DETECTING PHISHING WEBSITES USING DATA MINING ALGOs

By :

Reshma Krishnakumar

Santhosh Bodla

Surbhi Dogra

Tanya Gupta

TEAM - 11

OBJECTIVE

Phishing is one of the most dangerous threats to your online accounts and data, because these kind of exploits hide behind the guise of being from a reputable company or person, and use elements of social engineering to make victims far more likely to fall for the scam.

- In order to prevent the user from these practices we try to predict the data based on the URL of the website.

- Analyze different classification algorithms and choose the best model based on the performance metrics like F1 score.

BACKGROUND

- To give you an idea: 74% of U.S. organizations experienced a successful phishing attack in 2020, a 14% increase from 2019. According to the FBI, phishing is the most common type of cybercrime. In 2020, they saw 12 times more phishing attacks than in 2016.

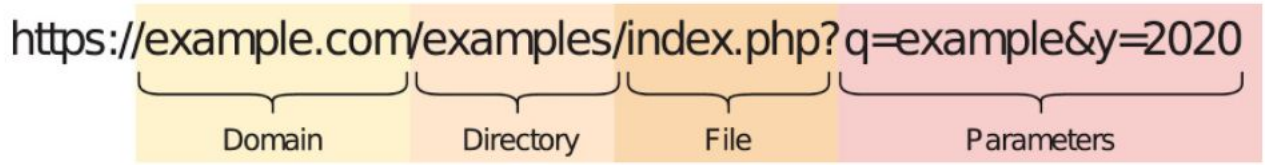
- COVID-19 phishing scams likely accelerated the increase in 2020. The pandemic introduced new opportunities for scammers, and some estimate that drove up cybercrime by 600%!

DATA SET DETAILS

- Total number of instances: 58,645
- Number of non-phishing instances: 27,998
- Number of phishing instances: 30,647 Total
- Number of features: 111
- Reason to choose: More Balanced

```
Total phishing URL 30647
Total legit URL 27998
```

