IC-272 Mini Project Presentation

Group 6 - Thursday Batch

Our Data..

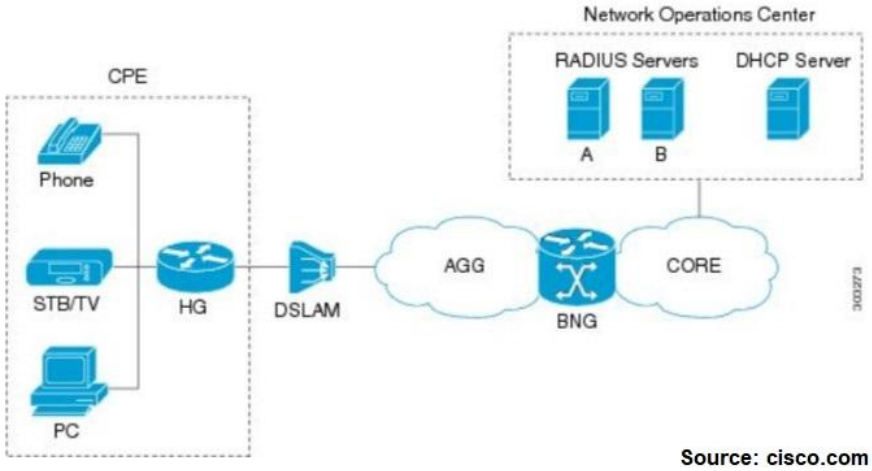
Performance record of a BNG device.

Tracked at intervals of 15 minutes. User count details, processor usage, memory usage, device temperature, bandwidth utilization and packet transmission details.

CSV file: 12385 tuples of data. Overall 14 attributes.

70:30 train-test ratio used.

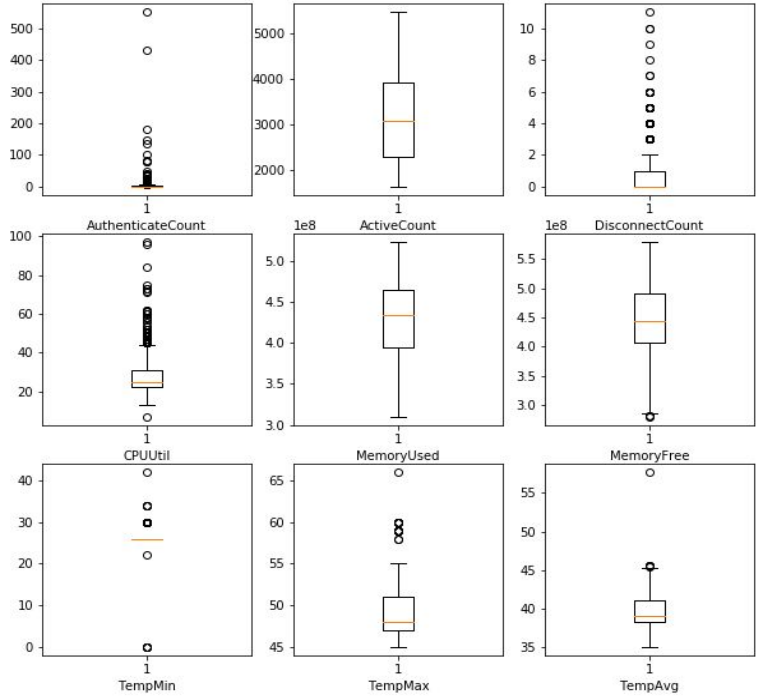
Figure 1: BNG Architecture

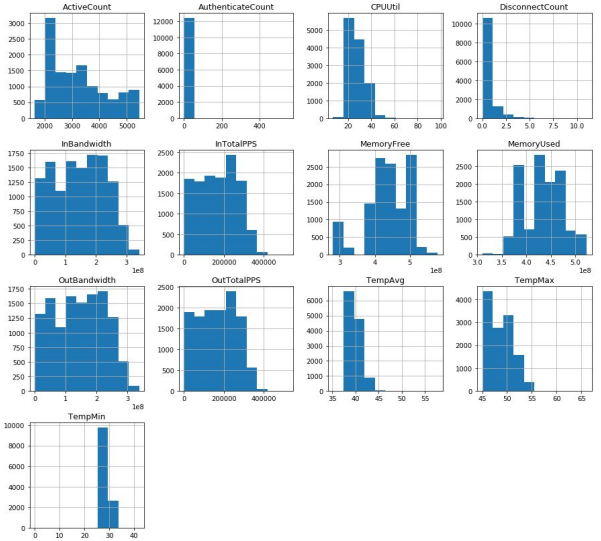


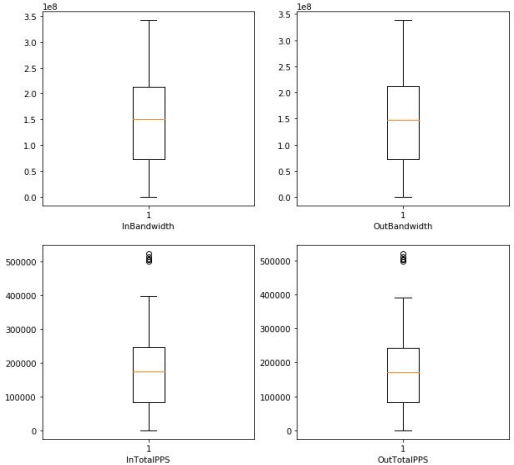
Source: cisco.com

Problem Statement: Predicting Memory Used attribute using various regression analysis techniques.

