Analyzing Restaurant Health Inspections and Reviews to Predict Food Safety Risks

Group 13: Siddhant Sukhani, Sydney Tang, Curran Mitra, Justin Zhang, Aman Patel, Chaitanya (Rinky) Yetukuri

1 INTRODUCTION

Restaurant food safety remains a pivotal challenge in our nation due to complex regulations, uneven enforcement, and limited public access to inspection data. Current literature tackling this problem fails to present their work in a unified, user-friendly format. Our project addresses this gap by combining historical inspection reports with advanced machine learning models to predict future health scores. In doing so, we provide an intuitive, map-based web platform where users can quickly evaluate a restaurant's safety rating within Atlanta. This integrated system fosters better-informed dining decisions and encourages proactive compliance, ultimately enhancing public health transparency.

2 PROBLEM STATEMENT

The predictive task is estimating the health inspection score of a restaurant based on its historical data and previous violations. The specific historical data chosen was the days between inspections, number of violations, previous health score, Follow Up Needed, Foodborne Illness Risk and Inspection Purpose.

Our approach integrates restaurant inspection data with machine learning to predict health scores and track compliance trends. We aim to uncover hidden relationships that provide actionable insights for consumers, public health officials, and restaurant owners. This method enhances transparency, making health inspection information more accessible while enabling businesses to improve compliance before inspections occur. Ultimately, this project benefits civilians by allowing them to make safer dining choices, assists public health officials in identifying trends, and helps restaurant owners better understand the factors influencing their ratings through a finalized interactive visualization similar to TabNet-GRA [3].

3 LITERATURE SURVEY

Current research on health inspections and food safety compliance has explored Artificial Intelligence [16], big data approaches [6], and Machine Learning [11] for predicting violations. Novel methods such as graph networks [18], sentiment analysis [4], and Naïve Bayes Algorithms [9] have also been observed in this domain. Researchers have analyzed consumer perceptions of food safety through Yelp (mentioned in [9]) and Facebook reviews [17], finding a correlation between negative reviews and poor health inspection results [13]. Additionally, research suggests that more stringent regulations do not always lead to better compliance [10]. However, despite these insights, restaurant health inspection reports remain difficult for the public to access and interpret [8]. While studies have examined food safety violations and consumer sentiment separately, few have combined multiple data sources to establish a predictive model for health scores incorporating factors like review sentiment, location, and rule complexity. This gap limits public awareness, timely interventions, and proactive compliance efforts.

Potential risks include bias in reviews [1], inspector subjectivity [7], and geographical limitations [14], which could limit the accuracy of our work. As used by [5, 12, 15], online reviews may reflect factors unrelated to food safety, such as service quality or pricing, while inspector subjectivity can lead to inconsistencies in health scores. However, the payoff is substantial—enhanced transparency in restaurant health standards and better-informed consumers. Ultimately, if successful, this project can transform how restaurants are evaluated by consumers by shifting the focus beyond subjective reviews to more data-driven insights, allowing diners to make informed choices based on both regulatory health scores and real-world customer experiences.Additionally, it can improve society's access to health safety measures by making inspection data more transparent, actionable, and easily accessible, fostering greater accountability among restaurants and encouraging higher food safety standards across the industry. We can measure gains in consumer awareness by tracking website/app traffic, including visits, page views, and interactions with the health inspection dashboard. Additionally, analyzing search queries, time spent on pages, and repeat users will help assess engagement, while

post-launch surveys can gauge changes in consumer awareness of restaurant health scores.

4 PROPOSED METHODS

4.1 Intuition

Quite a lot of work has been done with AI and ML for food handling and safety, but hardly anything has been done to predict restaurant safety scores. The novelty of our work is what will make it the new state-of-the-art in the area. Using powerful state-of-the-art models like random forests and gradient-boosting classifiers, we will be able to accurately predict future restaurant health and safety scores. This is what will make our work better than the current non-existent state of the art in the field.

In contrast to previous work that focuses on health inspection data alone or on sentiment analysis of reviews in isolation, our approach integrates multiple heterogeneous data sources, such as inspection records, violation history, and spatial attributes, into a unified predictive framework. This not only improves prediction accuracy,, but also provides actionable insights that are immediately interpretable through visual dashboards and geospatial maps. Moreover, unlike most existing solutions that remain in academic prototypes or are inaccessible to the public, our project delivers a fully functional, user-facing platform. This bridge between advanced analytics and public accessibility gives our work practical utility and positions it as a significant step forward in real-world food safety monitoring.

