Microbusiness Density Forecasting Based on XGBoost

Xing Guo*

Department of Stony Brook Institute, Anhui University, Hefei, China

*Corresponding author: r32114008@stu.ahu.edu.cn

Abstract. Modern American business culture heavily depends on microbusiness. Defined as a specific type of small business with an online presence and no more than ten employees, microbusiness faces certain challenges that distinguish it apart from its larger counterparts, including a lack of financial resources, sh279279ort of assistance and concerns from authorities. Recognizing the critical need to understand and support this vital segment of economy, the study will employ advanced machine learning methods to predict the next month's microbusiness density based on historical time series data set. Specifically, the utilization of XGBoost algorithm is coupled with appropriate feature engineering to create a robust predictive model. Delving into the algorithm structure, anomalies are identified and corrected through data smoothing, utilizing rolling window sum features to capture recent trends, and the process of creating lag features. Manually filtering out anomalies by setting boundaries, this study fits the model using the constructed features to make adequate predictions. On microbusiness prediction, multiple forecasting models are compared and their performance is assessed using a test set that is well-known. The result presents that XGBoost algorithm outperforms traditional time series models when predictions are solely based on previous observations. The approach of the research emphasizes the significance of leveraging cutting-edge machine learning techniques to gain insight into the microbusiness landscape, thereby enabling more targeted support and opportunities for this crucial sector of the U.S. economy.

Keywords: Microbusiness; machine learning; XGBoost; time series; density forecasting.

1. Introduction

Microbusiness is a term used to describe a small-scale firm or business that often targets a specific market and employs a relatively small number of people. Essential to local economies and create jobs, microbusinesses are small businesses with nine employees or fewer, including the owner. [1]. Microbusinesses, which are structured as sole proprietorships or small corporations, are commonly owned and operated by individuals. They can be found in various industries, including retail, food service, consulting, e-commerce, and artisanal or craft production. Due to their frequent inclusion in small business statistics despite the fact that they may differ significantly from their small business counterparts, microbusinesses are usually undervalued in a world dominated by massive multinational enterprises [2]. Advancements in technology, especially the internet and e-commerce platforms, have made it easier for microbusinesses to reach a wider audience and operate more efficiently. According to the report, 3.8 million microbusinesses operate in the United States, which represent 74.8% of all private sector employers and 10.3% of all private sector jobs [1]. According to studies, microbusiness could improve people's quality of life rather than harm it if both entrepreneurs and the government uphold their respective obligations [3]. American policymakers attempt to build more inclusive, recession-resistant economies. The ease of launching a business has never been greater than it is right now thanks to technology improvements. According to studies, more Americans are deciding to launch their own businesses in order to meet their financial objectives, whether it's to find a better work-life balance, to follow a passion, or as a result of unemployment. It is difficult for policymakers to research microbusiness since it is too small to be included in traditional economic data sources. However, modern technology may be able to close the knowledge gap and provide understanding of the factors affecting microbusiness.

Nevertheless, microbusiness density forecasting is a relatively new field with limited historical research on the topic. Historically, geographic data at the regional or sub-regional level has been used in spatial economic studies [4]. The simplest solution, which is obviously a linear combination of probability forecasts, is not the best one [5]. The majority of the time, Bayesian Vector Autoregressive (BVAR) models are employed to create real-time density projections, but these models frequently neglect to take data uncertainty brought on by forthcoming data revisions into consideration [6]. A new asymmetric continuous probabilistic score for evaluating and contrasting density projections is offered by historical research, which also generalizes the proposed score and offers a weighted variation [7]. Economics and urban planners have long been intrigued by the idea of forecasting the spatial distribution of microbusiness. One of the earliest studies on the topic was conducted in the 1970s by economist David Birch, who argued that small businesses were the primary drivers of job growth in the United States, and that they tended to cluster in specific geographic areas [8]. Since then, researchers have continued to investigate the variables that affect the spatial distribution of microbusinesses and have created models to forecast their concentration. Recently, forecasting the density of microbusinesses has gained increased attention as a result of the growth of microbusinesses and the expansion of data availability. Overall, modern models use econometric techniques, rely on publicly available internal and census data, and prioritize understanding fundamental causes. These approaches are sufficient, but there is room for improvement by adding more information and applying cutting-edge techniques to predictions and decision-making.

