**FINAL PROJECT REPORT**

**ON**

**TRIPADVISOR EUROPEAN RESTAURANTS ANALYSIS AND RATING PREDICTION**

Logo

Description automatically generated

Submitted by:

Group 18

Sushmitha Jogula – 001546751

Gautham Rajsimha Pulipati – 001572432

**Abstract**

This project report contains the findings and final results along with comparative analysis and performance assessment of various Supervised Machine Learning algorithms used for the prediction of the average rating value of restaurants from Tripadvisor across Europe.

The dataset [Tripadvisor European Restaurants](https://www.kaggle.com/stefanoleone992/tripadvisor-european-restaurants) used in this project has been taken from Kaggle. The data is then preprocessed, analyzed, and cleaned according to the requirements based on the initial Exploratory Data Analysis performed on the dataset. Then, several Machine Learning models are built in Jupyter environment and Anaconda 3 to examine and assess the performance of these models on certain specific parameters of the dataset.

Various machine learning algorithms and models such as Linear Regression, Decision Tree Regressor, Support Vector Regressor, Random Forest Regressor, and Gradient Boosting Regressor are implemented and researched as part of this study. There are 42 features in this dataset, making it very apt and suitable for solving it using the machine learning approach.

**Table of Contents**

1. Introduction
   * Background
   * Motivation
   * Goal
2. Methodology
   * Data set and data source
   * Dataframe description
   * Dataframe information
   * Shape and data types
   * Numeric and non-numeric columns
   * Software and Libraries used
   * Data Cleaning
   * Feature Selection
   * Target Selection
   * Exploratory Data Analysis
   * Models used
     + Linear Regression
     + Decision Tree Regressor
     + Linear Support Vector Regressor
     + Random Forest Regressor
     + Gradient Boosting Regressor

* Feature Selection
* Visualization

1. Initial Data Analysis
2. Exploratory Data Analysis
3. Data Preprocessing and Cleaning
4. Restaurant Rating Prediction
5. Comparative Analysis and Performance Measurement
6. Conclusion
7. References

**Introduction**

(Done by Sushmitha Jogula)

**Background**

From planning and booking to taking a trip, Tripadvisor is the world's largest travel website and guidance platform, assisting hundreds of millions of people in becoming better travelers by making trip planning easy, thanks to more than 988 million reviews and opinions from almost 8 million companies. The Tripadvisor European Restaurants dataset consists of information related to all the restaurants in the main European countries along with their attributes and features such as average rating, cuisine types, awards, popularity ranking, location information, number of reviews, open working hours etc. Although there are around 1 million restaurants all over Europe , the food industry business is still far from saturating, with new restaurants of different categories emerging every day across Europe. However, it is tough for these new restaurants to compete with well-established establishments due to several reasons and factors such as increasing food costs, high real-estate expenses, a fragmenting supply chain system, supply shortage of qualified labor etc. With an everlasting demand for restaurants in the economic and financial sector, it becomes more crucial than ever to research several factors about a restaurant such as its location demographics, price range, working hours etc. to assess its performance rating. This project can reveal what customers consider most important in a restaurant, generate insights on what combinations of features one should adopt when launching a new restaurant and how likely it can succeed.

**Motivation**

Being avid travelers and frequent users of the Tripadvisor website, we thought this would be the apt dataset to explore for our project and to understand various factors of restaurants that contribute in making a restaurant successful and appreciated by the users, while performing a comparative analysis of common features of restaurants across several European countries such as average rating, open hours, awards, locations etc. Every company wants to know if it will be successful or not, with the restaurant rating being one of the most essential factors in deciding that. It not only indicates the quality and services provided by that restaurant, but it also aids in attracting new customers. Also, it is imperative to consider some factors here such as an area's demographics, degree of influence of people that live in the area, restaurant's theme. This project aims to provide answers to these considerations as well as predict changes in restaurant ratings and popularity using the following machine learning algorithms:

* Linear Regression
* Decision Tree Regression
* Support Vector Regression
* Random Forest Regression
* Gradient Boosting Regression

**Goal**

Our goal is to analyze and predict the performance and success of a restaurant based on various factors such as cuisine types, location, price range, rating for food, service, and value, average rating, popularity index, awards etc. by applying several machine learning models.

