MACHINE LEARNING MODEL

COMPARISON BASED ON SOME METRICS

Safa ORHAN

Computer Engineering Student

Istanbul Kultur University

Istanbul, Turkey

Eyüp USTA

Computer Engineering Student

Istanbul Kultur University

Istanbul, Turkey

I. SUPPORT VECTOR MACHINE

Kernels = linear, poly, rbf, sigmoid

C = 1, 2, 3, 4, 5

Degree = 1, 2, 3, 4, 5, 6

Gamma = scale, auto

Decision function shape = ovo, ovr

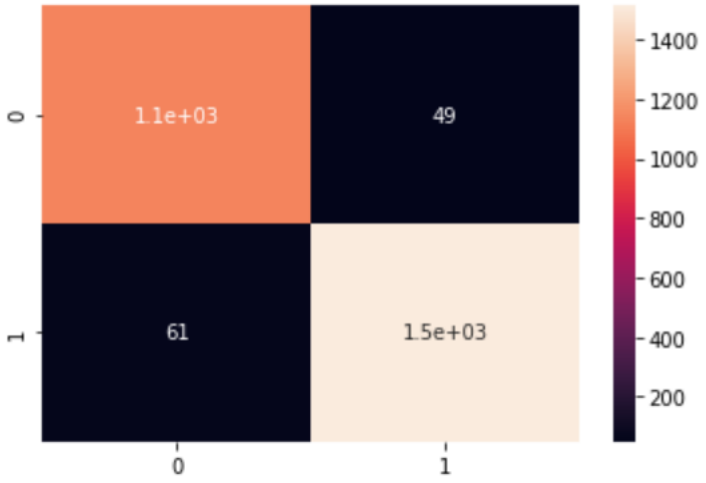
A. Results of SVM:

Best performed kernel: rbf

Worst kernel kernel: sigmoid

Linear: %93.31, Poly: %95.84, RBF: %95.01, Sigmoid: %83.73

The Average Score of SVC is %92.16 over 480 different model combinations. Best performed combination is “kernel = poly, C = 3, degree = 5, gamma = scale, decision function shape = ovo” with %97.53 accuracy.



II. LINEAR SUPPORT VECTOR MACHINE

Losses = hinge, squared hinge

Penalty = l2

C = 1, 2, 3, 4, 5

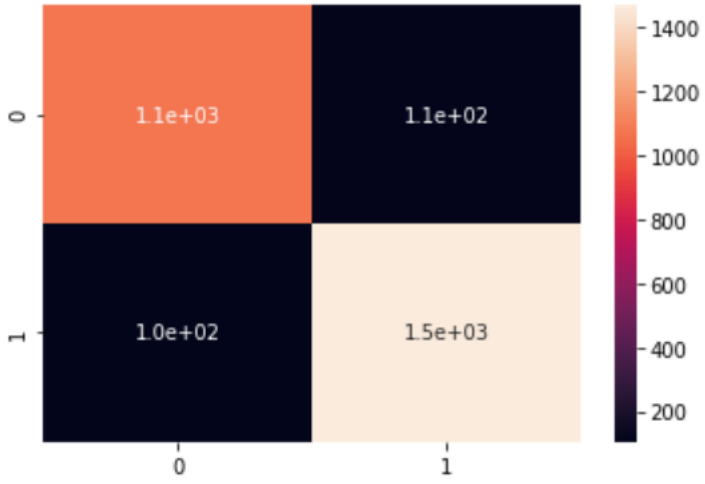
Multi Class = ovr, crammer singer

A. Resultf of Linear SVC:

Best performed loss function: squared hinge

Hinge: %93.10, Squared Hinge: %93.12

The average score of Linear SVC is %93.11 over 20 different models. Best performed combination is "loss function: hinge, penalty: l2, C: 1, Multi Class: ovr" with the accuracy of %93.198.



III. K – NEAREST NEIGHBORS

K = 1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20,21,22,23,24

Weights = uniform, distance

Metric = euclidean, manhattan, chebyshev, minkowski, wminkowski, mahalanobis, seuclidean.

