**MIS S381N Final Report**

1. **Project Goals**

Our project examines the rapid decline in the U.S. newborn population (1.6 children per woman in 2024), seeking to identify the root causes by leveraging machine learning algorithms - logistic and extreme gradient boosting. By consulting the US Census for Social, Economic, and Health Research database, we aim to uncover key characteristics of the underlying issue and answer the following problem statement: **How can we overcome and prevent the U.S. birth rate from falling below the replacement level necessary to sustain long-term population stability?**

We chose this problem since the declining U.S. birth rate has severe implications that hinders a country’s long-term stability. A continuous drop below the replacement level (2.1 children per woman) hinders economic growth, social security infrastructures, and balances in the workforce; Hence, our work is relevant not only to policymakers, but also to corporations, healthcare providers, and numerous entities that depend on human resources. By identifying the key drivers of fertility decisions, our analysis offers actionable insights for designing policies and support programs that encourage family growth, ultimately benefiting the nation.

1. **Data Collection**

### **1. Data Collection: Strategic Sourcing**

Our primary data source is the **Integrated Public Use Microdata Series** (**IPUMS**), a renowned resource for high-quality, individual-level US census data. We selected IPUMS for its well-documented and comprehensive nature, which uniquely allowed us to create a custom dataset by choosing specific variables. This approach was more efficient than downloading and merging multiple datasets. To further validate and enrich our findings, we also considered supporting sources like the **National Center for Health Statistics** (**NCHS**) for fertility data and the **Pew Research Center** for broader social insights.

### **2. Initial Data Cleaning and Filtering**

The raw IPUMS dataset began with a substantial **3,405,809** records *(Refer* ***Fig.******1.1*** *and* ***1.2*** *in appendix)*. To ensure a focused and relevant analysis, we performed several essential cleaning and filtering steps. Initially, we removed duplicate records to prevent redundancy. We then filtered the dataset to include only the relevant demographic: women of childbearing potential (ages between **15 and 55 years old**). Finally, we addressed and removed any invalid or incorrect values to ensure data integrity. Overall, our dataset was reduced to a sample of **763,296** records *(Refer* ***Fig. 1.3*** *in appendix)*.

### **3. Feature Engineering: Preparing the Data**

Feature engineering was a critical process of transforming for our machine learning models. The majority of this process consisted of converting existing string-based features into numerical or binary formats. For instance, the *GAVE\_BIRTH\_12MO* feature was converted to a binary variable (1 for 'Yes', 0 for 'No'), and text descriptions for *EDUCATION* were mapped to a numerical *education\_years*. The only feature that required a specific calculation was *years\_since\_marriage (2023 - YEAR\_LAST\_MARRIED)*. We also created dummy variables for our other categorical data (*has\_insurance, employed, gave\_birth\_12mo*). Ultimately, we were able to generate 11 distinct features for our analyses.

### **4. Key Libraries and Tools**

Our data collection and cleaning process was powered by multiple suites of Python libraries, leveraging Pandas extensively for all data manipulation tasks. The process of converting raw data into a model-ready format was facilitated by Patsy, which helped in creating design matrices. For our downstream analysis and modeling, we utilized Scikit-learn and XGBoost, with Matplotlib playing a key role in visualizing our findings.

1. **Exploratory Analysis**

The data used in our analysis consists of 763,296 women in the US, based on the latest research conducted in 2023. Approximately 30% of the women have 2 or more children, 19% having one child, and over half have no children *(Refer* ***1.4*** *in appendix)*. This class imbalance should be considered when interpreting the output of our models, as it may influence performance and bias towards the majority class. Furthermore, we find that married women tend to have significantly more children than any other type of marital status *(Refer* ***1.5*** *in* *appendix)*. This suggests that marital status may be a strong predictor in determining the possibility of bearing multiple children, which could play a key role in our modeling feature selection.

We also noticed that when grouped by education years, there is no group of women who have more than 2.1 children on average *(Refer* ***1.6*** *in appendix)*. We find that women with less education have more children compared to that of women with higher educational standing. According to our data, women with exactly 10 years of education have the lowest average number of children. This group, however, is a statistical outlier, merely representing 2.5% of the entire dataset. In contrast, women in higher education usually fall under the 1 child category - possibly due to career prioritization, later marriage, and greater access to family planning resources.

