Machine Learning Analysis of Microbusiness Densities Across US Counties

Leon Weingartner, Nicholas Amirsoleimani, Spencer Slaton, Zhengpu Zhao

***Abstract*—This project has been a two-part analysis using machine learning techniques to investigate the behaviors and driving features of microbusiness densities in the US on a county level. The project initially started as a Kaggle competition sponsored by the company GoDaddy which structured the competition as a forecasting problem with over 3000 teams competing. The training data provided *[1]* was on a monthly basis where many novel time series methods were used to reduce the error criteria SMAPE (symmetric Mean Absolute Percentage Error).**

I. Introduction

A. Project Overview

Our focus is on exploring county-level micro business density data, defined as businesses with an online presence and ten or fewer employees. Microbusiness density is determined by the number of micro-businesses for every 100 people in each county. We will cover 2 parts of this project with separate objectives. The first objective is to capture the monthly variability of microbusiness density as a forecasting problem. The second objective will involve predicting the driving features of the microbusiness density mean as an aggregate of the data from 2022. This pivot of our project objective is due to the ending of the Kaggle competition and the lack of useful insights from noisy time series data. The original dataset featuring the target variable is available on Kaggle as part of a competition sponsored by GoDaddy, a web domain sales company.

We denote our target variable as MBD (micro business density). Apart from its benefits to GoDaddy, policymakers can also utilize the prediction outcomes to gain insights into month-to-month microbusiness trends and understanding influential factors explaining MBD, influence resource allocation, and identify areas with low microbusiness densities. These predictions can potentially inform investment decisions as well. Local microbusiness knowledge can prove to be valuable information in determining the feasibility of starting a business in a specific county. Companies like Shopify, Turbo Tax, Office Depot, and MailChimp are some commercial interests that could benefit from such insights.

B. Forecasting Questions

For the Kaggle competition, we try to answer the following questions:

1. What features capture the month-to-month movement of MBD?
2. How can we utilize machine learning and GLM models to forecast MBD?
3. What insights can we obtain from this problem and how useful are the results to others?

C. Expected MBD Questions

The second portion of this project involved taking a mean of our data and performing prediction analysis on the resulting aggregated data. We became interested in the driving features of MBD as a regression problem rather than a forecasting one. We attempt to answer the following questions:

1. What features influence the expected MBD of a county?
2. How can we predict outlier counties as a classification problem?
3. What insights can be made from the expected MBD and how is it useful?

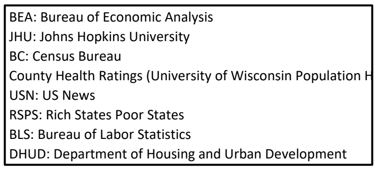
II. Data

A. Sources

The table below displays the various sources of data used in this project. Our target variable was monthly micro-business density (MBD) from 2019 to 2022, which we derived from a combination of survey data obtained by ASU, UCLA, and UIowa, as well as usage information of over 20 million GoDaddy registered micro-businesses *[3]*.

Table 1, Raw data sources and properties

Raw Data Sources and Properties			
Name	**Source**	**Longitude**	**Geography**
Microbusiness Density	GoDaddy	Monthly	County
Real & Sector GDP	BEA	Yearly	County
Population Census	BC	Yearly	County
Demographics Census	BC	Yearly	County
Covid-19 Death	JHU	Monthly	County
ChatGPT Dialogue	ChatGPT	Static	County
Google Search Trends	Google	Monthly	State
Google Search Trends	Google	Monthly	County
Business Tax	RSPS	Static	State
Health	CHR	Yearly	County
Education	CHR	Yearly	County
Crime	CHR	Yearly	County
Rent	DHUD	Yearly	County
Coastline	BC	Static	County
Nearest University	USN	Static	County
Unemployment	BLS	Month	State



BEA: Bureau of Economic Analysis
JHU: Johns Hopkins University
BC: Census Bureau
County Health Ratings (University of Wisconsin Population H
USN: US News
RSPS: Rich States Poor States
BLS: Bureau of Labor Statistics
DHUD: Department of Housing and Urban Development

In addition to the target variable, we also included various economic and demographic data from corresponding agencies such as the Bureau of Economic Analysis (BEA), Google Trends search frequency data, and county description dialogue data from ChatGPT as covariates.

