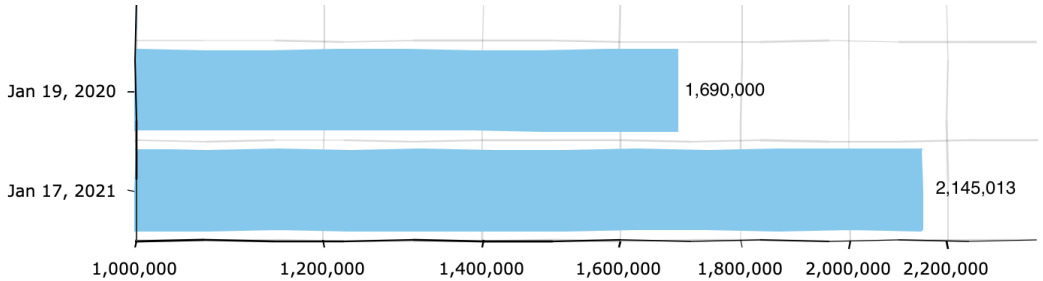
is it phishing?

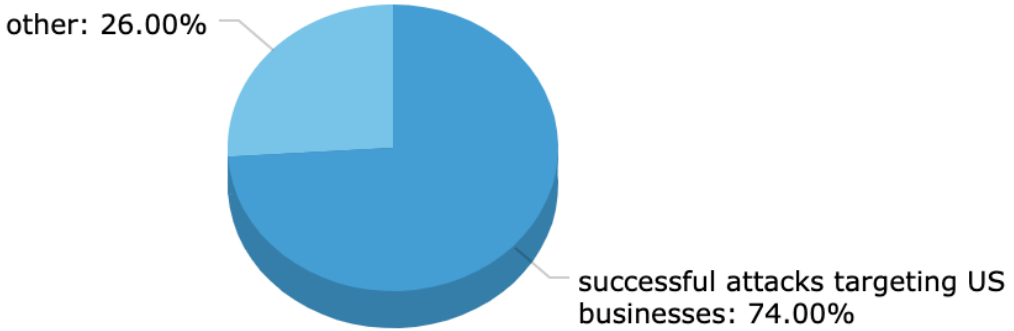
By: Batoul Alosaimi
Norah Alqahtani
Shroaq Almutiri

INTRODACTION AND SOME STATISTICE

- According to Google Safe Browsing; Google has registered 2,145,013 phishing sites as of Jan 17, 2021. This is up from 1,690,000 on Jan 19, 2020 (up 27% over 12 months).



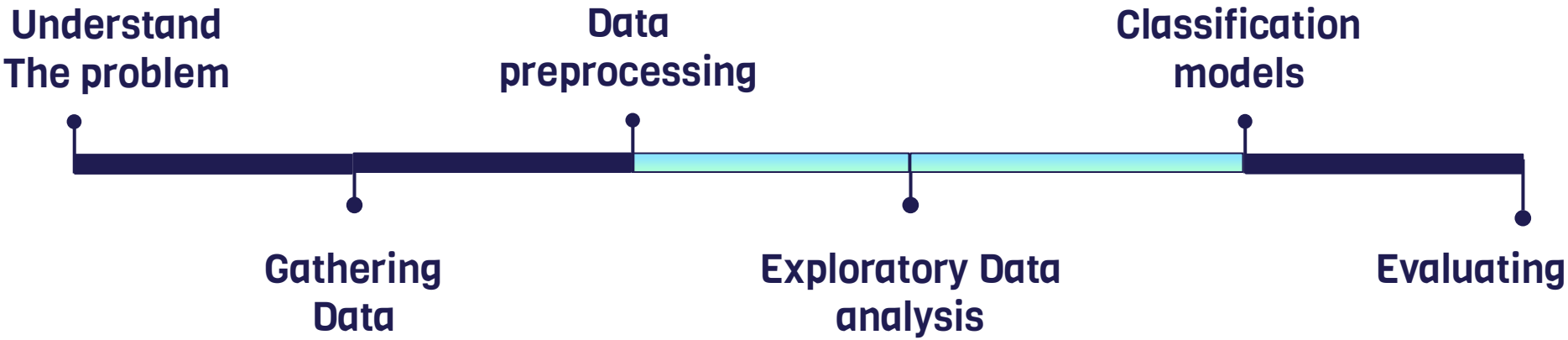
INTRODACTION AND SOME STATISTICE



INTRODACTION AND SOME STATISTICE

- IBM found that customers' personally identifiable information (PII) was both the most commonly compromised type of data and the costliest.

