**PREDICTING MICROBUSINESS DENSITY IN USA**

**Ketan Kapse G24637377**

**Sai Abhishree Pappusetty G26078396**

**INTRODUCTION**

Microbusinesses (MBs) are typically defined as businesses that operate on a very small scale and have an online presence. They have fewer than 10 employees and are usually managed directly by their owners. Examples of such businesses can be local cafes, restaurants, grocery stores, etc. These types of businesses are important for several reasons:

1. Economic growth: MBs play an important role in the growth of the local economy since they provide employment opportunities and are a means of income for individuals living in the local geographic area. They also play an important part in the country’s economy. According to the Small Business Association, there were approximately 30.7 small-scale businesses in the US in 2019, with MBs constituting about 89% of them [1].
2. Innovation: MBs are often run by entrepreneurs who are innovative and adapt quickly to changes in the market. They can develop new products and services that meet the needs of their customers.
3. Diversity: Many MBs have owners that hail from diverse backgrounds, who in turn also employ individuals from such backgrounds promoting diversity and inclusivity.

A major issue associated with MBs is that they are often too small or too new to show up in traditional economic data sources, making it nearly impossible for policymakers to study them and develop plans to promote their growth. Moreover, in 2022, these businesses were reported to employ around 66 million US citizens, which is approximately 46.4% of the working population [2] [3]. This is a significant number and, hence, the impact of microbusinesses can not be overlooked.

Regions with higher levels of microbusiness development and activity can benefit from policies and funding that would foster their growth, strengthening the economy. This project aims to predict the microbusiness density levels for US counties using multiple machine learning models. We believe that our project will act as a steppingstone for more in-depth analysis of MBs.

**DATASET**

We are using the GoDaddy microbusiness dataset [4] for this project. The dataset consists of 3135 timeseries data of counties in the US. Each of these series has a length of 39, with each datapoint corresponding to a single month ranging from August 2019 to October 2022 and its associated microbusiness density level. The dataset is split into two files, train.csv and test.csv as the training and testing data respectively.

For the test set, we need to make forecasts of the microbusiness density levels for the next eight months, i.e., November 2022 to June 2023. Since, each timeseries corresponds to a county, we need to make 3135 \* 8 predictions which is approximately 25000 predictions.

A screenshot of a computer

Description automatically generated with medium confidence

Figure 1. Training Dataset

A screenshot of a computer

Description automatically generated with medium confidence

Figure 2. Test Dataset

From the above images, we can gather what the training and test datasets look like. The description of each feature is as follows:

* row\_id: It is an ID code for the row.
* cfips: A unique identifier for each US county using the Federal Information Processing System. The first two digits correspond to the state FIPS code, and the remaining 3 represent the county.
* county\_name: The name of the county.
* state\_name: The name of the state.
* first\_day\_of\_month: The date of the first day of the month.
* microbusiness\_density: MBs per 100 people over the age of 18 in the county.
* active: The raw count of MBs in the county.

**METHODOLOGY**

We utilized the following regression models to predict the microbusiness densities for the counties:

1. Linear Regression: This method is used to model the relationship between a dependent variable and one or more independent variables. It assumes a linear relationship between the variables, and the goal is to find the best-fit line that minimizes the sum of squared errors between the observed and predicted values. The resulting equation can be used to predict the value of the dependent variable for a given set of independent variables.
2. Ridge Regression: This is a regularization technique used in linear regression to reduce the impact of multicollinearity and prevent overfitting. It adds a penalty term to the loss function to shrink the regression coefficients towards zero.
3. Lasso Regression: Lasso regression is another regularization technique used in linear regression to prevent overfitting and perform feature selection. It adds a penalty term to the loss function that encourages some regression coefficients to be exactly zero, effectively eliminating the corresponding features from the model.

The following statistical models were also used:

1. Auto Regression: This is a time series analysis method that models the relationship between a variable and its past values. It uses the variable's own lagged values as predictors in a linear regression model to make predictions about future values.
2. Moving Average: This is another time series modelling method that models the relationship between a variable and its past forecast errors. It calculates the mean of the errors over a rolling window of time and uses this as a predictor.
3. ARMA: This method combines autoregression (AR) and moving average (MA) models. It models the relationship between a variable and its own past values and past forecast errors, using both autoregressive and moving average terms in a linear regression model.
4. ARIMA: This combines autoregression (AR), differencing (I), and moving average (MA) models. It models the relationship between a variable and its own past values and past forecast errors, while also accounting for non-stationarity through differencing.

Every single one of the above models was trained on the training set and predictions for all 3135 counties were made using the trained models.