Descriptive Analysis



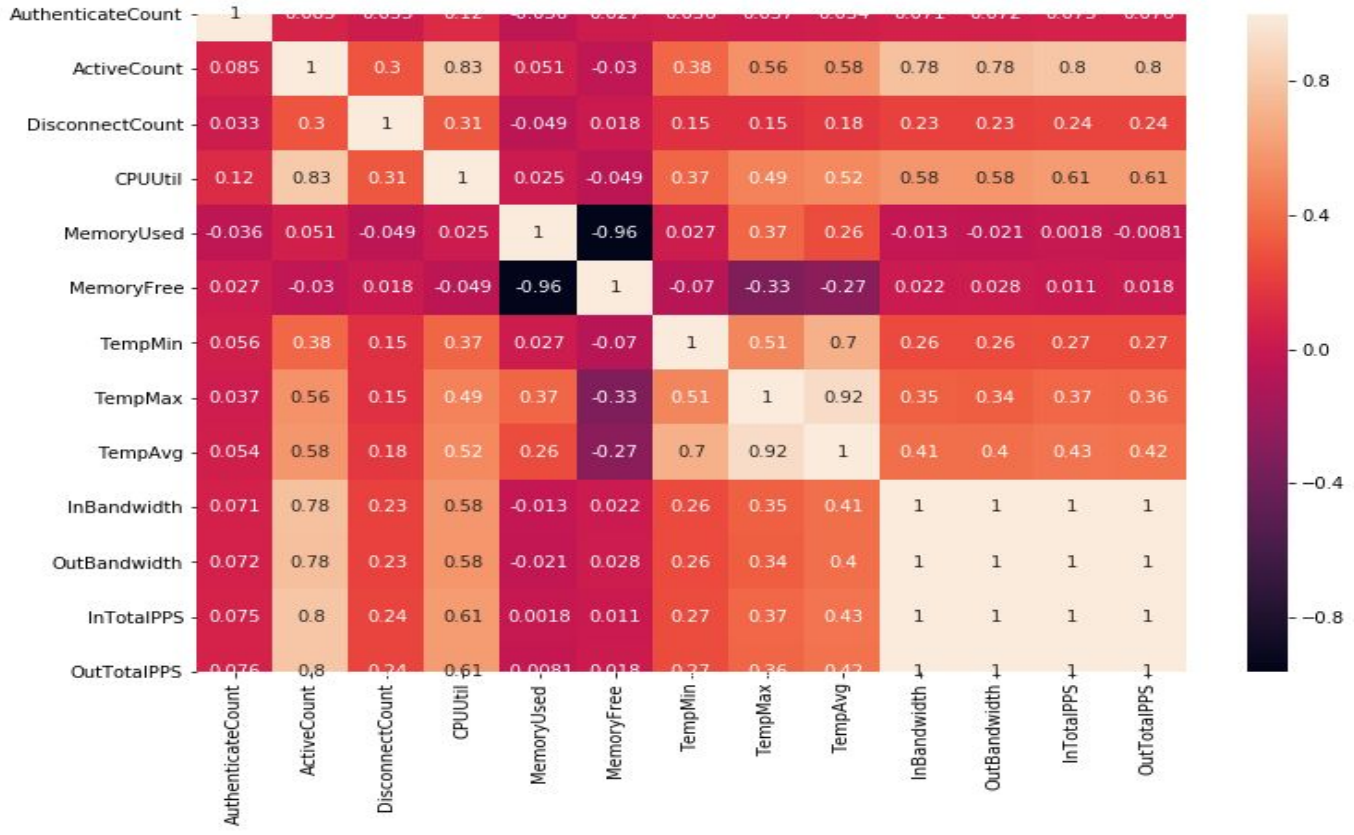




Descriptive Analysis

	Authenticate	ActiveCoun	DisconnectC	CPUUtil	MemoryUsed	MemoryFree	TempMin	TempMax	TempAvg	InBandwidth	OutBandwic	InTotalPPS	OutTotalPPS
count	12385	12385	12385	12385	12385	12385	12385	12385	12385	12385	12385	12385	12385
mean	1.42930965	3226.3922	0.64917239	26.7178	433324608	438226007	26.84344	49.06895	39.72559	147037386	145643412	170958.31	168756.896
std	7.15508777	1010.423	0.91977337	6.794919	39843765.7	58219533.3	1.733762	2.295181	1.577521	80944080.5	80338744	94143.73	93306.416
min	0	1646	0	7	310032000	280964000	0	45	35.14286	0	0	0	0
25%	0	2298	0	22	394372000	407900000	26	47	38.33333	73081599.5	72532770	83880.034	82388.7403
50%	1	3071	0	25	434268000	443640000	26	48	39.16667	150224690	148153456	173889.86	170680.592
75%	2	3923	1	31	464832000	491108000	26	51	41.16667	212951828	211782913	246522.03	244046.589
max	551	5450	11	97	523148000	578320000	42	66	57.66667	341378560	338059206	524028.62	521908.358

Descriptive Analysis



Data Preprocessing

1. Checking whether data is complete.
2. Replacing the outliers with median values..
3. Created normalized and standardized copies of the data.
4. PCA for dimensionality reduction.

