K-Means and ALS-Based Recommendation Engines for Yelp Data Using PySpark

Kabir Thakur
Syracuse University
kathakur@syr.edu

Pankaj Yadav
Syracuse University
pyadav05@syr.edu

Himanshu Mangal
Syracuse University
hmangal@syr.edu

Jenil Sheth
Syracuse University
jesheth@syr.edu

Abstract

In this paper, we present a comprehensive recommendation engine for Yelp data (Yelp, 2022) using PySpark, combining K-means clustering and ALS-based collaborative filtering techniques to provide personalized business suggestions to users. During the data preprocessing stage, we discovered that the highest concentration of restaurants and users was found in Pennsylvania (PA), which guided our focus on this specific region. Our approach uses restaurant categories as features, and we perform extensive data cleaning, transformations, and aggregations to improve the model's overall performance. By incorporating both user similarity and business characteristics in the recommendation process, we generate a diverse and context-aware set of recommendations. We propose a hybrid approach where the two models can be combined, showcasing improved effectiveness and scalability when working with large-scale Yelp data.

Keywords: Recommendation Engine, PySpark, K-means Clustering, ALS Collaborative Filtering, Yelp Data

1. Introduction

In the era of big data, recommendation systems play a crucial role in enabling businesses to effectively engage with their customers and personalize their experiences. The rapid growth of online platforms, like Yelp, has created a wealth of data about user preferences and business characteristics. We are leveraging this data to provide tailored recommendations to users. We explore the application of K-means clustering and Alternating Least Squares (ALS) Collaborative Filtering techniques to develop robust recommendation engines using Yelp data, with a focus on restaurants in the state of Pennsylvania. Our study employs PySpark, a widely-used big data processing framework, to handle the extensive Yelp dataset, which contains millions of reviews, users, and businesses. We perform rigorous data cleaning and preprocessing to extract valuable insights, such as the prevalence of restaurant businesses and user activity. By selecting restaurant categories as features, we create two distinct recommendation engines: one based on K-means clustering, and another employing ALS Collaborative Filtering. These recommendation systems aim to generate relevant suggestions for users, considering both user preferences and business attributes. In the following sections, we describe the methodology behind our approach, including data processing, feature engineering, and model development for both K-means and ALS-based recommendation engines. We then evaluate the performance of our proposed models and discuss their strengths, weaknesses, and potential improvements. Finally, we present our conclusions and explore future directions for research in the domain of big data recommendation systems. By investigating these advanced techniques, we aim to contribute to the ongoing development of more effective and personalized recommendation engines for large-scale datasets like Yelp.

2. Related Work

In recent years, several studies have explored the development of recommendation engines using various techniques, ranging from collaborative filtering to deep learning. In this section, we briefly review some of the most relevant research in the field, focusing on those works that have influenced our study and provided valuable insights into the application of K-means

clustering and ALS-based Collaborative Filtering for recommendation systems.

K-means clustering is a popular unsupervised learning algorithm that has been employed in various recommendation engine applications. Joshi and Dubey, 2020 used K-means to create user and item clusters for restaurant recommendations, while Bandyopadhyay et al., 2021 utilized the method for product recommendation in e-commerce platforms.

Collaborative filtering (CF) is one of the most popular methods employed in recommendation systems. It is based on the premise that users with similar preferences in the past will likely have similar interests in the future. Two primary approaches are commonly used in CF: user-based and item-based. Sarwar et al., 2001 introduced the item-based collaborative filtering algorithm, which has been successfully applied to large datasets, such as the Netflix Prize competition (Bell et al., 2007).

Matrix Factorization and Alternating Least Squares (ALS): Matrix factorization techniques have gained popularity for their ability to effectively handle sparse data and reveal latent factors. One such method, Alternating Least Squares (ALS), was introduced by Zhou et al., 2008. ALS has been widely used in recommendation systems, including the winning solution for the Netflix Prize competition (Bell et al., 2007).

Our study builds on this extensive body of research, with a focus on K-means clustering and ALS-based Collaborative Filtering applied to Yelp data. By exploring these techniques and their performance in recommending restaurants, we aim to contribute to the development of more effective and personalized recommendation engines.

