Predicting Food Insecurity at the County Level in New York State

Jonathan Burns

jonathan.burns54@spsmail.cuny.edu

0.1 ABSTRACT

This study utilizes two machine learning models to try and predict yearly food insecurity on a county level, which pulls in a range of economic, demographic and social predictors across each of New Yorks 62 counties from 2020-2024. The model comparison yielded one clear winner, the boosting algorithm XGBoost was far superior to the recurrent neural network LSTM, boasting an impressive R^2 of 0.94 and a mean absolute percentage error (MAPE) of 9.0%. In the most performant XGBoost model, childhood poverty, median income and school lunch enrollment accounted for the highest variance in food insecurity across counties. From there, predicted food insecurity levels in counties were classified as high risk, medium risk and low risk. Highlighting predicted high-risk counties can help targeted interventions on a county level instead of wide-sweeping state policies that may not work for every demographic in every county.

1. INTRODUCTION

Hunger numbers stubbornly high for three consecutive years as global crises deepen: UN report (2024, July 24) reported that in 2023, around 2.33 billion people faced moderate to severe food insecurity, with 864 million people going without food for over an entire day or more. And despite having largest economy for over a century, Feeding America reports that in 2023, 47 million people in the United States, including 14 million children, faced food insecurity. (2024, September 4)

COVID-19 led to a massive initial increase in food insecurity both globally and in the US. However, heightened governmental assistance across the board, as well as people returning to work through 2021 aided in quashing the initial explosion seen in 2020. Through 2023, as COVID-19 assistance dwindled and inflation soared, food insecurity among US households would have the same fate, Urell, A. (2023a, October 30).

Despite New York's economic dominance nationally, the New York State Department of Health estimates that about one in four adults reported experiencing food insecurity in New York State during 2023. New York also faces a unique challenge; high cost of living, severe income inequality and unequally dispersed resources make understanding food insecurity at a county level even more important.

As a basic human need, food insecurity cripples New Yorkers in every aspect of their lives. The National Institutes of Health (NIH) cite food insecurity as a major contributor to several chronic health conditions, mental health disorders and many other challenges (link since deprecated).

Access to quality food is also essential to economic growth both personally, statewide and nationally, as sustained food insecurity can lead to nearly insurmountable long-term physical, mental and behavioral health challenges, thus policymakers, community groups and politicians should all have vested interest in quashing food insecurity in New York.

While the intersection of food insecurity and machine learning has been explored, it is a relatively new endeavor and is seemingly untapped regarding its applications in New York State. Though machine learning techniques have been applied to food insecurity problems before, many of the applications have been in a global context, focusing on the poorest countries. The United Nations World Food Programme (WFP) transformed their remote hunger monitoring initiative into a machine learning backed tool to track and predict hunger in hard to access places globally. While this research is incredibly important on the global scale, there is high food insecurity in New York State and is often overlooked by policy makers. Thus, applying machine learning to food insecurity in the state brings a new discussion to the table and gives a new way to evaluate it.

This research project proposes to evaluate and predict food insecurity at the county level across New York, pulling from a range of economic, regional, demographic and social factors, spanning a yearly timeframe to rope in longitudinal impact as well. The primary source of the data is derived from the County Health Rankings & Roadmaps website which provide panel data on health outcomes for all U.S. states. For this analysis, only New York data is procured. Due to the structure of the data, the analysis uses an LSTM (Long Short-Term Memory) model to capture both the panel aspect of each county and the longitudinal nature of the 4-year span. XGBoost will also be considered alongside the LSTM model.

The implications of this research are extensive. Identifying key indicators of food security and providing accurate predictions from it, backed by both panel and longitudinal trends at the county level can assist policymakers, community health centers, food banks and many more stakeholders efficiently allocate and understand food insecurity in communities across the state.

1.1. LITERATURE REVIEW

There is a wide range of research surrounding food insecurity and the root causes since the topic is all-encompassing and involves a multitude of fields and mediums of study. Consequently, many of the studies referenced in this paper have unique perspectives surrounding root causes, methodology and data collection. These perspectives, nonetheless, were foundational in building the skeleton of the predictive model.

1.2. Food Insecurity Predictors

Lauren et al. (2021) examines the effects that race, income, relationship status and mental disorders had on food insecurity nationwide during COVID-19. The results indicate that people of color, those living with children and those with mental health disorders were individually, far more likely to experience food insecurity during COVID-19. While the study only looks at the pandemic period, the dynamics of food security existed before COVID and were only exacerbated during the period. These findings were the driving reason behind looking for demographic based indicators to add to the model.

Niles et. Al (2020) reported similar finding from their Vermont specific study, citing that "Food insecurity tracks closely with national and household economic conditions, with trends paralleling unemployment, poverty, and food prices." The authors then went on to report that "Respondents experiencing a job loss were at higher odds of experiencing food insecurity." They note that economic conditions, poverty, and CPI are critical to food insecurity. Drewnowski (2022) argues that food insecurity is strongly rooted in economic causes. However, a more holistic approach would contend that economics alone cannot explain all instances of insecurity. This is especially true when looking at something as exact and unique as county level. The Office of Disease Prevention and Health Promotion cites accessibility as a major factor in measuring food insecurity.

With a topic as intertwined as food insecurity is, it is imperative that a model be built to accompany that. Thus, this model draws on a lot of prior research which attributes rising food insecurity to a multitude of factors. Because of this, the model accompanies a range of demographic, health, geographic and economic predictors to best capture county level nuances.

1.3. Global Cases

Gholami et al. (2022) takes machine learning techniques to analyze and forecast food security in Southern Malawi. This case study used a set of twenty-one key predictors to predict food security outcomes accurately up to four months in the future. This study first utilizes Shapley additive explanations or SHAP to identify the top performing predictors in the model (called feature importance analysis in the paper) and then uses a neural network model to predict future outcomes of food insecurity. While this is applied specifically to Southern Malawi, this approach lends itself well to evaluating New York counties too.

