Forecasting Business Growth in San Diego Growth with Machine Learning

Maximilian Wei
mawei@ucsd.edu

Mariana Montoya
m2montoya@ucsd.edu

Michael Lue
mlue@ucsd.edu

Naomi Young
naomi.young@sandag.org

Abstract

Creating an effective and accurate business growth forecast is extremely valuable for local government planning organizations such as San Diego's SANDAG because it can provide insights and forecasts that can help local planners and policymakers make better decisions for the county and its economy. In this paper, we investigate how we can forecast business growth in San Diego. To answer this question, we first researched and determined the factors that influence business growth and the metrics for measuring and forecasting business growth in San Diego. Based on that research, we then decided what key factors and metrics would be best for our business growth forecast model. We found that job counts, population, average household income, percentage of the population working, total number of households, migration within the country, retirement, payroll, and number of establishments were significant factors in determining business growth. Using the Census Bureau Data, we tested three models: Ordinary Least Squares (OLS), Random Forest Regression, and ARIMA. We decided upon not directly forecasting business growth yet, and forecasting the number of employees in every zip code within San Diego County.

Website: https://abc.github.io/
Code: https://github.com/inno-apfel/DSC180A-Q1-Project

1 Introduction

1.1 Context

In a vast region like San Diego, California's second-most populous county, staying well-informed and taking a proactive approach to regional planning is crucial. This is where the importance of business growth forecasting becomes evident. Business forecasting holds immense relevance in areas such as strategic planning, risk management, market-entry, and expansion, among others. To create a region that not only benefits its residents but also positively reflects on SANDAG, we must employ proficient business growth forecasting. While this topic involves complexity, we should consider various factors when dealing with such a large region. To achieve the most accurate business growth forecasts, we must take into account a multitude of factors, ensuring a comprehensive understanding. For example, it's essential to understand the granularity at which we approach the problem, to determine whether we should approach the region by parcel clusters, unique cities, postal codes, or some other delineation. While we may not have all the answers yet, our research journey begins with a comprehensive understanding of both the business landscape and the unique characteristics of the San Diego region.

1.2 Literature Review

Recent work in "Empirical Analysis of Business Growth Factors Using Swedish Data" (Davidsson et al. 2002), compares past economic theory with real-world Swedish data, to quantify the impact of firm age, business size, industry, location, and legal form on predicting business growth. To this extent, they trained a simple OLS regression model on business-level data (each observation being a unique business), and used the converged weight coefficients to quantify a variable's impact. Their experiments concluded that the business age had the most significant impact on predicting business growth, with industry, legal form, and business size, coming behind in that order. Their results seem to contradict certain economic theories, for example: Gibrat's law, which theorizes that both small and large firms will on average have the same growth rate (Mansfield 1962). Although Davidsson et al.'s analysis is derived from business-level data, which we have restricted access to without government sanctions, it may still prove useful in determining which aggregate variables have the highest potential in predicting regional business growth.

In reading the paper "Profit as a Main Factor in Forecasting Agricultural Business Development" (Bal-Prylypko et al. 2023), we were able to gain insight into business development. This article specifically looked at the significance of profit in forecasting the development of agricultural production. The article highlights profit as the key factor in economic forecasting and extensively explores the DuPont model to understand the factors influencing return on equity. Employing this model enables you to identify the sources of profitability, whether it results from improved profit margins, more efficient asset utilization, increased financial leverage, or is experiencing a decline. Although the article primarily revolves

around profit and agricultural production, it allows us to gain insights into various aspects of a business to achieve a comprehensive understanding.

Among the notable prior projects is the 'Series 14: 2050 Regional Growth Forecast' (SANDAG 2017-2018). Completed in 2019, this projection by SANDAG entailed a comprehensive examination of anticipated growth in population, housing, employment, and income. This detailed report provides valuable background information, outlining the methodology employed and the rationale behind the selection of the 16 employment clusters within the region. The methodology delves into the specifics of each section, such as the datasets utilized for the following: Longitudinal Employer-Household Dynamics (LEHD) Origin-Destination Employment Statistics (LODES) 7.3 (2015), SANDAG Employment Estimates (2016), SANDAG Population and Housing Estimates (2016), and the SANDAG Activity-Based Transportation Model (2016).

