LABMENTIX

ZOMATO RESTAURANT SENTIMENT ANALYSIS & BUSINESS INSIGHTS

A Data Science & Machine Learning Project Report

NAME : VIJAY PULLABHOTLA

1. Introduction

Online food delivery platforms generate massive volumes of customer-generated content in the form of reviews, ratings, and engagement metadata. These reviews represent a valuable but underutilized source of business intelligence. This project leverages Zomato restaurant and review data to perform sentiment analysis, exploratory data analysis, statistical testing, and machine learning modeling to extract actionable insights. The project bridges customer experience analysis with business optimization by identifying best-performing restaurants, pricing inefficiencies, sentiment drivers, and influential reviewers.

2. Problem Statement & Objectives

The primary objective of this project is to analyze customer sentiment from restaurant reviews and derive insights that benefit both customers and businesses. Specifically, the project aims to:

1. Analyze customer sentiments from textual reviews using NLP techniques.
2. Segment restaurants based on ratings, sentiment consistency, cost, and popularity.
3. Help customers identify the best restaurants based on data-driven metrics.
4. Identify cost vs benefit inefficiencies for restaurants.
5. Detect influential reviewers (critics) who significantly impact public perception.
6. Build and evaluate machine learning models to classify sentiment accurately.
7. Ensure the solution is deployment-ready and reproducible.

3. Dataset Description

Two datasets were used:

3.1 Restaurant Metadata Dataset

- **Rows:** 105 restaurants
- **Columns:** Name, Links, Cost, Collections, Cuisines, Timings
- **Insights:**
    a. Wide diversity in cuisines (92 unique cuisines)
    b. Cost ranged across 29 distinct categories
    c. Collections column had significant missing values, indicating inconsistent tagging

3.2 Restaurant Reviews Dataset

- **Rows:** 10,000 reviews
- **Columns:** Restaurant, Reviewer, Review Text, Rating, Metadata, Time, Pictures
- **Insights:**
    a. Ratings ranged from 1 to 5 (10 discrete values)
    b. Reviewer influence varied significantly (7,446 unique reviewers)
    c. Textual data exhibited high variance in length and sentiment expression

First View: Zomato Restaurant Names and Metadata Dataset

	Name	Links	Cost	Collections	Cuisines	Timings
0	Beyond Flavours	https://www.zomato.com/hyderabad/beyond-flavou...	800	Food Hygiene Rated Restaurants in Hyderabad, C...	Chinese, Continental, Kebab, European, South I...	12noon to 3:30pm, 6:30pm to 11:30pm (Mon-Sun)
1	Paradise	https://www.zomato.com/hyderabad/paradise-gach...	800	Hyderabad's Hottest	Biryani, North Indian, Chinese	11 AM to 11 PM
2	Flechazo	https://www.zomato.com/hyderabad/flechazo-gach...	1,300	Great Buffets, Hyderabad's Hottest	Asian, Mediterranean, North Indian, Desserts	11:30 AM to 4:30 PM, 6:30 PM to 11 PM
3	Shah Ghouse Hotel & Restaurant	https://www.zomato.com/hyderabad/shah-ghouse-h...	800	Late Night Restaurants	Biryani, North Indian, Chinese, Seafood, Bever...	12 Noon to 2 AM
4	Over The Moon Brew Company	https://www.zomato.com/hyderabad/over-the-moon...	1,200	Best Bars & Pubs, Food Hygiene Rated Restauran...	Asian, Continental, North Indian, Chinese, Med...	12noon to 11pm (Mon, Tue, Wed, Thu, Sun), 12no...