1.4. New York State Food Insecurity

Azhar et al. (2023), research predictors of food insecurity and childhood hunger in the Bronx during the COVID-19 pandemic. Results of this paper suggest that one of the most critical predictors in food insecurity for Bronx residents during COVID was a lack of insurance, especially among families with children. The study also concluded that, though the funding of food pantries played a critical role in staving off food insecurity, there was a clear gap in what still needed to be provided and what the current state of food pantries could provide.

The New York State Comptroller's "New Yorkers In Need" report (2023) echoes similar research citing that, people of color, especially Black Americans and households with children were much more likely to be food insecure in New York.

The New York State Department of Health press release on Food Insecurity among adults (2024) quoted the current State Health Commissioner saying "Hunger stresses the body and mind, and can result in malnutrition, inability to concentrate, anxiety, and depression. In addition, adults who experience food insecurity are more likely to report chronic diseases such as diabetes, heart disease, asthma, and cancer." Food insecurity is not only about being hungry, but its roots go deeper, expanding to nearly every facet of someone's life.

These studies help to inform the reasoning behind why more than just a few economic related variables are selected for the model. Health, social and geographic variables are just as critical to the models and ensure predictions are made holistically considering many different perspectives, since food insecurity affects each person and family differently.

1.5. Food Insecurity & Machine Learning

The proposed methodology of this paper is rather niche, thus any literature surrounding the methods is critical to the production of the model. Li et, al (2023) provides an informative backbone for understanding XGBoost on a county level and how SHAP and multidimensional feature engineering can bolster the understanding and performance of a model. While the authors of the paper were predicting crop yield in the midwestern United States, the methodology was applied yearly, predicting crop yield down to the county level. Despite the differences in end goal, this paper's methodology assists in understanding how XGBoost can be applied at the county level, using data from counties in the United States.

1.6. COVID Specific Literature

COVID-19 caused a sort of paradigm shift in how experts viewed and forecasted food insecurity. Not only did it initially exacerbate problems for those who are statistically more

likely to be affected by it, food banks and other related programs reported a massive influx of new food insecure households and individuals. Meaning, COVID vastly widened the pool of food insecure people. Demographic disparities were also widened, (Morales et, al. 2021) demonstrates that racial and ethnic minorities were much more represented in food insecure households who reported that they were not confident they could afford to buy more food and were more likely to be afraid to go out. Additionally, racial and ethnic minority households were much less confident about their future (4 weeks out) food security than white food insecure households. However, possibly the most compelling trend to come out of the COVID era was what happened when federal and state programs began rolling back their bolstered programs and benefits. When programs like the expanded Child tax Credit, Supplemental Nutrition Assistance Program (SNAP) and the Special Supplemental Nutrition Program for Women, Infants, and Children (WIC), and universal free school lunches for all students returned to their pre-pandemic levels overall poverty as well as childhood poverty would skyrocket, doubling from 5.2% in 2021 to 12.4% in 2022. Wolfson et al. (2024) backs this claim, indicating that post expiration of the enhanced Child Tax Credit in 2021, the increase in childhood poverty was "the largest increase in more than 50 years."

1.7. Literature Review Summary

There is wide range of research surrounding food insecurity, however, that is the problem, a lot of the food insecurity research is painted with a broad stroke (global and national based approaches). Making policy decisions based on a study that was conducted nationally, or in a different country, does not leave any room for the nuance that can be found in state and county specific data. Thus, the gap in the research allows for the exploration of what food insecurity is affected by in New York State and then use these relationships to inform a predictive model of food insecurity on a county level. Using New York State specific data in a longitudinal panel format gives the project a unique avenue into the field filling two separate problems. Predicting food insecurity on a state specific county level, while also accounting for longitudinal regional changes in the variables. As a result of previous research, the two main models were worth comparing, XGBoost and LSTM (Long Short-Term Memory). XGBoost is able to extract information on the most important features, making it much more understandable than LSTM and other neural network approaches.

2. HYPOTHESIS

This research project posits that food insecurity at the county level in New York State can be predicted using a combination of economic, demographic, social, health and geographic predictors, modeled through XGBoost and LSTM.

Economic factors (e.g., median household income, unemployment rates, debt-to-income ratio) and social determinants (e.g., access to healthy foods, housing cost burden, childcare costs) will emerge as significant predictors of food insecurity across New York counties. Demographic variables (e.g., racial composition, percentage of uninsured adults, and rural vs. urban classification) will also play a critical role in explaining disparities in food insecurity.

Incorporating longitudinal data will improve the predictive accuracy of the model by capturing temporal trends and year-over-year changes in food insecurity. Panel data analysis will reveal county-specific patterns, encapsulating country specific nuances that may be lost in solely longitudinal predictions.

The comparison between XGBoost and LSTM models will provide an interesting and fruitful decision. LSTM benefits from capturing longitudinal temporal aspects of the data, while XGBoost predictions are much more interpretable with bult in feature importance. The comparison between the two models will weigh the costs and benefits heavily, however, the performance and accuracy of each model will be the deciding factor.

The model will help to identify specific counties in New York State as high-risk hotspots for food insecurity, particularly in regions with higher poverty rates, limited access to healthy foods, and severe housing cost burdens. Rural counties and those with higher percentages of minority populations will likely exhibit disproportionately higher rates of food insecurity compared to urban and predominantly white counties.

By testing these hypotheses, this research aims to demonstrate that machine learning models, particularly those leveraging longitudinal and panel data, can provide a robust framework for understanding and predicting food insecurity at a granular, county-level scale in New York State. This project looks to fill in gaps of understanding food insecurity in New York State, while liking together different areas of research into one model.

