Zomato Rating Prediction

Objective :

The main goal of this project is to perform extensive Exploratory Data Analysis(EDA) on the Zomato Dataset and build an appropriate Machine Learning Model that will help various Zomato Restaurants to predict there

respective Ratings based on certain features.

Benefits

1. Using organization data into real world Business use-case .
2. Predicting Restaurant Rating and other general objective.
3. Optimum Services provided by Restaurants.
4. Could create a good availability of Restaurants and services provided by them.
5. Helps in Increasing profit to organization.

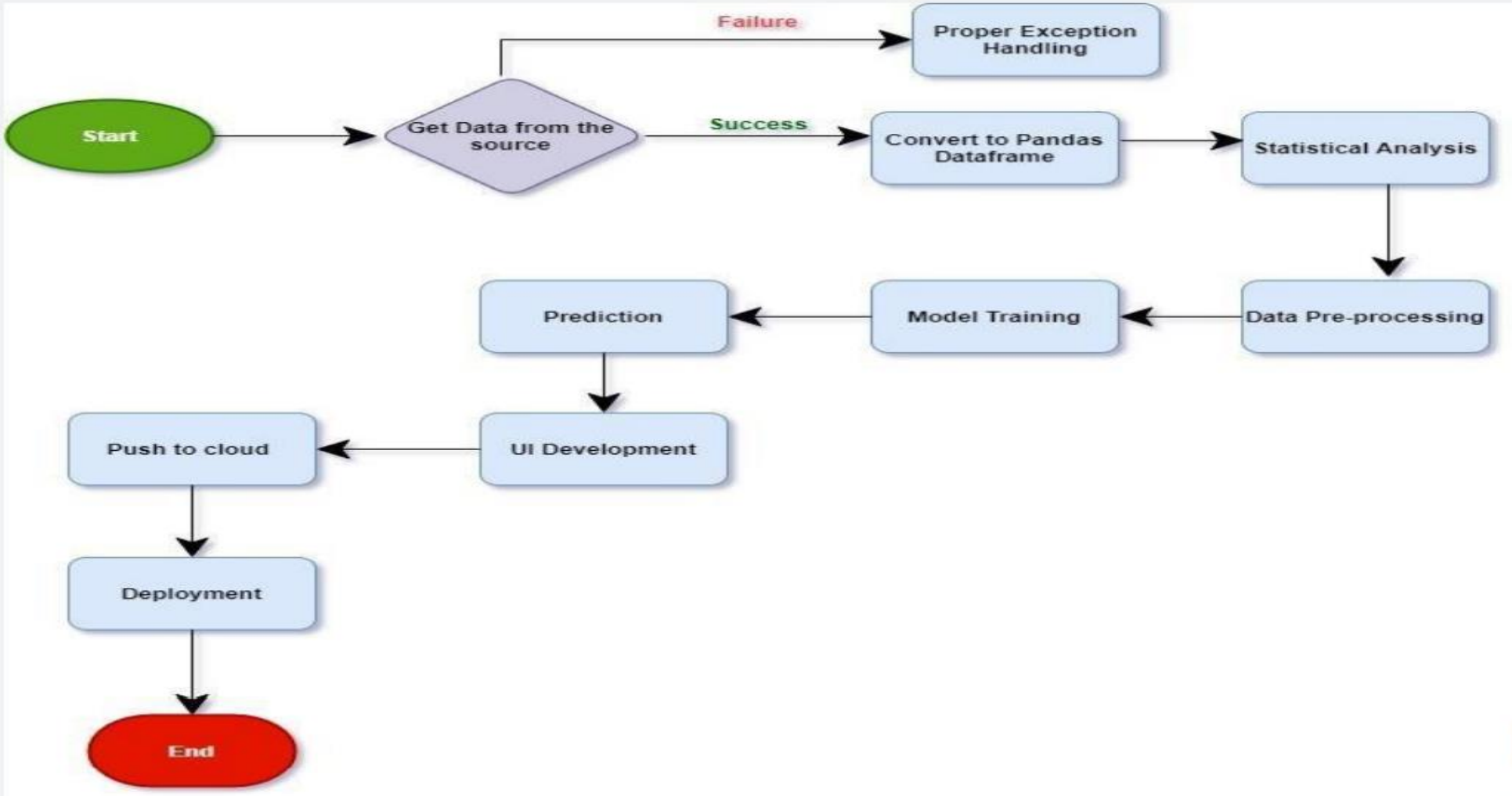
Data Sharing Agreement

- Sample file name (hours.csv) and source of the file is
- https://www.kaggle.com/datasets/himanshupoddar/zomato-bangalore-restaurants
- Shape of the data is 51717 x17
- 51717 Rows
- 17 Columns
- Colum datatypes where :-int64, object
- Where we have use only these 10 feature among 17

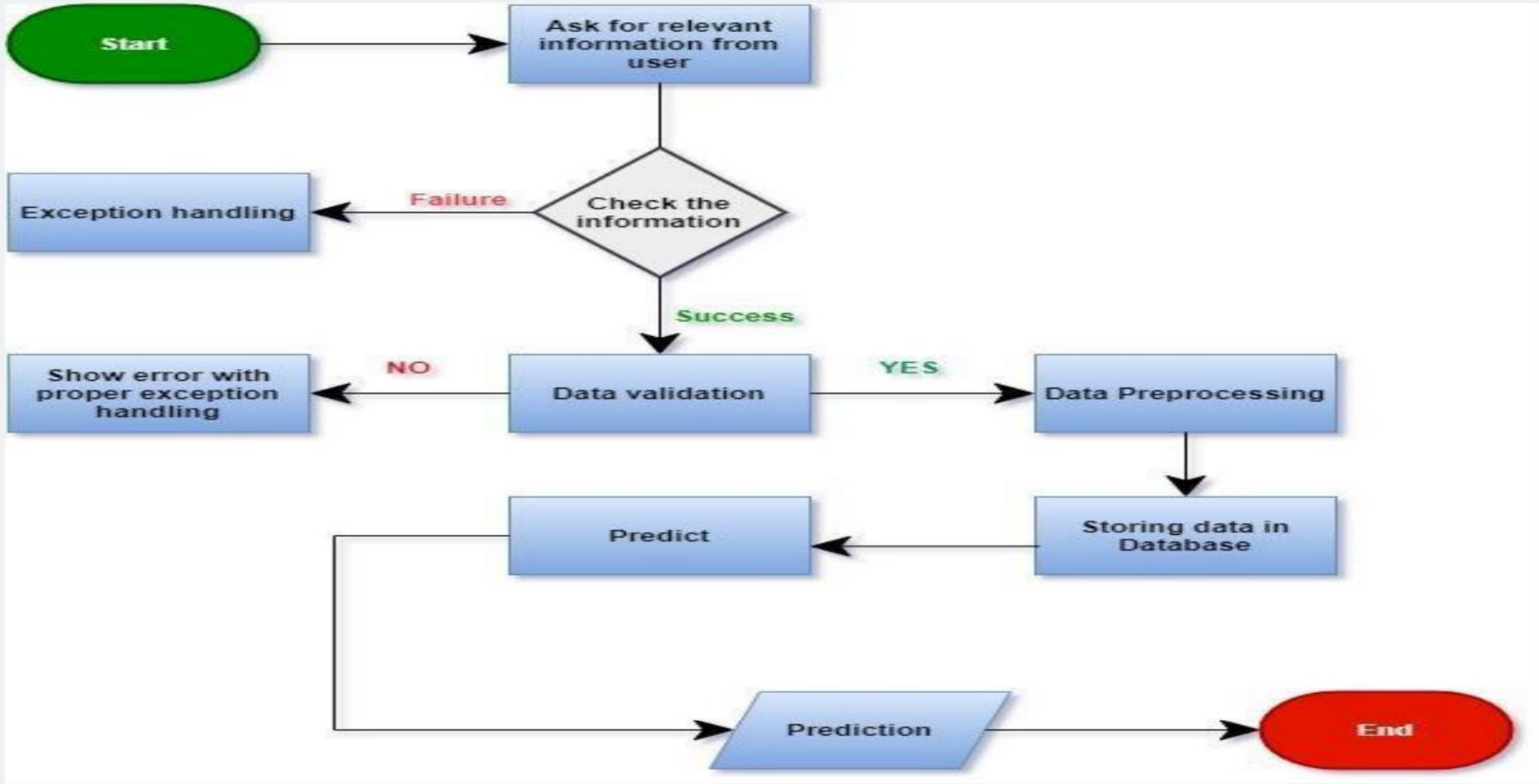
url	address	name	Online_order	Book_table	rate	votes	phone	location	Rest_type

