Microbusiness Density Forecasting

Project Report

Team 5 - Pramodh Gowda Poonadahally Shivadas, Smriti Bajaj, Sumukh Vasisht Shankar

INTRODUCTION

Any company that makes less than $10 million annually and typically has less than 10 employees is considered a microbusiness. 92% of all American businesses are microbusinesses and they employ more than 61 million people. The number of these micro-businesses per unit area (can be anything from a square kilometer to a zip code/county) is referred to as microbusiness density. It is crucial to forecast this density for several reasons, including but not limited to urban planning, targeted policymaking, company investment decisions, and market research.

Figure 1: (Left) Type of businesses in the USA and (Right) Employee count distribution in microbusiness.

DATA

Keeping our goal in mind, we have chosen the primary dataset from Kaggle which covers historic economic data at a county level. This contains information about county's name, microbusiness density, and active businesses as well as metadata from US census API about the population's citizenship and immigration rates, percentage of people with college degrees, access to broadband, etc. We combine both the datasets for analysis.

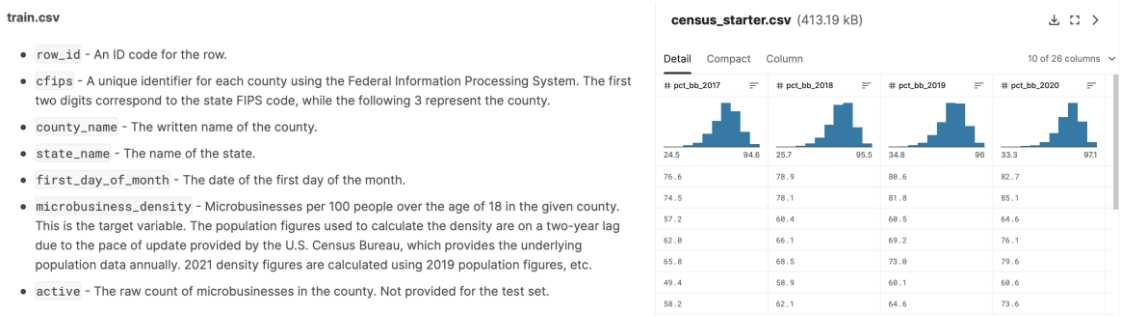


Figure 2: (Left) Columns in microbusiness density dataset and (Right) Columns in US census dataset.

LOW-RISK PROBLEM AND SOLUTION

The problem focused on understanding the relation between the different features of the combined datasets. We found that out of so many features, we can reduce the dimensionality using PCA and T-SNE. PCA is a linear technique for dimensionality reduction that caused a lot of data loss while t-SNE being a nonlinear

technique is particularly useful for visualizing high-dimensional data in a low-dimensional space, performed better than PCA while preserving the data.

```
[ ] pca = PCA()
    pca.fit(df)
    cumsum = np.cumsum(pca.explained_variance_ratio_)
    np.argmax(cumsum>=0.95) + 1

    8

[ ] pca = PCA()
    pca.fit(df)
    cumsum = np.cumsum(pca.explained_variance_ratio_)
    np.argmax(cumsum>=0.789) + 1

    3
```

To preserve 95% of the data, we need 8 dimensions but that would be too much to visualize. However, with 3 dimensions, we can preserve 78.9% of the data which looks like a good deal for analysis.

