EasyChair Preprint

№ 10920

# Ensemble Models for Forecasting Microbusiness Density: a Research Study

Tong Zhou



September 20, 2023

# Ensemble Models for Forecasting Microbusiness Density: A Research Study

Tong Zhou
*Department of Computer Science*
*Johns Hopkins University*
Baltimore, United States
tzhou11@jhu.edu

***Abstract*—Microbusinesses play a significant role in the economy, contributing to job creation and economic growth. Accurately forecasting microbusiness density is critical for policymakers and business owners in making informed decisions. However, forecasting microbusiness density is challenging due to the lack of reliable and comprehensive data. Ensemble models have emerged as a promising approach for forecasting microbusiness density by combining multiple models to improve accuracy. This research study aims to develop and validate ensemble models for forecasting microbusiness density and evaluate their performance using various metrics. The study's results have significant implications for policymakers and business owners in understanding the factors affecting microbusiness density and making informed decisions to promote economic growth. This paper provides an overview of the significance of microbusinesses in the economy, the challenges in forecasting microbusiness density, and how ensemble models can help address these challenges. The methodology for developing ensemble models, data sources, and performance metrics used for evaluating the accuracy of the models are also described. Finally, the paper presents the key findings of the study and their implications for policymakers and business owners.**

***Index Terms*—Microbusiness, Ensemble, Neural Network, Machine Learning**

## I. Introduction

The significance of microbusinesses in the economy has been a topic of interest among researchers. Studies have used various approaches to examine the density of microbusinesses as an indicator of their significance. For instance, some studies have employed ensemble predictions with different baseline models to provide an overview of time series forecasting problems, using microbusiness density data [1]. Other studies have introduced explicit measures of outlier detection to assess the local density of microbusinesses [2]. Researchers have also created ensembles of diverse technologies to develop a forecast model and related probability density, such as hidden Markov model ([3]). Moreover, density histograms have been used to evaluate risk in peer-to-peer lending, comparing it to starting a small business, with loan default rates being a central factor [4]. The introduction of Basel II has been noted as a response to criticisms of Basel Capital Accord, leading to the use of kernel density estimation and class overlap for prediction purposes . Additionally, ensembles have become popular tools for building powerful models, particularly in predicting firm birth rates, including concentration and pace of organization . Qualitative data has also been used to explore restaurant ownership turnover rates, which are considered significant in assessing the restaurant industry's overall health. Lastly, researchers have introduced metrics that measure rates of Gaussian mixture density and residual distribution for single GP prediction and ensemble output, respectively. Overall, these studies highlight the significance of microbusinesses in various sectors of the economy and underscore the importance of employing diverse approaches to understand their role in economic growth.

Forecasting microbusiness density is a challenging task that requires the use of multiple models and techniques. One approach is to use ensemble predictions with various baseline models. This provides an overview of the time series forecasting problem and helps researchers to create a forecast model based on a fusion of several diverse technologies and to research the related probability density. In order to create the ensembles, researchers chose to perturb two model parameters separately [5, 6]. Another approach is to introduce an explicit measure of outlier. of the closest neighbours themselves. If the local density of the data points is low, then they are considered outliers and are removed from the dataset. This approach can help improve the accuracy of the forecast model and reduce the impact of outliers on the final result [7]. Additionally, density histograms can be introduced for risk evaluation in peer-to-peer (P2P) lending. The density histograms of the loan applied for by the borrower can provide valuable information for risk evaluation in P2P lending. However, this approach has limitations due to class overlap, Kernel Density Estimation of predictions and other factors. Previous research on restaurant failures has focused mostly on quantitative factors and bankruptcy rates. A recent study explored restaurant ownership turnover rates using qualitative data, which can provide valuable insights into the challenges faced by small businesses and their owners. Finally, an overview of the methodology is shown in Figure 1, which includes variables such as pace, concentration and role in predicting firm birth. These variables play a key role in forecasting microbusiness density and must be taken into account when developing a forecast model.