3. DATA

As is was briefly mentioned in the introduction, the bulk of this data comes from the County Health Rankings & Roadmaps which is a project created by the University of Wisconsin Population Health Institute – School of Medicine and Public Health. The data is pulled down

from their website for each year of analysis (2020-2025) and then cleaned within an R program, taking only the variables of interest. From there the data is combined into one master file. Their 2025 data however, due to availability, are only estimates, as a result this year cannot reliably be used in cross validation in the models so 2024 is the last year that can be utilized to build and validate the model.

As mentioned previously, it is imperative to have a holistic approach when attempting to predict food insecurity. Thus, the model archives this by factoring in a plethora of data. The dependent variable is obviously going to be food insecurity and is defined as the percentage of the population who lack adequate access to food. While the data is aggregated and published by the University of Wisconsin Population Health Institute – School of Medicine and Public Health, this data is originated and measured by Map the Meal Gap, a national initiative headed by the group Feeding America. The County Health Rankings & Roadmaps dataset is utilized for almost every other independent variable, while every single variable will not be listed, it is helpful to understand the groupings that the variables are a part of.

Social and economic factors play a major role in building the model. Children in poverty, unemployment, income inequality and children in single-parent households are some of the socio-economic variables chosen for the model. Similar to the dependent variable (food insecurity), the data for these variables is procured by UW's Population Health Institute, but the data originates from Bureau of Labor Statistics (Unemployment) and the census county level surveys (childhood poverty, income inequality, etc).

Since this study has geographic components from the county level, variables that measure access are important. Thus, the food environment index, access to healthcare and access to healthy foods and rural percentage strengthen the underlying power of the model.

Lastly, this model leans on a few other, less cited predictors of food insecurity. As a major expense for most Americans, housing costs (and the burden they have) and childcare costs are a necessary addition to this model to capture an underrepresented aspect of people's ability to cover food cost.

4. EXPLORATORY DATA ANALYSIS

4.1. Missingness Challenges

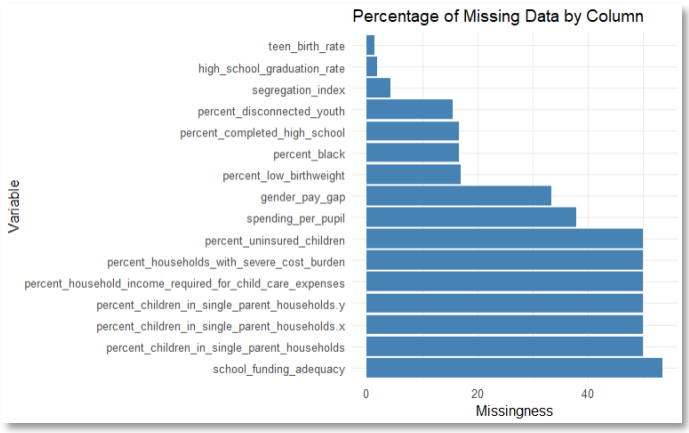

Figure 1.0 : Numerical Missingness (EDA)

After pulling together the yearly data and cleaning it the primary goal was to visualize and begin thinking about missingness present in the dataset. Opinions differ widely on missingness across academia and professional spaces thus there are no strict guidelines for a threshold of when its ok to impute missing variables and when its not. Given the nature of the data (longitudinal and cross county) any missingness under 50% will be considered for imputation. However, several iterations of imputing thresholds will be used in the models. So 50% will be the maximum and can be expected to be walked back throughout the process.

4.2. Distribution of Dependent Variable

Food insecurity is the primary dependent variable being predicted in the models, so looking at a baseline distribution is important for tailoring model selection and any other data manipulation that may need to be considered.

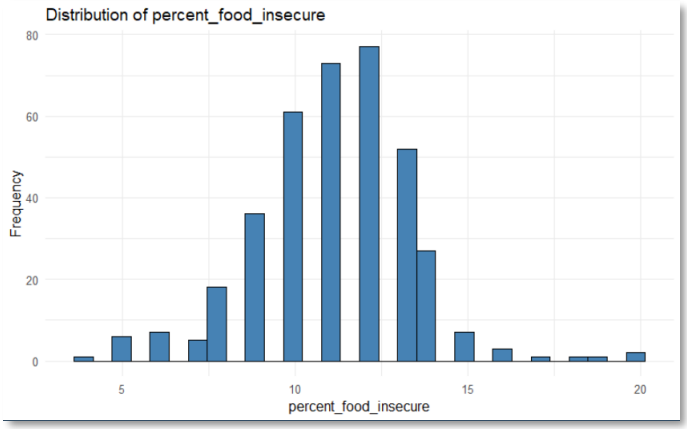

Figure 1.1: Primary Variable of Interest's Distribution

Outside of a few specific counties, food insecurity follows a mostly normal distribution with a very slight skewness to the right.

4.3. Overall Trends in Food Insecurity

From a bird's eye view, food insecurity does seem to be clearly on the rise throughout the five-year period.

