Analysis on Korean low birth rate, based on 2019 survey

Tianyu Wang, Somin Lee, ChenxinZhu 2024-04-29

```
knitr::opts_chunk$set(echo = FALSE, message = FALSE, warning = FALSE)
knitr::opts_chunk$set(fig.width=3.6, fig.height=3)

suppressPackageStartupMessages({
  library(readxl)
library(dplyr)
library(ggplot2)
library(caret)
library(randomForest)
library(e1071)
library(factoextra)
library(readr)
library(rpart)
library(gbm)
library(readr)
library(osmdata)
library(tidyverse)
library(tidyr)
library(rsample)
library(ggplot2)
library(pROC)
})
```

Abstract

The research question is around: why Korean people don't give birth? The report builds birth predicting models to look closely into the factors that affect the low birth rate in South Korea. It focus on the life of young people who are under 44 by using data from the national survey about time usage and lifestyle, because this can offer clues to understand why young people hesitate to have kids and what makes their life different from before when they don't have kids. It looks at meaningful variables to affect birth rate using heat maps and neural network analysis and interprets the variables in terms of work, housing, and social-related based on the recent social policy and reality of Korea. Then the report make GBM, linear, and CART models to predict whether the households have kids or not and enhance the model by selecting variables.

I. Introduction

1.1 Background and Motivation

"Korea is so screwed! Wow, I've never heard of that low fertility rate, 0.78(%)" Joan C. Wiliams, Professor of Law at UC Law San Francisco said in an interview by EBS(Korea Educational Broadcasting System).

East Asian countries such as China, Japan, and Taiwan show a lower birth rate (also known as fertility rate) than most other countries. The birth rate has decreased in most developed countries over the years, but the recent plunge in some Asian countries has attracted attention. Above all, the birth rate in South Korea hit

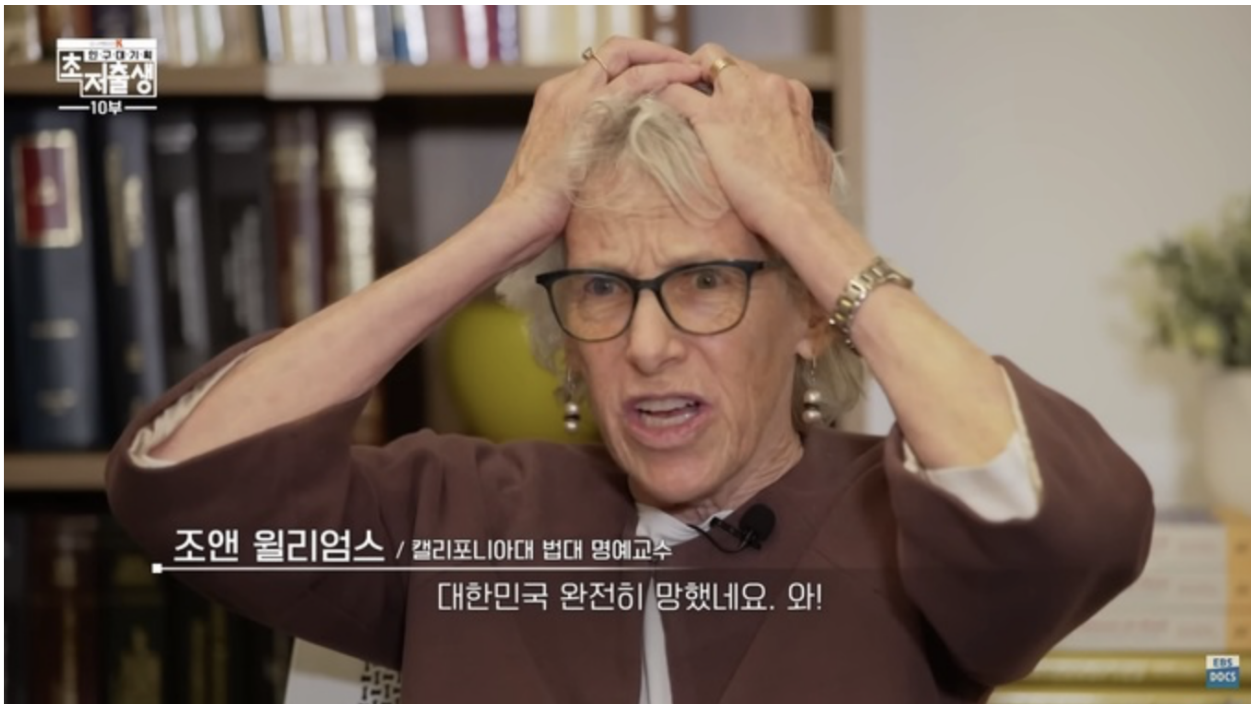


Figure 1: Alt text

the lowest 0.78% in 2021; it is even less than half of the US birth rate, 1.66%. It was quite shocking not only to Koreans but also to Joan C. Wiliams, who is a professor of law at UC Law San Francisco, and her response in disbelief to the birth rate has gone viral in Korea.

Due to the shocking number, the Korean government has tried to raise the birth rate, but it is only the beginning step, such as giving a subsidy. It should be identified why people don't give birth to successfully raise the birth rate, and that's the reason why we need to take a close look at the real lives of parents and non-parents.