3. Method

3.1. Data Description

The Yelp reviews dataset is a comprehensive collection of user-generated reviews for businesses listed on the Yelp platform. It comprises more than 8 million individual reviews, each associated with a rating ranging from 1 to 5 stars, indicating the reviewer's satisfaction with the business. The dataset provides additional valuable information such as the textual content of the reviews, the category or industry to which the businesses belong, and their geographical locations. To construct this dataset, we have utilized three distinct data files, namely the Businesses file containing 150,346 records, the Users file comprising 1,987,896 records, and the Reviews file encompassing 6,990,280 records. By amalgamating these data sources, we have created a rich and diverse dataset that facilitates in-depth analysis and exploration of user sentiments and preferences towards various businesses featured on Yelp.

3.2. Exploratory Data Analysis

The data preprocessing phase involved several steps to prepare the dataset for analysis. Irrelevant columns were removed, and the remaining columns were renamed appropriately. Any missing values were eliminated, and the necessary columns were typecasted to ensure proper data types. Once the datasets were cleaned, they were merged to create a master dataset, combining relevant information from each source.

To gain insights into the distribution of categories within the dataset, a barplot (Figure 1) was created. The analysis revealed that out of the 728 unique categories, a significant portion of businesses belonged to the Restaurants and food-related categories as seen in figure 1. This finding prompted the decision to filter the data and focus solely on Restaurants for further analysis, allowing for a more targeted investigation within the desired domain.

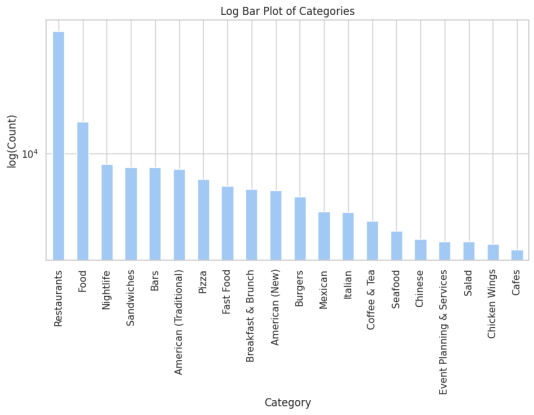


Figure 1. Top 20 Business Categories Distribution.

The barplot visualization played a crucial role in identifying the dominant categories within the dataset and guided the subsequent data filtering process. This approach ensures that the analysis focuses specifically on the Restaurants category, enabling more precise and relevant insights to be derived from the data.

The next step involved assessing the status of the businesses within the dataset. Specifically, we examined whether each business was open or closed. The evaluation revealed that a significant proportion, approximately 33%, of the businesses in the dataset were closed down. These findings as shown in Figure 2 prompted us to further refine the dataset by filtering

out the closed businesses. By narrowing our focus to only consider open businesses, we aimed to ensure that our analysis and insights are relevant to the current operational landscape, thus providing more accurate recommendations.

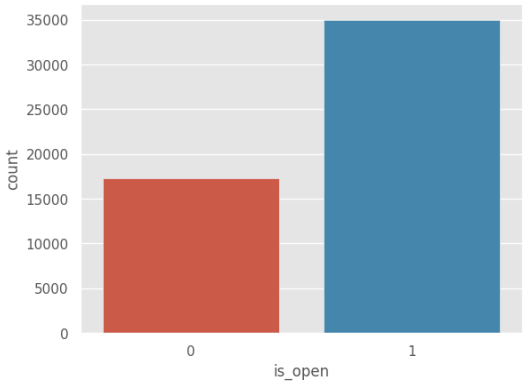


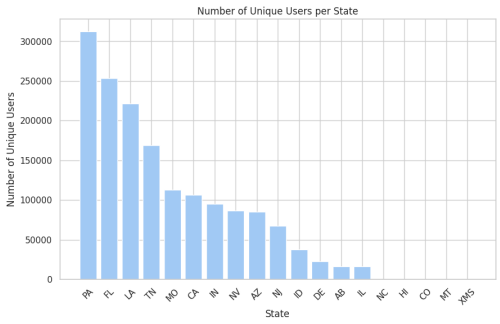
Figure 2. Open Businesses Distribution

To gain insights into user distribution, we investigated the number of unique users in each state. The analysis revealed that Pennsylvania (PA) had the highest count of unique users, indicating a substantial user presence in that state. Following Pennsylvania, Florida (FL) emerged as the second-highest state with a significant number of unique users. This information was visualized in Figure 3, providing a clear depiction of the user distribution across different states.

