MACHINE LEARNING

ASSIGNMENT 1

“WEATHER DATA ANALYSIS”

1. Using Linear Regression
    - Gradient Descent
    - Newton’s Method

2. Using Closed Form Solution

1. Task Introduction

- We have been provided with the weather dataset that contains weather data samples from various timestamps.
- Our task is to predict the “MinTemp” provided other data columns.
- In all, Dataset has 11 features and our assumptions are that these 11 features followes normal distribution, samples are Independently and identically distributed (IID) and that the features are linearly related to the target (MinTemp in this case).
- Following all these assumptions, we are ready to use linear regression to begin our analysis of this linear relationship that persists in our dataset throughout all samples.

2. Description the data and Preprocessing

- Pipeline of our preprocessing will be as follows.

    1. Look for duplicate rows and remove them.

    2. Look for potential null/nan values and use appropriate strategy to treat them.

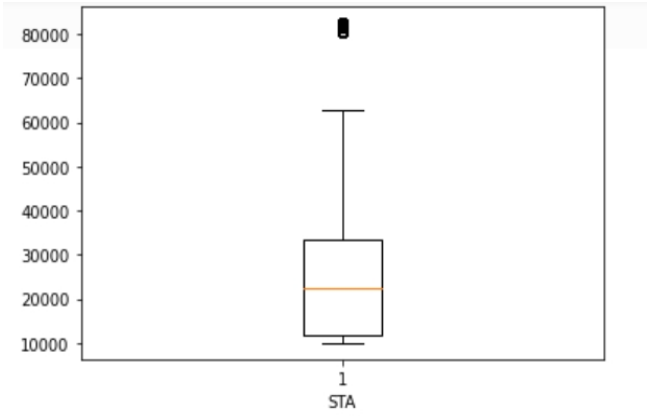
    3. Look for non adherent outliears and remove them or replace them with boundary values.

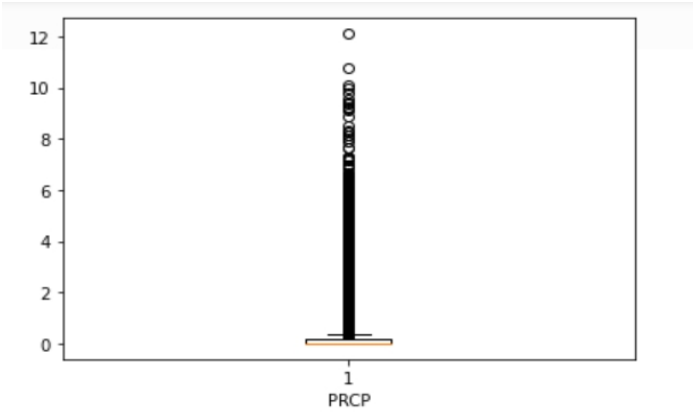
    4. Correlation of analysis of data columns with the target column and other features.

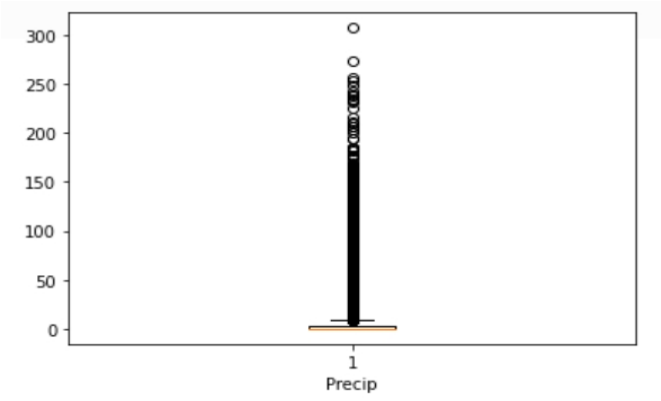
    5. Standardisation of data values to map them to standard scale.

