Factors Impacting Contraceptive Choice

Saketh Kollu, Anita Shen, Nicholas Wang

Final Project Notebook

Abstract

Contraceptive choice is an important aspect of a woman's life and this paper explores factors that contribute to the type of contraception used. The process used to build regression and decision tree models can help provide an understanding by trying to identify any interactions. This paper begins by exploring the relationship between number of children and contraceptive type, then building two models using Python's scikit-learn machine learning library and finally how we can select and categorize the given data to optimize our model.

Introduction

Motherhood is a beautiful thing and every woman has the right to decide when they want to experience the joys of motherhood through the use of contraceptives. In our project, we took a look at the Contraceptive Method Choice Data Set from the 1987 National Indonesia Contraceptive Prevalence Survey. The data consists of 1473 married women who were not pregnant at the time of the interview and their socio-economic status. The data also includes each woman's age, number of children she has, education on a scale of 4, whether the woman is muslim or not, whether she is working, her standard of living on a scale of 4, media exposure, and her husband's education and occupation. All data are categorical with the exception of the wife's age and number of children she has. Our goal is to predict the preferred contraceptive type of a woman based on her demographic and socio-economic capabilities.

Our group started out with some initial guiding questions to help us formalize our thinking:

1. How is the contraceptive method used related to the number of children?
2. Can we build a model to predict the type of contraceptive method used by an individual?
3. Which feature affects the contraceptive method used the most?

Our initial predictions are that there might be a higher distribution of number of children for mother's that don't use contraceptives than those that use them. In addition, we will build a multinomial logistic regression and a decision tree to help us predict the contraceptive method used by an individual. Lastly, we predicted that there are certain metrics that may have a larger influence on our model's prediction accuracy.

Data

The Contraceptive Method Choice Data Set includes 1473 married women with 10 attributes (including the contraceptive method classification). There is no missing data and each attribute is either categorical or numerical. The attribute information is in **Table 1**

Despite the 9 attributes given in **Table 1**, our group decided that having more clarity in what constitutes as “long-term” and “short-term” contraceptive types would be helpful in our analysis. We could also benefit from our data more by adding more categories in wife’s religion as well as specifying what each category of a husband’s occupation is. Lastly, since the data only surveys a small sample of Indonesian women and excludes unmarried women, our predictions may be too isolated and unable to generalize to a larger population.

Number of Children and Contraceptive Type

Data Processing

While the number of children a woman has is not indicative of how effective a contraceptive method is, our group is interested in seeing if there are any relationships between a mother’s choice of contraception and the usual number of children she might have. We predicted that there will be a difference between the type of method used and the distribution of number of children a mother has, more specifically, there will be a higher distribution of number of children for mothers that do not use contraception compared to those that use them.

Charts

We began by looking at the distribution of the number of children for ALL mothers in our sample data set (**Figure 1**). We noticed that the entire distribution is right skewed ranging from 0 to 16 and with a mean and median at 3 children.

We then proceeded to graph each individual contraceptive type using filled histograms and stack-stepped unfilled histograms. (**Figure 2 and Figure 3**)

For the type 1 contraceptive method, or mothers that do not use any contraceptives, we found a range between (0,12), a mean at 2.93 children, and a median of 2 children. Out of all the types, the blue graph in figure 2 has the highest count of mothers that have 2 or less children.

For the type 2 contraceptive method, or mothers that use long-term contraceptives, we found a range between (1, 13), a mean of 3.74, and a median of 3 children. The orange graph in figure 2 is the most obvious of the three that has a stepwise descent. This graph is the clearest in showing a decreasing number of type-2 mothers as the number of children increases.

For the type 3 contraceptive method, or mothers that use short-term contraceptives, we found a range between (0, 16), a mean of 3.35, and a median of 3 children. The green graph in figure 2 has the largest range of all the graphs and a clear peak at 3 children.

Surprisingly, the median number of children of mothers that don't use contraceptives is lower than those that use either short-term or long-term contraceptives. This may be due to the larger sample size of mothers that don't use contraceptives at 629 mothers versus mothers that use contraceptives(333 + 511) at 844 mothers. All of the distributions are heavily right skewed though again, the range of the number of children is largest for mothers that use short-term contraceptives. (0-16) short-term vs (1,13) long-term, and (0,12) no-use.

Lastly, we are interested in comparing mothers that use contraceptive vs those that do not (**Figure 4**). More specifically, those that do not use contraceptive will fall in the type 1 category with 629 mothers, and those that use contraceptive fall in the type 2 and 3 category with (333 + 511) = 544 mothers. Our results indicate a mean of 2.93 and median of 2 for mothers than do not use any contraceptives versus a mean of 3.50 and median of 3 for mothers than use contraceptives of some sort.

