Microbusiness Density Forecast Interim Report

Zhengpu Zhao | zhengpu@berkeley.edu

Introduction

Current aim of this project is to forecast the county -wise monthly microbusiness density (MBD), the number of microbusinesses (MBs) per 100 people. The project is adapted from a kaggle competition sponsored by GoDaddy. A MB is defined as a registered business with fewer than 10 employees, a discrete web domain, and an active website; there are around 45 million MBs in the United States. With the MBD survey[1] provided by GoDaddy, ASU, UCLA, UIowa, census data from agencies, and novel web data from ChatGPT and Google trends, we formulate statistical and machine learning models to forecast monthly MBD and gain insight into factors contributing to the regional emergence and decay of microbusiness.

Motivation

MBD forecasting has two key benefits. First, it benefits commercial interests as many companies view regions with growing MBD as emerging markets. MBs often have similar needs, especially in online services. Companies like GoDaddy, which sells web domains, Shopify, which provides an e-commerce platform for online stores, and Snackpass, which equips quick-serve stores and restaurants with online ordering systems, offer services that are in high demand by MBs. The forecast helps these firms to target their advertisement in regions of high MBD growth to efficiently attract new entrepreneur customers.

Second, academic, public, and nonprofit sectors are also interested in changes in MBD. The Milken Institute's 2022 white paper on MBD[2] showed that MBs are a major source of employment for those with lower education levels and contribute greatly to local economic and community development. Policymakers, NGOs, and economists are keen on tracking regional and longitudinal changes in MBD to gain a better understanding of local economic health and identify potential opportunities to act upon that are often unrepresented or underrepresented in traditional economic and demographic indicators.

Data Sources

Data for the project came from multiple sources, as shown in *table 1*. The target, monthly MBD from 2019 to 2022, was aggregated from survey data acquired by Godaddy, ASU, UCLA, UIowa economists along with the 20+ millions GoDaddy registered MBs' usage information across 3000+ US counties. The data is proprietary, and due to the nature of surveys, the data is only available and accurate to a certain level.

Listed below, the covariates include: various economic and demographic data from corresponding agencies such as Bureau of Economic Analysis (BEA), Google trends search frequency data, and ChatGPT county description dialogue data. (TABLE 1)

Raw Data Name	Source	Longitude	Geography
Microbusiness Density	GoDaddy	Monthly	County
Real & Sector GDP	BEA	Yearly	County
National Personal Income	BEA	Monthly	US
Population & Demographics	BC	Yearly	County
Covid-19 Death	JHU	Monthly	County
Business Tax Rate	RSPS	Static	State
Health & Education & Crime	CHR	Yearly	County
Rent Quantiles	DHUD	Yearly	County
Coastline Indicator	BC	Static	County
Google Search Trends	Google	Monthly	State/County
ChatGPT Dialogue	ChatGPT	Static	County

TABLE 1. BEA: Bureau of Economic Analysis **BC**: Census Bureau **JHU**: Johns Hopkins University **CHR**: County Health Ratings (UWisconsin) **RSPS**: American Legislative Exchange Council **DHUD**: Department of Housing and Urban Development.

Data Wrangling - Covariates

The team divided the task of scraping data from various sources for efficiency. I was assigned to collect economics and ChatGPT data.

The former are mostly available as csv files from economic agencies like BEA. Preprocessing was done to preselect features, check and fill missing values, introduce data lags, and transform the data for later analysis. For instance, 12 specific sectors were selected from yearly sector GDP data from 2019 to 2021, which included percentages of 34 economic sectors across 3169 state and county equivalents, such as:

1. Agriculture, forestry, fishing and hunting
2. Information services
3. Finance, insurance, real estate, rental, and leasing
4. Educational services, health care, and social assistance
5. Government and government enterprises

The data is mostly present, with 7-8 specific counties missing many columns, rendering the traditional zero, mean, or kNN imputation inaccurate. A manual kNN imputation was performed. For each county with a lot of missing value, the data is filled with the mean of 5 of economically similar counties in terms of sector makeup suggested by ChatGPT.

More interesting than the raw sector percentages, percentage changes were calculated. In addition, a two year lag was introduced to prevent leaking future data into the past features. The features were then left merged onto the targets. Similar procedure was done to other tabular formatted data to get covariates such as trends in population, trends in Covid-19 death, etc.