METHODOLOGY



DATA SET

- Public source from Declaration on Scientific paper named 'Phishing URL Classification using Machine Learning'

- Size

10000 records

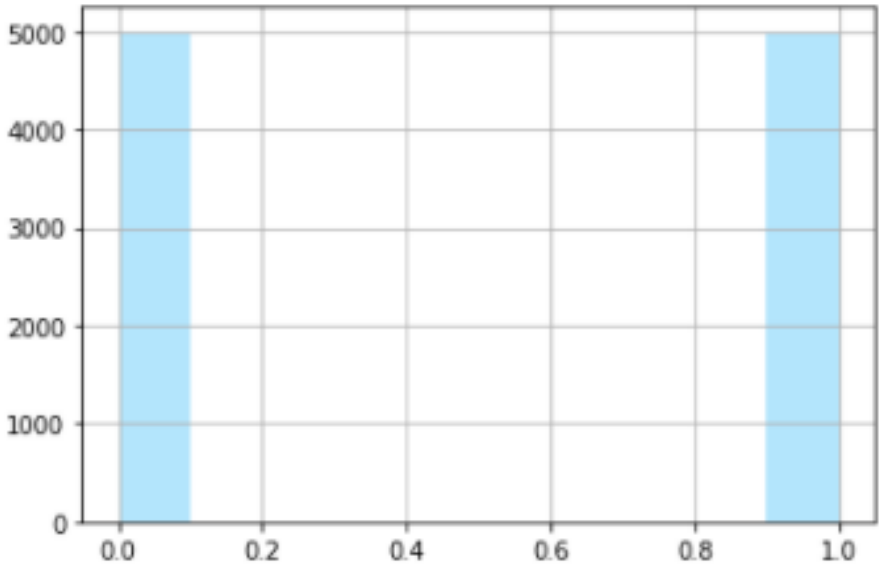
15 columns

- Target

1 : Phishing

0 : Benign

EXAPLORTY DATA ANALYISI

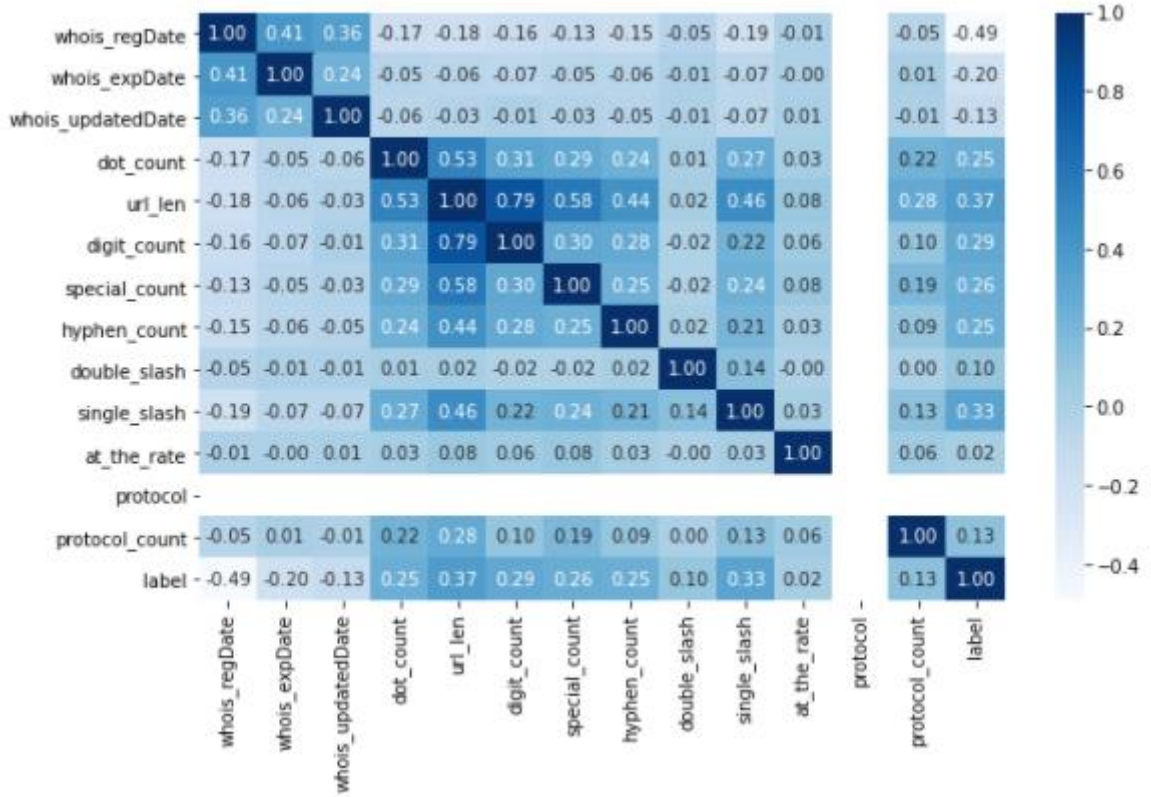


Target :