Interestingly, the distribution of mothers that use contraceptives is shifted to the right than those that do not use contraceptives. This corresponds with our findings from the earlier section. Both distribution are right skewed. Our initial prediction of mothers that use contraceptives relating with lower number of children has been proven incorrect in our sample data.

Predicting Contraceptive Type from All Features

Data Processing

Our intention is to create a multinomial logistic regression model to predict 3 response categories (No Contraception, Long Term Contraception and Short Term Contraception). Categorical features need to be One Hot Encoded for the logistic regression model, therefore Sci-kit Learn’s OneHotEncoder preprocessor was used on the 7 categorical features. Numerical features need to be normalized, however Sci-kit learn will take care of this when fitting the model to our data. Once the data has been processed, the one hot encoded categorical features and numerical features need to be merged into a single table which are then split into a training and test set with 80/20 ratio. The resulting process put 1178 instances in training and 295 instances in testing.

Logistic Regression Model

A multinomial regression model was created using the SAGA solver which is a variation of the Stochastic Average Gradient (SAG) solver that uses L1 regularization (LASSO), thus we can find a sparse solution. Training data was then fitted and then the model was evaluated on the training data, cross validation on the training data and finally on the test data.

Random Forest Model

Due to the number of categorical features in our data, we believed there was a case for a Random Forest Model as a Decision Tree can natively treat categorical data without any preprocessing. Thus we created a Random Forest and split the original Contraceptive Method Choice data set without any preprocessing using an 80/20 train/test split. Training data was then fitted and then the model was evaluated on the training data, cross validation on the training data and finally on the test data.

Accuracy

Logistic Regression Evaluation

According to **Table 2**, Training accuracy, Cross Validation and Testing Accuracy are all within a few percentage points of each other, suggesting that our model has reached the point where if we add more features we may go into the territory of overfitting. Although the accuracy is pretty low at 50%, since we are choosing between 3 categories, the random chance of guessing the right contraceptive category for a mother is 33%, while our model predicts on test data with 52%. Since we used all the features we had available in this model, we want to see what single feature or set of features can do better than using all our features.

Random Forest Evaluation

According to **Table 2**, training accuracy is significantly higher than Cross Validation and Testing Accuracy suggesting that our model has been overfit to our data, a drawback of decision trees in general but we had hoped the effect of which would have been reduced by using a Random Forest. This draws us to the conclusion that the Random Forest, although it has overfit our training data, does not perform significantly worse than our Logistic Regression model, sitting just shy of 50% accuracy, still above the 33% accuracy achieved by randomly guessing what contraceptive type a mother uses.

Selecting and Categorizing Data for Optimal Predictions

Data Processing

We initially wanted to examine how each individual column variable affects our prediction so that we might be able to extract some columns while not using others based on how well it does by itself. However, we quickly grew wary that this is not the appropriate approach to take to select data. This led us to find new ways to optimize the model by taking into account the numerical data in a new way. Instead of using a linear regression to model numerical data, we tried separating the numerical data into ranges so we can use them as categorical data. Then, we can one-hot-encode this category and perform logistic regression. Before we do any of this, let's see if transforming the numerical columns into categories actually helps our model predict better. We started off with predetermined ranges, and this new logistic model predicted about 3.5% better on the test set than the old logistic model.

Evaluation of New Model

A multinomial regression model was created using the lbfgs solver, an optimization algorithm for parameter estimation that uses L2 regularization. Training data was then fitted and then the model was evaluated on the training data, cross validation on 5 iterations, and test data.

We started off by testing the new categorized column variables with all categorical column data, and seeing if it made any improvement. Without the new variables, we were at 46.4%, but with the new variables, we were at 54.2%! So we concluded that we should stick with the categorical columns, and find a better subset of column variables to use our logistic regression model on.

After several trials, our best subset of column variables appended with the two new categorized columns has a test accuracy of 55.9%. We will continue evaluating this model in Table 3.

Accuracy

From **Table 3**, the transformation from numerical to categorical does in fact increase our prediction accuracy in both the training and testing set. However, a notable exception occurs on the cross validation score, where the prediction model actually suffers from this data modification. This surprised us because we would expect the cross validation score to reflect the testing set score, as it is of estimating testing accuracy. This sudden drop in cross validation accuracy reflects the nature of splitting numerical values into categorical values, since it only drops so low after adding our two new categorical data variables. After all, the sole purpose of the cross validation accuracy is to estimate the testing set accuracy, which is 15% higher. So is the cross validation score irrelevant? To make sure that the testing set always performs much better than cross validation, we ran the function on multiple train_test_split random_states, and confirmed this is true. While the cv score is always around 40%, the test accuracy is always around 55%. Thus, for the large part, we can disregard the cross validation accuracy.

