Yucong Ji, Jianxuan Yin

Data 100 Final Project

11 May, 2020

Social Studies with Data Science

Abstract

For this project, we are given a subset of the survey data from 1987 National Indonesia Contraceptive Prevalence Survey. We proposed two sets of questions and hoped to answer them with analysis of this dataset. For the first set, we investigated factors influencing women's fertility rate. We successfully built a linear regression model to predict the number of children a woman has based on her other information and found out that factors like a woman's education level could greatly influence the number of children she has. For the second set, we investigated factors influencing whether a woman is in the workforce. We tried using several models, including logistic regression, decision tree, and random forest but to no avail. After some analysis, we concluded that it was extremely difficult to predict a woman's employment status with the data given alone. To tackle this issue, we would need more information.

Introduction

Both fertility rate and female employment rate are important issues in a society. The former determines the population and social structure of a society, and the latter is a significant metric for evaluating women's rights. The dataset we are given includes personal information of over 1000 married women in Indonesia in 1987. While this dataset was originally meant to evaluate contraceptive methods used by these women, it does include many information that could influence female fertility and female employment, such as women's education levels and livings standards. We proposed two sets of questions. The first set is: Among the information collected, what are the factors that best determine the number of children a woman has? Can we build a model to accurately predict this information? If so, what does this inform us about the fertility rate in society in developing countries? The second set is: What are the factors that best help to classify whether a woman is working? Can we accurately classify the employment status of a woman based on her other information? If so, what does this tell us about what encourages or discourages a woman from joining the workforce?

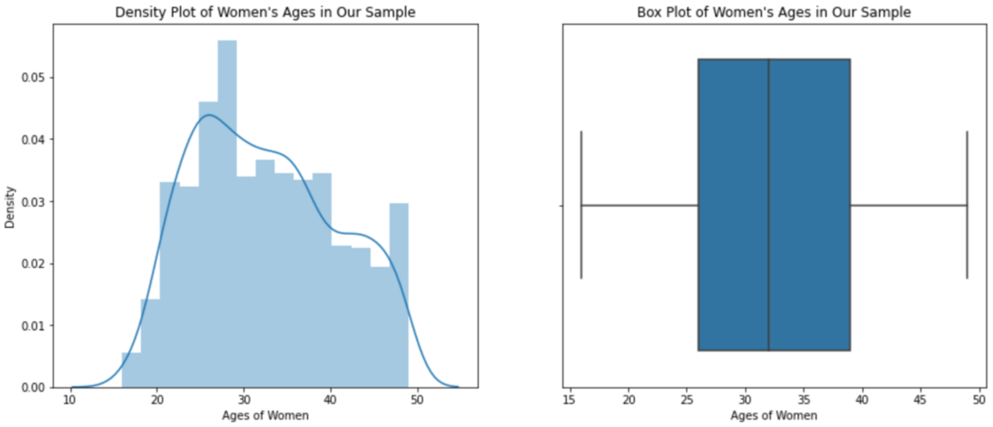
Description of Data

In our original dataset, there are 1473 rows and 10 columns, each row represents a specific woman in the survey, and each column represents some specific information about the woman. First, we investigated whether there are data transformations we need to make to our dataset.

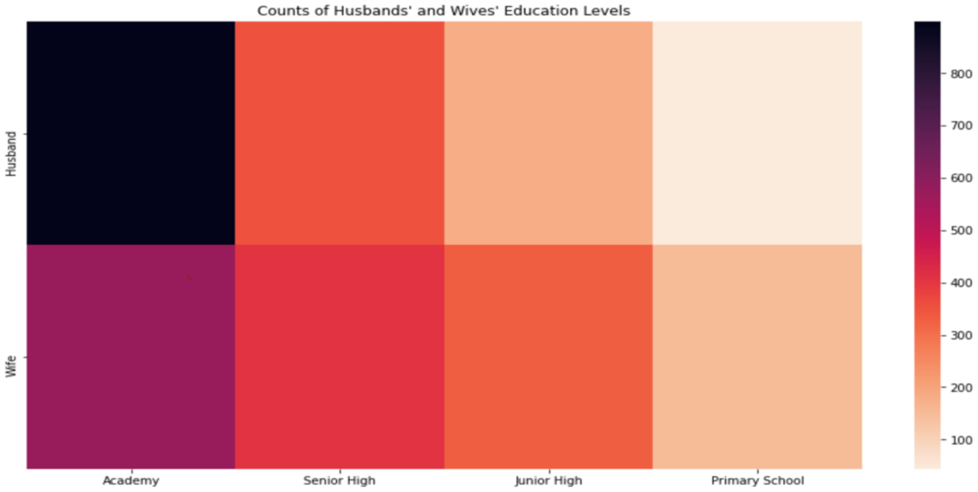
1. Missing values: We found there is not a single missing value in our dataset. So no imputation is needed.
2. Variable Types: In our dataset, there are 10 columns (variables), and only two of these, woman's age, and number of children a woman has, are numeric variables. The rest of the categorical variables, however, are all encoded with numbers, which is confusing and hard to interpret. Also, as we later build regression models on the data, we could use one-hot encoding to split categorical variables, increasing the number of columns and our level of confusion. Thus, we cleaned all the categorical variables by using texts that represent the column's meanings to replace numbers. For example, 4 in the column of husband occupation becomes "Agriculture".

After some initial cleaning, we moved on to make visualizations about distributions of values in different columns to understand the demographics of our sample before building models to investigate our issues.