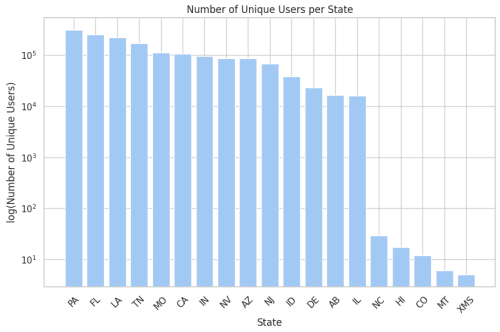
To further explore the restaurant landscape, we focused on open restaurants and examined their distribution across different states. Utilizing a boxplot visualization, we compared the restaurant counts in the top 10 states (Figure 4). Our analysis revealed that Pennsylvania exhibited the highest number of open restaurants among these states, indicating a thriving culinary scene in the region. Based on these insights we chose to work with restaurants and users in Pennsylvania.

Additionally, we delved into the restaurant counts in the top 10 cities across the United States (Figure 4). Notably, Philadelphia emerged as a standout city with a substantial number of restaurants, solidifying its position as a culinary hub. This finding emphasizes the significance of Philadelphia's dining culture and suggests a vibrant food scene within the city.

In order to gain further insights into the restaurant landscape, we analyzed the distribution of ratings across the top states in Figure 5. Our investigation revealed that Pennsylvania (PA) garnered the highest number of 5-star ratings, indicating a considerable number of highly-rated restaurants in the state. Florida (FL) followed closely behind, securing the second position



(a) Count of Unique Users

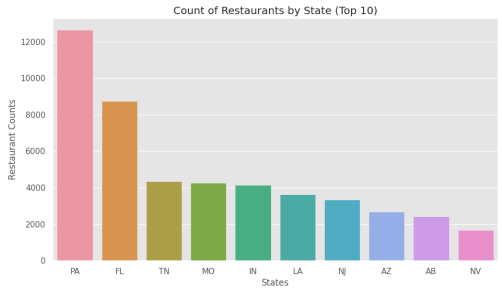


(b) Log Count of Unique Users

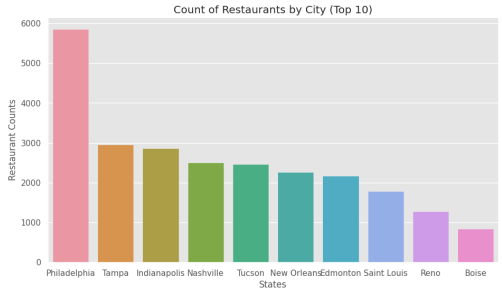
Figure 3. Count of Unique Users by State

in terms of 5-star ratings. These findings align with our earlier observation regarding the distribution of restaurants per state, where Pennsylvania and Florida ranked prominently. The correlation between high ratings and the prevalence of restaurants in these states suggests a positive reception of their culinary offerings, reflecting a vibrant and thriving dining scene.

In order to analyze user activity over time, we conducted a time series analysis based on the number of reviews. Our findings, as depicted in Figure 6, indicated that 2017 witnessed the highest influx of reviews. Notably, there was exponential growth in the number of reviews between 2010 and 2016, suggesting a significant increase in user engagement during that period. One plausible explanation for this surge is the growing popularity of online review platforms, which facilitated the ease of leaving reviews and encouraged user participation. However, a noticeable decline in the number of reviews occurred starting from 2019. This decline can potentially be attributed to the impact of COVID-19, as people began to adopt social distancing measures and reduce their dining-out activities. The pandemic's influence on consumer behavior likely contributed to the decrease in reviews during this period.