A. Results of KNN:

K-1: %96.28, **K-2**: %95.18, **K-3**: %95.21, **K-4**: %95.12, **K-5**: %95.17,

K-6: %95.26, **K-7**: %95.29, **K-8**: %95.14, **K-9**: %95.19, **K-10**: %95.04,

K-11: %95.07, **K-12**: %94.69, **K-13**: %94.77, **K-14**: %94.66,

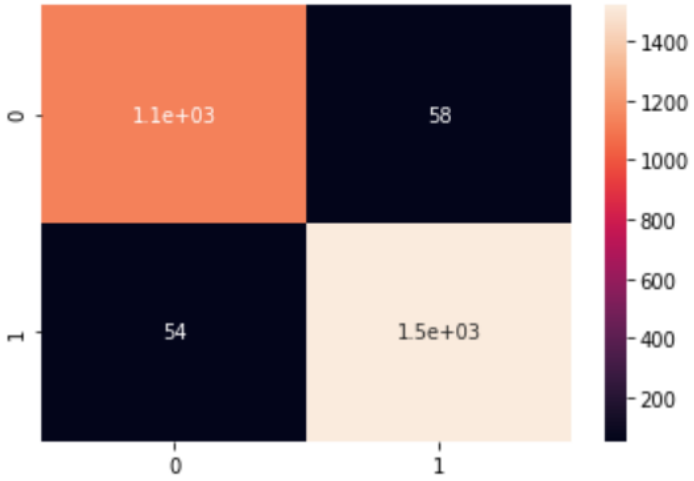
K-15: %94. 75, **K-16**: %94.59, **K-17**: %94.36, **K-18**: %9431,

K-19: %94.25, **K-20**: %9421, **K-21**: %94.20, **K-22**: %94.18,

K-23: %94.19, **K-24**: %94.16

Best performed k is 1 with %96.28.

The average accuracy of K-NN is %94.80 over 1344 model combinations. Best performed K-NN combination is “k: 6, weights: distance, algorithm: auto, metric: euclidean” with the accuracy %96.41.



IV. DECISION TREE CLASSIFIER

Max features = None, auto, sqrt, log2

Criterion = gini, entropy

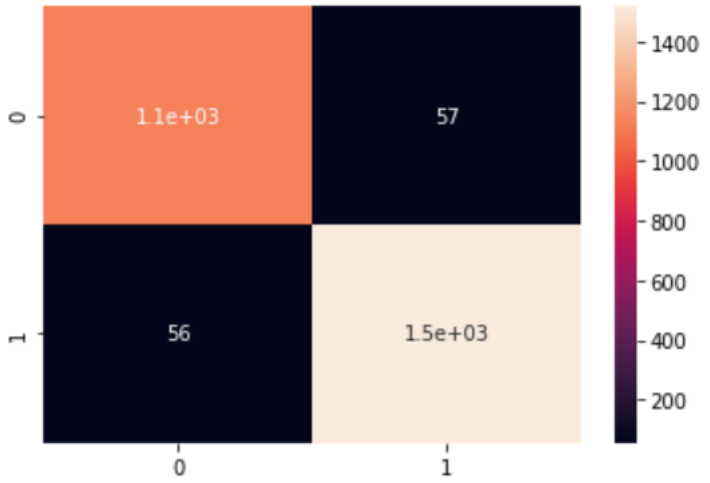
Splitter = best, random

A. Result of Decision Tree Classifier:

Best performed combination is “max_features = None, criterion = entropy, splitter = random” with accuracy of %96.52.

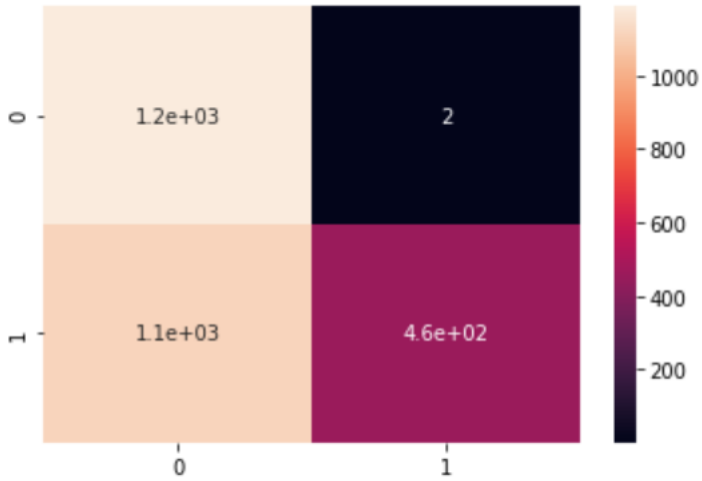
Worst performed combination is “max_features = log2, criterion = gini, splitter = random” with accuracy of %94.21.

The average accuracy of Decision Tree Classifer is %95.65 over 16 different model combinations.



V. GAUSSIAN NAIVE BAYES CLASSIFIER

The accuracy score of Gaussian Naive Bayes is %60.89.



