CPSC 531 Advanced Database Management

Spring 2024

Project Name: Doordash Analytics

Instructor: Dr. Tseng-Ching James Shen

Team Members:

Sr.	Name	Email
1.	Jinendra Gambhir	jinendra@csu.fullerton.edu

Introduction

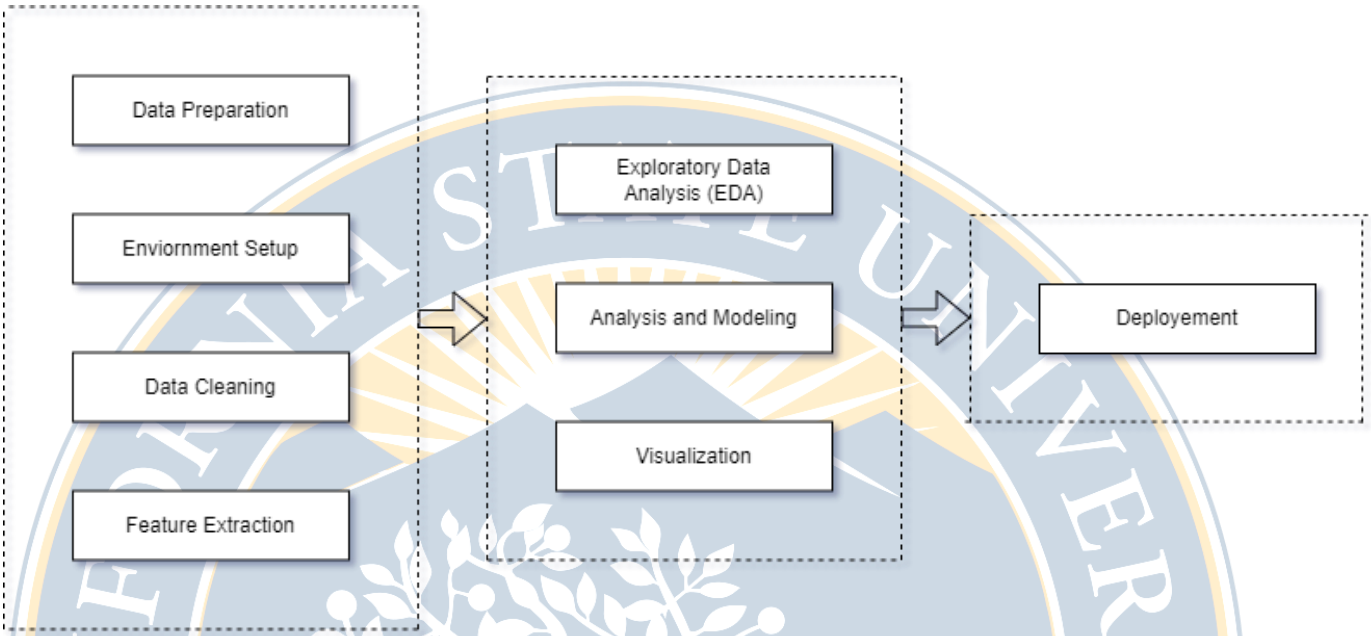
The food service industry has experienced tremendous growth in recent years thanks to consumer growth and technological advances. In this wonderful way, companies like Doordash strive to meet customer needs while maintaining efficiency and quality of service. To achieve these goals, the use of data analysis has become necessary. Analytics Project Doordash aims to leverage the power of data analytics to uncover valuable insights from customer behavior data and performance metrics. The project aims to address key issues and opportunities in the food delivery industry by analyzing large amounts of data collected from different locations, including restaurant partners, delivery activities and customer feedback. In this project, we embark on a journey to discover and understand Doordash 's core business processes, customer preferences, and market dynamics. Through comprehensive data analysis, we aim to uncover patterns, trends and relationships that can inform strategic decisions and improve business performance. Issues such as popular restaurants in the city and the impact of delivery time on customer satisfaction are examined and the factors affecting delivery time are revealed.