A picture containing text, screenshot, font

Description automatically generated

Figure 3. Workflow of the project

**RESULTS**

In case of linear, lasso and ridge regression, if the train error was found to be more than 15 percent (value selected arbitrarily), we assumed that the particular time series data did not follow a linear trend and we used the last known value in the train and test predictions. We evaluated these models using the MAPE metric. Upon further analysis, we observed that these three models performed almost identically on counties with linear trends and had similar MAPE values.

A picture containing text, line, plot, screenshot

Description automatically generatedA picture containing text, screenshot, plot, diagram

Description automatically generated

Figure 4. Lasso Regression on CFIPS 1003

A picture containing text, line, plot, diagram

Description automatically generated

A picture containing text, screenshot, diagram, plot

Description automatically generated

Figure 5. CFIPS 4025 does not show linear trends within the error threshold.

*Issue with these models:* A major issue with using these models for predictions is that only a small minority of the counties’ time series have linear trends within the error threshold (around 400), making it necessary to incorporate more models for the other counties.

For the AR, MA, ARMA and ARIMA models, imported from the statsmodels library, we first utilized the auto\_arima model from the pmdarima package to generate p,d and q (non-seasonal) orders for each county data. Auto ARIMA basically fits all of the above models onto the county data and generates the orders after taking into account the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) values. The lower these values, the better is the model. This allowed us to avoid manually fitting each of the above models one by one on the 3135 time series data.

A picture containing text, font, screenshot, algebra

Description automatically generated

Figure 6. Formulae for AIC, BIC and LogLikelihood.

A screenshot of a computer

Description automatically generated with medium confidence

Figure 7. Fit summary of the ARMA model on CFIPS 13035 using the order generated by Auto ARIMA.

Next, we stored these orders in a dictionary and used them to fit each of the models on the county data.

*Issues with these models:* Time complexity of these models is a big issue since we need to find the perfect orders for all the 3135 counties. Generating the orders for the counties using Auto ARIMA took approximately 2 hours on a CPU with 8 threads. Furthermore, fitting the models on the data also took a while, totaling around 2.5 hours for all the models.

Also, our approach of using the generated orders in the models brought up convergence warning during training, which indicated that the models did not fit their best orders for some counties.

Finally, we merged the predictions of all the 7 models into a single .csv file.

A screenshot of a computer

Description automatically generated with low confidence

Figure 8. First 20 rows of merged.csv

**FURTHER WORK**

Our project serves as a great baseline for developing further insights into microbusiness activity in different counties.

Results can be further improved if seasonal data is taken into account. For example, in the following figure, the green boxes highlight the presence of linear seasonal patterns in data for the county with cfips = 5007.

A picture containing line, plot, diagram, font

Description automatically generated

Figure 9. Seasonal Trends in CFIPS 5007

Considering seasonal variations will help provide more accurate forecasts especially with multiple seasonal patterns being present in the counties’ time series [8].

Furthermore, inclusion of more data about the counties themselves can be used to make better predictions. Features from US census data for each county can be used to fine tune the forecasts.

Models such as tree based regressors like XtremeGradientBoosting and LightGradientBoosting can also be utilized as they have shown to have great performance on time series problems [5] [6] [7].

**CONCLUSION**

We analyzed the microbusiness density dataset and trained several models on the time series data of each county in the dataset. Furthermore, we generated forecasts for the next 8 months present in the test set. Also, the drawbacks of each model were discussed, and possible future work mentioned.

**REFERENCES**

[1] <https://cdn.advocacy.sba.gov/wp-content/uploads/2021/08/30144808/2021-Small-Business-Profiles-For-The-States.pdf>

[2] https://www.chamberofcommerce.org/small-business-statistics/

[3] <https://advocacy.sba.gov/wp-content/uploads/2022/04/Small-Business-Job-Creation-Fact-Sheet-Apr2022.pdf>

[4] https://www.kaggle.com/competitions/godaddy-microbusiness-density-forecasting/data

[5] Tianqi Chen, et al. "XGBoost: A Scalable Tree Boosting System". *CoRR* abs/1603.02754. (2016).

[6] Ke, Guolin et al. "LightGBM: A Highly Efficient Gradient Boosting Decision Tree." *Proceedings of the 31st International Conference on Neural Information Processing Systems*. Curran Associates Inc., 2017.

[7] Shereen Elsayed, et al. "Do We Really Need Deep Learning Models for Time Series Forecasting?". *CoRR* abs/2101.02118. (2021).

[8] Phillip G. Gould, et al. "Forecasting time series with multiple seasonal patterns". *European Journal of Operational Research* 191. 1(2008): 207-222.