B. *Target Variable & EDA*

We initially investigate our target variable MBD by viewing the distribution *(top)* and standard deviation of MBD over time *(bottom)*.

Figure 1, MBD distribution

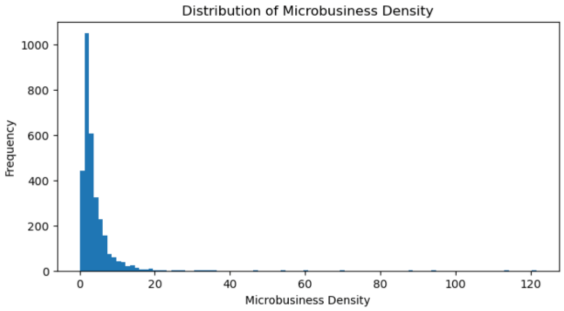
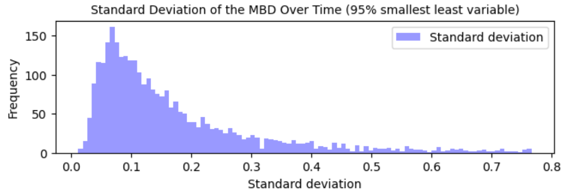
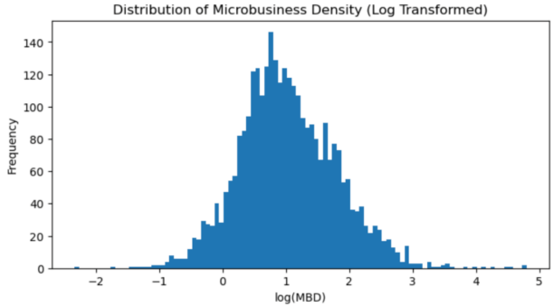


Figure 2, MBD Std. over time



We see from *Figure 1* that our target variable distribution is skewed to the right closely resembling a pareto distribution. We choose to take the log transformation of MBD to improve the performance of certain statistical analyses and machine learning models. The purpose of this transformation is to reduce the effect of extreme values and make the distribution of MBD more symmetric. From *Figure 2*, our standard deviation of MBD over time is very small indicating not much movement month-to-month.

Figure 3, MBD distribution (log transform)



C. *Collection and Wrangling*

We've included google search data following different string queries on both the county and state level to simulate the level of interest (ex. Alameda County Tax, how to start a small business, business loan, etc.). For some small counties with low population, we had little to no data to capture the high variance that was present. The movement of some of these counties seemed to be a result of noise rather than influence from our collected data.

Chat GPT responses were also used to obtain county-level data. The prompt given to Chat GPT followed as "Give me bullet points for why I should or should not start a small business in county, state, 3 pros and 3 cons". The results of this query gave us consistent responses making it easy to parse and encode into a neural network. This novel approach to feature engineering comes with several caveats. For quality assurance, we've estimated the response accuracy to produce 1 error for every 42 bullet points within the pros and cons list. This estimate is to be taken with a grain of salt as our manually checked sample size was small and is expected to produce more contradictory results with smaller counties that have little to no data relative to GPT. We also didn't take into account the dependency between different responses such that a previous query for a different county may influence the outcome of future prompts resulting in a higher error rate. The tokenized word vectors were put into a pre-trained BERT model. The output was then combined with K-means clustering that was tuned using ARI and NMI to obtain K=12 clusters. Each cluster involved our original bullet points that were similar to other bullet points from different counties. We manually named the 12 clusters and used these as a flag feature for all counties. Some examples: [Low Cost of Living, Low Pop Bad Weather, Tough Labor Market, etc].