Problem Statement

The modern food delivery sector is exceptionally competitive for organizations like Doordash who require analytics for optimizing their operational efficiency and enhancing customer satisfaction levels. This project's specific problem statement is the using data analytics techniques to get viable results such as Market Basket analysis from massive data on actual behavior of Doordash 's customers and their operational datasets. This project will seek to

1. Identify popular food and beverage trends in different cities and states to adapt menu offerings and marketing strategies.
2. Analyze the impact of delivery time on customer evaluation and satisfaction level to improve service quality and performance.
3. Develop a forecasting method to estimate delivery times based on various factors such as distance, configuration, and meter density to provide customers with accurate delivery estimates.
4. Implement a program to provide users with recommendations or personalized content based on their location and preferences to improve customer experience.

Flowchart



Block diagram of the Doordash project activity analysis explaining the sequential steps involved in the project execution process. Each chapter describes a different phase or task in the project, from preparing data to submitting and sharing project results. Arrows show the flow of information and workflow between different stages and show the structure of the project.

Tech Stack

In this Doordash analytics project, we use a variety of tools and technologies to perform data analytics, develop predictive models, generate insights, and deploy applications. The main tools and technologies used are:

Technology	Purpose
Pandas	Data manipulation and analysis.
PySpark	Big data processing and analysis.
Scikit-Learn	Machine learning algorithms and model development.
Matplotlib	Data visualization.
Seaborn	Enhanced data visualization.
Tableau	Creation of interactive visualizations and dashboards.
Streamlit	Deployment of interactive web applications.
GitHub	Version control and collaboration platform.
Jupyter Notebooks	Exploratory data analysis and documentation.
Google Colab	Cloud-based development and execution of Python code.
Python	Primary programming language for data analysis, machine learning, and application development.

Using these tools and technologies, we are equipped to perform comprehensive data analysis, develop machine learning models, create powerful visuals, and use interactive applications to provide valuable insights and actions to partners in the Doordash analytics project.

Database

Dataset Link: Kaggle Database

Cleaned Dataset Link: Cleaned Dataset

Data set information:

```
[ ]  print("Columns = ")
     df.columns
```

```
Columns =
Index(['searched_zipcode', 'searched_lat', 'searched_lng', 'searched_address',
       'searched_state', 'searched_city', 'searched_metro', 'city_slug',
       'latitude', 'longitude', 'distance', 'loc_name', 'loc_number', 'url',
       'address', 'cuisines', 'delivery_fee_raw', 'delivery_fee',
       'delivery_time_raw', 'delivery_time', 'service_fee_raw', 'service_fee',
       'phone', 'review_count', 'review_rating', 'RunDate'],
      dtype='object')
```

We did data cleaning, it involves handling missing values, segmenting data by food, allergy and culinary origin to improve the accuracy of the analysis.

Source: The dataset is derived from the Kaggle Database of publicly available data.

Functionality

1. **Famous dishes of the city:**
   - Analysis of the distribution of popular foods in different cities.
   - Group restaurants by cuisine type and count the number of each cuisine in each city.
   - Learn the top dishes popular with customers in each city so you can strategize marketing and menu selections.

2. **Best restaurants in the city:**
   - Discover the highest-rated restaurants in every city based on the rating scale.
   - Aggregation of restaurant rating points by cuisine type and city.
   - It allows you to have an idea about restaurant selection and customer satisfaction by constantly opening highly rated restaurants in certain cities.

3. **Most popular restaurants by state and highest-rated restaurants in states:**
   - Find popularity and rating of foods in different states.
   - Calculate the frequency and average of food consumption in each province.
   - Get to know your local cuisine and discover the interest and diversity of cuisines in the state.

4. **Effect of delivery time on evaluation:**
   - Investigate the relationship between delivery time and review scores to understand customer satisfaction levels.
   - Calculation of the correlation between delivery time and evaluation of points.
   - Find out if longer delivery times are associated with lower rankings, learn about strategies to improve delivery operations and improve customer experience.

5. **Relationship between distance and delivery time:**
   - Find the relationship between delivery distance and delivery time.
   - Calculate the relationship between distance and delivery time.
   - Understand how distance affects delivery time, process ease and efficiency.

6. **Rate the number of customers and average fee:**
   - Analyze the relationship between the number of customers reviews a restaurant has and its review rate.

- Try to understand the distribution of the number of repetitions and the average rate in the restaurant.
- Find out if restaurants with lots of reviews tend to receive high ratings, which indicates high levels of customer satisfaction and engagement.

7. **Compare city:**
   - Make a comparative analysis of key parameters such as review rates, delivery times and popular foods in different cities.
   - Try visualizing the sizes using interactive whiteboards so you can easily compare.
   - Improve targeting and decision-making by identifying differences in customer needs, service quality and efficiency between cities.