1.2 Research question

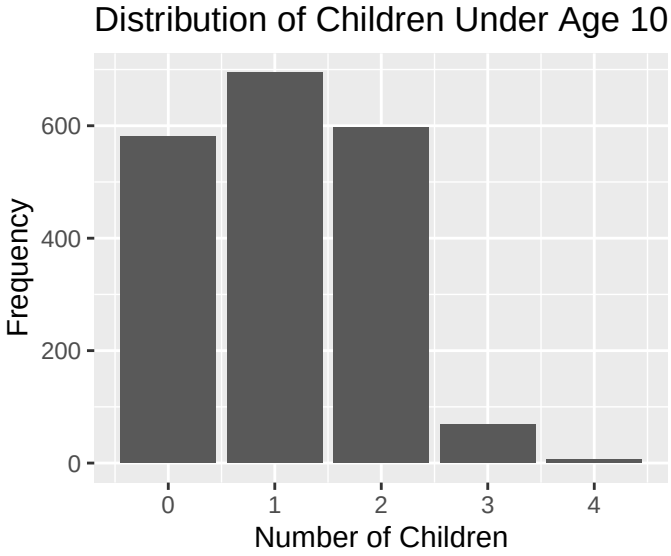
There are many surveys to identify why current young people don't have a baby. However, just asking why you do not have kids is so direct and limited that we can not catch the important reasons between the answers.

That's why we identify what is different between families with kids and without kids by selecting meaningful variables and making birth rate predicting models using data about lifestyle. Even though the models are designed to predict, their explanatory variables represent the differences between the exogenous families who have babies or not, and that could be a good starting point for making policies to improve the birth rate.

II. Method

2.1 Data set description

The data from the Time Use Survey in South Korea, which is for understanding people's lifestyle and quality of life by measuring how people spend their time, is used in this analysis. It was conducted by Statistic Korea, a government agency, in 2019, and microdata is also provided by the agency on request. We use data from married couples only, and the age of respondents is restricted under age 44 because we consider families with kids under 10, and the average age of women giving birth is 33.



2.3 Methods for Data Processing

2.3.1 Data Acquisition and Preprocessing

Handling of missing data was carried out, with separate strategies for numerical and categorical variables. Missing values were treated as meaningful, suggesting that the absence of data might itself be informative. Categorical variables were encoded as binary variables, facilitating the use of these predictors in modeling. This step involved converting categorical variables indicating yes/no responses to binary (1/0) representations.

2.3.2 Variable Selection

We first use several heatmaps and neural networks to do variable selection, which can visually demonstrate the correlation between different variables in the dataset, helping to identify which factors are most closely associated with birth rates.

```
With heatmap, we set a threshold for significant correlation: A threshold value of 0.1, which is the cu
```

2.3.3 Model Building and Evaluation

After variable selection, use different models including CART, GBM, and linear model to see which model has the better performance on the testing data, evaluating and comparing by RMSE.

Also, ROC curves and AUC statistics were generated for the models, providing a measure of their discriminative ability for the classification task.

2.3.4 Model Selection & Predictions

Based on performance metrics, the best models were selected for making final predictions, and see how each prediction behaves. The selected models were used to make final predictions on the test data. Predicted probabilities and classes were visualized using histograms, density plots, and bar charts to understand the model's performance.

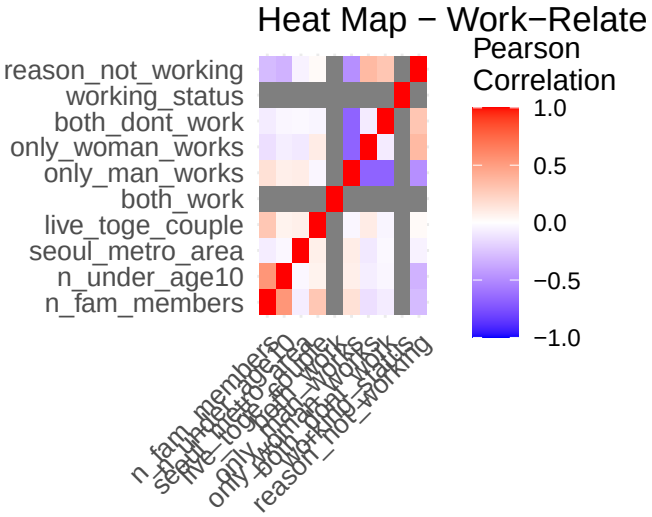
III. Results

3.1 Heatmap for Variable Selection

```
Heatmaps can visually demonstrate the correlation between different variables in the dataset, helping t
```

3.1.1 A. Work-related

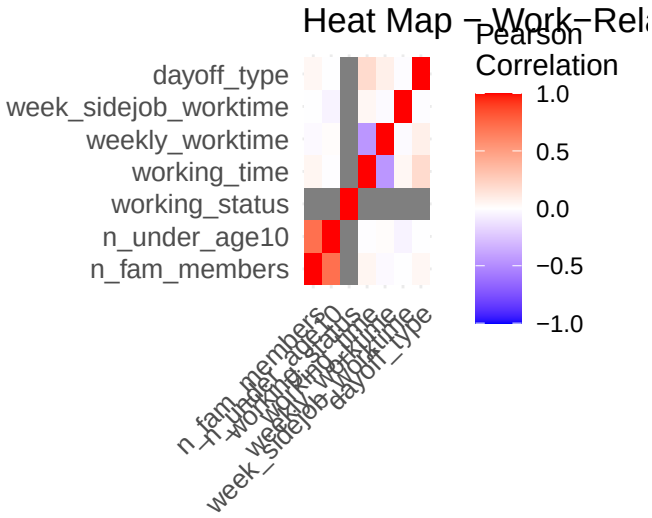
In the work-related heatmap, certain variables exhibited strong positive correlations, particularly between the different work status categories, suggesting a clear differentiation in the employment types within our dataset.



B. Work-related: Comparison between working people Additionally, we analyze only the people whose working status is yes. When it comes to working time, we can divide it into two parts.

