SDAIA
الهيئة السعودية للبيانات والذكاء الاصطناعي
Saudi Data & AI Authority

Phishing Websites Classification

Ghalia Maher
Jamila Rabeh
Ghadeer Ali

Supervised by Dr.Mejdal Alqahtani

WorkFlow Plan

01
Define the problem

02
Data Exploration and pre-processing

03
Classifier Models First approach

04
Classifier Models Second approach

05
Classifier Models Third approach

Is this link trusted?

Since the beginning of 2020 due to COVID-19 people have been social distancing and staying indoors as much as possible. Due to that the use of the Internet, E-commerce sites, and E-government operations have increased immensely and so have the attempts at phishing attacks.

His Excellency Dr. Khaled bin Abdullah Al-Sabti, Governor of the Cybersecurity Authority, spoke during the opening of the Global Cybersecurity Conference on April 7th, 2021, about the high increase in phishing sites by about 300% and the importance of being aware.

His Excellency

Dr. Khalid bin Abdullah Al-Sabti

Data Exploration and pre-processing

	qty_dot_url	qty_hyphen_url	qty_underline_url	qty_slash_url	qty_questionmark_url	qty_equal_url	qty_at_url	qty_and_url
0	3	0	0	1	0	0	0	0
1	5	0	1	3	0	3	0	2
2	2	0	0	1	0	0	0	0
3	4	0	2	5	0	0	0	0
4	2	0	0	0	0	0	0	0
...	...	...	...	...	...	...	...	...
88642	3	1	0	0	0	0	0	0
88643	2	0	0	0	0	0	0	0
88644	2	1	0	5	0	0	0	0
88645	2	0	0	1	0	0	0	0
88646	2	0	0	0	0	0	0	0

88647 rows × 112 columns

Data Exploration and pre-processing

Data cleaning

- Replace all negative values with zeros.
- Divide data frame into features and target.
- Check if the data contains missing value.
- Drop duplicated rows or records.

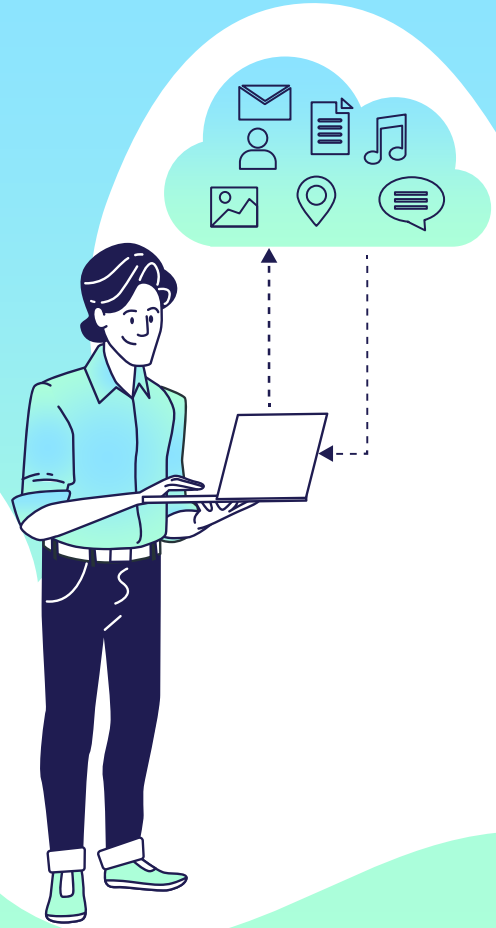
88647 rows × 112 columns

```
df.shape
```

```
(87199, 112)
```

Design

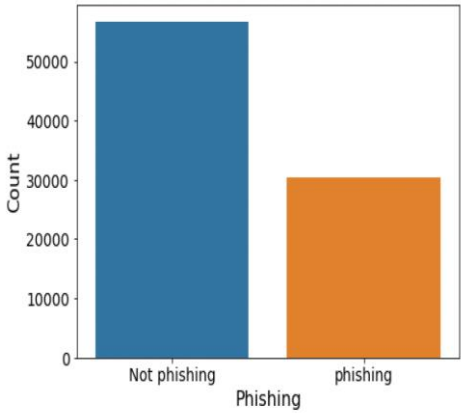
1. Decision tree.
2. Logistic regression.
3. Support vector machine.
4. Naïve bayes.
5. Random forest.
6. K-Nearest Neighbor.



