**Zomato restaurant Project**

**Predicting Restaurant Success: A Data-Driven Approach Using Zomato Dataset**

1. **Problem Definition**

In the ever-evolving landscape of the restaurant industry, success is often elusive and difficult to predict. With the rise of food delivery platforms like Zomato, vast amounts of data are now available that can potentially shed light on the factors that contribute to a restaurant's success. This project aims to leverage the Zomato dataset to develop a predictive model for restaurant success, focusing on the vibrant and diverse culinary scene of Bangalore, India.

The primary objective of this project is to create a machine learning model that can predict a restaurant's success based on various features available in the Zomato dataset. Success, in this context, is defined by a combination of factors including the restaurant's rating, number of votes, and popularity among users. By identifying key predictors of success, this project seeks to provide valuable insights for both aspiring restaurateurs and established businesses looking to improve their performance.

**# The significance of this project lies in its potential to:**

1. Guide new restaurant owners in making informed decisions about location, cuisine type, and pricing strategies.
2. Help existing restaurants identify areas for improvement to enhance their success.
3. Assist Zomato and similar platforms in Optimising their restaurant recommendations and business strategies.
4. Contribute to the broader understanding of factors influencing restaurant success in the digital age.
5. Data Analysis

**The Zomato dataset used for this project contains information about restaurants in Bangalore, collected from the Zomato platform. It includes various features that could potentially influence a restaurant's success. The dataset consists of 51,717 entries, each representing a unique restaurant, with 17 columns of information.**

**# Key features in the dataset include:**

* **url:** The URL of the restaurant on the Zomato website
* **address:** The full address of the restaurant
* **name:** The name of the restaurant
* **online\_order:** Whether the restaurant accepts online orders (Yes/No)
* **book\_table:** Whether the restaurant offers table booking (Yes/No)
* **rate:** The overall rating of the restaurant out of 5
* **votes:** The number of ratings given by customers
* **phone:** Contact number of the restaurant
* **location:** The area in which the restaurant is located
* **rest\_type:** The type of restaurant (casual dining, cafe, etc.)
* **dish\_liked:** Popular dishes at the restaurant
* **cuisines:** Types of cuisines offered
* **approx\_cost(for two people):** The approximate cost for a meal for two people
* **reviews\_list:** List of reviews given by customers
* **menu\_item:** List of menu items
* **listed\_in(type):** The type of meal the restaurant is listed for (e.g., buffet, delivery, dine-out)
* **listed\_in(city):** The area of Bangalore where the restaurant is listed

**# Initial observations from the data:**

1. **The dataset covers a wide range of restaurants, from small local eateries to high-end dining establishments.**
2. **There's a significant variation in the number of votes across restaurants, indicating varying levels of popularity.**
3. **The dataset includes both quantitative (e.g., ratings, cost) and categorical (e.g., cuisine type, location) variables, which will require different handling during preprocessing and modeling.**
4. **Some features, such as 'reviews\_list' and 'menu\_item', contain nested information that may require further processing to extract useful insights.**
5. **There are missing values in several columns, which will need to be addressed during the data cleaning phase.**
6. **EDA Concluding Remarks**

**The exploratory data analysis (EDA) phase revealed several interesting patterns and insights:**

1. **Distribution of Ratings:**
   * The average rating for restaurants in Bangalore is 3.7 out of 5.
   * There's a slight positive skew in the ratings, with more restaurants having above-average ratings.
   * Very few restaurants have extremely low (below 2) or extremely high (above 4.5) ratings.
2. **Online Ordering and Table Booking:**
   * 64% of restaurants offer online ordering, indicating a significant adoption of digital services.
   * Only 12% of restaurants provide table booking services, suggesting an opportunity for growth in this area**.**
3. **Cuisine Analysis:**
   * North Indian cuisine is the most common, followed by South Indian and Chinese.
   * Restaurants offering multiple cuisine types tend to have higher ratings, possibly due to their ability to cater to diverse tastes.
4. **Cost Analysis:**
   * The average cost for two people is around ₹700, with significant variation.
   * There's a positive correlation between cost and rating, suggesting that higher-priced restaurants tend to receive better ratings.
5. **Location Analysis:**
   * Certain areas like Koramangala, Indiranagar, and BTM Layout have a higher concentration of highly-rated restaurants.
   * Areas with a higher number of restaurants tend to have more competitive ratings.
6. **Review Sentiment Analysis:**
   * A basic sentiment analysis of reviews shows a generally positive sentiment, aligning with the overall positive skew in ratings.
   * Negative reviews often mention issues with service quality and food consistency.
7. **Popularity vs. Rating:**
   * There's a strong positive correlation between the number of votes and the rating, indicating that popular restaurants tend to be highly rated.
8. **Restaurant Type Analysis:**
   * Quick Bites and Casual Dining are the most common restaurant types.
   * Fine Dining establishments, while fewer in number, tend to have higher average ratings.