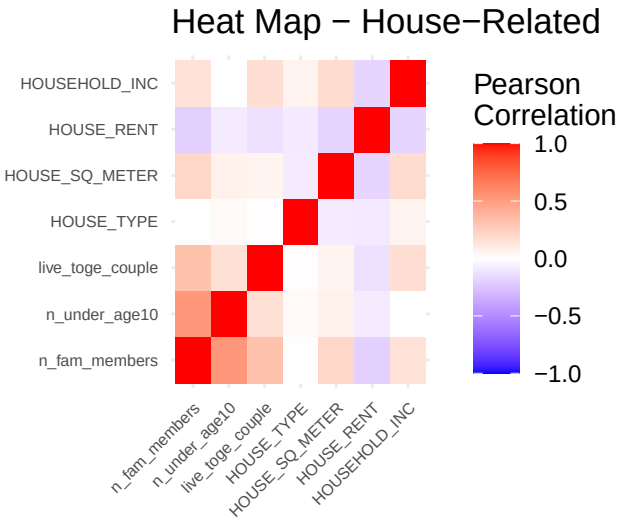
First, working_time(1:full time, 2:part time) represents the stability of a job. The labor market of Korea is not as elastic as the US, so shifting from a part-time to a full-time job is difficult, so a full-time job means a more stable source of income than a part-time job and people having a full-time job are more likely to have kids.

Second, working and side job work time show different correlations with having kids. Longer working time usually leads to more labor income so it is positively correlated. However, a side job means they are demanded to work more so it can be interpreted as instability of economic status. So, we think side job working hours have a negative correlation.



3.1.2 Housing-related

The house-related heatmap displayed a varied correlation landscape, with household income showing a notable correlation with house size, indicating an expected relationship between income and living space. Just live_toge_couple has some correlations.



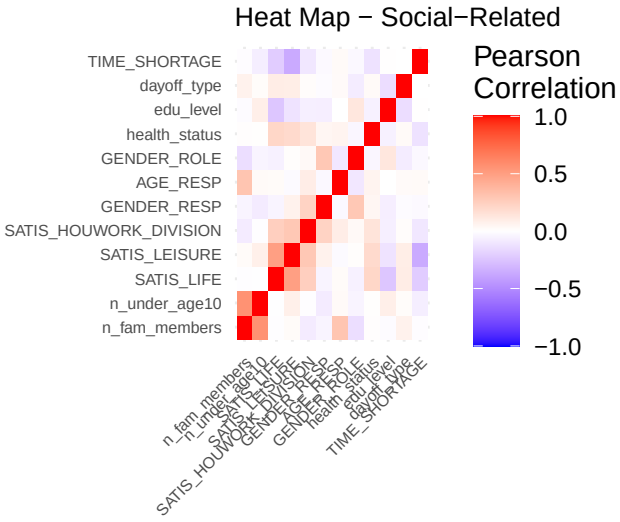
In terms of households with kids, household income shows a negative correlation with having kids, and it can be regarded as the opposite of the positive correlation with working hours. However, income includes not only labor income but also other non-labor income, so economic affluence does not necessarily guarantee having more babies.

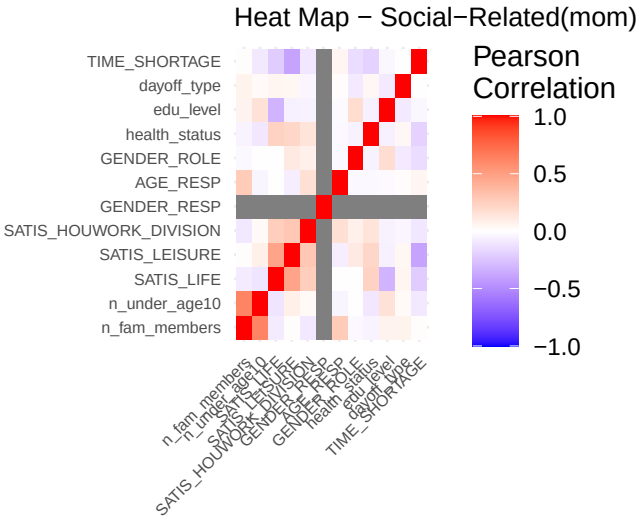
House rent (type) also gives us an interesting point. The bigger the number of house rents (code), the less stability in housing; for example, "house rent =1" represents owning a house, but as the number gets bigger, it becomes renting a house with a lower deposit and a higher monthly rent.

In Korea, rent is differentiated by the amount of deposit, and people prefer to lower their monthly rent by paying a larger deposit. Because it means they have enough money to pay a large amount of money that is closely related to the ability and credit of the tenant. Therefore, the negative correlation between house rent and the number of kids shows the importance of housing stability for the birth rate.

3.1.3 Social-related

The heatmap concerning subjective life satisfaction & others highlighted some intriguing relationships, such as a significant correlation between education level and life satisfaction, potentially alluding to the broader impact of education on the perceived quality of life. Conversely, the heatmaps also unveiled areas with minimal or negative correlations, guiding further inquiry into factors that might contribute to these inverse relationships.



