Malicious Website Detection

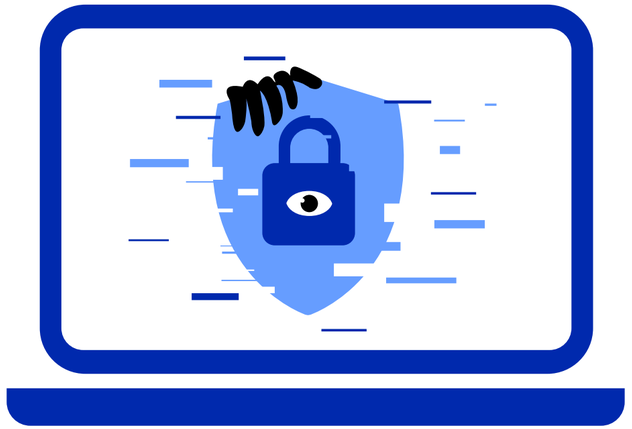
Name: Ritik Jangid
Enrolment No: 2018IMSCS017
Supervisor: Dr. Nishtha Kesswani

Agenda

- Aim
- Dataset
- Implementation
  - Feature Extraction
  - Machine Learning Algorithm
- Results
- Existing Work
- Implementing ANN
- References

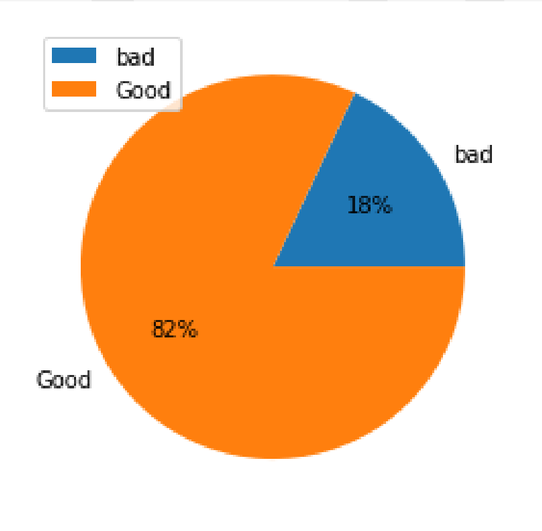
AIM

Predicting if a website is malicious or not by a given URL.



Dataset

Data containing URL and Label indicating good or bad URL.



Malicious_n_Non-Malicious URL

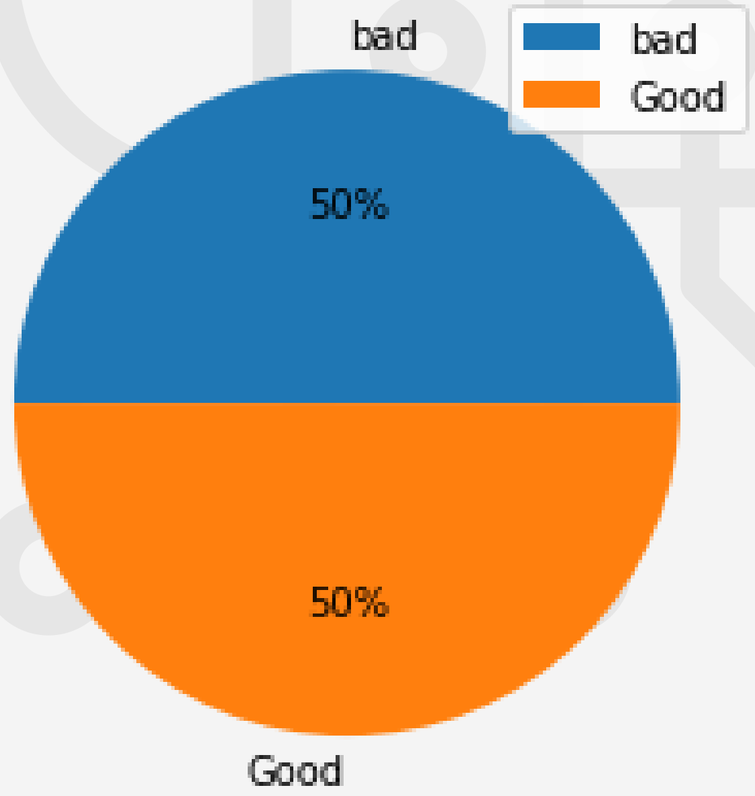
Supervised Machine Learning

kaggle.com

Total 420464 tuples

	url	label
0	diaryofagameaddict.com	bad
1	espdesign.com.au	bad
2	iamagameaddict.com	bad
3	kalantzis.net	bad
4	slightlyoffcenter.net	bad
5	toddscarwash.com	bad
6	tubemoviez.com	bad
7	ipl.hk	bad
8	crackspider.us/toolbar/install.php?pack=exe	bad
9	pos-kupang.com/	bad