Predictive Analysis

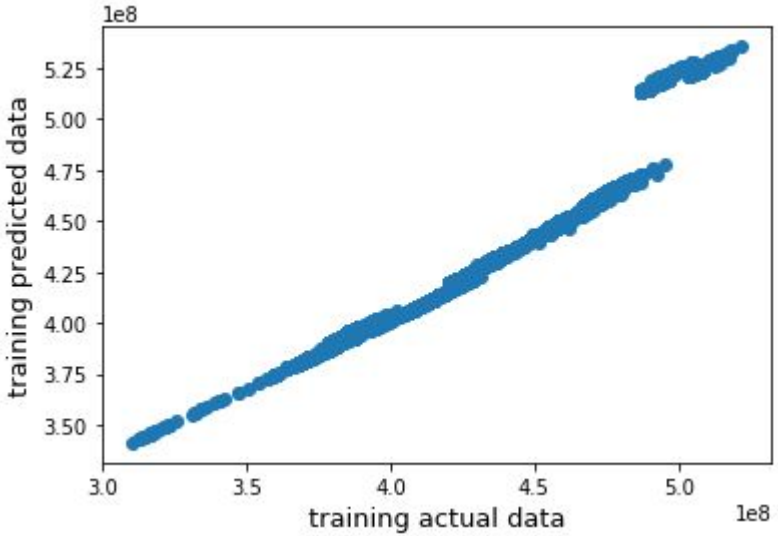
Linear Regression on one attribute

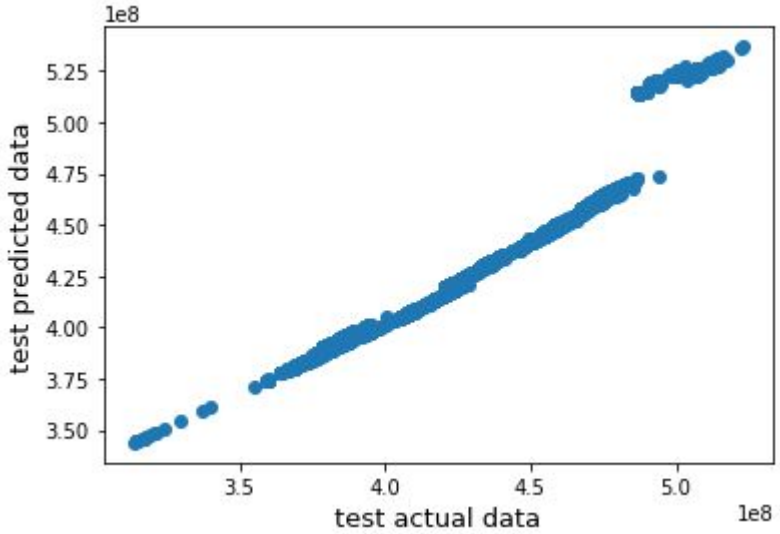
- Applying linear regression for Memory Used and input as Memory Free (correlation coefficient = -0.96).
- RMSE(train data): 1.07×10^7.
- RMSE(testing data): 1.05×10^7.
- R^2 train: 0.92.
- R^2 test: 0.93.
- Almost same results for Normalized and Standardized data.

Linear Regression

The coefficient and intercept of model: W_o=7.2e+8, W_1=-0.66.

Actual Vs Predicted plots:





Multivariate Linear Regression

- Drop least correlated attributes with “MemoryUsed” attribute using pca.
- RMSE train: 8.98×10^6.
- RMSE test: 8.91×10^6.
- R^2 train: 0.94.
- R^2 test: 0.95.

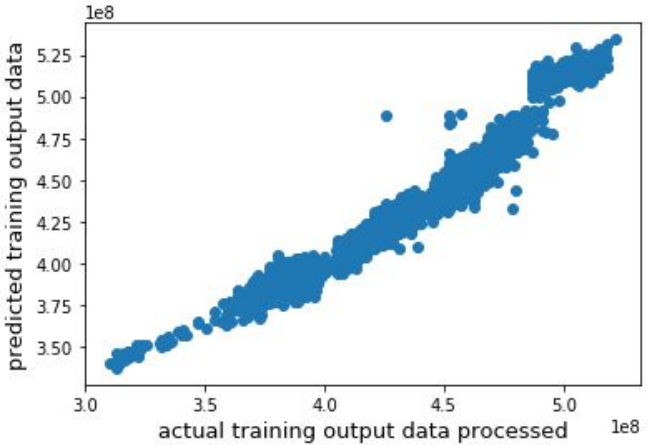
Standardized and Normalized data had identical results for train while for test data original was the best.

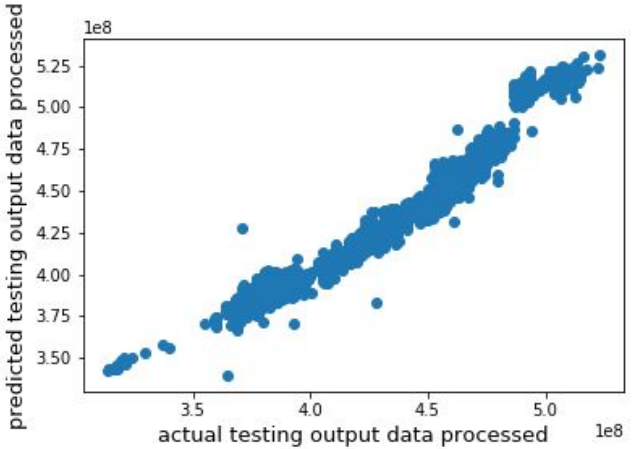
Multivariate linear regression

Actual Vs Predicted plots:

Coefficients and Intercept:

W = [-8.8e-03, 6.6e-01, 3.5e+00, 4.0e+01, -2.8e+03, 4.7e+01, -6.0e+04, -7.0e+05, 1.0e+05, 1.8e+06] Wo = 4.3e+08



