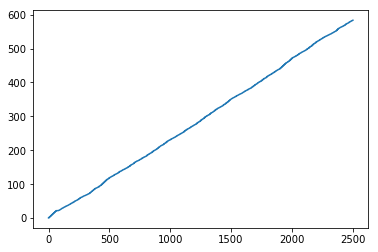
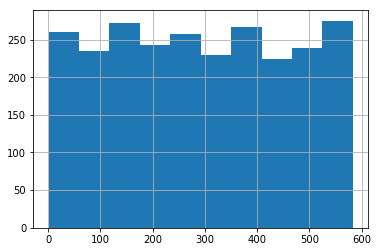
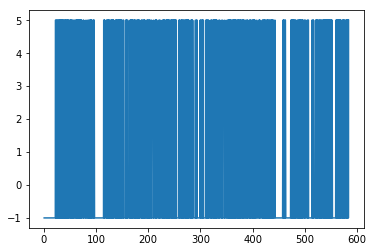
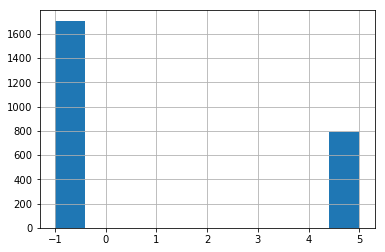
**Capsule8 Data Science Challenge**

**Data Wrangling and Exploration**

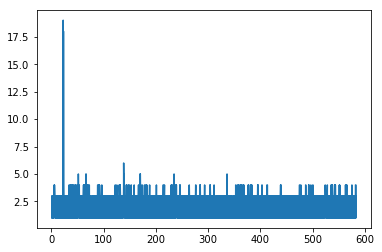
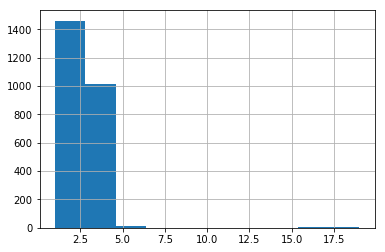
1. Plot & histogram for ‘ts’

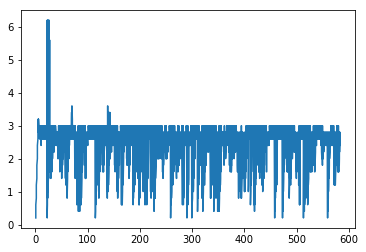
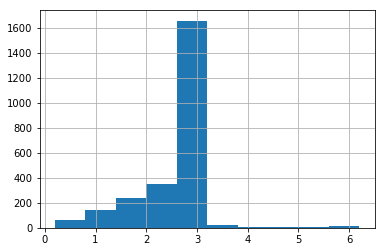
Time-series plot & Histogram for depth

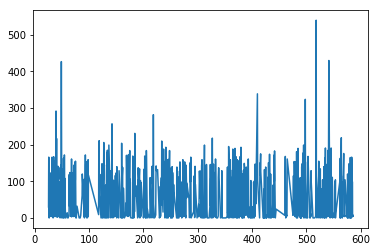
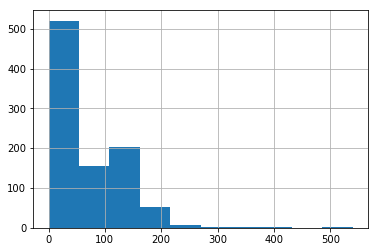
Time-series & histogram plot for rate\_1

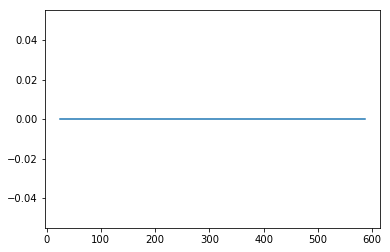
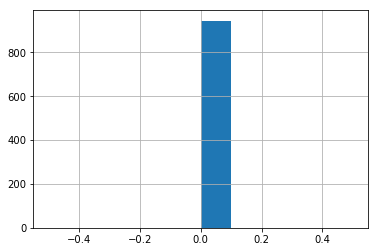
Time-series & Histogram plot for rate\_5

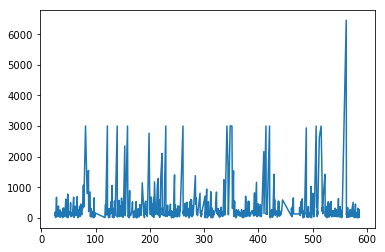
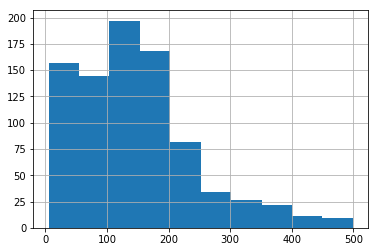
Time- series & Histogram plot for Rx

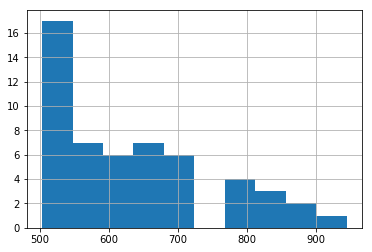
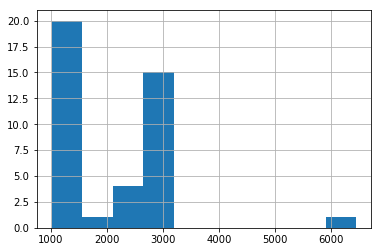
 

