Domain name detection

Problem definition

According to the latest information, there are more than 1.6 billion websites on the Internet distributed over 268 million domains . This ever-increasing expansion and growth has created the need for machine learning methods for researchers in this field. Domain and domain type recognition is an important issue that has various applications. For example, helping search engines, helping to build better tools for extracting information from websites , web filtering and advertising.

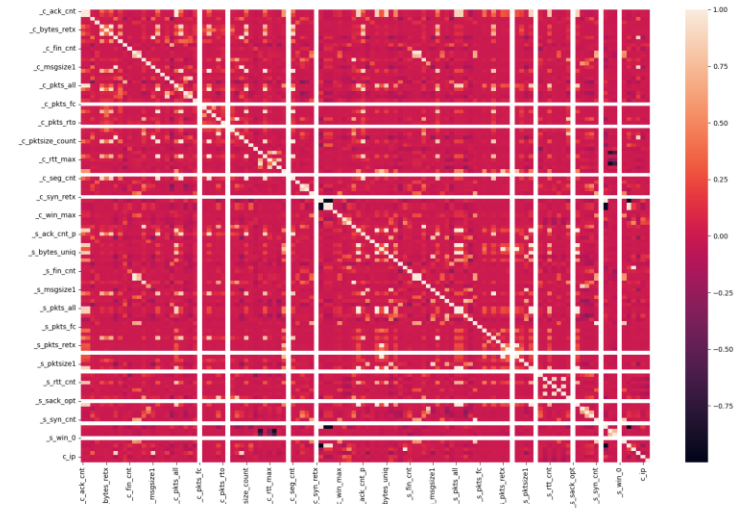
To use different machine learning methods on a data set related to network flows on the Internet and have a general view of the data set and the performance of different machine learning and selection models. Let's get a feature on this data set.

Data set and exploratory data analysis

The dataset contains information about network packets . For each sample, we have 122 featuresThese features are given in the table below. The training dataset contains 147863 samples and the test dataset contains 114580 samples Except for the c_ip attribute which represents the , IP address the rest of the attributes are numeric. We used the , ipaddress library to convert this feature . This library is written to work withIP addresses With its help, we converted this feature into a numerical feature.

Another issue is the label column. This column contains domain addresses as expected from the problem definition . We removed these addresses from categorical mode by label encoding method .and assigned a number to each domain.

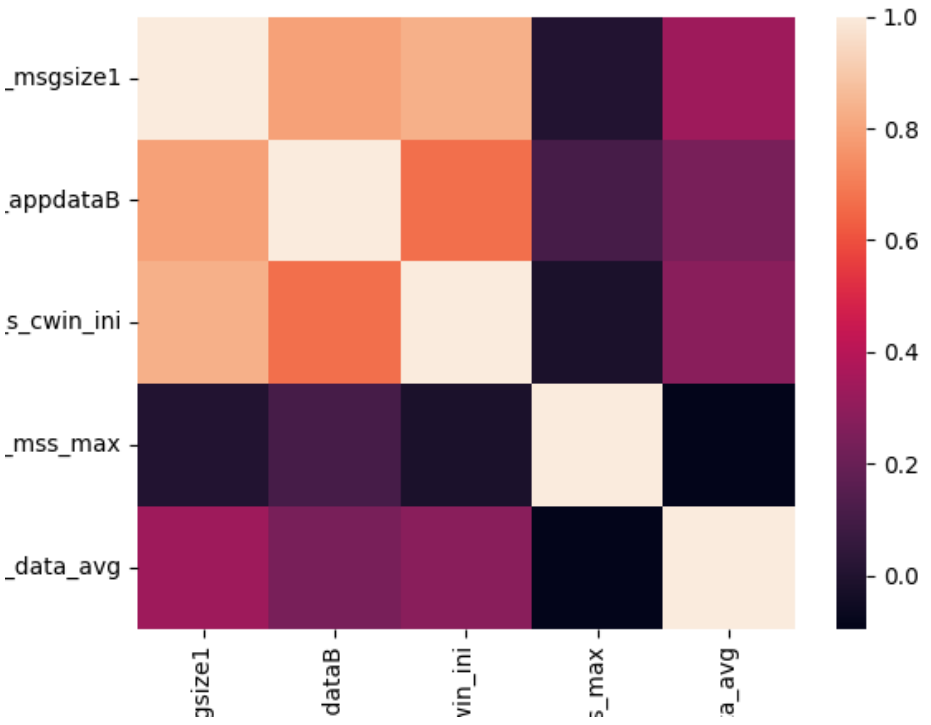
At first, to get a general view of the data, we plan to have a display of the data, but due to the large number of features , it is difficult to get useful information from the data. For example, the correlation . between data features is shown below.



