Phishing Detection Using Machine Learning and Deep Learning Techniques.

Revision Number: 2.0
Last Date Of Revision 28/09/2022

TABLE OF CONTENTS

Overview

Fraudsters send fake emails or set up fake websites that mimic Yahoo!'s sign-in pages (or the sign-in pages of other trusted companies, such as eBay or PayPal) to trick you into disclosing your user name and password. This practice is sometimes referred to as "phishing" — a play on the word "fishing" — because the fraudster is fishing for your private account information.

Effect of phishing

- Loss of data.
- Damage reputation.
- Direct monetary loss
- Loss of productivity.
- FInancial penalty.
- Intellectual theft

Definitions

Term	Definition	Purpose
GCP	Google Cloud Platform.	Will deploy our machine learning model on this platform.
CloudRun	Google Cloud Serverless Container Engine.	Cloudrun will serve as our serverless container engine to host our model.
Container Registry	Will server as a place for storing model image	
Cloud Storage Bucket	Google cloud storage bucket.	Will store our model and retrieve it during prediction using python SDK.
Service account		This service account will allow cloudrun access to cloud storage bucket to

		access our model.

Problem Definition

From the above definitions and disadvantages of phishing, We are going to build a machine learning model to predict if a URL contains malicious content or not.

Evaluation

If we are able to get an accuracy score between 70 and 95 and a good precision, recall and f1 score, then our model is ready to go into production.

Revolutional Neural Networks

- LSTM
- Embedding
- GRU
- Bidirectional
- Conv1D
- Transfer Learning techniques

Evaluation Metrics

- LSTM
- Embedding
- GRU
- Bidirectional
- Conv1D
- Transfer Learning techniques

Steps used to build model

1. The model was built using different steps. The first step was to collect the datasets. The dataset was a secondary datasets, which was collected from kaggle.

Deployment

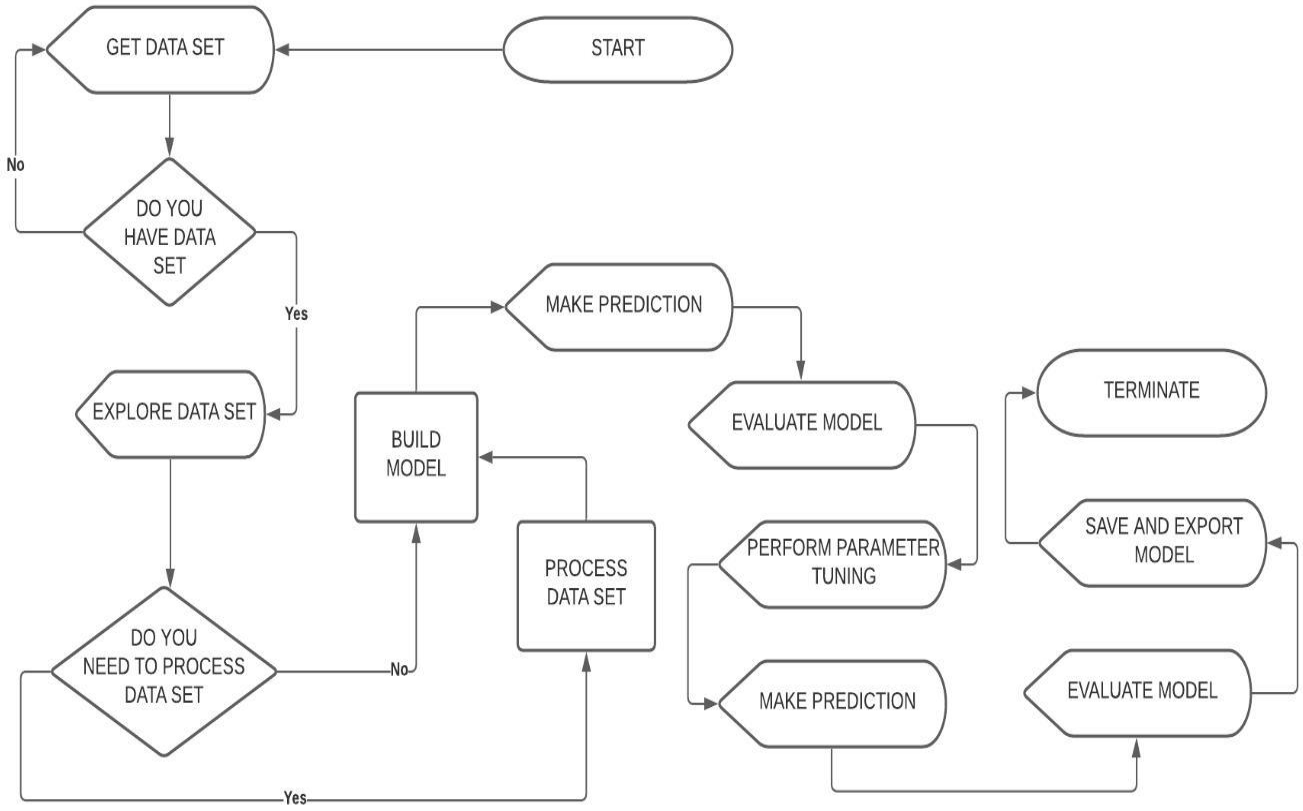
2. The second step was to explore the datasets. Here, activities such as checking null values, categorical and numeric features, outliers, highly correlated features, visualizing and many others.

