Capstone 2: Collaborative Recommendation Model based on User Profiles

Kenneth Truong

Background

- Client is building a food review site or looking to open a restaurant
- What factors entice users to rate a restaurant higher than others?

Deliverables

- Functioning Collaborative Recommendation Model
- Answers the question: what variables predict ratings the most.

Data Wrangling

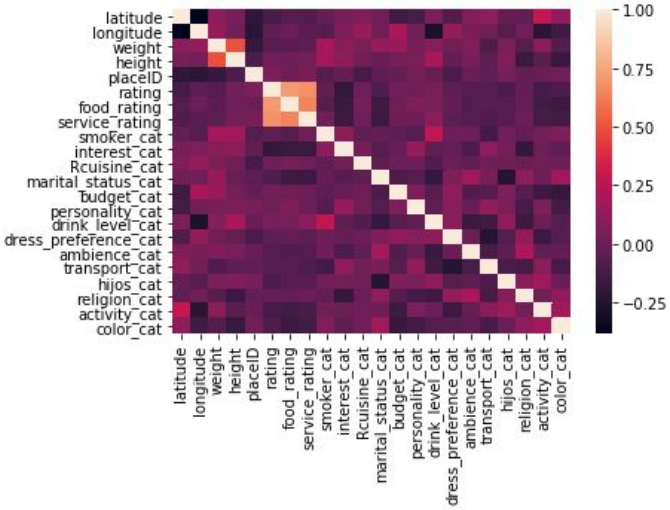
- Load datasets:
  - Used pd.read_csv to load in 3 datasets
- Check DataFrames
- Merge DataFrames
  - First, merged user_profile and ratings on userID, then with cuisine on placeID.

Data Wrangling(cont.)

- Treat missing values
  - Wrote a function to replace ‘?’ with a random value based on the proportion of the population.
- Feature Engineering
  - Wrote a function to change qualitative values to numeric ones
- Select columns that are relevant

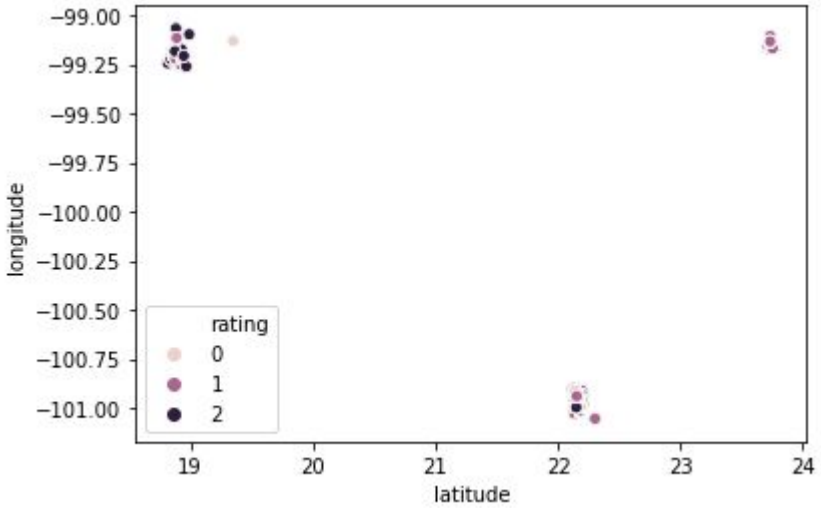
Data Visualizations

- Heat map based on correlation



Data Visualization

- Higher rated restaurants tend to congregate in the same location



Checking Distributions

- Using the chi-square test, I checked normality of height and weight.
- Height was normally distributed, but weight was not.

Machine Learning

- RandomForestClassifier and SelectFromModel
- GridSearchCV was used to tune the hyperparameters of the model.

Random Forest Classification

- R^2 score from the test was 0.559, with a RMSE of 0.826
- Tuned model: R^2 of 0.622 and RMSE of 0.693, with parameters {'max_depth': 155, 'max_features': 6, 'n_estimators': 19}

SelectFromModel

- SelectFromModel was used to select the most predictive variables.
- It selected 'userID', 'latitude', 'longitude', 'weight', 'height', 'placeID', and 'Rcuisine_cat'

Random Forest Classification(after SelectFromModel)

- R^2 score from the test was 0.576, with a RMSE of 0.815
- Tuned model: R^2 of 0.605 and RMSE of 0.753, with parameters {'max_depth': 153, 'max_features': 4, 'n_estimators': 45}

Recommendation System

- Wrote Class CollabReco to recommend restaurants based on other user profiles.
- The RMSE for this was 0.722.