Predicting Microbusiness Density Across US Counties: A Machine Learning Approach

Nicholas Amirsoleimani
University of California, Berkeley, namir@berkeley.edu

Abstract: This paper presents a machine learning approach to predict microbusiness density across US counties using a dataset collected from several online sources. The main objective of the study is to identify significant features that influence microbusiness density and to create a model that is able to accurately forecast these changes over time. The project employs a combination of supervised learning algorithms, feature selection techniques, and temporal cross-validation methods to achieve this goal, while addressing challenges related to the weighting of counties by population size and the focus on inference. The results provide valuable insights into the factors affecting microbusiness density, offering a better understanding of the underlying causes for variations in microbusiness density across counties. These findings can be used to inform policy decisions, target resources effectively, and support economic development efforts aimed at fostering microbusiness growth in regions with high potential.

1. INTRODUCTION

1.1 Background and Motivation

Microbusinesses, defined as businesses employing fewer than ten employees, are an essential component of the United States economy. They play a crucial role in fostering innovation, creating jobs, and contributing to the overall economic growth of the nation. The distribution and density of microbusinesses across the country are influenced by various factors such as the local economy, infrastructure, access to resources, and socio-economic conditions. Understanding these factors is critical for policymakers, urban planners, and businesses to support the growth and development of microbusinesses. For instance, policymakers can use this information to create supportive policies that foster microbusiness growth, while urban planners can design infrastructure that accommodates the needs of microbusinesses. Businesses that operate within the advertisement sphere, such as Go Daddy, can identify counties with an expected increase in microbusiness density and allocate advertising resources accordingly to target potential clients more effectively.

Machine learning, along with classical linear regression methods, provide powerful tools to analyze and predict microbusiness density and identify the factors that significantly contribute to their spatial distribution.

1.2 Problem Statement and Questions

This study aims to predict the change in microbusiness density, in terms of percentage, for each county and month. The primary questions addressed in this study are:

1. Which factors significantly influence microbusiness density in US counties?
2. How can machine learning algorithms and classical linear models be utilized to predict microbusiness density and identify key features?
3. How can the insights derived from the study be applied to support economic development and decision-making for various stakeholders?

1.3 Impact and Value of the Project

In-depth analysis and understanding of the factors affecting microbusiness density can provide valuable insights for various stakeholders, enabling them to make data-driven decisions and allocate resources optimally. Identifying the key factors influencing microbusiness density can also help uncover potential areas for intervention and support, contributing to the overall growth and development of microbusinesses. Furthermore, accurate predictions of microbusiness density can enable better resource allocation and decision-making to encourage economic development in regions with a high potential for microbusiness growth.

1.4 Relevant References to Previous Work

Several studies have explored the factors influencing microbusiness growth and distribution in different regions [1]. However, the use of machine learning techniques to predict microbusiness density across US counties remains relatively unexplored. This study builds upon existing research by leveraging machine learning algorithms to identify significant features and make predictions on microbusiness density. Previous studies have highlighted the importance of using machine learning algorithms, such as Random Forest and boosting methods, to reduce overfitting and handle correlated features effectively [2][3]. These methods offer an advantage in accurately predicting microbusiness density, especially when dealing with large, complex datasets.

1.5 Data Descriptions

The dataset used for this study was collected from various sources, including the US Census Bureau, Bureau of Economic Analysis, Go Daddy, Google Trends, and other relevant governmental and non-governmental organizations. The data includes information on microbusiness density, socio-economic factors, infrastructure, and other related features across US counties. The dataset is appropriate for addressing the research questions, as it contains comprehensive and relevant information that can support insights and answers on microbusiness density distribution, if deemed important.

1.6 Plan of Attack

The study followed a systematic approach that included data collection, exploratory data analysis (EDA), feature engineering, feature selection, modeling, optimization, and interpretation of the results. This process allowed for a comprehensive analysis of the factors affecting microbusiness density and the identification of significant features. The following sections of the paper will elaborate on each step, providing a detailed account of the methodology and findings and are organized as follows:

Section 2: Data Description, detailing the data collection, processing, and quality control steps.

Section 3: Exploratory Data Analysis (EDA), presenting relevant visualizations and patterns in the data to inform model selection.

Section 4: Preliminary Statistical and Machine Learning Model Analysis, discussing model choice, fitting, diagnostics, and interpretation of results.