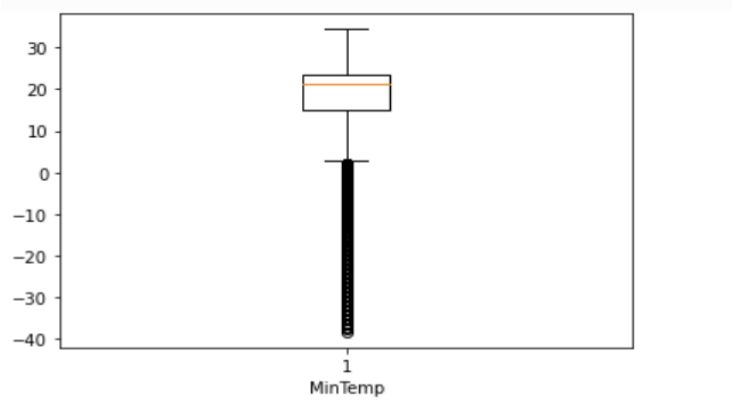
- Data can be described in various ways, We primarily be focussing on Box-Plots and Correlation matrix as our data is simple and numerically/contigously distributed.

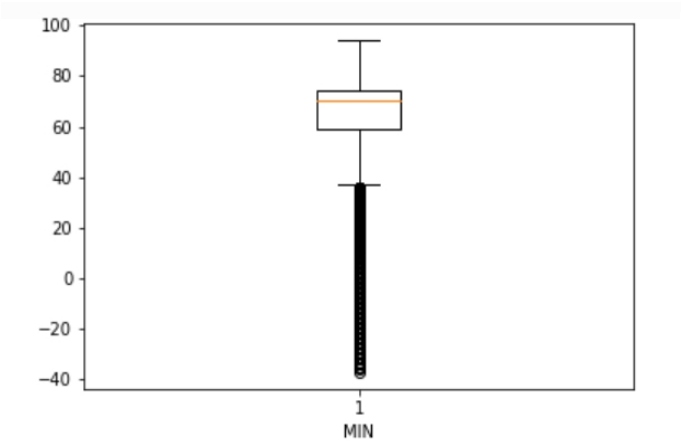
➢ Box plots of features having potential outliers

