A Framework for Microbusiness Density Forecasting

Eimantas Zaranka[1,2], Dmytro Klepachevskyi[1,2], Bohdan Krushelnytskyi[1,2] and Tomas Krilavičius[1,2]

[1] *Vytautas Magnus university, Kaunas, Lithuania*
[2] *Centre of Applied Research and Development, Lithuania*

Abstract

Microbusinesses are a vital part of a country's economy, and forecasting their density is crucial for both governments and hosting providers. Accurate predictions enable the government to plan future benefits for business owners, and hosting providers can efficiently allocate resources. In this study, we use data provided by Forward Venturer by GoDaddy, along with data collected from the U.S. Census Bureau to develop a model for microbusiness density forecasting. During the study we performed experiments using various machine learning techniques, including linear regression (LR), Ridge, Lasso, ElasticNet regression, decision tree (DT), random forest (RF), multilayer perceptron (MLP), gradient boosting, Ada boosting, support vector machine (SVM), XGBoost, LGBM, and TensorFlow decision forest (TFDF) regressors, as well as several neural network architectures such as multilayer perceptron (MLP), recurrent neural network (RNN), long short-term memory (LSTM), N-BEATS, and autoencoder. The performance of each model was evaluated using MAE and SMAPE metrics. This study highlights the potential of various machine learning and neural network algorithms for forecasting microbusiness density, which can aid in better resource planning for hosting providers and the government.

Keywords

Microbusiness, density forecasting, feature selection, machine learning, regression.

1. Introduction

A microbusiness is a small-scale enterprise with fewer than 10 employees, often operated by a sole proprietor or a small team. These small enterprises have been instrumental in creating jobs and generating economic growth in local communities. Microbusinesses are often too small or too new to show up in traditional economic data sources, but microbusiness activity may be correlated with other economic indicators of general interest.

In recent years, there has been a growing interest in the density of microbusinesses in different regions of the United States. County-level data on microbusiness density can provide valuable insights into the economic landscape of a region and help policymakers and entrepreneurs make informed decisions about investment and growth opportunities. In this context, analyzing the distribution of microbusinesses across different counties in the US can provide valuable insights into the factors that promote or hinder entrepreneurship and small business growth.

The Venture Forward team at GoDaddy were working on collecting data assets over the past years and launched the competition which goal is to predict monthly microbusiness density across the United States. This work will help policymakers gain visibility into microbusiness density as this often play a vital role in providing income and is a growing trend of small entities.

We were provided with the initial data of the access to broadband, population over age 25 with a 4-year college degree, percentage of foreigners, percentage of workers in information related industries, and the median household income in the county, that is publicly available at Census Bureau.

28th Conference on Information Society and University Studies (IVUS'2023), May 12, 2023, Kaunas, Lithuania
EMAIL: eimantas.zaranka@stud.vdu.lt (E. Zaranka), dmytro.klepachevskyi@stud.vdu.lt (D. Klepachevskyi), bohdan.krushelnytskyi@stud.vdu.lt (B. Krushelnytskyi) tomas.krilavicius@vdu.lt (T. Krilavičius)
ORCID: https://orcid.org/0000-0001-8509-420X (A. 4)

Additionally, to the initial data, we were encouraged to use any other useful data. We have found many useful variables that could make the prediction more accurate and help to realize more advanced approaches to improve predictions, and they are described in Section II.

There is a lack of comprehensive studies analysing microbusiness density in the United States. While there has been some research on small businesses [1], which generally include those with up to 500 employees, there has not been as much attention given to micro businesses, which typically have fewer than 10 employees. This is even though microbusinesses make up a significant portion of the overall business landscape in the US. As such, there is a need for more research on this important sector of the economy.

The rest of the paper is organized as follows. Literature review of the microbusiness density is presented in Section II. Section III provides description of the data that will be used for forecasting. Section IV presents data preprocessing with subsections of data combination, data preparation and features selection. Forecasting results are provided in Section V. Section VI describes the selected techniques. Finally, concluding remarks regarding forecasts are discussed in Section VII.

