Don't Be a Fool, Predict Your Tool

Exploratory Data Analysis and Prediction Model for Contraceptive Use

By Karina Martinez and Jose Jasso Jr.

Abstract

The purpose of this project is to perform exploratory data analysis (EDA) of datasets about contraceptive usage to develop a model that is able to predict which contraceptive method may be used by a woman based on socio economic characteristics. After identifying any correlations found through our EDA process, we were able to generate a model that is able to predict contraceptive method choice with 55.34% training accuracy and 47.43% test accuracy. Based on these results and observations found throughout EDA, we determined that the provided dataset was ineffective and created great limitations in our ability to effectively make predictions due to the inherent biases found in the dataset.

Introduction

Despite the numerous birth controls and contraceptives available for men and women, contraceptive use remains a taboo topic in women's sexual health. What works for one woman may not work for another as birth control side effects vary among many. Furthermore, birth control access and knowledge of remains an issue today as the Trump administration and anti-abortion polititions brought forth the defunding of Planned Parenthood and now an attempt to ban abortions nationwide. Now more than ever is it important to improve current methods of educating folks about sexual health and providing more guidance and clarity among the options available for people who are looking for contraceptives.

In this project, we aim to analyze the choice of contraceptive methods among surveyed women to be able to predict whether short-term, long-term or no contraceptive is the best choice for a woman based on socio economic characteristics they hold. Furthermore, we also analyzed reviews of contraceptives to identify which contraceptive method is preferred by women over

others. All in all, we hope to contribute to women's sexual health by providing more transparency of contraceptives while also highlighting the need for increased sexual health education and reform.

The Datasets

The main dataset we used is a subset of the 1987 National Indonesia Contraceptive Prevalence Survey. Though the actual survey includes various data attributes such as - demographic characteristics, respondent's background, reproduction, knowledge and practice of birth control (family planning), marriage, fertility preferences and husband's background and work, we used the provided dataset which includes the following attributes:

Attribute	**Information**
Wife's age	Numerical value
Wife's education	Categorical: 1=low, 2, 3, 4=high
Husband's education	Categorical: 1=low, 2, 3, 4=high
Number of children ever born	Numerical value
Wife's religion	Binary: 0=Non-Islam, 1=Islam
Wife's now working	Binary: 0=Yes, 1=No
Husband's occupation	Categorical: 1,2,3,4
Standard of living index	Categorical: 1=low, 2, 3, 4=high
Media exposure	Binary: 0=Good, 1=Not good
Contraceptive method used	Categorical: 1=No-use, 2=Long-term, 3=Short-term

Table 1: Attributes for the National Indonesia Contraceptive Prevalence Survey Dataset

```
contraceptive_data = pd.read_csv("https://github.com/kmart67/data_100_contraceptive_final_proj/raw/master/data/contrace
contraceptive_data.head()
```

	wife_age	wife_education	husband_education	num_child	wife_religion	wife_work	husband_occupation	standard_living	media_exposure	contraceptive
0	24	2	3	3	1	1	2	3	0	1
1	45	1	3	10	1	1	3	4	0	1
2	43	2	3	7	1	1	3	4	0	1
3	42	3	2	9	1	1	3	3	0	1
4	36	3	3	8	1	1	3	2	0	1

Figure 1: National Indonesia Contraceptive Prevalence Survey Dataset pre EDA

Furthermore, we also created our own dataset of all contraceptive reviews listed on WebMD along with the reviewer age range, contraceptive name, review comment, and treatment length. According to the survey program, "long-term methods [defined in the survey] are male or female sterilization, Norplant and IUO, while temporary are all others." Since female and male sterilizations are not drugs, we only marked norplants and IUO/IUDs listed on WebMD as long term and all others as short term. The final dataset created contained the following attributes:

Attribute	**Information**
Drug Name	String
Reviewer Age Range	Categorical Strings: '13-18', '19-24', '25-34', '35-44', '45-54', '55-64', '65-74'
Treatment Length	Categorical Strings denoting the treatment length
Comment	String
Long Term	Binary: 0=No, 1=Yes
Short Term	Binary: 0=No, 1=Yes

