# ANOMALY DETECTION & ANALYSIS IN MULTIVARIATE TIME SERIES DATA

Presented by:

ANKITA DUTTA GUPTA

# What is Anomaly & why is it important to detect it ?

- Anomalies/outliers are different from the norm with respect to their features.
- It is import for detecting the anomaly in power systems before it expands and causes serious faults such as power failures or system blackout .
- With deployment of phasor measurement units(PMUs), massive amount of synchrophasor measurement is collected .
- This makes it possible for the real time situation of awareness of the entire system.

Based on the categories, there are three kinds of Outliers:

1. **Point Outliers:** A single instance of data that is too far off from the rest. User Case: Detecting fraud on Credit Card based on “amount spent”.

2. **Contextual Outliers:** The outlier is context specific. User Case: High AC usage during the summer is normal but may be odd during the winter.

3. **Collective Outliers:** A subset of data that collectively deviate from the norm though individually those points might not be outliers. User Case: Multiple online orders for the same pizza at a very short span time from a specific block in a city. This might be a situation of market manipulation.

# About the data

- **Given** : Time series data from an electrical equipment, has various parameters associated and is not labelled.
- Total 84 columns, with 13 ‘Engineering Sensor’ columns, only these have been considered for the entire work as per instructions.
- Data has been labelled as ‘Anomalous’=True , ‘Non Anomalous’= False
- Data has 42070 instances(rows)

1

Let’s start with the first set of Visualizations

# FIRST STEPS in EDA

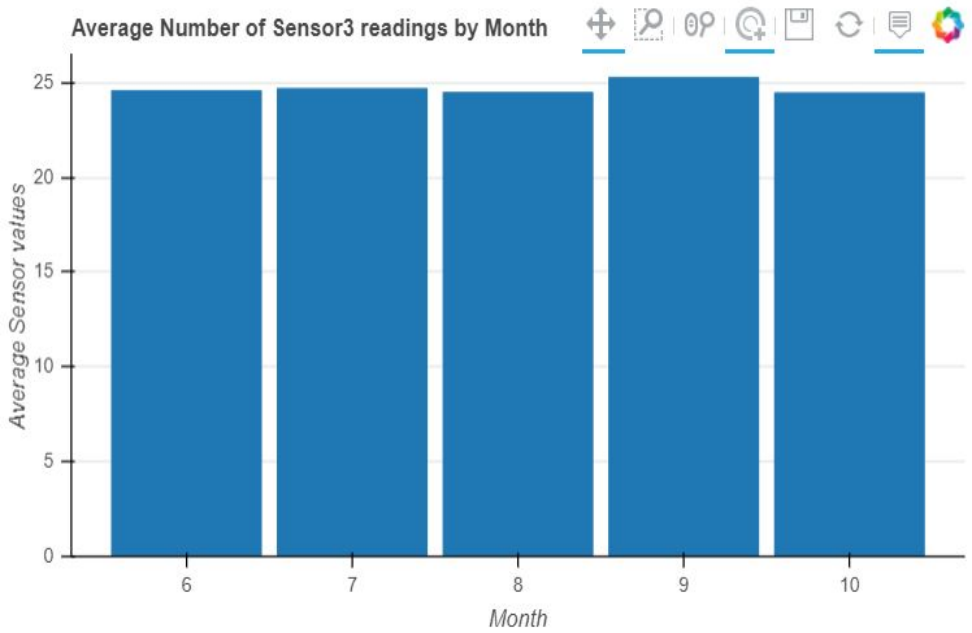
- Combined ‘date ’ and ‘ time’ and converted to pandas datetime format
- Dropped ‘date’ , ‘ time’ , ‘ timestamp’.
- Set the index as datetime
- Added columns like, hour, month, year, daylight, night, Dayoftheweek, Weekday, Weekend, Quarter to the dataset
- Observed that:
- Time period is from 21 september 2014 to 8 October 2017 with timestamps.
- Here is a sample screenshot of the data after cleaning.

| datetime | input current | input current (Min) | input current (Max) | input current (StdDev) | current data 0x1 | current data 0x2 |
|---|---|---|---|---|---|---|
| 2014-09-21 12:39:19 | 1.14 | 1.01 | 1.32 | 0.09 | 0.34 | 0.31 |
| 2014-09-21 12:39:29 | 1.16 | 1.01 | 1.27 | 0.08 | 0.37 | 0.30 |
| 2014-09-21 12:39:39 | 1.15 | 0.96 | 1.25 | 0.09 | 0.36 | 0.31 |