Section 5: Next Steps, outlining the gaps to be addressed and a brief roadmap to complete the project.

2. Data Description

This section provides details on data collection, processing, quality control, and other relevant aspects of the dataset used in this study.

2.1 Data Sources and Features

The dataset used for this study was collected from various sources, including the US Census Bureau, Bureau of Economic Analysis, Johns Hopkins University, and other relevant governmental and non-governmental organizations. The data includes information on microbusiness density, socio-economic factors, infrastructure, and other related features across US counties. A comprehensive table of data sources and the features obtained from each source is provided in Table 1.

TABLE 1
Types of features collected and their respective sources.

Raw Data Sources and Properties			
Name	**Source**	**Longitude**	**Geography**
Microbusiness Density	GoDaddy	Monthly	County
Real & Sector GDP	BEA	Yearly	County
Population Census	BC	Yearly	County
Demographics Census	BC	Yearly	County
Covid-19 Death	JHU	Monthly	County
ChatGPT Dialogue	ChatGPT	Static	County
Google Search Trends	Google	Monthly	State
Google Search Trends	Google	Monthly	County
Business Tax	RSPS	Static	State
Health	CHR	Yearly	County
Education	CHR	Yearly	County
Crime	CHR	Yearly	County
Rent	DHUD	Yearly	County
Coastline	BC	Static	County
Nearest University	USN	Static	County
Unemployment	BLS	Month	State

Sources: BEA (Bureau of Economic Analysis), **JHU** (Johns Hopkins University), **BC** (Census Bureau), **CHR** (County Health Ratings from the University of Wisconsin Population Health Institute), **USN** (US News), **RSPS** (Rich States Poor States), **BLS** (Bureau of Labor Statistics), **DHUD** (Department of Housing and Urban Development)

2.2 Data Collection and Web Scraping

Data from the aforementioned sources was downloaded in their respective formats, such as CSV, JSON, or web scraping for online websites. For websites that did not provide an API, the Beautiful Soup and requests libraries in Python were used to scrape them directly. For the data sources that did have an API, such as Google Trends, the data was directly downloaded using Python. This combination of techniques allowed us to gather a rich dataset containing relevant information for our analysis. The data was then shared among the team members using GitHub, which facilitated collaboration and version control.

2.3 Data Processing and Wrangling

After collecting the data, data processing and wrangling steps were performed to make the data amenable for our analysis. These steps included:

- Merging data from different sources based on common identifiers, such as county and date, to create a single dataset in the form of a 3D NumPy array, where the rows represented the dates, columns represented the features, and the 3D layers represented the counties.
- Handling missing values by either imputing them through values of nearby counties or removing the affected data points, depending on the nature and extent of missing data; features with over 50% missing values or those with low variance were deleted.
- Converting data types and units, as necessary, to ensure consistency and compatibility across different features.
- Creating new features or aggregating existing features to better represent the underlying relationships and trends

in the data, more in section 3.

- Using one-hot encoding to encode the category features into numerical data, keeping in mind to omit categories with low variance.
- Normalizing and scaling the data cross-sectionally to ensure that the features were on similar scales and suitable for the models.

2.4 Quality Control and Data Validation

To ensure the data sources were reputable and the collected data was accurate, several quality control (QC) and data validation steps were employed. These steps included:

1. Ensuring the data collected was reputable and reliable by cross-referencing from multiple sources to ensure accuracy and consistency.
2. Conducting sanity checks on the data by assessing the plausibility of the values and distributions for each feature.
3. Feature selection: Carefully selecting relevant features for the study to minimize noise and multicollinearity in the dataset.

These QC steps were crucial in ensuring the reliability and validity of our dataset, which in turn, contributed to the robustness and credibility of our analysis and findings.

3. Exploratory Data Analysis (EDA)

The Exploratory Data Analysis (EDA) stage involved generating relevant visualizations and descriptive statistics to gain a deeper understanding of the dataset and its underlying patterns. The primary goal of EDA was to highlight outliers, data errors, address skewness, and other data issues, while also identifying important patterns and relationships among the features. The insights gained from EDA informed the choice of appropriate statistical and machine learning methods for the project.

