A FIELD PROJECT REPORT

on

**Predicting Restaurant Rating Using Regression Analysis Approach**

**Submitted**

**by**

|  |  |
| --- | --- |
| 221FA04503  G. Joseph Anand Kumar | 221FA04006  N. Bala Sai |
| 221FA04575 221FA04591  V. Leela Venkata Mani Sai E. Sai Naga Lakshmi | |

**Under the guidance of**

**Sajida Sultana.Sk**

A red and black text on a white background

Description automatically generated

**DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING**

**VIGNAN'S FOUNDATION FOR SCIENCE, TECHNOLOGY AND RESEARCH**

**Deemed to be UNIVERSITY**

**Vadlamudi, Guntur.**

**ANDHRA PRADESH, INDIA, PIN-522213.**

A red and white sign

Description automatically generated

**CERTIFICATE**

This is to certify that the Field Project entitled **“****Predicting Restaurant Rating Using Regression Analysis Approach”** that is being submitted by 221FA04503 (G. Joseph Anand Kumar), 221FA04006 (N. Bala Sai), 221FA04575 (V. Leela Venkata Mani Sai), 221FA04591 (E. Sai Naga Lakshmi)for partial fulfilment of Field Project is a bonafide work carried out under the supervision of Sajida Sultana.Sk, Assistant Professor, Department of CSE.

|  |  |  |
| --- | --- | --- |
| Sajida sultana | A signature on a white background  Description automatically generated | A close up of a word  Description automatically generated  Dr.K.V. Krishna Kishore |
| Associate/Professor, CSE | HOD,CSE | Dean, SoCI |

A red and black text

Description automatically generated

**DECLARATION**

We hereby declare that the Field Project entitled **“Predicting Restaurant Rating Using Regression Analysis Approach”** is being submitted by 221FA04503 (G. Joseph Anand Kumar), 221FA04006 (N. Bala Sai), 221FA04575 (V. Leela Venkata Mani Sai), 221FA04591 (E. Sai Naga Lakshmi) in partial fulfilment of Field Project course work. This is our original work, and this project has not formed the basis for the award of any degree. We have worked under the supervision of Sajida Sultana.Sk, Assistant Professor, Department of CSE.

By

**221FA04503-(G. Joseph Anand Kumar),**

**221FA04006 (N. Bala Sai),**

**221FA04575 (V. Leela Venkata Mani Sai),**

**221FA04591 (E. Sai Naga Lakshmi)**

Date:

**ABSTRACT**

The establishment of a restaurant addresses the challenges of establishing a restaurant in a competitive market by introducing a framework for accurately predicting restaurant ratings, a vital tool for attracting customers and assessing venture success. By identifying and analyzing key factors that impact ratings, this research enables potential restaurant owners to make data-driven decisions before launching their business, reducing risks and saving time. The study employs seven regression models to compare performance metrics and determine the most reliable predictive model, ultimately providing a valuable resource that supports informed decision-making and improves the success prospects for new restaurant ventures.

**Key Words** — Consumer preferences, Data-driven decision-making, Performance metrics, Risk mitigation, Predictive modelling .

**TABLE OF CONTENTS**

1. [Introduction 01](#_bookmark0)
   1. [Why prediction of restaurant ratings are important? 02](#_bookmark1)
   2. What are the challenges in predicting restaurant ratings? 02
   3. [What is the role of Regression Analysis? 03](#_bookmark3)
   4. What are the Ethical Considerations for Predictions? 03
   5. What are the societal impacts? 03
2. [Literature Survey 04](#_bookmark5)
   1. [Literature review 05](#_bookmark6)
   2. [Key Steps in Conducting a Literature Survey 05](#_bookmark7)
   3. Motivation 08
3. [Proposed System 09](#_bookmark8)
   1. [Input dataset 10](#_bookmark9)
      1. Detailed features of dataset 11
   2. [Data Pre-processing 12](#_bookmark10)
      1. [Data Cleaning 12](#_bookmark11)
         1. Handling Missing Values 12
         2. Removing Duplicates 12
         3. Handling Outliers 13

3.2.2 Data Integration 13

3.2.3 Data Transformation 13

3.2.4 Feature Selection 13

3.2.5 Data Splitting 14

* 1. [Model Building 14](#_bookmark14)
  2. [Methodology of the system 17](#_bookmark16)
     1. Architecture of the Random forest 17
     2. Key Components 17
     3. Workflow Architecture 17
     4. Key Benefits 18
  3. [Model Evaluation 18](#_bookmark17)
  4. Model Deployment 20

1. [Implementation 21](#_bookmark19)
   1. Environment Setup 22

4.1.1Programming Language and Libraries 22

4.1.2 Development of Environment 22

4.1.3 Setup Workflow 22

4.1.3.1 Project Initialization 22

4.1.3.1 Workflow 23

1. Experimentation and Result Analysis 25
2. Conclusion 35
   1. Conclusion 36
   2. Future Enhancement 36
3. References 37

**LIST OF FIGURES**

|  |  |
| --- | --- |
| Fig – 4.1.3.2.1 : Importing Libraries |  |
| Fig – 4.1.3.2.2 : Upload Data |  | 23 |
| Fig – 4.1.3.2.3 : Label Encoding some of the Columns |  | 23 |
| Fig – 4.1.3.2.4 : Random Forest Before Parameter Tuning |  | 24 |
| Fig – 4.1.3.2.5 : Preprocessing technique in Scikit-learn |  | 24 |
| Fig – 4.1.3.2.6 : Random Forest After Parameter Tuning |  | 24 |
| Fig – 5.1 : Number of Restaurants accepting Online Orders |  | 27 |
| Fig – 5.2 : Number of Restaurants with Book Table facility |  | 28 |
| Fig – 5.3 : Type of Restaurant in Percentage |  | 28 |
| Fig – 5.4 : Average Cost in Restaurant for 2 people |  | 29 |
| Fig – 5.5 : Top 10 Liked dishes in Banglore |  | 30 |
| Fig – 5.6 : Average rating for Banglore restaurants |  | 30 |
| Fig – 5.7 : Restaurant Type vs Rate |  | 31 |
| Fig – 5.8 : Rate vs Online Order |  | 31 |
| Fig – 5.9 : True Rate vs Predicted Rate |  | 32 |
| Fig – 5.10 : Regression Models Results Before Tuning |  | 32 |
| Fig – 5.11 : Regression Models Results After Tuning |  | 33 |
|  |  |  |

**LIST OF TABLES**

Table 1. Description of Restaurant Data 11

# **CHAPTER-1 INTRODUCTION**

### **INTRODUCTION**

A restaurant in this competitive hospitality industry will heavily be determined to succeed based on customer ratings and reviews. In a world of such diversity with millions of places to dine at, it leaves the restaurants to be rated out of others and meeting up to the required standards of service. While attracting new customers is inevitable through good reviews, destructive criticism of a business operation will only deter people from coming in. This means that true future rating forecasting is crucial for new companies because the key information it gives them revolves around performance in locational, food-related, and price-sensitive areas. Opening up a new restaurant demands a vast amount of money as well as time. Most entrepreneurs are confused about determining the prospects of the project before the actual day of opening. While more traditional approaches like market research and customer surveys do provide some insight into the situation, they are generally without a data-driven predictive approach. This is where machine learning and data analytics come in. This article proposes a system using regression models to predict restaurant rating, making all controllable factors alterable even before its opening. The system analyzes various features of interest such as online ordering, reservations, and average cost for two diners in order to come up with a predicted rating. It allows for comparisons between various restaurant-related aspects and equips business owners with the information needed to make more informed decisions and mitigate risks concerning the opening of a new restaurant.

