Classifying Phishing Websites

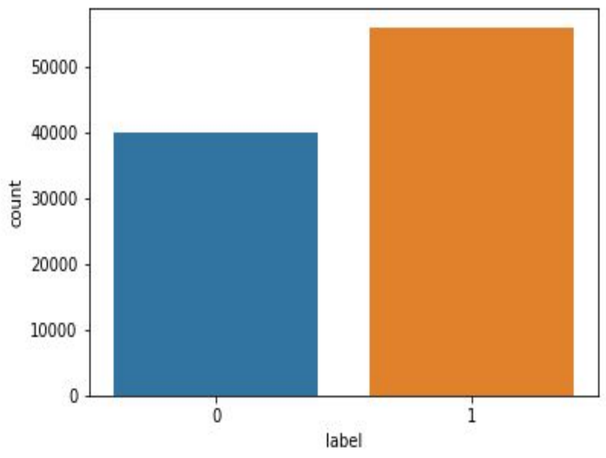
Akash Selvakumar | Drishti Arora | Parth Soni

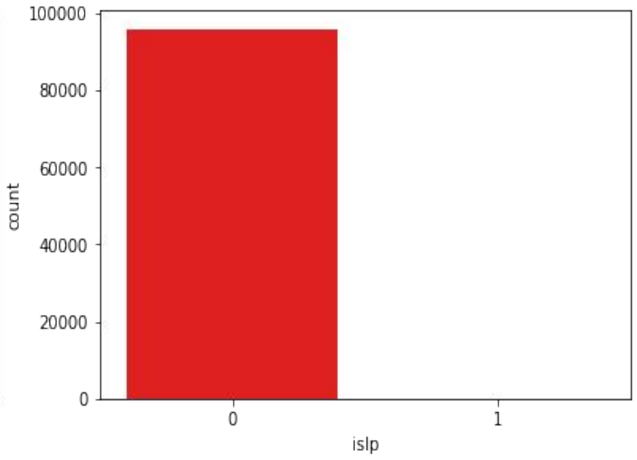
Course : IS 733 Data Mining
Advisor : James Foulds

