Birth Rate Forecasting Using Deep Learning Models

Yuya Jeremy Ong
SOC 423: Demography
Final Project Presentation

Outline

1. Introduction
2. Background & Related Work
3. Modeling Methodology
4. Dataset & Feature Analysis
5. Experiment Setup
6. Model Results
7. Discussion

Introduction

- The majority of academic work is often concentrated on *retrospective analytics* of population distributions, changes, and dynamics.

- Preston [1] outlines how trends in demographic study will improve with the following trends:
  A. Hardware and software for *data storage, retrieval, and processing are improving*.

  B. *Data collection, accessibility, and quality control* of data are improving with organized institutions and the internet through *Big Data, Social Media, etc...*

  C. *Statistical methods and data processing algorithms* are improving significantly.

- *Forecasting tools and algorithms* are crucial for various other sociological studies and decisions making guidelines for policy makers - especially issues pertaining to birth rates.

Projection Modeling

- Projection Modeling is a methodology applied to *taking past data* and generating a model which will describe the future demographic distributions and structure.

- Typically, we frame the modeling problem as a Time Series Model, where given a series of past historical values, we attempt to *predict the next sequence of values*.

- The United Nations was one of the very few organizations, in 1957 to release a projection model of various demographic information, such as birth rates.

Related Work

There are many published works and various models, which are often based on various mathematical and statistical based methods:

- United Nations' Population Division proposed first set of projections in 1957 and continuously updates models.
- Bongaarts proposed a **deterministic model** implementing various socio-economic features.
- Alkema et. al proposed a **probabilistic model** based on Bayesian methods.
- Shang et. al developed an **ARIMA forecasting** model.

There are a lot more work out there, *we are only scratching the surface of what has been done so far*.

Fundamental Challenges

Forecasting and predicting future trends is a non-trivial task due to the following reasons:

1. Data collection is often very challenging - regarding sampling strategies, quality control, and feature measurements.

2. Understanding and disentangling factors of proximate determinants of complex relationship is challenging.

3. Accounting for anomalies and sudden shifts in the distribution is very tricky (i.e. 1960 Baby Boom).

United Nations Medium Variant Population Projections, 1957 to 2000 (billions)

Year	Actual	1957	1963	1968	1973	1978
1950	2.52					
1960	3.02	2.91				
1970	3.70	3.48	3.59	3.63		
1980	4.44	4.22	4.33	4.46	4.37	4.42
1990	5.27	5.14	5.14	**5.44**	5.28	5.28
2000	6.06	**6.28**	**6.28**	**6.49**	**6.25**	**6.20**

United Nations, World Population Prospects (various issues), Population Division, Department of Economic and Social Affairs, New York.

ARIMA Model

The most commonly used method in time series modeling is ARIMA (Autoregressive Integrated Moving Average) or commonly known as the Box-Jenkins method.

- **Autoregressive (p):** The model regresses on its own lagged prior values.
- **Integrated (q):** Using differencing of raw observations to make values stationary.
- **Moving Average (r):** Considers the observation and residual error to factor in the lagged observations.

Advantages

- Easily interpretable parameters.
- Simple to implement and widely applied in various problem domains.
- Proven to be fairly robust for most cases - however with low precision.
- Computationally least intensive.

Disadvantages

- Assumes data to be stationary (mean, variance, and autocorrelation are consistent) - if not requires manual transformation.
- Assumes linear relationship temporally.
- Additional preprocessing and smoothing must be applied to data.
- Hand-Tuning parameters, p, q, and r is very difficult.

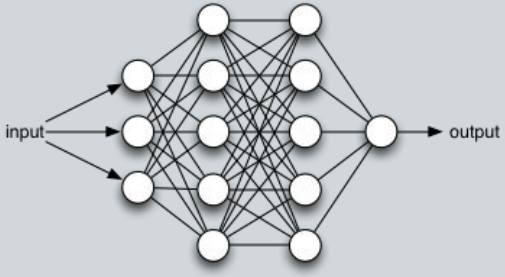
Deep Learning

- **Deep Learning** a specific type of **Machine Learning** algorithm inspired by the human brain, known as **Neural Networks**.

