**Base code model**:

step1: collect Phishing Attack Dataset from this URL

[Phishing website Detector | Kaggle](https://www.kaggle.com/eswarchandt/phishing-website-detector)

Step2: Data pre-processing

Step3: feature selection with PCA

Step4: ModelRandom Forest classifier

**Proposed code model**:

step1: collect Phishing Attack Dataset from this URL

[Phishing website Detector | Kaggle](https://www.kaggle.com/eswarchandt/phishing-website-detector)

Step2: Data pre-processing

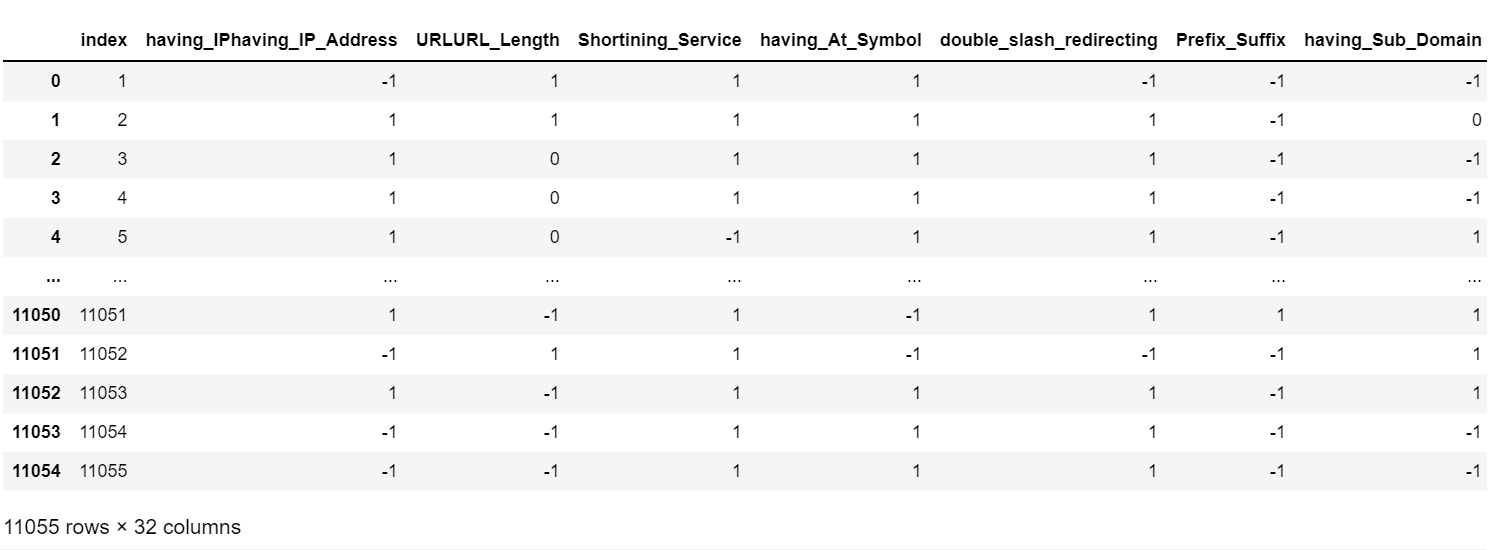
Step3: Feature selection SFS with random forest classifier