Balancing data



Techniques

- Oversampling
- Undersampling
- SMOTE

Implementation

Feature Extraction

Text Based Features

- Bag of Words
- Term Frequency — Inverse Document Frequency (TFIDF)

Domain Based Features

- Domain name
- Creation Time
- Last Update Time
- Expiration Time
- Country
- Top Level Domain

Obfuscation Based Features

- Comment Rows
- Redirection Number
- Links count
- Size of script
- Number of the plus operators

Feature Extraction

Text Based Features

Bag of Words using CountVectorizer

- It represents text documents to a matrix of the token count.
- CountVectorizer is provided by scikit-learn library to create Bag of Words.

TFIDF

- It transforms the text into a usable vector based on the importance of words in the document.
- It is calculated by multiplying Term Frequency (TF) and Inverse of Document Frequency (IDF).

$$w_{i,j} = tf_{i,j} \times idf_i$$

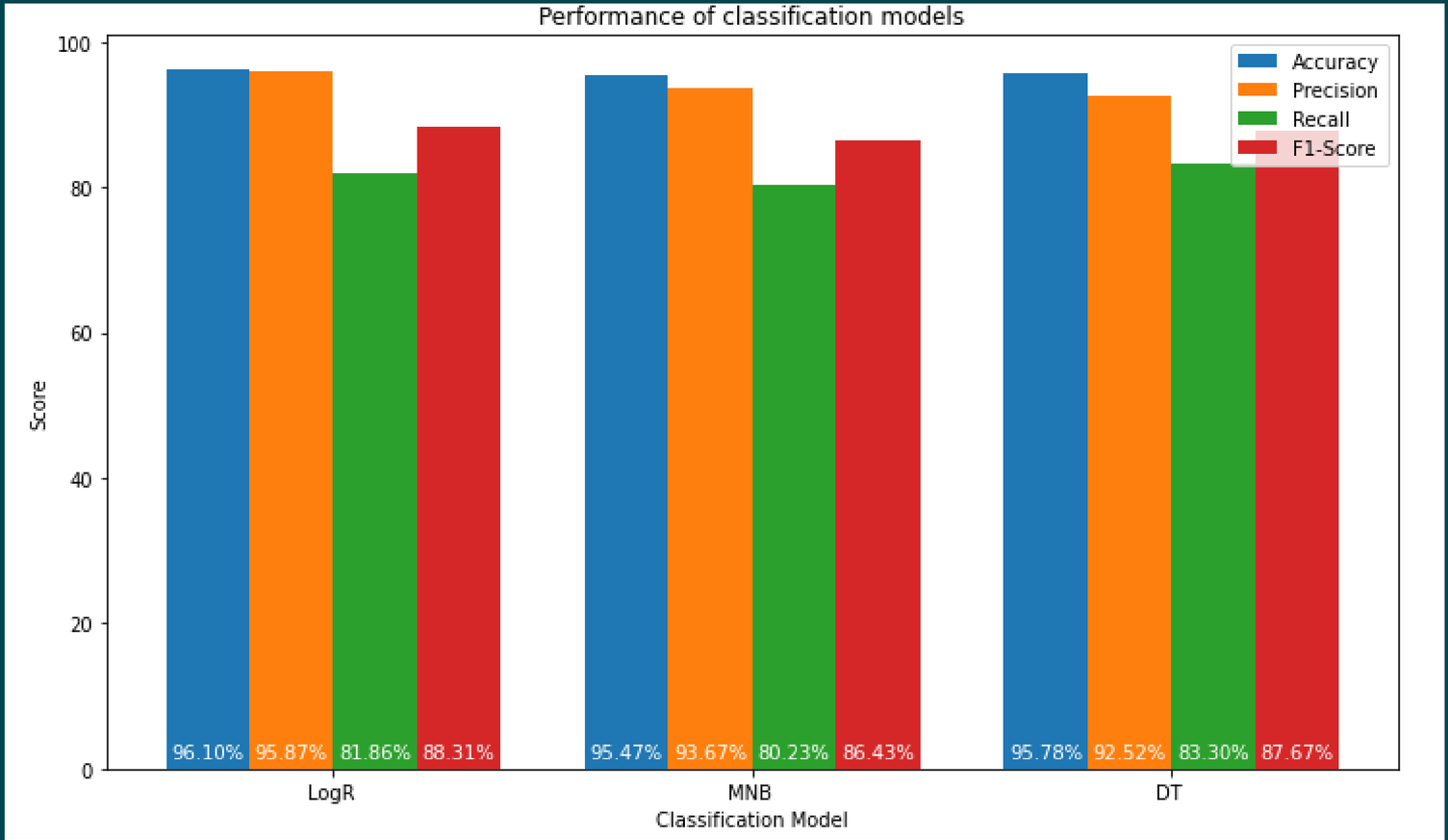
Implementing Machine Leaning Algorithms

A Binary Classification Problem

- **Logistic Regression**
- **Multinomial Naive Bayes**
- **Decision Tree**

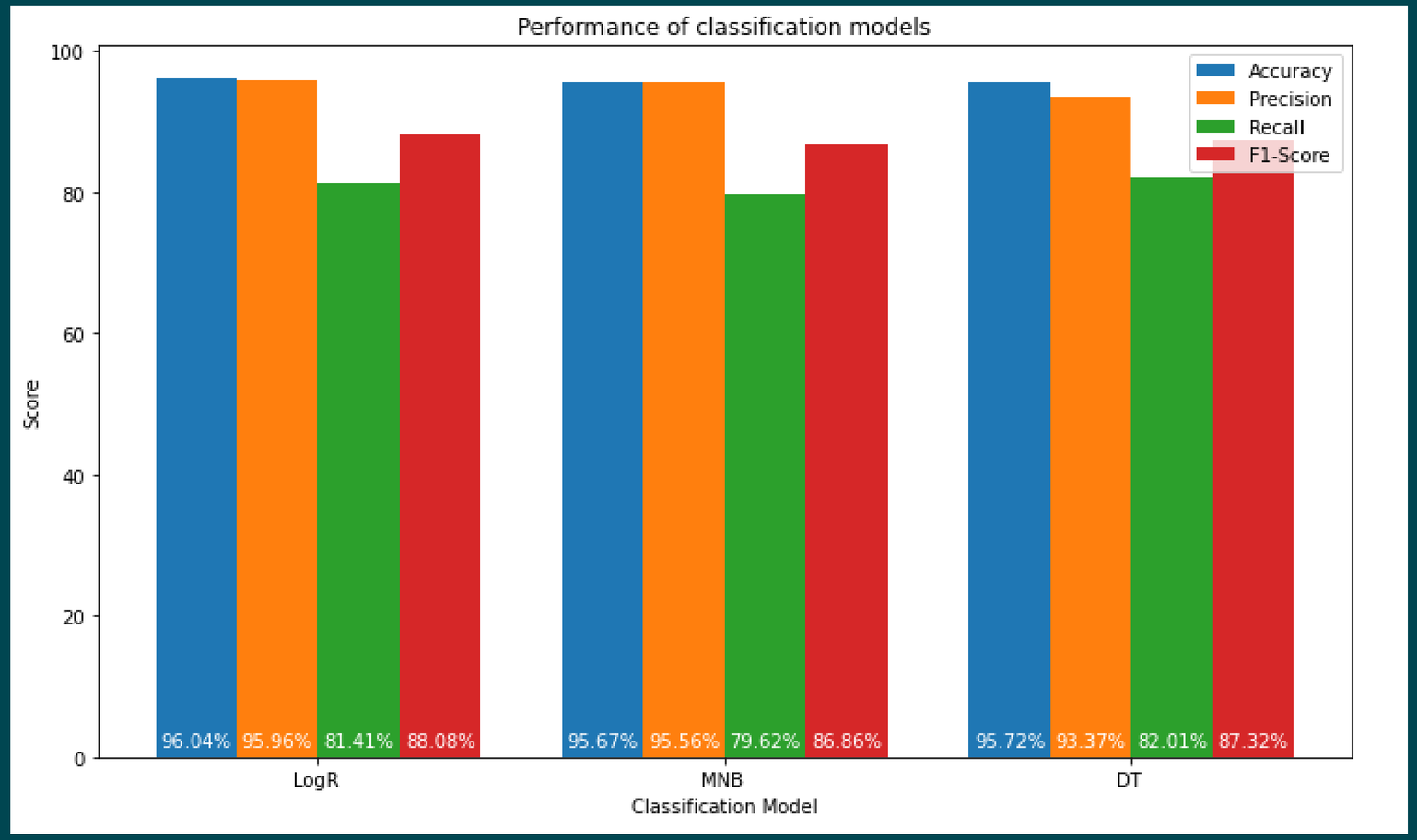
Results

CountVectorizer