# EDA

1. highest sensor values averages in the month of september

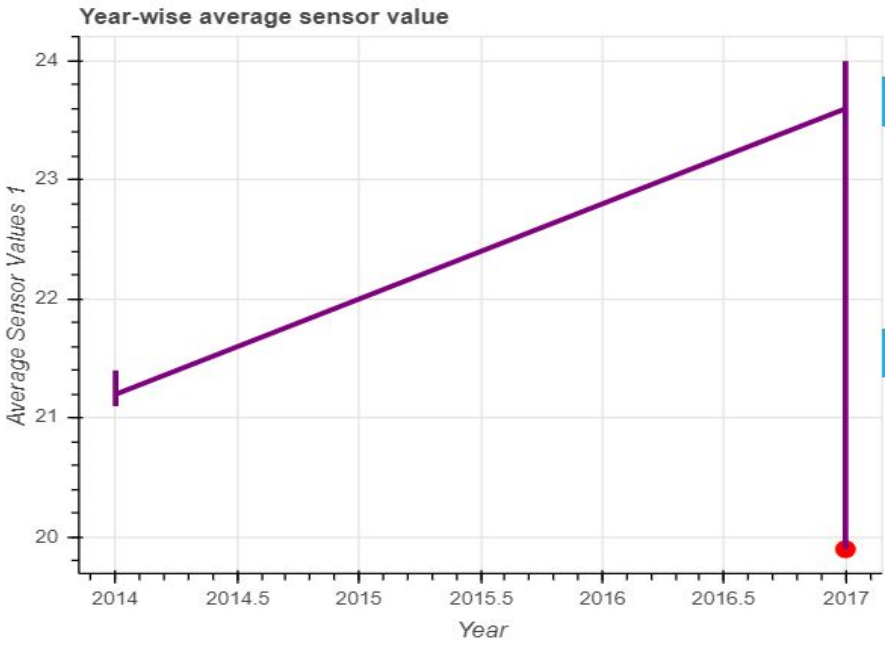
2. around 98% values recorded from the engineering sensors are from 2017 and rest from 2014

