IC272 project

Batch 8

Thursday Batch

Agenda

The goal of this project is to do descriptive analysis to understand and infer from data of performance of BNS device and then performing the regressive analysis on predicting the "***InBandwidth***" of the device using different regression techniques.

What did we do with the data ?

1. Import all the required libraries

For reference these were the libraries that we used for our data analysis:

Statistics, pandas, numpy, scipy.stats, sklearn.decomposition, matplotlib.pyplot, sklearn, sklearn.model_selection, math, sklearn.linear_model, sklearn.preprocessing, sklearn.metrics

2. Data Cleaning

- Checked for any missing values. *(there were none in our data)*
- Outlier detection *(there were 2178 outliers in our data)*
- We replaced the outliers with median *(after replacing 545)*
- Calculated the statistics of data *(mean,median etc.)*

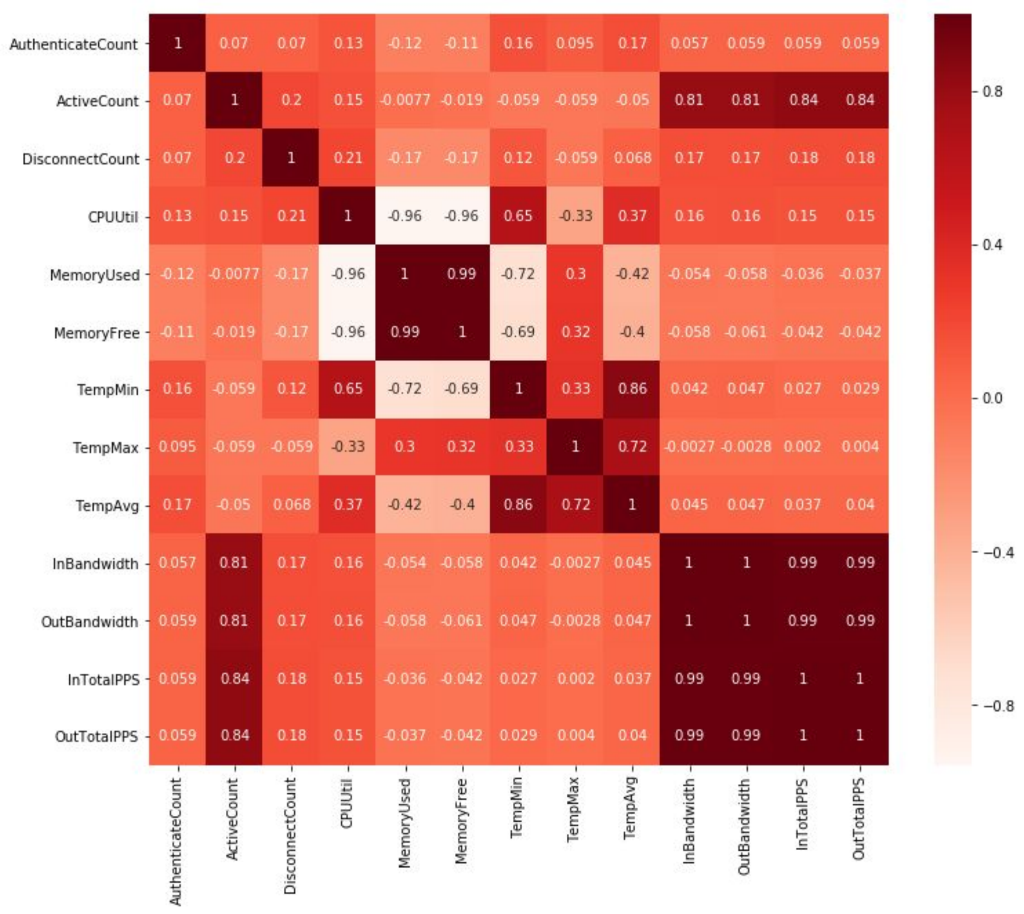
3. Data Preprocessing

The data is preprocessed by various methods like:

