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

Current research in business growth forecasting explores only "statistical modeling" techniques, with very few forays into "algorithmic modeling". However, recent work indicates that algorithmic modeling techniques may perform better with certain metrics, at the cost of interpretability, presenting the opportunity for "algorithmic modeling" forecasts to outperform current state-of-the-art techniques. Given this, we investigate how we 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 decided what key factors and metrics would be best for our business growth forecast model. Through our research, we identified key metrics including the total number of establishments, employed workers, industry distribution within zip codes, median household income, population demographics, and payroll data. Using the Census Bureau Data, we tested four models: ARIMA Model, Linear Regression Model, Long Short Term Model (LSTM), and Random Forest Regression Model. <insert brief about conclusion>

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

1 Introduction

Estimating business birth and death counts 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. Actual values are calculated by the U.S. Census Bureau using national surveys and IRS tax form information, and are only available with substantial lag, (currently, in 02/2024, only CBP records up to 2021 are available), presenting the need for effective and accurate business growth forecasts. Due to the statistical history of econometrics, previous works tend to approach the problem with a "Statistical Modeling" approach, with very few attempts to address "Algorithmic Modeling"(Breiman 2001). However, recent work indicates that algorithmic modeling techniques may perform better with certain metrics, at the cost of interpretability(Grieves, Mance and Witt 2023). This presents the opportunity for algorithmic modeling-based forecasts to serve as a higher accuracy auxiliary support to statistical modeling forecasts. Given this gap in research, we employed various machine learning models, including random forest regressors and LSTM RNNs against industry standard Ordinary Least Squares and ARIMA models to evaluate the potential of "Algorithmic Modeling" in business forecasting.

2 Methodology

2.1 Data

We approached the data sourcing issue from the perspective of the general public, without any special government clearances. To this end, we sourced and combined an array of publicly available datasets on business births/deaths and socio-economic demographics, mostly originating from the U.S. Census Bureau.

The data we primarily used were the County Business Patterns (CBP) and American Community Survey (ACS) 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 sub-national aggregation levels. The American Community Survey contains various demographic estimate datasets by year and also at various sub-national 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, ZIP Code Industry Details, ZIP Code Demographic and Housing Estimates, ZIP Code Selected Social Characteristics in the United States, and ZIP Code Income in the Past 12 Months 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. The ZIP Code Demographic and Housing Estimates dataset contained population estimates for various demographic subsets such as sex, age, and race

indexed by ZIP code. The ZIP Code Selected Social Characteristics in the United States dataset contained total household estimates and other household demographic subset estimates indexed by ZIP code. The ZIP Code Income in the Past 12 Months dataset contains various income estimates including median and mean income indexed by zip code. For simplicity's sake, we chose to use only data from 2012 to 2021. 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.

2.2 Data Preparation

We first began by pre-processing the data into data frames and reformatting the tables so that we could work zip codes as a column rather than a header within the data. We then concatenated all of the individual yearly sub-datasets into one dataset with all the years from 2012-2021 for each of the five data sources mentioned above. Afterward, we cleaned and filtered out only the features that were relevant to our task of predicting business establishment growth from the data tables we gathered. We decided to drop zip codes with incomplete observations as afterward we were still left with a majority of San Diego zip codes with complete observations across all datasets. After creating some additional features explained in the following section, we merged all of the datasets for each year into one master table. As our main predictive task was to predict the number of establishments for each zip code annually, we merged all the datasets by zip code and year.

To predict establishment 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 and may be busy focusing 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.

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

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

Additionally, using the Housing and Demographics dataset, we created a total_retirement population estimate feature by summing the estimates for populations greater or equal to age 65. We also chose to observe the mid-career working populations and split those features into the age groups 25-34 and 35-44 to observe any differences between them.

$$\text{total\_retirement} = \#\text{ population greater than 65 years of age}$$

$$\text{total\_midcareer\_25\_34} = \#\text{ population between 25 and 34 years of age}$$

$$\text{total\_midcareer\_35\_44} = \#\text{ population between 35 and 44 years of age}$$

After all feature transformations were finalized, we merged all of the remaining datasets with the previous combined ZBP Totals and Details master dataset by zip code and year pairs and lagged all variables excluding zip, year, and est.

Table 1: Master Dataset with Features

Name	Type	Definition
zip_xxxxx	bool	One-hot-encoding for ZIP code xxxxx
year	int	Year
est	int	Total number of establishments
qp1	float	Total First Quarter Payroll ($1,000) with Noise
ap	float	Total annual payroll ($1,000) with noise
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
median_hh_income	float	Median household income
total_population	int	Total population
total_retirement	int	Total population greater than 65 years of age
total_midcareer_25_34	int	Total population between 25 and 34 years of age
total_midcareer_35_44	int	Total population between 35 and 44 years of age

2.3 Models

Our primary objectie was to predict establishment counts for each ZIP-Code and year across our data. We simulated establishment count predictions for 2012-2021 with 4 different model architectures.

We considered two categories of approaches for predicting establishment counts:

1. ”Statistical Modeling”
    - Ordinary Least Squares
    - Autoregressive Integrated Moving Average (ARIMA)
2. ”Algorithmic Modeling”
    - Random Forest Regression
    - Long Short-Term Memory Recurrent Neural Net (LSTM)

All models made use of the same master dataset, with all fevatures included, and were trained to minimize root mean square error (RMSE).