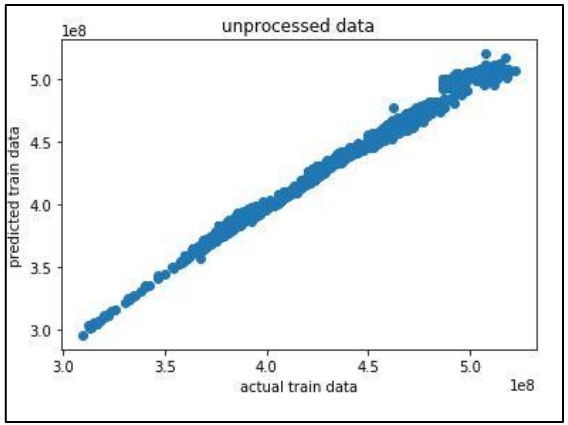


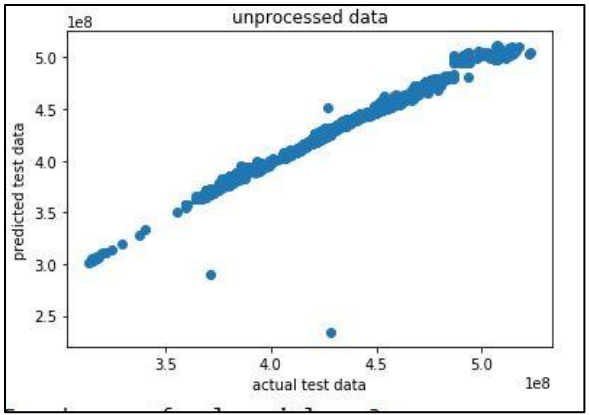
Inferences from Linear Regression

- Best results are obtained after applying pca with n_comp = 10.
- Memory Free attribute being highly correlated gives a good model but considering other attributes as well(even without pca) gives better results.

Polynomial Regression

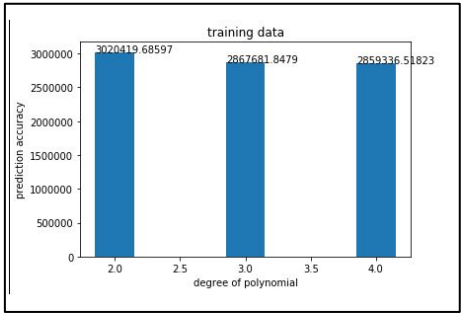
- Task of performing polynomial regression to predict the **Memory Used**
- Polynomial Regression was applied on Unprocessed data, Normalised Data, Standardised Data, Data reduced after using PCA and Feature Selection by correlation coefficient.





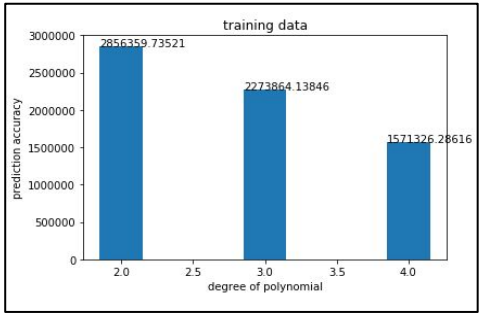
In above graphs, unprocessed data has been predicted at degree p=2

UNPROCESSED DATA



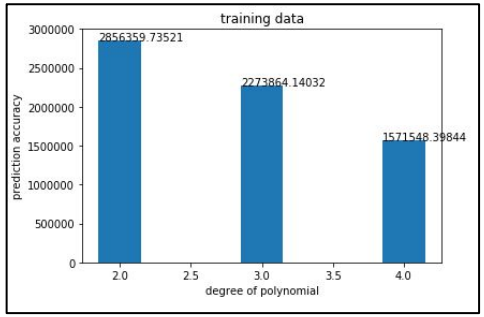
RMSE at degree 2 for unprocessed training data = 3.02 x 10^6

NORMALISED DATA



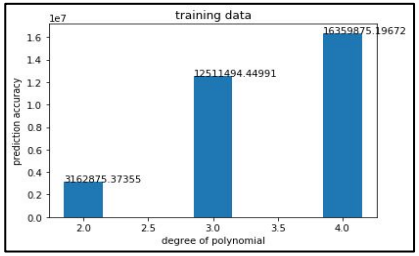
RMSE at degree 2 for normalised training data = 2.85 x 10^6

STANDARDISED DATA



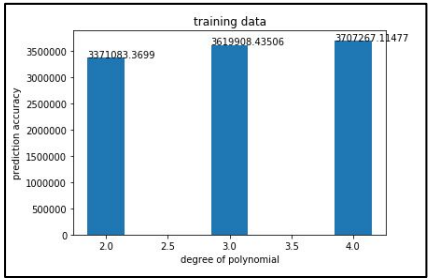
RMSE at degree 2 for standardised training data = 2.87 x 10^6

APPLYING PCA



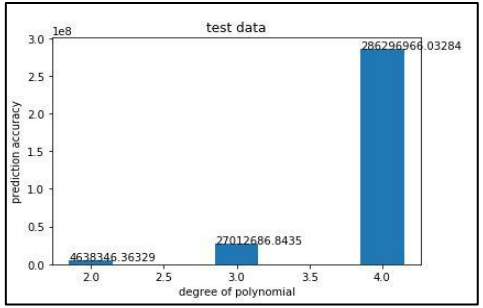
RMSE at degree 2 for reduced training data = 3.16 x 10^6