2. Literature review

We believe that the analysis of microbusiness is a relative new topic that has not received much attention so far. Microbusiness density has been the subject of interest only in [2]. The goal of this study was to determine the factors that influence local microbusiness venture density and the factors that are influenced by it. Authors employed several quantitative analysis techniques, including Ordered Least Square regression, Probit with the Huber-White sandwich estimator of variance, and Ordered Probit for the equivalent ordinal variable estimate. The results suggest that there is a significant relationship between microbusiness density and:

- employment level
- distribution of self-employed/wage-employee
- population density
- business turnover
- urban/rural area indicator
- gender distribution
- education
- prosperity indexes
- internet usage data
- distribution of ethnic groups.

3. Methodology

In this study, various models were utilized to predict the target variable. Models were selected based on their usage in previous competitions and included popular regression models such as Linear Regression, Decision Tree Regressor, XGB Regressor, and others. The performance of these models was evaluated using the symmetric mean absolute percentage error (SMAPE) as an accuracy metric, which was required by the competition evaluation rules and mean absolute error (MAE). These chosen metrics are defined as follows

$$SMAPE = \frac{100}{n}\sum_{t=1}^{n}\frac{|F_t - A_t|}{(|A_t| + |F_t|)/2}, \quad (1)$$

$$MAE = \frac{1}{n}\sum_{t=1}^{n}|A_t - F_t| \quad (2)$$

where A_t is an actual value and F_t is forecasted value.

To examine the relationships between variables, a correlation analysis was conducted using the Pearson correlation coefficient. This statistical method was selected for its ability to quantify the strength and direction of linear associations between variables. Pearson correlation coefficient of two variables *X* and *Y* is formally defined as

$$\rho_{X,Y} = \frac{cov(X,Y)}{\sigma_X \sigma_Y}, \tag{3}$$

where $cov(X,Y)$ is covariance of the two variables, σ_X is standard deviation of *X* and σ_Y is standard deviation of *Y*.

We applied data normalization to ensure that all attributes have an equal effect on the resulting variable. Min-max normalization was used to scale the variables between zeros to one:

$$x_i' = \frac{x_i - min(X)}{max(X) - min(X)} \tag{4}$$

where *min(X)* and *max(X)* represent the minimum and maximum values of *X* respectively.

To better understand the differences between counties we applied K-means clustering algorithm. The idea behind K-means is to partition the data $D = \{X_1, X_2, \ldots, X_n\}$ into k $(k < n)$ clusters so that so the objects within a cluster are more similar to each other than the objects in different clusters. In this case, objects similarity was measured based on Euclidean distance. Stepwise K-means clustering algorithm can be defined as follows.

1. Randomly select k initial objects as centroids.
2. Calculate the distance between each object and centroids.
3. Assign each object to the nearest cluster.
4. Calculate the mean of each cluster as new centroid.
5. Repeat steps 2 – 4 until convergence.

The objective of K-means algorithm is to minimize the squared error function:

$$E = \sum_{i=1}^{k} \sum_{X \in C_i} |X - c_i|^2 \tag{5}$$

where X – is a point in space representing a given object, c_i is the mean value of cluster C_i.

To define optimal number of clusters we used Davies-Buildin Score and Elbow Method. Davies-Buildin Score [3] is defined as the ratio between the within cluster scatter and the between cluster separation

$$DB = \frac{1}{N} \sum_{i=1}^{N} \max_{i \neq j} \frac{S_i + S_j}{M_{i,j}} \tag{6}$$

where S_i is a measure of scatter within the *i*th cluster defined as

$$S_i = \frac{1}{n_i} \sum_{X \in C_i} d(X, c_i) \tag{7}$$

and $M_{i,j}$ is a measure separation between *i*th and *j*th clusters:

$$M_{i,j} = d(c_i, c_j) \tag{8}$$

Lower value of Davies-Buildin Score indicates the better clustering.

The Elbow Method [4] is a partitioning method, where the goal is to define clusters such that the total intra-cluster variation is minimized

$$minimize\left(\sum_{i=1}^{k} W(C_i)\right), \quad (9)$$

where C_i is the ith cluster and $W(C_i)$ is the within cluster variation.

After the calculations, the results are plotted according to the number of clusters, and the location of a bend in a plot is usually considered an indicator of the appropriate number of clusters.

