Multivariate Classification of Contraceptive Method Choice

Violet Buxton-Walsh | COMPSCI C100, Principles & Techniques of Data Science | May 13, 2020

Abstract

The relationship between contraceptive method use, demographic, and socio-economic characteristics is of great interest to public health and family planning professionals, as well as to insurance companies, healthcare providers, and women using contraceptives. This study builds on previous work, examining how best to predict a woman's choice of contraceptive based on such factors. Using data from the 1987 National Indonesia Contraceptive Prevalence Survey, this study implements a Random Forest classifier for optimal performance in the predictive task. Only able to reach about 55% test accuracy, this work highlights previous findings indicating a limited predictive capability from within the dataset, and a tendency to overfit to training data.

Introduction

This study investigates the question of how to best predict the contraceptive method used by an individual based on their demographic and socio-economic information. To this day, global disparities in access to contraception are affected greatly by cultural and socio-economic factors, limiting the ability of many women to control their own reproduction (World Health Organization, Population Resource Bureau 2019). The problem of prediction is illuminating in that it reveals which factors are most impactful on a woman's choice of contraceptive method. Often, this underscores our intuitive knowledge that poverty and education are some of the strongest controls on women's reproductive outcomes, emphasizing inequitable access to contraceptives. Therefore, also investigated in this study, is the relationship of contraceptive type and standard of living.

The Contraceptive Method Choice Data Set comes from the UC Irvine Machine Learning Repository, and transfers cleanly to a pandas dataframe requiring minimal restructuring (limited to the renaming of column headers), making initial data cleaning (filling of NaN values, transposing columns, etc.) unnecessary.

Methods

Exploratory Data Analysis

Several exploratory data analysis showed that a woman's age and her number of previous children would be strong predictors of the contraceptive method used. Notably, the visible trends in Figure 1 are more clearly articulated, with slight differences, when broken down into categories for standard of living (1 lowest through 4 highest). These show a tendency for reliance on short-term contraceptives earlier in the reproductive years and a shift towards a higher frequency use of long-term methods in the middle years. This is potentially evidence of seemingly inevitable differences in health care options and affordability for women of different

backgrounds, with real consequences for those whose reproductive choices may be affected. The apparent trends, along with their variance across standard of living categories, inspired the modeling question of how to best predict contraceptive methods based on other available information about the individual women in the data set.