8. **Customer segment:**
   - Use technology to segment customers based on their calculations and evaluations.
   - Analyze the characteristics and needs of each customer segment.
   - Understand customer behavior and preferences, enabling personalized marketing strategies and service offerings.

9. **Conditional analysis:**
   - Analyzing short-term trends in key metrics such as delivery time, creditworthiness, and expected volume over time.
   - Identify trends, trends, and anomalies using trend analysis techniques.
   - Help you make decisions and allocate resources by understanding how key metrics change over time.

10. **Customer satisfaction measurements:**
    - Find additional customer metrics like customer feedback sentiment analysis and customer retention rates.
    - Try to understand customer standards to assess overall satisfaction level and know where you need to improve.
    - Accurately capture customer satisfaction levels and factors affecting customer experience, implement strategies to improve service quality and customer retention.

Version Control of Project

Link: GitHub

GitHub collaboration on Doordash Analytics project:

In Doordash analytics projects, GitHub has become the primary platform for collaboration, version control, and project management. The project's partners, Jayraj and I (Jinendra), have used GitHub extensively to streamline our work and enable better collaboration throughout lifecycle.

- The project's codebase is hosted on GitHub, providing a central repository to store and manage all project-related files, documents, and notebooks.
- We used GitHub version control to track changes, collaborate on code development, and maintain change history.

Test Result

1. Data Cleaning Results:

- **Before Cleaning**

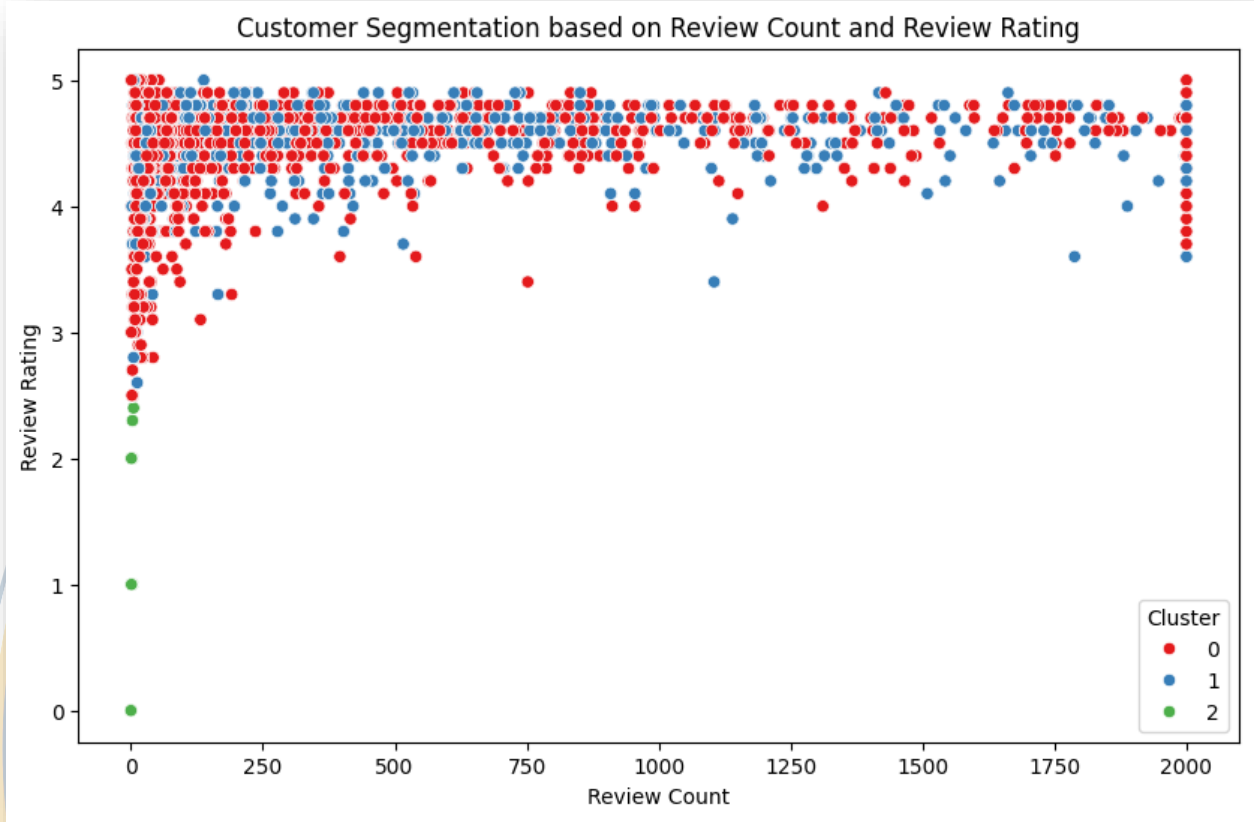
```
[ ]  print("Columns = ")
     df.columns

     Columns =
     Index(['searched_zipcode', 'searched_lat', 'searched_lng', 'searched_address',
            'searched_state', 'searched_city', 'searched_metro', 'city_slug',
            'latitude', 'longitude', 'distance', 'loc_name', 'loc_number', 'url',
            'address', 'cuisines', 'delivery_fee_raw', 'delivery_fee',
            'delivery_time_raw', 'delivery_time', 'service_fee_raw', 'service_fee',
            'phone', 'review_count', 'review_rating', 'RunDate'],
           dtype='object')
```