Figure 3: Data Loss in PCA

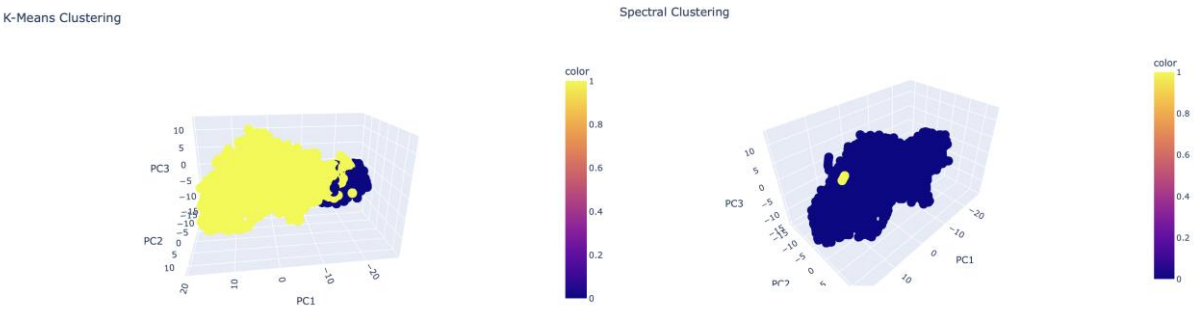


Figure 4: (Left) K-Means clustering and (Right) Spectral clustering

Finally, we clustered our data and identified correlations between the features that were not visible in the raw dataset using Spectral clustering. It is a graph-based clustering algorithm that can handle non-linearly separable data and is less sensitive to the shape of the clusters. The reason is because K-means clustering is a centroid-based algorithm, and it assumes that the clusters are spherical and have equal variance. We also experimented with DBSCAN which is a density-based algorithm that groups together data points that are close to each other in high-density regions, but it struggles with clusters of different densities and with datasets that have a varying density.

MEDIUM-RISK PROBLEM AND SOLUTION

As a part of the medium-risk problem, we predicted the microbusiness densities by modeling 2 time-series forecasting methods – *Autoregressive Integrated Moving Average (ARIMA)* and Prophet. ARIMA models aim to capture the patterns and correlations in a time series, and make predictions based on those patterns, whereas Prophet models time series data using an additive model that includes three main components: trend, seasonality, and holidays. ARIMA forecasted microbusiness densities holistically and Prophet was used to forecast microbusiness density for 50 counties in the US.

The performance of both the models were measured by various metrics such as SMAPE, MAPE, MSE and MAE. *SMAPE (Symmetric Mean Absolute Percentage Error)* measures the percentage difference between the actual and predicted values of a time series, and it is defined as the average of the absolute percentage errors (APE) of each point in the series. *MAPE (Mean Absolute Percentage Error)* measures the average absolute percentage difference between the actual values and the predicted values. *MSE (Mean Squared Error)* measures the average of the squared differences between the predicted & actual values. *MAE (Mean Absolute Error)* measures the average of the absolute differences between the predicted values and actual values.

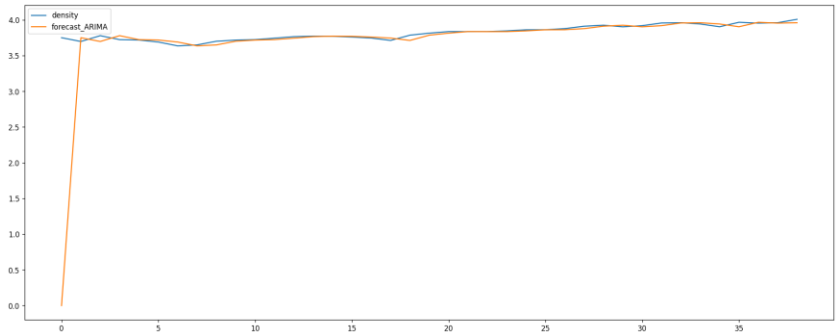


Figure 5: Forecast from ARIMA model on the test set

Model	SMAPE	MAPE	MSE	MAE
Prophet	1.02	0.01	0.001	0.03

Table 1: Performance metrics for Prophet on County 01001 (Autauga County, Alabama)

HIGH-RISK PROBLEM AND SOLUTION

We experimented by converting the time-series data into time-sensitive regression data with the help of three models- *Long Short-term Memory (LSTM)*, LightGBM(LGBM) and Linear regression to achieve better performance than the medium risk models. LSTM is a powerful deep learning model that can capture complex temporal patterns in the data. It can model sequential data by incorporating memory cells and gates, while linear regression is a statistical modeling technique used to find the linear relationship between a dependent variable (microbusiness density) and one or more independent variables(date). However, training LSTM models is computationally expensive and time-consuming. In our experiments, we found that LSTM gave similar results to linear regression, and linear regression was much faster to train.

Model	SMAPE
LSTM	1.1
Linear regression	1.09
LGBM	***1.06***

Table 2: Performance metrics for LSTM, Linear Regression and LGBM on all counties

Another efficient model is *LGBM (Light Gradient Boosting Machine)* which is a highly optimized performance-oriented framework developed by Microsoft. It is an ensemble learning method that uses decision trees as the base learner and employs a gradient boosting technique to iteratively train and refine the models.