**1.1 Why prediction of restaurant ratings are important?**

Regression analysis is essential for predicting restaurant ratings in order to enhance the dining experience for patrons and enable prompt, well-informed dining choices. It gives eateries information on the main elements influencing their ratings, allowing for focused enhancements and a competitive edge. By providing more accurate results, it improves user trust and search algorithms for online platforms. It also facilitates the comprehension of consumer behavior and market trends, which propels revenue growth. All things considered, it encourages proactive reputation management, data-driven decision-making, and strategic planning for long-term success in the food sector.

**1.2 What are the challenges in predicting restaurant ratings?**

Forecasting correctly restaurant ratings is still a challenging task, despite their growing significance. Accurate prediction is hampered by several issues:  
**∙ Subjectivity of Ratings:** Personal tastes, dining experiences, and cultural backgrounds are only a few of the variables that affect ratings, which are subjective.  
**∙ Diverse Data Sources:** Online review sites, social media, and restaurant websites are just a few of the places where pertinent data for predictions can be found. Processing and integrating this heterogeneous data might be difficult.  
**∙ Datasets that are unbalanced:** Frequently, datasets have a greater percentage of favorable ratings than negative or neutral ones. Prediction models may become skewed as a result.  
**∙ Changing Preferences and Trends:** It might be challenging to identify long-term patterns and modify models in response to the ever-changing dining and consumer preferences.

**1.3 What is the role of Regression Analysis?**

One of the most reliable statistical methods for simulating the relationship between a dependent variable and (restaurant rating) and independent factors (e.g., location, reviews, price range, cuisine, etc.). Regression algorithms can accurately predict future ratings by identifying patterns and trends in prior data.

**1.4 What are the Ethical Considerations for Predictions?**

**∙ Data privacy-**Talk about the moral ramifications of using consumer data, particularly private data like dietary restrictions or personal preferences.  
**∙ Fairness and Bias-**Discuss how to minimize the possibility of bias in the models and information in order to guarantee impartial and equitable forecasts.

**1.5 What are the societal impacts?**

**∙ Consumer Empowerment**-Emphasize how precise rating forecasts can enable customers to make knowledgeable choices and steer clear of bad dining experiences.

**∙ Business Optimization**-Talk about the possible advantages for eateries, like pinpointing problem areas, streamlining menus, and raising customer satisfaction.

# **CHAPTER-2**

# **LITERATURE SURVEY**

## **LITERATURE SURVEY**

#### **Literature Review**

A literature survey, in the context of academic research, is a comprehensive examination and analysis of existing scholarly literature related to a specific research topic or question. It involves systematically identifying, selecting, reading, and synthesizing relevant research articles, books, conference articles, and other scholarly sources.

* 1. **Key Steps in Conducting a Literature Survey:**

1. **Define the research topic and scope:** Clearly describe the research topic and scope. This will help to direct the literature search.
2. **Find pertinent databases and sources:** To find pertinent literature, locate pertinent databases (like Google Scholar, JSTOR, and PubMed) and other sources (like books and conference proceedings).
3. **Perform a thorough search of the literature:**To find pertinent articles and other sources, use search tactics and keywords.
4. **Pick pertinent research:** Choose articles that satisfy the inclusion requirements and are directly related to the research topic.
5. **Read and evaluate the chosen studies:** Give careful consideration to the study methods, results, and conclusions of the chosen studies as you read and evaluate them.
6. **Synthesize the results:** Compile the results of the chosen studies, looking for recurring themes, inconsistencies, and gaps in the body of knowledge.
7. **Write the literature review:** Write a succinct and understandable literature review that highlights the main conclusions and points of contention from the chosen studies.

**Suhas Somashekar and Suhas Mallesh** **[1]** The report describes a study that employed a variety of regression models to predict restaurant ratings. The authors describe the processing and analysis of data from more than 43,000 eateries in Bengaluru, India, and stress the significance of restaurant ratings for the development of new businesses. Using criteria like R2 score and mean absolute percentage error, seven regression models—including XGBoost, decision trees, and linear regression—were assessed. The best-performing models were Random Forest, ADA Boost, and XGBoost, indicating that these models are capable of accurately predicting ratings. The goal of the study is to maximize the chances of success for new restaurants by assisting them in making well-informed judgments.(2021)

**J. Priya** **[2]** Leveraging a dataset from Bengaluru's restaurant industry, the study describes how machine learning is used to predict restaurant evaluations. It entails preparing the data, visualizing it, and using eight distinct regression algorithms—such as Bayesian, Random Forest, Ridge, and

Linear Regression—to create predictive models. Metrics like regression score and error rates are

used to assess each model's performance. According to the results, Random Forest Regression performs better than the others, having the lowest error and the maximum accuracy. Restaurants can improve their services and business plans with the help of this analysis.(2020)

**Ibne Farabi Shihab et al.** **[3]** The article describes a machine learning method that uses Yelp data to assist business owners in finding the best sites for future restaurant openings. In order to predict restaurant ratings, four machine learning methods were used: Decision Tree, Logistic Regression, Decision Tree (presorted), and Support Vector Machine (SVM). Data pretreatment and feature selection were also part of the process. SVM proved to be the most efficient method for this problem, with the maximum accuracy of 97%. Based on projected ratings, the study's conclusion offered suggestions for the best cities to help businesses make judgments.(2019)

**Mara-Renata Petrusel and Sergiu George Limboi** **[4]** The study introduces a restaurant recommendation system that incorporates sentiment analysis to improve rating predictions. It uses machine learning classifiers such as Naive Bayes, SVM, and Logistic Regression to identify sentiment and categorize Yelp restaurant reviews as either positive or negative. In order to enhance rating predictions for restaurants that a user has not yet visited, the sentiment data is then integrated with collaborative filtering. The findings indicate that when compared to conventional techniques, this sentiment-enhanced approach yields more accurate recommendations. Polarity scores will be improved for even greater customisation in future research.(2019)

**Tugc ¸e Bilen et al. [5]** This study suggests the Business Location Estimator, a smart city program that employs machine learning to recommend the best places for businesses, particularly eateries. In order to estimate feature values for upcoming years with the least amount of error, the system uses regression models (such as SMOReg and MLP) to compile features from London data, including demographic and economic characteristics. In order to assist entrepreneurs in choosing appropriate locations, hierarchical clustering groups districts with comparable characteristics based on these projections. Users can examine suggested clusters on a map and interact with the model using a web interface.(2018)

**Ismam Hussain Khan et al.** **[6]** This study uses information gathered from Epicurious.com, such as servings, ingredients, and directions, to create a machine learning model that predicts ratings for culinary recipes. Classifiers like Naive Bayes, Logistic Regression, K-Nearest Neighbors (KNN), and Random Forest are trained and assessed following data preprocessing. The best accuracy and stability results are obtained with KNN and logistic regression. Prediction outcomes are further improved by an ensemble model that combines KNN, Random Forest, and Decision Tree. Future developments will involve investigating neural networks for rating prediction and incorporating additional features.(2021)

**Neha Vaish et al.** **[7]** The study used machine learning techniques to do sentiment analysis on hotel reviews. After preprocessing the 39,000 review dataset to eliminate noise, Latent Dirichlet Allocation (LDA) was used to extract aspects. Algorithms such as Support Vector Machine (SVM), Naive Bayes, Decision Tree, and Logistic Regression were used to classify the reviews. Based on metrics such as accuracy, precision, recall, and F1 score, the study revealed that Logistic Regression outperformed the other classifiers.(2022)

**Sandeep Bhatia et al.** **[8]** The study used machine learning to forecast restaurant success by analyzing Zomato restaurant data from Bengaluru. Eighty percent of the preprocessed and feature-selected dataset was utilized for training, while twenty percent was used for testing. Out of the five classification algorithms that were assessed, Decision Tree had the greatest accuracy (84.8%), followed by KNN, Random Forest, Logistic Regression, and Naive Bayes. The findings are intended to assist restaurant operators with strategic planning and the identification of success-influencing variables. (2023)

