Predicting Average Age of First Motherhood

Problem Identification

Approaching 30, many of my female peers are feeling pressure around motherhood. Some inside relationships feel rushed, some outside of relationships feel that time is running away from them. Looking outwards it *feels* like society is changing, it seems that more women are choosing motherhood later or not at all. But is this apparent cultural shift real?

This project investigates whether the data supports this perception: are women, on average, becoming mothers later in life? Specifically, I aim to analyse historical data for the age of first motherhood for women born in a given year and use machine learning techniques to forecast future trends.

While this question is personally relevant to me and my peer group, its implications extend far beyond the societal pressure felt by individual women. Accurately modelling the trend in the age of motherhood could inform:

- Healthcare e.g. adapting ante & post natal care for older mothers.
- Research e.g. understanding of risks for older mothers and their children, and anticipating increased demand for fertility treatment and assisted conception.
- Education e.g. reassessing current school infrastructure for changing population of children.
- Policy-making e.g. planning for aging population with fewer familial caregivers.
- Business e.g. anticipating market shifts as disposable income increases for households with fewer/no children.

While this assignment does not seek to explain *why* women may be delaying motherhood, forecasting the trend is a valuable step in planning for the consequences.

Dataset Selection & Exploration

Selection

The data set used to answer this question comes from the Office of National Statistics (ONS) and can be found here: https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/conceptionandfertilityrates/datasets/childbearingforwomenbornindifferentyearsreferencetable

In researching data on this topic, there were multiple papers but most referred to this same or similar ONS resources for their data sources. There was some ONS data available on birth characteristics and more detailed birth registration information, but given the aim of this project is focussed on the mother, this dataset was by far the most relevant.

One other source considered, also from the ONS, further breaks down motherhood statistics by local authority, which would allow analysis of whether there are regional differences. However, this data only covers 1993-2021 and categorises mothers as ‘under 20’ then by age in years up to ’40 and over’. This shorter time span and reduced specificity would lead to less reliable conclusions and so this dataset was not chosen. https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/livebirths/adhocs/1265livebirthsbyageofmotherandlocalauthorityenglandandwales1993to2021

About the Source

The ONS is the UK’s largest producer of independent research and a data source trusted by both government and academia. They hold internationally recognised standards and are transparent with their methods and quality assurance, meaning their data is trustworthy, accurate, and impartial. This particular dataset has been updated in 2025 and so as timely as possible.

About the Data

The metadata from the ONS must be understood to clearly define the scope and limitations. Full information can be found in the Notes and Metadata sheets of the excel file.

Key points are as follows:

- This data has been formed from ONS birth registrations data, mid-year population estimates and 2022 based national population projections.
- The data pertains to women born in England and Wales 1920-2008 and includes only live-births.
- Birth statistics by year of birth of child and age of mother have been available since 1938. Tables 1-4 show these statistics in birth cohort form, by year of birth of the mother. Years of birth of mother are necessarily approximate, since before 1963 the

data was only available for age of mother at the time of the child's birth. For example, a 30 year old mother giving birth in 1951 could have been born in 1920 or 1921, for ease this mother has been regarded as belonging to the 1921 cohort.

- Tables 1-3 show mother's age in exact years - by age 20 meaning up to and including the day before 20th birthday and table 4 shows mother's age in completed years - 20 meaning up to and including the day before 21st birthday.

- The data is presented from ages 16-45. Live births to women by age 16 includes all births up until a woman's 16th birthday.

- Statistics where age is given as 'Final' includes births after the 45th birthday occurring before the end of 2023 for women born in 1978 or earlier. For women born in 1979 onwards, 'Final' estimates include projected births to women beyond their 45th birthday. The number of births to women after their 45th birthday is very small.

- Tables 2 & 3 use True Birth Order (TBO) estimates. Until May 2012, birth registration data was collected on the number of the mother's previous children, live and still births, but only counting those births fathered by a current or previous husband. This meant that births occurring outside of marriage, whether the parents were subsequently married or not, were not counted. To allow for the fact that the proportion of births occurring outside marriage has risen steadily, the information collected on birth order at registration up until May 2012 was supplemented to give estimates of overall or true birth order. Amendments to the Population (Statistics) Act 1938 mean that since May 2012, information is now collected on the total number of all previous live and still births that the mother has had.

  The method of estimation using General Household Survey data can de found here https://webarchive.nationalarchives.gov.uk/ukgwa/20160105160709/http://www.ons.gov.uk/ons/rel/population-trends-rd/population-trends/no--108--summer-2002/population-trends.pdf

  Details on the changes to ONS birth statistics since May 2012 can be found here https://webarchive.nationalarchives.gov.uk/ukgwa/20160105160709/http://www.ons.gov.uk/ons/guide-method/user-guidance/health-and-life-events/quality-assurance-of-new-data-on-birth-registrations.pdf

- Obviously, male fertility and childbearing has influence on the data for women. However, the birth registration process does not collect data on the number of children a man has already had and without this information it is not possible to produce estimates for the average number of children a man will have or the proportion of men who become fathers. A man's reproductive span is also not as well defined as a woman's and so a much longer time series would be needed to calculate cohort measures.

- Tables 1b & 4b include projections derived from the 2022-based national population projections. These are not used in this assignment to avoid projected values having any influence on training. However, comparing with ONS projections may be useful in corroborating results from my model.

Possible Approaches

The data looks at childbearing for women born in different years from multiple perspectives, offering a few approaches to answering the question of this assignment.

Data	Possible Approach	Pros	Cons
Table 1a - average number of live-born children by age and birth cohort	Examine and predict, by birth year, at what age the average number of children first reaches 1.	This could be a measure for when we could expect an 'average' woman in the birth cohort to have had a child.	Impacted by women who have their first child at a younger age then having further children. So more a measure of how many children are born to the cohort rather than how many of the cohort become mothers.
Table 2 - proportion of women who have had at least one live birth, by age and year of birth of woman	Examine and predict, by birth year, at what age the proportion of women who have had a child first reaches 0.5. OR Examine and predict, by birth year, what proportion of women have had a child by age 30. OR Examine and predict, by birth year, whether a 30 year old woman would be 'expected' to have had a child (proportion >= 0.5)	This could be a measure for the age at which motherhood moves from uncommon to common within the birth cohort. More accurately representing when women are choosing motherhood. These ideas specifically target the age 30 which is common in culture as a reflection point at which women may feel motherhood becomes especially relevant.	1992 the last cohort for which this 0.5 threshold has been met, so only 72 past years to learn from, while motherhood appears to be changing rapidly in the years since the early 1990s. Again data is limited since 1994 is the last cohort to have turned 30. Limiting the investigation up to age 30 may also neglect some key changes in motherhood later in life. To truly answer whether motherhood is being delayed, may be better not to limit to 30. Historically, data is massively imbalanced as the proportion of women having a child by 30 has only fallen below 0.5 in recent cohorts for which there is data.

Table 3 contains data on the percentage of women having different numbers of children and table 4 pertains to age-specific fertility rates i.e. the number of children born per 1000 women. These are less relevant for answering this specific problem since they are not focussed on the first child born to a woman and so do not represent the beginning of motherhood in this context.

Problem Definition

Considering these options, an appropriate and practical approach to answer the question of whether women are becoming mothers later in life, is to use Table 2 data on the cumulative proportion of women from a birth cohort who have had a child, with 0.5 being the threshold for motherhood becoming common. Formally therefore the problem is defined as:

Predict the age at which 50% of women born in a given year (in England or Wales) are expected to have had their first child.

Data Pre-Processing

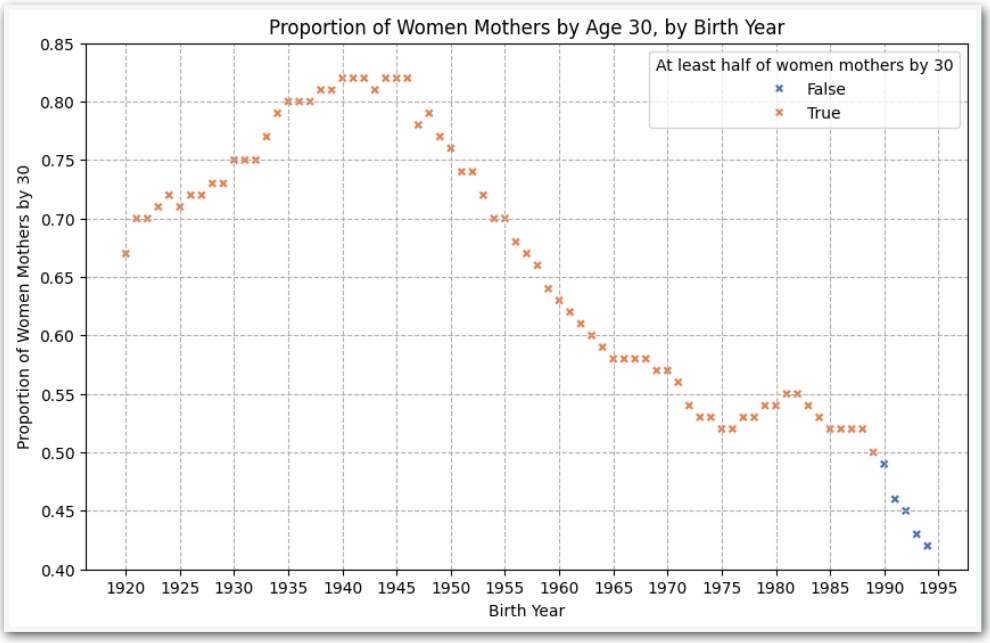
The data from the ONS is reliably complete and so not a lot of pre-processing has been necessary. A brief outline of the steps taken:

- Reading the file, selecting required sheet & trimming info above column headers
- Renaming columns for efficiency and with correct data types
- Checking NaN values follow expected pattern, not removed as informative at this stage
- Dropping childless proportion column as not needed for this analysis and can be easily calculated from other data anyway

There are no class imbalances to address here since approaching the problem from a regression perspective. However, if the problem was reframed as a classification e.g. 'Are more than half of women born in XXXX year mothers by 30?' then this would need addressing, since most historic years classify true but the trend (see below graph) shows that this is becoming false in recent years.

Since all the data is needed for exploration of the trends and to help with formulating how the data may need to be viewed in multiple ways by a model, the data has not yet been split into training and testing sets. In the modelling phase, a train-test split will be performed prior to any transformation involving the target variable in order to avoid data leakage. The data will be split chronologically rather than randomly so that the model can learn from complete curves and predict incomplete ones.

To view the pre-processing and EDA in full, please follow this link: https://colab.research.google.com/drive/1zUI9NId18xro9oRw5ICJQYHnxZNT7gzb?usp=sharing



This graph shows that the proportion of women becoming mothers by age 30 increased from 1920, peaking at over 80% for women born in the early 1940s.

Since then the proportion of women becoming mothers by 30 steadily decreased over the next 20 years to less than 60%. A plateau for the late 1960s cohorts is followed by a slight increase a decade later peaking at just over 55%.

