Cognifyz Machine Learning Internship

Full Project Report

Author: Adarsh V H
Project Type: Machine Learning · Data Engineering · Model Explainability · Recommender Systems

---

Abstract

This project delivers a complete machine learning pipeline built on a large international restaurant dataset.
The work includes exploratory data analysis (EDA), feature engineering, regression modelling for rating prediction, SHAP-based model interpretability, a content-based recommendation engine, and a multi-class cuisine classification system.

The final solution provides actionable restaurant insights, including predicted ratings, explanations for model decisions, and similarity-based recommendations.

---

1. Introduction

Restaurants generate rich but noisy datasets: cuisine types, ratings, votes, cost structures, locations, and delivery attributes.
This internship project focuses on transforming such data into **intelligent predictive and analytical systems**.

The project consists of:

1. Exploratory Data Analysis (EDA)
2. Preprocessing & Feature Engineering
3. Rating Prediction Model

4. Model Interpretability
5. Content-Based Recommendation Engine
6. Cuisine Classification Model

Each task builds upon the previous one to form a complete ML pipeline.

---

2. Task 1 — Exploratory Data Analysis

The raw dataset contained 9,551 entries across multiple countries. After cleaning invalid rows, 7,403 remained for machine learning.

2.1 Key Insights

- **Ratings** are compressed between 3.0–4.0 → bias toward mid-range.
- **Cost for Two** is heavily right-skewed → log transform reveals structure.
- **Votes strongly correlate** with rating → major predictive feature.
- **Top cuisines** include North Indian, Chinese, Fast Food, Bakery, Café.
- **Top cities** dominated by Delhi–NCR region → geographic imbalance.
- **Correlation heatmap** shows weak linear relationships → non-linear models necessary.

2.2 Visuals Created

- Rating histogram + KDE
- Boxplots for cost & rating
- Log cost distribution
- Votes vs Rating (linear & log-scale)

- Top cuisines & city frequency charts
- Correlation heatmap

All visuals stored in /visuals/.

---

3. Task 1 — Preprocessing & Feature Engineering

This stage transforms raw, messy data into clean numeric features usable for ML.

3.1 Cleaning Steps

- Removed unnecessary text fields (e.g., address, locality text)
- Converted Yes/No fields → binary 1/0
- Ensured zero missing values
- Removed non-rated rows

3.2 Feature Engineering

- Extracted **Primary Cuisine** from multi-cuisine strings.
- Grouped rare cuisines (< 10 samples) into **Other**.
- One-hot encoded cuisine groups → 45+ binary flags.
- Encoded Country Code to categorical integer.
- Created **City_Freq**, a frequency-based city importance signal.
- Preserved important numeric signals: cost, votes, rating, price range.

3.3 Final Output

A final dataset with:

- **7,403 rows**

- **53 engineered features**

Saved as:

data/processed/model_data.csv

This dataset powers every ML task in the project.

---

4. Rating Prediction Model

A supervised regression pipeline was built to predict restaurant ratings.

4.1 Baseline Models Evaluated

- Linear Regression
- Decision Tree Regression
- RandomForest Regression

RandomForest performed best due to:

- Handling non-linear relationships
- Robustness to feature noise
- Feature importance interpretability

4.2 Hyperparameter Tuning

Using RandomizedSearchCV, the best parameters were:

n_estimators = 700

max_depth = 20

min_samples_split = 10

max_features = 'sqrt'

bootstrap = True

4.3 Final Model Performance

R^2 = 0.626

MAE = 0.256

RMSE = 0.339

This is strong performance given rating compression (3.0–4.0) and limited numeric signals.

---

5. Model Interpretability (SHAP)

To understand *why* the model makes predictions, SHAP explainability was applied.

5.1 Global Insights

Top contributors:

1. **Votes**
2. **City_Freq**
3. **Average Cost for Two**
4. **Price Range**
5. **Cuisine_Group Flags**

This aligns perfectly with intuition: rating credibility increases with vote volume.

5.2 Local Explanation

For individual restaurants:

- High votes → pushes predicted rating upward
- Low cost / low votes → pushes rating downward

- Some cuisines act as minor positive or negative modifiers

Local force plots and waterfall plots were generated and saved.

---

6. Task 2 — Content-Based Recommendation Engine

A similarity-based restaurant recommender was developed.

6.1 Objective

Given a restaurant index:

- Predict its rating using the tuned RF model
- Compute cosine similarity between restaurants
- Retrieve **top 5 most similar restaurants**
- Display city, cuisines, and names
- Ensure mapping between processed index → raw dataset

6.2 Features Used for Similarity

- Average Cost
- Votes
- Delivery flag
- Table Booking flag
- Price Range
- Country Code
- Cuisine one-hot flags

Scaled values → cosine similarity.

6.3 Output Example

Selected Restaurant:

Name: Ikreate

City: New Delhi

Cuisine: Bakery

Predicted Rating: 3.17

Similar Restaurants:

1. A Pizza House (Similarity: 0.67)

2. Tpot (0.66)

3. Pandit Dhaba (0.65)

...

6.4 Deliverables

- Full modular code under /src/task2/
- Metadata loader
- Recommender functions
- CLI interface for interactive use

---

7. Task 3 — Cuisine Classification

A supervised multi-class classification pipeline.

7.1 Objective

Predict cuisine category based on numeric features alone.

7.2 Results

Accuracy ≈ 24%

Weighted F1 ≈ 23%

7.3 Interpretation

Low accuracy is expected due to:

- Extreme class imbalance
- Weak numeric representation of cuisine concepts
- Many cuisines having < 5 examples
- No text features used

Large cuisines (North Indian, Cafe, Chinese) performed best.

7.4 Improvement Strategies

- Collapse rare cuisines into "Other"
- Use TF-IDF cuisine embeddings
- Apply class balancing
- Train hierarchical cuisine models
- Use advanced models (XGBoost, LightGBM)

---

8. Conclusion

This internship project successfully built a complete machine learning ecosystem:

✓ Fully cleaned & engineered dataset
✓ Rating prediction model with strong performance
✓ SHAP explainability for transparency
✓ Content-based recommendation engine
✓ Cuisine classification system

✓ Modular and reusable ML codebase

✓ High-quality visualizations & documentation

The project demonstrates practical ML engineering, model interpretability, dataset handling, and structured problem-solving skills directly relevant to real-world data science and ML roles.