- **After Cleaning:**

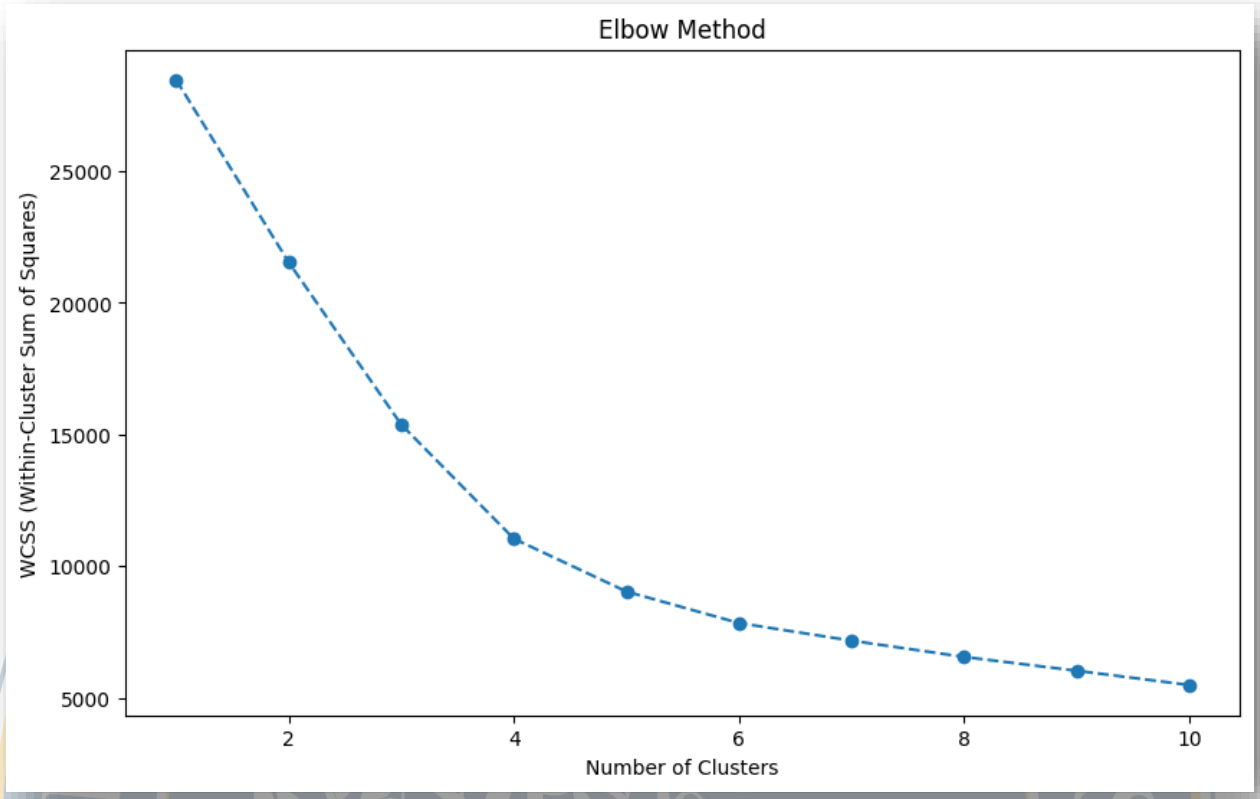
```
[ ]  print("Columns = ")
     df.columns

     Columns =
     Index(['searched_zipcode', 'searched_lat', 'searched_lng', 'searched_address',
            'searched_state', 'searched_city', 'searched_metro', 'latitude',
            'longitude', 'distance', 'loc_name', 'loc_number', 'address',
            'cuisines', 'delivery_time_raw', 'delivery_time', 'review_count',
            'review_rating', 'RunDate', 'Specialty Items', 'Meal Types',
            'Dietary Preferences'],
           dtype='object')
```

