# Machine Learning Engineer Nanodegree

## Capstone Project

Elissa Backas
August 26th, 2017

# I. Definition

## Project Overview

"Breast is Best." (1) As a new mother, this has been drilled into my head by healthcare providers, books, family, friends, random people on the internet. Breastmilk is the most nutritional choice for the first months of an infant's life. However, despite this knowledge, many woman don't breastfeed for the recommended 12 months or longer. The C DC reports only 30.7% of women continue breastfeeding for at least a year. The CDC reports one of the biggest factors in the success of breastfeeding is breastfeeding friendly hospitals and programs that support breastfeeding. (2) A woman's attitude towards breastfeeding and the support system around her, greatly influences if she will initiate breastfeeding and how long she will breastfeed for. (3)

I personally breastfed my son well past the 12-month mark, even after struggling with supply issues and supplementation, due to the support I received from an online community. Had I known about such community earlier, my problems could have been identified sooner. There are so many people out there who want to offer support in the form of these online communities, La Leche League, and other programs. I think it's important that we deliver this support to the women who want it and need it as early as possible so they can breastfeed for as long as they like.

1. https://www.aap.org/en-us/about-the-aap/aap-press-room/pages/aap-reaffirmsbreastfeeding-guidelines.aspx
2. https://www.cdc.gov/breastfeeding/data/reportcard.htm
3. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC 1595282/

## Problem Statement

What demographic factors predict how long a woman will breastfeed for? Given a woman's age, race, poverty level, education level, etc, how long is she likely to breastfeed her child? If we know a woman is at risk for not breastfeeding for as long as she would like, she can be given additional support from existing breastfeeding programs.

As I am trying to predict a specific, continuous value, a regression model is appropriate. Decision tree regressors seem a good fit here because I'm just as interested in interpreting the model to see how important the features are as I am in the model itself.  I would like to find the most improved model over the benchmark that uses a workable size of features one can input to make a prediction on the number of weeks a woman will breastfeed for.

## Evaluation Metrics

I expect age, born outside the US, marital status, working status, poverty level, education level and race to emerge as predictive features for this model based on the benchmark models.

I will split the NS FG data into training and testing sets, reserving 10% of my data for testing. I will compare regression models using the R^2 score and pick the one with the best score. R^2 score compares the mean squared error between the simplest model and our model. If the model isn't much better than just going by the average, the R^2 score will be close to 0, if the model is good, it will be close to 1.

R^2 score is appropriate for the problem since I am looking for the most improved model over the benchmark, I am interested in how much better the model performs than just using the mean. Looking to just minimize the error leaves off this comparison.

R^2 = 1 - residual sum of squares / total sum of squares. https://en.wikipedia.org/wiki/Coefficient_of_ determination

$$R^2(y, \hat{y}) = 1 - \frac{\sum_{i=0}^{n_{\text{samples}}-1} (y_i - \hat{y}_i)^2}{\sum_{i=0}^{n_{\text{samples}}-1} (y_i - \bar{y})^2} \text{ where } \bar{y} = \frac{1}{n_{\text{samples}}} \sum_{i=0}^{n_{\text{samples}}-1} y_i.$$

http://scikit-learn.org/stable/modules/model_evaluation.html

# II. Analysis

## Data Exploration

I'm using the National Center for Health Statistics (NCHS ). (2016). 2013-2015 National Survey of Family Growth Public Use Data and Documentation. Hyattsville, MD: CDC National C enter for

Health Statistics. Retrieved fromhttp://www.cdc.gov/nchs/nsfg/nsfg_ 2013_ 2015_ puf.htm
This survey contains a plethora of information relevant to family planning and pregnancy. I only plan on using the demographic data from the female pregnancy survey as well as the breastfeeding information including breastfeeding duration from that survey.

I am only looking at demographic data as just demographic data alone may help as doctors have access to this information and may be able to intervene early without having to administer a special survey. Also neighborhoods that contain more of a specific demographic group can be targeted for special programs.

Data in the survey was collected by female interviewers, in person, taking down responses on laptops, averaging 74 minutes. Interviewees were compensated. Respondents were given the opportunity to revise answers if they seemed inconsistent, but there may still be errors in the data due to human error. Values that were imputed manually or by regression for consistency are marked as so.

First I had to make sure the data was in a usable format, I found code to help me import it into a python pandas data frame. I then exported it to csv format so I could easily view the data in excel.

I studied the questions and possible answers to see which I thought were usable and which had no relevance to the problem. There are many columns with little data that will need to be dropped. There are also redundant columns, dates measured in both months and weeks for instance, and many columns towards the end of the data signifying if the data was edited that are also not needed. I list all of the columns I'm keeping below and notes for some of them.

I also need to drop the women who did not have a pregnancy end in a live birth or who are still breastfeeding. For multiples, I'm going to assume breastfeeding duration was equal and will take other features as needed for the first child only.