Novel sources of data required more wrangling and preprocessing. ChatGPT dialogues are requested from the provider's API, in the format of texts. With careful tuning of the prompt and API settings like temperature (chat stochasticity), reasonably stable and accurate responses were extracted. Following the prompt:

- Give me bullet points for why I should or should not start a small business in {county}, {state}, 3 pros and 3 cons.

For each county in each state, a consistent paragraph with bullet points are delivered:

Pros of starting a small business in Santa Barbara County, California:

1. High-quality lifestyle and access to recreational activities
2. Strong tourism industry
3. Access to investment capital and resources for small businesses

Cons of starting a small business in Santa Barbara County, California:

1. High cost of living and doing business
2. Complex regulatory environment
3. High taxes, rent, and wages.

From the perspective of quality control, a semi-hand-wavy-quantitative estimate of answer accuracy is about 1 mistake or contradictory point per 7 counties or 42 bullet points, or 97%, given the 20 county sample we manually checked. Noticeably, smaller counties like Iron County, Utah and Autauga County, Alabama are more likely to receive contradictory points.

Each bullet point of the county description is then tokenized into arrays of word vectors and fed into a pre-trained transformer model, distillBERT, and extracted out at a deep hidden layer as a vector in $\mathbb{R}^{768}$. K-means clustering is then performed on the vectors, where the hyperparameter K=12 was tuned with three criterions in mind, Adjusted Rand Index (ARI), Normalized Mutual Information (NMI), and human intuition on closeness of sentences within clusters when sanity checking. An example cluster would include very similar descriptions to different counties likes:

1. The county has a limited number of resources and services available to small businesses.
2. Limited access to capital and financing options due to the county's rural location.
3. Limited access to technology and infrastructure due to the county's rural location.

Clusters are then named based on interpretation of commonality of sentences within. For each county, 12 feature are added via one-hot encoding, including:

Large_Customer_Base	Tough_Labor_Market
Low_Cost_of_Living	Limited_Tech_n_Infra
Natural_Resources	Low_Pop_Bad_Weather

Exploratory Data Analysis - Targets

With targets and features aggregated into a 103455 by 169 shaped dataframe, we perform EDA and visual inspection over the targets and the features[3].

As we aim to predict change in MBD, the second moment, identifying outliers with high variance over time is very important. Figure 1 shows a histogram of the standard deviation of the MBD for all counties. We notice the existence of outliers with timewise standard deviation up 90 while the majority are around 1.

Figure 1 - Standard Deviation of the Microbusiness Density Over Time

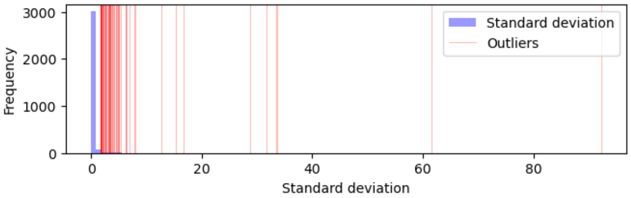


Figure 2 zooms in on the leftmost bar of figure 1 where 95% of the counties lie, showing that across the time axis, they have $\mu < 10$ and $\sigma < 1$.

Figure 2 - Mean of MBD Over Time (95% smallest least variable)

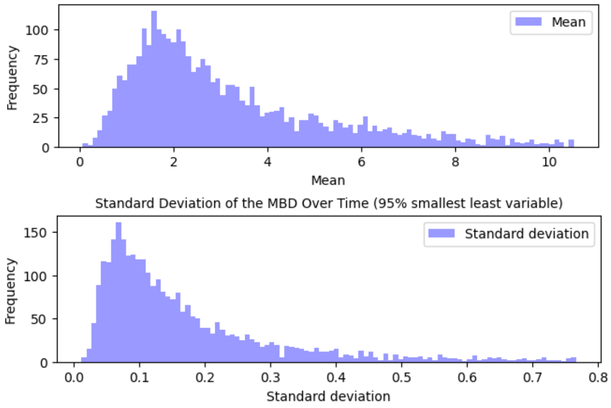
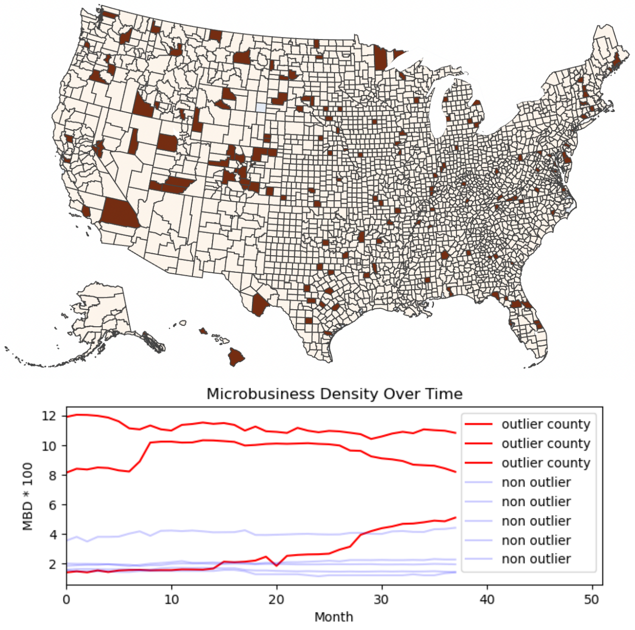


