Baby’s First Years: Which Maternal Factors Predict Child Health Outcomes?

Nasim Sheikhi, Olivia Floody, Riya Gurnani, Kelly Phalen

View full data and notebook file here:
https://github.com/nasim-sheikhi/DS4400-BFY-Study-Analysis

Abstract

The impact of poverty on child development is a key area of research that can be subdivided into various factors, including maternal stress levels and household income. This paper looks at the Baby’s First Years study, which examines poverty, maternal health, and child health throughout early development. The aim of our data analysis is to determine which socioeconomic and environmental factors can best predict child health and development. We used Decision Tree and Logistic Regression Classifiers and Multiple Regression to determine the effect of these aspects on poor or good child health outcomes at the age of one. The highest accuracy we achieved was 67.2% with 20 features and 66.8% with 36 features. These models were decently effective; however, our Multiple Regression model returned a low coefficient of determination ($R^2 = 0.20$), pointing to high variability between variables. Therefore, we believe that with our set of selected features, this survey can be narrowed down so that more data can be obtained, which can be used to inform on risk factors that affect a child’s health and development.

Introduction

The impact of poverty on child development is a key area of research in developmental psychology. Poverty has been linked to several outcomes in children, including stunted growth, cognitive deficits, abnormal brain development, and more. However, these studies are often limited in their cross-sectional design and breadth of their variables of interest.

The Baby’s First Years study (BFY) is a multi-site longitudinal clinical research study that examines the impact of poverty reduction on child development. This study builds on and adds to previous research in several ways. First, it breaks down poverty into several factors, including income, receipt of public benefits, food insecurity, and more. Additionally, it examines maternal health and satisfaction as a function of these poverty variables, which helps to address potential confounding factors in the relationship between poverty and adverse childhood outcomes. Finally, the study’s longitudinal design allows researchers to examine and draw conclusions about the cause-and-effect relationships between poverty and child health.

In the BFY study, 1000 mothers in low-income households are provided with cash gifts of $333/month or $20/month for the duration of their study participation, and qualitative and quantitative data are collected from the mother and child yearly for the first four years of the child’s life. In our project, we will analyze BFY quantitative data, collected from mothers at the 12-month time point, regarding family life as well as maternal and child health. We plan to examine which maternal and family factors (such as quality of mother’s relationship with a significant other, income, and parenting stress) best predict child health in the first year of life. Current publications have not used machine learning methods to analyze this data, and we believe the methods we implement will point to the predictive strength of these variables.

The BFY study, and our work with this data, are significant because they give us an understanding of the extent to which socioeconomic and environmental factors influence child development. Findings from these analyses can inform changes in public policy to better serve families from low-income backgrounds, which would improve child health outcomes. Such policy change may include more anti-poverty interventions, community support for survivors of domestic violence, improved public programs for early childhood education, and more.

Data Analysis

Participants include 986 mother-baby dyads. First, we examined demographic factors of the BFY dataset at age 1. As demonstrated in Figure 1, there is an equal distribution of male and female children, reducing any potential bias of sex differences in early development in future analyses.

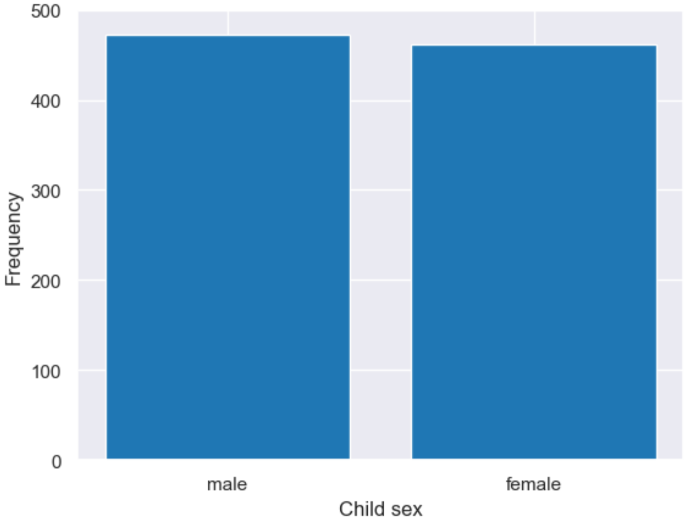

Figure 1. Frequency of male and female children across dataset.

Next, we examined the distribution of household income. Mothers' salaries spanned a range of $0 to $60,000, with 75% of mothers earning between $10,000 and $30,000 (*Fig. 2*).

