Forecasting Regional Aggregate Establishment Birth-Death Values: Using Algorithmic Modeling

Maximilian Wei
mawei@ucsd.edu

Mariana Montoya
m2montoya@ucsd.edu

Michael Lue
mlue@ucsd.edu

Mentor: Naomi Young
naomi.young@sandag.org

Abstract

Understanding the dynamics of establishment births and deaths is crucial for comprehending economic trends and informing policymaking. However, actual establishment data is often only available with considerable lag, necessitating effective forecasting methodologies. This paper investigates the potential of algorithmic modeling against traditional statistical data modeling in forecasting annual establishment counts at the ZIP Code level, focusing on the San Diego region. To augment our algorithmic modeling approaches we incorporate a diverse set of socio-demographic and economic variables as explanatory features, sourced from U.S. Census datasets. Despite limitations in data availability, our results reveal insights into the comparative performance of statistical and algorithmic models. While statistical models demonstrate superior forecasting accuracy in the absence of recessionary data, algorithmic models exhibit potential for capturing dynamic fluctuations, suggesting a potential to outperform statistical models, particularly during recessionary periods when traditional assumptions about establishment growth falter.

Website:
https://dsc-capstone-b10-3.github.io/algorithmic-business-forcasting
Code:
https://github.com/inno-apfel/regional-business-growth-forecasting

1 Introduction

Establishment birth and death data are highly significant for understanding the job market and business cycles. Birth data provide a measure of entrepreneurial activities and gauge new entries and reallocation of resources towards growing areas. Similarly, business death data measure failing enterprises and identify sectors from which resources are being shifted away. Accurate data on establishment birth and death values are valuable for local government planning organizations, providing significant insight into the state of national/regional entrepreneurship, and informing local planning and policy-making. As a local transportation planning organization, The San Diego Association of Governments (SANDAG), whom we collaborated with in partnership on this project, utilizes such forecasts to identify areas of future growth, correlated with increased transportation demand, to inform future transportation development.

However, 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, presenting a need for effective and accurate business growth forecasts. As of the time of writing this, (March 2024) establishment birth and death records are only available up to 2021.

To properly understand our goals for this project, one must first understand the differences between "Statistical Modeling" and "Algorithmic Modeling" approaches. As introduced in "Statistical Modeling: The Two Cultures"(Breiman 2003), we define "Statistical Modeling" as a modeling approach where some stochastic data model is chosen as the data generating process, and predictions are made by estimating model parameters with empirical data and feeding inputs (x) through that data model. On the other hand, we can define "Algorithmic Modeling" as a modeling approach where the data-generating process is considered an unknown black box, and predictions are made by finding a function that transforms inputs (x) into response variables (y) with the highest predictive accuracy.

Existing forecasting methodologies focus on "Statistical Modeling" techniques, which make various rigid assumptions about the data, mainly characterized by consistent seasonal patterns, which have broken down in times of extreme change, such as the recent COVID-19 recession. Recent work (Grieves, Mance and Witt 2023) hinted at the potential of "Algorithmic Models", which avoid the pitfalls of making rigid assumptions on the distribution of data in favor of iteratively discovering a function that best represents the data-generating process. Specifically, they concluded on the potential of recurrent neural networks to outperform traditional "Statistical Models" in forecasting establishment births and deaths.

Given this gap in research, we trained and evaluated various "Statistical" and "Algorithmic" modeling architectures on U.S. Census Bureau data, to assess the potential of "Algorithmic Modeling" in producing accurate birth-death forecasts. To increase applicability to SANDAG's current work, we focused on forecasting establishment births and deaths for ZIP Codes within San Diego County.

2 Methodology

2.1 Data Limitations

In a perfect world, to make forecasts on establishment birth-death values on a regional level, we would have records of establishment birth-death values by month for an expansive range of years, broken down by some sort of geographical area, such as census blocks. This would present a more realistic representation of historical trends, by accounting for multiple recession years where previous techniques have failed, while also accounting for changes caused by seasonality. However, for various security purposes, data at such a granularity, especially geographical, is generally restricted for use by government organizations, and not available to the general public. Without the necessary data clearances, we are left with limited data available to the general public.