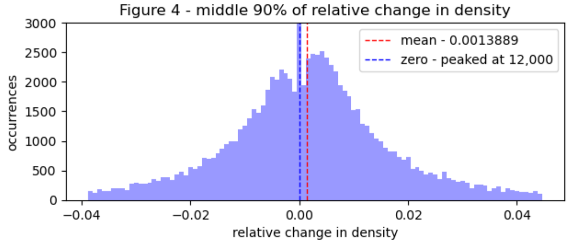
Figure 3 shows the 5% most variable counties in MBD geographically and plot MBD for the 5% outliers and the 95% in comparison. Visually, the midwest region has a higher ratio of outliers but not highly obvious. Based on this remark, we eliminated the idea of blacklisting certain states. However, it could be reasonable to separate out certain counties.

Figure 3 - Microbusiness Density Outliers



Aggregating the relative percent change in MBD over all counties and all months, in figure 4, we can see that the middle 90% of the changes in density are all nearly 0, with a mean of 0.0014 indicating a slight increase in MBD over time for all counties.

Figure 4 - middle 90% of relative change in density



Exploratory Data Analysis - Covariates

All 160+ covariates' histograms and correlations are plotted on and presented on this website[4], with a few shown by Figure 5. Some features are skewed, sparse, highly correlated, unbalanced, or missing data.

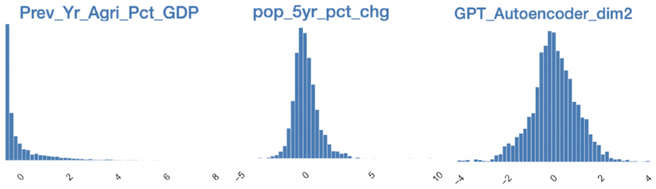


Figure 5 - Distributions of Selected Features

Several EDA informed decisions were made. First of all, highly imbalanced categorical features should be removed. The *Low_Cost_of_Living* column contains 97% ones, as there are only ~100 expensive counties among the 3000+, making it uninformative and model prone to bias. Secondly, highly skewed features like *Prev_Yr_Agri_Pct_GDP* causes the regressions to overfit to the high leverage tail, this can be addressed by transformations or removal. In addition, complex models such as tree based ensembles are generally more robust against skewed features.

Correlated covariates are common, especially in groups of data covering similar information. From the correlation map (Figure 5, 160+ features), we see at position ★$^{1-5}$, groups of features are highly correlated.

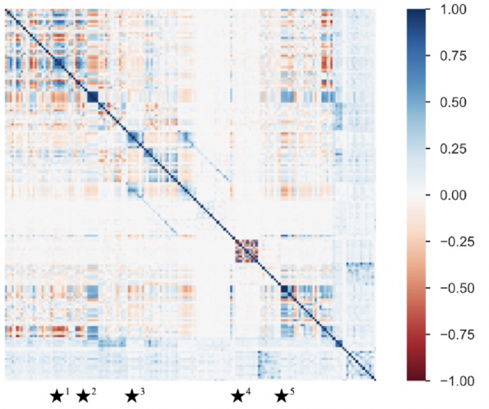


Figure 5

Group ★^{1} details health related features such as poor mental and physical health; using one or two columns should be enough. Group ★^{2} details rent quantiles, where using only the median should suffice. Group ★^{3} details google trends, where a clever combination of several trends into one could be considered. Group ★^{4} are paragraph-wise ChatGPT data reduced down to 10 dimensions by an Autoencoder; it should be removed to avoid extra noise and redundant information. Group ★^{5} details raw yearly population data and could be removed as population trend data is included.