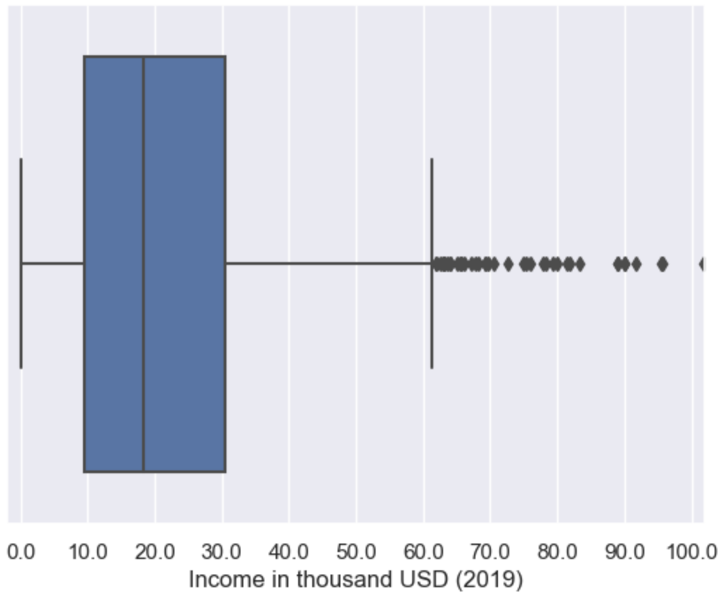

Figure 2. Boxplot distribution of household incomes in thousands of USD.

In this study, several facets of child development were measured: Child social-emotional development was measured by the Brief Infant-Toddler Social and Emotional Assessment (BITSEA); sleep health was measured by the Patient-Reported Outcomes Measurement Information System (PROMIS) sleep disturbance index; physical health was measured by the Poor Health Index; and language development was measured by Ages and Stages Questionnaire (ASQ-3) Communication Language Scale. Correlations were performed across these features of child development. It was observed that child social-emotional development, sleep disturbances, and physical health shared moderate associations with one another, but not language development (*Fig. 3*). These features were normalized and summed to collectively examine child health outcomes in predictive analyses.

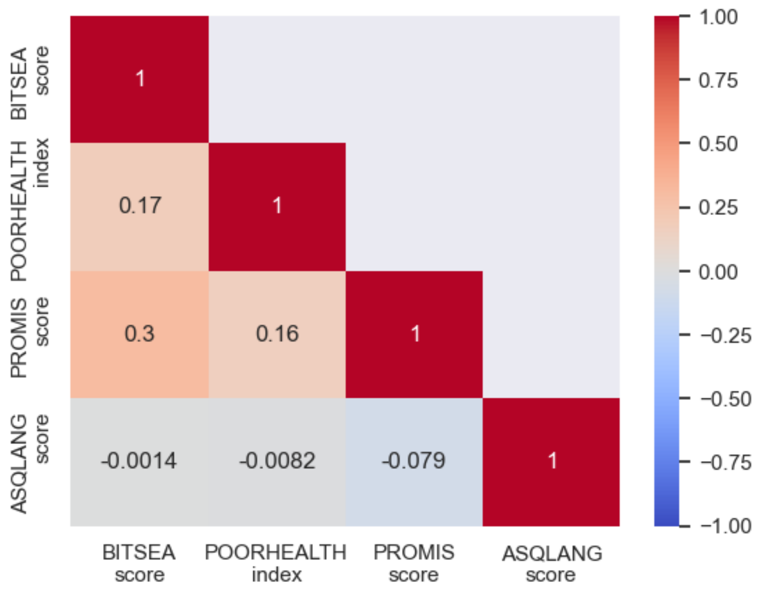

Figure 3. Correlation heatmap of child outcome variables (BISTEA score, Poor Health Index, PROMIS Sleep Disturbance Index, ASQ-3 Communication Language Scale). Values and colors demonstrate direction and magnitude of correlations.

Methods

Since our dataset contained 2,010 columns, we needed to prune which columns we used for our various models. The dataset comes from a questionnaire, so there is a lot of raw data, repeat questions, redacted information, and missing data. However, the researchers did add some aggregated data based on answers from multiple questions in the survey or collating the repeat questions. Therefore, we were able to narrow down many columns by reading the survey documentation to 36 feature columns. We chose this based on select categories we felt had the most impact: mother demographics, current relationships in the household, income and receipt of public program benefits, maternal health, economic stress, and parenting stress (the specific values of these columns can be found documented in the code). There was a percentage of rows in each of these columns that were missing values or were filled with an arbitrary value (-999 = refused to answer, -888 = don’t know). To maintain a high number of samples in our analyses, we imputed missing data with the mean value of the feature.

