MARKET BASKET PROJECT ON E-COMMERCE

DETAILED PROJECT REPORT

Irfan Razi

Data Science Intern @ iNeuron.ai

INTRODUCTION

With population size of 138 crores, India is the seventh-largest country by area, the second-most populous country, and the most populous democracy in the world. So, we have a resource which no other country is able to match, which is human resource. With the development of the technology, ease of online services, UPI transactions, India is way ahead than any other country. Competition among the vendors are high. They are working really hard to make the service as good as possible. In such cases, feedbacks and reviews play a major role.

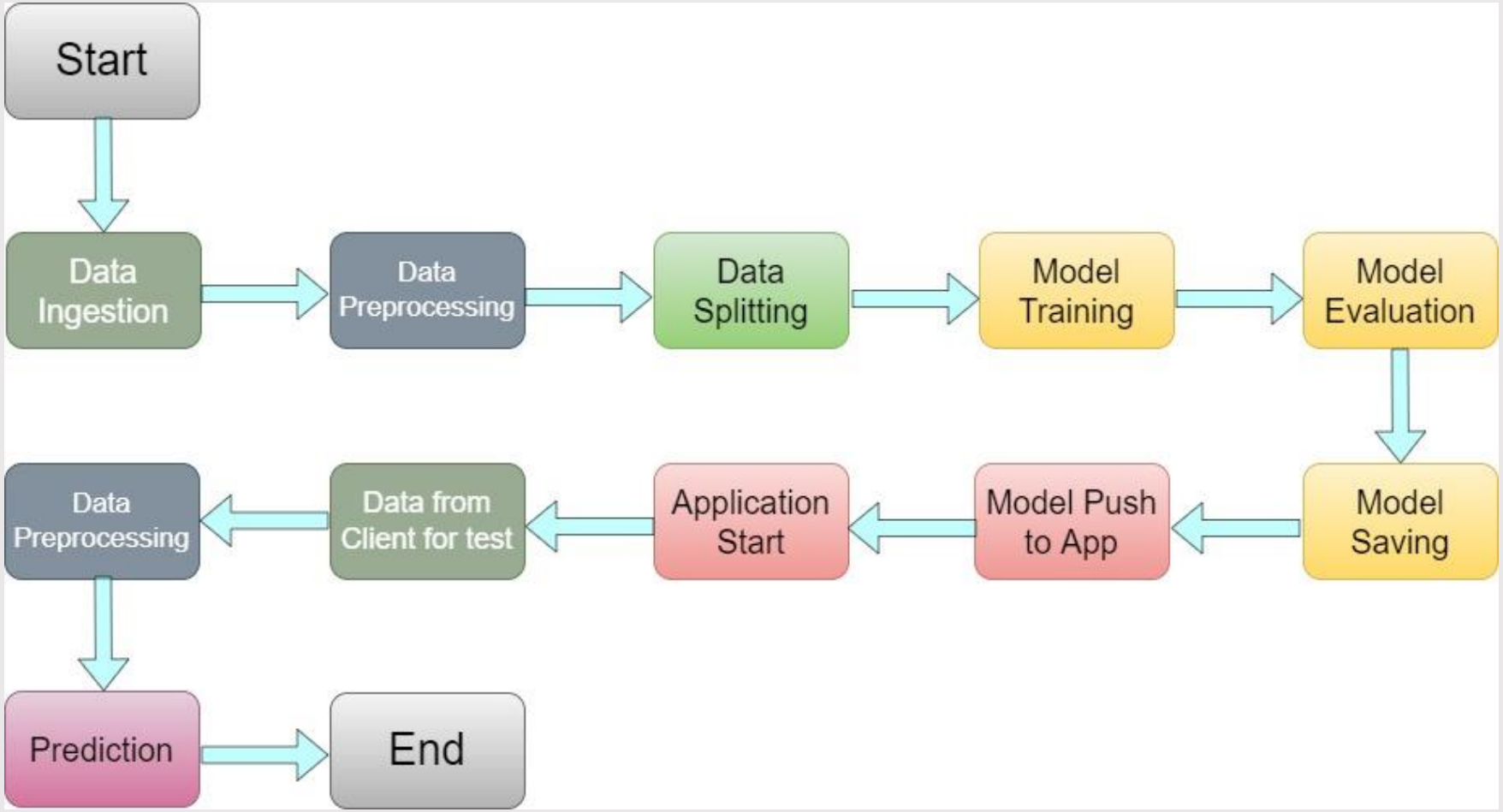
Most customers do not post a review rating or any comment after purchasing a product which is a challenge for any E-commerce platform to perform If a company predicts whether a customer liked/disliked a product so that they can recommend more similar and related products as well as they can decide whether or not a product should be sold at their end. This is crucial for E commerce-based company because they need to keep track of each product of each seller, so that none of products discourage their customers to come shop with them again. Moreover, if a specific product has very few ratings and that too negative, a company must not drop the product straight away, may be many customers who found the product to be useful haven't actually rated it. Some reasons could possibly be comparing your product review with those of your competitors beforehand, gaining lots of insight about the product and saving a lot of manual data pre-processing, maintain good customer relationship with company, lend gifts, offers and deals if the company feels the customer is going to break the relation.