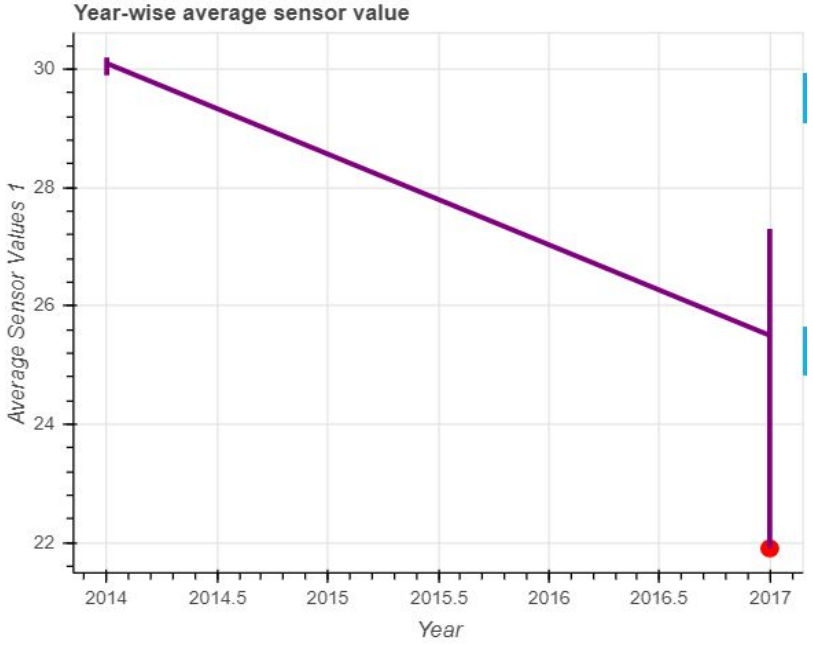


# EDA

1. Significant average decrease in sensor 1,2 value from 2014 to 2017 by 14%

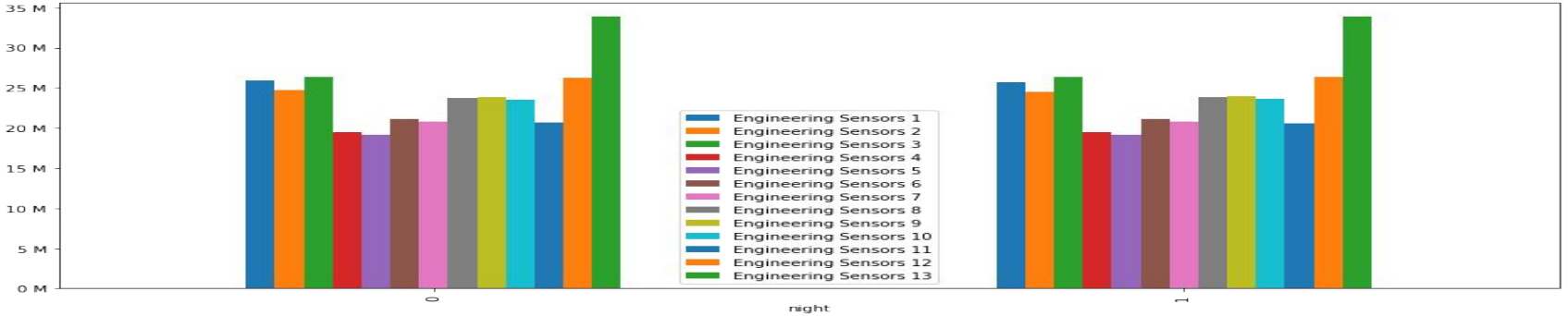
2. Slight average increase in sensor 10 value from 2014 to 2017 by approximate 9 %

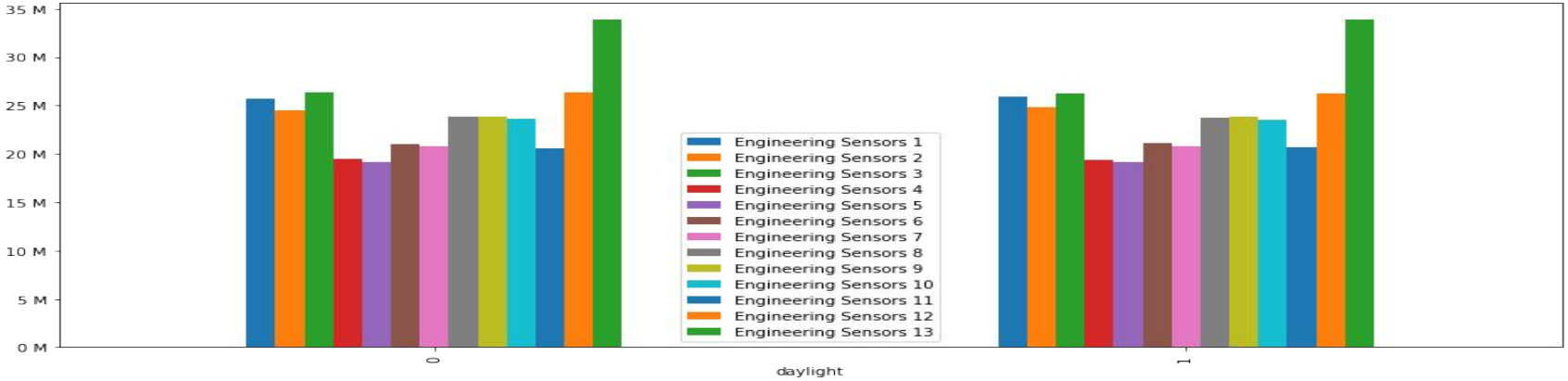


# EDA

No Significant average changes in all the sensor value in the day and night hours