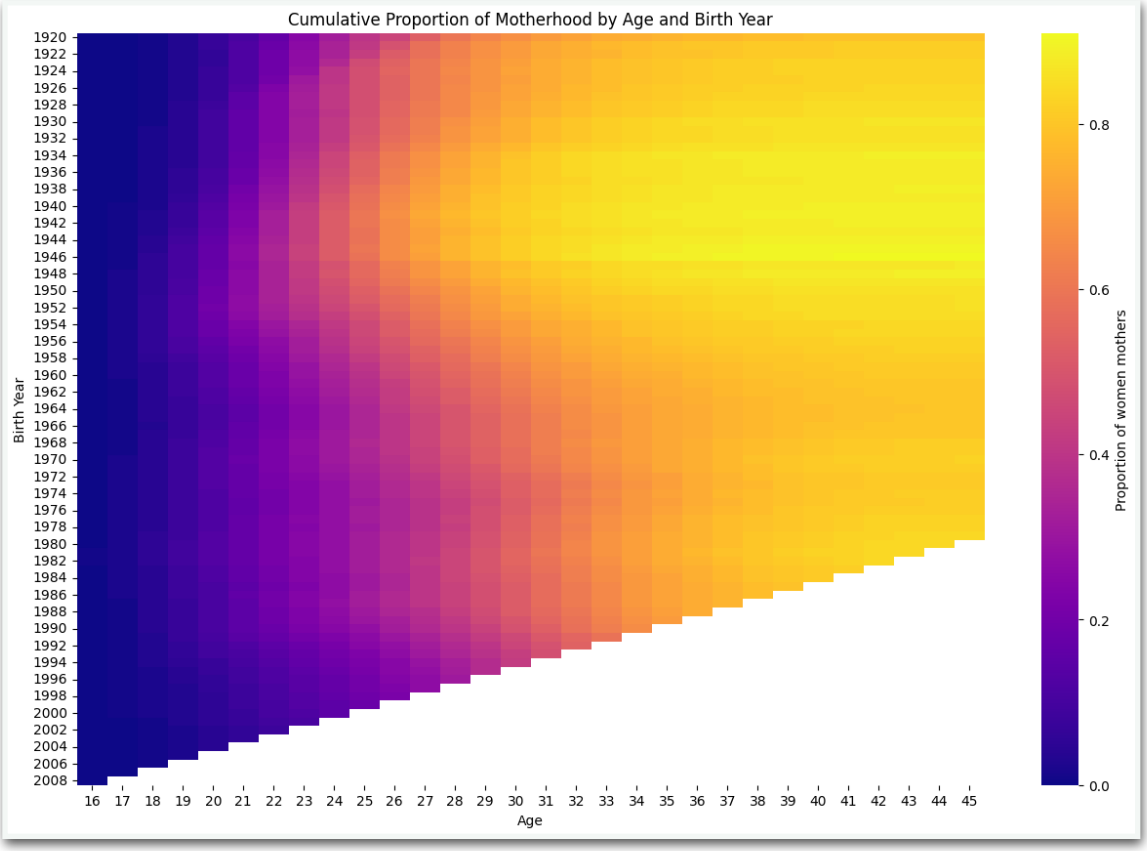
Since then, barring another plateau for the late 1980s cohorts, the proportion decreases again. Notably, in the last four cohorts to turn 30, 1990-4, the proportion of women becoming mothers has fallen below 50% and shows a declining trend.

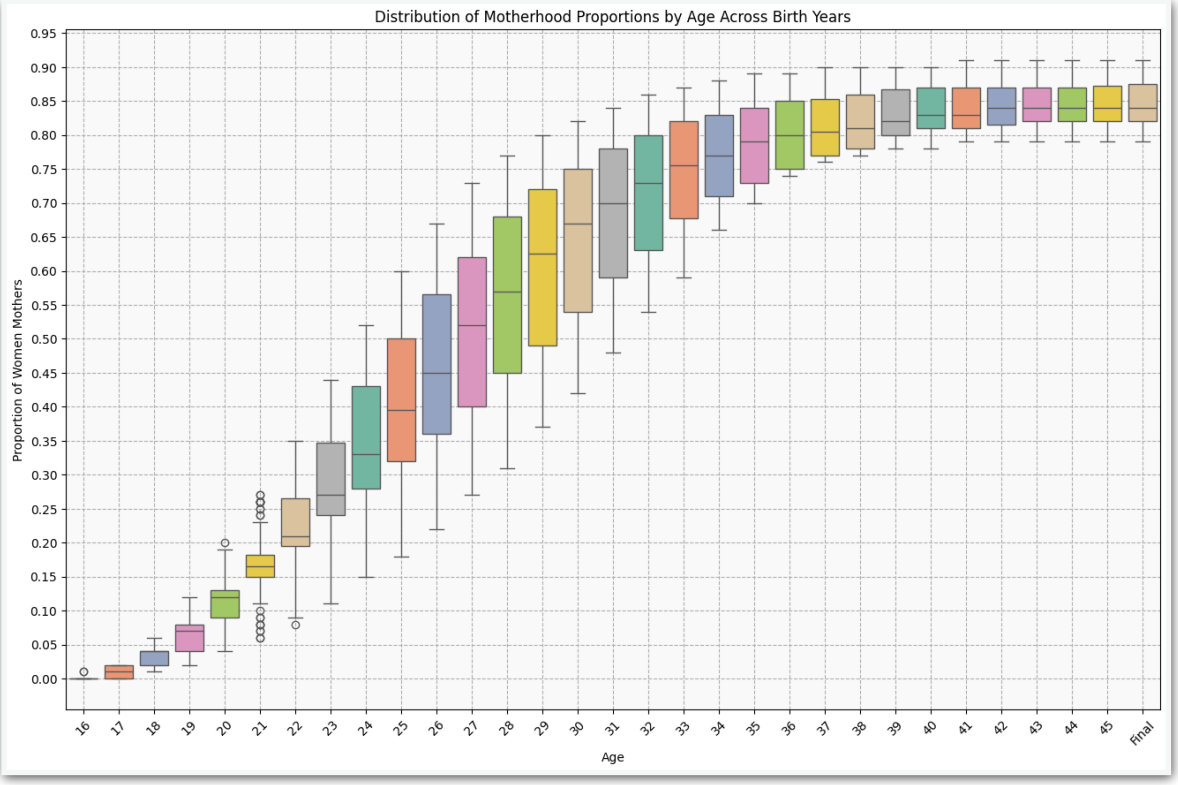
Exploratory Data Analysis

This heat map shows the cumulative proportion of women becoming mothers by age, across birth years from 1920-2008.

The brightest yellow portions indicate where more women have become mothers, we can see this occurs most and earliest for the cohorts in the early 1940s.

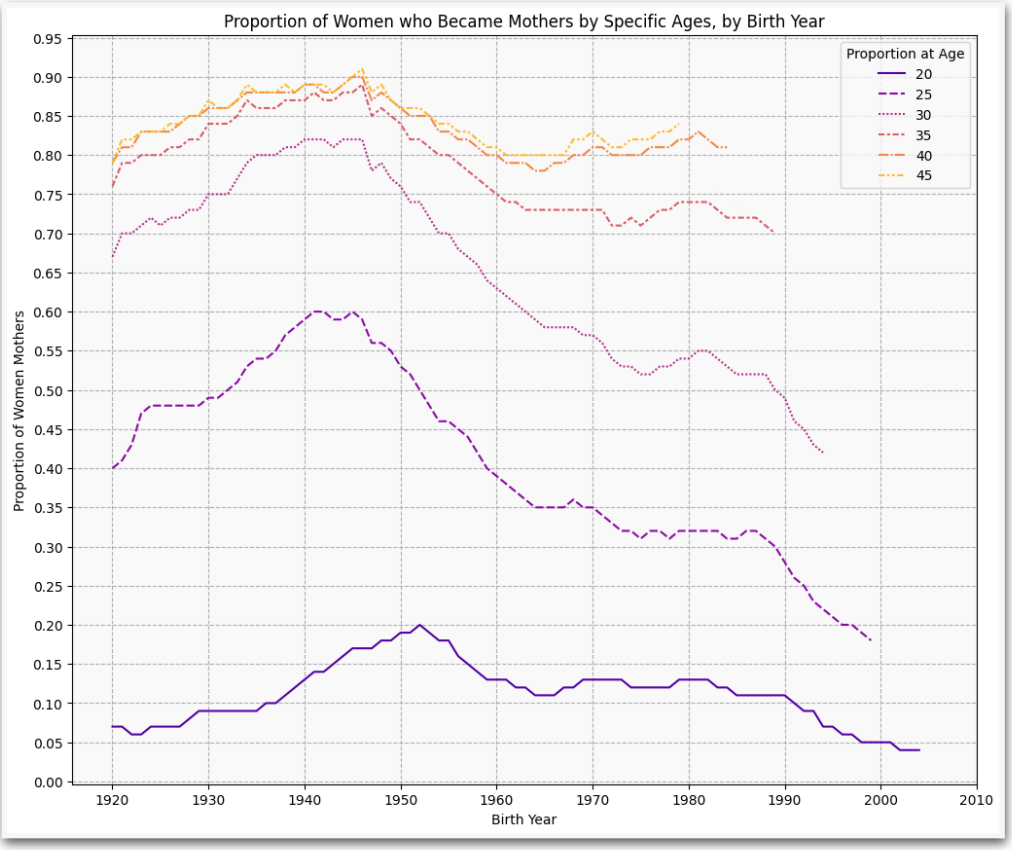
We can see from the colour gradient that, from the 1950s onwards, the darker blue purple and pink spread further along the map into the higher ages, indicating that motherhood is being delayed and, where the pink sections are particularly stretched, the transition of the cohort into motherhood is slower.





This graph shows the distribution of the proportions of women from all the birth cohorts who have become mothers at each age. The range and interquartile range are both relatively small for ages under 22, growing gradually through to the late 20s, then shrinking again and levelling out from 40 onwards. This suggests that the proportion of birth cohorts who became mothers before 22 and after 40 is fairly consistent over time, bearing in mind the 1978 cohort are the last to have completed all 45 years.

Both high and low outliers exist for age 21. Inspecting the data, these high outliers occurred in early 1950s cohorts with peak proportions 0.27 and low outliers in the most recent cohorts to reach 21, 2002 and 2003, with the 2002 cohort again the low outlier at age 22. We might expect that as the more recent cohorts age, these low outlying values may continue into the higher ages, or begin to affect the range. Since the proportion of mothers at 16 is almost always 0 in the data, the 1% reached in 1980 & 81 are automatically outliers.



Here we can see how the proportion of women becoming mothers by specific age milestones has changed over time.

All curves show very similar shape where they are complete. Increasing from 1920s to mid 1940s (1950s for age 20), decreasing most convincingly for the following 20 years and then bumpier decline from then on with a slight resurgence around the 1980s.

The most dramatic changes in proportion are in the age 25 and 30 curves with recent values about 40% lower than past peaks. This supports that women reaching these ages in the present day are delaying motherhood.

The 40 and 45 curves are much shallower and very close to one another, showing that the proportion of women who become mothers over the course of their lifetime has not changed as dramatically as is the case for motherhood at younger ages, and does not change much between the ages of 40 and 45.