3 Approaches

First Approach

Applying classifiers without scaling and modifying the data.

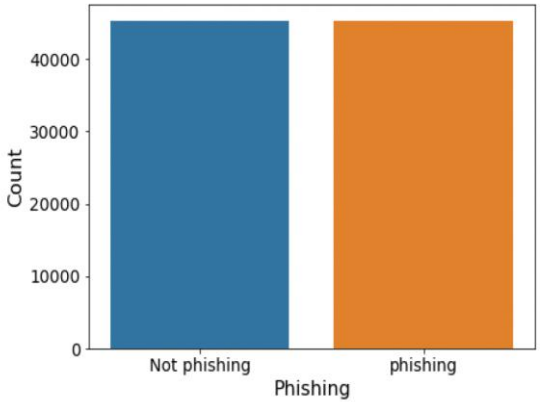


Second Approach

Applying smote to balance the data.

Third Approach

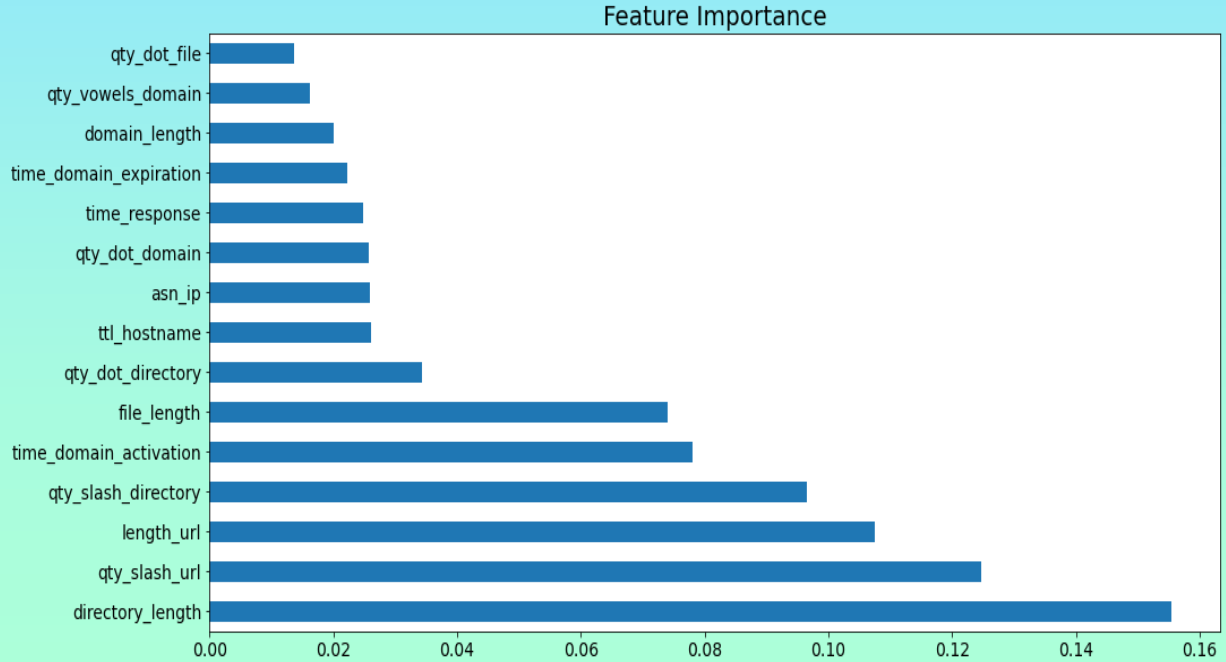
Balance the data and then scale.



