Exploring Predictive Models in Restaurant Inspection Outcomes and Ratings

Marco Ayala, Rohit Mohnani, Raghav Muthukumar

CSCI 1951A

Introduction

Recent studies in public health and consumer ratings have shown how restaurant characteristics can influence inspection outcomes and consumer satisfaction. Our research applies machine learning techniques, specifically Logistic Regression and K-Nearest Neighbors (KNN), to predict health violations and customer ratings based on multiple factors including neighborhood demographic and restaurant specifics.

Data

We utilized two primary datasets: health violations and restaurant ratings across various boroughs. Features included 'Borough', 'Cuisine Description', 'Community Board', 'Council District', 'Census Tract', and 'Borough's Median Income'. Data was cleaned and preprocessed for analysis, involving one-hot encoding of categorical variables and merging datasets on borough names.

Logistic Regression Methodology

Methodology for Logistic Regression (Health Violations Prediction):

- **Model Choice:** Logistic Regression was chosen for its suitability in binary classification tasks.
- D**ata Preprocessing:** Included binary encoding of the 'CRITICAL_FLAG' and merging datasets for income levels.
- **Model Configuration:** Employed a logistic regression model with a maximum of 1000 iterations and strong regularization (C=0.1), utilizing Sklearn's implementation.
- **Validation Technique:** K-fold cross-validation with 10 splits to ensure model robustness.

Logistic Regression Methodology

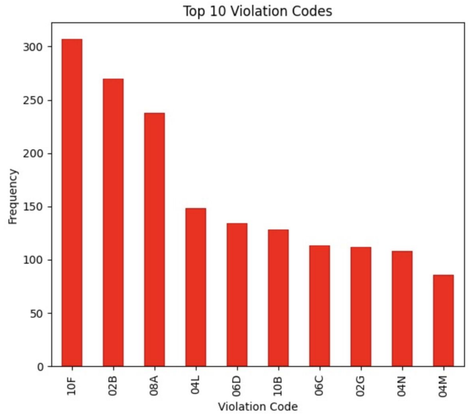
Methodology for KNN (Rating Prediction):

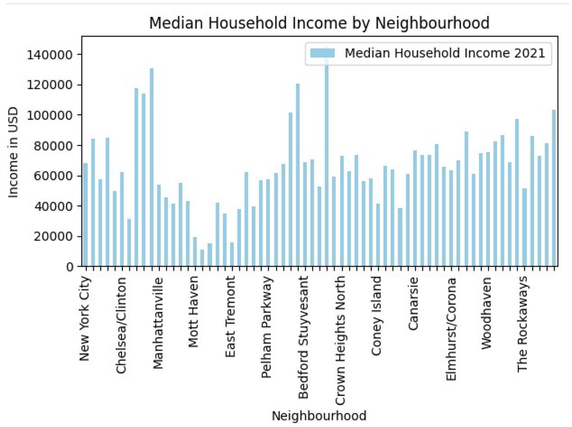
- **Model Choice:** KNN was selected due to its non-parametric nature and flexibility in handling a variety of relationships between features and the target variable.
- **Hyperparameter Tuning:** Adjustments made to the 'K' value to balance model accuracy and complexity.
- **Validation Technique:** Employed 10-fold cross-validation to assess the generalizability of the model.

