USING MACHINE LEARNING TO IMPROVE QUALITY OF SERVICE IN TRAVEL INDUSTRY

A report on training different machine learning models and evaluating final results

Abstract

This report outlines an almost complete machine learning workflow that uses a publicly available dataset. A companion Jupyter notebook is also available for anyone wishing to review the underlying Python code.

Prepared by: Babak Eslamieh (July 5, 2025)

Contents

Unsupervised Machine Learning – Clustering

This is the final assessment project for course 4 of IBM Machine Learning program.

About the data

We'll be working with Travel Reviews dataset provided by UCI Machine Learning Repository. This data set is populated by crawling TripAdvisor.com and it represents aggregated user ratings on various categories, ranging from restaurants and juice bars to museums and religious institutions. Each data instance represents user ratings for a specific travel destination in East Asia, averaged for each category. Each traveler rating is mapped as Excellent (4), Very Good (3), Average (2), Poor (1), and Terrible (0).

In the original dataset, category features are labeled using generic names (Category 1, Category 2, etc.) and as a first step, original labels were replaced with more descriptive names, guided by dataset description.

Dataset Feature

Note: The labels inside parenthesis represent new feature names.

- User ID : User identifier in string form (e.g. 'User 123')
- Category 1 : Average user feedback on art galleries (art_galleries)
- Category 2 : Average user feedback on dance clubs (dance_clubs)
- Category 3 : Average user feedback on juice bars (juice_bars)
- Category 4 : Average user feedback on restaurants (restaurants)
- Category 5 : Average user feedback on museums (museums)
- Category 6 : Average user feedback on resorts (resorts)
- Category 7 : Average user feedback on parks/picnic spots (parks_picnic)
- Category 8 : Average user feedback on beaches (beaches)
- Category 9 : Average user feedback on theaters (theaters)
- Category 10 : Average user feedback on religious institutions (religious_inst)

Objectives

ABC Company is experiencing a constant decline in sales figures of travel packages for East Asia. Intimidated by most competitors' clever use of AI-powered strategies, the company has hired a small Data Science team to figure out a solution. With a strong motivation to shine in the new department, the team gets busy with scraping user rating pages in tripadvisor.com and comes up with a promising dataset.

Stakeholders at ABC company aren't that impressed, but decide to give their new team a chance. They state some goals and objectives for a new pilot project. The project requirements are summarized below:

1. Using clustering analysis on Travel Reviews dataset, segment travelers based on common tastes and interests.

2. Domain experts will use these segments to design brand new travel packages, tailored to each segment.
3. Final clusters can have different sizes, but the team should keep the count small (4-7 clusters would be nice).
4. The main priority is boosting package sales and getting high customer ratings, so the team should focus on ordinary travelers and ignore possible outliers.

Exploratory Data Analysis

To get familiar with important dataset characteristics, an exhaustive data analysis was made before preparing dataset for modeling. The following sections briefly summarize the results of this analysis.

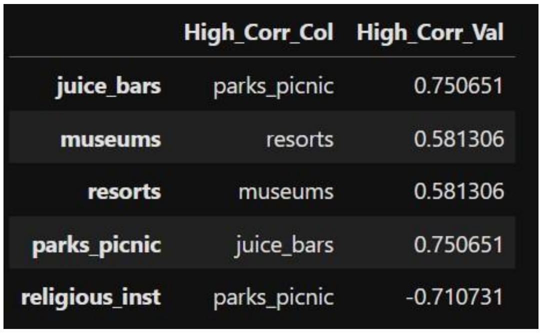
Initial data exploration

Preliminary examination of the dataset revealed the following characteristics:

- Dataset has 980 instances, so it's not a large dataset and training will be relatively fast in all models.
- All features are floating point numbers (float64) rounded up to 2 decimal points, except the first feature (User ID).
- Dataset has 10 numeric features with generic names (e.g., Category 1), meaning we should probably use more descriptive feature names prior to data analysis.
- Dataset does not have missing values in any feature, so data imputing is not required.
- Being averaged from integer values between 0 and 4, most numeric features have fairly similar ranges.
- Despite similar ranges in numeric features, scaling is required due to some features having small ranges (Category 7, Category 8 and Category 10).