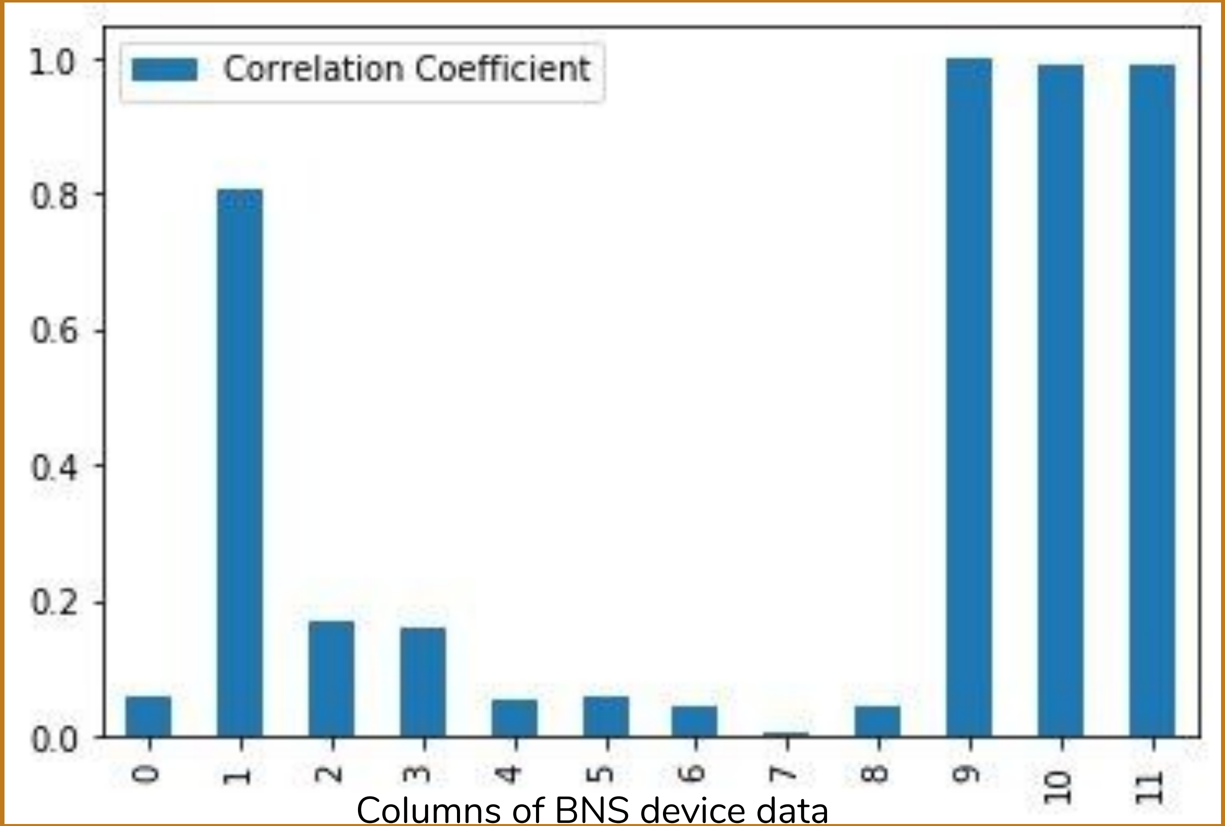
- ➢ Normalisation
- ➢ Standardisation
- ➢ Feature Selection
- ➢ PCA.

INFERENCES AND RESULTS FROM ANALYSIS

Correlation plot of each attributes

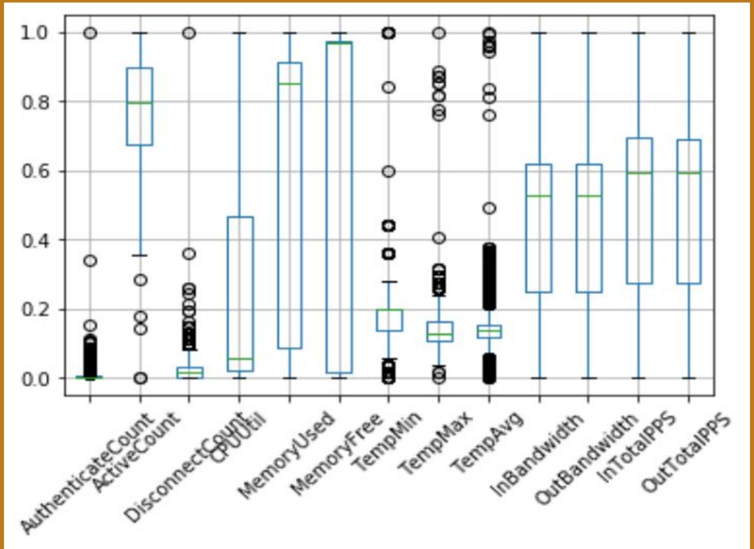


The inbandwidth is highly correlated to the OutBandwidth, InTotalPPS, OutTotalPPS, ActiveCount attributes.
Part of the reason can be because the number of active users determine the InBandwidth used for data transmission, more users would need more data bandwidth.