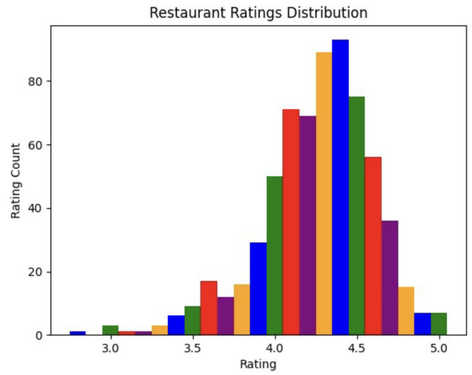
Hypothesis testing

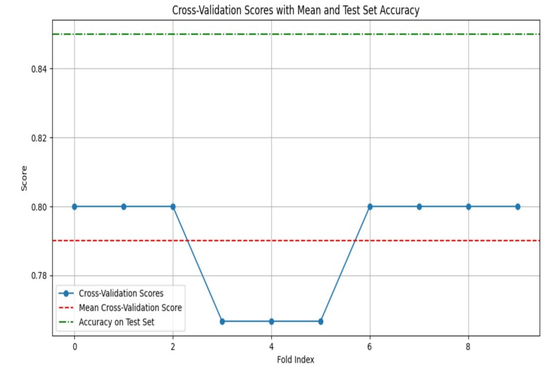
- Hypothesis 1: Cannot reject the Null (T test single sample))
- Hypothesis 2: Cannot reject the Null (Z test one sample)
- Hypothesis 3: Cannot reject the Null (Chi Square Independence))

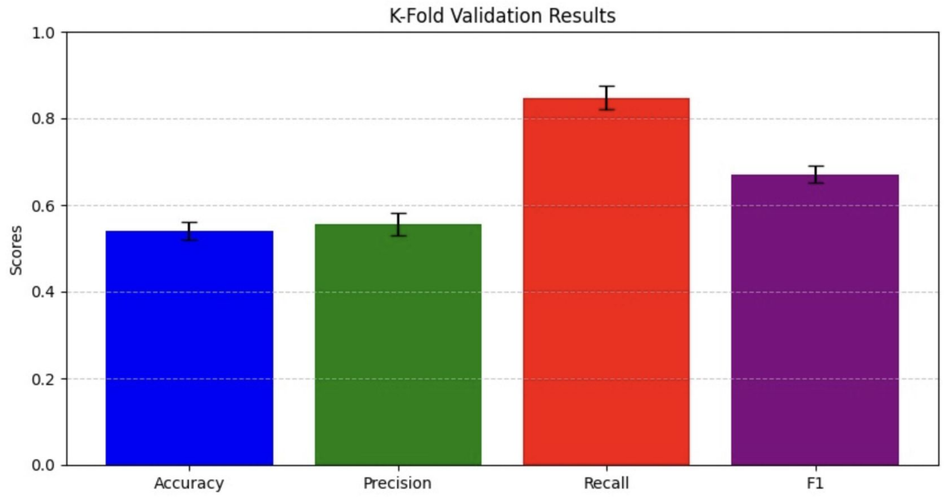
Results











Conclusion

Our investigation into the predictive capabilities of Logistic Regression and K-Nearest Neighbors (KNN) models on restaurant health violations and customer ratings yielded insightful findings that enhance our understanding of the factors influencing restaurant performance from a public health and consumer satisfaction perspective.

Predictive Performance of Logistic Regression:
The Logistic Regression model's results reveal that the model is somewhat effective but with room for improvement. The high recall rate suggests the model is capable of identifying a large proportion of actual positive cases (true health violations), yet the precision rate indicates a relatively high number of false positives. This suggests that while the model can predict violations, it often flags non-violating instances as problematic, which could lead to unnecessary inspections or concerns.

Effectiveness of KNN in Rating Predictions:
The application of the KNN model for predicting restaurant ratings showcased better performance, achieving an impressive accuracy of 85% on the test set and a mean cross-validation score of 0.79. This high level of accuracy suggests that KNN effectively captured the complex, multi-dimensional relationships among the features related to restaurant ratings. The model's robustness in handling diverse data inputs and its flexibility in feature relationship modeling contributed significantly to its success.

Methodological Insights

Impact of Feature Selection:
A critical takeaway from our research is the significant impact of feature selection on model performance. Initially, factors such as median income and cuisine type were presumed to have substantial influences on health violation predictions. However, these features did not perform as expected, suggesting that they may not be as predictive of health violations as hypothesized. This outcome emphasizes the necessity for a more nuanced understanding of which features most accurately reflect the underlying dynamics of health violations and customer satisfaction.

Methodological Insights:
The study also highlighted the importance of methodological considerations in predictive modeling. While Logistic Regression provided a solid baseline model, its limitations in handling the complex and nuanced nature of the data suggest that exploring more sophisticated algorithms or ensemble methods could potentially yield better results. Similarly, the success of the KNN model underscores the value of non-parametric methods in scenarios where traditional assumptions about data distribution do not hold.