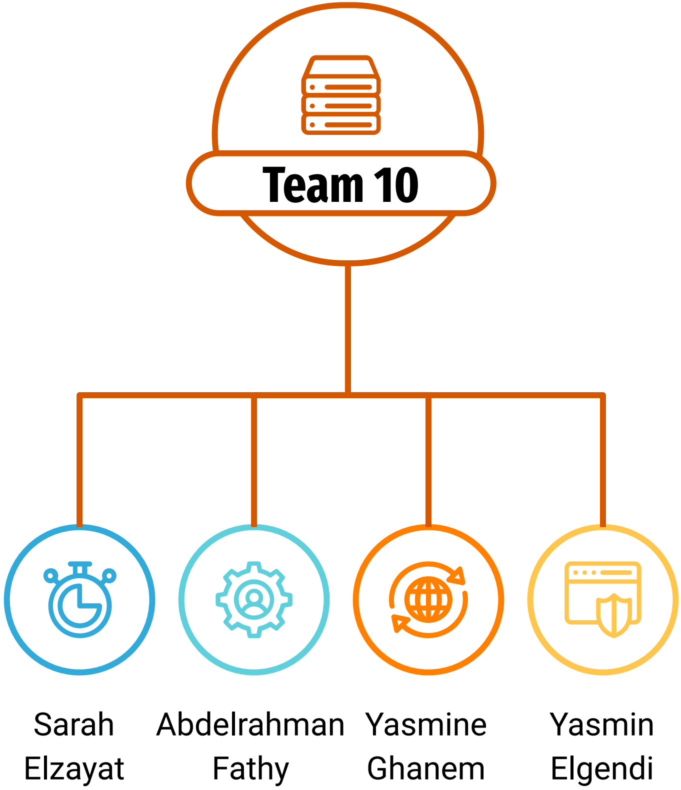
Big Data Project

PhiUSIIL Phishing URL (Website)



Eng/ Omar Samir

Agenda

Problem Intro

Definition of selected problem

Business Part

Business aspects of selected problem

Technical part

Technical aspect and approach solutions for problem

Conclusion

Results and Future Work

Agenda

Problem Intro

Definition of selected problem

Business Part

Business aspects of selected problem

Technical part

Technical aspect and approach solutions for problem

Conclusion

Results and Future Work

Problem Intro

Definition

Phishing

Phishing attacks involve malicious websites pretending to be legitimate with the aim of deceiving individuals into proclaiming personal and sensitive information.

Objective

Security

Develop a comprehensive framework for distinguishing between phishing and legitimate websites.

Agenda

Problem Intro

Definition of selected problem

Business Part

Business aspects of selected problem

Technical part

Technical aspect and approach solutions for problem

Conclusion

Results and Future Work

Business Part

Threat
Data breaches, financial losses, and damage to reputation.

Online Transactions
Robust cybersecurity measures is a MUST.

Brand Safety
Businesses may suffer long-term consequences from a single successful attack

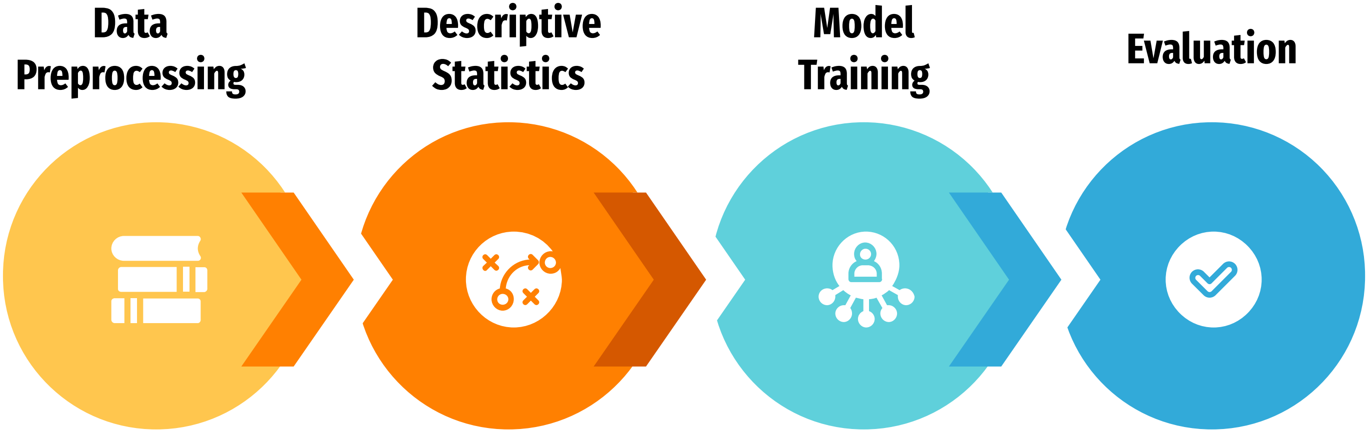
Impact
Benefits across across various sectors, including e-commerce, finance, healthcare, and beyond.