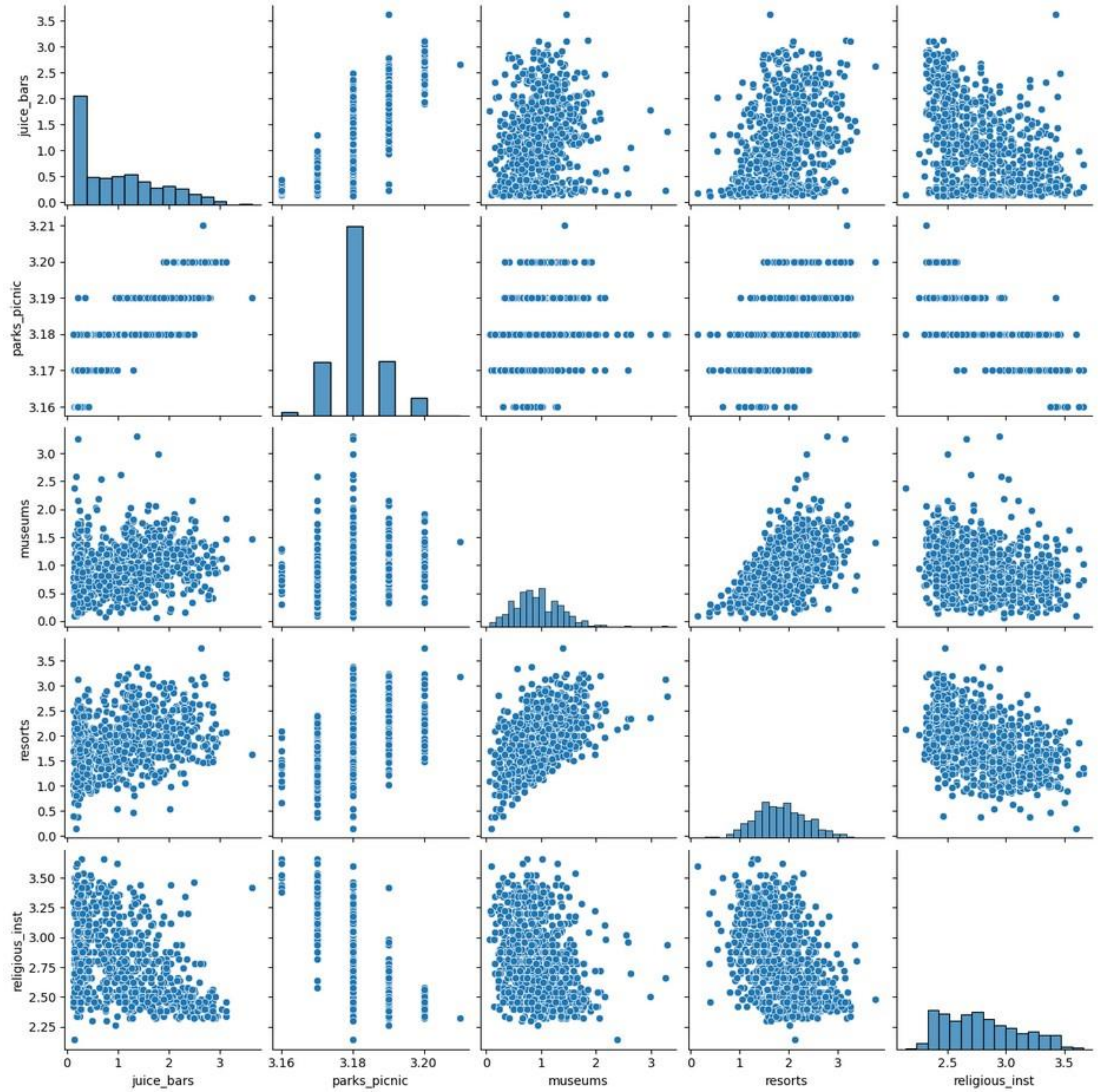
Correlation analysis

Correlation analysis revealed that some significant feature correlations exist, as shown in the following image:



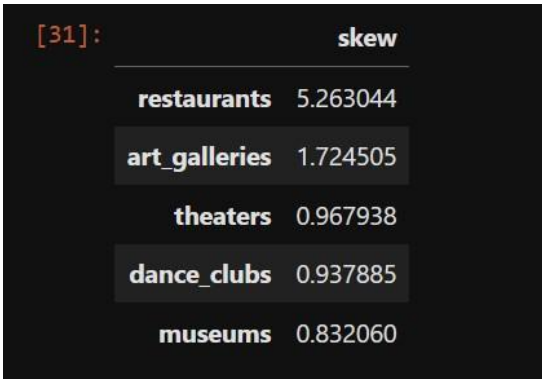
	High_Corr_Col	High_Corr_Val
juice_bars	parks_picnic	0.750651
museums	resorts	0.581306
resorts	museums	0.581306
parks_picnic	juice_bars	0.750651
religious_inst	parks_picnic	-0.710731

The clusters identified by our final model may (or may not) provide some insights about this, but it's important to keep a record of these correlations. We can visually confirm high correlations using the pair plot shown below.



Normality analysis

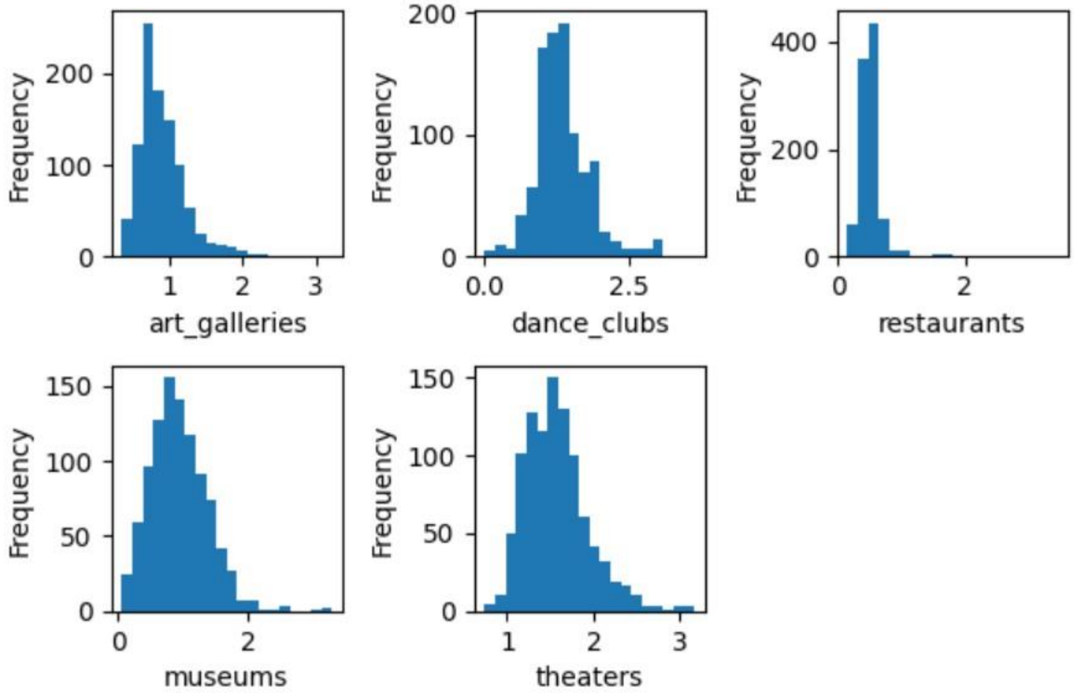
Distributions of a few features in this dataset are highly skewed and, according to the standard skewness threshold (0.75), half of the features are technically skewed, as shown in the following image:



[31]:

	skew
restaurants	5.263044
art_galleries	1.724505
theaters	0.967938
dance_clubs	0.937885
museums	0.832060

The following image shows different levels of skewness in histograms.



Outlier analysis

Since we're doing clustering analysis in this project, we're dealing with two types of outliers, both of which don't merit further action:

- Outliers in feature values: All features have a limited (small) range and extreme values doesn't happen in such small ranges.
- Outliers in clusters: Although DBSCAN will detect possible outliers, the ABC company is more focused on travel package sales and high satisfaction rates from typical customers. So, there is no specific plan for picky customers that may be hard to please.

