Cisco - Ariel University API Security Detection Challenge 2023

Dolev Abuhazira & Bar Luzon

Dataset 1

Preprocessing

1. Remove columns that have the same value for each record in the dataset.
2. Dealing with missing values by replacing them with ‘None’.

Feature Engineering :

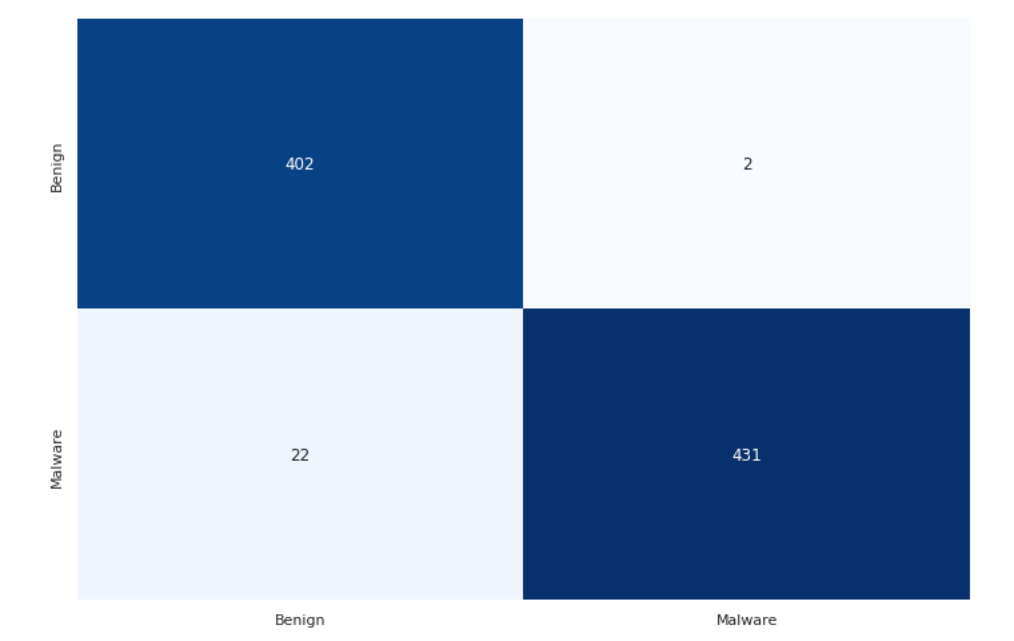
1. Extract information from ‘request.url’ by using 4 separate columns that represent the scheme, path, query,length.
2. Extract information from ‘request.headers.Cookie’ by using 3 separate columns that represent the name of the first user, second user, and the number of users in each record.
3. Transform the response features ‘Content-length’ and ‘status_code’ into a numeric representation.
4. Changing the request features ‘Accept-Encoding’,’Sec-Fetch-Dest’, and ’Sec-Fetch-Site’ into columns that contain 1 if the record in that column contains URL , else 0.
5. Encoding the categorical features by HashingVectorizer and LabelEncoder.

Model: Xgboost Classifier

This model works by training a number of decision trees. Each tree is trained on a subset of the data, and the predictions from each tree are combined to form the final prediction.

Train and test splitting :

Results - 0.9561, 7th place in the leaderboard right now.



	precision	recall	f1-score	support
Benign	0.94811	0.99505	0.97101	404
Malware	0.99538	0.95143	0.97291	453
accuracy			0.97200	857
macro avg	0.97175	0.97324	0.97196	857
weighted avg	0.97310	0.97200	0.97202	857

Dataset 2

Preprocessing (same as Dataset 1):

Feature Engineering (same as Dataset 1):

Dealing with an unbalanced dataset :

