Cyber-security: Phishing Domain Detection

Objective:

Development of a predictive model for identifying Phishing URL

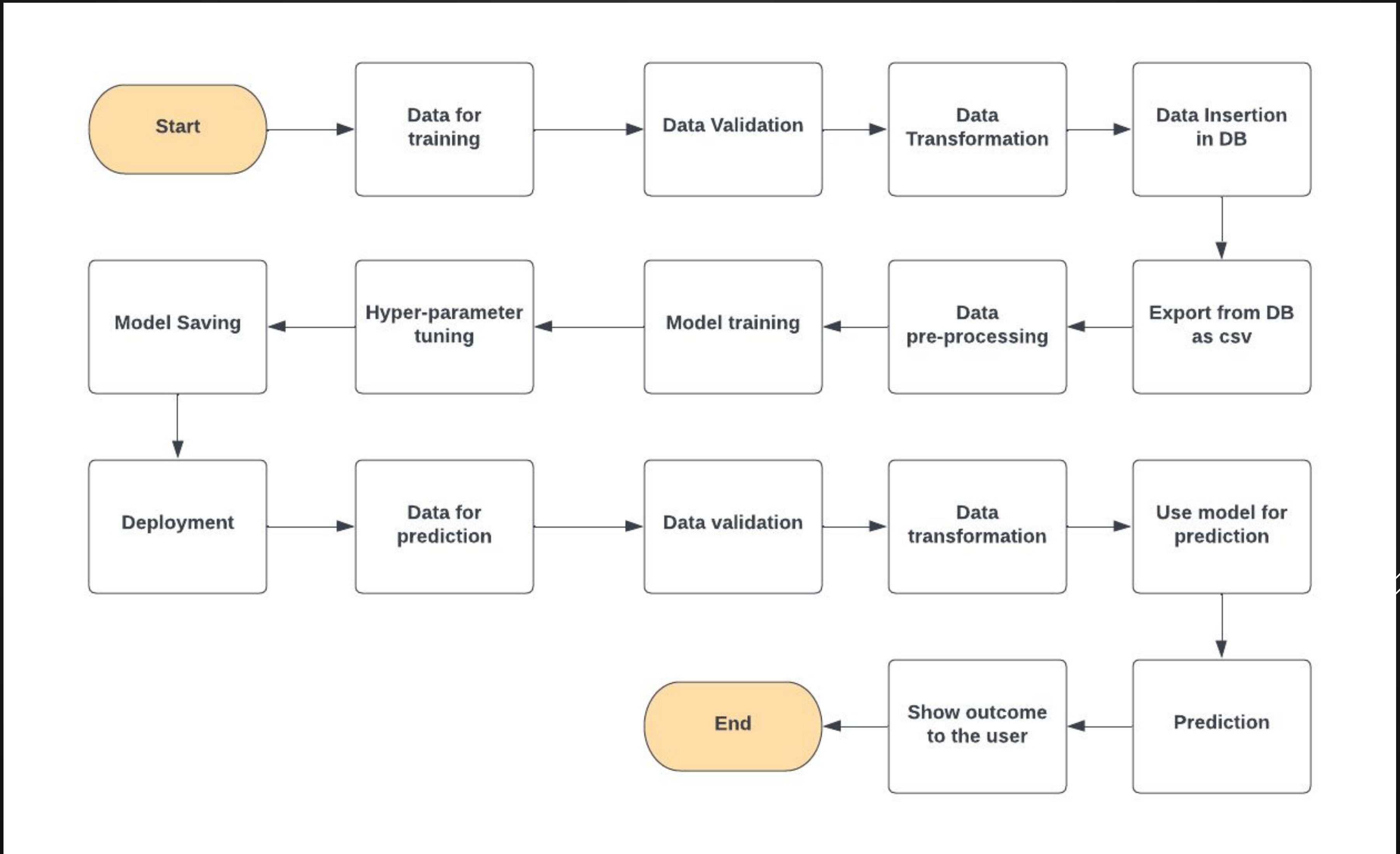
Benefits:

- Identifies harmful malicious URL
- Safeguard user data from being leaked
- Prevents the user system from getting hacked
- Prevents other cyber crimes related to Phishing

Data Sharing Agreement :

- Data file name (ex dataset_file.csv)
- Minimum length of URL: 11 characters
- Minimum mandatory attributes: protocol, domain
- Number of Columns
- Column names
- Column data type

Architecture



Data Validation:

- Takes data ingestion artifacts as input
- Validates if the data generated in the data ingestion phase is as per the findings in the EDA phase
- This is done by using the handling null values and checking for required columns
- We also check for data drift to ensure predictions in the future could be handled by the same model
- Generates a report for the same as an artifact in the artifact/data_validation

Data Transformation:

- We created the preprocessing pipeline
- This pipeline has Simple Imputer and RobustScaler
- This component takes train data from the Data Ingestion artifact and creates a trained pre-processing pipeline
- Using this we generated transformed the train and test data into test.npz and train.npz in artifact/data_transformation/transformed
- Our target feature was not numerical. We've used the LabelEncoder to encode the target feature. This is stored as an artifact in artifact/data_transformation/target_encoded

Data Insertion in Database:

- Table creation :- Table name “phishing_domain" is created in the database for inserting the files. If the table is already present then new files are inserted in the same table.
- Insertion of files in the table - All the files in the "cybersecurity" are inserted in the above-created table. If any file has invalid data type in any of the columns, the file is not loaded in the table

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.

- Takes transformed train_arr and test_arr as config
- Used RandomForest Classifier as the model and trained it
- Created a Model.pkl file as an artifact and saved in artifact/model_trainer/model

Prediction:

- The testing files are shared in the batches and we perform the same Validation operations ,data transformation and data insertion on them.
- The accumulated data from db is exported in csv format for prediction
- We perform data pre-processing techniques on it.
- Random Forest model created during training is loaded
- Once the prediction is done for all the data. The predictions are saved in csv format and shared.

Q & A:

Q1) What's the source of data?

The data for training is provided by the client in multiple batches and each batch contain multiple files

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?

- Extracting URL features from the URL using RegEx and inserting into new columns
- 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?

- ➢ The scaling was performed over training and validation data
- ➢ Random Forest algorithm was used and we saved that model

Q 8) How Prediction was done?

- ➢ The training data is from a research website
- ➢ Features are extracted from the URL in this file
- ➢ We perform the lifecycle of the until model training, then after model evaluation, predictions are made and the output is shown to the user

Q 9) What are the different stages of deployment?

- When the model is ready we deploy it in AWS EC2.

End