3.1 Histogram of the Target Variable

A histogram of the target variable, percent change in microbusiness density, was created to visualize its distribution and is available in Figures 1.1 and 1.2. There are many positive outliers in the histogram, due to the nature that relative percent change cannot be less than -1. However, this was the format of the competition and is therefore the target that was chosen to predict. Future considerations of this target are addressed in section 5. However, the data is symmetric once zoomed in and therefore no transformations were needed nor applied. In addition, the target is now stationary, which tends to perform better in models since that is many times an underlying assumption.

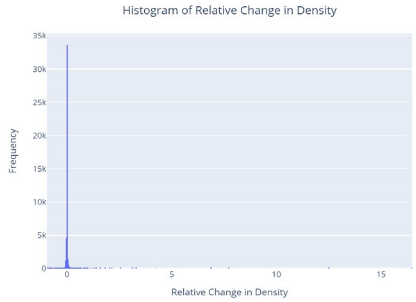


FIGURE 1.1
HISTOGRAM OF THE RELATIVE CHANGE IN MICROBUSINESS DENSITY.

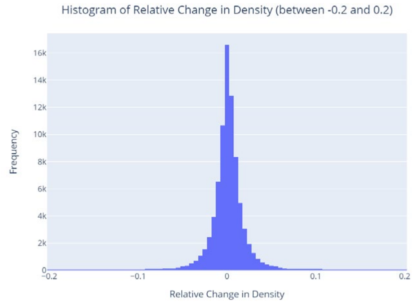


FIGURE 1.2
HISTOGRAM OF THE RELATIVE CHANGE IN MICROBUSINESS DENSITY FOR VALUES BETWEEN -0.2 AND 0.2.

3.2 Histograms and Scatter Plot of Important Features

Histograms and scatterplots were created to visualize the relationships between the most important features and the target variable, microbusiness density. These visualizations provided insight into potential linear or nonlinear relationships between the features and the target variable, helping to inform the choice of machine learning models.

To perform a cursory analysis of finding the most important features, an elastic net model was implemented. Given the temporal nature of the data, a time series cross-validation approach was employed to prevent data leakage during model training and validation. This approach ensures that the temporal structure of the data is preserved, and the models are trained and tested on non-overlapping periods. The TimeSeriesSplit function from the scikit-learn library was used to perform the time series cross-validation.

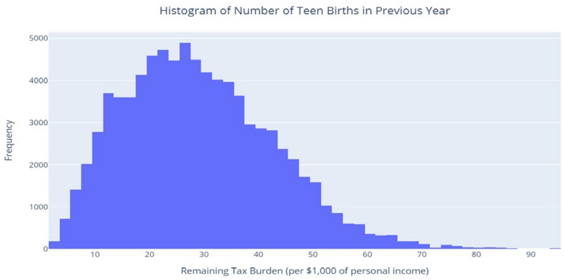


FIGURE 2.1
HISTOGRAM OF THE NUMBER OF TEEN BIRTHS IN THE PREVIOUS YEAR.

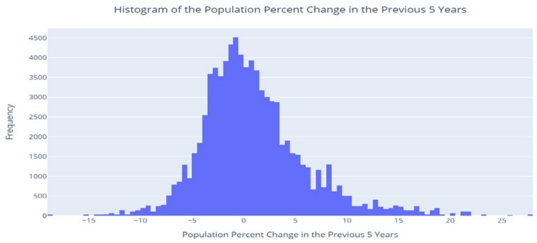


FIGURE 2.2

HISTOGRAM OF POPULATION PERCENT CHANGE IN 5-YEAR PERIOD.

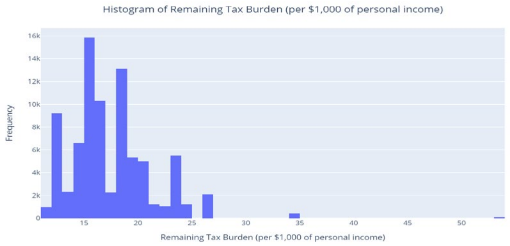


FIGURE 2.3

HISTOGRAM OF THE REMAINING TAX BURDEN PER $1,000 OF PERSONAL INCOME.

Figures 2.1, 2.2, and 2.3 show histograms of three important features according to the elastic net model. One can see that the features are slightly skewed and therefore do not possess normality. To address this, Yeo-Johnson and Box-Cox transformations were applied to all of the skewed data in order to make the data more Gaussian. One can also see that the data in Figure 2.3 is multimodal, which suggests that binning might be an appropriate strategy. However, due to the loss of information a binning strategy would pose, the decision was made against it.

