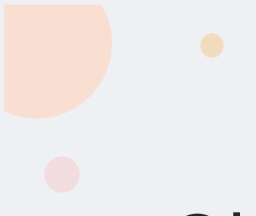


Zomato Rating Prediction

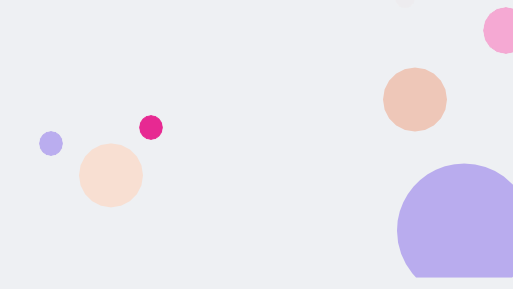




• Objective

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 their 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
- 

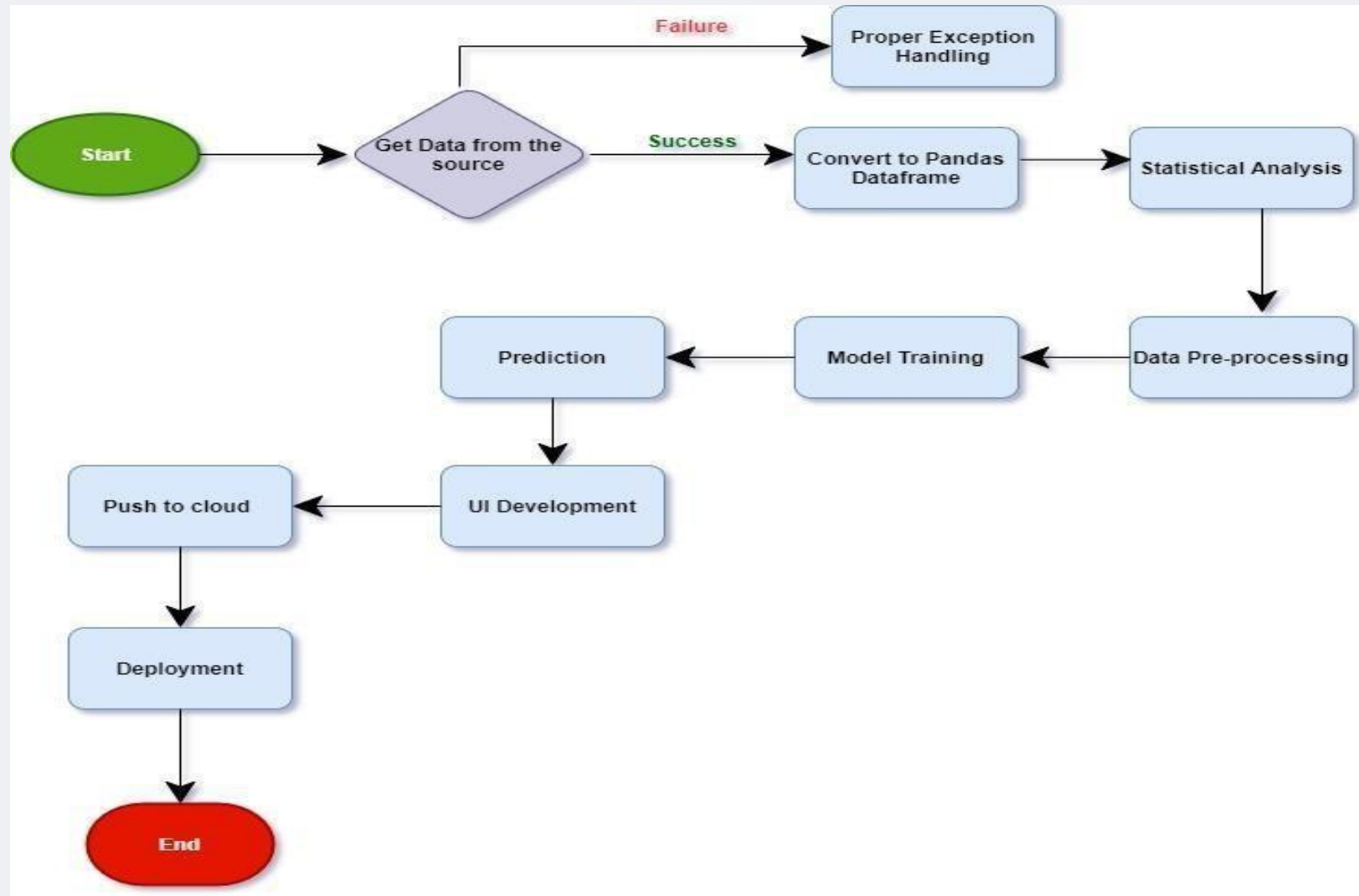
Data Sharing Agreement

- Sample file name (hours.csv) and source of the file is <https://www.kaggle.com/himanshupoddar/zomato-bangalore-restaurants>
- Shape of the data is 51717 x 17
- 51717 Rows
- 17 columns
- Column data types where:-int64,object
- Where we have use only these 10 feature among 16