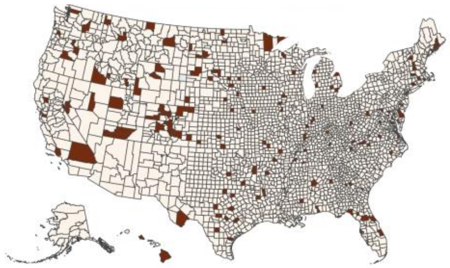
We've extracted other features primarily by web scraping. A list of features used in this report can be found in the appendix of this paper.

III. Part 1 (Forecasting MBD)

A. *Outliers and Collinearity*

Geographically, the lesser populated counties show a higher standard deviation over time. This provides insight into highly volatile counties that could potentially be separated from more stable/populated counties that only require minimal data to accurately predict micro business density in the upcoming months.

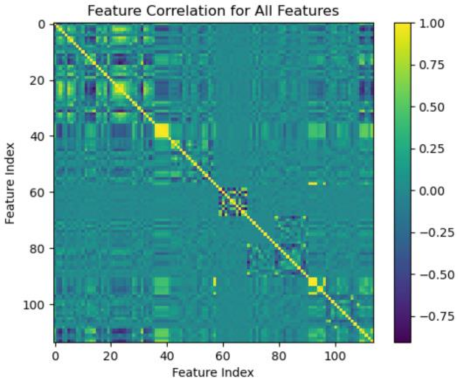
Figure 4, top 5% of most variable counties



B. Collinearity

We use a collinearity heatmap to identify potential multicollinearity issues in our dataset. In such cases, it becomes difficult to distinguish the individual effects of each variable on MBD, and the regression results may become unstable and unreliable. For features with higher levels of collinearity, regularization models such as ridge and LASSO methods are used in the preliminary model analysis section.

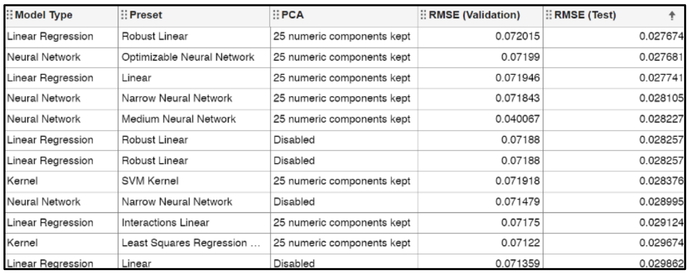
Figure 5, Collinearity Heat Map



C. Model Analysis

In this section, we explore the use of various regression models, decision trees, ensembles, and neural networks to predict micro-business density in different counties. To evaluate the performance of these models, we calculate the root mean squared error (RMSE) for both training and testing sets. Our train-test-split is broken down by a training set of 34 months from Jan-2020 to Oct-2022, and a test set of 2 months from Nov-2022 to Dec-2022. Additionally, we compare the performance of each model to a baseline model that only uses the previous month’s data to predict MBD. By conducting this preliminary analysis, we aim to identify the most effective models so we can establish a foundation for future model development. A table of the top 12 models used sorted by decreasing test RMSE can be found below *(Table 2)*. Note that our baseline validation RMSE is .1599 and our baseline test RMSE is .0728.

Table 2, Summary of results for the initial top 12 model by RMSE



Model Type	Preset	PCA	RMSE (Validation)	RMSE (Test) ↑
Linear Regression	Robust Linear	25 numeric components kept	0.072015	0.027674
Neural Network	Optimizable Neural Network	25 numeric components kept	0.07199	0.027681
Linear Regression	Linear	25 numeric components kept	0.071946	0.027741
Neural Network	Narrow Neural Network	25 numeric components kept	0.071843	0.028105
Neural Network	Medium Neural Network	25 numeric components kept	0.040067	0.028227
Linear Regression	Robust Linear	Disabled	0.07188	0.028257
Linear Regression	Robust Linear	Disabled	0.07188	0.028257
Kernel	SVM Kernel	25 numeric components kept	0.071918	0.028376
Neural Network	Narrow Neural Network	Disabled	0.071479	0.028995
Linear Regression	Interactions Linear	25 numeric components kept	0.07175	0.029124
Kernel	Least Squares Regression ...	25 numeric components kept	0.07122	0.029674
Linear Regression	Linear	Disabled	0.071359	0.029862