Step4: Model: Xgbhoost classifier, Light Gradient Boosted Machine

**Base code:**

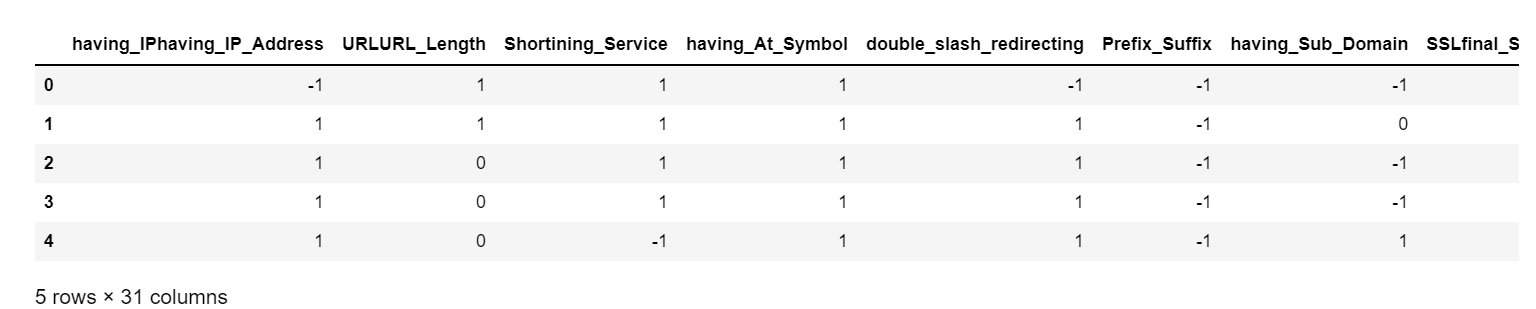
**Collect dataset**

Phishing attack dataset collected from Kaggle

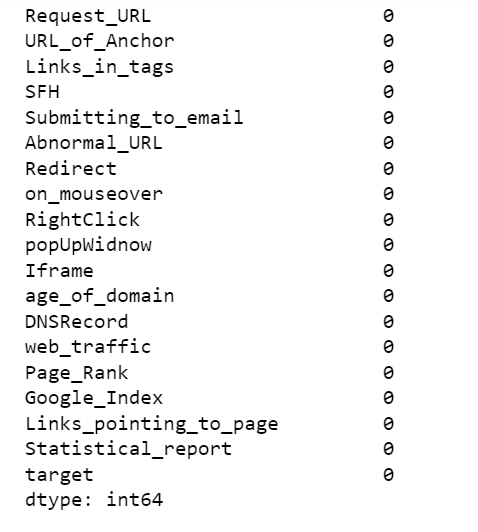
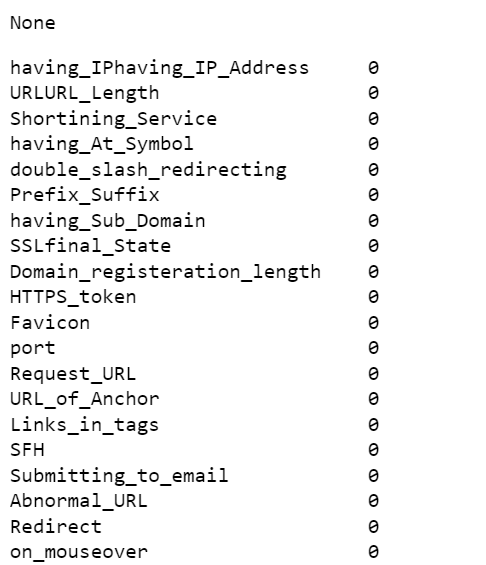