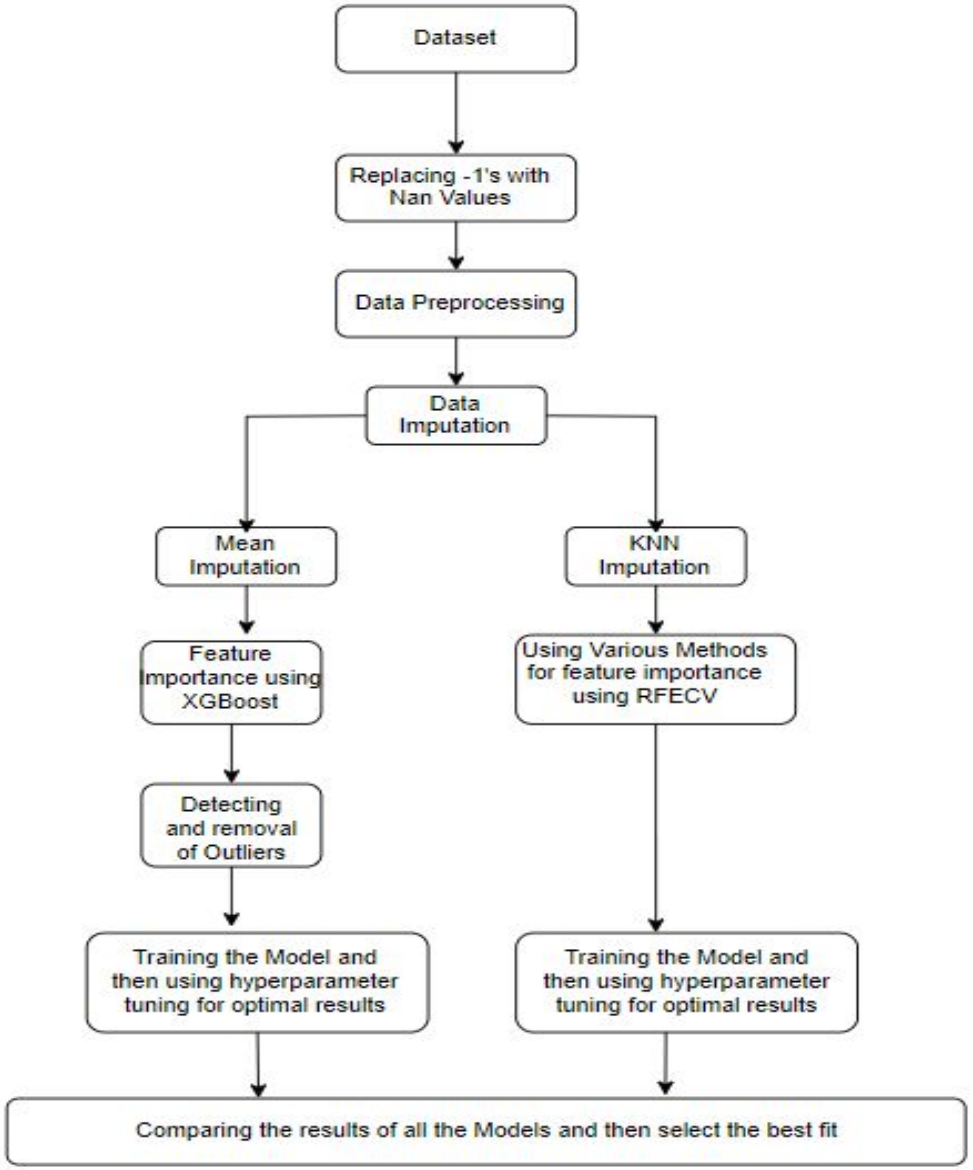
DATA SET DESCRIPTION



Our Data set is divided into 6 Different parts as

- Attributes based on the whole URL properties
- Attributes based on the domain properties
- Attributes based on the URL directory properties
- Attributes based on the URL file properties
- Attributes based on the URL parameter properties
- Attributes based on the URL resolving data and external metrics