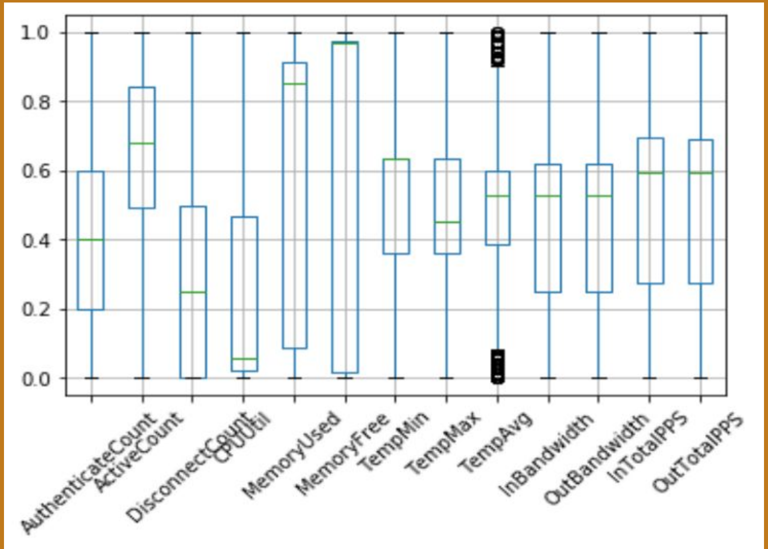

Correlation Coefficient
1.0
0.8
0.6
0.4
0.2
0.0
0
1
2
3
4
5
6
7
8
9
10
11
Columns of BNS device data

Box-plot of outliers

Actual normalised data with outliers
Total outliers: 2178



After outlier-removal with median the normalised data boxplot
Total outliers: 545



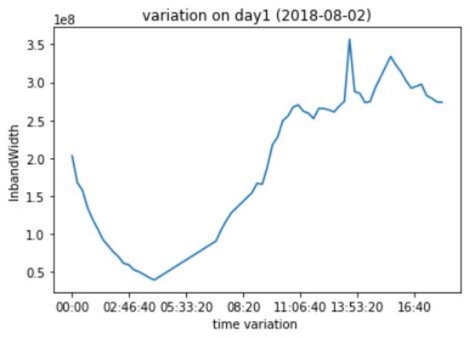
Results seen from box-plots

After replacing outliers with median more data comes under the IQR, so more data can be used for proper data analysis.

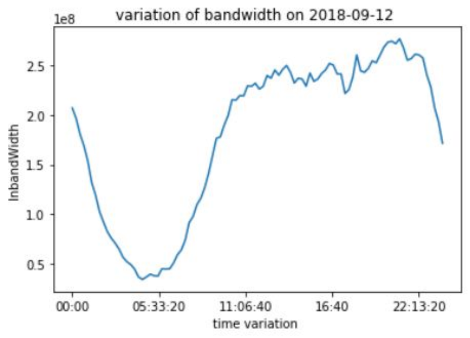
So a lot of data which would have been useless before can be of use now.

Target attribute variations on different months

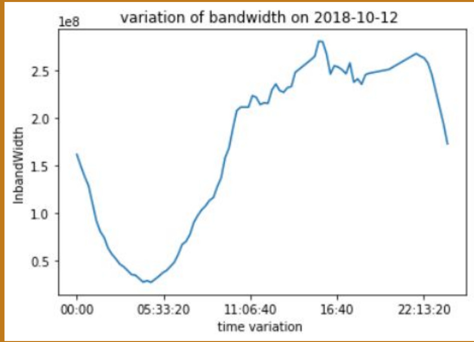
August



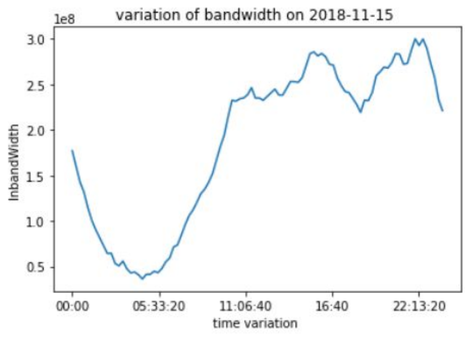
September



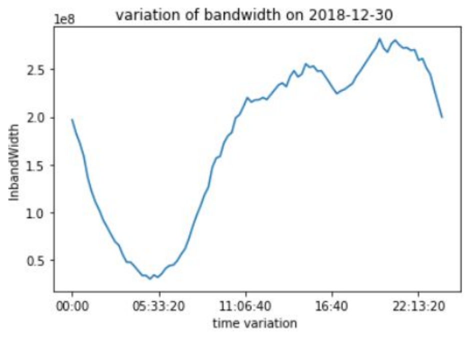
October



November



December



Inferences from variation of Inbandwidth with time

We see that when the device start working on a particular day the inbandwidth of the device mainly decreases at first and then it increases to a maximum and then decreases.