* Using machine learning models to predict the rating value of the restaurant
* Compare the predicted rating value with the true rating value and identify the most efficient algorithm with the highest accuracy rate for our dataset
* Visualize and analyze the performance and accuracy of all the applied models

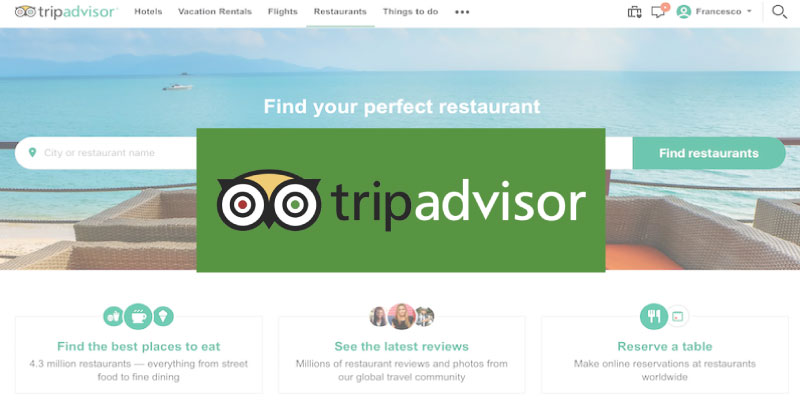
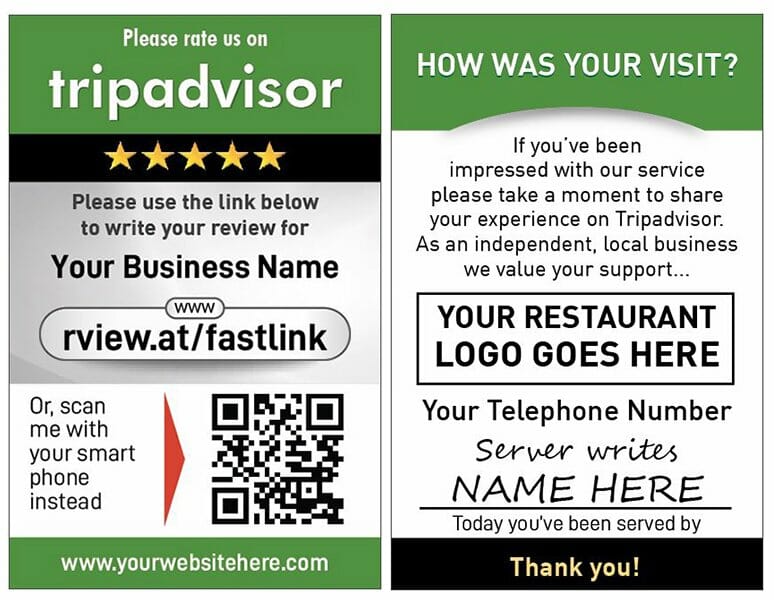
**Methodology**

(Done by Sushmitha Jogula)

**Data Set and Data Source**

The [Tripadvisor European Restaurants](https://www.kaggle.com/stefanoleone992/tripadvisor-european-restaurants) dataset has 1083397 unique restaurant links with 42 columns. The size of data is 679.68 MB. This dataset has been taken from Kaggle. The data is retrieved from the publicly available [Tripadvisor](https://tripadvisor.com/) website by scraping the data in early May 2021.