WORKFLOW

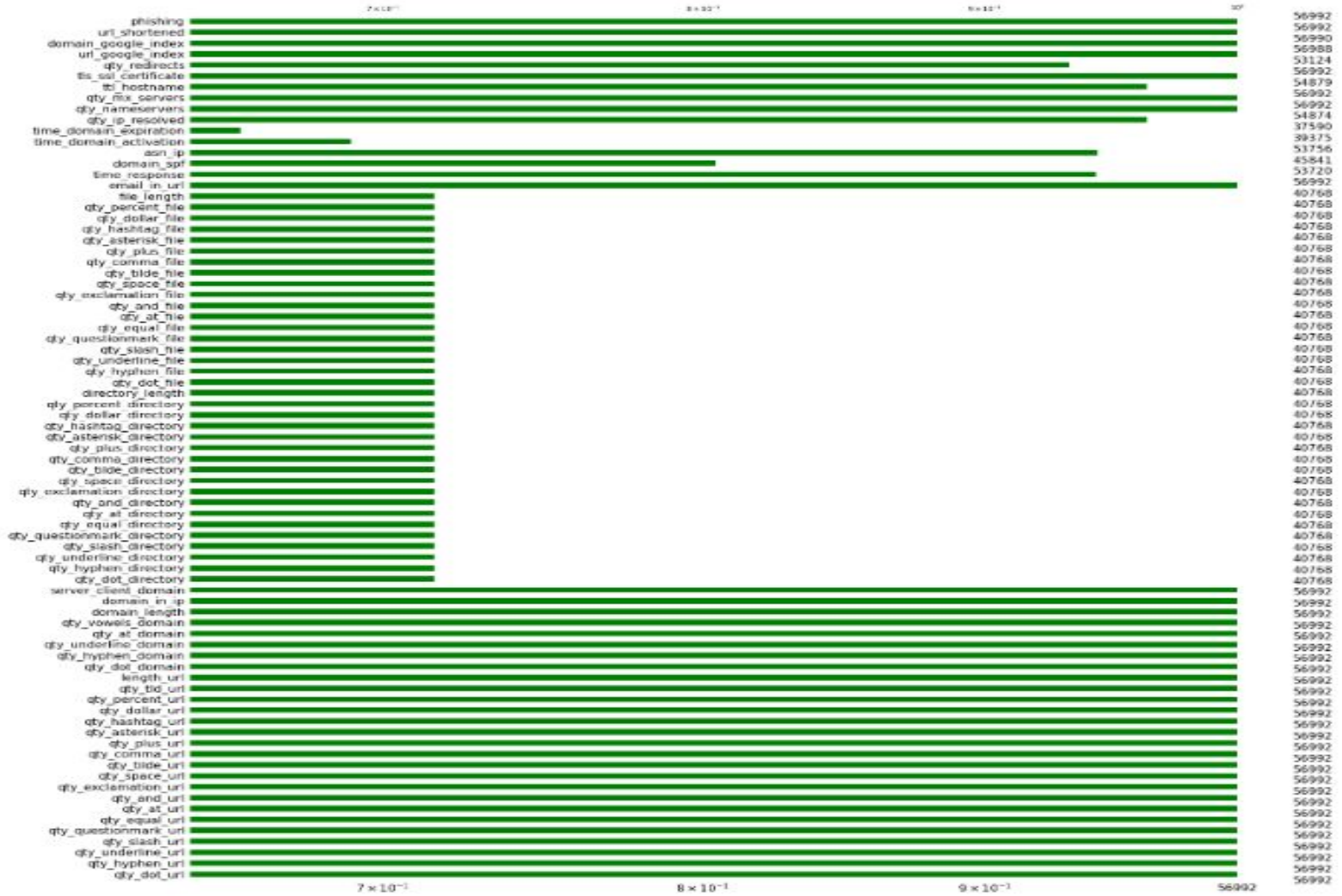

Dataset
Replacing -1's with Nan Values
Data Preprocessing
Data Imputation
Mean Imputation
KNN Imputation
Feature Importance using XGBoost
Using Various Methods for feature importance using RFECV
Detecting and removal of Outliers
Training the Model and then using hyperparameter tuning for optimal results
Training the Model and then using hyperparameter tuning for optimal results
Comparing the results of all the Models and then select the best fit

DATA PRE-PROCESSING

- Dropping all the Columns whose that have the same value in all samples
- Delete Duplicate rows (if any)
- Convert the -1’s to NAN values
- Dropping the columns having value -1 in more than 80% of the cells

VISUALIZATION

Visualize the Missing data using a bar chart from missingo library.