4.2 Approaches

Data Overview: The initial dataset consisted of health inspection records with the following columns: `Inspection ID`, `Item`, `Type`, `Facility`, `Address`, `City`, `State`, `Zip Code`, `Date`, `Permit Number`, `Score`, `Grade`, `Purpose`, `Risk Type`, `Last Score`, `Last Grade`, `Last Date`, `Prior Score`, `Prior Grade`, `Prior Date`, `Follow Up Needed`, `Follow Up Date`, `Foodborne Illness Risk`, `Date Time In`, and `Date Time Out`.

Feature Engineering: To enhance model performance, we engineered three new features and refined existing ones. The first is the days between inspections, which is calculated as the number of days since the last inspection. The second is the number of violations, which is engineered from individual violation records per inspection. The third is the last score, which is included directly from historical data.

Final Features and Target: The final feature set used for modeling comprised a mix of encoded, categorical, and numeric variables. The independent variables included the `Inspection ID`, along with binary-encoded indicators for `Foodborne Illness Risk` and whether a `Follow-Up` was needed. The `Item` field was one-hot encoded, while `Risk Type` remained as a categorical variable. The `Purpose` of inspection was grouped into two categories: 'Routine' and 'Others'. Additional numeric variables included the `Number of Days Between Inspections`, the `Last Score` received, and the `Number of Violations` noted. The target variable for the model was the `Score`, which was treated as a continuous numeric value.

Handling Missing Data: The dataset initially contained 71,986 records. After removing entries with missing values, the final dataset comprised 70,715 records, ensuring consistency and reliability in model training.

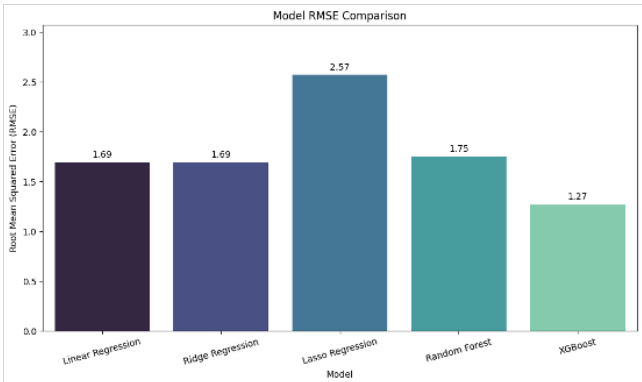

Figure 1: Different scores of the models we tested

Model Selection: To predict the restaurant health scores, we trained and evaluated several models including: Linear Regression (Regular, Lasso, Ridge), Random Forest, and XGBoost. We employed hyperparameter tuning for all our models to improve performance. We see the results below. We compare all the models on their root mean squared error, for which we notice considerably excellent performance for XGBoost (`n_estimators=250`, `learning_rate=0.1`, `max_depth=6`). This model will hence be used as our predictive model for all our visualizations. XGBoost [2] often outperforms other methods because it uses a gradient-boosting

framework with decision trees, capturing complex non-linear relationships while efficiently handling missing data. Its built-in regularization helps prevent overfitting, and the iterative boosting process refines predictions in each round, making it highly effective for noisy, real-world datasets like health inspection records.

Feature Importance: We can see the importance of the various features with respect to the model's performance below. The most important factor, `Foodborne Illness Risk`, is pivotal as we notice that this significantly impacts the health rating of a given restaurant. The other features that we see being important are the number of violations and whether a follow-up is needed or not.

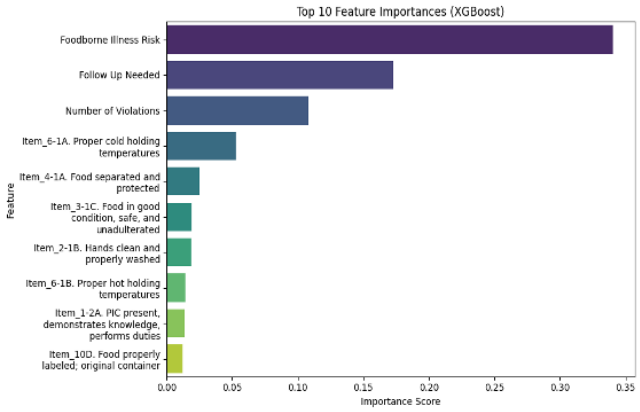

Figure 2: Feature importance visualised

Map Visualization: Our team parsed raw data using Python, then used the US Government's Census Bureau and Google Maps API Geocoders to assign latitude and longitude values to all full street addresses from the data source. We then aggregated the new, enhanced data in CSV format for visualization in Tableau. From this, we created a color-coded map of Atlanta, where users can explore the safety scores of nearby restaurants at a glance. Markers indicate individual restaurants, colored by predicted or actual score ranges, and users can filter by rating, zip code, or cuisine. This spatial representation of food safety data improves both usability and comprehension.

