DINING GUIDE: PERSONALIZED RESTAURANT RECOMMENDATION DATA REPORT

Business Understanding.

Picking the ideal restaurant can be both thrilling and daunting at a time when eating out and a diverse cuisine have become essential components of our social fabric. It's harder than ever to choose where to eat, with so many alternatives ranging from charming bistros to unique restaurants. Conventional restaurant websites have long depended on filters to provide consumers with a myriad of options to choose from based on amenities, geography, or cuisine varieties. But as the restaurant business changes and culinary scenes broaden, it is clear that finding the right restaurant requires a more sophisticated and customized strategy. Now enter the era of restaurant recommendation systems, a marvel of technology that does more than just narrow down restaurant options based on features. These systems provide personalized meal recommendations based on your interests and preferences by utilizing the power of data science, machine learning, and user preferences.

Restaurant recommendation systems provide an excellent solution in a world where choices are plentiful and time is of the essence. They improve the eating experience in ways that traditional filters just cannot match. This piece explores the realm of restaurant recommendation systems, examining their significance, usefulness, and revolutionary effects on how we find and savor food. We'll show you how these clever algorithms are changing the culinary scene, accommodating diners' ever-changing tastes, and transforming the process of choosing a restaurant. Come with us on this voyage as we explore the future of dining discovery and uncover the mysteries of restaurant recommendation systems.

Research Question.

- What disparities exist between the dataset's user and business rating distributions, and what do these variations suggest about user preferences?
- Which restaurant categories are the most common, and how does the distribution of these categories affect the choices and preferences of customers in the restaurant industry?
- Which cities occur most frequently, and how does the distribution of restaurants in these cities affect the restaurants that customers choose?
- Which states are the most prevalent, and how do restaurant distribution patterns within these states affect the kinds of food that are offered and the dining preferences of users?

Problem Statement.

This project aims to address the challenge faced by individuals in making informed choices about restaurants and dining experiences by developing a user-friendly restaurant

recommendation system that empowers individuals to make informed dining decisions, ultimately enhancing their overall restaurant experience

Objectives.

Main Objective

- To develop an interactive and user-friendly restaurant recommendation system.

Specific Objectives

- Analyze key factors for restaurant ratings, identifying and evaluating the key attributes and factors that significantly influence restaurant ratings and customer preferences using data analysis techniques.
- Develop content-based recommendation algorithms, creating and implementing advanced content-based recommendation algorithms that can generate personalized restaurant recommendations based on user-defined text, restaurant names, and other user preferences.
- Integrate interactive maps to create an interactive mapping feature within the recommendation system. This map will allow users to explore geographic trends in restaurant recommendations, providing a visually engaging way to discover dining options based on location.
- Build an intuitive user interface that allows users to easily access and interact with the restaurant recommendation system**.**

Data Understanding.

Data Source.

The dataset used in this project was extracted from the Yelp Restaurant database, which contains data about businesses across various locations. The dataset contains 908,915 tips/reviews by 1,987,897 users on 131,930 businesses in the form of JSON files. There are five JSON files namely **business.json, checkin.json, review.json, tips.json,** and **user.json**, but only two files were found to contain the relevant required information;

1. business.json: this JSON file has data on various businesses all spread over different US states and their relevant attributes.
2. review.json: this JSON file contains information on reviews made by different users on various businesses they were served.

Data Description.

Due to the large nature of our dataset, only 54,380 rows and 14 columns were extracted from the two above-stated JSON files. The columns were;

- *User_id*: A unique identifier for each user who submitted a review
- *Business_id*: A unique identifier for each business being reviewed
- *Name*: string, the business's name
- *Address*: string, the full address of the business
- *Stars*: The rating given by the user in terms of stars (e.g., 1.0, 2.0, 3.0, 4.0, 5.0),
- *Text*: The actual text content of the review and
- *Review_count*: number of reviews the business has received
- *City*: string, the city eg "San Francisco",
- *State*: string, 2-character state code, if applicable eg" CA",
- *Latitude*: float, the latitude of the business
- *Longitude*: float, longitude of the business
- *Attributes*: business attributes and features
- *Categories*: a list of the business categories
- *Hours*: hours in when the business is open, hours are using a 24-hour clock

Data Preparation.