Time-series & Histogram for tx

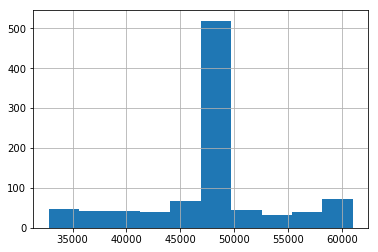
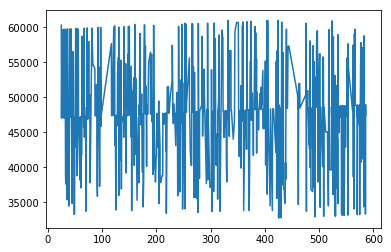
Time-series & Histogram for Dur

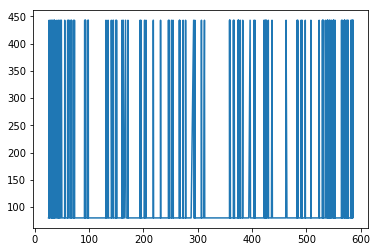
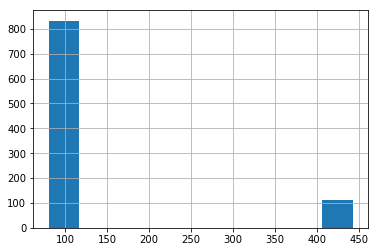
 

I broke the distribution of ‘dur’ as follows 0 – 500, 500 – 1000 & 1000 – 6000 and plotted the histogram. The distribution is long right tailed distribution and the no. of breaks in the distribution is 2.

Time-series & Histogram for lport

Time – series & Histogram for Rport

**Data Statistics**

df2['dur'].describe()

count 944.000000

mean 261.219121

std 479.331879

min 5.040000

25% 85.512500

50% 144.650000

75% 226.797500

max 6454.020000

The mean is centered on 260 and median is 145 which tells us it skewed right a great deal and there is the presence of outliers in it which we can even infer from the histogram plots of ‘dur’.

I used the concept of Inter-Quartile ranges to

upper = 226.797500

lower = 85.512500

IQ = upper – lower #141.28500000000003

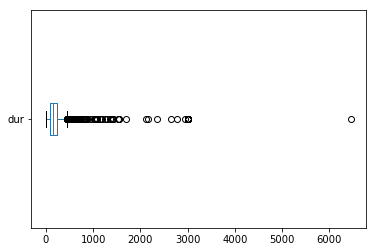
lower\_inner\_fence = lower - 1.5\*(IQ) # -126.41500000000003

upper\_inner\_fence = upper + 1.5\*(IQ) # 438.725

lower\_outer\_fence = lower - 3.0\*(IQ) # -338.3425000000001

upper\_outer\_fence = upper + 3.0\*(IQ) # 650.6525000000001

From the IQ ranges, values beyond the 438.725 are considered mild outliers whereas values beyond 650.6525 are considered extreme outliers. I plotted a box plot for the column ‘dur’ to see this.



sum(df2['dur']>650)

Out: 63

There are total 63 values that are outliers in the ‘dur’ column and these duration values can be different from the durations of the regular Linux processes that take place in the system. They can be a breach in the security or different processes that are taking place in the system.

**Data Modelling**

**Prediction of ‘Dur’**

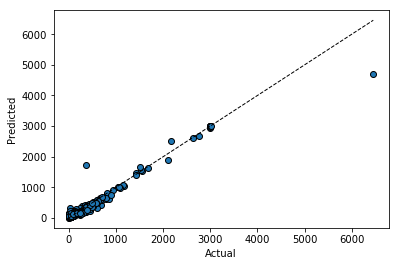
Random Forest Algorithm

Random Forest Algorithm is an ensemble method where each classifier individually a ‘weak learner’ but when all classifiers are taken together they are strong learners. The random forest is based on decision trees, in a decision tree an input is entered at the top and it traverses down the tree. Here our tree is the weak learner and by the notion of ensemble we put together the trees to get a forest of trees or in other words a random forest which is our strong learner. When an input enters a random forest, it traverses down all the trees, the result may be an average or weighted average of all the terminal nodes it reaches or in case of categorical variables voting is done.

I chose this algorithm since it is a rule based technique and it would be good for my prediction.

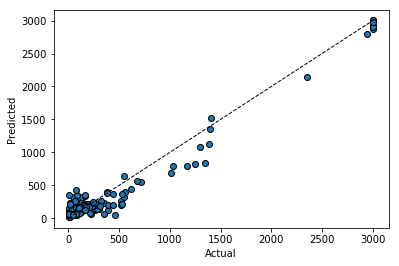
Metrics:

In-sample Plot



The model is fitting well on the training data points this means that the model is working well in predicting ‘y’

Out-sample Plot



In the testing set, the model can differentiate between predicting lower and higher values which also tells us that the model is fitting well with the training set as well.

MAPE (Mean Absolute Percentage Error)

In-sample MAPE: 23.616331968039106 ~ 23%

Out-sample MAPE: 40.601990851939846 ~ 40%

**Extra Credit**

The data in the ‘exec’ represent Linux processes and processes that take place on docker containers and the data in ‘tcplife’ talks about port connections(VPN) of the process that is taking place. Lport (local port) is the listening port and rport is the remote port. ‘rx’ is the file size. ‘443 ‘ is HTTPS port & ‘80’ is HTTP port. Duration is the time taken for each process. Historyport, Histosize and histotimes are histories that contain min and max functions of the columns rport, rx and duration. Prediction & Predictquality tells us if there is a breach in security of the docker containers.

**Questions**

**Prediction of ‘dur’**

I have selected the model to predict ‘dur’. But, I need features for the next 25 rows in tcp\_life to predict values of duration. Can you kindly give me the values for the features in tcp\_life for the next 25 rows without ‘dur’ & ‘lport’.

**Prediction of ‘lport’**

My lport has 919 unique classes out of the 944 rows and so the frequency of each class is also less. Can you kindly tell me how do I approach this problem in predicting lport for the next 25 values.