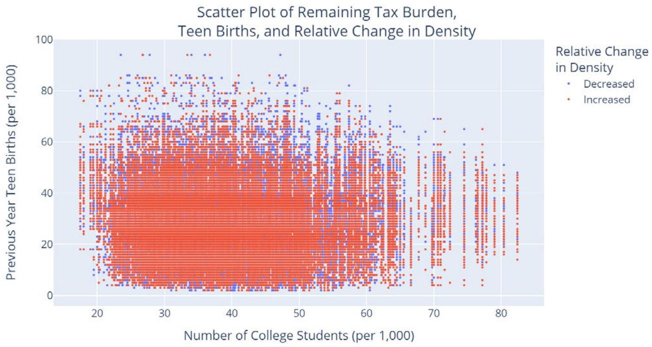


FIGURE 3

SCATTER PLOT OF THE REMAINING TAX BURDEN AND TEEN BIRTH DENOTED BY COLOR OF WHETHER THE MICROBUSINESS DENSITY INCREASED OR DECREASED FROM THE PREVIOUS MONTH.

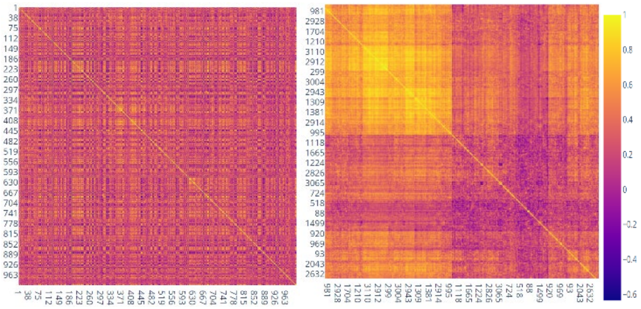
From Figure 3, one can see that as the number of teen births from the previous year increases, there is a slight increase in microbusiness densities decreasing more often. However, this relationship looks nonlinear, implying that a standard linear regression model would be insufficient. Therefore, different models with kernels that create nonlinearities within the data were tested.

3.2 Clustered Correlation Heatmap

A correlation heatmap, shown in Figure 4.1, of the microbusiness densities across counties was generated to visualize the intensity of correlations within counties. The heatmap was then clustered using a hierarchical clustering scheme (Figure 4.2) in order to determine why these correlations exist. The dendrogram in Figure 4.3 shows that there are three primary clusters that exist within the data. Upon further inspection, these clusters are mostly comprised of counties that have similar population, suggesting that population size is pertinent to change in microbusiness density.

With this knowledge gained, several additional features pertaining to population size were created, in the hopes of increasing the performance of the models. These features include but are not limited to the percent change in population over the past 1, 3, and 5 years, the relative population of counties compared to nearby counties, as well as an indicator of whether the population has consistently gone up or gone down each year over the past 5 years.

The heatmap also helped identify potential multicollinearity issues among the features and guided the feature selection process. By identifying highly correlated feature pairs, the number of features used in the analysis was able to be reduced, thus improving the interpretability and stability of the models.



FIGURES 4.1 AND 4.2

UNCLUSTERED AND CLUSTERED CORRELATION HEATMAPS OF THE MICROBUSINESS DENSITIES OVER COUNTIES.

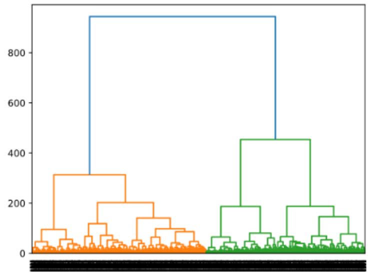


FIGURE 4.3

DENDROGRAM OF THE CLUSTERS SHOWN IN FIGURE 4.2.

3.4 Geographic Visualization of University Data

A geographic map visualization was created to display the distribution of universities across US counties, shown in Figure 5. This visualization helped highlight the potential impact of educational institutions on microbusiness density and facilitated the identification of regional patterns and clusters. From this data, several features regarding universities were created, including an indicator variable of whether a large university was within 20 miles of the county, the number of universities within a certain radius from the county, the weighted sum of the size of nearby universities.

