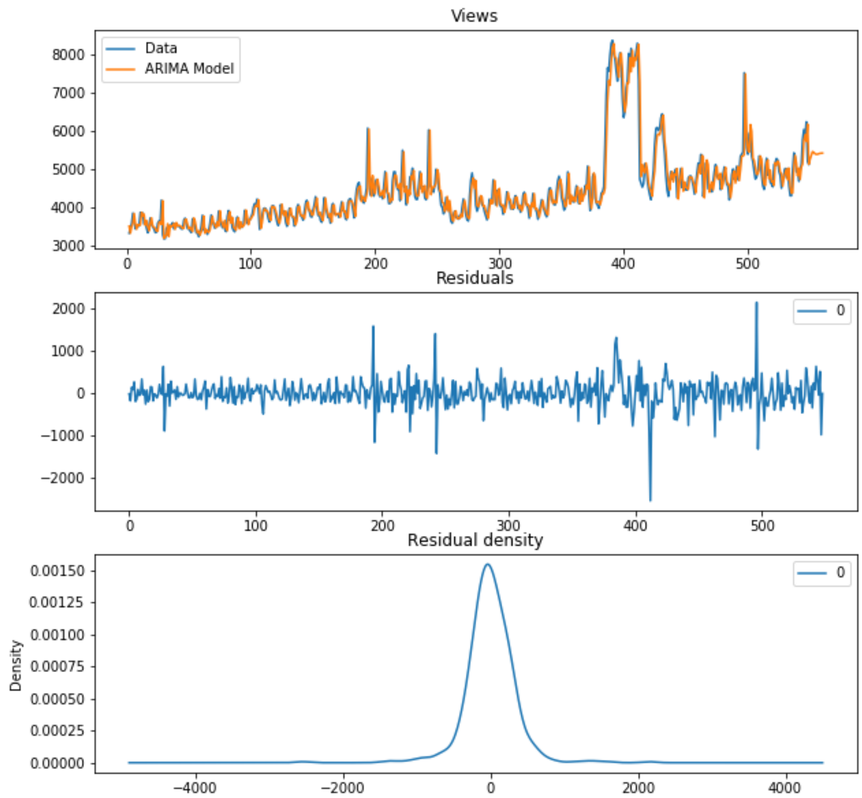

Figure 8.1 :ARIMA prediction sample

Root mean squared error for prediction with ARIMA model on the given dataset is **337.862382**

8.2 LSTM

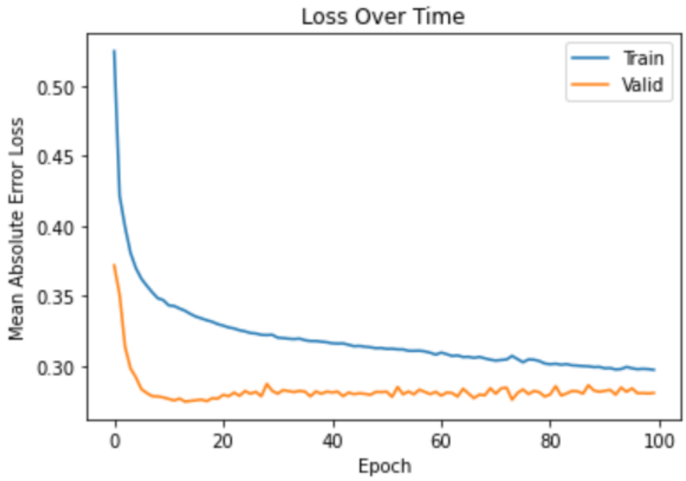

Figure 8.2: Loss graph of train vs validation for LSTM

Root mean squared error for prediction with LSTM model on the given dataset is **287.845766**

8.3 CNN

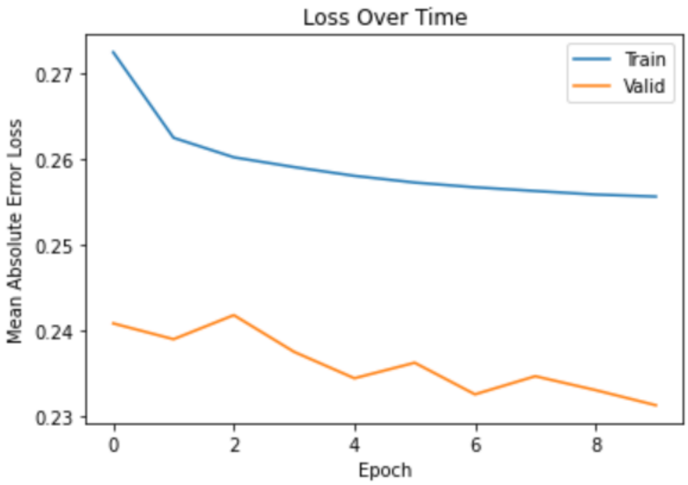

Figure 8.3 : Loss graph of train vs validation for CNN

Root mean squared error for prediction with CNN model on the given dataset is **271.942186**.

8.4 Article Trend Analyser Web-App

For the web-app , wikimedia rest api [10] is used to fetch real-time data for any Wikipedia article whose title is entered into the site and then predictions are made with different models.