2.2 Data

Given the aforementioned data limitations, we sourced data on the count of establishments broken down by year and ZIP Code Tabulation Area (ZCTA). While birth-death data is available by month through the U.S. Bureau of Labor Statistics's Current Employment Statistics program and by county or state from the U.S. Census' Business Dynamics Statistics program, we sought to look for smaller geographic areas, to stay within the San Diego County. To this end, ZCTAs were the smallest level of geographical area for which data we could get our hands on. However, since birth-death data were not available by ZCTA we substituted it by using establishment counts, which is roughly equivalent to the previous year's establishment count plus total births minus total deaths.

This data on establishment counts by ZTCA can be found in the U.S. Census Bureau's ZIP Code Business Patterns Totals datasets, collected as part of the County Business Patterns (CBP) program. As ZBP data was only available up to 2021 at the time of our project, we used data from 2012 to 2021. While ZBP records back until 1984 exist, a lack of auxiliary data, explained below, restricted our data to 2012 and after.

Given that currently deployed birth-death forecasts(Grieves, Mance and Witt 2023), built upon the auto-regressive integrated moving average (ARIMA) architecture, can only learn from trends and patterns within birth-death values, we sought to source auxiliary data to act as explanatory variables for our "Algorithmic Models", speculating that more nuanced patterns could be learned from them, improving forecast accuracy in recession years.

In exploring potential explanatory variables, we referenced prior analyses on regional variation in business births and deaths (Reynolds, Miller and Maki 1995). Based on their work we collected and processed data representing the processes they found to have "major" or "strong" explanatory power. We sourced this data from the following socio-demographic and economic datasets from the U.S. Census:

- **ZIP Code Business Patterns Totals (ZBP Totals)**
  - Total employment counts
    - *hypothesis*: more total employees → more establishments
- **ZIP Code Business Patterns Details (ZBP Details)**
  - Distribution of establishments by industry
    - *hypothesis*: certain "volatile" industries may be prone to larger changes in establishment births/deaths
  - Distribution of establishments by establishment size
    - *hypothesis*: small businesses experience higher churn rates → more births/deaths
- **American Community Survey DP02 (ACS DP02)**
  - Total household counts
    - *hypothesis*: more households → more employees → more establishments
- **American Community Survey DP05 (ACS DP05)**
  - Population age demographics
    - *hypothesis*: more citizens of midcareer age → more entrepreneurship → more establishment births
- **American Community Survey S1901 (ACS S1901)**
  - Distribution of median household incomes
    - *hypothesis*: higher median household income → more spending power → increase in niche businesses

2.3 Data Preparation

Our collected data was in the form of multiple datasets, with an individual CSV file for each year and dataset pair. To simplify model development and training, we would prefer a single master CSV file dataset to work off. To achieve this, we cleaned and processed all the individual datasets, concatenating and merging them when necessary to create a single dataset indexed by (ZCTA, year) pairs containing all our collected variables. In cleaning the data, we filled in missing values for establishment counts, representing data that was either not available or not comparable, with zeros. While merging the datasets, inner merges were used to ensure data completeness, leading to the loss of certain ZIP Codes with incomplete data in any of our datasets. For purposes of analysis, these lost ZIP Codes accounted for less than 5% of our total dataset.

To properly represent the processes we are replicating for certain explanatory variables we made a couple of transformations to our collected data to create our final explanatory variables.