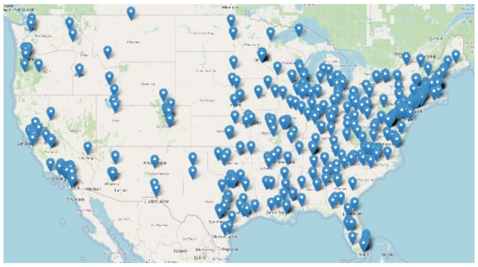


FIGURE 5
MAP SHOWING LARGE UNIVERSITY LOCATIONS WITHIN THE U.S.

4. PRELIMINARY STATISTICAL AND MACHINE LEARNING MODEL ANALYSIS

4.1 Model Selection

Several machine learning models were tested in this study, including decision trees, regular and robust linear regressions, neural networks, random forest, XGBoost, and support vector regressors. These models were chosen due to their diverse underlying assumptions, ability to handle different types of relationships between features and the target variable, and their unique strengths in addressing specific challenges in the data.

For instance, random forest and XGBoost are effective at handling highly correlated features and can capture complex relationships between variables. These models reduce overfitting by leveraging an ensemble of decision trees and can provide more accurate predictions when dealing with intricate patterns [3].

On the contrary, support vector regressor is robust to outliers and can model nonlinear relationships by employing kernel functions [3]. This characteristic makes it suitable for datasets with potential outliers or nonlinearity in the relationships between features and the target variable.

Lastly, neural networks can learn highly complex and nonlinear patterns from the data, making them a powerful tool for capturing intricate relationships between features that may not be easily captured by other models [3]. Through evaluating a variety of diverse models, the goal was to determine the optimal approach for accurately predicting microbusiness density across US counties, given the available dataset.

4.3 Hyperparameter Optimization

A Bayesian search method was used to optimize the hyperparameters of the models. This approach efficiently searches the hyperparameter space to find the best combination of hyperparameters for each model through cross-validation and is more efficient than exhaustive grid search [4].

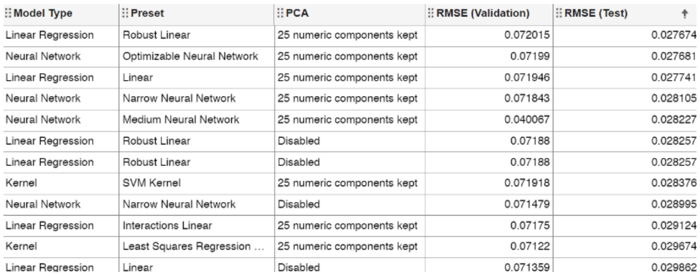
4.4 Principal Component Analysis (PCA)

For some of the models, PCA was applied to the dataset to reduce the dimensionality and reduce correlation. After slight optimization, it was found that 25 components was optimal. The PCA-transformed data yielded better results compared to the original dataset, indicating that the reduction of correlated features and noise in the data was problematic in model accuracy.

4.6 Model Fitting and Output

The selected models were fitted to the original and PCA-transformed and datasets, and their performances were evaluated the time series cross-validation scheme mentioned in section 3.2. The grid of results is presented in Table 2.

TABLE 2
SUMMARY OF RESULTS FROM THE TOP 12 MODELS IN TERMS OF TEST RMSE



Model Type	Preset	PCA	RMSE (Validation)	RMSE (Test) ↑
Linear Regression	Robust Linear	25 numeric components kept	0.072015	0.027674
Neural Network	Optimizable Neural Network	25 numeric components kept	0.07199	0.027681
Linear Regression	Linear	25 numeric components kept	0.071946	0.027741
Neural Network	Narrow Neural Network	25 numeric components kept	0.071843	0.028105
Neural Network	Medium Neural Network	25 numeric components kept	0.040067	0.028227
Linear Regression	Robust Linear	Disabled	0.07188	0.028257
Linear Regression	Robust Linear	Disabled	0.07188	0.028257
Kernel	SVM Kernel	25 numeric components kept	0.071918	0.028376
Neural Network	Narrow Neural Network	Disabled	0.071479	0.028995
Linear Regression	Interactions Linear	25 numeric components kept	0.07175	0.029124
Kernel	Least Squares Regression …	25 numeric components kept	0.07122	0.029674
Linear Regression	Linear	Disabled	0.071359	0.029862

4.7 Model Diagnostics and Assumptions

For each model, assumptions and diagnostics were checked, including the evaluation of residuals. Figure 6 shows the residual from the top performing model, the robust linear regression.

