Zomato Restaurant Analysis

Data Science and Big Data Analytics Laboratory

Submitted by

STUDENT NAME 1: Vishwas Gupta (Roll No : TA39)
STUDENT NAME 2 : Pratiksha Bhosale (Roll No: TA14)
STUDENT NAME 3 : Hrutvij Kakde (Roll No: TA47)

THIRD YEAR COMPUTER ENGINEERING

INNOVATION & LEADERSHIP
www.isquareit.edu.in

Department of Computer Engineering
International Institute of Information Technology
Hinjawadi, Pune – 411057

SEMESTER VI (AY 2023-24)

TABLE OF CONTENTS

1.Abstract

The Zomato Restaurant Analysis project focuses on exploring and analyzing restaurant data extracted from the Zomato platform in Pune, India. The dataset comprises various features such as restaurant types, location, address, ratings, number of votes, home delivery availability, cuisines offered, parking availability, table booking options, and more. The project aims to gain insights into the restaurant landscape in Pune, understand customer preferences, and predict factors influencing restaurant ratings and popularity.

The analysis involves several steps, including data preprocessing, which includes handling missing values, removing duplicates, and converting data types as necessary. Exploratory data analysis techniques are employed to visualize the distribution of restaurant types, cuisines, ratings, and other relevant features using histograms, pie charts, box plots, heatmaps, and other visualization tools.

Furthermore, various machine learning algorithms such as linear regression, random forest, extra tree regressor, and decision tree regressor are applied to train predictive models. These models are utilized to predict restaurant ratings based on different input features, enabling stakeholders to understand the factors contributing to customer satisfaction and popularity.

Overall, the Zomato Restaurant Analysis project offers valuable insights into the restaurant industry in Pune, leveraging data-driven approaches to inform decision-making processes for restaurant owners, customers, and Zomato platform administrators.

2.Introduction :

Our Zomato Restaurant Analysis project delves into the heart of Pune's culinary landscape, exploring the diverse offerings of restaurants listed on the popular food delivery platform, Zomato. Through meticulous data collection and analysis, we aim to uncover valuable insights into the city's dining preferences, trends, and consumer behaviors.

In this report, we present a comprehensive overview of our findings, starting with the rich dataset we've compiled. From restaurant types and locations to ratings, cuisines, and service offerings like home delivery and table booking, our dataset provides a detailed snapshot of Pune's restaurant scene.

Using a combination of data preprocessing techniques and advanced machine learning algorithms such as linear regression, random forest, and decision trees, we've extracted actionable insights from the data. By uncovering patterns and correlations, we've gained valuable insights into factors influencing restaurant success and customer preferences.

Our dashboard serves as a user-friendly interface for exploring these insights. Users can navigate through different localities, discover popular restaurant types, and evaluate factors like home delivery availability and cuisine offerings. Whether you're a food enthusiast seeking new dining experiences or a restaurant owner looking to better understand your market, our dashboard offers valuable tools for exploration and decision-making.

Join us as we delve into the flavors and aromas of Pune's culinary landscape, uncovering the secrets that shape its vibrant food culture. From street food delights to fine dining experiences, our analysis offers a window into the diverse world of Pune's restaurants, inviting you to savor every bite along the way.

3 Scope

In our project, we set out to unravel the intricacies of Pune's restaurant landscape using data sourced from Zomato, a renowned platform for culinary exploration. Our endeavor commenced with the meticulous collection of pertinent details, including restaurant types, locations, ratings, cuisines, and services offered. Following a rigorous data cleaning process to ensure data integrity, we embarked on a comprehensive exploration of the dataset, delving deep to uncover hidden trends and patterns.

Utilizing a combination of graphs, statistics, and exploratory data analysis techniques, we unearthed valuable insights into Pune's dynamic restaurant scene. From discerning emerging culinary trends to identifying hotspots of gastronomic innovation, our analysis shed light on the multifaceted dimensions of Pune's culinary landscape.