Interestingly, our analysis found no significant differences in birth rate across different household income brackets or geographical regions within the U.S. This suggests that, at least within the scope of our dataset, economic and regional contexts do not influence the number of children women have.

1. **Solutions and Insights**

Our solution aims to answer the research question: **How can we predict whether a woman will have two or more children, and which demographic and socio-economic factors drive this decision?** Understanding these drivers can help inform policies and initiatives aimed at encouraging higher birth rates to support population replacement goals.

The analysis begins with defining the target variable, *above\_2child*, as a binary indicator representing whether a woman has two or more children. Then, we engineer a set of features that we believe can capture relevant aspects of a woman’s demographic and life history that might affect childbearing behavior. Key engineered features included: *ever\_married*, *years\_since\_marriage*, *education\_years*, and *gave\_birth\_12mo*. We combine these features with others like age, household income, employment, and insurance status.

To classify children status, we test two main models. We first fit a **Logistic Regression** model, as we hope to see interpretable coefficients with direction and magnitude of each factor’s association with having two or more children. We set the maximum number of iterations to 10,000, observing that the model converged around the 7,000th iteration. We also then train an **XGBoost Classifier** to capture potential nonlinear relationships and complex interactions among variables. Through cross-validation, we identify an optimal tree depth of 4 and select 200 trees (n\_estimators) to balance model complexity and performance.

The logistic regression model achieves a solid training and testing accuracy of approximately 71%. The XGBoost model outperforms logistic regression, achieving a training and **test accuracy of 78%**, showing its ability to grasp the underlying patterns without overfitting the data. Both models achieve a substantial lift over the baseline accuracy of 30%. From the confusion matrix, the false positive rate is only about 13%, indicating that the model is highly effective at correctly identifying women with fewer than two children. Since our policy priority is to target support toward this group, maintaining a low false positive rate is critical to avoid misallocation of resources.

Figure 1.7 shows the feature importance of Logistic Regression vis-a-vis XGBoost classifier. We found that some factors were particularly important like whether the **woman has ever gotten married *(ever\_married)***, and whether she has **given birth in the past year *(gave\_birth\_12mo)***. The odds of having two or more children **increases 13 times** (e**2.6** = 13.5) when a woman is married, and **3 times** (e**1.04** = 2.8) when she has given birth in the past 12 months. XGBoost result also corroborates the finding from Logistic Regression. *ever\_married*, and *gave\_birth\_12mo* are among the top three features that XGBoost uses for making predictions. The other finding was factors like number of marriages, employment, and number of years of education decrease the odds of having a second child by 40%, 22%, and 4% respectively.