Results

TFIDF



Existing Work

Malicious URL Detection using NLP, Machine Learning and FLASK [1]

Uses Kaggle dataset with 500000+ urls

Algorithm	CV	TF-IDF
K-NN	89%	60.9%
DT	96.8%	96.4%
RF	97.1%	97.4%
LR	96.5%	96.1%

Accuracy Comparison

Existing Work

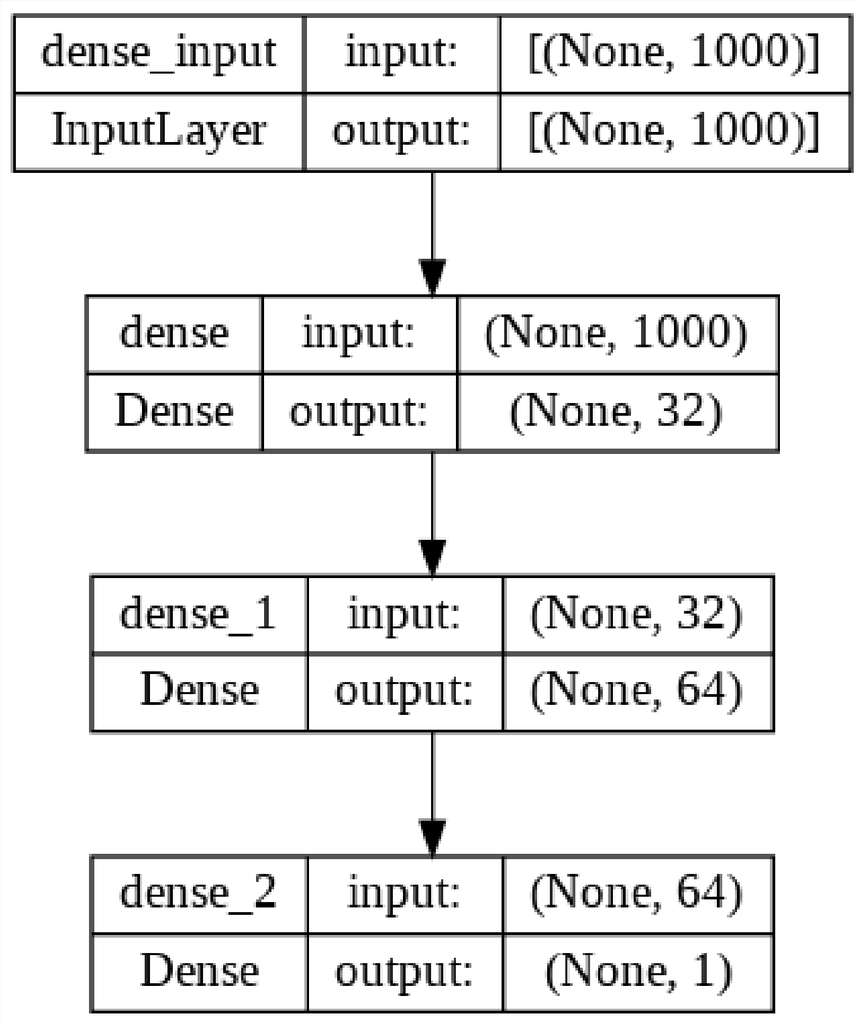
Intelligent Malicious URL Detection with Feature Analysis [2]

- Used domain-based, Alxea based and Obfuscation based features for classification using XGBoost classifier with 99% accuracy.
- The dataset contains 5 million URLs from Alxea ranking, urlquery.net, urlscan.io and Github.

