An Investigation of Contraceptive Use Among Married Women in Indonesia

Authors: Jennifer Senta

Abstract

In this study, we investigate the relationship of contraceptive use with demographic and socio-economic variables. Our data is based on a survey of married women in Indonesia collected in 1987. We use exploratory analysis to visualize interesting relationships between contraceptive use and education level, contraceptive use and woman's age, and various other relationships in the data. To predict which category of contraceptive use a woman falls into based on this data, we train both a one-vs-rest logistic regression model and a random forest decision tree model using a training subset of 75% of the data. Although we see good training fit on the latter model (due to methodological over-fitting), neither model reaches 50% predictive accuracy on the withheld 25% test set, indicating that additional information may be necessary to accurately understand women's contraceptive use choices. Future work may include additional data on unmarried women; data that has been collected more recently given advances in contraceptive methods; and data with greater geographical and religious diversity.

Introduction

Contraceptive use among women around the world varies widely according to a number of factors, including (but not limited to) demographic information and socio-economic status. Understanding contraceptive use decisions is of vital importance for public health, and can help researchers decide who may be good contraceptive candidates or who may need more education and information about contraceptive use. In this analysis, we attempt to investigate some of the correlates of various contraceptive use levels. For these purposes, we will use a data set based on the 1987 National Indonesia Contraception Prevalence Survey.

In our analysis, we assess the relationship of available data points such as a woman's age, religion, work status, and other features, with the decision to use contraception. We begin by exploring the relationship between attributes of the data, and then use these features to build a multi-class logistic regression classification model. We train our model on a subset of the data, and test its accuracy at classification of contraceptive method using a held-out test set. We also perform 5-fold cross-validation to more robustly verify our model does not overfit to our training set, and examine a confusion matrix of precision versus recall in testing classification.

As an alternate model, we train a random forest classifier on the same set of training data, and use precision/recall curves to assess the impact of hyperparameter tuning.

Ultimately, we find that the data available in this study is not sufficient to classify contraceptive use very accurately. Further research in this area might include assessing new demographic and socio-economic variables for a more meaningful relationship with contraceptive choices. Additionally, expanding the data set beyond a single country might help with generalizability and reveal new features, such as country-prevalent religion, that bear a relationship to contraceptive use.

Description of Data

This data set is publicly available via the University of California, Irvine Machine Learning online data repository. Our data is based on a survey of married women living in Indonesia, and contains the following attributes:

Attribute Information

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 ever born (numerical)
5. Wife's religion (binary) 0=Non-Islam, 1=Islam
6. Wife's now working? (binary) 0=Yes, 1=No
7. Husband's occupation (categorical) 1, 2, 3, 4
8. Standard-of-living index (categorical) 1=low, 2, 3, 4=high
9. Media exposure (binary) 0=Good, 1=Not good
10. Contraceptive method used (class attribute) 1=No-use, 2=Long-term, 3=Short-term

*Source: https://archive.ics.uci.edu/ml/datasets/Contraceptive+Method+Choice

Description of Data: Exploratory Data Analysis

Our first task in examining this data was to ascertain whether any data points were missing, NaN, or unreasonable values. To investigate this, we calculated a data summary of the fields in the data, as well as a count of missing/NA values. The data appeared clean, with no missing values, and reasonable min and max values for each category. The only value which was questionable was the maximum of 16 in the field denoting number of children. However, while this seems high, it is not outside the realm of possibility, and so we left this data point as-is in our analysis.

Once we had verified whether the data was clean, we performed an exploratory data analysis to assess potential patterns and relationships within the data. Our first and strongest suspicion was that religious status would be strongly correlated with birth control use. Our data only contained information as to which women were Islamic vs. non-Islamic, so to investigate this, we first graphed the percentage of women in each contraceptive use category who were Islamic.