1Correlation betweenfeatures

In order to have better information from the data set, we are going to calculate the importance of features with the help of decision tree. The decision tree is a good method for calculating feature importance due to its nature (which performs classification by calculating the most influential feature in order) . The importance of features is determined by cross entropy. After applying Extra Tree Classifier for the first 5 lines and showing their correlation, it is given

	_s_msgsize1	_s_appdataB	_s_cwin_ini	_c_mss_max	_s_pkts_data_avg
0	160	161	160	590	516.5
1	6774	7033	0	362	980.625
2	160	161	160	1384	756.5
3	128	129	0	506	276
4	128	0	0	401	128



2

Correlation of theselected features

Below is a description of the statistical characteristics of these features . The average, standard deviation, minimum and maximum amount and values related to the quartiles can be seen in the table below.

	_s_msgsize 1	**_s_appdataB**	**_s_cwin_ini**	**_c_mss_max**	**_s_pkts_data_avg**
count	147863	147863	147863	147863	147863
means	3731.7	3300.306094	3270.468988	612.4803568	885.8335787
std	2617.648	2836.851932	2719.584411	316.2012066	380.7012008
min	0	0	0	117	7
25%	160	155	156	517	703.6
50%	3948	4000	3767	517	933.5
75%	6585.5	5600	5468	605	1207.894737
max	11241	11500	11241	1448	1448

Another step is to decide to work withnull fields The existing data set is complete and no field .

Classification

In this part of the project, six supervised methodshave been implemented. The feature sets are scaled before input to each model.

1. Logistic Regression

Accuracy	53%
Precision	48%
Recall	35%
F1-score	37.5%

2. Naïve Bayes

Accuracy	6%
Precision	15%
Recall	18%
F1-score	9%

3. Decision Tree

Accuracy	80%
Precision	72%
Recall	76%
F1-score	74%

4. AdaBoost

Accuracy	39%
Precision	21%
Recall	11%
F1-score	11%

5. Random Forest

Accuracy	85%
Precision	89%
Recall	75.5%
F1-score	81%

6. SVM

To implementSupport Vector Machine it is usually necessary to set various parameters. Among , these parameters, we canmention kernel, regularizationparameter (C) andgamma . To adjust these parameters, wehave tried to optimizethem with the help of a grid search By optimizing . precision andrecall parameters we have found the best ,SVM model and reported the results on it.

In addition, because it is time- consuming and non-optimal to runSVM on high-dimensional data using the ,Extra Tree method (next section), we selected five features with high importance and .classified with this set. We made a new feature

Accuracy	89%
Precision	88%
Recall	79%
F1-score	83.5%

summary

As it is clear from the shape of the correlation between the features , the features of the dataset are highly correlated, and methods such asNaïve Bayes which assume that the features are , independent, do not provide acceptable results. The decision tree method is known as a good . parameter-free classification method . This method works well on high-dimensional data sets by calculating the importance of features . As expected , the use of random forest, which actually uses multiple decision trees , has resulted in higher accuracy. AlthoughAdaBoost uses several trees asa weak learner like random forest, it has a higher probability of over-fitting than random forest . SinceAdaBoost had an acceptable result on the training data set, this drop in accuracy can be seen as a signof over fitting the model. Support vector machine is a good classification method for all kinds of classification problems . The only challenge thatSVM faces is the setting and selection of parameters. Since we tried to find the model with the most optimal settings with the help of a search in the parameter space, the obtained result has been improved compared to other . methods

Feature selection

In this part of the project, three feature selection methods have been implemented . In the .following, we will describe each one

1. Extra Tree Classification

This method, in fact, is a classification method thattrains a number of trees on different subsets of the data setand obtains the bestaccuracy by averaging The way of making the forest in the . mentioned method makes it a suitable method for feature selection . Each tree withk features is built from the total features and this helps to determine the best feature. According to the amount ofloss that occurs in the use of each feature, a Gini importance measure can be calculated for each feature. To select a feature, you can sort the feature in descending order of this criterion and select higher features. implemented this method with five features.

2. Recursive Feature Elimination

RFE is a feature selection method thatfits a model on a datasetand removes the weakest feature (s) . InRFE implementation we used ,SVM as a model and first implemented the algorithm with features. Selected features are listed in the table below 5.