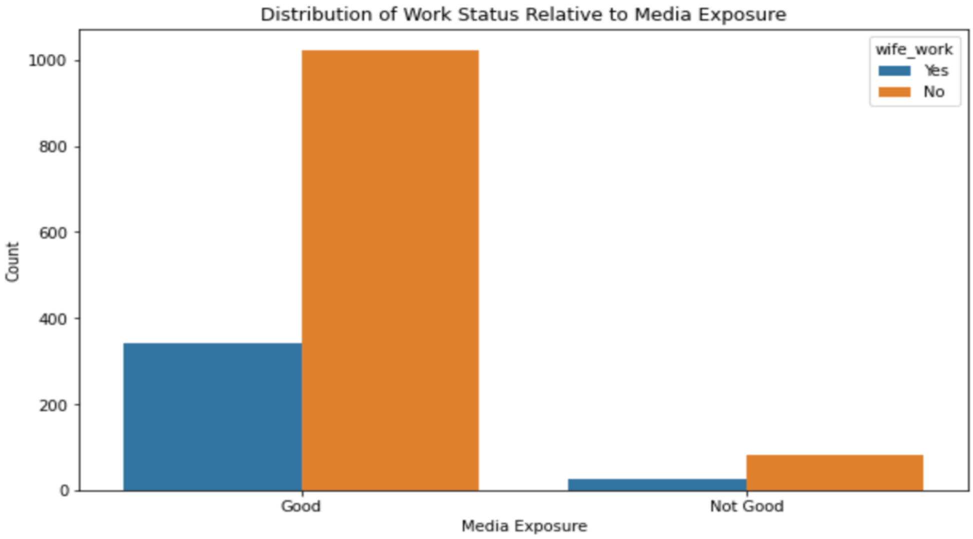
1. Ages of women:



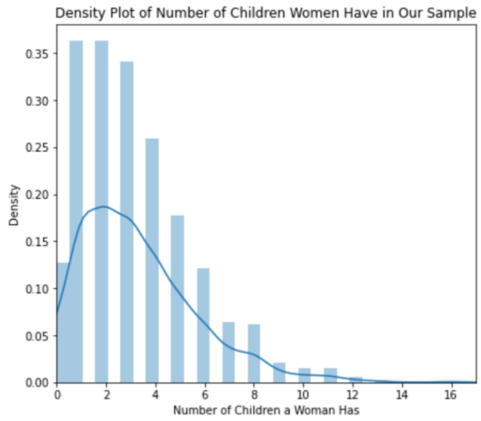
2. Education levels of wives and husbands:

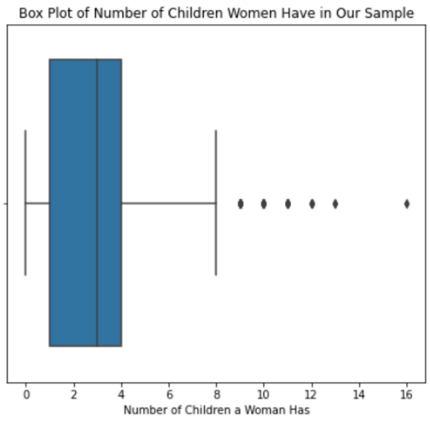


3. Work status and media exposure of women:

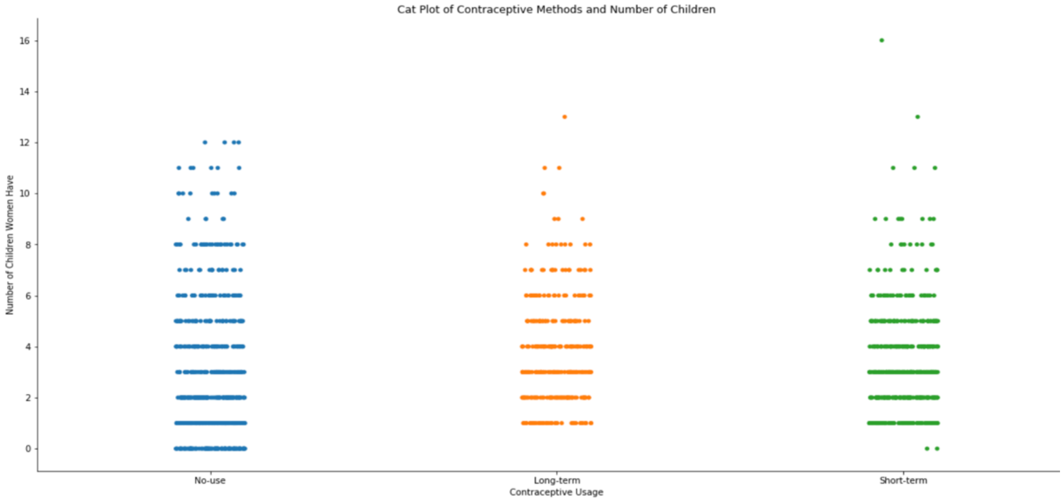


4. Number of children women have:

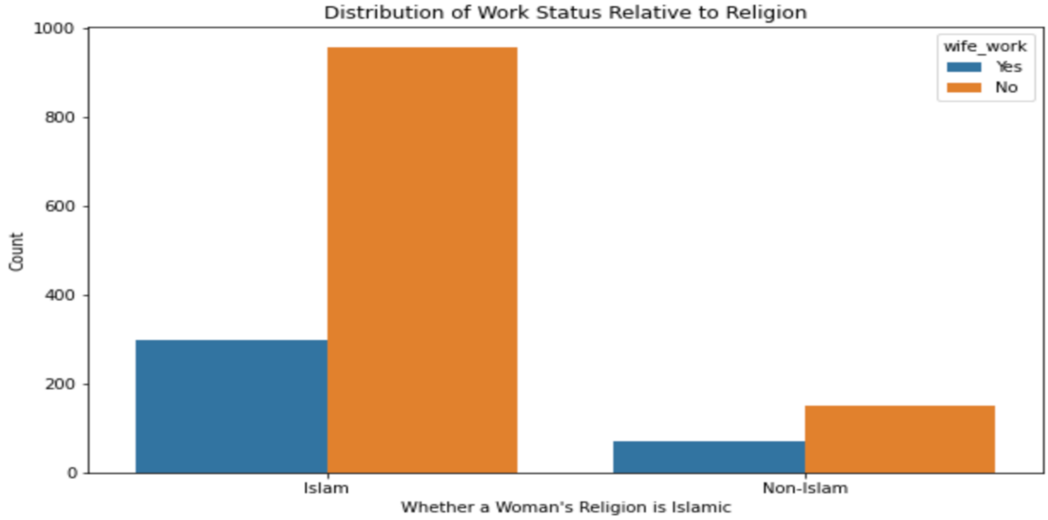




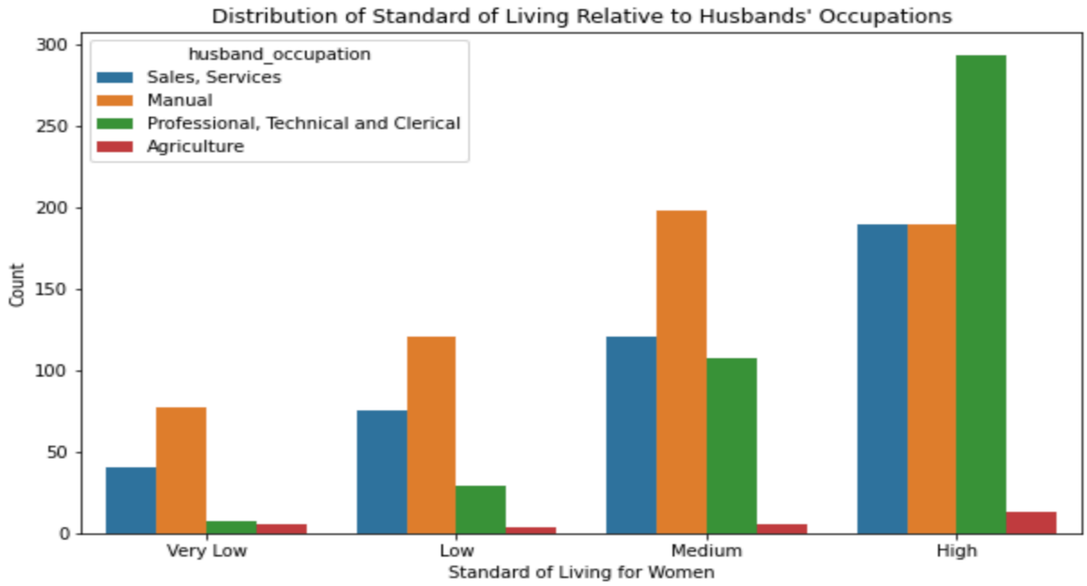
5. Number of children and contraceptive methods



6. Work status and religion of women

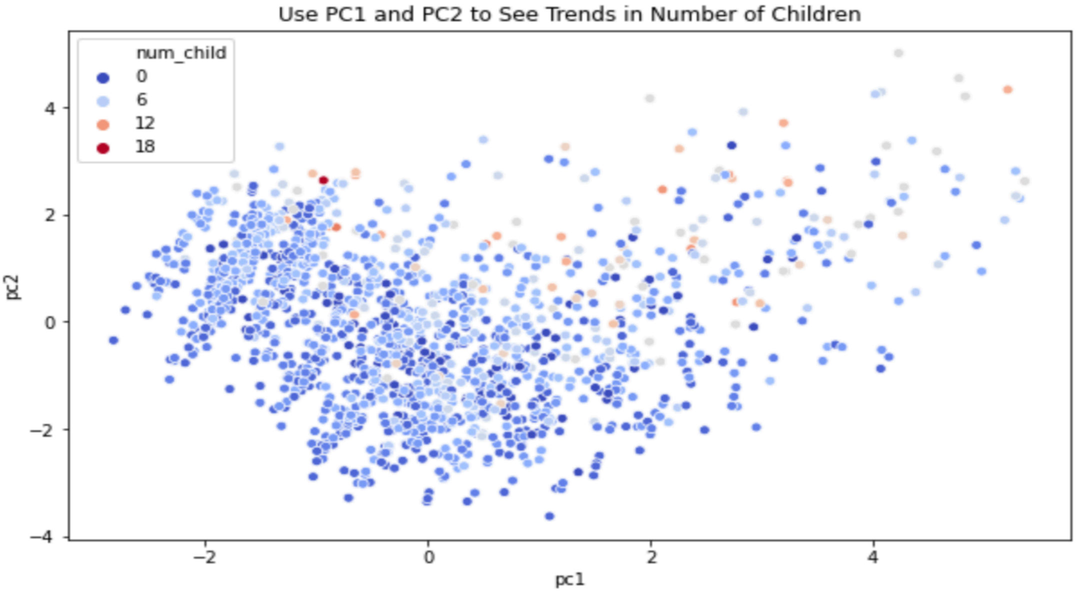


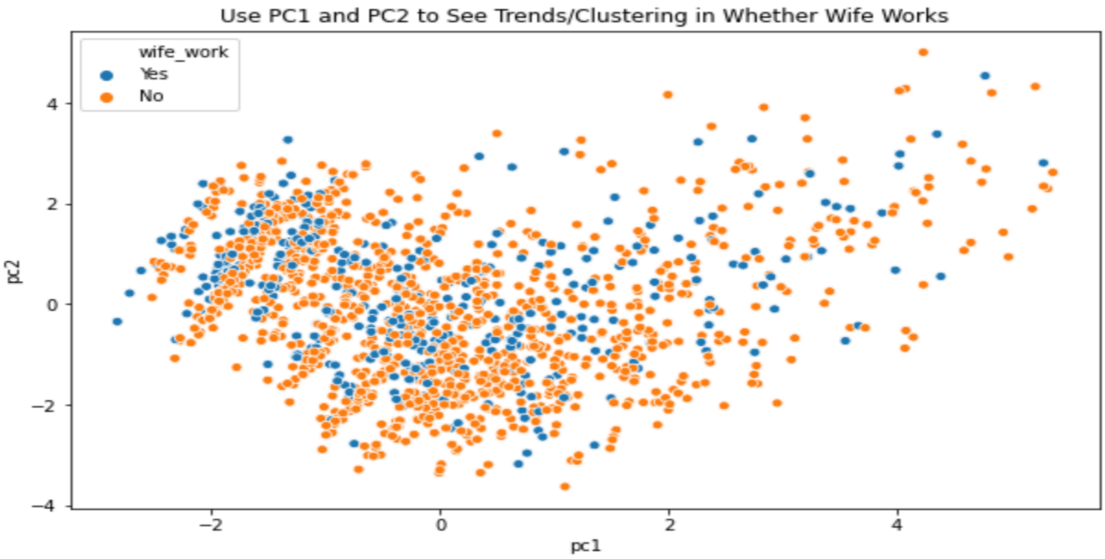
7. Husbands’ occupations and living standards of women