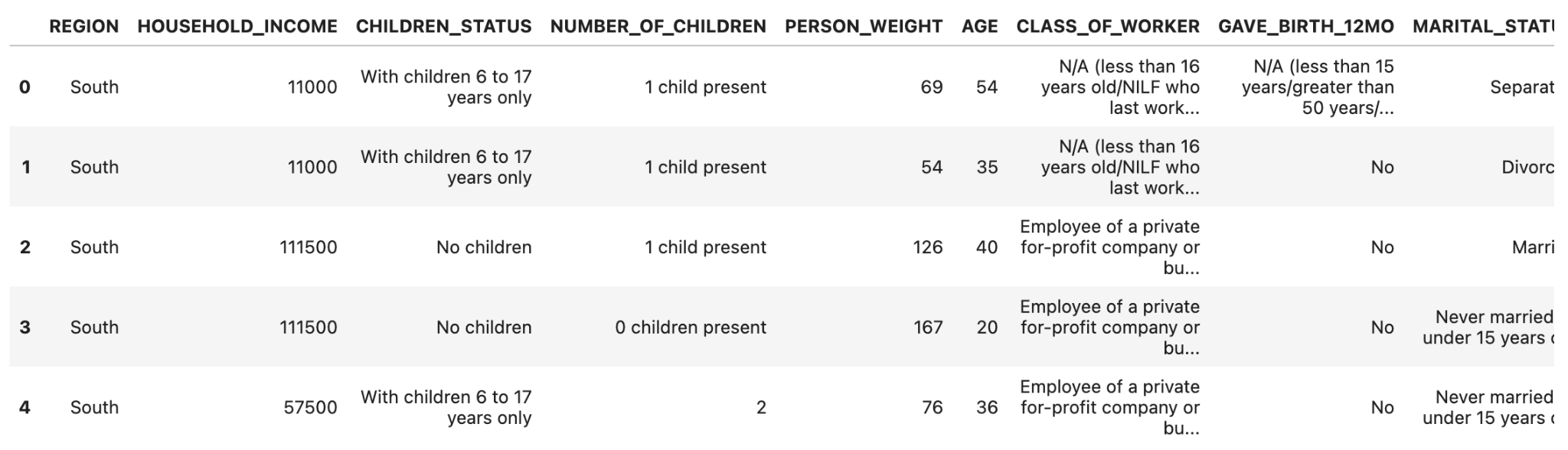
We believe fertility rate can be increased if the government provides a robust support system for childbirth and child rearing both. The purpose is relieving the cost burden of women of childbearing potential. The measures like sharing childbirth costs and easy access to trained staff for postnatal care can be helpful. Moreover, initiatives like free day-care for households having 2+ children can be an incentive to bearing more children, as proven in other countries. In respect to private sectors, corporations should come up with win-win policies such as guaranteed full wage compensation during maternity leave/paternity leaves, and welcome back programs for women upon return to the workforce.

1. **Conclusion**

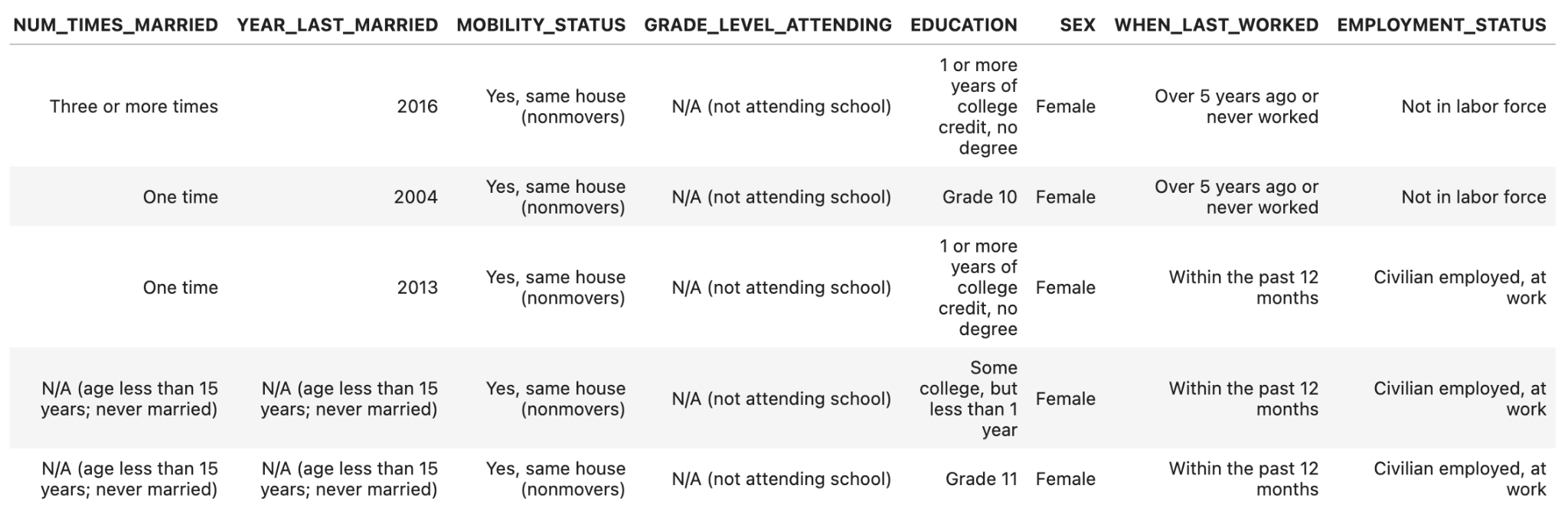
The primary objective of our project was to explore the rapid decline in the U.S. birth rate, which is now at 1.6 children per woman in 2024. We aim to find the key factors influencing whether a woman has two or more children, the threshold needed to sustain long-term stability. Using IPUMS data, we were able to predict with 78% accuracy whether a woman would have two or more children, and the demographic and socio-economic factors strongly linked to fertility outcomes. Accordingly, we recommended a few initiatives that could help reverse the downward trend in the fertility rate and keep us on the path to attain the target of 2.1 children per woman required for maintaining the population level in the long run.