4. Data description

The initial dataset used in this study was provided by GoDaddy [5]. It consisted of information on microbusiness density for 3135 counties. The data covered the period from August 2019 to December 2022. Thus, each county was described by 41 monthly observations. In total, the dataset had 128 535 observations and 7 features. The description of the dataset can be seen in Table 1. We also used additional information obtained from U.S. Census Bureau and covering the period from 2017 to 2021. A detailed description of this dataset is shown in Table 2.

The target variable represented the number of microbusinesses per 100 people age over 18 in the given county. Due to the ACS update window, the population figures used to calculate the microbusiness density are on a two-year lag. This means that the microbusiness density for 2022 was calculated using population figures from 2020.

Table 1
Venture Forward Survey Data

Feature	Description
row_id	ID code consisting of cfip and first day of the month columns
cfips	A unique county identifier using Federal Information Processing System, where first two digits corresponds to the state FIPS code, while following three numbers represents the county
county_name	Name of the county
state_name	Name of the state
first_day_of_month	The date of the observation. Consists of year, month, and day
microbusiness_density	Number of microbusinesses per 100 people age over 18 in the given county. This is a target variable.
active	The microbusiness in the county (not provided in future forecasting)

Table 2
ACS Survey Data

Feature	Description
pct_bb_[year]	The percentage of households in the county with access to any type of broadband. Derived from ACS table B28002
cfips	County identifier
pct_college_[year]	The percentage of the population in county over age of 25 with a 4-year college degree. Derived from ACS table S1501
pct_foreign_born_[year]	The percentage of population that was born outside of United States. Derived from ACS table DP02
pct_it_workers_[year]	The percentage of population that is employed in IT related fields. Derived from ACS table S2405
median_hh_inc_[year]	The median income of household. Derived from ACS table S1901

In addition to training data, a testing dataset was also provided, which had the same structure and features as Table 1, but with missing target feature *microbusiness_density* and *active* column, which

is yet unknown. The forecasts should cover the period from January to June 2023 for all 3135 counties.

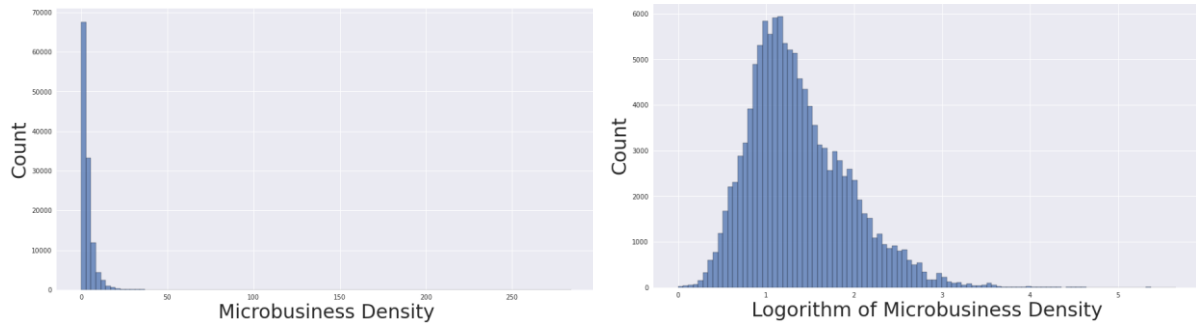
The organizers strongly encouraged the use of external data sources that might help to improve prediction performance. Therefore, we enriched the initial data using publicly available sources such as the ACS website. Every externally collected dataset had the same structure as Table 2, where each column contained information for the given year. The following information was gathered:

- Business turnover [6].
- Population estimates [7,8].
- Demographic of population in counties [7,8].
- Internet usage [9].
- Unemployment [10].
- Education of populous [11].
- Ethnicity of counties [12].
- Geographical location of counties [13].

4.1. Explanatory data analysis

To get a better understanding of the initial data, we performed the following steps: we checked the distribution of target values, performed correlation analysis, and checked for possible seasonality.

The microbusiness density distribution revealed that most observations are centered around zero, see Figure 1(a). Moreover, data distribution is skewed to the right. For this reason, we applied the log transformation, as shown in Figure 1(b). Based on these results, we can conclude that the microbusiness density follows a close-to-log-normal distribution.



(a) (b)

Figure 1: (a) original microbusiness density distribution, (b) logarithmically transformed microbusiness density.