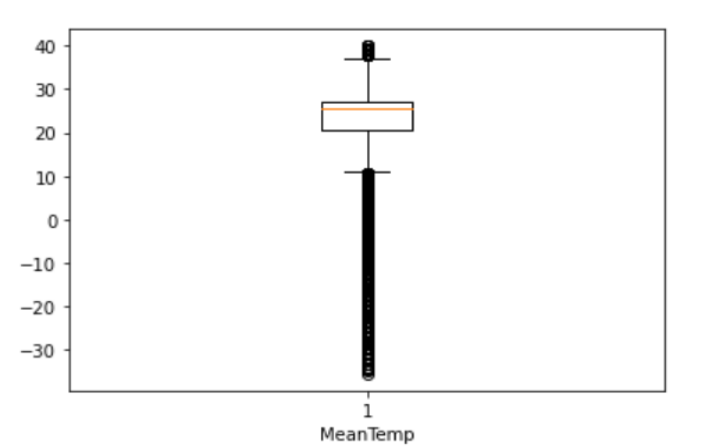


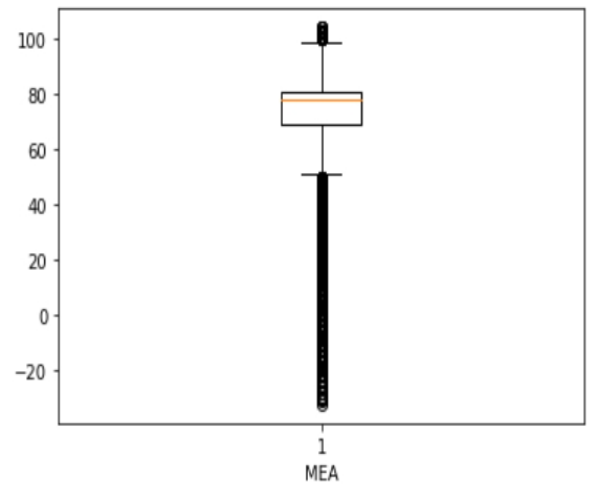


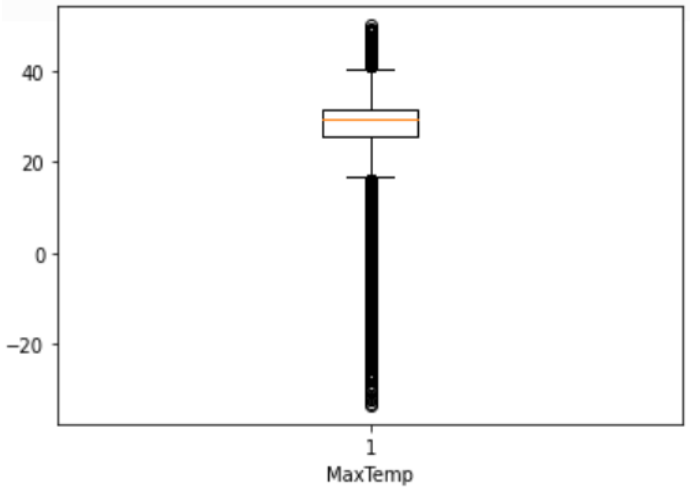


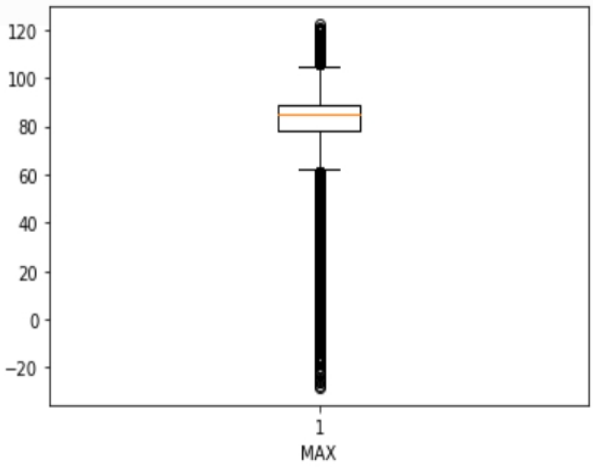







100
80
60
40
20
0
−20
1
MEA


40
20
0
−20
1
MaxTemp


120
100
80
60
40
20
0
−20
1
MAX

➢ Statistics for Dataset.

