COURSE NAME: Hillel - Machine Learning. Lesson 15 (17.04.2023 19:15,) 15. Ensemble Models, Random Forest
LECTURER: Yevhen Krasnokutsky
STUDENT: Maryna Lysenko

HW 12 REPORT “Feature importance”.

Investigate feature importance for a tabular dataset of your choice (regression or classification problem):

- use at least 3 feature importance approaches on a dataset of your choice
- compare feature importance results
- make report

Датасет Classifying Phishing websites from Legitimate ones
https://www.kaggle.com/datasets/aman9d/phishing-data

- Domain: The URL itself.
- Ranking: Page Ranking
- isIp: Is there an IP address in the weblink
- valid: This data is fetched from google's whois API that tells us more about the current
- status of the URL's registration.
- activeDuration: Also from whois API. Gives the duration of the time since the
- registration up until now.
- urlLen: It is simply the length of the URL
- is@: If the link has a '@' character then it's value = 1
- isredirect: If the link has double dashes, there is a chance that it is a redirect. 1-> multiple
- dashes present together.
- haveDash: If there are any dashes in the domain name.
- domainLen: The length of just the domain name.
- noOfSubdomain: The number of subdomains preset in the URL.
- Labels: 0 -> Legitimate website , 1 -> Phishing Link/ Spam Link

Model

```
model = CatBoostClassifier(iterations=50,
                           learning_rate=0.2,
                           od_type='Iter',
                           verbose=25,
                           depth=10,
                           random_seed=42)
```

CatBoost Accuracy Score is **0.94396**

	precision	recall	f1-score	support
0	0.94	0.93	0.93	7810
1	0.95	0.96	0.95	11372
accuracy			0.94	19182
macro avg	0.94	0.94	0.94	19182
weighted avg	0.94	0.94	0.94	19182

Важность фич

Сначала проанализирую фичи на корреляцию с таргетом:

activeDuration -0.523114

valid -0.266774

ID 0.004193

isIp 0.012811

is@ 0.039777

isredirect 0.073247

nosOfSubdomain 0.113249

domainLen 0.231828

haveDash 0.239623

urlLen 0.396519

ranking 0.516873

label 1.000000

(ID далее исключаю из обучения и оценки важности)

Feature Importance from Tree-Based Models

Feature ranking:

1. urlLen (29.2231)

2. ranking (22.4543)

3. activeDuration (17.0556)

4. domainLen (14.1205)

5. nosOfSubdomain (11.6093)

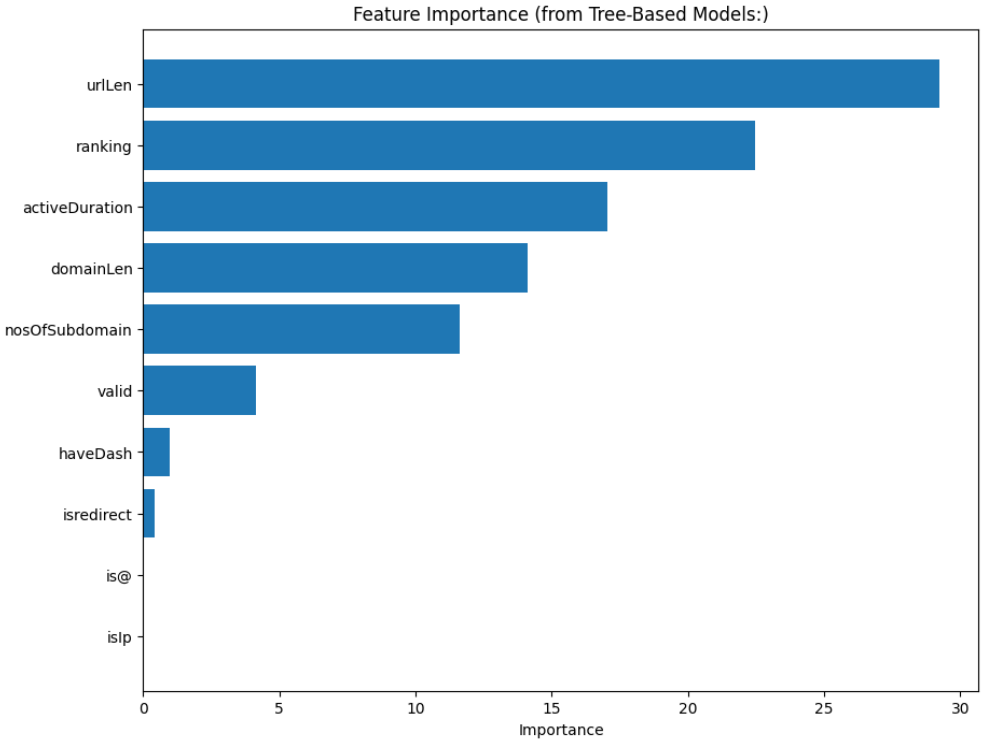
6. valid (4.1356)