- 0 is benign
- 1 is Phishing

EXAPLORTY DATA ANALYISI

correlation between the Target and features, and between each features



Data Preparation

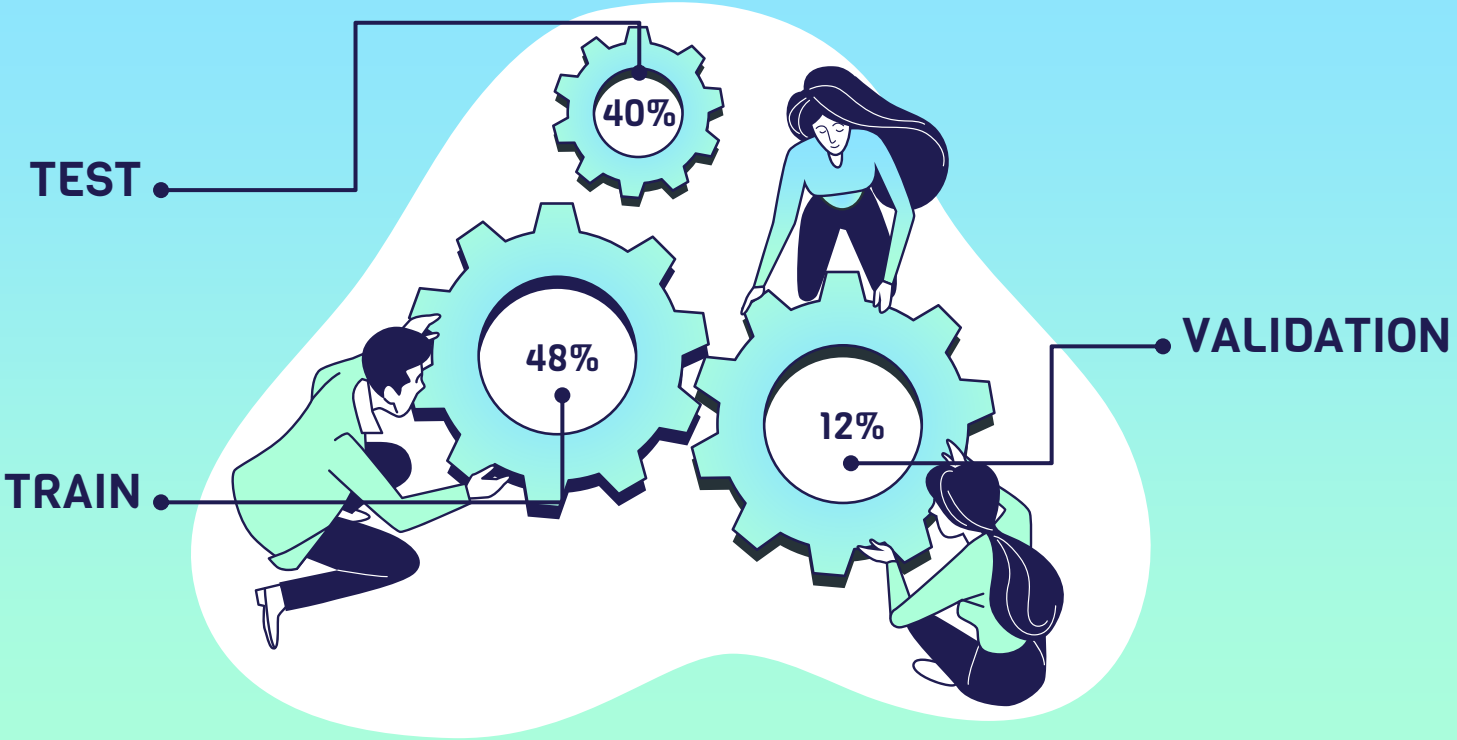
Data Cleaning :

- Find and drop null values.
- Find duplicated.

