Response Time Analysis and Prediction

Project report submitted in partial fulfillment

of the requirements for the degree of

Bachelor of Technology

in

Computer Science Engineering

by

Shivam Sharma - 15UCS130

Vivek Kumar - 15UCS162

Under Guidance of

Prof. Vikas Bajpai

Department of Computer Science and Engineering

The LNM Institute of Information Technology, Jaipur

May 2018

The LNM Institute of Information Technology

Jaipur, India

CERTIFICATE

This is to certify that the project entitled Alumni Association Web Application , submitted by Shivam Sharma (15UCS130) and Vivek Kumar (15UCS162) in partial fulfillment of the requirement of degree in Bachelor of Technology (B. Tech), is a bonafide record of work carried out by them at the Department of Computer Science and Engineering, The LNM Institute of Information Technology, Jaipur, (Rajasthan) India, during the academic session 2018-2019 under my supervision and guidance and the same has not been submitted elsewhere for award of any other degree. In my/our opinion, this thesis is of standard required for the award of the degree of Bachelor of Technology (B. Tech).

Date Adviser: Prof. Vikas Bajpai

Acknowledgments

This project would not have been conceived without the kind support and help of many individuals. We would like to express our deepest gratitude to our supervisor Prof. Vikas Bajpai whose expertise, inspiring ideas, understanding and patience, added considerably to our ongoing B.Tech project experience. This project has helped us in enriching our experience and has given us an opportunity to learn new techniques and apply them.

Abstract

In this project we have taken out the data of the response time for 100 websites. We have used open source websites for this project. We have used localhost server as the hosting server so that the response time does not gets affected by the internet speed. We have used Badboy a website testing software. We have made several iterations of the flow path . Taking different paths in different websites so as to collect random data. Taking out iterations consisting of 1, 20 and 30 sets of iterations . Taking out the Response time graph of the 20 and 30 iterations.

We used XAMPP and WAMP software to host the websites and calculated response time using Badboy. It took us 3 months to do this work .We downloaded an average of 1500 websites in the process. With the probability of success was 1/15 . But in the end we were able to take out the data . We will try to automate the task by using automation code. The data collected right now will be used for the prediction of the response time of Important Websites which is our future goal

More details about the project can be found here –

https://github.com/jack17529/ResponseTimeAnalysis

TABLE OF CONTENTS

Chapter-1

Introduction

The smaller the response time the better is the Website's economy.When we talk about E-Commerce the first thing that comes into our mind is the money made by the E-Commerce websites. All big companies like Amazon, Flipcart etc put a lot of money to check that the response time of the website remains small at all the time of a day.

The smaller the response time the smaller is the time taken to get from one request to other request , thus more people will stay on the site. Which help converting a viewer into a customer. Hence directly affecting the sales of the products in a shopping website.

Chapter-2

Response Time

2.1 What is Response Time ?

Response time refers to the amount of time Enterprise Server takes to return the results of a request to the user. The time taken to make HTTP GET request to a URL. More requests per minute can be performed if the response time of a website is small. Response time increases if the number of users on the system increases, even though the number of requests per minute declines.

The time that passes between the first byte of information to last byte of every image, style sheet or java file during a user's request is known as Response Time.

It consists of 3 parts –

1. Time to first byte
2. Time to receive headers
3. Time to load HTML of the site

2.2 Calculating Response Time

After the peak load point the response time calculations becomes inaccurate thus we are trying to calculate response time at peak load point.

The response time is inversely proportional to the requests made per minute. The sharper the decline in requests per minute, the steeper the increase in response time.

The formula for calculating the response time is -

$T_{response} = \mathbf{n/r} - T_{think}$

where

- **n** is the number of concurrent users
- **r** is the number requests per second the server receives
- $\mathbf{T_{think}}$ is the average think time (in seconds)

The think time in the equation is included so as to obtain an accurate response time. In this experiment the peak load is used as the bottleneck to calculate the response time.

2.3 Usage / Importance

Everyone in the business world understands that the importance of having fast response time of websites. The more time a customer spends on your website the more likely he/she is to spend money. The behavior of the website should influence the users to spend more time on the website.

Response time is critical to influencing buying behavior. The performance of your website is directly related to the level of your business skills. Checking the website performance the way in which customers will see, is what matters the most.