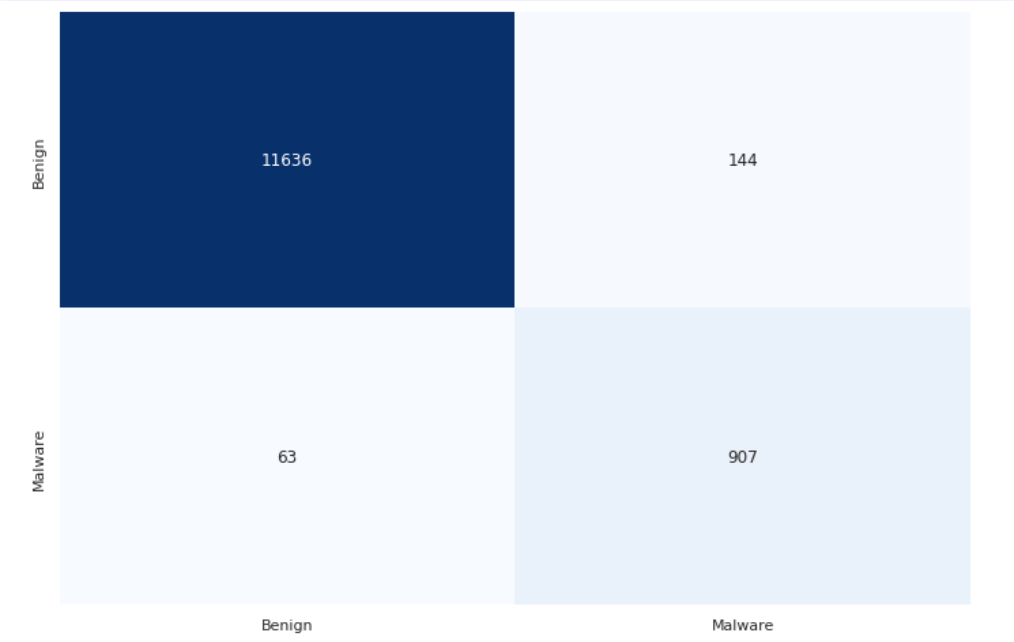
We used SMOT which creates synthetic samples of the minority class by the KNN algorithm and thus the data set is balanced with an equal number of samples.

Model: Xgboost Classifier

This model works by training a number of decision trees. Each tree is trained on a subset of the data, and the predictions from each tree are combined to form the final prediction.

Train and test splitting :

Results: 0.960, 7th place in leaderboard right now.



	precision	recall	f1-score	support
Benign	0.99461	0.98778	0.99118	11780
Malware	0.86299	0.93505	0.89758	970
accuracy			0.98376	12750
macro avg	0.92880	0.96141	0.94438	12750
weighted avg	0.98460	0.98376	0.98406	12750

Dataset 3

Preprocessing (same as Dataset 1):

Feature Engineering (same as Dataset 1) excepts :

1. Changing the request feature ‘Accept-Encoding’ which now includes three types of values into three columns that represent those values.
2. Selection of the most relevant features by Random Forest (18 selected in total)

Dealing with an unbalanced dataset (same as Dataset 2):

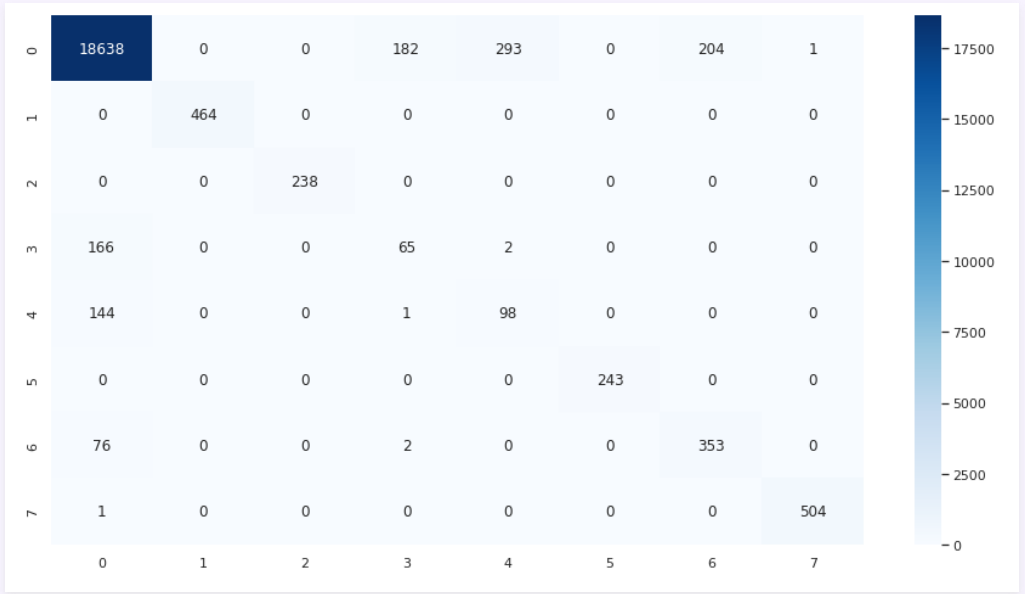
Model: Random Forest Classifier

This Model consists of a large number of individual decision trees that operate as an ensemble. Each individual tree in the random forest spits out a class prediction and the class with the most votes becomes our model’s prediction

