# Classifying Phishing Websites

UMBC

Akash Selvakumar | Drishti Arora | Parth Soni
Advisor : James Foulds
Course : IS 733 Data Mining

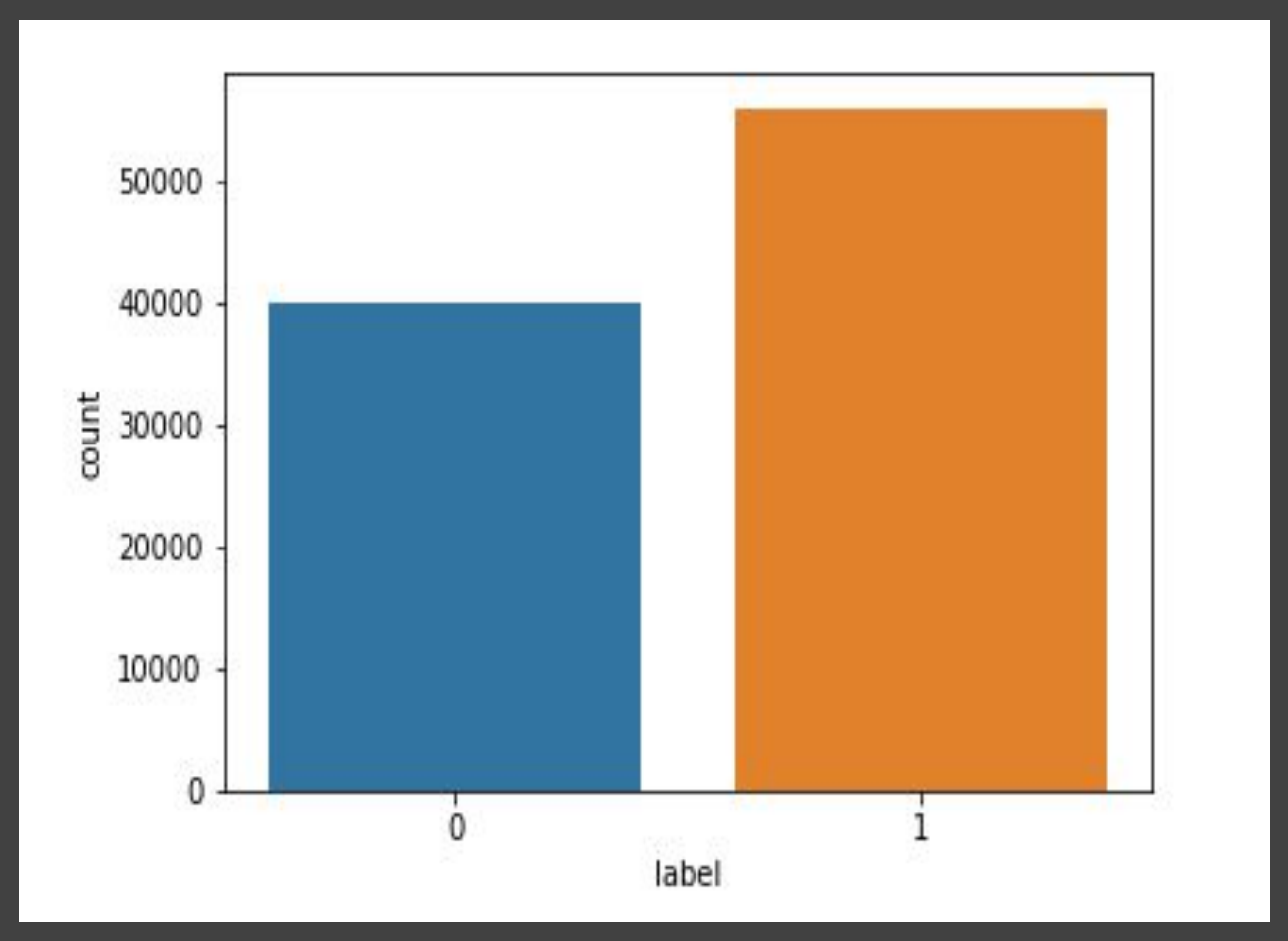
## Introduction

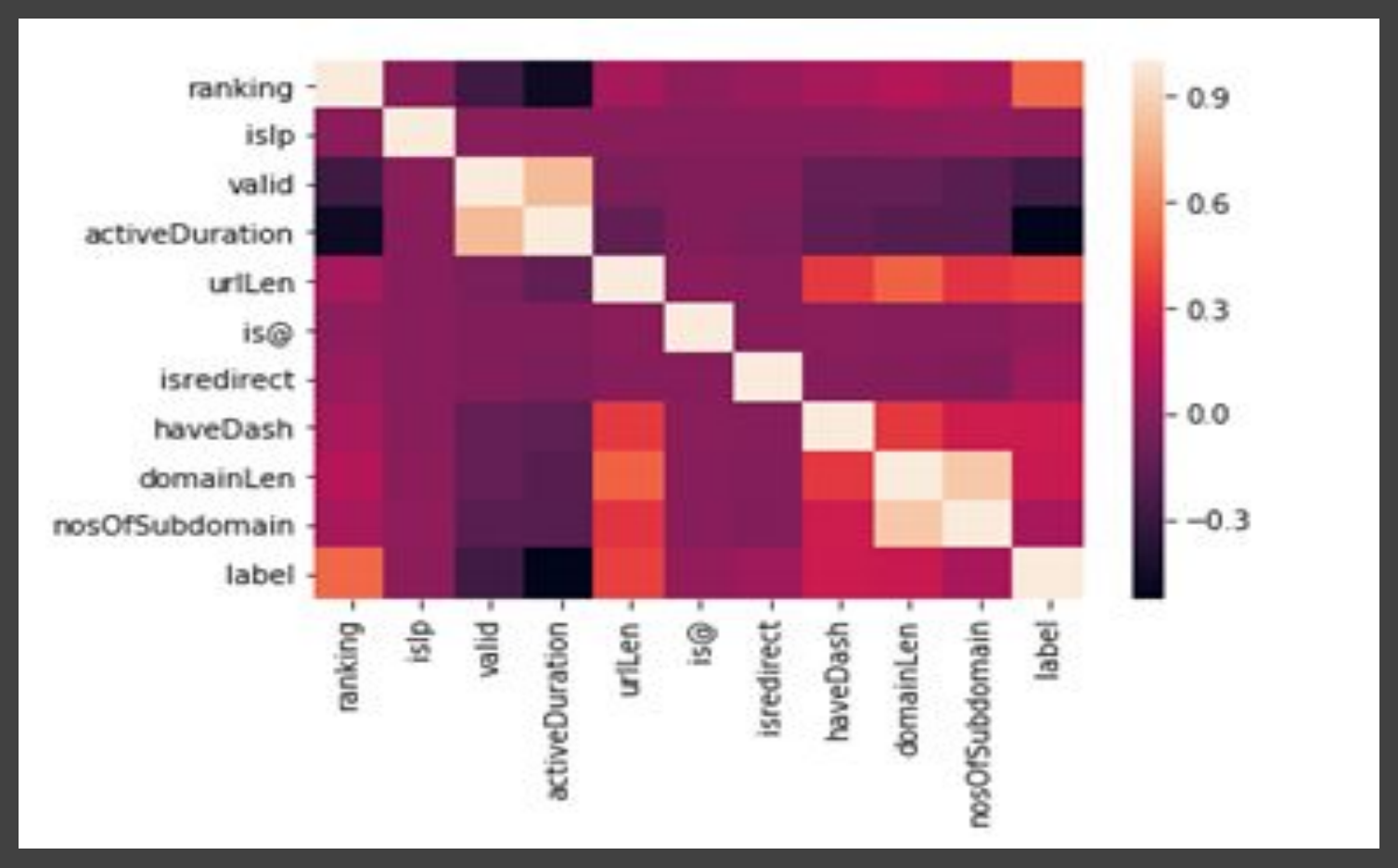
With the sudden growth in technology, e-commerce has become a necessary part and phishing has become a common practice for fishers to gain personal information, passwords etc of the user through spoofed emails or phishing software. This project will classify the websites as legitimate or non-legitimate based on their URLs using machine learning techniques with the help of few parameters like page ranking, IP address in the link, URL length, if the link has double dashes, domain length, number of subdomains etc. The fine-tuned parameters help in selecting the most appropriate machine learning model for classifying phishing websites.