Trust
safeguard their customers' data, preserve trust in online platforms

Agenda


Problem Intro
Definition of selected problem
Business Part
Business aspects of selected problem
Technical part
Technical aspect and approach solutions for problem
Conclusion
Results and Future Work

Project Pipeline



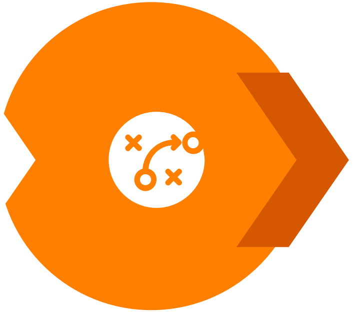
Project Pipeline

Data Preprocessing

1. Read Dataset
2. Check for missing or null values
3. Check for the unique values of some columns {URL}
4. Transfer categorical features to ONE-HOT encoding vector
5. Splitting the data into *Training, Testing & Validation*

Project Pipeline

Descriptive Statistics



To show the distribution of each column over the labels (phishing or legitimate). We used 2 types of plots

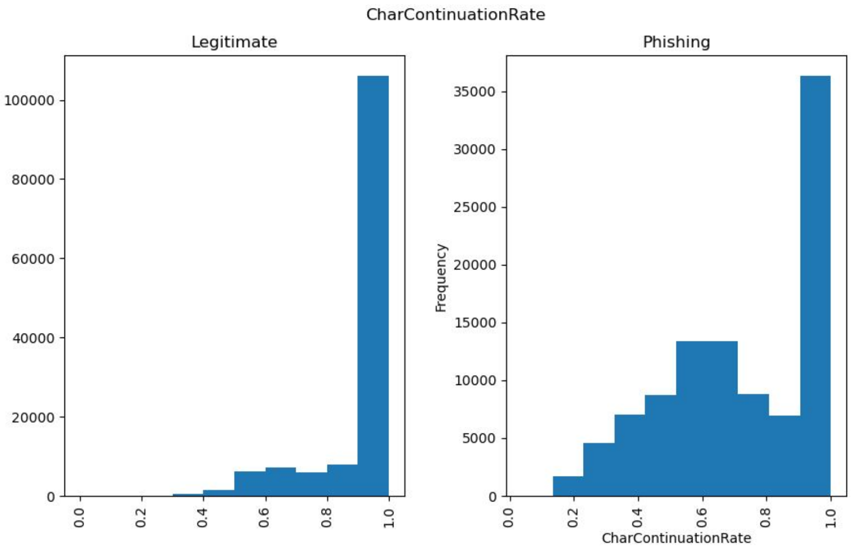
For **numeric features**, we used histograms.