The provided image shows the results obtained after data cleaning in the Doordash analytics project. Through the process of carefully refining and implementing the steps, qualitative insights are generated, leading to the creation of insight pillars that provide useful information for analysis and decision-making.

2. Customer Segmentation (K-Means):

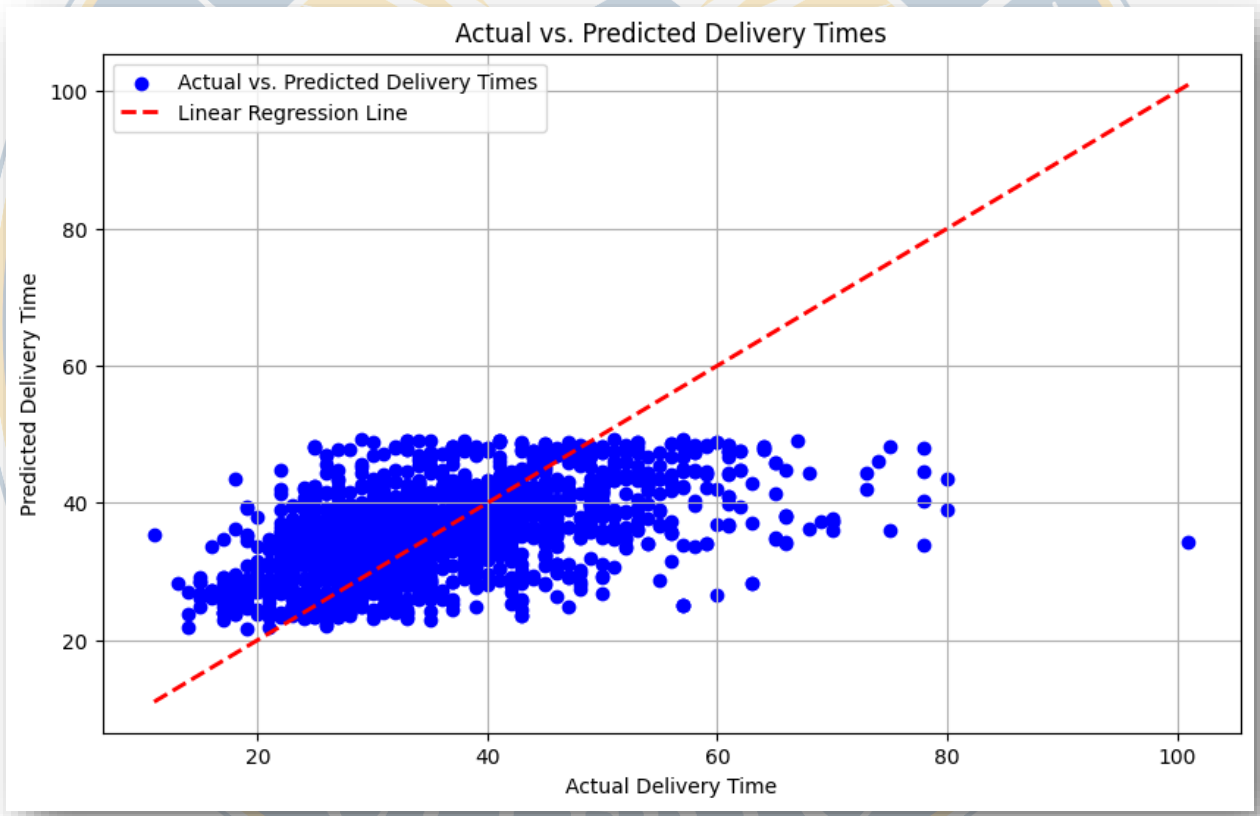



```
features = ['distance', 'review_count', 'review_rating', 'delivery_time']

# Determine optimal number of clusters using the Elbow Method
wcss = []
for i in range(1, 11):
    kmeans = KMeans(n_clusters=i, init='k-means++', random_state=42)
    kmeans.fit(data_scaled)
    wcss.append(kmeans.inertia_)

# Plot the Elbow Method to find the optimal number of clusters
plt.figure(figsize=(10, 6))
plt.plot(range(1, 11), wcss, marker='o', linestyle='--')
plt.title('Elbow Method')
plt.xlabel('Number of Clusters')
plt.ylabel('WCSS (Within-Cluster Sum of Squares)')
plt.show()
```


