Cocktail Recommender System

Using Real & Synthetic Data for Recommendations Evaluation

Using Machine Learning to Suggest Cocktails

Dharti Seagraves

Date: May 22, 2024

Github

Problem Statement

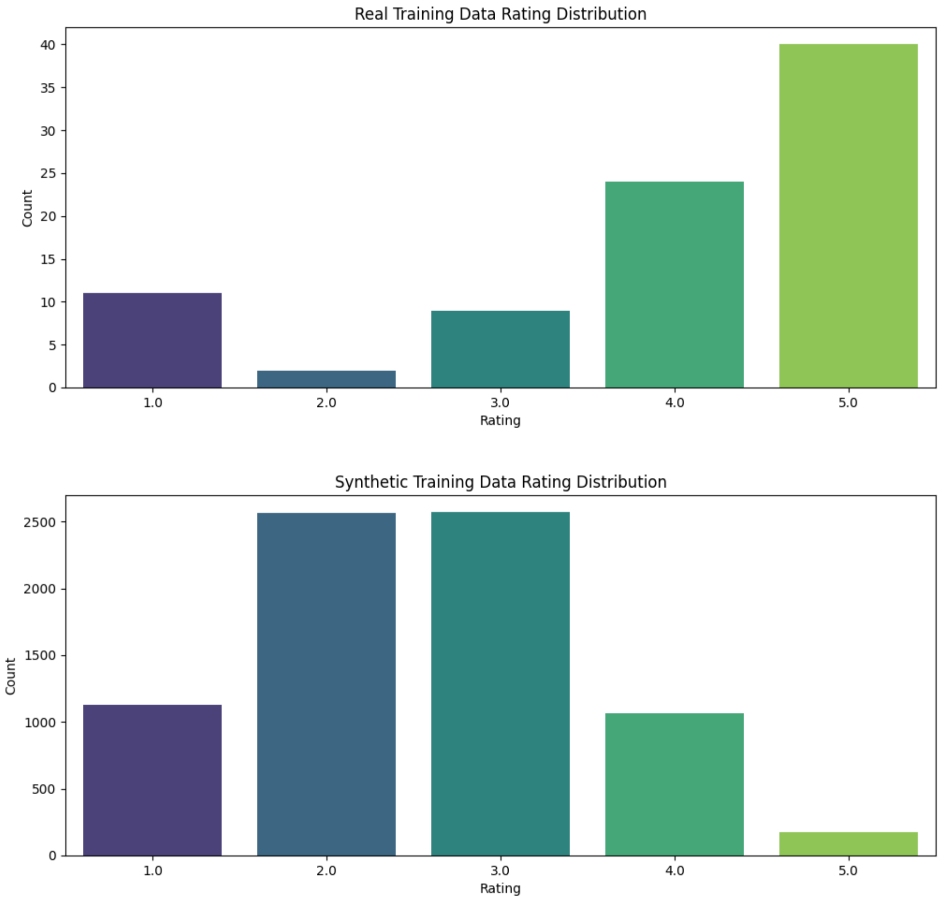
- **Objective**: Create a cocktail recommender system using cocktail features and user ratings data
- **Importance**:
  - Personalized recommendations can enhance user experience and satisfaction in cocktail selection
  - Helps users discover new cocktails tailored to their tastes
  - Can help generate revenue for bars and restaurants
- **Motivation**: Improve customer engagement and satisfaction in the beverage industry

Dataset

- **Cocktail Data Source**: Web scraped from CocktailViz
  - **48** cocktails, **22** features
  - Details on ingredients, alcohol content, and flavor profiles
- **User Ratings Data Source**: Collected via a Google Form survey
  - Included user preferences and ratings on various cocktails
- **Dataset Size**:
  - **Real Data**: 115 records (actual user ratings)
  - **Synthetic Data**: 10,000 records (generated to compare with a larger dataset)