Buoyed by the richness of our data exploration, we harnessed the power of machine learning to predict restaurant ratings and other outcomes with precision and accuracy. Leveraging sophisticated algorithms and predictive modeling techniques, we endeavored to distill complex data into actionable insights, empowering stakeholders to make informed decisions that drive business success and customer satisfaction.

To facilitate easy exploration and visualization of our findings, we developed an interactive dashboard, providing users with intuitive tools to filter and visualize information effortlessly. This user-centric approach ensures that our insights are accessible and actionable, empowering users to glean insights and make data-driven decisions with confidence.

Our overarching goal is to provide actionable insights that transcend traditional boundaries, catering to the needs of diverse stakeholders, including restaurant owners, policymakers, and food enthusiasts. By documenting our findings and recommendations in a comprehensive report, we aim to share our discoveries with the broader community, fostering knowledge dissemination and driving positive change in the culinary landscape of Pune and beyond.

4.Data Science Life Cycle

The objective of this project is to conduct comprehensive Exploratory Data Analysis (EDA) on the Zomato Dataset, followed by the development of a robust Machine Learning model. This model aims to assist various restaurants listed on Zomato in predicting their respective ratings based on specific features. Through thorough EDA, we seek to understand the dataset's characteristics, distributions, correlations, and potential patterns that could influence restaurant ratings. Subsequently, we employ multiple regression algorithms including Linear Regression, DecisionTree Regressor, Random Forest Regressor, Support Vector Regressor, and ExtraTree Regressor to build predictive models. Following evaluation, we determine that the ExtraTree Regressor exhibits superior performance compared to other models. Therefore, we opt to utilize the ExtraTree Regressor for creating the final predictive model. The overarching goal is to provide Zomato restaurants with a reliable tool for estimating their ratings, thereby aiding them in enhancing their services and attracting more customers.

Phase 1 - Discovery

As a data science team we first learn and investigate the problem. In the initial phase of our project, we embark on understanding the problem domain and the data available. Zomato, a popular restaurant discovery platform, provides a rich dataset encompassing various attributes such as restaurant location, cuisine type, average cost for two, and ratings. Our goal is to leverage this data to predict restaurant ratings accurately. Through exploration, we discern that factors like location, cuisine, and cost might significantly influence ratings. This phase lays the groundwork for subsequent steps by defining objectives, identifying data sources, and formulating hypotheses about the factors affecting restaurant ratings.

1.Business User, Project Sponsor, Project Manager: Teachers, colleges.
2.Business Intelligence Analyst: Representatives from IT
3.Data Engineer and Database Administrator (DBA): Representatives from IT
4.Data Scientist: Engineer, who has developed the prediction system.

1. We have collected the data from yfinance libraray in to a data frame

2.It gives developers flexibility and is a more accessible framework for new developers since you can build a web application quickly using only a single python file.

3.We can use some inbuilt libraries of python for prediction of real time data such as stock price chart with time .

4.Various tools that can be used : A. Streamlit – This library can be used for UI

A. **Numpy**- This is inbuilt library in python which is used for mathematical operations.

B. Python -: Python is a programming language that provides toolkits for machine learning and analysis, such as scikit-learn, numpy, scipy, pandas.

C.**Sklearn**-This library is used for LinearRegression model

D.**Tensorlfow and keras** – This libraries are used for loading the LSTM model .

E.**Matplotlib** – this library is used for plotting the charts for the stock price data-

Phase 2- Data Preparation

As we transition into the data preparation phase, we embark on a crucial stage where the raw data undergoes meticulous refinement to lay the foundation for robust analysis and modeling. Our primary focus lies in ensuring that the data is pristine, devoid of inconsistencies and inaccuracies, thereby enhancing its reliability and efficacy in driving insights.