**Yi Luo and Xiaowei Xu** **[9]** The study employed localized linear regression models to predict Yelp reviews' usefulness. From user and business reviews, it extracted a variety of elements, such as text-based, geographic, statistical, and qualitative data. To improve performance, a localized weighted regression technique was added, along with regression models such as Ridge Regression and LASSO. According to RMSLE, this method significantly increased prediction accuracy. The study found that integrating popularity, geography, and language characteristics produced superior outcomes to conventional techniques.(2019)

**Xiaochen Wang, Yanyan Shen, Yanmin Zhu** **[11]** The study suggests a way to forecast new restaurant ratings without depending on patron feedback. In order to estimate factors influencing ratings, it pulls pertinent information from restaurant and urban data (such as points of interest, road networks, and satellite light). These features are combined by MR-Net, a deep learning method that emphasizes spatial interdependence, to predict scores. The technique performs better than eight baseline models, demonstrating how well restaurant and urban data can be combined to predict ratings.

**Yifan Chen; Fanzeng Xia [12]** The purpose of the research is to project Yelp restaurant ratings for 2019 by evaluating data from 2018. It makes use of both non-textual (such as review counts and restaurant type) and textual (such as review sentiment) information. A maximum accuracy of 82.5% was attained by testing five machine learning models, including Decision Tree and Neural Network. The findings imply that enhancing data and techniques could increase forecast accuracy even more, which would be advantageous to investors and restaurant owners.​

**Nabiha Asghar [13]** In order to forecast Yelp review ratings based on textual reviews, the article tackles the "Review Rating Prediction" problem. By combining four machine learning algorithms (support vector classification, logistic regression, Naïve Bayes, and perceptrons) with four feature extraction techniques (unigrams, bigrams, trigrams, and Latent Semantic Indexing), it investigates sixteen prediction models. According to the study, logistic regression works best when combined with bigrams and unigrams. It also talks about possible enhancements, such as the use of non-linear models, sophisticated feature engineering, and extra validation techniques.

**Nanthaphat Koetphrom [14]** Three filtering strategies—content-based, collaborative, and hybrid filtering—for a restaurant recommendation system are compared in this study. Regression based on restaurant and customer attributes is used in content-based filtering. Collaborative filtering predicts ratings based on the ratings of comparable users by using weighted averages, cluster analysis, and similarity tests. Predictions from both approaches are combined in hybrid filtering. When measured by mean absolute error (MAE), hybrid filtering outperformed content-based or collaborative filtering alone in terms of predicting satisfaction.

**Sanjukta Saha; A. K. Santra [15]** The study offers a recommendation system that assigns ratings to Kolkata restaurants based on Zomato user reviews. It employs sentiment analysis to categorize evaluations as neutral, negative, or positive. These ratings are then calculated to get an overall score for every business. By examining the evaluations of comparable users, the method incorporates collaborative filtering to forecast consumer preferences. Based on thorough reviews, the system provides a computed ranking for every restaurant, assisting customers in choosing where to eat.

**F. M. Takbir Hossain; Md. Ismail Hossain; Samia Nawshin [17]** In this article, a machine learning-based sentiment prediction model for restaurant reviews is presented. After gathering reviews from a website in Bangladesh, the authors parsed the content and used classification techniques (such SVM and Naïve Bayes) to classify the reviews as either positive or unfavorable. After testing a number of classifiers, they discovered that logistic regression was the most successful, with an accuracy of over 77%. In order to assist restaurant owners in understanding consumer mood and market positioning, the model additionally computes a positive-to-negative review ratio. To improve insights, future work will involve developing multi-level sentiment classifiers.

#### **Motivation**

Regression analysis is mostly utilized for forecasting restaurant ratings in order to use information-driven knowledge to enhance decision-making for both patrons and restaurant owners. We can create precise predictive models that enable customers to make knowledgeable decisions and assist restaurant operators in streamlining their operations by examining past data and determining the major elements impacting ratings. By offering insightful information about consumer preferences, market trends, and the effects of many factors on customer satisfaction, this strategy has the potential to completely transform the food sector.

# 

# **CHAPTER-3**

# **PROPOSED SYSTEM**

### **PROPOSED SYSTEM**

The proposed system in the article, "**Predicting Restaurant Ratings Using Regression Analysis Approach**" is a comprehensive predictive framework designed to forecast restaurant ratings using a variety of regression models. The goal of the system is to enable restaurant owners and start-ups to make data-driven, strategic decisions based on variables that have a big influence on customer happiness and overall business performance. This solution uses predictive analytics to assist reduce the risks associated with setting competitive pricing, maximizing service quality, and selecting a company location—all of which are critical components in the fiercely competitive restaurant industry.

Customer reviews and ratings are crucial in deciding a restaurant's performance in the highly competitive hospitality sector. Low ratings and unfavorable reviews can turn off potential consumers, who mostly rely on ratings when choosing where to eat. Consequently, a major strategic advantage can be gained by correctly forecasting a restaurant's rating prior to launching or altering operations. Market research and surveys, two conventional methods of assessing possible restaurant company success, are imprecise and lack the insight that data-driven forecasts may bring. This suggested solution gives restaurant owners a methodical approach to find and address important elements that contribute to high ratings by examining trends in past data.

To forecast ratings based on a dataset with several pertinent factors, such as location, service quality, cost, and dining options, the system makes use of a variety of regression models. This allows restaurant owners to concentrate on insights that can be put into practice, assisting them in making well-informed decisions regarding important aspects even prior to the opening of their facility. In the end, this data-driven strategy improves a restaurant's chances of success by maximizing important controllable factors that affect rating results and customer happiness.

The goal of system extension for the suggested restaurant rating prediction model is to improve its predicted accuracy, boost its usefulness for restaurant owners, and expand its applicability in a variety of scenarios. Adding more data sources, improving prediction models, and investigating novel use cases are all part of system expansion. Adding more characteristics to the dataset can yield deeper understanding of the variables affecting restaurant evaluations.

#### **Input Dataset**

The input dataset comprises thorough compilation of restaurant data, describing a range of characteristics that are crucial to comprehending the elements affecting restaurant evaluations. There are 9,551 records in this collection, each of which represents a distinct restaurant. A variety of attributes are included in each record, arranged in 21 columns that offer comprehensive details about the restaurant's location, offerings, and patron reviews. To determine the exact location and setting of each restaurant, the columns contain a unique identification, restaurant name, and geographical indications like city, country code, address, locality, and a descriptive description of

the locality. Longitude and latitude coordinates are also included in the dataset, which enables spatial analysis and mapping of restaurant locations. This could reveal geographic trends and their impact on customer ratings.

This dataset provides a strong basis for examining consumer preferences, service aspects, and location-specific trends by encompassing a variety of criteria that affect restaurant evaluations. Regression analysis and other predictive modeling approaches can be used to determine the important factors that influence customer evaluations because of the mix of quantitative and qualitative data, which offers important insights into restaurant business consumer behavior.