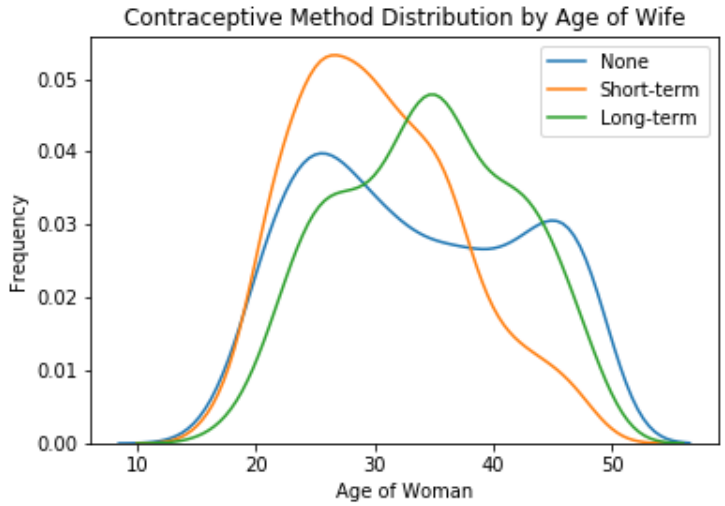

Figure 1. Age Distribution for Contraceptive Types in Women

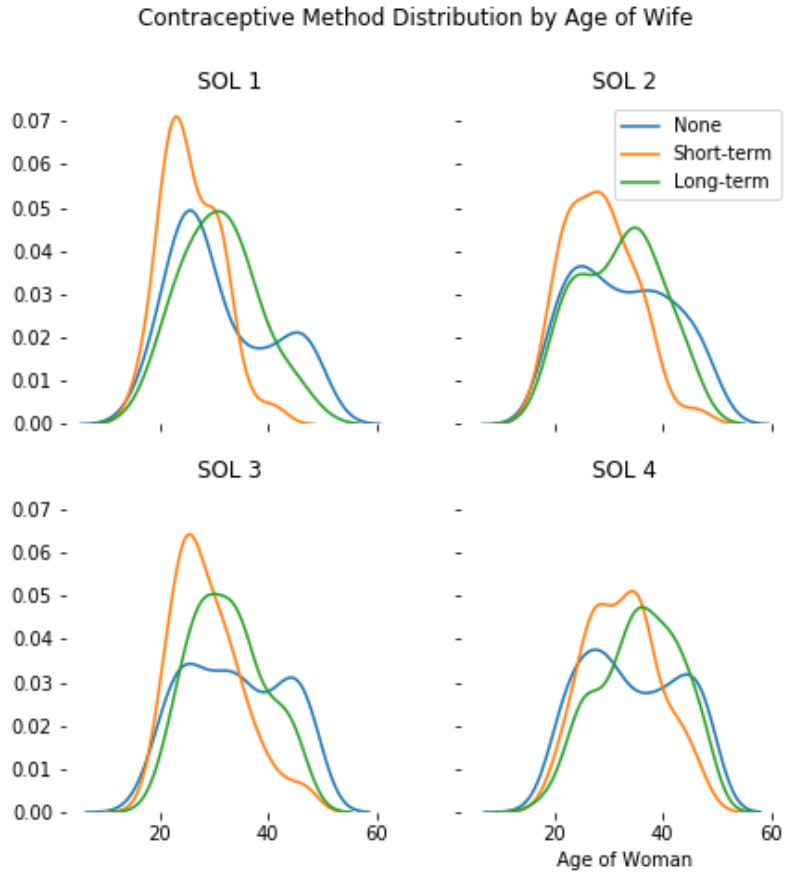

Figure 2. Age Distribution for Contraceptive Types in Women based on Standard of living

Simplifying Figure 2, and only viewing the relationship between standard of living (SOL) and the type of contraceptive used, it's apparent that women who use no contraceptive are twice as likely to be from SOL 2-3 than their counterparts using long or short-term contraceptives (Figure 4). This is a significant finding, seeing as the dataset is weighted towards women of a higher standard of living and education level, implying these results are plausibly even less indicative of disparities than the overall population sample would be. Data from higher standards of living, which correlate with education, may also be more representative than contraceptive data from women in with lower SOL due to a larger sample size, furthering the concern that contraceptive data within the set is less reliable for the women with a low SOL (see Figure 3, Figure 4).