Below are the details of this dataset:

|  |  |
| --- | --- |
| **Column Name** | **Column Description** |
| restaurant\_link | Unique TripAdvisor Link |
| restaurant\_name | Restaurant Name on TripAdvisor |
| original\_location | Original Location displayed on TripAdvisor |
| country | Country name retrieved from original\_location |
| region | Region name retrieved from original\_location |
| province | Province name retrieved from original\_location |
| city | City name retrieved from original\_location |
| address | Address displayed on TripAdvisor |
| latitude | Latitude coordinate |
| longitude | Longitude Coordinate |
| claimed | Restauarnt business claimed on TripAdvisor |
| awards | Award Names |
| popularity\_detailed | Popularity detailed ranking |
| popularity\_generic | Popularity generic ranking (among all places to eat in the area) |
| top\_tags | Top tag names |
| price\_level | Level of price in current currency |
| price\_range | Range of price in current currency |
| meals | Types of meals |
| cuisines | Types of Cusines |
| special\_diets | Types of special diets |
| features | Restaurant features |
| vegetarian\_friendly | Is the restaurant vegetarian friendly? (Yes or no) |
| vegan\_options | Does the restaurant have vegan options? (Yes or no) |
| gluten\_free | Does the restaurant have gluten free options? (Yes or no) |
| original\_open\_hours | Original open hours on trip advisor |
| open\_days\_per\_week | Number of open days per week retrieved from original\_open\_hours |
| open\_hours\_per\_week | Number of hours per week retrieved from original\_open\_hours |
| working\_shifts\_per\_week | Number of working shifts per week retrieved from original\_open\_hours |
| avg\_rating | Average restaurant rating |
| total\_reviews\_count | Total count of the reviews |
| default\_language | Default language displayed while scraping |
| reviews\_count\_in\_default\_language | Total reviews count in default language |
| excellent | Excellent reviews count in default language |
| very\_good | Very\_good reviews count in default language |
| average | Average reviews count in default language |
| poor | Poor reviews count in default language |
| terrible | Terrible reviews count in default language |
| food | Food rating |
| service | Service rating |
| value | Value rating |
| atmosphere | Atmosphere rating |
| keywords | Popular keywords |

**Dataframe Description**

**Table

Description automatically generated**

**DataFrame Information**

Table

Description automatically generated

**Shape and data types**

**Table

Description automatically generated**

**Numeric and non-numeric columns**

**Text

Description automatically generated**

**Software and Libraries used**

The [Tripadvisor European Restaurants](https://www.kaggle.com/stefanoleone992/tripadvisor-european-restaurants) dataset is downloaded from Kaggle. Libraries used:

* Numpy
* Pandas
* Matplotlib
* Matplotlib.pyplot
* Seaborn
* Sys
* Collections
* Pandas\_profiling
* Sklearn.preprocessing
* Sklearn.feature\_selection
* Sklearn.model\_selection
* Sklearn.metrics
* Sklearn.linear\_model
* Sklearn.tree
* Sklearn.ensemble
* Sklearn.neighbors
* Sklearn.svm
* Missingno
* Sklearn.feature\_extraction.text
* Mean Squared Error
* Scikitlearn Tree
* DecisionTreeRegressor
* train\_test\_split
* Standard Scaler
* r2\_score
* accuracy\_score
* RandomForestRegressor
* LabelEncoder

**Data Cleaning**

(Done by Sushmitha Jogula)

* Perform data wrangling and data preprocessing to clean the raw data after it is read and before training the data
* Identify any missing or inconsistent values in the target columns or feature matrix and handle it accordingly
  + Either discard the column from the prediction process
  + Or fill with mean values

The dataset contained much sparse data with many missing values for many feature columns. Almost 9% of data was missing for target column avg\_rating. There were many columns with more than 50% of data missing such as awards, price range, features, keywords etc. Therefore, many rows had NAN and ‘’ values which can affect the performance of the models, so to sort it, most of the columns with more than 70% of missing data were dropped and null rows in some other columns were dropped.

Text

Description automatically generated

**Feature Selection**

(Done by Sushmitha Jogula)

#### The dataset originally had 42 columns with several columns such as awards, price range, features, keyword having more than 70% missing values. There are some columns with up to 50% of missing data such as meals, original\_open\_hours, open\_days\_per\_week etc. The rest of the columns had comparatively less proportion of missing data such as region, top\_tags, cuisines, default\_language etc. There were some columns which had no missing data at all such as restaurant\_link, restaurant\_name, country, claimed, address etc.