==========

First View: Zomato Restaurant Reviews Dataset

	Restaurant	Reviewer	Review	Rating	Metadata	Time	Pictures
0	Beyond Flavours	Rusha Chakraborty	The ambience was good, food was quite good . h...	5	1 Review , 2 Followers	5/25/2019 15:54	0
1	Beyond Flavours	Anusha Tirumalaneedi	Ambience is too good for a pleasant evening. S...	5	3 Reviews , 2 Followers	5/25/2019 14:20	0
2	Beyond Flavours	Ashok Shekhawat	A must try.. great food great ambience. Thnx f...	5	2 Reviews , 3 Followers	5/24/2019 22:54	0
3	Beyond Flavours	Swapnil Sarkar	Soumen das and Arun was a great guy. Only beca...	5	1 Review , 1 Follower	5/24/2019 22:11	0
4	Beyond Flavours	Dileep	Food is good.we ordered Kodi drumsticks and ba...	5	3 Reviews , 2 Followers	5/24/2019 21:37	0

4. Data Cleaning & Wrangling

Data preprocessing included:

1. Removal of duplicate reviews (36 duplicates removed)

2. Imputation of missing values using:

    a. Mode for categorical columns
    b. Median-based strategies for numerical features

3. Conversion of ratings from object to numerical scale
   Feature derivation including:

    a. Review length (number of characters)
    b. Review posting hour
    c. Sentiment score and sentiment label
    d. Reviewer follower count

Post-wrangling, the cleaned datasets contained **9,954 reviews** and were fully analysis-ready.

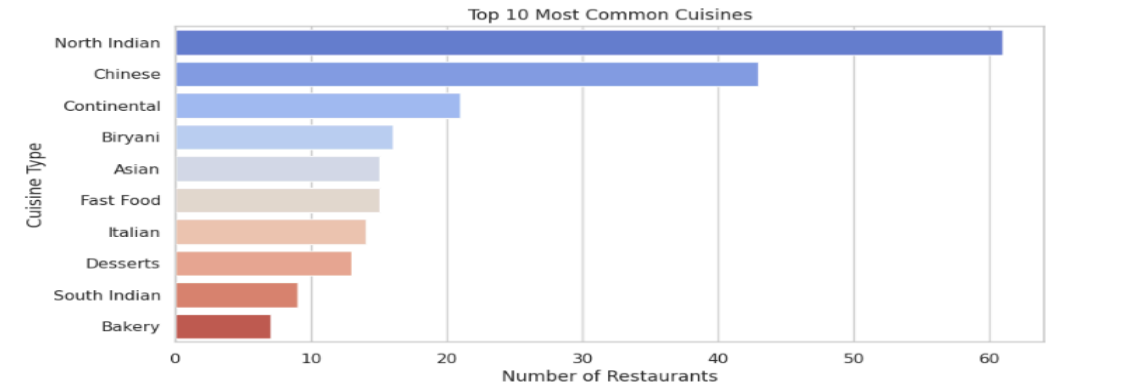
5. Exploratory Data Analysis & Visualization

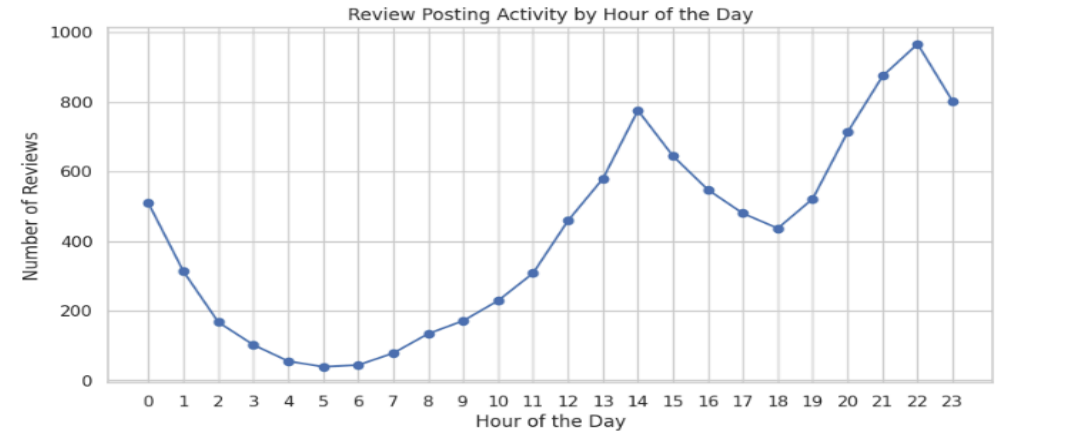
A structured visualization strategy following the **UBM framework** was adopted.