Architecture

Machine Learning Model



Input Output Flow of Project



Model Training

- Data Export From CSV:

  Loading CSV data using python pandas and extracting all the data into dataframe in python file

- 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.
  - Perform Feature Selection and extract all the necessary features from the data

- Feature Selection:

  In Feature Selection we have Selected the required feature from the dataset on Three main basis : -

  1. Based on co-relation of input variable with output variable
  2. Based on common input feature which user can select
  3 . And which input variable should not have same dependency on output Feature (avoid multicolinearity)

- Train and Test Split:

  - Train data is 70% of whole data which 16396 records
  - Test data is 30% of full record which is 7028
  - Data is randomly splited in train and test
  - There is only train and test data available there is no validation data

- Model Selection:

  As this is the regression problem use case we have used linear regression and followed by the other regression algorithms such as ensemble algorithm. Where linear regression was not giving accuracy more than 25% so we use Ensemble algorithm such as Extra Tree Regressor and Random Forest among both Extra Tree Regressor was giving better result approximate (91%) accuracy and least error comparison to Random Forest.

- Prediction:

  1. Loading CSV data using python pandas and extracting all the data into python file
  2. We are perform data pre-processing techniques on the data loaded.
  3. We have use Extra Tree regression algorithmfor creating model for prediction .
  4. Based on the Extra Tree algorithm respective model is loaded and is used to predict the outcome from the data
  5. Prediction of Model is done on the specific features as available in dataset as input variable
  6. Prediction of the Model is done given specific amount of records(16396)
  7. We cannot add any other feature in same running Model without getting this model application down and have to retrain the Model on new feature
  8. Model is giving approx 91% accuracy with Extra tree algorithm

Data Insertion In Database

- After the model is created and prediction is done the result is inserted into Cassandra database in order to keep records of result.
- Database : - Database is be created with name (Zomato), scalable,flexible format.

Q&A

- Q1) What is the source of data

- The data for training is provided by the client (ineuron) in form of csv and source of the file is

https://www.kaggle.com/datasets/himanshupoddar/zomato-bangalore-restaurants

Q2) what are the type of data.

The data was the combination of numerical and Object values. There are some null Values present in this data

Q3) What's the complete flow you followed in this Project

Refer slide 4th and 5th for better Understanding

Q4) 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.

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

- Visualizing relation of independent variableswith each other and output variables
- Removing unwanted attributes or feature
- Checking and changing Distribution of continuous values
- Cleaning data according to dataset.
- Selecting Feature .

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

- Before diving the data in training and test set we performed pre processing to order to get better data.
- As per model the training and test data were divided.
- The random split was performed over training and validation data.
- Algorithms like Linear Regresssion , Extra Tree regrssion were used based onthe accuracy and MSE,RMSE final model was used .
- saved that model for further use

Q 8) How Prediction was done?

- The data files are shared by the client then we have done some pre-processing techniques on data and trained data using Extra tree regressor algorithm and which is loaded and performed prediction.In the end we get the prediction.

Q 9) What are the different stages of deployment?

- When the model is ready we deploy it in Local environment ,Where UAT is performed over it.
- Then project will be added to github for cloud deployment
- Finally The model has been deployed in Heroku cloud platform.