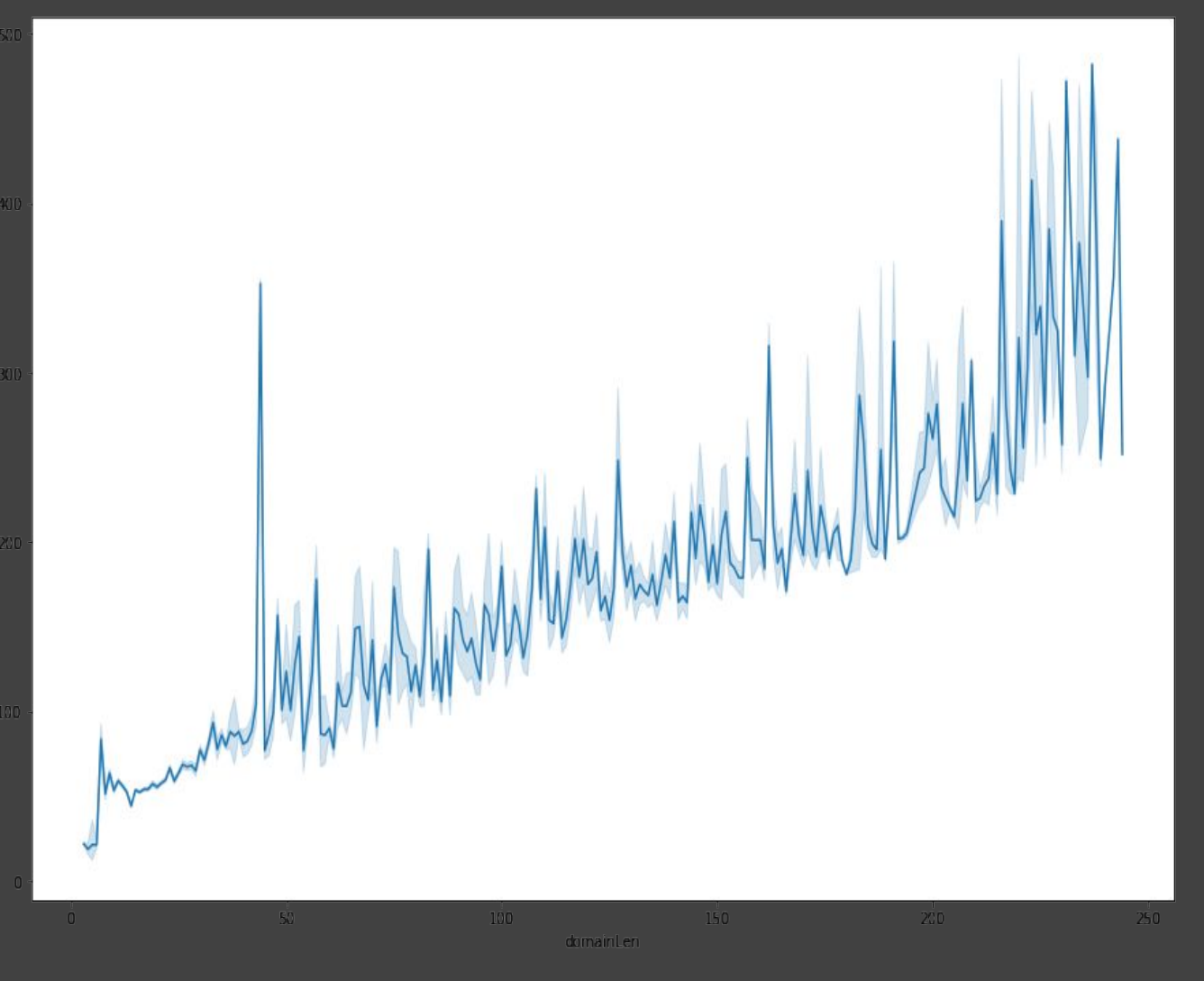
## Data Pre-processing

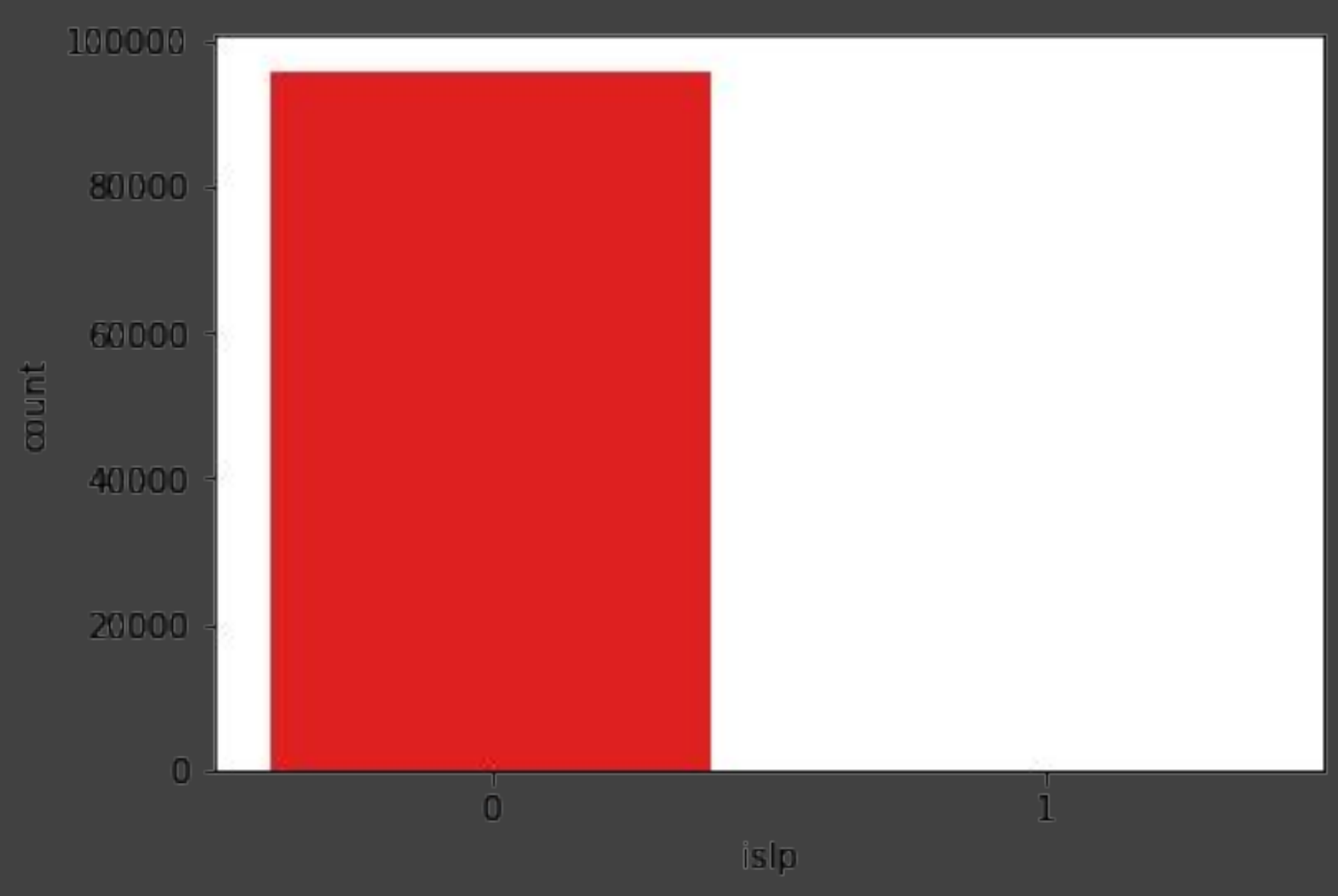
- Checked for null values and categorical values.
- Removed special characters and converted all characters to small letters so that machine learning models can perform better.
- Checked correlation of the parameters to the target/label to ignore the lowest ranked parameters.