2 Methods

2.1 Data

The data we primarily used was the County Business Patterns (CBP) datasets obtained from the U.S. Census Bureau. The County Business Patterns is a set of annually updated datasets that provides economic data on establishments and employees at various subnational aggregation levels. For our purposes, we chose to focus on zip-code level data as it was the smallest level of aggregation available. Specifically, we chose to utilize the ZIP Code Totals and ZIP Code Industry Details datasets, choosing to only focus on zip codes within the San Diego region. The ZIP Code Totals (ZBP Totals) dataset contained employee counts and payroll information indexed by ZIP Code, while the ZIP Code Industry Details (ZBP Details) dataset contained information on the number of establishments by number of employees indexed by a ZIP Code and industry pairing. For simplicity's sake, we also chose to use only CBP data after 2012, despite available data going back as far as 1984. As a comparison to extant business growth forecasts, we also utilized SANDAG Series 14 forecasts (SANDAG 2017-2018), specifically their forecasts on Jobs by ZIP Code.

Table 1: ZBP Details Dataset

Name	Type	Definition
zip	str	ZIP Code
year	int	year
naics	int	2-digit NAICS code for industry
est	int	total number of establishments
ni_j	int	number of establishments with between i and j employees

Table 2: ZBP Totals Dataset

Name	Type	Definition
zip	str	ZIP Code
year	int	year
emp	int	total number of employed workers
q1	int	total first quarter payroll ($1,000)
ap	int	total annual payroll ($1,000)

2.2 Data Preparation

Since the data was stored in a neat columnar fashion and contained essentially no missingness, there was not much work done for preparation aside from merging the datasets for each year into two master tables, one for ZBP Totals and Details. As our main predictive task was to predict the number of employees for each ZIP Code, the ZBP Totals dataset, indexed by ZIP Code was used as the main dataset, with information from ZBP Details, being used only to build ZIP Code level features for modeling use.

2.3 Feature Transformation

In order to predict employment counts in each ZIP Code, we first needed our data to be observations on the ZIP Code level. To do this, we took the ZBP Totals dataset as our master table and merged in ZIP Code level features we transformed from the ZBP Details dataset. The ZBP Details dataset included key information about the distribution of industries and establishment sizes within ZIP Codes. We hypothesized that this information would play a significant role in predicting employment growth in certain ZIP Codes, as certain industries or business types may see slower/faster growth. For example, San Diego's booming biotech industry may see more growth than the mining sector. Or that small companies may see more growth than large and old corporations that may be busy focusing more on company politics rather than innovation.

To make sense of this information, we created features encoding the proportion of establishments of a certain industry or establishment size within each ZIP Code. For every naics industry code in our data, we created a feature called naics_x_pct, and for every establishment size bin in our data, we created a feature called ni_j_pct.

$$\text{naics\_x\_pct} = \frac{\text{\# establishments of industry x in the ZIP Code}}{\text{total \# of establishments in the ZIP Code}}$$

$$\text{ni\_j\_pct} = \frac{\text{\# establishments with between i and j employees in the ZIP Code}}{\text{total \# of establishments in the ZIP Code}}$$

After merging the ZBP Totals data with our transformed features, we ended up with a dataset containing the following columns, indexed by (ZIP Code, year) pairs.

Table 3: Master Dataset with Features

Name	Type	Definition
zip	str	ZIP Code
year	int	year
emp	int	total number of employed workers
naics_x_pct	float	proportion of establishments in the ZIP Code that are in industry x
ni_j_pct	float	proportion of establishments in the ZIP Code that have between i and j employees

2.4 Models

Our primary objective was to predict employment counts in each zip code. To do this we chose to employ three distinct models: a simple multivariate linear regression, an autoregressive integrated moving average (ARIMA) model, and random forest regressor. Unless otherwise specified, all our models were trained on the above master dataset, containing ZBP Totals observations and our transformed features.

From what we have seen, simple linear regression (OLS) and general equilibrium models are popular solutions to population and establishment distribution problems within econometrics literature (Carlino and Mills 1987). Due to a lack of econometrics expertise, we have chosen to forgo implementing a general equilibrium model for this problem, instead choosing to implement a simple linear regression model as a baseline to replicate current work.

Since simple linear regression models cannot inherently address time-series data without major adjustments to the structure of input data. We were incentivized to implement another model popular with time-series analysis, particularly stock trend predictions, an autoregressive integrated moving average (ARIMA) model. The idea here is that, after accounting for stationarity, the ARIMA model is able to leverage a number of past timestamp observations when making predictions for the future. This would in theory allow the model to adapt to changes in employment trends throughout the years, whereas a simple linear regressor would simply assume either monotonic growth or decay in employment over the years. A major flaw in the ARIMA model for our problem is that it can only take in univariate observations at each timestamp. Because of this, we would be unable to leverage the

majority of our data, aside from employment counts. Understanding that, without at least including ZIP Code information, the model would end up predicting the same employment count for every ZIP Code, we chose to instead train one ARIMA model for each ZIP Code.