VI. NAIVE BAYES CLASSIFIER

The accuracy score of our Naive Bayes classifier is %56.54.

A. Bernoulli Naive Bayes Classifier:

Alpha = 0, 1, 2, 3, 4, 5, 7, 9, 11

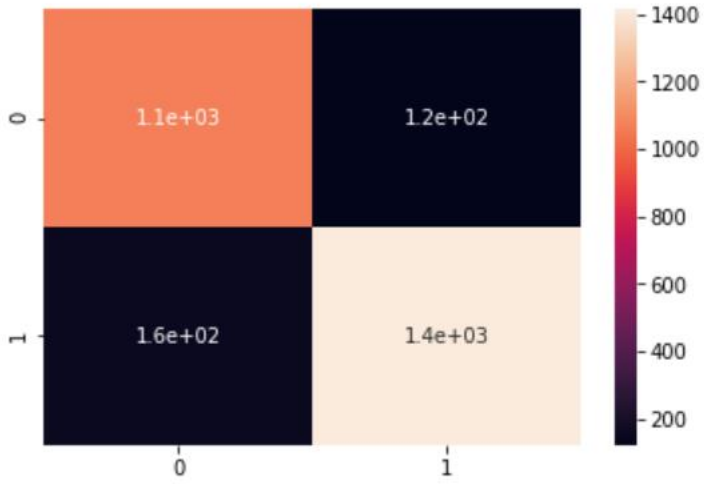
Binarize = 0, 1, 2, 3, 4, 5, 7, 9, 11

Fit prior = True, False

B. Results of BernoulliNB:

Best performed alpha value: 0,1,2 performed same accuracy value of %60.36; 3 and 4 performed %60.36; 5, 7 and 9 performed %60.35 accuracy; 11 performed %60.34 accuracy. As we can see the accuracy tends to perform worse as the alpha value increases.

The average performance of BernoulliNB is %60.36 over 162 different models. The best performed combination is “alpha = 0, binarize = 0, fit prior = True” with the accuracy of %90.95.



VII. SUPPORT VECTOR MACHINE

The accuracy of Support Vector Machine is %38.72 on testing set.

A. Random Forest Classifier:

Max features = None, auto, sqrt, log2

Criterion = gini, entropy

Class_weight = None, balanced, balanced_subsample

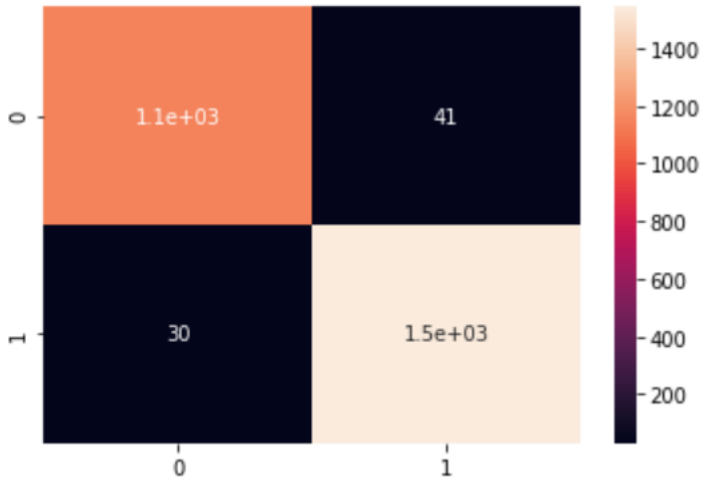
Warm start = True, False

B. Result of Random Forest Classifier:

Best performed combination is “max_features = log2, criterion = entropy, class_weight = balanced_subsample, warm_start = True” with accuracy of %96.96.

Worst performed combination is “max_features = auto, criterion = gini, class_weight = balanced, warm_start = False” with accuracy of %96.12.

The average accuracy of Random Forest Classifier is %96.60 over 48 different model combinations.



VIII. DEEP LEARNING WITH TENSORFLOW

Input Layer: Flatten (input_shape = (30,2))

Deep Layers: Dense (64, activation = relu), Dense (128, activation = relu), Dense (128, activation = relu)

Output Layer: Dense (1, activation=softplus)

Optimizers = 'sgd', 'rmsprop', 'adam', 'adadelta', 'adagrad', 'adamax', 'nadam', 'ftrl'

Loss = 'binary_crossentropy', 'categorical_crossentropy', 'hinge', 'squared_hinge', 'huber'

A. Results of the Deep Network:

Best performed combination is “optimizer = rmsprop, loss = huber” with the accuracy of %76.59 on validation set.