Ensemble modeling techniques have shown promise in various types of forecasting problems, including those related to microbusiness density. In order to improve the accuracy of predictions, ensembles combine the outputs of multiple

models, often with varying assumptions and approaches [8]. One novel approach to ensemble modeling involves using outlier detection to identify unusual patterns in the data and adjust the models accordingly. For example, in one study, an explicit measure of outlier distance was incorporated into the ensemble model to better capture the local density of the data. Another study used a fusion of diverse technologies to create an ensemble forecast model for probability density estimation, and perturbed two models to create the ensembles. Ensemble models have also been used in risk evaluation for peer-to-peer lending, where density histograms of loan applications were compared to those of small businesses, and kernel density estimation was used to make predictions. Additionally, the introduction of Basel II has prompted the use of ensemble models in financial forecasting, where class overlap and other factors can complicate predictions. In the context of small business ownership turnover rates, qualitative data has been used in ensemble models to improve accuracy [9]. Finally, ensemble models have been used in Gaussian mixture density prediction for reliability estimation, and a metric has been introduced to measure the rate of false positives and false negatives in such predictions. Overall, ensembles provide a powerful tool for improving forecasting accuracy by leveraging the strengths of multiple models and adjusting for outliers and other complexities in the data.

## II. Methods

In order to develop ensemble models, various data sources are utilized. In the SDF forecast system, multi-model rainfall forecasts are employed to enhance the quality of streamflow forecasts [10]. The rainfall forecasts used in the SDF service are ECMWF and PME, with CHyPP model used to calibrate ECMWF forecasts in the absence of ACCESS-GE data [10]. The data sources used for developing ensemble models in this study include NWP rainfall products, CHyPP model for post-processing, GR4H hydrologic model, and ERRIS streamflow post-processor inbuilt in the SWIFT package [10]. The data is aggregated to daily values and hourly streamflow data is used [10]. Climate forecasts from the Australian Bureau of Meteorology are used as input for ensemble models, and a range of streamflow model inputs, including rainfall, temperature and soil moisture, are utilized to develop ensemble models [10]. The Ensemble streamflow forecasting service for Australia uses historical streamflow data from over 200 catchments to refine their predictions [10]. In summary, multiple data sources are utilized to create ensemble models, including climate forecasts, hydrological models, and streamflow data.

The development and validation of ensemble models are crucial for assessing forecast performance [10]. The proposed methodology for this study is an ensemble-based co-training scheme for binary classifications problems, in which the ensemble classifier is imposed as a base learner within the co-training framework [11]. The structure of the ensemble classifier is determined by a static ensemble selection approach from a pool of candidate learners, which include state-of-the-art deep learning models such as LSTM, Bi-directional LSTM, and convolutional layers [12]. The efficacy and efficiency of these models were evaluated through various classical benchmarks and reported statistical analysis, showing accurate and reliable cryptocurrency forecasting models. The models were evaluated on forecasting the cryptocurrency price on the next hour (regression) and the prediction of the next price directional movement (classification) with respect to the current price. To examine the reliability of all ensemble models as well as the efficiency of their predictions, the authors examined for autocorrelation of the errors. Additionally, criteria for accepting forecast locations for operational service are developed in consultation with key stakeholders based on model performance and forecast skill. The first criterion is that the Nash-Sutcliffe efficiency (NSE) of simulated streamflow is 0.6 or greater for a forecast location in the model validation, and if only the first criterion is satisfied, then forecasts may be released only to registered users based on stakeholder requirements and the social and economic importance of forecasts at the location. This emphasizes the importance of consistently maintaining forecast quality to avoid miscommunication of flow conditions with the public, which can impact the reputation of the service and organization.

