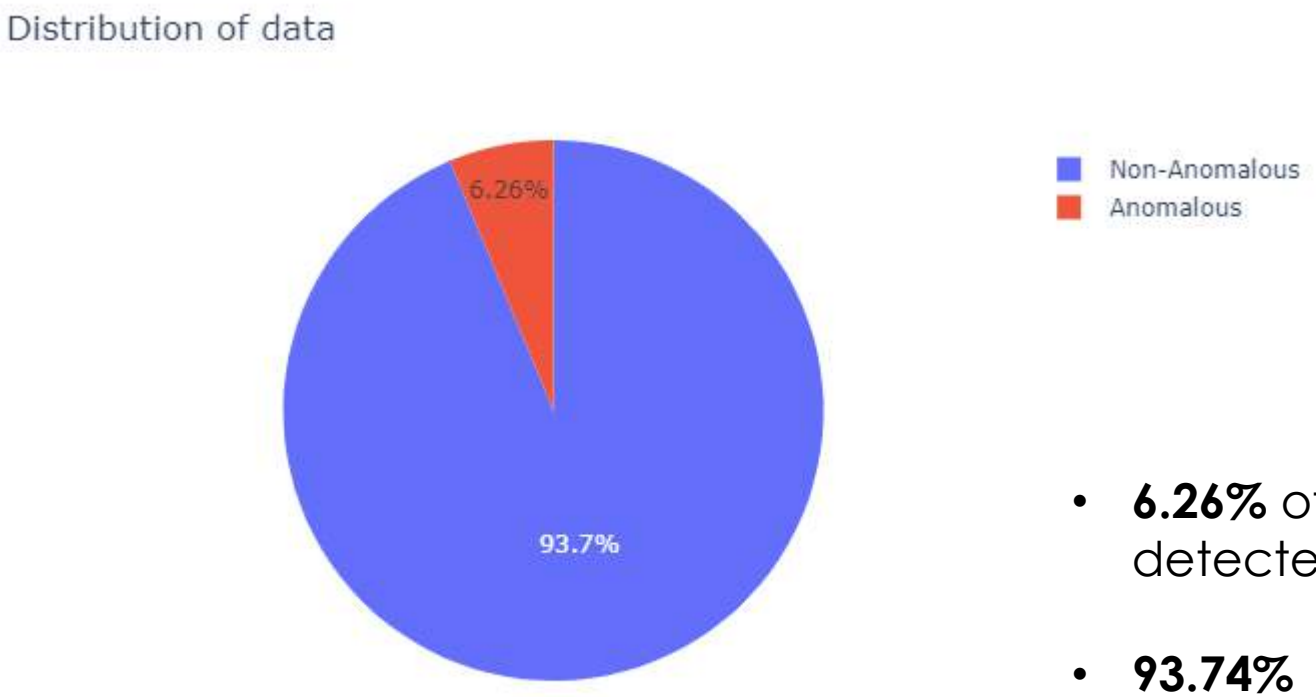
ANOMALY DETECTION

- BY RAYBHAN PAWAR

An Overview of the data

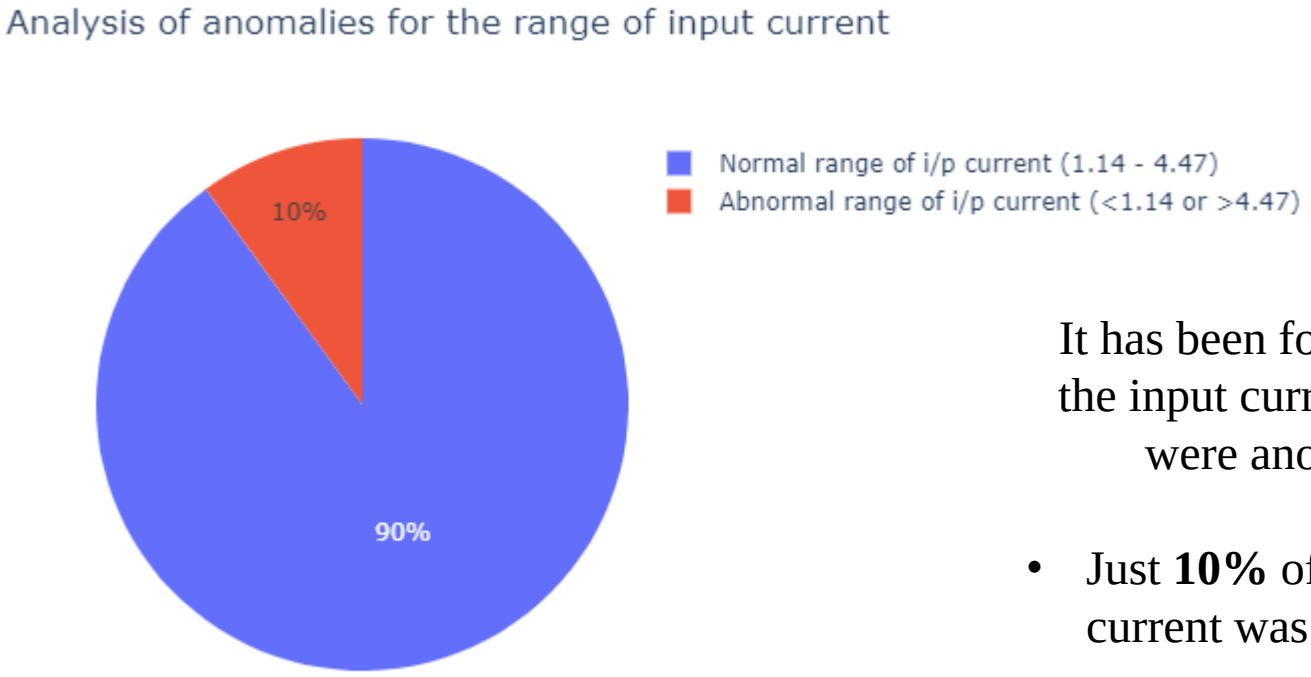
- The dataset consists of 42070 rows and 84 columns.
- 16 columns (timestamp, input current, Heat Sink Temp, Engineering Sensors 1 – 13) have been used in the project for the purpose of Anomaly Detection.
- The sensor readings are captured at an interval of 10 seconds.
- The dataset consists of sensor readings captured for 6 days – 21/Sept/2014 and from 6/Aug/2017 to 10/Aug/2017.

Overall Data Distribution



- **6.26%** of the captured sensor readings have been detected as anomalous.
- **93.74%** of the captured readings are non-anomalous