The reason for this is that the Inbandwidth depends on the number of active users,input packets at that time so at the no. of users connected to the device is less and after sometimes it increases so inbandwidth also increases.

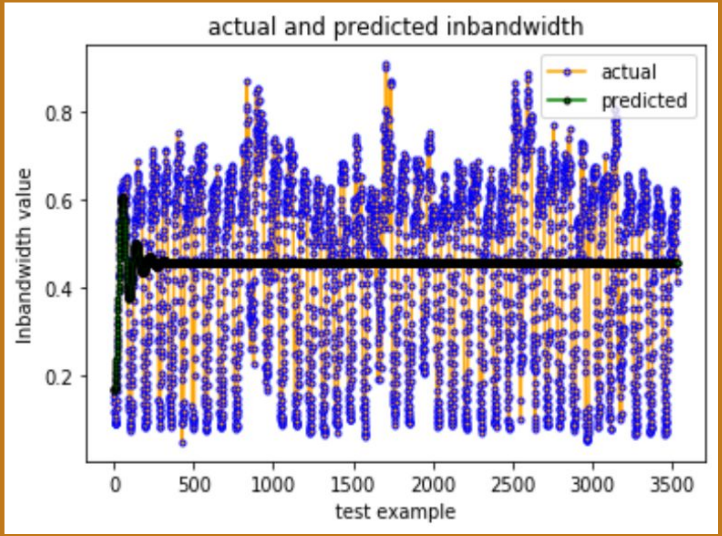
Auto Regression

Here we have 70% data as training data and 30% as test data

The optimal lag value which we obtain is 36

R_squared_score we get is 0.02476

data	Rmse values	R2-score
Actual data	95353821.3	0.0247
Normalise data	0.2099	0.0247
Standardised data	0.9506	0.0247

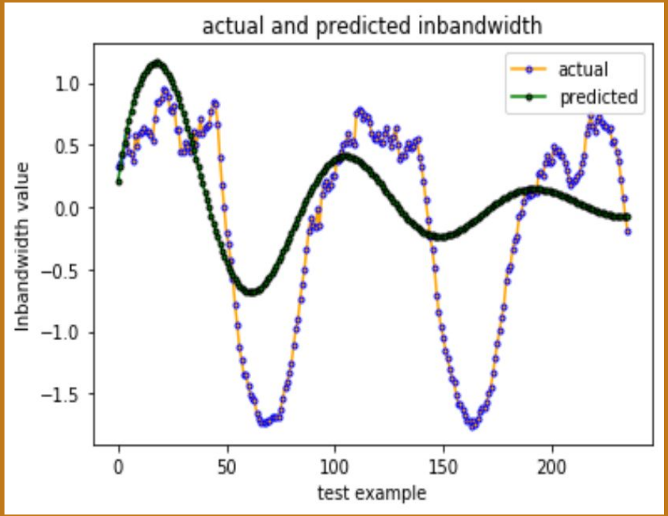


Auto-regression

In this case we take last 250 data as test data and predict the values using autocorrelation analysis.

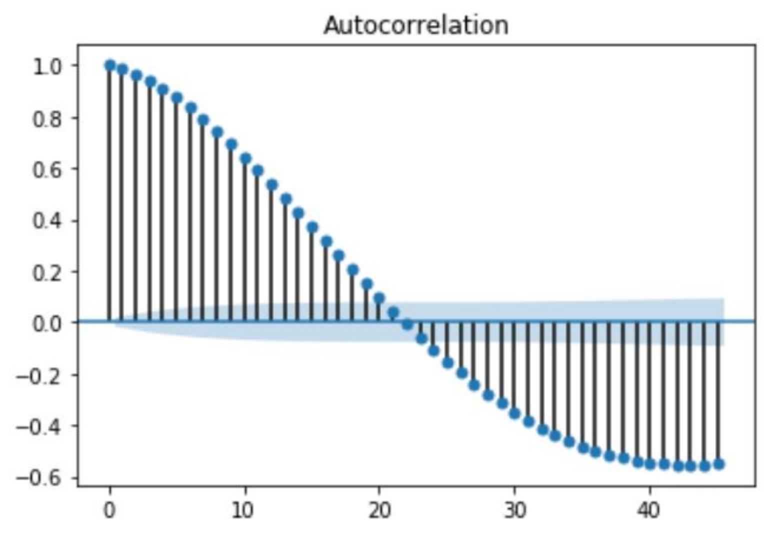
The optimal lag value which we obtain is 39.