## Exploratory Data Analysis



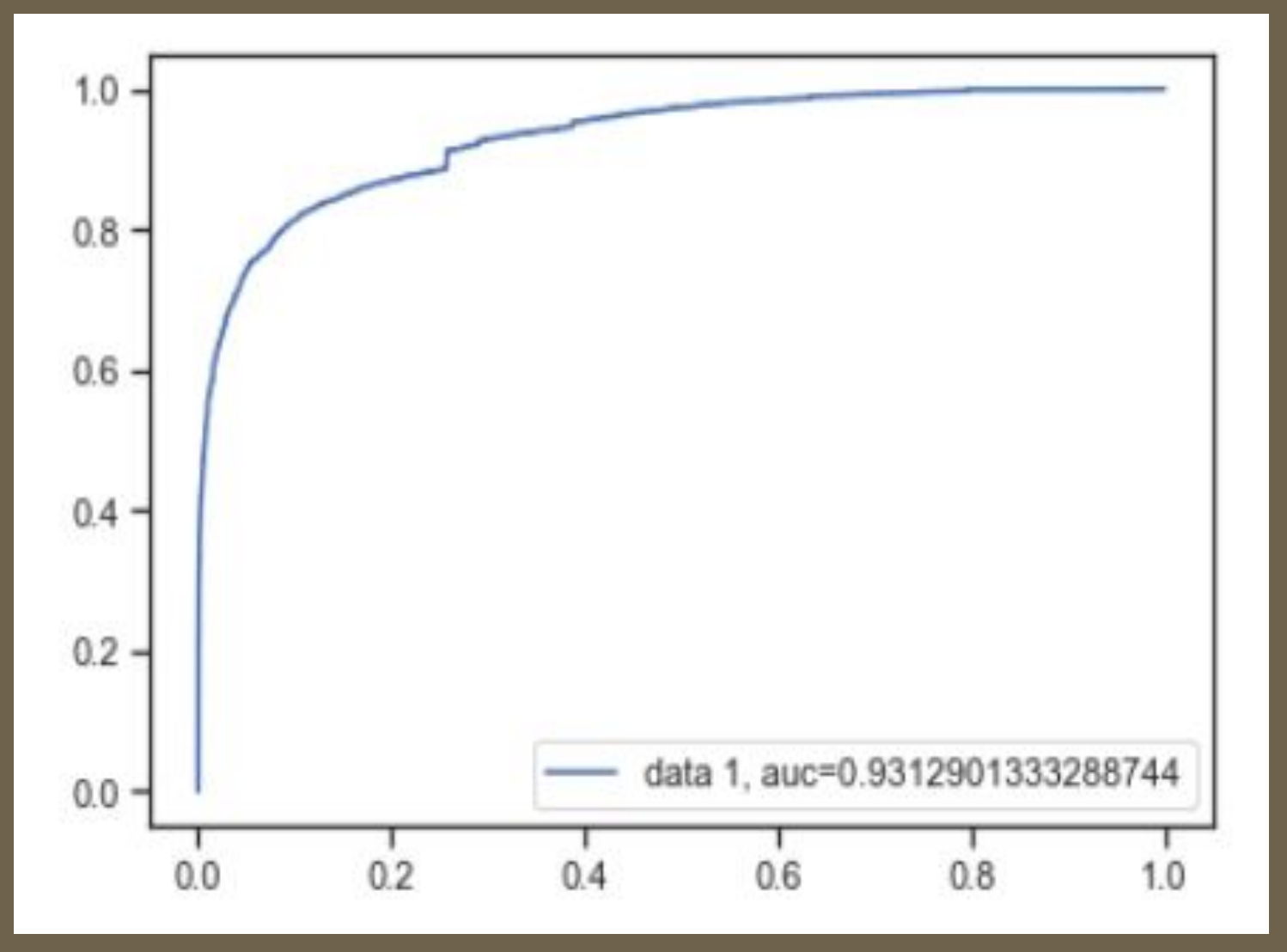




