Title: Multivariate Forecasting of Scotland's Monthly Birth Rates

1. Introduction

Accurate forecasting of birth rates is crucial for healthcare planning, educational infrastructure, social services, and economic policy. While univariate models based solely on historical birth counts can capture trends and seasonality, they often miss important drivers—such as economic conditions, public health indicators, and sociocultural factors—that influence family-planning decisions. This proposal outlines a **multivariate time-series forecasting** approach that integrates additional predictors to enhance the accuracy and interpretability of birth-rate forecasts for Scotland.

2. Problem Statement

Existing forecasts for Scotland's monthly birth volumes rely primarily on past birth data (univariate models). However, fertility behavior is also shaped by:

- **Economic indicators** (unemployment rate, consumer price index)
- **Health factors** (maternal access to prenatal care, average maternal age)
- **Education and socio-demographic variables** (female tertiary-education enrollment, average household income)

Gap: Without these covariates, forecasts may under-react to sudden economic shocks (e.g., recessions) or public-health events (e.g., pandemics).
Goal: Develop a multivariate forecasting framework that leverages multiple data sources to produce more robust, real-time birth-rate predictions.

3. Methodology

1. **Data Integration & Preprocessing**
    - Collect and merge monthly series for births, unemployment rate, inflation (CPI), maternal-health metrics, education enrollment, etc.
    - Impute missing values and align on a common date index.
    - Perform exploratory data analysis (EDA) to visualize correlations and seasonality across variables.
2. **Feature Engineering**
    - Create lagged features (e.g., 1-, 3-, 6-month lags) for each predictor.
    - Generate rolling aggregates (e.g., 3-month rolling average of unemployment).
    - Encode calendar effects: month-of-year dummies, holiday flags.

3. **Model Selection**
    - **Baseline:** Multivariate ARIMA (VAR or ARIMAX).
    - **Machine Learning:**
        - Gradient-boosted trees (XGBoost, LightGBM) with lagged features.
        - Random Forest regression with recursive feature elimination for interpretability.
    - **Deep Learning:**
        - LSTM/GRU networks to capture complex temporal dependencies.
        - Temporal convolutional networks (TCNs) for longer-range interactions.
4. **Training & Validation**
    - Use a **rolling-window evaluation** (e.g., expanding window: train on 1998–2018, validate 2019–2020, test 2021–2022).
    - Optimize hyperparameters via cross-validation focused on forecast accuracy (MAE, RMSE, SMAPE).
5. **Interpretability & Explainability**
    - Apply SHAP or PDP (Partial Dependence Plots) to quantify each feature's contribution.
    - Compare feature importance over time to identify shifting drivers.
6. **Deployment**
    - Package the best model into a **Streamlit/Dash dashboard** for interactive forecasting by policymakers.
    - Automate monthly data ingestion and model retraining.

4. Source Data

Data Category	Source	Frequency	Period
Monthly Birth Registrations	National Records of Scotland (NRScotland.gov.uk)	Monthly	Jan 1998 – Dec 2022
Unemployment Rate	UK Office for National Statistics (ONS)	Monthly	Jan 1998 – Dec 2022
Inflation (CPI)	ONS	Monthly	Jan 1998 – Dec 2022
Maternal Health Indicators	Public Health Scotland	Monthly	Jan 2010 – Dec 2022
Female Tertiary Enrollment	Higher Education Statistics Agency (HESA)	Annual*	1998 – 2022
Holiday Calendar	UK Government holiday API	Date-level	1998 – 2025

Annual series will be interpolated to monthly granularity.

5. Tools and Technologies

- **Data Wrangling & EDA:** Python (Pandas, NumPy), Jupyter Notebooks
- **Statistical Modeling:** Statsmodels (VAR/ARIMA), scikit-learn
- **Machine Learning:** XGBoost, LightGBM, scikit-learn
- **Deep Learning:** TensorFlow (Keras), PyTorch
- **Hyperparameter Tuning:** Optuna or scikit-optimize
- **Visualization:** Matplotlib, Seaborn, Plotly
- **Dashboard Deployment:** Streamlit or Dash
- **Version Control & Collaboration:** GitHub, GitHub Actions (for CI/CD)
- **Workstation:** Google Colab / local GPU-enabled environment

6. Expected Outcomes

- **Improved Forecast Accuracy:** Demonstrated reduction in MAE/RMSE compared to the univariate baseline.
- **Driver Analysis:** Clear insights into which external factors most influence birth-rate volatility.
- **Interactive Tool:** A live dashboard for policymakers to explore "what-if" scenarios (e.g., rising unemployment).

7. Conclusion

By incorporating economic, health, and educational predictors into a multivariate forecasting framework—and by leveraging both traditional and modern ML/DL methods—we aim to deliver **more accurate**, **explainable**, and **actionable** birth-rate forecasts for Scotland. This approach will empower decision-makers to anticipate demographic shifts and allocate resources more effectively.