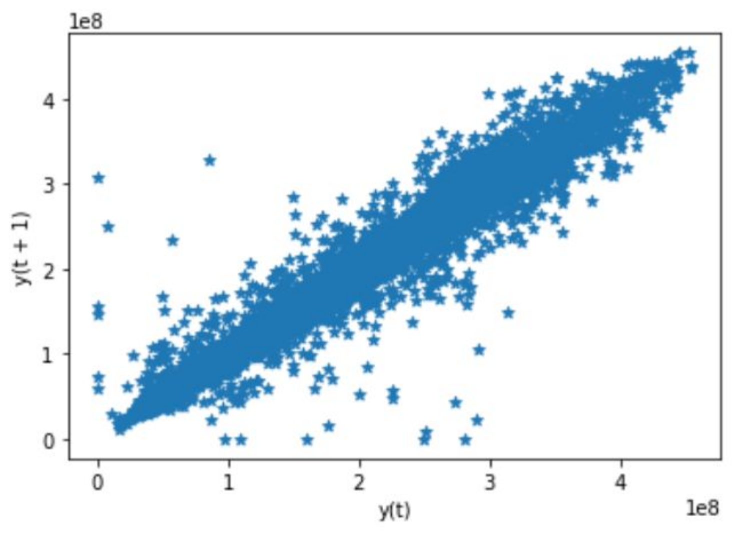
The rmse values obtained are:



data	RMSE values	R2-score
Actual data	69412366.069	0.33947
Normalised Data	0.15281	0.33947
Standardised Data	0.69204	0.33947

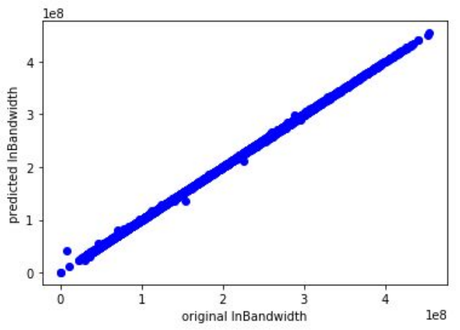
Auto-regression plot



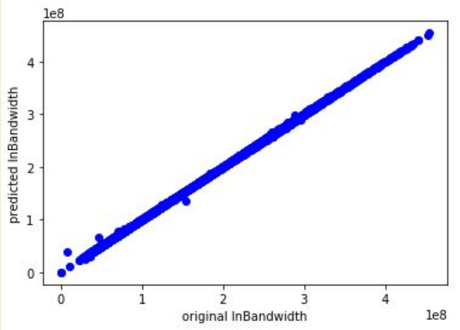


POLYNOMIAL CURVE FITTING (degree 2)