First Approach

Applying classifiers without scaling and modifying the data

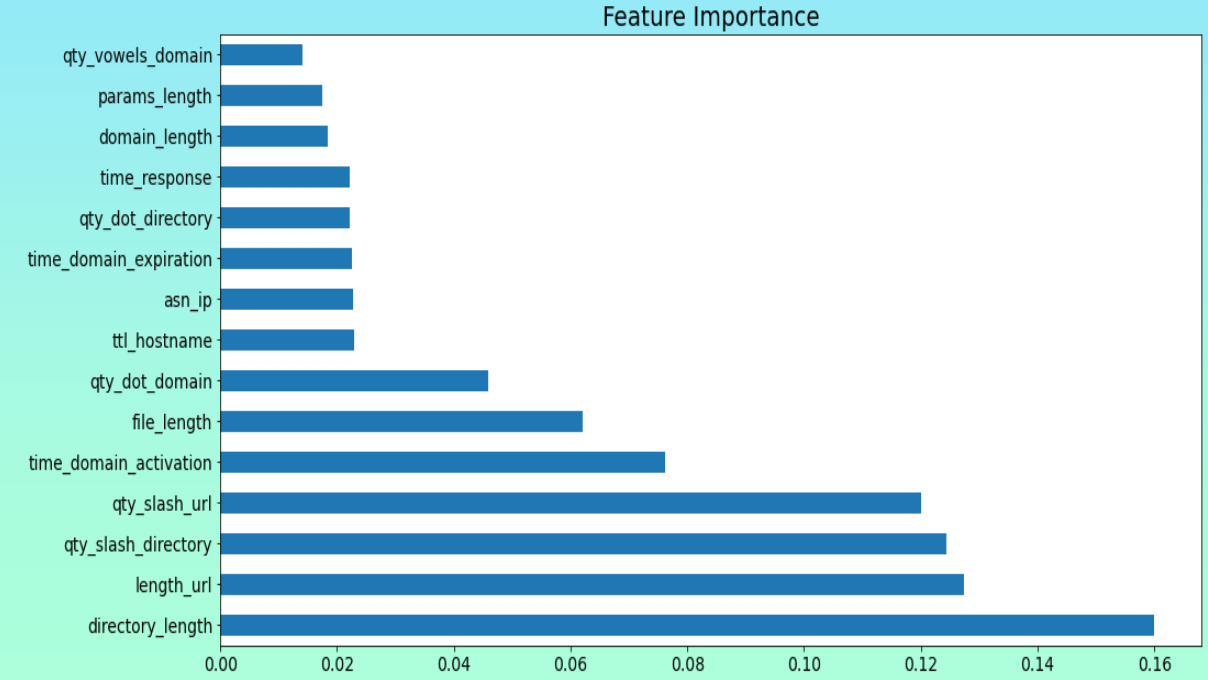
feature importance of the high classifier score (Random Forest)



Second Approach

Applying smote to balance the data

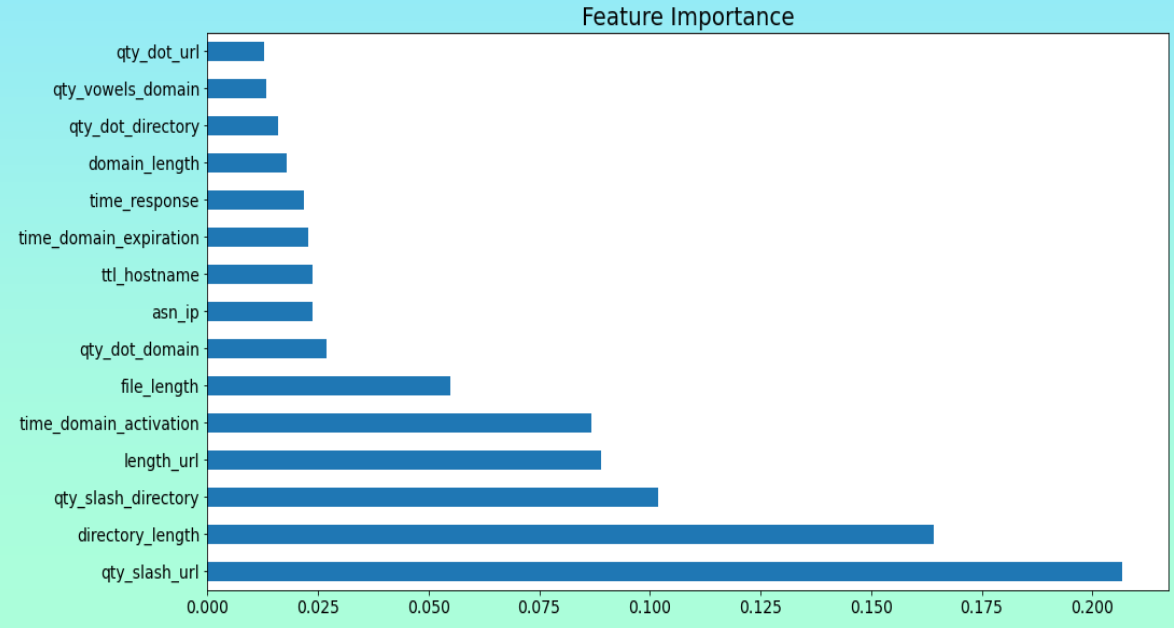
feature importance of the high classifier score (Random Forest)