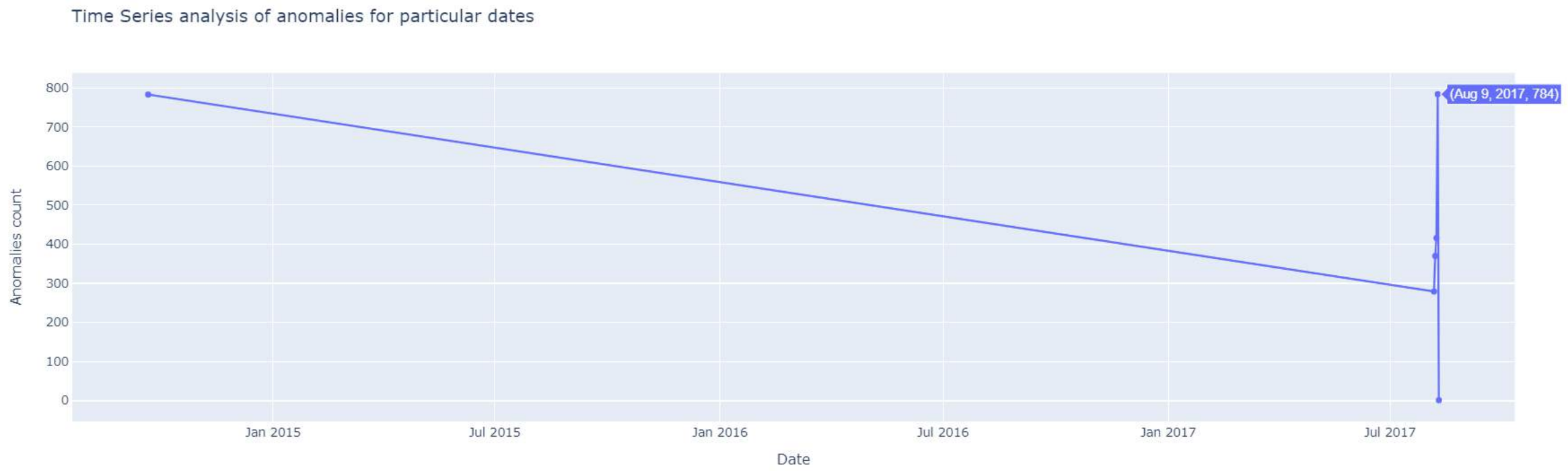
Anomalous Data Distribution



It has been found out that when the readings were non anomalous, the input current range was 1.14 - 4.47 whereas when the readings were anomalous, the input current range was 1.05 – 7.28.

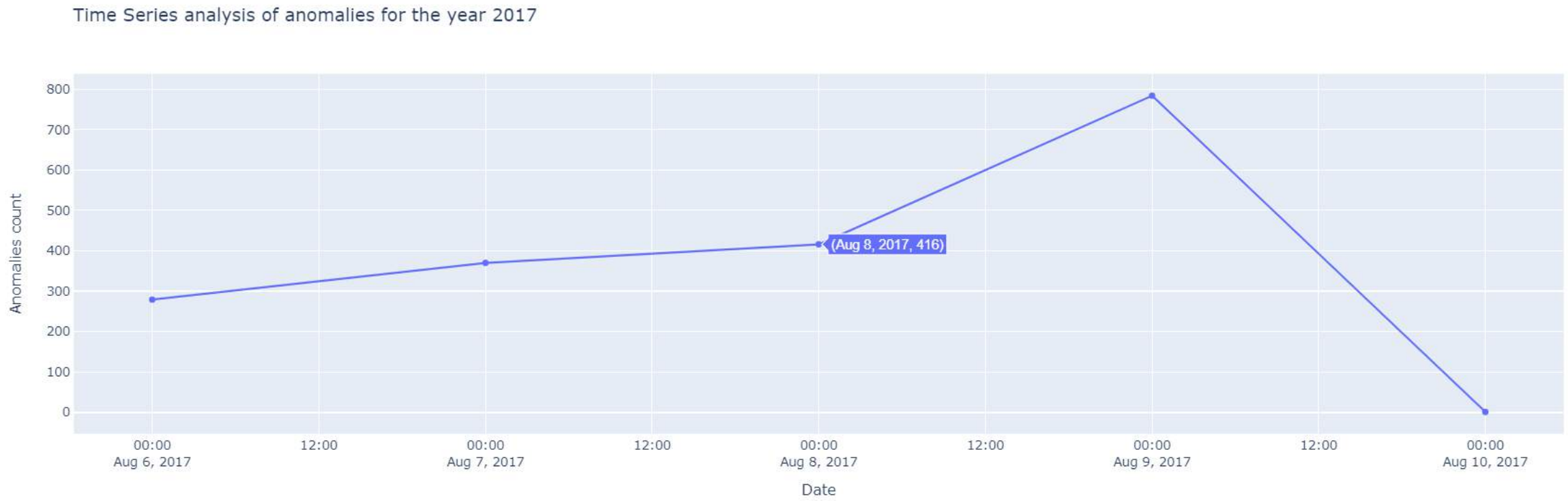
- Just **10%** of the total anomalies were detected when the input current was not in the normal range (1.14 – 4.47).
- Also, it can be inferred that whenever the input current is below 1.14 or 4.47 then the captured reading would be anomalous.