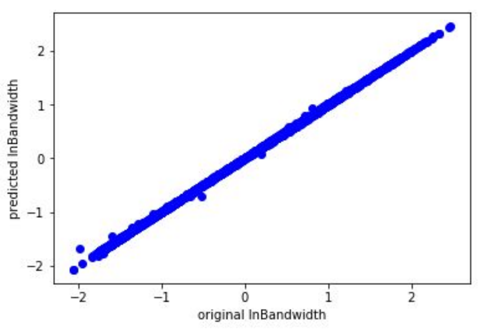
On untreated data



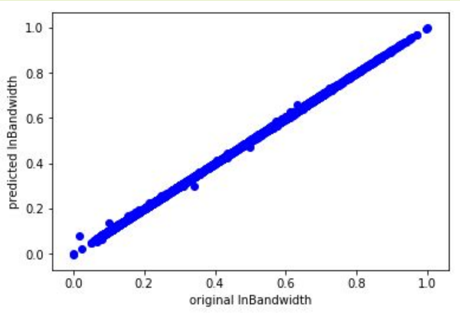
On data without ouliers



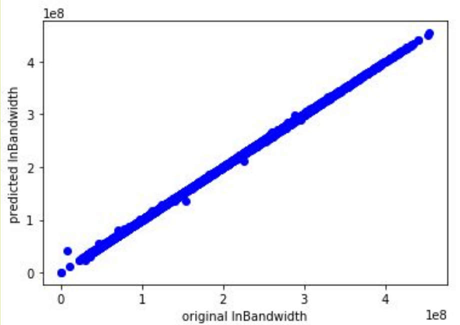
On standardised data



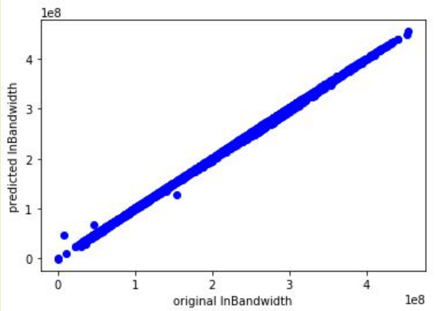
On normalised data



After feature selection

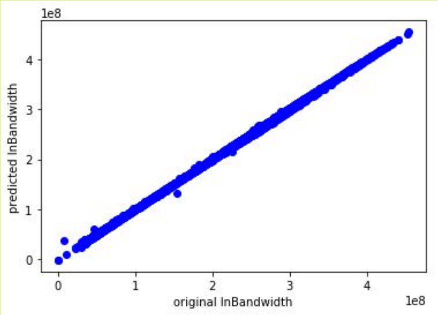


After PCA

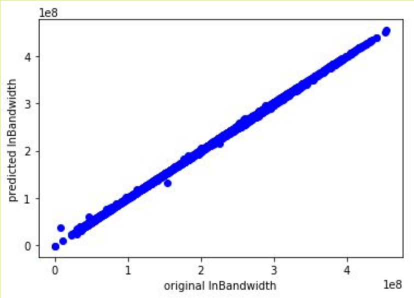


MULTIPLE LINEAR REGRESSION

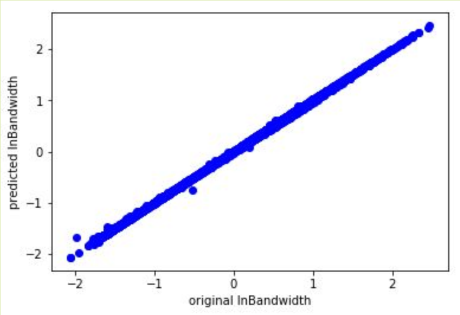
On untreated data



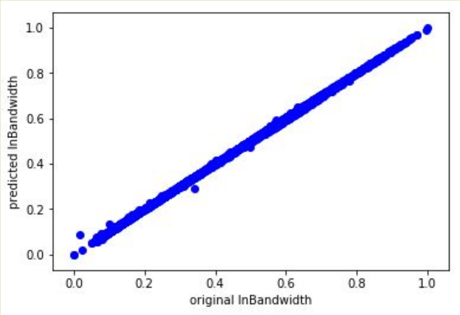
On data without ouliers



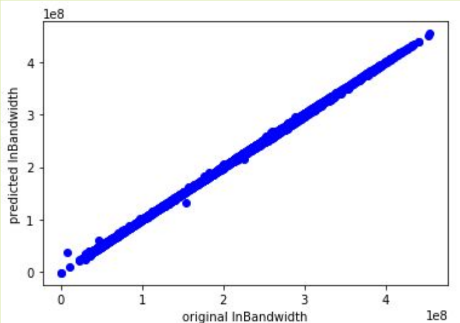
On standardised data



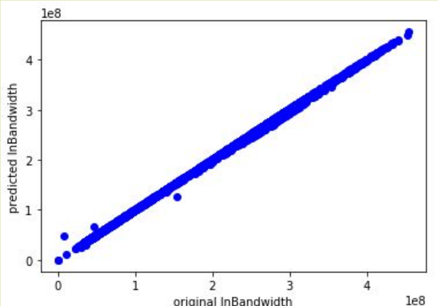
On normalised data



After feature selection

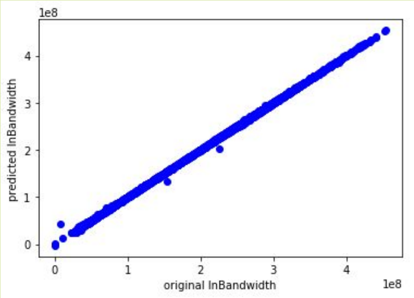


After PCA

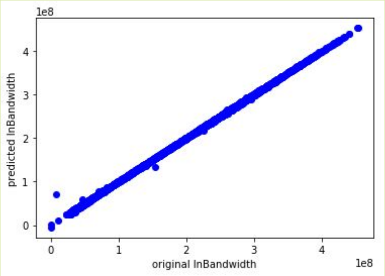


POLYNOMIAL CURVE FITTING (degree 3)