OBJECTIVE

The main goal of this case study is centred around predicting customer satisfaction with a product which can be deduced after predicting the product rating a user would rate after he makes a purchase.

The classical machine learning tasks like Exploratory Data Analysis, visualization, data cleaning, Feature Engineering, Feature Selection, model building, hyper-parameter tuning, model evaluation and testing.

ARCHITECTURE



DATASET

1. order_item_id: This attribute is the quantity of the product purchased. It is int type.

2. product_weight_g: This attribute is the weight of the item in grams. It is int type.

3. payment_installments: This attribute is the number of installments in which customer paid for the product. This is int type.

4. order_delivered_customer_time_in_days: This attribute is created from other attirbutes which is the timestamp for purchase order placed and delivery date. This is int type.

5. product_volume: This attribute is created from other attributes which are length, width and depth of the product. This is int type.

6. customer_seller_distance: This attribute has been created from the latitudes and longitudes of the customer and seller. This is int type.

7. order_purchase_year: We have data from year 2016 to 2018. This is int type.

DATASET

8. total_payment: This is combined attribute of price and freight value for the product. This is int type.

9. order_status: The order might be delivered, in-processing, cancelled, etc. This attribute has been encoded to int type.

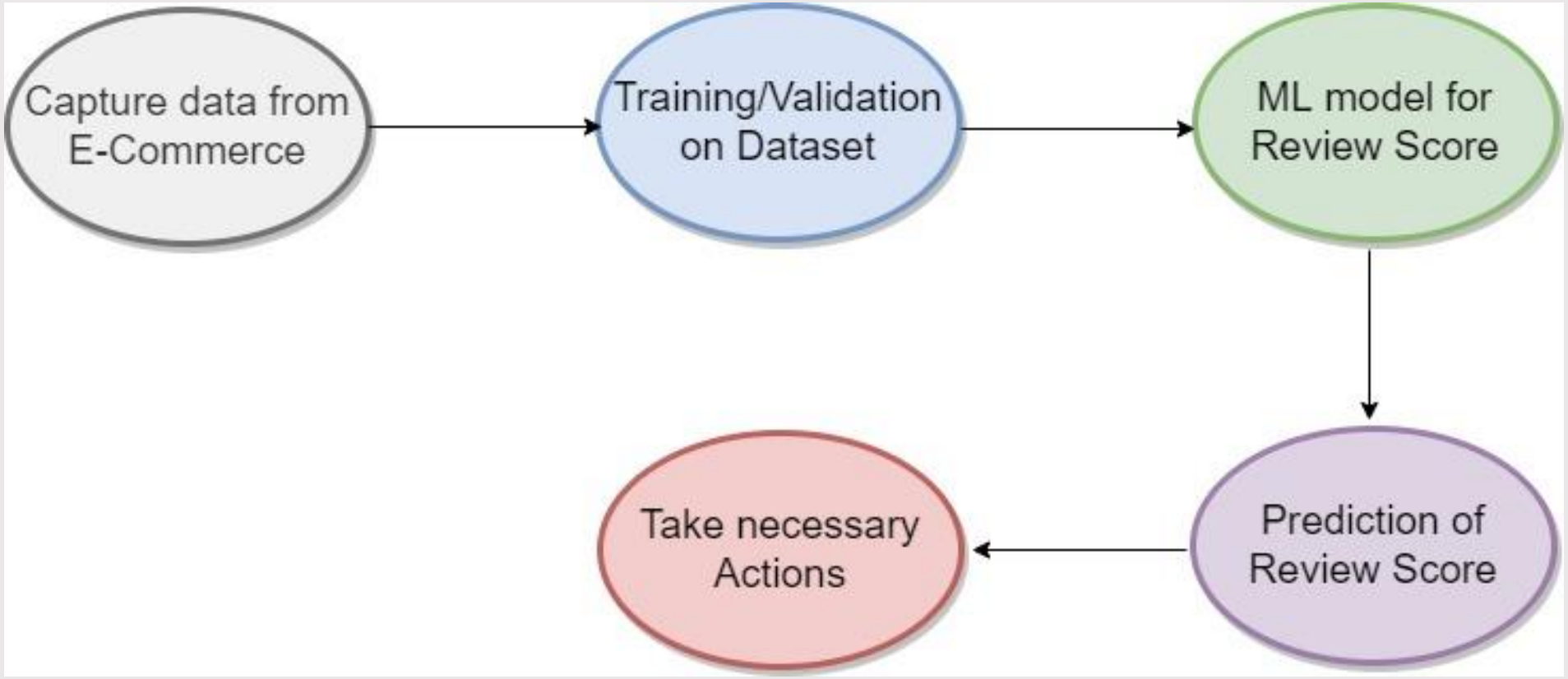
10. payment_type: Customer may make the payment via credit card, debit card, voucher, etc. This attribute has been encoded to int type.