Microbusiness density forecasting is driven by the desire to support regional economic growth from multiple aspects and to provide relevant information for decision-making [9]. In the first place, it helps governments make better policies. Governments can develop tailored policies to promote business activity in areas with low density and provide essential infrastructure by anticipating the density of microbusiness in various regions, which might encourage balanced regional economic growth. Secondly, it helps companies make strategic decisions. Companies might develop new branches or introduce new goods and services in areas with a high population of microbusinesses. To identify potential consumer groups, they might also examine the different types of microbusinesses in various areas. In addition, accurate forecasting of microbusiness density can help job seekers choose places with higher employment rates and provide investors with useful information to make investment decisions. The study will help policymakers in their efforts to improve the world for microbusiness owners by better educating them on how adverse business settings usually have a disproportionately detrimental influence on microbusiness and inhibit their development [10]. Communities all around the nation will be affected significantly by this, and it will also aid in the adaptation of our whole economy to a rapidly changing global environment.

Sec. 2 contains the introduction of data and model. The dataset used for microbusiness density forecasting typically contains information about microbusiness in different regions. The research delves into the algorithm structure of XGBoost, detailing the process of creating lag features to identify anomalies, correcting these anomalies through data smoothing and utilize SMAPE as evaluation metric. Sec. 3 contains results and discussion. For feature engineering, the research utilizes rolling window sum features to capture recent trends, manually filtering out anomalies by setting boundaries, and finally fitting the model using the constructed features to make adequate predictions. Besides, it will include a comparative analysis of various forecasting models on microbusiness density prediction, utilizing a known test set to evaluate their effectiveness and preceding the explanation of the meaning of the research. Sec. 4 contains the limitations and future outlooks. Finally, Sec. 5 presents the conclusion of the research.

2. Data and Method

2.1. Data

A great deal of data is publicly available about countries and this research utilizes the datasets provided in Kaggle to train and test the prediction model. The data consists of three individual datasets, including training set, test set, and census starter set. The American Community Survey provides examples of valuable columns that are included in the census starter set. The raw counts provided were converted into percentage fields. Every field has a two-year lag to correspond with the data that was available when a specific microbusiness data update was published. In training set and test set, apart from the ID code for the row, the day of the month's first day and the written name of the county, each county has a special identification number. The state's FIPS code is represented by the first two digits, while the county is represented by the final three digits. Moreover, it contains the initial number of microbusinesses of each county. Not provided for the test set. In addition, the census starter set includes the proportion of county homes with access to any kind of broadband and the proportion of county residents over 25 who have completed a college degree, the percent of county residents who were born outside of America, the proportion of county workers working in information-related industries, and the county's average household income.

2.2. Model

2.2.1 XGBoost algorithm

The machine learning algorithm XGBoost, which is based on decision trees, builds the model incrementally through a process known as gradient boosting. At each stage a new model is added that focuses on the errors of the previous model. XGBoost improves upon traditional gradient boosting by using a more regularized model formalization called regularized boosting to control overfitting. Some key aspects of XGBoost include handling missing data, weighted quantile sketch for approximating feature contributions, and support for various objective functions like regression, classification, ranking etc.

$$\widehat{y}_i^{(0)} = 0 \tag{1}$$

$$\widehat{y}_i^{(1)} = f_1(x_i) = \widehat{y}_i^{(0)} + f_1(x_i) \tag{2}$$

$$\widehat{y}_i^{(2)} = f_1(x_i) + f_2(x_i) = \widehat{y}_i^{(1)} + f_2(x_i) \tag{3}$$

$$\widehat{y}_i^{(t)} = \sum_{k=1}^{t} f_k(x_i) = \widehat{y}_i^{(t-1)} + f_t(x_i) \tag{4}$$