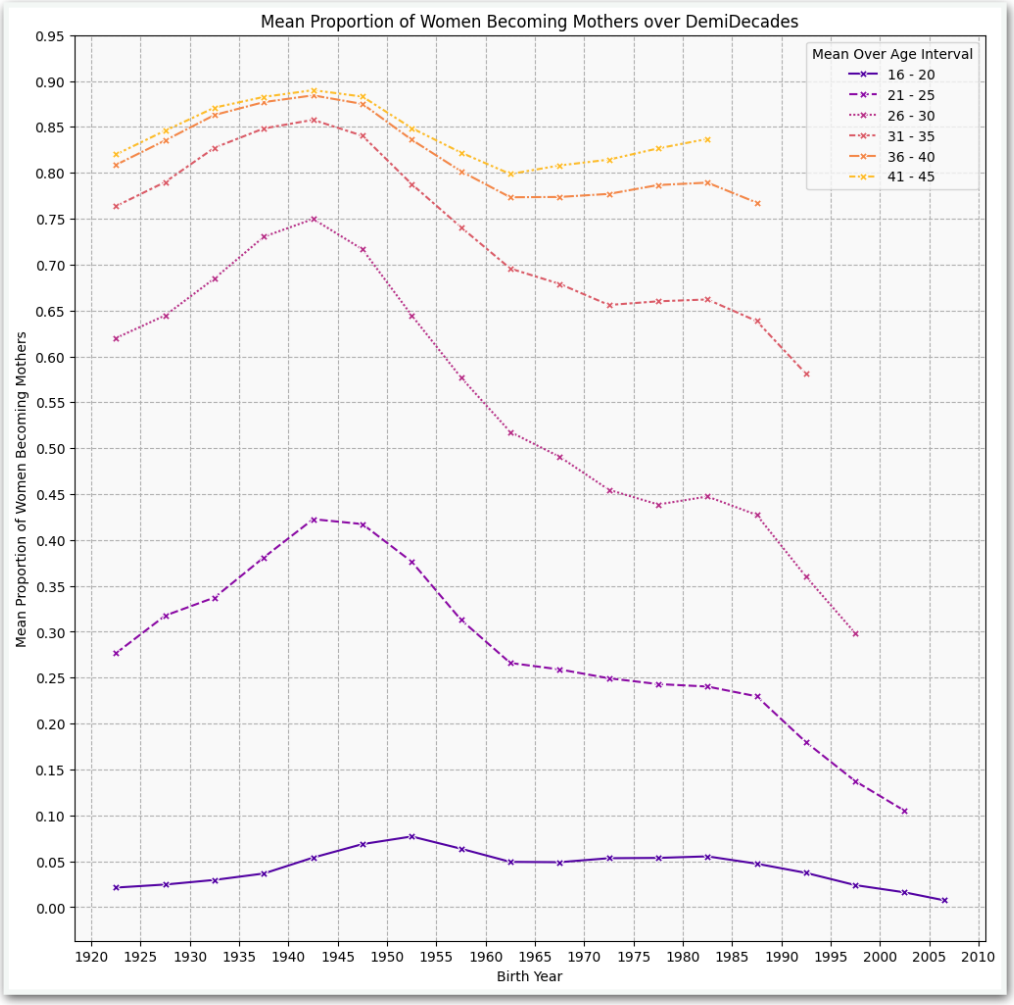
This graph shows how the proportion of women becoming mothers has changed over time, this time taking a mean over the 5 year demi-decades and a further mean over 5 year age ranges.

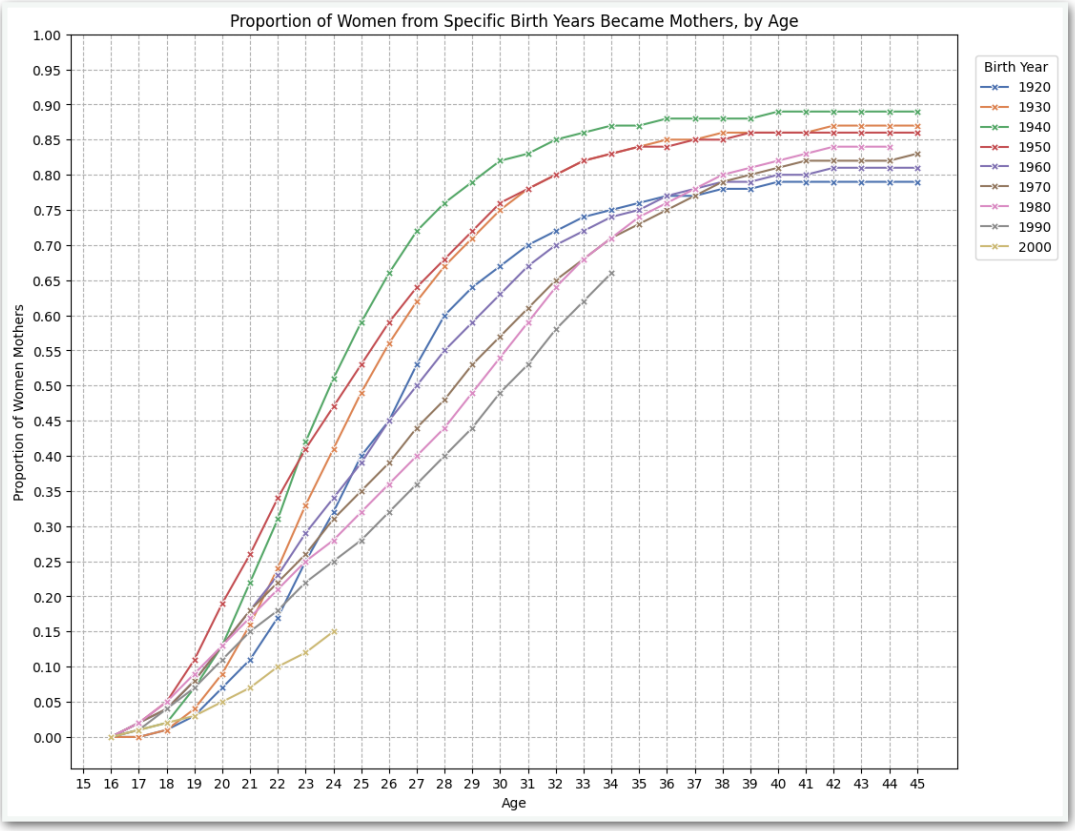
We therefore see the shape of the above curves averaged out to show the more general trend.

In this averaged version, the 26-30 curve is the one with the most extreme differences in maximum and minimum.

The increasing difference between the 36-40 and 41-45 curves here show that since the 1950s more women have been becoming mothers at these later ages. Time will tell if the cohorts from the 1980s onwards show similar trends in these older curves as they do in the younger.

Ideally the ML model would learn how the shape of these curves is changing over birth cohorts and also how the shape of the curve changes from one age range to the next, as a further angle from which to predict how the incomplete and future cohorts may behave.



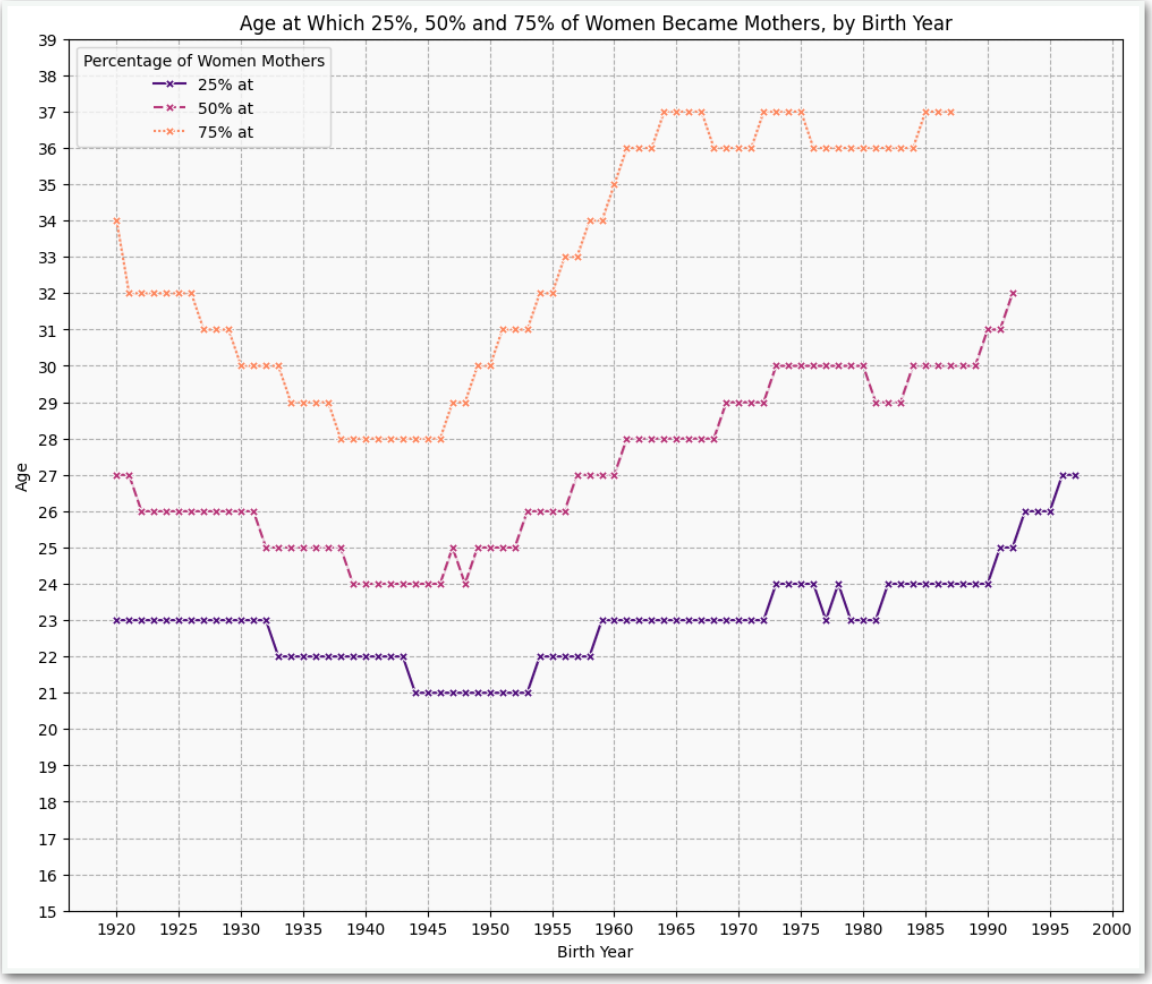


This graph shows the proportion of women who became mothers by each age for the first birth cohort in each decade.

The 1940s cohort grows fastest from age 20 and reaches the highest proportion for all ages after 23. The 1960s and 1970s are similar to the 1920s curve, all of which are outgrown by the 1930s and 1950s. This mirrors what we saw in the heat map, that childbearing increased and occurred younger from 20s-40s before declining again.

The 1980s curve grows more slowly than the 1970s curve though it catches up and overtakes at age 34.

While the 1990 and 2000 curves are incomplete, they are at lower proportions than their closest cohort at every age from 18 and, particularly in the case of the 2000 cohort, are growing more slowly getting further behind the other curves.



This graph shows how the age at which 25%, 50% and 75% of women become mothers has changed over the birth cohorts.

Each proportion was reached earliest for women born in the 1940s, and early 50s in the case of 25%. Since then these proportions have been reached at steadily increasing ages. There is some levelling out in the 75% curve since the 1960s, while the 50% and 25% curves continue to climb in age, so time will tell whether this increasing is carried forward as these younger cohorts age.