Time Series Analysis of anomalies



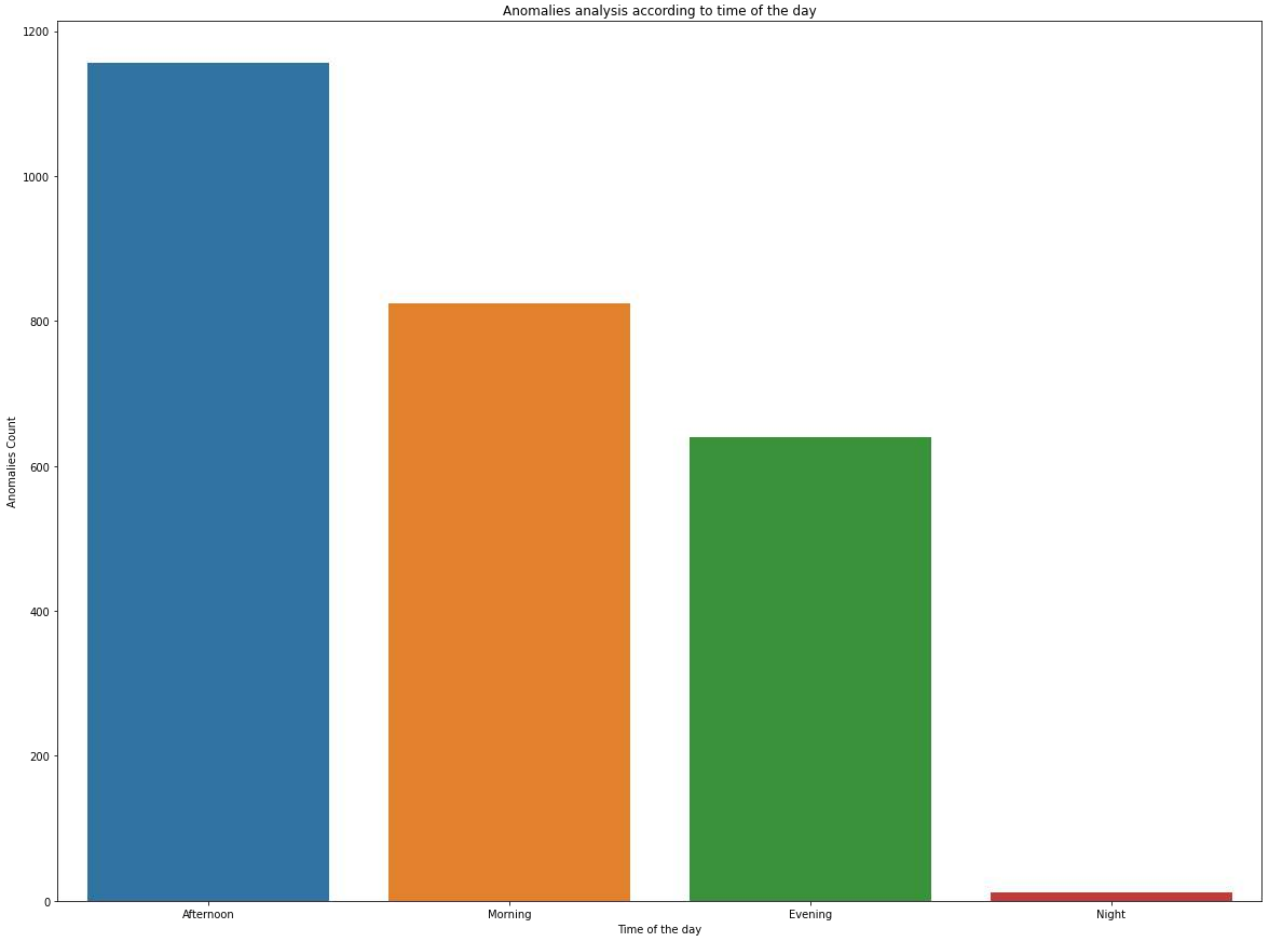
The number of anomalies detected on 21st Sept 2014 were far greater than that detected during the initial days of August 2017.