3. The third step was to balance our dataset. Our dataset was not balanced. One class contained more values than the other hence we had to balance the datasets in order to have an equal number of values in each class.

4. The fourth step was to encode our label feature from category to numeric. After that we built our base model, made predictions and evaluated our model using a machine learning algorithm (Logistic Regression).

5. Eight different models were built until we chose the best model that had a good accuracy score, precision score, recall and f1 score.

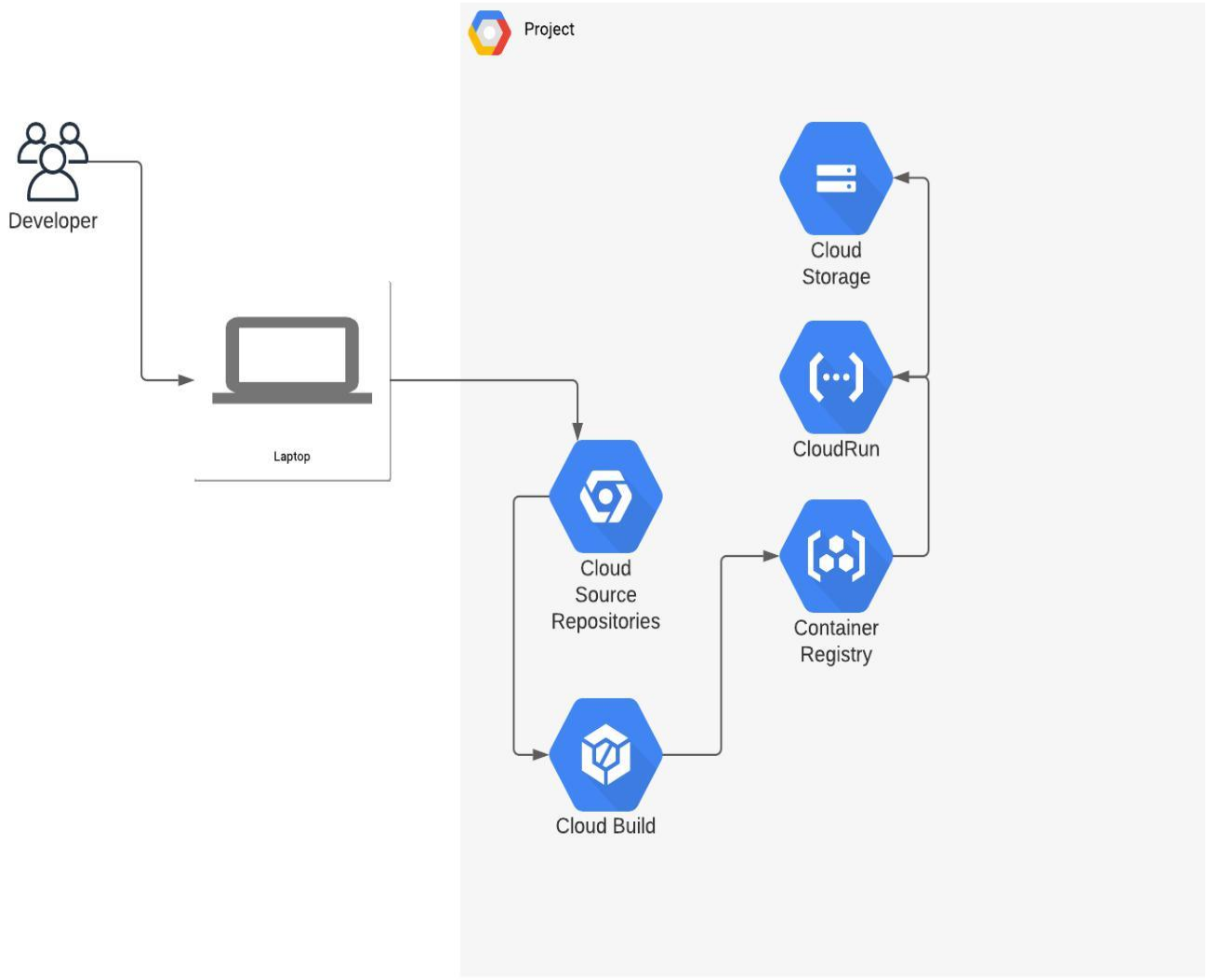
The diagram below Illustrate how the model was built



Deployment

The model was successfully deployed on Google Cloud Platform. The model was containerized and pushed to google container registry and deployed to google Serverless Container Services(CloudRun) using a script to automate the entire process.

The diagram below illustrates how the model was deployed on Google Cloud Platform.



Conclusion

The model can be deployed and be consumed by web applications, mobile applications and many other applications. Also the model can be retrained and redeployed using MLOPs technique