_s_msgsize1
_s_appdataB
_s_cwin_ini
_c_mss_max
_s_pkts_data_avg

3. Recursive Feature Elimination with Cross-Validation

REF is usually used withcross-validation to find the best number of features . In this part, we used thelogistic regression category to get the best number of features that does not reduce the performance of the system and preserves more information. REFCV recommends to keep 12 features.

For comparison, after selecting these features with the helpof REF andREFCV we ,use the Random Forest method once again to categorize these data to compare with the previously obtained results.

	Before Feature Selection	After REF (with 5 features)	After REFCV (with 12 features)
Accuracy	85%	82.5%	86%
Precision	89%	78.5%	87%
Recall	75.5%	76%	80%
F1-score	81%	77%	83%

Considering that the removal of features leads to loss of information , it is expected that the accuracy will decrease with the removal of features . As seen with 5 features . But when we use the featuresthat REFCV offers , the accuracy increases . This shows that not only having these features does not decrease the accuracy , but the absence of other features also helps the random 12 .forest to choose a better feature

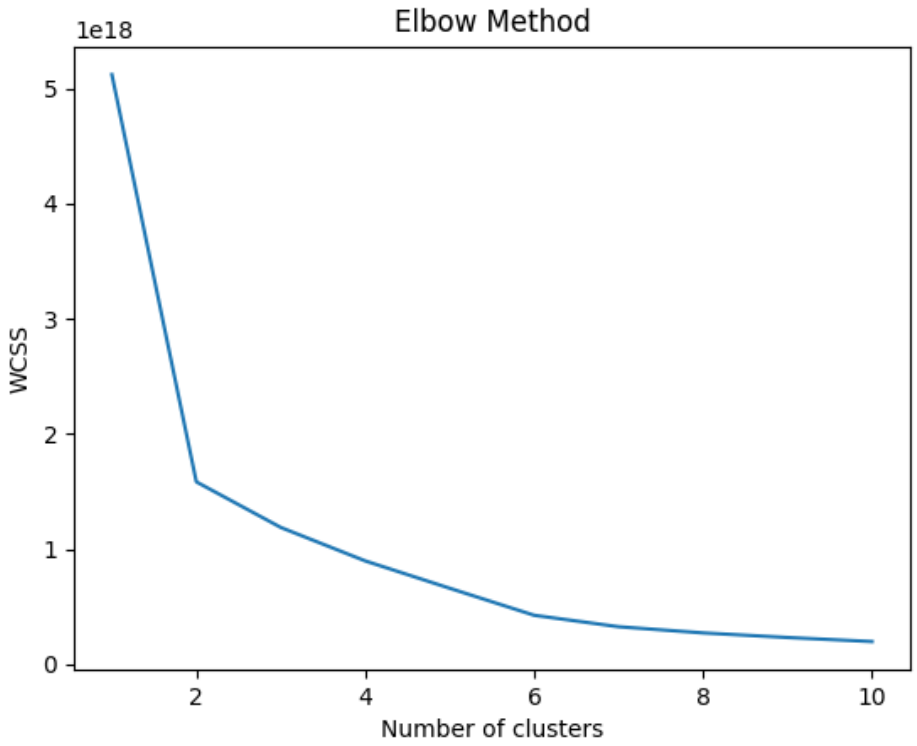
4. Principal Component Analysis

PCA is a feature selection method that tries to reduce the dimension while preserving as much information as possiblePCA has the ability to create a new feature by combining the original features . Since principal component analysis is applied to the covariance matrix , the data must be standardized. We selected five features with the helpof PCA and applied the Random Forest method . on the new data set Feature selection with PCA has resulted in loss of information and .reduced accuracy

	Before Feature Selection	After PCA (with 5 features)
Accuracy	85%	51%
Precision	89%	40%
Recall	75.5%	33%
F1-score	81%	35%

Clustering

We are going to use KMeans for clustering . Before clustering , we used a classifier (decision tree here) toselect the best features for clustering . Then, based on the importance of the features we chose five features to continue. Choosing the number of clusters is a challenge in using any , clustering method . In order to find the optimal number of clusters, we used theElbow method with theWithin Cluster Sum of Squares parameter. Based on this method, the number of clusters is determined where theWCSS criterion decreases at a point (the curve is broken) According . .to the form and nature of the problem, the logical choice is 6



Obtaining the optimal number of clusters by theElbow method

The result of clustering can be seen in the image below.