Discussion

One of our original ideas was to select certain column variables to include in our logistic regression model by order of priority of the prediction accuracy of individual column variables. For example, if wife_education predicted best, we would include the column in our final logistic regression. However, we quickly realized that this was not the appropriate approach to take to select variables. Oddly enough, some columns even have the same prediction percentage, which was interesting and definitely surprised us! Another interesting feature we came across was how the cross validation score was relatively low with respect to both the training and test data (about a 15% differential!). We were taught that cross validation reflects the test accuracy to measure the degree of overfitting, but this clearly wasn't the case for our model. It was definitely interesting, but for the large part, we disregarded it because we concluded that the cross validation score was irrelevant and perhaps even the wrong way of simulating tests because it wasn't accurately reflecting the test score.

We originally believed that the number of children will be an effective feature for predicting the type of contraceptive a mother will choose, but from figure 2 and 3, most of the mothers have overlapping numbers of children regardless of their preferred contraceptive type. Furthermore, figure 4 describes that the curve of mothers that do use contraceptives is similar to that of mothers that do not. Number of children has proven to be an ineffective feature because it does not consider the choice of the mother to have the child. There is no causal relationship between the number of children and type of contraceptive a mother chooses.

Although the schema of our data was easy to use, the biggest trouble we had was in increasing our logistic regression model accuracy past around 50%. We believe the cause of this issue was a combination of the features we choose and a nature of the phenomenon the data presents. A common theme we had in our figures was a lot of overlapping regions of our data points no matter how we split, take figure 3 and figure 4 for example where the response we are trying to pick does not differ very

much by the features we had chosen. Even switching to the decision tree did not yield better test accuracy due to overfitting of our training data, suggesting that supervised learning techniques to predict the contraceptive type based on the features we have at hand may not be the best approach to this problem.

The mindset we went into the project was assuming that less children had a relationship with long term contraceptive, hoping to see this in addition to the other features we had yielded a good model. However this assumption is not true all the time based on figures 1 through 4 and the results of our models. Another aspect is that birth control is an extremely personal choice based on a vast number of situations going on in a person's life, so trying to predict based on 9 features the type of contraception a person uses is bound to have limitations.

As a result of the nature of the data being from a specific group of people from a specific culture and country, we wanted to keep in mind that any results or ideas we take away from this research is not representative of any individuals or groups. We want to make sure that our results do not reinforce any biases or stereotypes.

Having more information on the husband feature would be interesting because contraception type may be influenced by partners and even society. To generalize this information, it would be useful to get more samples from people of all religions and backgrounds. We also hoped that the resolution of the dataset would be larger, including more specific information about jobs and contraception type rather than broad categories that can be misinterpreted. Also having contraception usage information (ie. how often people take it) would be helpful as we could analyze the effectiveness of various contraception types given more information about how many children the couple planned for versus actually have.

While analyzing the data, it felt peculiar taking into account the religion variable as either Muslim or non-Muslim, as this categorization obviously has some ethical concerns to it. Not only is it misleading to single out a single religion and predict contraceptive type based on that, it's also unethical for any data produced to say "this family used contraception type X because they are Muslim". We shouldn't be associating this data with any religious connotation. Alternatively, there could also be more representation in the dataset by adding in other religions so it doesn't single out a single religion.

From a series of modeling and exploration, we see that our predictive models have low testing and cross validation accuracy with the risk of overfitting. In the future, we would like to explore unsupervised learning models such as clustering to discover hidden patterns in our data. We can have our algorithm identify and group structures in our data that may be undetectable through regression models or decision trees. Furthermore, our analysis can also benefit more from increasing the number of features and increasing the sample size. Lastly, societal views and medical technologies have significantly evolved since the survey was conducted in 1987. It might be interesting to join a more recent dataset and compare how women's statuses and choices have changed throughout the years.