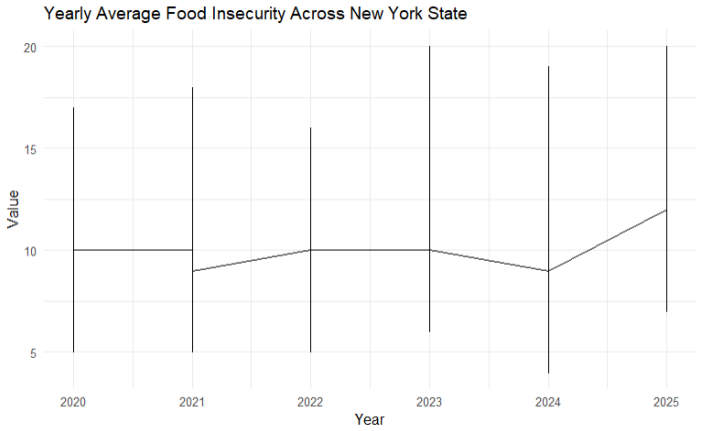

Figure 1.2: Total Statewide Yearly Change in Food Insecurity

This trend is also present across most, if not all of the New York State counties during the five-year period too. Regardless of the average level of food insecurity for a given region, there is a clear dip in 2024 followed by a sharp expected increase throughout the end of 2024.

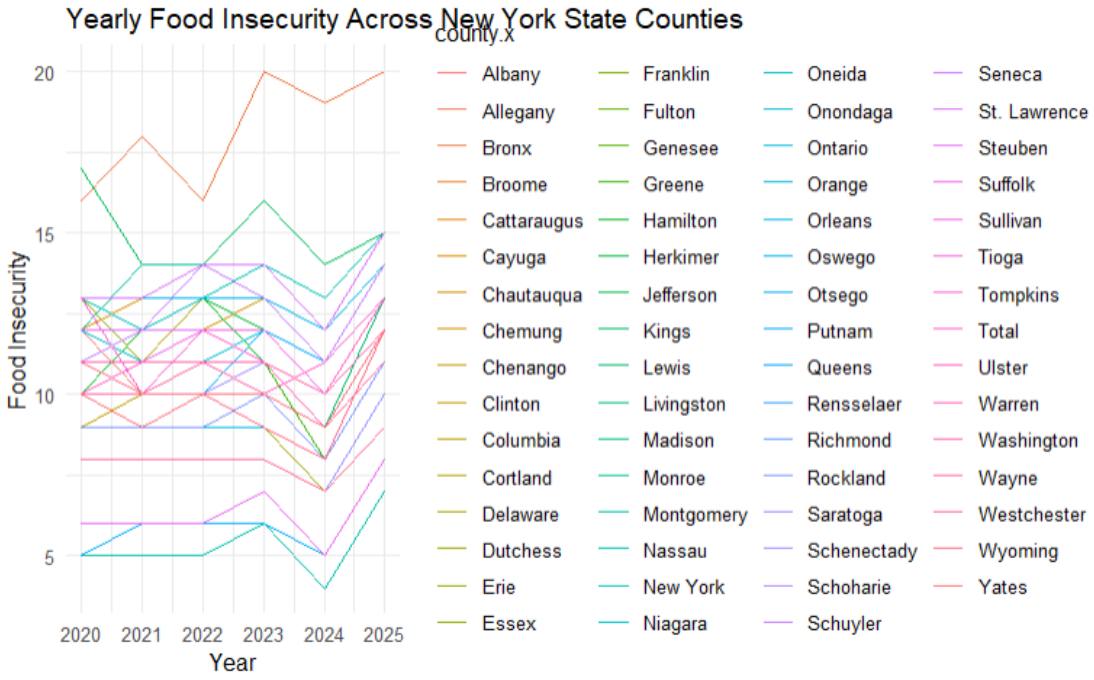

Figure 1.3: Yearly Change in Food Insecurity Within Counties

5. METHODS

The methods employed in this paper attempt to address the unique challenges of predicting food insecurity at the county level, while accounting for yearly temporal changes. The statistical methodology begins with data acquisition and preparation, feature engineering and eventual model comparison and selection.

5.1. Data Acquisition and Preparation

As mentioned above the data for this paper was acquired from the University of Wisconsin Population Health Institute – School of Medicine and Public Health and utilizes their County Health Ranking and Outcomes data, specifically for New York State for the time frame 2020 – 2025. The dataset itself contains hundreds of variables surrounding health outcomes in NYS, however only food insecurity, economic indicators, demographic, social and geographic factors were hand-selected for the analysis.

The data tables are only downloadable for one year, so there is some leg work required to get the five years of data into one working dataset. Each downloadable excel sheet has two main data tabs, the selected (primary) set of variables and then a supplemental tab which had some more niche, yet equally important variables. Each variable of interest is handpicked, from there an R program pulls in the variables from each of the tabs, gives it a year variable and merges the variables together. After completing this step for each year of analysis the data is stacked on top of one another for further cleaning.

Prior to conducting any data transformation, an exploratory data analysis (EDA) exercise provides invaluable insight into the structure and characteristics of the data. Missingness testing, correlations, skewedness and variable distributions all aid in informing next steps in feature engineering and eventually model selection and deployment.

Part of the data preparation process was deciding what to do about missing values. The Long Short Term Recurrent Neural Network requires NAs to be pre-processed beforehand as they do not have the capability to deal with missingness within the model like XGBoost does. Thus, part of the preprocessing steps for the LSTM models consisted of imputing missing numeric variables using mean and computing missing categorical variables using mode imputation.

On the other hand, XGBoost left a few options to experiment with regarding missing values. The XGBoost machine learning model is designed to handle missingness naturally on its own. The problem here is that the model treats a variable with missingness of 50% and 10% the same, so instead of leaving it to the model to sort out, an additional preprocessing step was added.

The debate about what percentage missingness is statistically sound to impute is widespread across the sciences. Some studies cite that 50 percent missingness is a valid amount to impute without introducing too much bias, while others cite five percent as the ceiling. To err on the side of caution, this study used a modest 25 percent missingness threshold where any variable below it will be imputed, and any variable above will be dropped prior to the model training. This step found seven variables that met the requirements to compute and nine that did not. From there the nine high missingness variables got dropped and the remaining seven were imputed using median imputation if they were numeric value and mode imputation if they were categorical. It is also worth noting that there were no categorical variables that needed to be imputed at the moment, that code was left in as a placeholder if it was decided that new data needed to be pulled in or the missingness threshold increased. Following this, there were 19 predictor variables with no missing data ready for the model. This step was added for XGBoost model 3.