FEATURE SELECTION



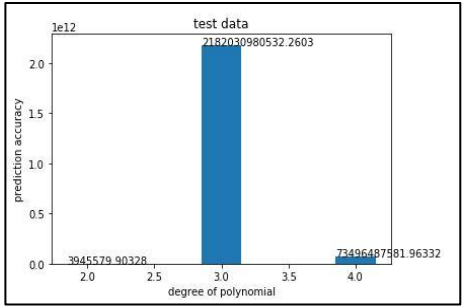
RMSE at degree 2 for feature selected training data = 3.37 x 10^6

UNPROCESSED DATA



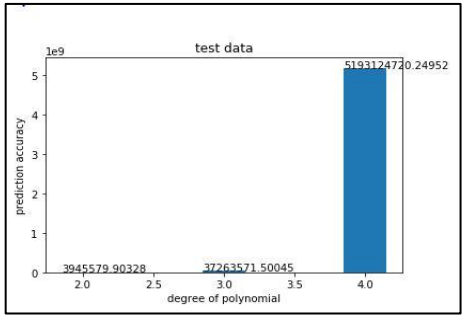
RMSE at degree 2 for unprocessed test data = 4.63 x 10^6

NORMALISED DATA



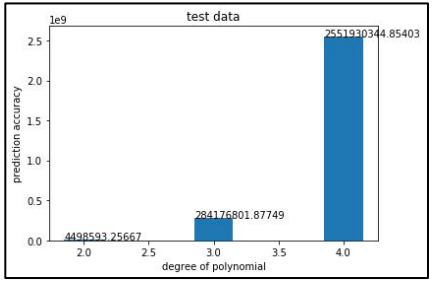
RMSE at degree 2 for normalised test data = 3.94 x 10^6

STANDARDISED DATA



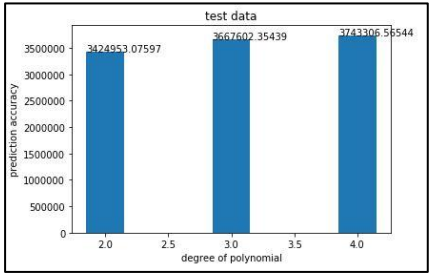
RMSE at degree 2 for standardised test data = 3.96 x 10^6

APPLYING PCA



RMSE at degree 2 for reduced test data = 4.49 x 10^6

FEATURE SELECTION



RMSE at degree 2 for feature selected test data = 3.42 x 10^6

Polynomial Regression

Coefficients (n=5, p=2):

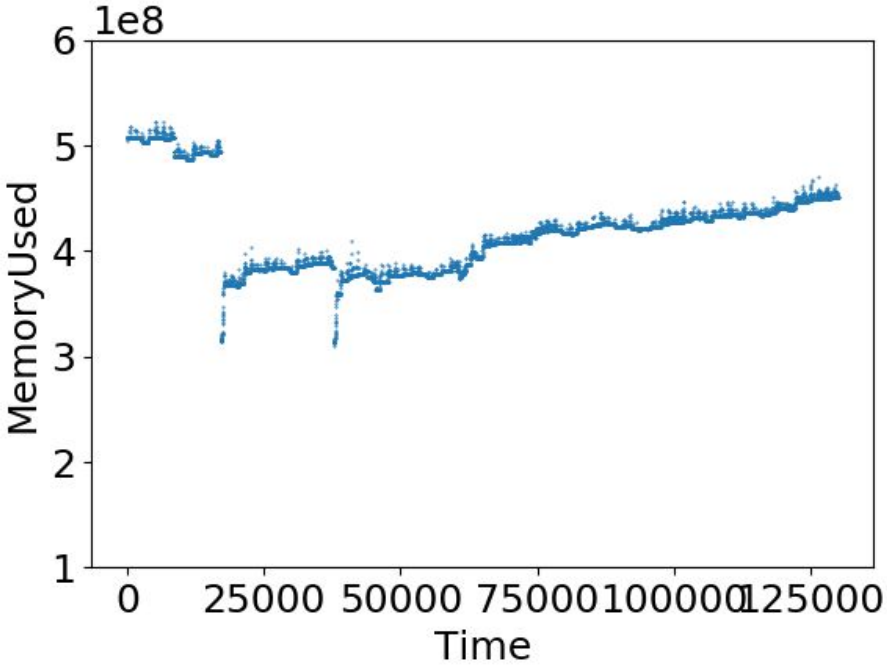
[0.0e+00, -1.4e-02, 7.5e-01, 2.1e+00, 2.8e+01, -9.2e+01, 4.8e-12, 7.3e-11, -7.4e-09, 1.8e-07, -4.7e-06, -2.1e-09, 4.9e-09, 7.2e-07, 1.7e-05, -2.1e-07, 2.1e-05, 1.4e-04, -6.8e-05, 4.6e-03, -4.1e-03]