The main idea is to continuously create new trees, learning from each one based on how different it is from the target value and the previous tree, hence minimizing model bias. The output of the final model is:

$$y_i^t = \sum_{k=1}^{t} f_k(x_i) \tag{5}$$

The input training set will be $\{(x_i, y_i)\}_{i=1}^{N}$ and a differentiable loss function $L\big(y, F(x)\big)$, one initializes model with a constant value:

$$\hat{f}_{(0)}(x) = argmin \sum_{i=1}^{N} L(y_i, \theta) \tag{6}$$

And computes the gradients and hessians:

$$\hat{g}_m(x_i) = \left[\frac{\partial L\big(y_i, f(x_i)\big)}{\partial f(x_i)}\right]_{f(x)=\hat{f}_{(m-1)}(x)} \tag{7}$$

$$\hat{h}_m(x_i) = \left[\frac{\partial^2 L\big(y_i, f(x_i)\big)}{\partial f(x_i)^2}\right]_{f(x)=\hat{f}_{(m-1)}(x)} \tag{8}$$

Then fits a base learner using the training set $\left\{x_i, -\frac{\hat{g}_m(x_i)}{\hat{h}_m(x_i)}\right\}_{i=1}^{N}$ by solving the optimization problem below:

$$\hat{\phi}_m = argmin \sum_{i=1}^{N} \frac{1}{2} \hat{h}_m(x_i) \left[\phi(x_i) - \frac{\hat{g}_m(x_i)}{\hat{h}_m(x_i)} \right]^2 \tag{9}$$

$$\hat{f}_m(x) = \alpha \hat{\phi}_m(x) \tag{10}$$

Updates the model

$$\hat{f}_{(m)}(x) = \hat{f}_{(m-1)}(x) + \hat{f}_m(x) \tag{11}$$

with output

$$\hat{f}_{(x)} = \hat{f}_{(M)}(x) = \sum_{m=0}^{M} \hat{f}_m(x) \tag{12}$$

Firstly, the research provides historical microbusiness activity over time, with key metrics like microbusiness density and raw count of active microbusinesses. Then, integrate with additional indicators (i.e., broadband access, educational attainment, foreign-born population percentage, IT workforce density, and median household income). Afterwards, XGBoost mechanism parallelly constructs a series of decision trees on data subsets. Each subsequent tree refines and corrects the errors of its predecessor, with greater emphasis placed on more accurate models, culminating in a consolidated final prediction.

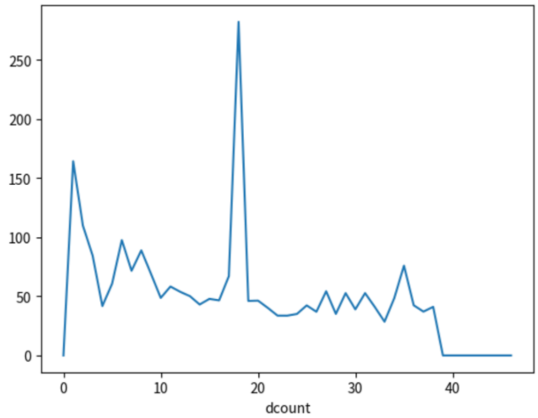

Fig. 1 Percentage increment of microbusiness density.

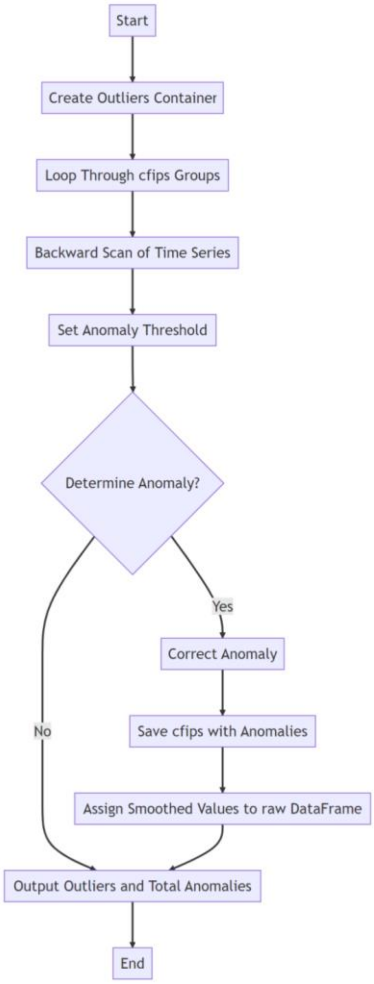

Fig. 2 Flow chart for anomaly identification and correction.