Table 2: Attributes for WebMD Contraceptive Reviews

```
contraceptive_reviews = pd.read_csv("https://github.com/kmart67/data_100_contraceptive_final_proj/raw/master/data/webmc
contraceptive_reviews.rename(columns={'Unnamed: 0': 'Review_Id'}, inplace=True)
contraceptive_reviews['Comment'].fillna("", inplace=True)
contraceptive_reviews.head()
```

	Review_Id	Drug_Name	Long_Term	Short_Term	Comment	Reviewer_Age_Range	Treatment_Length
0	0	Mirena Intrauterine Device	1	0	I did initially experience excruciating, breat...	25-34	5 to less than 10 years
1	1	Mirena Intrauterine Device	1	0	Memory Loss!!! Never again on any form of birt...	35-44	2 to less than 5 years
2	2	Mirena Intrauterine Device	1	0	Horrible experience would not recommend to nob...	NaN	NaN
3	3	Mirena Intrauterine Device	1	0	I had a Mirena placed in me in 2006. I had maj...	45-54	5 to less than 10 years
4	4	Mirena Intrauterine Device	1	0	The Mirena IUD was the worst thing to ever hap...	19-24	2 to less than 5 years

Figure 2: WebMD Contraceptive Reviews Dataset

All methods used for review and reviewer information extraction can be seen in our *Contraceptives_Reviews.ipynb* file in our GitHub repository. When creating this dataset, we hoped to identify any correlations between reviewer information, treatment length, and review comment as well as see if we are able to utilize it to our advantage for our model.

Exploratory Data Analysis

Before performing any data analysis on the survey dataset, we first one-hot encoded all categorical attributes. Since the provided dataset did not indicate what null values represented, we decided to drop any null values to ensure our model is able to predict the most efficiently. We then created the following correlation map from the one hot encoded dataset:

