Restaurant Rating Prediction

Detailed Project Report

Amritanshu Sharma

INTRODUCTION

The basic idea of analyzing the Zomato dataset is to get a fair idea about the factors affecting the establishment of different types of restaurant at different places in Bengaluru, aggregate rating of each restaurant, Bengaluru being one such city has more than 12,000 restaurants with restaurants serving dishes from all over the world. With each day new restaurants opening the industry has'nt been saturated yet and the demand is increasing day by day. Inspite of increasing demand it however has become difficult for new restaurants to compete with established restaurants. Most of them serving the same food. Bengaluru being an IT capital of India. Most of the people here are dependent mainly on the restaurant food as they don't have time to cook for themselves. With such an overwhelming demand of restaurants it has therefore become important to study the demography of a location. What kind of a food is more popular in a locality. Do the entire locality loves vegetarian food. If yes then is that locality populated by a particular sect of people for eg. Jain, Marwaris, Gujaratis who are mostly vegetarian. These kind of analysis can be done using the data, by studying the factors such as
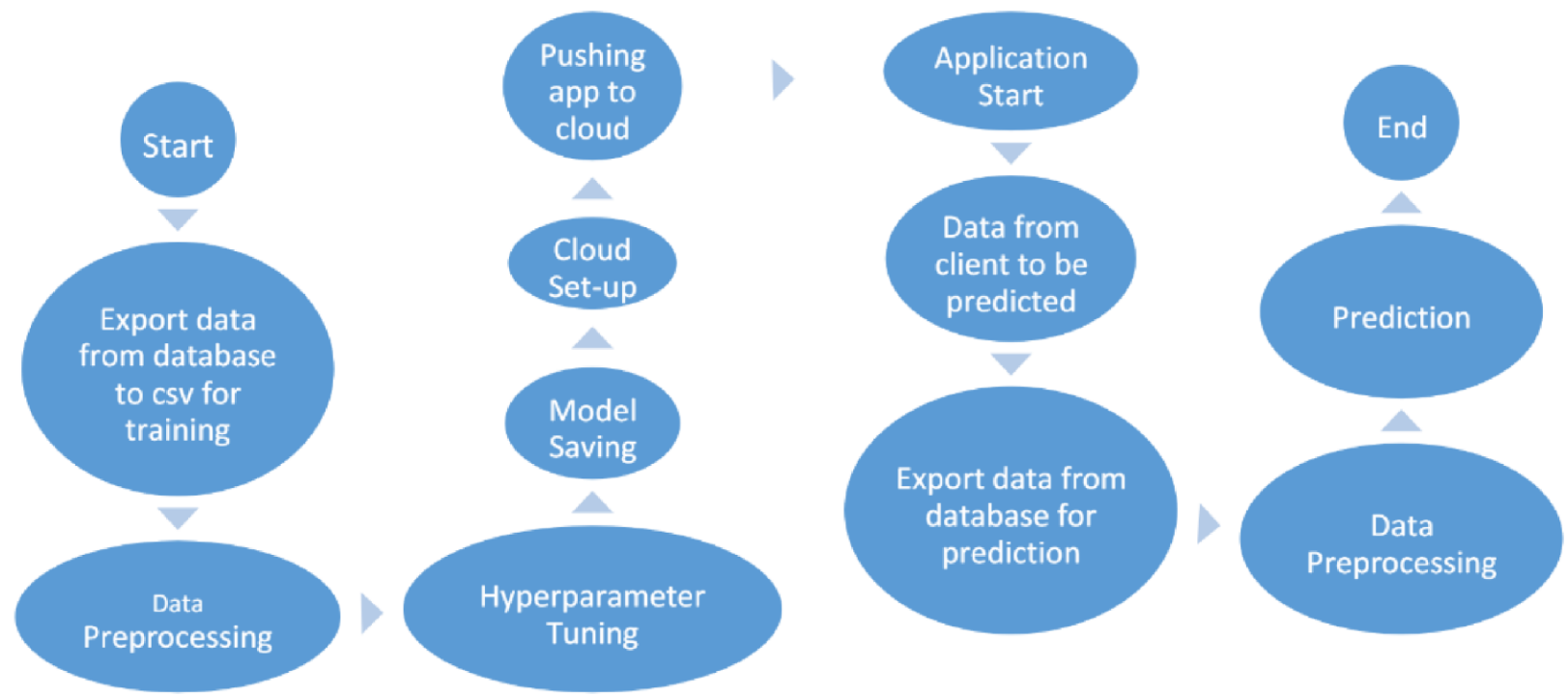
- Location of the restaurant
- Approx Price of food
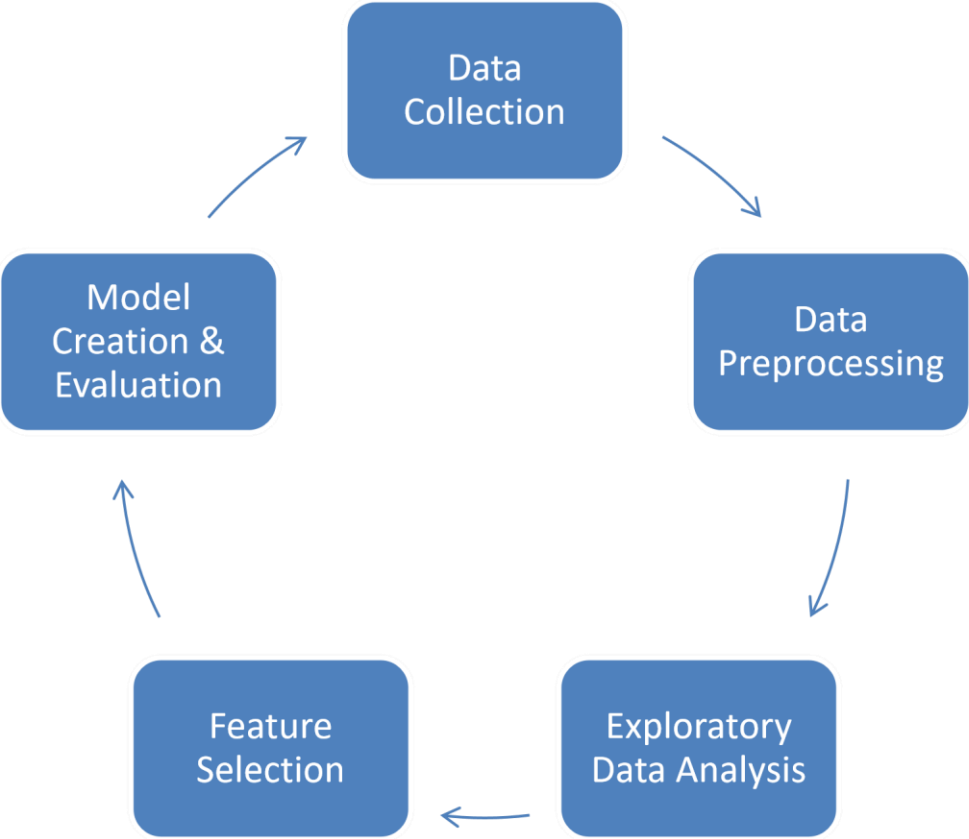- Theme based restaurant or not