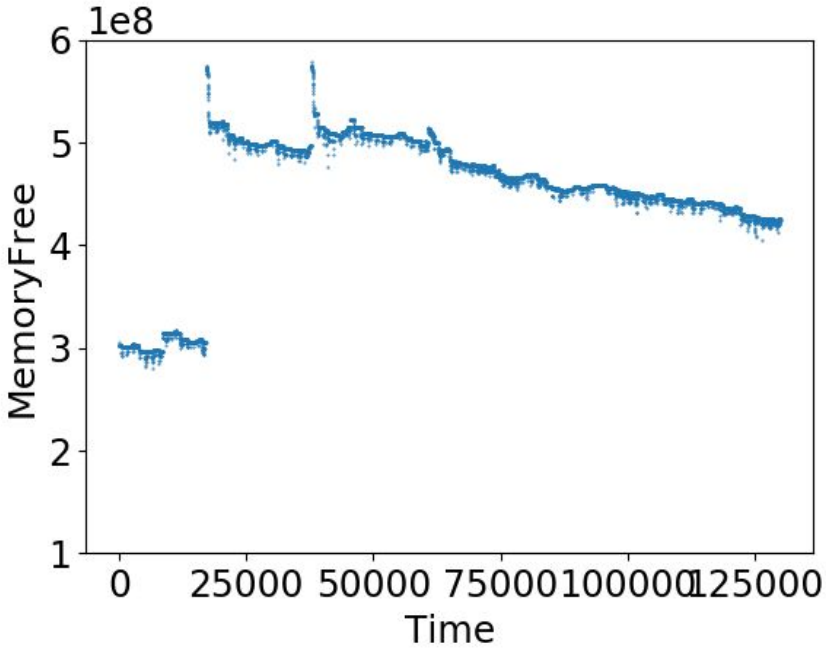
Inferences from Polynomial Regression

After applying Polynomial Regression , we have learnt the following :

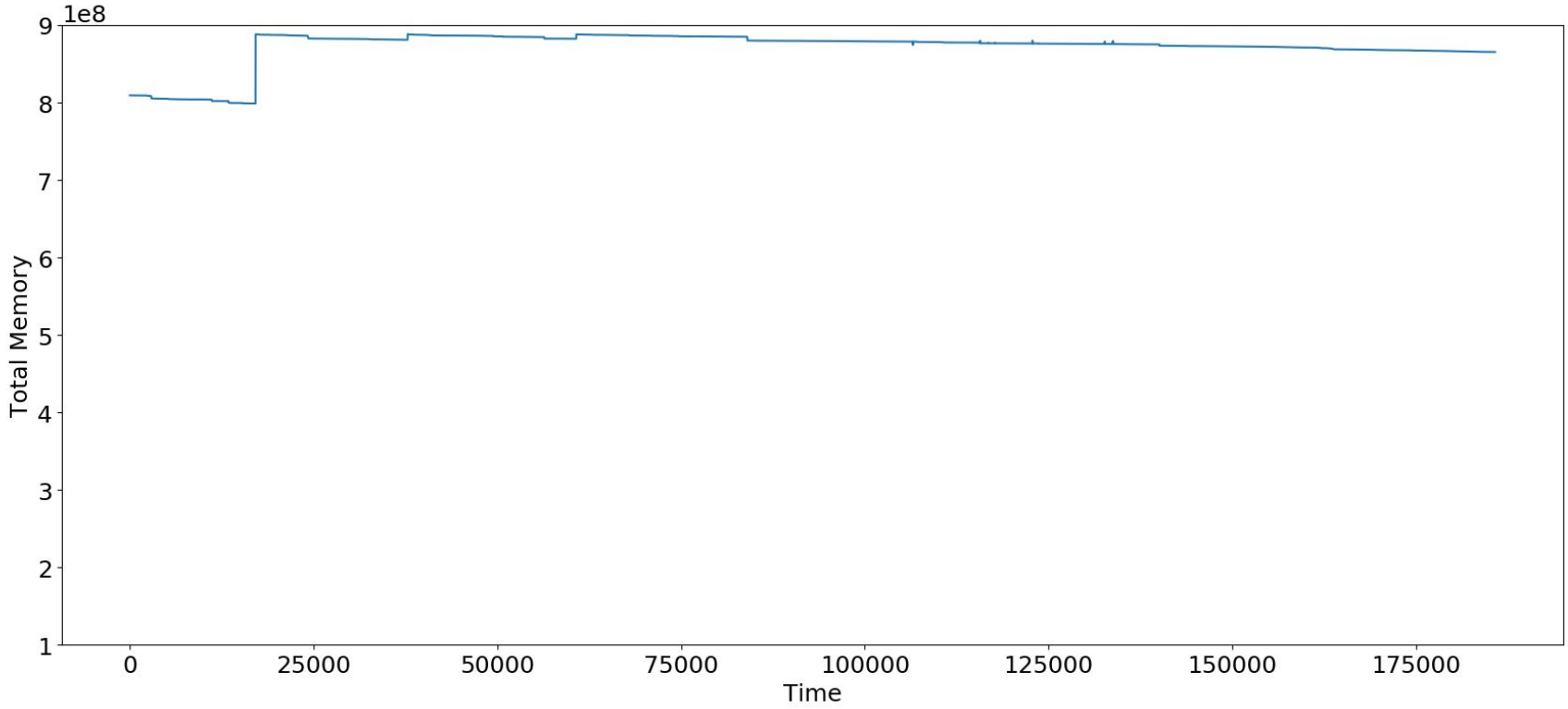
- The best fit polynomial has degree = 2 .
- At higher degrees , the model is getting Overfitted.
- Upon applying PCA for varying number of components, we found the minimum RMSE at n=5.
- The columns that are strongly correlated with MemoryUsed column are: MemoryFree, TempMax and TempAvg.
- the minimum RMSE occurs for feature selected data.