There are various tools to measure the response time –

1. Selenium.
2. JMeter
3. Badboy

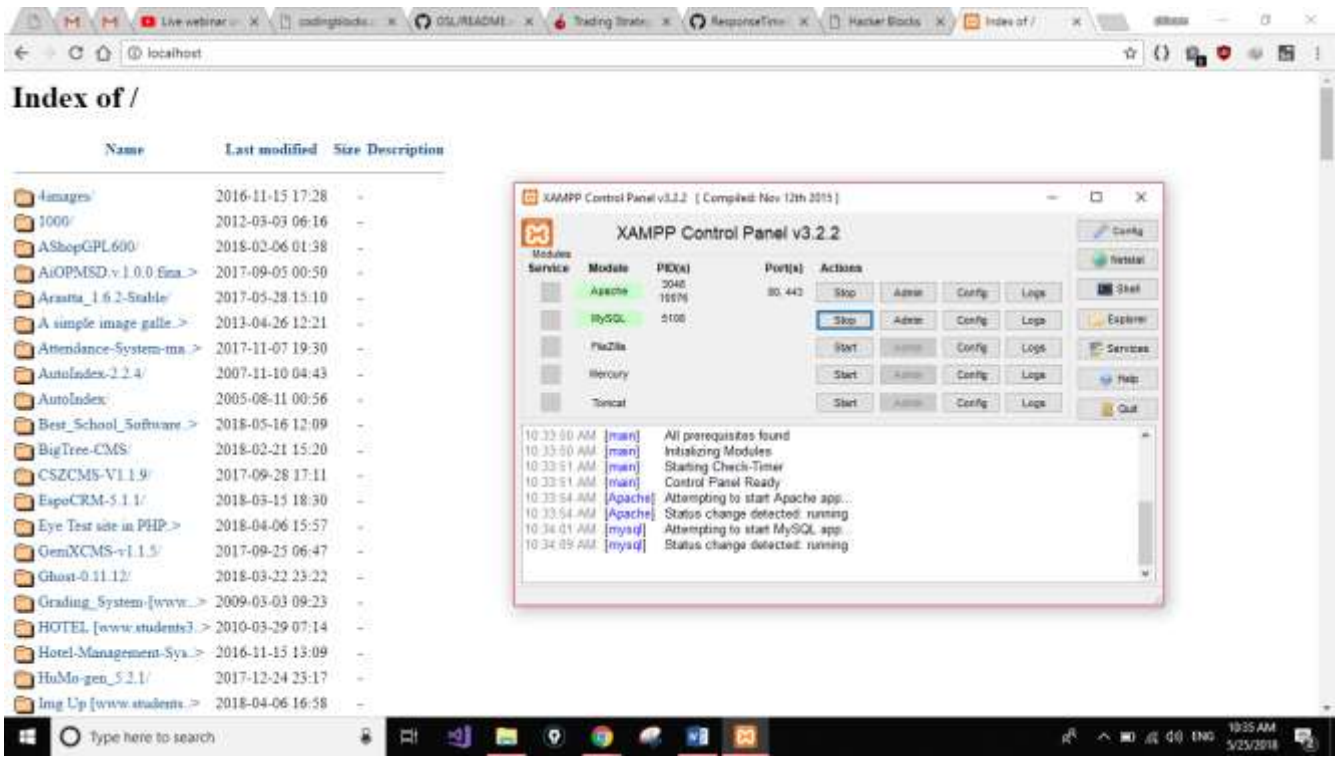
Chapter – 3

Our Work

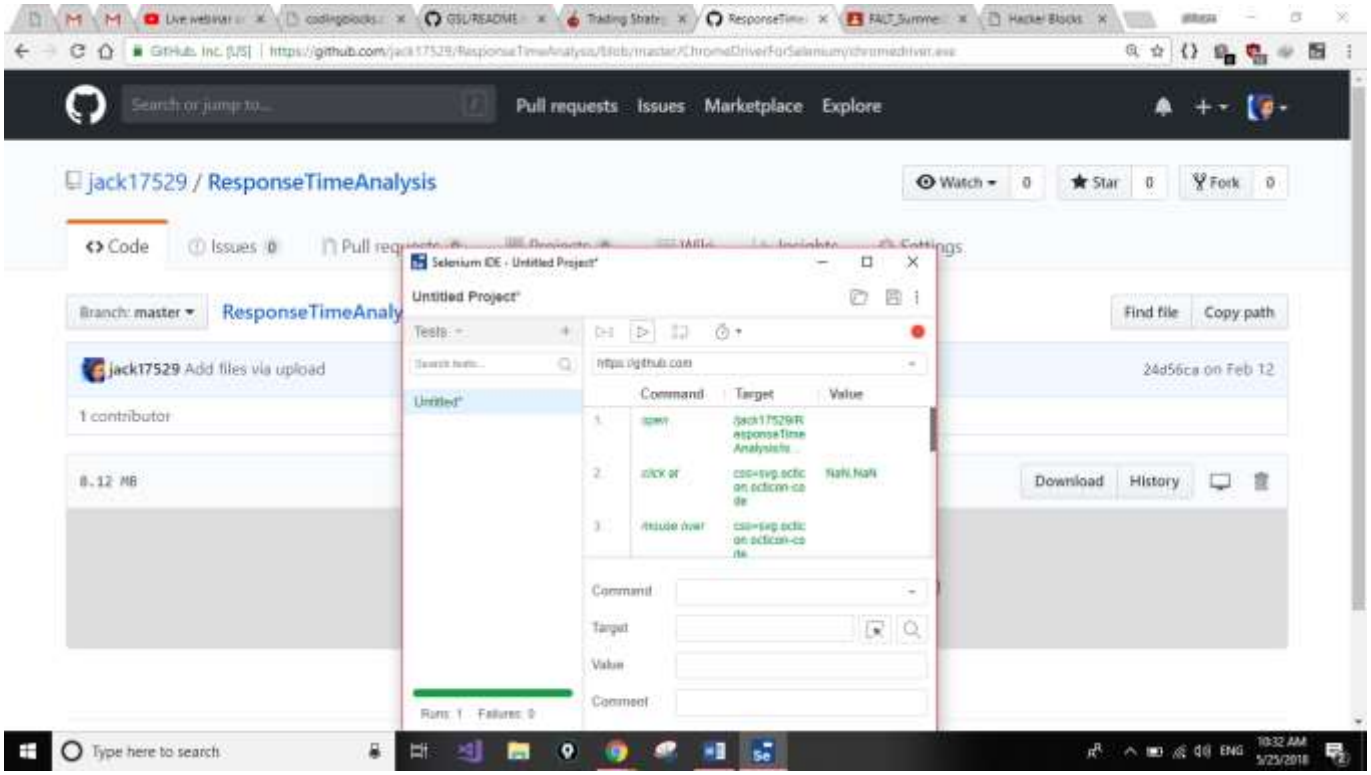
In the first week , we did a good research on what does response time mean , we checked several definitions of it on the internet. We did calculations of the response time with the formula.

In the second week, we read research papers to know what is the current trend on the research happening on the topic.