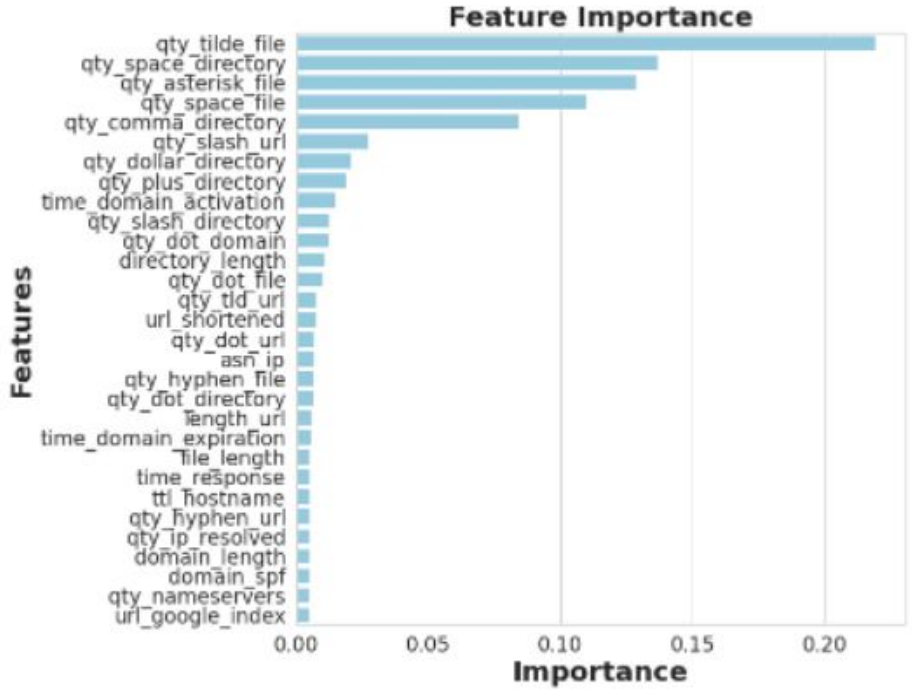


METHODOLOGIES - MEAN IMPUTED DATA

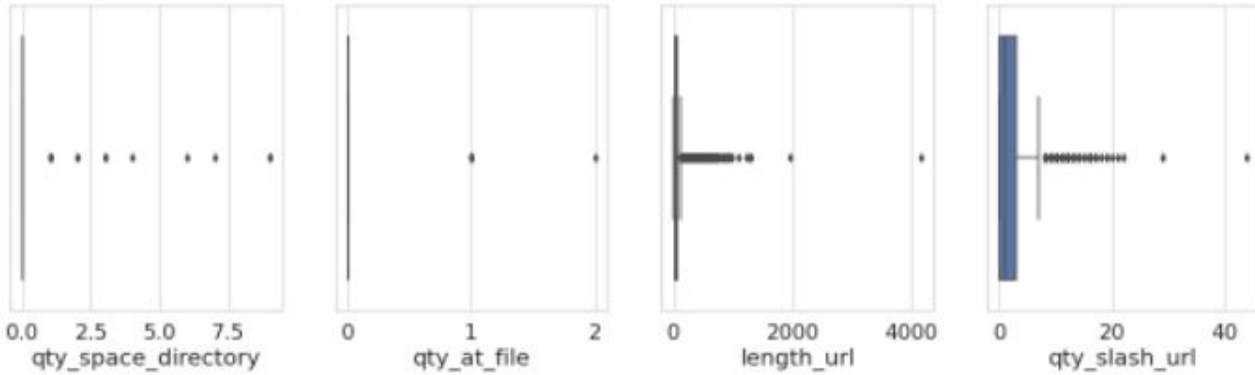
- Feature Importance using XGBoost
- Hyper Parameter Tuning of XGBoost
- Detection and Removal of Outliers
- Visualization and Analysis of data
- Model Training by leveraging Hyperparameter tuning for Obtaining Optimal Results