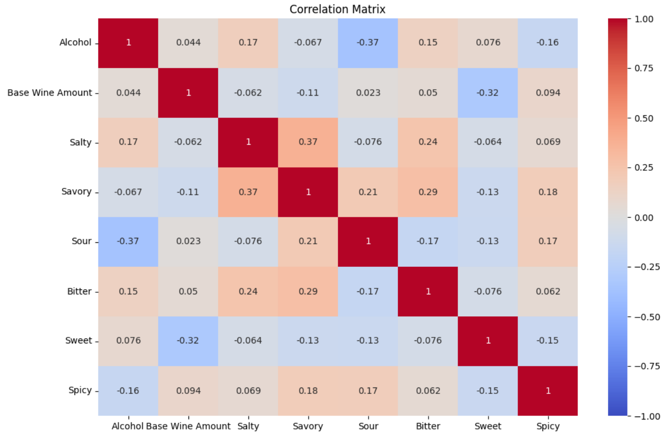
Assumptions and Hypotheses

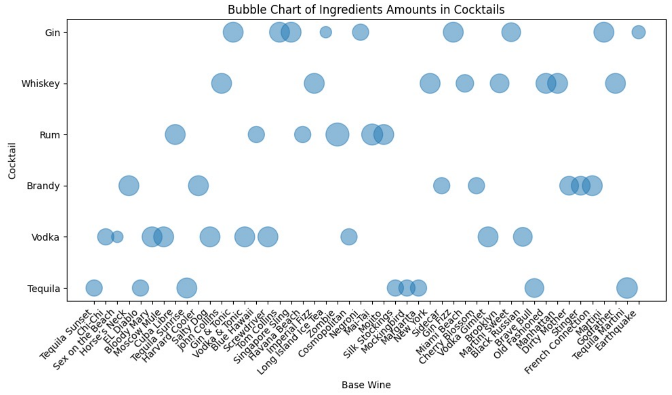
- Assumptions
  - User preferences are consistent over time
  - Survey bias does not exist
  - Synthetic data is a reliable approximation of real user behavior
- Hypotheses
  - Feature importance aligns with user preferences
  - Synthetic data will not replicate real user data because it lacks real-world patterns (used primarily for the quantity of records)



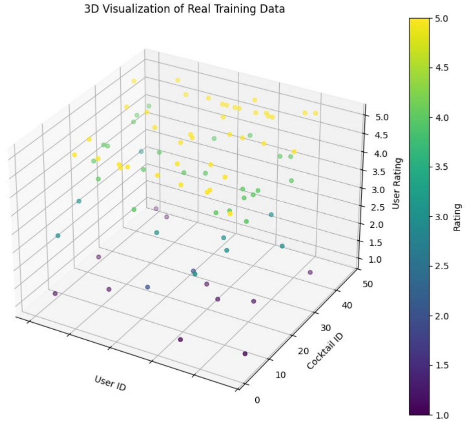
THE UNIVERSITY OF CHICAGO

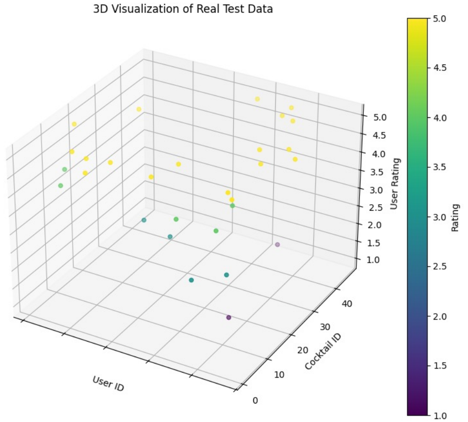
Exploratory Data Analysis





- **Correlation Matrix** (left top)
  - No multi-collinearity
- 117 features after **Feature Scaling**
  - Even distribution of liquors and alcohol content (left bottom)
- **3D distribution** of users, cocktails, and ratings (right)
- Synthetically generated user ratings:
  - Train/test: 7500, 2500
  - Unique Users = 2k





Feature Engineering & Transformations

- **Feature Engineering:** None conducted
- **Transformations**: Normalization, encoding, and scaling techniques applied to improve model performance
  - Standard Scaling numerical values
    - ['Alcohol', 'Base Wine Amount', 'Salty', 'Savory', 'Sour', 'Bitter', 'Sweet', 'Spicy']
  - One-hot encoding categorical values
    - ['Category', 'Making', 'Base Wine', 'Liquor', 'Liquor Amount', 'Juice', 'Juice Amount', 'Spice', 'Spice Amount', 'Soda', 'Soda Amount', 'Others', 'Taste', 'Type of Glass']

- **Feature Importance**:
  - Sweet: 0.000569
  - Juice_Pineapple: 0.000374
  - Alcohol: 0.000354
  - Liquor_Green Mint: 0.000147
  - Base Wine_Rum: 0.000140

Model Development

- **Approaches**:
  - **Item-based Collaborative Filtering**: Recommends items similar to those the user liked
  - **User-based Collaborative Filtering**: Recommends items liked by similar users
  - **Content-based Filtering**: Recommends items similar to those the user has liked based on features
  - **SVD-based Collaborative Filtering**: Uses matrix factorization to predict user preferences
- **Overfitting and Underfitting Checks**:
  - **Learning Curves**: Analyzed to determine if models are learning effectively.