The Kaggle competition uses a SMAPE error function to quantify the performances of each submission. The baseline model is a reasonable, naive approach since there is minimal movement in micro business density as illustrated in the distribution of standard deviations from Figure 2. The last value prediction serves as a decent method when compared to the top-performing teams participating in the Kaggle competition. We are interested in the future time-step SMAPE score using past data. In other words, we want to minimize $SMAPE_{t+1}$ with a model trained up until time t.

Our initial approach was to perform simplistic autoregressive models such as simple linear regression, regularized cubic spline, and LSTM on appropriately lagged data. These did not perform better than the baseline which indicates our movement stems elsewhere. Before moving to more complex models, we utilized LASSO which is a common feature selection technique to determine the influential features on our micro business density variable. The choice to use LASSO over other feature selection techniques is due to computational costs. We attempted to use Gradient Boosted Decision Trees, Support vector machine regression, and random forest. A cross-validation grid search was used to fine-tune the corresponding hyperparameters. From Table 2, we see that the Robust Linear regression model under a PCA transformation of 25 principal components performed the best under unseen data.

D. More Model Results

We experimented with some more complex models including random forest, gradient boosted decision tree, and support vector regressor. Below shows the results of forecasting a single month ahead.

Table 3, Summary of results under more complex models by SMAPE

Model	SMAPE_t+1
Support Vector Regressor	1.3638
Gradient Boosted Decision Tree	1.3677
Random Forest	1.3790
Naïve (Last Value)	1.3792

Despite the presence of a hidden test set in Kaggle, the best SMAPE score achieved by any team so far is 1.2366. Out of the 3500 participating teams, the 100th team, which represents the top 3% of all teams, achieved a score of 1.3717, which is comparable to our results. The 1000th team scored 1.3791, which barely exceeded the baseline *[2]*. This suggests that 70% of the models were unable to outperform the naive model that relies on last month's micro-business density value. However, it should be noted that the actual performance of our model may differ from the Kaggle results due to the presence of the hidden test set.

E. Project Pivot

Due to the difficulties of evaluating and predicting variability, and the uncertain ties with signal and noise influence, our team decided to pivot this project in a different direction. A natural question for this project is to predict the mean of micro business densities by county instead of the movement under a time component. Predicting the mean can provide a more stable and reliable insight into the economic and demographic characteristics of a region, as it is less affected by short-term fluctuations and noise.

IV. Part 2 (Expectation of MBD)

A. Motivation

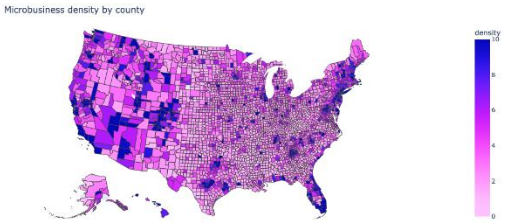
The decision to shift the project's direction towards predicting the mean of micro business densities by county instead of movement under a time component primarily stems from the interest of interpretability. This study provides useful insights into the features that can drive growth and development of microbusinesses. Policymakers, government officials, urban planners, economic developers, entrepreneurs, and small business owners can all benefit from understanding the underlying factors that drive microbusiness growth and development.

We've broken up this portion of our project into three parts. Firstly, we want to identify important factors contributing to MBD. We then structure a regression problem to predict a county's expected MBD. Finally, we investigate the top 10% of outliers posed as a classification problem.

