



**Architecture**



**Data Preparation**

**Model Development**

**Deployment**

**Deployment Process**

**Architecture Description**

**Data Preparation**

The features variables include address, name, book table, online order, rate, votes,

phone, cuisines, reviews list, menu item, approximate cost for 51717 reviews available

for different restaurants. The data is taken from Kaggle Platform. It can be used for

review analysis using natural language processing, predicting restaurant rating using

regression algorithm, drawing valuable insights on about the factors affecting the

establishment of different types of restaurant at different places.

Data Preprocessing

In data preprocessing step, we check if there missing data, duplicate values, and

mixed datatypes presence in feature variables. In our dataset, feature variable that

contain missing value are rate(15.2%), approximate cost(0.88%), dish liked(54.2%),

restaurant type(0.43%) etc. There was no duplicate values but datatype of multiple

columns was “object” type, therefore they was converted to proper datatype.

Data Cleaning

This step involves outliers removal such extremely high approximate cost, missing

values data imputation using median, Datatype correction performed on features

variable containing integer variable but having object type as object.

Exploratory Data Analysis

This step includes data visualization of features and target variables. Visualization utilizes bar plot,dist plot, histograms, countplot,vilion plot ,heatmap .

Feature Engineering

New features were created utilizing provided information such as price category, that categories prices into cheap, affordable and expensive categories, no of top selling varieties. Lastly, unnecessary columns were dropped.

**Model Development**

Algorithm Spot Checking

Algorithm spot checking was performed on cleaned dataset by training the different models and evaluating their performance using Grid Search CV having 5 folds. The different model are Random Forest Regressor, Xtreme Gradient Boosting Regressor, Decision Tree Regressor , Linear Regression. Their r2 score were obtained. The highest score is achieved by Random Forest Regressor.

Hyper-parameter Tuning

The best model is Random Forest Regressor from algorithm spot checking, random search on no of estimators, learning rate, max depth and subsample is performed using Grid Search CVto get the best hyper parameters for the model. Those parameters is then used on the model to get better result.

Model Evaluation

Test dataset is used to evaluate the model. 20% of dataset was separated for testing. Predicted results of the model are compared with the actual data to check the amount of error.

**Deployment**

Streamlit Web App

For this project, the trained is deployed for inference using an interactive web app. Interactive web app is created streamlit. Streamlit is a popular open-source framework used for model deployment by machine learning and data science teams. The web application contains three pages introduction, prediction and visualization.

Designing Web App

The web application contains three pages introduction, prediction and visualization. Prediction pages provides inference services for user defined features variables. Dynamic visualization offers valuable insights about various factors affecting the restaurant establishment at different locations.

**Deployment Process**

After creating the streamlit app and testing it on local machine, code is pushed to github and linked to streamlit cloud. Specifying the branch to be utilized for creating web app, we will in creating an app hosted on streamlit cloud.