**Original Phishing attack dataset**

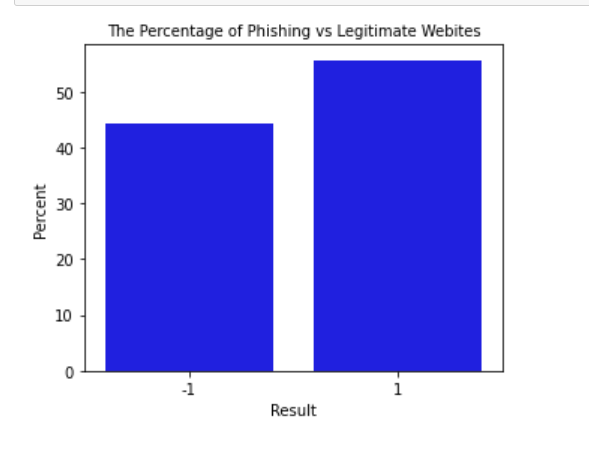
**Data pre-processing**

****

**Drop Index column**

****

**Check null value**

****

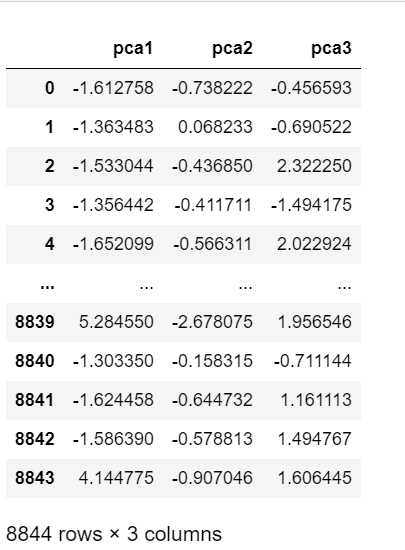
**Fig.1**

**-1 and 1. As a result, -1 denotes phishing websites and 1 denotes legitimate websites.**

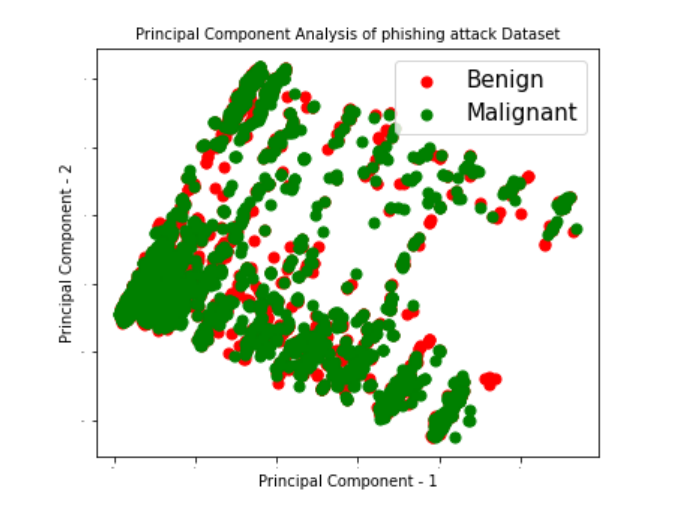
****

**Dataset split into train and test set**

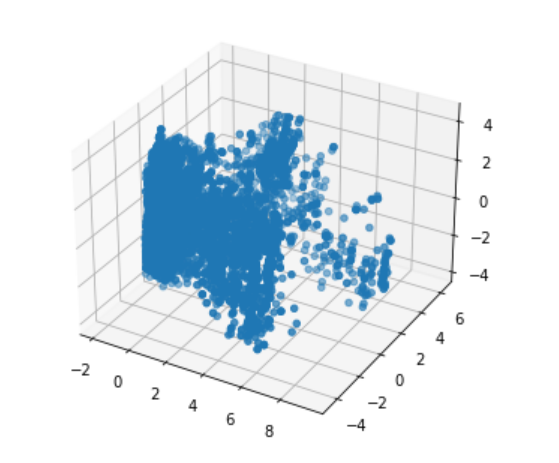
**Feature selection: PCA**

****

**Feature selection PCA**

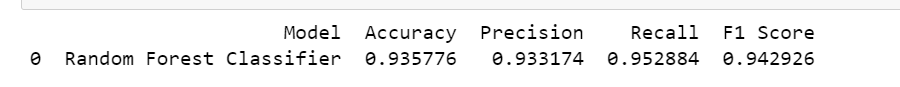