B. EDA

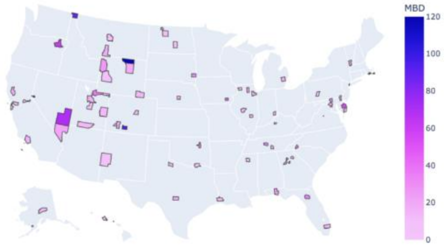
In the EDA section, a geographical map of micro business density by county is shown in *Figure 6* to provide a visual representation of the distribution of micro businesses across different regions *[4]*. This map can help identify areas with high concentrations of micro businesses, which may indicate favorable economic conditions and potential areas for growth and investment. The map can also reveal areas with lower densities of micro businesses, which may indicate areas that require additional resources and support for economic development.

Figure 6, Geographical map of MBD mean (2022)



Before we classify outliers, in *Figure 7* the top 10% of MBD counties are visible.

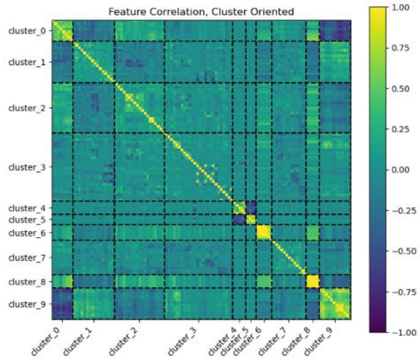
Figure 7, top 10% of MBD outliers



C. Decorrelation of Feature Space

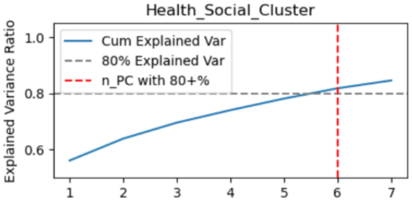
As part of preprocessing our data, we look into fixing feature correlation. High feature correlation can be detrimental to many models such as OLS. This is a necessary step we must take care of before we begin modeling. We can see the initial correlation heat map in *Figure 5.* We use k-means clustering to group our highly correlated features into their own cluster separated by dashed lines in *Figure 8.* We see that the light yellow and deep purple get grouped together as expected.

Figure 8, Clustered correlation heat map



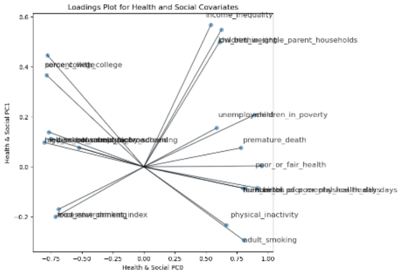
Each cluster manually receives an interpretable name based on the features contained in them. We perform PCA on each cluster as an attempt for feature reduction. The criteria for the number of principal components requires 80% of the variance to be explained. If only less than 3 components are needed then we handpick features, otherwise keep the number of components to transform. We will use a cluster of features titled Health_Social_Cluster to illustrate our process of decorrelation. In *Figure 9*, we show the cumulative explained variance as a function of the number of principal components. We see our target of 80% is reached using 6 components.

Figure 9, PCA explained variance



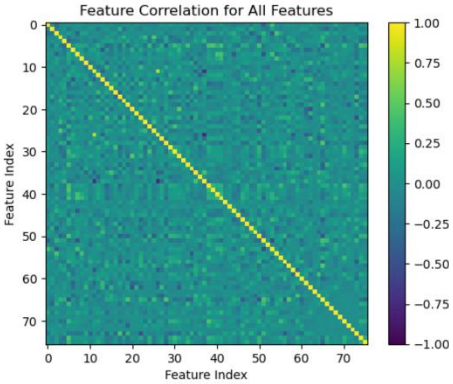
For more context, we utilize a biplot in *Figure 10,* which is a type of graphical representation that allows for simultaneous visualization of the principal components and the features of the cluster Health_Social_Cluster. The principal components are represented on the axis of the graph, while the features are represented as lines extending from the origin. The length and direction of the lines indicate the strength and direction of the relationship between the features and the principal components

Figure 10, Biplot of relationship of PC0 and PC1 with health and social features



The group related blocks with PCA performed on each cluster results in a very nice-looking correlation feature heat map that showcases the effectiveness of using this method of decorrelation.