(a) Restaurants Count by State



(b) Restaurants Count by City

Figure 4. Restautants Distribution

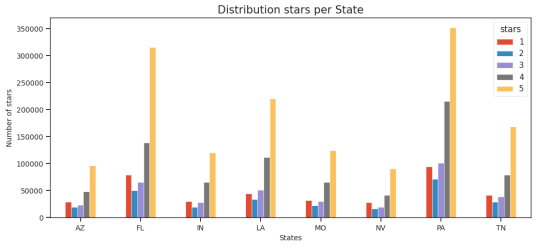


Figure 5. Stars Distribution

3.3. Feature Engineering

3.3.1. K-means

To begin, we carefully select the relevant features for our model. Our initial selection includes the columns containing business IDs and categories. There are approximately 8,000 restaurants and 450 categories in this dataset. However, this creates an extensive feature space, so we filter out categories which were present in fewer than 10 restaurants. This refined dataframe will be used for clustering restaurants based on the types of categories they serve.

In order to create a feature set for users, we join the restaurant features dataframe with the reviews table, retaining only the user IDs from the reviews table. This merge allows us to identify which users reviewed specific restaurants. Next, we aggregate the category

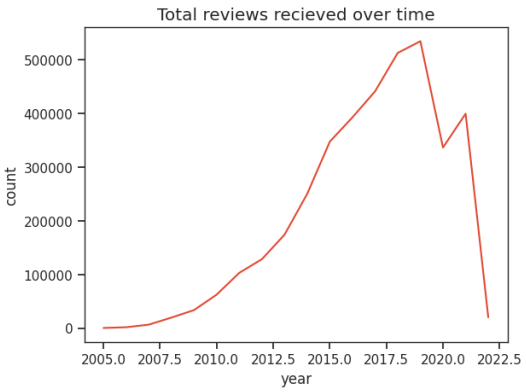


Figure 6. Over time Trend

columns by summing them for each unique user ID. By doing this, we gain insight into each user's preferences based on the weighted categories. It is important to note that the category columns, which serve as the dataset's features, remain consistent across both feature sets.

The final training data for kmeans similar restaurant model has 8069 restaurants and 155 features. The final training dataset for kmeans similar users has 269462 users and the same 155 features.

3.3.2. Alternating Least Squares

In the case of collaborative filtering with the ALS algorithm, the primary features are derived from the user-item interaction data. This data usually contains user IDs, item IDs (in our case, business IDs), and the ratings or preferences users have expressed for specific items(in our case, review star ratings).

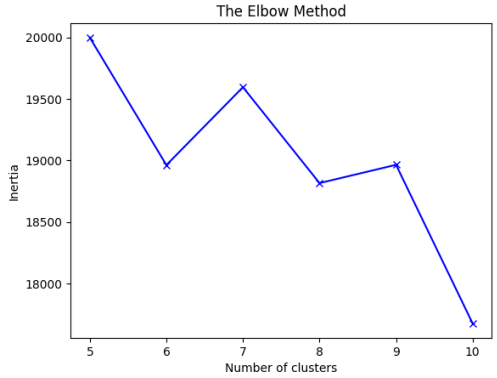
For the ALS model, we need to transform the raw data into a format that the algorithm can use efficiently. This involves converting the original string IDs for users and items into integer indices.

3.4. K-means Models

We subject both feature sets to our K-means pipeline and run it for k values ranging from 5 to 10. We then analyze the inertia values and silhouette scores to determine the optimal number of clusters. The elbow points and silhouette scores pointed towards a k value of either 6 or 7 for both restaurant and user clusters.

Additionally, we examined the count and log(count) plots for restaurants and users per cluster, aiming to select the optimal value for k that yields fairly even distributions. As a result, we choose k=6 to generate the final two K-means models.

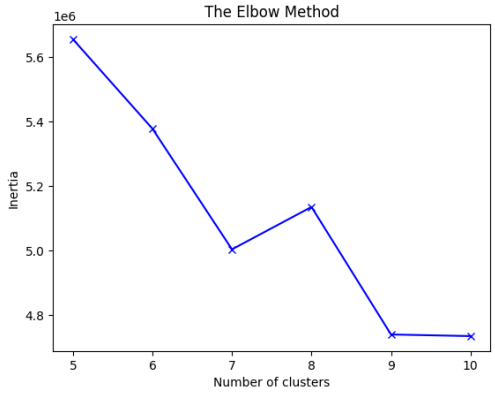
Applying these models to our data provides us with the cluster membership number of each restaurant and user. Utilizing this information, we create separate