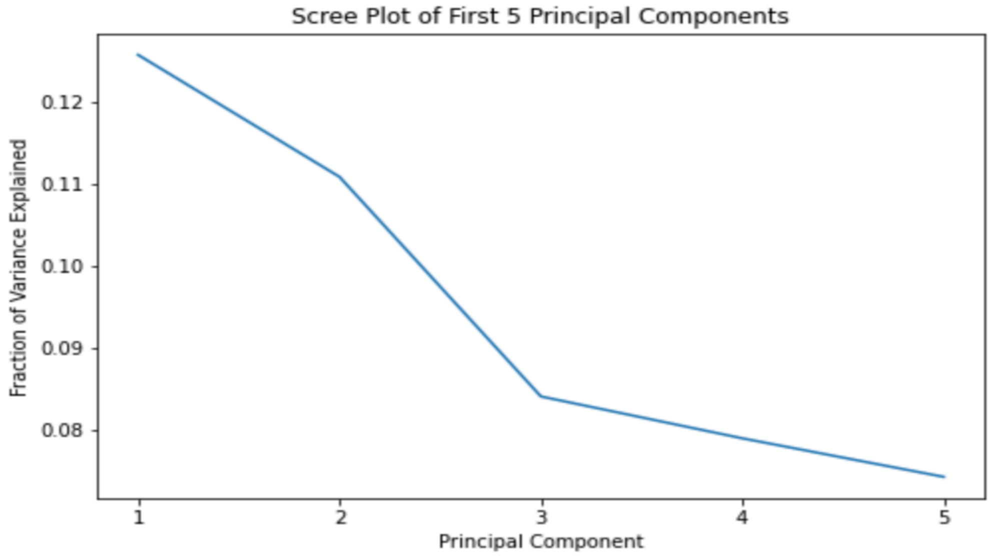
After these data visualizations, we now have some images of what our sample looks like. They are married women with ages from 15-49 in Indonesia in 1987. Most of them are in their 20s to 40s. The average education level is around senior high school, but not as many of them studied at academy-level institutions compared to their husbands. The media exposure rate among them is really good, but only a few of women in the sample are in the workforce. Most of the women have less than 5 children and believe in Islam. However, non-Islamic women seem to be more likely to work than Islamic women. Most of these women have medium to high living standards. Except in the group with high living standards, most of these women are married to husbands that do manual work. Many of them use long-term or short-term contraceptive methods. But the vast majority of women with no children use no contraceptive method.

Since there are 10 features in our dataset, it is hard to plot and interpret their relationships with number of children and employment status one by one. Thus, we went on to do a principal component analysis. First, we did one-hot encoding to transform our categorical data to dummy variables, because PCA requires numeric variables. If we successfully reduce dimensions to only two variables, we can then see clearly on a 2-D plot the trends or clustering in our data.





We extracted the first two principal components, plotted them with a scatterplot, and colored the points with colors representing number of children and employment status. However, it seems that we couldn't see any trends/clustering with the variables we are trying to classify and predict.



We went on to draw a screen plot showing the fraction of variance captured by the first five principal components. According to this plot, the first two principal components together only captured less than 25% of the variance. So reducing dimensions for this dataset is not a good strategy since it removes too much variance.

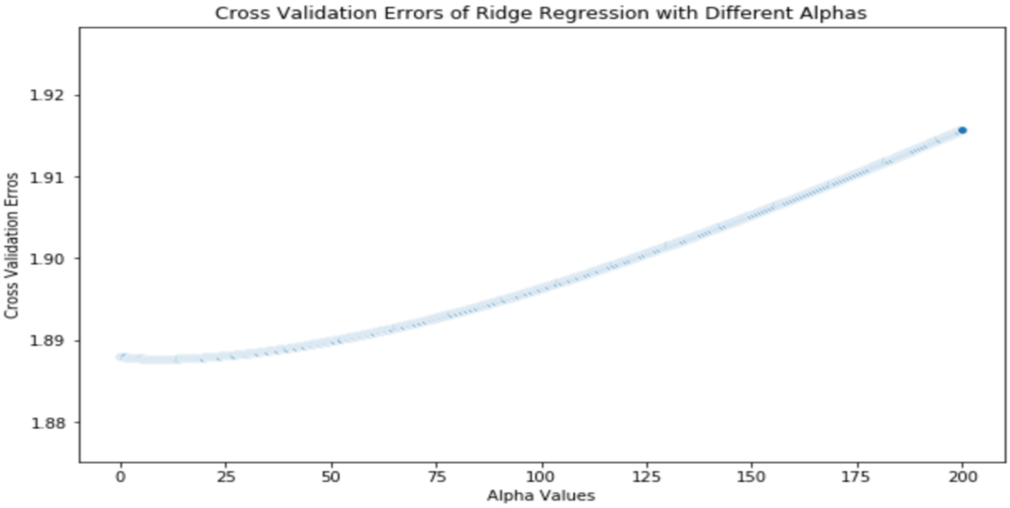
Description of Methods

Fertility: To predict the number of children a woman has, we decided to do linear regression for interpretable result. Below are the steps and the rations behind them.

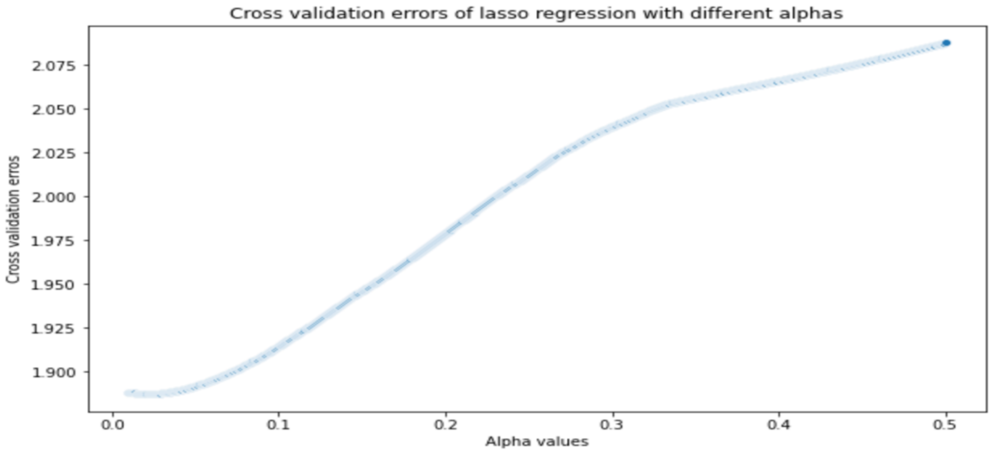
1. Train/Test split and error definition: According to stack overflow, usually 8/2 is a good split.[1] However, our dataset has only 1473 rows. 20% means less than 300 women. A test set that small could fail to represent the feature space. Thus, we randomly chose 30% of our data to be the test set. For error definition, we used the traditional error estimation method for linear regression, root mean squared error (RMSE).