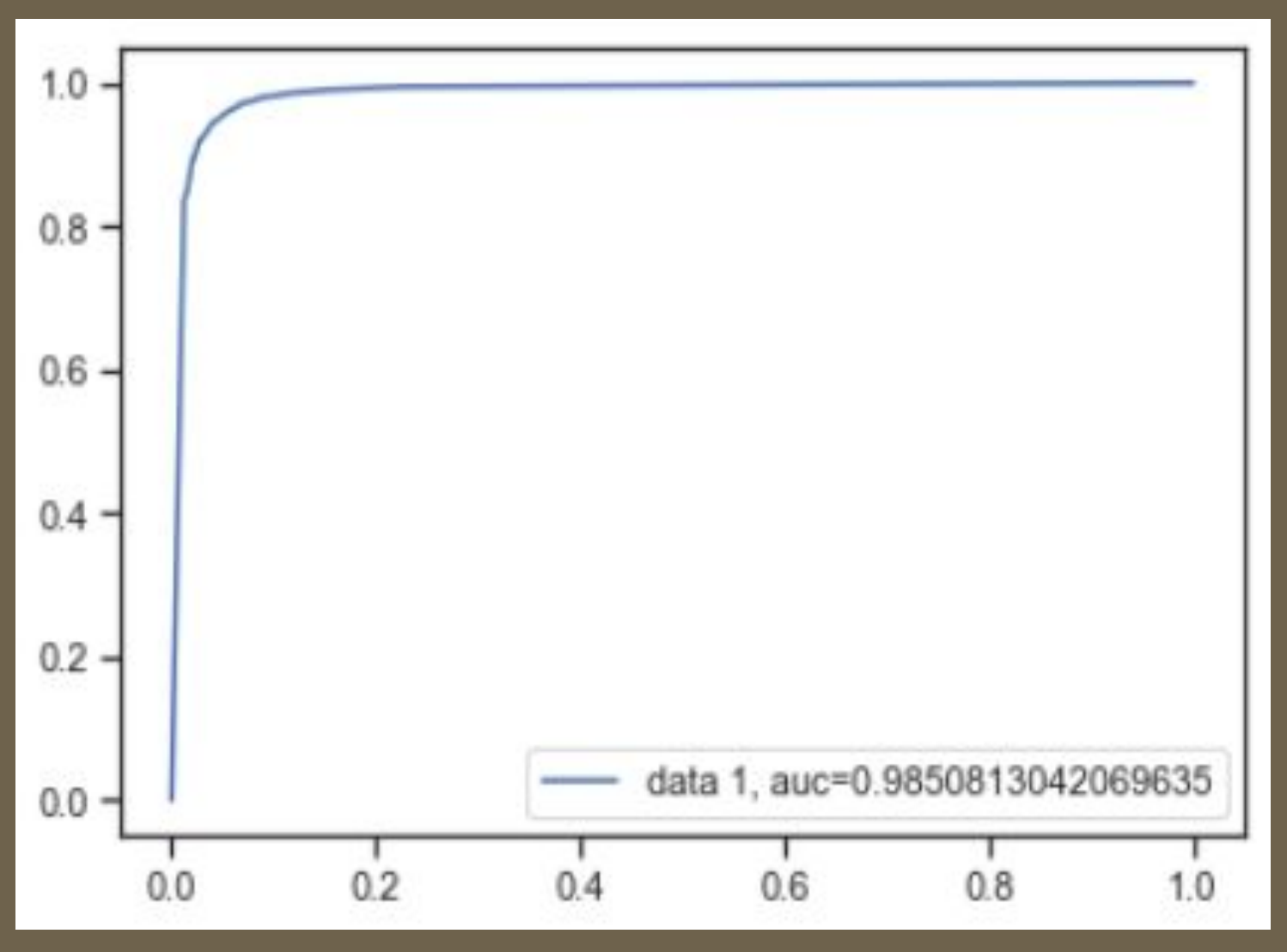


## Experimental Results