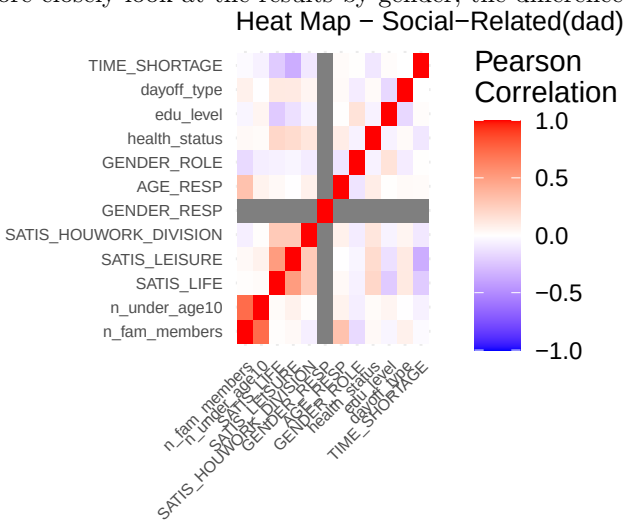


We figure out three variables: education level, leisure, and time shortage. First, education level shows a negative correlation with kids, so it breaks the stereotype that the more educated, the less likely you are to have kids.

The satisfaction of leisure is positively correlated with kids, and it seems counterintuitive. Even though the definition of leisure could be a little ambiguous, parents with kids have higher satisfaction with leisure time. It represents that having kids does not necessarily deteriorate parents' leisure time, they can enjoy family leisure together.

Lastly, time shortages are negatively correlated with having kids, it means that parents feel their time is less abundant than a married couple without kids. It shows that looking after kids is such a time-consuming and labor-intensive job.

Then, when we more closely look at the results by gender, the difference is detected in the satisfaction of life and gender roles.



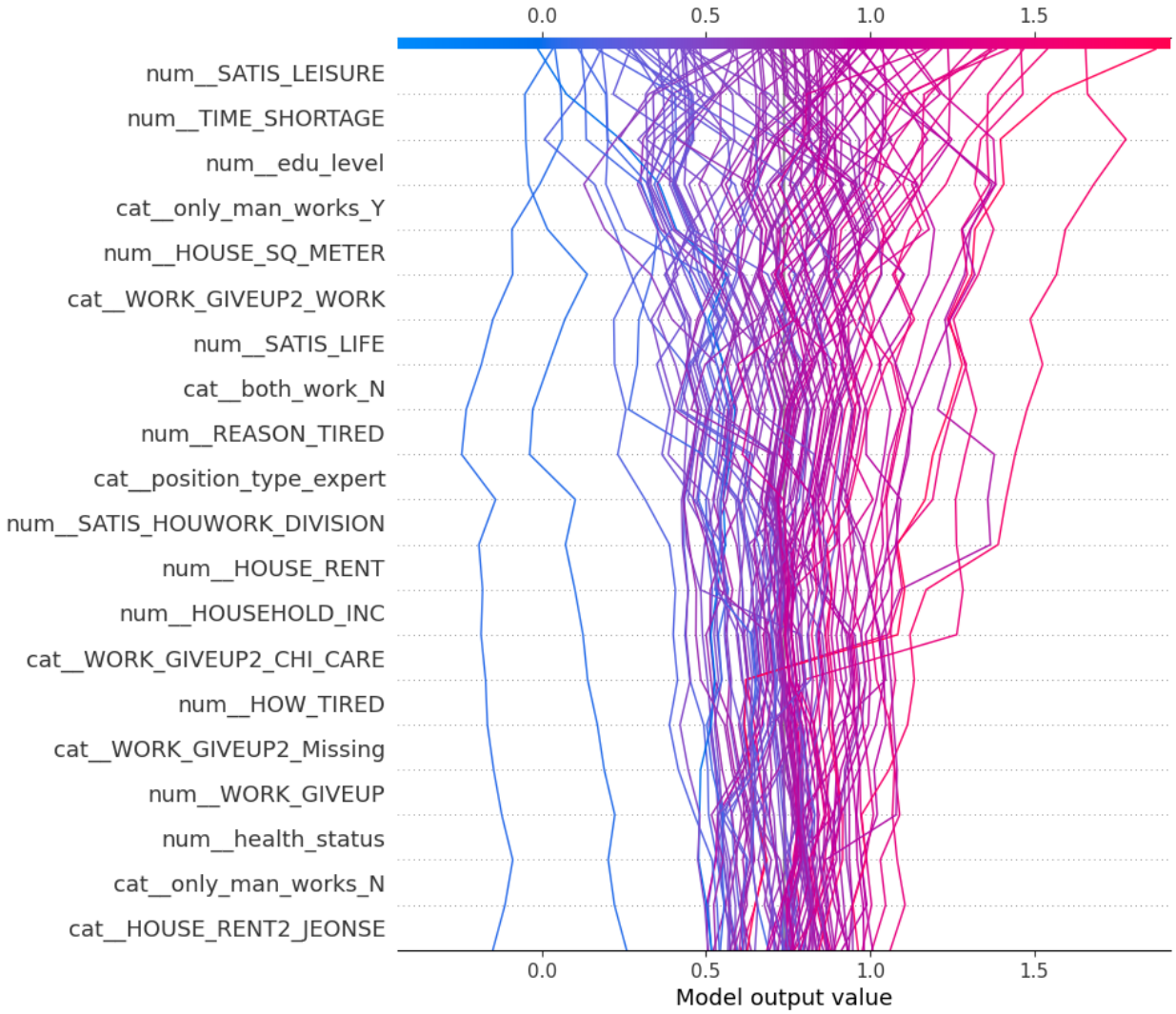
Fathers' satisfaction with life seems neutral to kids, close to zero, but that of mothers increases as they have kids (the smaller the number, the more satisfied with life). It shows women's life satisfaction is more positively related to their kids than men's, and when men disagree on the gender role, which reflects the traditional role allocation such as "man works outside, woman cares for home," the married couple is more likely to have children.