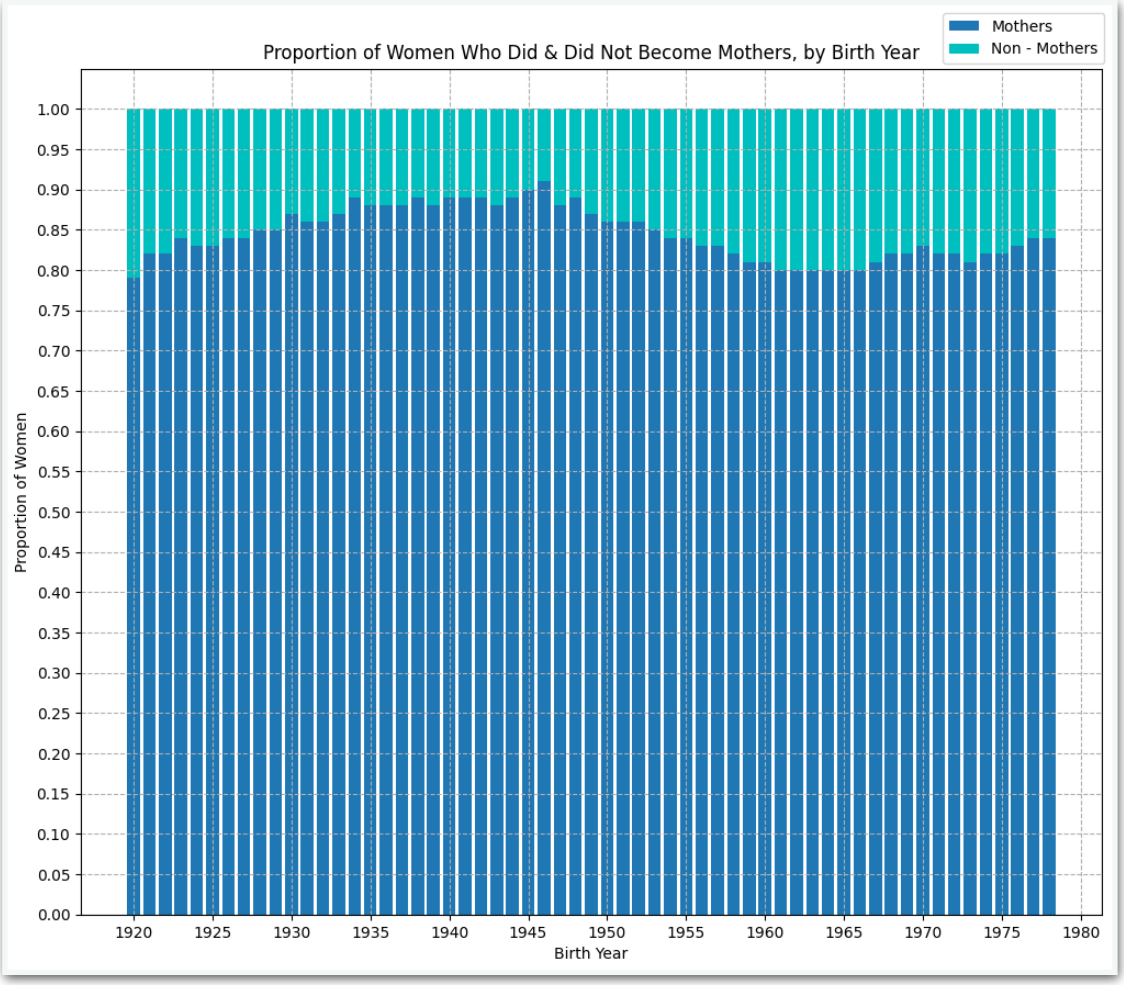
Notably, 75% of 1938-1946 cohorts were mothers by age 28, by the early 60s this is the 50% age and most recent cohorts are only at 25% by age 26, so the proportions reached by the late twenties have changed dramatically over time.

Ideally the ML model would learn how these curves are changing over time, particularly the 0.5 curve as this is the threshold determined for when motherhood becomes common in a cohort.

Finally, this graph shows how the proportion of women from the complete cohorts (those which have turned 45) who have become mothers vs not has changed over time.

While the aim of this assignment is to look at when half of women become mothers and not the final proportion, it is worth noting how the general shape mirrors what we have seen in the rest of the data, though much shallower changes occur.

Due to age, this data is still experiencing the slight increase from the late 1970s cohorts, so may yet decline as the younger curves have. The rest of the data does suggest that more recent cohorts, especially what limited information we have since 1990, may be behaving differently from women who came before them.



Machine Learning Approach

The aim of this project is to model motherhood patterns across birth cohorts of women in England and Wales using machine learning, and hence predict the age at which 50% of women in a cohort are expected to have become mothers. The target variable (age) is therefore a continuous value. The data is sequential, both within the time-series for individual birth cohorts, and in the sense that the trends are changing as one birth cohort follows another. It is therefore necessary to consider machine learning approaches appropriate for a time-series regression task.

Challenges

Time-Series Data

The limited availability of data (only 72 cohorts have reached the 50% threshold) means that some models which require large training datasets may not be useful. Data augmentation, which can be employed in other contexts to increase the size of

training datasets, is not feasible here as the chronology of the time-series must be respected and so it is not an option to invent new years to provide more training data.

Cross-validation in the traditional sense is also not possible here, again because the time-series must be understood chronologically so random shuffling of the data would not work. Data leakage in the context of time-series can be referred to as 'lookahead' as the model must learn only from the past and not look forwards to the validation/ testing series. Time series cross-validation though can still be employed: the training set is split chronologically with each validation set taken from the next consecutive block. Each new fold is a superset of the previous one, ensuring that the model observes the chronology of the time series data, looking at increasing portions of the past while never looking ahead.

This Time-Series Data

Autocorrelation, which is a measure of how the current value is affected by those that come before it, is usually an important consideration for time-series data analysis. High positive autocorrelation means that high values are likely followed by high values, and high negative autocorrelation means that high values are likely followed by low values.

Since this data is measuring a cumulative proportion of women becoming mothers, values cannot decrease and are directly affected by the preceding values, so autocorrelation will always be high within the birth cohort. Measuring autocorrelation for this data is therefore not explicitly necessary, but feature selection should ensure that this property is captured.

This particular dataset can be viewed as a time-series of time-series data. Therefore, the model needs to learn how the values change within an individual cohort *and* how this changes over time. This presents a number of challenges.

Firstly, the data is non-stationary as the distribution is shifting over time and the patterns in more recent cohorts are different from those in older ones that the model will see during training.

Another complexity is that the target variable is not a fixed distance into the future, as the EDA shows the age at which the 50% threshold is reached changes across the cohorts. Unlike typical time-series forecasting which may seek to predict say 10 years ahead, the model here must appreciate the changing forecast horizon.

It is therefore necessary to ensure feature selection captures rates of change numerically so the model can see and learn from these, while also using validation to quantify the shift.

Pre-Trained Models

Due to the complexity of the problem, it makes sense to consider transfer learning and as such seek a model pre-trained in time-series data with similar non-seasonal, non-stationary properties.

Unfortunately, my research returned models pre-trained on things like energy consumption or sales which work over continuous time and predict only the next value in a sequence. Such models are suited to short-term forecasting of values at regular time-intervals and therefore not applicable to this task of determining the varying point at which a cumulative threshold is reached. Most pre-trained options also require fixed length inputs, which is another challenge presented by this dataset as we want to be able to predict based on varyingly incomplete curves as well.

A possible exception could be the use of a pre-trained model for predicting the shape of whole curves combined with another further model to manage the incomplete aspect, as discussed further below. In the traditional sense of transfer learning though, no public model appears to be available that would easily transfer to the task of this time-series of time-series.

Model Comparison

Model	Best For	Pros	Cons
Linear Regression	Predicting single values with simple linear trends	Fast and simple to interpret	Not suitable for non-linear relationships
Random Forest Regressor	Non-linear relationships in small datasets	Captures non-linear element, still easy to interpret	Suited to interpolation, so may not extrapolate well to make forecasts
Gradient Boosting Regressor	Complex non-linear relationships	Can outperform Random Forest when tuned well, pays attention to difficult examples	Sensitive to over-fitting, requires careful tuning
Exponential Smoothing/ ARIMA	Traditional time-series prediction of a single value	If the problem sought to predict proportion at 30 in a cohort, useful to assign more weight to recent values (women now more in common with near past than far past)	These don't model curves or patterns across curves
Shallow Neural Network	Predicting whole curves from sequence-style data	Flexible modelling more complex, non-linear trends	Requires more data than is available, harder to interpret, not appropriate for predicting the single value wanted here
Convolutional Neural Network (CNN)	Predicting full curve shapes	Sees the whole curve and can learn from smaller trends in segments of the data	Requires fixed-length input for time-series data, harder to interpret
Recurrent Neural Network (RNN)	Short sequential data with simple patterns	Designed for sequential data, using loops to remember and so can learn temporal dependencies.	Struggles to retain early information in longer sequences, replaced by LSTM or Transformer models

Model	Best For	Pros	Cons
Long Short Term Memory	Predicting full time-series curves	Improves on RNN, understanding sequential patterns and time-series data as full curves, handles longer and variable-length sequences	Requires more data than is available, very computationally expensive and slow
Transformer	Using long sequences to make forecasts with long horizons	Sees the whole time-series at once, can focus on early and late parts	Requires more data than is available to avoid overfitting, computationally complex, memory intensive

Model Selection

Given the time constraints of this project, a Random Forest Regressor is a realistic and effective approach. This model captures non-linear relationships and can be tuned to mitigate overfitting, especially with small datasets. Using numerical features that capture the changing rates within and across birth-cohorts (like change in proportion of mothers age 20-25) will enable this model to learn meaningful characteristics of the curves from numerical data, even though it can not ‘see’ the curves fully. Gradient Boosting could also be explored as an alternative ensemble method, though since the training dataset is small and this method is more prone to over-fitting, the time required for careful tuning may be beyond the scope of this assignment.

There may be some advantage in also training a Convolutional Neural Network (CNN) to model full curve shape. This requires fixed-length input, so the CNN would be trained on truncated curves from the earlier complete cohorts. Features extracted from the CNN would then be passed to a Random Forest Regressor, in a kind of hybrid transfer learning, to further develop the model to generalise to incomplete curves. However, the pre-processing required would be extensive and this enhanced model still may not outperform a standalone Random Forest Regressor.

More advanced models such as LSTM or Transformer may be more powerful in the general context of time-series analysis and in theory better suited to this problem of variable length inputs and changing forecast horizon, but their complexity and resource demands are unlikely to be justified by performance due to the small size of the dataset.