Industry distribution data available within the ZBP details dataset was represented as establishment counts by 5-digit North American Industry Classification System (NAICS) industry codes. To reduce the total number of features we would use to represent industry, we aggregated this data up to the more generalized 2-digit NAICS codes. We then created a feature called naics_x_pct for every 2-digit NAICS code we had.

$$\text{naics\_x\_pct} = \frac{\text{\# establishments of industry x in the ZIP Code}}{\text{total \# of establishments in the ZIP Code}} \tag{1}$$

Establishment size distribution data available within the ZBP details dataset was similarly represented as establishment counts by establishment size bins, where ni_j represents the number of establishments with between i and j employees. Similarly, we created a feature called ni_j_pct for every establishment size bin.

$$\text{ni\_j\_pct} = \frac{\text{\# establishments with between i and j employees in the ZIP Code}}{\text{total \# of establishments in the ZIP Code}} \tag{2}$$

Age demographic information in the ACS DP05 dataset was represented as population counts by age bins. For our purposes, we arbitrarily chose to select the age bins: 25 to 34, 35 to 44, and 65 and over, to represent the mid-career and retirement age populations, defining mid-career as two separate age bins for further granularity.

$$\text{total\_midcareer\_25\_34} = \text{\# population between 25 and 34 years of age} \tag{3}$$

$$\text{total\_midcareer\_35\_44} = \text{\# population between 35 and 44 years of age} \tag{4}$$

$$\text{total\_retirement} = \text{\# population greater than 65 years of age} \tag{5}$$

After making the above transformations, we are left with our final master dataset, which is used as input for all models in the following sections. (see A.1 for further details)

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
emp	int	Total number of employees
qp1	float	Total First Quarter Payroll ($1,000)
ap	float	Total annual payroll ($1,000)
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

2.4 Data Analysis

In this section, we explore our collected data to gain an understanding of how the data is distributed, temporally, and geographically.

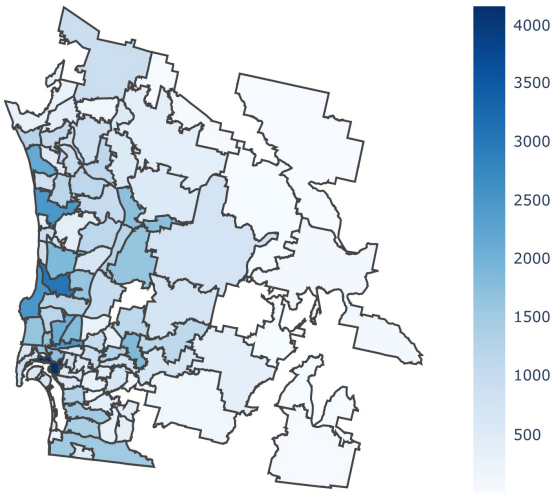

Figure 1: Establishment Counts by ZIP Code

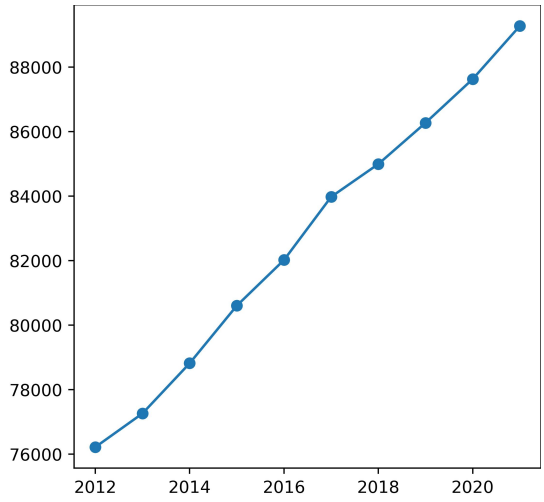

Figure 2: Establishment Counts by Year

Figure 1 shows, that most ZIP Codes have around 1,500 or fewer establishments, with a few outlier ZIP Codes with significantly more establishments such as Downtown San Diego (92101) with 4,000k establishments. This tells us that a select group of ZIP Codes contains the majority of businesses in the region, and, unless accounted for, will have a disproportionate impact on our model training and evaluation.

With our task inherently being one of time series analysis, we are chiefly interested in how our features are distributed over time, as it will influence which models perform best, especially when it comes to "Statistical Models" which assume much about the distribution of data.

Looking at Figure 2, we see the impact of our data limitations once again. In aggregate, establishment counts across the county exhibit trends of monotonic growth, that is, they increase year after year.