OBJECTIVE

The main goal of this project is to predict the insurance premium and tell the client for their health benefits that uh can cover this much by taking insurance premium and this is good for their health and wealth.

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.
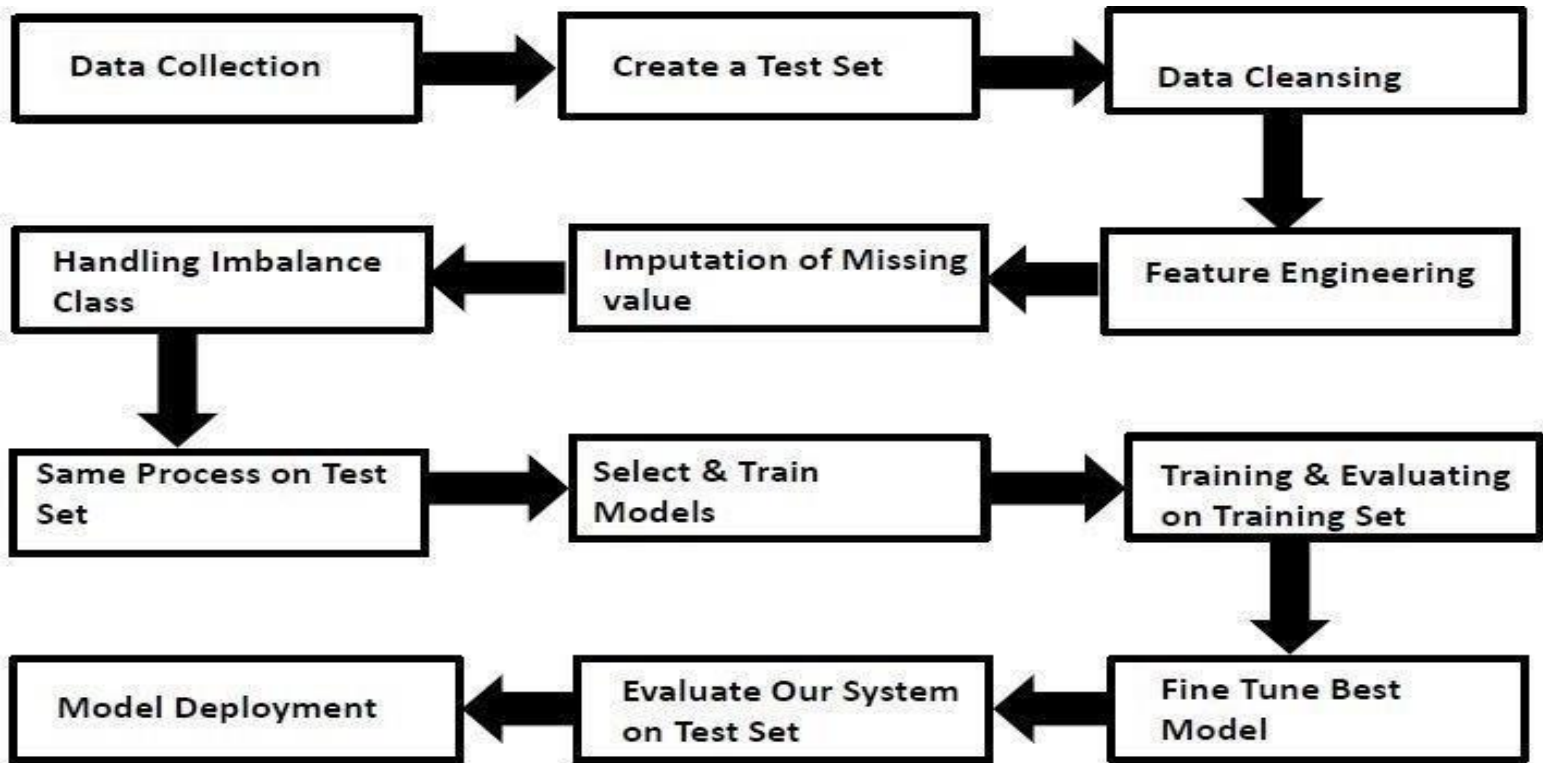
ARCHITECTURE

Data Analysis Steps

```
        Data
      Collection

  Model                    Data
  Creation &           Preprocessing
  Evaluation

        Feature          Exploratory
        Selection       Data Analysis
```

MODEL TRAINING
AND VALIDATION WORKFLOW

Data Collection	→	Create a Test Set	→	Data Cleansing

Handling Imbalance Class ← Imputation of Missing value ← Feature Engineering

Same Process on Test Set → Select & Train Models → Training & Evaluating on Training Set

Model Deployment ← Evaluate Our System on Test Set ← Fine Tune Best Model

MODEL TRAINING

AND VALIDATION
WORKFLOW

Data Collection

- Insurance Premium Prediction data set from Kaggle
- For Data Set: https:https://www.kaggle.com/datasets/himanshupoddar/zomato-bangalore-restaurants?resource=download

Data Pre-Processing

- Missing values handling by Simple imputation (Used KNN Imputer)
- Outliers' detection and removal by boxplot and percentile methods
- Categorical features handling by ordinal encoding and label encoding
- Feature scaling done by Standard Scalar method
- Imbalanced dataset handled by SMOTE -Over sampling
- Drop unnecessary columns

MODEL TRAINING
AND VALIDATION

WORKFLOW

Model Creation and Evaluation

- Various classification algorithms like Random Forest, XG Boost, KNN, etc. tested.

- Random Forest, XGBoost and KNN all were given better results. Random Forest was chosen for the final model training and testing.

- Hyper parameter tuning was performed.

- Model performance evaluated based on accuracy, confusion matrix, classification report.

Random Forest Regression Model

INTRODUCTION

It is a decision-tree-based ensemble Machine Learning algorithm which combines the output of multiple decision trees to reach a single result.

The Random Forest Regressor is a supervised learning algorithm which we can use for regression and classification problems. It is among the most popular machine learning algorithms comes under bagging ensemble technique.

Random Forest Regressor being ensemble algorithm tends to give more accurate result. This is because it works on the principle i.e., it creates the random forest by combining N decision tree, and make predictions for each tree created in the first phase. Even if one or few decision tree are prone to noise, overall results would tend to be correct.

Reason to use Random Forest Regressor model:

• It takes less training time as compared to other algorithms. •
It gives better model performance.

THANK YOU