Feature selection and engineering

Our exploratory data analysis revealed the need for some data preparation steps. These steps are summarized below.

Feature selection

The User ID feature was removed, as it doesn't have any modeling value.

Feature engineering

- All features were renamed according to the main dataset description. This step was performed prior to data analysis.
- All feature values were scaled using MinMaxScaler.

Clustering models

In order to find the suitable number of clusters suggested by business objectives (i.e., 4-7) several models were trained. Modeling steps can be summarized as follows.

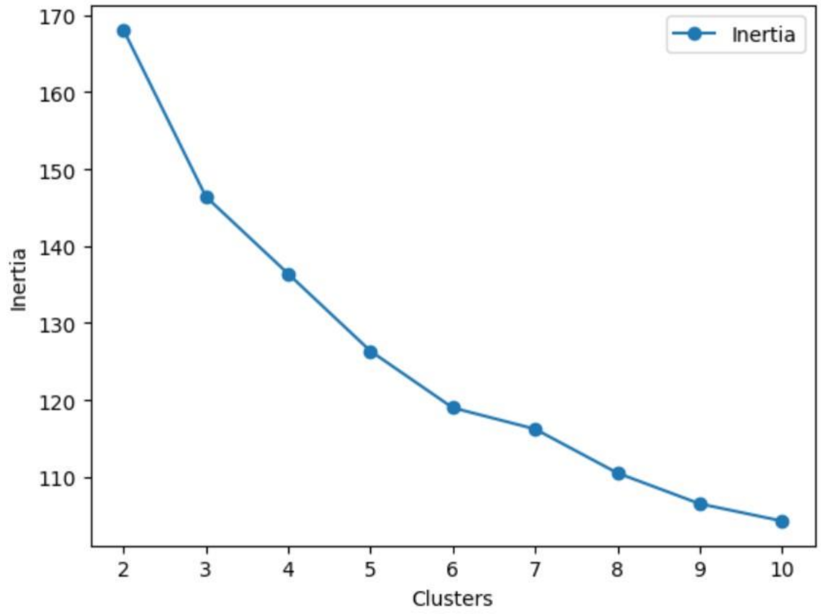
- A range of cluster counts (from 2 to 10) were used to train several K-Means models. Using the Elbow Method, the optimal cluster count was estimated.
- A wide range of epsilon and n_clu values were used to train several DBSCAN models. The counts of clusters and outliers were aggregated and final models were evaluated based on arbitrary criteria: n_clusters < 10 and n_outliers < 10 (roughly 1% of all data points)
- A single Mean Shift model was trained using the estimated bandwidth.

Results

Our modeling result is summarized in the following sections.

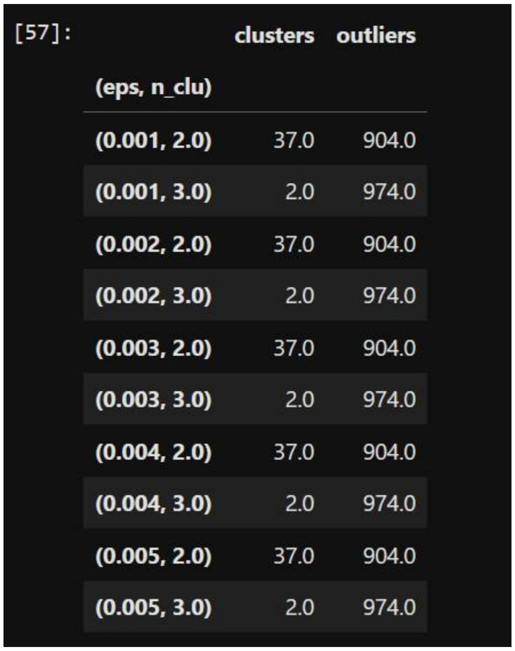
K-Means

As shown in the following image, we couldn’t easily find an inflection point for elbow method. However, to better align with our business objectives, we can choose K = 6 as our best K-Means model.



DBSCAN

One can verify from the following (sample) results that DBSCAN could not lead to optimal clusters with this dataset. This can be due to the infamous Curse of Dimensionality. Many ranges of smaller *epsilon* values were also used, always with similar results.