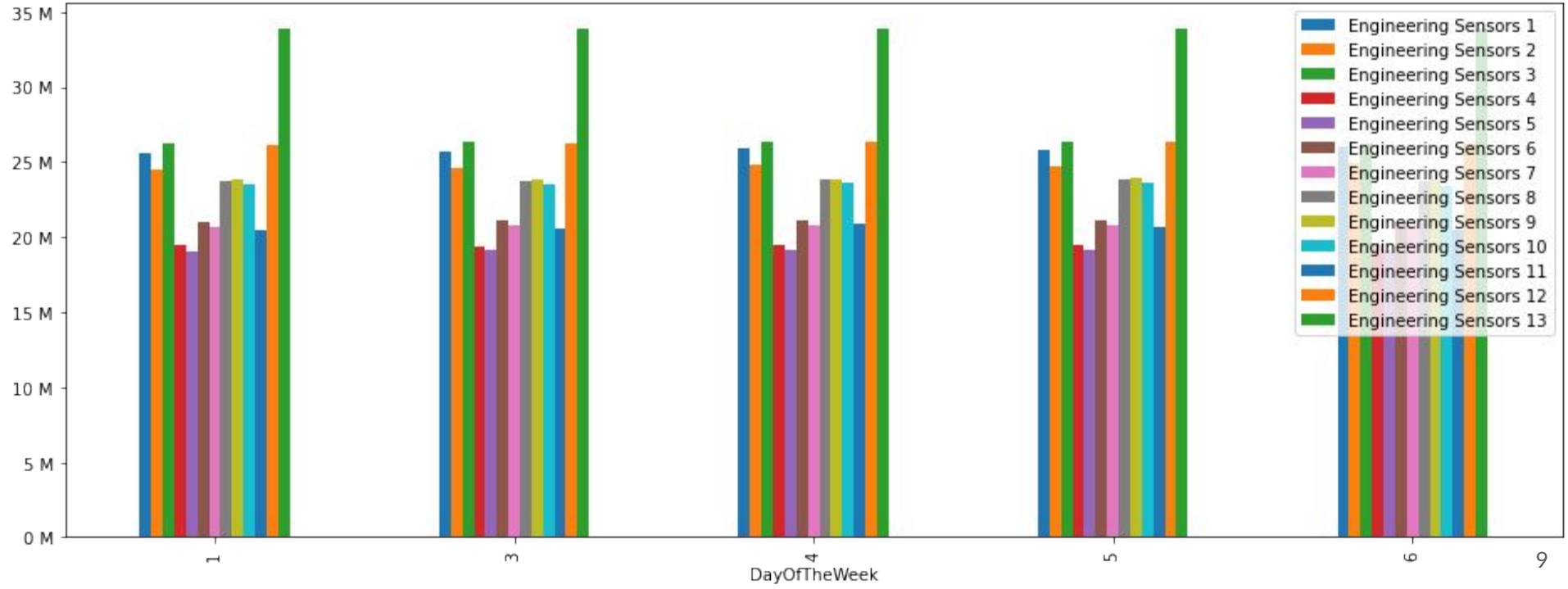


# EDA

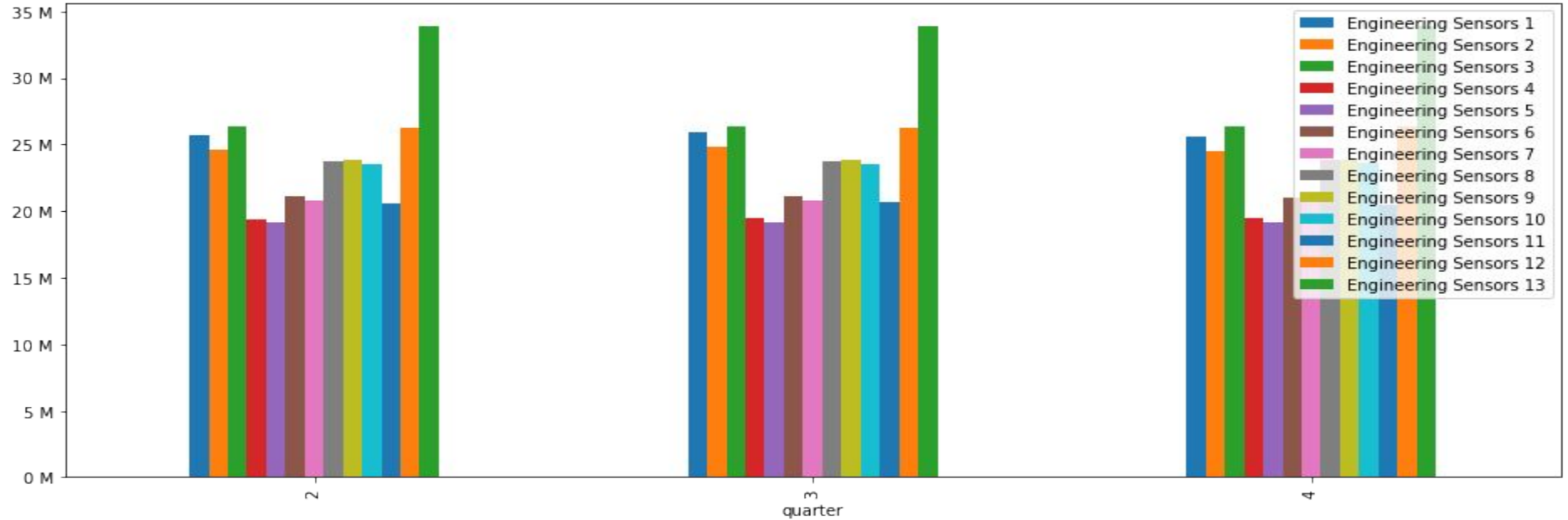
1. Slight variations in average sensor 1 values on weekly basis

2. No major variations in average sensor values for other 12 sensors on weekly basis



# EDA

No significant variations in average sensor values for all 13 sensors on quarterly basis