Finally, as the previous two models we have mentioned are both from the data modeling culture, which makes numerous assumptions about the data, we chose to explore a random forest model as an allude to the problem-solving capabilities of the algorithmic modeling space, which assumes little to nothing about the data, treating the data generation process as a black box (Breiman 2001). The random forest model is a subset of ensemble models and is based on the idea of a decision tree, which iteratively learns and applies a set of rules to the data to make a prediction, and makes predictions by averaging the predictions of multiple decision trees trained on random subsets of the data.

3 Results

3.1 Model Evaluation

For simplicity, we evaluated our models using Root Mean Squared Error (RMSE). We evaluated our models for two different scenarios: immediate next-year forecasting, and sustained forecasting. We did this to investigate the performance differences for different modeling approaches, specifically the hypothesis that an OLS model would perform better on immediate next-year forecasting, while an ARIMA model might perform better for sustained forecasting due to its ability to learn more from trends over time. To evaluate our models' immediate next-year forecasting capabilities, we trained them on data from 2012 to 2020 and evaluated them on observations for 2021. For sustained forecasting, we trained our models on data from 2012 to 2018 and evaluated them on observations from 2019 to 2021. For a comparison to currently to current forecasting techniques used by government planning organizations, we also compared our models against SANDAG's Series 14 forecasts on Jobs by ZIP Code (SANDAG 2017-2018).

Table 4: Model RMSEs

Scenario	Immediate Forecast	Sustained Forecast
ARIMA	1681.94	2147.66
Random Forest	3843.99	2550.29
OLS	2008.98	2117.05
SANDAG Series 14	7835.81	8046.42

3.2 Discussion

Our evaluations indicate that, for the prediction of employment, ARIMA models perform similarly to linear regressors in immediate next-year forecasting, but outperform them when it comes to sustained forecasting. The ability of our ARIMA models to perform similarly to linear regressors in terms of next-year forecasting despite lacking most of the data indicates that a combination of the two to both efficiently utilize time-series trends and multivariate data could significantly improve the accuracy of our forecasts. We also see that, perhaps due to the structure of our data, random forest models are outperformed by both ARIMA and linear regression models. However, at least in terms of RMSE, all three model types outperform SANDAG's current forecasting. Though, to keep in mind SANDAG's Series 14 Forecasts were built exclusively on data from 2016.

4 Conclusion

While our current models demonstrated poor performance, we are optimistic about refining our approach in the next quarter. Our plan involves incorporating additional machine-learning techniques and introducing more variables to enhance the accuracy of our forecasts. Specifically, we intend to include overarching variables like GDP, aiming to provide a more comprehensive understanding of the factors influencing business growth. In our initial attempt, we focused solely on predicting job growth. However, in the upcoming quarter, we plan to extend our model to predict business establishment growth. This expansion is a strategic step toward improving the overall efficacy of our forecasting model. The assessment of our models provided valuable insights into their performance across different aspects, highlighting areas where the data proved effective and where it fell short. This understanding is crucial for informing our approach to future projects. Looking ahead, we are considering testing an LSTM (Long Short-Term Memory) model and exploring feature transformations. This includes incorporating population metrics (such as population number and the percentage of the population working), household metrics, and broader economic metrics like the country's GDP. Additionally, we may introduce a metric describing the distance of each zip code from downtown to other sections, such as Oceanside. Recognizing the need for refinement, we acknowledge that adjustments can be made to the variables we initially selected, such as industry percentage and establishment size. We intend to narrow the focus to variables specifically relevant to the San Diego region, ensuring our models are tailored to the unique characteristics of the local economy.

References

Bal-Prylypko, L, O Cherednichenko, L Stepasyuk, and Z Titenko. 2023. "Profit as a main factor in forecasting agricultural business development." *IOP Conference Series: Earth and Environmental Science* 1150 (1), p. 012008. [Link]

Breiman, Leo. 2001. "Statistical Modeling: The Two Cultures (with comments and a rejoinder by the author)." *Statistical Science* 16 (3): 199 – 231. [Link]

Carlino, Gerald A., and Edwin S. Mills. 1987. "THE DETERMINANTS OF COUNTY GROWTH*." *Journal of Regional Science* 27 (1): 39–54. [Link]

Davidsson, Per, Bruce Kirchhoff, Abdulnasser Hatemi–J, and Helena Gustavsson. 2002. "Empirical Analysis of Business Growth Factors Using Swedish Data." *Journal of Small Business Management* 40 (4): 332–349. [Link]

Mansfield, Edwin. 1962. "Entry, Gibrat's Law, Innovation, and the Growth of Firms." *The American Economic Review* 52 (5): 1023–1051. [Link]

SANDAG. 2017-2018. "Series 14 Regional Growth Forecast Documentation and Baseline Subregional Allocation." [Link]