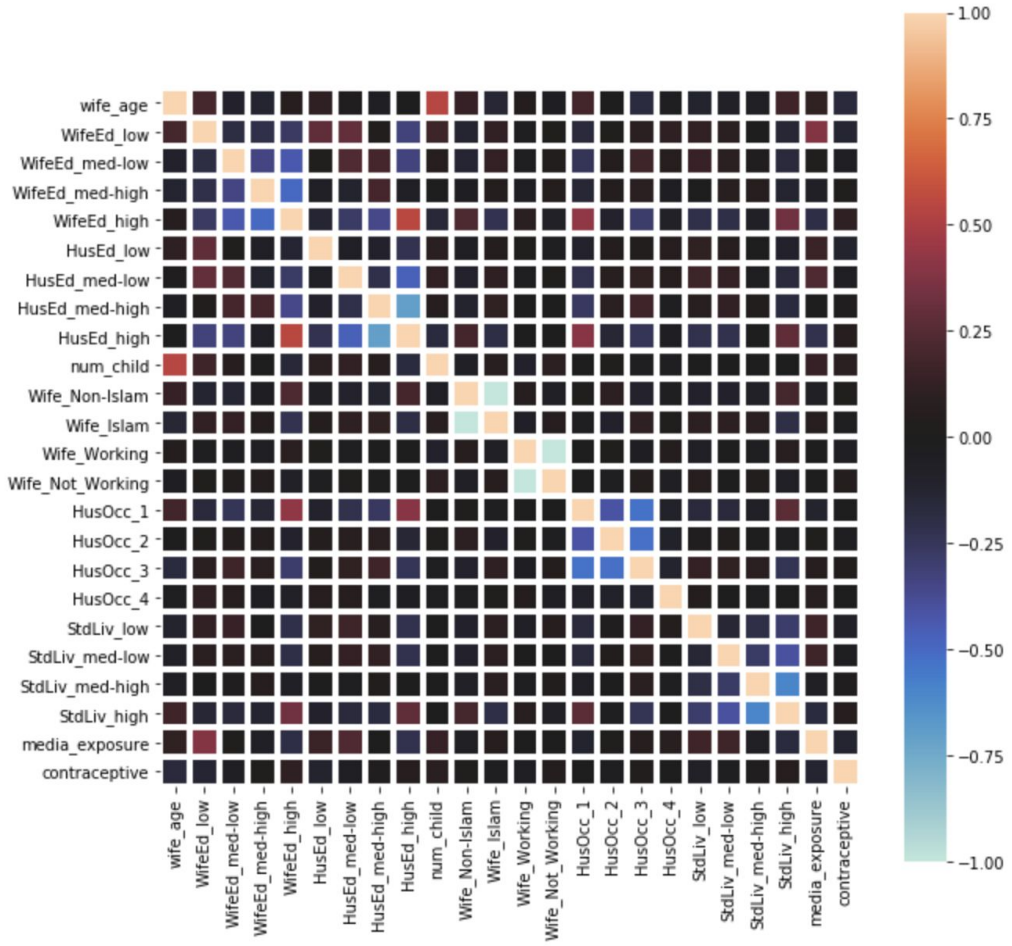

Figure 3: Survey Dataset Attribute Correlations

As we can see, the majority of attributes do not correlate with each other. This ends up being noteworthy in demonstrating how accurate our model was actually able to perform - explained more in the next section.

Furthermore, we were interested in seeing the correlation between wife's age and long term contraceptive usage. Below are two plots indicating the number of women who used a long term contraceptive according to their age and age range:

