HP Machine Learning Challenge-Approach

Machine learning model to detect the phishing attacks..

1.Simple Exploratory Data Analysis

- Pandas, seaborn, matplotlib libraries are used in Exploratory data analysis.

2.Data Pre-Processing

- No pre-processing is done for the data.

3.Model

- Selected columns for model,
    - 0_key
    - 1_having_ip
    - 2_url_length
    - 3_shortining_service
    - 4_having_at_symbol
    - 5_double_slash_redirecting
    - 6_prefix_suffix
    - 7_having_sub_domain
    - 8_sslfinal_state
    - 9_domain_registeration_length
    - 10_favicon
    - 11_port
    - 12_https_token
    - 13_request_url
    - 14_url_of_anchor
    - 15_links_in_tags
    - 16_sfh
    - 17_submitting_to_email
    - 18_abnormal_url

- 19_redirect
- 20_on_mouseover
- 21_rightclick
- 22_popupwidnow
- 23_iframe
- 24_age_of_domain
- 25_dnsrecord
- 26_web_traffic
- 27_page_rank
- 28_google_index
- 29_links_pointing_to_page
- 30_statistical_report

- Compared multiple classifiers using pycaret’s compare_models function

	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	TT (Sec)
rf	Random Forest Classifier	0.9643	0.9933	0.9747	0.9622	0.9684	0.9273	0.9274	1.7960
et	Extra Trees Classifier	0.9644	0.9885	0.9722	0.9648	0.9685	0.9276	0.9277	1.8480
catboost	CatBoost Classifier	0.9647	0.9951	0.9722	0.9654	0.9688	0.9283	0.9284	4.2500
lightgbm	Light Gradient Boosting Machine	0.9647	0.9946	0.9702	0.9673	0.9687	0.9283	0.9285	0.6020
dt	Decision Tree Classifier	0.9547	0.9624	0.9600	0.9595	0.9597	0.9079	0.9080	0.3840
knn	K Neighbors Classifier	0.9461	0.9846	0.9591	0.9458	0.9524	0.8902	0.8905	1.7280
gbc	Gradient Boosting Classifier	0.9475	0.9903	0.9557	0.9512	0.9534	0.8933	0.8933	1.2880
ada	Ada Boost Classifier	0.9363	0.9867	0.9509	0.9369	0.9438	0.8704	0.8706	1.0340
ridge	Ridge Classifier	0.9298	0.0000	0.9503	0.9268	0.9384	0.8568	0.8573	0.6340
lda	Linear Discriminant Analysis	0.9298	0.9822	0.9503	0.9268	0.9384	0.8568	0.8573	0.6880
lr	Logistic Regression	0.9389	0.9866	0.9495	0.9423	0.9459	0.8757	0.8758	0.4560
svm	SVM - Linear Kernel	0.9324	0.0000	0.9211	0.9576	0.9386	0.8633	0.8648	0.7240
nb	Naive Bayes	0.8615	0.9758	0.7707	0.9784	0.8622	0.7269	0.7480	0.6580