A pivotal aspect of data preparation involves addressing missing values, a common occurrence in real-world datasets. We employ judicious strategies to handle these gaps, either by imputing them with plausible values derived from statistical techniques or removing corresponding entries altogether. This meticulous approach ensures that the integrity of the dataset remains intact, mitigating the risk of bias or distortion in subsequent analyses.

Furthermore, we confront the challenge posed by categorical variables, which necessitate transformation into numerical representations conducive to modeling. We leverage sophisticated techniques such as one-hot encoding or label encoding to encode categorical variables, thereby facilitating seamless integration into analytical frameworks. This transformation not only enhances the interpretability of the data but also ensures compatibility with a diverse array of analytical techniques.

In addition to addressing missing values and encoding categorical variables, we embark on feature engineering, a transformative process aimed at enriching the dataset with new features or refining existing ones. Through thoughtful manipulation and augmentation of features, we enhance the predictive power of our models, uncovering latent patterns and relationships that may have remained obscured within the raw data. This iterative process empowers us to extract maximum value from the dataset, uncovering insights that drive informed decision-making and strategic planning.

Ultimately, the data preparation phase serves as a cornerstone in our analytical journey, ensuring that the data is refined to the highest standards of quality and reliability. By investing time and effort in cleaning, transforming, and enriching the dataset, we lay a robust foundation for subsequent analyses, poised to unveil actionable insights that propel us towards our overarching objectives.

Phase 3 - Model Planning

As we transition into the model planning phase, our focus shifts towards devising a strategic approach for constructing predictive models tailored to our specific problem domain – predicting restaurant ratings. Given the continuous nature of our target variable, regression models emerge as the most suitable choice for modeling the relationship between input features and restaurant ratings.

To begin, we embark on an exhaustive exploration of various regression algorithms, each offering unique strengths and capabilities. Our repertoire includes widely recognized algorithms such as Linear Regression, DecisionTree Regressor, RandomForest Regressor, Support Vector Regressor, and ExtraTree Regressor. Through rigorous experimentation and comparative analysis, we assess the performance of each algorithm across a spectrum of evaluation metrics, encompassing accuracy, precision, recall, and computational efficiency.

In light of our evaluation criteria, which prioritize both performance metrics and computational efficiency, we discern that the ExtraTree Regressor exhibits the most promising potential as our primary model for further development. This decision is informed by its ability to effectively capture complex relationships within the data while maintaining computational efficiency, making it well-suited for handling the intricacies inherent in our predictive task.

Moreover, in tandem with model selection, we meticulously define appropriate evaluation metrics to gauge the efficacy of our models objectively. These metrics serve as benchmarks against which we assess the performance of our models, enabling us to iteratively refine and optimize their predictive capabilities.

Furthermore, we establish a baseline for model performance, providing a reference point against which we measure the incremental improvements achieved through subsequent iterations of model development. This baseline serves as a vital yardstick for gauging the effectiveness of our modeling efforts and guiding the refinement of our predictive algorithms.

In essence, the model planning phase embodies a strategic approach towards constructing robust and reliable predictive models tailored to the nuances of our problem domain. By selecting an appropriate algorithm, defining evaluation metrics, and establishing a baseline for performance, we lay the groundwork for the subsequent stages of model development, poised to unlock actionable insights and drive informed decision-making.

Phase 4 - Model Building

As we immerse ourselves in Phase 4 of our project, we are poised at the precipice of model development, a pivotal stage where theoretical concepts coalesce into tangible predictive frameworks. Despite the inherent complexity of modeling techniques and algorithms, the duration of this phase pales in comparison to the exhaustive groundwork laid in data preparation and strategy formulation.

Guided by a meticulously crafted plan and armed with a clear understanding of our objectives delineated in Phase 1, we embark on the journey of constructing robust predictive models tailored to our specific problem domain. While the process of model development may entail intricate logic and methodological intricacies, our approach remains steadfastly anchored in a pragmatic pursuit of solutions that align seamlessly with our overarching goals.