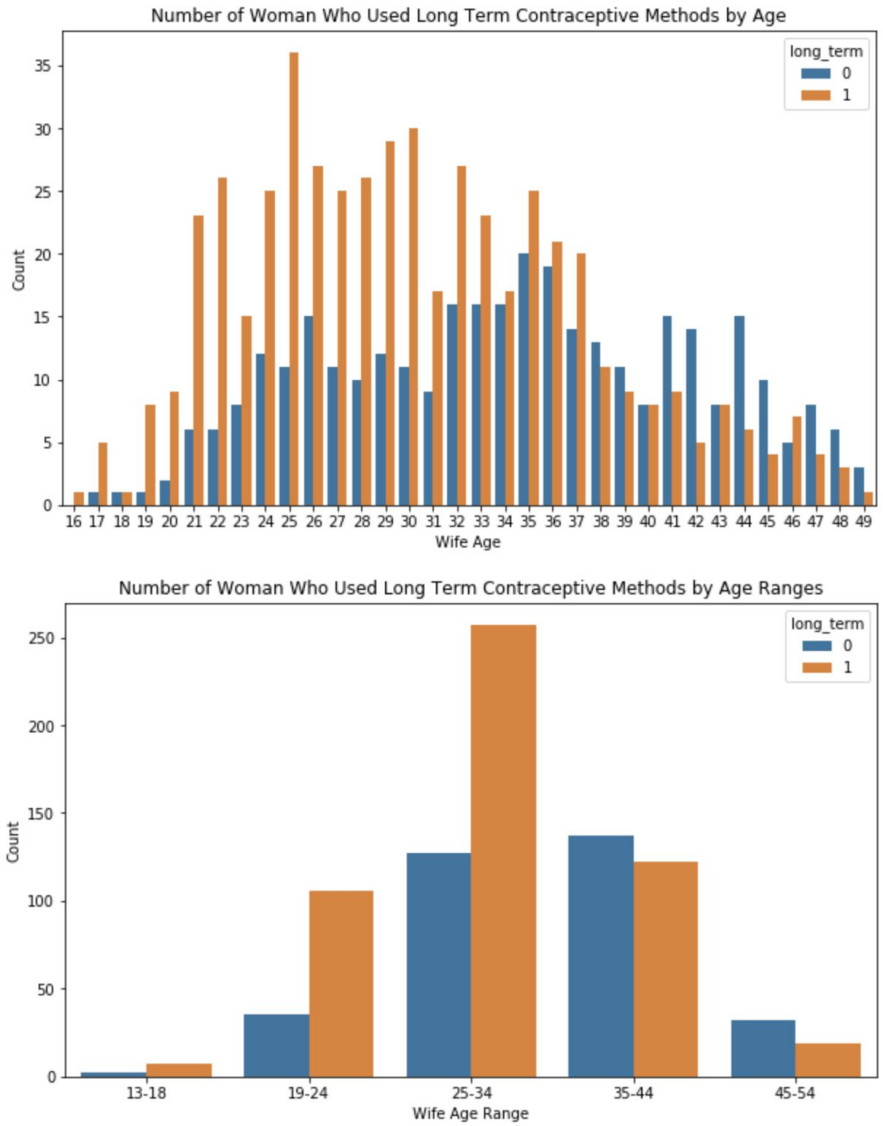

Figure 3: Surveyed Women's Age and Their Choice of Contraceptive Method

To provide consistency, we used the same age ranges as in the WebMD dataset. As we can see, women who were surveyed that were under the age of 38 greatly preferred long term contraceptives over short term contraceptives. Since all women surveyed were heterosexual married woman, it is important to note the bias demonstrated through these plots. Due to the lack of single and homosexual women represented in this dataset - in other words, due to the selection

bias of this dataset - we cannot confidently say that all woman under the age of 38 prefer long term methods over short term methods.

From the correlation heatmap above, we noticed a hotspot demonstrating that the wife's age positively correlated with the number of children ever born. We decided to explore that more through this joint plot:

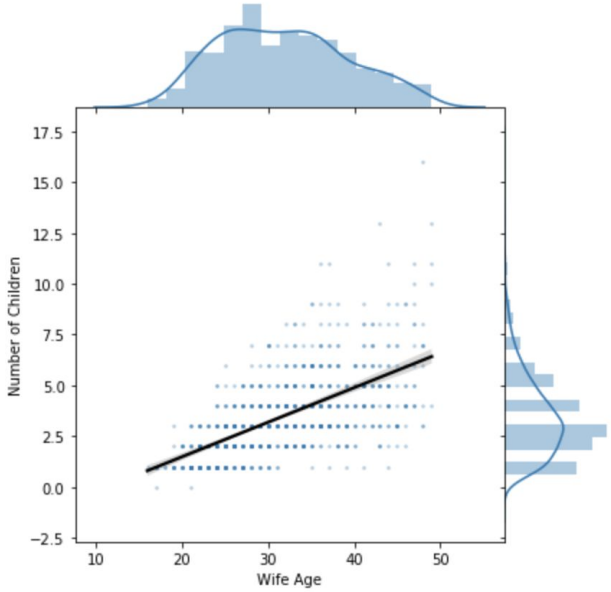

Figure 4: Joint plot between wife's age and the number of children

As we can see the greater the wife's age, the more children she bears. Unfortunately the dataset lacks to provide us the husband's age and we are unable to conclude whether the same correlation remains true for men.

Now, before performing exploratory data analysis of the WebMD contraceptive reviews dataset, we first cleaned up the dataset by filling null comments with empty strings and dropping any rows with null age ranges and treatment length. Since the WebMD platform did not give users the ability to rate drugs, or contraceptives in this case, on a numerical scale, we decided to utilize VADER used from Homework 4 to determine the polarity of the review comments. To

calculate the polarity, we followed the same steps as we did in the homework to calculate the polarity of the comments then merge it to the original dataframe to generate a new polarity column. We then generated the following plot to identify the difference in review polarity among long term and short term contraceptives:

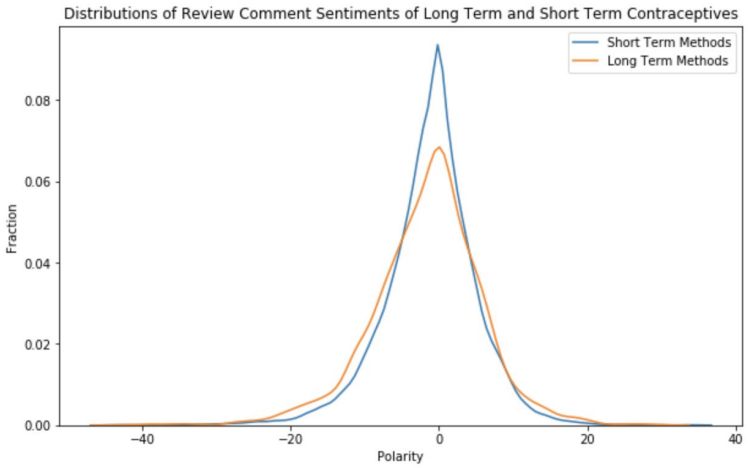

Figure 5: WebMD Contraceptive Reviews Polarity of Short Term and Long Term Contraceptives

