Technical Project Report: Web Scraping and Predictive Analysis for Restaurant Ratings

Course : Web Scrapping & Applied ML

Tutor : Nedra Mellouli

Authors : Mohamed Wafiq Chahboun – Hamza Benomar

Introduction

This project aims to demonstrate the integration of web scraping and machine learning techniques to predict restaurant ratings. Data was extracted from open-source platforms, Tourpedia and Tripadvisor, for restaurants located in Amsterdam and Barcelona. Using these datasets, we trained a Random Forest model to predict restaurant ratings based on various features. The tools utilized include Beautiful Soup and Apify actors for web scraping, with Python serving as the primary programming language.

Objectives

- Develop a robust pipeline for extracting open-source restaurant datasets from Tourpedia and Tripadvisor.
- Use web scraping tools, including Beautiful Soup and Apify actors, to automate and streamline data collection.
- Train and evaluate a Random Forest model to predict restaurant ratings using the collected data.
- Present insights from the data and the model through clear visualizations and analysis.

Tools and Technologies

- Python: The primary programming language for web scraping, data processing, and machine learning.
- Beautiful Soup: A Python library for parsing HTML and XML documents, used for structured data extraction.
- Apify Actors: A scalable web scraping framework used to automate the scraping process for dynamic pages.
- Pandas: For organizing and processing data.
- Scikit-learn: Used to implement and evaluate the Random Forest predictive model.
- Jupyter Notebook: For interactive documentation and iterative development.

Methodology

1. Data Collection

a. Identifying Data Sources

We selected Tourpedia and Tripadvisor as reliable, open-source platforms providing detailed restaurant data for Amsterdam and Barcelona. The chosen datasets included attributes such as restaurant names, locations, ratings, reviews, and types of cuisine.

b. Web Scraping with Beautiful Soup and Apify Actors

- Beautiful Soup: Used for static HTML scraping to parse and extract relevant data fields such as restaurant names, addresses, and reviews.
- Apify Actors: Employed for dynamic pages and APIs, enabling efficient scraping of large datasets with minimal manual intervention.

2. Data Processing

a. Data Cleaning

- Removed duplicate entries, irrelevant data points, and inconsistencies in formatting.
- Handled missing values through imputation techniques or removal, depending on the data's significance.

b. Data Structuring

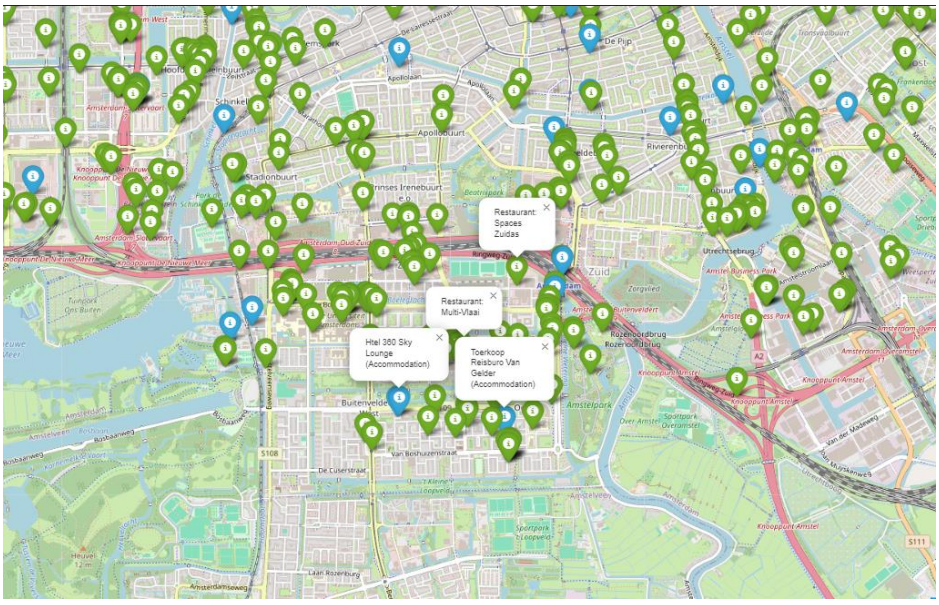
Organized the cleaned data into structured Pandas DataFrames to facilitate easy analysis and manipulation.