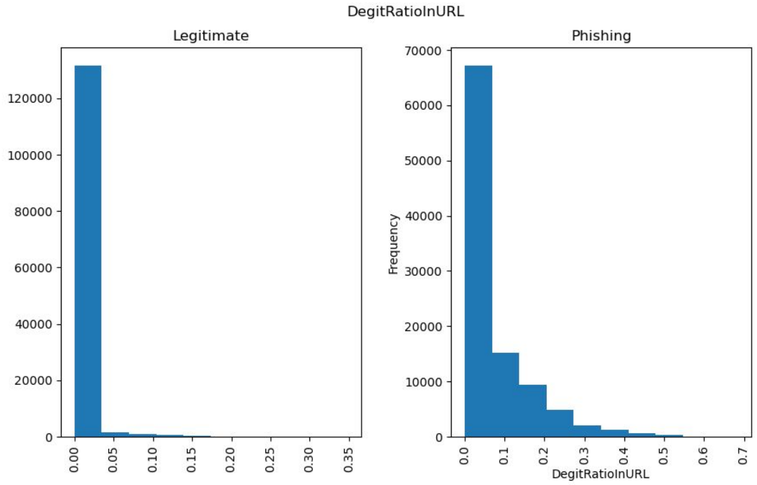
For **binary features**, we used count plots.

By observing the data distributions, we **dropped some features** we found that don't differentiate enough between the classes.

