Malicious Url Detection with PySpark

Big Data Computing - Final Project - 2021/2022

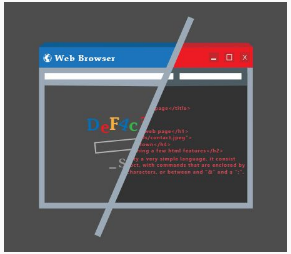
Clizia Giorgia Manganaro
2017897

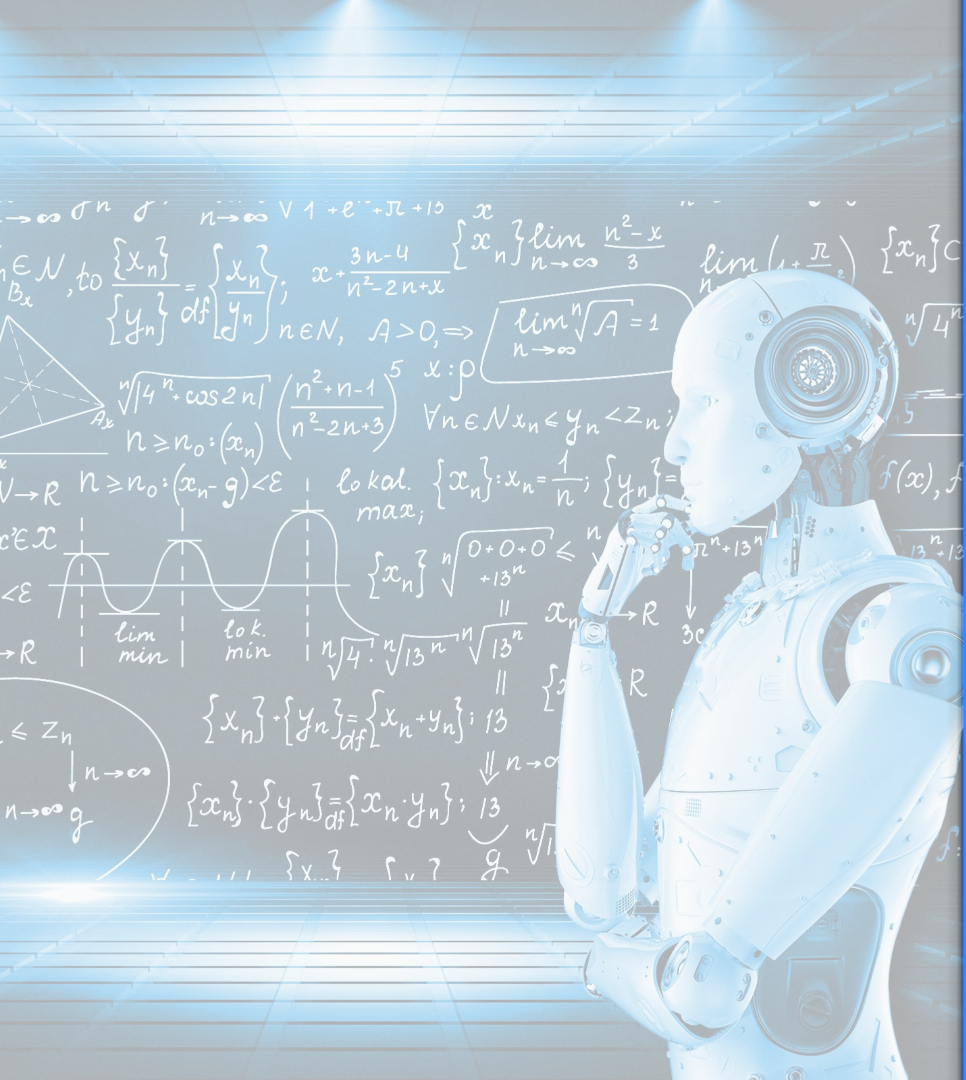
1. Introduction
2. Task
3. Dataset
4. Feature Engineering
5. Exploratory Data Analysis
6. Binary Classification
7. MultiClass Classification

Introduction

Malicious URLs or malicious website is a very serious threat to cybersecurity.

The Web has long become a major platform for online criminal activities. URLs are used as the main vehicle in this domain.