Overall, these visualizations provided a groundwork for identifying variables that might influence each other and warranted further examination in our subsequent analyses. The next step is to set a threshold and exclude some factors that have little or no influence on birth rates.

3.2 Threshold Setting & Filtering for Variable Selection

`With heatmap and threshold filtering: For the work-related category, the selected variable is "reason_n`

3.3 Variable Selection by Neutral Network



In this section, we use neural networks as a means of filtering variables. We manually eliminated some of the variables that we thought would be noisy, such as region and reason_not_working, to reduce the probability of overfitting the model. We built a three-layer neural network, the first layer acquires the dataset and outputs 128 features, the second layer reduces the number of features from 128 to 64, and the third layer outputs a single value. The reason why we only built a simple three-layer neural network is the too narrow dataset. We chose ReLU as the activation function.

We then use the SHAP Decision plot as a visual interpretation of the model. We used the neural network to fit 100 times and calculate the SHAP value of each prediction separately, then averaged to hedge the randomness of the model output caused by the small data set, and finally used the decision plot to show the degree of influence of the variable on the output result.

From the top to bottom, we could see that SATIS_LEISURE(Satisfaction to leisure life), TIME_SHROTAGE(busy or not, 5 levels), and edu_level(education level) have greater effects compared with the other variables.

3.3 Model Comparison

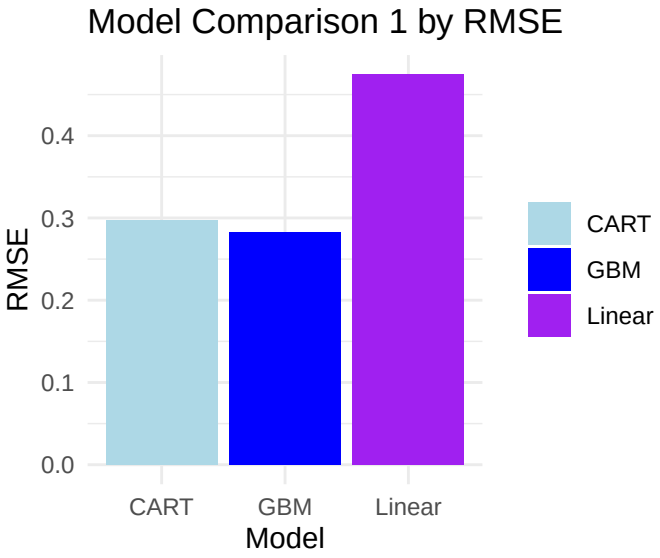
3.3.1 RMSE Enhancement

After variable selection, the LM model is the best model with the lowest RMSE. The RMSE does not show much improvement, which might be because of overfitting reduction. Without variable selection, models may have access to more information and could potentially overfit the data, especially complex models like GBM (Gradient Boosting Machine). Overfitting occurs when the model learns the noise in the training data instead of the actual signal, leading to lower performance on unseen data. Variable selection helps by removing irrelevant features, which may lead to a better generalization and, thus, better performance on the test set. Next step is to do model selection under different criteria.

```
## [1] "CART RMSE: 0.297063613495298"
```

```
## [1] "Gradient Boosting RMSE: 0.28270041592142"
```

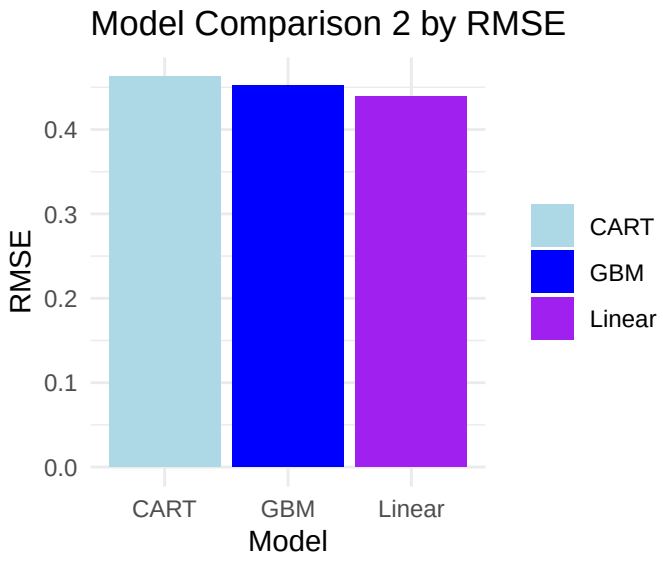
```
## [1] "LM RMSE: 0.474323956427037"
```




```
## [1] "CART RMSE: 0.462336642541877"
```