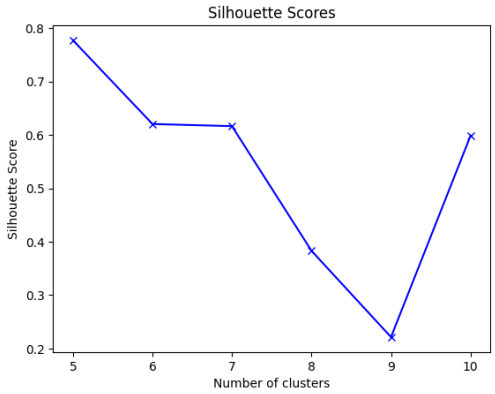
(a) Restaurants clusters Elbow Plot

(b) Restaurant cluster Silhouette

Figure 7. Restaurants cluster analysis



(a) User clusters Elbow Plot



(b) User cluster Silhouette

Figure 8. User cluster analysis

dataframes for each cluster and compute the distance of each data point from its respective centroid. We then sort the points according to their distances from the centroids and generate tables of recommendations for both the user and restaurant models. Each business receives recommendations for two other businesses, while each user is recommended two other users. This is accomplished by sorting each cluster by distance from the centroid and selecting the two nearest points as the recommendations.

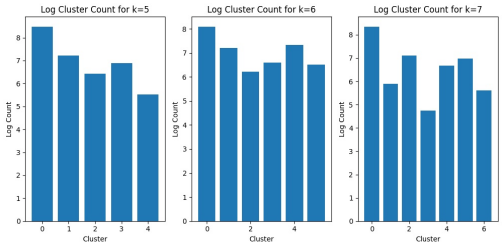
To offer restaurant recommendations to users, we consider the establishments they have visited and recommend similar ones based on the restaurant recommendations. Furthermore, we identify similar users based on user recommendations and compile a list of restaurants they have visited. Finally, we combine these two sources to generate a comprehensive set of recommendations for each user.

3.5. Alternating Least Square model

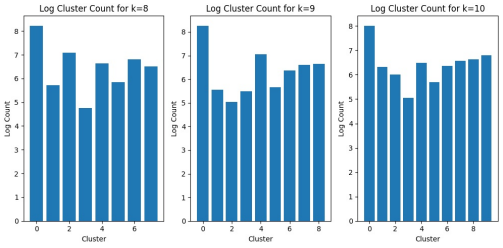
The ALS model learns latent features or factors for both users and items (in this case, restaurants) from the user-item interaction data (ratings provided by users). By learning these latent features, the model can predict the preferences or ratings that users would give to previously unrated items. This allows the recommendation system to suggest new items (restaurants) to users based on their past preferences and the preferences of other users with similar tastes.

Using the training data the ALS algorithm learns the latent features or factors for both users and items by minimizing the difference between the observed ratings and the product of user and item factor matrices. The algorithm uses an alternating least squares optimization process, where it first fixes the user factors and updates the item factors, and then fixes the item factors and updates the user factors. This alternating process continues until the specified number of iterations is reached or the model converges.

We have limited our ALS model to provide 3 recommendations per user.