The accuracy of ensemble models is evaluated using various performance metrics. The Continuous Ranked Probability Skill Score (CRPSS) is the primary metric used in this study to evaluate the streamflow forecast skill. The skill score is used as a measure of expected forecast skill, and sensitivity analysis is conducted to select the optimum block size and check the effect of the number of bootstrapping iterations on forecast skill. The bias is also used as a performance metric for evaluating the accuracy of ensemble models. PIT-alpha is used as a reliability metric for evaluating the accuracy of the ensemble models. In addition, calibration adds value to raw NWP rainfall forecasts, and it substantially improves forecast reliability for all lead times for the tested rainfall products [13]. The degree of improvement of forecast skill provided by error modeling decreases with lead time, and the forecast skill (CRPSS) reduces with lead time for both raw and error-modeled streamflow forecasts. Moreover, thorough evaluation is recommended before selecting NWP products to use, as the relative improvement in accuracy of rainfall forecast products is different for each product. The selection of rainfall forecast products for the operational forecasting system can affect the quality of streamflow forecasts, and the availability of the product at the BoM, hindcast period, and ease of use are criteria for selecting NWP rainfall forecast products. Catchments with Nash-Sutcliffe Efficiency (NSE) values lower than 0.6 contain intermittent or ephemeral rivers, while 97 out of 100 forecast locations exceeded the NSE value of 0.6, which is used as a performance metric for evaluating the accuracy of ensemble models.

## III. Results

The study aimed to predict and prevent bank failure in the Eurozone using Extreme Gradient Boosting (XGBoost) methodology. It was found that the use of this methodology

in the banking sector is appropriate, and XGBoost can be used to predict bank failure effectively. Leading indicators, such as the FL-Score, can facilitate timely recognition of risk management control failures and ultimately prevent banks' financial distress. The study identified different profiles of failure risks according to firm size and presented relevant considerations regarding using metrics related to the current and expected generation of economic value for the modeling of distress and recovery prediction scores. Furthermore, eye-tracking data during a decision-making task of product selection can provide insights into important information for consumers when making a purchase-related decision. The study also found that country product image and affective country image significantly influence corporate image, with important theoretical and practical implications. However, it is important to note that the study did not use market-wide competition measures such as the Herfindahl index to link indicators of bank failure. Overall, the study provides valuable insights into predicting bank failure, understanding consumer decision-making processes, and the impact of country image on corporate image.

Ensemble models have become increasingly popular in recent years as a means of building powerful forecasting models. In a study investigating the accuracy of ensemble models in forecasting microbusiness density, the authors found that combining multiple baseline models enhanced the predictive accuracy and robustness of the resulting model. The study utilized microbusiness density data to explore the factors that impact firm birth, finding that organizing effort played a crucial role in predicting firm birth, along with rate, concentration, and duration of focus on a particular area. The results of the experiment were discussed in Section 5, which provided implications for agricultural SMEs, LEs, and FSPs in confidently facing the challenges of predicting microbusiness density. The study employed a tree ensemble model, which consists of a set of classification or regression trees, and found that adding the prediction of multiple trees improved the accuracy and robustness of the prediction results. The ensemble models' weighted weak prediction rules over the model ensemble, as well as past profitability, were found to be effective predictors of future results. Density plots for predicted probabilities were also shown to improve significantly with the tree ensemble model. Overall, the study's findings highlight the potential benefits of using ensemble predictions with various baseline models in accurately forecasting microbusiness density. The study's findings have significant implications for both policymakers and business owners. Policymakers can use the results to develop policies that support small- and medium-sized businesses in financial distress. Bankruptcy prediction is a crucial aspect of financial decision-making for managers, financial institutions, and government agencies. The study provides a solution for overcoming the challenging scenario of imbalanced learning in bankruptcy prediction for these companies, which can help them make informed decisions about their future finances. The study identifies the most relevant financial attributes for bankruptcy prediction, which policymakers can use to design effective regulations and policies. Moreover, the study's results are validated on datasets from the manufacturing and construction industries, which makes them even more relevant to policymakers in these sectors. For business owners, the high prediction performance of $91\%$ in terms of geometric mean score is an important takeaway from the study, as it can help them predict bankruptcy threats and take necessary action to avoid it. However, it is important to note that the literature on private company failures is relatively small and fragmented. Policymakers and business owners may need to carefully consider the limitations of the existing literature when making decisions about private company failure. Overall, the study's findings can significantly impact policymakers and business owners in their decision-making processes, and can provide them with valuable insights into bankruptcy prediction for small- and medium-sized businesses.