Analyzing microbusiness data on a state level, we observed that several states had higher density of microbusinesses, i.e., California, Colorado, Delaware, Wyoming, Utah, Nevada, and Florida. The average microbusiness density per county in these states was 8.38, while the rest of the U.S. states had an average of 3.41, with the standard deviations of 13.21 and 3.06 respectively. The two most outstanding states were Delaware with an average of 18.74 and Nevada with 12.42 microbusinesses density per county, as shown in Table 3. To account for these fluctuations, we decided to perform a clustering analysis. To identify the most appropriate number of clusters, we performed experiments with a different number of clusters $k = 2,3,\dots 30$. The maximum number of clusters was based on the rule of thumb, i.e.

$$k \sim \sqrt{n/2}, \qquad (10)$$

where n is a number of observations.

Table 3

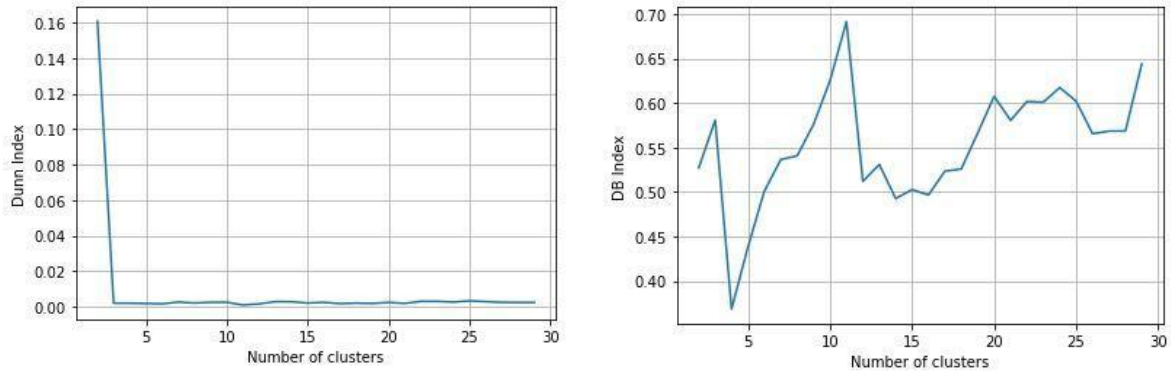
Highest density states information

State	Avg. microbusiness density	Standard deviation of microbusiness density	Avg. active microbusinesses	Standard deviation of

				active microbusinesses
California	7.5628	4.5661	57343.3090	161610.6671
Colorado	8.7500	9.7034	7451.0907	17404.2814
Delaware	**18.7419**	**15.0966**	**48959.5772**	**37695.6337**
Florida	6.9452	5.6750	29589.2766	63021.4941
Nevada	**12.4261**	**31.0850**	**26526.8292**	**86315.5518**
Utah	8.3480	6.6135	8550.141295	20837.4785
Wyoming	9.3745	19.7700	2142.0381	4998.8983
All states	3.8278	5.0593	6461.1692	33117.5875

To compare clustering results obtained using different *k* values, we used Davies-Bouldin Score (DB Score) and the Elbow Method. The results are illustrated in Figure 2. In this case, the Elbow Method did not show a clear indication of an elbow point (see Figure 2(a)), while the DB Score indicated that the optimal number of clusters for this dataset is four (see Figure 2(b)). A more detailed analysis revealed that extracted clusters quite well represent the distribution of microbusiness density among counties. For instance, cluster number two represents counties having the largest density, i.e., the average value of microbusiness density in this cluster is 62.13 with a standard deviation of 51.19. In contrast, Cluster 3 has the lowest results, with an average of 2.20 and a standard deviation of 1.02. For more details, please refer to Table 4.

Based on these results, we included an additional variable representing the counties' membership among clusters.