Unfortunately, the plots for both contraceptive methods are not significantly different therefore, we also generated another plot to identify the correlation between review length for both contraceptive methods:

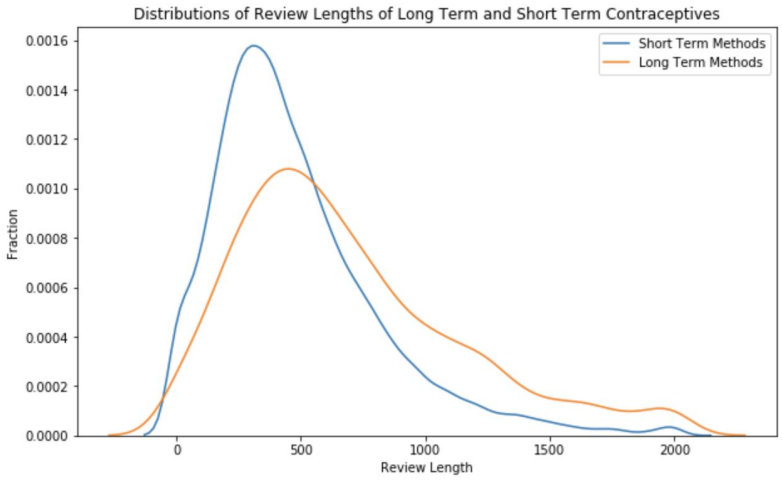

Figure 6: WebMD Contraceptive Review Length of Short Term and Long Term Contraceptives

As we can see, the plot for long term methods is slightly more skewed to the right indicating that long term methods had longer reviews than short term contraceptive methods. We also wanted to identify if treatment length expressed any correlation with contraceptive method choice, as shown by this plot:

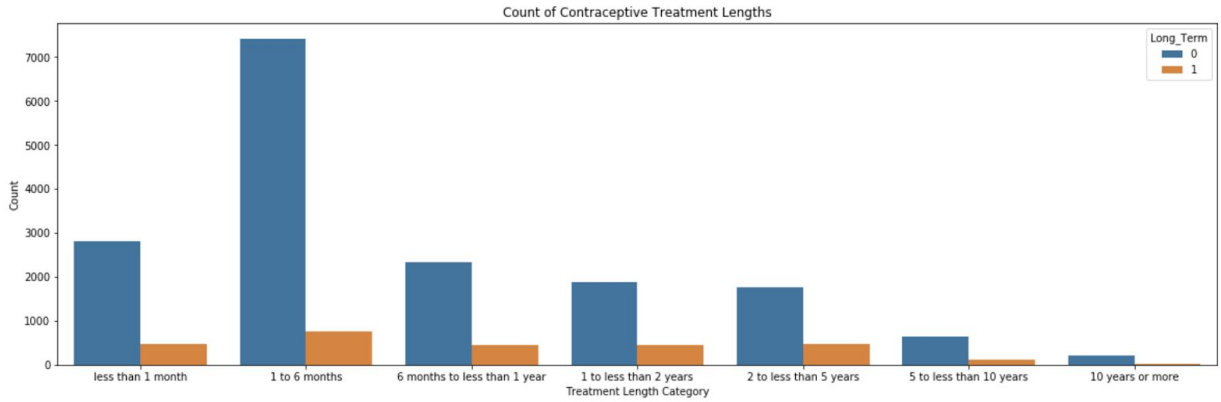

Figure 7: WebMD Contraceptive Treatment Lengths

Although treatment length for long term contraceptive methods didn't vary widely, we see that short term methods were most commonly used for 1 to 6 months.

As mentioned before, the survey dataset presented a bias towards married women. On the other hand, WebMD is a resource available to all women as long as they have access to the internet. Therefore, we found it interesting to see if the age range and method choice plot was any different for our WebMD dataset compared to the survey data.