- Recently, Neural Networks have shown significant improvement in predictive modeling of complex features in various tasks like Image Recognition, Self-Driving Cars, Natural Language Processing, etc - achieving state-of-the-art results.

- Models demonstrate a high rate of accuracy and precision over previous deterministic and statistical methods.

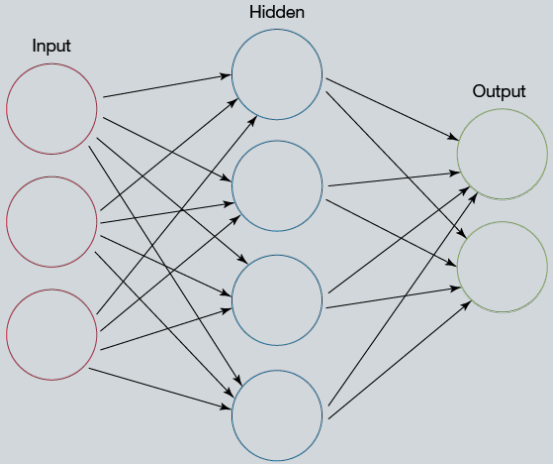
- Models are capable of learning ***complex non-linear features*** from the data in an end-to-end fashion.



The Proximate Determinant Framework

The graphical topology of Neural Networks can help to model Bongaarts' Proximate Determinant Framework in an end-to-end manner.

Indirect Determinants
Socioeconomic
Cultural
Environmental
Political
Technological
etc...

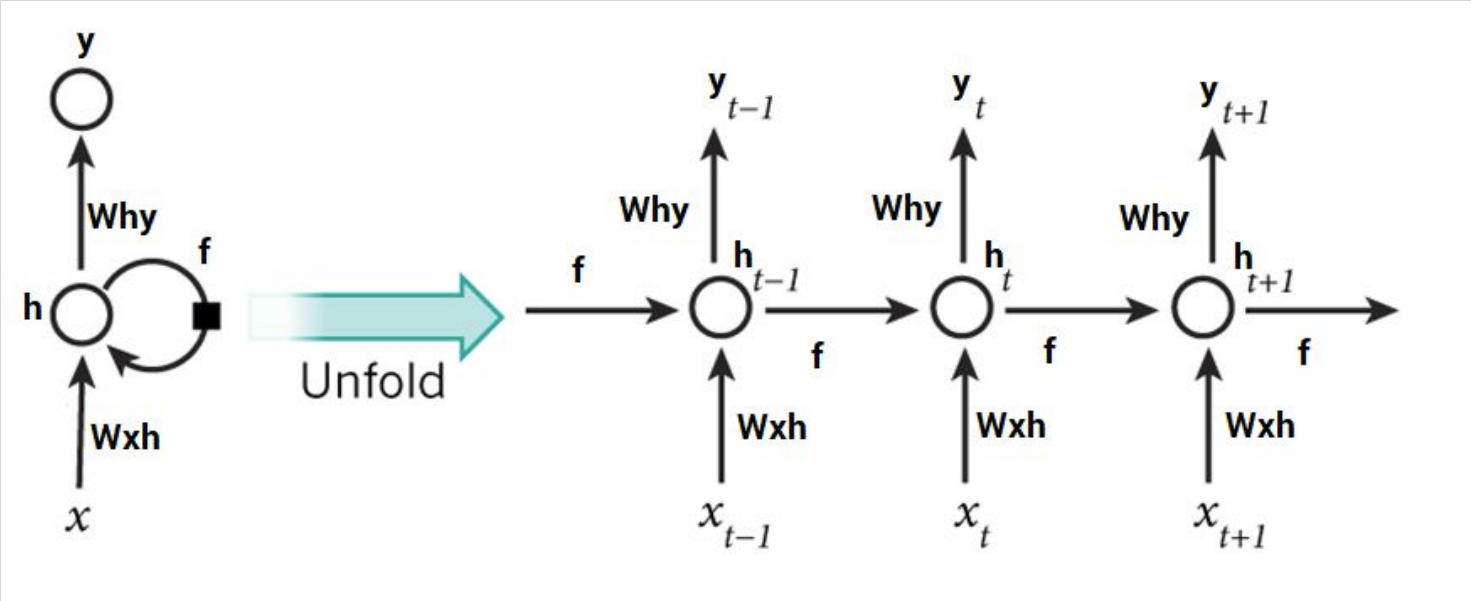


Direct Determinants
Fertility
Marriage
Reproductive Health
etc...

