Abstract

Fifa Merchants, [mayala7, rmhonani, rmuthuk1]

Goal:

The objective of this project is to predict the likelihood of restaurant health violations and the ratings customers give to restaurants using logistic regression and K-nearest neighbors (KNN). We hypothesize that factors such as neighborhood socio-economic status, cuisine type, and administrative district boundaries have significant predictive power over both health inspection outcomes and customer ratings.

Data:

The dataset, obtained from the New York City Health Department and various online rating platforms, spans several years and includes attributes like borough, cuisine type, and health violation records. Additional socio-economic factors were incorporated to enhance predictive accuracy. Data challenges we ran into involved cleaning duplicate records, encoding categorical variables, and mitigating biases related to inspection frequency and socio-economic status, which required careful preprocessing to ensure robust model inputs.

Model+Evaluation Setup:

For predicting health violations, logistic regression was used for its proficiency in binary classification tasks, with model performance measured through accuracy, precision, recall, and F1 score metrics. K-nearest neighbors (KNN) was chosen for predicting customer ratings due to its ability to handle non-linear relationships among variables, evaluated through accuracy and cross-validation scores to assess generalizability. The division of data into training and test sets was strategically planned to include unseen data in the test phase to evaluate the models' predictive power on new, unobserved data.

Results and Analysis:

Claim #1: Logistic Regression demonstrates substantial predictive capability with an 80.42% recall in identifying true health violations.

Support for Claim #1: This high recall rate is crucial for public health safety, indicating the model's strength in correctly identifying actual violations, which is vital for regulatory purposes. The performance was substantiated through confusion matrix analysis and ROC curve assessments, highlighting the model's sensitivity.

Claim #2: Precision for the Logistic Regression model is relatively low at 56.30%, indicating a notable number of false positives.

Support for Claim #2: The precision-recall trade-off suggests the model might be overly conservative, marking non-violating instances as violations. This observation necessitates further model tuning or exploration of additional features to enhance precision without sacrificing recall.

Claim #3: KNN achieved impressive accuracy, predicting restaurant ratings with 85% accuracy on the test set.

Support for Claim #3: This high level of accuracy, supported by consistent cross-validation scores averaging 0.79, indicates that KNN effectively captures the complex dynamics influencing customer satisfaction. This model's success demonstrates its suitability for handling diverse and multifaceted data landscapes.

Claim #4: Logistic Regression showed signs of overfitting, with degraded performance on the test set compared to the training set.

Support for Claim #4: Learning curves indicated increasing divergence between training and test accuracy as training progressed, pointing to overfitting. This finding suggests the need for implementing regularization techniques to balance model complexity and training data fidelity.

This project not only underscores the utility of logistic regression and KNN in predictive analytics within the public health and service sectors but also highlights the importance of sophisticated data handling and model evaluation strategies to tackle real-world problems effectively

Socio-Historical Context and Impact Report.

Socio-Historical Context

Our project on predicting restaurant health violations and customer ratings via Logistic Regression and K-Nearest Neighbors (KNN) intersects with significant socio-historical factors that influence both the collection and interpretation of data. Historically, health inspection data has been a critical component in maintaining public health standards. An effective health inspection regime can prevent disease outbreaks and ensure food safety, which are key concerns in urban settings where population density is high and public health risks are significant.

The major stakeholders in this project include restaurant owners, patrons, health inspectors, and the broader community living within the neighborhoods studied. For restaurant owners, predictions of health violations can inform better practices and compliance, while for consumers, these predictions are vital for making informed choices about where to eat. Health inspectors could use such predictive models to optimize inspection schedules, focusing resources on higher-risk establishments based on predictive factors.

Existing research, such as the studies on the impact of neighborhood demographics on health code compliance, shows that socioeconomic factors significantly influence health inspection outcomes. This understanding is crucial for our project as it highlights the importance of considering these factors in our predictive models. The societal impact of such research is profound, offering potential for more targeted and efficient public health interventions.

The socio-historical context significantly affects our project's framing. For instance, understanding that neighborhoods with lower median incomes may experience different rates of health violations suggests a need for tailored interventions that address not just individual restaurants but community-wide issues. This recognition impacts how we analyze our data, urging a consideration of socio-economic factors as potential confounders or moderators in our analysis.

Ethical Considerations

This project raises several ethical considerations, particularly concerning data bias and privacy. Given the historical and societal biases possible within our data sources:

Data Biases: There is a potential for inherent biases in health inspection data, which may reflect historical socio-economic discrimination. Poorer or ethnically diverse neighborhoods might have been subjected to stricter scrutiny or higher frequencies of inspections, not solely based on objective health standards but influenced by biased policies.

Data Collection Systems: The systems used for collecting health inspection data and customer ratings may also possess biases. It is crucial that these systems do not disproportionately target or neglect specific groups or areas. Ensuring uniform application across all neighborhoods is necessary to prevent these biases.

Interpretation Biases: Interpretations of data could perpetuate stereotypes, particularly if certain types of cuisines or restaurant setups are assumed to be more prone to health violations. Our analytical methods must carefully consider these aspects, ensuring that any predictive modeling does not reinforce existing stereotypes.

Data Privacy: The aggregation and analysis of data must respect privacy. Ensuring data anonymization and preventing any possible re-identification of individuals or specific establishments is critical, complying with privacy laws and ethical guidelines.

Misuse of Data: There is a risk that our project's results could be misinterpreted or misused, potentially leading to reputational damage for restaurants or unnecessary public alarm. Clear communication regarding the limitations and scope of our predictions is essential to mitigate these risks.

Own Perspective's Influence: Our own identities, backgrounds, and perspectives could influence our analysis. Recognizing and critically assessing how these factors shape our research approach and interpretations is essential for maintaining objectivity and ethical integrity.

Citations

Pechey, Rachel, and Pablo Monsivais. “Socioeconomic inequalities in the healthiness of food choices: Exploring the contributions of food expenditures.” Preventive medicine vol. 88 (2016): 203-9. doi:10.1016/j.ypmed.2016.04.012

(OCR), Office for Civil Rights. “Standards for Privacy of Individually Identifiable Health Info.” *HHS.Gov*, 28 June 2021, www.hhs.gov/hipaa/for-professionals/privacy/guidance/standards-privacy-individually-identifiable-health-information/index.html.

“Health Information Privacy Law and Policy.” *HealthIT.Gov*, 1 Sept. 2022, www.healthit.gov/topic/health-information-privacy-law-and-policy.

“Astho Legislative Prospectus: Data Modernization.” *ASTHO*, 9 Dec. 2022, www.astho.org/advocacy/state-health-policy/legislative-prospectus-series/data-modernization/.

Venkataramani, Atheendar S., et al. “Economic Influences on Population Health in the United States: Toward Policymaking Driven by Data and Evidence.” *PLOS Medicine*, Public Library of Science, journals.plos.org/plosmedicine/article?id=10.1371%2Fjournal.pmed.1003319. Accessed 15 May 2024.

Guo, H., Li, X., Li, W. et al. Climatic modification effects on the association between PM1 and lung cancer incidence in China. BMC Public Health 21, 880 (2021). https://doi.org/10.1186/s12889-021-10912-8