(a) (b)

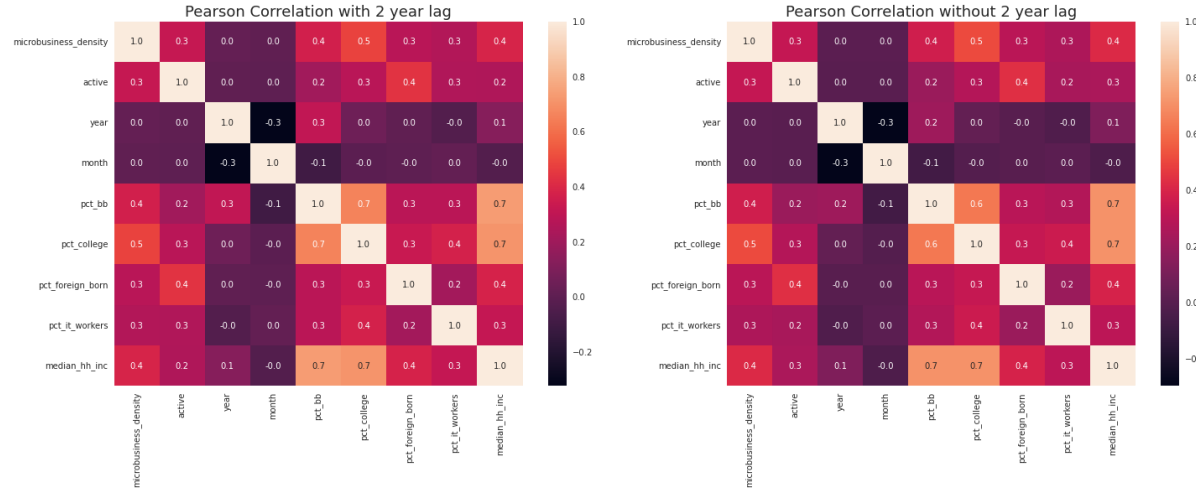
Figure 2: (a) The Elbow Method, (b) Davies-Bouldin Score results

Table 4
Clusters density information

Cluster	Avg. microbusiness density	Standard deviation of microbusiness density	Avg. active microbusinesses	Standard deviation of active microbusinesses
0	6.72	1.99	14277.95	28691.87
1	16.62	8.61	73920.18	141360.26
2	**62.13**	**51.19**	**13116.62**	**20907.04**
3	2.20	1.02	824.60	1847.07
All states	3.83	5.06	6461.17	33117.59

To analyze relations between microbusinesses density and other factors we performed correlation analysis. Due to the nature of microbusiness density, we focused on two possible variations. The first approach sought to understand how a 2-year lag impacts the correlation, and second, how features correlate with each other. As can be seen in Figure 3, a 2-year lag has a positive impact on the correlation

between features. A weak-to-medium linear dependency can be observed between features, with the correlation coefficient ranging between 0.2 and 0.6. The target value correlates the most with the college education feature (value 0.5), followed by broadband usage and median income features (values 0.4), as shown in Figure 3(a)



(a) (b)

Figure 3: (a) Pearson correlation with 2-year delay, (b) Pearson correlation results without 2-year delay.

Lastly, the seasonality analysis was conducted. It was noticed that the majority of the counties has a noticeable increase in microbusinesses during festive periods and decrease during inter-holiday period. The rest of the counties had either no seasonality or very small seasonality.

4.2. Data preparation

The process of data preparation is essential for machine learning model performance. The following preprocessing steps were performed to prepare the data for the modelling stage:

1. **Combination of training and testing datasets.** Because a one-year lag will be introduced into the dataset, it is important to have a temporary full dataset. Before the combination two additional features were introduced: *is_test* column that marked the original sets and *dcount* that marked the sequence of observations in each county. This step ensures that the first row of each county will have a full set of features.
2. **Introduction of lag terms.** The dataset was enriched using series' own past values, so-called lags. In this case, we included 1-12 month lags representing the microbusiness density of the previous year.
3. **Imputation of missing values.** The merging external dataset introduced some missing values that were imputed using the mean value of the corresponding county feature.
4. **Splitting the combined dataset back into training and testing sets**. To prevent data leakage full dataset was split back into training and testing datasets using *is_test* feature introduced in the first step of data preparation.
5. **Removal of features with incomplete lag values in the training dataset.** The first eleven entries of each county in the training dataset were dropped due to the missing lag values.
6. **Numerical feature scaling.** All numerical features, except the target and lag values, were scaled to be in the range of [0, 1]. This ensures that machine learning models interpret all features on the same scale.
7. **Log-transformation**. Two different variations of datasets were created. One with original target and lag values, and the second, where target and lag values were logarithmically transformed.