Data Splitting Strategy

As previously mentioned, this time series data must be split chronologically for training, validation, and testing sets. Data leakage in the form of ‘look ahead’ must be avoided and it is crucial that the training set involves only birth cohorts which are complete up to and including the target variable, age at which 50% of the cohort have become mothers. NaN values are not removed in incomplete curves as they

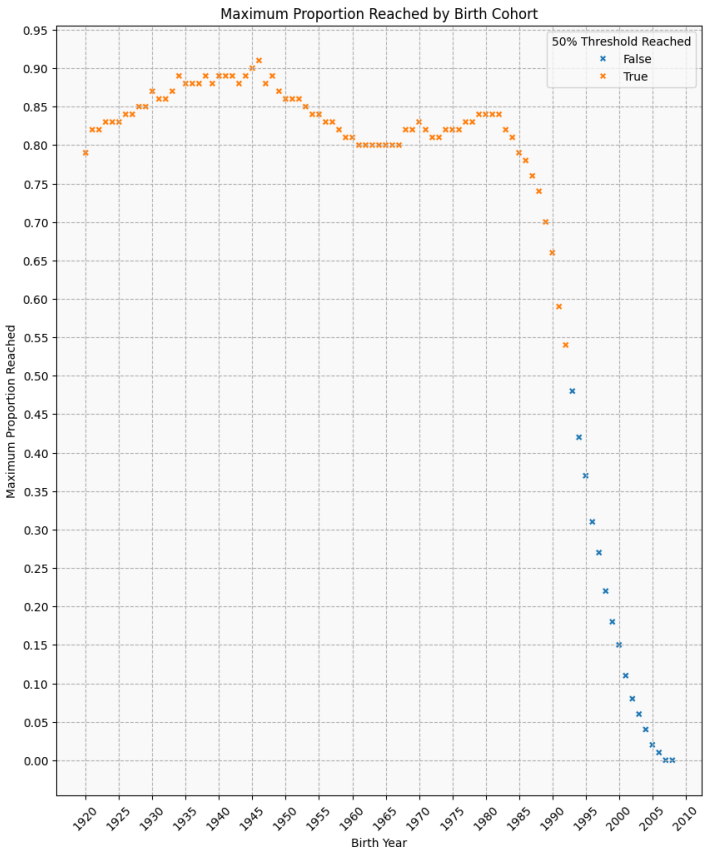
meaningfully represent the age standing of that cohort and so should not be replaced by predicted values.

A column 50_reached is added to the DataFrame to explicitly filter based on the condition that this target variable is present in the data.

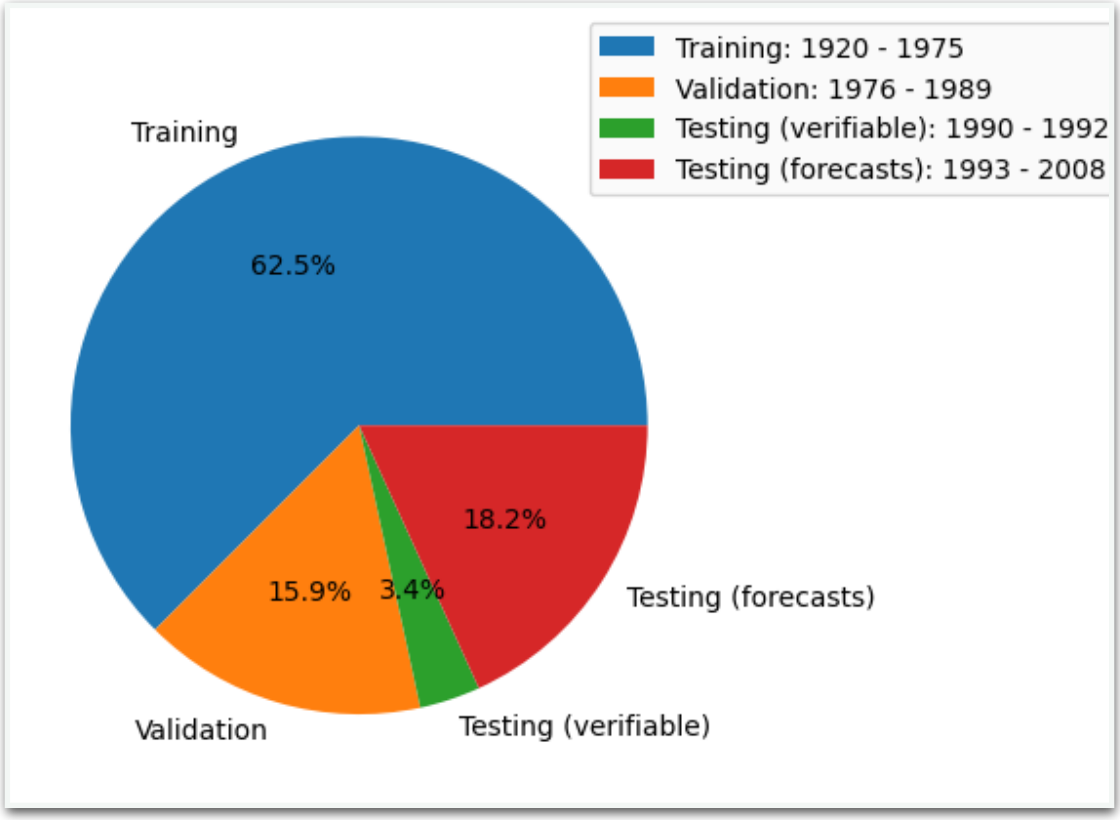
This graph shows that cohorts 1993-2008 have not yet reached the 50% threshold and so can not be used for training or validating the model, since predictions can not be evaluated by measuring against true values.

Common train:test ratios are between 50:50 and 80:20, where the training set contains validation. Given the small dataset and how the trends change in more recent years, it makes sense here to use as much data as possible for training and validation.

After some inspection, the sets are as follows.



Test	Birth Cohorts	Justification
Training	1920 - 1975	All of these curves are complete and so provide stable training data. Covering almost 65% of the available data, these birth cohorts contain enough variation for the model to learn meaningful relationships between the features and the target. Including more recent cohorts ensures that the model can observe shifts in patterns in the later cohorts.
Validation	1976 - 1989	All of these curves reach the target 50% threshold so can be used to evaluate the model's performance against true values. This set covers the late 1970s resurgence in earlier motherhood and the drop-off beginning in the mid 1980s which continues moving forward. This makes these cohorts a challenging but useful validation set to tune the model and mitigate overfitting.
Testing (verifiable)	1990 - 1992	These curves are complete and so there is still data available for final unseen evaluation of the model's prediction performance. While this subset is small, it will still allow evaluation of the model's generalisation while making the most of verifiable data for training and validation.
Testing (forecasts)	1993 - 2008	These curves are incomplete so represent genuine forecasts, mirroring real-life deployment of such models. There is still some scope to qualitatively assess these results by comparing with the ONS predictions for average family size by birth cohort (contained in Table 1b of the dataset).



Feature Engineering

The Random Forest Regressor learns from tabular, numerical input data. It is important to capture important characteristics of the curves in a format the model can interpret.

Target Variable

Age at which 50% of the birth cohort have become mothers.

Input Features

1. Year of birth

   - Captures the trend over time, allowing the model to learn how motherhood patterns shift and the target variable changes over time.

2. Age at 25% motherhood

   - Provides an early indicator of the 50%. EDA showed that the 25% and 50% curves have very similar shape.

   - A percentage threshold is used instead of a fixed age (e.g. proportion of mothers at age 25) because this would not be equally meaningful across cohorts - some earlier cohorts have exceeded 50% by age 25, introducing risk of lookahead if fixed ages were used.

3. Change in proportion (age 20-25)

   - Captures the early (16-20 less informative) rate of increase in motherhood.

- Meaningful across all cohorts, even where 50% is reached within this range (mid 1930s-1940s). High rate combined with low age at 25% helps the model to learn that early steep growth leads to lower age at 50%.

4. Change in age (25%-35%)

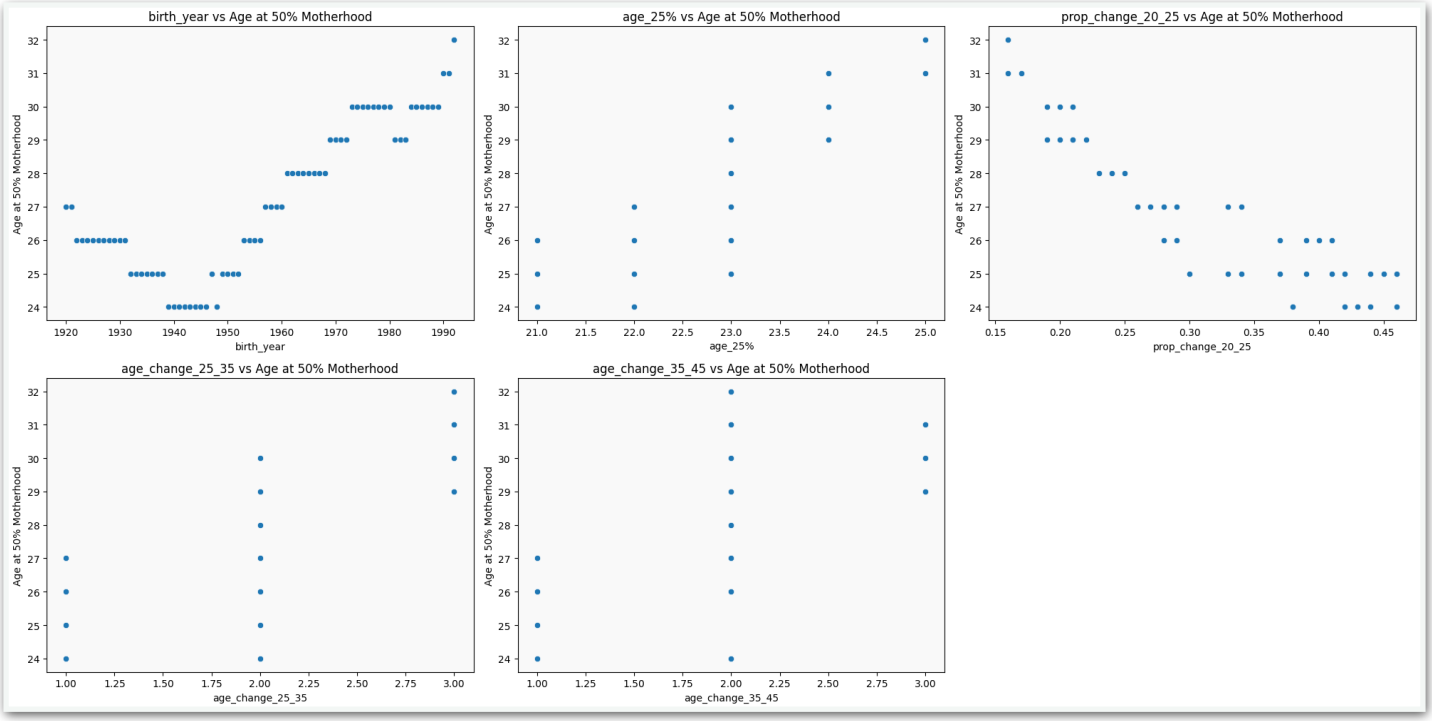
- Inverting the slope represents the time required for the proportion of mothers to increase from 25% to 35%, capturing the acceleration of motherhood uptake.

5. Change in age (35%-45%)

- Similarly, capturing the acceleration of motherhood closer towards the 50% threshold. The 50% threshold is intentionally excluded to avoid leakage of target-related data.

- These proportion based ranges have been chosen to be distinct and non-overlapping to avoid data redundancy and enable calculation in more recent and incomplete cohorts.

The following plots show the relationship between each individual feature and the target variable. While there are some clear relationships (birth year and proportion change from age 20-25) others appear more dispersed. This highlights the importance of using a model that can learn how combinations of features influence the target.



Data Integrity

All of these features are calculated within individual birth cohorts, so there is no data leakage when feature engineering is performed before splitting the data into training, validation and testing sets.

Training

The training data set contains birth cohorts 1920-1975, where the true age at 50% motherhood is known.

The chosen model is a Random Forest Regressor, which fits a number of decision trees on different sub-samples of the dataset and uses averages to improve predictions and mitigate over-fitting.