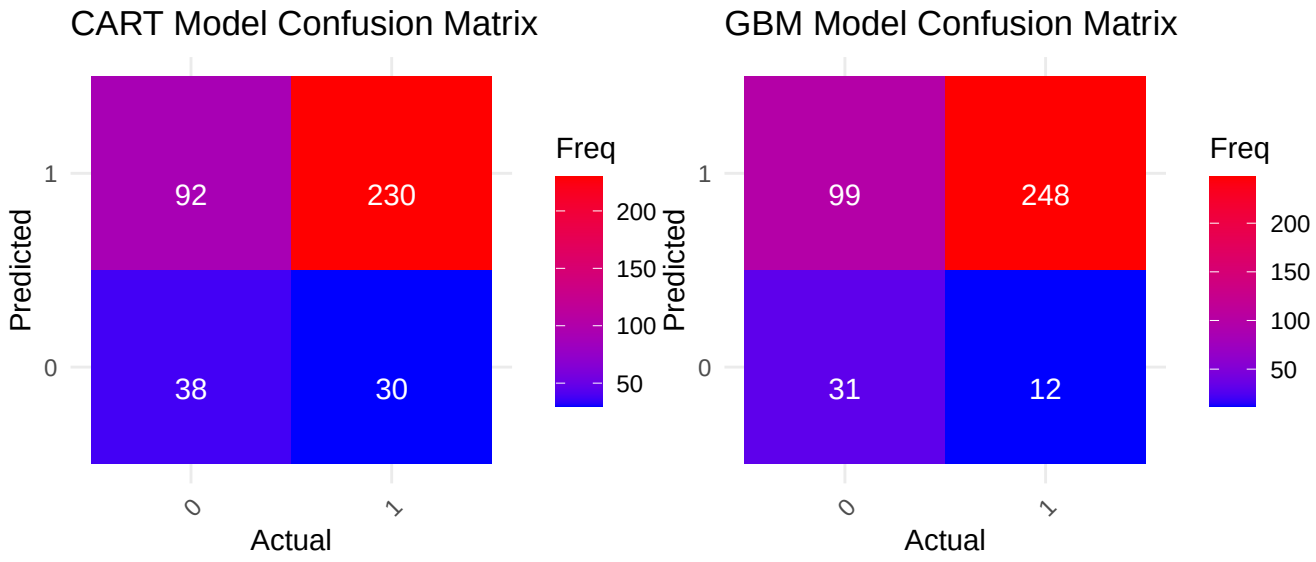
```
## [1] "Gradient Boosting RMSE: 0.452119118881464"
```

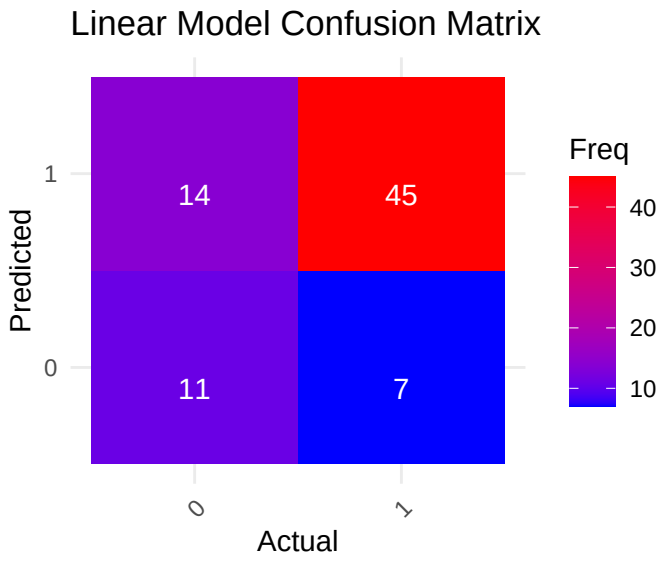
```
## [1] "LM RMSE: 0.438872492917845"
```



3.3.2 Confusion Matrix analysis

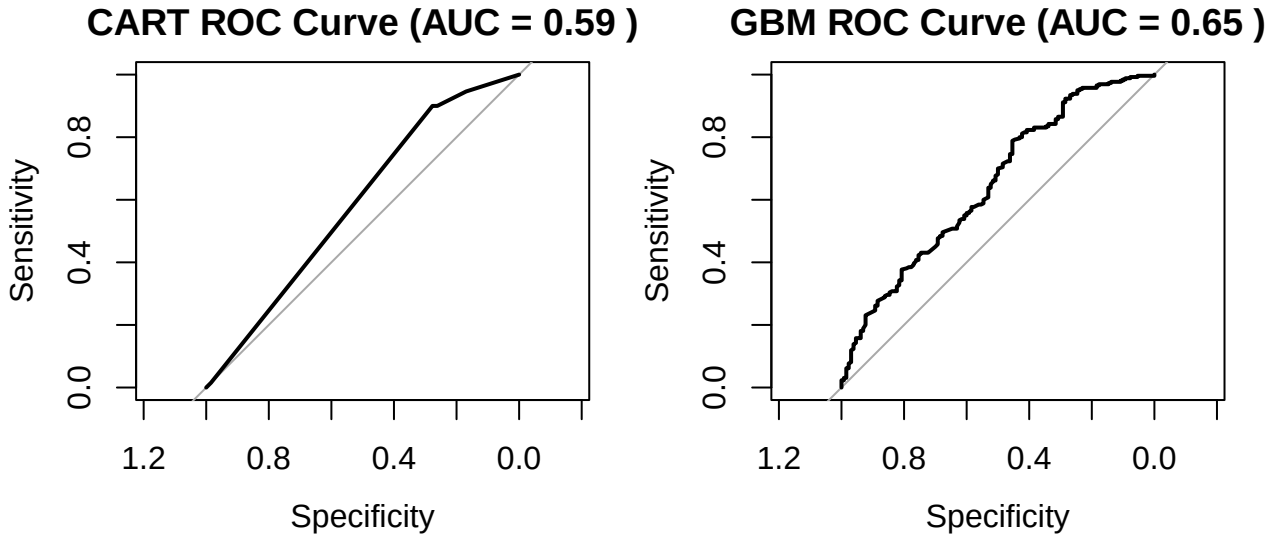
From these matrices, the CART model seems to strike a better balance between identifying both classes correctly compared to the other models, but it also has a high FP rate. The high FP and FN across all models suggest there may be challenges with the models' classification abilities or with the inherent difficulty of the task.