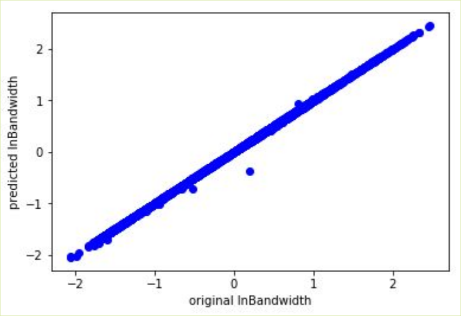
On untreated data



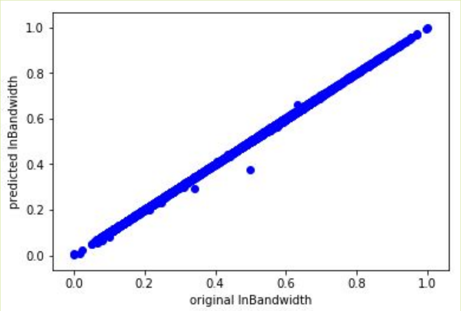
On data without ouliers



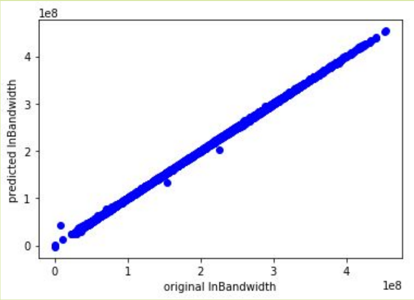
On standardised data



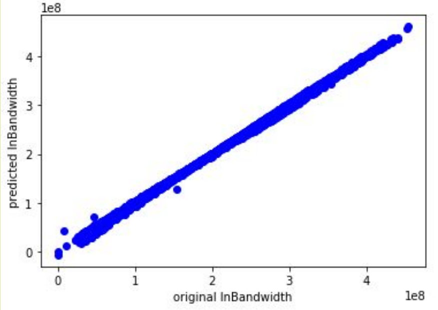
On normalised data



After feature selection

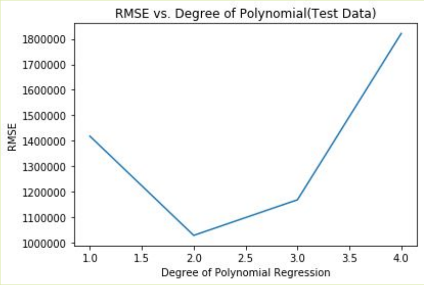


After PCA

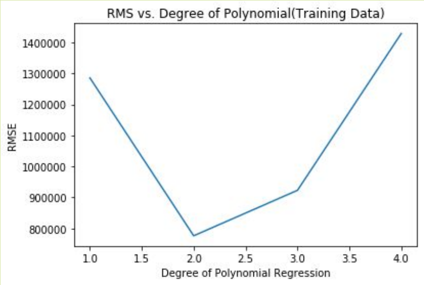


RMSE vs degree of polynomial

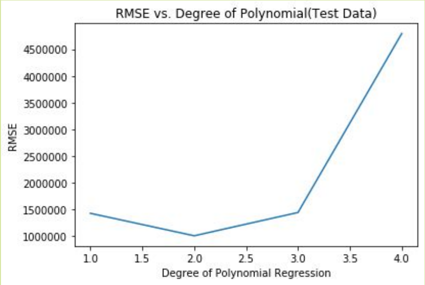
For feature selection



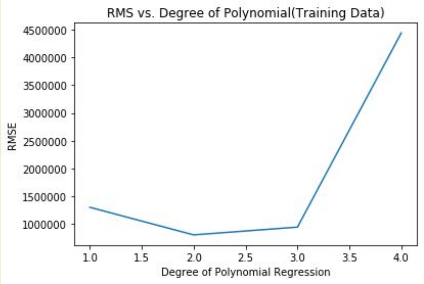
Feature selection



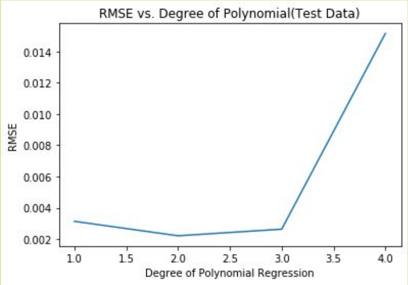
No outliers



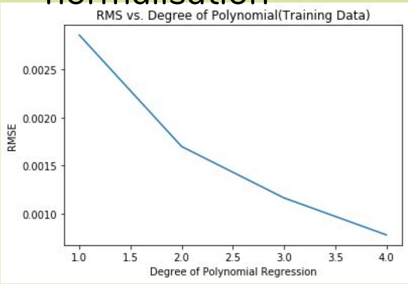
No outliers



normalisation

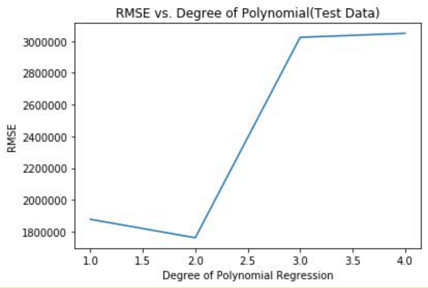


normalisation

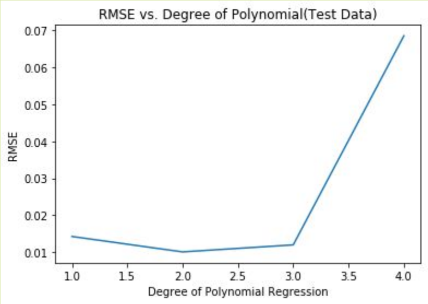


RMSE vs degree of polynomials

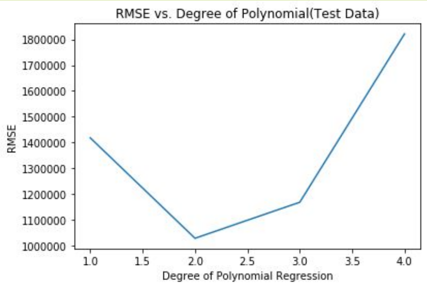
PCA Test



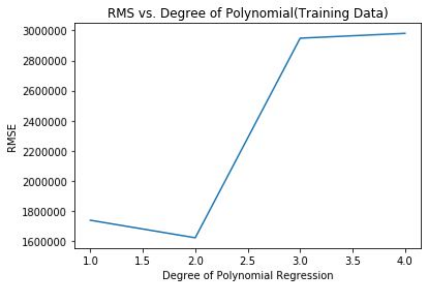
Normalisation Test



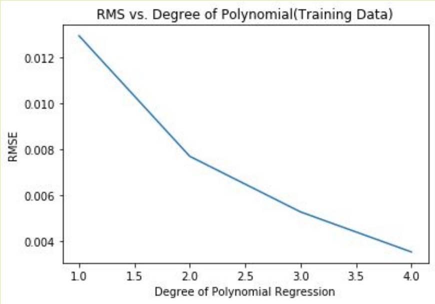
Original Data Test



