CS 412 Spring 2023 - Final Project Report

Paolo Ferrari
paolof2@illinois.edu
Mechanical Science and Engineering, UIUC
Champaign, IL, USA

ABSTRACT

This report aims to give an overview and the main results of the project for CS 412. The goal of the project, originated from a Kaggle challenge, was to predict the density of microbusiness across different counties in the US. To address this problem, we used 3 different methods, all relying on linear regression between a time variable and the density of microbusiness. In the first model, we use the whole time span of the data in the linear regression; this is the model that was analyzed in the Midpoint review of the Project. As shown here, this model often has poor performance because of some discontinuities in the data. Thus, for the second model, we only use the last 12 months of the data in the linear regression. In the third model, we use a variable criterion to select the time span where the data is taken for the regression. Overall, linear regression models were chosen as our methods due to their simplicity and because, generally, the density of microbusiness is a slowly-changing function of time. While the models are simple, they provide very good predictions for the microbusiness density in the test data. The average relative error of the best model was 2.2 %.

1 INTRODUCTION

This project is proposed by the website GoDaddy.com and is outlined in Kaggle [2]. The goal was to use public survey data to predict the density of microbusiness in the US at different counties at a given month of the year. The density of microbusiness is defined by the number of microbusiness divided by every 100 inhabitants of a County.

The data is supplied by Kaggle [2] in the form of csv files. The input to the algorithm is a vector with the relevant information of a county. The output is a prediction (a number) representing the density of microbusiness for that county.

The organization of this report is as follows. First, we describe the both training and testing data followed by an exploratory analysis to get insights on which model could be appropriate and any potential limitations of them. Next, based on this analysis, we state the baseline model: simple linear regression. Then, we describe the two more advanced models adopted in this work and review their main results. Next, we analyze the main assumptions of the third model and tune its main parameter to optimize the performance. Then, we state the main implications of this Project. Finally, we point out possible directions for further improvement.

2 DATA

The basic goal of the challenge is to predict the density of microbusiness across 3135 Counties in the US for months starting in November of 2022. Thus, we will build models that compute the density of microbusiness in a month given a timestamp and a county ID.

The data provided in the Kaggle challenge consists of a csv file containing 6 columns:

(1) The ID number for each county
(2) The real name of each county
(3) The US state of the county
(4) The date where the microbusiness density was taken
(5) The microbusiness density in the county
(6) The raw count of microbusiness in the county

The train data contains monthly values of the microbusiness density and raw count of each county from 8/2019 to 10/2022. Thus, each county contains 39 sequential data points. Similarly, the test data contains the same information begging in 11/2022. However, because the data is pulled from the US Census [1], so far, only the months of November and December of 2022 are available at the time of this writing. Nonetheless, we will use the data from these two months as the test for our model.

3 EXPLORATORY ANALYSIS

Before doing any analysis on the data, we checked whether there was any missing data across the 3135 Counties and 39 months. We found that 2 counties had many zeros recorded for their microbusiness density, namely Issaquena County (MI) and Loving County (TX). Thus, we excluded these 2 counties from our analysis, redcuing the total number to 3133.

A typical the plot of the microbusiness denisty versus time for an arbitrary county is shown in Figure 1 along with its linear regression line.

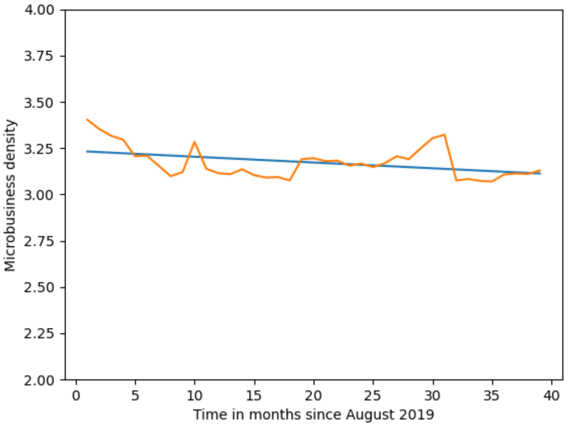

Figure 1: Typical plot of microbusiness density versus time. The orange curve is the data, the blue line is the linear regression fit.