****

**Two principal Component**

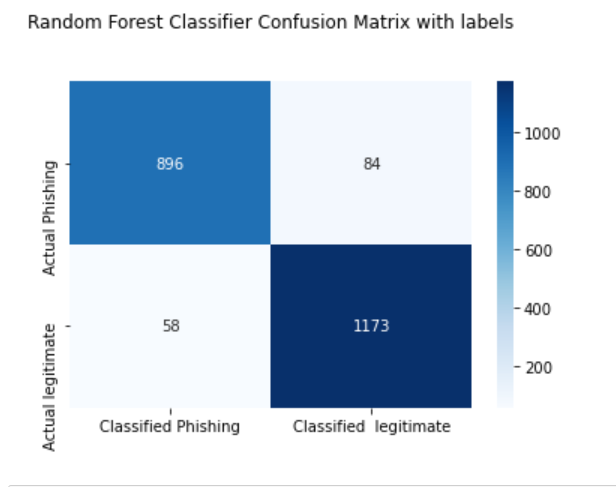
****

**3D Scatter by PCA**

**Model:  Random Forest Classifier**

****

**Classification report display, precision, recall. F1-score, support**

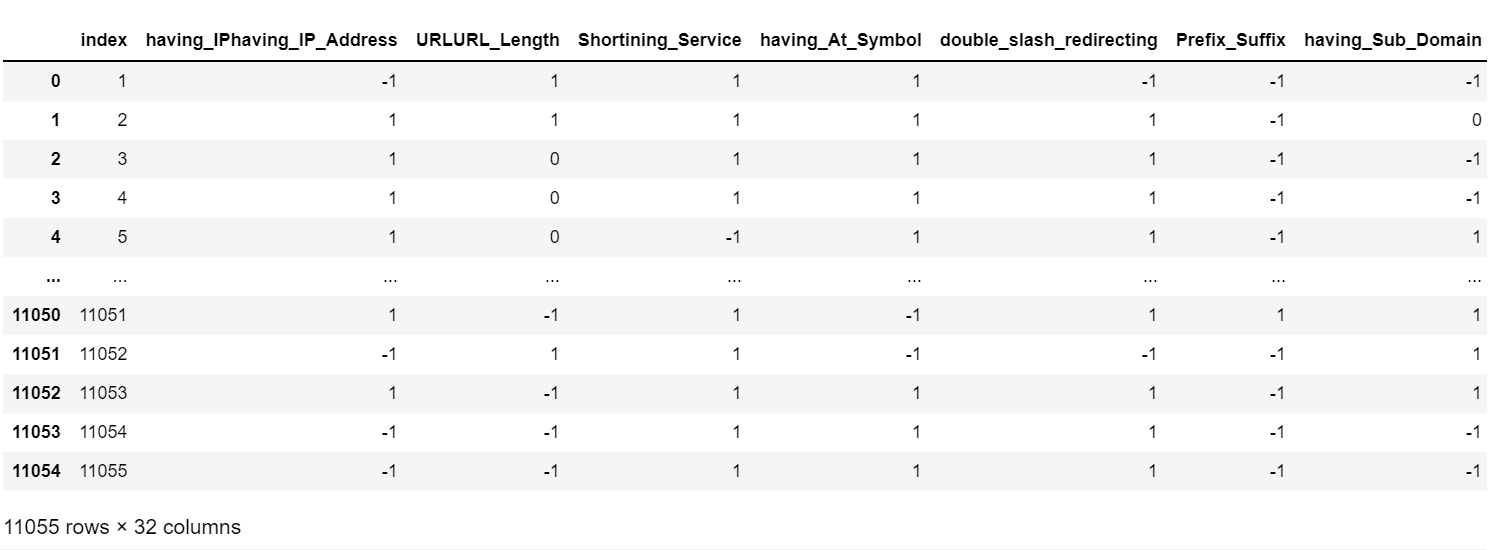
****

**Random forest Classifier Confusion Matrix**

**Proposed code:**

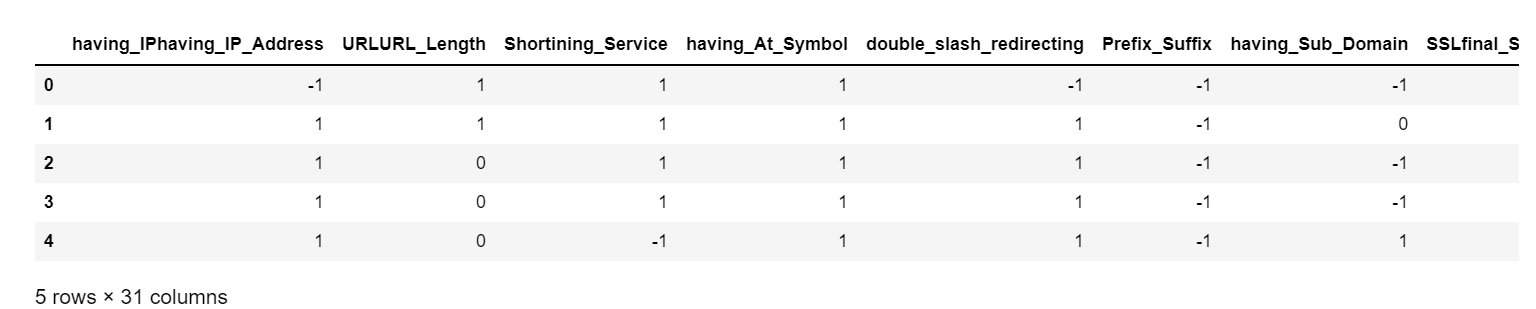
**Collect dataset**

Phishing attack dataset collected from Kaggle

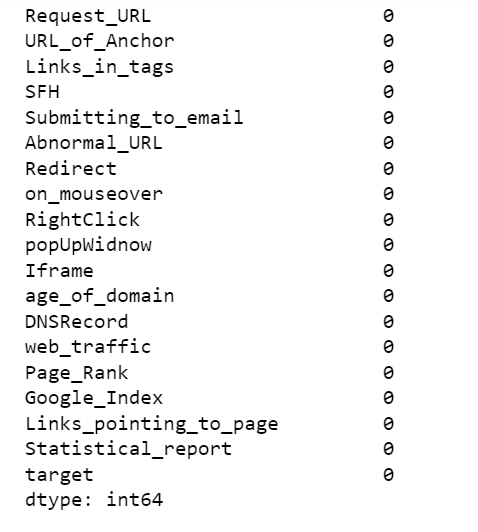