MEAN IMPUTED DATA - FEATURE IMPORTANCE AND OUTLIERS

Finding out the most important features which contribute to detecting phishing

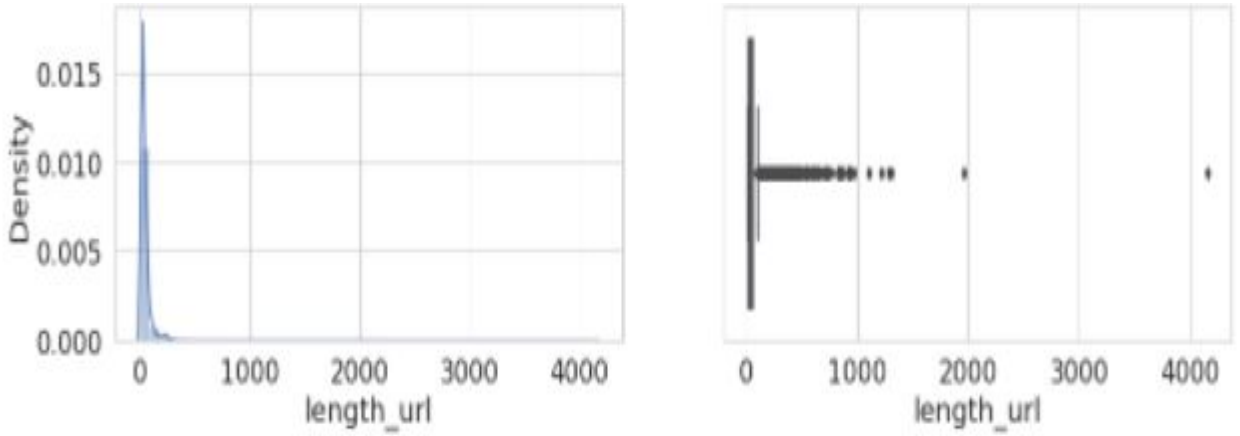


Visualizing Outliers for certain features

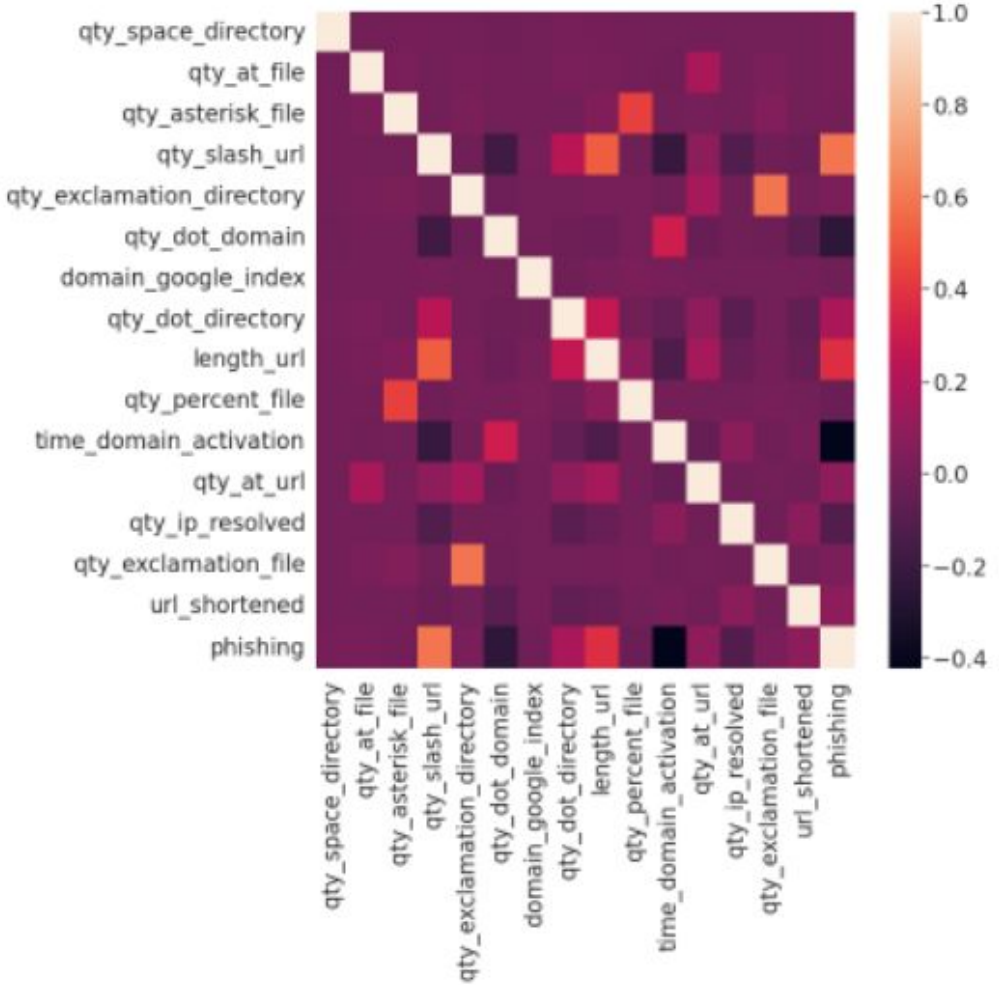


MEAN IMPUTED DATA

Post Removal of Outliers using IQR



Correlation between Phishing and other features



RESULTS OBTAINED MEAN IMPUTED DATA

Classifier	Accuracy
Logistic Regression	99.94% (Overfitting)
KNN Classifier	94.42%