Tables and Charts

Attribute	Type	Notes
1. Wife's Age	numerical	
2. Wife's Education	categorical	1 = low, 2, 3, 4 =high
3. Husband's Education	categorical	1 = low, 2, 3, 4 =high
4. Number of Children	numerical	
5. Wife's Religion	categorical	0 = Non-Islam, 1 = Islam
6. Wife's Work Status	categorical	0 = Currently Working, 1 = Not Working
7. Husband's Occupation	categorical	Sectors 1, 2, 3 , 4
8. Standard of Living	categorical	1 = low, 2, 3, 4 =high
9. Media Exposure	categorical	0 = Good Exposure, 1 = Low Exposure
10. Contraceptive Method Used	Class attribute	1 = No use of Contraceptive 2 = Long-term Contraceptive 3 = Short-term Contraceptive

Table 1

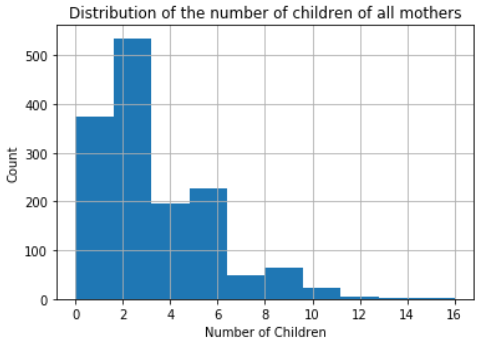

Figure 1

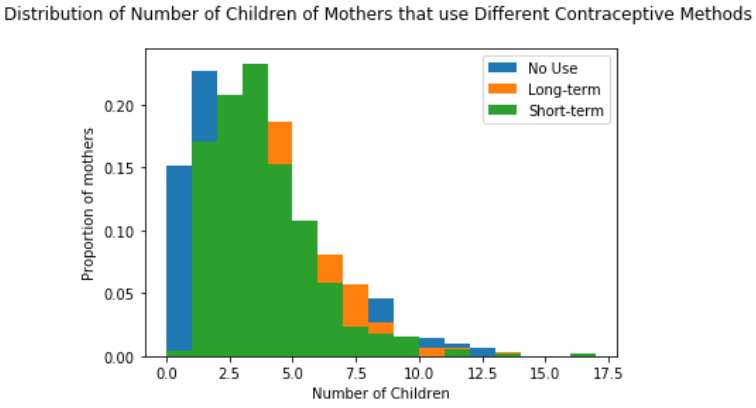

Figure 2

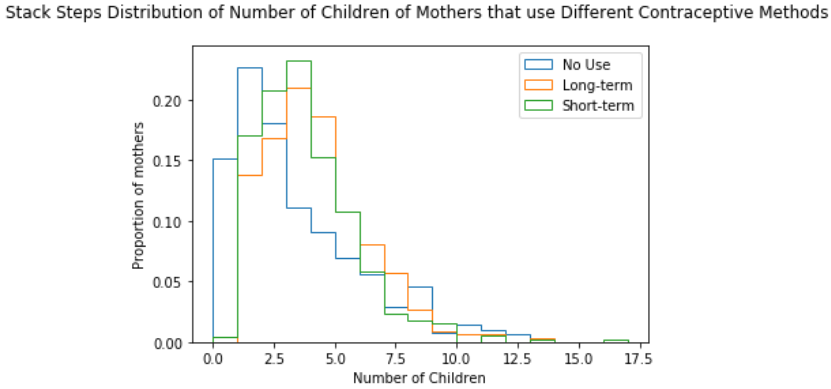

Figure 3

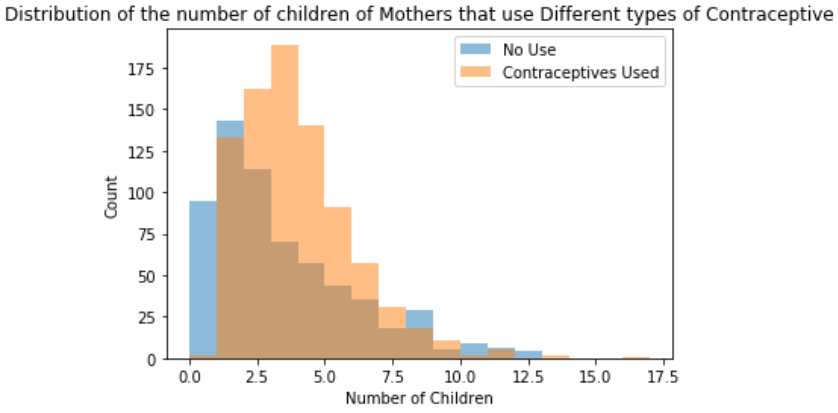

Figure 4

Model	Training Set Accuracy	Cross Validation Score	Testing Set Accuracy
Logistic	0.52462	0.51868	0.51864
Random Forest	0.94312	0.51946	0.52542

Table 2

Model	Training Set Accuracy	Cross Validation Score	Testing Set Accuracy
Before Data Modification	0.52462	0.51868	0.51864
After Data Modification	0.56197	0.39895	0.55932

Table 3