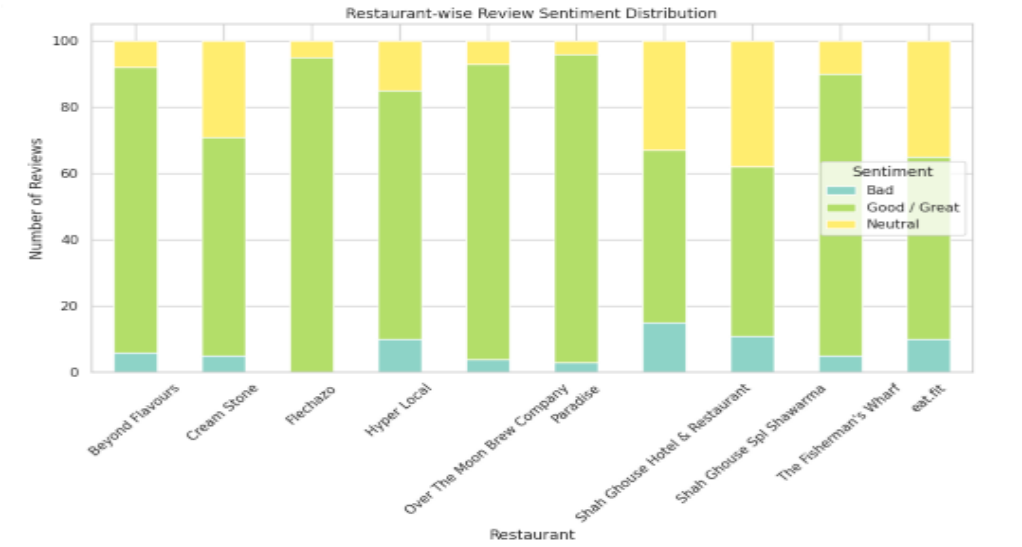
Key Findings:

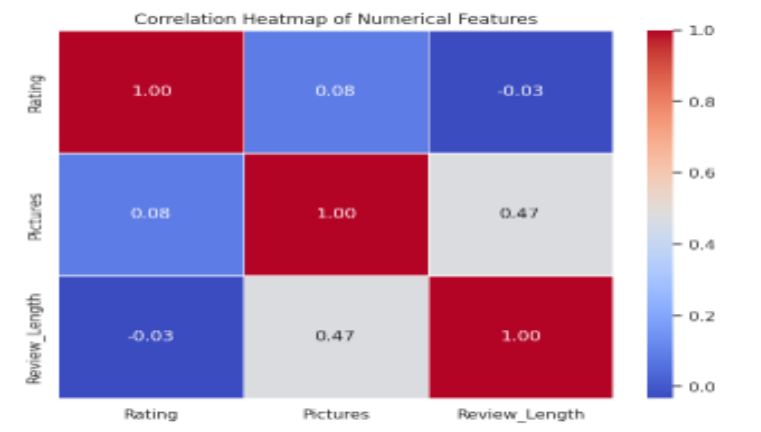
1. Rating distribution was positively skewed, with most reviews rated **4 or 5**.
2. North Indian and Chinese cuisines dominated in volume, while Mediterranean and European cuisines had higher average ratings.
3. Longer reviews did not necessarily correlate with lower ratings, disproving common assumptions.
4. Review posting peaked during evening dining hours, indicating engagement patterns.