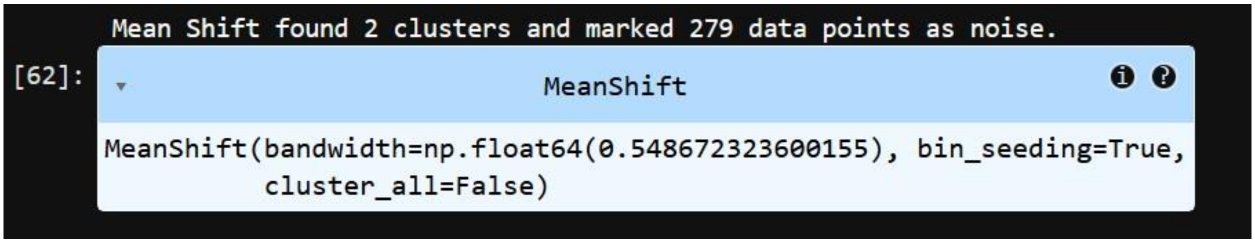
[57]:

(eps, n_clu)	clusters	outliers
(0.001, 2.0)	37.0	904.0
(0.001, 3.0)	2.0	974.0
(0.002, 2.0)	37.0	904.0
(0.002, 3.0)	2.0	974.0
(0.003, 2.0)	37.0	904.0
(0.003, 3.0)	2.0	974.0
(0.004, 2.0)	37.0	904.0
(0.004, 3.0)	2.0	974.0
(0.005, 2.0)	37.0	904.0
(0.005, 3.0)	2.0	974.0

Ultimately with current dataset shape, DBSCAN fails to provide good results.

Mean Shift

Finally, our Mean Shift clustering result is shown in the following image.



```
Mean Shift found 2 clusters and marked 279 data points as noise.
```

```
MeanShift(bandwidth=np.float64(0.548672323600155), bin_seeding=True,
          cluster_all=False)
```

Comparing our business objectives and the result shown above, we can verify that Mean Shift found too few clusters and too many outliers. So, our Mean Shift model is not acceptable.

Best model

Based on the above analysis, our best clustering model is **K-Means with 6 clusters**.

Insights and key findings

After training different models and comparing results, we can make some notes about the quality of our modeling workflow and hidden patterns in this dataset.

- Several high correlations between some features may have degraded the overall performance of our models. Also, repeating the process after transforming skewed features may lead to better results in all models.
- K-Means model resulted in a final model that aligns well with our business objectives. However, with a 10-dimensional feature space, we can't get a sufficiently accurate visual cue about the shape and density of final clusters.
- Thinking about the inner logic of DBSCAN clustering, the constant output of the algorithm may be due to varying cluster densities in the dataset. Also, high dimensionality of our dataset may force data points to appear too far apart to distance-based algorithms. Repeating the process with lower dimensions (2D or 3D) may be worth trying.
- As implemented in Scikit-Learn, the Mean Shift model doesn't allow hyperparameter tuning and we're forced to stick with a single estimated bandwidth. To get the most out of this model, some more study and research may be helpful.

Next steps

Based on the trained models and data preprocessing steps, we may be able to improve data quality and plan for some more experiments. Hopefully, these actions will lead to models that perform better and make room for easier interpretation.

Possible faults

No dimensionality reduction was incorporated in our data preprocessing. This may be vital for the following reasons:

- Our dataset may be noisy and it's usually good practice to apply techniques like PCA analysis before modeling.
- Because most clustering algorithms are distance-based, using a small feature space may provide much better results.

Action plan

Repeating the whole modeling with a more carefully refined dataset is definitely worth trying. The following actions can be taken to improve our preprocessing pipeline:

- Apply PCA with a range of components to account for high correlations and possible noise.
- To enable cluster visualization, prefer a PCA transform with 2 or 3 components.
- Try other approaches, like Multi-Dimensional Scaling (MDS), to shrink dataset’s feature space.

Acknowledgements

Thank you so much for taking the time to review and grade this project.

I’d also like to thank our brilliant instructors (especially Dr. Joseph Santarcangelo) for their excellent learning materials. It’s been a real pleasure taking this course.