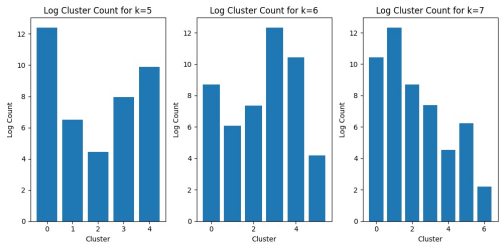


(a) Cluster with k=5,6,7

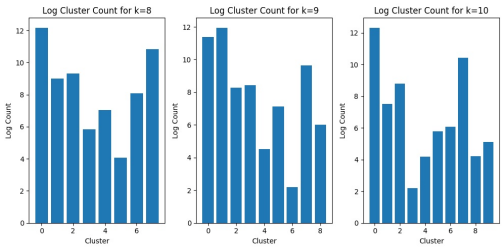


(b) Cluster with k=8,9,10

Figure 9. Users per cluster



(a) Cluster with k=5,6,7



(b) Cluster with k=8,9,10

Figure 10. Users per cluster

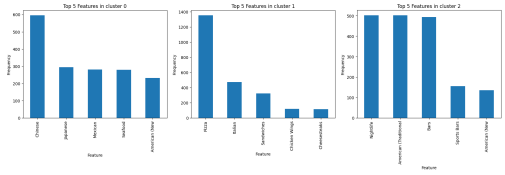
4. Results

The final K-means model for restaurants had 6 clusters as we can see in image 11. From the plots we can see that cluster 0 has a lot of Asian and Mexican cuisine, cluster 1 has mostly Italian restaurants that serve pizza and chicken wings. Cluster 2 and 3 mostly consist of restaurants which offer a vibrant nightlife and alcohol. Cluster 4 is for sandwiches and tea and cluster 6 is for fast food.

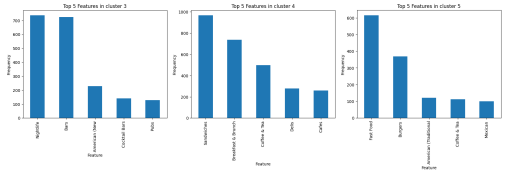
Let us consider a username Desirae. We can see from figure 12 that even for a user which has not visited many restaurants our K-means model can generate a lot of recommendations. From Desirae's recommendations we can also clearly see that she likes sandwiches or has visited restaurants that serve sandwiches.

Let us look at another user Brett who has visited over a thousand restaurants. We can see from figure 13 that our model generated a lot of recommendations. Upon comparing the recommended restaurant with the restaurants the user has already visited we saw that there are 2 restaurants in common. This means that the user already visited these two restaurants.

The ALS model generates a rating for each restaurant user pair and then shows all restaurants which would get a high rating by the user. To reduce computational costs we only limit the recommendation per user to 3. We can see the restaurants recommended by our ALS system in figure 14. The RMSE value for the ALS model is about 1.30 which suggests that our model was quite accurate in predicting user ratings for restaurants.



(a) Restaurants features per cluster



(b) Restaurants features per cluster

Figure 11. Restaurants features per cluster

5. Conclusion

In conclusion, this study has demonstrated the potential of using K-means clustering and ALS-based Collaborative Filtering for constructing recommendation engines with Yelp data. The proposed hybrid approach takes advantage of both K-means clustering to segment the dataset into similar groups based on restaurant categories and the power of the ALS method for collaborative filtering to provide personalized restaurant recommendations to users.

By employing these methods on Yelp's restaurant data, we were able to generate relevant and meaningful recommendations that cater to users' preferences, as seen by the examples. Furthermore, the data cleaning

name_user	name_restaurant	cluster_type
Desirae	Charley's Grilled Subs	u
Desirae	FreshWorks of Levittown	u
Desirae	Walt's Steak Shop	u
Desirae	Charley's Grilled Subs	u
Desirae	FreshWorks of Levittown	b
Desirae	La Casa Del Sandwich	b

(a) Recommendations for Desirae

name
Cosmos Fine Nail ...
Jeans Cafe

(b) Restaurants visit by Desirae

Figure 12. User - Desirae

process provided valuable insights into the distribution of restaurants and users, with the highest concentration being in Pennsylvania.

The combined use of K-means and ALS can demonstrate promising results in addressing the challenges of data sparsity and scalability, which are common in real-world datasets. This study contributes to the ongoing research in the field of recommendation engines and highlights the benefits of integrating multiple techniques to improve recommendation quality.

Future research can further investigate the optimal combination of parameters and techniques for different datasets and explore the benefits of this hybrid approach, additionally incorporating content-based filtering, to enhance the performance of recommendation engines.

6. Discussion

While our study has shown the potential of using K-means clustering and ALS-based Collaborative Filtering for creating recommendation engines using Yelp data, it is essential to acknowledge that there is always room for improvement. In the context of recommendation engines, various techniques can be incorporated to enhance the system's overall performance, and no single method or combination of methods can guarantee optimal results for every scenario.