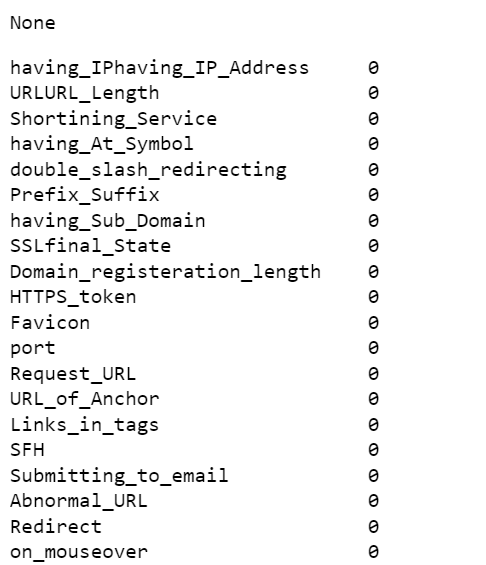


**Original Phishing attack dataset**

**Data pre-pre-processing**

****

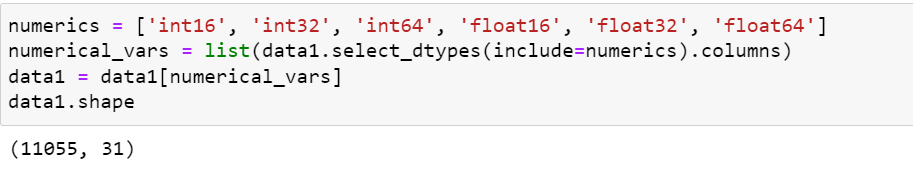
**Drop Index column**

****

**Check null value**

**Feature selection: Sfs with random forest classifier**

**Step1: I will only utilize numerical variables to choose numerical columns**

****

**Covert float value into an integer**

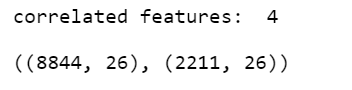
**Step2: split dataset into training and testing part**