**Target Selection**

(Done by Sushmitha Jogula)

#### The target variable here is “avg\_rating” column as we are trying to predict the column using the values of other features in the dataset.

**Exploratory Data Analysis**

(Done by Sushmitha Jogula)

After performing initial data cleaning and visualizing missing values using various Pandas and Matplotlib commands. Then, we have performed Exploratory Data Analysis (EDA) on the entire raw dataset to analyze and generate insights, and visualize the fundamental comprehensive attributes of the selected data set. This process is essential to make conclusions about the predictions made by the models. We visualized missing values in the dataset using several plot formats such as heatmap, bar plot, matrix from the Missingno library. Then, we converted the categorical features into numeric values for statistical computation. Then, we added new measures such as Number of Awards, Average Price, Restaurant Category, Number of Features created using manipulating the existing columns. Then, we have plotted several graphs analyzing and plotting the relationship between the features and visualized the results. At the end, we have plotted a heatmap showing the correlation between all the features of the dataset.

#### **Models Used**

(Done by Gautham Rajsimha)

The Dependent Target Variable is avg\_rating which contains float digits from 1 to 5, which means, **1** being the least rating value given by a customer in their review and **5** being the highest rating value given by a customer in their review based on their dining, service, value, and atmosphere experience in the restaurant.

The prediction model used here is Regression as the target variable has many classes where the analysis is a form of predictive modelling technique which investigates the relationship between a dependent (target) and independent variables (predictor).

The dataset with all the features (independent variables) is trained on various regression models to predict the class value of the dependent target variable.

Regression models used in the project are:

* Linear Regression
* Decision Tree Regresison
* Random Forest Regression
* Support Vector Regression
* Gradient Boosting Regression

1. **Linear Regression**

Linear Regression analysis is used to predict the value of a variable based on the value of another variable. The variable you want to predict is called the dependent variable. The variable you are using to predict the other variable's value is called the independent variable. Linear Regression was not much success with the accuracy of the prediction of the target value of our dataset. The TripAdvisor Restaurants dataset has numerous columns which can be used to predict the target (Y), which here is the average rating of the restaurant and we go with the hypothesis that the Linear Regression model will train to generate coefficients that would predict the average rating of the restaurant.

1. **Decision Tree Regression**

Decision Trees are a non-parametric supervised learning method used for both classification and regression tasks. The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features. A decision tree is represented as upside down where its root is at the top of the tree then it splits into branches and when it cannot further split then the end branch is called as decision. We use Grid Search for hyperparameter tuning on the given dataset which evaluates the input vector with every combination to give the best result. We can generate the best parameters of the 100 input parameters and predict using this to improve our accuracy.

1. **Support Vector Regression**

Support Vector Regression method uses the Support Vector Machine(SVM, a classification algorithm) algorithm to predict a continuous variable. While other linear regression models try to minimize the error between the predicted and the actual value, Support Vector Regression tries to fit the best line within a predefined or threshold error value.

As Linear regression is a simple model, we applied Support Vector Regressor on the model to get a decision boundary in order to improve the accuracy

Chart, line chart

Description automatically generated

The model is expected to fit similar to linear regression but better performance is expected as we now have a decision boundary.

The simple equation looks like: -a < Y- mX + c < +a

Where X is the input feature matrix, Y is the output target vector, and a is the measure (distance) if boundary plane

1. **Random Forest Regressor**

Random forest is one of the most popular algorithms for regression problems (i.e. predicting continuous outcomes) because of its simplicity and high accuracy. In decision tree only one tree is made but in random forest, our algorithm randomly creates a specified number of decision trees. And chooses tree which is best for our model. It worked the best and retured the best results out of the other results. For the given dataset, we have numerous features, and a random forest can be generated by constructing multitude of decision trees while training the model and this gives us the mean prediction of the individual trees. So, this works as an improvement to Decision Trees by aggregating all of them and ensures the ensemble model does not rely heavily on any individual feature. The randomness also prevents overfitting and is expected to give perfect results.