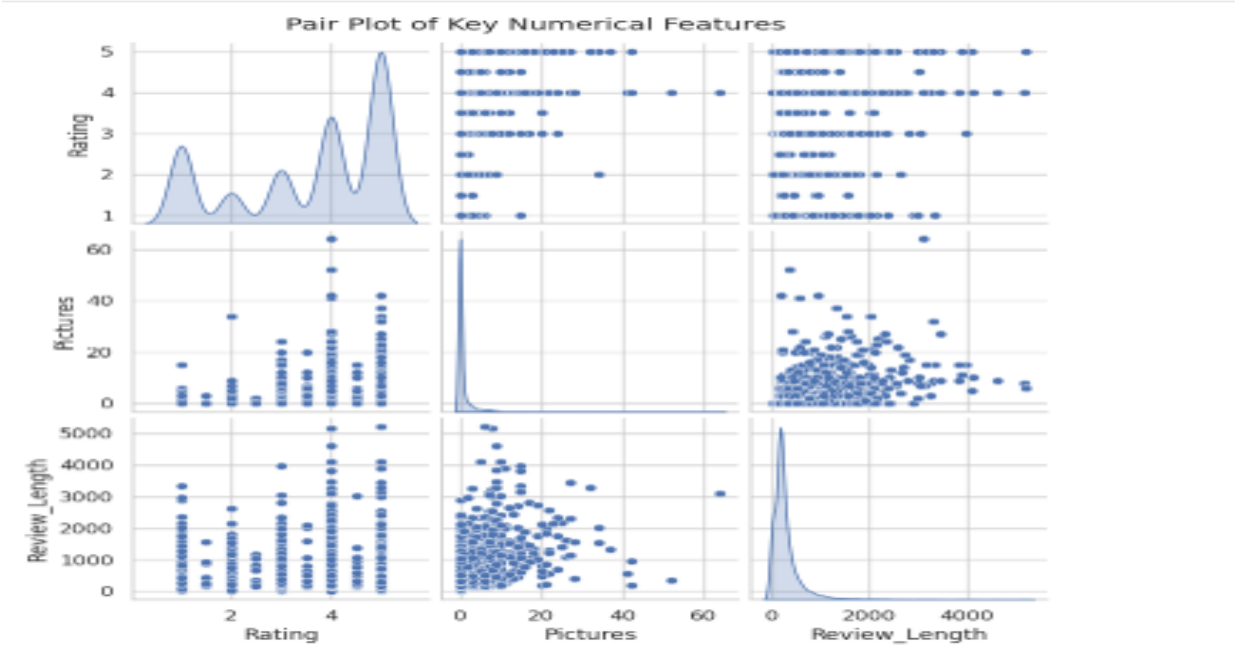
Advanced plots such as violin plots, stacked bar charts, bubble plots, correlation heatmaps, and pair plots were used to uncover multivariate relationships.




Review Posting Activity by Hour of the Day
Number of Reviews
Hour of the Day
0
200
400
600
800
1000
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23


Restaurant-wise Review Sentiment Distribution
Number of Reviews
Restaurant
Sentiment
Bad
Good / Great
Neutral
Beyond Flavours
Cream Stone
Flechazo
Hyper Local
Over The Moon Brew Company
Paradise
Shah Ghouse Hotel & Restaurant
Shah Ghouse Spl Shawarma
The Fisherman's Wharf
eat.fit


Correlation Heatmap of Numerical Features
Rating
Pictures
Review_Length
1.00
0.08
-0.03
0.08
1.00
0.47
-0.03
0.47
1.00
1.0
0.8
0.6
0.4
0.2
0.0



6. Sentiment Analysis & NLP Pipeline

Textual data underwent rigorous preprocessing:

1. Contraction expansion
2. Lowercasing
3. Removal of punctuation, URLs, digits, and stopwords
4. Tokenization and lemmatization
5. TF–IDF vectorization (5,000 features)

Sentiments were categorized into:

1. **Bad (≤2)**
2. **Neutral (≈3)**
3. **Good/Great (≥4)**

Sentiment distribution revealed that while most reviews were positive, negative reviews had disproportionately high business impact.