5.2. Feature Engineering

The EDA step informed most of what is done for feature engineering. Regardless of what is found in EDA, temporal features for the LSTM model is necessary. Lagged variables for food insecurity and a handful of other features will strengthen the predictive power of LSTM, however, more in depth EDA is necessary to figure out which features these should be. Lastly, an argument could be made about combining similar features into single variables.

Included in the data preparation for both models was the preparation of lag variables, as mentioned above. Since this data spans 2020 – 2024, the creation of two lag variables was warranted, one measuring food insecurity lagging one year, and one measuring food insecurity measuring two years. The problem of food insecurity is a slow changing, slow burning pandemic, thus creating two new features to try and reflect the past aids in improving its forward-looking accuracy.

An additional step in feature engineering was adding a semi-geographic piece to the model, as geography is known to be a crucial piece to many socio-economic problems. Building off of the percent_rural variable from the original dataset, the new rural_urban feature categorizes the numeric values into completely rural, mostly rural and mostly urban to add greater geographic context to the models, unmasking more nuance to geographic regions and food insecurity.

Additional features were created for the XGBoost model in addition to the rural_urban categories. Food_risk_score is a combination feature which contains a weighted average of

several food related variables in the data, like percent_enrolled_in_free_or_reduced_lunch and food_environment_index. The final combination feature built for the model is an economic_stress variable which combines and takes an average of percent of unemployed, percent of children in poverty and percent of the population with a severe housing problem. Adding this variable aims to capture a smoother number for the three figures.

5.3. Model Selection and Development

Due to the structure and nature of the data, there are two models that needed to be evaluated. XGBoost and LSTM (Long Short-Term Memory) were the clear choices. Both models carry several strengths and weaknesses. XGBoost is an ensemble machine learning method, which means the algorithm iteratively improves the performance of weak learners (Uhunmwangho, 2024). While LSTM is a subsection of recurrent neural networks and excels in capturing temporal properties across sequential data. XGBoost is more interpretable because of the built variable importance feature which identified the most important predictors in the food insecurity model. On the other hand, LSTM may capture complex longitudinal relationships that are crucial to predicting food insecurity.

Performance metrics to compare the XGBoost model and LSTM model consist of mean squared error (MSE), mean absolute percentage error (MAPE), root mean squared error (RMSE) and R-squared for inter-XGBoost. The variable importance function is also employed to better interpret the results and identify key predictors within the XGBoost model.

5.4. Model Structure and Training

Across every model, temporal (lag) features were built to try and capture historical trends in food insecurity as the country level. In addition to this, the lag variables were calculated within groups (FIPS codes are just numbered representations of counties).

5.5 XGBoost Model

The XGBoost model is structured to use 2020 through 2023 for the training dataset and have 2024 set as the test data. Following this, non-predictive variables were removed (fips, county, state) from the x_train and x_test sets and percentage variable (percent_food_insecure) in the y_train and y_test sets. Although the parameters and preprocessing of the data changed across the three experimental XGBoost models, they were all done locally using cross-validation, shifting hyperparameters, a refined XGBoost tuning step.

While the datasets were rather small, some of the hyperparameters and crossvalidation settings made this computationally expensive because it was done locally. Consequently, parallel processing via the doParallel package was utilized in R. Experiments 2 and 3 used 4 cores of parallel processing to get the models trained, speeding up the process as a result.

5.6 Long Short-Term Memory (LSTM) Model

Due to complications with the Keras and Tensorflow libraries in R, this part of the model analysis and comparison shifted to Python. The goal was to set up and run very similar types of models between LSTM and XGBoost. Two lag variables at 1 and 2 years were created to capture temporal patterns in the data. For missing variables, fillna.mean() was used for missing numeric variables, while fillna.mode() was used for missing categorical variables.

Sequence creation is an important aspect to capture temporal patterns in the data. Thus, each county has a set of sequences created, admittedly there may not have been an adequate number of years available for proper sequencing of the data, however this is discussed more in depth in the limitations section of the paper.

The model architecture for the best performance consists of two main layers (3 if the input shape layer is included). There are two LSTM layers stacked on one another with 50 units per layer (ReLU activation). Both layers contain ReLU activation, L2 regularization (to combat overfitting by penalizing large weights) and dropout at 15% (to also combat overfitting). The final layer is a single unit Dense layer for regression.

An early stopping module was also created to prevent overfitting in the model. With the module set to monitor validation loss and patience at 15 epochs to stop training the model past this arbitrary number.

6. RESULTS:

6.1. XGBoost Results:

XGBoost models demonstrated superior predictive power, accuracy, and consistency; Model 1 (a baseline XGBoost model, with minimal feature engineering and preprocessing) outperformed even the best performing LSTM model by a wide margin. With an R-Squared of 85% a MAPE of 10.3%, MSE of 1.15 and RMSE of 1.07, XGBoost model 1 was anywhere from .5 to 4x more performant than any LSTM.

XGBoost models 2 and 3 showed notable improvements in performance over the first model (and any LSTM consequently). Growing R-squared metrics across the two models and a concurrent drop in MSE, MAPE% and RMSE can be attributed to a few changes. XGBoost model 2 brings two specific changes from model 1. This model is tuned to be faster, using a tuning grid and parallel processing. In addition to the grid and the parallel processing mechanics, model 2 also features a sped up cross-validation section, reducing the stock amount of folds from model 1, from 10 to 5 folds, in turn producing a faster, more performant model.