Third Approach

Balance the data and then scale

feature importance of the high classifier score (Random Forest)



Model Evaluation

1. Accuracy.
2. Precision.
3. Recall.
4. Roc_auc.
5. F1 scores.
6.ROC curves

Metrics results

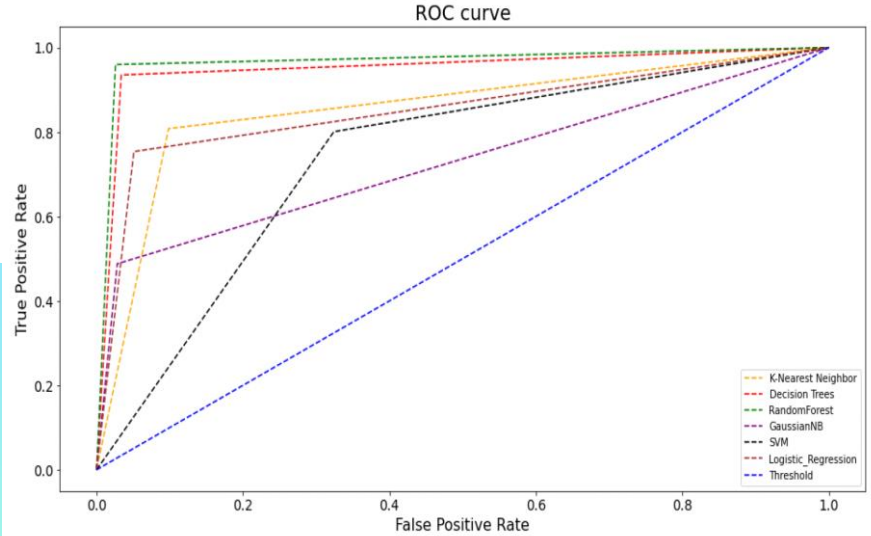
1	Accuracy	Precision	Recall	ROC_AUC	F1
Logistic Regression	0.881193	0.753673	0.887269	0.882960	0.815033
Support Vector Machines	0.765596	0.949645	0.603251	0.782334	0.737814
Decision Trees	0.954014	0.934126	0.933509	0.949222	0.933817
Random Forest	0.969209	0.961532	0.950392	0.964897	0.955929
Naive Bayes	0.803727	0.487865	0.902015	0.841504	0.633237
K-Nearest Neighbor	0.869037	0.807826	0.813737	0.855936	0.810771

2	Accuracy	Precision	Recall	ROC_AUC	F1
Logistic Regression	0.891628	0.818062	0.862789	0.884289	0.839831
Support Vector Machines	0.759690	0.703318	0.640216	0.736806	0.670286
Decision Trees	0.953727	0.945022	0.923524	0.946951	0.934149
Random Forest	0.968291	0.969292	0.941167	0.962281	0.955022
Naive Bayes	0.815195	0.525343	0.901416	0.847365	0.663816
K-Nearest Neighbor	0.862844	0.861483	0.770639	0.846000	0.813533