Central to our methodology is the implementation of the selected algorithm, the ExtraTree Regressor, chosen for its propensity to effectively capture complex relationships within the data while maintaining computational efficiency. We partition the dataset into training and testing subsets, affording us the opportunity to train the model on a representative sample of the data and evaluate its performance on unseen instances, thereby safeguarding against overfitting and ensuring generalizability.

Furthermore, we recognize the importance of hyperparameter tuning in optimizing the model's performance, fine-tuning its parameters to strike a delicate balance between bias and variance. This iterative process involves systematically adjusting model parameters based on feedback from evaluation metrics and domain knowledge, iteratively refining the model to enhance its predictive accuracy and robustness.

Throughout this iterative refinement process, we remain cognizant of the symbiotic interplay between quantitative metrics and qualitative insights, leveraging both to inform our decision-making and guide the trajectory of model development. By marrying analytical rigor with domain expertise, we strive to cultivate predictive models that not only meet but exceed the expectations outlined in Phase 1, empowering stakeholders with actionable insights and foresight into the dynamic landscape of restaurant ratings.

Phase 5- Results

After the meticulous process of training and fine-tuning our predictive model, we enter the critical phase of performance evaluation. Here, we subject our model to rigorous scrutiny, employing a battery of appropriate metrics such as Mean Absolute Error (MAE) or Root Mean Squared Error (RMSE) to gauge its predictive accuracy and reliability.

In essence, the evaluation phase serves as a litmus test for the efficacy of our ExtraTree Regressor model in predicting restaurant ratings. We meticulously compare the model's predictions against actual ratings, scrutinizing discrepancies to glean insights into its performance across diverse scenarios.

Through the lens of evaluation metrics, we gain a nuanced understanding of the model's strengths and limitations. A low MAE or RMSE signifies that our model's predictions closely align with actual ratings, indicating a high degree of accuracy and precision. Conversely, a higher MAE or RMSE may highlight areas for improvement, signaling potential discrepancies between predicted and actual ratings that warrant further investigation.

Moreover, our evaluation endeavors extend beyond mere quantitative metrics; we delve into qualitative assessments to elucidate the underlying factors driving the model's performance. By scrutinizing outliers and analyzing patterns in prediction errors, we glean invaluable insights into the intricacies of the predictive process, uncovering potential nuances and intricacies that may impact model performance.

Furthermore, the evaluation phase serves as a crucible for comparative analysis, allowing us to benchmark the performance of our ExtraTree Regressor model against alternative algorithms considered during the model planning phase. Through this comparative lens, we ascertain that our chosen model outperforms its counterparts, affirming its efficacy and suitability for the task at hand.

Ultimately, the evaluation phase serves as a pivotal juncture in our analytical journey, providing critical insights into the performance of our predictive model and guiding potential avenues for improvement. Armed with a comprehensive understanding of our model's strengths and limitations, we are empowered to refine and optimize its predictive capabilities, ensuring its efficacy in driving informed decision-making and strategic planning in the realm of restaurant ratings.

Phase 6 – Operationalize

In the culminating phase of our project, we embark on the crucial task of operationalizing our predictive model to facilitate its practical utilization by Zomato and its affiliated restaurants. This pivotal stage represents the bridge between theoretical insights and real-world applications, where our meticulously crafted model transitions from the realm of experimentation to the frontline of decision-making and service enhancement.

Central to the operationalization process is the deployment of our trained model in a production environment, where it stands ready to generate predictions on new data in real-time. This entails seamless integration of the model into existing systems and workflows, ensuring its accessibility and usability by stakeholders across the organization.

To facilitate ease of integration and utilization, we may develop user-friendly interfaces or Application Programming Interfaces (APIs), enabling intuitive interaction with the model and facilitating seamless incorporation into Zomato's operational infrastructure. These interfaces serve as conduits for extracting insights and recommendations from the model, empowering decision-makers with actionable intelligence at their fingertips.