7. haveDash (0.9631)

8. isredirect (0.4337)

9. is@ (0.0043)

10. isIp (0.0006)



Permutation Importance:

1. activeDuration (0.1396)

2. urlLen (0.1109)

3. ranking (0.1038)

4. domainLen (0.0616)

5. nosOfSubdomain (0.0350)

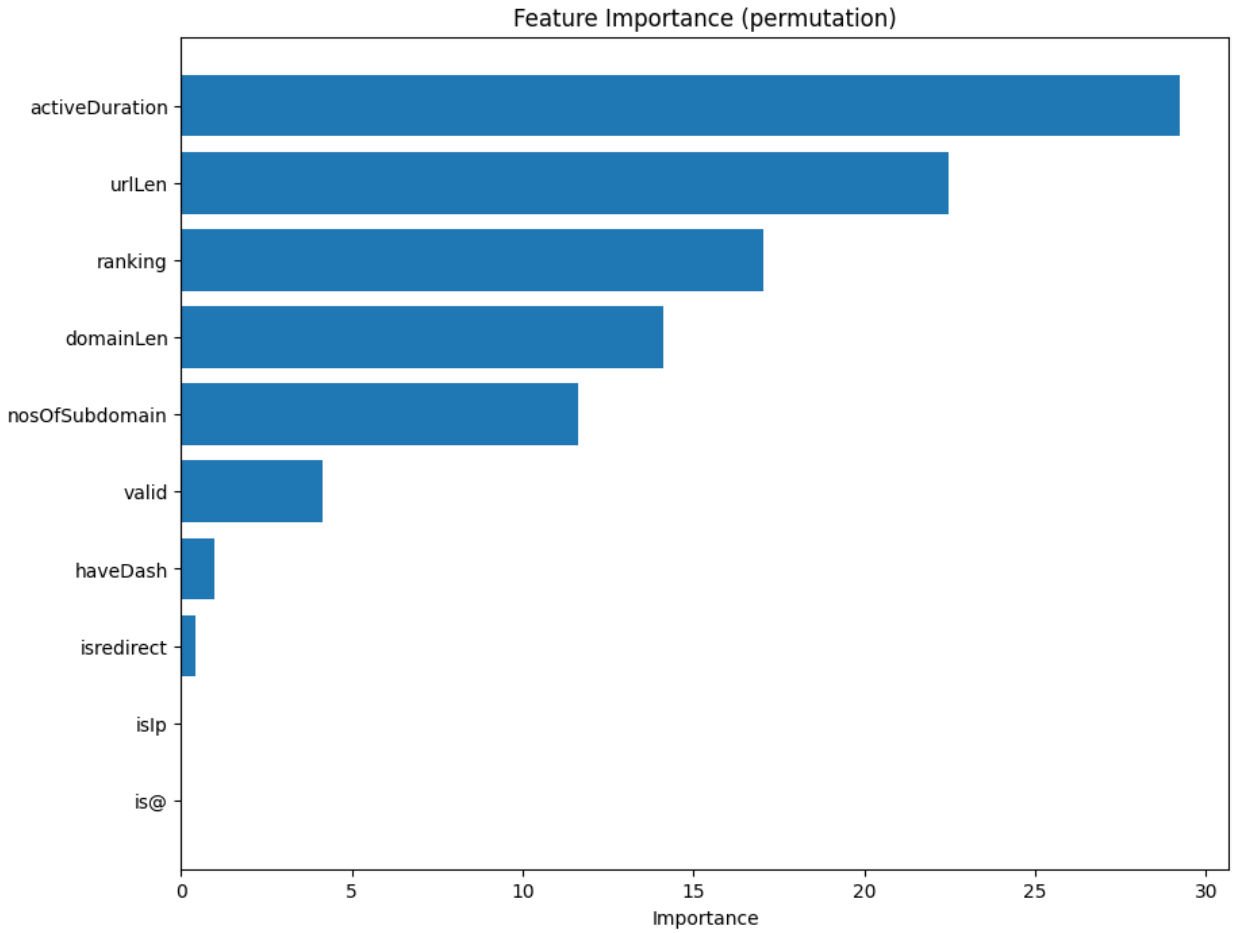
6. valid (0.0204)

