DOCUMENTATION

**INTRODUCTION:**

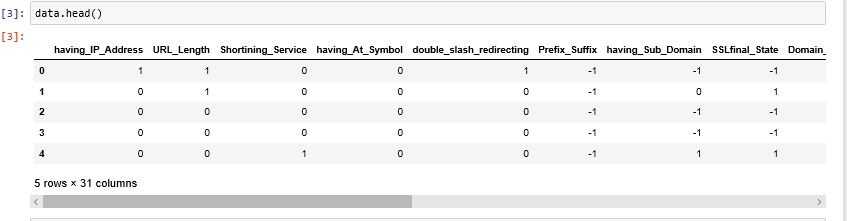
I have built a classification model for classifying whether a website is a phishing website or not. The models used were all supervised learning algorithms.

The packages used are:

* SKLEARN
* PANDAS
* NUMPY
* XGBOOST
* PLOTLY

**ABOUT DATA:**

The Phishing dataset used for creating the model is having 30 variables and 1 target variable. There are 2456 records in total.



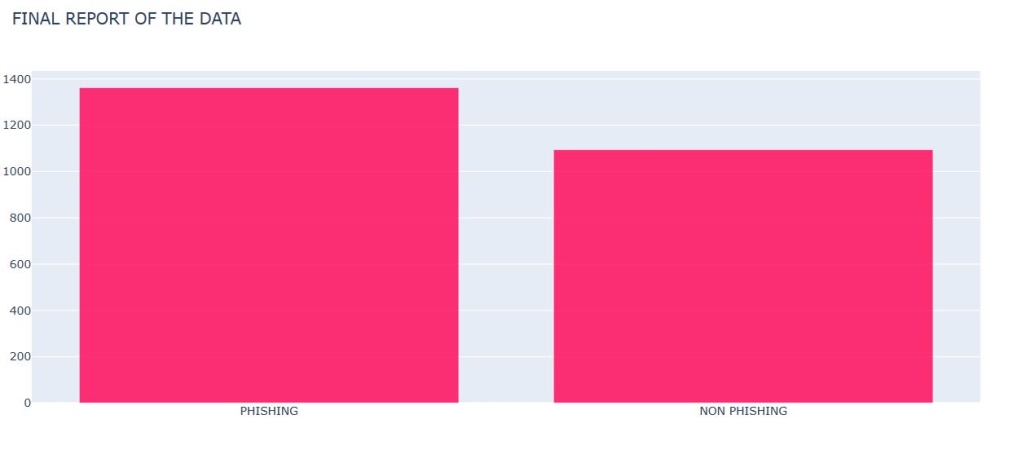
The variables present in the data are as follows:



It is noted that there was no missing value present in the data. The target variable in the dataset consist of two classes which are as follows:

* 1 indicates the website is a phishing website.
* 0 indicates the website is a non-phishing website.

The following plot gives better understanding of the target variable:



**METHODOLOGY USED:**

1.The dataset was divided into two parts: Feature and Target variable.

2.Then the dataset was divided into train data, test data and validation data.

3.The following classification models were used for training, testing and validation:

* Support vector classifier (linear kernel)
* Support vector classifier(rbf kernel)
* K nearest neighbor
* Logistic regression
* Naïve Bayes
* Decision tree classifier
* Random forest
* XGboost classifier

The accuracy and the corresponding validation accuracy for each model is shown in the following table**:**

|  |  |  |  |
| --- | --- | --- | --- |
| S.no | MODEL | ACCURACY (%) | VALIDATION ACCURACY (%) |
| 1 | Support vector classifier (linear) | 92.40 | 94.18 |
| 2 | Support vector classifier(rbf) | 92.67 | 94.37 |
| 3 | Logistic regression | 92.13 | 94.57 |
| 4 | K nearest neighbor | 95.43 | 93.6 |
| 5 | Naïve Bayes | 92.19 | 89.6 |
| 6 | Decision trees | 99.75 | 95.93 |
| 7 | Random forest | 96.76 | 95.54 |
| 8 | XG Boost | 97.59 | 96.32 |

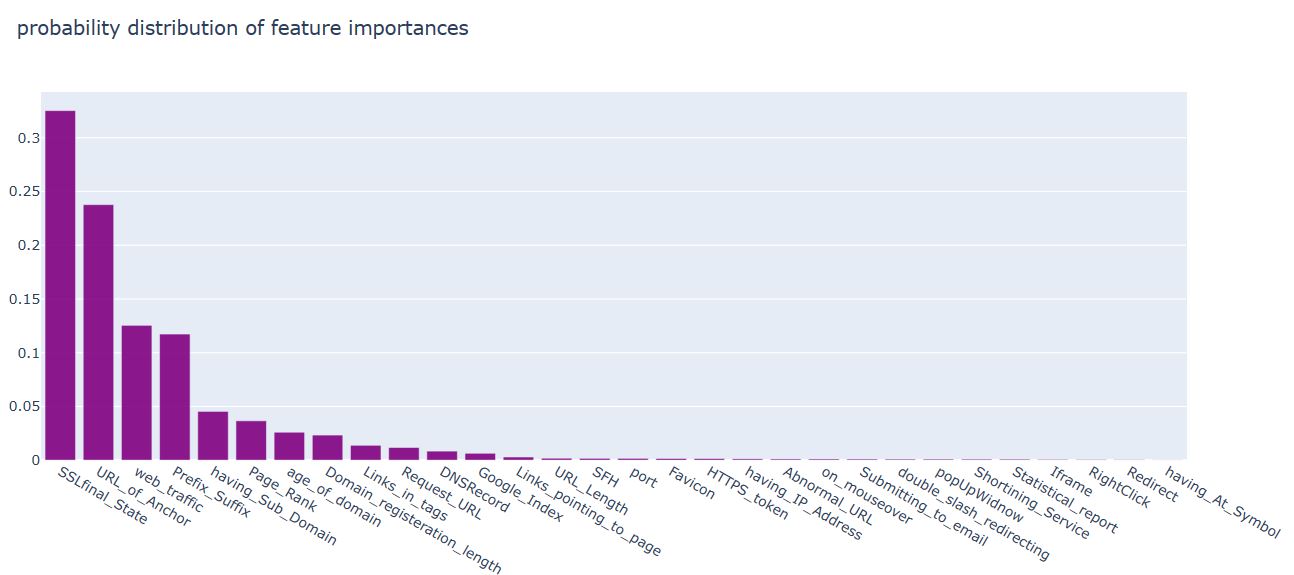
**OBSERVATIONS:**

From the above table we can infer that decision tree model gives the maximum accuracy i.e. 99.75%. But even though decision tree is giving pretty good accuracy , it suffers from the problem of over fitting. The difference in the accuracy on validation and test shows this.

Random forest don’t suffer over fitting, but it was not as good as XGboost classifier. From the random forest model we were able to identify the important features in the dataset. It is noted that the following variables influences the target variable more while model building:

* SSLfinal\_state
* URL\_of\_anchor
* Web\_trafic
* Prefix\_suffix

The following plot is a plot to show the feature importance:



**XGBoost** model has the best combination of prediction performance and processing time compared to other algorithms. This algorithm was giving 97.59% accuracy on training and 96.32% on validation.

So, the final selected model is XGBoost.

.