Output:
Forecasted Fertility Rates

Recurrent Neural Networks

Recurrent Neural Networks are a variant of Neural Networks which learns a temporally dependent set of parameters, which allows it to learn a sequence of values.



Experiment Setup

- For simplicity, we frame the problem as an **univariate time series model**
  - Consider most simplest model.
  - Useful model for countries with insufficient data.
  - Easiest baseline to construct, given the short amount of time for this project.

- Construct two models for evaluation:
  - ARIMA [Facebook's Prophet Library]
  - RNN (LSTM) Model [PyTorch Deep Learning Library]

Dataset

Dataset: Birth Rate, Crude (per 1,000 People)
Source: The World Bank

Year Ranges: 1960 - 2016

Total Countries: 235

“Crude birth rate indicates the number of live births occurring during the year, per 1,000 population estimated at midyear. Subtracting the crude death rate from the crude birth rate provides the rate of natural increase, which is equal to the rate of population change in the absence of migration.”

Model Evaluation

- We build the model based on a temporal validation method, and correspondingly perform a **20 year projection of the future birth rates**.

- Input: Data between the years of 1960 to 1996

- Output: Predictions from 1996 to 2016

- Error Metric for Evaluation: Mean Absolute Percentage Error

$$\mathrm{M} = \frac{100\%}{n} \sum_{t=1}^{n} \left| \frac{A_t - F_t}{A_t} \right|,$$

A = Actual Value
F = Forecasted Value
N = Number of Total Elements

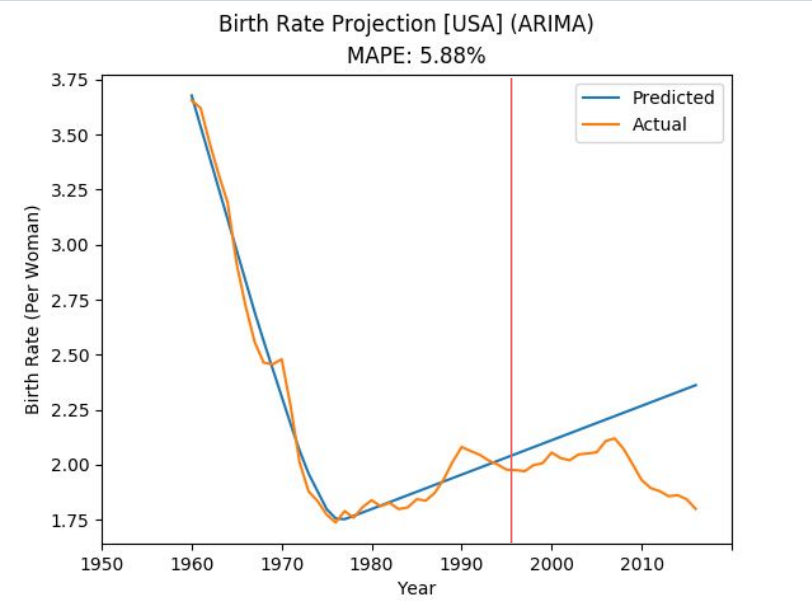
Model Results (MAPE Distribution)

