CHAPTER 1

Introduction

This chapter focuses on introduction of Time Series Forecasting, necessity, objectives of the proposed system, theme and report organization

1 Introduction

1.1 Introduction

Data is the new fuel. Today everything is becoming digital from office work to homework, we are surrounded by digital equipment where ever we go. These digital equipment generate a tremendous amount of data every day. Today, condition is so critical that if we stop optimizing this data then storage will run out of capacity in just 6 months. The generated is not waste at all. Some of this data is very useful if we mine it properly. This data is very useful from a business perspective. This huge amount of data generation is called as Big Data. Now, this data cannot be handled manually so we use modern methods for this purpose called Data Analysis.

"Time Series" is a series of values of a quantity obtained at successive times, often with equal intervals between them.We are analyzing "Time Series" data of Wikipedia articles which will be useful to predict future hits on that articles. This is quite challenging task as articles hits are going to be influenced by languages and dates. Through this, we can identify which article is getting maximum hit at the specific period.

1.2 Necessity

With the rapid rise of real time data sources, prediction of future trends and the detection of anomalies is becoming increasingly important. Accurate time series forecasting is critical for business operations for optimal resource allocation, budget planning, anomaly detection and tasks such as predicting customer growth, or understanding stock market trends

1.3 Objective

1. Data Transformation.

2. To generate parameters and visualize the patterns.

3. To implement different forecasting methods.

4. To critique the performance of different forecasting methods.

1.4 Report Organization

The project work involves the forecasting of a time series data of various Wikipedia articles and predict the future hits on those article:

1.4.1 Chapter 1: Introduction

This chapter contains basic idea of Time Series Forecasting which uses different Forecasting Models to achieve results.

1.4.2 Chapter 2: Literature Survey

This chapter describes the sources which has been referred for better understanding of Time Series Forecasting.

1.4.3 Chapter 3: System Design

This chapter describes the design of the basic architecture of the project work, data flow diagram.

1.4.4 Chapter 4: Experimental Setup and Results

This chapter describes the experimental setup along with the details required for development of the project and the results achieved from the project work.

1.4.5 Chapter 5: Conclusion

This chapter describes the conclusion obtained from the proposed system and also describes the future work that can be done in the proposed system. It also describes the various applications of the project.

1.4.6 Chapter 6: References

This chapter describes the various websites and articles we studied for the understanding and the development of the project work.

CHAPTER 2

Literature Survey

This chapter describes the various websites we had studied for understanding various concepts and development of the project work.

2 Literature Survey

[1] **Time series forecasting using improved ARIMA**

The authors Soheila Mehrmolaei and Mohammad Reza Keyvanpour provides brief insights on the need of time-series forecasting, propose a novel approach to improve ARIMA model by applying a mean of estimation error for time series forecasting. Experimental results indicate that the proposed approach can improve performance in the process of time series data forecasting.

[2] **A Survey on Forecasting of Time Series Data**

The authors G.Mahalakshmi, Dr.S.Sridevi and Dr.S.Rajaram describe processes used for forecasting models. This survey covers the overall forecasting models, the algorithms used within the model and other optimization techniques used for better performance and accuracy. The various performance evaluation parameters used for evaluating the forecasting models are also discussed in this paper. This study gives the reader an idea about the various researches that take place within forecasting using the time series data.

CHAPTER 3

System Design

This chapter describes the design of the basic architecture of the project work, data flow diagrams, flow charts, sequence diagram and class diagram.