Some of the aforementioned aspects, such as subsetting and transforming, are not yet implemented in modeling. This issue will later affect the preliminary modeling and feature selection. In the next section, we see that *prev_year_sexually_transmitted_infections*, a highly correlated feature in Group ★^{1} that describes the number of sexually transmitted infections in the prior year, be strongly selected for by a lasso model, to characterize MBD change.

Preliminary Modeling - Introduction

Due to the lengthy data scraping, cleaning, and aggregation process, preliminary modeling was done on an individual basis and shared in meet-ups when progress was made. Team members simultaneously carried out feature selection and model exploration. The following may not include all the notable and applaudable work by other members.

Evaluation Metric & Baseline Benchmark

To benchmark the accuracy of the MBD forecast, we evaluated the *SMAPE* score, *Symmetric Mean Absolute Percentage Error,* that is defined as:

$$SMAPE(\hat{y}, y) = \frac{100}{2N} \sum_{t=1}^{N} \frac{|\hat{y} - y|}{|\hat{y}| - |y|} \ \& \ SMAPE(0, 0) = 0$$

Before formulating any statistical models, we create a baseline to benchmark performance. Due to the high dimensionality of the feature space and unknown signal to noise ratio, simple linear regression may not be the best bet. As previously observed, most of the MBD changes over time are around 0, alluding to the reasonable 'last value' baseline:

$$\Delta_{\mathrm{MBD}}(t) = \frac{\Delta_{MBD}(t)}{MBD_t} = 0 \ \Leftrightarrow \ \mathrm{MBD}_{t+1} = \mathrm{MBD}_t$$

For any model that either predicts:

$$\Delta_{\mathrm{MBD}}(t) \quad \text{or} \quad \frac{\Delta_{MBD}(t)}{MBD_t} \quad \text{or} \quad \mathrm{MBD}_{t+1}$$

we transform the prediction into MBD_{t+1} and calculate the score of the predicted and actual MBD.

We define the train-test-split with a training set of 34 months from Jan-2020 to Oct-2022, and a 2 months test set of Nov-2022 and Dec-2022. With the baseline and the benchmark set up, we find for the immediate 2 following time steps:

$$SMAPE_{baseline,\, t+1} = 1.3792$$

$$SMAPE_{baseline,\, t+2} = 2.0890$$

With the baseline in mind, we first tried models that only used the autoregressive nature of the targets; namely, a simple linear regression on lagged data, a regularized cubic spline, and a LSTM. Unsurprisingly, none of the models outperforms the last value baseline, and often defaulting to $\Delta_{\mathrm{MBD}}(t) = 0$, indicating that the signal might not be in the autoregressive-ness, and a need for more and engineered features.

Feature Selection via LASSO

With 160+ plus features that are likely not linearly related to Δ_{MBD}, sometimes correlated and potentially uninformative, and unsatisfactory simple models, we face the feature selection problem on complex models. A naive forward selection on SVM, random forest, or gradient boosted decision trees are computationally too time consuming to be implemented.

A computationally cheaper method is to bootstrap Lasso for a sparse group of significant features. Here, interaction terms between categorical and numerical features are also included. A bootstrap of 100 straps and lasso regression is run, each time cross validating for best λ and β'*s*. We get a bootstrapped distribution of Lasso β'*s* with 100 samples for each β. There are 27 β'*s* with 68% statistical significance, $0 \notin \{\mu \pm \sigma\}$. We can reasonably use these features as a starting point for more complex models, including:

- *MBD trends from previous 1,2,3,4,5 time steps*
- *Two years prior population up or down trend*
- *Previous yearly unemployment trend*
- *Previous yearly normalized sexually transmitted infections*
- *Normalized nation personal income*
- *Previous yearly government sector GDP trend*
- *Google search trend on keyword 'loan', 'office' at state level*
- *ChatGPT large customer base × MBD trends prior 1,2,3,4,5*

Note that here the Lasso model consistently selects for the covariate 'previous yearly normalized sexually transmitted infections'. However from a first principles perspective, we realize that such phenomena is due to the high correlation between the covariate and other health related covariates aforementioned in Group ★^{1}. It causes the model to arbitrarily select for 'sexually transmitted infections' over other related features, increasing bias. In future iterations, we could drop some of features or implement more correlation robust regularized models like elastic net or group lasso on highly correlated covariates in Group ★$^{1\text{-}5}$.