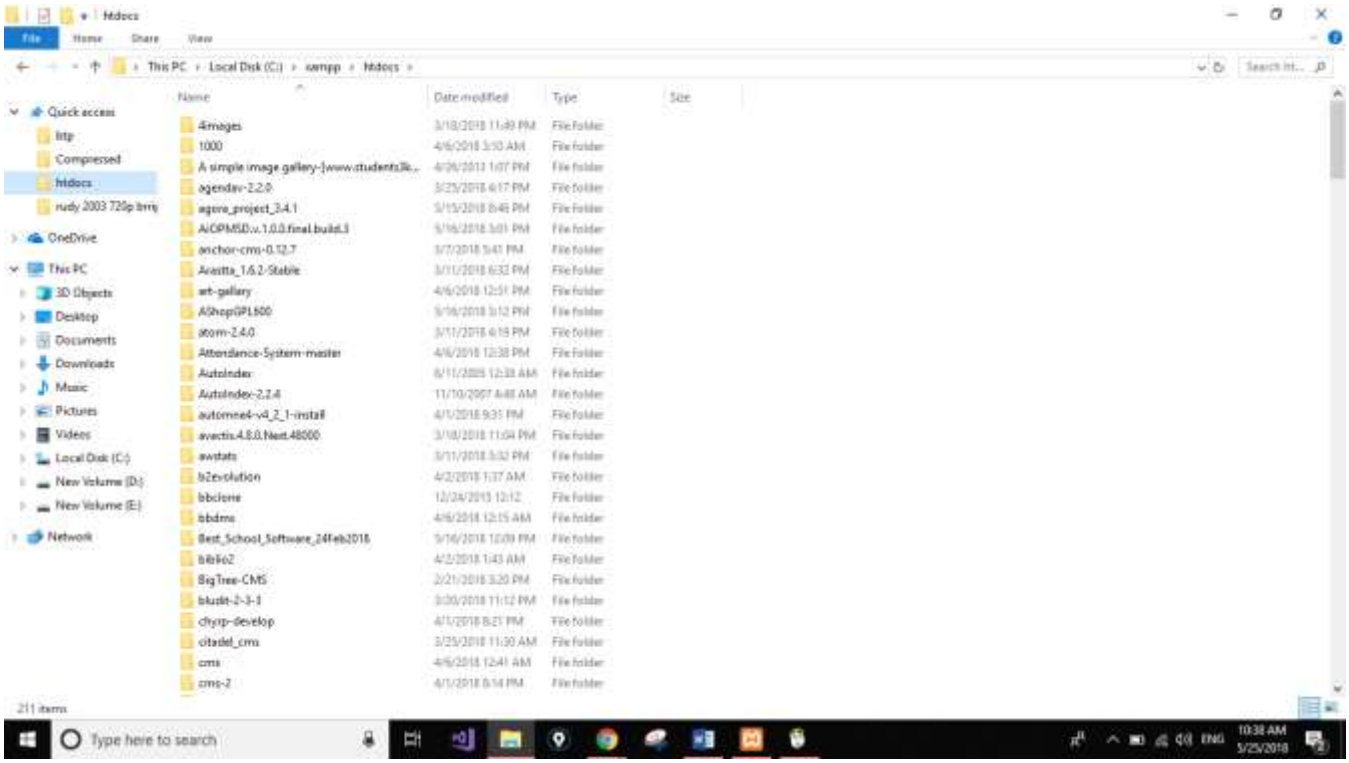
In the third week , we were asked to host 10 websites locally. So we first learnt how to use local servers to host your website then we tried XAMPP and WAMP. These are the most popular servers to host your website locally for windows users .We learnt how to host websites locally. We struggled in finding self hosting websites on the internet.

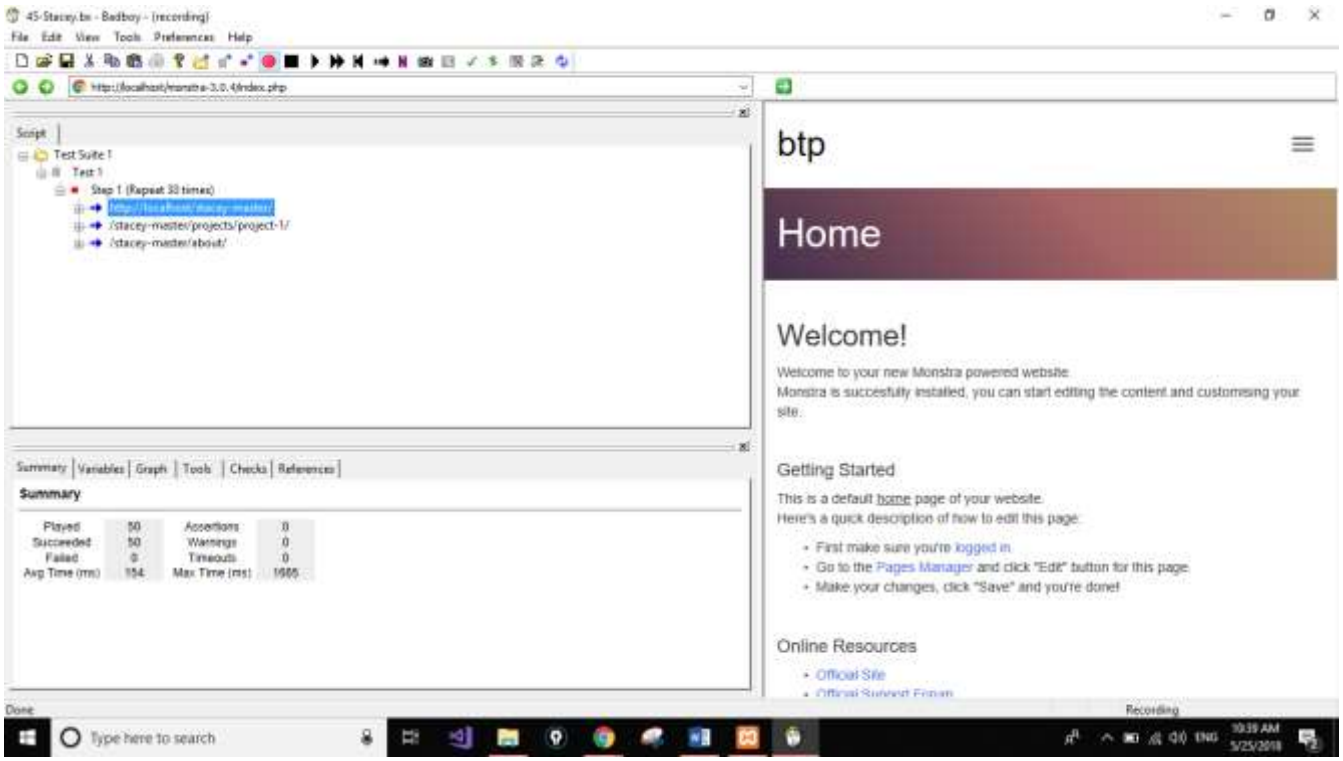


In the fourth week, we were asked to find the software to calculate response time , we tried different softwares including Selinum , JMeter . Shivam used the Selinum to find the response time and even automated the task , Vivek calculated the response time with JMeter.