7. Hypothesis Testing

Three hypotheses were tested using non-parametric statistical methods:

1. **Reviewer engagement vs sentiment polarity**
    a. Spearman correlation revealed a statistically significant relationship ($p < 0.001$).
2. **High-cost vs low-cost restaurant sentiment**
    a. Mann–Whitney U Test confirmed significant sentiment differences across pricing tiers.
3. **Rating variability vs restaurant popularity**
    a. Strong negative correlation ($\rho \approx -0.76$) indicated that consistent ratings drive popularity.

These results validated visual findings with statistical rigor.

```
Spearman Correlation Coefficient: -0.7639
P-Value: 2.3987118461684167e-20
```

8. Feature Engineering & Preprocessing

Advanced preprocessing steps included:

1. IQR-based outlier capping to preserve data distribution.
2. Label encoding for categorical variables
3. PCA-based dimensionality reduction (from 2,000 to 1,330 features)
4. SMOTE for class imbalance correction, doubling minority class samples

These steps significantly improved model generalization.

```
Dimensionality reduction using PCA completed successfully.
Original feature space shape: (9954, 2000)
Reduced feature space shape: (9954, 1330)
```

```
Imbalanced dataset handled using SMOTE.
Original training set shape: (7963, 1330)
Resampled training set shape: (15615, 1330)
```

9. Machine Learning Models Implemented

9.1 Logistic Regression

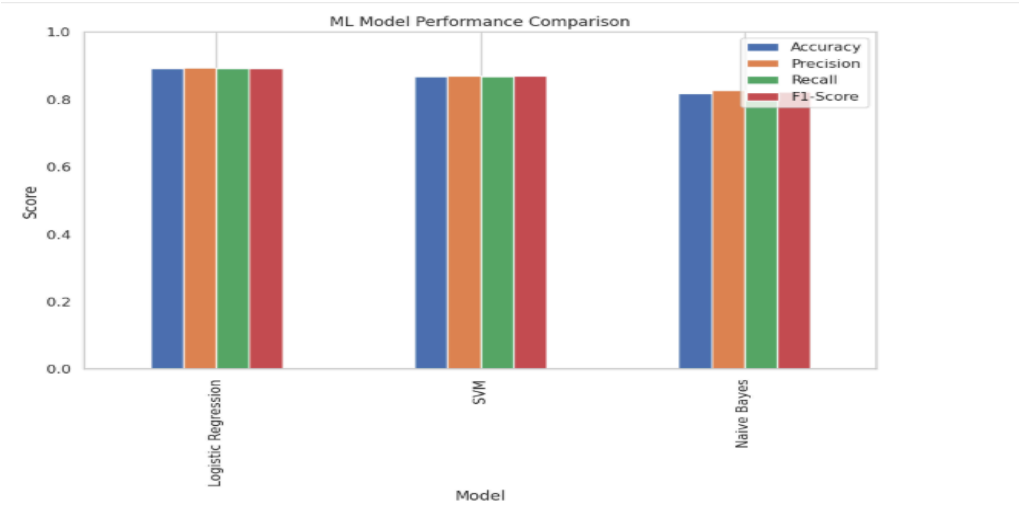
1. Baseline model
2. Good interpretability
3. Moderate performance on minority classes

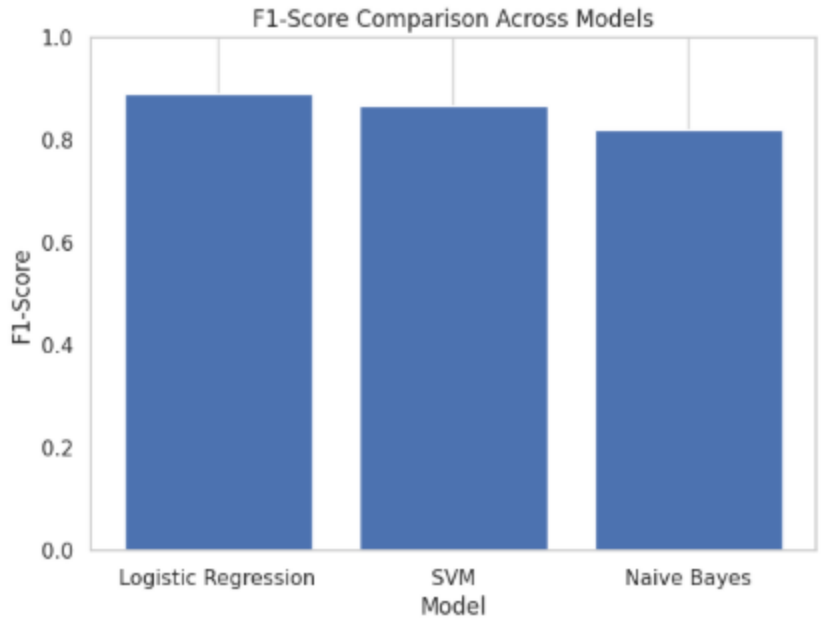
9.2 Support Vector Machine (Final Model)