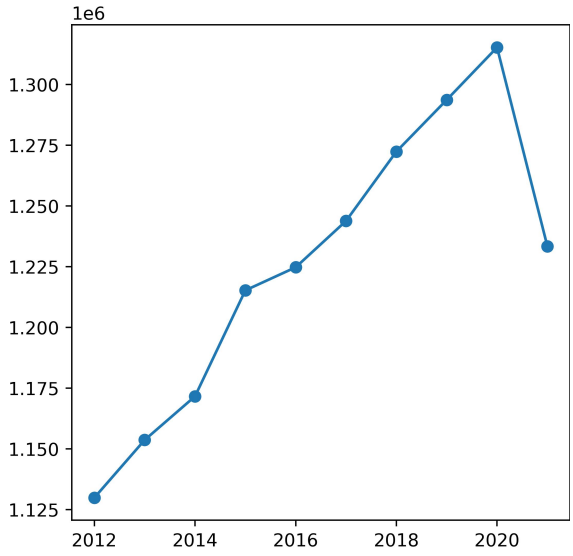

Figure 3: Total Employees by ZIP Code

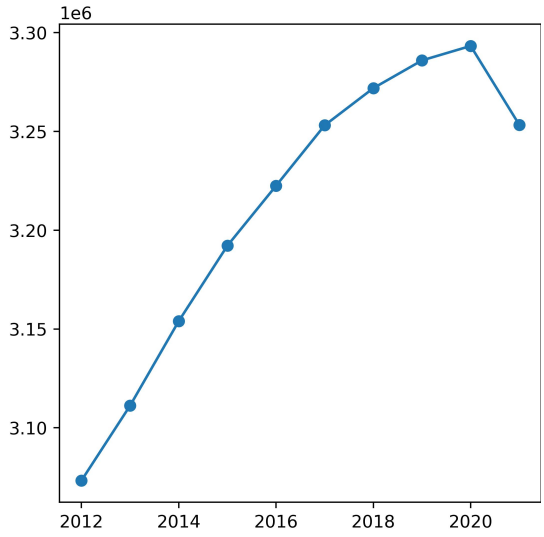

Figure 4: Total Population by Year

In Figure 3 and 4 we sharp dips in employment and population between 2020 and 2021, signaling the start of the COVID-19 pandemic. Though it is interesting to note that this dip was not reflected in establishment counts. This could likely be attributed to efforts including the California Microbusiness COVID-19 Relief grant program, which funneled money back into the economy in attempts to keep small businesses afloat.