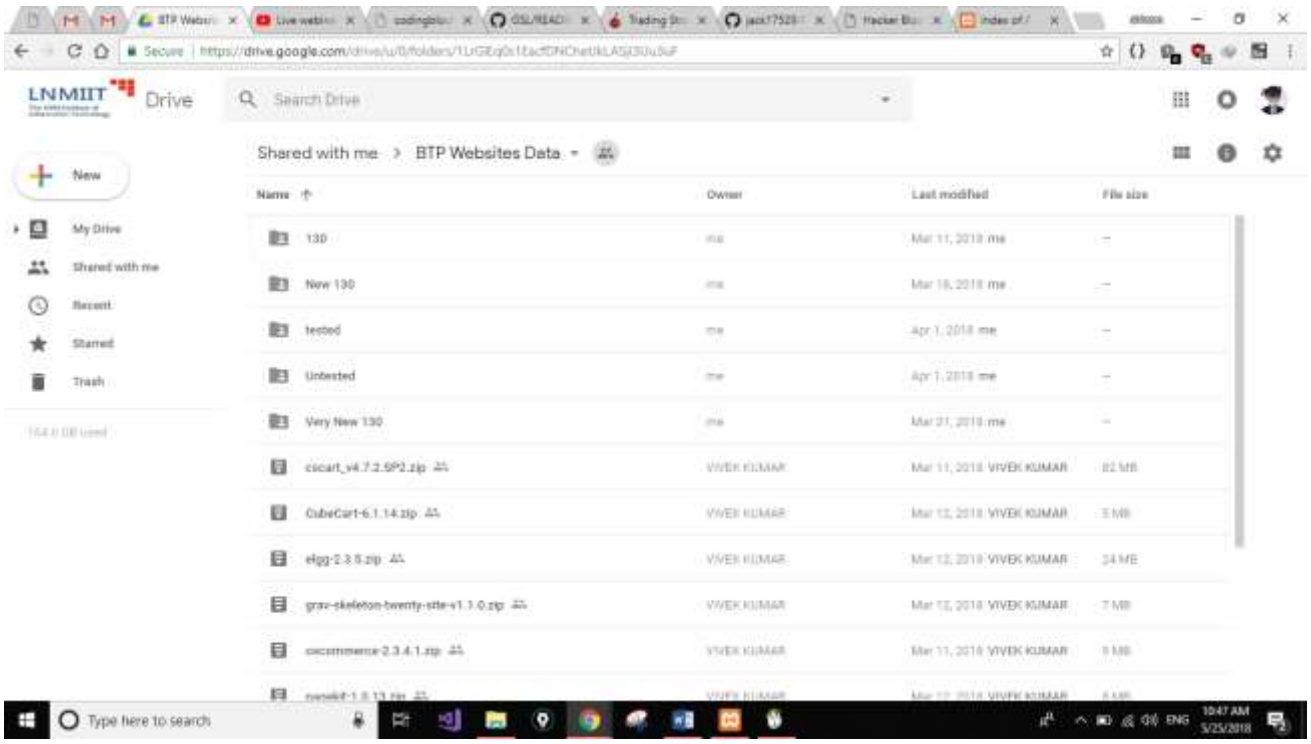


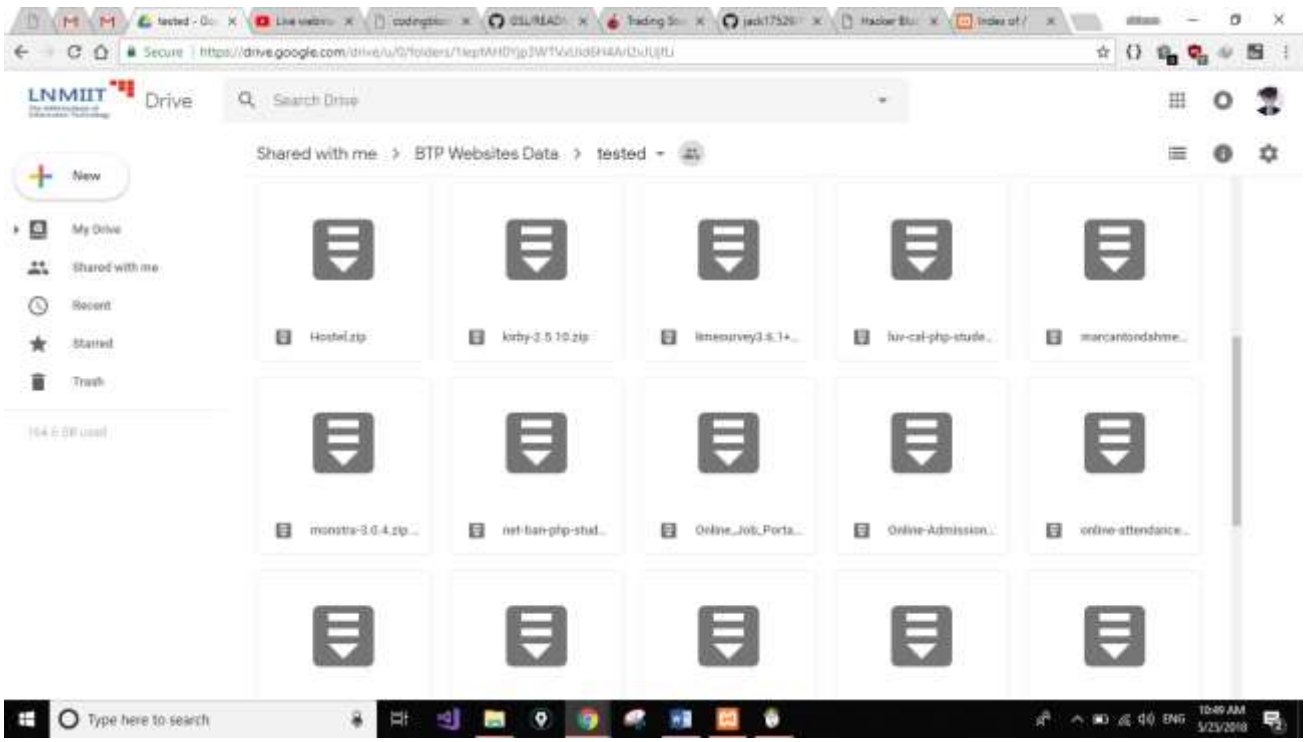
In the fifth week, we were told that we are not on the right track and have to bring data of 20 websites , then we used Badboy to calculate the response time of the websites that we collected. Most of the websites gave Php version error , we tested over a 100 websites yet only 20 worked . We brought both the data of the websites of 1,20,50 iterations and graphs. We even showed the DFD of the websites . After a lot of struggle our work was approved.


htdocs
File Home Share View
This PC › Local Disk (C:) › xampp › htdocs ›
Search ht...
Name
Date modified
Type
Size
Quick access
http
Compressed
htdocs
rudy 2003 720p brrip
OneDrive
This PC
3D Objects
Desktop
Documents
Downloads
Music
Pictures
Videos
Local Disk (C:)
New Volume (D:)
New Volume (E:)
Network
4images
1000
agendav-2.2.0
agore_project_3.4.1
AiOPMSD.v.1.0.0.final.build.3
anchor-cms-0.12.7
Avastta_1.6.2-Stable
art-gallery
AShopGPL600
atom-2.4.0
Attendance-System-master
AutoIndex
AutoIndex-2.2.4
automne4-v4_2_1-install
avactis-4.8.0.Next.48000
awstats
b2evolution
bbclone
bbdms
Best_School_Software_24Feb2018
bibliO2
BigTree-CMS
bludit-2-3-1
chyrp-develop
citadel_cms
cms
cms-2
211 items
Type here to search
ENG
10:38 AM
5/25/2018

