Phishing Domain Detection

Objective:

Development of a predictive model for detecting website is phishing or not. The model will determine whether a website link is malicious or real.

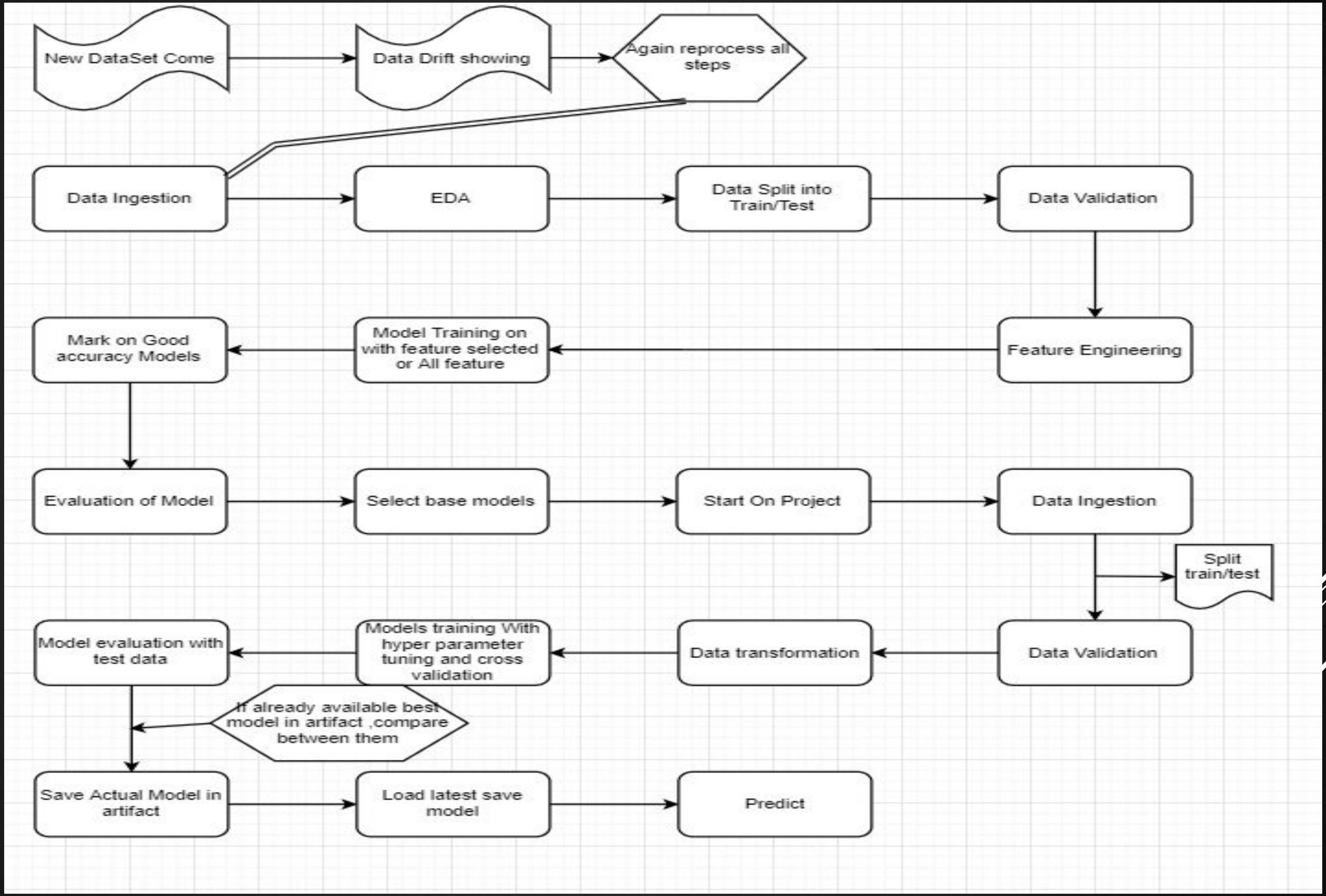
Benefits:

- Detection between phishing or legitimate site.
- User will be aware when site is riskier.
- This model also use as api in multiplatform after some modification like make widget to identify site whenever user click on links in social media messaging apps eg. in whatsapp.

Data Sharing Agreement :

- Sample file name (ex dataset_full.csv)
- Number of Columns
- Column names
- Column data type

Architecture


New DataSet Come
Data Drift showing
Again reprocess all steps
Data Ingestion
EDA
Data Split into Train/Test
Data Validation
Feature Engineering
Model Training on with feature selected or All feature
Mark on Good accuracy Models
Evaluation of Model
Select base models
Start On Project
Data Ingestion
Split train/test
Data Validation
Data transformation
Models training With hyper parameter tuning and cross validation
Model evaluation with test data
If already available best model in artifact ,compare between them
Save Actual Model in artifact
Load latest save model
Predict

Data Validation and Data Transformation :

- Name of Columns - The name of the columns is validated and should be the same as given in the schema file. If not, then the file is not accepted for re-training.
- Data type of columns - The data type of columns is given in the schema file. It is validated when we insert the files into Database. If the datatype is wrong, then the file is not accepted.
- Null values in columns - If any of the columns in a file have all the values as NULL or missing, then we use simple imputer and replace by -1 according to experiment in feature support.

Model Training:

- Data Export from Csv:

  The accumulated data from 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.
  - Check collinearity of features and remove high collinear features.
  - Perform Power transformer (yeo-johnson) to scale down the values.

- Model Selection –

  After the feature engineering on data , we find the best model for dataset.

  For getting base models we experiment all classification algorithms and find the base best models are XG Boost classifier and Decision tree classifier and Random Forest Classifier. After apply cross validation and hypertunning we will find that best actual model is Random Forest Classifier.

Prediction:

- After submit the url for phishing detection prediction.
- I extract all the features of url according the training dataset like domain property, Whole url property like length , count of something of url, get parameters, get directory, get file name and its properties, some extra features like google index number of url.
- Some property which are not found we replace with -1 as in dataset and good experiment accuracy.
- After the get all features we are doing data preprocessing and then after predict the result and show the message.
- Batch prediction of urls is also supported with csv,excel etc. Function is provided in app class. By default only 200 urls indexing available. As a developer you will aso modify the function. Its takes time to extract all the urls properties and predict the result.

Q & A:

Q1) What’s the source of data?

The data for training is provided by the research paper in multiple datasets.h contain multiple files

Q 2) What was the type of data?

The data was the numerical.

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

Refer slide 4th 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 Artifact Folder and a list of these files has been shared with the client and we not accept the bad file for data training.

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?

- Checking and changing Distribution of continuous and discrete values
- Removing high collinear features.
- Cleaning data and imputing if null values are present.
- Scaling the data

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

- As per dataset full which are imbalance we are using, convert into balance dataset using imbalance library as SMOTE after that the training and validation data were divided.
- The scaling was performed over training and validation data
- Algorithms like Random boost classifier was used based on the accuracy of final model.
- and we saved that model .

Q 8) How Prediction was done?

We use another dataset from another source (Kaggle) for testing purpose to check real world accuracy. In accuracy we will give focus to both precision and recall.

And do some testing on phis tank database of urls which are manually.

After that we finally select that model for phishing domain detection.

- Q 9) What are the different stages of deployment?
  - When the model is ready we deploy via github actions as per requirement into aws vs heroku.
  - We also divide and modify according to functionality requirements onto cloud. Like minimum resource or good resource instances. Eg. Live training , live update configuration.