****

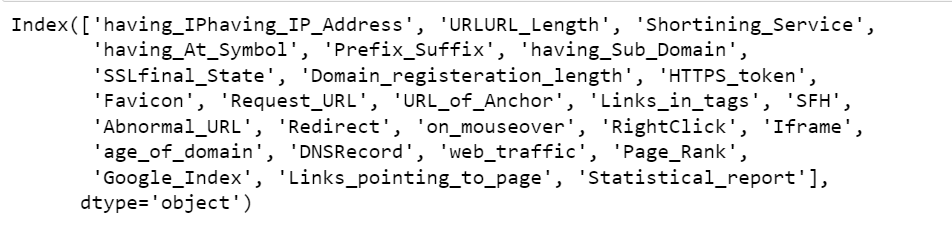
**Step3:**

**1) find out correlated features**

**2) removed correlated features**

****

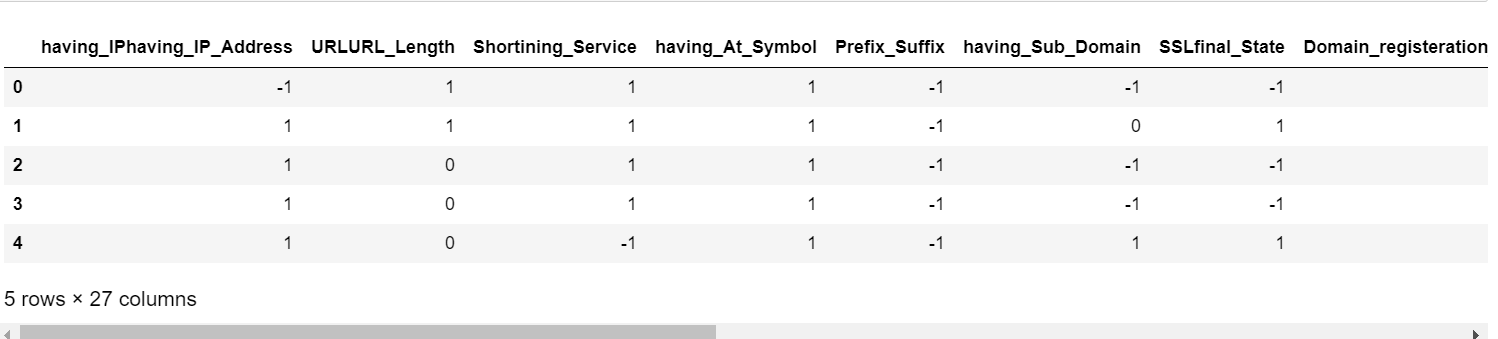
**Select 26 features. Next, we'll create a classifier and a sequential forward feature selector with a random forest classifier.**

****

**Feature Selected**

**Again, working for model with important feature**

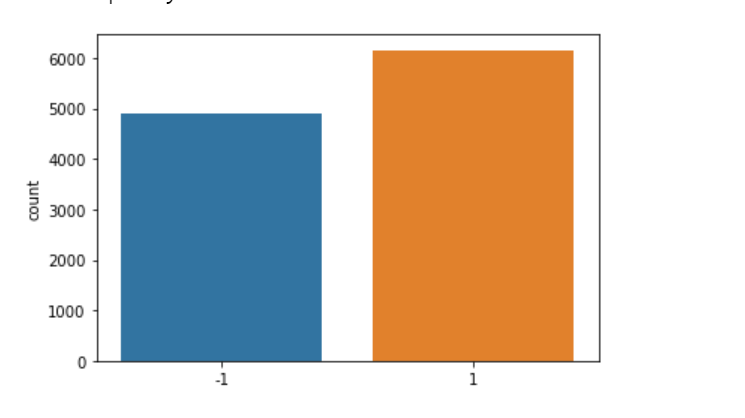
**Step1: upload an original dataset and select important features to find out using Sfs with random forest classifier to make a new CSV file.**