#### **Detailed Features of the Dataset**

The dataset seems to be a thorough compilation of restaurant data, encompassing a broad range of characteristics associated with different restaurants.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| |  | | --- | | **Column Name** |  |  | | --- | |  | | **Data Type** | **Description** |
| Restaurant ID | int64 | Unique identifier for each restaurant |
| Restaurant Name | object | Name of the restaurant |
| Country Code | int64 | Code representing the country where the  restaurant is located |
| City | object | City where the restaurant is located |
| Address | object | Full address of the restaurant |
| Locality | object | Specific area or neighborhood of the restaurant |
| Locality Verbose | object | Detailed description of the locality |
| Longitude | float64 | Longitude coordinate of the restaurant |
| Latitude | float64 | Latitude coordinate of the restaurant |
| Cuisines | object | Types of cuisines offered by the restaurant |
| Average Cost for two | int64 | Average cost for two people |
| Currency | object | Currency used for pricing |
| Has Table booking | object | Indicates if table booking is available (Yes/No) |
| Has Online delivery | object | Indicates if online delivery is available (Yes/No) |
| Is delivering now | object | Indicates if the restaurant is currently delivering (Yes/No) |
| Switch to order menu | object | Option to switch to the order menu (Yes/No) |
| Price range | int64 | Price range category (1 to 4, with 4 being the highest) |
| Aggregate rating | float64 | Overall rating of the restaurant |
| Rating color | object | Color associated with the rating (e.g., Dark Green) |
| Rating text | object | Text description of the rating (e.g., Excellent, Good) |
| Votes | int64 | Number of votes the restaurant has received |

#### 

#### Table-1 : Description of Restaurant Data

#### **Data Pre-processing**

Data preprocessing is an essential step in the data analysis and machine learning pipeline, serving as the bridge between raw data and a strong analytical model. Considering the link between unprocessed data and a robust analytical model, data preparation is a crucial stage in the pipeline for data analysis and machine learning. This stage involves several steps to convert unstructured and often flawed raw data into a refined version suitable for further analysis or modeling. The primary goal of data preprocessing is to improve the quality and reliability of the data because this enhances model performance, accuracy, and the overall credibility of the study.

Through data preparation, raw, unstructured data is converted into an understandable, structured format for effective model processing. It lays the foundation for reliable and insightful analysis by minimizing any risks that could compromise the model's accuracy and generalizability. By addressing the complexities of real-world data through cleaning, transformation, feature selection, and other preparatory procedures, preprocessing ensures that the subsequent stages of data analysis or machine learning can grow upon a solid and dependable foundation.

**3.2.1.Data Cleaning**

**3.2.1.1 Handling Missing Values**

Missing data is a frequent problem in datasets that can result from a number of things, including missing survey responses, equipment failures, and data entry errors. Missing values must be handled carefully since they can skew results or lower an analysis's statistical power. The following are a few methods for handling missing data:

1. Eliminating Columns or Rows
2. Using Statistical Techniques to Impute Values
3. Complex Imputation Methods

**3.2.1.2 Removing Duplicates**

Errors in data entry, the combining of data from several sources, or repeated measurements can all result in duplicates in a dataset. In order to prevent biased analysis or exaggerated data, duplicates must be handled carefully.

* Identifying Duplicates
* Removing Duplicates

**3.2.1.3 Handling Outliers**

Data points that substantially deviate from other observations are known as outliers, and they can have a negative impact on statistical studies like regression models and mean computations.

1. Detection Methods

* Z-Score Method
* Interquartile Range (IQR)
* Visual Inspections

1. Mitigating Outliers

* Trimming
* Capping
* Transformation
* Robust Methods

1. Advanced Techniques

* Clustering Algorithms
* Isolation Forest
  + 1. **Data Integration**

In order to generate a single, consistent dataset, data integration entails combining information from several disparate sources. This procedure is essential for creating thorough models that forecast results like restaurant ratings. The quality and dependability of predictive models can be increased by analysts adding diverse insights to their datasets through the integration of data from various sources.The benefits of data integration are

* Comprehensive Insights
* Improved Model Accuracy
* Enhanced Feature Variety
  + 1. **Data Transformation**

For machines to extract significant characteristics and ensure uniform scales and formats, data transformation is crucial for preparing data for machine learning algorithms. Models converge more quickly and precisely when normalization and standardization are done correctly. Models may comprehend qualitative data by encoding categorical variables, and feature engineering offers fresh perspectives that can greatly enhance prediction performance. Gaining proficiency in these methods produces datasets that are cleaner and more informative, which can improve machine learning projects' efficacy.

* + 1. **Feature Selection**

A crucial component of data preparation is feature selection, which entails locating and selecting the most significant and pertinent features for constructing a machine learning model. This procedure' primary objective is to simplify the dataset by concentrating solely on the variables that significantly affect the model's ability to predict outcomes. The risk of overfitting, in which the model becomes unduly complicated and captures noise instead of the underlying data patterns, is lessened by feature selection, which lowers the number of features. This simplification increases the prediction capacity of models by allowing them to generalize more effectively when exposed to unknown input.

Feature selection is crucial because it can strike a balance between model performance and complexity. Models may become computationally costly and less interpretable when datasets contain an excessive number of characteristics, particularly those that are redundant or useless. For jobs where model explainability is critical, choosing a subset of relevant characteristics makes models easier to understand and faster to train.

* + 1. **Data Splitting**

Through partitioning the available dataset into discrete subsets—training, validation, and testing—data splitting, a basic machine learning technique, guarantees the resilience and dependability of a model. The main function of this division is to evaluate a model's ability to generalize to new, untested data. By separating the data, practitioners can replicate real-world situations in which a model comes into contact with data it has never seen before, giving a more realistic assessment of the model's predictive power.

The learning algorithm uses the training data to fit the model and modify its internal parameters. Here, the model discovers the links and underlying patterns in the data. Overfitting, in which the model becomes overly adapted to the particular subtleties of the training data and performs poorly on fresh data, might result from testing the model just on the training set.

#### **Model Building**

The process of choosing and creating a mathematical or computational framework to depict relationships in data for analytical or predictive purposes is known as model development. To guarantee accurate and dependable results, this procedure entails describing the issue, choosing suitable methods or approaches, training the model on data, and fine-tuning it. Developing a model that can effectively generalize to new data by generating predictions or insights from patterns discovered during the training stage is the aim of model construction.

Data preparation, feature selection, selecting a model type, training the model, evaluating its performance, and fine-tuning it based on evaluation metrics are some of the essential stages that are usually included in model creation. Iterative in nature, this procedure frequently calls for feature engineering or hyperparameter tuning to enhance model performance. A model that is appropriate for the particular issue it is meant to address and strikes a balance between accuracy, complexity, and generalizability should be the end product.

The nine predictive model-building models listed with an emphasis on their traits, uses, benefits, and possible drawbacks are :

* **Linear Regression**

A basic statistical technique called linear regression fits a linear equation to the observed data in order to model the connection between a dependent variable and one or more independent variables. The formula is,

where stands for the error term and β values are coefficients that measure the influence of each predictor X on the result Y. The accuracy of the model may be impacted by outliers, as linear regression is sensitive to them and presumes a linear connection between variables. Although it is simple to use and understand, it has trouble with multicollinearity among predictors and intricate, non-linear interactions.

* **Random Forest**

An ensemble learning technique called Random Forest builds several decision trees during training and aggregates their results to improve prediction accuracy and manage overfitting. The mode (classification) or mean (regression) of all trees is the final prediction made by each tree, which is constructed using a random subset of features. For managing complicated interactions and non-linear relationships in data, Random Forest is incredibly reliable and efficient. In comparison to a single decision tree, it is also comparatively resistant to overfitting. However, as the number of trees increases, the model may become less interpretable and more computationally demanding.

* **Ridge Regression**

A regularization factor is incorporated into Ridge Regression, a kind of linear regression, to avoid overfitting, particularly when multicollinearity among features is present. By adding a penalty proportionate to the sum of the squared values of the coefficients, which is managed by the hyperparameter λ , the model alters the cost function. By decreasing the size of the coefficients, this penalization stabilizes the solution and enhances generalization. Because Ridge Regression retains all predictors in the model, although with reduced coefficients, it performs well on datasets with a high number of associated features but is unable to do feature selection.