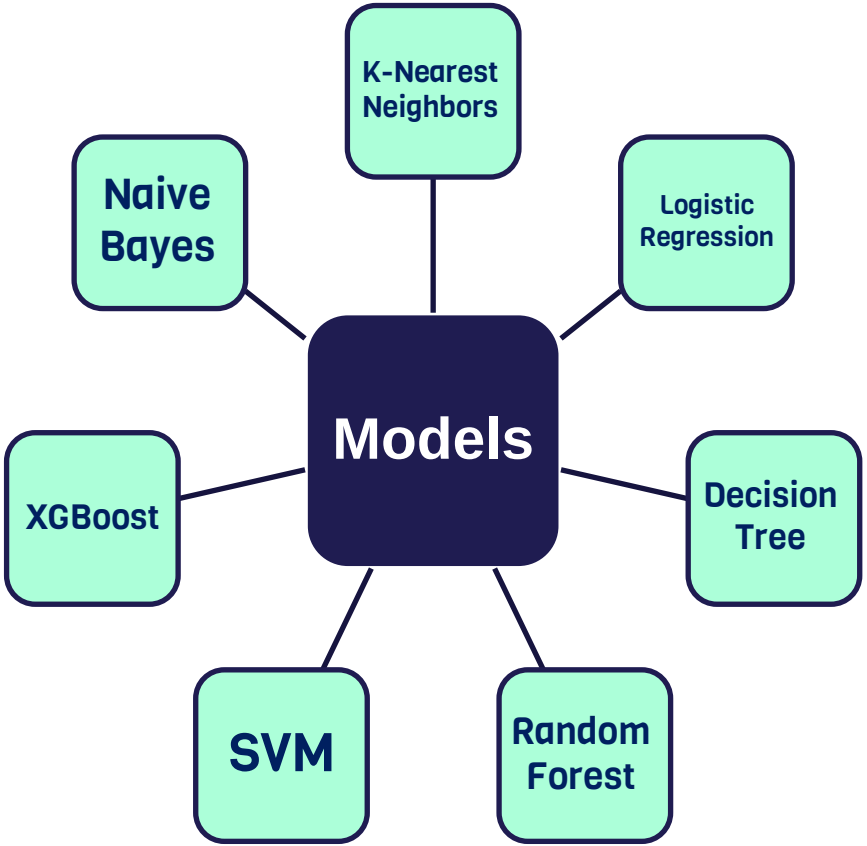
Feature Engineering :

- we can see that 'protocol' column has no values but 0 in it so we will drop it.
- The scale of at_the_rate column is deferent than the other columns.