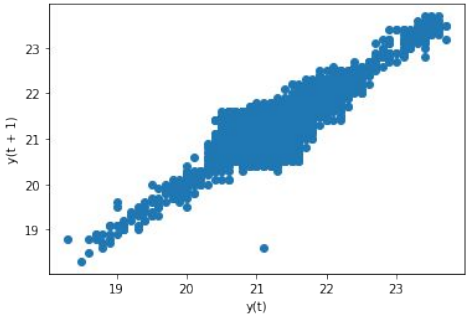


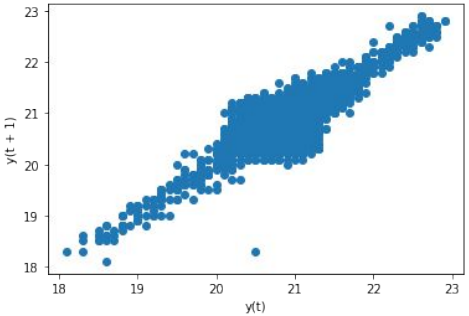
# EDA

LAG PLOTS:

SENSOR 6,7 relatively strong positive correlation between observations and their lag1 values.SENSOR 11,4  highly uncorrelated

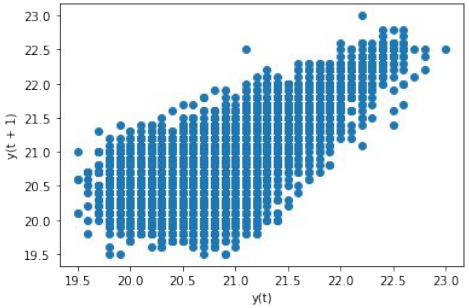
SENSOR6

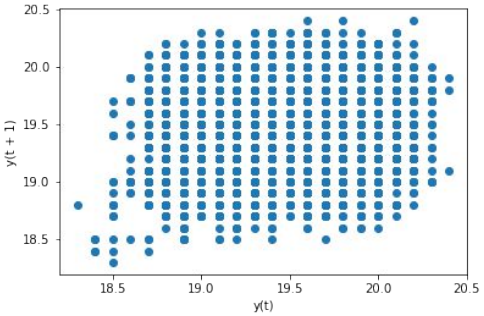



SENSOR7

SENSOR11




SENSOR4

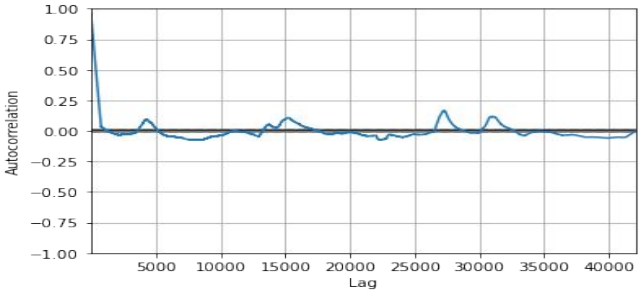
# EDA

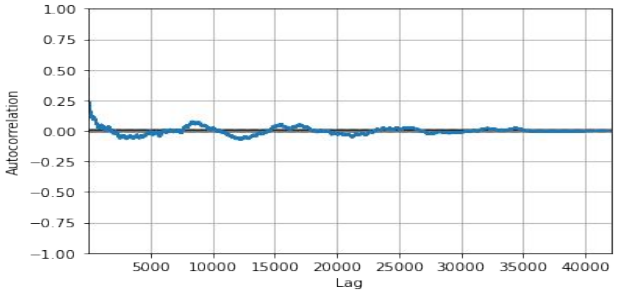
1. The resulting plot shows lag along the x-axis and the correlation on the y-axis. Dotted lines are provided that indicate any correlation values above those lines are statistically significant (meaningful).

2. sensors 1,2,3,10 show few strong positive correlations

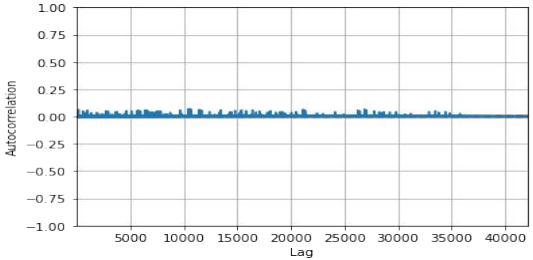
3. sensor 4,5 doesn't show much autocorrelation

SENSOR1




SENSOR4




SENSOR13

# Anomaly Detection:

Models tried : Isolation Forest, OneClassSVM

1. Normalized the 13 sensor columns using StandardScaler()

2. Used model : IsolationForest(n_estimators=100,max_samples='auto',max_features=scaled_time_series.shape[1],n_jobs=-1,random_state=42,verbose=0)