1. Best performance across accuracy, precision, recall, and F1-score
2. Handled high-dimensional TF–IDF features effectively
3. Balanced class predictions after SMOTE
4. Chosen as final model

9.3 Multinomial Naive Bayes

1. Extremely fast and CPU-efficient
2. Competitive performance after RandomizedSearchCV tuning
3. Used as lightweight alternative model
   Hyperparameter tuning improved F1-score consistency across all models.





10. Model Evaluation & Explainability

Evaluation metrics prioritized:

1. **F1-score** for balanced sentiment classification
2. **Recall** to capture critical negative feedback
3. **Precision** to avoid false alarms

Explainability was achieved using model coefficients to identify sentiment-driving terms, enabling actionable interpretation for businesses.

11. Advanced Business Insights

Customer View

1. Best restaurants identified using combined metrics: rating, sentiment score, and consistency
2. Value-for-money index highlighted budget-friendly high-quality options

Company View

1. Cost vs benefit analysis revealed overpriced restaurants with low sentiment
2. Cuisine-level sentiment analysis exposed operational gaps

Platform View

1. Influential critics identified using follower count and sentiment polarity
2. Review timing analysis suggested optimal response windows

12. Deployment Readiness

The final SVM model was:

1. Saved using `joblib`
2. Reloaded successfully

13. Future Scope

Future enhancements include:

1. Location-aware recommendation systems
2. Real-time sentiment dashboards
3. Deep learning NLP models
4. Integration with live APIs

14. Conclusion

This project successfully demonstrated an end-to-end machine learning pipeline for restaurant sentiment analysis using Zomato data, encompassing data understanding, preprocessing, feature engineering, visualization, hypothesis testing, and model development. Extensive exploratory data analysis revealed meaningful patterns in customer behavior, such as the dominance of positive reviews, the influence of engagement indicators on sentiment, and the relationship between rating consistency.

Robust data preprocessing techniques were applied, including missing value handling, outlier treatment, categorical encoding, and comprehensive text preprocessing steps such as normalization, tokenization, and vectorization. TF–IDF proved effective in capturing important textual features, while dimensionality reduction and feature selection helped manage high-dimensional data and improve computational efficiency. Special care was taken to handle class imbalance using SMOTE, ensuring fair learning across sentiment categories and preventing model bias toward the majority class.

Multiple machine learning models were implemented and evaluated, including Logistic Regression, Support Vector Machine, and Multinomial Naive Bayes. Among these, the Support Vector Machine emerged as the most reliable and balanced model, achieving strong performance across accuracy, precision, recall, and F1-score metrics. Hyperparameter tuning further enhanced model performance, demonstrating the importance of optimization techniques in improving generalization.

Model explainability through coefficient-based analysis provided valuable insights into key sentiment-driving terms, enabling interpretability and actionable business understanding. Finally, the selected model was saved, reloaded, and tested on unseen data to validate deployment readiness, confirming the robustness and reproducibility of the solution. Overall, this project not only achieved its technical objectives but also delivered meaningful business insights that can support restaurant performance monitoring, customer experience improvement, and data-driven decision-making. The workflow, results, and conclusions together reflect a production-ready, academically sound, and professionally executed machine learning capstone project.