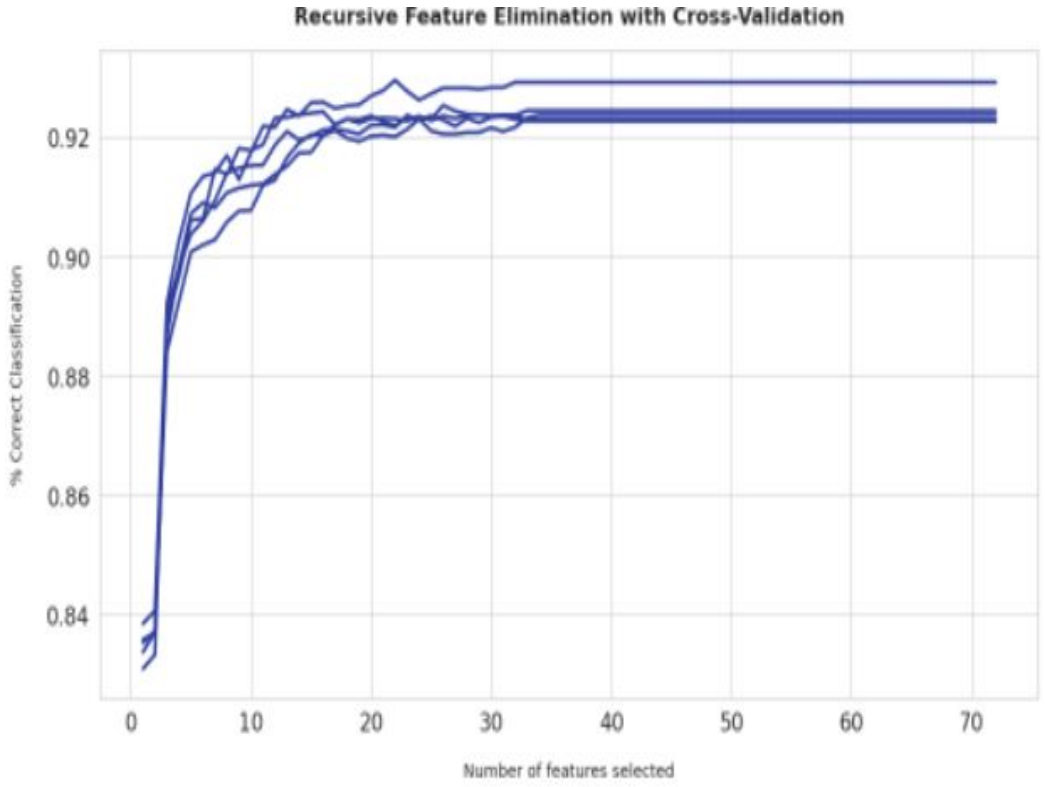
METHODOLOGIES - KNN IMPUTED DATA

- Data Analysis on Categorical Features
- Correlation between dependent Feature and Independent feature
- Feature Selection using Recursive feature elimination with cross-validation
- Model Training by leveraging Hyperparameter tuning