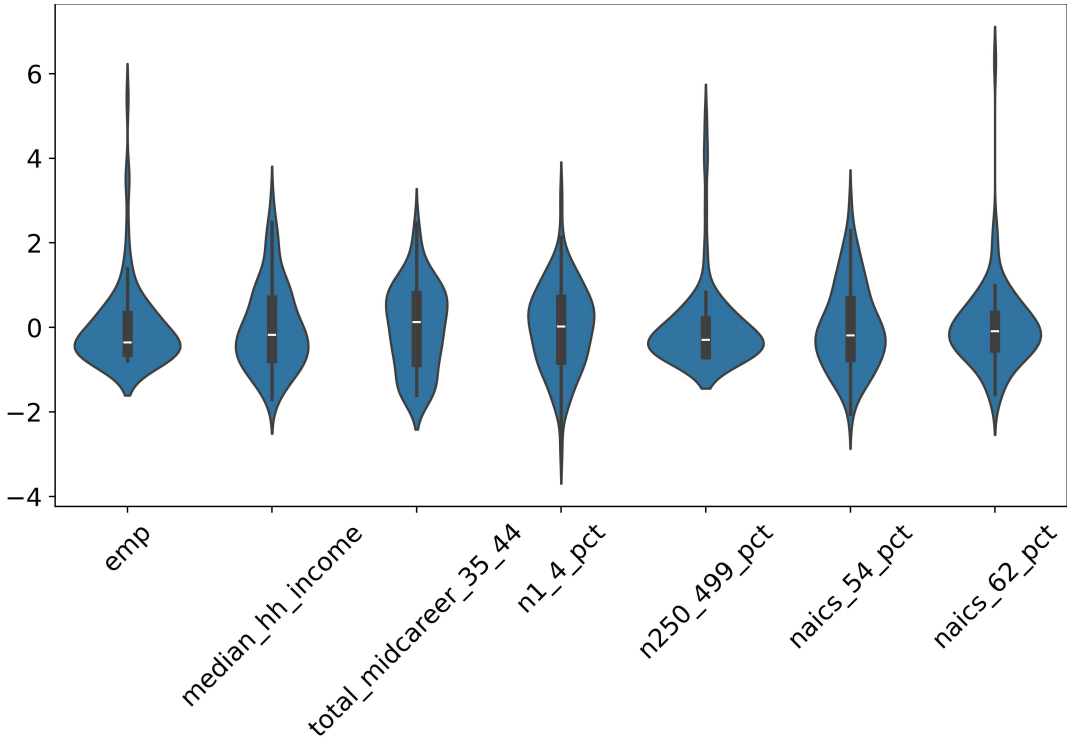

Figure 5: Value Distributions of Choice Features (normalized)

The skewed distributions of values for certain features as we see in Figure 5 indicate that certain features tend to prefer certain types of ZIP Codes over others. We see that smaller businesses (n1_4_pct) are mostly normally distributed across ZIP Codes, while big businesses (n250_499_pct) see a couple large outliers, telling us that some specific ZIP Codes either develop or attract most large businesses in the region, while smaller business thrive everywhere. We see similar patterns of right-skewed distributions with industry distributions, indicating preferences for some businesses to establish themselves in ZIP Codes with other businesses in the same industry. We see that the health sector (naics_62) seems to be particularly affected by this "grouping behavior" while the management services sector (naics_54) cares less. A reasonably being that the management services sector, mostly focusing on desk work, may be less dependent on physical contact between partnered establishments, than say the health sector, which may value geographical closeness for transfer of samples and patients.

2.5 Models

With our goal being to assess the potential of "Algorithmic Modeling", we split our modeling architectures into two categories: "Statistical" and "Algorithmic". For each category, we evaluated one simpler "baseline" model and a "theoretically best" model.

- Statistical Models
  - Ordinary Least Squares Regression (OLS)
    - Least Absolute Shrinkage and Selection Operator Regression (LASSO)
  - Autoregressive Integrated Moving Average (ARIMA)
- Algorithmic Models
  - Random Forest Regression
  - Long Short-Term Memory Recurrent Neural Net (LSTM)

2.5.1 Ordinary Least Squares (OLS and LASSO)

Ordinary least squares, or in other words, multiple linear regression, is a simple statistical model commonly used for interpretability. An important point regarding the data modeling approach of ordinary least squares models is that they can only predict a monotonic increase or decrease after the training data, making them less dynamic than other models. This rigidity leads to recession years in training to significantly drag down lines of best fit, while recession years in testing may be completely over-shot, increasing testing error. However, given that the data we are working with contain no recession years, it may perform better than it should.

With ordinary least squares being the only model we considered that doesn't have some kind of feature selection strategy built in, we explored two different paths for feature selection, in an attempt to avoid the curse of dimensionality and "trash in trash out". We first trained the model using L1 regularization, selecting regularization weight using n-fold rolling time series cross-validation, turning the model into a least absolute shrinkage and selection operator (LASSO) regression model. L1 regularization essentially iteratively reduces the weights of uninformative features to zero, turning them off. We then chose to explore a few sets of choice-selected features by fitting ordinary least squares models using each of the following feature sets and selected the best-performing feature set using n-fold rolling time series cross-validation:

- Top 30 features by Pearson correlation with establishment counts
- Top 30 features selected by forward feature selection
  - Iteratively adds features to a set by incrementally selecting the feature that reduces RMSE the most.
- Top 30 features by lowest mean decrease in impurity (MDI) calculated from our random forest model
  - MDI counts the times a feature is used to split a node, weighted by the number of samples it splits