[1] https://stackoverflow.com/questions/13610074/is-there-a-rule-of-thumb-for-how-to-divide-a-dataset-into-training-and-validation

2. Base model: We built a base model, so our linear regression model has a benchmark to compare against. Our base model will always predict the median number of children of all women in the train set, regardless of other information. It has a train error of 2.43, and a test error of 2.23.
3. Cross validation: After building our linear regression model, we used 5-fold cross validation to select features. We first used only the leftmost feature, then the first 2 features on the left etc. We calculated cross validation error for each iteration. After considering every possible number of leftmost features, we shuffled the columns and did the same process for 1000 times. We recorded the combination of features that gave us the lowest cross validation error. Our cross validation considered 1000 * 18 (number of columns after one-hot encoding - 1) = 18000 possible combinations of features. Our final selected model has a train error of 1.88 and a test error of 1.76.
4. Regularization: We also tried to use Lasso and Ridge regression models to select features. If there is an overfitting, using regularization can potentially give us better performances. For Ridge regression model, we tried 1000 different alpha values from 0 to 30 and selected the alpha with lowest cross validation error.



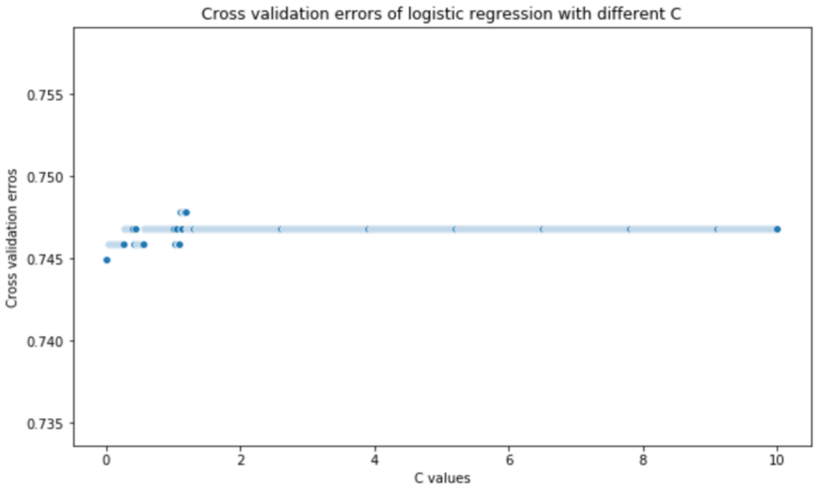
As shown in the plot, the optimal alpha for Ridge regression is around 9. And the train error and test error outputted by our best Ridge model is 1.88 and 1.79. We also did the same parameter tuning process for Lasso regression, examining cross validation errors with 1000 alpha values from 0.01 to 0.5.



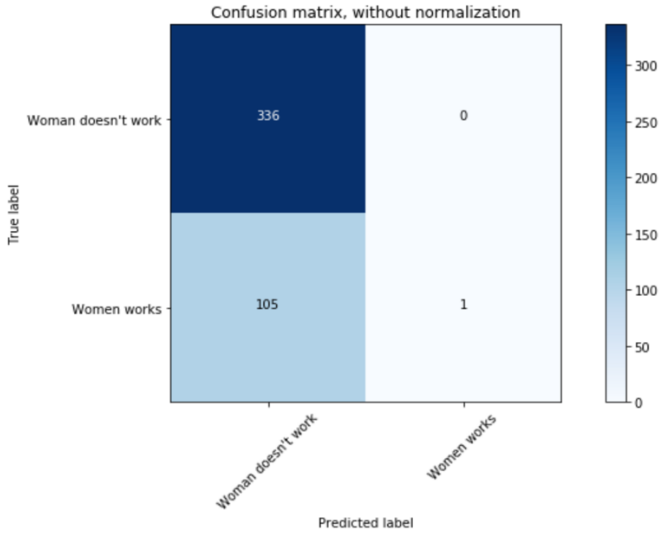
The optimal alpha value for our Lasso regression model is around 0.02. With our best Lasso model, the train and test error are 1.88 and 1.77. Using regularization, in this case, didn’t produce a better result.

Employment: For this topic, our goal is to classify whether a woman works based on her other information. We tried several models, including logistic regression, decision tree and random forest. All of these models are suitable for classification tasks.

1. Train/Test split and evaluation metric: The method of train test split is the same as we did when studying fertility. The metric we used to evaluate our models is classification accuracy, which is the percentage of observations our model guessed right.
2. Base model: As our previous plot shows, majority of women in our sample don't work. Thus, a simple base model is to predict woman not working for all cases. This model has a train accuracy of 0.74 and test accuracy of 0.76.
3. Cross validation and regularization: We first built our logistic regression model. We used the same cross validation process as we did when studying fertility. For logistic regression in sklearn, we adjusted parameter C to change regularization strength. We evaluated 500 C values from 0.01 to 10.

