Phishing Attack domain detection

Objective:

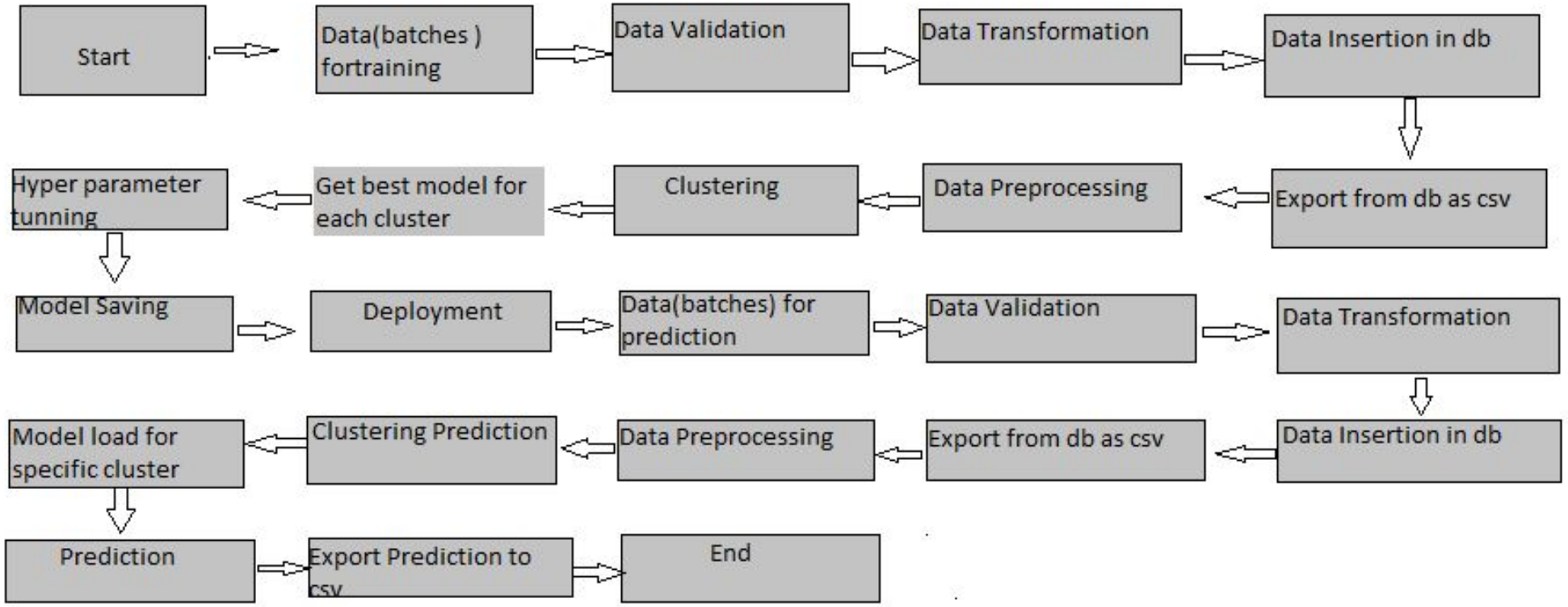
Phishing is a type of fraud in which an attacker impersonates a reputable company or person in order to get sensitive information such as login credentials or account information via email or other communication channels.Phishing is popular among attackers because it is easier to persuade someone to click a malicious link that appears to be authentic than it is to break through a computer's protection measures.

The main goal of this project is to create a domain authentication system that would detect if a given domain url is legit or fake website created to perform fraud. Multiple ML models will be tested for this problem. A web Interface along with suitable Rest-API's will be created for commercial use.

Benefits:

- Detection fraudulent scams
- Prevent cyber attacks
- Avoid user’s from malicious sites

Architecture



Model Training:

- Data Export from Db :

  The accumulated data from db is exported in csv format for model training

- Data Preprocessing
  - Performing EDA to get insight of data like identifying distribution , outliers ,trend among data etc.
  - Check for null values in the columns. If present impute the null values.
  - Encode the categorical values with numeric values.
  - Perform Standard Scalar to scale down the values.

- Model Selection –

  After the data has been processed,we train multiple models using that data and choose the best one for production. Some of the models we used include Decision tree,Random Forest,Neural Network.

Q & A:

Q1) What's the source of data?

The data has been collected from various open-source datasets on Internet.

Q 2) What was the type of data?

The data was the combination of numerical and Categorical values.

Q 3) What's the complete flow you followed in this Project?

Refer slide 5th for better Understanding

Q 4) After the File validation what you do with incompatible file or files which didn't pass the validation?

Files like these are moved to the Achieve Folder and a list of these files has been shared with the client and we removed the bad data folder.

Q 5) How logs are managed?

We are using different logs as per the steps that we follow in validation and modeling like File validation log , Data Insertion ,Model Training log , prediction log etc.

Q 6) What techniques were you using for data pre-processing?

- Removing unwanted attributes
- Visualizing relation of independent variables with each other and output variables
- Checking and changing Distribution of continuous values
- Removing outliers
- Cleaning data and imputing if null values are present.
- Converting categorical data into numeric values.
- Scaling the data

Q 7) How training was done or what models were used?

- Before diving into training,the data was preprocessed and cleaned
- After that the data was split into train and test sets
- The data was scaled or normalized before giving it as input to the model
- The models used were Decision Tree,Random Forest,Neural Network

Q 8) How Prediction was done?

The given input URL is first preprocessed to extract required features,then these numerical features are given to the model which then makes a prediction.It returns the probability value of URL being malicious or legitimate.

- Q 9) What are the different stages of deployment?
  - When the model is ready we deploy it in Heroku.
  - The model has been deployed as both REST API and a web interface is also been created for it.