Elbow Method
WCSS (Within-Cluster Sum of Squares)
25000
20000
15000
10000
5000
2
4
6
8
10
Number of Clusters

3. Predictive model using Linear Regression:

- Database Coding and Initial Processing
- Feature Engineering
- Model Training with Scikit-Learn
- Model Evolution
- Predicted and Actual Delivery Times Alongside Test Features

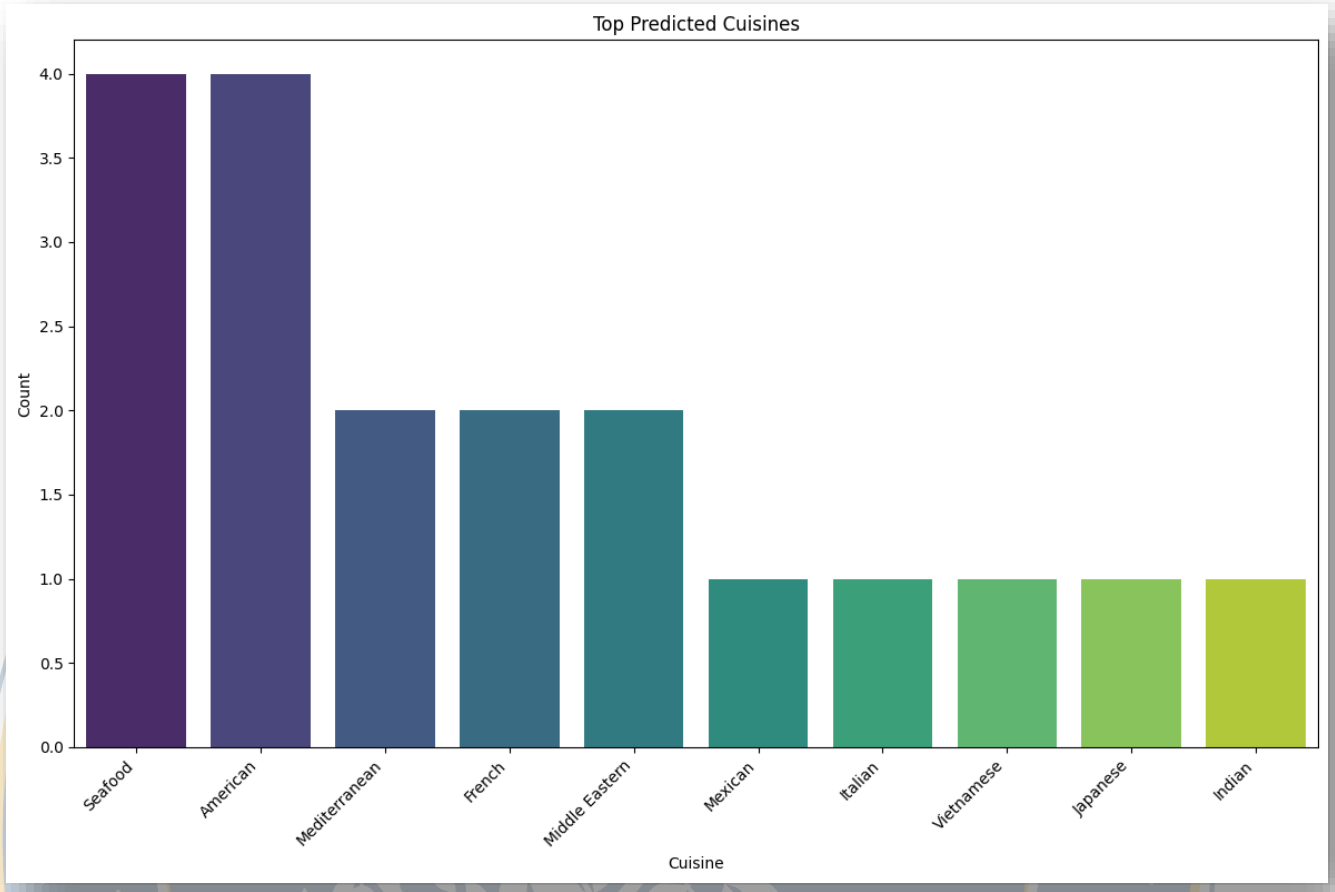


4. Cuisine Recommendation System:

A recommendation system that suggests cuisines or specialty items to users based on their location and preferences:

	searched_city	cuisines	Specialty Items	Meal Types	Dietary Preferences	review_count	review_rating	prediction
0	Manhattan Beach	Mexican	Tacos	Lunch	Vegetarian	2000	4.5	4.696221
1	Manhattan Beach	Mexican	Sandwiches	Beverages	Gluten-Free	2000	4.7	4.696221
2	Manhattan Beach	Mexican	Tacos	Beverages	Vegan	2000	4.7	4.696221
3	Burbank	Mexican	Tacos	Breakfast	Vegetarian	13	4.5	4.678928
4	Burbank	Mexican	Tacos	Lunch	Vegan	193	4.6	4.678928
5	New York	Mexican	Burgers	Breakfast	Halal	58	4.5	4.375225
6	New York	Mexican	Tacos	Desserts	Vegan	1778	4.7	4.375225
7	Boston	Mexican	Tacos	Lunch	Vegetarian	2000	4.1	4.264778
8	Glendale	Mexican	Tacos	Beverages	Vegan	991	4.8	4.636184
9	Glendale	Mexican	Tacos	Beverages	Vegetarian	991	4.8	4.636184
10	Jersey City	Mexican	Tacos	Beverages	Gluten-Free	1779	4.5	4.531206

- ALS (Alternating Least Squares) algorithm to generate recommendations for restaurants based on user review ratings.
- Predict the review ratings for different cuisines, specialty items, meal types, and dietary preferences in each city, leveraging the trained ALS model.



- Extracts the most common dietary preferences, specialty items, meal types, and cuisines for each city and merges them with the top predicted cuisine recommendations.

- Utilizing top predicted cuisine recommendations alongside expected customer preferences can inform decisions on opening new restaurants, potentially increasing profitability by aligning offerings with local tastes and demand.

Deployment

Link: Streamlit

In the Doordash analytics project, we used a simple deployment method to move their code from a development environment like Google Colab to a production environment for end users. The implementation involved using Streamlit, a popular Python library for building interactive websites, to create a web-based, easy-to-use search engine for research and development projects.