Time Series Analysis of anomalies for the year 2017



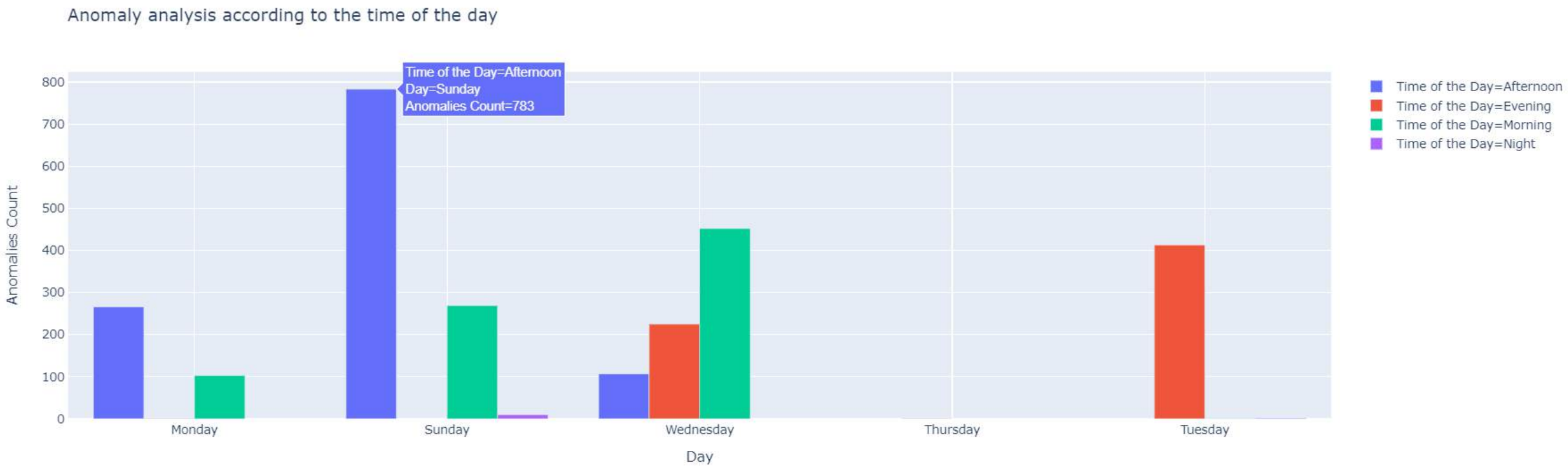
The number of anomalies kept on increasing from 6th to 9th August whereas it reached an all time low on 10th August 2017.