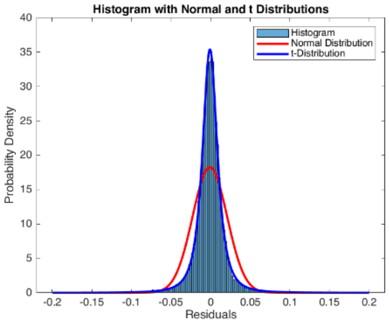


FIGURE 6
HISTOGRAM OF RESIDUAL FROM ROBUST LINEAR REGRESSION MODEL

4.8 Interpretations and Relation to Goals and Questions

The baseline estimate used in this study was the previous month's change in density as the prediction for the next month's change in density. For the training and test sets, the baseline achieved an RMSE of 0.1599 and 0.0728, respectively. As shown in Table 2, all of the models outperformed the baseline, with the Robust Linear Regression emerging as the best performer on the test set. Intriguingly, the second-best model (performing almost equally well) was a Neural Network. These two models are quite distinct in terms of their assumptions and complexities, which makes the analysis somewhat nuanced. This outcome suggests that both fundamental and complex relationships exist within the data, with the linear model capturing the fundamental aspects and the neural network uncovering the complex ones.

Other notable models that demonstrated strong performance were the Kernel SVM (with a quadratic kernel) and the Linear Regression with interaction terms, which is reasonable considering they both introduce nonlinearities. It appears that all of the models capitalized on different aspects of the data, implying that blending them together, potentially through a weighted average, could create a robust ensemble.

It is also worth mentioning that most of the top-performing models utilized PCA with a small subset of features, indicating that correlated features posed a significant challenge.

Through a stepwise feature selection process of the Robust Linear Regression, the most important features identified were the lagged 1, 2, and 3-month densities, the number of teen births in the previous year, the number of college students, the percent change in population over 5 years, the change in the previous year's Food Environment Index, and the total number of businesses. Some of these features make intuitive sense, as factors such as population change, education level, and the number of businesses can directly influence microbusiness density. However, it is crucial to note that due to the high correlation between features, the importances of the correlated features might have been aggregated, causing an insignificant feature to appear significant. This highlights the importance of carefully interpreting the results and considering the impact of feature correlations on the analysis.

The residuals from the Robust Linear Regression model are shown in Figure 6, which reveals that they followed a t-distribution instead of a normal distribution. This observation aligns with the theoretical expectation that the epsilon hats in a linear model are t-distributed. Understanding the distribution of residuals is essential for model evaluation and helps in determining the validity of the model's underlying assumptions.

This residual analysis, combined with the findings from various machine learning models, contributes to addressing the research questions posed at the beginning of the study:

1. The identification of significant factors influencing microbusiness density in US counties was achieved through a stepwise feature selection process of the Robust Linear Regression model. The analysis revealed that variables such as lagged densities, teen births, college students, population change, Food Environment Index, and the total number of businesses play a crucial role in shaping microbusiness density. However, it is important to consider the potential aggregation of importances due to high correlation between features when interpreting these results.
2. The study demonstrated the effectiveness of both machine learning algorithms and classical linear models in predicting microbusiness density and identifying key features. The diverse range of models tested, including robust linear regression and neural networks, revealed fundamental and complex relationships within the data. The use of PCA and elastic net feature selection further enhanced the models' performance by addressing issues related to correlated features.
3. The insights derived from this study can be applied to support economic development and decision-making for various stakeholders, such as policymakers, urban planners, and businesses. The identification of key factors influencing microbusiness density can guide the formulation of targeted policies and infrastructure plans that foster microbusiness growth. In addition, the accurate predictions of microbusiness density enable efficient resource allocation and strategic planning to encourage economic development in regions with high potential for microbusiness growth. Forecasting also enables businesses to spend additional capital in advertising in counties that are estimated to increase in microbusiness density in subsequent months.

5. Next Steps

While the current study has provided valuable insights into the factors affecting microbusiness density, there are several gaps that need to be addressed to fully realize the potential of the project. The following steps outline a roadmap to address these gaps.