The final XGBoost model adds five changes to aid performance (maximizing R-squared and minimize error metrics). The first change was just ensuring that the two lag variables created were only backward looking and not accidentally pulling in future data into the calculations of the lags. Second was preprocessing missing variable instead of dropping them (model 1) or letting the XGBoost algorithm handle it (model 2). Imputing variables that had 25% missingness or less allowed the model to consider 7 more variables than it had in the previous two models. This model looked to reduce overfitting by adding L2 regularization into the hyperparameters, in turn penalizing larger coefficients, distributing weights more evenly across every feature instead of only the large outliers. For example, L2 regularization helps to smooth over the model when federal and state COVID-19 benefits (increased SNAP, unemployment funding, etc) were rolled back, sending millions back into poverty and food insecurity (ie, outlier years in the data).

The performance of these models alone indicates that XGBoost was the right machine learning model for the task at hand, however some nuances about the data were discovered when implementing LSTM that must be addressed.

6.2. LSTM Results:

The LSTM models were an interesting trial; however, it was realized toward the end of the experimentation with different methods to minimize performance metrics that there might not be enough data to successfully derive a statistically sound and meaningful temporal pattern from the data. The dataset has a lot of data points, 62 counties across 4 years spanning anywhere from 8-28 predictors depending on the model. Despite this there were not nearly enough yearly variables for LSTM's temporal piece to truly have an impact on the outcome of the model.

Traditionally anything below 10% MAPE is acceptable and considered highly accurate, thus anything below 10% is not seriously considered for this research. The improvements in the

LSTM models did in fact make a difference in lowering MAPE, MSE and RMSE, however, not enough to move through with any real evaluation against of on the better performing XGBoost models.

Model Type	Mean Squared Error (MSE)	Mean Absolute Percentage Error (MAPE) %	R_Squared (XGBoost Only)	RMSE (XGBoost Only)
XGBoost 1	**1.158309**	**10.31290%**	**0.8556726**	**1.076248**
XGBoost 2	**0.7858386**	**8.44709%**	**0.9354144**	**0.8864754**
XGBoost 3	**0.8419701**	**9.04140%**	**0.9449322**	**0.9175893**
LSTM 1	**5.413426**	**16.57202%**	***Null***	**2.326677**
LSTM 2	**11.901971**	**25.72257%**	***Null***	**3.449923**
LSTM 3	**4.735223**	**15.37036%**	***Null***	**2.176057**
LSTM 4	**5.05472**	**16.05878%**	***Null***	**2.24827**

Table 1.0: Cross comparison table highlighting predictive power, accuracy and error of the two models

6.3. XGBoost Variable Importance Discussion:

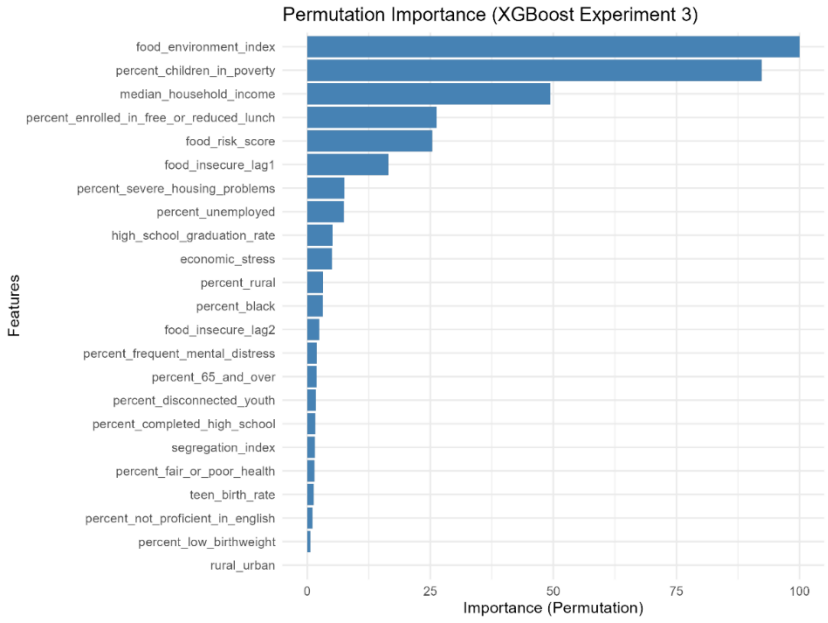

Figure 1.4: Explainable variance within XGBoost model 3 (most performant)

The top 5 features represent most of the explainable variance in the best performing model (model 3). percent_children_in_poverty, median_household_income and percent_enrolled_in_free_lunch are all seemingly critical to the predictive power of the model,

making up a majority of the explained variance. An interesting finding was that the food_insecure_lag1 variable (1 year of lag) seemed to matter a lot more and explain much more variance in the model comared to lag2. This indicates that with food insecurity, recent history matters a lot more than even two years back. Suprisingly, there were some demographic and socioeconomic features that did not carry a lot of weight. As mentioned in the literature review, there are severerly heightened rates of food insecurity for people of color, especially the black population, more data may be needed to truly see this problem more prevalently in the model.

7. LIMITATIONS TO THE RESEARCH

This was a comparative paper looking to see if LSTM or XGBoost models could better predict inter-county food insecurity. Overall setting up both models to even do inter-county predictions was admittedly difficult because of the fact the model needed to take into account a panel analysis of predictors, per county, all while taking into account the temporality of yearly changes in food insecurity. Although the XGBoost model predicted food insecurity outcomes fairly well, with good accuracy (MAPE), the lack of yearly data for both models showed in the outcome of both analyses.