User Profiles (Item-Based)

user_id	Cocktail	user_rating
2	Tom Collins	5
2	Gin & Tonic	4
2	Mojito	4
2	Black Russian	4
2	Manhattan	3
2	Screwdriver	3
2	Cosmopolitan	1
2	Bloody Mary	1
2	Tequila Martini	1

User A (2)	Recommendation 1	Recommendation 2	Recommendation 3
Item Based CF	Blue Hawaii	Vodka & Tonic	Martini
User Based CF	Martini	Moscow Mule	Long Island Ice Tea
Content Based	Horse's Neck	Moscow Mule	Vodka & Tonic
SVD	Margarita	Old Fashioned	Moscow Mule

user_id	Cocktail	user_rating
38	Blue Hawaii	4
38	Tom Collins	4

Alcohol	Name	Category	Making	Taste	Sour	Sweet	Spicy
16	Tom Collins	Long	Shake	Mild	17	17	0
14	Blue Hawaii	Long	Shake	Mild	17	17	0

- **Blue Hawaii:** This cocktail is very similar to the Tom Collins rated highly by user 2
- **Vodka & Tonic:** This drink is similar to Gin & Tonic in the simplicity, base spirit, and carbonation
- **User-similarities:** Item-based also considers how other users rated their drinks; user 38 rated both Tom Collins and Blue Hawaii highly

User Profiles (User-Based)

user_id	Cocktail	user_rating
2	Tom Collins	5
2	Gin & Tonic	4
2	Mojito	4
2	Black Russian	4
2	Manhattan	3
2	Screwdriver	3
2	Cosmopolitan	1
2	Bloody Mary	1
2	Tequila Martini	1

user_id	Cocktail	user_rating
38	Gin & Tonic	5
38	Moscow Mule	5
38	Bloody Mary	5
38	Margarita	5
38	Blue Hawaii	4
38	Tom Collins	4
38	Long Island Ice Tea	4
38	Manhattan	4
38	Tequila Sunset	3
38	Vodka & Tonic	3
38	Screwdriver	1

User A (2)	Recommendation 1	Recommendation 2	Recommendation 3
Item Based CF	Blue Hawaii	Vodka & Tonic	Martini
User Based CF	Martini	Moscow Mule	Long Island Ice Tea
Content Based	Horse’s Neck	Moscow Mule	Vodka & Tonic
SVD	Margarita	Old Fashioned	Moscow Mule

- User 2 is most similar to **User 38** based on the similarity matrix
- **User 2 and User 38** both have highly rated **Tom Collins** and **Gin & Tonic**
- **Moscow Mule** and **Long Island Ice Tea** are recommended because User 38 enjoys these drinks and they share similar preferences with User 2

User Profiles (Content-Based)

- Here we can see the similarities between User 2’s highly rated drinks and the recommended drinks (Category, Making, Taste)

User A (2)	Recommendation 1	Recommendation 2	Recommendation 3
Item Based CF	Blue Hawaii	Vodka & Tonic	Martini
User Based CF	Martini	Moscow Mule	Long Island Ice Tea
Content Based	Horse’s Neck	Moscow Mule	Vodka & Tonic
SVD	Margarita	Old Fashioned	Moscow Mule

User Rating	Alcohol	Name	Category	Making	Base Wine	Base Wine Amount	Soda Amount	Others	Taste	Type of Glass	Salty	Sour	Sweet
5	16	Tom Collins	Long	Shake	Gin	45	200	Lemon Piece	Mild	Highball	17	17	17
4	14	Gin & Tonic	Long	Build	Gin	45	240	Lemon Piece	Mild	Highball	17	83	33
4	25	Mojito	Long	Build	Rum	45	-	Lime/Mint	Mild	Highball	17	33	17
4	32	Black Russian	Long	Build	Vodka	40	-	-	Mild	Rock	0	0	50
-	**10**	**Horse,Äôs Neck**	**Long**	**Build**	**Brandy**	**45**	**200**	**Lemon Piece**	**Mild**	**Old Fashioned**	**17**	**83**	**17**
-	**12**	**Moscow Mule**	**Long**	**Build**	**Vodka**	**45**	**220**	**Lime Piece**	**Mild**	**Highball**	**0**	**83**	**17**
-	**14**	**Vodka & Tonic**	**Long**	**Build**	**Vodka**	**45**	**200**	**-**	**Mild**	**Old Fashioned**	**0**	**83**	**33**

User Profiles (SVD)

- **SVD-based** collaborative filtering leverages matrix factorization to reveal the latent relationships between users and items other than ratings
- **Margaritas** usually have a sweet, sour, and salty flavor profile, similar to User 2's preferences which is why it was highly recommended

User A (2)	Recommendation 1	Recommendation 2	Recommendation 3
Item Based CF	Blue Hawaii	Vodka & Tonic	Martini
User Based CF	Martini	Moscow Mule	Long Island Ice Tea
Content Based	Horse's Neck	Moscow Mule	Vodka & Tonic
SVD	Margarita	Old Fashioned	Moscow Mule