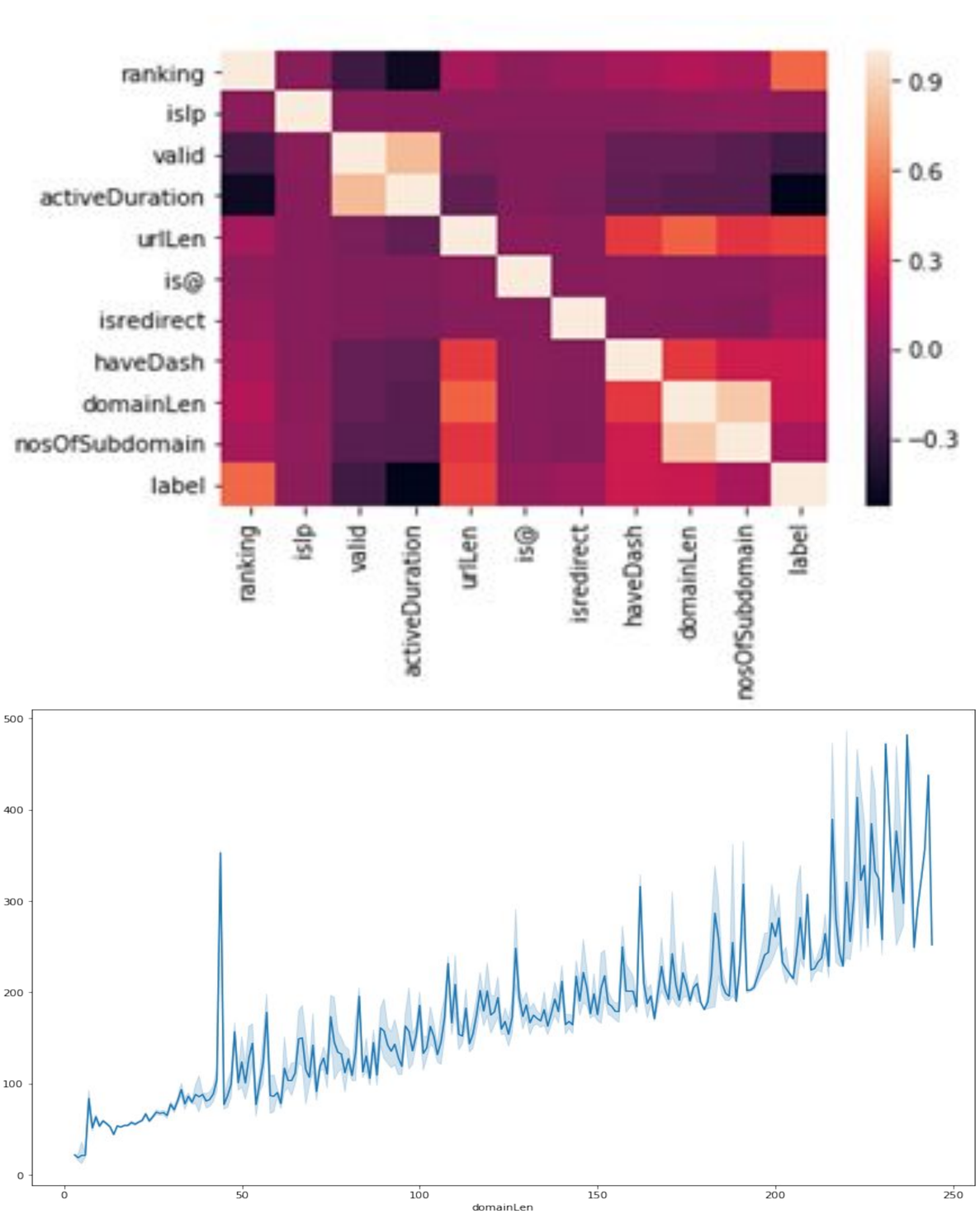
Introduction

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.

Exploratory Data Analysis

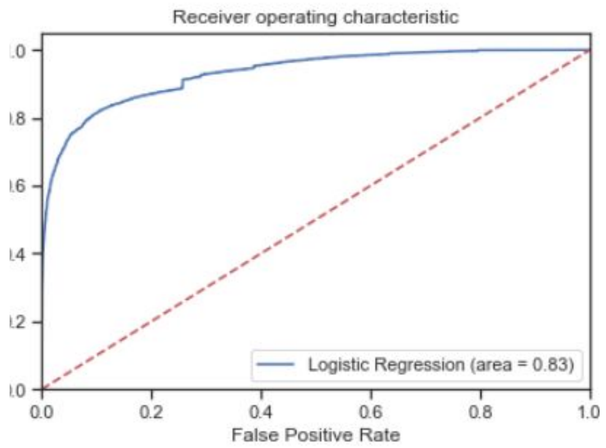




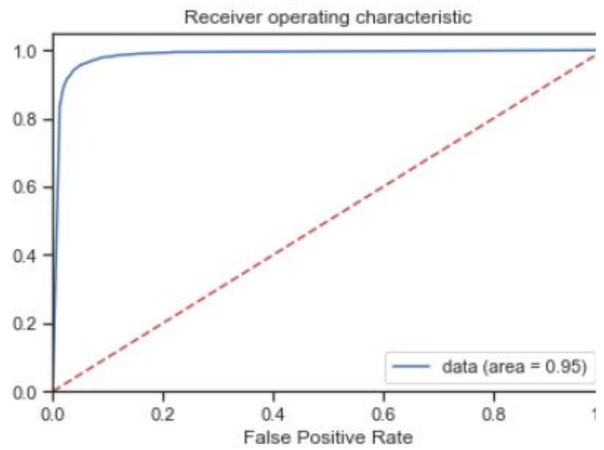


Experimental Results

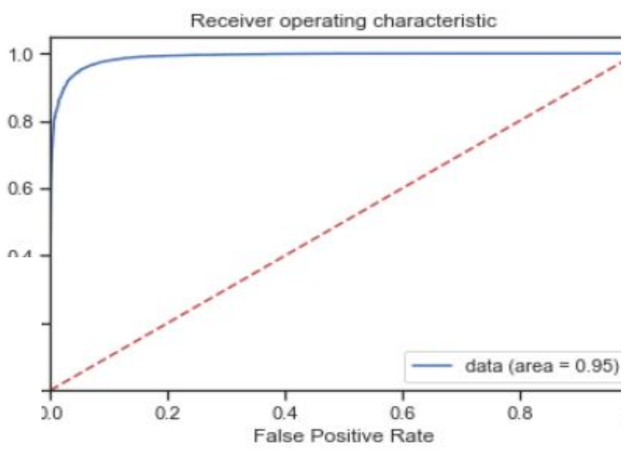
1) Logistic Regression Classifier:



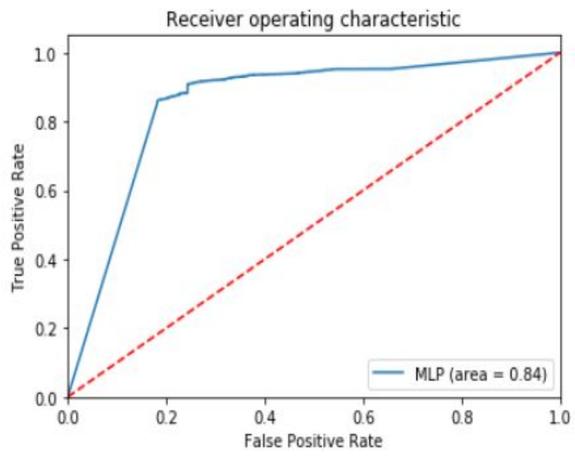
2) Random Forest Classifier:



3) XGBoost Classifier:



4) Multiple Layer Perceptron Classifier:



Model Results::

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

Conclusion :

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.