Machine Learning Capstone Project Report

Project Title: Restaurant Rating Prediction using Machine Learning and Natural Language Processing

1. Introduction

This project presents a comprehensive end-to-end machine learning solution aimed at predicting restaurant ratings using a combination of structured restaurant metadata and unstructured textual customer reviews. The motivation behind this project is to showcase practical skills required in real-world data science roles, including data cleaning, feature engineering, natural language processing, statistical analysis, model building, evaluation, and interpretation. By integrating NLP with classical machine learning models, the project bridges the gap between text analytics and predictive modeling.

2. Business Problem & Objective

Online food delivery and restaurant discovery platforms depend heavily on user-generated ratings to influence consumer behavior and restaurant visibility. However, ratings are often subjective, noisy, or inconsistent due to individual bias. The core business objective of this project is to predict restaurant ratings more reliably using historical data, thereby enabling platforms to improve ranking algorithms, recommendation systems, and customer trust. From a business perspective, accurate rating prediction helps platforms highlight quality restaurants and allows restaurant owners to understand drivers behind customer satisfaction.

3. Dataset Overview

The dataset contains detailed information about restaurants, reviewers, costs, cuisines, timestamps, and textual customer reviews. It combines both structured and unstructured data, making it suitable for advanced feature engineering and NLP tasks. Key attributes include restaurant name, reviewer identity, cost for two, cuisine types, review text, rating values, and review timestamps. The diversity and richness of the dataset allow the model to capture both quantitative trends and qualitative customer sentiments.

4. Data Preprocessing

Data preprocessing was a critical step in ensuring model reliability and robustness. Missing numerical values were handled using appropriate statistical imputation, while categorical variables were cleaned and encoded. Textual reviews underwent extensive preprocessing including contraction expansion, lowercasing, punctuation removal, URL elimination, stopword removal, tokenization, normalization, and part-of-speech tagging. These steps ensured that the text data was transformed into a clean and analyzable format.

5. Feature Engineering

Several domain-driven and statistical features were engineered to improve model performance. Numerical features such as review length, logarithmic cost, and engagement metrics were derived. Cuisine information was transformed using one-hot encoding, while high-cardinality features like restaurant and reviewer names were label encoded. Textual data was vectorized using TF-IDF to capture word importance across reviews. This multi-level feature engineering enabled the model to learn complex relationships in the data.

6. Statistical Analysis & Hypothesis Testing

Statistical techniques were applied to validate assumptions and uncover meaningful patterns. Independent Samples t-tests (Welch's t-test) were conducted to examine whether factors such as cost categories or review engagement levels had a statistically significant impact on ratings. These hypothesis-driven analyses strengthened the analytical foundation of the project and complemented machine learning findings.

7. Machine Learning Models

Multiple machine learning models were implemented and compared, including baseline regression models and advanced tree-based ensemble techniques. The goal was to evaluate trade-offs between interpretability and predictive power. Hyperparameter tuning was performed to optimize model performance, and cross-validation was used to ensure generalization across unseen data.

8. Model Evaluation & Explainability

Models were evaluated using metrics such as Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R-squared score. Feature importance analysis helped interpret the final model by identifying key predictors such as cost, review length, and cuisine type. This explainability aspect is crucial for stakeholder trust and business decision-making.

9. Deployment Considerations & Future Scope

The trained model can be serialized and deployed as a backend service for real-time rating prediction. Potential future enhancements include incorporating deep learning-based NLP models, sentiment analysis, real-time streaming data, and recommendation system integration. Scalability and model retraining pipelines can further enhance production readiness.

10. Conclusion

This project demonstrates a complete machine learning lifecycle, integrating data preprocessing, statistical analysis, NLP, feature engineering, and predictive modeling. It highlights practical problem-solving skills, analytical thinking, and readiness for real-world machine learning roles. Overall, the project effectively balances technical depth with business relevance.