2.2.2 Identification of anomalies

In the analysis of microbusiness density data, the author has set the lag value to 1, reflecting the interest in the data of previous month. This lag value enables the research to make month-to-month comparisons, which is a critical aspect of time series analysis in this paper. The creation of lag features involves grouping the data by each county and shifting the microbusiness density column down by one. This allows for a direct comparison of density between the current month and the previous month value for each county. The missing value of the first row of each county is handled through a backward fill method. Then, the author introduces a metric called the difference ratio, representing the percentage increment in current business density compared to the previous month (Fig. 1). It serves as an essential tool for identifying anomalies or significant changes in microbusiness density. The most important part is handling edge cases and special attention is given in the creation of lag features. Notable examples include a significant increment of business density at the 18th time step related to a lag value of 0. These cases should be handled with specific conditions to ensure accuracy.

2.2.3 Anomaly correction and data smoothing

Anomalies in time series data can distort the analysis and lead to inaccurate predictions. The approach in this paper involves scanning the time series data backward to identify and correct potential anomalies, ensuring a more accurate representation of the underlying trends. The anomaly threshold is set at 20% of the mean value of the time series up to the current iteration. This threshold helps in determining whether a data point is an outlier based on the absolute difference between two subsequent time points. Once an anomaly is identified, correct it by lifting all previous values to the current level. This correction is achieved by multiplying time points before the anomaly with the instantaneous slope at that specific point. This method ensures that the previous relative differences are preserved while mitigating the influence of the anomaly. An essential part of the process in this paper is saving and tracking the total number of each county with outliers. This tracking allows the research to understand the prevalence of anomalies within the data and ensures a clear record of the corrections made. The flow chat is given in Fig. 2.

3. Results and Discussion

3.1. Feature Engineering

Feature engineering plays a crucial role in improving the performance of XGBoost model and understanding past events to predict future occurrences. Identifying and selecting relevant features that capture the key factors influencing microbusiness density can enhance the model's predictive capabilities. Domain knowledge and understanding of the microbusiness environment are essential for feature engineering. In the research, the author considers both lag features (ranging from 1 to 5 months) and rolling sum features (with windows of 2, 4, and 6 months). The author iterates through the past 1 to 3 months of density information, creating new feature columns for microbusiness density corresponding to lag columns. It allows to track the increase in microbusiness density over past iterations, providing valuable insights into trends and patterns. The rolling window sum operation is vital in indicating the recent trend or seasonality in microbusiness density. With a window size of 3, a rolling sum for the first window might be 30, and for the second window, 45. An increasing sum indicates an upward trend, and a consistent increment within certain months in a year may signal a seasonal effect. Different window sizes reveal various aspects of the data. A smaller window size uncovers short-term fluctuation, while a larger window size smooths out those fluctuations and emphasizes longer-term trends. Utilizing different window sizes to create a diverse set of features captures different facets of the data’s temporal structure, potentially enhancing model performance.

3.2. Model Performance

XGBoost is known for its high accuracy and predictive power. When applied to microbusiness density forecasting, XGBoost model can capture complex relationships in the data and generate accurate predictions. However, the accuracy of the model heavily relies on the quality and representativeness of the data used for training. In addition, XGBoost models have several hyperparameters, such as learning rate, maximum depth, and regularization parameters, can significantly impact the model's performance. Conducting a systematic hyperparameter search and using techniques like cross-validation can help fine-tune the model and improve its performance. It's important to note that the performance of an XGBoost model in microbusiness density forecasting can vary based on the specific context and data characteristics. Careful experimentation, model refinement, and domain expertise are crucial for optimizing the model's performance and ensuring accurate forecasts. In this paper, the author loops through regions and plot the corresponding predicted value, which offers a more intuitive understanding of the model’s prediction. The typical results are given in Fig. 3 and Fig. 4.

