Phishing Domain Detection

Vahe Charchyan

01.04.2022

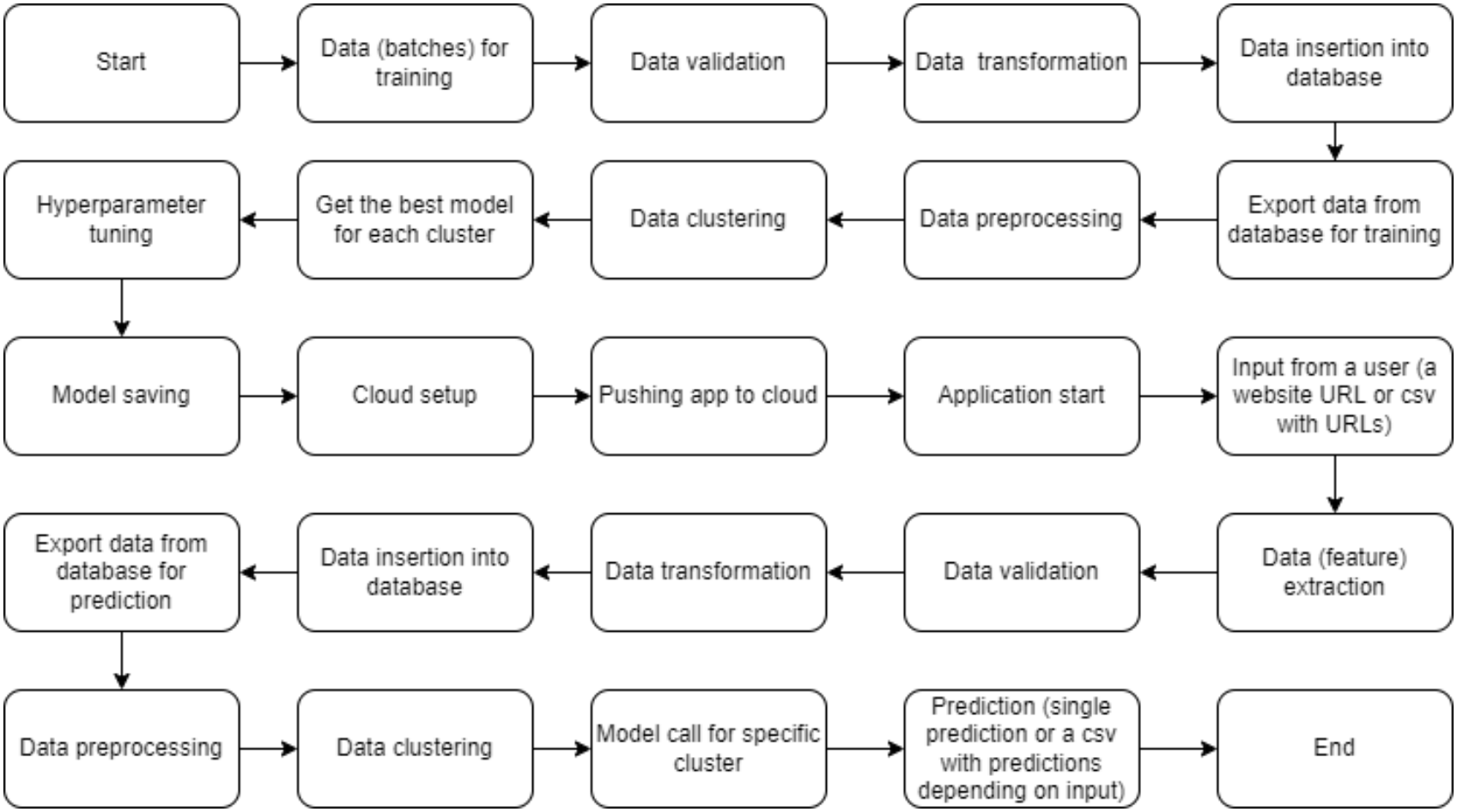
Objective

Developing a phishing detection solution to indentify malicious websites

Benefits

- Detection of phishing websites
- Risk reduction and safe surfing over the web

Architecture



Data ingestion

- Data validation
- Data insertion into database

Export data from the database

The data is exported from the database using CQL. Then it is converted to pandas dataframe.

Data preprocessing

- Replace the invalid values with numpy nan to able to use imputer on such values
- Check for null values in the columns and impute them using the KNN imputer
- Replace -1 values with -999 for continuous/numeric features
- Use one-hot encoding for categorical features

Data clustering

- KMeans clustering

Model building

Model building includes finding the best model for each cluster. We use Support Vector Machine, XGBoost, Naive Bayes, Logistic Regression, Random Forest, and LightGBM algorithms. For each cluster, the algorithms use the best parameters derived from GridSearch/BayesSearch. We calculate the AUC scores for the models and select the model with the best score for each cluster.

Deployment

The model is deployed to Heroku. CircleCi is used for CI-CD.

Data from a user

At this stage a user inserts a website URL or a csv/xlsx file with URLs.

Model call

The estimated model is called to identify malicious websites.

Q & A

- What's the source of data ?
  - The data is hosted here.
- What is the type of data ?
  - The data contains both numeric and categorical values.
- What's the complete flow you followed in the project ?
  - Please, look through 4th slide for better understanding.
- How logs are managed ?
  - The logging is applied in every process, but logging results will not be accessible for end users. Moreover logging is made accessible via console and a file for developers.

Q & A

- What techniques are used for data pre-processing ?
  - KNN imputer to replace missing values
  - One-hot encoding for categorical values
  - Some values' replacement

Q & A

- How training was done or what models were used ?
  - Data clustering
  - Finding the best model out of Support Vector Machine, XGBoost, Naive Bayes, Logistic Regression, Random Forest, and LightGBM for each cluster
  - Hyperparameter tuning for each model
- How prediction was done ?
  - A user provides inputs and clicks on check button. The rest happens automatically.
- What are the different stages of deployment ?
  - The deployment was done using Heroku. We pushed code to github and connected Heroku to the github. Then we used CircleCi for CI-CD.