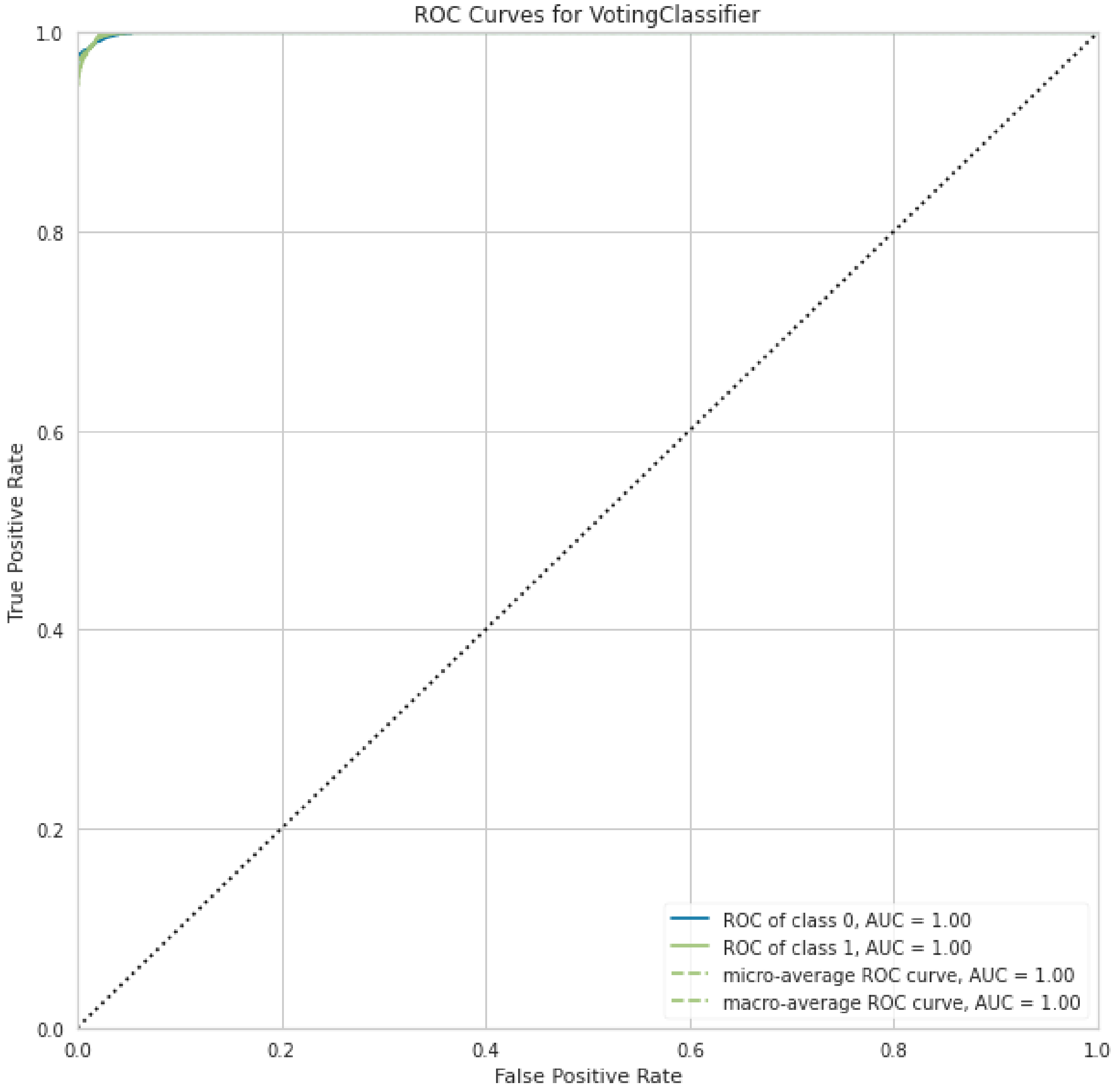
	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	TT (Sec)
qda	Quadratic Discriminant Analysis	0.5684	0.6162	0.2324	1.0000	0.3769	0.2096	0.3420	0.6360
dummy	Dummy Classifier	0.4378	0.5000	0.0000	0.0000	0.0000	0.0000	0.0000	0.3600

.