Our cross-validation on the above feature sets concluded with selecting the "Top 30 features

selected by forward feature selection"; ('est_lag_1', 'total_midcareer_35_44', 'total_population', 'n500_999_pct', 'n250_499_pct', 'naics_56_pct', 'naics_54_pct', 'naics_48_pct', 'naics_11_pct', 'emp', 'zip_92109.0', "zip_92106.0', 'zip_92105.0', 'zip_92101.0'...), as inputs for our ordinary least squares model. This confirms a couple of our previous speculations, indicating that the number of mid-career citizens plays a role in new establishment growth, that certain ZIP Codes, including those with more large businesses, dominate our prediction results, and that establishment growth rates differ by industry.

2.5.2 Auto-regressive Integrated Moving Average (ARIMA)

The auto-regressive integrated moving average (ARIMA) is a classical model used for statistical time-series analysis which makes predictions by averaging a select number of previous observations, after accounting for stationarity, and cannot utilize multiple explanatory variables. This makes it well-suited for cases where time-series forecasts must be made without clean, structured explanatory variables, making it, currently, one of the most widely used forecasting model architectures in the industry. For that reason, we implemented a basic ARIMA model for the "Statistical Modeling" category to serve as a stand-in for the current state of the industry. However, it is to be mentioned that the lack of seasonality in our data may negatively affect its forecasting accuracy.

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.

2.5.3 Random Forest Regression

Random forest models (Breiman 2001) are a popular "catch-all" type of model and a key representative of "Algorithmic Modeling"(Breiman 2003). Because random forest models assume little to nothing about the data, and iteratively split the data to make predictions, it often performs well on all sorts of datasets, making it a potential candidate for our problem. However, random forest models do require observations to be independent and identically distributed, going against the time-series nature of our data, making forecast accuracy almost solely dependent on the explanatory power of our input features.

As a trade-off, however, random forest models, and decision trees in general, are especially prone to overfitting. To address this problem, we employed careful hyper-parameter selection through n-fold rolling time series cross-validation on a diverse set of parameter ranges to optimize model generalizability. Chiefly within the parameters considered was maximum tree depth, which serves as a proxy for feature selection. The lower the maximum tree depth is set, the fewer splits there are in the trained trees, on average reducing both overfitting and the number of features considered.

2.5.4 Long Short Term Memory Recurrent Neural Network (LSTM)

Neural networks are the second key representative of "Algorithmic Modeling" mentioned by Breiman in "Statistical Modeling: The Two Cultures(Breiman 2003)", of which recurrent neural networks are specifically designed for time series analysis. The long short-term memory network(Hochreiter and Schmidhuber 1997), is an adjustment of the basic recurrent neural network, specifically on how hidden states and calculated and used, which allows the model to retain "memory" of previous time steps over extended time intervals.

This capability of retaining information on data from hundreds or thousands of timestamps in the past makes LSTMs a good candidate for forecasting recession years, which are often significantly spaced apart and independent of seasonality. As a variant of recurrent neural networks, it is also capable of taking in an arbitrary number of inputs at each time step, allowing it to take advantage of the variance present in explanatory features, combining the benefits of both OLS and ARIMA models, making it the optimal candidate for our problem.

Additionally, it is possible to modify the traditional LSTM architecture to introduce an autoregressive feedback loop(Graves 2014), allowing the model to make long-term forecasts without the accompanying input features. However, it is important to note that this is done by predicting all input and output features at each timestamp, increasing prediction error, which compounds exponentially the further forecasts are made from the training data.

3 Results

To simplify training and maintain consistency with previous works, we trained and evaluated our model performances using root mean squared error (RMSE).

For forecasting evaluation, we considered two different use case scenarios: immediate next-year forecasting and long-term forecasting. To evaluate our models' immediate next-year forecasting capabilities, we trained them on data from 2012 to 2020 and evaluated them against observed data for 2021. For long-term forecasting, we would preferably evaluate forecasting performance tens of years after our training data, as SANDAG does with their Series 14 Forecasts(SANDAG 2018), forecasting out to 2050 using a base year of 2016. However, with a limited range of years available to us, we chose to train our models on data from 2012 to 2018 and evaluate them against observed data from 2019 to 2021.