3 System Design

3.1 System Architecture

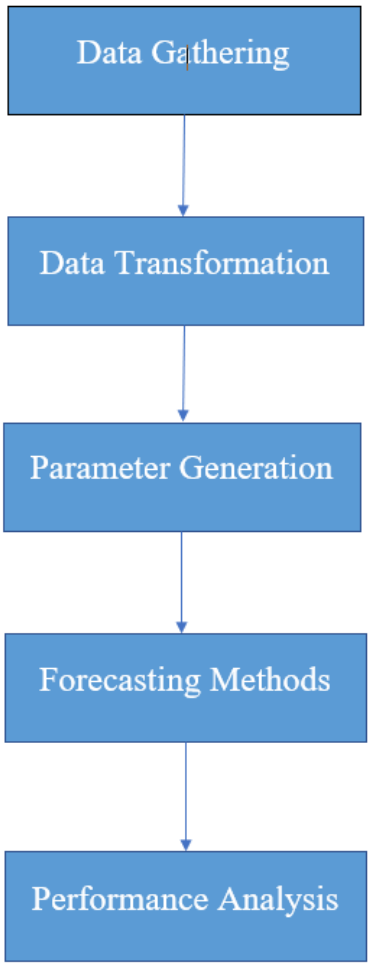

Figure 3.1 : System Architecture

Following figure 3.1. describes the system architecture of project. The first step is to gather data, which will have daily hits on various Wikipedia articles. This gathered data might be linear or nonlinear thus, in the next stage, we are serializing this gathered data. That serialized data is going to be applied on various forecasting models. In the next stage, we use various forecasting models with the help of Python and R languages. In later stage we analyse the forecasted results from the different forecasting models. These results will be helpful to predict hits on various Wikipedia articles for next two months.

3.2 Data Flow Diagram

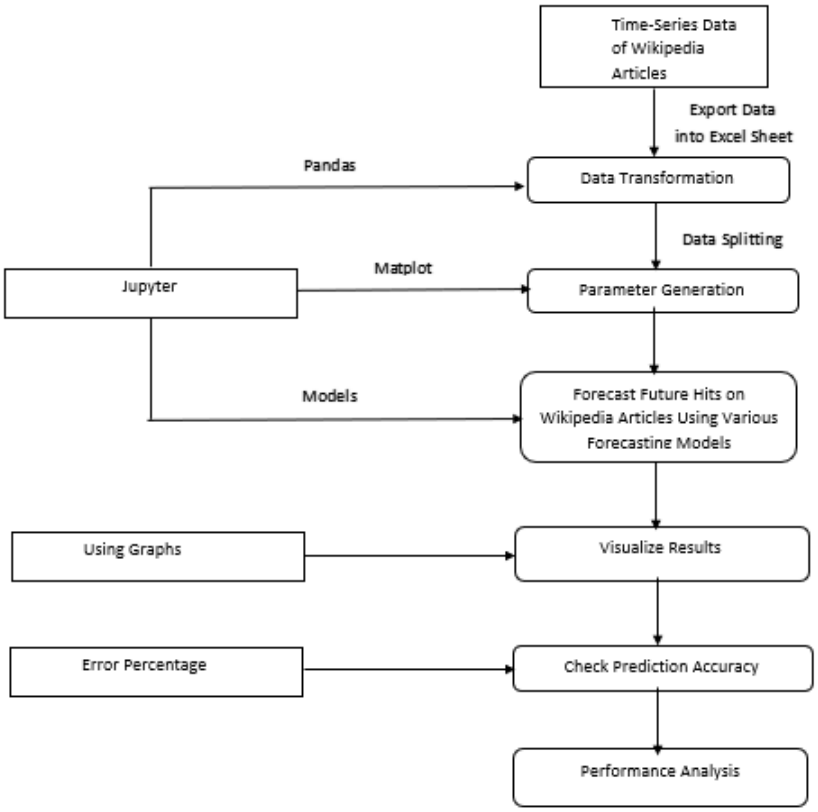

Figure 3.2.1 : Data Flow Diagram

The figure 3.2.1 shows the data flow diagram which describes the data propagation throughout the system during execution of various request.

3.3 Module 1: Data Transformation

The article information contains page and number of visit to each page in a particular time. In this module, almost all the needed libraries are imported and then dataset is being seperated into two sets i.e, in train.csv and validation.csv. The dataset is being imported and being cleaned The main purpose of this module is a cleaning of the data according to forecasting models and then split them into two sets.

3.4 Module 2: Parameter Generation

In this module, we mainly perform parameter generation and visualization of the data. These generated parameters are used to verify results from various forecasting models. Parameters generated in this process are Agent, Project, Access, and Language. By using these parameters we perform further operations.

3.5 Module 3: Forecasting Methods

The various forecasting methods are as follows.

- Median Model: In situations where the location (level) and variability is not really changing over time, and the serial dependence between consecutive observations is weak, means or medians (or various other possible estimates of location) may be quite reasonable as forecasts.
- ARIMA Model: ARIMA stands for Autoregression integrated moving average. An ARIMA model is a class of statistical models for analyzing and forecasting time series data.It uses standard structure, and provides a simple yet powerful method for making skillful time series forecasts.[2]
- Prophet: Prophet model is developed by Facebook and is freely available for use.It works best with time series that have strong seasonal effects and several seasons of historical data.It is accurate and fast,fully automatic and available in both R and Python.
- SMAPE: Symmetric mean absolute percentage error (SMAPE or sMAPE) is an accuracy measure based on percentage (or relative) errors.