Project Pipeline

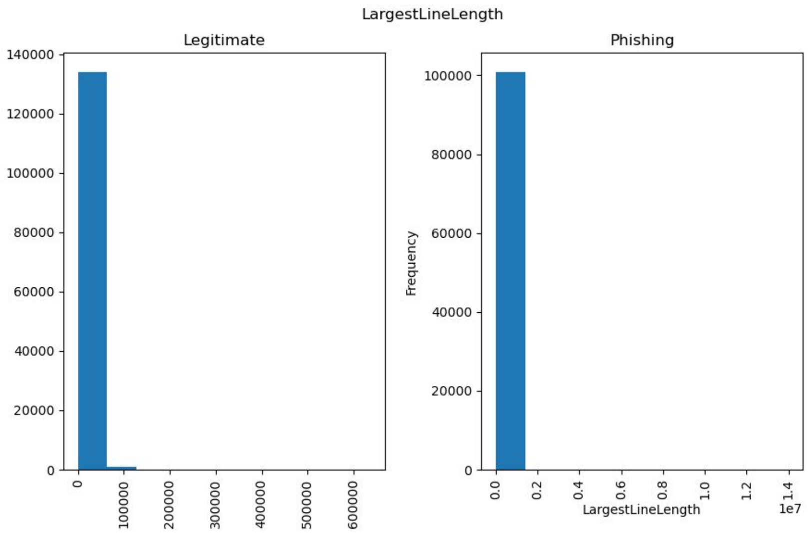
Some **numeric features**

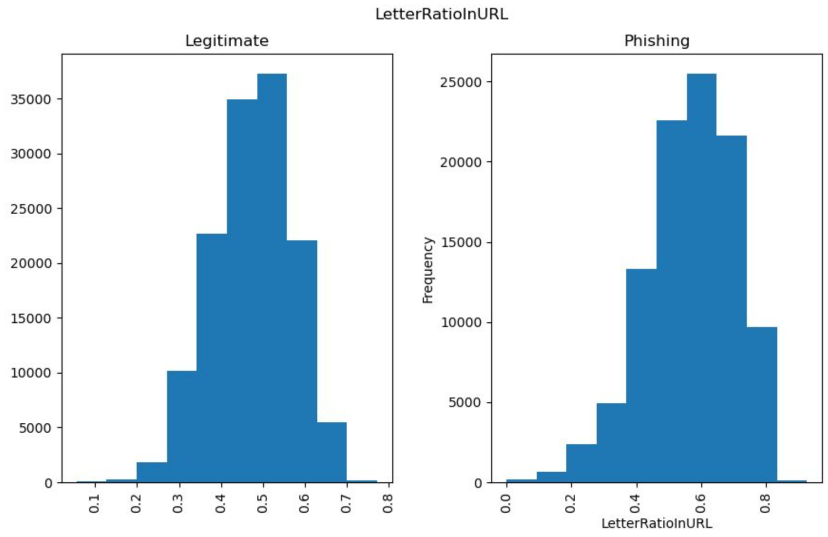




Project Pipeline

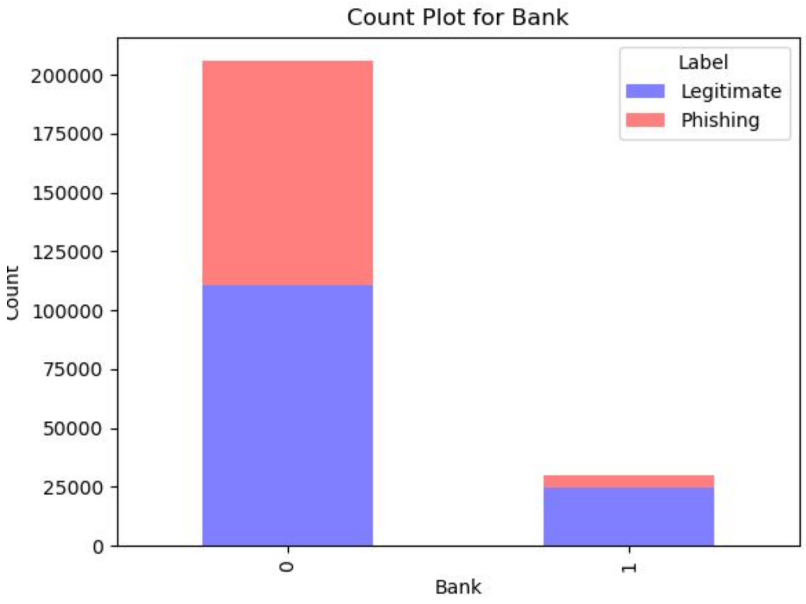
Some **numeric features**

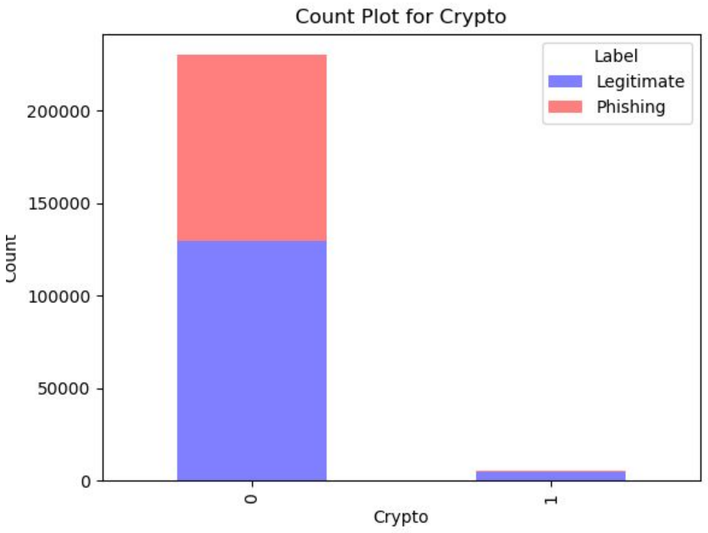




Project Pipeline

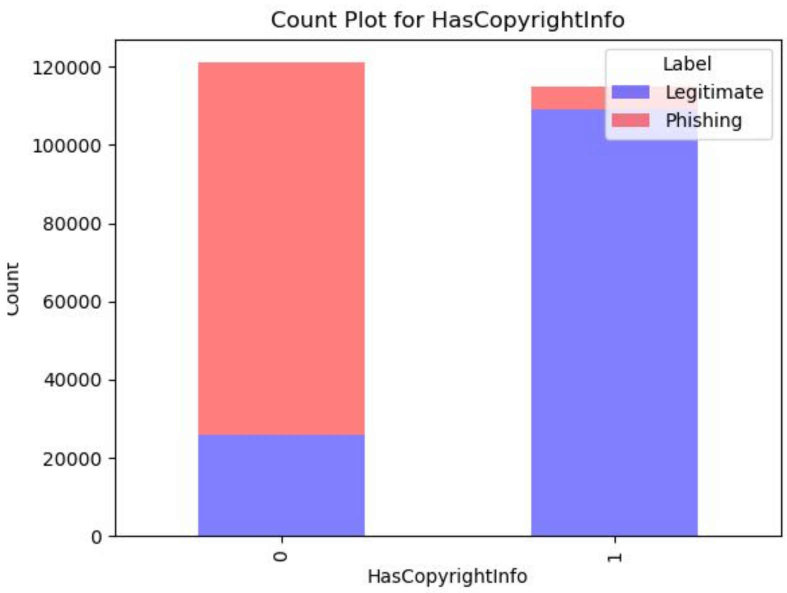
Some **binary features**

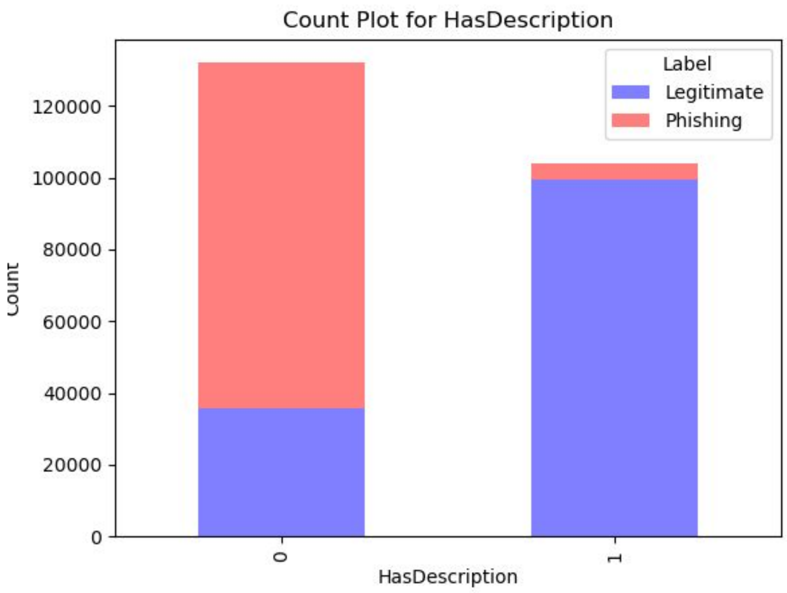




Project Pipeline

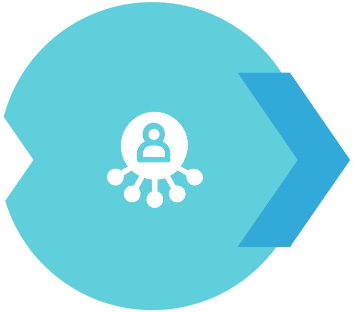
Some **binary features**





Project Pipeline

Model Training