One major challenge in evaluating recommendation engines is the lack of quantitative metrics that accurately measure their performance. Typically, performance metrics like RMSE or precision and recall can provide insights into the effectiveness of the engine. However, these metrics may not necessarily align with the actual user satisfaction or their actions based on the recommendations provided.

name_user	name_restaurant	cluster_type
Brett	QDOBA Mexican Eats	u
Brett	QDOBA Mexican Eats	u
Brett	Biryani City	u
Brett	Miss Winnie's	u
Brett	Rio Brazilan Steak Truck	u
Brett	Meat Wagon BBQ	u
Brett	Concerto Fusion	u
Brett	Tran's Chinese Food Cart	u
Brett	Sake Hana	u
Brett	Kapow Kitchen	u
Brett	Rebel Taco	u
Brett	Buena Onda	u
Brett	Bibou	u
Brett	Landolfi's Cafe & Deli	u
Brett	Ummi Dee's burger bistro	u
Brett	WIBS	u
Brett	Delorenzo's Tomato Pies	u
Brett	Diamante Pizzaria	u
Brett	Keshet Kitchen	u
Brett	Poke Bros	u
Brett	Rocky's Pizza and Grille	u
Brett	Via Veneto Pizza	u
Brett	A'Dello Vineyard & Winery	u
Brett	La Collina	u
Brett	Restaurant Alba	u
Brett	Avola Kitchen + Bar	u
Brett	Panera Bread	u
Brett	Mercer Cafe	u
Brett	Stove and Tap	u
Brett	IHOP	u
Brett	Forsythia	u
Brett	WOOJUNG Sushi	u

(a) Recommendations for Brett

name
Kung Fu Tea
Shake Shack

(b) Previously visited restaurants also in recommendations

Figure 13. User - Brett

Incorporating additional data sources or features can potentially improve the recommendation engine's performance. For example, the Yelp dataset also contains check-in data, which records the instances where users have visited businesses. This information can be used to validate the model's recommendations and provide an additional layer of insight into user preferences and behavior. By including check-in data, we can analyze whether users are more likely to visit the recommended restaurants, providing a more practical measure of the recommendation engine's performance.

Moreover, it is crucial to consider the ever-evolving nature of user preferences and the restaurant landscape. Continuous monitoring and updating of the models are necessary to maintain their accuracy and relevance. Exploring other techniques, such as incorporating

	user_name	Restaurant_name	rating
46803	Desirae	Steel Penny Cafe	5.619937
46804	Desirae	Otolith Sustainable Seafood	5.552305
46805	Desirae	The Chilly Banana	5.521576

(a) ALS Recommendations for Desirae

	user_name	Restaurant_name	rating
743433	Brett	Academic Bistro	4.982213
743434	Brett	Otolith Sustainable Seafood	4.952850
743435	Brett	El Primo Produce	4.901747

(b) ALS Recommendations for Brett

Figure 14. ALS recommendations for Desirae and Brett

content-based filtering or leveraging deep learning models, could further enhance the recommendation engine's ability to provide personalized and accurate suggestions.

References

Bandyopadhyay, S., Thakur, S., & Mandal, J. (2021). Product recommendation for e-commerce business by applying principal component analysis (pca) and k-means clustering: Benefit for the society. *Innovations in Systems and Software Engineering*, *17*(1), 45–52.

Bell, R. M., Koren, Y., & Volinsky, C. (2007). Modeling relationships at multiple scales to improve accuracy of large recommender systems. *Knowledge Discovery and Data Mining*.

Joshi, S., & Dubey, J. (2020). Restaurant recommendation system based on novel approach using k-means and naıve bayes classifiers. *Proceedings of the Global AI Congress 2019*, 609–620.

Sarwar, B. M., Karypis, G., Konstan, J. A., & Riedl, J. (2001). Et almbox. 2001. *Item-based collaborative filtering recommendation algorithms. WWW*, *1*.

Yelp. (2022). *Yelp dataset*. https://www.kaggle.com/datasets/yelp-dataset/yelp-dataset

Zhou, Y., Wilkinson, D., Schreiber, R., & Pan, R. (2008). Large-scale parallel collaborative filtering for the netflix prize. *Algorithmic Aspects in Information and Management: 4th International Conference, AAIM 2008, Shanghai, China, June 23-25, 2008. Proceedings 4*, 337–348.