The model was trained iteratively, using the validation set to assess performance and guide improvement. The Random Forest Regressor was first trained using its default settings as a benchmark against which other configurations were compared. There are many hyper parameters which can be tuned for this type of model. This assignment focussed on a small number relevant to model complexity and generalisation:

Hyperparameter	Purpose	Default	Range (to RandomizedSearch)	Tuned
n_estimators	Number of trees in the forest. Lower values (50 - 100) can prevent over-fitting on small datasets.	100	[50, 100, 150, 200]	50
max_depth	Maximum depth of the tree. Shallower trees help reduce overfitting. Depths of 5-10 are often effective for small datasets.	None	[None, 5, 10, 20]	10
min_samples_split	Minimum number of samples needed to split an internal node. Higher values help control tree complexity.	2	[2, 4, 5]	2
min_samples_leaf	Minimum number of samples required to be at a leaf node. More data represented at leaves smooths predictions. Low range preserves tree flexibility for small datasets.	1	[1, 2, 3]	1
max_features	The number of features to consider when looking for the best split. Choosing fewer features per split introduces randomness which helps generalisation. (Limited by 5)	1	[2, 3, 5]	5

Hyperparameter	Purpose	Default	Range (to RandomizedSearch)	Tuned
bootstrap	Use random samples with replacement of the training dataset to build each tree. Bootstrap can worsen overfitting on small data, so worth investigating False to use all data for each tree.	True	[True, False]	False
random_state	Controls randomness of bootstrapping and sampling. Fixed for reproducibility, not tuned.	None		
n_jobs	Set to -1 for parallel training, not tuned.	None		

To evaluate different hyperparameter settings, TimeSeriesSplit was used to achieve cross-validation. Time-series data can not be randomly shuffled as in traditional cross-validation. TimeSeriesSplit uses increasing subsets of the training dataset, validating with the next chronological observations. This enables multiple evaluations of the model learning from the past to predict the future. This technique mimics real-world deployment and mitigates over-fitting to a particular split, while avoiding data leakage. Since this dataset is small, 4 folds were used which is typical of time-series problems.

Although the range of hyperparameters wasn't huge, combining all values across the folds requires 3456 model fits. This is too many for an exhaustive GridSearch to be practical, so RandomizedSearch was used for a more efficient investigation of possible configurations. The tuned model was then fit again to the full training dataset and its performance compared with the benchmark from the default model. Full performance diagnostics for these models are covered in the next section.

Experiments

Initially, a wide GridSearch of the RandomForestRegressor hyperparameters was used in an attempt to optimise configuration. This was computationally expensive and ultimately returned a model that performed worse than the default, likely due to overfitting. This supported switching to RandomizedSearch which was much more efficient and performed similarly.

To investigate this suggested overfitting, a smaller grid focused on max_depth and min_samples_leaf was evaluated using 4-fold and 3-fold. It was hoped that these complexity controls and increasing the amount of data available in each fold might improve the model's generalisation. In both cases, the default model was returned as best performing, further suggesting that it is already well-calibrated to this dataset.

Gradient Boosting was also briefly explored as an alternative ensemble method. While it showed lower MAE and RMSE, functional accuracy collapsed. Visual inspection revealed that the model was predicting a single value for the target, indicating underfitting. A subsequent RandomizedSearch of more trees and lower learning rates failed to improve this, and further tuning was not feasible within the time constraints.

These experiments are included (commented out) in the code submission for reference, though not incorporated into the final pipeline due to redundant performance.

Validation

The validation set covers birth cohorts 1976 - 1989, where the age at 50% motherhood is known and so can be compared with predictions to measure the model's performance.

Evaluation Metrics

Since this is a regression problem, Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) are appropriate validation metrics. These are interpretable as they maintain the input units: MAE of 3 means the prediction is out by 3 years. MAE is the main metric, with RMSE a supplementary measure to emphasise any large errors.

To reveal the workings of the model, ._feature_importances was used to show which features have the most bearing on the predictions made.

Error Analysis

Residuals measure how far the prediction is from the true value in the validation set. In this project, residuals are defined as prediction - true value, hence a positive residual indicates an overestimate by the model and negative an underestimate. Examining residuals illuminates how and where the model is under or overestimating.

A plot of predicted vs true values is used to visualise these differences and the distribution of residuals plotted to ensure the mean is zero indicating that the model is unbiased. (Note: the mean of residuals is different from MAE as equal over and underestimates cancel one another out, so this mean can be close to zero even where errors are large.)

To investigate whether the model tends towards particular erroneous predictions, residuals are plotted against predicted values to expose any pattern which would suggest that the model is missing some signal in the data.

In time series data it is particularly important to investigate whether the residuals vary over time. A further plot of residuals against birth year shows whether the model performs poorly for particular birth cohorts, indicating that the model may be failing to learn underlying trends. Any particular pattern would suggest that there is useful information left in the residuals which could be used to improve the model's predictions.

The ONS data gives the age in complete years, whereas the model predicts continuous values. This means that a prediction of 29.8 years by the model would be functionally equivalent to age 29 in the source data, even though it would be measured by the above metrics as incorrect with a residual of 0.8. To account for this, an accuracy measure compares the floored prediction values with the true values as integers. This gives an interpretable picture of how well the model predicts the age in complete years.

Model Evaluation & Comparison

As a benchmark, the Random Forest Regressor with default settings was fit to the training data and used to predict on the validation set. As described earlier, the Random Forest Regressor model was then tuned using RandomizedSearch to find optimal hyperparameter configuration, using TimeSeriesSplit with 4 folds as the cross-validation method. Training and validation was then repeated.

Evaluation metrics for predictions by each model are as follows:

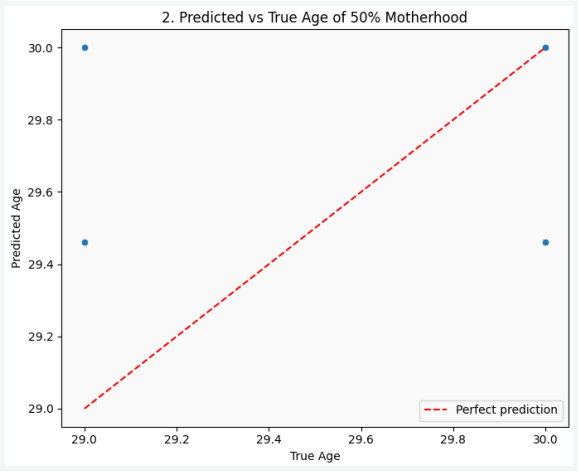
Model	Mean Absolute Error	Root Mean Square Error	Accuracy Score
Tuned Random Forest	0.283	0.46	64.286
Default Random Forest	0.291	0.47	64.286

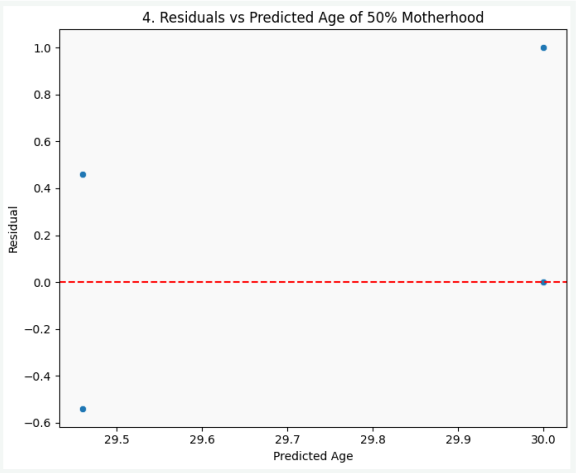
The default model was only slightly outperformed by the tuned model in both error metrics, with equal accuracy representing 9/14 correct predictions. The following two pages show greater detail on the errors of each model.

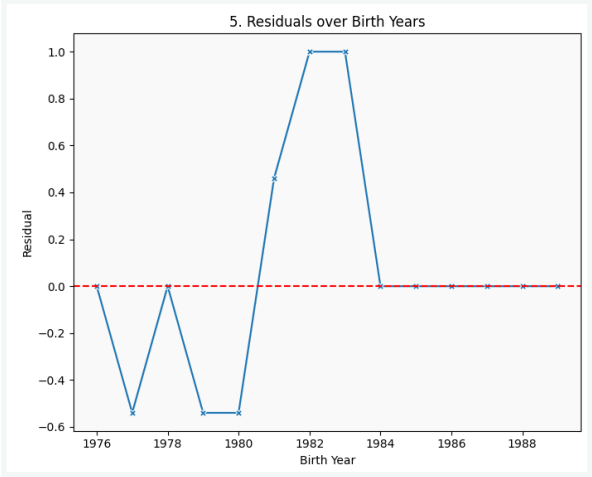
Default Random Forest Diagnostics (Validation)

The points in graphs 2 and 4 are not too far from the perfect prediction lines, indicating that the model's predictions as continuous values are reasonably close to the true values.

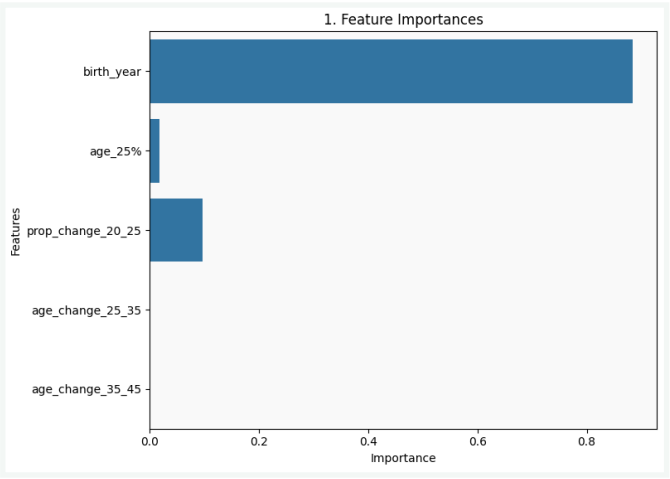
Graph 2 reveals that the model overestimates true ages of 29 and underestimates ages of 30, suggesting that there is some signal in the data it may be missing. The scatter in **graph 4** is relatively flat, implying consistent error magnitude across the predictions.



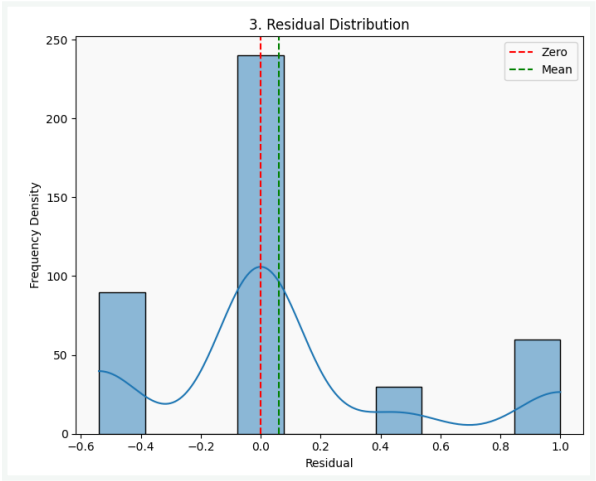




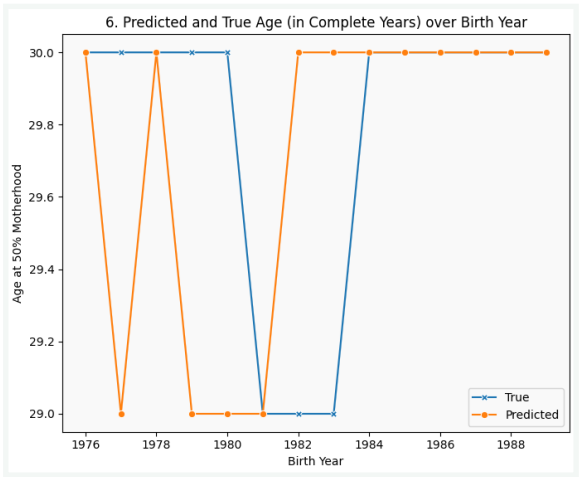
Graph 1 reveals that birth year dominates the model's decision about the age at which a cohort reaches 50% motherhood. The change in proportion from age 20 to 25 makes a much smaller contribution with the other features largely neglected. This may present a problem for the more recent cohorts as the higher age at 25% motherhood and slower adoption of motherhood represented by the age change features are the signals that age at 50% motherhood should be higher, but the model is not paying attention to these features.