Table 2: Model Evaluations (RMSE)

Model	**1-Year Out RMSE**	**3-Year Out RMSE**
OLS	22.97	19.70
LASSO	22.83	20.04
ARIMA	21.86	33.47
Random Forest	38.80	48.99
LSTM	30.38	32.51

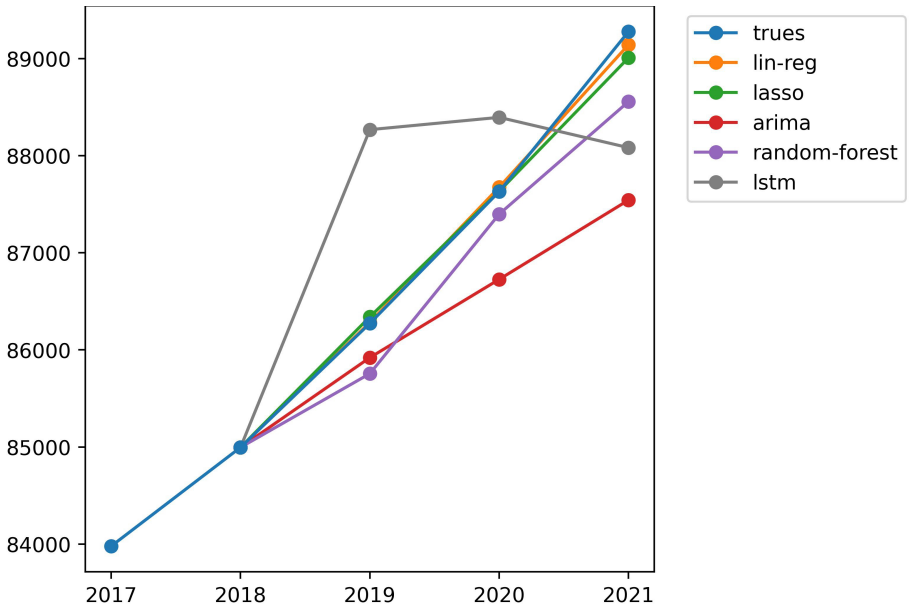

Figure 6: Regional Aggregate Establishment Count Forecasts by Year and Mode

Table 2 shows model performances by RMSE for immediate next year (1-year out) and long-term (3-year out) forecasting, while Figure 6 shows our model's long-term forecasts, aggregated up to the regional level.

We see that our ordinary least squares models, including LASSO, perform the best in terms of RMSE, and predict closest to actual establishment growth, likely because their assumptions of monotonic growth are reflected in both our training and testing data.

Like our ordinary least squares models, our ARIMA model appears to underestimate establishment growth year after year. However, due to it making predictions by averaging the last few timestamps, each under-prediction further reduces the next prediction, resulting in significant under-prediction compared to our other models. In actuality, though, this result of conservatively underestimating establishment growth may be preferred to an alternative of compounding overprediction when it comes to the forecast's influence on planning decisions.

While our "Algorithmic Models", random forest and LSTM, perform poorly in terms of RMSE, they exhibit more "dynamic" predictions than our "Statistical Models", in that, their forecasts change more drastically than our "Statistical Models", which are mostly predicting the same number of new establishments every subsequent year. This indicates some potential for the "Algorithmic Models" to outperform "Statistical Models" in cases where recession years are accounted for. Another potential reason for their higher RMSEs could simply be due to the number of records we used for training, as without enough diverse training data, our "Algorithmic Models" resort to overfitting on the train set.

Figure 7 shows establishment count forecasts for the entire San Diego County up to 2050 generated by our auto-regressive feedback LSTM models, forecasting establishment growth which quickly decays then returns to exponential growth.