* **Lasso Regression**

Similar to Ridge Regression, Lasso (Least Absolute Shrinkage and Selection Operator) Regression regularizes the linear regression model. Nevertheless, Lasso penalizes the absolute sum of the coefficients by using an L1 regularization term. By eliminating less significant predictors, this results in some coefficients being set to zero, so carrying out feature selection. Because of this characteristic, Lasso is particularly helpful for models that need to be easy to understand. When a dataset has a lot of features, Lasso helps find and keep only the most important ones. When features are highly connected, Lasso may have trouble since it might choose one at random while disregarding others.

* **Support Vector Machine(SVM)**

A strong and adaptable supervised learning model for classification and regression applications is the Support Vector Machine (SVM). By identifying the hyperplane that optimizes the margin between data points, the regression model (SVR) seeks to match the data within a certain tolerance margin. Combining SVM with kernel functions (such as a polynomial or radial basis function) makes it efficient for high-dimensional spaces and non-linear data. SVM's performance is dependent on the selection of hyperparameters such as the kernel type and regularization parameter, and it can be computationally demanding, particularly when dealing with huge datasets.

* **K-Nearest Neighbors(KNN)**

A straightforward, non-parametric model for regression and classification is K-Nearest Neighbors (KNN). By averaging the values of a data point's k closest neighbors, as measured by a distance metric like Euclidean distance, it forecasts the value of the data point. KNN can adjust to different data distributions and model intricate non-linear relationships. It does, however, have a number of drawbacks, including the potential for computational expense for big datasets, the need for careful k selection to prevent underfitting or overfitting, and potential susceptibility to noise or unbalanced data. Furthermore, if the data contains a large number of redundant or irrelevant characteristics, KNN performance may suffer.

* **Decision Tree**

A decision tree is a model that divides data into branches according to feature values, with each leaf node producing a forecast or decision. Because it is simple to understand and intuitive, it is frequently used for predictive modeling and exploratory data analysis. Non-linear correlations and interactions between features can be captured using decision trees. They are vulnerable to overfitting, though, as they have the potential to become intricate and deep, over-modeling the training data. To manage this problem, methods like pruning and establishing the maximum tree depth are frequently employed. Even while decision trees offer precise guidelines for making decisions, using ensemble techniques like Random Forest or boosting algorithms frequently improves their predictive effectiveness.

* **AdaBoost**

Adaptive Boosting, or AdaBoost, is an ensemble learning strategy that trains weak learners—usually shallow-depth decision trees—sequentially to increase their accuracy. AdaBoost modifies the weights of the training data points at the end of each iteration, giving special attention to those that were incorrectly classified or badly forecasted in earlier iterations. By concentrating the model's attention on data that is more difficult to predict, this method enables the model to grow over time and learn from its errors. AdaBoost provides greater accuracy than single learners and works well for both classification and regression problems. However, because it keeps making outliers and noisy data more significant over iterations, it may be vulnerable to them.

* **XGBoost**

A sophisticated gradient boosting implementation made for speed and performance is called XGBoost (eXtreme Gradient Boosting). Iteratively, it creates a group of decision trees, each of which fixes the mistakes of the one before it. Regularization terms in XGBoost improve model generalizability by preventing overfitting. It is well-liked for competitive machine learning projects and large-scale data due to its fast handling of missing information, parallel processing capabilities, and streamlined algorithms. Although XGBoost is renowned for its exceptional speed and accuracy, tuning it can be more difficult than with simpler algorithms. To attain best performance while preserving computing economy, rigorous hyperparameter adjustment is necessary.

In recognition of their resilience and adaptability, random forests are frequently chosen in a variety of modeling settings. They are excellent at managing both numerical and categorical variables in intricate datasets with many of characteristics. Random forests' ensemble nature, which aggregates the predictions of several decision trees, improves generalization on unknown data and reduces overfitting. Furthermore, random forests come with built-in feature importance metrics that let users determine which predictors in their models are the most important. Random forests are a popular choice in both academic research and real-world applications across a variety of disciplines because of their comparatively low processing cost when compared to other ensemble approaches and their capacity to function well with little adjustment.

#### **Methodology of the system**

Having discussed the foundational elements in the preceding sections, we now venture into the core of our traffic congestion prediction system. In this section, we embark on a journey through the inner workings of our model, unveiling the methodology that drives our system's ability to forecast traffic congestion. Just as a well-orchestrated symphony requires each instrument to play its part harmoniously, our methodology combines data, pre-processing, modelling, and evaluation to create a seamless and efficient prediction system. In all the models used the performance of Random Forest is good.

* + 1. **Architecture of Random Forest**

Building several decision trees during training and combining their predictions is the foundation of the Random Forest ensemble learning technique. It works well for jobs involving both regression and classification. The architecture of the model is made up of many trees, each of which processes a random portion of the dataset's attributes and samples to produce a variety of predictions that are frequently more correct.

* + 1. **Key Components**

The fundamental units of a Random Forest are decision trees. Every tree in the forest operates on its own, categorizing information or forecasting outcomes based on data divides.  
**Bootstrap Aggregation (Bagging):** This method uses sampling with replacement to create multiple subsets of data. Decision trees are trained independently for each subset.

**Feature Randomness:** Random Forest reduces overfitting and increases tree diversity by choosing a random subset of characteristics for every tree.

**Voting Mechanism:** Each tree casts a vote for the output class in classification tasks, and the class with the most votes is chosen. The forest uses the average of each tree's output for regression.

* + 1. **Workflow of Architecture**

**Data Sampling:** To add diversity and lessen correlation between trees, Random Forest chooses samples and features at random for every tree.

**Tree Building:** Every decision tree is trained using a subset of collected data to identify particular patterns.

**Prediction Aggregation:** After training the forest, all tree outputs are combined to create predictions. This is accomplished in regression by averaging outputs, and in classification by majority vote.

**3.4.4 Key Benefits**

**High Accuracy**: By averaging predictions from several models, the ensemble technique lowers overfitting and raises accuracy.

**Robustness to Overfitting:** On complicated datasets, overfitting is lessened by using random samples and subsets of features for every tree.

**Performs Well with High-Dimensional Data:** By taking into account only subsets of features at each split, Random Forests are able to efficiently handle enormous feature sets.  
Addresses Outliers and Missing Values: Random Forest can handle a variety of data problems since decision trees are comparatively resilient to outliers and missing information.

#### **Model Evaluation**

A machine learning model's performance and efficacy are evaluated through a theoretical and practical procedure called model evaluation. Assessing a model's ability to generalize to new data and make sure it correctly represents the underlying patterns in the data without overfitting or underfitting are the two main objectives of model evaluation. Understanding the model's advantages, disadvantages, and applicability for use in practical applications depends on this procedure.

Assessing a model's prediction performance entails using particular indicators and methodologies. Metrics like accuracy, precision, recall, F1-score, and AUC-ROC are frequently used for classification tasks in order to assess how well the model differentiates between various classes. Metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and the R-squared (R²) score are used to evaluate how well the model predicts actual values for regression tasks.

A different test dataset that the model hasn't encountered during training or validation is usually used for model evaluation. This guarantees that the model's behavior when exposed to fresh data is reflected in the performance measures. A more thorough grasp of the model's stability and dependability across several data samples is also provided by methods like cross-validation, which divides the dataset into several subsets for training and validation.

The model evaluates the performance of different regression models using the R-squared (R²), Accuracy, and Root Mean Squared Error (RMSE) metrics.

* **R2 Score**

A statistical metric called the coefficient of determination, or R-squared (R²) score, is used to evaluate the percentage of the dependent variable's variance that can be predicted from the independent variables. Based on the percentage of total variation described by the regression model, it shows how effectively the model replicates the actual results.