We used various models as:

1. **Random Forest**
2. **SVM**
3. **KNN** (With MapReduce and without)
4. **Naive Bayes** (With MapReduce and without)

Project Pipeline

Random Forest

Initial run

Using cross validation, with 5 folds.

```
Test Accuracy = 0.9999
Test Precision = 0.9999
Test Recall = 0.9999
Test F1 Score = 0.9999
+-----+----------+-----+
|label|prediction|count|
+-----+----------+-----+
|    0|       0.0|24956|
|    1|       1.0|33370|
```

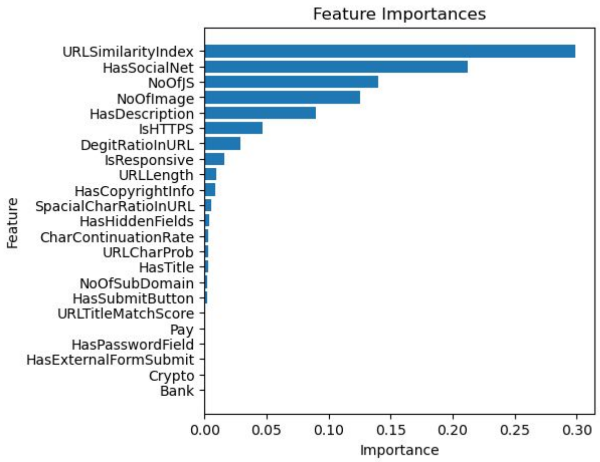
Train with all features, default parameters