We also decided to narrow down our features even more by running recursive feature elimination to find the 20 best features from the 36 we chose. This algorithm iteratively builds a model, first with all 36 features, and each time discards the least important feature until there are only 20 left.

We chose to implement three machine learning models: multiple linear regression, logistic regression, and decision tree. Our target variable was determined by creating an overall child health outcome based on normalizing the three features with correlations shown in Figure 3. This outcome score ranges from 0 to 3 where a lower score implies there are less issues with the 3 categories. Due to the continuous nature of this variable, we decided to implement multiple linear regression between the continuous features and our outcome. We implemented K-Fold Cross Validation on our models which splits our dataset into training and testing sets, where each fold is used once as a validation set and the k-1 remaining folds are used as the training set. We ran cross validation before and after feature selection. Finally, we implemented grid search to tune linear regression hyperparameters, ultimately identifying the best features and hyperparameters to improve model performance.

For classification analyses, we decided to bin our outcome variable into two categories of a child having a 'Good Health' outcome (class 0) or a 'Bad Health' outcome (class 1). This was done based on the median of the outcome variable so that all of the outputs less than or equal to the median are placed in class 0 and the rest in class 1. This allowed us to also explore the classification models of logistic regression and decision tree. We thought since we had many features to work with, logistic regression and decision tree would be good classifiers to try to predict poor or good health in babies after their first year of life.

We again implemented K-Fold Cross Validation before and after feature selection for our classifiers. Feature selection initially did not improve the base-model's accuracies for logistic regression or decision tree, but later after hyperparameter tuning, the 20 features did have higher accuracy than those that did not go through RFE.

We implemented grid search in order to conduct hyperparameter tuning using both the 20 features selected after running RFE and the 36 features we had chosen at the start to compare the results. The results were better with the 20 selected features as opposed to the 36 features. For decision tree, we looked for the best parameters in the criterion, maximum depth, minimum samples split, maximum leaf nodes, and maximum features. For logistic regression, we looked for the best parameters in solver, inverse of regularization state (C), maximum number of interactions taken for the solvers to converge, and the norm of the penalty.

Analysis

The highest accuracy for Logistic Regression was 66.8% and the highest accuracy for Decision Tree was 67.2%. Since the dataset is fairly small, running the models results in slightly different accuracies for each classifier before and after feature selection, but overall, the highest accuracy appears to be near 70%.

Running recursive feature elimination to get the top 20 features for the classification algorithms revealed the significant categories to predict the overall child health outcome. These were household income, mother's physical and mental health, parenting stress, and household member relationships (such as whether a mother's partner lives in the household). It would make sense that these environmental factors affect our outcome variable and this provides insight in how this survey can be compressed to produce better data. It would lessen the quantity of survey questions which would result in better quality data since less answers would be missing.

Our Multiple Regression model had a mean squared error (MSE) of 0.08 with a coefficient of determination (R^2) of 0.20. A low MSE, coupled with low R^2, means that, while our model has low prediction error, only 20% of the variation in child health outcomes is explained by maternal health and income. This may be due to low to moderate correlations between maternal variables, a typical confound in research on human behavior. Interestingly, our model returned the same MSE and R^2 before feature selection and hyperparameter tuning as after; this suggests that the selected features, including maternal depression, sleep quality, parenting stress, and household chaos, are the most important in predicting child health outcomes, rendering other variables like income negligible.

A summary of these results can be seen in Figure 4, where our best model is the decision tree post-feature grid search (has been hyperparameter tuned) with an accuracy of 67.2%.