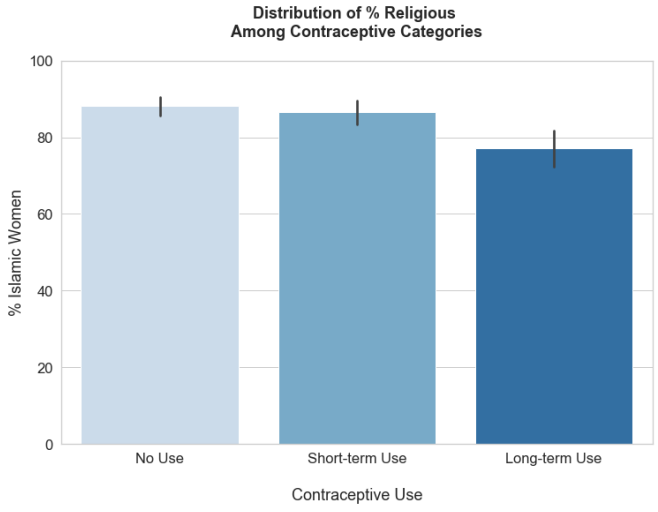

Figure 1:
Percentage of women in each contraceptive category who are Islamic

This figure was somewhat surprising, in that it shows there is not a large difference in the percentage of women who are Islamic across the different contraceptive use categories (although there is a small yet significant relationship which shows that Islamic women comprise a smaller portion of the long-term use contraceptive group than the other groups). To see this same data in another way, we created an alternate visualization of the relationship between the wife's religion and contraceptive use.

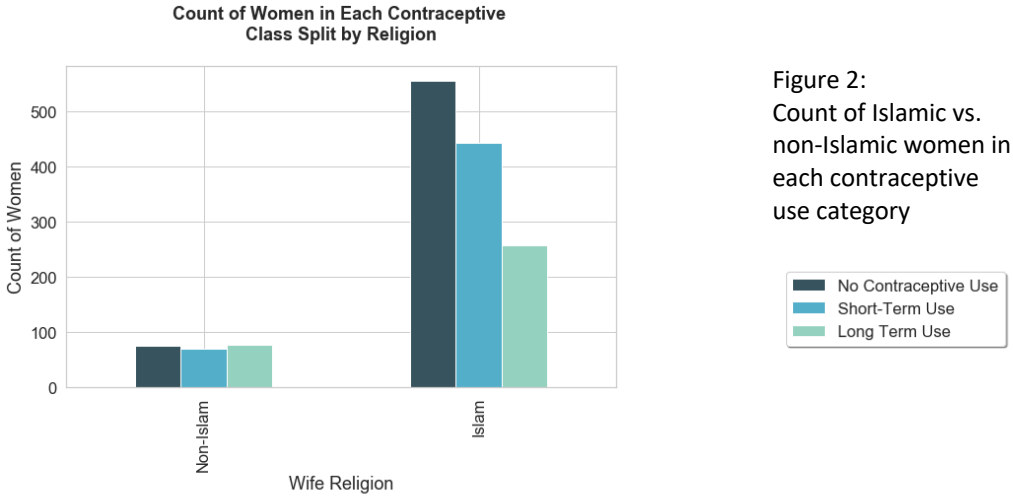

Figure 2:
Count of Islamic vs. non-Islamic women in each contraceptive use category

This alternate visualization reveals that, while the number of non-Islamic women is relatively evenly split among contraceptive use, the Islamic women have a clear trend towards no or short-term use. This will be a good feature to test in our classification model.

Next, we wanted to investigate the relationship that education level has with number of children. We looked first at wife education levels, and next at husband education levels. While we do observe that increasing levels of education are generally associated with smaller number of children, this relationship may not be statistically significant. We also note with interest that the outlier data point with 16 children occurred at the highest possible level of education for both husband and wife.