5 WEBSITE

Public health data is typically presented in dense formats like large databases or technical reports that are not intuitive for everyday users. To enhance public access to food safety information, our team developed an interactive platform that combines data-driven analysis with user-friendly visualization. Our approach integrates advanced predictive modeling, geospatial mapping, and intelligent recommendation systems to empower both consumers and health inspectors. It also features a minimalist, responsive design with clear navigation. Users can toggle between map view, score lookup, and recommendation tools. We also included a "Report an Issue" form to collect user-reported discrepancies or violations, which can be used to improve model accuracy or flag unreported safety concerns.

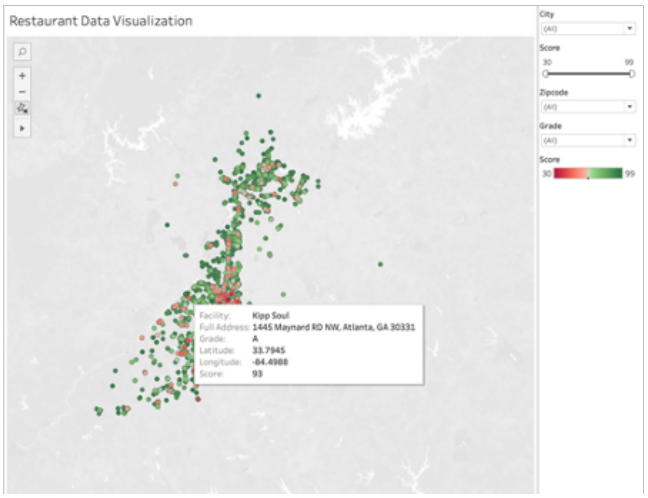

Figure 3: Map of our results of restaurants across Atlanta.

6 EVALUATION

To evaluate the success of our platform, we will track user engagement metrics such as page views, time spent on predictive and mapping tools, and the frequency of restaurant recommendations used. Additionally, the team designed a Qualtrics survey and distributed it to students, friends, and family members. We collected 60+ responses across the following categories: User Demographics, User Motivation, and User Feedback. Some of the key results from the survey are displayed in the below figures.

From figure 4, we see the vast majority of users reported that they found the website intuitive and easy to navigate, which is a good sign for the general release of our tool. Another interesting finding in figure 6 is that

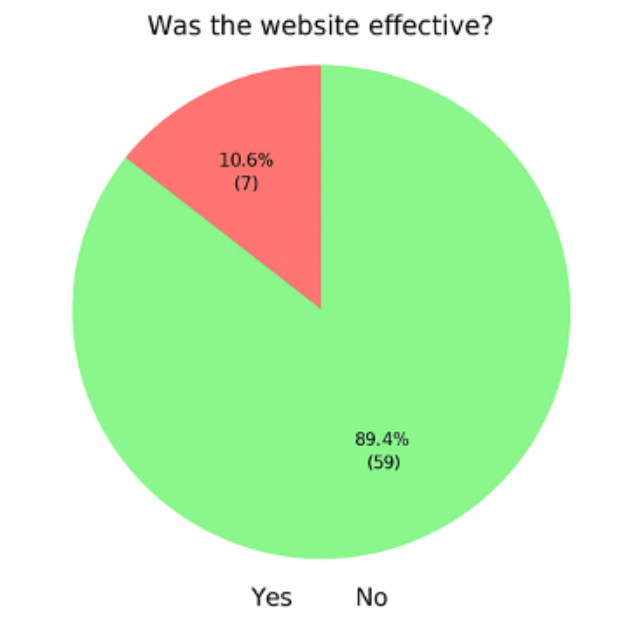

Figure 4: Pie chart of our survey

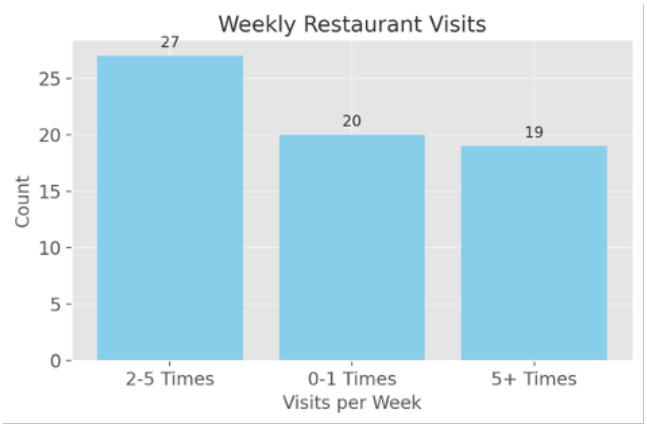

Figure 5: Visits to restaurants on a weekly basis.