2.3.1 Linear regression / OLS

Linear models, including ordinary least squares, are simple statistical models commonly used in the econometrics domain for their interpretability (Reynolds, Miller and Maki 1995). Ordinary least squares models, specifically multi-variate regression models, are one of the simplest modeling techniques that effectively exploit multiple explanatory variables for a single prediction. For this purpose, we chose to implement a multi-variate regression model as a baseline for the"Statistical Modeling" category.

We trained multiple linear regression models using different sets of features. Our choice of features were selected in two ways. The first was using features from our random forest model, and the second was creating a correlation matrix based on our ZBP lagged dataset in which we had correlation coefficient values between each feature and the target variable (business establishment growth) and choosing the top 5 and 10. Selecting those top features as predictors in our linear regression models. We split the data into training and testing sets to assess the models' performance. We choose mean squared error (MSE) as a measurement of accuracy.

Additionally, we explored a fixed effect model and a random effect model. The fixed effect model was used to account for individual-specific effects. Dummy variables were created for each ZIP code, and the model was fitted using linear regression. The results from the fixed effects model showed statistically significant features including employee count (emp), average payroll (ap), and the various employee size categories (ni_j_pct).

The random effect model looked for unobserved heterogeneity at the individual level. Just as the fixed effects model it was fitted with a linear regression. The random effect model saw significance in average payroll (ap) and employee size categories. There was a high adjusted R-squared value indicating a high explanatory power when compared to the random effect model. Based on the evaluation and statistical significance of predictors, we concluded that the fixed effect model provided the best fit for our data, capturing both observed and unobserved heterogeneity at the ZIP code level.

2.3.2 Autoregressive Integrated Moving Average (ARIMA)

The autoregressive integrated moving average (ARIMA) is a classical model used for statistical time-series analysis. The ARIMA model makes predictions purely off trends in the independent variable and is unable to utilize multiple explanatory variables. However, the trade-off is that, unlike OLS models, it's able to effectively exploit the temporal structure of time-series data. This makes it well-suited for cases where time-series forecasts must be made without clean, structured explanatory variables, making it, currently, the most widely used model architecture for business forecasting in the industry (Grieves, Mance and Witt 2023). As such, we implemented a basic ARIMA model for the "Statistical Modeling" category to serve as a stand-in for the current state of the industry. Because of the inability of ARIMA models to take in explanatory variables, we were unable to feed it any information regarding individual ZIP codes. As such, we chose to train one ARIMA model for each ZIP-code, to avoid a singular ARIMA model predicting the same establishment count for every

ZIP-code.

2.3.3 Random Forest Regression

Random forest models are a popular "catch-all" type of model and a common example of "Algorithmic Modeling". 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. Because it assumes little to nothing about the data, it often performs well on all sorts of datasets, making it a good candidate for our problem.

2.3.4 Long Short Term Memory Recurrent Neural Network (LSTM)

The long short-term memory network is a recurrent neural network introduced by Hochreiter and Schmidhuber in 1997 (Hochreiter and Schmidhuber 1997). It is an adjustment of the basic recurrent neural network which allows the model to retain "memory" of previous time-steps over extended time intervals. The model makes predictions at each time step, with information from the previous time-step, and a vector of relevant information from all previous time steps. As a variant of recurrent neural networks, the LSTM model is critically capable of taking in an arbitrary number of explanatory variables at each time step. The ability of the LSTM model architecture to exploit both multi-variate data and temporal information, combines the benefits of both OLS and ARIMA models, making it a good candidate to surpass both, creating a convincing argument for "Algorithmic Modeling".

The LSTM network's ability to easily make autoregressive predictions also makes it possible to forecast establishment counts in the distant future, without the need to first acquire data on explanatory variables for those years.

3 Results

3.1 Model Evaluation

We evaluated our model performances by computing their prediction root mean squared errors (RMSE) in two different scenarios: short-term forecasting and long-term forecasting. To evaluate our models' short-term forecasting capabilities, we trained them on data from 2012 to 2020 and evaluated them against observed data for 2021. For long-term forecasting, we trained our models on data from 2012 to 2018 and evaluated them against observed data from 2019 to 2021. For a comparison 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 2: Model RMSEs

Scenario	Immediate Forecast	Sustained Forecast
ARIMA	X	X
OLS	X	X
Random Forest	X	X
LSTM	X	X
SANDAG Series 14	X	X

3.2 Discussion

<insert_discussion>

4 Conclusion

<insert_conclusion>

References

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

Grieves, Chris, Steve Mance, and Collin Witt. 2023. "Predicting the Effect of Business Births and Deaths on the Current Employment Statistics Survey: Using Sample Information to Minimize Coverage Error." *Office of Survey Methods Research*. [Link]

Hochreiter, Sepp, and Jürgen Schmidhuber. 1997. "Long Short-term Memory." *Neural computation* 9: 1735–80. [Link]

Reynolds, Paul D., Brenda Miller, and Wilbur R. Maki. 1995. "Explaining Regional Variation in Business Births and Deaths: U.S. 1976-88." *Small Business Economics* 7(5): 389–407. [Link]

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

Appendices

A.1 Additional Details

Quarter 2 Proposal [Link]