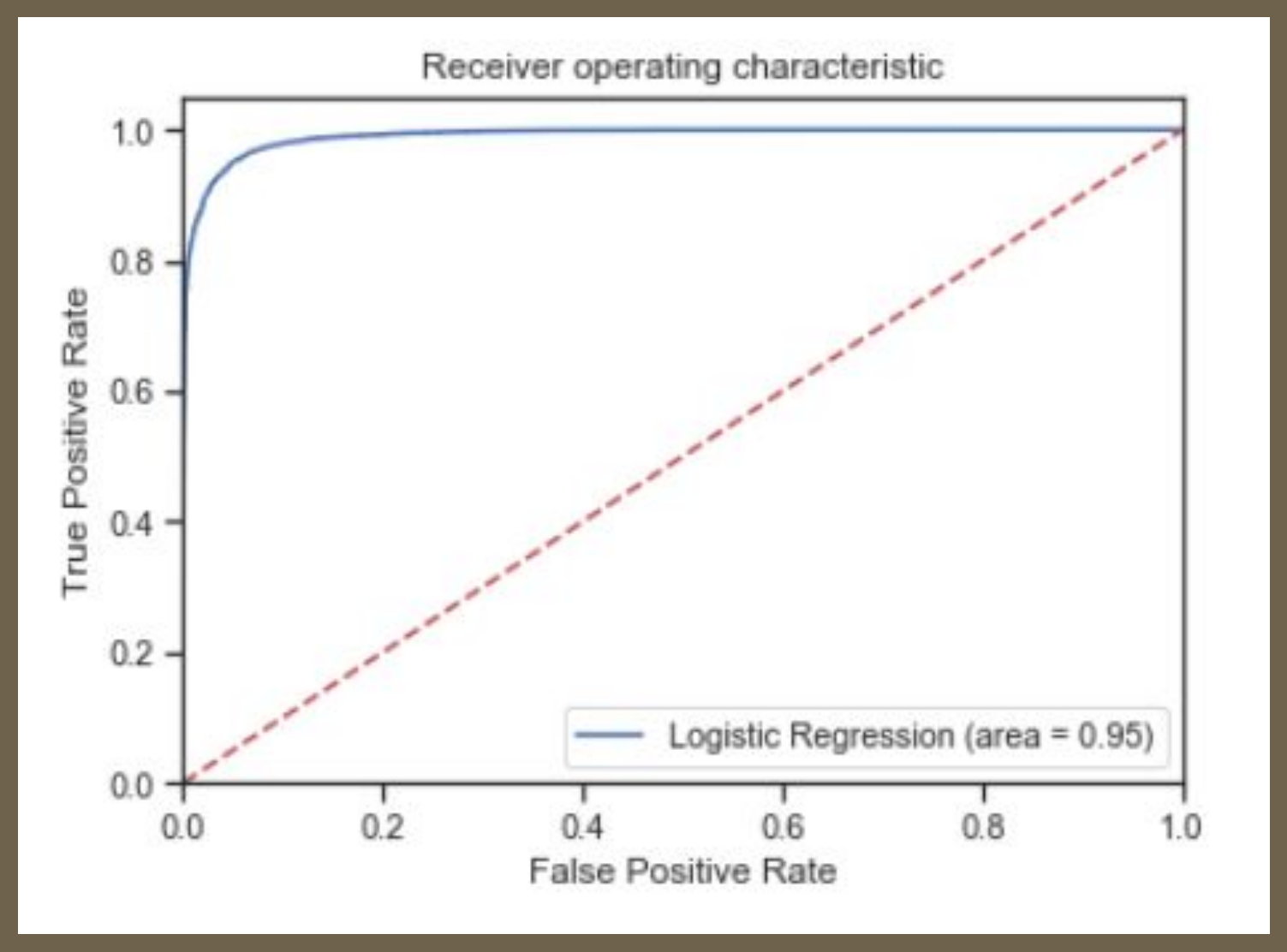
### Logistic Regression Classifier:



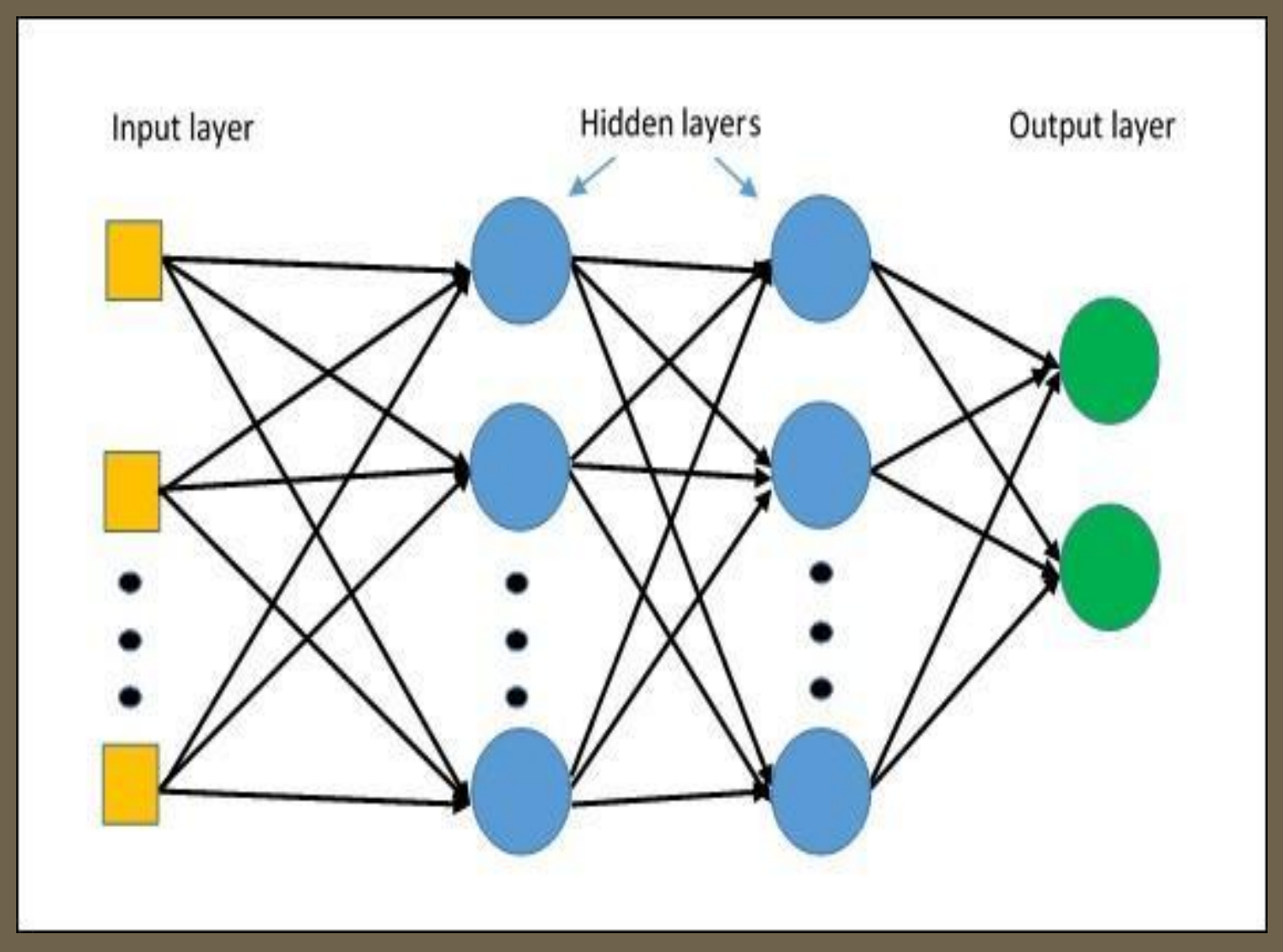
### Random Forest Classifier:



### XGBoost Classifier:



### Multiple Layer Perceptron Classifier:



## Methods

1. Logistic Regression Classifier:

Logistic regression models the data using the sigmoid function. The sigmoid function gives an ‘S’ shaped curve that can take any real-valued number and map it into a value between 0 and 1 for a particular instance. All predictions having a sigmoid function value >= 0.5 can be considered as phishing websites and values < 0.5 can be considered as legitimate websitess.

2. Random Forest Classifier:

It analyses the features of the given data by splitting them into numerous decision trees. Each decision tree result is obtained and based on the majority votes the final result is calculated. The accuracy increases as the number of trees in the forest increases.

3. XGBoost Classifier:

XGBoost (Extreme Gradient Boosting) belongs to a family of boosting algorithms and uses the gradient boosting (GBM) framework at its core. It uses gradient boosting decision tree algorithm to combine a set of weak learners and delivers improved prediction accuracy. Stronger models are created by capitalizing on the misclassification errors produced by the previous model.

4. Multiple Layer Perceptron Classifier:

It is a Multi-Layered feed-forward neural network where the input nodes each have a numerical value. Each layer has a connection to the next layer which transmits the weights of the nodes which is used to calculate the weighted sum The activation function based on the results of the weighted sum will predict the website's legitimacy.

## Conclusion:

| S. no | Models | Results |
|---|---|---|
| 1 | Logistic Regression | Default accuracy: 72.97 %<br>Hyper-parameter tuned accuracy: 83.85 % |
| 2 | Random Forest | Default accuracy: 88.98 %<br>Hyper-parameter tuned accuracy: 95.36 % |
| 3 | XGBoost | Default accuracy: 90.38 %<br>Hyper-parameter tuned accuracy: 95.14 % |
| 4 | MLP | Default accuracy: 81.04 %<br>Hyper-parameter tuned accuracy: 86.27 % |

Therefore, we can see that tuned Random Forest classifier has the highest accuracy of 95.36 % in detecting phishing websites compared to other classifiers such as Logistic Regression, XGBoost and MLP.

## Future Implementations:

- Additional features like Website’s domain can also be analyzed and applied to these classifiers.
- Nested Cross-validation can be used for hyper-parameter tuning to improve the performance of the models.