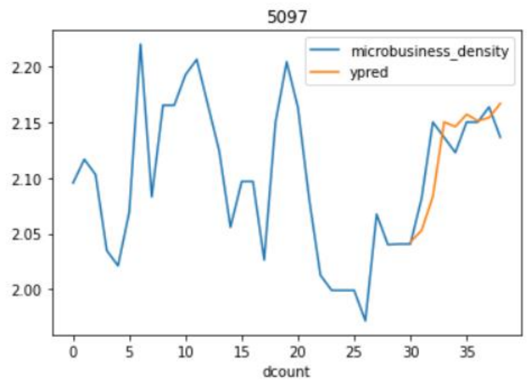

Fig. 3 Prediction of the microbusiness density of county 5097

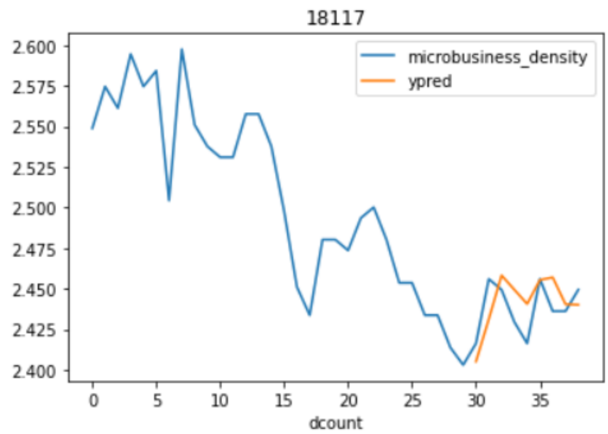

Fig. 4 Prediction of the microbusiness density of county 18117.

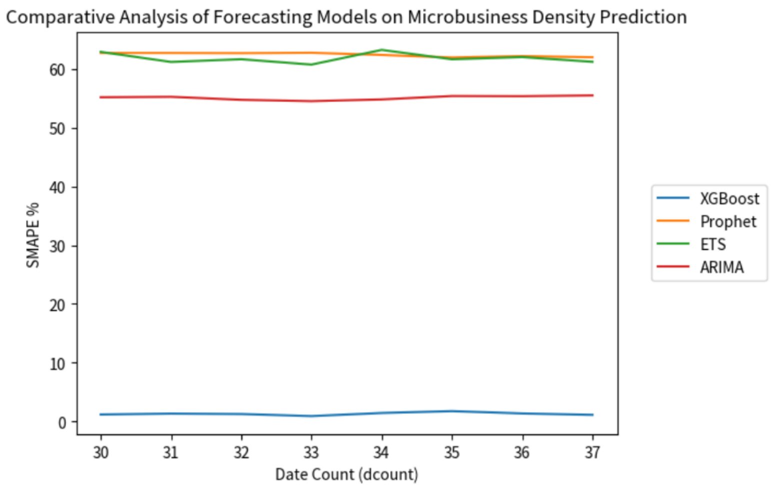

Fig. 5 Comparative analysis of forecasting models.

3.3. Explanation

Traditional time series models have been widely used in forecasting but often struggle with datasets that lack clear seasonal patterns or trends. These models also have limitations in handling direct or indirect influences on target values, which can lead to inaccuracies in predictions. In contrast, the XGBoost machine learning model demonstrates superior performance, particularly with large, complicated datasets. Its success in handling multiple variables and complex relationships between them makes it a robust choice for predicting microbusiness density. When making the comparative analysis of forecasting models on the microbusiness density prediction, the author introduced SMAPE as an evaluation metric:

$$SMAPE = \frac{100}{N}\sum_{t=1}^{N}\frac{|y_t - \widehat{y_t}|}{\frac{|y_t| + |\widehat{y_t}|}{2}} \tag{13}$$

The comparison results are given in Table 1 and Fig. 5, where XGBoost shows the best performance.

Table 1. Comparative analysis of forecasting models

Model	Mean prediction SMAPE %
XGBoost	1.311397
Prophet	62.397259
ETS	61.800931
ARIMA	55.072995

4. Limitations and Future Outlooks

Microbusiness density forecasting relies on the availability of accurate and comprehensive data. However, obtaining such data for microbusinesses can be challenging, as they often operate on a smaller scale and may not have extensive data records. Limited data availability can affect the accuracy and reliability of the forecasting model. Moreover, the quality of data used for microbusiness density forecasting is crucial. Inaccurate or incomplete data can lead to biased or unreliable forecasts. Ensuring data quality, including data cleaning and validation, is essential for improving the accuracy of the forecasting model. XGBoost is a powerful machine learning algorithm, but it can be complex to implement and tune. Developing an effective XGBoost model for microbusiness density forecasting requires expertise in feature engineering, hyperparameter tuning, and model validation. The complexity of the model may pose challenges for users without sufficient knowledge and experience in machine learning.