1. **Gradient Boosting Regression**

Gradient Boosting can be used to produce a predictive model from an ensemble of weak predictive models. Gradient boosting can be used for regression and classification problems. The accuracy score of gradient Bossting was 51% was the least among all the regression methods. Like how assumed our dataset to fit decision trees, we expect this model to train with numerous input features of the data and result a better performing model compared to decision trees.

We also tuned the hyperparameters of this Gradient Boosting Regressor using grid search to select the best parameters and predict the test dataset using this as this is expected to perform better.

**Feature Selection**

We selected features based on their degree of contribution to our output, thus decreasing the computing time of our model.

**Visualization**

We have created visualizations and graphs of different types for both Exploratory Data Analysis (EDA) and implementation of machine learning models. We have used the libraries Matplotlib, Seaborn and Pandas Profile Reporting to build our visualizations.

**Initial Data Analysis**

(Done by Gautham Rajsimha)

**Data Loading**

After the dataset csv file is downloaded to the local system and after importing all the required libraries, we are loading the dataset into a Pandas dataframe directly without specifying the column names as they are already provided in the dataset csv file.

**Table

Description automatically generated**

**Pandas Profile Reporting**

(Done by Sushmitha Jogula)

We have executed the Pandas Profile Reporting module to generate a HTML profile report from the Pandas restaurants\_df dataframe for a quick data analysis.

**Text

Description automatically generated with medium confidence**

**Visualizing missing values in the dataset**

(Done by Gautham Rajsimha)

We have used the Missingno library to create visualizations for indicating the missing values in the dataset of different plot kinds such as heatmap, matrix, and bar plot.

**Heatmap:**

Chart

Description automatically generated

Chart

Description automatically generated

**Bar Plot:**

(Done by Gautham Rajsimha)

Chart, surface chart

Description automatically generated with medium confidence

**Matrix:**

(Done by Gautham Rajsimha)

A picture containing diagram

Description automatically generated

From the above plots, we were able to categorise the columns based on their proportion of missing values:

|  |  |  |
| --- | --- | --- |
| **More than 70% missing values** | **40%-70% missing values** | **Less than 30% missing values** |
| awards | meals | region |
| Price\_Range | Special\_diets | province |
| features | Original\_open\_hours | city |
| atmosphere | Open\_days\_per\_week | Popularity\_detailed |
| keywords | Open\_hours\_per\_week | Popularity\_generic |
|  | Working\_shifts\_per\_week | Top\_tags |
|  | food | cuisines |
|  | service | Avg\_rating |
|  | value | Total\_reviews\_count |

**These columns have no missing values in them:** restaurant\_link, restaurant\_name, original\_location, country, address, latitude, longitude, claimed, vegetarian\_friendly, gluten\_free, vegan\_options

**New measures added using existing features**

(Done by Gautham Rajsimha)

1. **Number of Awards**

Graphical user interface, text, application, email

Description automatically generated

1. **Average Price**

Text

Description automatically generated

1. **Restaurant Category**

Text

Description automatically generated

Correlation matrix is formed with dataset features.

A picture containing text

Description automatically generated

We can observe from the heatmap that some factors are positively correlated while others are negatively correlated. High positive correlation could indicate redundant data. As a result, one of the features may be dropped. At this point, we're primarily interested in the correlations between our target column and the other features. The more a feature is associated to the target (positively or negatively), the more likely it is to aid in target prediction. And from the correlation matrix, we can see that the 2 top positively correlated features to the target are ‘Number of Awards’ and ‘Total\_Reviews\_Count.’ This implies that more the number of awards received by the restaurant and more the number of reviews for the restaurant, the avg\_rating value increase.