Task

Classification Task:

- Binary classification
- MultiClass classification

Datasets

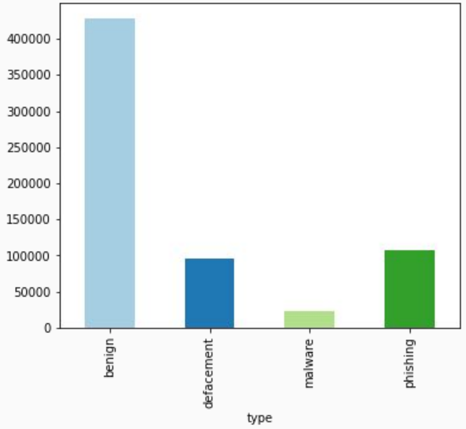
schema → (url, type)

651206 urls

- defacement: 96457
- phishing: 94108
- malware: 32520
- benign: 428103

14268 phishing url

Unbalanced Dataset



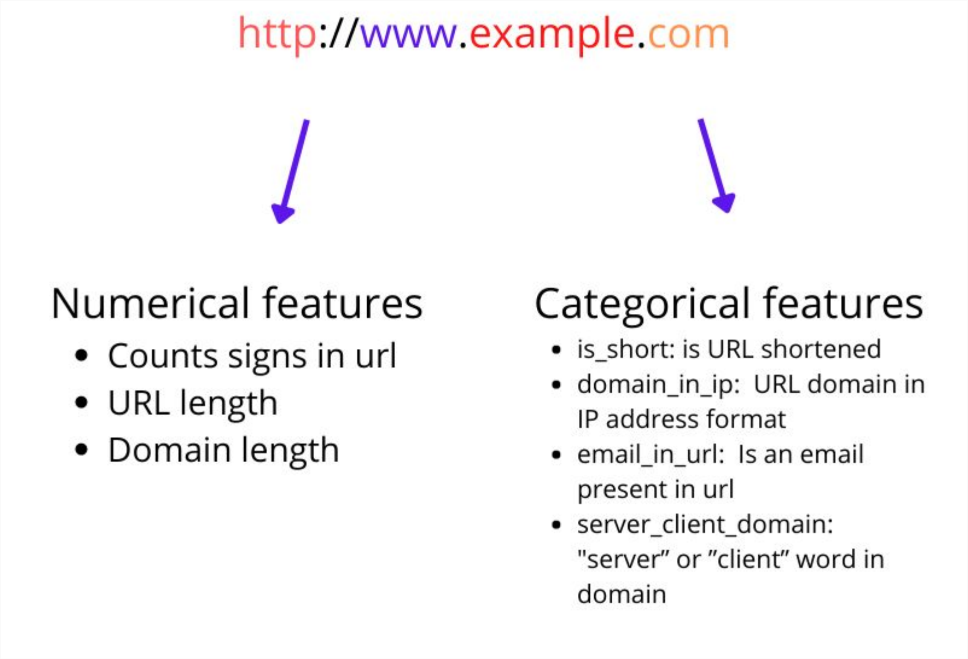
Undersampling Benign urls

type	count
malware	23645
defacement	95308
phishing	107119
benign	226309

452.381 urls

Feature Engineering

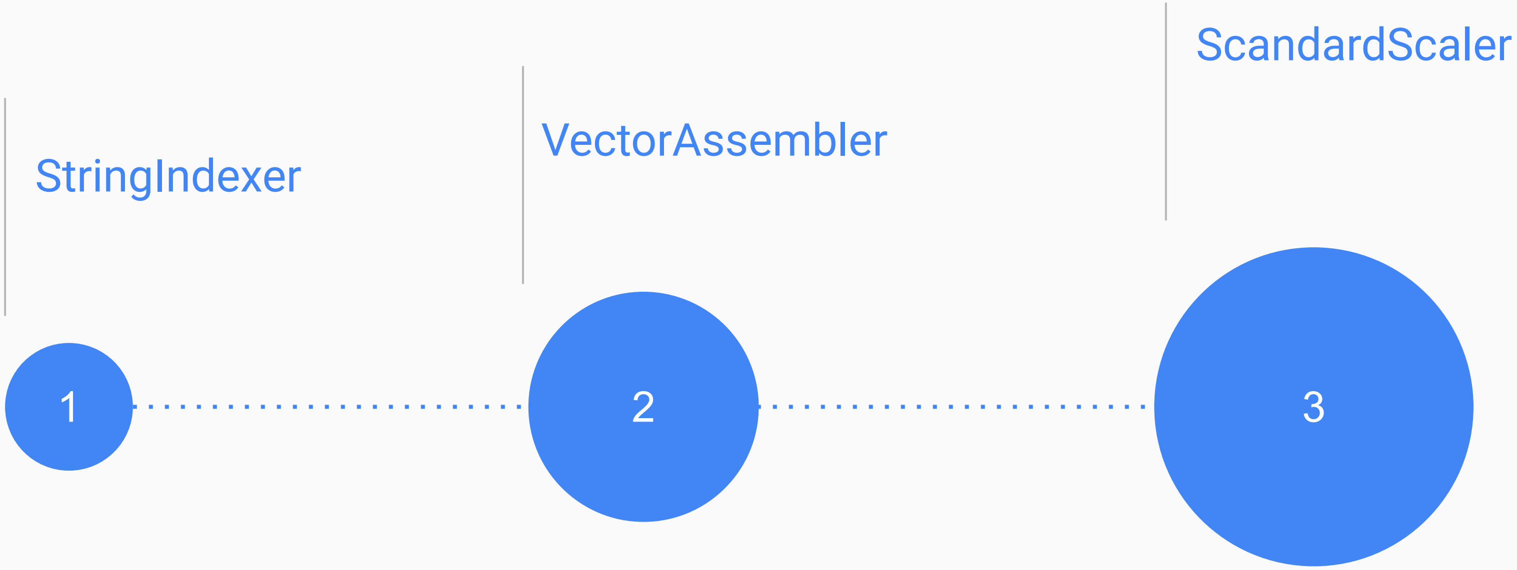
- 28 new features have been added.
- The new dataset will consist of:
  - 23 numerical features
  - 4 categorical features
  - target variable.