Figure 12, Correlation heat map after decorrelation method



D. Models and Tuning

For model selection and hyperparameter tuning, we utilize grid search with cross-validation on the training set, and the test set performances are then evaluated.

Both forward and bi-directional feature selection are performed with OLS and KNN, as they can suffer from the influence of useless covariates. To balance classes for classification, upsampling with SMOTE is performed. SMOTE, which stands for Synthetic Minority Oversampling Technique, is used to overcome the class imbalance issue by creating synthetic samples of the minority class.

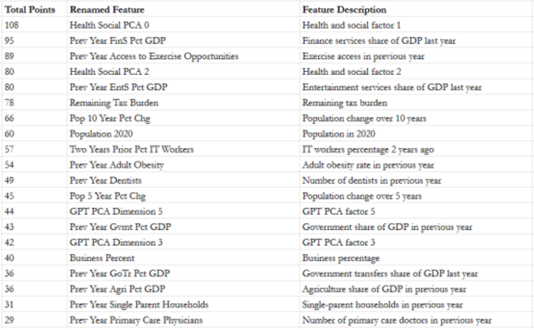
For predicting a county's micro business density (MBD) using regression, the models that are considered for this task include Linear Regression (OLS), Lasso, KNN, SVM, MLP, Decision Tree, Random Forest, and LightGBM.

For predicting if MBD > 10%, a classification task is performed. The models that are considered for this task include Logistic Regression, SVM, Random Forest, and LightGBM. By considering multiple models and tuning their hyperparameters, we can identify the most effective models for predicting MBD and classifying outlier counties.

E. Feature importance

In order for us to discern influential features, we introduce a ranking system of feature importance across the following models: OLS, KNN, decision tree, random forest, Bayesian ridge, XGBoost, LightGBM, CatBoost, and bootstrapped Lasso. We rank the features by statistical significance by assigning points to the top 20 performing features. This provides us with an aggregate sum of points for us to identify statistically significant features across different models.

Figure 11, Feature ranking by importance



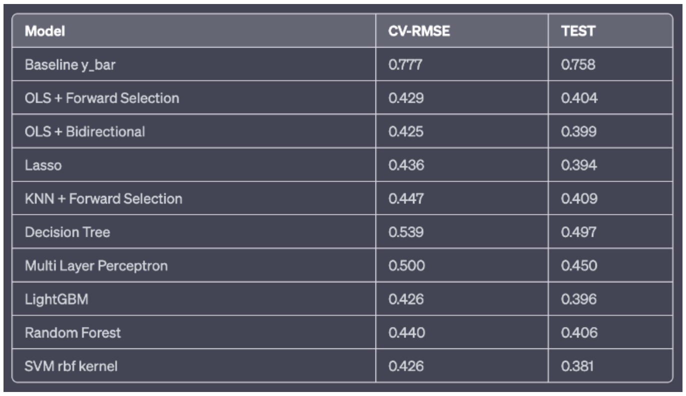
Total Points	Renamed Feature	Feature Description
108	Health Social PCA 0	Health and social factor 1
95	Prev Year FinS Pct GDP	Finance services share of GDP last year
89	Prev Year Access to Exercise Opportunities	Exercise access in previous year
80	Health Social PCA 2	Health and social factor 2
80	Prev Year EntS Pct GDP	Entertainment services share of GDP last year
78	Remaining Tax Burden	Remaining tax burden
66	Pop 10 Year Pct Chg	Population change over 10 years
60	Population 2020	Population in 2020
57	Two Years Prior Pct IT Workers	IT workers percentage 2 years ago
54	Prev Year Adult Obesity	Adult obesity rate in previous year
49	Prev Year Dentists	Number of dentists in previous year
45	Pop 5 Year Pct Chg	Population change over 5 years
44	GPT PCA Dimension 5	GPT PCA factor 5
43	Prev Year Gvmt Pct GDP	Government share of GDP in previous year
42	GPT PCA Dimension 3	GPT PCA factor 3
40	Business Percent	Business percentage
36	Prev Year GoTr Pct GDP	Government transfers share of GDP last year
36	Prev Year Agri Pct GDP	Agriculture share of GDP in previous year
31	Prev Year Single Parent Households	Single-parent households in previous year
29	Prev Year Primary Care Physicians	Number of primary care doctors in previous year