While many of the counties show this type of stable microbusiness density profiles, many of them show discrete jumps as shown by Figure 2. The Figure also shows the linear regression model.

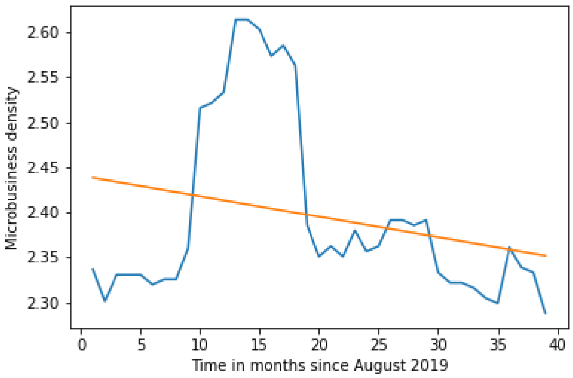

Figure 2: Microbusiness density (blue) and prediction (orange) versus time for the case when a clear distinct behavior is present in the months 10-18.

The reason for these changes in behavior could be many: a business closing or opening its doors on a given time, a change in the registered location of the headquarters of a business, etc. Thus, it is very difficult or even impossible to predict when these sudden changes will occur.

To address this challenge, instead of trying to predict when sudden changes occur, we decide to perform linear regression only on the final months of the data. This is reasonable because, in many instances, the microbusiness density at any given time on a county should only be correlated to the microbusiness density a few months to a year before, not on the density a few years before. This way, the influence of these events that cause sudden changes in the microbusiness density influence is minimized.

4 MODELS

4.1 Model 1

As mentioned before, because of the clear time dependence of the data, we decided to model the microbusiness density as linear function of time. To treat the specificities of each county, each linear regression is run independently for the 3133 counties. The time is counted in units of month, beginning in August of 2019 (month 1) to October of 2022 (month 39).

In summary, we assume that for each county C, the microbusiness density Y_C depends on time X as

$$Y_C(X) = \alpha_C X + \beta_C \tag{1}$$

Here $X = 1, 2, ..., 39$ is the time in months since 8/2019. As explained earlier, the train data correspond to the months 8/2019 to 10/2022; while the test is evaluated on the months 11 and 12/2022.

Fitting is done with ordinary least square estimation for α_C and β_C. That is, for each county, we seek the parameters α_C and β_C that minimize the objective function

$$SS = \sum_{i=1}^{39} (y_i - \alpha_C X_i - \beta_C)^2 \tag{2}$$

In terms of implementation, the linear regression model is first fitted to the train data and then evaluated on the test data. Both train and test data have similar formats, with the only difference that the test data lacks the response variable, that is, the microbusiness density.

4.2 Model 2

Model 2 (M2) has the same format of Model 1 (M1), except that only the last 12 months of the train data are used for regression. As explained previously, this prevents any events happening before an year of the months of November and December of 2022 to have any effect on their prediction. That is, in this case, we seek the parameters α_C and β_C that minimize the objective function

$$SS = \sum_{i=28}^{39} (y_i - \alpha_C X_i - \beta_C)^2 \tag{3}$$

4.3 Model 3

A more rigorous choice for the time span used in the prediction is done in Model 3 (M3). Here, we adopt the following algorithm:

(1) Fit a linear model using only the last N months of data. At the first iteration, $N = 6$.
(2) Compute the residuals of this fit in the N months and their standard deviation σ_N.
(3) Extrapolate the model to a month before the fit and compute the difference between the prediction and the actual value.
(4) If the absolute value of this difference is smaller than $\lambda \sigma_N$, where λ is a tuning parameter, include this month in the model and go back to Step 2. If not, stop here and use the current linear model.

The motivation for M3 is the following. We start assuming that only the months $34, 35, ..., 39$ contain relevant information for predicting the microbusiness density for months 40 and 41 (the months of the test data). Then, we go backwards and start including more months in the regression if the data looks qualitatively the same.