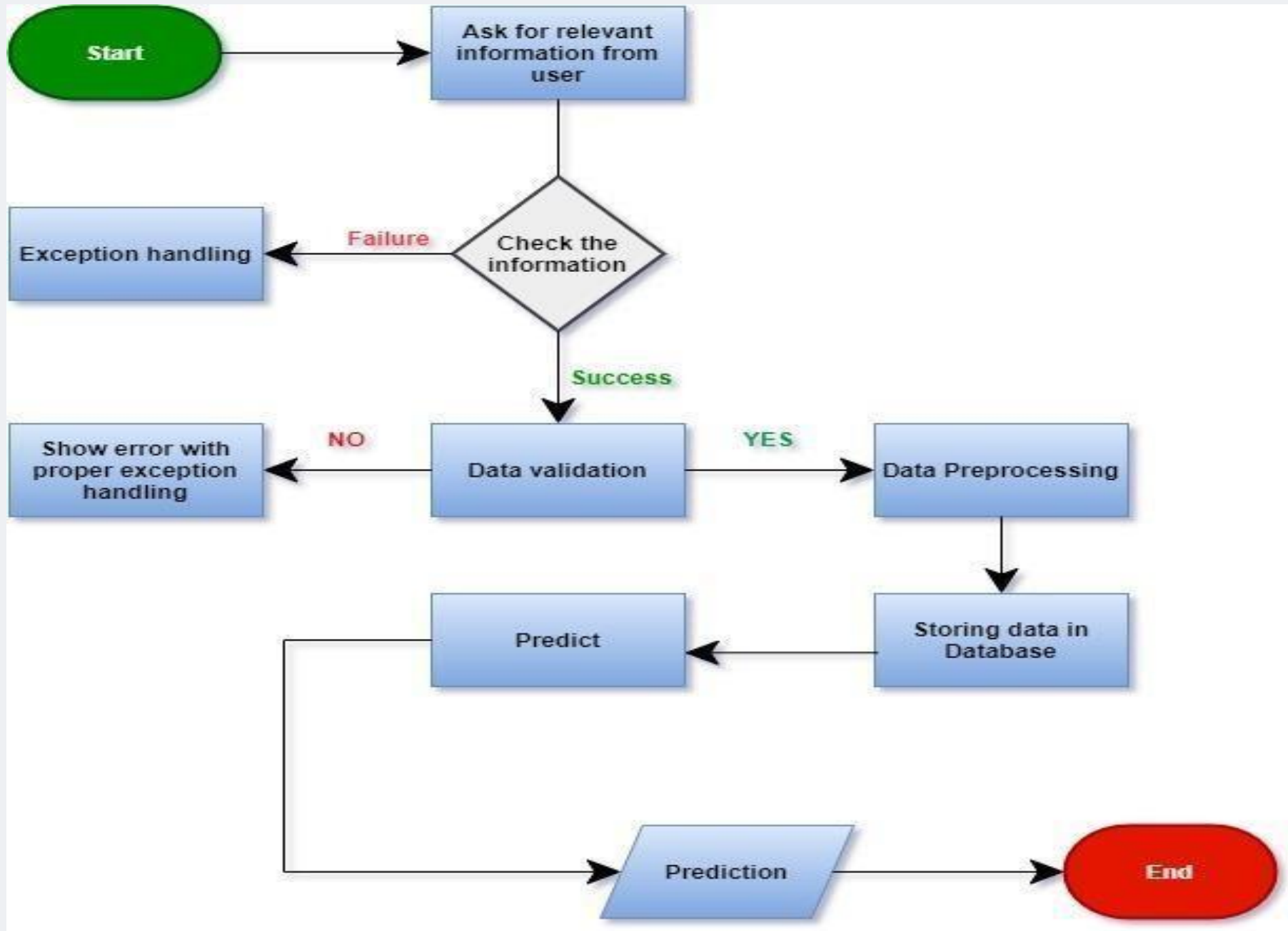
url	address	name	Online_order	Book_able	rate	votes	phone	location	Rest_type
-----	---------	------	--------------	-----------	------	-------	-------	----------	-----------

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 dataset.
- 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 multi-collinearity)

- 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 split 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
9. Once the Prediction is Done it will save in Mongo DB database.