**Student**: Emma Delehanty

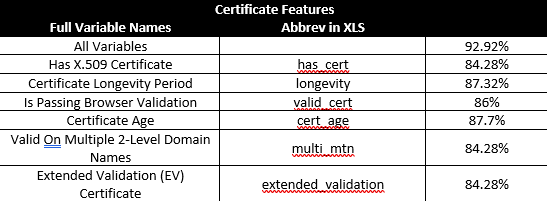
**Algorithm**: KNN(K-nearest neighbors algorithm)

* The KNN algorithm is a type of instance-based learning and a non-parametric method used for classification and regression, it assumes that similar things exist in closed proximity. In other words, similar things are near to each other. The k-NN algorithm is among the simplest of all machine learning algorithms.

(<https://en.wikipedia.org/wiki/K-nearest_neighbors_algorithm>)

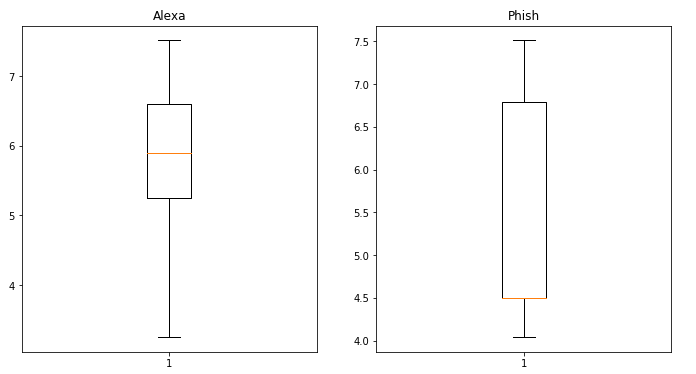
* A peculiarity of the KNN algorithm is that it is sensitive to the local structure of the data.
* After evaluating best k from 3-10, the model using 5 neighbors, which seemed to return the highest KNN score.
* The train\_test\_split() function was used with test\_size=0.2, random\_state=42.
* The target of this data was the type with alexa=True and phish=False.

**Results**

****

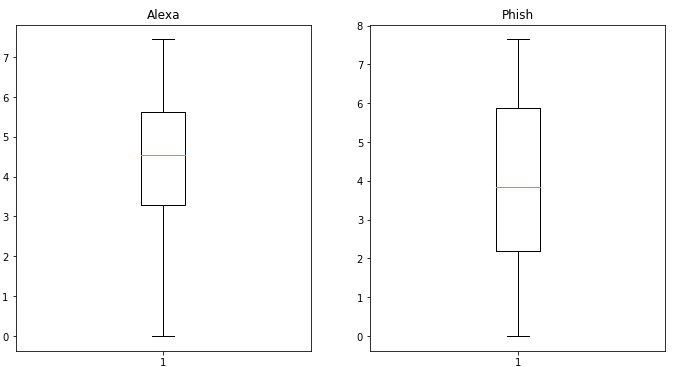
**Comparison of Certificate Longevity Period between legitimate and phishing websites**

(This feature indicates the number of days a given certificate is valid.)



**Comparison of Is Passing Browser Validation between legitimate and phishing websites**

(This feature indicates whether the provided X.509 certificate is valid and is issued for the requested domain. A valid certificate must not be expired or revoked at the time of access. It also must have a valid signature. Moreover, the certificate of CA that signed the certificate must be valid. This chain of valid certificates must end with a valid root certificate that is trusted by the validator. To determine the validity of a certificate, we rely on the internal web browser used by PhishMon)



**Notes**

* The accuracy for all the considered certificate features has highest 92.92%.
* The certificate age had the highest certificate feature accuracy score (87.7%) while 4 features (Has X.509 Certificate, Valid on Multiple 2-Level Domain Names, and Extended Validation (EV) Certificate) were tied for the lowest certificate feature accuracy score at 84.28%
  + The fact that these 3 values have the same accuracy score could indicate that they are strongly correlated and would be represented by the same principle component in PCA.
* From the comparison of Certificate Longevity Period between legitimate and phishing websites box plot, we can see the legitimate websites has longer days of valid certificate.
* From the comparison of Is Passing Browser Validation between legitimate and phishing websites box plot, we can see the legitimate websites has higher Passing Browser Validation.

KNN adding features

# [ipaddress](https://docs.python.org/3/library/ipaddress.html#module-ipaddress) — IPv4/IPv6 manipulation library in python

The functions and classes in this module make it straightforward to handle various tasks related to IP addresses, including checking whether or not two hosts are on the same subnet, iterating over all hosts in a particular subnet, checking whether or not a string represents a valid IP address or network definition, and so on.

<class 'pandas.core.frame.DataFrame'>

Int64Index: 7792 entries, 0 to 1318

Columns: 2209 entries, domain to ip

dtypes: int32(4), int64(2202), object(3)

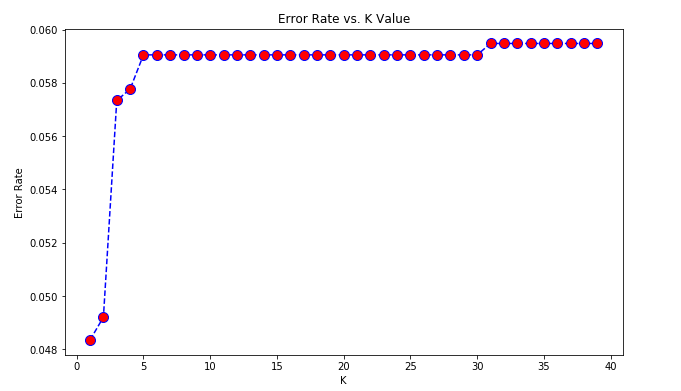
memory usage: 131.3+ MB

* New features: 'Network\_private','Network\_reserved','Network\_link\_local','Network\_loop', 'Network\_unspecified', 'Network\_global','Network\_class\_IPv4Network'



(data frame after OneHot encode )

**Best K Value evaluation**: from graph, we can see low K value has lower error rate, we choose K value (K-nearest neighbors=1), which seemed to return the highest KNN score 95.16%



Confusion matrix and report(K=1,Alexa=0,Phish=1): Confusion matrix and report(K=1,Alexa=0,Phish=1): the accuracy precision for Alexa is 96%,the accuracy precision for Phish is 95%