This is reflected in the distribution of the residuals shown in **graph 3**, as the mean is close to zero indicating an unbiased model. The tails suggest misses are more often small underestimations with the fewer overestimations being slightly larger errors.

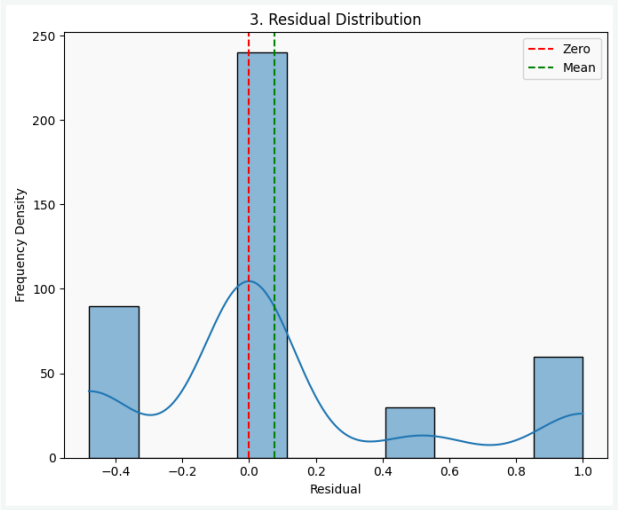


Graph 5 shows that the model underestimates 1977-80 and then overestimates 1981-3, before becoming a perfect predictor. **Graph 6** represents the model's predictions as ages in complete years, since functionally 29.4 means age 29. The same pattern is observed here, with just 1981 rounding to a correct prediction. This emphasises how small continuous errors in the model's prediction lead to misclassification in real-world deployment.



Given that the range of observed ages at 50% motherhood is only 8 years (24-32), the mis-estimation of a full year at 12.5% of the range is significant. This suggests that the model is not correctly interpreting how the interaction between features influences the target, perhaps due to the imbalanced importance of features in its decision making.

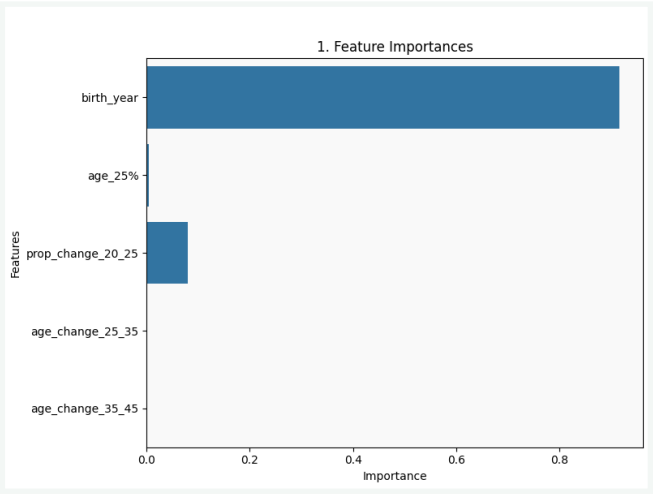
Tuned Random Forest Diagnostics (Validation)



Graph 3 also shows minimal difference in the residual distribution, reflecting that this tuned performs almost identically well.

As indicated by the near enough equal evaluation metrics, the tuned Random Forest Regressor model is very similar to the default, with only marginal changes in structure and prediction success.

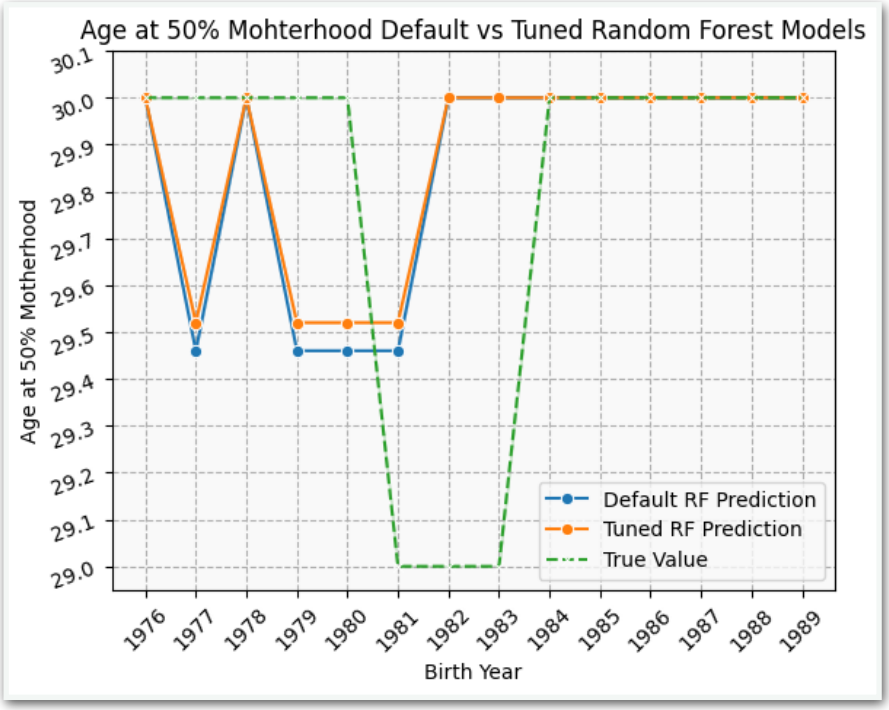
Graph 1 shows almost identical feature importances for this model, with birth year still dominating the decisions and age at 25% motherhood in fact having even less influence than before.



Random Forest Comparison

This visualisation reveals that the tuned model is only slightly improved predictions in four instances, hence the marginal decrease in MAE and RMSE. This improvement is only apparent when interpreting the predictions as continuous values though, as in real terms when the ages are viewed in complete years, both models perform equally well.

Despite the minimal difference, the tuned model was selected for testing as it still represents the best performing configuration identified.



Testing

The test set covers birth cohorts 1990 - 2008. However the age at 50% motherhood is only known for the first three cohorts (1990-2), making them the only verifiable cases. As the best performer in the validation set, the tuned RandomForestRegressor was used for all test predictions.

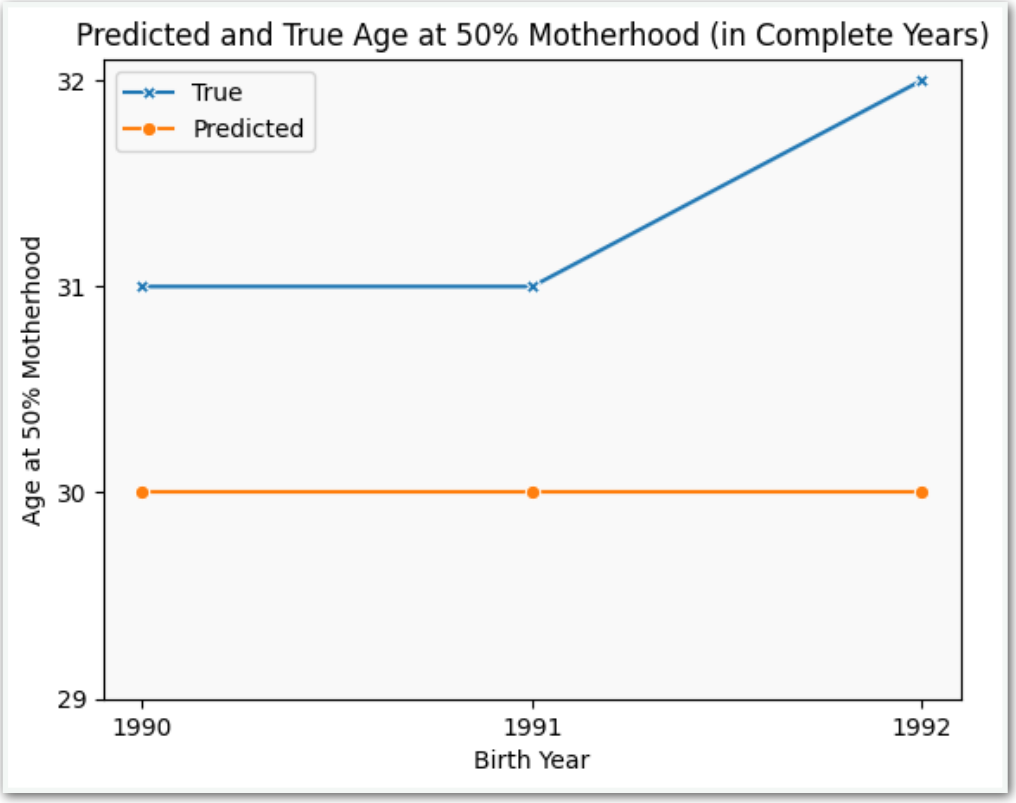
Verifiable Evaluation

Compared with the true values for 1990-2, the model performed poorly:

MAE: 1.33

RMSE: 1.41

Functional accuracy: 0%



In all three cases, the model under-predicted the age at 50% motherhood, returning a flat prediction of age 30.

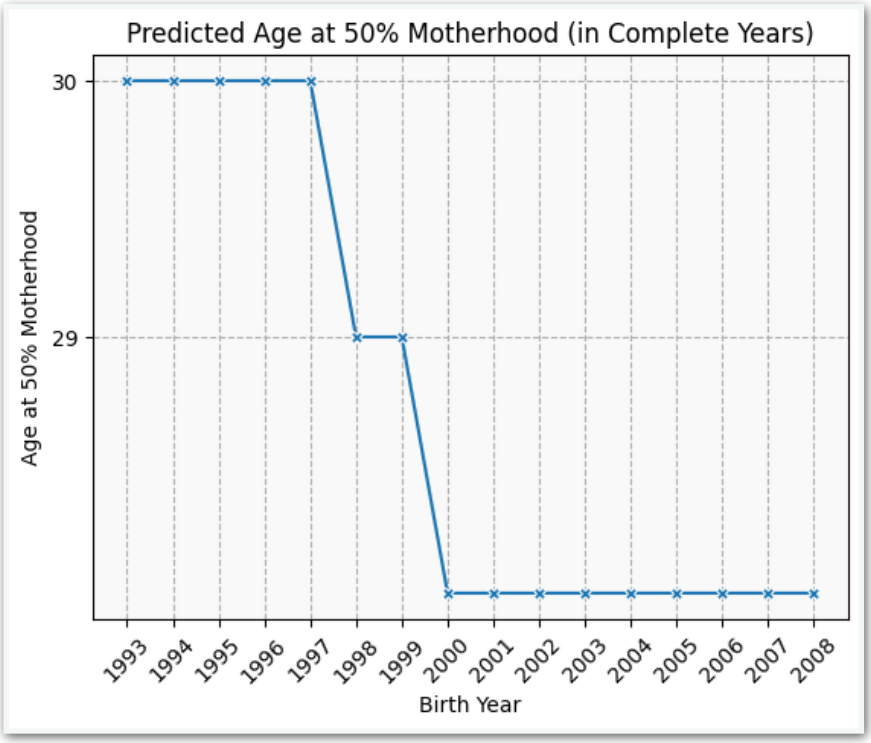
This indicates that, though the model performed reasonably well on the validation set, it has not generalised well to newer cohorts. This may be due to

the continued and increasing upward trend in motherhood age which was evident in EDA but which was not yet present in the earlier cohorts included in the training data. The model has not been exposed to cohorts older than 30 at 50% motherhood which may explain its failure to extrapolate outside this historical range. These results may also suggest that a more expressive feature set, incorporating more curve dynamics, may be needed to help the model detect changing trends in recent cohorts.