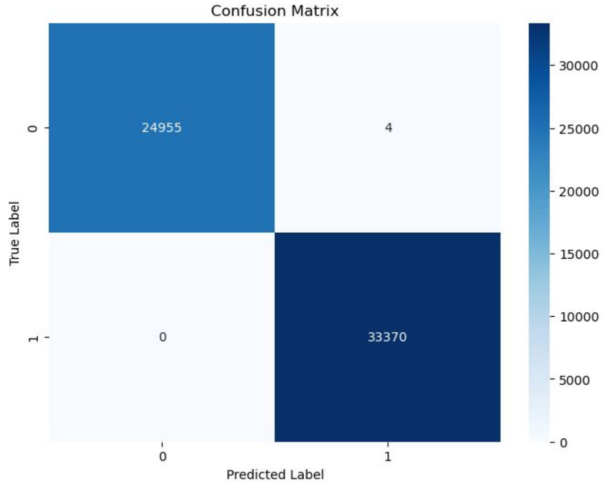
```
Test Accuracy = 0.9999314234771725 6513
Test Precision = 0.99993143169632792144
Test Recall = 0.99993142347717256513
Test F1 Score = 0.99993142278429758552
Validation Accuracy = 0.99991483321504615045
Validation Precision = 0.99991483415042059502
Validation Recall = 0.99991483321504603943
Validation F1 Score = 0.99991483285983306928
```

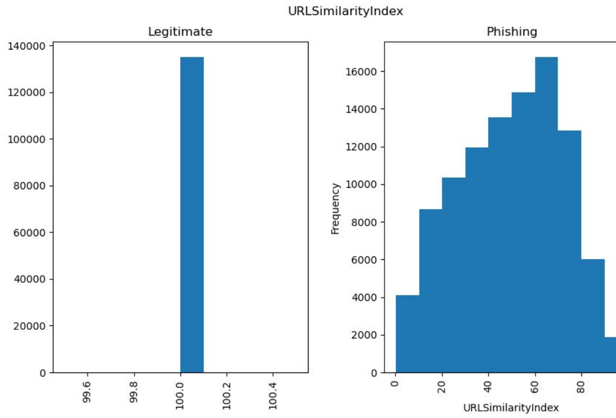
Project Pipeline

Random Forest

By observing the histogram distributions, the URL similarity index almost completely classifies the URL's correctly. Few features have significant weight