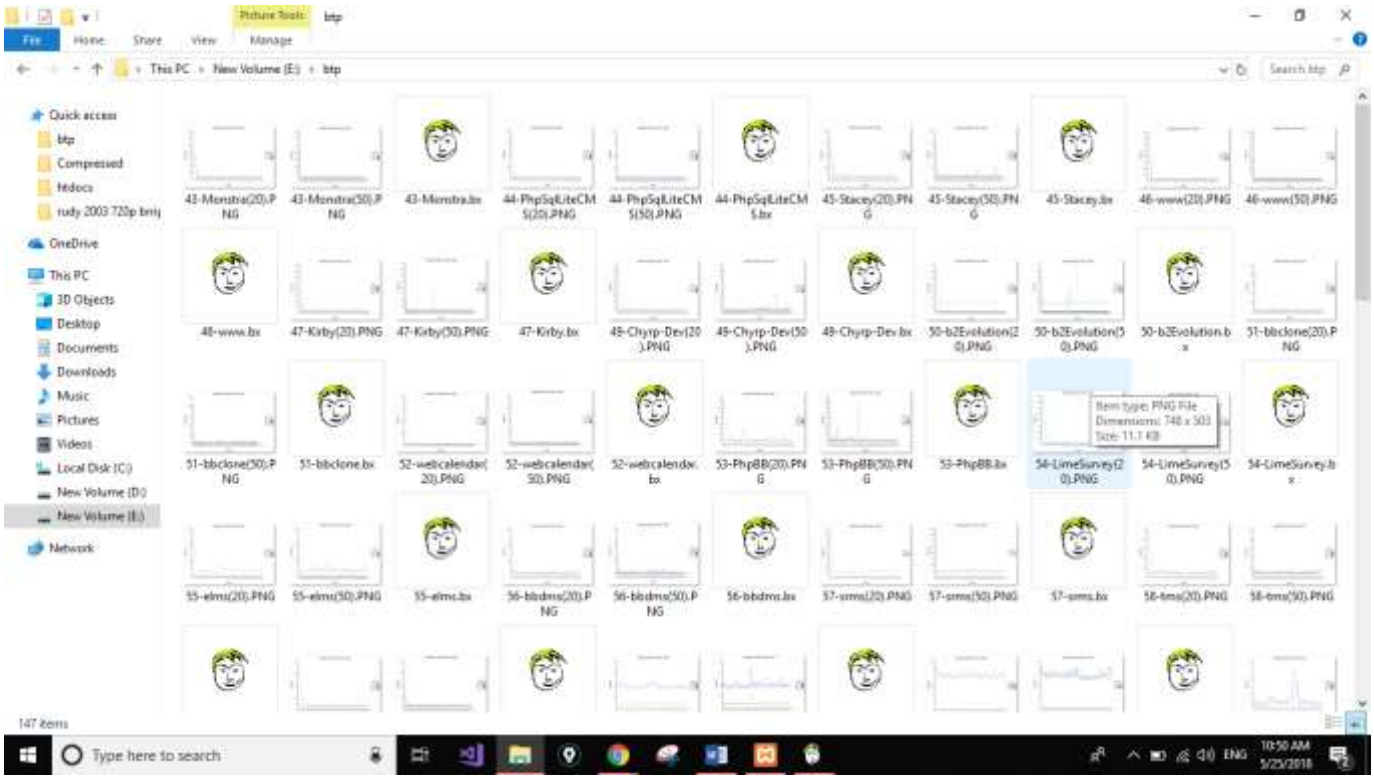

45-Stacey.bx - Badboy - (recording)
File Edit View Tools Preferences Help
http://localhost/monstra-3.0.4/index.php
Script
Test Suite 1
Test 1
Step 1 (Repeat 30 times)
http://localhost/stacey-master/
/stacey-master/projects/project-1/
/stacey-master/about/
Summary
Variables
Graph
Tools
Checks
References
Summary
Played 50
Succeeded 50
Failed 0
Avg Time (ms) 154
Assertions 0
Warnings 0
Timeouts 0
Max Time (ms) 1665
Done
Recording
btp
Home
Welcome!
Welcome to your new Monstra powered website.
Monstra is succesfully installed, you can start editing the content and customising your site.
Getting Started
This is a default home page of your website.
Here's a quick description of how to edit this page.
First make sure you're logged in
Go to the Pages Manager and click "Edit" button for this page
Make your changes, click "Save" and you're done!
Online Resources
Official Site
Type here to search
ENG
10:39 AM
5/25/2018