8. **Categorical feature encoding.** All categorical features, i.e., state, county, cluster number were transformed using one-hot encoding.

The final dataset had a total of 3162 features. Seeking to avoid unnecessary complexity caused by the dataset's dimensionality we performed feature selection.

4.3. Feature selection

The data preparation step introduced many new features that negatively impacted the model. Therefore, statistical feature significance tests, specifically ordinary least squares regression tests [14], were conducted for feature selection. A stepwise feature selection procedure [15] was performed to ensure that all features had p-values less than 0.05. Based on this procedure, we identified 21 statistically significant features. Please refer to Table 5 for more details.

Table 5
Statistically Significant Features

Feature	p-value	Feature	p-value	Feature	p-value
lag_1	0.000	lag_9	0.000	state_Nevada	0.000
lag_2	0.000	lag_10	0.006	state_Wyoming	0.000
lag_4	0.000	lag_11	0.002	state_Delaware	0.000
lag_5	0.000	pct_college	0.000	state_Colorado	0.046
lag_6	0.014	broadband_usage	0.024	month_1	0.003
lag_7	0.000	numEstab	0.001	month_2	0.017
lag_8	0.000	18-24 some college or associate's degree	0.010	month_6	0.014

5. Results

Both original and logarithmically transformed datasets were split into training and validation sets. Training data contained observations from August 2019 to July 2022, and validation data from August 2022 to December 2022. Experiments were performed using various regression techniques, i.e., linear, Ridge, Lasso and ElasticNet regression; decision tree, random forest, multilayer perceptron, gradient boosting, Ada boosting, support vector machine, XGBoost, LGBM and TensorFlow decision forest regressors. Additionally, the following neural network architectures were trained: recurrent, multilayer perceptron, LSTM, N-BEATS and autoencoder. Finally, a validation test was used to estimate the model's performance. The obtained results are presented in Table 6.

Table 6
Validation dataset results

Model	Original target		Log target	
	MAE	SMAPE	MAE	SMAPE
Linear Regression	0.080	2.624	**0.057**	**1.710**
Ridge regression	0.079	2.600	**0.057**	**1.710**
Lasso regression	0.139	5.300	2.290	56.297
ElasticNet	0.117	4.359	2.290	56.297
DecisionTreeRegressor	0.095	2.720	0.088	2.651
KNeighborsRegressor	0.100	2.630	0.136	2.900
MLPRegressor	0.140	4.430	0.178	3.883
RandomForestRegressor	0.067	1.966	0.068	1.931
GradientBoostingRegressor	0.082	2.476	0.095	1.979

AdaBoostRegressor	3.560	80.690	1.038	29.384
SVR	0.178	2.463	0.110	2.658
XGBRegressor	0.074	2.064	0.073	1.930
LGBMRegressor	0.010	2.197	0.092	1.932
LSTM	0.058	1.759	0.072	1.888
RNN	0.054	1.690	0.066	1.821
N-BEATS	0.102	2.985	0.088	2.375
Tensorflow Decision Forest	0.072	1.885	0.061	1.790
AE+LSTM	0.849	22.838	0.080	1.847
MLP	**0.055**	**1.696**	0.451	9.056

The best performing model with original target values was a multilayer perceptron, achieving MAE of 0.055 and SMAPE of 1.696. On the other hand, the least accurate model was the AdaBoost Regressor, with MAE of 3.560 and SMAPE of 80.690. However, when using logarithmically transformed target values, the best performing models were the linear regression and Ridge regression, both with an MAE of 0.057 and SMAPE of 1.710. The worst models were ElasticNet and Lasso regressor, both models achieved an MAE of 2.290 and SMAPE of 56.297. It is important to note that the best performing models for the final submission may differ from the best performing models on the validation dataset.

Models trained on logarithmically transformed data showed superior results compared to those trained on original target values, as seen in Table 6. Consequently, the logarithmic transformation was selected for the final submissions. It is worth noticing, however, that the best performing models for the Kaggle competition differed from the models in experimentation phase. The subset of model's forecasting, that showed the highest results, can be seen in Table 7.

The results were evaluated on data that contained only January 2023 density values. The highest performing model is the XGBoost regressor with a SMAPE of 3.3159, followed closely by the Random Forest regressor with an SMAPE of 3.3189. The least accurate model is the recurrent neural network with SMAPE of 3.8251. The final and full results will be known on June 14th, 2023.