3.6 Module 4: Performance Analysis

In this module, analysis of results will be performed. This leads to give better results to forecast the hits on articles.

CHAPTER 4

Experimental Setup and Results

This chapter describes the experimental setup along with the details of the installation of software required for development of the project and the results achieved from the project work.

4 Experimental Setup and Results

4.1 Tools Used

Anaconda Distribution 4.0

The Most Trusted Distribution for Data Science. Anaconda is a package manager, an environment manager, a Python distribution, and a collection of over 1,500+ open source packages. Anaconda is free and easy to install, and it offers free community support. Anaconda gives the User ability to make an easy install of the version of python he/she wants. Removes bottlenecks involved in installing the right packages while taking into considerations their compatibility with various other packages as might be encountered while using pip. Over 200 packages are automatically installed with Anaconda. Over 1500 additional open source packages can be individually installed from the Anaconda repository with the conda install command.

Thousands of other packages are available from Anaconda Cloud.

If there is one tool which every data scientist should use or must be comfortable with, it is Jupyter Notebooks (previously known as iPython notebooks as well).

Jupyter Notebooks are powerful, versatile, shareable and provide the ability to perform data visualization in the same environment. Jupyter Notebooks allow data scientists to create and share their documents, from codes to full blown reports. They help data scientists streamline their work and enable more productivity and easy collaboration. Due to these and several other reasons you will see below, Jupyter Notebooks are one of the most popular tools among data scientists.

Jupyter Notebook is an open-source web application that allows us to create and share codes and documents.

It provides an environment, where you can document your code, run it, look at the outcome, visualize data and see the results without leaving the environment. This makes it a handy tool for performing end to end data science workflows â data cleaning, statistical modeling, building and training machine learning models, visualizing data, and many, many other uses.

Jupyter Notebooks really shine when you are still in the prototyping phase. This is because your code is written in indepedent cells, which are executed individually. This allows the user to test a specific block of code in a project without having to execute the code from the start of the script.

These Notebooks are incredibly flexible, interactive and powerful tools. They even allow you to run other languages besides Python, like R, SQL, etc. Since they are more interactive than an IDE platform, they are widely used to display codes in a more pedagogical manner.

4.2 Anaconda Distribution 4.0 System Requirements

1. License: Free use.

2. Operating system: Windows 7 or newer, 64-bit macOS 10.10+, or Linux, including Ubuntu, RedHat, CentOS 6+, and others.

3. If your operating system is older than what is currently supported, you can find older versions of the Anaconda installers in there archive that might work for you.

4. System architecture: Windows- 64-bit x86, 32-bit x86; MacOS- 64-bit x86; Linux- 64-bit x86, 32-bit x86, 64-bit Power8/Power9.