the top motive for users to use our tool was to find new restaurants, rather than using it for its primary purpose of preventing foodborne illness. This may require an additional survey to understand why users are using our tool to find restaurants instead of major sites like Google Maps and Yelp. As for users' health conditions and concerns which are displayed in Figure 5, it turns out that most users did not report having a health issue that may be a risk when eating out, implying that they may be using our tool to choose restaurants with generally good sanitation and hygiene. Lastly, the majority of our users as indicated in Figure 7 eat at restaurants at least twice a week, indicating that there is a large audience for our tool to reach.

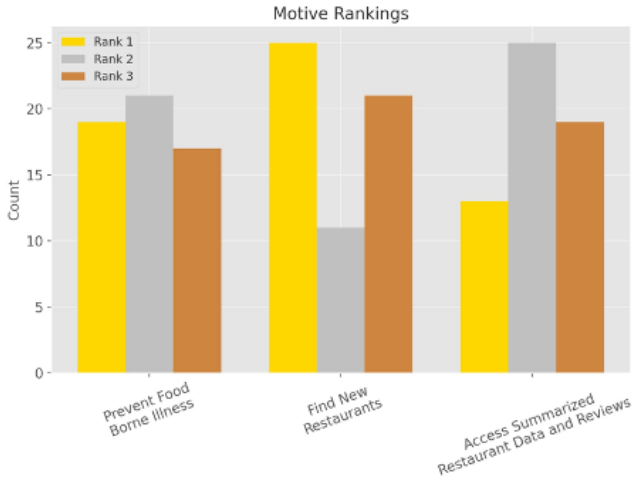

Figure 6: Motivation for using our platform.

It is too early to fully evaluate the accuracy of the predictive tool outside of its test set error, as health inspections in Atlanta are typically conducted once or twice per year per restaurant. Since our model was recently deployed, we will need to wait for more health inspection data to be released to accurately validate our tool. Moving forward, we plan to track inspection results and user engagement over the coming months and refine our model accordingly.

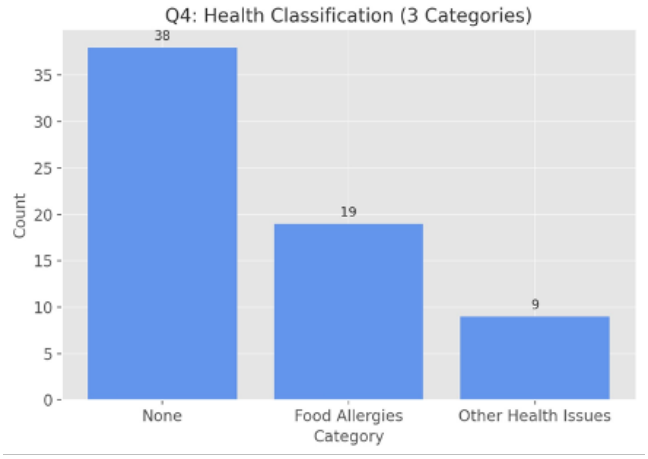

Figure 7: Health risks classified based on survey

7 CONCLUSION

This project demonstrates how machine learning and geospatial visualization can be used to predict restaurant health inspection scores and present food safety information in a more accessible, interactive format. Using features like `Foodborne Illness Risk`, violation count, and follow-up needs, our XGBoost predictive model achieved strong performance with an average error of just 1.4 percentage points. The interactive web interface and map-based dashboard allow users to explore local health data intuitively, supporting safer dining decisions and proactive compliance by restaurant owners. However, our system's full impact is contingent on the frequency of public inspection updates, meaning long-term validation will require additional time and data. Future work may involve incorporating social media signals, expanding the geographic scope beyond Atlanta, and refining user engagement strategies based on ongoing feedback. Additionally, the model may need further analysis to prevent biases from arising. Overall, our work lays the foundation for a scalable, data-driven approach to improving food safety awareness and public health transparency.

All team members have contributed a similar amount of effort.

REFERENCES

[1] S. Abdullah, P. Van Cauwenberge, H. Vander Bauwhede, and P. O'Connor. 2024. Review ratings, sentiment in review comments, and restaurant profitability: Firm-level evidence. *Cornell Hospitality Quarterly* 65, 3 (2024), 378–392. https://doi.org/10.1177/19389655231214758

[2] Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 785–794.