11. product_category_name: There are wide range of product categories. This attribute has been encoded to int type.

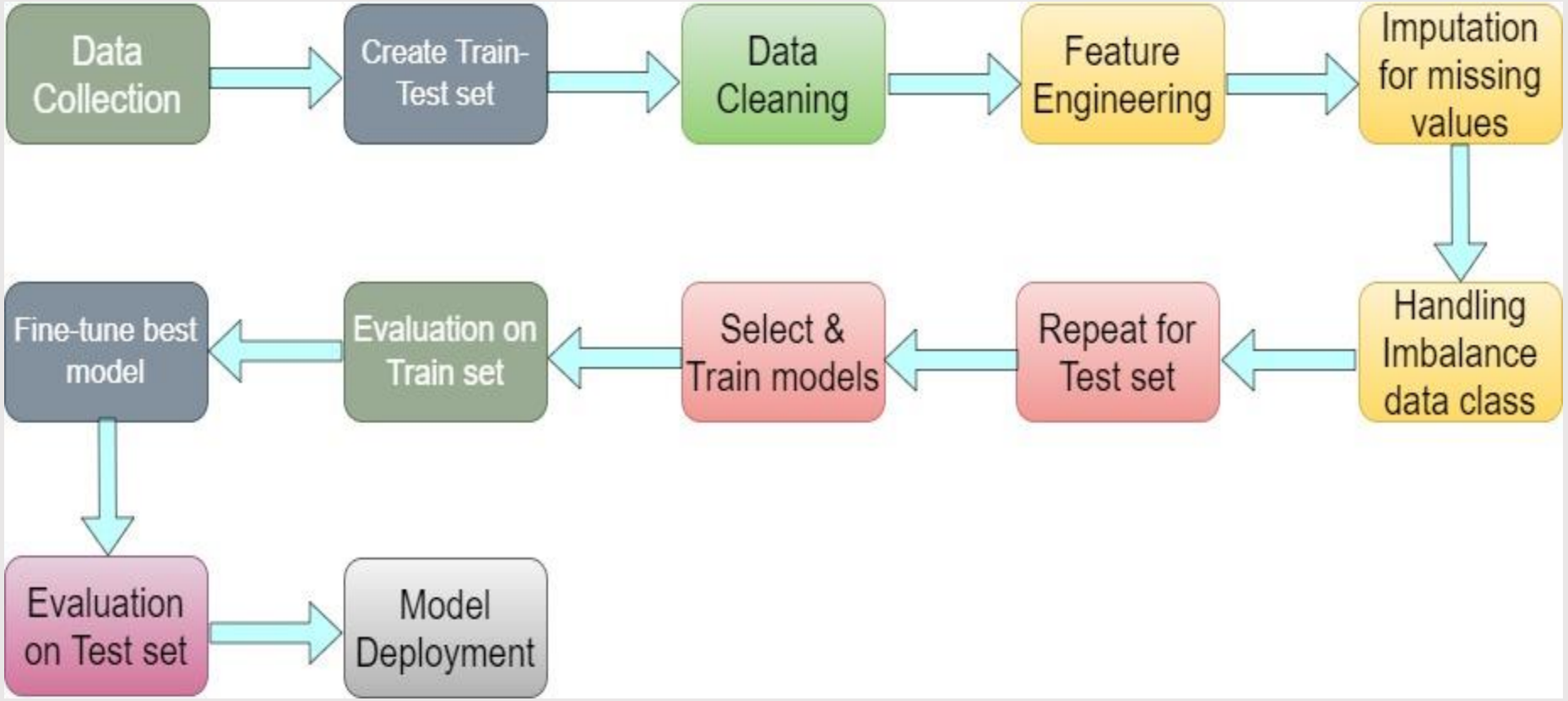
12. Timing: This attribute is to understand at what time customer has placed the order. It could be Morning, Evening, Afternoon and Midnight. This attribute has been encoded to int type.