5. Minimum 5 GB disk space to download and install.

4.3 Result

4.3.1 Data Transformation

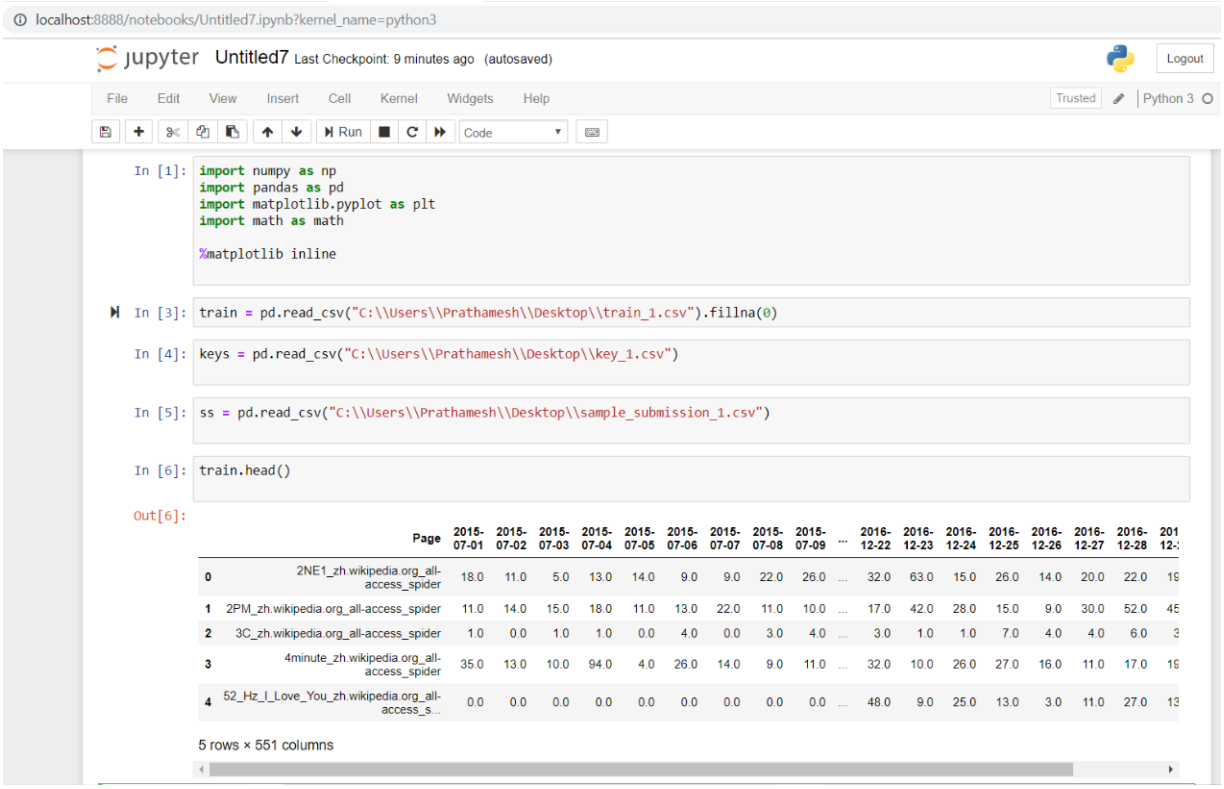

Figure 4.3.1 : Data Transformation

The concept of missing values is important to understand in order to successfully manage data. If the missing values are not handled properly by the researcher, then he/she may end up drawing an inaccurate inference about the data. Figure 4.3.1 shows handling of missing values. We replace the "NAN" values with "0". This process is called as data cleaning.