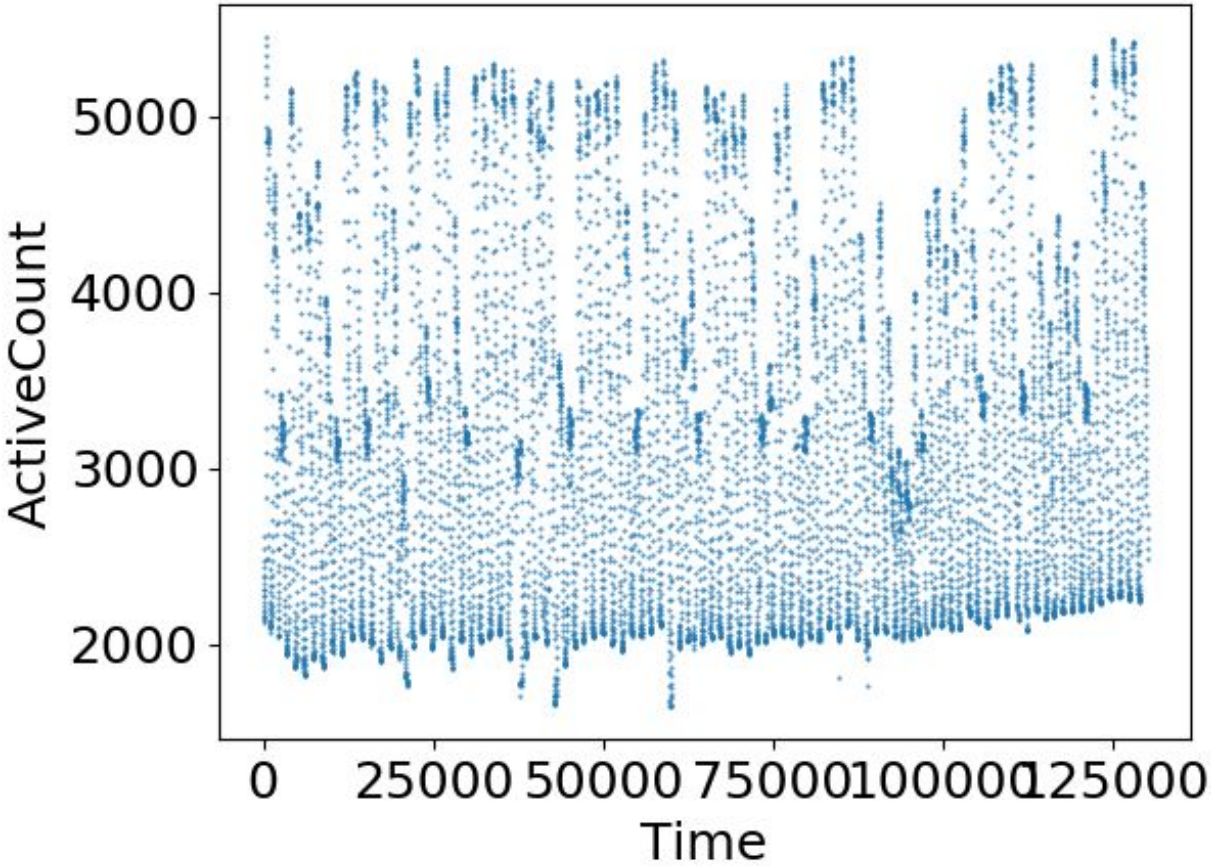
AutoRegression





Sudden drop/rise : 2018-08-17 0:15:00 --> 2018-08-17 2:30:00
Total Memory : ~8 x 10^8 bytes to ~9 x 10^8 bytes




5000
4000
3000
2000
ActiveCount
0
25000
50000
75000
100000
125000
Time

Inferences from time series plots

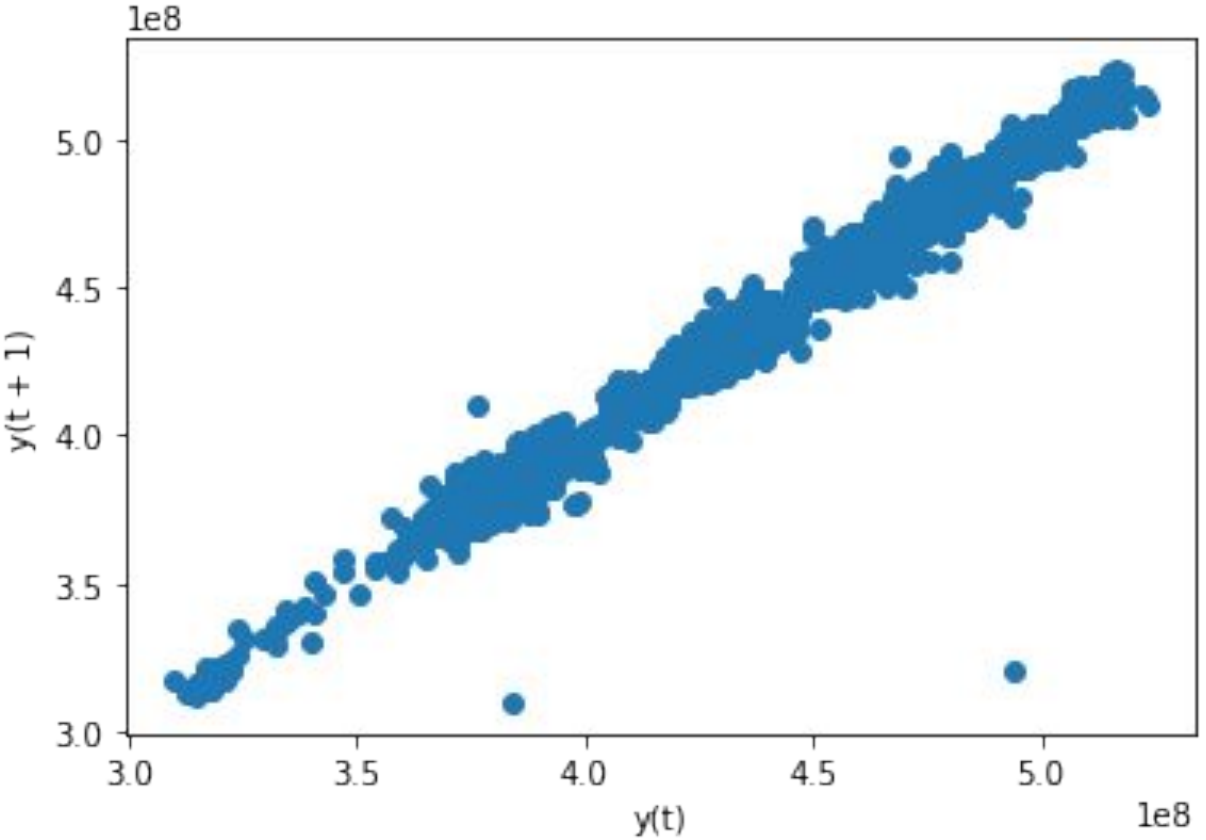
1) Memory used -> drop
2) Total memory -> increase
3) Increase in efficiency of hardware -> Repair work