Anomaly Analysis according to the time of the day



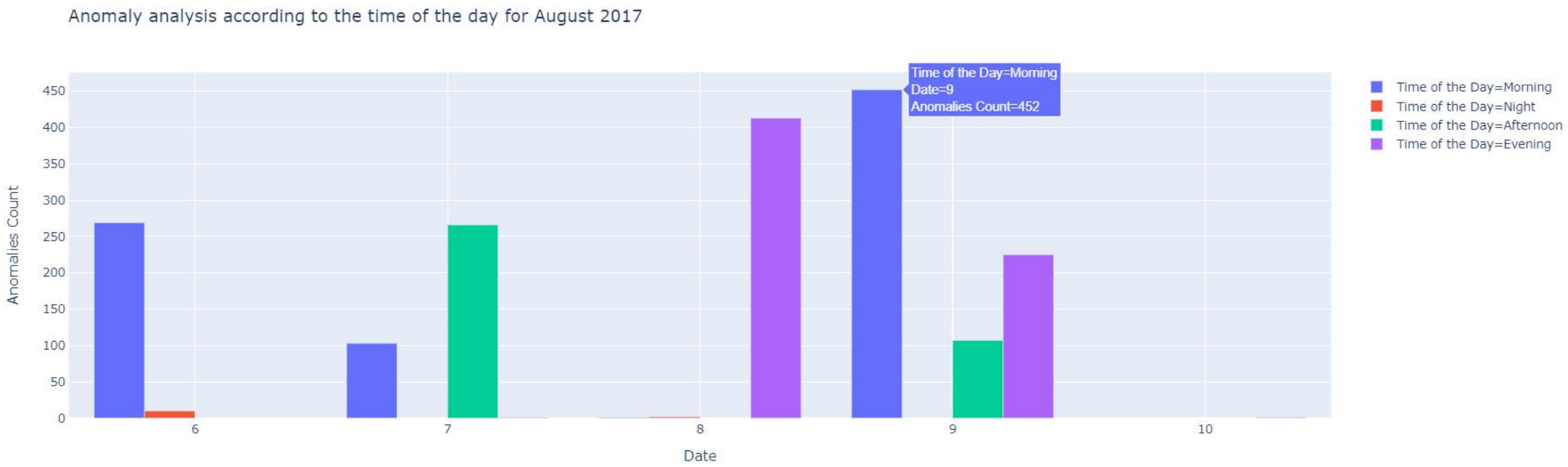
- Most of the anomalies in sensor readings occurred during the afternoon period i.e. 12:00 pm to 4:59 pm.
- The least number of anomalies occur during the night time i.e. during 8:00 pm to 5:59 am.

Anomaly Analysis according to the time of the day for particular days



Sunday afternoon (i.e. between 12:00 pm to 4:59 pm) recorded the highest number of anomalies (783).

Anomaly Analysis according to the time of the day for August 2017



During the 5 days of August most of the anomalies were recorded during the morning period i.e. between 6:00 am to 11:59 am.

Machine Learning

Machine learning models used for predicting anomalies after labelling the data

- K Means Clustering
- Decision Tree Classifier
- Random Forest Classifier
- XGBoost Classifier

Metrics Used for model evaluation

- F1 score
- RMSE
- Accuracy

The Random Forest Classifier model had a better performance amongst others.

Following features were selected which gave the best model performance for RandomForestClassifier

- input current
- Heat Sink Temp
- Engineering Sensors 1
- Engineering Sensors 2
- Engineering Sensors 3
- Engineering Sensors 4
- Engineering Sensors 6
- Engineering Sensors 7
- Engineering Sensors 8
- Engineering Sensors 9
- Engineering Sensors 10
- Engineering Sensors 11
- Engineering Sensors 12
- Engineering Sensors 13
- Hour
- Minutes
- August

(The subset of the original given dataset consisted of 16 features. Thirteen new features were created while feature engineering resulting in a total of 29 features in the dataset)

RandomForestClassifier Model Performance Achieved

F1 Score	RMSE	Accuracy
0.9557	**0.0739**	**0.9945**

Reasons for Feature Selection

- A combination of Correlation score and Feature Importance was used for selecting the features.

For instance

- Of the newly created features, ‘Hour’, ‘Minutes’ and ‘August’ were selected because they had significant feature importance for the model.
- 10 out of 14 of the newly created features were dropped because they had a low feature importance for the model. ‘Date’ was dropped because it was redundant.
- The ‘Engineering Sensors 5’ was dropped because it had a low correlation with the target variable.
- ‘Heat Sink Temp’ and ‘input current’ amongst others had a greater correlation with the target variable.

THANK YOU!