4.3.2 Parameter Generation

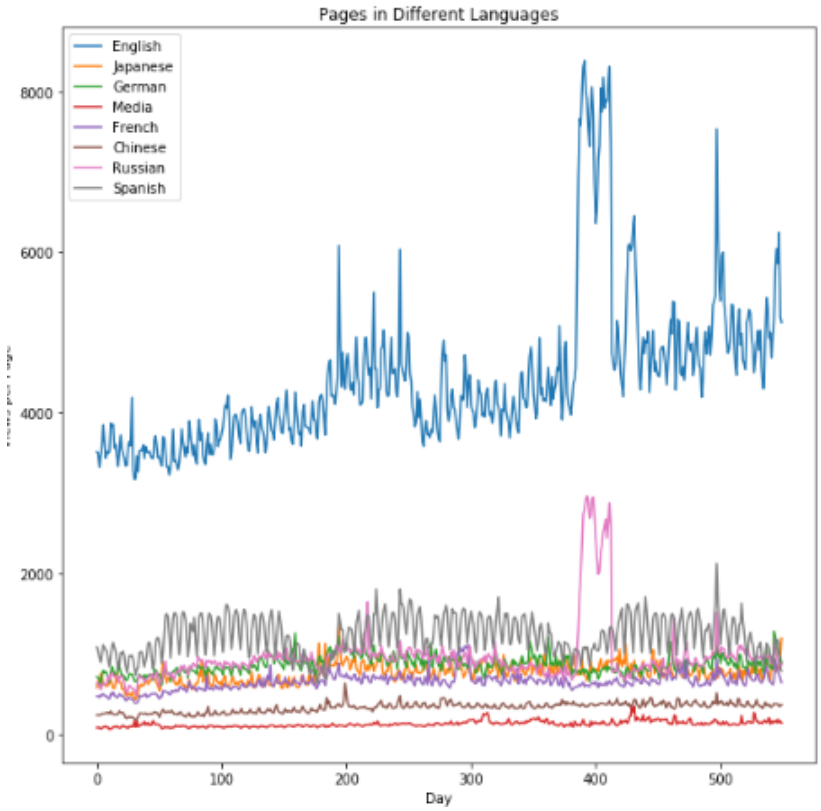

Figure 4.3.2 : Wikipedia Traffic By Page Language

Different languages used in the Wikipedia articles affect the dataset as shown in above Figure 4.3.1.So, we use a simple regular expression to search for the language code in the Wikipedia URL.There are also a number of non-Wikipedia URLs that will fail the regex search.These are Wikimedia pages, so we give them the code 'na' since we haven't determined their language.English shows a much higher number of views per page, as expected since Wikipedia is a US-based site. The English and Russian plots show very large spikes around day 400 (around August 2016), with several more spikes in the English data later in 2016.In the Spanish data there is clear periodic structure there, with a 1 week fast period and what looks like a significant dip around every 6 months or so.

4.3.3 Wikipedia Pages by Access

```
In [17]: df_access = train.groupby(['Access'])[cols].mean()
         df_access = df_access.T

In [29]: f, (ax1) = plt.subplots(1, 1, figsize = (7, 7), sharex=True)
         df_access.plot(ax=ax1)
         plt.show()
```

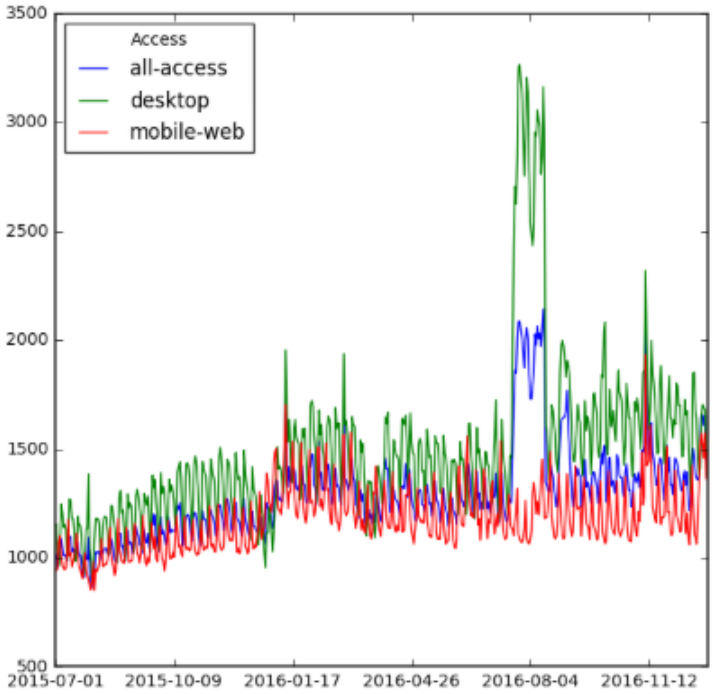

Figure 4.3.3 :Wikipedia Pages by Access

As the above Figure 4.3.3 shows Wikipedia pages get affected by its access, It categorized in to three parts desktop,mobile-web and all .Desktop access has higher number of views per page.All access views plots show medium level spikes around day.

4.3.4 Wikipedia pages by Agent

```
In [6]: df_agent = train.groupby(['Agent'])[cols].mean()
        df_agent = df_agent.T

In [7]: f, (ax1) = plt.subplots(1, 1, figsize = (7, 7), sharex=True)
        df_agent.plot(ax=ax1)
        plt.show()
```