Table 7
Public Leaderboard Results

Model	SMAPE
Linear Regression	3.3711
RNN	3.8251
Ridge Regression	3.4195
Random Forest Regressor	3.3189
XGBoost Regressor	**3.3159**
XGBoost Regressor Ensemble	3.6199

6. Conclusions

In this paper, we presented the results of our investigation on microbusiness density forecasting using various machine learning techniques. Experiments were performed using the dataset from the GoDaddy Kaggle competition, which was enriched using external data sources. Explanatory data analysis revealed a significant positive relationship between microbusiness density, college education, broadband usage, and median income. Moreover, we observed that the distribution of microbusiness density is not uniform across counties, i.e., California, Colorado, Delaware, Wyoming, Utah, Nevada, and Florida had a higher density of microbusinesses. To account for these fluctuations, we performed a clustering analysis based on the K-means algorithm. As a result, four clusters were extracted and included in the analysis as an additional feature.

The final dataset consisted of 3162 features. To reduce the dimensionality of the data, we conducted feature selection using a statistical significance test. As a result, we identified 21 statistically significant features that were later used for modelling experiments. A detailed analysis of various regression

techniques revealed that the XGBoost method performed the best, with a SMAPE of 3.3159. We also discovered that logarithmically transforming microbusiness density values produced better validation results. Consequently, we chose the logarithmically transformed dataset to forecast unseen data. However, forecasting in an open system is unpredictable due to many factors that can positively or negatively impact the results. Therefore, such forecasting requires continuous model retraining to achieve the best possible results for the upcoming months.

7. References

[1] U.S small business administration, Office of Advocacy. URL: https://advocacy.sba.gov/category/research/.

[2] G. Saridakis, N. Litsardopoulos, C. Hand, Great Britain Microbusiness White Paper, 2022. URL: https://www.godaddy.com/ventureforward/wp-content/uploads/2022/03/GoDaddy_Great_Britain_Microbusiness_White_Paper_2022.pdf.

[3] U. Aickelin, I. Dent, T. Craigy and T. Roddenz. An approach for assessing clustering of households by electricity usage. In proceeding of: UKCI 2012, 12th Workshop on Computational Intelligence, 2012

[4] B. Boehmke, K-means Cluster Analysis. URL: https://uc-r.github.io/kmeans_clustering#kmeans

[5] K. J. Gracey, Dataset Description, 2023. URL: https://www.kaggle.com/competitions/godaddy-microbusiness-density-forecasting/data.

[6] United States Census Bureau, County Business Patterns, 2020. URL: https://data.census.gov/table?q=CBP2020.CB2000CBP&g=010XX00US$0500000&tid=CBP2020.CB2000CBP.

[7] United States Census Bureau, County Population Totals: 2010-2019. URL: https://www.census.gov/data/datasets/time-series/demo/popest/2010s-counties-total.html

[8] United States Census Bureau, County Population Totals and Components of Change: 2020-2022, 2022. URL: https://www.census.gov/data/datasets/time-series/demo/popest/2020s-counties-total.html.

[9] A. Thomas, Broadband Usage in US, 2022. URL: https://data.world/amberthomas/broadband-usage-in-us.

[10] United States Census Bureau, Employment Status, 2021. URL: https://data.census.gov/table?q=S2301&g=010XX00US$0500000.

[11] United States Census Bureau, Educational Attainment, 2021. URL: https://data.census.gov/table?q=S1501&g=010XX00US$0500000.

[12] B. Dill, County level population by race ethnicity 2012-2019, 2020. URL: https://data.world/bdill/county-level-population-by-race-ethnicity-2010-2019.

[13] Latitude Longitude Team, States in United States, 2012. URL: https://www.latlong.net/category/states-236-14.html.

[14] Lumivero, XLSTAT: Ordinary Least Squares Regression (OLS). URL: https://www.xlstat.com/en/solutions/features/ordinary-least-squares-regression-ols

[15] M. Kuhn, K. Johnson, Feature Engineering and Selection: A Practical Approach for Predictive Models, 2019. URL: https://bookdown.org/max/FES/greedy-stepwise-selection.html