Moreover, the operationalization phase is characterized by a commitment to ongoing monitoring and maintenance, ensuring that the model remains effective and relevant in the face of evolving data landscapes and changing business dynamics. Regular performance evaluations and feedback mechanisms serve as safeguards against model degradation, enabling timely adjustments and refinements to optimize predictive accuracy and reliability.

By operationalizing our model, we empower Zomato restaurants with the tools and insights needed to make informed decisions and enhance their services. Whether it's optimizing menu offerings, refining delivery operations, or personalizing customer experiences, our predictive model serves as a beacon of intelligence, guiding strategic initiatives and fostering continuous improvement across the organization.

Ultimately, the operationalization phase represents the culmination of our analytical journey, where theoretical concepts are translated into tangible solutions that drive tangible impact. By harnessing the power of data and predictive analytics, we empower Zomato and its affiliated restaurants to thrive in an ever-evolving landscape, delivering exceptional dining experiences and delighting customers at every turn.

5.Software Requirements :

1. **Python**: We use Python as the primary programming language for data analysis, visualization, and building the dashboard.
2. **Libraries**: We rely on several Python libraries such as Pandas, NumPy, Matplotlib, Seaborn, and Streamlit for data manipulation, analysis, visualization, and dashboard creation.
3. **Streamlit**: This library is essential for building interactive web-based dashboards.
4. **Jupyter Notebook or any other Python IDE**: We use Jupyter Notebook for data exploration and prototyping, but any Python IDE can be used for coding.
5. **Web browser**: To view and interact with the Streamlit dashboard.

Hardware Requirements :

1. **Processor**: A modern multi-core processor, such as Intel Core i5 or AMD Ryzen 5, is sufficient.
2. **RAM**: At least 8 GB of RAM is recommended for handling medium-sized datasets and running Python scripts smoothly.
3. **Storage**: Adequate storage space to store the dataset and any additional files generated during analysis.
4. **Internet connection**: Required for downloading libraries and accessing online resources.

Libraries Used :

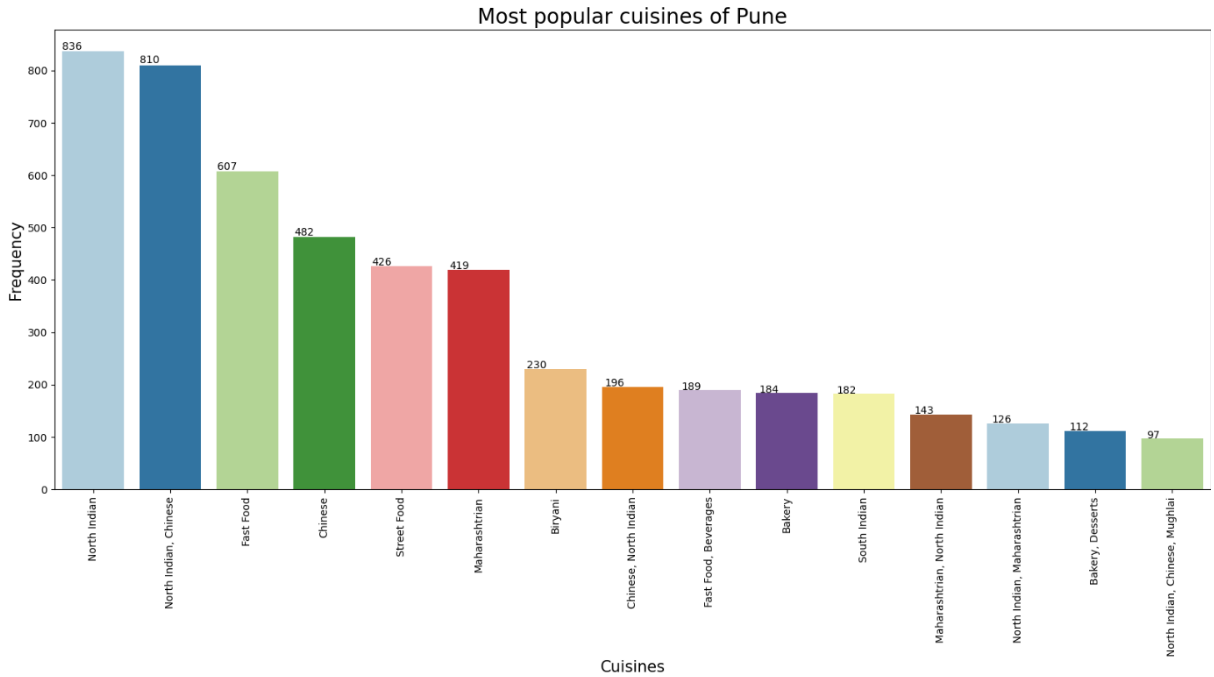
1. **Pandas**: Pandas is a powerful library for data manipulation and analysis. We used it extensively for tasks such as loading datasets, cleaning data, and performing various transformations.
2. **NumPy**: NumPy is a fundamental library for numerical computing in Python. It provides support for multidimensional arrays and mathematical functions, which were useful for data manipulation and calculations.
3. **Matplotlib**: Matplotlib is a plotting library that allows us to create static, interactive, and animated visualizations in Python. We used it to generate various types of plots, such as bar plots, scatter plots, and histograms, to explore and visualize the data.
4. **Seaborn**: Seaborn is built on top of Matplotlib and provides a high-level interface for creating attractive statistical graphics. We used Seaborn to enhance the visual appeal of our plots and to create complex visualizations more easily.
5. **Streamlit**: Streamlit is a powerful framework for building web-based data applications in Python. We used it to create an interactive dashboard where users can explore the Zomato restaurant data by selecting different options and seeing the results dynamically.