13. Season: This attribute gives the idea during which season the purchase has been made. It could be summer, winter, etc. This attribute has been encoded to int type.

PROCESS FLOW



MODEL TRAINING AND EVALUATION WORKFLOW



MODEL TRAINING AND EVALUATION

Data Collection:

- Brazilian E-Commerce Public Dataset by Olist
- Link for Dataset: https://www.kaggle.com/datasets/olistbr/brazilian-ecommerce

Data Pre-processing:

- Drop insignificant columns.
- Feature selection using Mutual Gain Information technique.
- Handling Categorical features.
- Handling missing values by using KNN imputation.
- Handling imbalanced dataset using resample technique.

MODEL TRAINING AND EVALUATION

- Classification Algorithm: Random Forest Classifier has been used since this model give the best result. It has been chosen for model training and testing.
- Model performance has been evaluated based on Recall and F1 score using classification report.

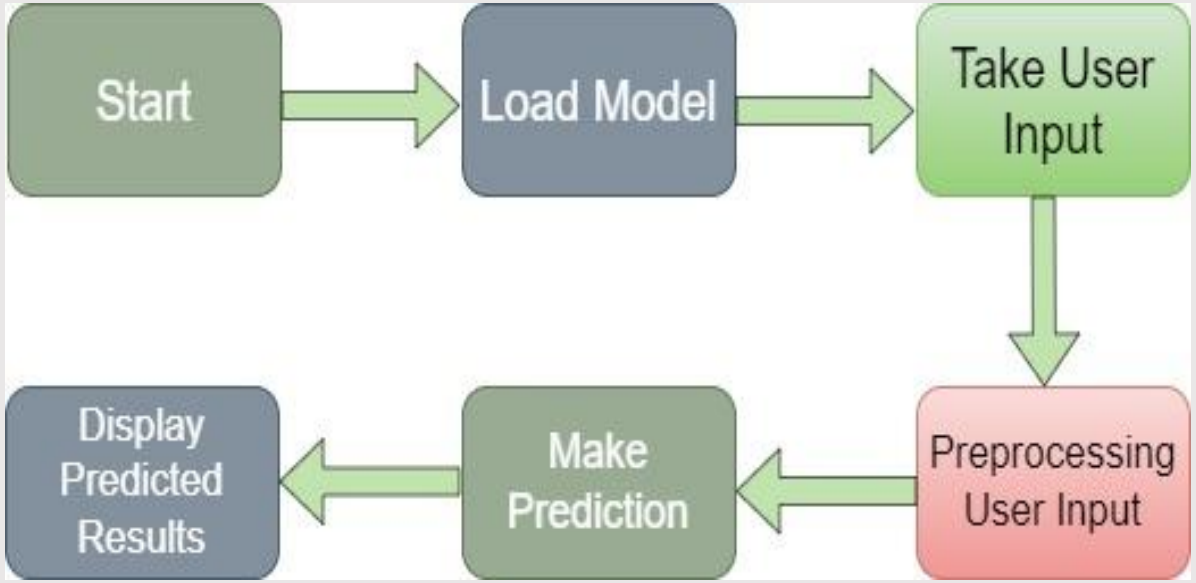
WORKFLOW

- **Data Description:** We will be using Brazilian E-Commerce Public Dataset by Olist. This is a Brazilian ecommerce public dataset of orders made at Olist Store. The dataset has information of 100k orders from 2016 to 2018 made at multiple marketplaces in Brazil. Its features allow viewing an order from multiple dimensions: from order status, price, payment and freight performance to customer location, product attributes and finally reviews written by customers. We also released a geolocation dataset that relates Brazilian zip codes to lat/lng coordinates.
- **Data Ingestion:** Here, we will be ingesting all the batches of data from Cassandra database to our machine in csv format.
- **Data Pre-processing:** We will do Exploratory Data Analysis of the data in Jupyter Notebook to get the complete understanding of the data. Based on that we can decide the strategy for Data Processing and validation. We may have to drop insignificant columns, handle missing values, handle imbalanced data, etc so that we can get a clean data for model training. For this, we have to write separate modules as per our need.
- **Data Splitting:** We split the data for model training and model validation.
- **Model Training:** We train our data with various ML models. Among those, Random Forest Classifier is the best fit model.

WORKFLOW

- **Model Evaluation:** Model evaluation is done by classification report. Since, this is a problem of imbalanced data, we have to analyse and improve Recall score and F1-score, not just Accuracy.
- **Model Saving:** After model training and evaluation, we will save the model for production.
- **Model Push to App:** We are going to do the cloud setup for our model deployment. We are going to create Flask App and User Interface. We will integrate our model with it.
- **Data from Client for Testing:** Now, our Web-Application is ready and deployed to clouds. We can get the data from our clients and start testing the model.
- **Data Pre-processing:** Client-data is also required to go through the same process as our train data has gone for model training.