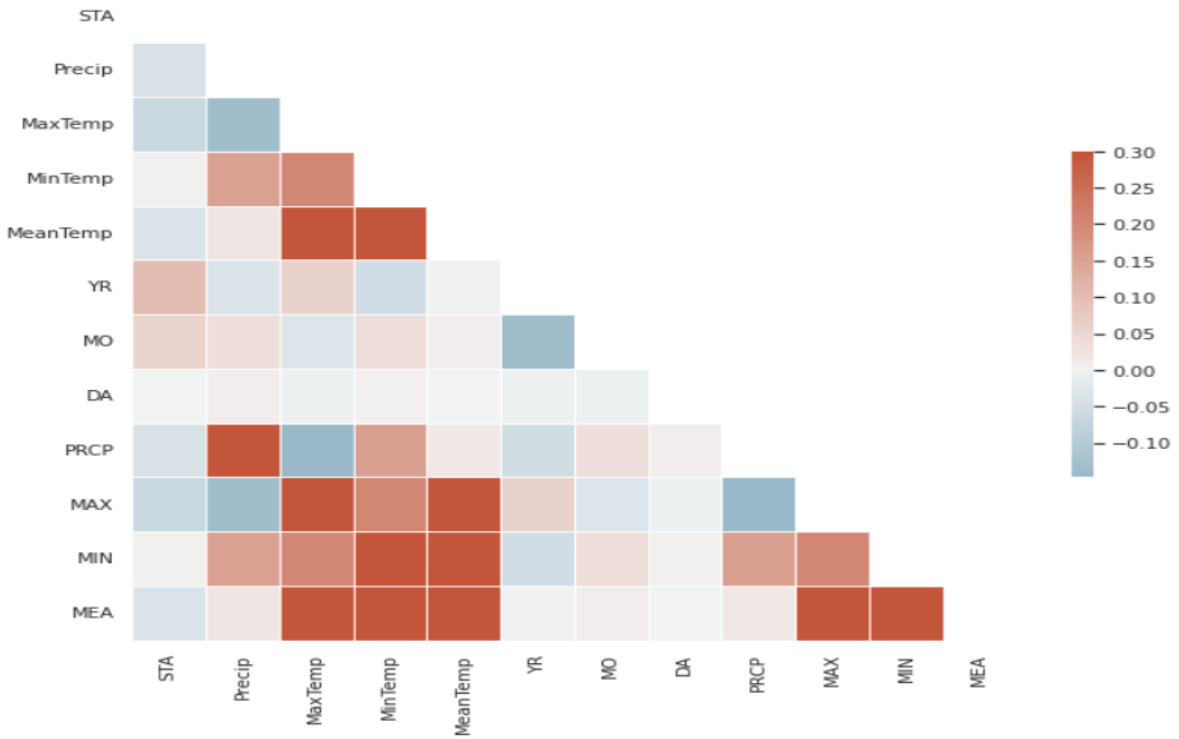
	STA	Precip	MaxTemp	MinTemp	MeanTemp	YR	MO	DA	PRCP	MAX	MIN
count	103426.000000	103426.000000	103426.000000	103426.000000	103426.000000	103426.000000	103426.000000	103426.000000	103426.000000	103426.000000	103426.000000
mean	30340.621094	1.788722	28.803896	19.234772	24.022203	43.874104	6.742135	15.793098	0.073153	83.847008	66.622589
std	21243.585938	3.234394	5.577626	5.815077	5.319395	1.035674	3.385328	8.788745	0.127652	10.039728	10.467139
min	10001.000000	0.000000	10.000000	1.111111	7.777778	41.000000	1.000000	1.000000	0.000000	50.000000	34.000000
25%	11704.000000	0.000000	26.666666	16.666666	21.666666	43.000000	4.000000	8.000000	0.000000	80.000000	62.000000
50%	31101.000000	0.000000	30.000000	21.111111	25.555555	44.000000	7.000000	16.000000	0.000000	86.000000	70.000000
75%	34005.000000	3.754205	31.666666	23.333334	27.222221	45.000000	10.000000	23.000000	0.150649	89.000000	74.000000
max	82506.000000	18.542000	46.666668	34.444443	39.444443	45.000000	12.000000	31.000000	0.730000	116.000000	94.000000

OUTLIERS Here we see the sone data columns has values that are more than 3 times the standard deviation from the mean.

Two most promising techniques are, To remove the entire row containing the non adherent outlying data and To replace the data column value with the boundry value i.e. the 3rd quantile data value of the respective column. We are using the former one, as the data samples are not sparse and data shows no sensitivity with minimal loss of samples.

Correlation Analysis:

- With reference to “Statistical Learning by Trevor – Stanford publication” There should be minimum correlation between features and quite a large correlation with the regression target.
- Reason for former is to explain as much variance in regression target, We need to have diversified and uncorrelated features as extreme correlation between features doesn’t account any additional info to explain the target, perhaps only make regression task harder by worsessing the over-fitting and increasing complexity of model.



- Fortunately, Our correlation matrix outrages all the flaws concluding from high positive/negative intra-sample correlation.

➢ Dealing with Null/Na/Nan values

- Following figure depicts the amount of null values consitituted in our dataset.