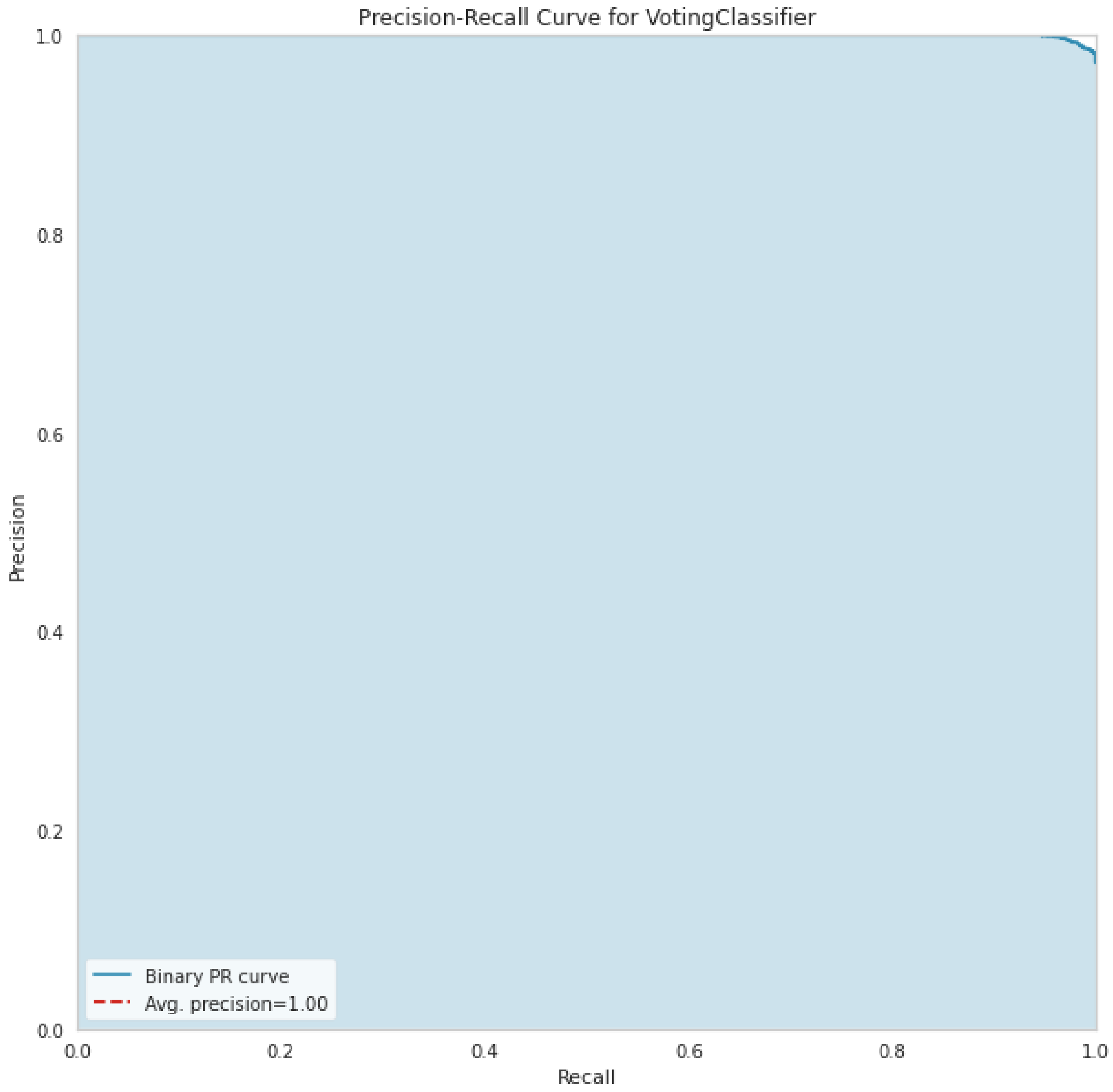
- Then took the top 3 models based on the recall then blend the model by using pycaret blend_models function

Fold	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC
0	0.9665	0.9956	0.9787	0.9623	0.9705	0.9318	0.9320
1	0.9665	0.9958	0.9702	0.9702	0.9702	0.9320	0.9320
2	0.9681	0.9944	0.9702	0.9730	0.9716	0.9352	0.9352
3	0.9673	0.9956	0.9759	0.9662	0.9710	0.9334	0.9335
4	0.9665	0.9954	0.9787	0.9623	0.9705	0.9317	0.9319
Mean	0.9670	0.9953	0.9747	0.9668	0.9708	0.9328	0.9329
Std	0.0006	0.0005	0.0038	0.0042	0.0005	0.0014	0.0013