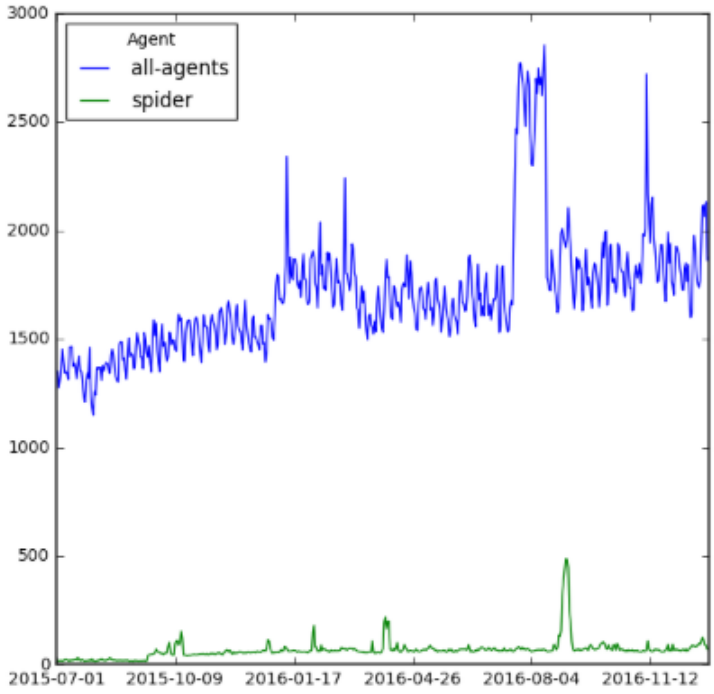

Figure 4.3.4 : Wikipedia pages by Agent

As the above Figure 4.3.4 shows the web articles are of which agent types. There are two agents all-agents and spider. Articles having all-agent type shows higher spikes in the graph. Articles of all-agent type has higher views .

4.3.5 Wikipedia pages by Project Patterns

```
df_project = train.groupby(['Project'])[cols].mean()
df_project = df_project.T
```

```
f, (ax1) = plt.subplots(1, 1, figsize = (7, 7), sharex=True)
df_project.plot(ax=ax1)
plt.show()
```

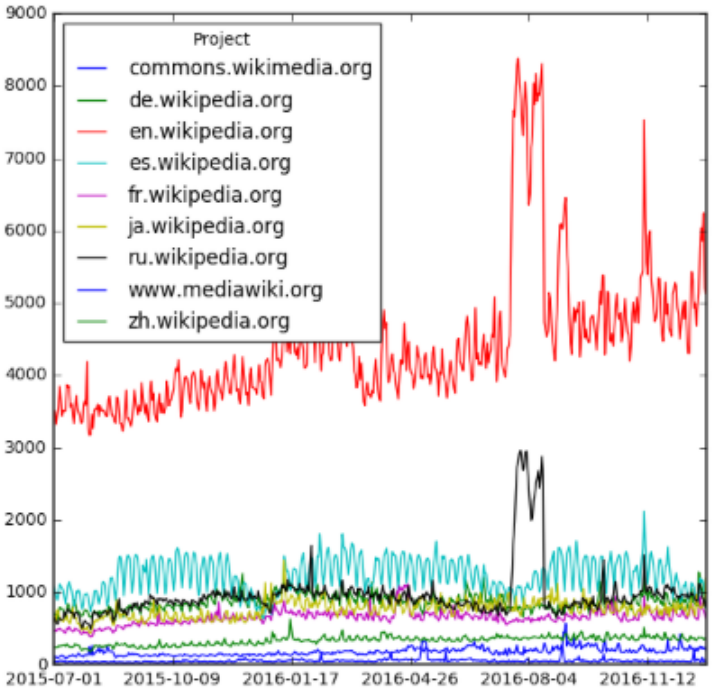

Figure 4.3.5 : Wikipedia pages by Projects Patterns

Wikipedia page traffic got influenced by the different project patterns.There are 7 languages plus the media pages of the project patterns.The languages used here are: English, Japanese, German, French, Chinese, Russian, and Spanish and the media pages as well.As the Figure 4.3.5: shows English project pages shows higher number of views per page.