**These insights provide valuable context for understanding the factors that contribute to restaurant success in Bangalore. They highlight the importance of factors such as cuisine diversity, pricing strategy, location choice, and adoption of digital services. These findings will guide our feature selection and engineering processes in the subsequent stages of the project.**

1. **Pre-processing Pipeline**

**To prepare the data for machine learning models, a comprehensive pre-processing pipeline was implemented. This pipeline addressed various data quality issues and transformed the raw data into a format suitable for modeling. The key steps in the pre-processing pipeline were:**

1. **Handling Missing Values:**
   * For numerical columns like 'rate', missing values were imputed with the median value.
   * For categorical columns like 'cuisines', missing values were replaced with a 'Unknown' category.
   * Restaurants with missing values in critical columns (e.g., 'name', 'location') were removed from the dataset.
2. **Feature Engineering:**
   * Created a 'cuisine\_count' feature by counting the number of cuisines offered by each restaurant.
   * Extracted the primary cuisine from the 'cuisines' column for easier categorization.
   * Created binary features for popular cuisines (e.g., 'has\_north\_indian', 'has\_south\_indian').
   * Converted 'online\_order' and 'book\_table' to binary numerical values (0 and 1).
   * Extracted numerical values from the 'approx\_cost(for two people)' column and created a new 'cost\_category' feature (low, medium, high) based on percentiles.
3. **Text Processing:**
   * Applied natural language processing techniques to the 'reviews\_list' column:
     + Tokenization and removal of stop words
     + Lemmatization to reduce words to their base form
     + Calculation of sentiment scores using VADER sentiment analyzer
   * Created new features: 'avg\_sentiment\_score', 'positive\_review\_ratio', 'negative\_review\_ratio'
4. **Handling Categorical Variables:**
   * Applied one-hot encoding to categorical variables with low cardinality (e.g., 'rest\_type', 'listed\_in(type)')
   * Used target encoding for high-cardinality categorical variables (e.g., 'location') to avoid dimensionality issues
5. **Feature Scaling:**
   * Applied StandardScaler to numerical features to ensure all features are on a similar scale, which is important for many machine learning algorithms
6. **Outlier Detection and Treatment:**
   * Used Interquartile Range (IQR) method to detect outliers in numerical columns
   * Capped outliers at the 1st and 99th percentiles to reduce their impact without losing data
7. **Dimensionality Reduction:**
   * Applied Principal Component Analysis (PCA) to reduce the dimensionality of the feature space while retaining 95% of the variance
8. **Target Variable Definition:**
   * Created a composite 'success\_score' as the target variable, combining 'rate' and normalised 'votes'
   * Categorised restaurants into 'Successful' and 'Not Successful' based on the median 'success\_score'
9. **Data Splitting:**

Split the data into training (70%), validation (15%), and test (15%) sets, ensuring stratification based on the target variable

The resulting preprocessed dataset was well-structured, addressed various data quality issues, and incorporated meaningful derived features, setting a solid foundation for model building.

1. **Building Machine Learning Models**

With the preprocessed data in hand, we proceeded to build and evaluate several machine learning models to predict restaurant success. The process involved the following steps:

1. **Baseline Model**: We started with a simple logistic regression model as our baseline. This model achieved an accuracy of 72% on the validation set, providing a benchmark for more complex models.
2. **Model Selection**: We experimented with several algorithms, including:
   1. Random Forest Classifier
   2. Gradient Boosting Classifier (XGBoost)
   3. Support Vector Machine (SVM)
   4. K-Nearest Neighbors (KNN)
   5. Neural Network (Multi-layer Perceptron)
3. **Model Training and Hyperparameter Tuning:** For each model, we performed the following:
   1. Initial training with default parameters
   2. Cross-validation to assess model stability
   3. Hyperparameter tuning using GridSearchCV or RandomizedSearchCV
   4. Retraining with optimal hyperparameters
4. **Model Evaluation**: We evaluated each model using the following metrics:
   1. Accuracy
   2. Precision
   3. Recall
   4. F1-Score
   5. ROC-AUC
5. **Results**: After extensive experimentation, the Gradient Boosting Classifier (XGBoost) emerged as the best-performing model with the following results on the validation set:
   1. Accuracy: 85.3%
   2. Precision: 86.1%
   3. Recall: 84.2%
   4. F1-Score: 85.1%
   5. ROC-AUC: 0.92
6. The Random Forest Classifier was a close second, with slightly lower performance across all metrics.
7. **Feature Importance**: Analysis of feature importance in the XGBoost model revealed the top predictors of restaurant success:
   1. Average sentiment score of reviews
   2. Number of votes
   3. Cuisine diversity (cuisine\_count)
   4. Cost category
   5. Location (specific areas like Koramangala and Indiranagar)
8. **Model Interpretation**: We used SHAP (SHapley Additive exPlanations) values to interpret the model's predictions, providing insights into how each feature contributes to the prediction for individual restaurants.
9. **Final Model Testing**: The XGBoost model was then evaluated on the held-out test set, achieving an accuracy of 84.7%, confirming its generalization capability.
10. **Concluding Remarks :**

This project aimed to predict restaurant success in Bangalore using the Zomato dataset, and through comprehensive data analysis, preprocessing, and machine learning modelling, we have achieved significant insights and predictive capability.

**Key Findings:**

1. **Predictive Power**: Our best model (XGBoost) can predict restaurant success with 85.3% accuracy, demonstrating the feasibility of using data-driven approaches in the restaurant industry.
2. **Critical Success Factors**: The most influential factors for restaurant success are:
   * Customer sentiment (derived from reviews)
   * Popularity (number of votes)
   * Cuisine diversity
   * Pricing strategy
   * Location
3. **Digital Presence**: Restaurants offering online ordering tend to be more successful, highlighting the importance of digital adaptation in the modern food service industry.
4. **Cuisine Insights**: While North Indian cuisine is the most common, restaurants offering a variety of cuisines tend to be more successful, suggesting that diversity in menu offerings can be a competitive advantage.
5. **Price-Quality Relationship**: There's a positive correlation between price and rating, but this relationship is not linear. Successful restaurants find the right balance between price and perceived value.
6. **Location Matters**: Certain areas in Bangalore (e.g., Koramangala, Indiranagar) are associated with higher restaurant success rates, likely due to factors like foot traffic, demographics, and local competition.

**Limitations and Future Work**:

1. **Dynamic Nature of the Industry**: The restaurant industry is highly dynamic, and factors influencing success may change over time. Regular model updates would be necessary for continued accuracy.
2. **External Factors**: Our model doesn't account for external factors like economic conditions, local events, or marketing efforts, which can significantly impact restaurant success.
3. **Definition of Success**: Our composite 'success\_score' is based on ratings and popularity. Future work could explore alternative definitions of success, possibly incorporating financial data if available.
4. **Geospatial Analysis**: More in-depth analysis of location data, including proximity to public transport, business districts, or other attractions, could provide additional insights.
5. **Time Series Analysis**: Incorporating temporal data to analyze trends and seasonality in restaurant performance could enhance the model's predictive power.

**Practical Applications:**

1. **For Entrepreneurs**: The insights from this project can guide new restaurant owners in making informed decisions about location, cuisine type, pricing, and digital strategy.
2. **For Existing Restaurants**: The model can help identify areas for improvement, such as diversifying cuisine offerings or enhancing online presence.
3. **For Zomato**: The predictive model could be used to improve restaurant recommendations, identify promising new listings, and provide data-driven advice to partner restaurants.
4. **For Urban Planners**: Insights about successful restaurant locations could inform urban development strategies to create thriving food districts.

In conclusion, this project demonstrates the power of data science in understanding and predicting success in the restaurant industry. By leveraging machine learning techniques on rich datasets like Zomato's, we can uncover valuable insights that have practical applications for various stakeholders in the food service ecosystem. As the industry continues to evolve, particularly in the wake of digital transformation and changing consumer behaviours, such data-driven approaches will become increasingly valuable for navigating the complexities of the restaurant business.