7. haveDash (0.0013)

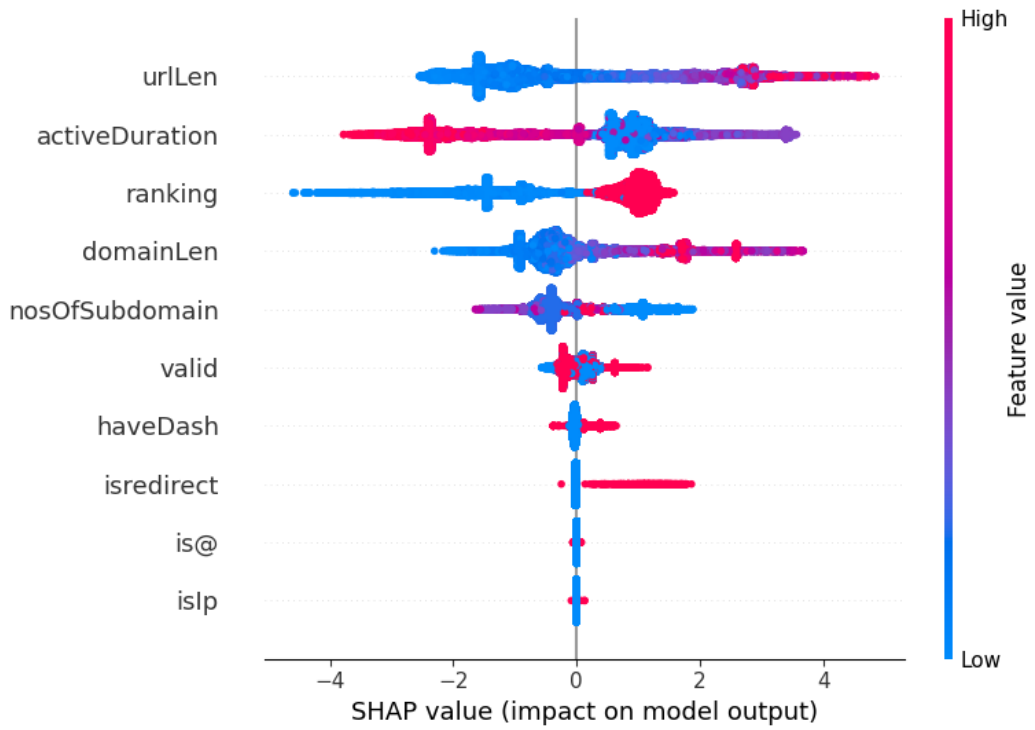
8. isredirect (0.0005)

9. is@ (0.0000)

10. isIp (0.0000)



Shap



feature	importance
urlLen	1.490992
activeDuration	1.347269
ranking	1.198582
domainLen	0.70693
nosOfSubdomain	0.592153
valid	0.176335
haveDash	0.049944
isredirect	0.018018
is@	0.000262
isIp	0.000028

Conclusions:

На основе результатов анализа важности признаков с использованием трех различных подходов можно сделать следующие выводы:

1. Самый важный признак для предсказания, является ли сайт легитимным или фишинговым, это длина URL, как указывается во всех трех подходах: корреляция с целевым значением, важность признака из моделей на основе деревьев и важность перестановки. Сайты с более длинными URL более вероятно будут классифицироваться как фишинговые ссылки.
2. Ранжирование страниц - второй наиболее важный признак, согласно корреляции с целевым значением и важности признака из моделей на основе деревьев. Это указывает на то, что сайты с высоким рангом страниц более вероятно будут легитимными.
3. Длительность с момента регистрации (activeDuration) - третий по важности признак согласно важности признака из моделей на основе деревьев и важности перестановки. Сайты, которые существуют дольше, более вероятно будут легитимными.
4. Длина домена, количество поддоменов и наличие тире в имени домена являются признаками средней важности для определения легитимности сайта.
5. Признаки, связанные с наличием IP-адресов, символа "@" и перенаправлений менее важны при предсказании легитимности сайта, согласно всем трем подходам.
6. В целом результаты указывают на то, что длина URL, ранжирование страниц и длительность с момента регистрации являются наиболее важными факторами при определении, является ли сайт легитимным или фишинговым. Эти выводы могут быть полезны при разработке лучших классификаторов для

обнаружения фишинговых сайтов и улучшения безопасности интернета.