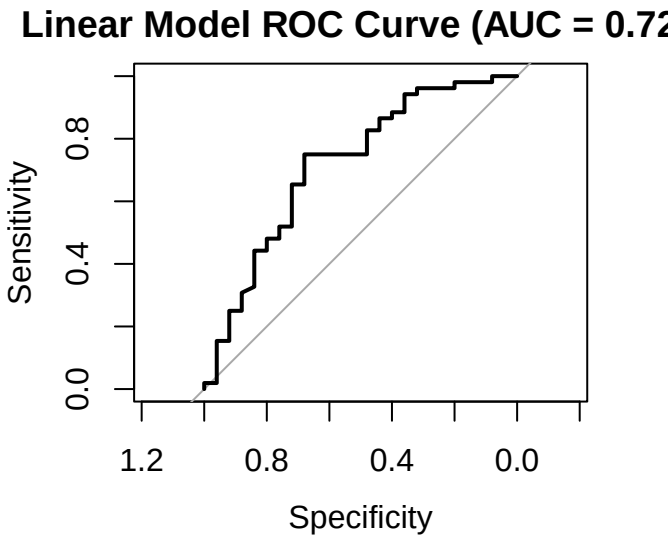




3.3.3 AUC-ROC curves analysis

Overall, the Linear Model is likely the best choice among the three given its higher AUC score and its higher predictive accuracy as reflected in the confusion matrix. The Linear Model stands out with the highest AUC score, which indicates a superior ability to distinguish between the positive and negative classes compared to the other two models. Despite having fewer predictions, those it makes are more likely to be correct.





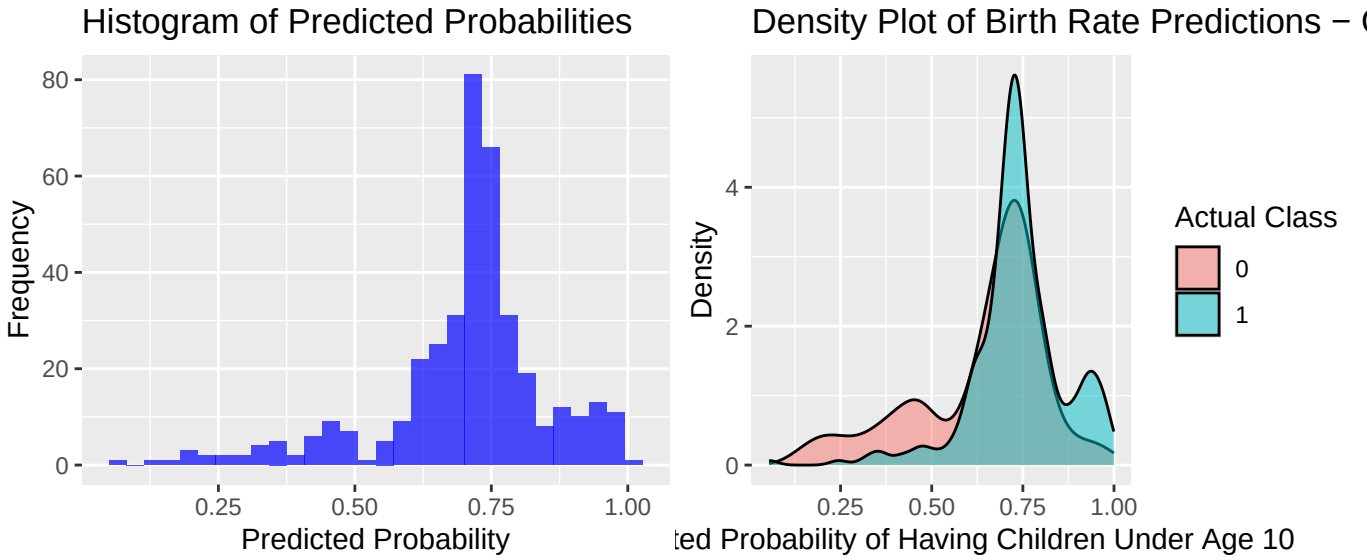
IV. Using selected models to do prediction

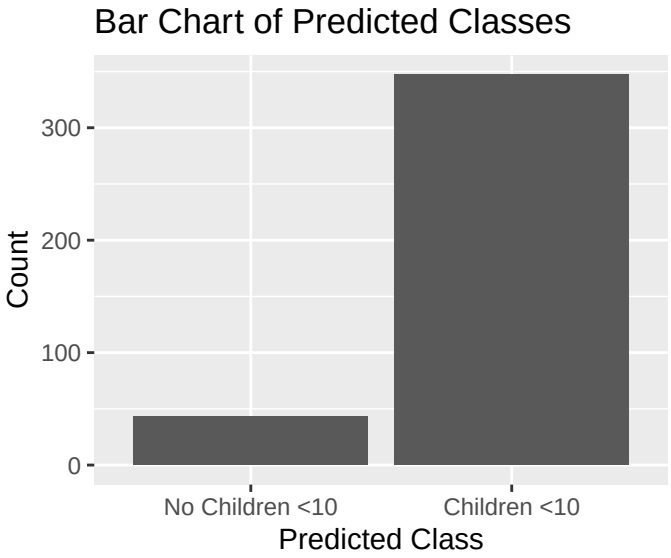
To better see how different models behave when prediction, we do not fix on one model to do prediction, but do predict and then see which model can better do the prediction. Main conclusion: GBM is a better model here for the following reasons.

The GBM model shows better separation between the classes with less overlap, indicating a higher confidence in distinguishing between households with and without children. The peak for Class 1 is much closer to 1, and for Class 0 closer to 0, which means the GBM model has a higher confidence level in its predictions.