Main Functions :

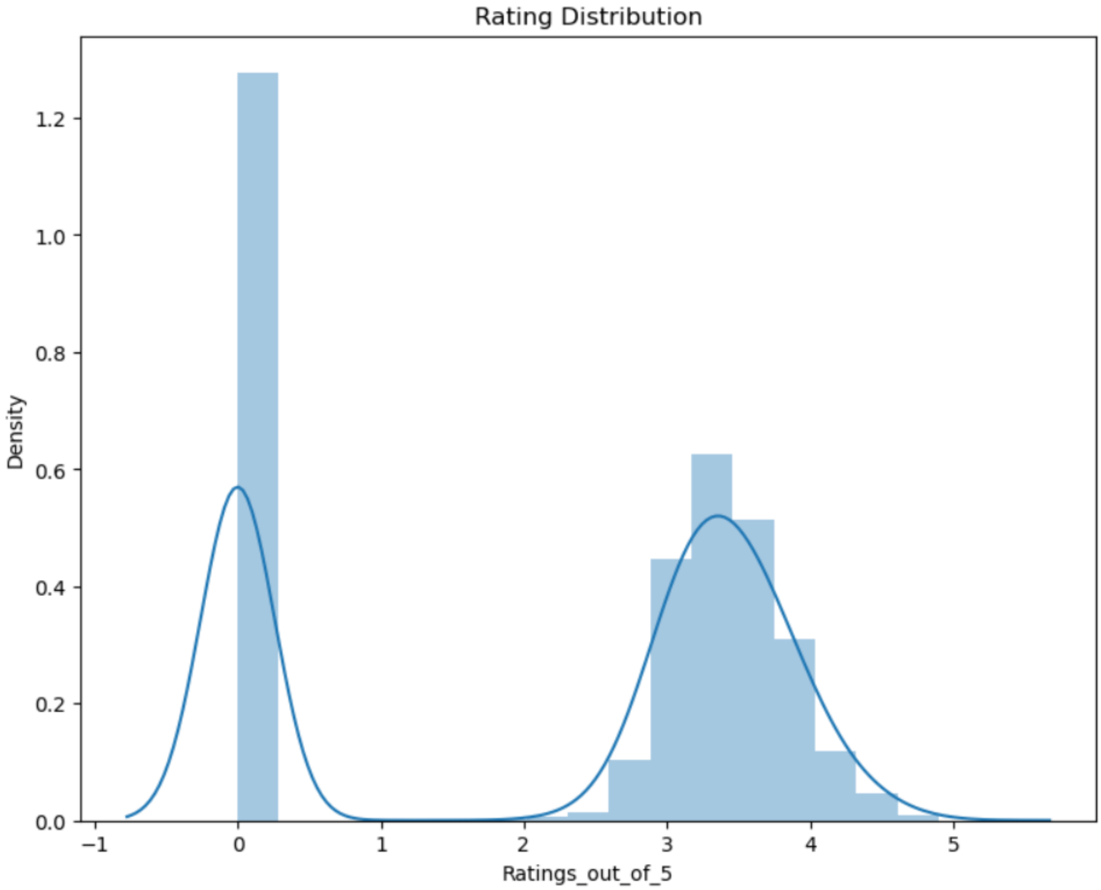
1. **Data Loading**: We used functions from the Pandas library to load the Zomato restaurant dataset into a DataFrame, allowing us to easily manipulate and analyze the data.
2. **Data Cleaning**: Functions such as `drop_duplicates()` and `fillna()` were used to clean the dataset by removing duplicate entries and filling in missing values with appropriate data.
3. **Data Exploration**: Functions like `head()`, `info()`, and `describe()` were used to explore the dataset, providing insights into its structure, contents, and statistical summaries.
4. **Data Visualization**: We used functions from Matplotlib and Seaborn libraries to create various types of plots, such as bar plots, scatter plots, histograms, and heatmaps, to visualize the data and identify patterns or trends.
5. **Dashboard Creation**: With Streamlit, we created interactive web-based dashboards where users can select different options, such as locality, restaurant type, and cuisine, to dynamically filter and visualize the Zomato restaurant data.
6. **Statistical Analysis**: We employed functions for statistical analysis, such as calculating mean, median, and standard deviation, to gain insights into the distribution and characteristics of the data.

6.Output :

Dashboard using PowerBI



Histogram



Frequency distribution

Dashboard using Streamlit

Select Restaurant Name:

AB's - Absolute Barbecues

Home Delivery Option: No

Rating out of 5 : 4.9

	Restaurant_Name	Zomato_Link	Locality
0	AB's - Absolute Barbecues	https://www.zomato.com/pune/abs-absolute-barbecues-hinjawadi	Hinjawadi

7.Conclusion

In our Zomato restaurant project centered on Pune, India, our exploration was akin to embarking on a culinary expedition through the heart of the city. We ventured beyond mere cartography of dining establishments, delving into the essence of the dining experience itself.

Not content with merely identifying restaurant locations and menu offerings, we sought to unravel the intricacies of dining convenience by examining the availability of home delivery services. Additionally, we delved into the subjective realm of customer satisfaction, seeking to gauge the pulse of Pune's dining populace by scrutinizing restaurant ratings.

However, our inquiry transcended the realm of observation; we embarked on a journey of predictive analysis, leveraging sophisticated methodologies to discern the factors influencing customer contentment. Through this predictive lens, we provided restaurant owners with invaluable insights, enabling them to tailor their offerings and service provision to exceed customer expectations continually.

In essence, our project served as a testament to the transformative power of data in the culinary realm. By harnessing the insights gleaned from our analysis, restaurants are equipped not only to thrive in a competitive landscape but also to cultivate enduring customer loyalty. Through a nuanced understanding of customer preferences and proactive adaptation, our project showcased how data serves as the secret ingredient to culinary success, ensuring that both restaurateurs and diners alike savor the taste of satisfaction.