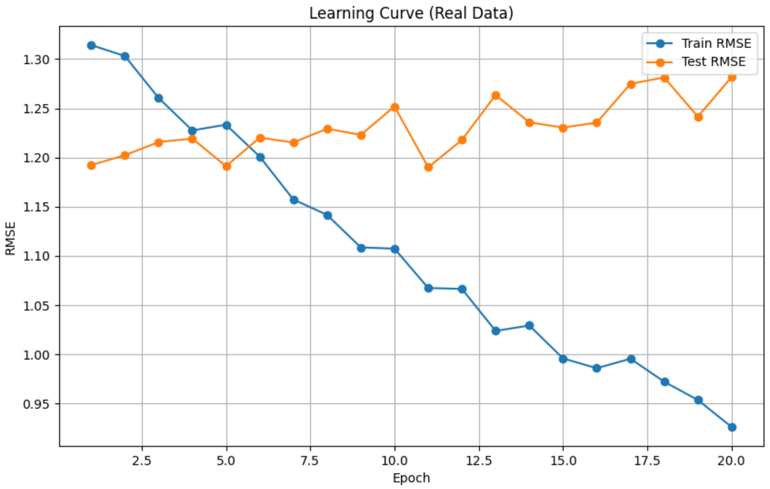
user_id	Cocktail	user_rating
2	Tom Collins	5
2	Gin & Tonic	4
2	Mojito	4
2	Black Russian	4
2	Manhattan	3
2	Screwdriver	3
2	Cosmopolitan	1
2	Bloody Mary	1
2	Tequila Martini	1

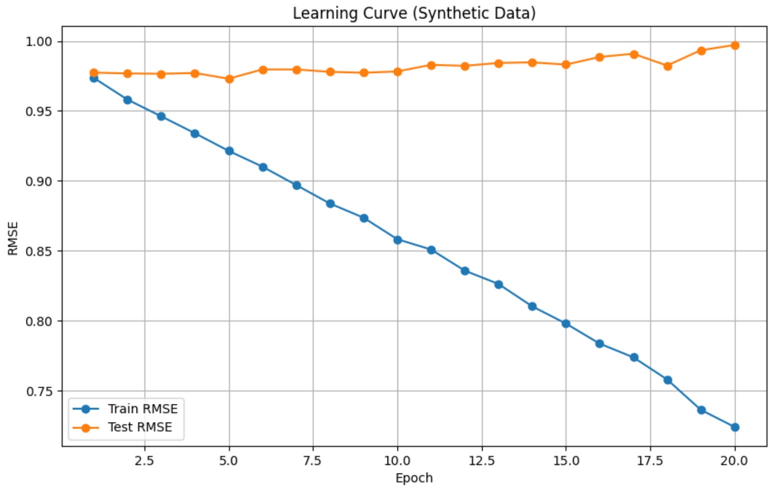
$$\mathbf{A} = \mathbf{U\Sigma V}^{\mathrm{T}}$$

A = User Ratings
Σ = diagonal matrix of singular value
U = User "features" matrix
V^T= Cocktail "features" matrix

Model Evaluation

- **Visualization**: Below are the learning curves for the SVD model on both the real and synthetic datasets for 20 epochs
- **Metrics**: RMSE for training and testing sets
  - Real data - Train RMSE: 1.31 Test RMSE: 1.17
  - Synthetic data - Train RMSE: 0.97 Test RMSE: 0.97
- **Observations**:
  - Real: The training RMSE is steadily decreasing over epochs indicating the model is learning the data set. The testing RMSE remains stable with a slight increase towards the end indicating the data is starting to overfit.
  - Synthetic: Compared to the real data the synthetic data train and test RMSE start at about the same point, due to the nature of less noise/variability in the synthetic data. The test RMSE has a steeper curve indicating the model is learning quicker than with the real data. The test data shows very good generalization with a steady curve.





Results & Learnings

- **Summary of RMSE Values for Each Model:**
  - Item-Based Collaborative Filtering RMSE: 3.86
  - User-Based Collaborative Filtering RMSE: 3.83
  - Content-Based Filtering RMSE: 2.96
  - SVD-Based Collaborative Filtering RMSE: 1.13
- The **best model** for the given real dataset was the **SVD** model with the lowest RMSE. The recommendations made by the SVD model are the closest to the user’s actual ratings
- The latent factors in SVD influence the user’s preference to make more accurate recommendations
- Based on this information the SVD model should be chosen to make cocktail recommendations

Future Work

- **Increase the volume of data**
- **Reduce user bias by randomly showcasing a few cocktails for users to rate versus allowing users to choose**
- **Try advanced algorithms**
  - Regularization and Hyperparameter Tuning
  - Neural Collaborative Filtering
  - Hybrid Models
- **Incorporate Time-Series Elements as user preferences tend to change over time**
- **A/B Testing and continuous improvement**
- **Add a more robust cocktail selection, including non-alcoholic options**

THANK YOU!