**Exploratory Data Analysis**

(Done by Sushmitha Jogula, all of it)

1. What are the most popular cities in Europe for restaurant business?

**Chart

Description automatically generated**

We can see that Paris and Rome are leading with most number of restaurants across Europe. Hence, the competition in these cities amongst restaurants will be very high.

### What are the most favorite cuisines liked by Europeans?

Chart, bar chart

Description automatically generated

We can see that Italian, European, and Mediterranean are the top 3 most favorite cuisines amongst the people of Europe. So, we can conclude that a restaurant with any of these cuisines is more likely to succeed.

1. What are the top 10 provinces with most spending in restaurants?

Chart, waterfall chart

Description automatically generated

We can see that the Province of Frosinone and North Ayrshire single-handedly dominate the entire restaurant business amongst all the provinces with more spending and hence, we can conclude that a restaurant in these provinces can price their items at a higher level without significantly affecting the customer experience and potentially generate more revenues.

1. Regions with highest number of active restaurant customers

Chart, line chart

Description automatically generated

We can see that the regions of Lazio, Lombardy, and Tuscany have the customers actively giving most number of rating values in their reviews. Hence, we can conclude that the customers in these regions are highly inclined towards visisting restaurants and are foodies. Restaurants in these areas can see highest customer engagement rate with the restaurant experience.

1. What are the top provinces with most awarded restaurants?

Chart

Description automatically generated

In the graph, we can see that Italy is the destination for highly awarded and prestigious restaurants all across Europe followed by England. So, in these provinces, it might become difficult for new restaurants to match up to the reputation created by well-established restaurants.

1. Does restaurant category affect its rating?

Chart, line chart

Description automatically generated

From this graph, we can see that the best ratings are for Fine Dining restaurants closely followed by Deli restaurants. It is to be noted here that Quick Bites restaurants have the least rating values in this case, so it can be deduced that Fine Dining restaurants are more popular and Quick Bites restaurants are least popular amongst Europeans.

1. Restaurant Category Distribution Analysis

Chart, bar chart

Description automatically generated

From this graph, we can see that Mid-range restaurants are dominating the restaurant business scene in all the European countries followed by Cheap Eats restaurants. So, it might be profitable to establish restaurants of these categories.

1. How do the features claimed\_by\_tripadvisor, vegetarian\_friendly, gluten\_free, vegan\_options affects the restaurant rating value?

Chart, pie chart

Description automatically generated Chart, pie chart

Description automatically generated

Chart, pie chart

Description automatically generated Chart, pie chart

Description automatically generated

Chart, box and whisker chart

Description automatically generated

From the above graph, it can be seen that all the restaurants having the 4 features have avg\_rating value ranging from 3.5-5, falling into the best category whereas restaurants not having these features have their avg\_rating value starting from 1, falling into bad category. Hence, we can conclude that having these features positively impacts the rating value.

1. What are the popular meal types in European restaurants?

Chart, bar chart, funnel chart

Description automatically generated

From the graph, we can see that dinner and lunch are the predominant meal types across European restaurants. So, a restaurant can consider including these meal types in their menus to attract more customers and generate more revenue.

1. Do customers prefer food quality or quantity?

Chart, bar chart

Description automatically generated

From this comparison, we can see that McDonalds is one of biggest restaurant chain across Europe with around 4500 McDonalds restaurants across Europe. But, McDonalds has the worst rating value with more than 2000 user reviews giving it a rating of less than 2.5. Hence, we can conclude that customers prefer quality over quantity, and having more number of restaurants doesn’t guarantee good ratings and customer experience; the food quality, and service, value, and atmosphere ratings play a significant role here.

1. How does number of working days affect the number of reviews received by restaurant?

Chart

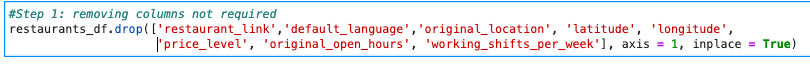
Description automatically generated

From the graph, we can clearly see that number of reviews is directly proportional to the number of days the restaurant is open in a week. That is, a restaurant that is open all the days in a week receives more number of customers, hence more number of reviews, when compared to restaurants which are not open all days in a week.