Exploratory Data Analysis:

1. Visualizing the distribution of the Numerical Features
2. Histograms of individual categorical features
3. Correlation Matrix

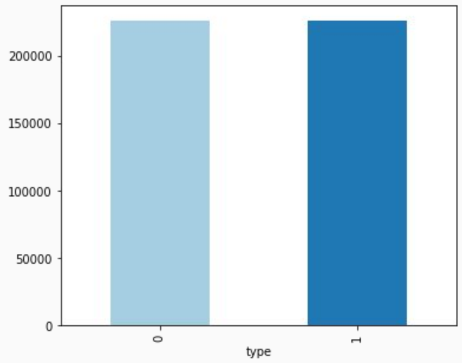
ML Pipeline



Binary Classification

type	count
malware	23645
defacement	95308
phishing	107119
benign	226309

type	count
malicious	226072
benign	226309



Dataset Splitting: Training and Test Set

- **Training set:** 80% of the total number of instances
- **Test set** : 20% of instances

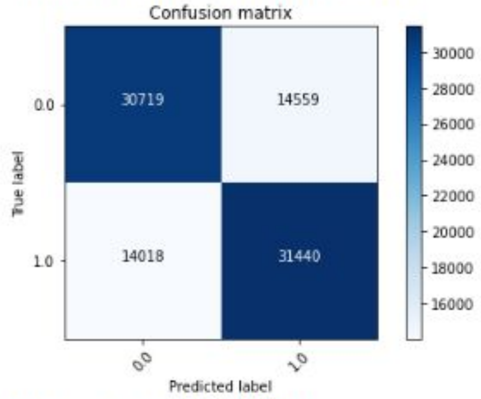
Four classification models:

1. Logistic regression
2. Decision tree
3. Random forest.
4. Gradient boosted decision tree

Results:

Logistic Regression

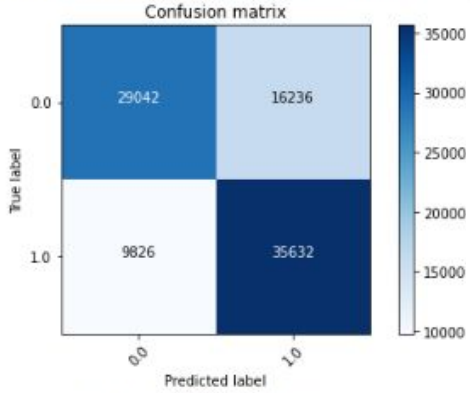
```
metrics: log_reg_malben_std
Confusion matrix without normalization
```




```
Accuracy : 0.6850533415623347
Precision : 0.6850753479591554
Recall : 0.6850402741007038
F1-score : 0.6850578105809983
Test Under ROC Curve (ROC AUC):: 0.759
Area Under Precision-Recall Curve: 0.761
```

Decision Tree

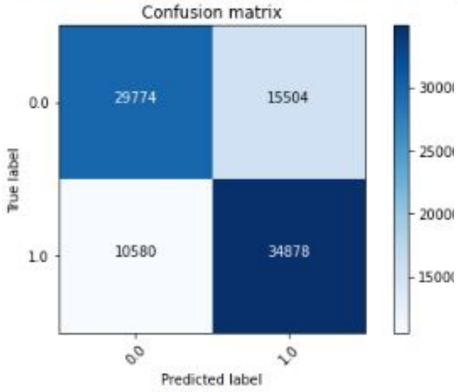
```
Metrics decision_tree_malben_std
Confusion matrix without normalization
```




```
Accuracy : 0.7127711162052548
Precision : 0.7170851322074618
Recall : 0.7126298423405601
F1-score : 0.7148505454680736
Test Under ROC Curve (ROC AUC):: 0.651
Area Under Precision-Recall Curve: 0.594
```

Random Forest

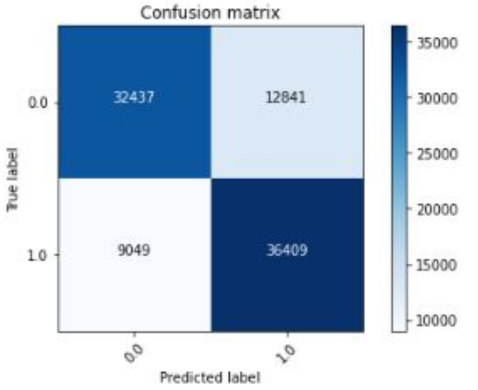
```
Metrics random_forest_malben_std
Confusion matrix without normalization
```




```
Accuracy : 0.712528654558279
Precision : 0.7150456698069648
Recall : 0.7124198685463711
F1-score : 0.7137303541158223
Test Under ROC Curve (ROC AUC):: 0.803
Area Under Precision-Recall Curve: 0.830
```

Gradient Boosted Decision Tree

```
Metrics gradient_malben_std
Confusion matrix without normalization
```




```
Accuracy : 0.7587506612590372
Precision : 0.7605736297560843
Recall : 0.758666806360925
F1-score : 0.7596190214159675
Test Under ROC Curve (ROC AUC):: 0.852
Area Under Precision-Recall Curve: 0.866
```

Multiclass Classification

Problem: Unbalanced Dataset

Solutions:
1- Random Stratified Sampling
2- Prefer other evaluation metrics over accuracy, so f1-score, precision and recall

Three Multiclassification models:

1. Logistic regression
2. Decision tree
3. Random forest.

type	count
malware	23645
defacement	95308
phishing	107119
benign	226309

Random Stratified Sampling

TRAIN	
type	**count**
malware	18925
defacement	76343
phishing	85865
benign	181239

TEST	
type	**count**
malware	4720
defacement	18965
phishing	21254
benign	45070

Results:

Logistic Regression	
Precision	0.72
Recall	0.53
F1-Score	0.60
ROC AUC	0.60

Decision Tree	
Precision	0.83
Recall	0.52
F1-Score	0.64
ROC AUC	0.60

Random Forest	
Precision	0.85
Recall	0.53
F1-Score	0.65
ROC AUC	0.62

Conclusion and Future works:

- Tuning Hyperparameters (with K-fold Cross Validation)
- Increase the amount of data of minority classes by using the SMOTE technique