[3] Y. Chen, H. Li, H. Dou, H. Wen, and Y. Dong. 2023. Prediction and visual analysis of food safety risk based on TabNet-GRA. *Foods* 12, 16 (2023), 3113. https://doi.org/10.3390/foods12163113

[4] Q. Gan, B. H. Ferns, Y. Yu, and L. Jin. 2016. A text mining and multidimensional sentiment analysis of online restaurant reviews. *Tourism Analysis* 21, 5 (2016), 465–492. https://doi.org/10.1080/1528008X.2016.1250243

[5] Y. Jiang. 2020. Restaurant reviews analysis model based on machine learning algorithms. In *2020 Management Science Informatization and Economic Innovation Development Conference (MSIEID)*. IEEE, Guangzhou, China. https://doi.org/10.1109/MSIEID52046.2020.00038

[6] C. Jin, Y. Bouzembrak, J. Zhou, Q. Liang, L. M. van den Bulk, A. Gavai, N. Liu, L. J. van den Heuvel, W. Hoenderdaal, and H. J. P. Marvin. 2020. Big data in food safety: A review. *Current Opinion in Food Science* (2020). https://doi.org/10.1016/j.cofs.2020.11.006

[7] A. C. Johnson, B. A. Almanza, and D. C. Nelson. 2014. Factors that influence whether health inspectors write down violations on inspection reports. *Food Protection Trends* 34, 4 (2014), 226–237. https://www.foodprotection.org/files/food-protection-trends/Jul-Aug-14-Johnson.pdf

[8] T. F. Jones and K. Grimm. 2008. Public knowledge and attitudes regarding public health inspections of restaurants. *American Journal of Preventive Medicine* 34, 6 (2008), 510–513. https://doi.org/10.1016/j.amepre.2008.01.035

[9] H. Kang, S. J. Yoo, and D. Han. 2012. Senti-lexicon and improved Naïve Bayes algorithms for sentiment analysis of restaurant reviews. *Expert Systems with Applications* 39, 5 (2012), 6000–6010. https://doi.org/10.1016/j.eswa.2011.11.107

[10] D. W. Lehman, B. Cooil, and R. Ramanujam. 2019. The effects of rule complexity on organizational noncompliance and remediation: Evidence from restaurant health inspections. *Journal of Management* 46, 8 (2019), 1436–1468. https://doi.org/10.1177/0149206319842262

[11] R. A. Oldroyd, M. A. Morris, and M. Birkin. 2021. Predicting food safety compliance for informed food outlet inspections: A machine learning approach. *International Journal of Environmental Research and Public Health* 18, 23 (2021), 12635. https://doi.org/10.3390/ijerph182312635

[12] B. Shin, S. Ryu, Y. Kim, and D. Kim. 2022. Analysis on review data of restaurants in Google Maps through text mining: Focusing on sentiment analysis. *Journal of Multimedia Information Systems* 9, 1 (2022), 61–68. https://doi.org/10.33851/JMIS.2022.9.1.61

[13] M. Siering. 2020. Leveraging online review platforms to support public policy: Predicting restaurant health violations based on online reviews. *Decision Support Systems* 143 (2020), 113474. https://doi.org/10.1016/j.dss.2020.113474

[14] S. Singh, B. Shah, C. Kanich, and I. A. Kash. 2021. Fair decision-making for food inspections. arXiv. https://arxiv.org/abs/2108.05523

[15] S. Wong, H. Chinaei, and F. Rudzicz. 2016. Predicting health inspection results from online restaurant reviews. arXiv. https://arxiv.org/pdf/1603.05673

[16] W. Yu, Z. Ouyang, Y. Zhang, Y. Lü, and C. Wei. 2024. Research progress on the artificial intelligence applications in food safety and quality management. *Trends in Food Science & Technology* 112 (2024), 104855. https://doi.org/10.1016/j.tifs.2024.104855

[17] K. Zahoor, N. Z. Bawany, and S. Hamid. 2020. Sentiment analysis and classification of restaurant reviews using machine learning. In *2020 21st International Arab Conference on Information Technology (ACIT)*. IEEE. https://doi.org/10.1109/ACIT50332.2020.9300098

[18] Y. Zhang and J. Li. 2022. Food inspection data analysis system based on knowledge graph. In *2022 7th International Conference on Intelligent Computing and Signal Processing (ICSP)*. 1120–1124. https://doi.org/10.1109/ICSP54964.2022.9712685