Another fact to note is that the Lasso models do not outperform the baseline up to 2 decimal places in the *SMAPE* benchmark. Alluding to the issue of low signal to noise ratio contained by the covariates especially under the linear model assumptions.

More Complex Models

With the general idea of which set of covariates contains useful information, models with higher complexity are tried; namely, Support Vector Machine Regression (SVR), Random Forest (RF), and Gradient Boosted Decision Trees (LGBM). In a manual and arduous process, covariates such as population trend, unemployment trend, sector GDP trends are iteratively added to the models while basic hyperparameter tuning is done each time. While we are able to decrease the *SMAPE* score, test set performance increases are not too significant[4], as shown in the table:

Model	SMAPE_{t+1}	SMAPE_{t+2}
Last Value Baseline	1.3792	2.0890
Random Forest	1.3790	2.0800
Gradient Boosted Decision Tree	1.3677	2.0407
Support Vector Regressor	1.3638	2.0259

Relativize to Goals - Next Steps

The results achieved are underwhelming. However, before moving on to further model exploration and tuning, a more alarming fact is luring in sight. Given the current stage of the Kaggle competition, top scorers with complex, uninterpretable stacked models do not outperform the baseline by large margin[5]. Practical upper bound for model improvement seems low.

Although kaggle includes a hidden test set and our performance may not fully translate, still the current best *SMAPE* score is only 1.2366 and out of all 3548 teams, the 100th team, or the top 3% teams, achieve a score of 1.3717 which is just on par with our results, while the 1000th team scored 1.3791 justin beating the baseline, meaning that 70% of the models cannot outperform the last value.

With such observations, we question the overall signal to noise ratio in census data relating to Δ_{MBD}. In other words, we question whether changes are mostly dominated by noise in the short term, if so, we consider a change of objective for the project.

An intuitive objective stands, to predict the first moment, or the expected value of MBD county wise rather than second moment county and time wise. For such a problem, we answer the question:

What is a County's Microbusiness Density Given its Economic and Demographic Characteristics?

With such a question statement, most sources of our current data stay useful while our target value becomes less stochastic when taking the moving average. The team is in discussion to move forward with this proposed direction as an option.

Acknowledgement

I would like to express my gratitude to my teammates, Nicholas Amirsoleimani, Leon Weingartner, and Spencer Slaton, who all have supported and collaborated with diligently throughout this project, in many aspects not limited to problem formulation, data acquisition, exploratory data analysis, preliminary modeling, and model checking. Their contributions and hard work have been invaluable to the progress of the group.

I would also like to express gratitude to professor Libor and Thomas for their invaluable guidance and advice, especially the insightful interim feedback which suggested data source checks, quality control measures, and the potential change of objective in case of low signal to noise ratio.

Lastly, I express gratitude to classmates that lectured across the duration of the current progress, many of their topics proved useful as tools or intuition checks for the project.

References

[1] GoDaddy Microbusiness Density Survey:
https://www.godaddy.com/ventureforward/microbusiness-datahub/
https://www.anderson.ucla.edu/about/centers/ucla-anderson-forecast/projects-and-partnerships/godaddy
[2] Milken Institute Microbusiness Impact Report:
https://milkeninstitute.org/report/best-performing-cities-microbusiness-activity
[3] Exploratory Data Analysis Visualization with Normalized Data:
https://zhengpu-berkeley.github.io/MBD_Repo/
[3] Exploratory Data Analysis Visualization with Raw Data:
https://zhengpu-berkeley.github.io/MBD_repo_raw/
[4] More Preliminary Modeling Results by Other Members:
https://colab.research.google.com/drive/128YAffffZP81xzy8CmEkbYS_ItlfM_oA?usp=sharing
[5] Kaggle Competitors' Performances:
https://www.kaggle.com/competitions/godaddy-microbusiness-density-forecasting/leaderboard

Appendix

[1] Github Repository:
https://github.com/zhengpu-berkeley/GoDaddy_Interim_1
[2] Interim Presentation:
https://docs.google.com/presentation/d/1yPi9pdm2HWKUM
VMgjQPl5hc9-fV6QUtQEcDoKZVrvuk/edit?usp=sharing