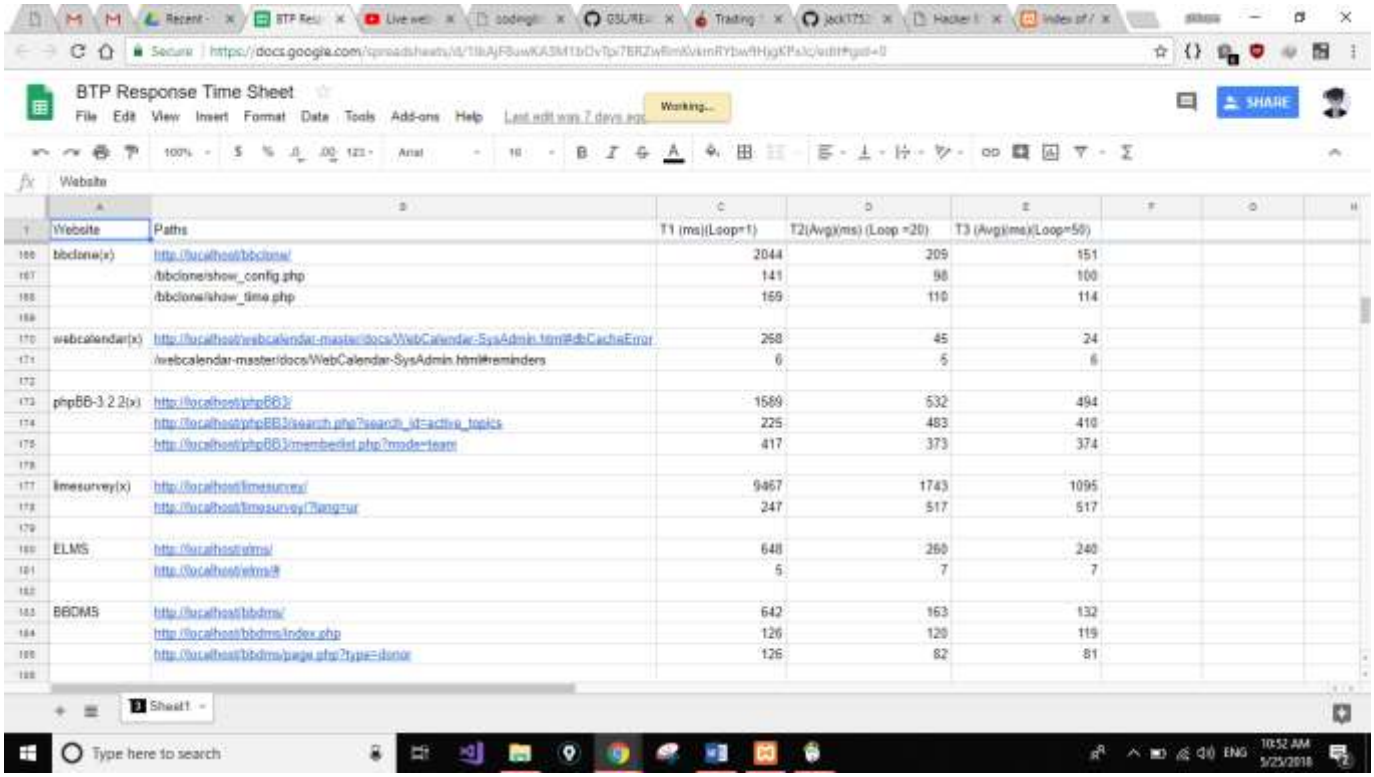
In the sixth week, we were told to collect the data of 50 websites . At this time Vivek was not present in LNMIIT and Shivam was struggling to collect the data alone , and thus we mailed for time extension. But somehow it took us 2 weeks to collect the data.





In the eighth week , we were asked to collect data for 100 websites and were given full time until the semester ends to complete the work. We struggled really hard during this time , but somehow in the end we were able to fetch the data along with the graphs.





BTP Response Time Sheet

Website	Paths	T1 (ms)(Loop=1)	T2(Avg)(ms) (Loop =20)	T3 (Avg)(ms)(Loop=50)
bbclone(x)	http://localhost/bbclone/	2044	209	151
	/bbclone/show_config.php	141	98	100
	/bbclone/show_time.php	169	110	114
webcalendar(x)	http://localhost/webcalendar-master/docs/WebCalendar-SysAdmin.html#dbCacheError	268	45	24
	/webcalendar-master/docs/WebCalendar-SysAdmin.html#reminders	6	5	6
phpBB-3.2.2(x)	http://localhost/phpBB3/	1589	532	494
	http://localhost/phpBB3/search.php?search_id=active_topics	225	483	410
	http://localhost/phpBB3/memberlist.php?mode=team	417	373	374
limesurvey(x)	http://localhost/limesurvey/	9467	1743	1095
	http://localhost/limesurvey/?lang=ur	247	517	517
ELMS	http://localhost/elms/	648	260	240
	http://localhost/elms/#	5	7	7
BBDMS	http://localhost/bbdms/	642	163	132
	http://localhost/bbdms/index.php	126	120	119
	http://localhost/bbdms/page.php?type=donor	126	82	81

Chapter 4

Prediction Of Response Time

4.1 Which Algorithm?

For prediction of response time we search for a Machine Learning algorithm and came up with Regression. Regression is part of Supervised Machine Learning algorithms. It uses the target and input properties for prediction of the results. The algorithm first trains with the input features provided in the input dataset. Then it constructs a model on the properties of training dataset and predicts it using the model. In Machine Learning according to "No Free Lunch", there is no perfect algorithm so we have to use "Trail And Error" technique. We used trial and error on many other Regression algorithms like Multi Variable Linear Regression which based on multiple independent input features, Polynomial Regression based on non-linear combination of input features etc. According to the size of dataset and heuristic analysis the best algorithm for us was SVR(Support Vector Regression).

4.2 What is SVR?

SVR(Support Vector Regression) uses similar principles as SVM(Support Vector Machines). The both algorithms try to make a model that minimizes the error. The SVM is used for Classification on the other hand SVR is used for Regression to predict one of the many infinite values. It is one of the regression algorithms in which we try to minimize the error rate, keeping in mind that the error is in range and hence tolerable.

There can be two types of SVR on basis of Linearity –

1. Linear SVR
2. Non-Linear SVR

4.3 Types Of SVR?

On the basis of linearity, we chose non-linear SVR because out target value is non-linear for the dataset. There can be many kernels for a SVR.

Types of Kernels –

1. linear
2. poly
3. rbf
4. sigmoid
5. precomputed
6. a callable

We removed “linear” kernel from the list because by plotting the dataset we knew that it would not be linear. We used trial and error method on the remaining kernels to find the kernel with most accuracy. The best fit was to use “rbf” kernel. Radial Basis Function kernel usually makes good default kernel too. As suggested by the experts on the internet we can use an automated version by minimizing the model selection by the use of Nerd-Mead method. Although there is always a risk of overfitting in the model selection. It has also been advised to use grid selection along with cross validation. We didn’t do any of these because risks were more then the actually coding work that we have to pay for it. Although we have put the links to these papers in the bibliography.