1. **Appendix**

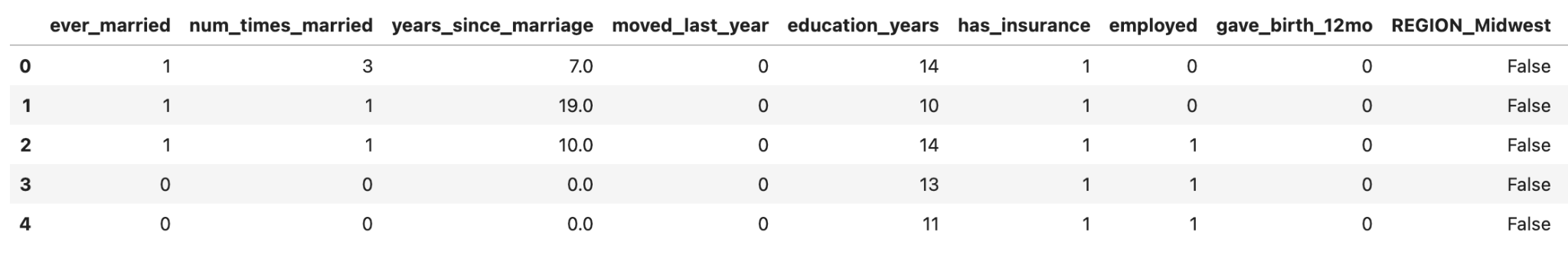
**Fig. 1.1 Raw Data collected - 1**

****

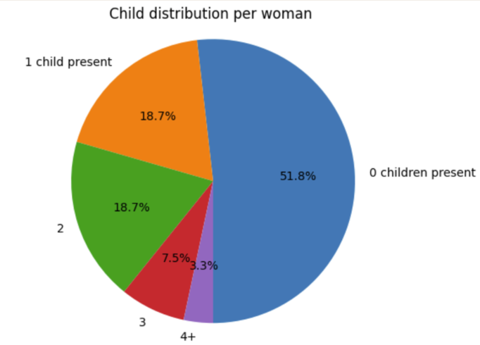
**Fig. 1.2 Raw Data collected - 2**

****

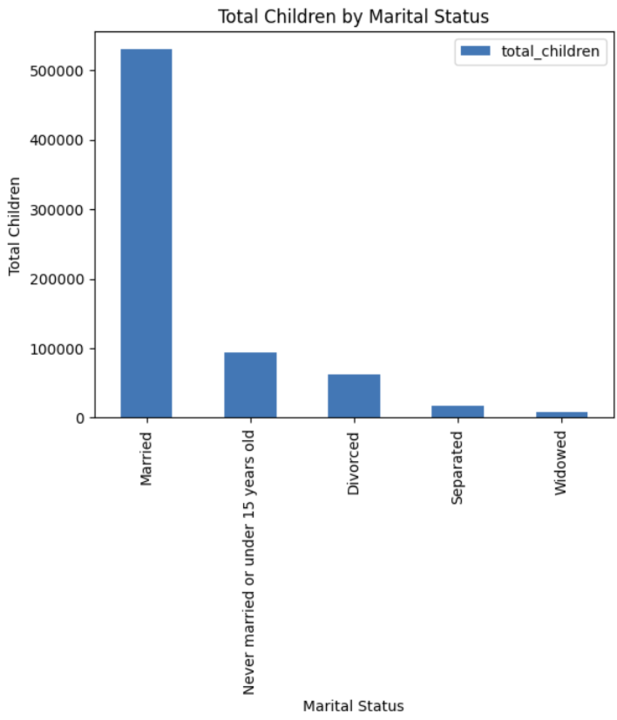