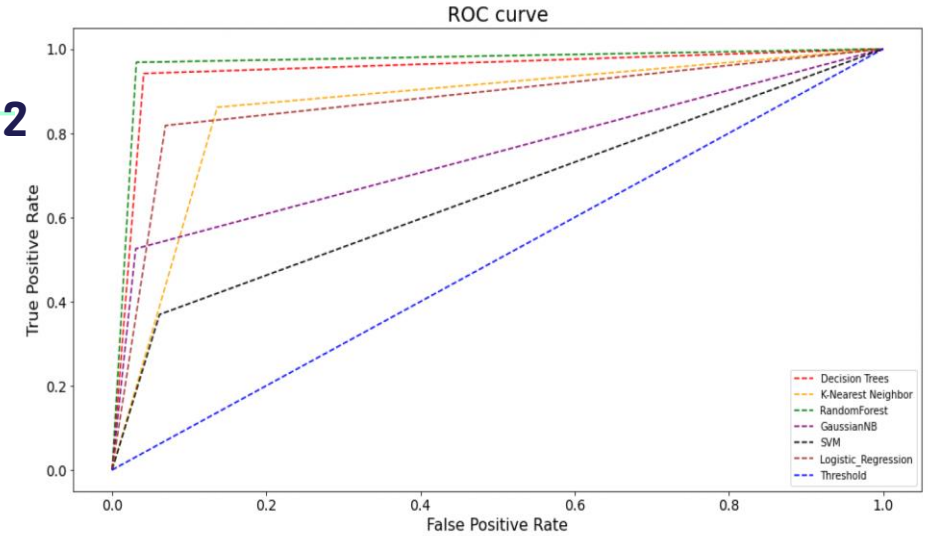
3	Accuracy	Precision	Recall	ROC_AUC	F1
Logistic Regression	0.933773	0.923890	0.889666	0.924160	0.906455
Support Vector Machines	0.932569	0.916460	0.892300	0.923599	0.904219
Decision Trees	0.952867	0.939739	0.925679	0.946676	0.932656
Random Forest	0.968922	0.966815	0.944973	0.963548	0.955770
Naive Bayes	0.728211	0.231138	0.944032	0.826093	0.371353
K-Nearest Neighbor	0.947248	0.940069	0.910894	0.939226	0.925252

ROC curves visualization

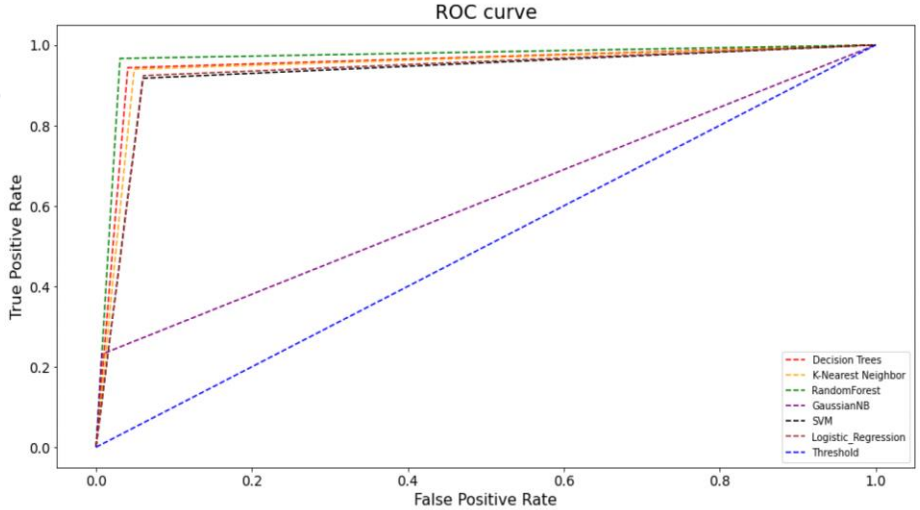
1



2



3



After All..

After applying the three approaches :

- The best classifier was Random Forest in all approaches with the Out of Bag score equal 0.97

	Accuracy	Precision	Recall	ROC_AUC	F1
Random Forest	0.969209	0.961532	0.950392	0.964897	0.955929

```
In [35]: model3.oob_score_
Out[35]: 0.975308369558022
```

- The classifiers that are sensitive to feature scaling got a good score in the third approach

THANKS!

Time to have fun..