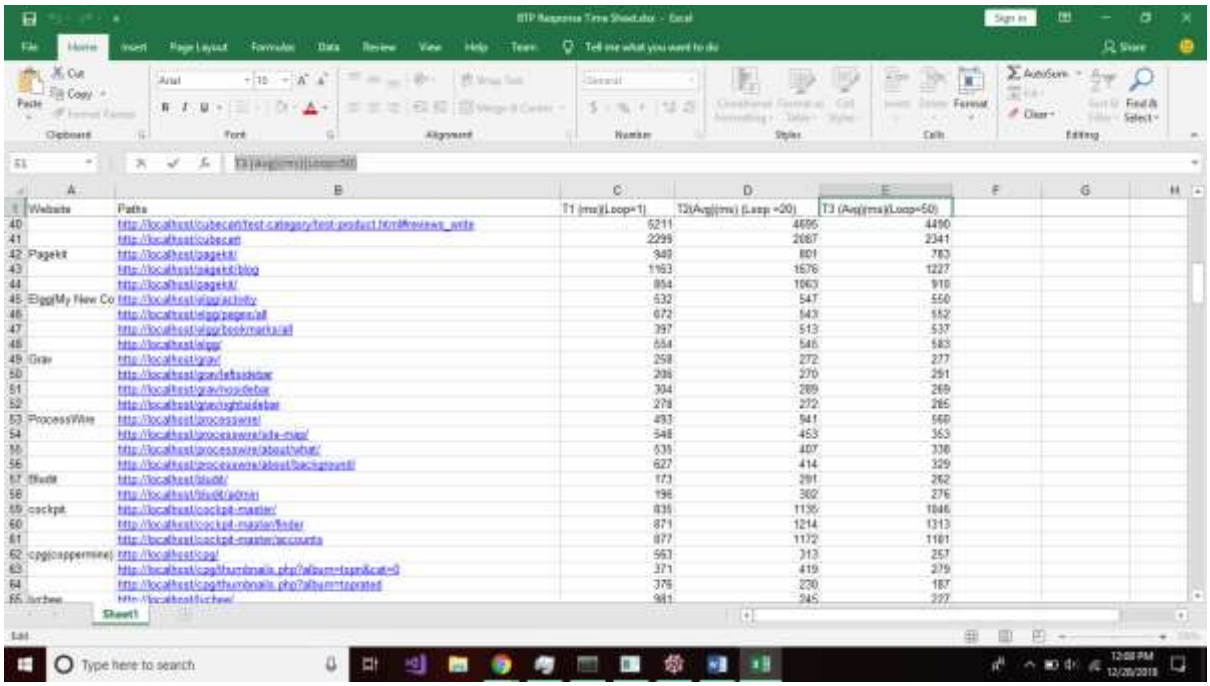
Chapter 5

Our Work

5.1 Why and How to use SVR?

SVR is used when we have less data points to train on usually less then 1000 points. So we first made a model on SVR and then we used repeated K Fold Cross Validation technique on data set to train it well.

We used the two columns in the data set as the features which are “T1 (ms)(Loop=1)” and “T2(Avg)(ms) (Loop =20)” as features and the third and last column as target which is “T3 (Avg)(ms)(Loop=50)”. We then trained the model along with K Fold Cross Validation.



```
# -*- coding: utf-8 -*-
#SVR

#Importing Libraries
import numpy as np
import matplotlib.pyplot as plt
import pandas as pd
from sklearn.svm import SVR
from sklearn.metrics import mean_squared_error
from math import sqrt

```

```

file='BTP Response Time Sheet.xlsx'
data = pd.ExcelFile(file)

data = data.parse('Sheet1')

data = data.fillna(value=0)

urls = range(len(data['Paths']))

X = np.array(data[data.keys()[2:4]].dropna())
y = np.array(data[data.keys()[-1]].dropna())

```

5.2 Repeated K Folds Cross Validation

We used 5 Fold Cross Validation on the dataset, we decided the value of K according to the number of features present in the data set. Our dataset had 2 features so we used 5 fold cross validation as we wanted low bias. Cross validation is an important technique which trains the model well on the data. The training results in less variance and low bias in the model.

Important points to keep in mind while using –

1. The lower the value of K the more is the bias. More the value of K the less is the bias.

2. Drawback of choosing large value of K is that model can fall pray to large variability.

3. We used 5 splits with 5 random states to initialize each time.

4. We trained the model on the 4 dataset each time and validated on the 5th data set.

5. Choosing the validation set each time for the 5th dataset as different.

```
54
55 # Do repeated K Fold Cross Validation.
56 from sklearn.model_selection import RepeatedKFold
57 rkf = RepeatedKFold(n_splits=5, n_repeats=10, random_state=None)
58
59 trmse=0
60 t=0
61
62 for train_index, test_index in rkf.split(X):
63       #print("Train:", train_index, "Validation:",test_index)
64       X_train, X_test = X[train_index], X[test_index]
65       y_train, y_test = y[train_index], y[test_index]
66
67       #Fitting SVR to the dataset
68       regressor=SVR(kernel='rbf')
69       regressor.fit(X_train,y_train)
70
71       #Predicting a new result
72       y_pred = regressor.predict(X_test) #value
73
74       #Calculating Root Mean Squared Error(RMSE).
75       rms = sqrt(mean_squared_error(y_test, y_pred))
76       #print(rms)
77       trmse+=rms
78       t+=1
79
80 print("So average Root Mean Square Error Of the Model Is ")
81 print(trmse/t)
82
```

5.3 Root Mean Square Error

We found the root mean squared error using the average of all the RMSE values of the model while iterating. We used the validation split of the data set and took out RMSE for each split. Then we took the average and it. The lower the value of RMSE the better is the model. We then found the percentage of the RMSE which came out to be 22%.

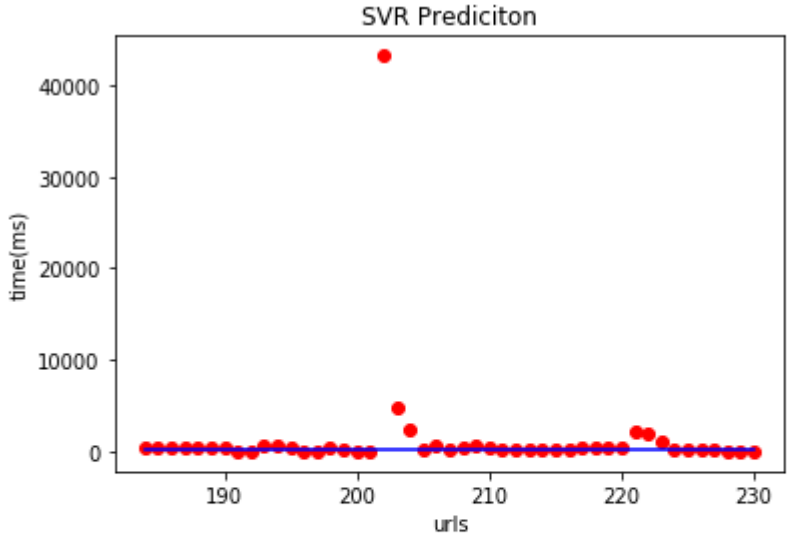
```
49 #Calculating Root Mean Squared Error(RMSE).
50
51 initial_rms = sqrt(mean_squared_error(y_test, y_pred))
52 print(initial_rms)
53
```