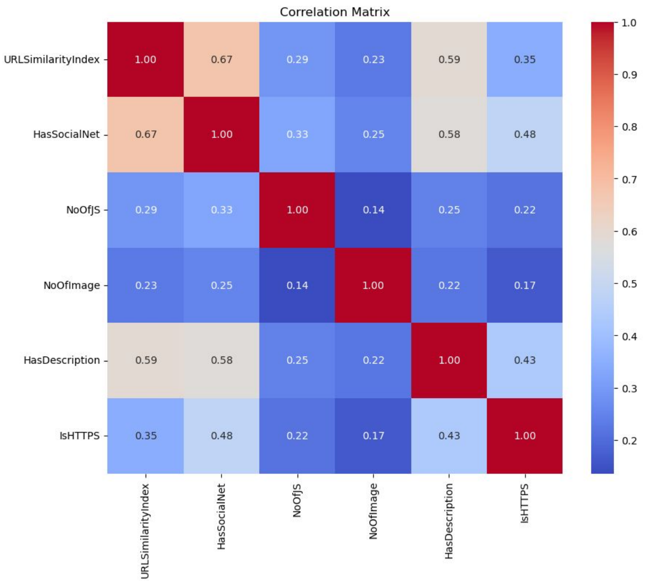


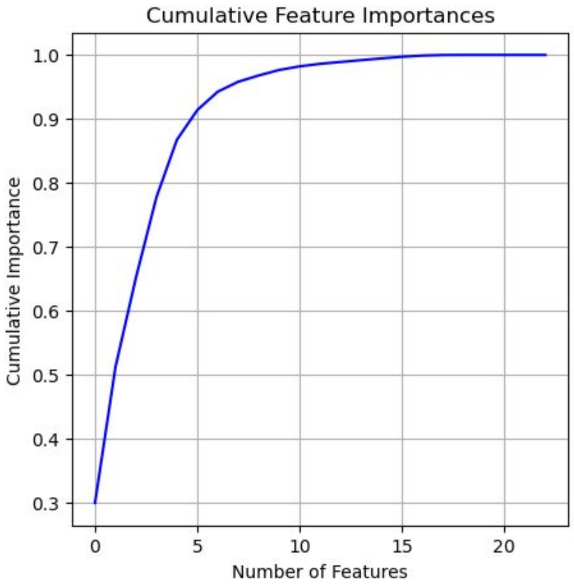


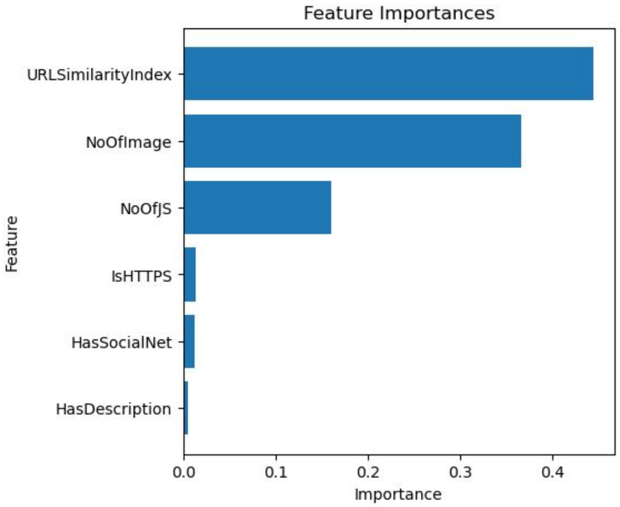
Project Pipeline

Random Forest

Final Model: we capped it down to the features that support up to 90% of the importance, which narrowed it to about 6 features





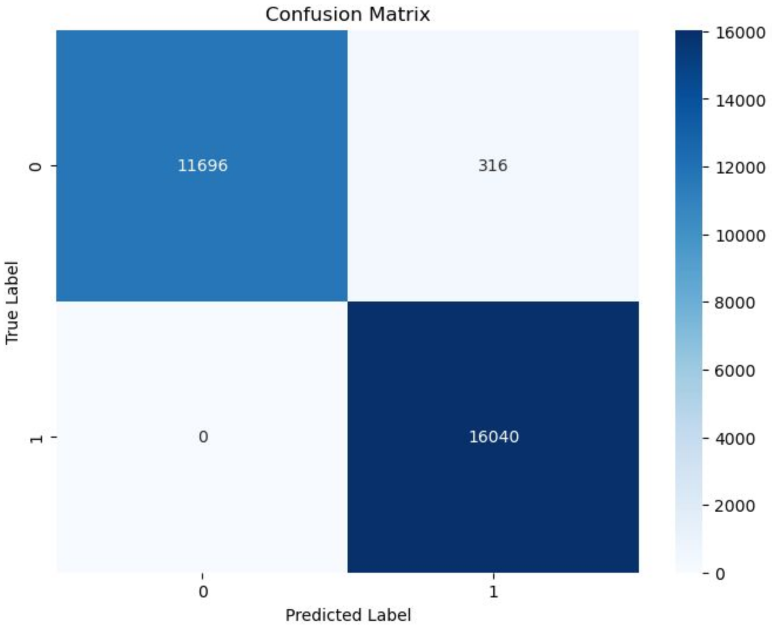


Project Pipeline

SVM

Validation data

```
Validation LinearSVC Accuracy = 0.98873520604591469407
Validation LinearSVC Precision = 0.98895284329765664744
Validation LinearSVC Recall = 0.98873520604591469407
Validation LinearSVC F1 Score = 0.98871507279793557910
```