Loading the data.

After that, the datasets were added to the Jupyter Notebook, where both of them were previewed for a better understanding of their columns and the relationships that exist between them. Then the two datasets were merged into one dataset, using the **business_id** column as the primary key which was contained in both of the datasets, to obtain all the features in one dataset for easier analysis i.e. a **LEFT JOIN** was used.

Cleaning data.

From the new combined data frame data cleaning was done on it. For easier comprehension, the following columns **stars_x and stars_y** were **renamed** into **rating** and **b/s_rating** columns. Subsequently, the dataframe was scrutinized to check for **missing values** and since the missing values were not in the core columns of our dataset they were imputed with a **"Not-Available"** character which wouldn't have an impact on our analysis. After the data was examined for missing values, duplicates were not discovered.

Feature Engineering.

Feature Engineering work was done to create a new column **location** by combining the city, state, and address columns. Another column **Price** was created by extracting price information from the attributes column. Then the **User_id** column was converted from string datatype to integer datatype and finally, **data splitting** took place to only acquire businesses that were restaurants, given that our data included several businesses other than restaurants. Next, we **dropped** uninformative **irrelevant columns** like date, is_open, hilarious, useful, and cool.

External Data Source Validation

Our Yelp dataset comes from the official Yelp Fusion API, so we can be sure we have the right API key and are abiding by Yelp's license agreements and conditions of use. Data Consistency: The consistency of the data was examined. We verified that all anticipated fields including company name, address, ratings, reviews, and timestamps were present and appropriately formatted.

Relevance of the Data: To ensure that the dataset is relevant to our study, we have filtered the data to only include companies and evaluations located in our designated geographic area of interest.

Data Privacy and Ethics: We are dedicated to protecting Yelp users' privacy. Under Yelp's privacy policy and applicable data protection laws, personally identifiable information (PII) is neither gathered nor disclosed without appropriate consent.

EDA

Performing a comprehensive exploratory data analysis (EDA) is crucial for developing an interactive restaurant recommendation system. The analysis focused on key dataset features, examining rating, category, and restaurant distributions across cities and states, including popular restaurants. Visualizations like histograms, box plots, and hexbin plots provided insights. Histograms revealed distributions of ratings, categories, and restaurant counts. Box plots depicted business ratings against price ranges, and hexbin plots showed the relationship between ratings and review counts. Word clouds highlighted common words in reviews and a map visualized restaurant locations. This EDA illuminated essential dataset aspects: a majority of food-related establishments with diverse cuisines, prominent nightlife venues, and fewer fast-food establishments.

Modeling

The project started with a sentiment analysis, which uses machine learning models to perform text analysis of human language. We performed Text Preprocessing which included processes such as :

1. Tokenizing the text: Split reviews into individual words or tokens.
2. Removing stop words: Common words like "and," "the," and "in" are often removed as they don't carry much sentiment information.
3. Perform stemming to reduce words to their root forms.
4. The TF-IDF (term frequency inverse-document frequency) algorithm was used to calculate the uniqueness of words in various restaurants.

We then moved to create a **Content-Based Model**, based on user input of either text, location, category, or name, which offered recommendations that met the entered user specifications.

The outputs from this model were filtered based on inputted category and location. The model majorly relied on the following;

1. **A Cosine Similarity** matrix that calculated the similarity scores between one restaurant and others and the restaurants with the top scores were offered as recommendations.
2. **Text Input** which was entered in the form of user input, was analyzed and compared with restaurant reviews, and the restaurants that contained those specifications were recommended.