That is, if the microbusiness density for a month m, not included in the fitting, deviates from its prediction by more than λ standard deviations of the fit, the month m should not be included since there could be a qualitative change from month m and backwards. Thus, only the months $m + 1, m + 2, ..., 39$ should be included for fitting.

In this case, similar to models M1 and M2, the regression coefficients α_C and β_C minimize the objective function

$$SS = \sum_{i=m+1}^{39} (y_i - \alpha_C X_i - \beta_C)^2 \tag{4}$$

The value for λ is chosen to be 3 for the purpose of comparing the models. However, later we perform change it from 1 to 10 and check that 3 is a reasonable number.

Table 1: Statistics of the relative error for the train data

Model	Mean	Median	Maximum
M1	0.045	0.026	3.54
M2	0.015	0.009	1.64
M3	0.011	0.006	1.94

Table 2: Statistics of the relative error for the test data

Model	Mean	Median	Maximum
M1	0.067	0.034	14.44
M2	0.027	0.012	3.04
M3	0.027	0.011	5.55

5 CODE IMPLEMENTATION

The code for the model is written in Python using the standard Sckikit-learn, Pandas and Numpy packages. The linear fit is done using Scikit-learn modules.

The code is run in a Jupyter notebook which is attached to the assignment page.

6 RESULTS

The original challenge ranks the individual submission by computing the mean absolute percentage error. Thus, this is the metric we adapt to test the performance of the linear regression model.

We compute the error err_C for county C as

$$err_C = \sum_{i=1}^{N_{months}} \frac{|Y_C(i) - Y_C^{pred}(i)|}{Y_C(i)} \tag{5}$$

where $Y_C(i)$ is the recorded microbusiness density in month i and $Y_C^{pred}(i)$ is its predicted value. This metric is evaluated both in the train and test data.

Tables 1 and 2 show the calculated mean, maximum and median of err_C across all counties for the three models tested.

We see all linear model have reasonably good fitting. The average of the relative error is of M1 4.5% for the train data and 6.7 % for the test data, whereas these values are significantly smaller for M2 and M3. For M2, the errors for train and test data are 1.5% and 2.7 %; for M3 they are 1.1% and 2.7%.

Figure 3 qualitatively shows the enhancement brought by M2 and M3 vs M1. We see that both $M2$ and $M3$ adequately capture the relevant trend near the final months of the time-series, whereas M1 suffers from irrelevant events happening before.

Interestingly, as Table 2 shows, M2 has very similar performance to M3, even though M3 only considers, on average, fewer months for prediction. In fact, Figure 4 is a histogram of the number of months that are included when using M3. We see that in the majority of times, only the final 6 to 12 months of data are used for regression in M3. Although, in some cases, the number of months can become up to 39, in cases where the microbusiness density is regular across the whole time. In addition, the peak near 20 months could indicate a common feature occurring in multiple counties, such as a change in the way the microbusiness density is recorded in a state or region on the US at around that month.

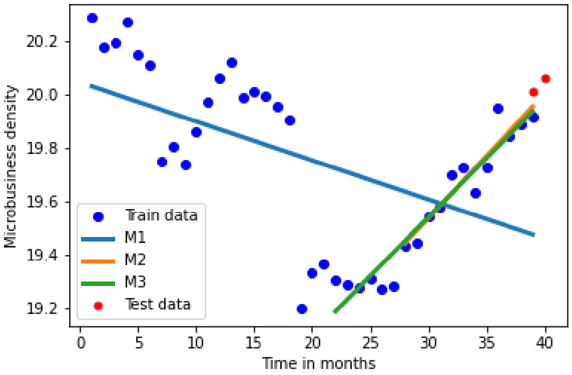

Figure 3: Microbusiness density train data (blue dots) and predictions (solid lines) versus time. The two test data points are shown in red.