3. OneClassSVM(nu=nu_estimate, kernel="rbf", gamma=0.01)

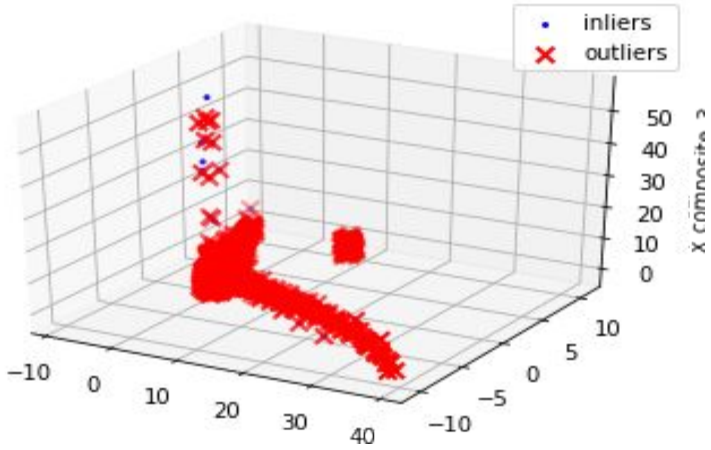
   where,outliers_fraction = 0.05 and nu_estimate = 0.95 * outliers_fraction + 0.05

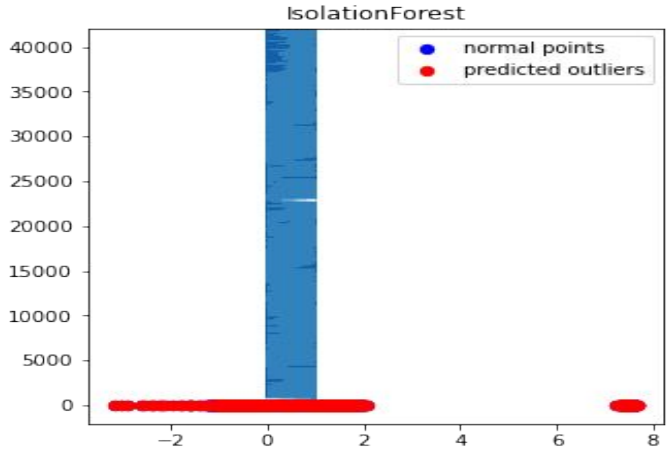
MODEL CHOSEN: ISOLATION FOREST

THEN, WHY ISOLATION FOREST FOR AD?

- Isolation Forest works by selecting a sub-set of features and sub-set of data from the entire dataset and running a Decision tree algorithm.
  - 
- It isolates the outliers by randomly selecting a feature from the given set of features and then randomly selecting a split value from a uniform distribution between the lowest and highest values of that feature.
- The idea behind isolation forest is that outlier values needs lesser number of splits and will thereby produce shorter paths.

# RESULT VISUALISATION




**2D plot gives a better picture,Anomalies are highlighted as red edges and normal points are indicated with blue points in the plot.**

# DATA MODELING

# DATA MODELING

WHY DID I TRY THE FOLLOWING MODELS ?

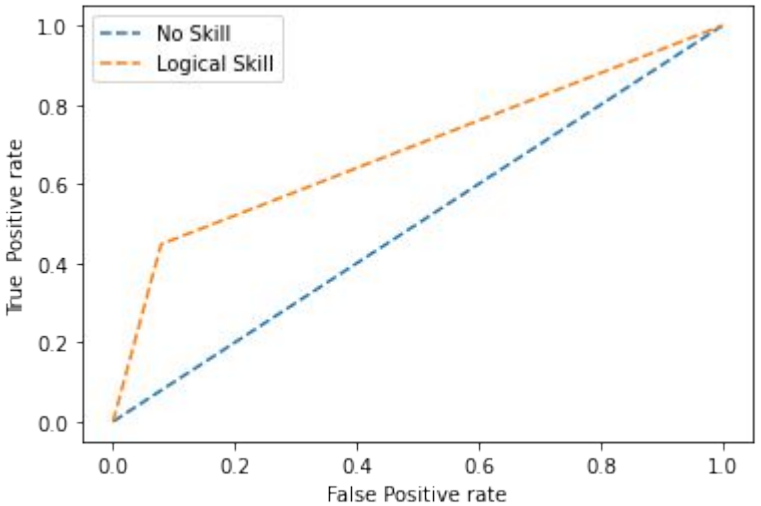
1. CLASS CLASSIFICATION : ONE CLASS  SVM