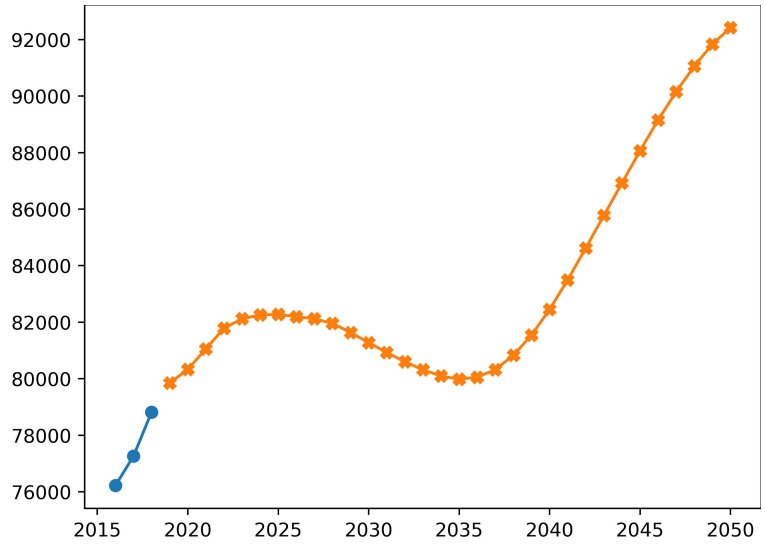

Figure 7: Forecasted Regional Aggregate Establishment Count by Year (LSTM)

4 Conclusion

Our results conclude that "Statistical Models" outperform "Algorithmic Models" in forecasting establishment counts by ZIP Code, with LASSO regression coming onto with an immediate next year and long-term testing RMSE of 22.83 and 20.04 respectively. However, we believe that the data we resorted to using, due to various data limitations, significantly influenced this outcome, as a lack of recession years coincidentally fit the assumptions of our ordinary least squares models, where the real-world phenomena may not. Replicating our project with a longer time series, containing multiple recession years, and introducing monthly seasonality may better reveal the potential of "Algorithmic Modeling", particularly LSTMs, for business growth forecasting.

Further conclusions about our input features can also be made considering the performance of our random forest model. Our random forest model performs the worst out of all our models, with an RMSE of 38.80 for immediate next-year forecasting and 48.99 for 3-year out forecasting, which indicates that, while our input features may perform well in explaining variation in establishment births/deaths geographically, they do not provide sufficient explanatory power temporally. As such exploring different avenues for explanatory features may prove fruitful.

While our work used RMSE for evaluation simplicity, we note that this can lead to our models overfitting on the few ZIP Codes that contain the majority of businesses in the region. This may potentially lead to comparably worse prediction errors for up-and-coming, developing, ZIP Codes, which are especially in need of developmental support from planning projects informed by these forecasts. Future work into evaluation metrics for forecasting models can be explored to mitigate such prediction biases to ensure certain areas are not disproportionately favored and avoid compounding algorithmic injustice.

References

Breiman, L., J. Friedman, C.J. Stone, and R.A. Olshen. 1984. *Classification and Regression Trees*. Taylor & Francis. [Link]

Breiman, Leo. 2001. "Random Forests." *Machine Learning* 45 : 5 – 32. [Link]

Breiman, Leo. 2003. "Statistical Modeling: The Two Cultures." *Statistical Science* 16 (3): 199 – 231. [Link]

Graves, Alex. 2014. "Generating Sequences With Recurrent Neural Networks." [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. 2018. "Series 14 Regional Growth Forecast Documentation and Baseline Subregional Allocation." [Link]

Appendices

A.1 Training Details

Modifications to Master Dataset for Certain Models:
Since the multiple linear regression and random forest regression models require data on input variables at inference, they were trained on a modified dataset where all input variables were lagged by 1 year, reducing training data by 1 year.

Explanation of Mean Decrease in Impurity:
"In scikit-learn, we implement the importance as described in [1] (often cited, but unfortunately rarely read...). It is sometimes called "gini importance" or "mean decrease impurity" and is defined as the total decrease in node impurity (weighted by the probability of reaching that node (which is approximated by the proportion of samples reaching that node)) averaged over all trees of the ensemble."(Breiman et al. 1984)