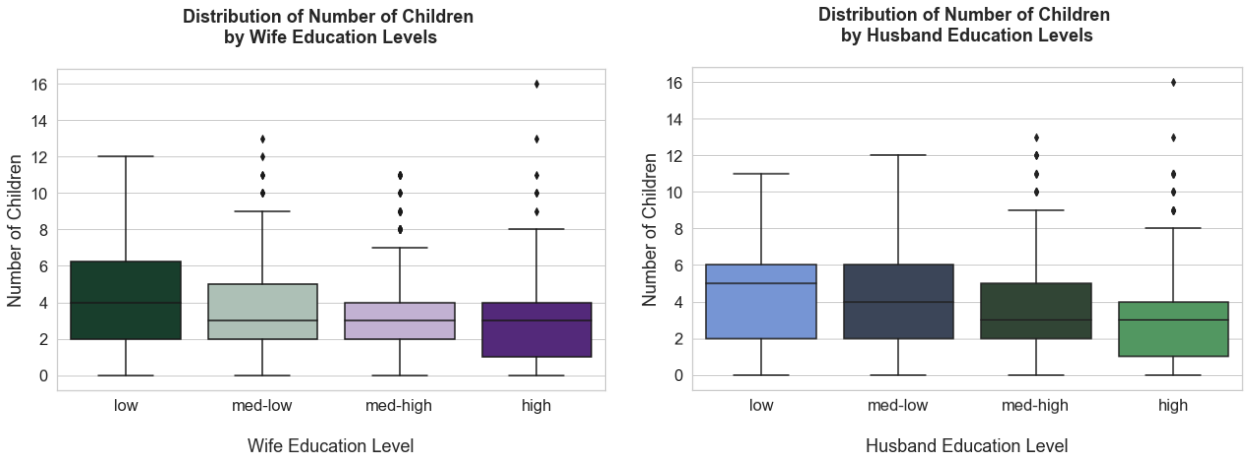

Figure 3: Distribution of number of children by education level of husband and wife

This trend for higher education to be associated with a lower number of children may indicate that higher education levels are associated with increased contraceptive use. We therefore next examined the distribution of contraceptive use categories across each level of wife education provided by the data.

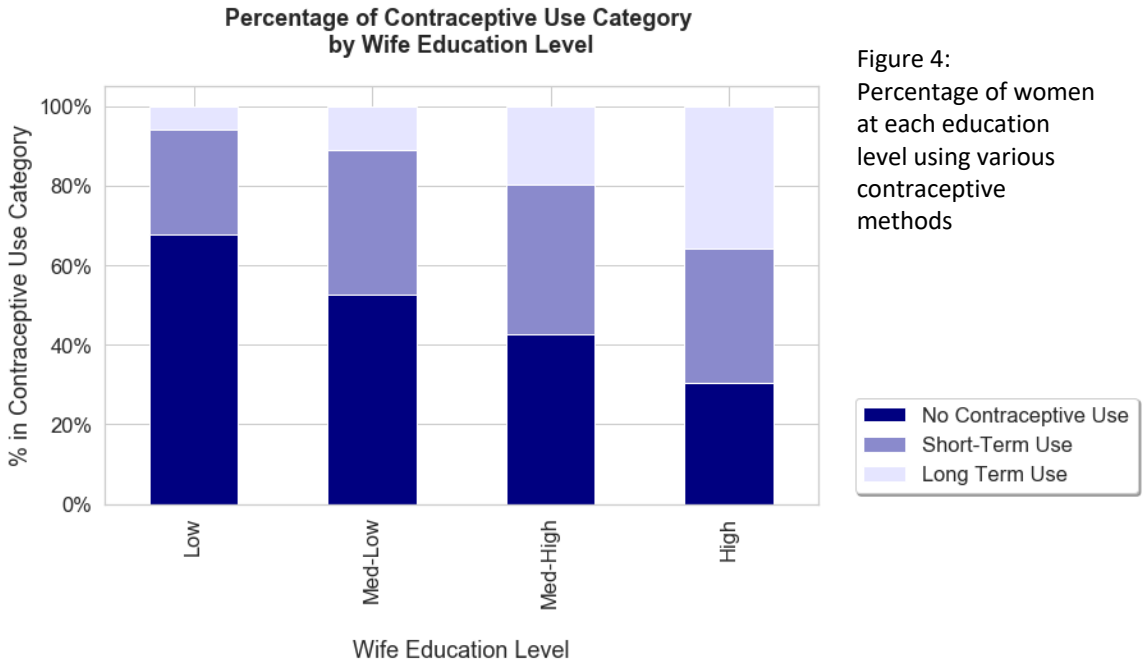

Figure 4: Percentage of women at each education level using various contraceptive methods

Here, we observe a clear decrease in proportion of the "no contraceptive use" category, and a clear increase in proportion of "long-term" contraceptive use, with increasing wife education level. This may therefore be a key driver of contraceptive choice, and will be a good feature to test in our model (see Description of Methods section for one-hot-encoding procedure for categorical variables).

Another potential driver of contraceptive use might be wife age. We visualized the relationship between wife age and number of children (see next page for Figure 5), and we observed an increasing linear relationship between these variables (as we would expect given that older women have had more time to have children).

We also separately visualized the distribution of each contraceptive use category by wife age (see next page for Figure 6), and can see that short-term contraceptive use appears more prevalent among younger women, while no contraceptive use has peaks at a younger and older wife age. This feature (wife age) will also be added to our classifier for testing (see Description of Methods section for standardization procedure for numerical variables).

