Group 21 Final Report

1 Introduction

American policy leaders aim to foster more inclusive and resilient economies in light of technology advancements that have made entrepreneurship increasingly accessible. Many Americans are choosing to start their own businesses for various reasons, but these "microbusinesses" (businesses with fewer than 10 employees, typically based online) often elude traditional economic data sources, posing a challenge for policymakers. Understanding where these microbusinesses are centrally located can help spur policy creation to incentivize growth, allocate resources such as capital and business development services, and focus recovery efforts in times of economic downturn. GoDaddy has collected microbusiness survey data over the past 4 years, and hosted a Kaggle Competition from December 2022 to June 2023 asking competitors to find advanced approaches to predict the microbusiness density (Microbusinesses per 100 people) for a given county.

2 Data

For our microbusiness density prediction project, two primary datasets were utilized. The Microbusiness Density (MBD) dataset comprised 128,535 observations across 7 features. This data had information about the raw count of microbusinesses in each county and the target variable, microbusiness density. The Census dataset, consisting of 26 features and 3143 observations, had examples of useful columns from the Census Bureau's American Community Survey (ACS) at data.census.gov. The two datasets were joined using a unique county ID.

The data cleaning phase was predominantly focused on addressing missing values within the Census dataset. Leveraging mean imputation, the 14 missing values in the Census dataset were filled in, ensuring a complete and consistent dataset for subsequent analysis. No such missing values were present in the Microbusiness data. By merging the datasets based on temporal correspondence on a 2 year lag, a unified dataset of 128,535 observations and 12 columns was formed.

To train the model, we got rid of 'row_id','active','cfips' as we think row id won't provide any information, active column is not included in the test set, and cfips provides the same information as state and county. To encode categorical features, we tried one hot encoding but it took too much RAM. As such, we switched to target encoding for state and county, replacing each with the average microbusiness density by the given state/county respectively. The other numerical features remained the same except for month. As month is a cyclic variable, we decided to represent it as a coordinate on a unit circle, with sin(2 * PI * month/12) being the y coordinate and cos(2 * PI * month/12) being the x coordinate.

Lastly, to ensure that we were not using highly correlated features with microbusiness density, we calculated each feature's correlation coefficient with microbusiness density, and no feature has a r-value higher than 0.7 (county), meaning that no features need to be dropped.

3 Exploratory Data Analysis

First, we found geographical features could influence MBD. By states, the three states with the highest mean MBD are Delaware (18.74), District of Columbia (13.53), and Nevada (12.43), whereas the lowest three are Mississippi (1.70), West Virginia (1.86), and Arkansas (2.00). A negative correlation between the number of counties and MBD suggests that as the number of counties increases, microbusiness density tends to decrease.

The further analysis of mean Microbusiness Density (MBD) across U.S. counties reveals significant trends and patterns. The distribution of mean MBD by county, ranging from 0.06 to 83.48, exhibits a skewed right distribution. Combining the Figure(top county) and Figure(by county) we could find notable variation with numerous outliers exceeding the third quartile such as Carson City and Lincoln County, Nevada, standing out with the highest mean MBD, while Issaquena County, Mississippi, showcasing the lowest. Besides, the counties with extreme values in Nevada may explain its high ranking in mean MBD by state. Surprisingly, despite Delaware having the highest state-level MBD, it lacks representation in the top 10 counties, with Sussex County, Delaware, ranking 11th-highest.

Lastly, correlation heatmap analysis indicates moderate positive correlations between variable pairs. Scatterplots reveal heteroscedasticity in the "active" feature, suggesting potential considerations for excluding this feature before training machine learning models. Outliers are present across all features, necessitating careful partitioning of counties by size or population for a nuanced analysis.

4 Modeling

4.1 Methods

Baseline Model: Linear Regression, comparing L1, L2, Elastic-Net
For hyperparameter tuning, we chose to use grid search and cross validation of size 5. Alpha values tried in the grid search are 0.001, 0.01, 0.1, 1, 10, 100. L1 ratios tried in the grid search are 0.2, 0.5, 0.8.

After hyperparameter tuning, the modeling performance for lasso regression is 17.7893, ridge regression is 17.7827, and elastic net is 17.9729. The target variable has a range 284.34003 thus MSE less than 20 is relatively small in this case.

Gradient Boosting/XGBoost:
Extending from linear regression and regularized linear regression, ensemble methods such as Gradient Boosting and XGBoost are more complex and able to capture non-linear variance present in the data.

To enable comparison, we continue using the test/train dataset from our base linear regression model. After hyperparameter tuning, the modeling performance has a clear edge over linear regression at 13.58 for XGBoost and 17.69 for Gradient Boosting.

Given that XGBoost wins over Gradient Boosting heavily, we proceed by tuning XGB hyperparameters. We are able to get MSE down to 12.31.
As for feature importance, the three most important features are county, the percentage of

households in the county with access to broadband, and the percent of the population in the county over age 25 with a 4-year college degree.

Deep Learning:
Time Series models like ARIMA cannot be used on this dataset. This is because there are 1871 different counties and 50 different states, each with their own time series. Instead we will be using TensorFlow to model the data via Deep Learning.

The model we ended up using was a basic Neural Network that takes in 10 features, and using Dense layers to reduce the dimensions down to 64, 32, etc. all the way to 1 layer, making sure to dropout 25% of the data in each layer. Activation functions will be ReLU for all layers except for the last, which will have linear activation as this is a regression problem. The neural network was trained for 20 epochs, producing a Mean Squared Error of 15.5576 on the training data and 19.4430 on the validation set. Feature Importances can be derived via SHAP values (since Keras does not provide them by default)

We also tried creating a Recurrent Neural Network because this is still Time Series data. More specifically, we replace the first Dense layer with a GRU or LSTM layer. However, since both of them require 3D arrays as inputs, we reshape the arrays to have dimensions [a1, 1, a2], where [a1, a2] are the original dimensions of the array. GRU has an MSE of 14.8811 on training, 18.0119 on validation, while LSTM has an MSE of 23.4607 on training, 33.2017 on validation.

Lastly, feature importances were not easily accessible for these models; the best course of action would be to use SHAP values, but the GRU and LSTM models are in 3 dimensions and SHAP takes 2 dimensions at most, while for the original neural net, matrix multiplication fails.

4.2 Results
XGBoost has the best performance of MSE on the test set.

Among L1, L2, and elastic net regressions, Lasso regression has the lowest MSE. Top 3 important features for the lasso regression model are counties with absolute coefficient 2.748755, the percent of the population in the county over age 25 with a 4-year college degree with absolute coefficient 0.824260, and states with absolute coefficient 0.470095. This indicates high correlation between microbusiness and geographical locations and education, which is reasonable intuitively. Month and year information is least important in the regression models. As for the best performing model XGB, feature importance similarly marks an important correlation between microbusiness density and geographical location/education/access to the internet or information in general.

Findings of our model suggest that microbusiness density is strongly influenced by geographic locations and education levels. These factors can affect economic opportunities and entrepreneurial skills, which are crucial for micro business success. Additionally, our XGB model highlights the importance of internet access, enabling a great extent of information, communication, and productivity. These findings offer direct insights into how to develop policy and infrastructure that support microbusiness growth, focusing on education and digital connectivity in specific regions.