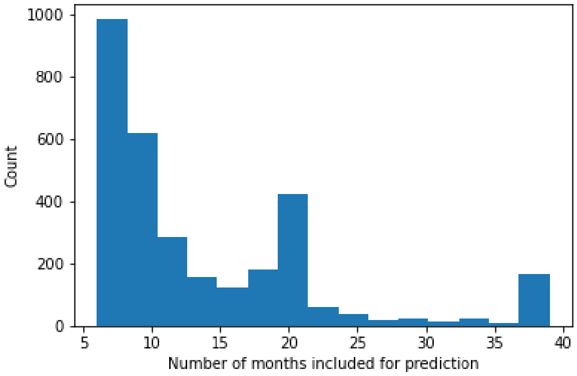

Figure 4: Histogram of number of months used during the regressions of Model 3.

In some cases, such as shown in Figure 5, the predictions for Model 1 and Model 2 are better than M3. This is specially the case when there is a lot of random variation around the data but no consistent change in behavior.

Table 3 shows the relative train and test errors for different values of λ. Surprisingly, both in terms of the train and test data, using a smaller λ seems to be better on average. This indicates that most of the information for predicting the microbusiness density is contained in only 6 months prior to the moment in time.

7 CONCLUSION AND IMPLICATIONS

Using simple linear regression models, we were able to achieve very good prediction performance for the microbusiness density across different Counties in the US. The major feature of the data that allowed us to do that is the slowly changing nature of the

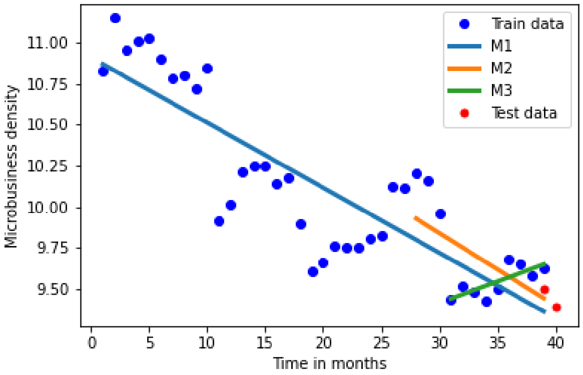

Figure 5: Microbusiness density train data (blue dots) and predictions (solid lines) versus time. The two test data points are shown in red.

Table 3: Statistics of the relative error for the train and test data under different λ

λ	Mean Train Error	Mean Test Error
1	0.0103	0.022
2	0.0102	0.025
3	0.0121	0.027
4	0.0158	0.034
5	0.0197	0.039
6	0.0231	0.043
7	0.0251	0.046
8	0.0266	0.048
9	0.0278	0.049
10	0.0289	0.051

microbusiness denisty with respect to time. In fact, as indicated by our results, using only 6 to 12 months worth of data is enough to give a prediction of about 2% accuracy for the microbusiness density at two months in the future.

Going forward with this project, it would be good to check the validity of the model against a larger number of months in the test data. A major drawback of the linear models, as seen in multiple Counties, discrete events suddenly change the microbusiness density.

In order to account for these events, as well as other shifts in the linear trends of the data, it would be necessary to use other geographical information as explanatory variables. The reason why we decided not to include this is because they are published only once every year by the US Bureau. [1] In addition, the data for the year 2022 is absent because it is published only with to a 2-year lag. Thus, in order to have a relevant amount of information, it would be necessary to look other data banks to extract monthly data for the period 2019-2022.

8 OUTLOOK

To develop further more accurate models, including more explanatory data is necessary. This data would have to be published monthly, have significant month-to-moth variation and contain good indicators for economic activity in the counties. Examples include county data for number of houses, traffic data, number of households with broadband access, number of IT workers, number of people in hosted in hotels.

REFERENCES

[1] United States Census Bureau. 2023. Access microdata. (Feb. 2023). Retrieved Feb 16, 2023 from https://data.census.gov/.

[2] kaggle.com. 2023. Godaddy - microbusiness density forecasting. (Feb. 2023). Retrieved Feb 16, 2023 from https://www.kaggle.com/competitions/godaddy-microbusiness-density-forecasting/overview.