1. **Work on Colab:**
   - Initially, we developed and refined our analysis and performance analysis methods in Google Colab, Colab provided us a simple collaboration platform for coding, analyzing and testing a variety of methods and techniques.

2. **Streamlit Application Development:**
   - After the analysis and modeling work in Colab was completed, we modified the structure of the Streamlit application to present the findings of our project.
   - We used Streamlit's Python API to develop websites based on existing code bases without the need for front-end development.

3. **Creating user interfaces:**
   - With Streamlit, we created a user interface that allows end users to seamlessly integrate project charts, dashboards, and insights.
   - We include web widgets such as slides, drop-down menus, and buttons so that users can easily search for different types of information and analyze the results.

4. **Deploying the Streamlit app:**
   - Once the Streamlit software was developed and tested locally, we began putting it into production and rolling it out to the masses.

5. **Continuous monitoring and updates:**
   - ➢ After implementation, we monitored the performance of the Streamlit app and quickly resolved any issues and errors.
   - ➢ We constantly update the app with new features, enhancements, and enhancements based on user requests and project goals.

Advantages:

We ran the project on Streamlit, ensuring analytical insights and results were accessible to a wide range of people, including stakeholders, decision makers, and end users.

Streamlit software provided an easy-to-use and engaging experience, allowing users to discover project insights and tailor their analysis to their preferences.

Using Streamlit for distribution made integration possible and allowed the program to meet increasing user and data storage needs as the project grew in popularity and popularity.

The integration of this implementation plan into a project report demonstrates Jinendra and Jayraj's efforts to make their analytics project feasible, interactive, and effective for stakeholders and end users.