Worse performed combination is “optimizer = adam, loss = categorical_crossentropy” with the accuracy of %43.45 on validation set.

The average accuracy over 45 different models is %53.43 on validation set with this architecture of the network.

IX. DATA SET FEATURES

1. *Using the IP Address:*

Rule: *IF* $\begin{cases} \text{If The Domain Part has an IP Address} \rightarrow \text{Phishing} \\ \text{Otherwise} \rightarrow \text{Legitimate} \end{cases}$

2. *Long URL to Hide the Suspicious Part*

Rule: IF $\begin{cases} URL\ length < 54 \rightarrow feature = \text{Legitimate} \\ else\ if\ URL\ length \geq 54\ and\ \leq 75 \rightarrow feature = Suspicious \\ otherwise \rightarrow feature = \text{Phishing} \end{cases}$

3. *Using URL Shortening Services “TinyURL”*

Rule: IF $\begin{cases} \text{TinyURL} \rightarrow \text{Phishing} \\ \text{Otherwise} \rightarrow \text{Legitimate} \end{cases}$

4. *URL’s having “@” Symbol*

Rule: IF $\begin{cases} \text{Url Having @ Symbol} \rightarrow \text{Phishing} \\ \text{Otherwise} \rightarrow \text{Legitimate} \end{cases}$

5. *Redirecting using "//"*

Rule: IF $\begin{cases} \text{ThePosition of the Last Occurrence of "//" in the URL} > 7 \rightarrow \text{Phishing} \\ \text{Otherwise} \rightarrow \text{Legitimate} \end{cases}$

6. *Adding Prefix or Suffix Separated by (-) to the Domain*

Rule: IF $\begin{cases} \text{Domain Name Part Includes } (-) \text{ Symbol} \rightarrow \text{Phishing} \\ \text{Otherwise} \rightarrow \text{Legitimate} \end{cases}$

7. *Sub Domain and Multi Sub Domains*

Rule: IF $\begin{cases} \text{Dots In Domain Part} = 1 \rightarrow \text{Legitimate} \\ \text{Dots In Domain Part} = 2 \rightarrow \text{Suspicious} \\ \text{Otherwise} \rightarrow \text{Phishing} \end{cases}$

8. *HTTPS (Hyper Text Transfer Protocol with Secure Sockets Layer)*

Rule: IF $\begin{cases} \text{Use https and Issuer Is Trusted and Age of Certificate} \geq 1 \text{ Years} \rightarrow \text{Legitimate} \\ \text{Using https and Issuer Is Not Trusted} \rightarrow \text{Suspicious} \\ \text{Otherwise} \rightarrow \text{Phishing} \end{cases}$

9. *Domain Registration Length*

Rule: IF $\begin{cases} \text{Domains Expires on} \leq 1 \text{ years} \rightarrow \text{Phishing} \\ \text{Otherwise} \rightarrow \text{Legitimate} \end{cases}$

10. *Favicon*

Rule: IF $\begin{cases} \text{Favicon Loaded From External Domain} \rightarrow \text{Phishing} \\ \text{Otherwise} \rightarrow \text{Legitimate} \end{cases}$

11. *Using Non-Standard Port*

Rule: IF $\begin{cases} \text{Port \# is of the Preffered Status} \rightarrow \text{Phishing} \\ \text{Otherwise} \rightarrow \text{Legitimate} \end{cases}$

12. The Existence of "HTTPS" Token in the Domain Part of the URL

$$\textit{Rule: IF}\begin{cases}\text{Using HTTP Token in Domain Part of The URL} \rightarrow \text{Phishing}\\ \text{Otherwise} \rightarrow \text{Legitimate}\end{cases}$$

13. Request URL

$$\textit{Rule: IF}\begin{cases}\%\text{ of Request URL} < 22\% \rightarrow \text{Legitimate}\\ \%\text{of Request URL} \geq 22\% \text{ and } 61\% \rightarrow \text{Suspicious}\\ \text{Otherwise} \rightarrow \text{feature} = \text{Phishing}\end{cases}$$

14. URL of Anchor

$$\textit{Rule: IF}\begin{cases}\%\text{ of URL Of Anchor} < 31\% \rightarrow \textit{Legitimate}\\ \%\text{ of URL Of Anchor} \geq 31\% \text{ And} \leq 67\% \rightarrow \text{Suspicious}\\ \text{Otherwise} \rightarrow \text{Phishing}\end{cases}$$