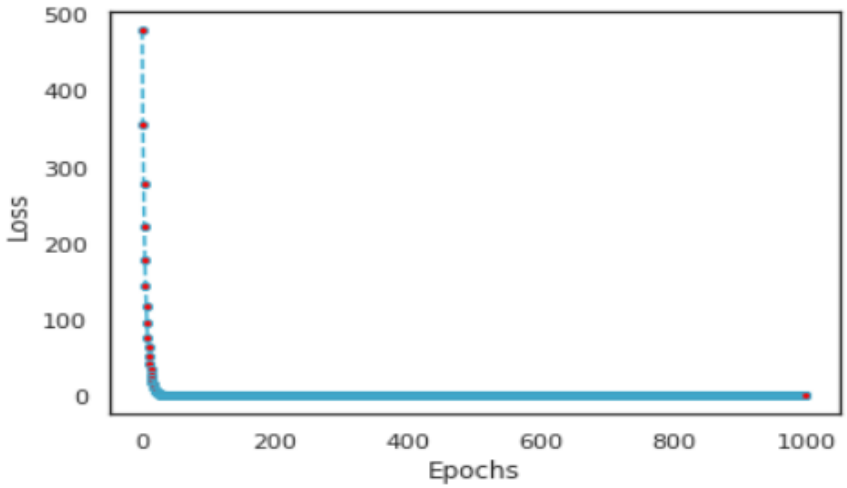
```
0   STA       119040 non-null
1   Precip    119040 non-null
2   MaxTemp   119040 non-null
3   MinTemp   119040 non-null
4   MeanTemp  119040 non-null
5   YR        119040 non-null
6   MO        119040 non-null
7   DA        119040 non-null
8   PRCP      117108 non-null
9   MAX       118566 non-null
10  MIN       118572 non-null
11  MEA       118542 non-null
```

We see that few columns have null values, So how do we deal with them?

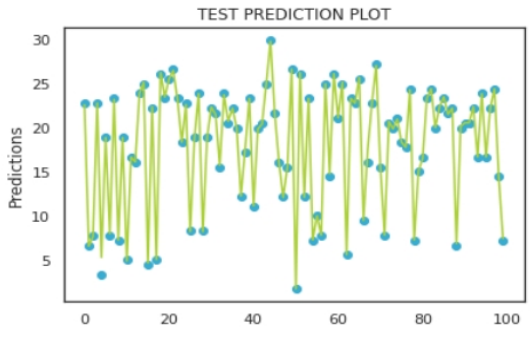
- Simplest approach will be to remove the rows containing null values but that will result in loss of valuable information and should not be performed when data generation is costly.

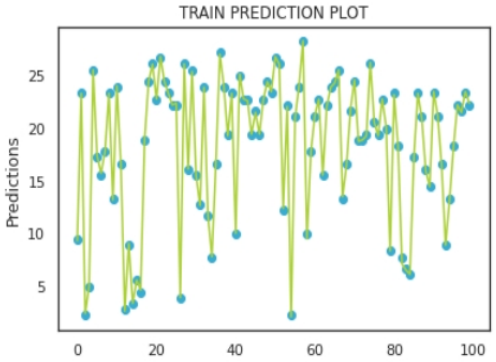
- Second approach, The most promissing and perhaps that's what we have used is to replace the null values with mean of respective column.

➢ Training Model and Analysis of results.