Forecast Evaluation

For the remaining test cohorts 1993-2008, predictions represent genuine forecasts, as these birth cohorts have not yet reached 50% motherhood.

Given that the model did not accurately predict on the verifiable test data, it was anticipated that predictions here may be implausible, which is demonstrably the case.



This visualisation of the model's predictions for 1993-2008 shows a decreasing age at 50% motherhood from 30 to 28. This contradicts previous trends and what we already know, for instance, the 1993 & 4 cohorts have already turned 30 without reaching 50% motherhood. The closest verifiable cohort (1992) were 32 at 50% motherhood and trends revealed in EDA suggest that cohorts should be getting older when reaching this threshold, yet the model is predicting even younger ages.

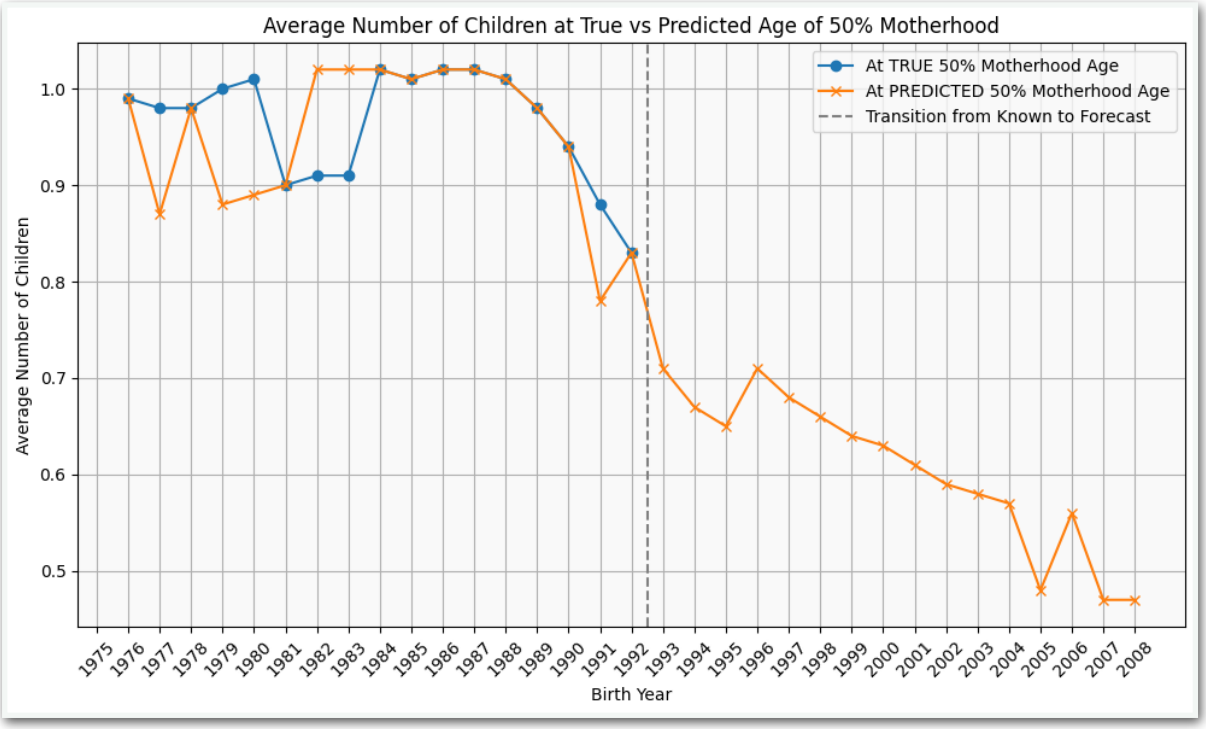
The model's failure to generalise with these more recent test cohorts may reflect several limitations:

- struggling to predict a changing forecast horizon based on incomplete curves,
- Random Forest is not well suited to extrapolating beyond past seen values, and
- the model's lack of focus on early indicators: age at 25% motherhood and change in age as proportion increases from 25% to 35%. The model is not interpreting these clues present in more recent cohorts as it is focussed predominantly on birth year.

The model may be misinterpreting the unfamiliar behaviour of recent cohorts as resembling the earlier historical patterns of younger motherhood. This highlights the risk of using a tree-based model for evolving time series data.

Comparison with ONS Projections

The implausibility of these predictions can be further corroborated by comparison with the ONS predictions. The ONS do not directly provide predictions for the age at 50% motherhood, but offer predictions for the average number of live born children. These two variables are strongly linked, since the average number of children will increase roughly in line with the proportion of mothers, allowing for some variability due to some mothers having large numbers of children. These predictions therefore provide a useful proxy.



The above graph shows, by birth cohort, the average number of children at the true and predicted 50% motherhood age.

This comparison highlights three key observations:

- The predicted values show reasonable alignment to the true values for the validation set 1976-89 and are certainly within a sensible range, supporting that the model learnt some meaningful relationships.
- The slight decline in average number of children at the 50% motherhood age 1989-92 might reflect diverging social expectations: fewer women are having children early, but those that do possibly having more children.
- Crucially, the average number of children at the 50% motherhood age remains above 0.8 for all true values. The drop below 0.7 and beyond predicted by the model is implausible and even illogical, since the average number of children at 50% motherhood should be at least 0.5 (as there must be at least one child born to every new mother).

Together, these testing results confirm that the model is failing to generalise to the evolving patterns in more recent cohorts and should not be relied upon for future forecasting in its current form.

Conclusion

Goal

This project aimed to predict the age at which 50% of a birth cohort of women have become mothers, using data spanning 1920-2008. This target was selected to measure whether a cultural feeling that motherhood is being delayed is reflected in real data.

Model Choice

The time series dataset posed several challenges: it was small, non-stationary and involved a changing forecast horizon. A Random Forest Regressor was chosen due to its ability to model non-linear relationships and robustness on small datasets. The default model was compared with a model tuned using RandomizedSearch and TimeSeriesSplit. Following evaluation using MAE, RMSE and a functional accuracy measure, the tuned model was selected for

testing. Gradient Boosting was briefly explored but discounted due to lower practical accuracy.

Model Performance

On the validation set, the tuned model achieved an MAE of 0.28 and 64% functional accuracy (9/14 correct predictions in complete years). While this validation performance was promising, results on the verifiable test set were entirely inaccurate. The model's predictions for the forecasts then contradicted observed trends and ONS projections.

This suggests that the model learned relationships from training data well enough to align with reality for the validation set but failed to recognise the signals for the evolving behaviour of more recent cohorts. Consequently, the model is not a reliable forecaster for birth cohorts beyond 1989.

Strengths of the Approach

Critique of available data led to selection of an appropriate and reliable source. Exploratory analysis exposed trends within and across birth cohorts, leading to insights for how different rates of change influence the age at 50% motherhood. This led to feature selection which captured these rates numerically, compensating for the model's inability to see curve shapes. Avoiding data leakage was prioritised in the data splitting strategy, feature engineering and use of TimeSeriesSplit. While the model underperformed in testing, visual diagnostics and comparison with reliable predictions enabled credible evaluation of its limitations.

Limitations of the Approach

The key limitation stems from the Random Forest model's inability to extrapolate beyond the observed range. Due to the chronological nature of this small dataset, the model had not been exposed to cohorts with higher ages at 50% motherhood. The model did not assign sufficient importance to the features that might serve as early indicators to signal the shifting trends in recent cohorts. The variability and influence of these features was less apparent in the earlier training cohorts, perhaps accounting for the model's lack of focus on them.

Proposed Improvements

Given that this data will grow slowly, one cohort each year, it will be a long time before recent cohorts, which appear to be diverging from historical patterns, can be used in training. An alternative modelling approach would therefore require improved extrapolation capacity, for instance polynomial regression. However, overfitting remains a risk and so the model may not perform well predicting shifts not present in training data.

A more promising avenue may involve learning from the shape of the curves themselves. Such methods require fixed length inputs meaning truncated complete curves, again excluding recent cohorts from training. A CNN could be used to interpret curve shapes (e.g. age 16-30) extracting embedded representations of shape dynamics. These could then be passed to a polynomial regression model for forecasting. This hybrid approach may allow the model to interpret and generalise curve behaviour while extrapolating predictions beyond the training range.

Whilst social factors are difficult to quantify, the dataset could be enhanced by including variables such as the proportion of women in the cohorts with different levels of academic qualifications and income. Such factors may influence the timing of first motherhood and provide valuable signals for changing behaviour, helping the model to better interpret and forecast trends in more recent birth cohorts.

Final Reflection

While the model ultimately failed as a forecasting tool, the process exposed many challenges in modelling time-evolving demographic data. Meaningful patterns have been observed and evaluated, with a more advanced strategy proposed to better model them.

Research Sources

https://preset.io/blog/time-series-forecasting-a-complete-guide/ - Time series forecasting.

https://otexts.com/fpp3/ - Working with time-series data including cross validation, train test splitting and residual diagnostics.

https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.TimeSeriesSplit.html - Train test split for time series data.

https://towardsdatascience.com/random-forest-regression-5f605132d19d/ - Random forest regression models.

https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html - Random forest regressor

https://medium.com/@byanalytixlabs/random-forest-regression-how-it-helps-in-predictive-analytics-01c31897c1d4 - Random forest regression in predictive analysis.

https://towardsdatascience.com/the-reasonable-effectiveness-of-deep-learning-for-time-series-forecasting-60e2c8affb9/ - Deep learning techniques for time series forecasting.

https://www.mathworks.com/discovery/lstm.html - Long short term memory

https://journalofcloudcomputing.springeropen.com/articles/10.1186/s13677-023-00576-7/tables/2 - Advanced methods for time series forecasting for non-stationary series (beyond me at this stage).

https://neptune.ai/blog/select-model-for-time-series-prediction-task - Model selection for time-series tasks.

https://github.com/qianlima-lab/time-series-ptms/blob/master/README.md - Pre-trained models for time-series.

https://medium.com/@hassaanidrees7/gradient-boosting-vs-random-forest-which-ensemble-method-should-you-use-9f2ee294d9c6 - Random Forest v Gradient Boosting