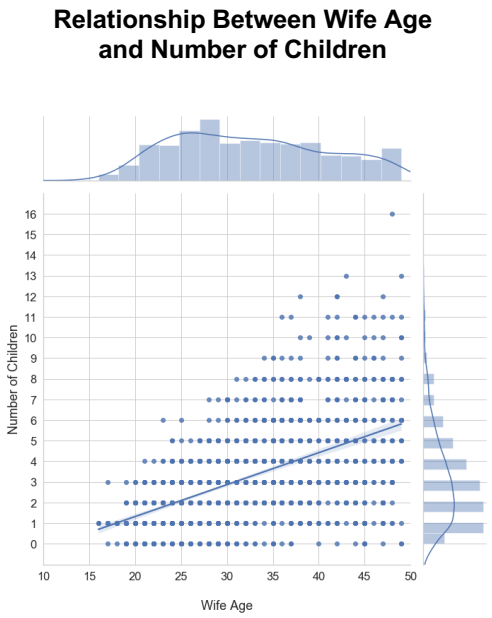

Figure 5:
Wife age vs. number of children

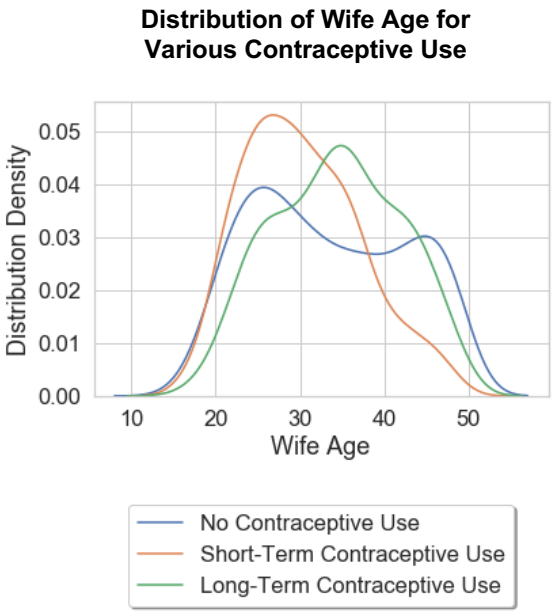

Figure 6:
Distinct distributions of contraceptive categories by wife age

Description of Methods and Summary of Results

After performing exploratory data analysis, we decided to first use a one-vs-rest logistic regression classifier to fit and predict our data. This model type is appropriate for use in categorization tasks with more than two outcome categories, and therefore is well-suited for our contraceptive data with three use categories.

Our first step was to divide the data into a training set of 75% of the data, and a testing set of 25% of the data, using the Sklearn model_selection module's train_test_split function (with a random state = 100 for reproducibility). This enabled us to train our classifier on one set of data, and then test our model's out-of-sample predictions to ensure generalizability.

We fit our one-vs-rest logistic regression model on the training data in several iterations, each time adding additional features to assess improvement of predictive training accuracy. In every case, we attempt to predict contraceptive use category using various features of the data, and calculate our predictive accuracy. Each of the categorical features was one-hot-encoded by a reusable custom function that uses the Sklearn feature_extraction module's DictVectorizer. The two numeric variables were standardized by subtracting mean and dividing by the standard deviation. We created another custom function which used the Sklearn standardizer operation to accomplish this, so it could be reused in later data processing if needed.

Our one-vs-rest logistic regression model is fully specified using the Sklearn built-in logistic regression function, with "ovr" multi-class option. This model includes an intercept fit by default, and uses limited-memory BFGS for optimization by approximating the matrix of second derivatives via gradient estimations. Our model assumes that the data are independent, and that the relationships between the dependent and the independent variables are largely linear (given that we have not used any polynomial or logarithmic transformations to our data).

We first tried to train our model using only the following data features: wife religion, wife education level, and husband education level. After one-hot-encoding these categorical variables and training the model on this 1st attempt training data set as described above, we tested the classification accuracy on the training data, and found we had an accuracy of 47%. While better than a chance model, we would prefer a higher predictive accuracy on our training data.

We next decided to add wife work status as an additional predictor, and accordingly we one-hot-encoded this categorical variable and added it to our data. After retraining the classifier on this 2nd attempt training data set, we found that our training predictive accuracy did not improve (the accuracy was still at 47%).