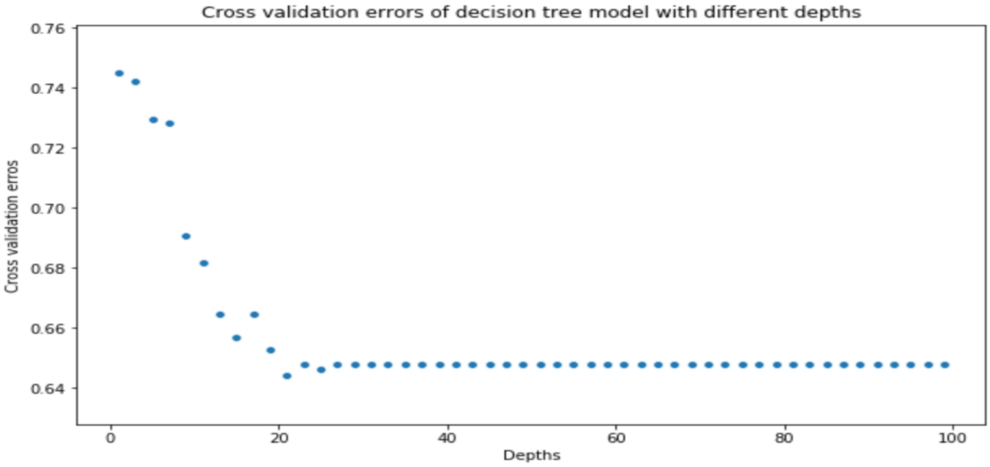


Our optimal value for C is around 1, which is just the default value for C. With train and test accuracy of 0.74 and 0.76, our tuned model shows no improvements from the base model. To further evaluate this, we looked at the precision and recall of our model. Here's the confusion matrix for our predictions for the test set:



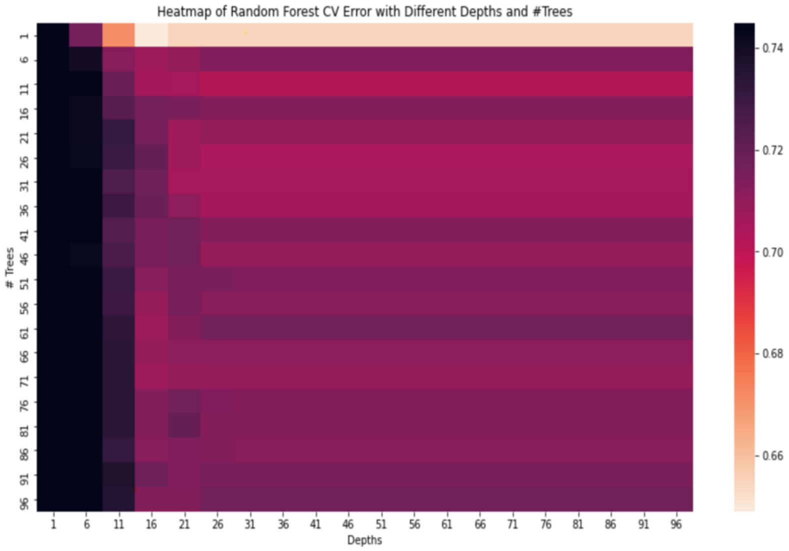
Looking at the graph, our logistic regression model's behavior is almost the same as our base model. There are almost not any positives in our predictions, meaning almost no true positive, false positive or false negative. So evaluating precision and recall here won't give us much information.

4. Decision tree: We moved on to building and evaluating a decision tree model. Without pruning, the decision tree model has a train accuracy of 0.99 and a test accuracy of 0.64. Apparently, our model is overfitting. Thus, we changed the maximum depth our tree is allowed to have. We tried around 50 depths from 1 to 100, and calculated 5-fold cross validation accuracy for each depth. We selected the model with highest CV accuracy.



Our cross validation accuracy is highest when depth is 1, meaning that our model is overfitting a lot. At depth 1, our model has a train accuracy of 0.74 and a test accuracy of 0.76. Unfortunately, the performances are the same as those of our base model.

5. Random forest: We then tried random forest model. Without tuning parameters, we have a train accuracy of 0.99 and a test accuracy of 0.72. It is also overfitting. To tune our model, we tried 20 different numbers for maximum depth each tree is allowed to have and 20 different numbers for number of trees, resulting in a total of 400 combinations. We calculated 5-fold CV accuracy for each combination of parameters:



In this heatmap, the deeper the color is, the higher the cross validation accuracy is. We realize that to improve our CV accuracy, we have to prune our tree so that it only has a few levels. After pruning, number of trees almost has no impact. Using our tuned random forest model, both train and test accuracy becomes 0.75. This is no noticeable improvement from our base model.

Summary of Results

Fertility: After building our models, we went on to try to answer the questions we framed.

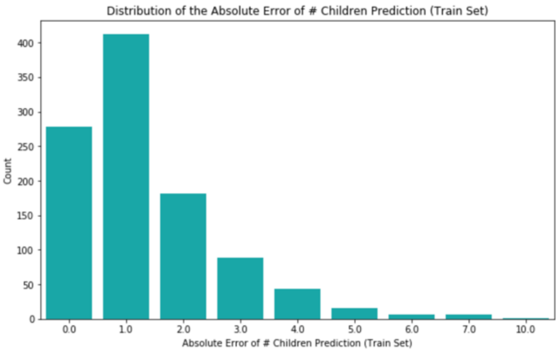
1. What are the factors that best determine the number of children a woman has, based on her other information?

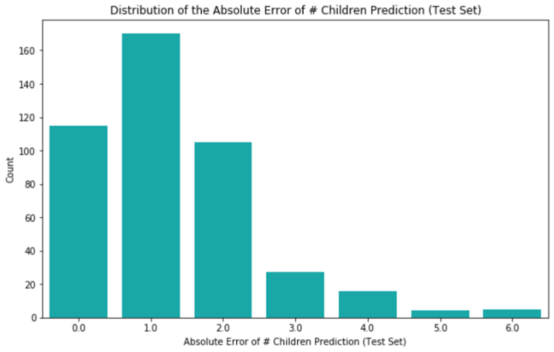
We printed out the features selected by our best linear regression model. They are the standard living index of a woman, whether the husband's education level is Senior High or Junior High, the types of contraceptive methods used, the woman's education level, whether the woman works, whether the woman's religion is Islamic and the age of the woman. Noticeably, features about husband's occupation or woman's media exposure are not selected. The selected features can be considered factors that best determine the number of children a woman has.

2. Can we build a model to accurately predict this information?

   To further examine the performances of our best linear regression model, we calculated the proportion of predictions that are correct. 27% of predictions are exactly right in the train set, and 26% in the test set. Moreover, we plotted the absolute differences between our predictions and the actual numbers in train set and test set.