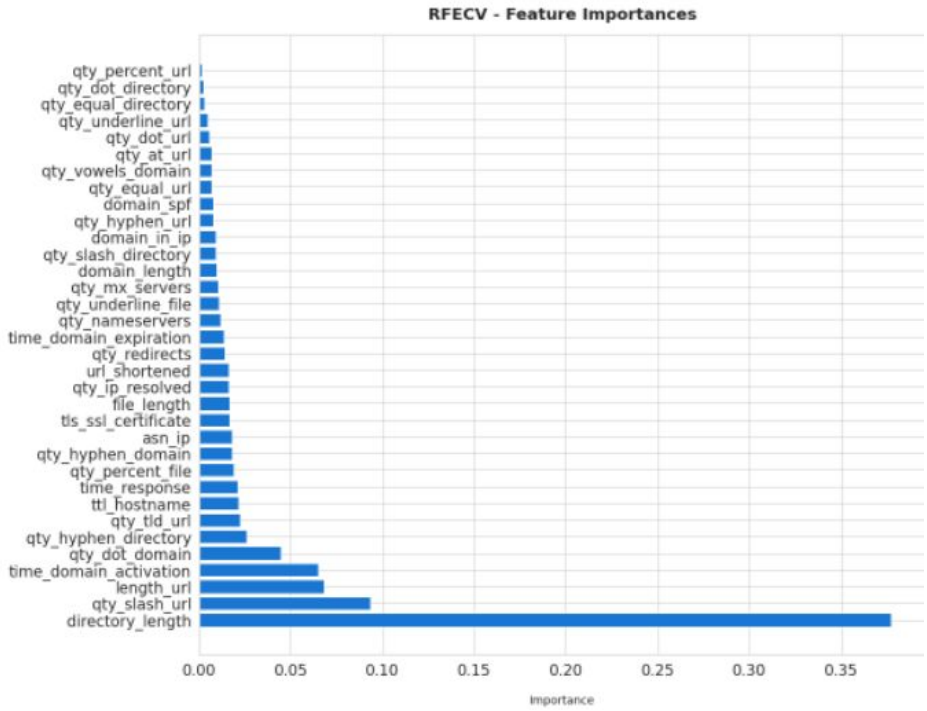
for Obtaining Optimal Results

KNN IMPUTED DATA

- **Accuracy vs Features selected**



- **Feature Importance using RFECV**



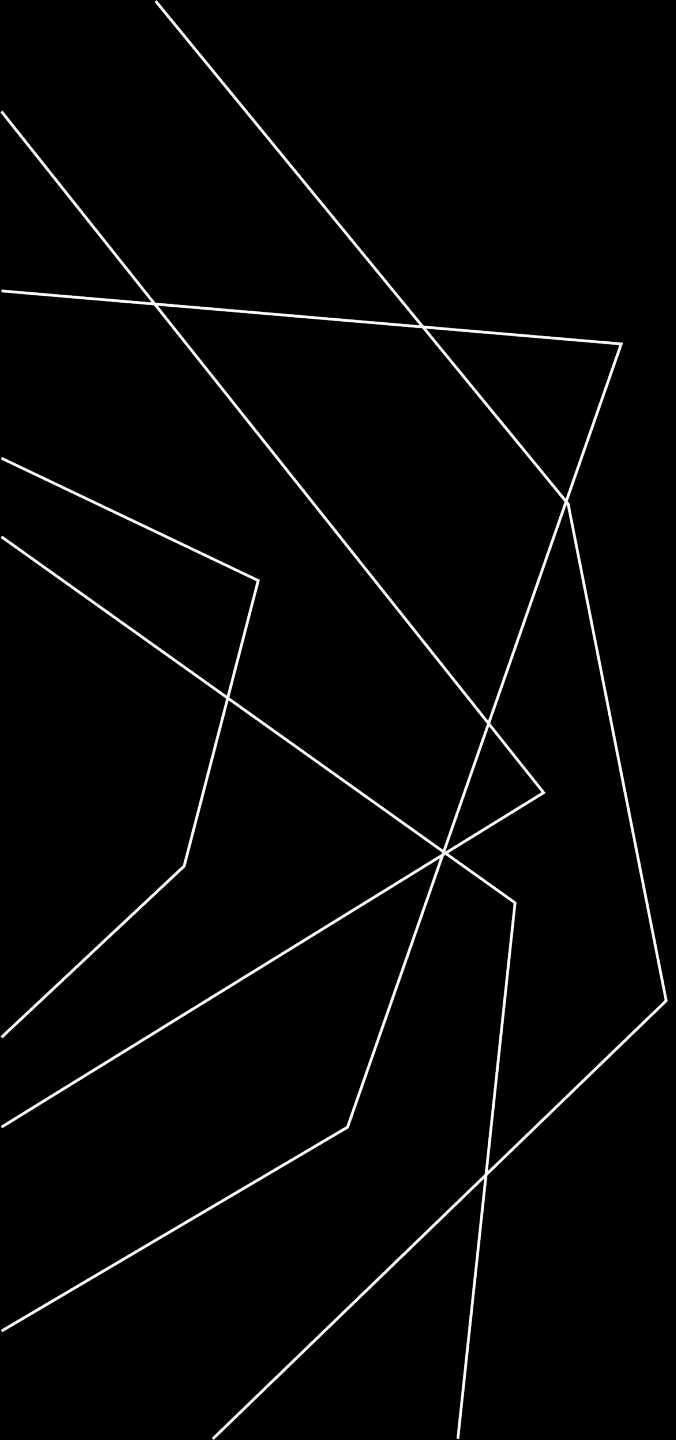
We can observe from the above plot as no of selected features increase our accuracy initially increased and then saturated. After approximately 30 selected features the accuracy is flattened.

RESULTS OBTAINED FOR KNN IMPUTED DATA

Classifier	Accuracy
Logistic Regression using hyperparameter tuning	89.07%
XGBoost hyperparameter tuning	92.77%
RandomForest hyperparameter tuning	95.63%

CONCLUSION

- As Mean Imputation doesn't incorporate relation among the features, despite dropping the outliers
  - **Logistic regression classifier** was **overfitting** with accuracy of **99%**
  - And, **KNN Classifier** gave accuracy of **95.34%**
- To overcome the issue encountered with Mean Imputation, we performed KNN imputation
  - **Logistic regression** gave accuracy of **89.07%(overcame overfitting)**
  - **Random Forest Classifier** gave accuracy of **94.63%**
  - **XGB Classifier** gave accuracy of **92.77%**
- Therefore, we achieved **best accuracy** using **KNN imputation.**
- It can be observed that the important features for both **KNN imputed analysis** and **Mean imputed analysis** is based on the **attributes relating to URL and External services**.



THANK YOU