**Project Report: Zomato Bangalore**

## **Introduction**

**Background**

The dining industry has significantly evolved due to digital platforms which have changed consumer preferences and the global expansion of culinary cultures. Utilizing Zomato, a leading restaurant discovery and review platform, we have access to a rich dataset that captures various factors related to dining establishments worldwide. The use of data analytics allows a nuanced understanding of factors such as location, cuisine, ratings, and reviews correlating with restaurant popularity and success, shifting away from traditional metrics like word-of-mouth and critical reviews in print media.

**Motivation**

The drive behind our project is the potential of data analytics to enhance the dining experience, support restaurant owners with informed decisions, and predict future dining trends. By dissecting the Zomato dataset, our aim is to identify key factors contributing to a restaurant's success, understand global dining trends, and provide valuable insights into customer preferences.

**Goals**

Our objectives are:

Identify patterns and trends in customer dining preferences.

Determine key factors influencing restaurant ratings and success.

Forecast emerging culinary trends on a global and regional scale.

Provide actionable insights to restaurant owners for strategic decision-making.

## **Methodology**

Our methodology for analyzing the Zomato dataset involved several steps:

We used the below machine learning algorithms for ***classification***:

Logistic Regression:

Logistic regression was chosen for its effectiveness in making predictions where the outcomes can be split into two clear categories. In our project analyzing Zomato deliveries, we specifically applied logistic regression to determine whether a restaurant falls into the 'expensive' category or not. This method is not only straightforward but also excels in handling binary classification tasks, making it a natural fit for answering questions like this where there are only two possible outcomes.

Decision Tree:

We opted for a decision tree due to its straightforward approach to decision-making. In our Zomato project, it helped us classify restaurants as 'expensive' or 'not expensive' by breaking down the decision process into simple, easy-to-follow steps. This method excels in handling categorical data and visually represents decision paths, making it clear why each classification is made.

Random Forest:

Random forest was chosen for its enhanced accuracy and robustness. By using multiple decision trees, this method strengthens predictions and reduces the risk of overfitting. In our analysis, it helped ensure that our classifications of restaurant costs were reliable, drawing on a wider range of data insights than a single tree could offer. This ensemble approach aggregates predictions to provide a well-rounded decision.

We used the below machine learning algorithms for ***regression analysis****:*

Linear Regression:

We started with linear regression because of its straightforwardness and clarity. It's a basic model that assumes a direct relationship between the features (like location, cuisine type) and the restaurant ratings. It acts as our starting point, helping us see how different variables might linearly affect a restaurant's rating.

Ridge Regression:

We picked Ridge regression because it's really good at dealing with situations where the characteristics of restaurants, like their location or type of cuisine, might be very similar to each other. This similarity can make it tricky for basic models to predict ratings accurately because they might overly focus on these similar traits. By using cross-validation, we make sure our Ridge regression model is really well-tuned. Cross-validation works by testing the model's accuracy on different parts of our data, not just one single part.

Lasso Regression:

Lastly, we used Lasso regression because it's excellent for refining our model. Lasso helps by potentially eliminating less important features entirely, simplifying the model by keeping only those variables that have a substantial impact on the ratings.

## **Data preprocessing and cleaning**

Ensuring data accuracy and usability by removing irrelevant or duplicate data, addressing missing values, and standardizing formats.

* Removing ‘/5’ from the ratings column
* Removing comma ‘,’ from the cost\_for\_two columns
* Renaming certain columns for better understanding
* Dropping the unnecessary columns

## **Exploratory Data Analysis (EDA)**

Using visualization and statistical techniques to explore the dataset, identify trends in restaurant popularity, cuisine preferences, and the impact of location on success.

* Showing the number of restaurants that accept table booking, online ordering
* Most liked cuisines
* Number of restaurants that take online ordering in each area
* Comparing the distribution of votes, ratings, cost for two between restaurants that take online ordering and the ones that do not.

1. % of restaurants taking online orders.

A screenshot of a pie chart

Description automatically generated

1. % of restaurants taking table booking facilities

A purple circle with black text

Description automatically generated

1. Most common rating:

A graph of a rating

Description automatically generated with medium confidence

1. Which cuisines do customers like the most?

A screenshot of a computer

Description automatically generated

1. Top 10 common restaurant types.

A screenshot of a computer

Description automatically generated

1. Area vs count of restaurants taking online orders

A screenshot of a computer

Description automatically generated

1. Average price for two people based on service type

A screen shot of a graph

Description automatically generated

1. Distribution of rate/vote/cost for online accepting & non-accepting restaurants

A screenshot of a graph

Description automatically generated

A screenshot of a computer screen

Description automatically generated

1. Histogram plotting

A screenshot of a computer screen

Description automatically generated

## **Feature engineering**

Selecting and engineering features indicative of restaurant success, informed by domain knowledge and statistical analysis.

* Created a new feature ‘expensive’ that tells if a restaurant is considered expensive or not based on the cost\_for\_two column

A screenshot of a computer

Description automatically generated

## **Sentiment Analysis**

Based on the reviews\_list for each restaurant, added a column ‘sentiment’ which states if the overall reviews are positive/negative.

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

**Model Selection and Evaluation:** Applying machine learning models such as Logistic regression, Random Forest, Decision Trees and Linear regression, Ridge regression and Lasso regression to predict restaurant’s rating and to consider if a restaurant is expensive or not. Model performance was evaluated to ensure reliability and accuracy.

## **Results and Analysis**

Our analysis covered various models and methods:

**Classificaton Models- If a restaurant is expensive or not**

**A screenshot of a computer program

Description automatically generated**

**A screenshot of a computer program

Description automatically generated**

Used Logistic regression, Decision tree classifier and Random Forest for this, while the Logistic regression gave the highest accuracy, Decision tree classifier gave the lowest accuracy.

Hence, tuned the hyper-parameters of the decision tree classifier model to check if it will give a better accuracy, and it did. The accuracy drastically changed from from 84% to 95%.

**Regression Models -To predict the Ratings**

**A screenshot of a computer program

Description automatically generated**

Used Linear, Ridge and Lasso regression models to predict the ratings.

Ridge regression model gave the least RMSE indicating that this model fits the data well, while the Lasso regression model gave the highest RMSE.

So, tuned the hyper-parameters of Lasso regression model and it yielded better RMSE results.RMSE has decreased from 0.479 to 0.296.

## **Conclusion**

The project successfully utilized data analytics to identify critical factors impacting restaurant success, highlighted through the analysis of the Zomato dataset. Our findings can help restaurant owners make more informed decisions and anticipate future trends in the culinary world.

In short,

Classification Models: Logistic regression performed the best in determining if a restaurant is expensive. Despite initially poor performance, tuning the decision tree classifier greatly improved its accuracy from 84% to 95%.

Regression Models: Ridge regression was the most effective for predicting restaurant ratings, as indicated by its lowest RMSE. The Lasso regression model initially had the highest RMSE, but after hyper-parameter tuning, its performance improved significantly, reducing the RMSE from 0.479 to 0.296.

These findings suggest that both model choice and hyper-parameter tuning are crucial in optimizing model performance for specific tasks in the dataset.

## **References**

* Zomato Dataset: <https://www.kaggle.com/datasets/rishikeshkonapure/zomato/data>
* Professor Liu’s teaching slides
* For ML algorithms: <https://scikit-learn.org/stable/>
* For Sentiment Analysis, referred to another project: https://www.analyticsvidhya.com/blog/2021/06/nlp-sentiment-analysis/