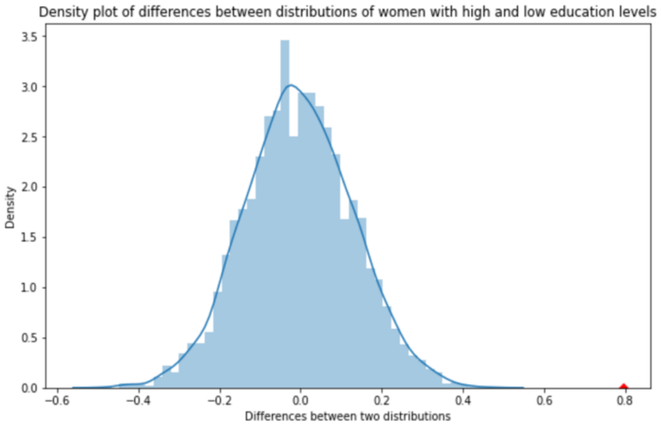
   We found that in both the train and the test set, most of the predictions are at most 1 off. Under the circumstance, we are willing to say that we can build a model to accurately predict the number of children a woman has.





3. If so, what does this inform us about the fertility rate in our society?

   We believe that the features selected all have some impact on the fertility of a woman and looked at the coefficients assigned to each feature to see how much impact they have. Two of the most important features are whether the woman's education is primary school or junior high (There are 4 levels of education in the dataset, the rest 2 are senior high and academy). According to our model, the higher education a woman has, the less children she tends to have. Is this really a reasonable conclusion to draw on a broader level? We decided to do A/B testing using a P-value of 0.01. The null hypothesis is: In the population, the distribution of number of children for mothers is the same for mothers who have high-level education as mothers who don't. The difference in the sample is due to chance. The alternative hypothesis is: In the population, the number of children of the mothers who have low level education is more, on average, than the number of children of the mothers who have high-level education. The test statistics is: Average # of children of the mothers who have low education levels − average # of children of the mothers who have high education levels. We define high education level as senior high or academy, and low education level as junior high or primary school. We did 5000 simulations under the null hypothesis and got a p-value of 0!



This is sufficient evidence for us to reject the null hypothesis. Thus, a low education level is related to the number of children a woman has and knowing the education level of a woman is beneficial for predicting her number of children. On a broad level, for developing countries like Indonesia, as the education levels of women improve, fertility rate is likely to drop. For countries that want to reduce population pressure on environment and society, improving the education levels of women may be a good way.

Employment: After building our models, we went on to try to answer the questions we framed.

1. What are the factors that best help to classify whether a woman is working?

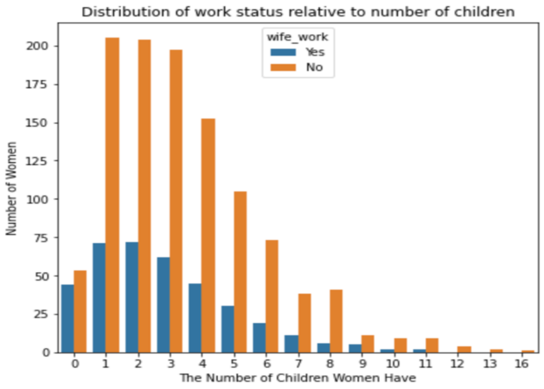
   To answer this question, we looked at the features selected for logistic regression model by our cross validation process. The selected features for our model are the age of woman, the number of children a woman has, the education level of woman, the education level of the husband, and whether the woman's religion is Islamic. However, considering the performances of our logistic regression model, their effectiveness for classification is also limited. We haven't really discovered factors that best help to classify whether a woman is working.

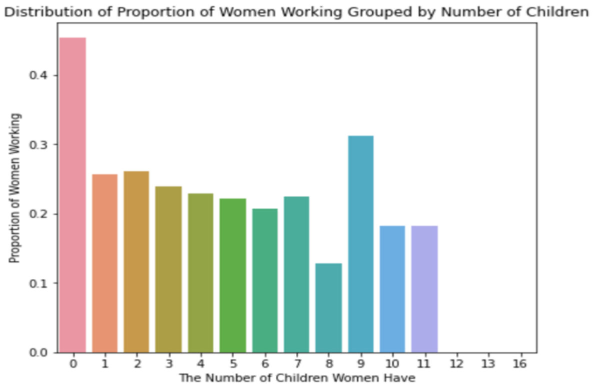
2. Can we accurately classify the employment status of a woman based on her other information?

   Since all our models perform almost no better than the base model, our efforts to classify the employment status of a woman based on given information is not successful. While it certainly is possible that there are features that can help us to do classification better, we find it difficult to perform better than a base model with the given information.

3. If so, what does this tell us about what encourages or discourages a woman from joining the workforce?

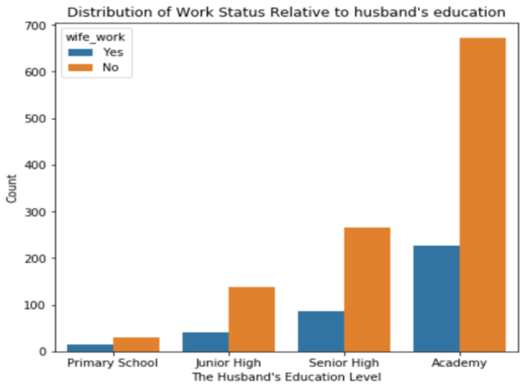
   Again, we examined the coefficients our logistic regression assigned to selected features to see which features are more important. However, since our model performs poorly, one guess we have is that none of the features we extracted really had a significant impact on the employment status of women. We selected two seemingly important features: husband's education level and number of children a woman has and drew some graphs that show the relationship between them and a woman's work status.

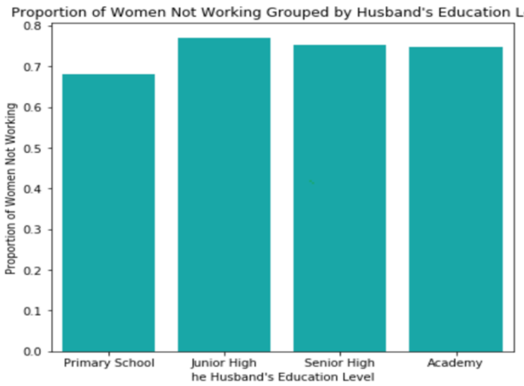




The right plot above shows us the proportion of women who works grouped by number of children they have. From the plot, we can see that for women with 1 -7 children, the proportion of them working are almost the same. However, if we look at the left plot above, we can see that most of women in the dataset have 1-7 children. This tells us that number of children may only help to clearly distinguish a small subset of the data.