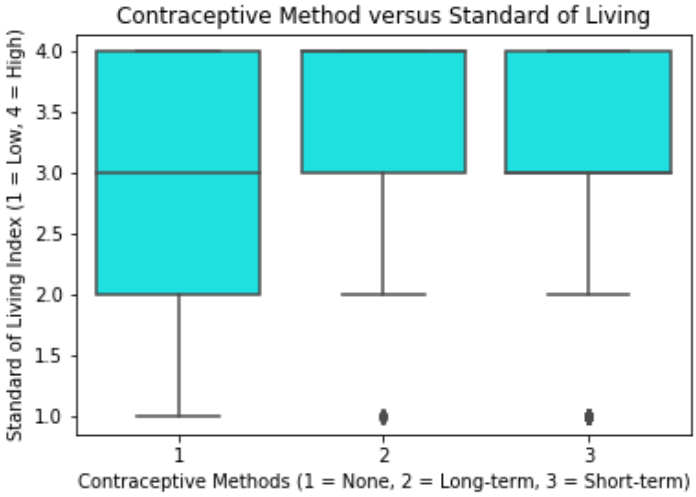

Figure 4. Typical SOL Based on Contraceptive Type

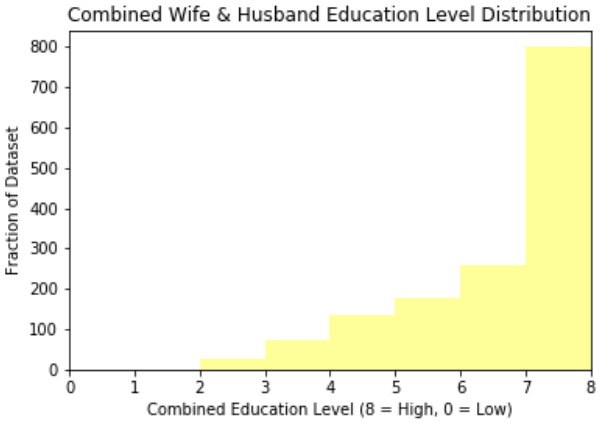

Figure 3. Distribution of Education Level

SOL	Number in Data Set
1	129
2	229
3	431
4	684

Figure 5. Distribution of SOL

Features

The utility of any potential transformation was first diagnosed using a heat map to visualize feature correlation with the target value, the contraceptive method of choice. This revealed generally weak correlations between attributes, with one of the highest being the correlation value -0.16 of the woman's age and contraceptive method. The three categories with the next strongest correlations were media exposure (0.12), the number of children had previously (0.083), and SOL (0.091). It seems media exposure is also a strong proxy for standard of living and education level (0.35 and 0.25, husband and wife's education respectively) in addition to contraceptive methods, making it difficult to distinguish the effects of standard of living and education level from the effects of media exposure alone on contraceptive method choice.

Based on the presence of weak correlations, several features were established in an attempt to transform data and improve the correlations between the existing data and predicted attributes. Features included in the final model included combined education level, accounting in one metric for the combined education level of each woman and her husband, and several indicator features. Indicators were established for women who: were in the lowest category for SOL, were within the ages of 20 and 30 and most likely to have children, and above the age of forty and less likely to have children. Lastly, the number of children a woman has was given a log transformation. This transformation is sensible based on the intuitive assumption that each additional child is less impactful on a woman's life, and on contraceptive choices; the log scale weights this appropriately. These additional features proved to have similar or stronger correlations with the chosen contraceptive than other features, leading to their incorporation in the final model. Lastly, categorical features were modified with one-hot-encoding in order for the model to weight them properly.

Modelling

The modelling process for this experiment involved testing several models on preliminary data, before the addition of features and transformations. Based on preliminary results, it was clear that a Random Forest Classifier provided the best fit to test data and one of the highest cross validation scores, leading to its selection. Random Forests have been established of some of the most accurate estimators of variable importance, an important aspect of fitting to data riddled with weak correlations without overfitting (Breiman 2001, Breiman and Cutler 2000).

Hyperparameters of the model were tuned using a randomized search across 100 different parameter combinations, and evaluating its results against the base model. Fitting again to a range of values set around the results of the randomized search, and again evaluating against the base model allowed for the final selection of parameter values (Koehrsen 2018). Despite a typical accuracy score increase of about 1% from the base model, a complete re-run of the classifier with tuned hyperparameters may show increases in test accuracy upwards of 4%.

Results

The final model used was a Random Forest Classifier with hyperparameters set to: number of estimators = 488, minimum sample split = 4, minimum sample leaves = 4, maximum features = 'sqrt', maximum depth = 50, and bootstrap set to true. **This yielded 70% training accuracy, 55% test accuracy, and a fifteen fold cross validation score of 0.56.** It is notable that other models explored (including various regressions, a decision tree classifier, LASSO, and more) sometimes yielded higher training accuracy than the ending model. However, test accuracies and cross validation scores were higher or the same for the Random Forest, evidencing that other models were probably overfitting to attain their high training accuracy.

Numerical parameters included in the final model included each woman's age, her number of previous children, and the log value of her number of previous children. Categorical values which were converted to binary indicators through one-hot-encoding and then included in the final model are: wife's religion, a low standard of living indicator, ages forty years old, ages between twenty and thirty years old, wife's work status, husband's occupation, overall standard of living, and the combined education level of husband and wife. The inclusion of a woman's media exposure was found to increase training accuracy but decrease test accuracy by up to 5%, leading to its exclusion from the final model.