However, advancements in data collection methods, such as the use of IoT devices, mobile applications, and online platforms, can provide more accurate and real-time data for microbusiness density forecasting. This can enhance the reliability and timeliness of the forecasts. Incorporating external factors, such as economic indicators, demographic data, and social media trends, into the forecasting model can improve its accuracy. By considering these factors, the model can capture the broader context and potential influences on microbusiness density. Enhancing the interpretability of the forecasting model can help users understand the underlying factors driving the forecasts. Techniques such as feature importance analysis and model visualization can provide insights into the key drivers of microbusiness density. Ongoing research and development efforts can lead to advancements in XGBoost and other machine learning algorithms, making them more suitable for microbusiness density forecasting. This includes addressing limitations, refining algorithms, and developing new techniques specifically tailored to microbusiness forecasting.

In summary, while microbusiness density forecasting based on XGBoost has limitations related to data availability, quality, and model complexity, there are promising future outlooks. Improvements in data collection, integration of external factors, ensemble models, interpretability, and continuous model improvement can enhance the accuracy and applicability of microbusiness density forecasting.

5. Conclusion

To sum up, the paper has successfully demonstrated the effectiveness of the XGBoost machine learning model in predicting microbusiness density, outperforming traditional time series model. Through comprehensive feature engineering, including lag features, anomaly correction, and rolling window sum features, the model has shown superior performance in handling complex datasets. However, there are areas for future improvement, such as optimizing hyperparameters, incorporating additional features, enhancing anomaly detection, and exploring other advanced machine learning algorithms. The success of this research has opened up several avenues for future research, including conducting a more granular analysis of microbusiness density at the regional or city level, studying the impact of governmental policies, and extending the forecasting horizon for long-term predictions. By leveraging sophisticated techniques and identifying areas for continued exploration, this study paved the way for innovation in the field of microbusiness density prediction, providing valuable insights and robust forecasting tools for understanding and supporting the microbusiness landscape.

References

[1] Krueger G B. Microbusiness Marketing with No Time, No Money, and No Expertise-Research and Recommendations for Improving Marketing Results for Microbusinesses. William Howard Taft University, 2019.

[2] Campbell J M, Kubickova M. Agritourism microbusinesses within a developing country economy: A resource-based view. Journal of Destination Marketing & Management, 2020, 17: 100460.

[3] Ogunbode T O, Oyelude O, Oyebamiji V O. Evaluation of the impacts of micro-business operations on the quality of urban environment: a case study of Iwo, Southwestern Nigeria. Frontiers in Sustainable Cities, 2022, 4: 1027450.

[4] Piacentino D, Arbia G, Espa G. Advances in spatial economic data analysis: methods and applications. Spatial Economic Analysis, 2021, 16(2): 121-125.

[5] Pauwels L, Vasnev A. Forecast combination for US recessions with real-time data. The North American Journal of Economics and Finance, 2014, 28: 138-148.

[6] Iacopini M, Ravazzolo F, Rossini L. Proper scoring rules for evaluating density forecasts with asymmetric loss functions. Journal of Business & Economic Statistics, 2023, 41(2): 482-496.

[7] Clements M P, Galvão A B. Density forecasting with Bayesian Vector Autoregressive models under macroeconomic data uncertainty. Journal of Applied Econometrics, 2023, 38(2): 164-185.

[8] Birch D G W. Job creation in America: How our smallest companies put the most people to work. University of Illinois at Urbana-Champaign's Academy for Entrepreneurial Leadership Historical Research Reference in Entrepreneurship, 1987.

[9] Rocha E A G. The impact of the business environment on the size of the micro, small and medium enterprise sector; preliminary findings from a cross-country comparison. Procedia Economics and Finance, 2012, 4: 335-349.

[10] White S. Creating better business environments for micro and small enterprises. Technical Report, 2018.