4) Memory used increase -> Updates or advancement in technology
5) Can predict future time for repair work (as nearly line is fitted)

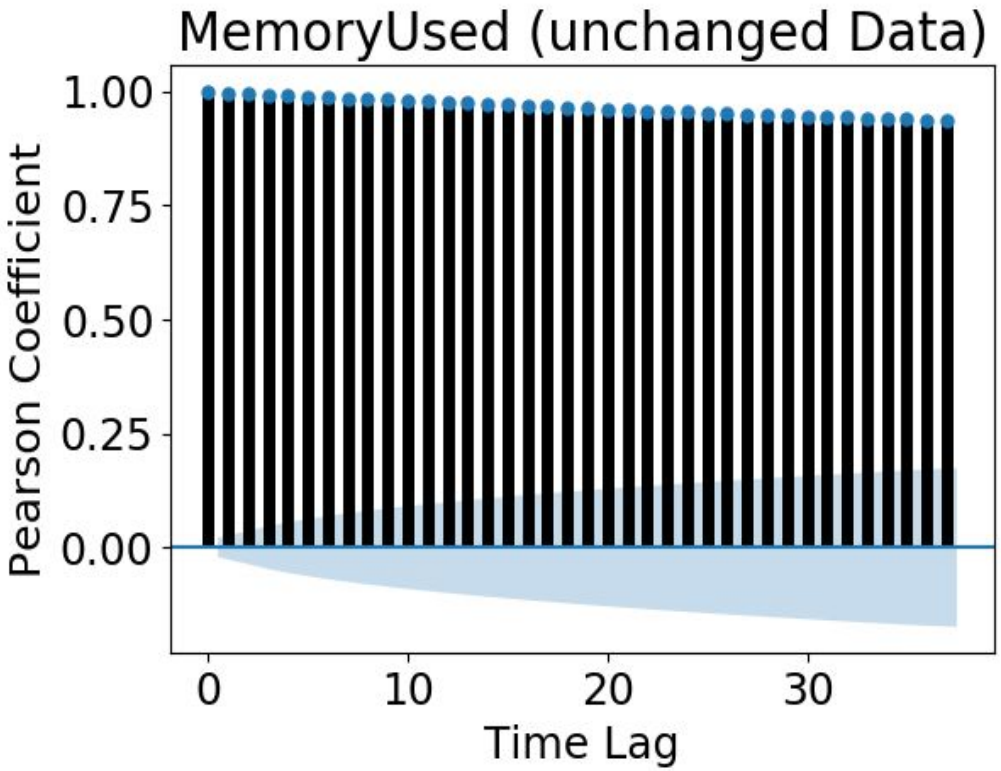
AutoRegression

- **Note** : Time interval of 2hrs 15mins being very small has been treated continuous.
- Criteria for choosing optimal lag : **Bayesian Information Criterion (BIC)**
- **BIC = k log(n)- 2log(L(θ̂))**
  - n - sample size
  - k - no.of parameters to estimate
  - θ - set of parameters
  - L(θ) - likelihood of model, evaluated at maximum likelihood values of θ
- Model with the **lowest BIC** is considered the **best**.
- Prevents overfitting

Scatter plot of MemoryUsed(t+1) vs MemoryUsed(t)



AutoRegression



Optimal lag:
9

Test data

RMSE (non-dynamic model):
46842165

AutoRegression

Optimal lag: 9

Parameters:
[1.07651499e+06 7.00780448e-01 1.90937500e-01 5.24506807e-02
9.71145282e-02 -6.11853094e-02 1.11652613e-02 -2.93620542e-02
8.97791239e-02 -5.42678901e-02]

[const,t-1,.....t-9]

Inferences from Auto Regression

Considering tuples only from 5000 to 12385 (break point)

Lag=12

Coefficients = [1.21786557e+06, 2.82006323e-01, 1.36022290e-01, 5.88638609e-02,2.09100186e-01, -3.94611302e-02, 4.52385734e-02, 1.21531779e-02, 1.79434597e-01, -6.86169994e-03, 2.15051626e-02, 1.36505054e-03, 9.79412506e-02]

RMSE: 16458319 (<46842165)

Percentage error: 3.105

Conclusion

BEST MODEL

Polynomial regression of degree 2 with 5 attributes (on applying PCA).