We then returned to our exploratory data analysis, and assessed the relationships between wife age, number of children, and contraceptive use, as outlined above in our Data Analysis section. We decided to add both wife age and number of children as predictors to our classifier, standardizing each attribute prior to use. We then re-trained our classifier using this 3rd attempt training data, and obtained a new training accuracy of 52%, which represents a sizable improvement over our previous models.

Next, we added one-hot-encoded categorical variables for socio-economic status and media exposure, to see if these external factors could contribute to our classification without overfitting. Our new training accuracy for this 4th attempt was 54%, which was the best predictive training accuracy so far. At this point, we decided to add the last remaining variable (husband occupation), which we one-hot-encoded based on occupation descriptions as provided on Piazza (1=Professional, 2=Sales, 3=Manual, 4=Agruculture). This raised the predictive accuracy on the training set to just over 55%, making it the most effective subset of data from a training perspective.

However, we had not yet assessed whether using this number of predictors would result in overfitting our training data.

To explore this possibility, we performed a 5-fold cross validation on our final (5th attempt) training data to assess whether we had overfit to the training data. This analysis indicated that we had not overfit to our training data, as each cross-validation fold produced a similar predictive accuracy to our earlier result, and the mean across the cross-validation folds was also very close to 55%.

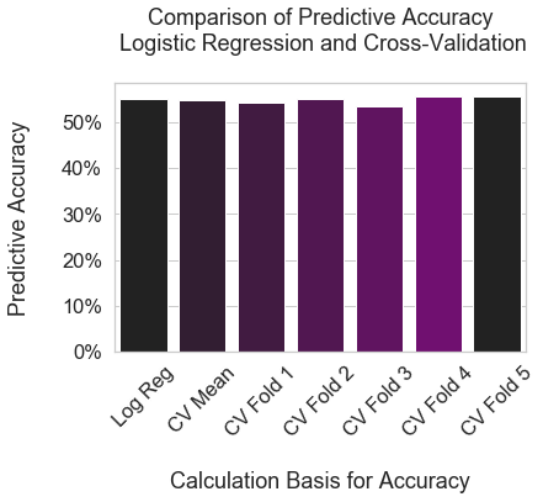

Figure 7:
Predictive accuracy scores for cross-validation folds, average cross-validation, and original training score

The final step of our logistic regression classifier was to calculate predictive accuracy on our held-out test data. As expected, this testing accuracy of 47% was not as high as the training accuracy.

We then visualized the confusion matrix for our final model. We normalized the classifications to show the percentage of each category's predicted values. As we can see, although we have relatively good categorization for the "no-use" category, we have lots of incorrect classifications (as seen in the off-diagonal entries). Ultimately, this indicates that our data set is probably not sufficient to accurately classify contraceptive use categories among this population.

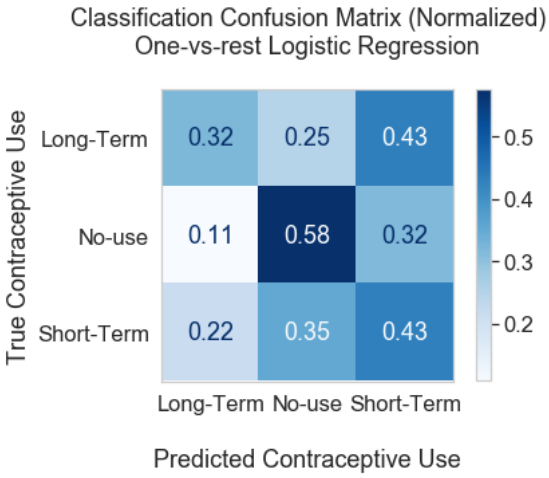

Figure 8:
Confusion matrix of logistic regression classification precision vs. recall on testing data

We then decided to see whether another classifier type might perform better. Decision trees are useful models for categorization tasks, such as this categorical classification problem. We trained a random forest decision tree model on our final (5th attempt) training data from above,

and found that this model had a training accuracy of approximately 92%. This was therefore an extremely accurate classifier for the data on which it was trained; however, decision trees tend to heavily overfit training data by their structure of multi-stage categorization.

We generated the following plot of one of our random forest model estimators to visualize the sheer complexity of this model after fitting to the training data. Although not useful to actually read or observe the decision boundaries drawn, this picture illustrates the large number of decision nodes and connections required in a random forest model for our data set, which had many predictive variables.