R2 assesses a regression model's goodness of fit. It falls between 0 and 1, where:  
  
**a.** **R2 = 1**: Shows that there is no prediction error and that the model fully explains the variance in the dependent variable.

**b.** **R2 = 0**: Indicates that the model does not explain any of the variance in the dependent variable, which means it is no more effective than making predictions based on the data mean.

**c. Negavtive R2 :** When the model fits the data less well than a horizontal line that represents the dependent variable's mean, a negative R2 may be observed.

Mathematically R2  score is represented as ,

A mathematical equation with numbers and symbols

Description automatically generated

Regarding assessing and contrasting the performance of regression models, the R-squared score offers a rapid and clear indicator of how well a model fits the data. For a thorough grasp of model efficacy, it should be interpreted in conjunction with additional metrics and diagnostic tests.

* **Root Mean Squared Error**

A popular metric for assessing a predictive model's accuracy, especially in regression tasks, is Root Mean Squared Error (RMSE). The average magnitude of the discrepancy between expected and actual values is measured by RMSE. In essence, it is the average of the squared discrepancies between the actual and expected values, squared as the square root.

Mathematically RMSE is represented as,

A mathematical equation with numbers and symbols

Description automatically generated

Inevitably require a metric that is simple to understand and shows the average forecast error on the same scale as the target variable, RMSE can be helpful. Because it emphasizes huge errors more than other measures like Mean Absolute Error (MAE), it is particularly useful when outliers are a particular issue.

* **Accuracy**

Simply calculating the frequency with which a model's predictions come true, accuracy is a metric used to assess a model's performance. Especially for classification tasks, it is among the most straightforward and widely used measures. The ratio of accurately predicted observations to total observations is the definition of accuracy in binary or multi-class classification.

Mathematically Accuracy is represented as,

A black text with black text

Description automatically generated

The term "accuracy" is not commonly used in regression tasks in the same manner as it is in classification tasks. Performance measures such as Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R-squared (R²) are more frequently used for regression. The percentage of predictions that fall within a specific error threshold from the actual values, however, can be used to modify accuracy for regression tasks.

In order to have a thorough understanding of a model's performance, accuracy should be utilized in conjunction with other evaluation measures, particularly when working with regression tasks or imbalanced data.

**3.6 Model Deployment**

The process of making a trained machine learning model usable in a production setting so that it can provide insights or predictions based on fresh input data is known as model deployment. In order for a model to engage with people or systems and provide value from its predictions, it must be deployed from a development or experimental stage into an actual application.

The process of turning a machine learning model into a workable application is known as model deployment. It guarantees that development efforts have practical applications, enabling businesses to use predictive models for improved automation and decision-making. In order to maintain the model's applicability and efficacy, proper deployment procedures include both initial setup and continuing maintenance.

**CHAPTER – 4**

**IPLEMENTATION**

**4 . IMPLEMENTATION**

**4.1 Environmental Setup**

For the researcher to create, train, and assess the regression models for restaurant rating prediction, the environment setup and execution are essential.

**4.1.1 Programming Languages and Libraries**

Considering Python's broad support for machine learning and data analysis through modules like

* **Pandas:** It is used for data manipulation and analysis, the implementation most likely uses Python.
* **Numpy:** For numerical operations, use NumPy.
* **Scikit-learn:** Several machine learning models, such as Linear Regression, Ridge, Lasso, and KNN, can be implemented with Scikit-learn.
* **XGBoost:** To execute the XGBoost model, use the XGBoost library.
* **Seaborn and Matplotlib:** Seaborn and Matplotlib are used to visualize data.
* **SciPy:** For other statistical operations, use SciPy.

**4.1.2 Development of Environment**

* A possible choice for interactive coding, testing, and visualization is Google Colab or Jupyter Notebook.
* Additional complicated coding and project management could be done with Integrated Development Environments (IDEs) like VSCode or PyCharm.

**4.1.3 Setup Workflow**

Establishing the workflow for a machine learning project entails following a set of organized procedures to transform raw data into a deployed, assessed model.

* + - 1. **Project Initialization**
* Create a Project Directory
* Version Control
  + - 1. **Workflow**

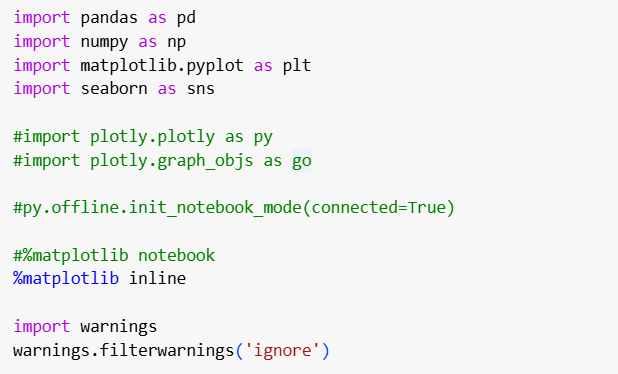
****

Fig – 4.1.3.2.1 : Importing Libraries

**A close-up of a word

Description automatically generated**

Fig – 4.1.3.2.2 : Upload Data

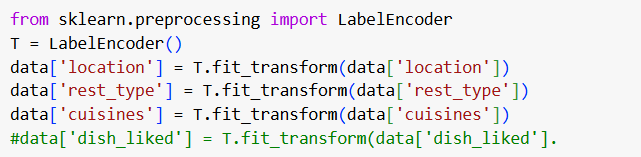


Fig – 4.1.3.2.3 : Label Encoding some of the Columns

A screenshot of a computer code

Description automatically generated

Fig – 4.1.3.2.4 : Random Forest Before Parameter Tuning

A screen shot of a computer code

Description automatically generated

Fig – 4.1.3.2.5 : Preprocessing technique in Scikit-learn



Fig – 4.1.3.2.5 : Random Forest After Parameter Tuning

**CHAPTER – 5**

**EXPERIMENTATION AND RESULT ANALYSIS**

**5. Experimentation and Result Analysis**

In order to predict restaurant ratings, the experiment applies multiple regression models on a dataset of restaurant data (the "Zomato Bangalore Restaurants" dataset).A systematic methodology comprising feature engineering, model training, data preprocessing, and hyperparameter tweaking was used to train and evaluate each model. As input features for the models, the dataset itself included details on restaurants, including names, locations, average prices, and reviews.

**Random Forest Performance:** When compared to the other models assessed in the study, the Random Forest model performed extraordinarily well, as evidenced by its remarkable 93% accuracy. With a 93% accuracy rate, the Random Forest model was able to predict restaurant ratings for a sizable chunk of the dataset that were extremely close to the real ratings. This high accuracy shows how well the model represented the connections between the target variable and the input features.

Several decision trees are constructed using the Random Forest ensemble learning technique, which then combines the results to generate predictions. The ensemble is resistant to overfitting and able to simulate non-linear interactions since each decision tree in the forest is constructed using a random selection of characteristics and training data. However, a Random Forest model's hyperparameters, which regulate things like feature selection, tree depth, and the number of trees, can have a big impact on how well the model performs.

* Decrease in Overfitting
* Increased Stability and Diversity
* Making the Best Use of Data

**Comparison between Models:**

Random Forest excels at striking a balance between ease of use and efficacy when compared to other models like XGBoost, SVM (Support Vector Machine), Decision Trees, and Linear Regression. Because of its ensemble of decision trees, Random Forest is able to simulate complicated interactions, unlike Linear Regression, which assumes a linear relationship between variables and may struggle with complex, non-linear patterns. A single decision tree has the ability to capture non-linear relationships, but it has a tendency to overfit, which results in predictions with a high variance. Random Forest reduces overfitting and produces more reliable, broadly applicable predictions by combining several trees trained on different subsets of the data and attributes.