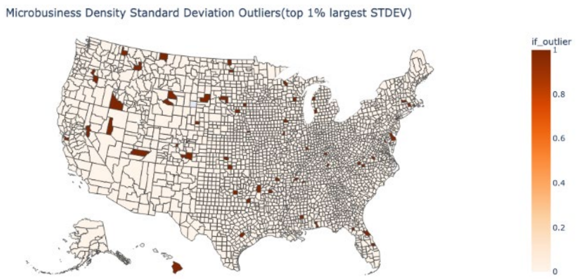

Microbusiness Density Standard Deviation Outliers(top 1% largest STDEV)
if_outlier
1
0.8
0.6
0.4
0.2
0

FIGURE 7
Map showing the microbusiness density standard deviation outliers.

5.1 Revising the Model to Account for County Size

The current model predicts the percentage change in microbusiness density, treating each county's density equally, regardless of its size. This approach does not account for the actual impact of a change in microbusiness density in counties with larger populations compared to smaller ones. Figure 7 illustrates this fact by showing that most of the deviations in microbusiness density are from smaller, less populated counties. To address this issue, the model can be revised to predict the expected number of microbusinesses per county, taking into account both the density and the population size. This approach will provide a more meaningful measure of the impact of changes in microbusiness density for policymakers and businesses.

5.2 Focusing on Inference and Identifying Underlying Causes

Adopting the model to predict the total number of microbusinesses averaged over the past year, rather than predicting densities, allows for the identification of underlying causes for variations in microbusiness density across counties. By shifting the focus from forecasting to interpretability, the study provides valuable insights into human nature and the factors driving microbusiness growth and development. This information is particularly relevant for behavioral scientists, psychologists, policymakers, and local government officials who seek to understand and promote microbusiness growth.

For entities such as governments and local authorities aiming to improve the number of microbusinesses, it is crucial to comprehend the underlying causes of variations in microbusiness density. By identifying the factors that contribute to the growth and development of microbusinesses, these stakeholders can make informed decisions, formulate targeted policies, and allocate resources effectively to create an environment conducive to microbusiness expansion. Ultimately, the insights derived from this study can help drive economic development and foster a thriving microbusiness ecosystem.

Urban planners can use the findings to design policies and infrastructure that support microbusiness development, while marketing agencies can target advertising and promotional efforts towards areas with high microbusiness growth potential. Financial institutions can utilize the results to identify regions with high microbusiness density for investment and lending opportunities.

By addressing these gaps and expanding the applicability of the project, the study will provide valuable insights for a wide range of professionals and industries, contributing to a better understanding of the factors driving microbusiness density and supporting the growth and development of microbusinesses in regions with high potential.

ACKNOWLEDGMENT

I would like to express my sincere gratitude to my teammates, Zhengpu Zhao, Leon Weingartner, and Spencer Slaton, for their valuable collaboration and unwavering dedication throughout this project. Their contributions and insights have been instrumental in shaping and driving the research forward, making this work a truly collaborative effort.

I am also grateful to Professors Thomas Bengtsson and Libor Pospisil for providing excellent guidance, advice, and feedback during the interim project. Their expertise and support have been invaluable in refining our approach and ensuring the quality of our research.

Lastly, I would like to extend my appreciation to all the academics who have laid the foundation for the research we have built upon. Their tireless efforts in developing algorithms and advancing knowledge have enabled us to access a wealth of resources and tools for free. Without their contributions, our work would not have been possible.

REFERENCES

[1] Acs, Z.J., Szerb, L. Entrepreneurship, Economic Growth and Public Policy. Small Bus Econ 28, 109–122 (2007). https://doi.org/10.1007/s11187-006-9012-3

[2] Breiman, L. Random Forests. Machine Learning 45, 5–32 (2001). https://doi.org/10.1023/A:1010933404324

[3] Hastie, T., Friedman, J. & Tisbshirani, R. (2017). The elements of Statistical Learning: Data Mining, Inference, and prediction. Springer.

[4] Jasper Snoek, Hugo Larochelle, Ryan P. Adams: “Practical Bayesian Optimization of Machine Learning Algorithms”, 2012; [http://arxiv.org/abs/1206.2944 arXiv:1206.2944].

[5] Dweck, Carol S. “Messages That Motivate: How Praise Molds Students’ Beliefs, Motivation, and Performance (In Surprising Ways).” In Aronson, Joshua (ed.), 2006, Improving Academic Achievement: Impact of Psychological Factors on Education. New York: Elsevier Science, pp. 37 – 60.