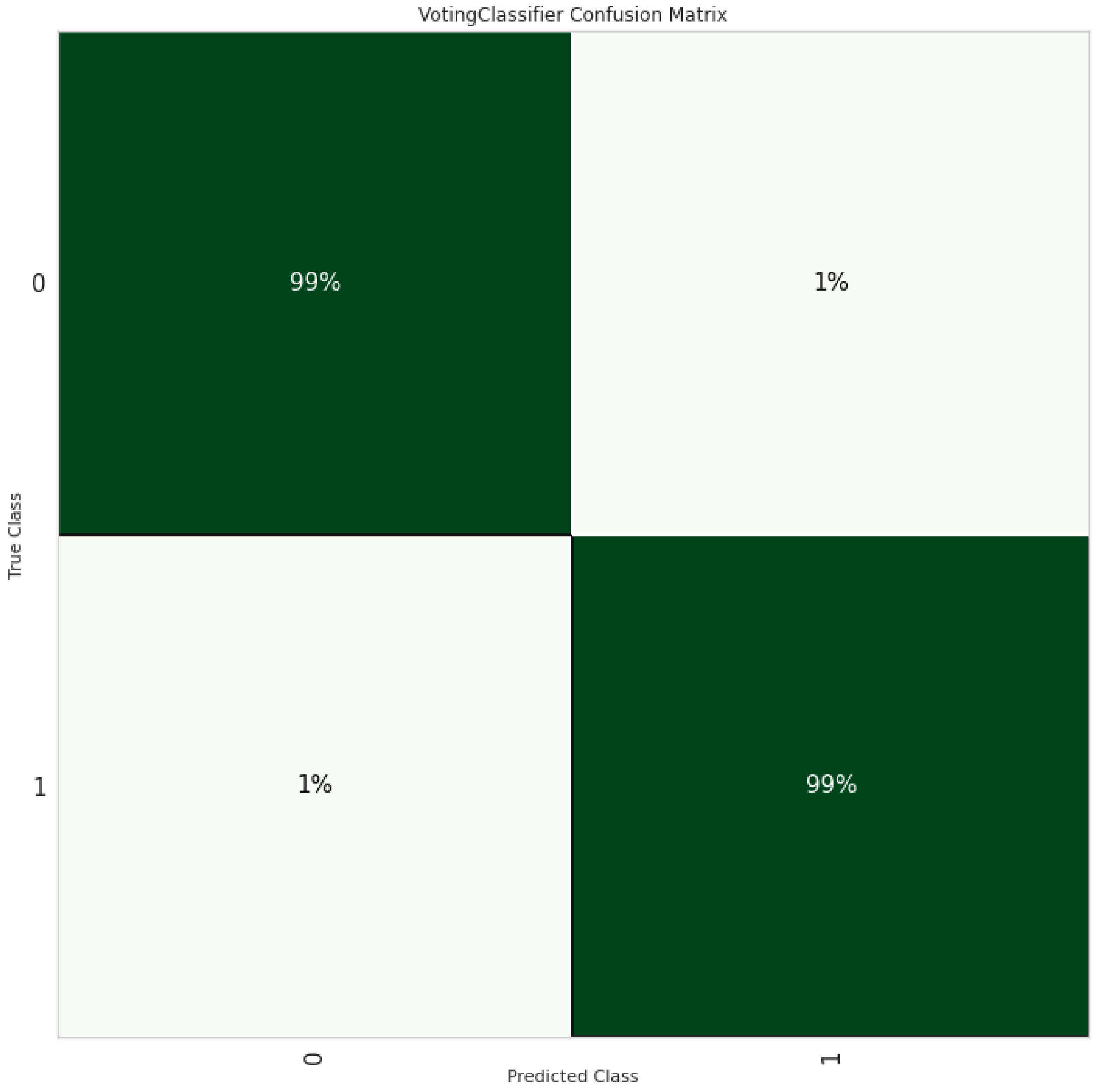
- ROC curve


ROC Curves for VotingClassifier
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
ROC of class 0, AUC = 1.00
ROC of class 1, AUC = 1.00
micro-average ROC curve, AUC = 1.00
macro-average ROC curve, AUC = 1.00