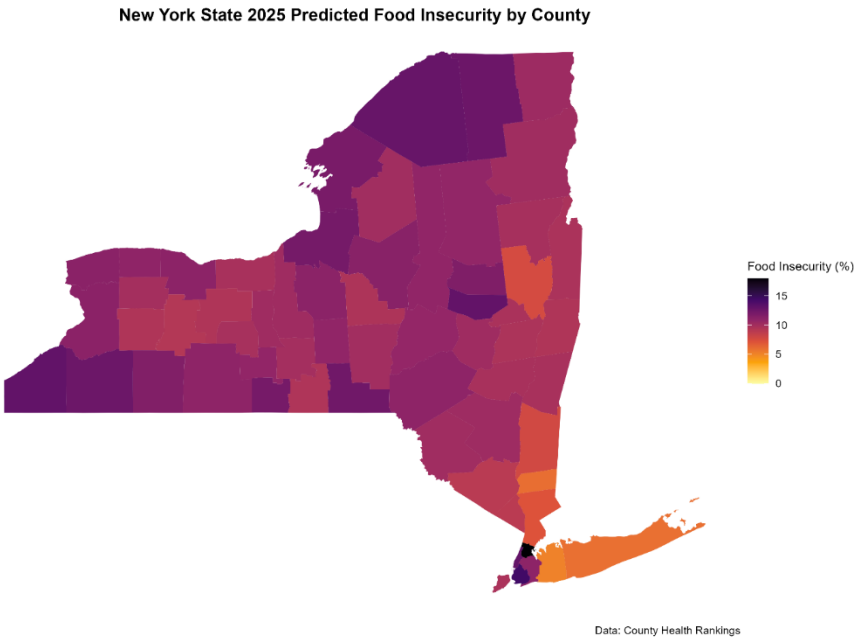

Figure 1.5: 2024 Predicted county-level food insecurity heatmap

8. CONCLUSION

Given the magnitude of food insecurity across New York counties, this research represents a much-needed analysis of an issue that has plagued generations of New Yorkers. By leveraging longitudinal panel data and the advanced models XGBoost or LSTM neural networks, food insecurity can be predicted more accurately, enabling a better understanding of current and future challenges for communities. Lastly, the geographical view of predictions will enable policymakers, community organizations, and other stakeholders to identify high-risk areas and allocate resources more effectively.

8.1. Superior Performance of XGBoost

Overall, the three XGBoost experiments solidified that for this set of data, XGBoost was the far superior model. An R-Squared of .94, a MAPE of 9.0% and an RMSE below 1.0 deems XGBoost model 3 as a highly accurate, trustworthy and statistically sound model to predict county level food insecurity in New York State.

The feature importance analysis also cemented that poverty rates (especially among children), median household income and free/reduced price lunch are the strongest predictors of food insecurity in the model, which is also backed by numerous papers in the literature that cite income and food access as main contributors. The one-year lag variable aimed at capturing extra temporal patterns from the data outperformed the two-year lag variable, indicating that shorter term windows of food insecurity are much more important than looking at the long-term window.

8.2. Challenges with LSTM for Short Time Series

While XGBoost outperformed LSTM across the board, theoretically with more data, LSTM should be on par. Thus, high MAPE and RMSE metrics for all LSTM experiments meant that the model would not be used for the primary analysis. As mentioned above, it was not exactly given a fair chance to compete with the boosting model, a lack of data makes it very difficult for a neural network to derive any meaningful temporal patterns.

8.3. Policy implications

Looking back at the yearly, county level line graph, there is jarring evidence to just how well government programs like SNAP, enhanced child tax credit, etc. aid in keeping people from being food insecure. Ending toward the tail end of 2023, the CARES act lifted millions of New Yorkers up, helping to keep enough food on the table. Consequently, finding high risk counties

through this data, where the loss of these benefits among other shocks could be disastrous is an important tool in helping to fight food insecurity.

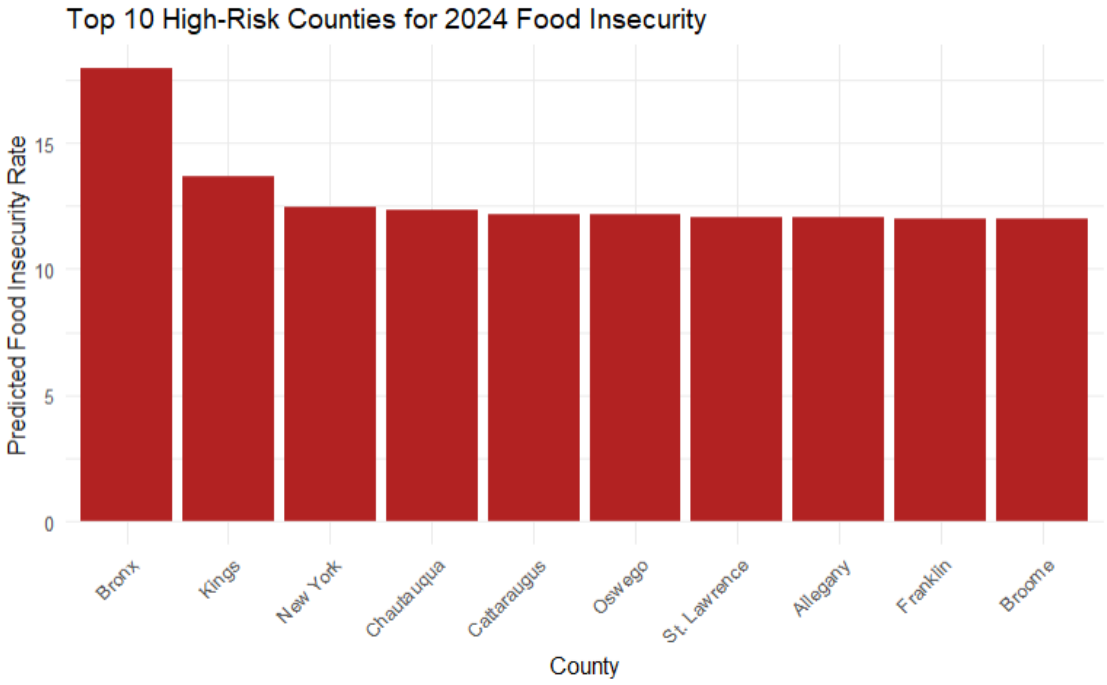

Figure 1.6: Highest risk counties identified by their change from 2023 and overall level of food insecurity