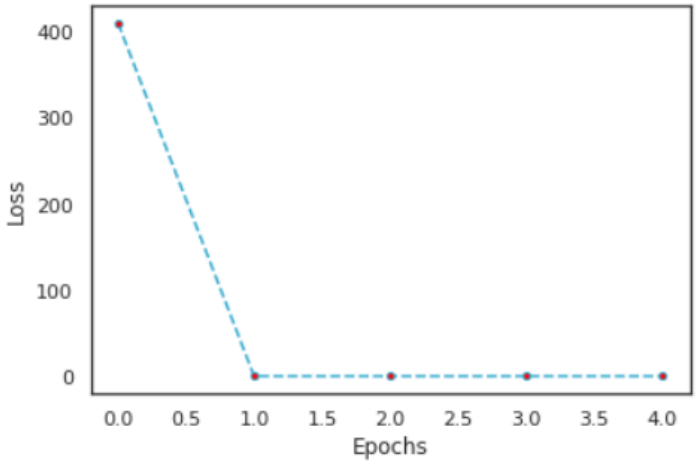


Test MSE Loss = 0.028, Test Accuracy : 99.79%

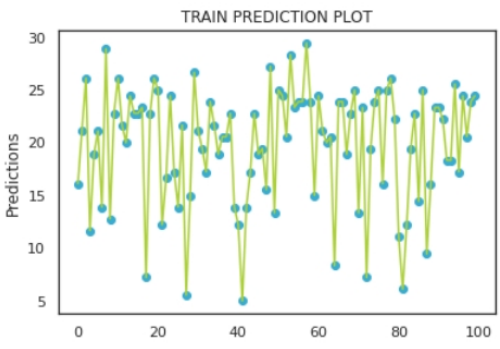


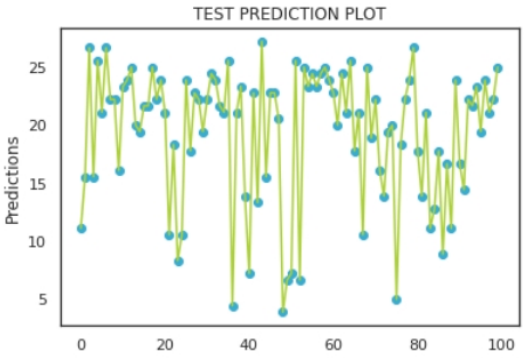


- Training with Newton's optimisation method

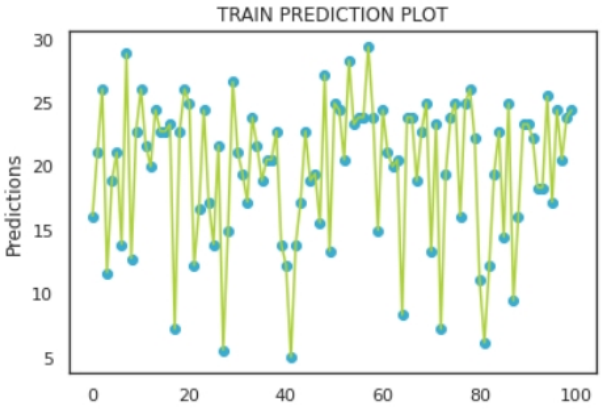


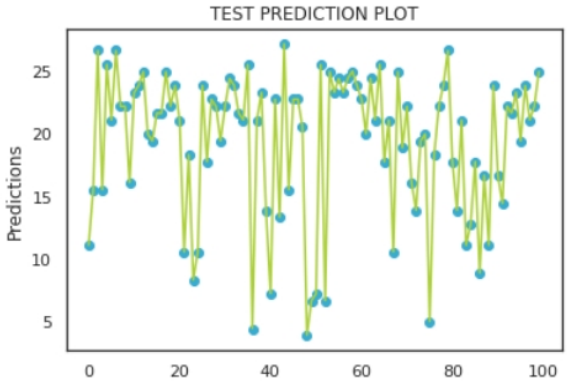
Test MSE : 0.005 , Test Accuracy : 99.98%





- # Closed form solution





Closed Form Loss : 0.0007, Accuracy Nearly 100%

Here, we end our analysis of weather dataset.

For Reader,

After going through the python notebook and this report describing analysis process, if you really found this analysis intuitive and worth appreciating, I would say this task couldn't be accomplished without immense efforts made by our T.A. Shourabh Payal and Prof. Dinesh Babu.

Deepak Nandwani
MT2021037
Mtech. CSE