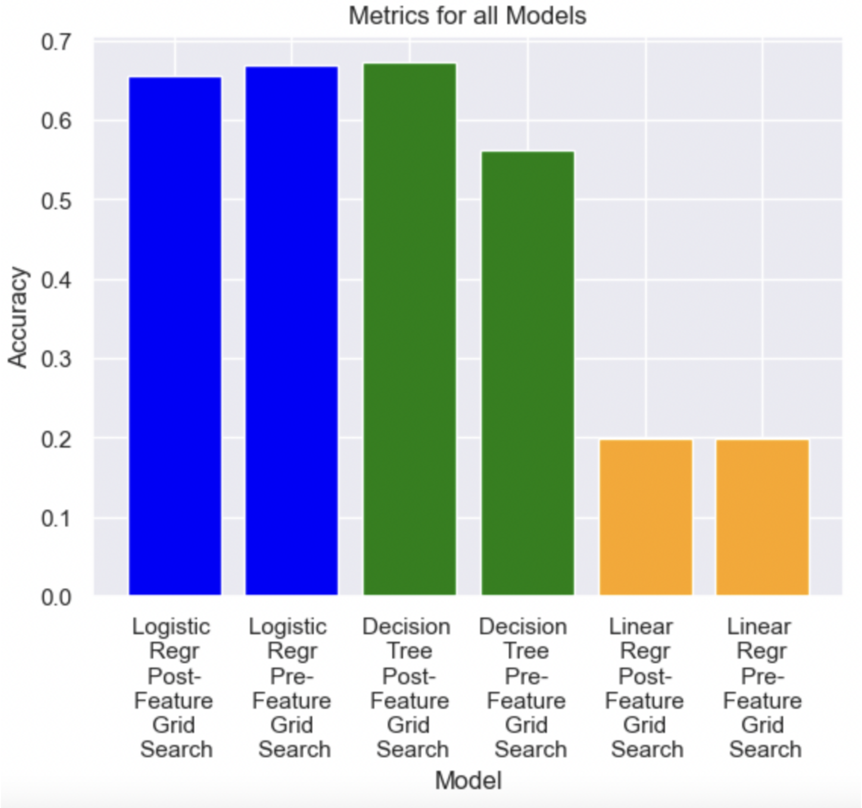

Figure 4. Comparison of accuracies between all the models both pre and post feature selection.

Conclusion

With about a 70% accuracy for our best model, we believe that this survey can be decreased drastically from the 2,000 questions asked to determine effects on the baby's health and development.

We were surprised that with such a complex data set our decision tree had really low values for the optimal max_leaf_nodes parameter. By running grid search, we were able to determine the optimal decision tree hyper parameters. Oftentimes the trees were a lot shallower than we had anticipated indicating that there were strong decision indicators that separated the children in each family into good health or bad health. Depending on the split of random between training and test data, we get different column nodes that the tree splits on. The most common columns that the trees have generated are mother's sleep quality, mother's depression score, mother's anxiety score, and mother's perceived stress. Therefore these are the factors that will best indicate whether a child will have relatively good or poor health based on the 37 features we selected from the data set. The model tested with anywhere from a 61%-69% accuracy and poor health had a slightly higher f1-score than good health outcomes.

We recognize that this data has many limitations, the greatest being the small sample size and the missing data. We believe that cutting down this survey to include far fewer questions in the features we identified would be the best way to collect more data, leading to more valid and reliable analyses on child health. Part of the reason there was so much missing data was due to the fact that people skipped sections, so reducing this dataset would decrease the length of the survey, while still being able to predict the outcome. With an expanded dataset, these metrics of mothers can be taken into account to determine which babies and mothers need their most support in their babies' first years of life.

Ultimately, we identified several important factors that contribute to adaptive or maladaptive child development, including income, supporting household contributors, maternal sleep health and depression, and more. Upon future research and examination, decision tree models can serve as powerful diagnostic tools to predict child support needs given various features of poverty. In addition, Local governments can use this information to create and better allocate resources to the mothers that meet criteria in this survey.

Author Contributions

We had a lot of data cleaning work to do in the beginning by combing through the extensive documentation and pulling out which features would be most useful and interesting to examine through our models. Once we decided on which category types to focus on, we split the data cleaning amongst Kelly, Riya, and Olivia so each person worked on a set of columns while Nasim focused on creating our target variable and exploratory data analysis. Then, Kelly and Riya added the model code functions to create the models, run cross validation, run feature selection, and do grid search from sklearn. Nasim added code for the multiple linear regression analysis. Olivia worked on hyperparameter tuning and analyzing the decision tree. We discussed our findings and added them to the final report and poster slide.

References

Troller-Renfree, Sonya V., et al. "Associations among Stress and Language and Socioemotional Development in a Low-Income Sample." *Development and Psychopathology*, vol. 34, no. 2, May 2022, pp. 597–605. *DOI.org (Crossref)*, https://doi.org/10.1017/S0954579421001759.

Yoo, Paul Y., et al. "Unconditional Cash Transfers and Maternal Substance Use: Findings from a Randomized Control Trial of Low-Income Mothers with Infants in the U.S." *BMC Public Health*, vol. 22, no. 1, Dec. 2022, p. 897. *DOI.org (Crossref)*, https://doi.org/10.1186/s12889-022-12989-1.