****

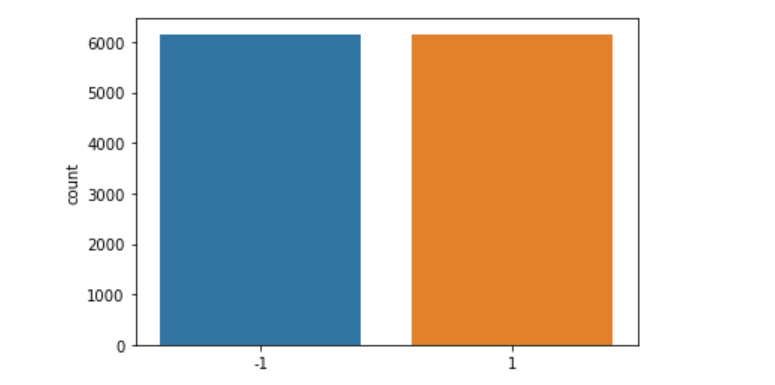
**New feature selected**

**Unbalanced target value, as you can see from the figure above, there are two target values:**

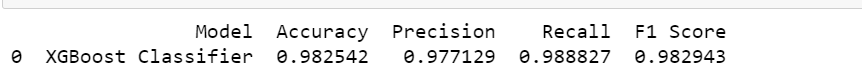
**-1 and 1. As a result, -1 denotes phishing websites and 1 denotes legitimate websites.**

****

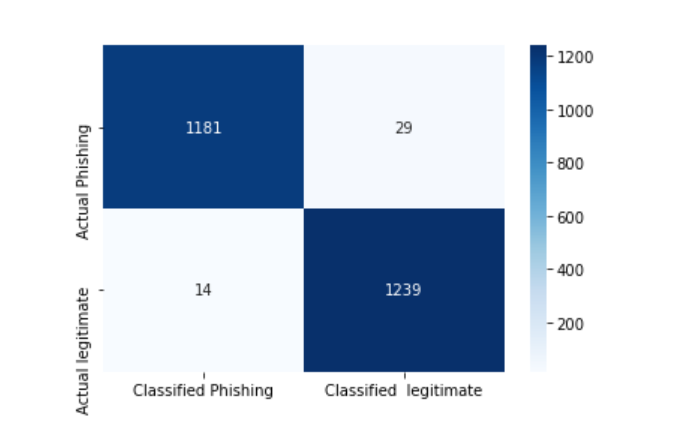
**For balanced dataset, I used SMOTE technique**

****

**Model: XGBClassifier**

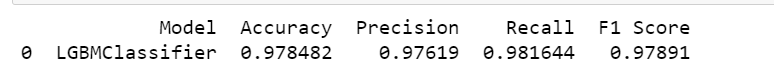
****

**Classification report display, Accuracy, precision, recall. F1-score, support**

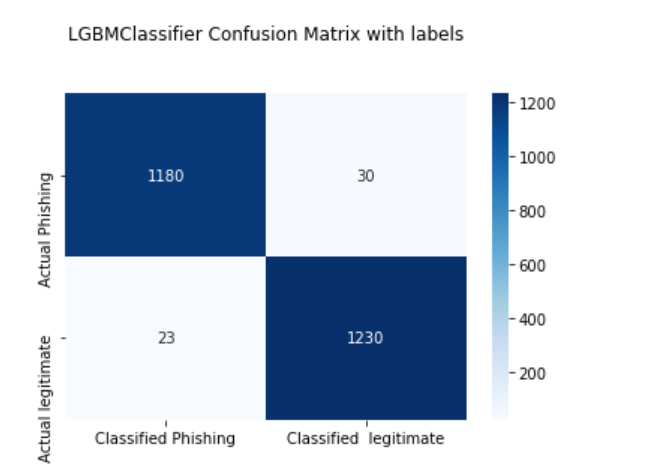
****

**Xgbhoost Classifier Confusion Matrix**

**LGBM Classifier:**

****

**Classification report display, Accuracy, precision, recall. F1-score, support**

****

**LGBM Classifier Confusion Matrix**

**Proposed Model:**

|  |  |  |  |
| --- | --- | --- | --- |
| Based code | | Proposed code | |
| Result | Random forest classifier | **Xgbhoost classifier** | **Light Gradient Boosted Machine** |
| Accuracy | 94% | 98% | 97% |
| precision | 93 | 97 | 97 |
| recall | 95 | 98 | 98 |
| f1-score | 94 | 98 | 97 |