2. PROXIMITY / DENSITY BASED= CLUSTER BASED LOCAL OUTLIER(CBLOF)  & KNN

3. PROBABILISTIC APPROACH = ANGLE BASED OUTLIER DETECTOR

4. OUTLIER ENSEMBLES & COMBINATION= ISOLATION FOREST

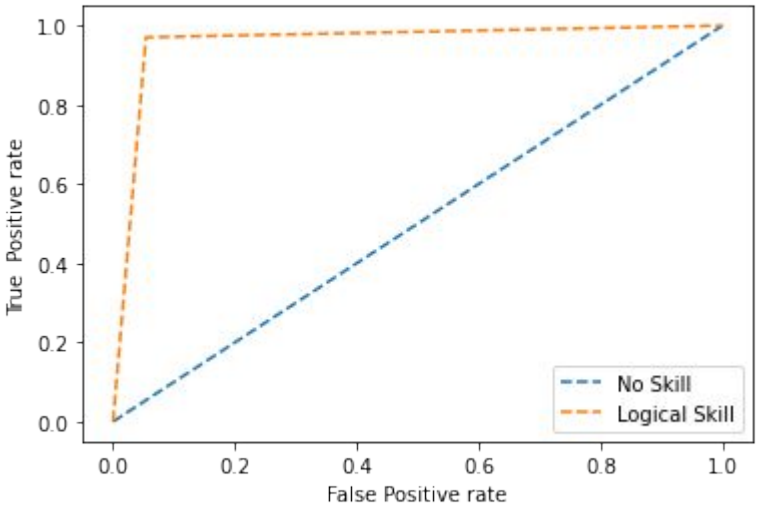
# ANGLE BASED OUTLIER DETECTOR (ABOD)