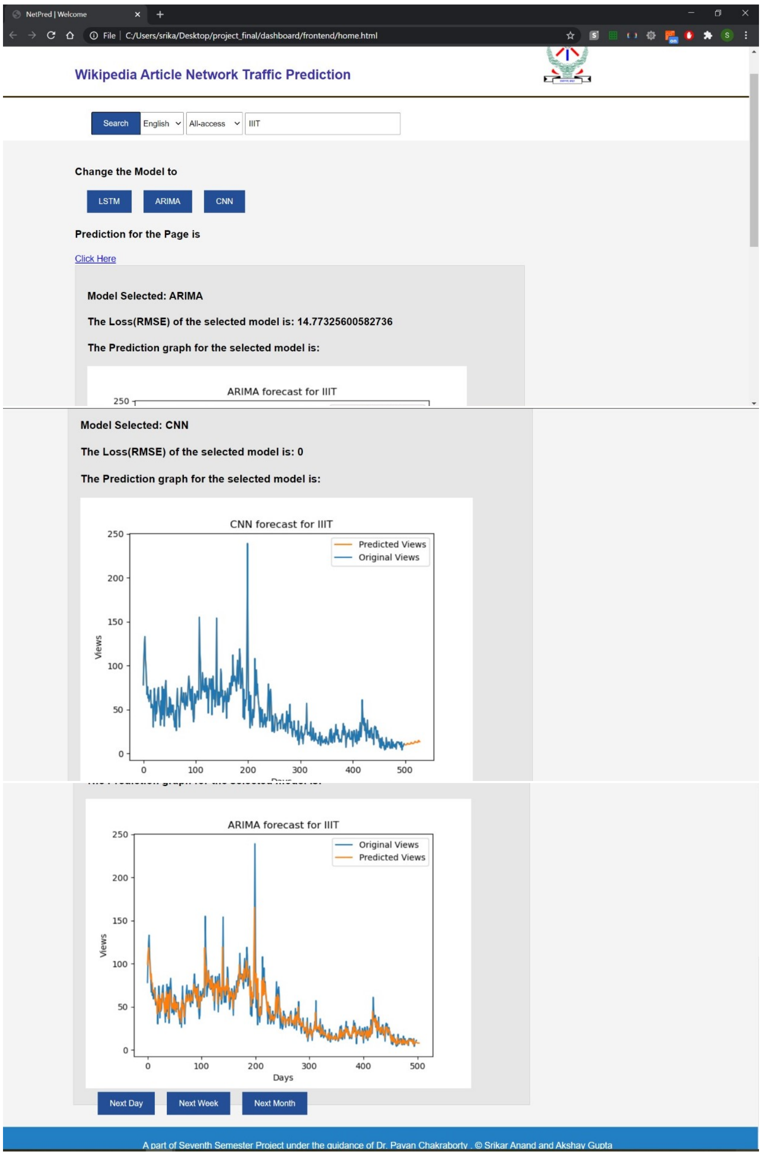

Figure 7.2.2 : Article Trend Analyser with different models for IIIT wikipedia page

9 Requirements

- Intel i5 or above processor.
- RAM - 8GB minimum.
- Operating System - Linux (Ubuntu), Windows
- Python 3.6 and above
- Tensorflow v1.14 with Keras API
- Pandas , Numpy , SKlearn statsmodel python libraries
- matplotlib and seaborn libraries for plotting purposes
- jupyter notebook or spyder IDE
- Google Colab resources
- Google Chrome web browser for Google Collaboratory
- Flask framework

10 Conclusion

Time series analysis is a must for every organization to understand seasonality, cyclic,trend and randomness in the data . In this project , three different models for time series analysis is done namely ARIMA , LSTM , and CNN.
The root mean square error for ARIMA model was found to be 337.862382.
The root mean square error for LSTM model was found to be 287.845766.
The root mean square error for CNN model was found to be 271.942186.
The CNN model was found to be great at picking recurring patterns in time-series. CNN predictions was also found to be more accurate and expressive when compared to LSTM and ARIMA model.

11 References

[1] Forecast future traffic to Wikipedia Pages [Online],Available: https://www.kaggle.com/c/web-traffic-time-series-forecasting [Accessed : 10-09-2020]

[2] N. Petluri and E. Al-Masri, "Web Traffic Prediction of Wikipedia Pages," 2018 IEEE International Conference on Big Data (Big Data), Seattle, WA, USA, 2018, pp. 5427-5429, doi: 10.1109/BigData.2018.8622207.

[3] Jia Li,Andrew W. Moore, "Forecasting Web page views : Methods and Observation" , Journal of Machine Learning Research 9 (2008) 2217-2250

[4] 0oord, Aaron Dieleman, Sander Zen, Heiga Simonyan, Karen Vinyals, Oriol Graves, Alex Kalchbrenner, Nal Senior, Andrew Kavukcuoglu, Koray. (2016). WaveNet: A Generative Model for Raw Audio.

[5] Auto correlation and partial auto correlation [Online]. Available: http://sfb649.wiwi.hu-berlin.de/fedc_homepage/xplore/tutorials/sfehtmlnode59.html. [Accessed: 16-Sep-2020]

[6] Jason Brownlee, "How to Decompose Time Series Data into Trend and Seasonality" Posted on January 30, 2017. [Online]. Available: https://machinelearningmastery.com/decompose-time-series-data-trend-seasonality/ [Accessed: 16-Sep-2020].

[7] Shi Yan, "Understanding LSTM and its diagrams," Medium, 14-Mar-2016. [Online]. Available: `https://medium.com/mlreview/understanding-lstm-and-its-diagrams-37e2f46f171` [Accessed: 16-Sep-2020].

[8] Eugine Kang, " Time Series : Arima Model", Medium , 26-Aug-2017. [Online]. Available `https://medium.com/@kangeugine/time-series-arima-model-11140bc08c6` [Accessed: 16-Sep-2020]

[9] https://github.com/farizrahman4u/seq2seq [Accessed : 01-11-2020]

[10] *https://wikimedia.org/api/rest_v1/*