However, Random Forest might not always achieve the same degree of accuracy and computing efficiency as more sophisticated ensemble models like XGBoost. Compared to Random Forest, XGBoost uses gradient boosting, which iteratively fixes mistakes from earlier trees and incorporates regularization strategies that better avoid overfitting. This makes it possible for XGBoost to perform better and attain more accuracy on intricate datasets, particularly when adjusted. But compared to Random Forest, which is simpler to set up and frequently works well enough with less tweaking, XGBoost usually takes longer for training and hyperparameter tuning. Because of this, Random Forest is a great option in situations where reliability and ease of use are important, whereas XGBoost might be better suited for applications requiring greater precision.

A bar graph with numbers and a number of restaurants accepting online orders

Description automatically generated

Fig – 5.1 : Number of Restaurants accepting Online Orders

In Fig 5.1 A horizontal bar chart displaying the number of eateries that take online orders is displayed in the image. On the y-axis of the chart are two bars with the labels "Yes" and "No," which indicate whether or not establishments accept online orders. Comparing restaurants that allow online orders versus those that do not, the "Yes" bar is longer. Restaurant counts are displayed on the x-axis, with numbers up to roughly 30,000. With the caption "Number of Restaurants accepting online orders," the graphic illustrates the distribution of eateries that offer this service versus those that don't.

A bar graph with a red and blue rectangle

Description automatically generated

Fig – 5.2 : Number of Restaurants with Book Table facility

In Fig 5.2 The number of restaurants that provide and do not offer table reservations is shown in the bar chart. The x-axis shows the total number of restaurants, and the y-axis displays two categories: "Yes" for eateries that accept reservations and "No" for those that don't. The chart's blue bar, which is noticeably longer than the red bar, shows that a substantially greater percentage of establishments do not allow reservations. This implies that table reservations are not very common among the eateries in the sample.

A pie chart with different colored circles

Description automatically generated

Fig – 5.3 : Type of Restaurant in Percentage

In Fig 5.3 The distribution of various restaurant kinds in percentage terms is depicted in the pie chart. With 47.7% of the total, "Quick Bites" is the largest section, followed by "Casual Dining" with 25.8%. "Cafe" (9.3%), "Delivery" (6.5%), "Dessert Parlor" (5.6%), and "Takeaway, Delivery" (5.1%) are further categories. This suggests that "Quick Bites" is the most prevalent restaurant type, accounting for about half of all establishments; other varieties have comparatively lower proportions.

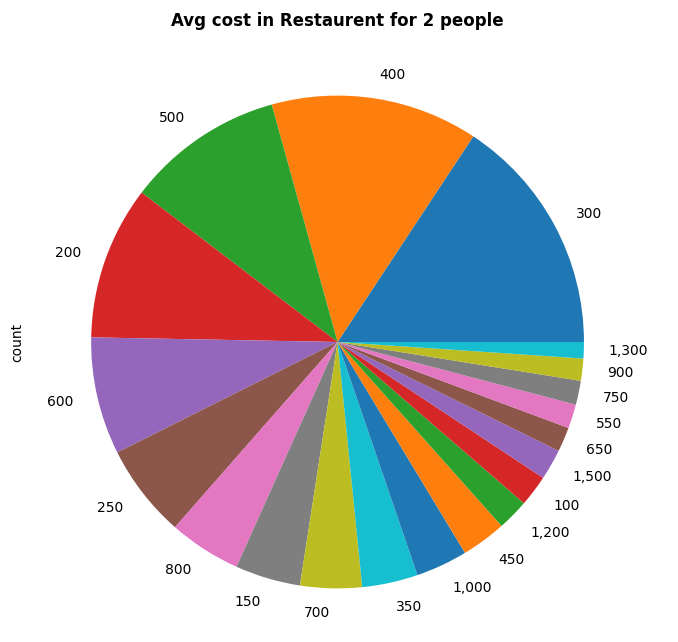


Fig – 5.4 : Average Cost in Restaurant for 2 people

In Fig 5.4 The distribution of average restaurant expenses for two persons, broken down by price range, is displayed in the pie chart. Labeled by the average cost, each slice represents a distinct price range. Certain pricing ranges—like ₹200, ₹300, and ₹400—occupy greater segments, suggesting that there are more eateries in these price categories. Other price points, such as ₹1,500 and ₹1,300, occupy smaller market niches, indicating that they are less prevalent. With a focus on mid-range prices, this chart illustrates the spectrum of restaurant prices.

A graph of a number of different colored bars

Description automatically generated with medium confidence

Fig – 5.5 : Top 10 Liked dishes in Banglore

In Fig 5.5 A bar graph of the top ten foods enjoyed in Bangalore is shown in the image; pasta has the highest number, followed by burgers and cocktails; pizza, biryani, and coffee are also highly popular; mocktails, sandwiches, paratha, and noodles round out the list. The graph vividly illustrates Bangalore's culinary preferences.

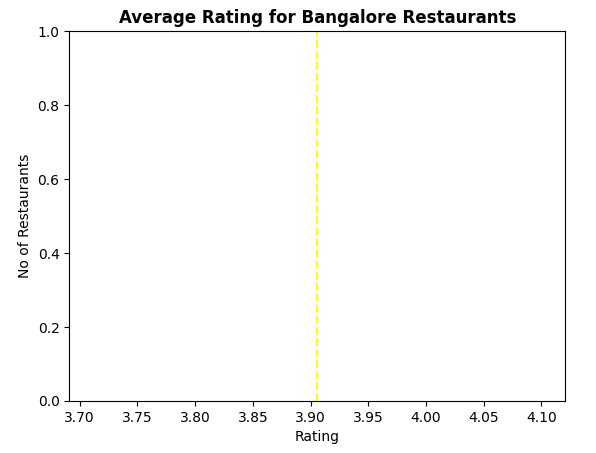


Fig – 5.6 : Average rating for Banglore restaurants

In Fig 5.6 The average rating of Bangalore restaurants is shown in the graphic as a bar graph. The rating falls between 3.7 and 4.1. Nevertheless, the graph seems to be blank, with no bars denoting a particular grade. This implies that there might not be any restaurants in the designated rating range or that the data used to build the graph may be insufficient. It's crucial to remember that this graph makes it hard to make any intelligible inferences in the absence of the actual data points.

A graph with blue lines and text

Description automatically generated

Fig – 5.7 : Restaurant Type vs Rate

In Fig 5.7 The picture shows a line graph that shows how average rating and restaurant type relate to one another. The average rating is shown on the y-axis, while the x-axis shows the different sorts of restaurants, such as cafes, casual dining, quick bites, etc. With some restaurant categories having higher average ratings than others, the graph displays a changing pattern. As an illustration, fine dining establishments typically receive higher ratings than food trucks. The graph is an illustration of how various restaurant kinds are viewed in relation to patron satisfaction.

A graph of a number of bars

Description automatically generated

Fig – 5.8 : Rate vs Online Order

In Fig 5.8 The graph investigates the connection between online order acceptance and restaurant ratings. It demonstrates that online ordering is more common at restaurants with higher ratings. Online ordering may be offered in lower-rated restaurants, but the trend indicates that higher-rated companies are more likely to provide this practical feature. Numerous reasons, including client demand, restaurant management effectiveness, and overall business strategy, may be to blame for this.

A chart with blue and yellow dots

Description automatically generated

Fig – 5.9 : True Rate vs Predicted Rate

In Fig 5.9 A scatter plot comparing actual and anticipated restaurant ratings is displayed in the image. The genuine rating is shown on the x-axis, and the anticipated rating is shown on the y-axis. Instances where the anticipated rating is greater than the actual rating—a sign of overestimation—are indicated by the yellow dots. Conversely, underestimation is suggested by the blue dots, which show cases where the expected rating is lower than the actual rating. The map indicates regions where the prediction model may be overestimating or underestimating ratings and aids in visualizing the model's accuracy.