Train and test splitting :

Results - 0.932 (Binary) 0.922 (Multiclass)

8th place in the leaderboard right now.



	precision	recall	f1-score	support
0	0.98029	0.96918	0.97470	19240
1	1.00000	1.00000	1.00000	488
2	1.00000	1.00000	1.00000	239
3	0.45736	0.50644	0.48065	233
4	0.27487	0.40385	0.32710	260
5	1.00000	1.00000	1.00000	245
6	0.66426	0.76190	0.70974	483
7	0.99795	0.99795	0.99795	487
accuracy			0.95483	21675
macro avg	0.79684	0.82991	0.81127	21675
weighted avg	0.96044	0.95483	0.95737	21675

Dataset 4

Preprocessing (same as Dataset 3)

Feature Engineering (same as Dataset 3)

Dealing with an unbalanced dataset (same as Dataset 2):

Model: Xgboost Classifier using RandomForest, GradientBoosting, SVC (Boosting Ensemble)

Boosting algorithms train the individual models sequentially, where each model attempts to correct the mistakes of the previous model. The final prediction is made by combining the predictions of the individual models using a weighted sum, where the weights are determined by the accuracy of each model.

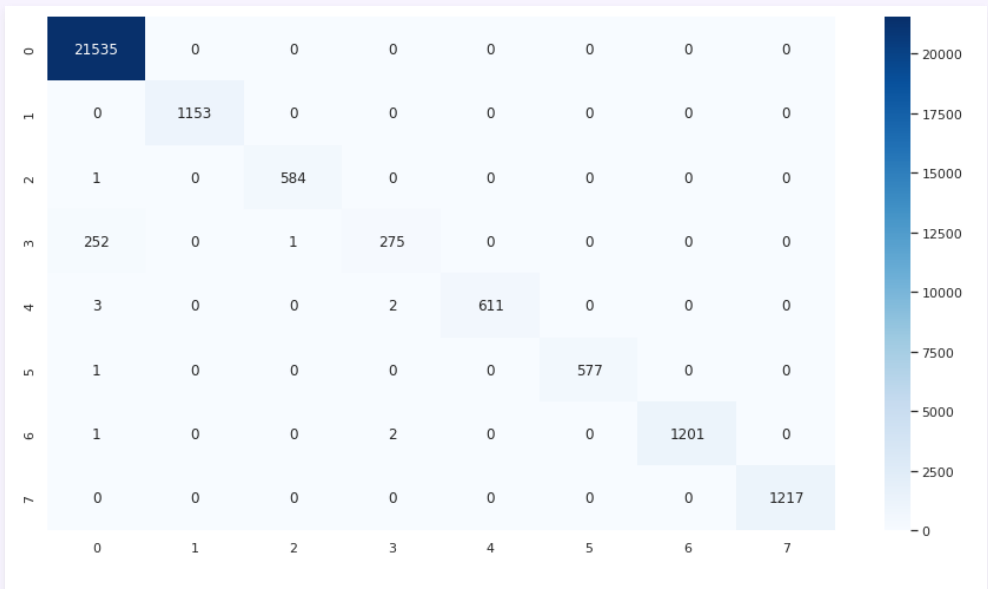
Hyperparameter optimization: using GridSearch with 5 folds cross-validation

Learning rate = 0.3 , maximum depth = 5 , subsample = 0.6.

Train and test splitting :

Results - 0.974 (Binary) 0.931 (Multiclass)

8th place in the leaderboard right now.



	precision	recall	f1-score	support
0	0.98816	1.00000	0.99405	21535
1	1.00000	1.00000	1.00000	1153
2	0.99829	0.99829	0.99829	585
3	0.98566	0.52083	0.68154	528
4	1.00000	0.99188	0.99593	616
5	1.00000	0.99827	0.99913	578
6	1.00000	0.99751	0.99875	1204
7	1.00000	1.00000	1.00000	1217
accuracy			0.99041	27416
macro avg	0.99651	0.93835	0.95846	27416
weighted avg	0.99039	0.99041	0.98899	27416