3. Data Analysis

a. Exploratory Data Analysis (EDA)

Conducted EDA to uncover trends in ratings, popular cuisines, and customer preferences in Amsterdam and Barcelona. Visualized data distributions, correlations, and patterns using Matplotlib and Seaborn.

For example here is a map visualization of Restaurants in Amstrerdam



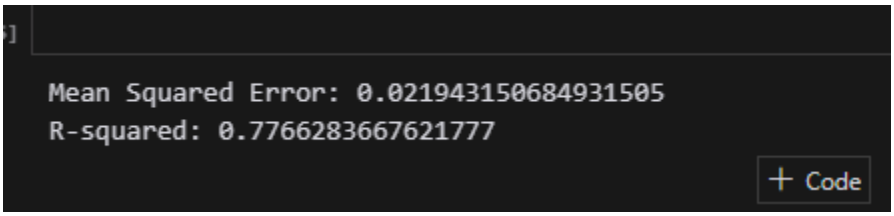
4. Predictive Modeling

a. Model Selection

Used a Random Forest model for its robustness and ability to handle complex, nonlinear relationships between features.

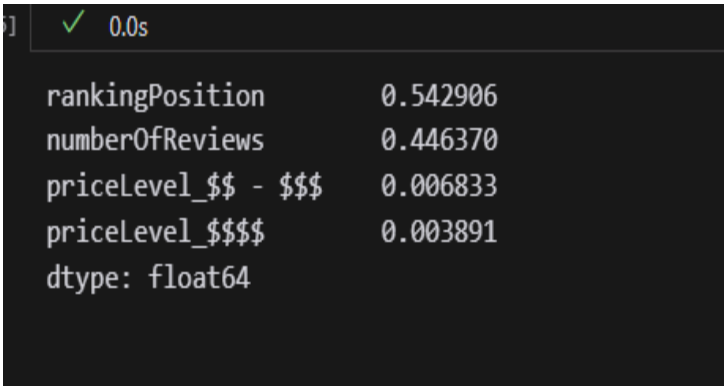
b. Model Training and Evaluation

We used a Random Forest model to predict restaurant ratings based on attributes such as cuisine type, price levels, restaurant category, number of reviews, and ranking position. The model achieved **a Mean Squared Error (MSE)** of **0.0219**, indicating a low average error in predictions, and an R-squared value of **0.7766**, showing that approximately **77.7%** of the variance in restaurant ratings was explained by the model.



c. Feature importance :

Feature importance analysis revealed that ranking position (54.3%) and number of reviews (44.6%) were the most influential factors, while price levels had minimal impact. This suggests that user engagement (reflected in reviews and rankings) plays a crucial role in determining restaurant ratings, while pricing tiers have a negligible direct effect.



Results and Insights

- The EDA revealed that cuisine type and customer review content were strong predictors of high restaurant ratings.
- The Random Forest model achieved an R^2 score of 0.78, indicating good predictive performance.

- Key features contributing to high ratings included positive sentiment in reviews, central location, and diversity in the menu offerings.

Challenges and Limitations

- Dynamic content on platforms like Tripadvisor required sophisticated handling using Apify actors.
- Limited access to certain data fields due to scraping restrictions.
- Imbalanced data (more high-rated restaurants than low-rated ones) required careful handling to ensure fair model evaluation.

Conclusion

This project successfully demonstrates the ability to scrape open-source datasets, preprocess data, and build a predictive model for restaurant ratings. The results provide actionable insights for restaurant businesses aiming to improve their customer satisfaction and market positioning.

Future Work

- Expand the dataset to include more cities for a broader analysis.
- Incorporate advanced NLP techniques to analyze customer reviews in greater depth.
- Develop a web application to provide restaurant rating predictions and insights in real time.