We then moved to **Collaborative Filtering models**, which are popular techniques in recommendation systems and machine learning that help predict a user's preferences or interests based on the preferences and behaviors of other users. It relies on the idea that users who have agreed on or liked similar items in the past are likely to agree on or like other similar items in the future. There are two types of collaborative filtering models;

1. **Neighborhood-Based Models**: The similarity metric that outperformed the others with a small RMSE was the **Pearson Similarity**. The model in this category that had the lowest **RMSE** was the **KNNBaseline** model outperforming the KNNWithMeans and KNNBasic models.

2. **Model-Based Models:** The **SVD** model was used in this category and the model was tuned with different parameters to improve its performance.

We then developed a **Deep Neural Network** intending to improve the RMSE scores for our model predictions, with the best neural network model having a test RMSE of 1.3018.

Hence, out of all the above models created, SVD performed the best with the lowest RMSE score and hence it was to form our recommender system. The **Cold-Start Problem** was solved by combining the SVD model with the content-based model, hence forming our final model.

Conclusions

In conclusion, this project has successfully achieved its main objective of developing an interactive and user-friendly restaurant recommendation system. This system not only provides personalized dining suggestions but also takes into account various factors that influence restaurant ratings and user preferences. The integration of an advanced recommendation algorithm ensures that users can access tailored restaurant recommendations, enhancing their overall dining experiences. Throughout this project, we also met specific objectives. We designed and developed a user-friendly website, making it convenient for users to interact with our recommendation system. Additionally, we conducted in-depth analyses of the factors that significantly impact restaurant ratings and user preferences. This understanding was crucial in refining our recommendation algorithms, ensuring they provide valuable and relevant suggestions to users. Furthermore, we harnessed the power of geographical data visualization using Folium. By creating interactive maps, we were able to explore geographic trends related

to restaurant recommendations. These maps not only make our recommendations more engaging but also help users discover new dining experiences in their preferred locations. In summary, this project's multifaceted approach aimed at delivering a holistic restaurant recommendation system has proven successful. Users can now access personalized dining recommendations, taking into account various influencing factors and geographical trends. This project not only achieves its specific objectives but also offers a valuable service that enhances the dining experiences of users.

Recommendation

a) Integration of user feedback: Actively seek and integrate user feedback to refine and improve the recommendation system

b) Enhanced user profiles: Use this data to provide more personalized restaurant recommendations.

c) Enhance recommendation algorithms: Continue to refine and enhance the recommendation algorithms. Explore more advanced machine learning techniques, including deep learning, to improve recommendation accuracy and personalization.

d) Expand geographical coverage: Gradually expand the geographical coverage of the recommendation system to include more regions and cities, providing users with a broader range of dining options.

Future Improvement Ideas

a) Enhanced visuals: Incorporate images and visual content, such as restaurant photos and dishes, to make the recommendations more visually appealing and informative.

b) Community and social sharing: Encourage users to share their dining experiences and reviews within the system. Implement social sharing features to build a user community and facilitate restaurant recommendations from peers.

c) Real-time updates: Enable real-time updates of restaurant information, including opening hours, special offers, and menu changes. Users should receive the most current and accurate data.

d) Integration with food delivery services: Collaborate with food delivery services to allow users to place food orders for delivery or pickup directly through the recommendation system.

e) Advanced machine learning algorithms: Explore the use of advanced machine learning algorithms to further enhance recommendation accuracy.

Deployment

After a comprehensive testing and analysis process, we have chosen the most high-performing model to power our restaurant recommendation system. To ensure the model's future reusability, we will serialize it using the 'pickle' technique before deploying it. This approach will preserve its current state and enable its use in subsequent applications. The model will be seamlessly integrated into a user-friendly web interface designed to enhance the overall user experience. This interface will include a recommendation feature that tailors restaurant suggestions based on users' past restaurant ratings. Our deployment strategy involves exploring various deployment options, including web and mobile applications, to ensure easy accessibility across a wide range of devices.