The most important features are sorted by total points achieved from each of the models. The description column gives a more detailed explanation for each feature.

F. Regression

After collecting and processing the necessary data, we conducted a regression analysis to predict the mean micro-business density in each county. Our aim was to develop a precise model that could accurately predict the average micro-business density using our predictors derived from decorrelation and feature selection. In this analysis, we used the RMSE metric to assess the accuracy of our model's predictions with the results of a naive baseline model which makes a prediction on the sample mean of y for comparison. With appropriate feature selection OLS performs especially well, even on par with complex models like kernel SVM and Gradient Boosted Trees. Apart from considering RMSE, we can also evaluate the MAE score, which is often more intuitive. With the OLS model, we were able to attain a mean absolute error of 0.284 (baseline: 1.167) for the bottom 95% of counties where MBD is less than 10.

Figure 13, Regression results



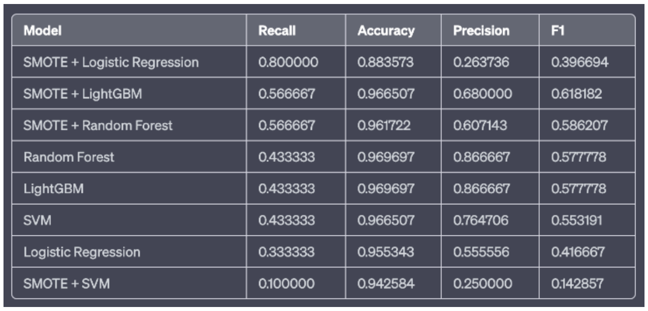
Model	CV-RMSE	TEST
Baseline y_bar	0.777	0.758
OLS + Forward Selection	0.429	0.404
OLS + Bidirectional	0.425	0.399
Lasso	0.436	0.394
KNN + Forward Selection	0.447	0.409
Decision Tree	0.539	0.497
Multi Layer Perceptron	0.500	0.450
LightGBM	0.426	0.396
Random Forest	0.440	0.406
SVM rbf kernel	0.426	0.381

Looking a little deeper into our OLS results, we obtain a correlation coefficient R^2 of 0.711 which is surprisingly good for economics data. Some of the log transformations were incorporated into our model as significant features which serves as validation. Interestingly, we obtained many ChatGPT terms. Despite the inaccuracies and dependence concerns regarding ChatGPT outputs, they seem to be statistically significant in our model.

G. *Classification*

In order to identify counties with MBD > 10 for targeted advertisements and economic profit, it is crucial to prioritize recall as a metric to avoid overlooking high MBD counties. Despite having the lowest accuracy, the logistic regression model with SMOTE demonstrates the highest recall and is therefore the optimal choice for the task.

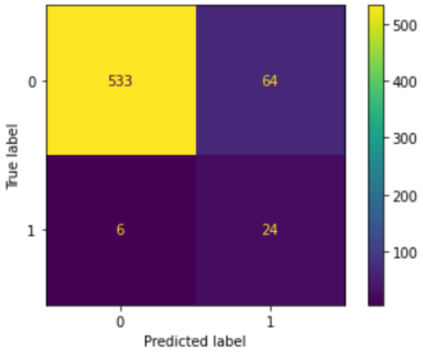
Figure 14, Classification Results



Model	Recall	Accuracy	Precision	F1
SMOTE + Logistic Regression	0.800000	0.883573	0.263736	0.396694
SMOTE + LightGBM	0.566667	0.966507	0.680000	0.618182
SMOTE + Random Forest	0.566667	0.961722	0.607143	0.586207
Random Forest	0.433333	0.969697	0.866667	0.577778
LightGBM	0.433333	0.969697	0.866667	0.577778
SVM	0.433333	0.966507	0.764706	0.553191
Logistic Regression	0.333333	0.955343	0.555556	0.416667
SMOTE + SVM	0.100000	0.942584	0.250000	0.142857