No Skill : ROC AUC = 0.500

Logistic : ROC AUC = 0.684

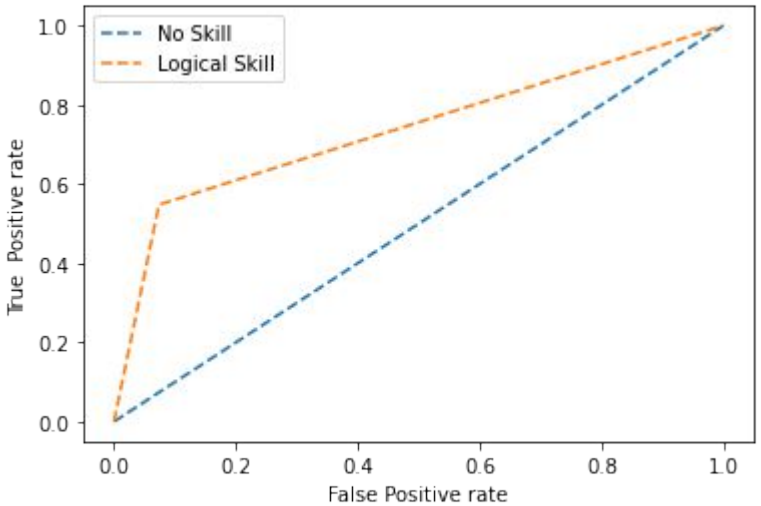
# CLUSTER BASED LOCAL OUTLIER FACTOR




No Skill : ROC AUC = 0.500

Logistic : ROC AUC = 0.958

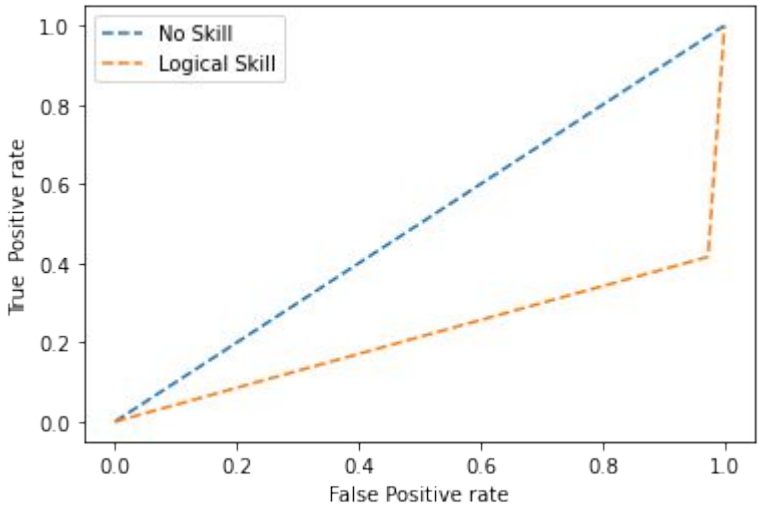
# K NEAREST NEIGHBOURS




No Skill : ROC AUC = 0.500

Logistic : ROC AUC = 0.737

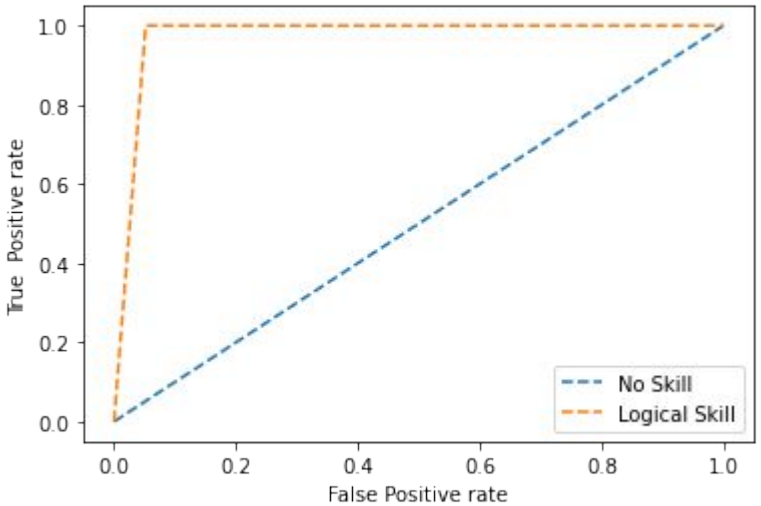
# ONE CLASS SVM




No Skill : ROC AUC = 0.500

Logistic : ROC AUC = 0.222

# ISOLATION FOREST




No Skill : ROC AUC = 0.500

Logistic : ROC AUC = 0.974

THEREFORE MODEL CHOSEN : ISOLATION FOREST

# Model Scores

| Model | Weighted F1 Score |
|---|---|
| Angle Based Outlier Factor | 0.684 |
| Cluster based Local Outlier Factor | 0.958 |
| KNN | 0.737 |
| OneClassSVM | 0.222 |
| Isolation Forest | 0.974 |

# HYPERPARAMETER TUNING OF Isolation Forest

TRIED RandomizedsearchCV with the model, here is the parameter grid:

'n_estimators': [10, 100],

'max_samples': [100, 500, 5,'auto'],

'contamination': [0.1, 0.2, 0.3, 0.4, 0.5],

'max_features': [10,13,5],

'bootstrap': [True, False],

'n_jobs': [5, 10, 20, 30,-1]

Best parameters chosen where :
'n_jobs': -1,

'n_estimators': 10,

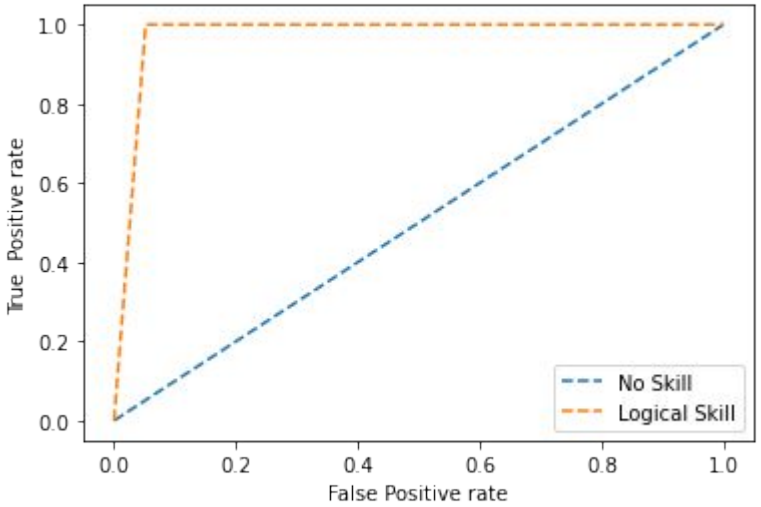
'max_samples': 5,

'max_features': 10,

'contamination': 0.3,

'bootstrap': True

# FINAL MODEL AFTER HYPERPARAMETER TUNING




Classification Report:

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| -1 | 0.00 | 0.00 | 0.00 | 0 |
| 0 | 0.00 | 0.00 | 0.00 | 7943 |
| 1 | 0.02 | 0.24 | 0.03 | 471 |
| accuracy | | | 0.01 | 8414 |
| macro avg | 0.01 | 0.08 | 0.01 | 8414 |
| weighted avg | 0.00 | 0.01 | 0.00 | 8414 |

Precision Score : 0.001
Recall Score : 0.013

No Skill : ROC AUC = 0.500
Logistic : ROC AUC = 0.974

# HOW CAN I IMPROVE FURTHER?

1. Better modelling
2. Better Eda and insight generation
3. Better outlier Detection techniques

# THANKS!