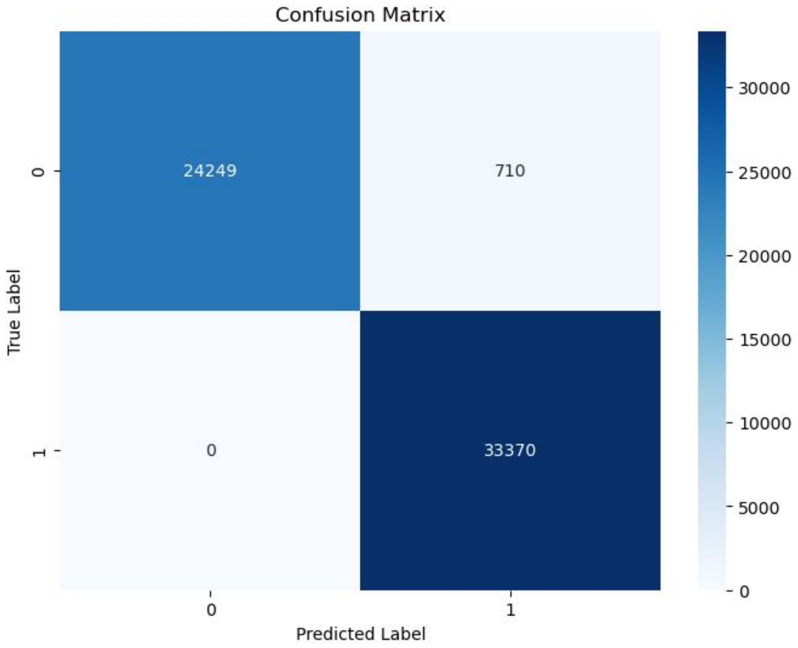


Project Pipeline

SVM

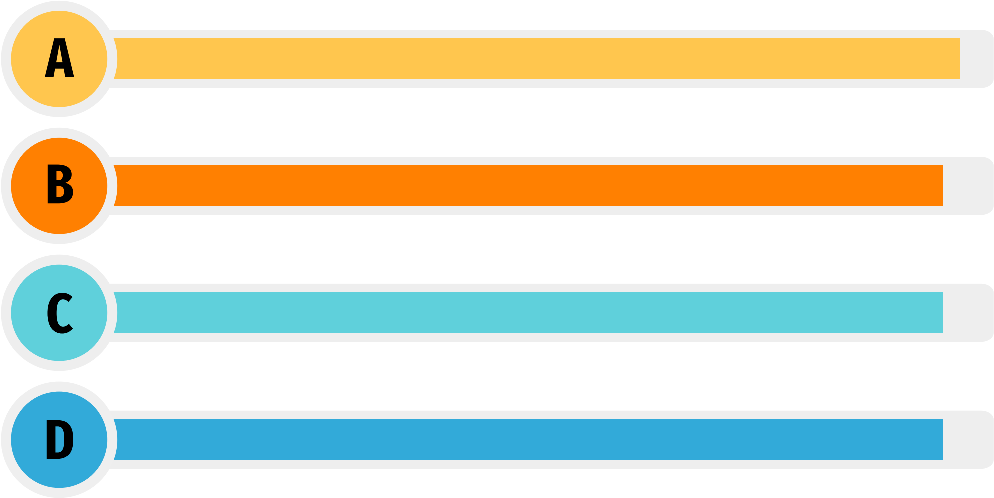
Test data

```
Test LinearSVC Accuracy = 0.987827667198
Test LinearSVC Precision = 0.98808125746
Test LinearSVC Recall = 0.98782766719813
Test LinearSVC F1 Score = 0.987803917562
```



Project Pipeline

Evaluation



A RF 99.64 %

B SVM 98.78%

C KNN 98.62%

D Naive Bayes 99.98%

Agenda

Problem Intro

Definition of selected problem

Business Part

Business aspects of selected problem

Technical part

Technical aspect and approach solutions for problem

Conclusion

Results and Future Work

Conclusion

Data

Data had few features with actual importance/weight that is significant to classification. Therefore it was biased to an extent

Models

All models used got an accuracy above 98 % in classifications.

THANK YOU!

Any Questions?

Eng/ Omar Samir