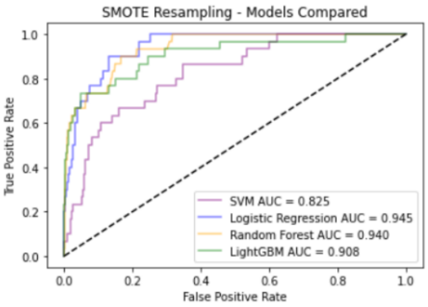
From our results we see that the SMOTE + logistic model performs best. This is especially preferable since it is more easily interpretable. We can view the accuracy of this model in the form of a truth/prediction table.

Figure 15, truth/prediction table for logistic regression with SMOTE



We showcase an ROC chart to measure the model's ability to distinguish between outlier and non-outlier classes. The diagonal dashed line indicates an AUC of 0.5 which is equivalent to randomly guessing. In this case, the logistic regression model has the highest AUC value of .945, indicating that it has the best overall performance in detecting outlier counties. The random forest and lightGBM models also have high AUC values of .94 and .908, respectively, indicating good performance. The SVM model has the lowest AUC value of .825, which suggests that it may not be as effective in distinguishing between outlier and non-outlier classes compared to the other models.

Figure 16, ROC graph of binary classifier models



H. *Model Selection*

We were able to develop linear models that were comparable or superior to complex models like Support Vector Machines (SVMs) and Gradient Boosted Trees for both regression and classification tasks, while keeping specific metrics in mind. This makes model selection relatively straightforward, particularly when the focus is on inference. Even if there is a slight loss in accuracy, it is acceptable if it results in a considerably more interpretable model.

I. *Conclusion*

In conclusion, the important features that were identified in this study include finances and entertainment shares of GDP, health and social factors, obesity rates, overall population and population change, remaining tax burden, obesity, and engineered ChatGPT features. The regression models showed promising results with an R^2 of 0.711 under OLS. However, the complexity of the data resulted in some amount of noise. Classification models were successful, particularly with the help of SMOTE sampling which improved the recall. We obtain a pseudo R^2 of 0.752. Overall, these findings highlight the importance of considering a wide range of factors when analyzing data and building predictive models.

V. Acknowledgement

I would like to extend my thanks and appreciation to my teammates, Nicholas Amirsoleimani, Zhengpu Zhao, and Spencer Slaton, for their unwavering support and collaboration throughout this project. They have contributed immensely in various aspects, including problem formulation, data collection, exploratory data analysis, preliminary modeling, and model checking.

References

[1] GoDaddy, *Micro Business Density Forecasting*. URL https://www.kaggle.com/competitions/godaddy-microbusiness-density-forecasting
[2] Kaggle Competitors' Performances https://www.kaggle.com/competitions/godaddy-microbusiness-density-forecasting/leaderboard
[3] *GoDaddy Microbusiness Density Survey*. URL: https://www.godaddy.com/ventureforward/microbusiness-datahub/
[4] Google Collab, *Preliminary Modeling,* URL https://colab.research.google.com/drive/128YAffffZP81xzy8CmEkbYS_ItlfM_oA?usp=sharing

Appendix

Github Repository:
https://github.com/zhengpu-berkeley/Stat_222_Proj

Final Presentation Slides:
https://docs.google.com/presentation/d/1QvixeRWfGMFNOeGSfpwcsJQqGbWuhqX3yhojFQebD0Q/edit#slide=id.g242ac6119d6_0_532