Splitting the data



MODELLING

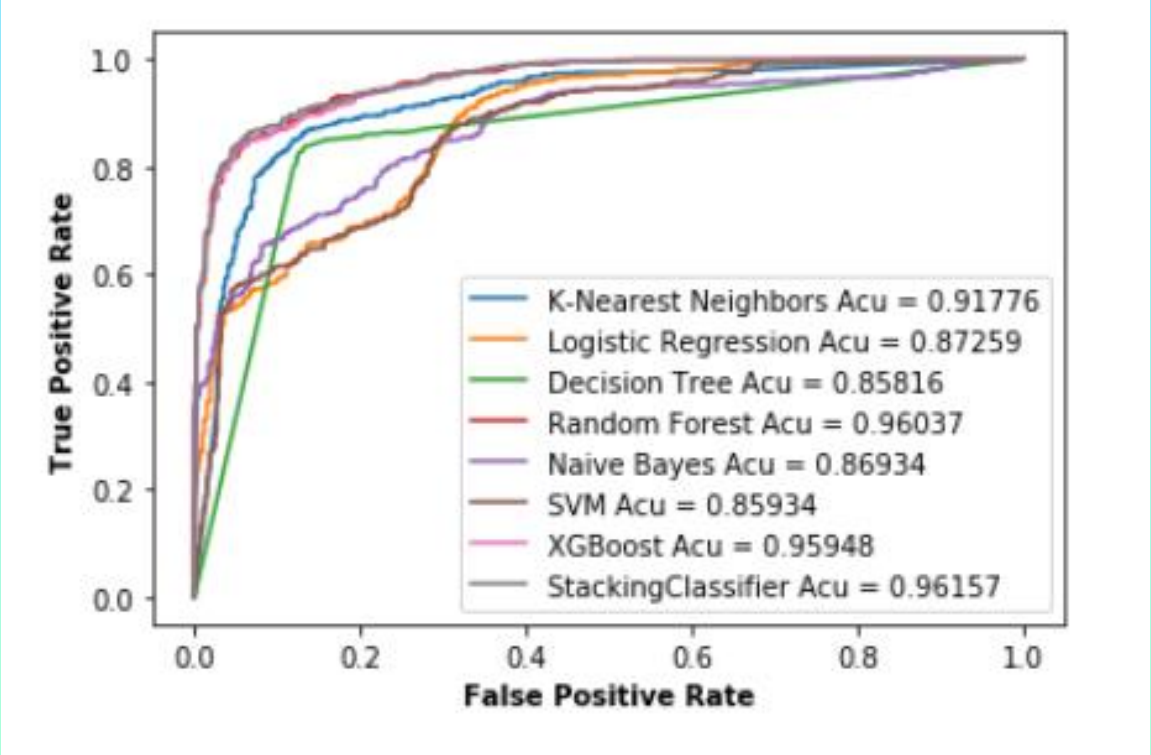

K-Nearest Neighbors
Naive Bayes
Logistic Regression
Models
XGBoost
Decision Tree
SVM
Random Forest

EVALUATING

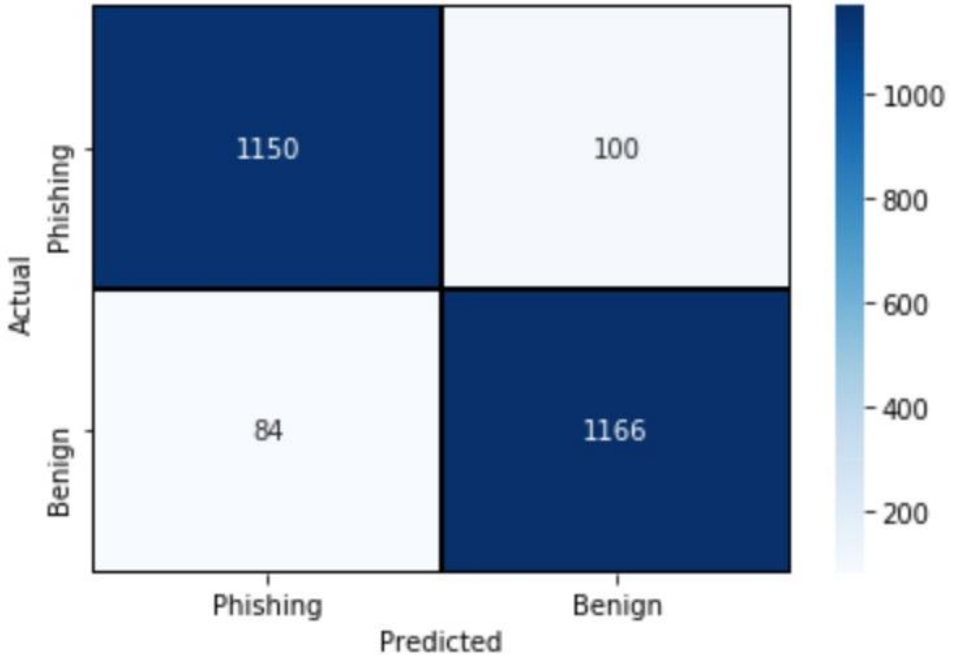
Table showing all confusion matrix values of the models.

	Model	Accuracy	Recall	Precision	F1 score
0	KNN	0.860833	0.873355	0.855072	0.864117
1	Logistic Regression	0.755833	0.789474	0.744186	0.766161
2	Decision Tree	0.851667	0.837171	0.865646	0.851171
3	Random Forest	0.884167	0.881579	0.888889	0.885219
4	Naive Bayes	0.744167	0.523026	0.949254	0.674443
5	SVM	0.775833	0.856908	0.741110	0.794813
6	XGBoost	0.877500	0.875000	0.882255	0.878613
7	StackingClassifier	0.889167	0.891447	0.889984	0.890715

ROC CURVE


True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
K-Nearest Neighbors Acu = 0.91776
Logistic Regression Acu = 0.87259
Decision Tree Acu = 0.85816
Random Forest Acu = 0.96037
Naive Bayes Acu = 0.86934
SVM Acu = 0.85934
XGBoost Acu = 0.95948
StackingClassifier Acu = 0.96157

Best model for our project (Random Forest)



RF Accuracy= 0.93
RF F1 score= 0.93

Flask API

- Save the Model
- Create fanction which take URL and extract its features
- Builds website using Flask API
- http://127.0.0.1:5000/

Conclusion

- Our model solve the problem of classifing URLs with accuricy(0.93) and F1(0.93)
- By comparing our results with the results of the previous work; our accuracy is better where theirs is 0.87

FUTUER WORK

- Publish a website
- Improving the Model until F1= 0.96
- Create ourown Dataset

THANKS!