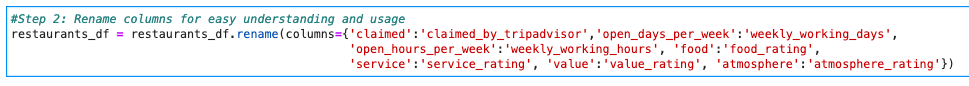
**Data Preprocessing**

(Done by Gautham Rajsimha)

1. Dropping the columns not required.



1. Renaming columns for ease of understanding.



1. Cleaning data

Text, letter

Description automatically generated

1. Changing categorical features for statistical computation



**Data Cleaning**

(Done by Gautham Rajsimha)

1. Converting Nan values to fit in Count Vectorizer

Text

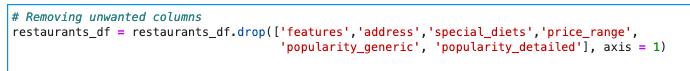
Description automatically generated

1. Converting text into dummies for cuisines, top tags, meals, and category

Table

Description automatically generated

1. Removing unwanted columns



1. Removing dummy columns



1. Dropping null rows and duplicate rows

Graphical user interface

Description automatically generated with low confidence

**Best features to apply in the models for prediction:**

Graphical user interface, text

Description automatically generated

Text

Description automatically generated

**Restaurant Rating Prediction**

(Done by Sushmitha Jogula)

**Training and Testing Data**

Data is separated in Training set as well as Testing set and then machine learning models are applied on this training data and testing data.



**Machine Learning Models**

1. **Linear Regression**

**Graphical user interface, text, application, email

Description automatically generated**

(Done by Gautham Rajsimha)

**Result :**The accuracy score of the Linear Regression is 31% and not the accurate model for the dataset.

**Text

Description automatically generated**

1. **Decision Tree**

**Text

Description automatically generated**

(Done by Gautham Rajsimha)

**Graphical user interface, text, application, email

Description automatically generated**

**Result :**The accuracy score of the Decision Tree is 60% and this is not the accurate model for the dataset.

**Graphical user interface, text, application

Description automatically generated**

1. **Support Vector Regression**

**Graphical user interface, text, application, email

Description automatically generated**

(Done by Gautham Rajsimha)

**Result :** The accuracy score of the Support Vector Regression is 32% and this is not the accurate model for the dataset.

Graphical user interface, text, application, email

Description automatically generated

1. **Random Forest Regression**

**Graphical user interface, text, application, email

Description automatically generated**

(Done by Sushmitha Jogula)

**Result :**The accuracy score of the Random Forest Regression is 81% and as of now, this is the best performing model for the dataset.

**Text

Description automatically generated**

1. **Gradient Boosting Regressor**

**Graphical user interface, text, application, email

Description automatically generated**

(Done by Sushmitha Jogula)

**Result :**The accuracy score of the Gradient Boosting Regressor is 71% and this is a good performing model for the dataset.

Text

Description automatically generated

1. **Gradient Boosting with GridSearch**

**Graphical user interface, text, application, email

Description automatically generated**

(Done by Gautham Rajsimha)

**Result :**The accuracy score of the Gradient Boosting Regressor with GridSearch is 77% which is a significant improvement over traditional Gradient Boosting Regressor method and this is the second best performing model for the dataset.

**Graphical user interface, text, application, email

Description automatically generated**

**Comparative Analysis and Performance Measurement**

(Done by Sushmitha Jogula)

**Text

Description automatically generated with medium confidence**

**Chart, bar chart

Description automatically generated**

Here, from the above graph, we can see that Random Forest Regression gives the best accuracy rate for the dataset with R2 score of 81%, followed by Gradient Boosting Regressor with GridSearch with an accuracy rate of 77%. The worst performing model in this case is Linear Regression with an accuracy rate of 31% which does not provide an accurate fit for the given dataset.