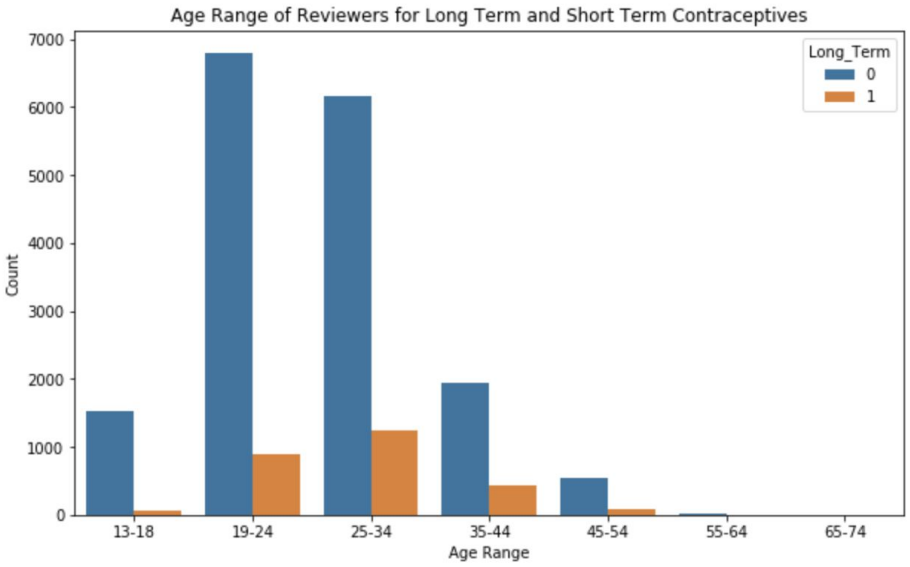

Figure 8: WebMD Reviewers Age Ranges

Interestingly, this plot is far more skewed to the right compared to the provided dataset, expressing the inherent bias that exists since WebMD is not solely for married women. Although the survey plot indicated that women under the age of 37 preferred long term methods over short term methods, this plot indicates otherwise. This may be due to multiple reasons including lack of knowledge of long term contraceptives, accessibility, and affordability.

Results & Discussion

Due to the conflicting results of our datasets and the inability to easily merge the two datasets together, we decided to move forward with creating a model solely based on the survey dataset. For this model, we prepared the data by One-Hot Encoding all categorical attributes then performing a test-train split to be able to play around with our model. Our final model included all one-hot encoded columns along with original columns that were not one-hot encoded, resulting in a 55.34% train accuracy and 47.43% test accuracy. Though we were able to fit a

model that generated a 99% train accuracy, we found that it greatly overfitted the train data and performed badly in testing data.

There were many challenges and limitations faced when developing our prediction model. As mentioned throughout the paper, there was a selection bias among the survey dataset that inherently became our greatest limitation in creating an effective prediction model. Biased data equals bad data equals bad model. Despite spending numerous hours searching for any datasets surrounding women's sexual health, family planning, and/or any other relevant topics, we reached a dead end as we realized we would not be able to attempt to *fix* the given bad dataset. Future work is desperately needed to not only improve this model but also improve the state of available data for folk in similar research fields. This can be through outsourcing more data that is not currently readily available, changing the way in which this survey is conducted by opening the survey to *all* women - single, married, cis-, trans-, heterosexual, homosexual, etc. women. This really sheds light to the need for more transparency and accessibility to information regarding women's sexual health beyond this outdated and biased datased. We hope that through this project, we are able to highlight the ethical responsibility for folk to come together to address inaccessibility and lack of knowledge of sexual health.

Resources Used

Central Bureau of Statistics Jakarta, Indonesia, et al. *National Contraceptive Prevalence Survey 1987*. Jan. 1989, dhsprogram.com/pubs/pdf/FR19/FR19.pdf?fbclid=IwAR2qrdVasVTj8gxC4N9G6vLXhmVG4bl4DdVIoJ2K3ATlMvcTPjyB67d8Mh4.

"Drugs & Medications A-Z." *WebMD*, WebMD, www.webmd.com/.

Dua, D. and Graff, C. (2019). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.