5.4 Final Model Testing and Plotting

We used the 20% of the data set for testing which is the 47 remaining urls. We used matplotlib for plotting the scatter plot. We used the X axis as the number of urls and Y axis as the time in milli seconds. We used the red color for the actual value of the time in ms and blue is the predicted value using our trained model.

```
41 #Fitting SVR to the dataset
42
43 regressor=SVR(kernel='rbf')
44 regressor.fit(X_train,y_train)
45
46 #Predicting a new result
47 y_pred = regressor.predict(X_test) #value
48
```

```
84
85 #Visualising the SVR results
86 plt.scatter(urls_test,y_test,color='red')
87 plt.plot(urls_test,regressor.predict(X_test),color='blue')
88 plt.title('SVR Prediciton')
89 plt.xlabel('urls')
90 plt.ylabel('time(ms)')
91 plt.show()
92
```



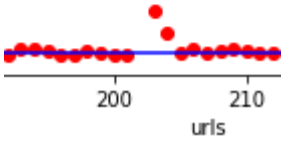

```
In [3]: len(y_train)
   ...:
Out[3]: 184

In [4]: len(y_test)
Out[4]: 47

In [5]:
```

5.5 Explaining The Outlier

The outlier came because the website was made like that. Going from one webpage to the other may be done by the developer of the webpage must be programmed badly. It can't be the mistake of the server that we were using because if it was then the red points which are its neighbors would also be high (45000ms) but they are low (50ms and 5000ms). As the disk usage and memory consumption by the CPU in the computer does not go down in 10-15 seconds or so. It takes a minute or so to go down.



5.6 What can go wrong

1. As the data was still less so our model may fall to prey to underfitting.

2. We hosted the websites on our computer with other programs working and the performance of the server may be different at different points of time.

Chapter 6

Comparison

Comparison with a Research Paper

As we wanted to make the best of the best project thus we compared our model with the best model made yet in this field of research of response time prediction. We have mentioned the link to the paper in the bibliography.

1. The other research paper we studied included the network bandwidth also in the response time calculation which we did not as we hosted it on our own server so we get better results in calculations.

2. They had more data then us so the prediction made by their SVR model can have low bias and low variance as compared to our model but the accuracy score of our model and their model are approximately same 78%.

Chapter 7

Conclusions & Results

We were able to get Response Time data of 100 websites. All websites from different domains item selling websites, mailing websites, software application websites etc. Taking websites from different domain fetched us a variety of data. We have 2 graphs for each of these 100 websites. We learnt a lot about Php errors. The security of websites related to databases. We even wrote the issues on Github based on the website. We even read a lot of websites documentations. We made different users in MyPhpAdmin for different websites. We learnt how to create databases and also import databases for different websites. In the end we actually figured out why our laptops were slowing down as we take out more data. We gained a profound knowledge of XAMPP and WAMP and how to use them. We learnt a lot about Response time from Research papers.

7.1 Further Scope

We will try to make a model so as to predict the response time of any sample website. We will try to predict the response time of big websites like Amazon, Facebook etc. We will try hard to publish a paper related to the data gathered and the model made by us to predict response time.

We can use ensemble technique and test it on different regression algorithms to improve on our accuracy score. We can use bagging and boosting technique too to increase our model accuracy.

We can even collect more data to test our model on to get low bias and low variance thus improving our model.

Bibliography

1. *Measuring website's response time –* https://www.websitepulse.com/blog/how-to-measure-website-response-time

2. *Average response time -*

   https://docs.oracle.com/cd/E19316-01/820-4342/abfch/index.html

3. *Difference between Response time and Page Load time -*

   https://help.pingdom.com/hc/en-us/articles/115001228925-Difference-between-Response-Time-and-Page-Load-Time

4. *Uptime report calculations –* https://help.pingdom.com/hc/en-us/articles/211847325-Response-time-and-calculations-in-the-uptime-report

5. *Prediction of website response time based on support vector machine –* https://ieeexplore.ieee.org/abstract/document/7003908/

6. *Website's speed (need for speed 1997 article) -* https://www.nngroup.com/articles/website-response-times/

7. *Nielson Norman Group Article –* https://www.nngroup.com/articles/response-times-3-important-limits/

8. *20 Factors Influencing website's response time –* http://www.apmdigest.com/website-response-time-1

9. *What is a good response time ?* https://stackoverflow.com/questions/164175/what-is-considered-a-good-response-time-for-a-dynamic-personalized-web-applicat

10. Support Vector Regression Or SVR.
https://medium.com/coinmonks/support-vector-regression-or-svr-8eb3acf6d0ff

11. Radial Basis Function Kernel
https://en.wikipedia.org/wiki/Radial_basis_function_kernel

12. Top 6 Regression Algorithms
https://www.analyticsindiamag.com/top-6-regression-algorithms-used-data-mining-applications-industry/

13. Math Of RBF Kernel
https://calculatedcontent.com/2012/02/06/kernels_part_1/

14. Difference between SVM Linear, polynomial and RBF kernel?
https://www.researchgate.net/post/Diffference_between_SVM_Linear_polynmial_and_RBF_kernel

15. Difference between SVM and SVR in implementation.
https://stats.stackexchange.com/questions/198199/how-different-is-support-vector-regression-compared-to-svm

16. Implementation of SVR.
https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVR.html

17. All kinds of SVM algorithms
https://www.quora.com/How-many-kinds-of-SVM-algorithms-exist

18. 15 types of regression
https://www.r-bloggers.com/15-types-of-regression-you-should-know/

19. How to decide kernel for SVM?
https://stats.stackexchange.com/questions/18030/how-to-select-kernel-for-svm

20. Kernel functions which give better results in support vector regression?
https://www.researchgate.net/post/Can_anyone_tell_me_how_to_decide_which_kernel_function_gives_better_results_in_Support_Vector_Regression

21. Different Cross Validation and their implementation
https://scikit-learn.org/stable/modules/cross_validation.html

22. Steps from making model to final evaluation.
https://scikit-learn.org/stable/modules/cross_validation.html

23. Improving model performance using cross validation
https://www.analyticsvidhya.com/blog/2018/05/improve-model-performance-cross-validation-in-python-r/

24. Choosing values of epsilon, C and gamma
https://medium.com/@univprofblog1/support-vector-regression-matlab-r-and-python-codes-all-you-have-to-do-is-preparing-data-set-1d8e4333f831

25. RMSE
https://stackoverflow.com/questions/17197492/root-mean-square-error-in-python

26. Choosing model after Cross Validation
https://stats.stackexchange.com/questions/52274/how-to-choose-a-predictive-model-after-k-fold-cross-validation