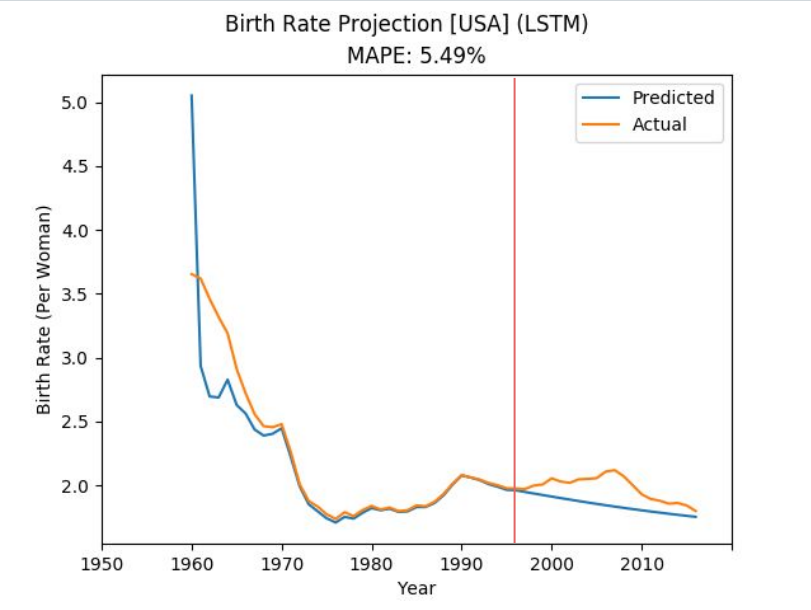
We find that the RNN (LSTM) model performs better than the ARMIA model.

ARIMA		**RNN (LSTM)**	
Mean:	**9.127874**	Mean:	**8.790662**
Standard Deviation:	18.151828	Standard Deviation:	9.395344
Min:	2.484999	Min:	0.249976
25%:	4.745813	25%:	2.506318
50%:	5.788534	50%:	5.299511
75%:	7.811595	75%:	12.230355
Max:	208.356131	Max:	51.121002