A table with numbers and letters

Description automatically generated

Fig – 5.10 : Regression Models Results Before Tuning

In Fig 5.10 The picture displays a table that summarizes how well different regression models performed on a particular dataset. SVM, KNN, Decision Tree, AdaBoost, XGBoost, Random Forest, Ridge Regression, Lasso Regression, and Linear Regression are among the models. The R2 score, which quantifies the percentage of the dependent variable's variation that the model can account for, and the RMSE (Root Mean Square Error), which shows the average difference between expected and actual values, are both shown in the table. Decision Tree and XGBoost seem to be the best models according to these criteria; their high R2 scores and low RMSE values indicate that they can predict the target variable with accuracy.

A table with numbers and text

Description automatically generated

Fig – 5.11 : Regression Models Results After Tuning

In Fig 5.11 A table showing the performance of different regression models after reevaluating their parameters is displayed in the image. SVM, KNN, Decision Tree, AdaBoost, XGBoost, Random Forest, Ridge Regression, Lasso Regression, and Linear Regression are among the models. Each model's accuracy, RMSE, and R2 score are shown in the table. Accuracy is the percentage of accurate predictions, RMSE is the average variation between expected and actual values, and R2 score quantifies the amount of variance explained by the model. With the highest R2 score and accuracy, XGBoost seems to be the best-performing model according to these measures, indicating that it can better predict the target variable following parameter reevaluation.

**CHAPTER – 6**

**CONCLUSION AND FUTURE ENHANCEMENT**

**6.1 Conclusion**

A combination of its outstanding accuracy and capacity to manage intricate data structures, XGBoost stands out as an excellent model for restaurant rating prediction. The strength of this model lies in its use of gradient boosting, in which the errors of the previous trees are corrected by each new tree added to the ensemble. A highly optimized model with reduced prediction errors is the result of this iterative procedure. XGBoost is more resilient than standalone Decision Trees or simple ensemble models like Random Forests because it also includes sophisticated features like regularization (L1 and L2) that assist avoid overfitting. Furthermore, XGBoost is especially adaptable for large-scale and computationally demanding problems because to its parallel processing capabilities and effective handling of missing data. Its proficiency in adjusting hyperparameters, like learning rate and tree depth, enables accurate model modifications and enhances predictive performance.

Decision trees are straightforward and easy to understand, but they frequently lack the complexity required to meet XGBoost's predictive accuracy. Despite their ease of visualization and potential for modeling non-linear interactions, decision trees are prone to overfitting when used alone. The ensemble nature of XGBoost excels in this situation, fusing the advantages of several decision trees and continuously enhancing through mistake correction. New variables like seasonal patterns, customer sentiment scores, and other pertinent data could help catch more subtle trends and improve model performance even more. Furthermore, ensemble methods that integrate models (such as stacking or combining XGBoost with other algorithms) may improve prediction resilience and accuracy even more. The properties of the dataset and the project's business objectives ultimately determine which model is optimal, highlighting the significance of ongoing model evaluation and tuning to sustain peak performance.

**6.2 Future Enhancement**

Through adding more contextual and dynamic data, the model's predicted accuracy can be greatly increased by adding elements like consumer sentiment analysis from reviews and real-time variables like seasonal trends or promotional events. By identifying more intricate patterns in the data, using cutting-edge machine learning techniques—such as deep learning models or hybrid ensemble methods—could improve the model's performance even further. The model's usefulness would increase if its scope was expanded to predict ratings across various regions or cuisines. This would make the model more relevant and adaptive for a larger range of restaurant kinds and geographic areas. Furthermore, by creating an intuitive tool or API based on this improved model, restaurant owners would be equipped with data-driven, actionable insights that would help them make better decisions that would maximize customer pleasure and economic success.

**CHAPTER – 7**

**REFERENCES**

**7. REFERENCES**

[1] Somashekar, S., & Mallesh, S. (2021, December). Restaurant Rating Prediction Using Regression. In 2021 5th International Conference on Electronics, Communication and Aerospace Technology (ICECA) (pp. 1139-1144). IEEE.

[2] Priya, J. (2020, February). Predicting restaurant rating using machine learning and comparison of regression models. In 2020 International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE) (pp. 1-5). IEEE.

[3] Shihab, I. F., Oishi, M. M., Islam, S., Banik, K., & Arif, H. (2018, December). A machine learning approach to suggest ideal geographical location for new restaurant establishment. In 2018 IEEE Region 10 Humanitarian Technology Conference (R10-HTC) (pp. 1-5). IEEE.

[4] Petrusel, M. R., & Limboi, S. G. (2019, September). A restaurants recommendation system: Improving rating predictions using sentiment analysis. In 2019 21st International Symposium on Symbolic and Numeric Algorithms for Scientific Computing (SYNASC) (pp. 190-197). IEEE.

[5] Bilen, T., Erel-Özçevik, M., Yaslan, Y., & Oktug, S. F. (2018, June). A smart city application: Business location estimator using machine learning techniques. In 2018 IEEE 20th International Conference on High Performance Computing and Communications; IEEE 16th International Conference on Smart City; IEEE 4th International Conference on Data Science and Systems (HPCC/SmartCity/DSS) (pp. 1314-1321). IEEE.

[6] Khan, I. H., Khan, M. H. U., & Howlader, M. M. (2021, April). An Intelligent Approach for Food Recipe Rating Prediction Using Machine Learning. In 2021 1st International Conference on Artificial Intelligence and Data Analytics (CAIDA) (pp. 281-283). IEEE.

[7] Vaish, N., Goel, N., & Gupta, G. (2022, January). Machine learning techniques for sentiment analysis of hotel reviews. In 2022 International Conference on Computer Communication and Informatics (ICCCI) (pp. 01-07). IEEE.

[8] Bhatia, S., Arya, C., Verma, S., Gautam, D., Naib, B. B., & Kumar, A. (2023, July). Predict Success of a Zomato Restaurant using Machine Learning. In 2023 World Conference on Communication & Computing (WCONF) (pp. 1-7). IEEE.

[9] Luo, Y., & Xu, X. (2019). Predicting the helpfulness of online restaurant reviews using different machine learning algorithms: A case study of yelp. Sustainability, 11(19), 5254.

[10] Wang, X., Shen, Y., & Zhu, Y. (2018, November). A Data Driven Approach to Predicting Rating Scores for New Restaurants. In Asian Conference on Machine Learning (pp. 678-693). PMLR.

[11] Chen, Y., & Xia, F. (2020, August). Restaurants’ rating prediction using Yelp dataset. In 2020 IEEE International Conference on Advances in Electrical Engineering and Computer Applications (AEECA) (pp. 113-117). IEEE.

[12] Asghar, N. (2016). Yelp dataset challenge: Review rating prediction. arXiv preprint arXiv:1605.05362.

[13] Koetphrom, N., Charusangvittaya, P., & Sutivong, D. (2018, July). Comparing filtering techniques in restaurant recommendation system. In 2018 2nd International Conference on Engineering Innovation (ICEI) (pp. 46-51). IEEE.

[14] Saha, S., & Santra, A. K. (2017, August). Restaurant rating based on textual feedback. In 2017 International conference on Microelectronic Devices, Circuits and Systems (ICMDCS) (pp. 1-5). IEEE.

[15] Hossain, F. T., Hossain, M. I., & Nawshin, S. (2017, December). Machine learning based class level prediction of restaurant reviews. In 2017 IEEE Region 10 Humanitarian Technology Conference (R10-HTC) (pp. 420-423). IEEE.