4.3.6 Implementation of Median Model

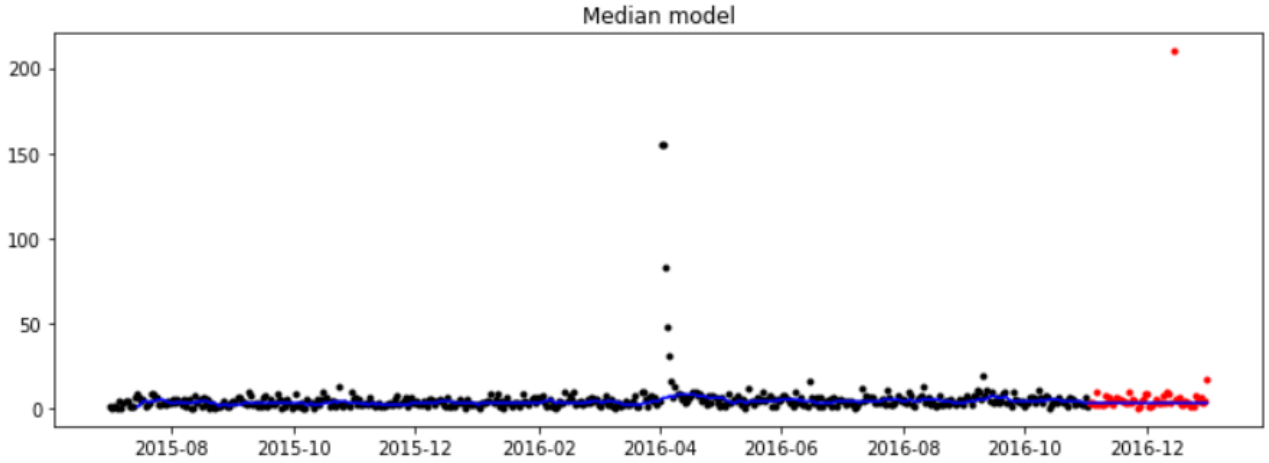

Figure 4.3.6 : Implementation of Median Model

Above Figure 4.3.6 represents average hits for an particular article. The article name which is used in median model for forecasting the results is "3C$_z$*h.wikipedia.org*$_a$*ll*−*accesses*$_s$*pider*".3*Cwordisusedforcon* 04*and*2016−05*whereChinesenewyearison*1*stofmayofeveryyear.Wepredictingthat*3*Carticleisgettingmaximur Commerceorganizesbigdiscountsalesduringthesefestivemonths.*

CHAPTER 5

Conclusion

This chapter describes the final conclusion obtained from the proposed system and also describes the future work that can be done in the proposed system.It also describes the various applications of the project.

5 Conclusion

5.1 Conclusion

Today data is generated in huge amount and this data is very useful if we clean it properly. Most of the time the data generated has common factor called Time. Time series is a sequence taken at successive equally spaced points in time. Time Series Forecasting is very useful to draw out useful insight from this data. Several time-series forecasting models can be useful to predict accurate results from data.

5.2 Future work

These are some of the amendments which we are going to fulfill in next semester:

1. Forecast Results using various forecasting models.
2. Compare parameters with forecasted results.

5.3 Applications

General : Time series are used in signal processing, pattern recognition, mathematical finance, weather forecasting, earthquake prediction, electroencephalography, control engineering, astronomy, communications engineering, and largely in any domain of applied science and engineering which involves temporal measurements.

Specific : Using the time series data we calculate the future hits on Wikipedia articles.

References

[1] W. Fan, âSystematic data selection to mine concept-driftig data streamsâ, Proc.of ACM KDD,pp. 128-13

[2] C. Xu, Z. Li, W. Wang. âShort-term traffic flow prediction using a methodology based on autoregressive moving average and genetic programmingâ, Transport, vol.31, pp.343- 358,2016.

[3] G. P. Zhang. âTime series forecasting using a hybrid ARIMA and neural network model.â Neurocomputing, vol.50, no.1, pp.159-175, 2003.

[4] G.Mahalakshmi, Dr.S.Sridevi and Dr.S.Rajaram. A Survey on Forecasting of Time Series Data, 2016.

[5] Soheila Mehrmolaei, Mohammad Reza Keyvanpour. Time series forecasting using improved ARIMA, April 2016.

Websites

[1] https://searchbusinessanalytics.techtarget.com/definition/data-exploration
[2] https://www.datascience.com/blog/time-series-forecasting-machine-learning-differences
[3] https://otexts.org/fpp2/forecasting-on-training-and-test-sets.html
[4] https://machinelearningmastery.com/backtest-machine-learning-models-time-series-forecasting/
[5] https://otexts.org/fpp2/case-studies.html
[6] https://pandas.pydata.org/pandas-docs/stable/10min.html
[7] https://pandas.pydata.org/pandas-docs/stable/cookbook.htmlcookbook
[8] https://www.analyticsvidhya.com/blog/2018/02/time-series-forecasting-methods/
[9] https://www.kdnuggets.com/2018/03/time-series-dummies-3-step-process.html
[10] https://matplotlib.org/gallery/lines$_b$ars$_a$nd$_m$arkers/csd$_d$emo.htmlsphx − glr − gallery − lines − bars − and − markers − csd − demo − py
[11]*https* : //*matplotlib.org*/*users*/*index.html*
[12]*https* : //*matplotlib.org*/*gallery*/*lines*$_b$*ars*$_a$*nd*$_m$*arkers*/*cohere.htmlsphx* − *glr* − *gallery* − *lines* − *bars* − *and* − *markers* − *cohere* − *py*
[13]*https* : //*eng.uber.com*/*neural* − *networks*/

Search Engine
www.google.co.in