Figure 6: Forecasts by LGBM for county 01001 (Left) and county 01003 (Right)

As a part of the high-risk solution, we also developed a data dashboard to showcase the data analysis and forecasting. The dashboard is important for our project case for several reasons including, but not limited to, visualizations (to communicate insights to stakeholders who may not be familiar with the underlying data), monitoring, efficiency (by streamlining data analysis and reducing the need for manual data manipulation), and collaboration (shared with other team members, which can facilitate collaboration and improve communication). Our dashboard showcases the exploratory data analysis, clustering and forecasting across 3 different pages, along with filtering capabilities. The dashboard has been developed

with a Python-Flask server for the backend, and HTML, CSS, and JavaScript along with frameworks such as Bootstrap and Bulma for the front end. Refer to the figure below for the complete architecture of our dashboard. (The dashboard has not been deployed yet on the internet.)

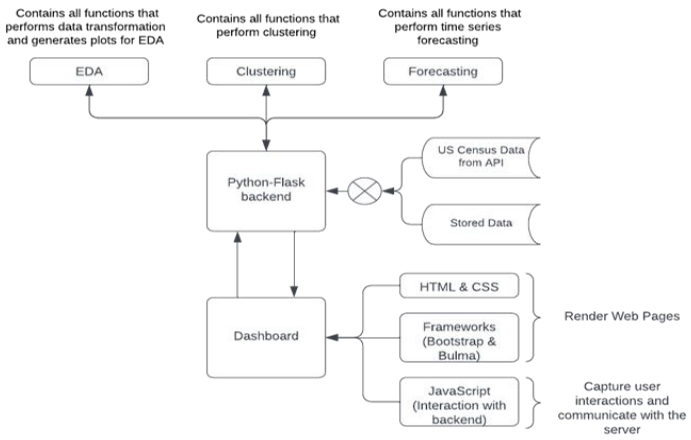
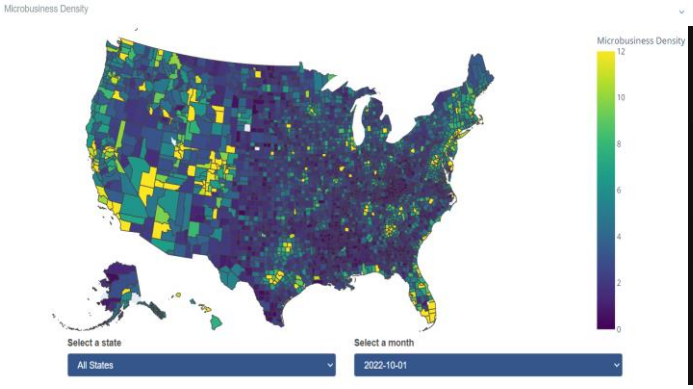


Figure 7: Architecture of dashboard

CONCLUSION

Our techniques of forecasting microbusiness density can help to improve estimates of the density for these businesses in different regions of the USA. By determining the ideal location, Unique Selling Point (USP), and other economic criteria required to benefit these microbusinesses, our tool can boost the revenue and their market position. These small enterprises can increase the number of direct, indirect, and induced jobs by remaining in business and so raising the country's GDP. Governments can develop targeted policies that support entrepreneurship and economic development. By analyzing the concentration of microbusinesses, investors can identify untapped markets and potential acquisition targets.



row_id	cfips	first_day_of_month	state	county	microbusiness_density
1001_2022-11-01	1001	2022-11-01	Alabama	Autauga County	3.470966
1003_2022-11-01	1003	2022-11-01	Alabama	Baldwin County	8.375717
1005_2022-11-01	1005	2022-11-01	Alabama	Barbour County	1.236247
1007_2022-11-01	1007	2022-11-01	Alabama	Bibb County	1.291599
1009_2022-11-01	1009	2022-11-01	Alabama	Blount County	1.835726

Figure 8: (Left) Visualization from the dashboard - Microbusiness densities at a county level across the USA as of October 2022 (Right) Tabulation of these microbusiness densities

REFERENCES

Dataset- https://www.kaggle.com/competitions/godaddy-microbusiness-density-forecasting/data
Codebase- https://github.com/pramodhsway/DS-5550-Capstone-Project-Phase-1