- **Prediction:** Finally, when we complete the prediction process with client's data, we convert it into csv format and share it to the client.

MODEL DEPLOYMENT



- Final model has been deployed on Heroku using Flask framework.
- Docker Hub is used for Dockerisation.
- Circleci is used to build CI-CD pipeline.

HEROKU Flask circleci

FREQUENTLY ASKED QUESTIONS

Q1. What is the source of data?

Ans: The data for training is obtained from Brazilian E-Commerce Public Dataset by Olist.

Link for Dataset: https://www.kaggle.com/datasets/olistbr/brazilian-ecommerce

Q2. What was the type of data?

Ans: The data was the combination of numerical and categorical values.

Q3. What is the complete flow followed in this project?

Ans: Refer Slide No 8, 9 and 10.

Q4. After the File validation what you do with incompatible file or files which didn't pass the validation?

Ans: Such files are moved to the archive data folder. A list of these files are shared to the client. Then, we remove the bad data.

Q5. How logs are managed?

Ans: We are using different logs as per the steps that we follow in training and prediction like model training log and prediction log etc. And then sub log are inside those folder.

FREQUENTLY ASKED QUESTIONS

Q6. What techniques were you using for data pre-processing?

Ans: Following techniques were used:

- Dropping insignificant columns.
- Feature selection using Mutual Gain Information technique.
- Correlation visualization of features with each other and the target variable.
- Handling Categorical features by Count-Frequency Encoding technique.
- Handling missing values by using KNN imputation.
- Handling imbalanced dataset using resample technique.

Q7. How training was done or what models were used?

Ans: First, Data validation was done on raw data and then good data insertion was done in the Cassandra Database. Then, Data pre-processing was done on the final CSV file received from the Database.

- Various model such as Decision Tree, Random Forest, AdaBoost, Gradient Boost and XGBoost models were trained on all the data and based on performance, model is saved.

FREQUENTLY ASKED QUESTIONS

Q8. How prediction was done?

Ans: The testing files are shared by the client. We Perform the same life cycle till the data is clean .

Then model is loaded and prediction is preformed. In the end, we get the accumulated data of predictions.

Q9. What are the different stages of deployment?

Ans: After model training and finalizing the model, we manifest required files for deployment.

CI-CD pipeline has been created by the Circleci using github.

When the code is pushed to the Github, Circleci starts the deployment process. It orchestrates the dockerization of the complete model from Docker Hub and then push it to Heroku for the final deployment.

Q10. How is the User Interface present for this project?

Ans: For this project, I have designed one Interface for User testing and prediction. UI is user-friendly and easy to use.

THANK YOU