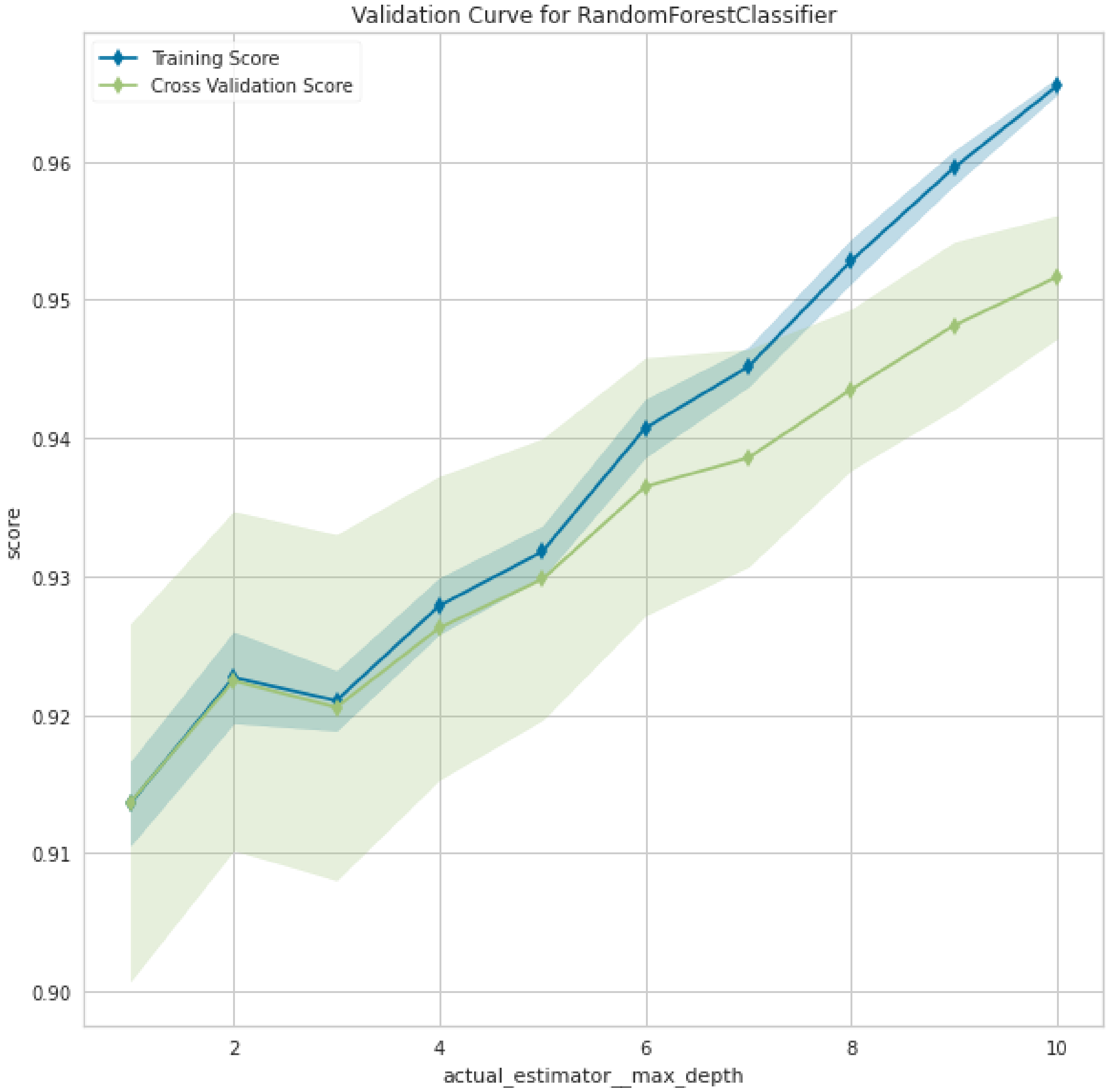
- Precision & Recall curve



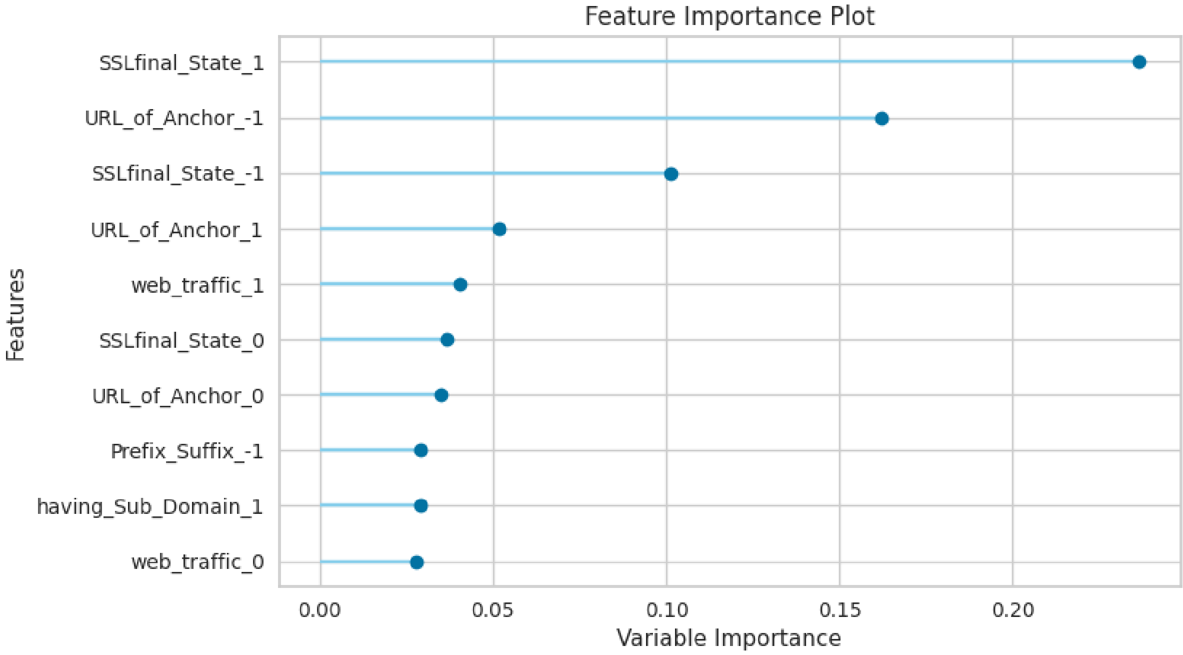
- Confusion Matrix



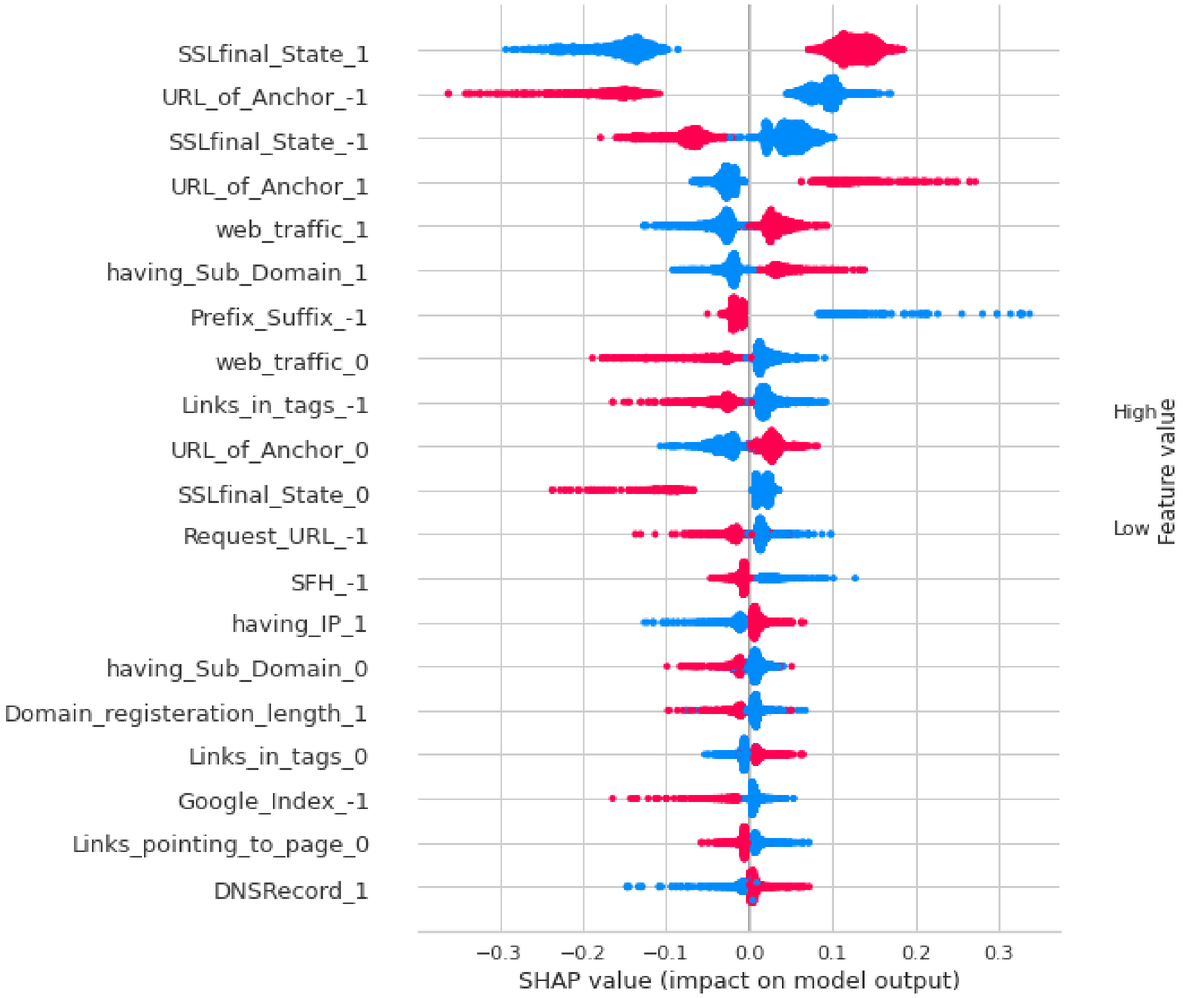
- Random Forest model validation curve



- Random Forest model feature Importance



- SHAP - Random Forest model feature Importance



- Final score is 97.9892