15. Links in <Meta>, <Script> and <Link> tags

$$\textit{Rule: IF}\begin{cases}\%\text{ of Links in "} < \text{Meta} > \text{","} < \text{Script} > \text{" and "} < \text{Link>"} < 17\% \rightarrow \text{Legitimate}\\ \%\text{ of Links in } < \text{Meta} > \text{","} < \text{Script} > \text{" and "} < \text{Link>"} \geq 17\% \text{ And} \leq 81\% \rightarrow \text{Suspicious}\\ \text{Otherwise} \rightarrow \text{Phishing}\end{cases}$$

16. Server Form Handler (SFH)

$$\textit{Rule: IF}\begin{cases}\text{SFH is "about: blank" Or Is Empty} \rightarrow \text{Phishing}\\ \text{SFH Refers To A Different Domain} \rightarrow \text{Suspicious}\\ \text{Otherwise} \rightarrow \text{Legitimate}\end{cases}$$

17. Submitting Information to Email

$$\textit{Rule: IF}\begin{cases}\text{Using "mail()" or "mailto:" Function to Submit User Information} \rightarrow \text{Phishing}\\ \text{Otherwise} \rightarrow \text{Legitimate}\end{cases}$$

18. Abnormal URL

This feature can be extracted from WHOIS database. For a legitimate website, identity is typically part of its URL.

Rule: IF $\begin{cases} \text{The Host Name Is Not Included In URL} \rightarrow \text{Phishing} \\ \text{Otherwise} \rightarrow \text{Legitimate} \end{cases}$

19. *Website Forwarding*

Rule: IF $\begin{cases} \text{ofRedirect Page} \leq 1 \rightarrow \text{Legitimate} \\ \text{of Redirect Page} \geq 2 \text{ And} < 4 \rightarrow \text{Suspicious} \\ \text{Otherwise} \rightarrow \text{Phishing} \end{cases}$

20. *Status Bar Customization*

Rule: IF $\begin{cases} \text{onMouseOver Changes Status Bar} \rightarrow \text{Phishing} \\ \text{It Does't Change Status Bar} \rightarrow \text{Legitimate} \end{cases}$

21. *Disabling Right Click*

Rule: IF $\begin{cases} \text{Right Click Disabled} \rightarrow \text{Phishing} \\ \text{Otherwise} \rightarrow \text{Legitimate} \end{cases}$

22. *Using Pop-up Window*

Rule: IF $\begin{cases} \text{Popoup Window Contains Text Fields} \rightarrow \text{Phishing} \\ \text{Otherwise} \rightarrow \text{Legitimate} \end{cases}$

23. *IFrame Redirection*

Rule: IF $\begin{cases} \text{Using iframe} \rightarrow \text{Phishing} \\ \text{Otherwise} \rightarrow \text{Legitimate} \end{cases}$

24. *Age of Domain*

Rule: IF $\begin{cases} \text{Age Of Domain} \geq 6 \text{ months} \rightarrow \text{Legitimate} \\ \text{Otherwise} \rightarrow \text{Phishing} \end{cases}$

25. *DNS Record*

Rule: IF $\begin{cases} \text{no DNS Record For The Domain} \rightarrow \text{Phishing} \\ \text{Otherwise} \rightarrow \text{Legitimate} \end{cases}$

26. Website Traffic

$$\textit{Rule: IF}\begin{cases}\text{Website Rank} < 100{,}000 \rightarrow \text{Legitimate}\\ \text{Website Rank} > 100{,}000 \rightarrow \text{Suspicious}\\ \text{Otherwise} \rightarrow \text{Phish}\end{cases}$$

27. PageRank

$$\textit{Rule: IF}\begin{cases}\text{PageRank} < 0.2 \rightarrow \text{Phishing}\\ \text{Otherwise} \rightarrow \text{Legitimate}\end{cases}$$

28. Google Index

$$\textit{Rule: IF}\begin{cases}\text{Webpage Indexed by Google} \rightarrow \text{Legitimate}\\ \text{Otherwise} \rightarrow \text{Phishing}\end{cases}$$

29. Number of Links Pointing to Page

$$\textit{Rule: IF}\begin{cases}\text{Of Link Pointing to The Webpage} = 0 \rightarrow \text{Phishing}\\ \text{Of Link Pointing to The Webpage} > 0 \text{ and} \leq 2 \rightarrow \text{Suspicious}\\ \text{Otherwise} \rightarrow \text{Legitimate}\end{cases}$$

30. Statistical-Reports Based Feature

$$\textit{Rule: IF}\begin{cases}\text{Host Belongs to Top Phishing IPs or Top Phishing Domains} \rightarrow \text{Phishing}\\ \text{Otherwise} \rightarrow \text{Legitimate}\end{cases}$$