The right plot below shows us the proportion of women who doesn't work grouped by their husbands' education levels. From the plot, we can see that for women whose husbands have different education levels, the proportion of them not working are almost the same. The left plot below shows us that this is true even though the proportions of husbands with different education levels differ a lot.





As a conclusion, we understand that at least in our sample, no single feature is very useful for predicting whether a woman works. This might tell us that on a societal level in Indonesia, to look for what greatly encourages or discourages a woman form joining the workforce, we need to look at more than a woman's age, number of children, education level, husband's education level, and her religion.

Discussion

1. What were two or three of the most interesting features you came across for your particular question?

For the issue of fertility, the two most interesting features, in my opinion, are a woman's education level and a woman's religion. Through analysis of our model, we found it very interesting that having a high level education discourages fertility, and Islamic women are less likely to have more children. For employment, the two most interesting features for us are husband's education level, and number of children a woman has. It's interesting that women with no children are more likely to work, and women with 1-7 children are almost equally likely to work. It seems that having children, no matter how many, is a discouragement to a woman going to work. For husband's education level, it's interesting that it hardly has any impact on whether a woman works.

2. Describe one feature you thought would be useful, but turned out to be ineffective.

For both fertility and employment, a feature we thought would be important is media exposure of the woman. However, this feature was not selected by models in both questions. Intuitively, a woman with good media exposure may know more job opportunities, and becomes more likely to join the workforce. She may also learn more about the costs of having more children through media, and thus be discouraged from having more children. However, this feature turned out to be ineffective and was not included.

3. What challenges did you find with your data? Where did you get stuck?

One challenge we found with our data was the failure to see pattern and clustering in the dataset for the issues we want to explore. There were many features in the dataset, and we drew a lot of graphs to investigate relationships between different features, but there was too much information for us to process. We tried to use principal component analysis to reduce dimensionality, but the first two principal components captured too little of the variance, and we couldn't see any clustering or pattern. Another challenge was we got stuck at producing a good model that could classify whether a woman works with our data. We tried several different models and tuned their parameters but to no avail.

4. What are some limitations of the analysis that you did? What assumptions did you make that could prove to be incorrect?

We did our analysis with linear regression. However, linear regression is limited to a linear context. It is highly likely that the relationship between our features and the variable we want to predict is not linear. Also, we used logistic regression for analysis, which requires us to identify the important independent features to work well. But it is likely that we never had good features in our data in the first place.

For both of our issues, the population we wanted to investigate is women in developing countries. We assumed that the subset of data given to us is representative of the population. And we analyzed the population based on our model results. However, this assumption is likely to be incorrect. First of all, women in Indonesia in 1987 may not represent women in society of developing countries. The structure of the society may have changed through time. Different developing countries may have different social structures. Moreover, there might be selection bias with the 1987 survey. I looked at the original questionnaire. It's very long and complicated. Women who can successfully answer the questionnaire may be those who receive good education in the population. And they may not be representative of the population we want to study.

5. What ethical dilemma did you face with the data?

In the data, there is detailed information about the personal background of women in Indonesia. Looking at the original questionnaire, we found that there are a lot of questions that may be triggering to them. We didn't know how the interviews were conducted and whether the women were given a safe and private environment to answer.

6. What additional data, if available, would strength your analysis, or allow you to test some other hypothesis?

In many of the features, the distribution of women is really unbalanced. For example, far more women in our sample have good media exposure than bad. It is questionable whether this sample is chosen randomly and is representative of the population. Also, we have only a limited number of features and couldn't build a good model to classify employment status with these features. We can include more features that may relate to employment status, for example, the age of a woman's youngest child, a woman's health conditions, and the employment status of other women in her family. Furthermore, the sample size we are given is too small, increasing vulnerability to variability and bias. Ideally, larger sizes of data of women, with more features and selected more randomly, can help us to strengthen our analysis and improve the performances of our models.

7. What ethnical concerns might you encounter in studying this problem? How might you address those concerns?

In our study, we aim to predict personal information about women, including their employment status and number of children. If models like ours have high accuracies, then they have risks of being misused by others. For example, employers could use models to predict the number of children a woman is going to have, and decide whether to hire her or not, even though it is discriminatory to do so. To address those concerns, I believe the best way is to establish consent. Any use of personal data must be authorized by owner of that data. Use of predictions from models must be authorized by the owner of the models. This way, we can prevent misuse of information.

Limitation of our methods

Fertility: We used linear regression to predict number of children of a woman. However, there are several limitations with our approach. Linear regression can only model a linear relationship. However, the relationship between number of children and a combination of other information for a woman may not necessarily be linear. Also, linear regression is very sensitive to outliners. For number of children of women in our sample, there are several outliners. For example, one woman has as many as 16 children. These outliners may make our linear regression less accurate than it could be.

Employment: We used logistic regression to predict whether a woman works. A major limitation is that logistic regression requires us to already identify the important independent variables. However, we may not even have the important features that greatly influence employment status since we have a limited number of features. So our logistic regression model performed poorly. We also used decision tree and random forest. But both of these models are limited by their tendency to overfit. We tried to tune their parameters to reduce overfitting, but it then damaged their train accuracy too much, and their test accuracy only increased a little. Essentially, both decision tree and random forest could perform well neither on test set nor train set with tuning, and overfit on train set without it.

Surprising Discoveries and Future work

Fertility: When studying fertility, a surprising discovery we made is that in our model, if a woman doesn't use any contraceptive methods, she tends to have less children. This seems counter-intuitive. After investigating the distribution of number of children relative to contraceptive methods used, we found that a lot of women with no contraceptive method have no children. Maybe these women never intend to have children, so there is no need for them to use any contraceptive method. However, we still don't understand why women with short-term and long-term contraceptive methods don't tend to have less children than women who have children and don't use any contraceptive method. This is an interesting topic that could use more research.

Employment: We were really surprised to discover that with the features given to us, we couldn't produce a good model that classify employment status better than base model. For us, future work on this topic is about what features are actually good for this classification. And with good features, how accurate our model can become.