This research paper focused on the use of ensemble models for forecasting microbusiness density. The study demonstrated that ensembles have become popular tools for building powerful models, particularly in predicting firm birth rates, including concentration and pace of organization. Various approaches were used to examine the density of microbusinesses as an indicator of their significance, including the use of kernel density estimation, class overlap, and density histograms. The accuracy of ensemble models was evaluated using various performance metrics, and qualitative data was used to improve accuracy in certain contexts, such as small business ownership turnover rates and restaurant failures. However, the study also identified limitations and gaps in the literature, such as the need to carefully consider the limitations of existing research when making decisions about private company failure. The study suggests that policymakers and business owners need to be aware of these limitations and consider future research directions in this area. Overall, this research highlights the importance of ensemble models in forecasting microbusiness density and provides a valuable contribution to the ongoing advancement of knowledge in this field.

## REFERENCES


[1] T. Zhou, "Improved sales forecasting using trend and seasonality decomposition with lightgbm," *arXiv preprint arXiv:2305.17201*, 2023.

[2] S. Kessioui, M. Doumpos, and C. Zopounidis, "A bibliometric overview of the state-of-the-art in bankruptcy prediction methods and applications," in *Governance and Financial Performance: Current Trends and Perspectives*. World Scientific, 2023, pp. 123–153.

[3] T. Zhou, "Nonparametric identification and estimation of earnings dynamics using a hidden markov model: Evidence from the psid," *arXiv preprint arXiv:2306.01760*, 2023.

[4] H. A. P. Hapuarachchi, M. A. Bari, A. Kabir, M. M. Hasan, F. M. Woldemeskel, N. Gamage, P. D. Sunter, X. S. Zhang, D. E. Robertson, J. C. Bennett *et al.*, "Development of a national 7-day ensemble streamflow

forecasting service for australia," *Hydrology and Earth System Sciences*, vol. 26, no. 18, pp. 4801–4821, 2022.

[5] T. Zhou, "Improved sales forecasting using trend and seasonality decomposition with lightgbm," in *2023 6th International Conference on Artificial Intelligence and Big Data (ICAIBD)*, 2023, pp. 656–661.

[6] M. Huo, S. Wang, T. Xu, D. B. Huang, and T. Zhou, "Jpx tokyo stock exchange prediction with lightgbm," in *Proceedings of the 2nd International Conference on Bigdata Blockchain and Economy Management, ICBBEM 2023, May 19–21, 2023, Hangzhou, China*, 2023.

[7] A. Thelen, M. Li, C. Hu, E. Bekyarova, S. Kalinin, and M. Sanghadasa, "Augmented model-based framework for battery remaining useful life prediction," *Applied Energy*, vol. 324, p. 119624, 2022.

[8] P. Koumbarakis and T. Volery, "Predicting new venture gestation outcomes with machine learning methods," *Journal of Small Business Management*, pp. 1–34, 2022.

[9] A. Malakauskas and A. Lakštutienė, "Financial distress prediction for small and medium enterprises using machine learning techniques," *Engineering Economics*, vol. 32, no. 1, pp. 4–14, 2021.

[10] J. Duan, "Financial system modeling using deep neural networks (dnns) for effective risk assessment and prediction," *Journal of the Franklin Institute*, vol. 356, no. 8, pp. 4716–4731, 2019.

[11] S. Figini, F. Bonelli, and E. Giovannini, "Solvency prediction for small and medium enterprises in banking," *Decision Support Systems*, vol. 102, pp. 91–97, 2017.

[12] B. Khattalov, M. Murphy, T. Fuller-Rowell, J. Boisvert, and E. R. T. E. B. CO, "Long-term ionospheric forecasting system," Tech. Rep., 2004.

[13] H. Parsa, J. T. Self, D. Njite, and T. King, "Why restaurants fail," *Cornell Hotel and Restaurant Administration Quarterly*, vol. 46, no. 3, pp. 304–322, 2005.