The better model can be choosed based on how close the mean square error (MSE) value is to 0. In the above applied machine learning models, it is evident that Random Forest Regressor has the lowest mean squared error and its close to 0 than any other models.

**Actual Average Rating value prediction after applying Random Forest Regression:**

**Text

Description automatically generated**

**Chart, bar chart

Description automatically generated**

**Conclusion**

(Done by Sushmitha Jogula & Gautham Rajsimha)

Here, after applying all the models, we can see that Random Forest Regression gives the best accuracy rate for the dataset with R2 score of 81%, followed by Gradient Boosting Regressor with GridSearch with an accuracy rate of 77%. The worst performing model in this case is Linear Regression with an accuracy rate of 31% which does not provide an accurate fit for the given dataset.

The better model can be choosed based on how close the mean square error (MSE) value is to 0. In the above applied machine learning models, it is evident that Random Forest Regressor has the lowest mean squared error and its close to 0 than any other models.

Insights that can be generated from this study are:

* Paris and Rome are the top 2 most popular cities in Europe for restaurant business.
* The five most favorite cuisines liked by Europeans are Italian, European, Mediterranean, Pizza, and Café.
* The Province of Frosinone and North Ayrshire single-handedly dominate the entire restaurant business amongst all the provinces with more spending.
* The regions of Lazio, Lombardy, and Tuscany have the customers actively giving most number of rating values in their reviews.
* Italy is the destination for highly awarded and prestigious restaurants all across Europe followed by England.
* The best ratings are for Fine Dining restaurants closely followed by Deli restaurants. It is to be noted here that Quick Bites restaurants have the least rating values.
* Mid-range restaurants are dominating the restaurant business scene in all the European countries followed by Cheap Eats restaurants.
* Dinner and lunch are the predominant meal types across European restaurants.
* Number of reviews is directly proportional to the number of days the restaurant is open in a week. That is, a restaurant that is open all the days in a week receives more number of customers, hence more number of reviews, when compared to restaurants which are not open all days in a week.

These are the above points to be considered before establishing any restaurant in Europe.

This study's purpose is to predict the average rating of restaurants in Europe recorded by Tripadvisor. We experimented with various machine learning models and methods. The Random Forest Regressor is the best model we found out of all the models we have implemented. This is a fascinating opportunity for business regions to assess restaurant features and customer preferences in Europe. With this knowledge, it is possible to examine the attributes of a restaurant before its establishment or during the initial stages of its establishment itself, and this information is highly significant for both businessmen and customers.

**References**

* <https://www.kaggle.com/stefanoleone992/tripadvisor-european-restaurants>
* <https://towardsdatascience.com/exploratory-data-analysis-with-pandas-profiling-de3aae2ddff3>
* <https://www.ibm.com/topics/linear-regression>
* [Decision Tree Tutorials & Notes | Machine Learning | HackerEarth](https://www.hackerearth.com/practice/machine-learning/machine-learning-algorithms/ml-decision-tree/tutorial/)
* [The Ultimate Guide to Random Forest Regression (keboola.com)](https://www.keboola.com/blog/random-forest-regression)
* [Support Vector Regression Made Easy(with Python Code) | Machine Learning | Artificial Intelligence Online Course (aionlinecourse.com)](https://www.aionlinecourse.com/tutorial/machine-learning/support-vector-regression)
* [Gradient Boosting regression — scikit-learn 1.0.1 documentation](https://scikit-learn.org/stable/auto_examples/ensemble/plot_gradient_boosting_regression.html)
* <https://machinelearningmastery.com/linear-regression-for-machine-learning/>
* <https://www.nbshare.io/notebook/312837011/Decision-Tree-Regression-With-Hyper-Parameter-Tuning-In-Python/>
* https://www.analyticsvidhya.com/blog/2020/03/support-vector-regression-tutorial-for-machine-learning/
* <https://medium.com/swlh/random-forest-and-its-implementation-71824ced454f>