Discussion

(i) The most interesting features for the predictive question of what contraceptive method a woman will use were selected from those which possessed the strongest correlation with that value. For this experiment those were standard of living, and education level. (ii) One feature which initially seemed relevant, but was not particularly useful, was a woman's religion. This attribute was collected as a binary indicator for Islam, the predominant religion in Indonesia. The vast majority of the women counted were Muslim, weakening the strength of the correlation between religion and contraceptive method choice.

(iii) A challenge inherent to this data set is the low correlation of many collected features with my attribute of interest, the contraceptive method of choice, leading towards an initial modeling tendency to overfit training data. Because of the low correlation values, deciding at what point to include or exclude a feature or data attribute was a difficult decision, and may have not been done as analytically as desired. (iv) Essential limitations of this analysis are data based, preventing the model from reaching high test accuracy, and overfitting test data. Further analysis based limitations include slow processing of hyperparameter searches, preventing broad and repeated runs and affecting the utility of hyperparameter selection. Because of a lack of rigor in the feature and attribute selection, the assumptions made about the utility of features based on intuition and social science knowledge could prove to be incorrect. Specifically, the strength of combined husband and wife's education level as a feature, as well as the exclusion of media exposure from the final model due to findings suggesting it worsened the model's predictive capabilities.

Despite an attempt to follow best practices, the approach taken by this study did not achieve particularly high test or training accuracies. However, this seems to be the case even within other scholarly works, so it seems likely that this rudimentary analysis is relatively good even when evaluated against other, more sophisticated, models of the contraceptive method choice data set (Lim et al. 2000).

(vi) One potential improvement to the understanding gained from this model would be sound understanding of the healthcare system in Indonesia in 1987, including how access to contraceptives is related to income and socio-economic background, as well as other factors. Furthermore, statistics on the distribution of society into different standard of living categories and education levels would be highly relevant for interpreting how representative this particular data set is. In order to explain and explore connections between women of disadvantaged socio-economic backgrounds and lack of contraceptive use, overall population data giving a national or even international context would provide useful and relevant information. Ultimately, if the objective is equitable access to high quality family planning, more qualitative information about why women make the contraceptive choices they do is necessary. Examples of this may include cultural or family norms, lack of access, personal preference, or a multitude of other reasons.

(v) Whenever women's reproduction is considered academically, ethical issues are likely to arise. Primary concerns worth consideration for this data set are how this data and its analysis may be used by the national government collecting it, and that as typical, women's reproductive rights may become a political stance controlled by a ruling body dominated by men. It is impossible to write about such issues from an unbiased perspective, and this analysis should disclose its own stance towards contraceptives as an essential component of women's health

(vii) Beyond the usage of this data set, the manner in which it was collected may also be an ethical concern. It is unclear what form of consent was given by participants, if how their data would be used was fully disclosed, and if they were compensated for participation in any way. In order to address this type of concern, researchers must be transparent about data collection practices in studies with human or animal participants, and a description of the methods used in this case would ideally be reported alongside the data set and available for consideration within this report.

Future work should undoubtedly consider ethical concerns more thoroughly, and incorporate other factors to increase the robustness of the data set. Potential considerations would be demographic data on national rates of poverty or access to contraceptives, information on the healthcare system, and a context of previous and later observations in which to place the results of this study and gain context as to how the factors affecting women's choice of contraceptive may be changing, and why.

References

Breiman, Leo. "Random forests." *Machine learning* 45.1 (2001): 5-32.

Koehrsen, Will. “Hyperparameter Tuning the Random Forest in Python.” *Towards Data Science*, Medium, 10 Jan. 2018.

Loh, W., L. Tim, and Y. Shih. "A comparison of prediction accuracy, complexity, and training time of thirty three old and new classification algorithms." *Machine Learning* 40.3 (2000): 203-238.