8.4. Limitations and Future work

The fact that there was not enough data for proper utilization of the LSTM models is disappointing, however this opens the possibility and the need for further research. Using the County Health Rankings data catalogue to its fullest extent might be exactly what a full LSTM model would need. There are clearly geographical nuances to this data and this problem overall, therefore warranting a more ironed-out approach to classification of geographical differences is needed. The final direction this project could stem out of would be a more generalized model which can take in not only New York State County data but also generalize the model to fit and work with every county in the United States. An interface like Rshiny or a python equivalent, built out with either an LSTM model or XGBoost in the backend which people can use to generate a dashboard for the state they want to look at food insecurity predictions for.

9. REFERENCES:

World Health Organization. (n.d.). ***Hunger numbers stubbornly high for three consecutive years as global crises deepen: UN report*****. World Health Organization. https://www.who.int/news/item/24-07-2024-hunger-numbers-stubbornly-high-for-three-consecutive-years-as-global-crises-deepen--un-report**

Feeding america urges bold, collective action in face of increase in food insecurity**. Feeding America. (2024, September 4). https://www.feedingamerica.org/about-us/press-room/usda-food-security-2023**

Food insecurity is on the rise in New York | News | dailygazette.com. (n.d.). https://www.dailygazette.com/the_recorder/leader_herald/news/food-insecurity-new-york/article_071f7cc6-abfa-11ee-8f91-0fd76f71a888.html

Data and resources**. County Health Rankings & Roadmaps. (n.d.). https://www.countyhealthrankings.org/health-data/new-york/data-and-resources**

https://www.nimhd.nih.gov/resources/understanding-health-disparities/food-accessibility-insecurity-and-health-outcomes.html (DEPRECATED)

Azhar, S., Ross, A. M., Keller, E., Weed, J., & Acevedo, G. (2023). Predictors of Food Insecurity and Childhood Hunger in the Bronx During the COVID-19 Pandemic. ***Child & adolescent social work journal : C & A*****, 1–14. Advance online publication. https://doi.org/10.1007/s10560-023-00927-y**

Lauren, B. N., Silver, E. R., Faye, A. S., Rogers, A. M., Woo-Baidal, J. A., Ozanne, E. M., & Hur, C. (2021). Predictors of households at risk for food insecurity in the United States during the COVID-19 pandemic. ***Public Health Nutrition*****,** ***24*****(12), 3929–3936. doi:10.1017/S1368980021000355**

New Yorkers in need - New York State comptroller. (n.d.-b). https://www.osc.ny.gov/files/reports/pdf/new-yorkers-in-need-food-insecurity.pdf

State Health Department Releases Report On Food Insecurity Among Adults – Department of Health : https://www.health.ny.gov/press/releases/2024/2024-01-03_food_insecurity.htm

Gholami, S., Knippenberg, E., Campbell, J., Andriantsimba, D., Kamle, A., Parthasarathy, P., … Lavista Ferres, J. (2022). Food security analysis and forecasting: A machine learning case study in southern Malawi. ***Data & Policy*****,** ***4*****, e33. doi:10.1017/dap.2022.25**

Using big data and machine learning to Monitor Food Security. **World Food Program USA. (2023a, December 13). https://www.wfpusa.org/articles/leveraging-big-data-and-machine-learning-monitor-global-food-security-in-real-time/**

Uhunmwangho, O. P. (2024). Comparing XGBoost and LSTM Models for Prediction of Microsoft Corp's Stock Price Direction. *Mountain Top University Journal of Applied Science and Technology*, *4*(2). **https://doi.org/https://mujast.mtu.edu.ng/storage/issues/Year_2024_Vol_4/Number_2/1729800557_MUJAST_240801.pdf**

Li , Y., Zeng, H., Zhang, M., Wu, B., Zhao, Y., Yao, X., Cheng, T., Qin, X., & Wu, F. (2023, March 27). *A county-level soybean yield prediction framework coupled with XGBoost and Multidimensional Feature Engineering*. **International Journal of Applied Earth Observation and Geoinformation.**

Urell, A. (2023a, October 30). *Millions more U.S. households are experiencing food insecurity*. **Equal Justice Initiative. https://eji.org/news/millions-more-u-s-households-are-experiencing-food-insecurity/**

Wolfson, J. A., & Leung, C. W. (2024). Food insecurity in the COVID-19 ERA: A national wake-up call to strengthen snap policy. *Annals of Internal Medicine*, *177*(2), 255–256. **https://doi.org/10.7326/m23-3363**

The State of rural New York Report. (n.d.-c). https://ruralhousing.org/wp-content/uploads/2023-State-of-Rural-New-York-Report.pdf

Junaid, K. P., Kiran, T., Gupta, M., Kishore, K., & Siwatch, S. (2025). How much missing data is too much to impute for Longitudinal Health Indicators? A preliminary guideline for the choice of the extent of missing proportion to impute with multiple imputation by chained equations. *Population Health Metrics*, *23*(1). **https://doi.org/10.1186/s12963-025-00364-2**

PUBLIC GITHUB REPOSITORY LINK:

https://github.com/jonburns2454/DATA_698