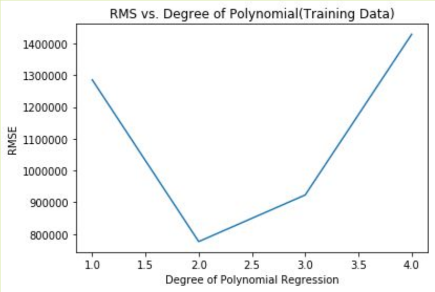
PCA Train



Normalisation Train



Original Data Train



RMSE of test data comparison table

data	Linear regression	Polynomial regression (best degree=2)	auto-regression
Preprocessed data	1426613.40	1005208.07	69412366.069
Normalised data	0.0031	0.0022	0.15281
Standardised data	0.0142	0.0100	0.69204

Polynomial regression is the best regression model among these as per given data.

R2_score of test data comparison table

data	Linear regression	Polynomial regression (best degree=2)	auto-regression
Preprocessed data	0.99979	0.99990	0.33947
Normalised data	0.99979	0.99990	0.33947
Standardised data	0.99979	0.99990	0.33947

Polynomial regression is the best regression model among these as per given data.

CONCLUSIONS

- Normalised data gave the better results.
- Polynomial curve fitting was better regressive analysis
- Feature Selection and PCA improved RMSE (but order is same), reduced processing time significantly.