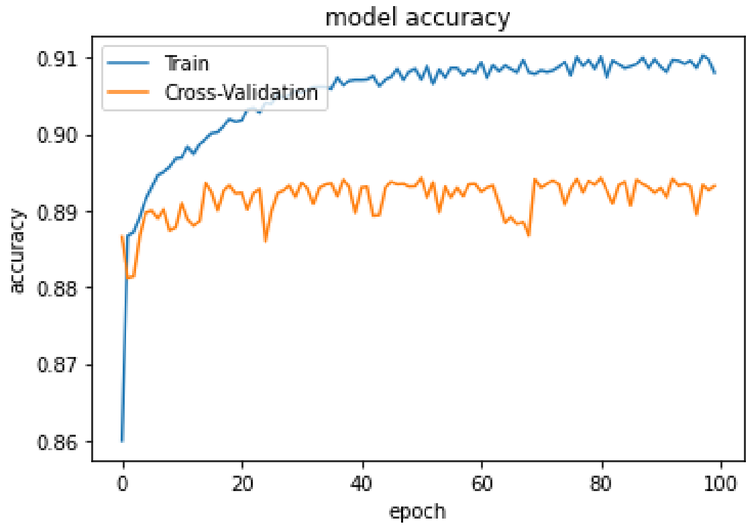
Discovering features for detecting malicious websites [3]

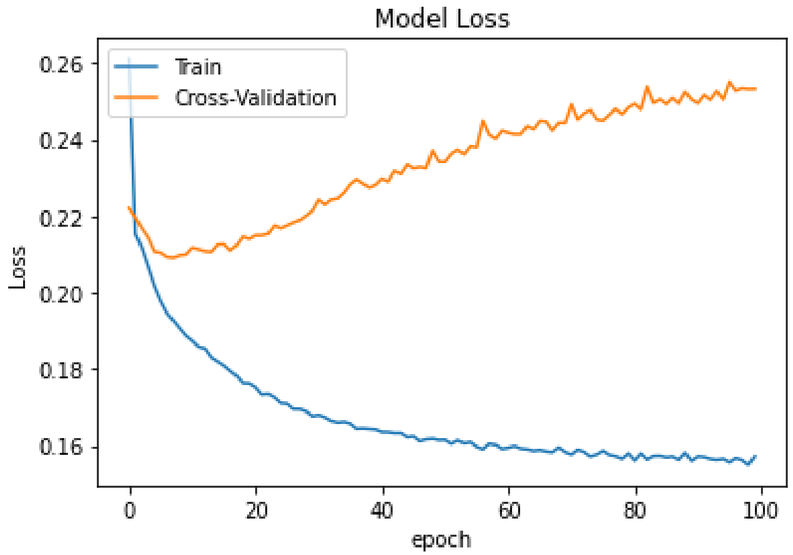
- The dataset contains 34,742 Alexa Top Domains, and 4,441 malicious entries provided by Cisco Talos Intelligence Group.
- Extracted Features from URL, webpage content, and HTTP response header.
- 97 % accuracy with Random Forest Algorithm.

Artificial Neural Network



Artificial Neural Network -Results





Accuracy of ANN model is: 0.8948012030273986

	precision	recall	f1-score	support
0	0.85	0.96	0.90	14916
1	0.96	0.83	0.89	15341
accuracy			0.89	30257
macro avg	0.90	0.90	0.89	30257
weighted avg	0.90	0.89	0.89	30257

References

1. Lakshmanarao, A., M. Raja Babu, and MM Bala Krishna. "Malicious URL Detection using NLP, Machine Learning and FLASK." 2021 International Conference on Innovative Computing, Intelligent Communication and Smart Electrical Systems (ICSES). IEEE, 2021.
2. Chen, Yu-Chen, Yi-Wei Ma, and Jiann-Liang Chen. "Intelligent malicious url detection with feature analysis." 2020 IEEE Symposium on Computers and Communications (ISCC). IEEE, 2020.[1]
3. McGahagan IV, John, et al. "Discovering features for detecting malicious websites: An empirical study." Computers & Security 109 (2021): 102374.[2]
4. McGahagan IV, John, et al. "Discovering features for detecting malicious websites: An empirical study." Computers & Security 109 (2021): 102374.

Any Question ?