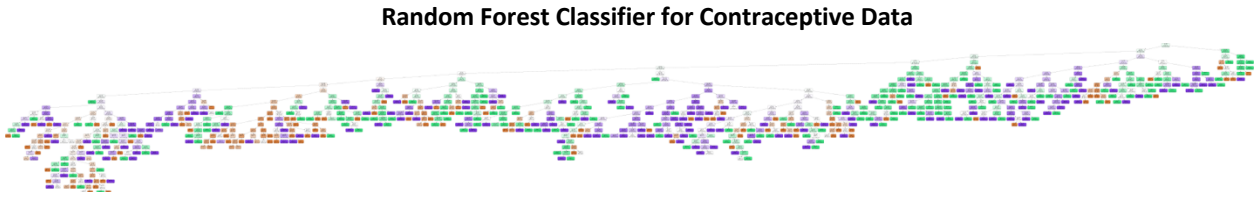

Figure 9: A visualization of the complexity of our random forest decision tree model

In general, we can tune hyperparameters of our models to potentially achieve a better fit to our data. For our random forest classifier, we examined precision vs. recall curves for two differently specified random forest classifiers; one with 20 iterations and a maximum tree depth of 10, and one with 50 iterations and a maximum tree depth of 3. As expected, a higher number of estimators is consistent with slightly improved precision/recall tradeoff. However, ultimately the adjusted hyperparameters do not do much to increase the predictive capability of our model.

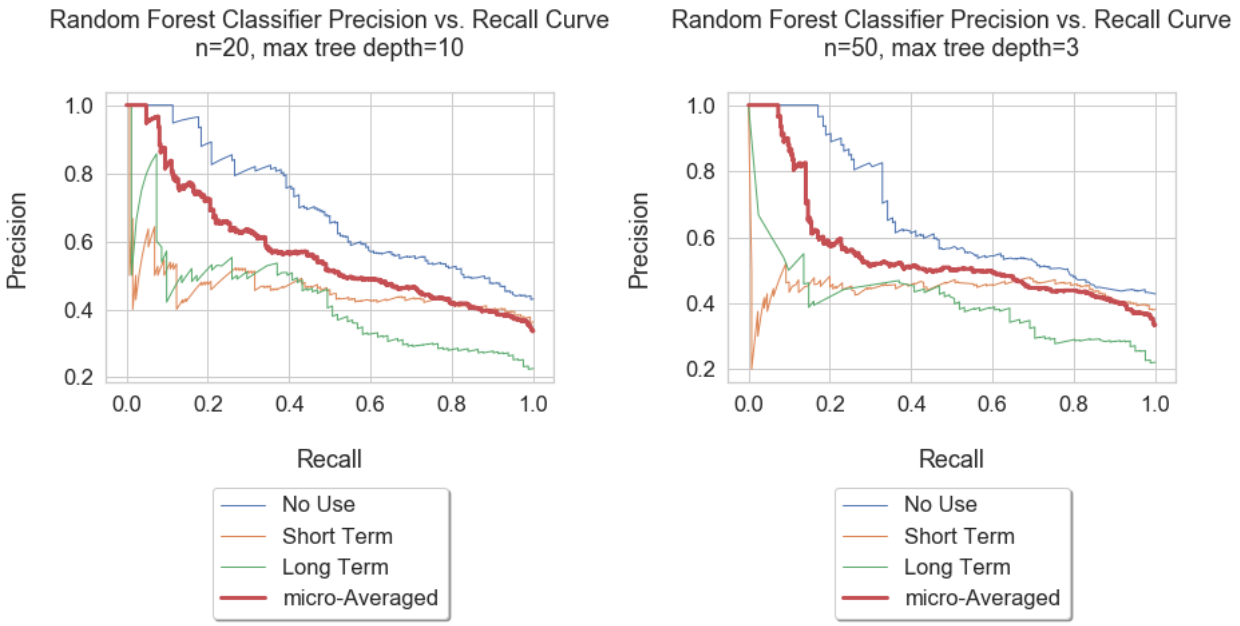

Figure 10: Precision-recall curves for two random forest model hyperparameter specifications

The predictive accuracy of the random forest decision tree model on the held-out test data set was 48%, which was much worse than its training accuracy. Moreover, this is slightly more effective than our final logistic regression classifier model, which had a final testing predictive accuracy of 47%. However, the results of the two models are close, and ultimately we cannot predict categorical contraceptive use with much accuracy in either model based on this data.