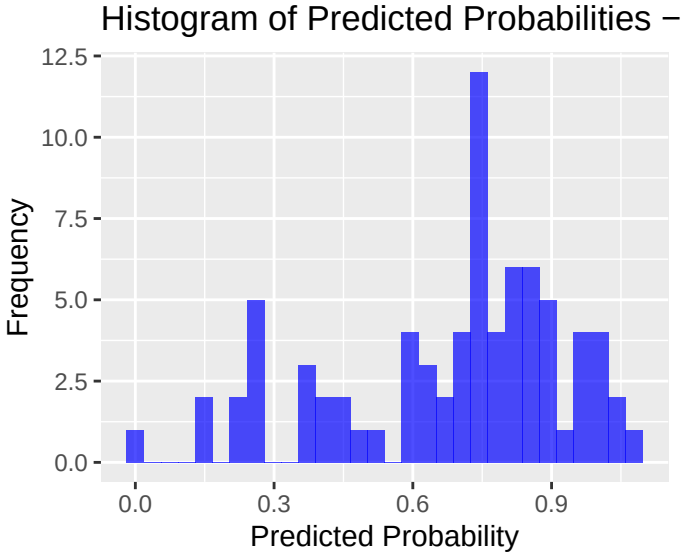
With the issue of low birth rates in Korea, a model like GBM that can better differentiate between households could be critical in targeting social support, optimizing resource allocation for child services, and planning community development.

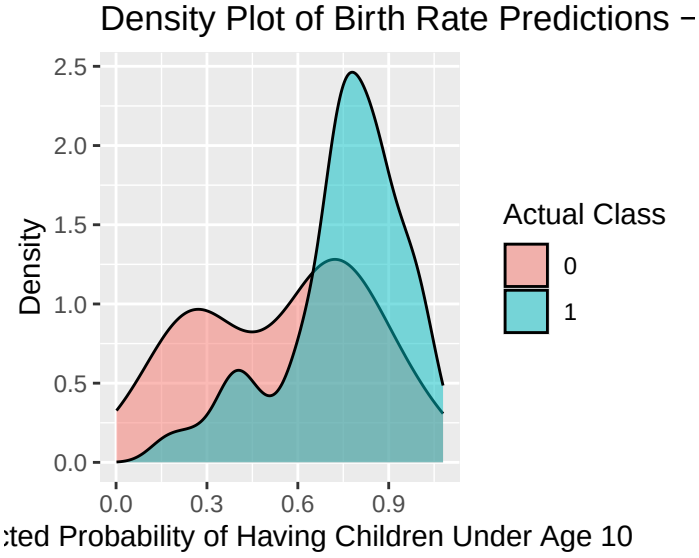
Regarding confidence and reliability, the GBM model appears to be more confident in its predictions, as shown by the higher peak and less overlap in the density plot, and would likely be more reliable in a real-world setting for making predictions about birth rates, which is essential for planning and policy-making.

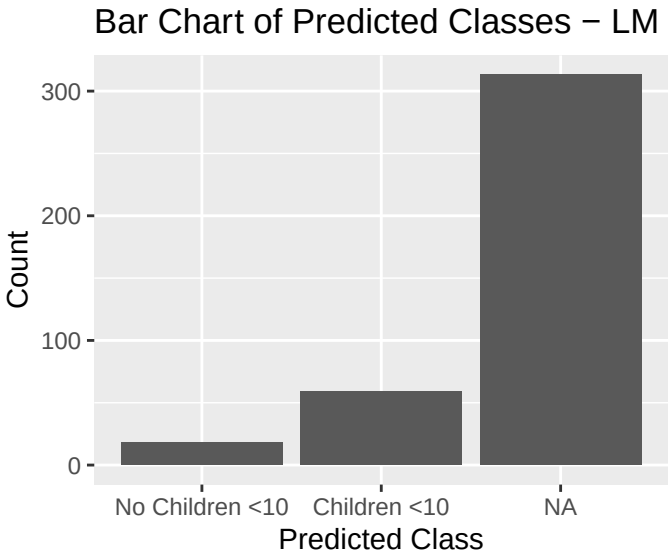





```
## null device
##           1
```







V. Conclusion

The methodology adopted in this analysis involved deploying three distinct predictive models: CART, GBM, and Linear regression, to classify and predict outcomes in a binary setting. The most contributing explanatory variables help us understand the difference between households with kids and those without kids, and we can deduce implications for the policy for boosting fertility rates based on those variables. We split these into three parts: work, life, and economics-related.

Firstly, when it comes to work, whether or not a woman works is more influential than whether a man does. It represents that the role of the mother is more significant in caring for kids, and the difference in weight in infant care between husband and wife makes women hesitate to have kids. Therefore, systematic support from firms for balancing the role of caring for kids is needed, and the government should induce that systemic change. Second, in the light of life, the cohabitation of a husband and wife and satisfaction with life are related to the birth rate. The stability of living with a spouse matters in the birth rate, so the relocation plan for public corporations and government ministries should be carefully developed considering the living conditions of people. The plan is to lower the concentration in the Seoul metropolitan area by the government.

Also, it needs to change the recognition of life satisfaction. We can see that the satisfaction of leisure is higher with kids, but people usually think that having kids requires parents to give up their leisure. But parents can have family-oriented leisure with kids, and their satisfaction with it is not lower than in households without kids. Lastly, economic stability is crucial. Working hours and some house-related variables, such as house size, can be interpreted as economic states. But it does not necessarily mean that economic affluence is absolutely important. As we see before, stability seems more important than just higher income, and it might be related to the need to reserve time to care for children.

However, even though it shows the difference between households with kids or not, it does not necessarily show the causation of not having kids. It is just an analysis of the current state, but the government can support households with kids by a new policy to make them as free as households who don't have kids based on the analysis and it can help people to recognize that having kids is not worse than not having kids.