Empirical Results

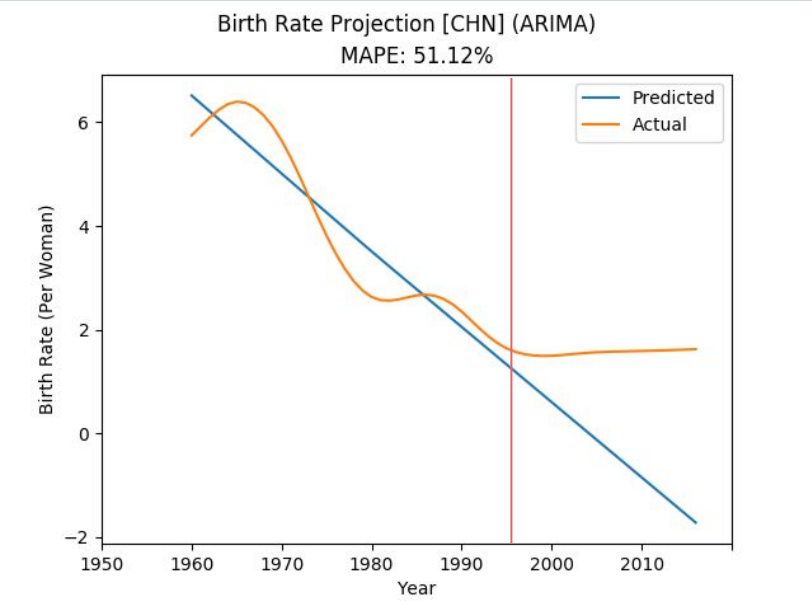
United States

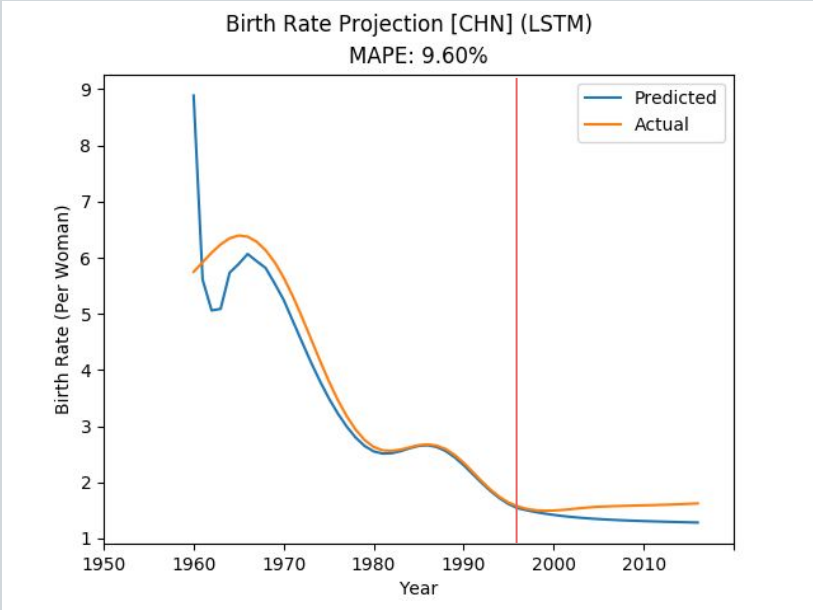




Empirical Results

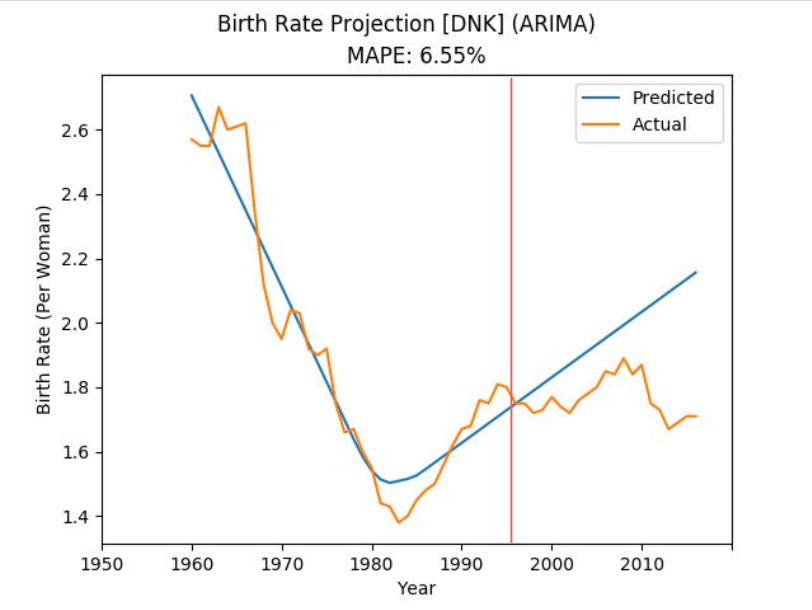
China

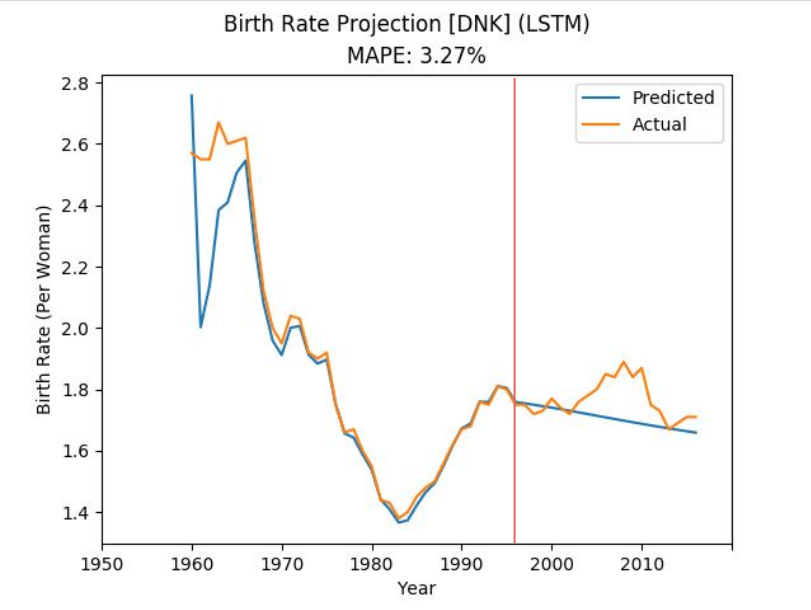




Empirical Results

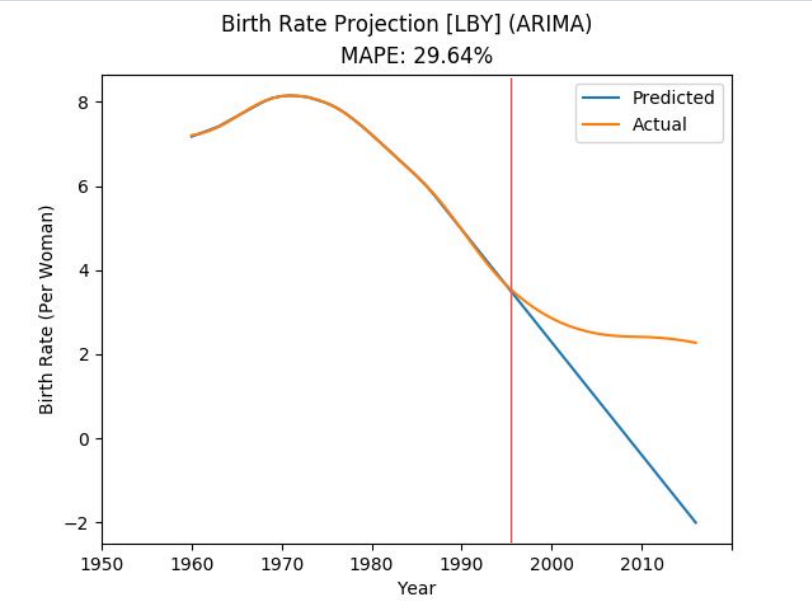
Denmark

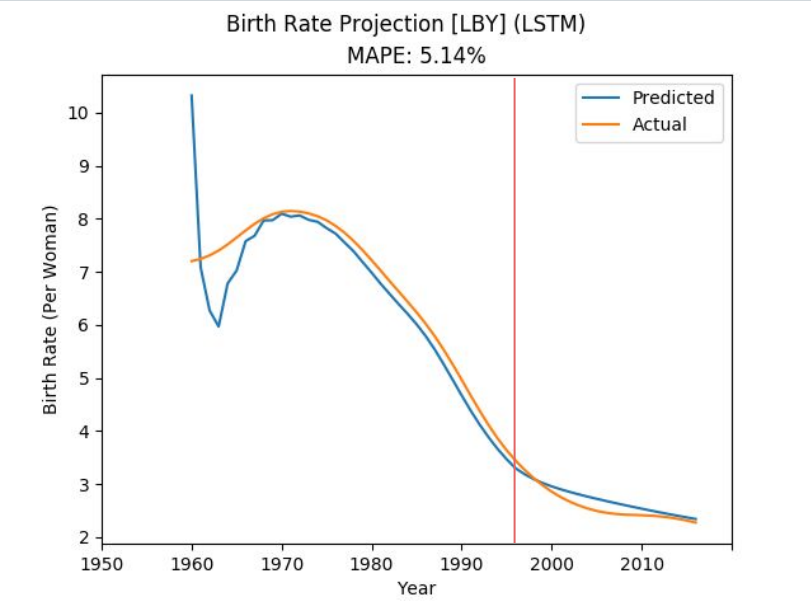




Empirical Results

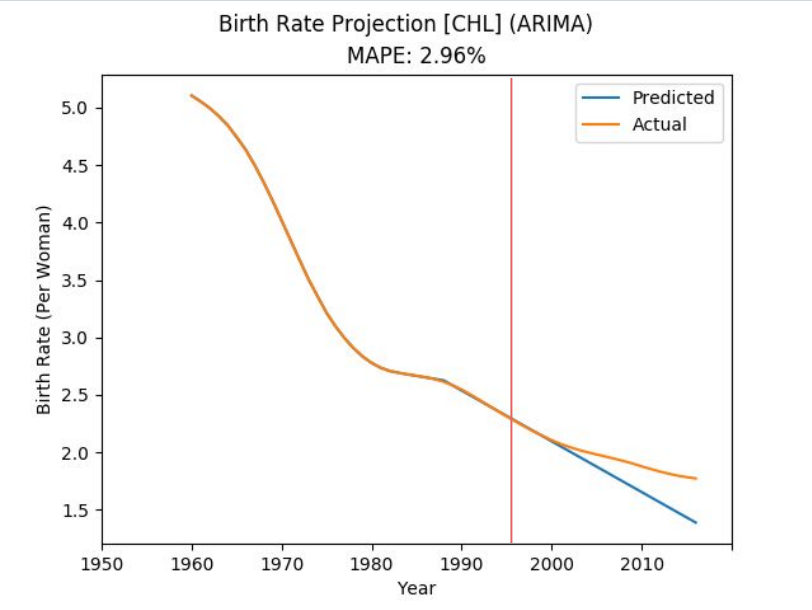
Libya

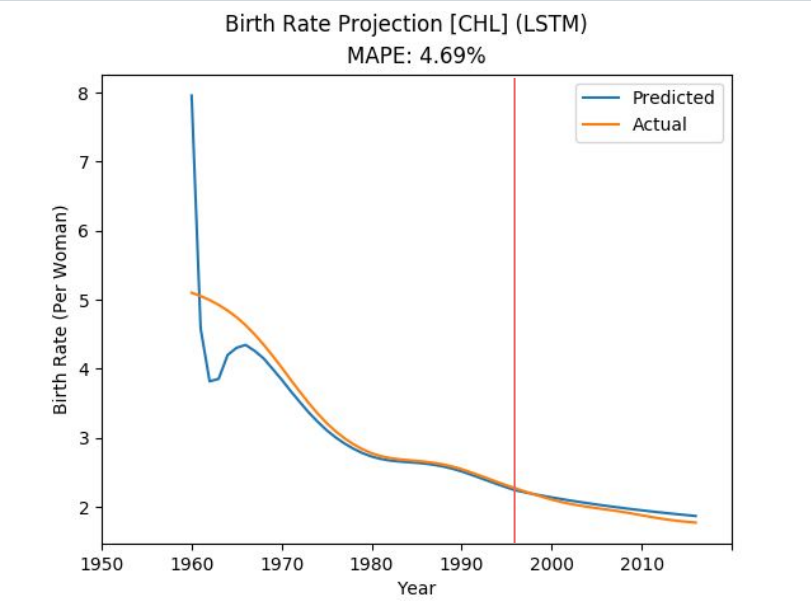




Empirical Results

Chile





Discussion

- Multiple models should be utilized when weighing different scenarios of projections.
- Quality of data sources will matter greatly when modeling.

Below we consider some of the advantages and disadvantages of the Deep Learning based model:

Advantages

- Requires no preprocessing or any strong statistical assumptions necessary.
- Higher level of precision & accuracy of prediction.
- Ability to develop model in an end-to-end fashion without too much effort in parameter tuning.
- Able to capture nonlinear dynamics of the dataset.

Disadvantages

- Complex development and maintenance process.
- Potential for overfitting the data - better architectures can improve this.
- Interpretability issues model - utilize as black box system.
- Computationally expensive - however new GPU hardware can increase speeds.

Future Work

To extend the capabilities of the model, we can consider the following potential ideas:

- Use much more complex Deep Learning models which are known to have good state-of-the-art results.
- We can consider a larger set of features and frame the problem as a multivariate time series modeling problem.
- Consider data from social media and other various features from other sources which may potentially work towards understanding the factors of fertility.

Conclusion

1. Evaluated key challenges in birth rate forecasting.
2. Introduced Deep Learning in the context of birth rate forecasting.
3. Developed a univariate time series model using Deep Learning.
4. Empirically demonstrated the effectiveness of Deep Learning methods in birth rate forecasting through experimental results.

Questions, Comments, or Suggestions?

References

[1] Contours of Demography