Discussion

We attempted to model contraceptive use categories based on provided demographic and socio-economic information. There were several very interesting features of the data; in particular, the percentage of women at each education level who used each of the various contraceptive methods clearly varied between educational group (figure 4, above), and showed a strong decrease in "no use" of contraception as education level increased. Additionally, it was fascinating to see how the density distributions of each contraceptive use category varied differently by wife age (figure 6, above). This feature revealed that long-term contraceptive use tended to skew towards older women, while no contraceptive use had peaks at both younger and older ages.

One surprise during our analysis was that adding wife's working status to our logistic regression classifier did not prove effective in increasing classification accuracy. Intuitively, we assumed that a woman's choice to work might be associated with a higher likelihood of using contraception, but this attribute did not make a significant difference to our classification model.

The data was challenging in its presentation of nearly all features as categorical. In addition, one binary variable (wife work status) was reverse coded, with 0 = Yes and 1 = No. However, perhaps the ultimate challenge in working with this data was that the provided data attributes were simply not sufficient for a reasonably accurate classification of contraceptive use categories. Although we ultimately used all the data fields provided, and assessed multiple classification algorithms, ultimately we could not exceed 50% accuracy for out-of-sample test set predictions.

The analysis in this report contains many limitations. From a demographic perspective, the maximum value of wife age in our data was 49 years old; however, many women begin or complete menopause in their 40s, and these women would have no need to use birth control, even if they previously had used contraception at some point in the past. This might bias our data toward more women falling into the "no use" category (although using wife age as a predictor, which we did, might help to account for this). Additionally, we implicitly assume a linear relationship between our predictors and our categorization outcomes by using a logistic regression classifier. However, some features of our data, such as wife education level, might have more curvature/non-linearity to their relationship with contraceptive use, and so our final model assumptions might not be correct. Finally, the data in use was collected in 1987, in a predominantly Muslim country. Since this time, there have been medical advances in contraceptive methods as well as advances in the dissemination of accurate scientific

information about contraception. This study is therefore limited by both the outdated nature and the overly specific sample group of the data used herein.

It was important to be attentive to ethical considerations while exploring this data. A woman's contraceptive choice is a highly personal matter and private health decision; such data requires thorough de-identification. Additionally, it may be difficult for a woman who is religious, or whose husband is religious, to honestly and accurately report her birth control use if religious or other pressure is exerted on her. This self-report may therefore place some women in a difficult position where they feel pressure to answer in a certain way. Additionally, as the study author (self) may have had strong pre-existing intuitions about which data attributes would be related to contraception, it was ethically very important to ensure that the analysis was unbiased in its examination and evaluation of the data attributes.

Although the available data was informative, a further study of contraceptive use would certainly benefit from additional data. First, our sample of data is taken solely from married women. It would be important to look at both married and unmarried women to better determine the influence of our demographic and socioeconomic predictors after controlling for married status, and it would be interesting to assess the relationship of marriage with contraceptive use. Additionally, data collected from geographically diverse regions of the world would be helpful. There may be strong cultural influences inherent in our data, which was all collected in Indonesia, that we are unable to observe. Also, a country's predominant religion may exert an influence on women's contraceptive choices, even if the women themselves do not report being religious.

In broadening the scope of this study for future research, several ethical concerns might arise. For example, women may be hesitant to self-report birth control use if they believe it can be associated with them personally. One method to address this would be to ensure anonymity by assigning random study IDs to survey participants, which are not linked with name or personal information. Another ethical concern might arise if certain relationships are discovered between polarizing features, such as specific religion and contraceptive use, for example. One method to address this might be to break reporting of religion down into two separate features: 'religious/non-religious', and 'religion opposes contraceptive use' (Y/N). This could provide adequate information regarding the overall relationship of religion with contraceptive use, while avoiding polarizing implications for any particular religion.

Ultimately, we were surprised that contraceptive use could not be more accurately predicted based on variables like socio-economic status, education level, and religion. Future work in this area should use broader and more current data as outlined above. Additionally, future research might consider related yet alternative questions, such as using these same demographic variables to predict the number of children a woman is likely to have, rather than her contraceptive use.

To view a brief presentation of these results, please refer to the video included in the separate .zip file submission.