**Fig. 1.3 Data Feature Engineering**



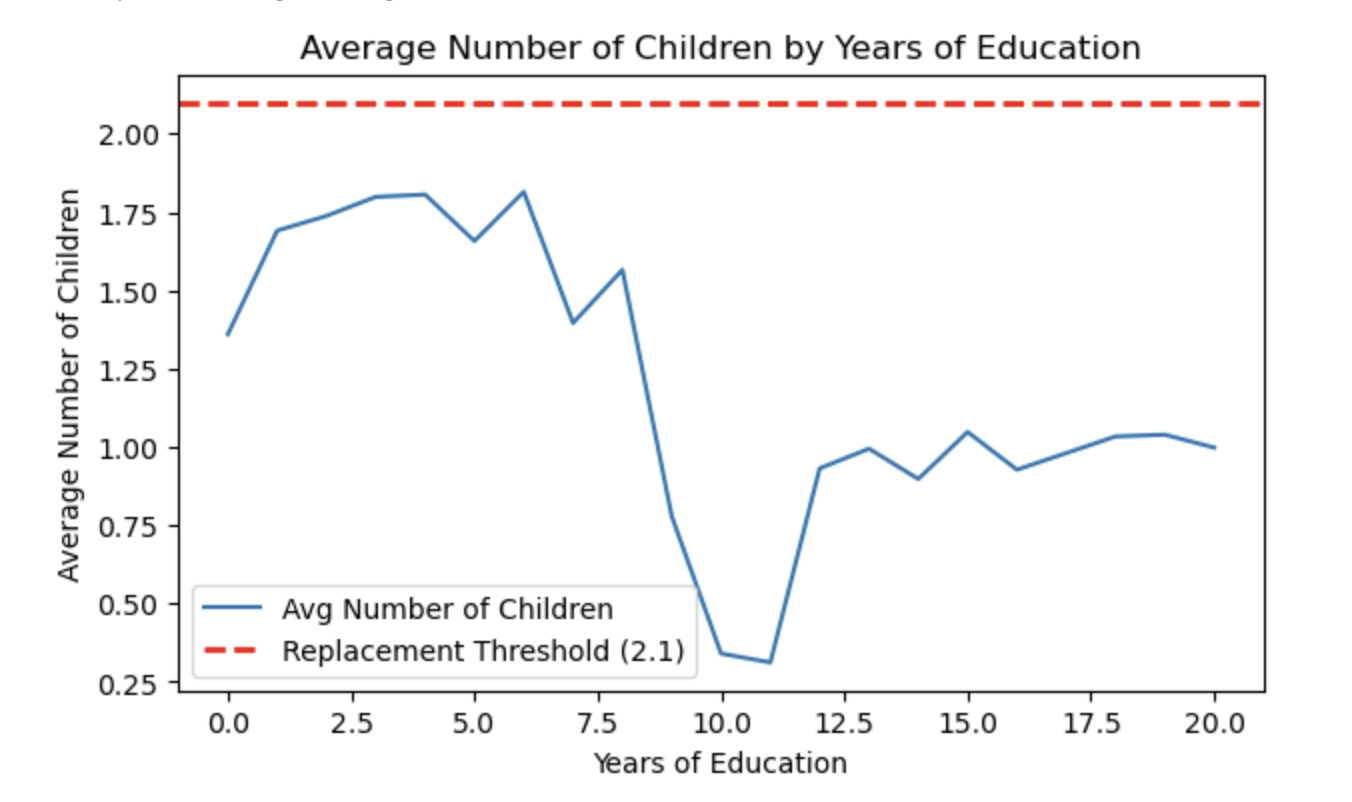
**Fig. 1.4: Exploratory Analysis**



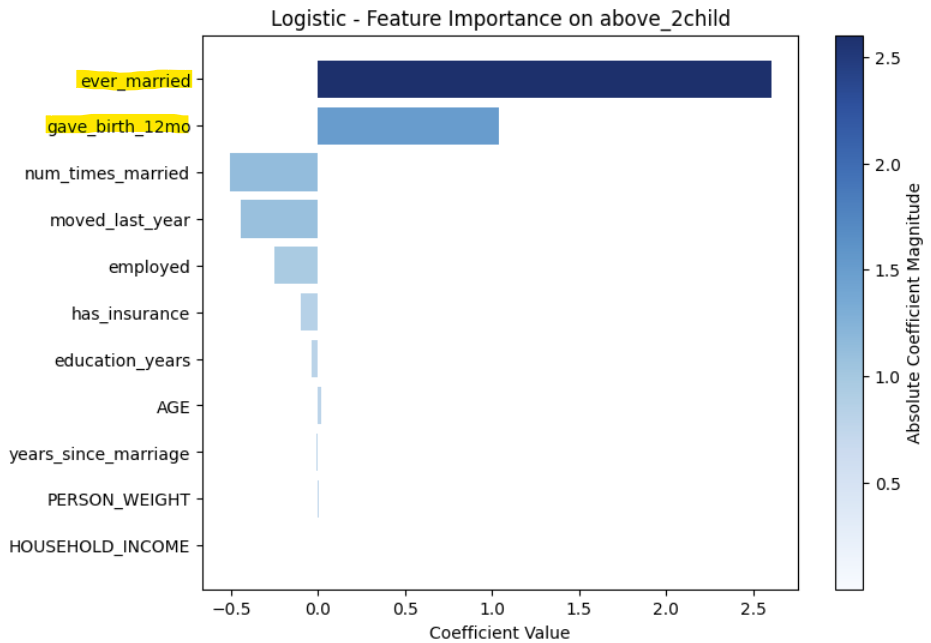
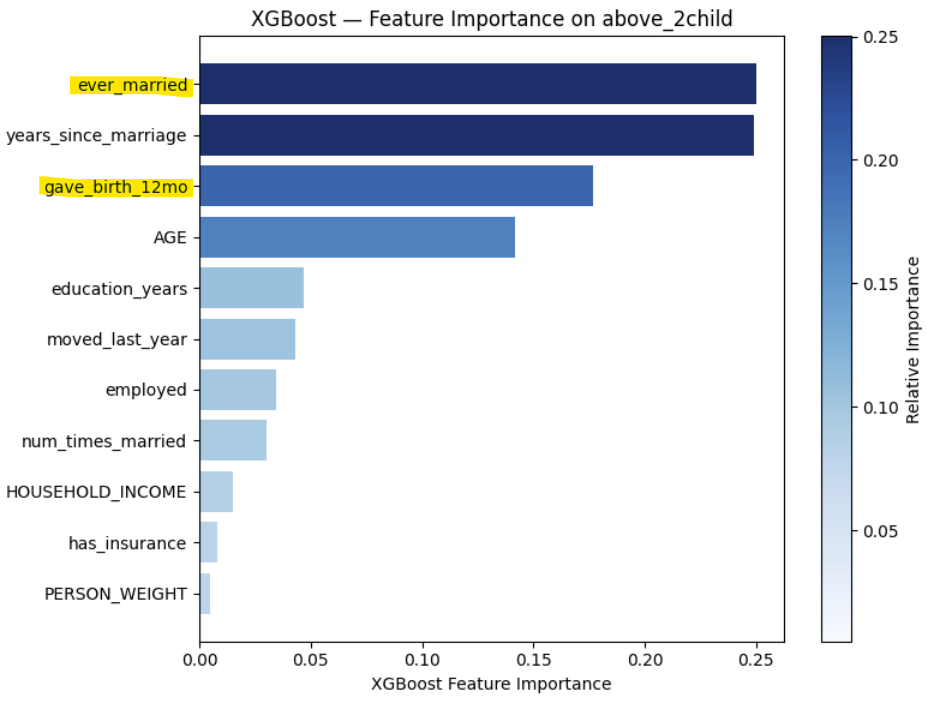
**Fig. 1.5 Exploratory Analysis**

****

**Fig. 1.6 Exploratory Analysis**

****

**Figure 1.7: Feature Importance Logistic Regression vs. XGBoost**

1. **References**

“U.S. Census Data for Social, Economic, and Health Research.” *IPUMS USA*, usa.ipums.org/usa/. Accessed 10 Aug. 2025.