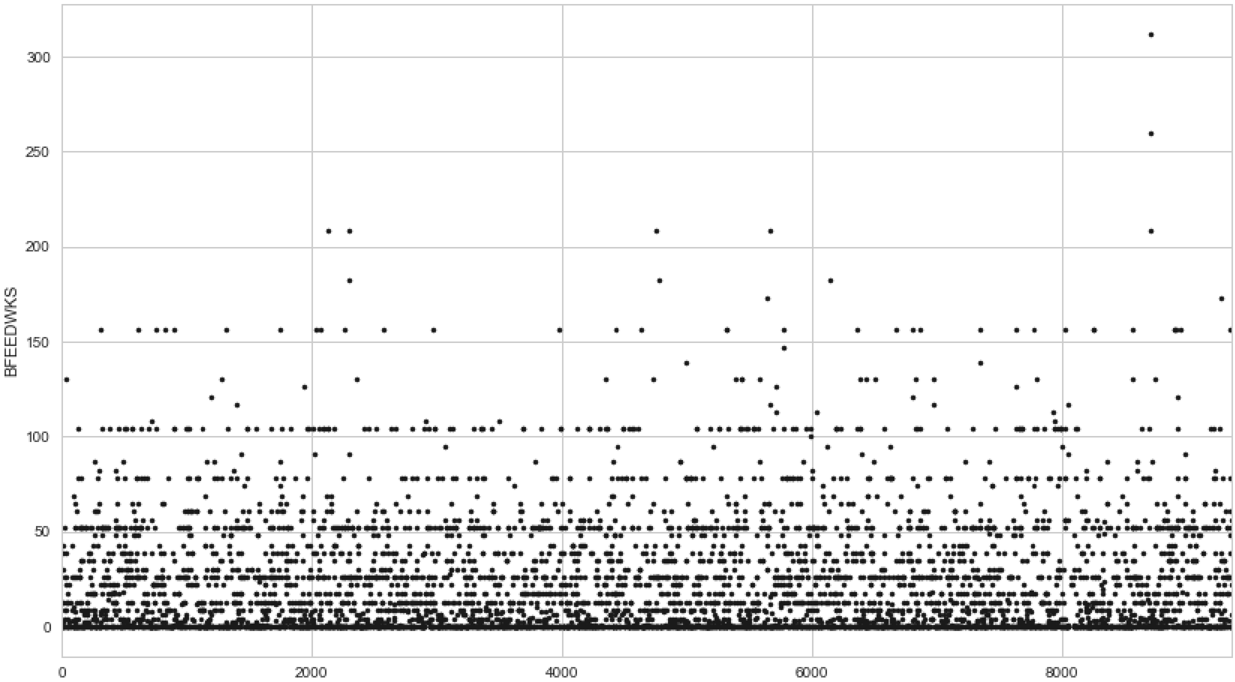
After dropping unnecessary data, I am going to scale my continuous features such as age and one hot encode my discrete features.

These are the fields in the data I'm using or considered using:

- CASEID "Case identification number" #id number to correlate with the other survey, this is an index
- PREGORDR "Pregnancy order (number)" #continuous
- PREGEND1 "BC -1 How Pregnancy Ended - 1st mention" #discrete, vaginal or c section
- WKSGEST "Gestational length of completed pregnancy (in weeks)" #continuous
- BPA_BDSCHECK1 "Whether 1st liveborn baby from this pregnancy was BPA or BDS " #drop babies who died or were given away for adoption
- BABYSEX1 "BD-2 Sex of 1st Liveborn Baby from This Pregnancy" #discrete
- CMBABDOB "CM for baby's or babies' date of birth (delivery date)" #continuous, need to fill in values "not ascertained", "refused", "don't know" answers with mean

- HPAGELB "BD-6 Father's age at time of children's birth" #continuous, need to fill in values "not ascertained", "refused", "don't know" answers with mean
- PRIORSMK "BE-3 Amount R smoked in 6 mos before R knew she was pregnant" #I'd like to use this, but not enough data
- NPOSTSMK "BE-5 Amount R smoked during pregnancy after R knew she was preg" #I'd like to use this, but not enough data
- GETPRENA "BE-6 Any prenatal care for this pregnancy" #I'd like to use this, but not enough data
- CMKIDIED1,2,3 "C M for child's date of death - 1st from this pregnancy" #after removing women with no data for breastfeeding weeks, all of the children left were still alive
- NBRNLV_ S "# of babies born alive from this preg (based on CCSD)" #drop women with NaN for this
- COHPBEG "EG -18a Was R living w/father of preg at beginning of preg" #discrete
- COHPEND "EG -18b Was R living w/father of preg when preg ended/baby was born" #discrete
- BIRTHORD "Birth order" #continuous
- AGEPREG "Age at pregnancy outcome" #continuous
- DATECON "CM date of conception" #continuous
- AGECON "Age at time of conception" #continuous
- FMAROUT5 "Formal marital status at pregnancy outcome" # same as informal, except with one less category, will drop
- PMARPREG "Whether pregnancy ended before R's 1st marriage (premaritally)" #discrete
- RMAROUT6 "Informal marital status at pregnancy outcome - 6 categories" #discrete
- FMARCON5 "Formal marital status at conception - 5 categories" # same as informal, except with one less category, will drop
- RMARCON6 "Informal marital status at conception - 6 categories" #discrete
- PAYDELIV "Payment for delivery" #discrete
- LBW1 "Low birthweight - 1st baby from this preg" #discrete
- **BFEEDWKS "Duration of breastfeeding in weeks" #trying to predict this, continuous**
- EDUCAT "Education (completed years of schooling)" #continuous
- HIEDUC "Highest completed year of school or degree" #discrete, captured in EDUCAT, will drop
- HISPRACE2 "Race & Hispanic origin of respondent - 1997 OMB standards (respondent recode)" #discrete, this captures all of the race features, will drop the others
- PREGNUM "CAPI-based total number of pregnancies" #continuous
- PARITY "Total number of live births" #continuous
- CURR_ INS "Current health insurance coverage" #discrete
- PUBASSIS "Whether R received public assistance in prior calendar year" #discrete
- POVERTY "Poverty level income" #continuous
- LABORFOR "Labor force status" #discrete
- RELIGION "Current religious affiliation" #discrete
- METRO "Place of residence (Metropolitan / Nonmetropolitan)" #discrete
- BRNOUT "IB-8 R born outside of US " #discrete
- YRSTRUS "Year R came to the United States" #This isn't relevant to enough of the women to use

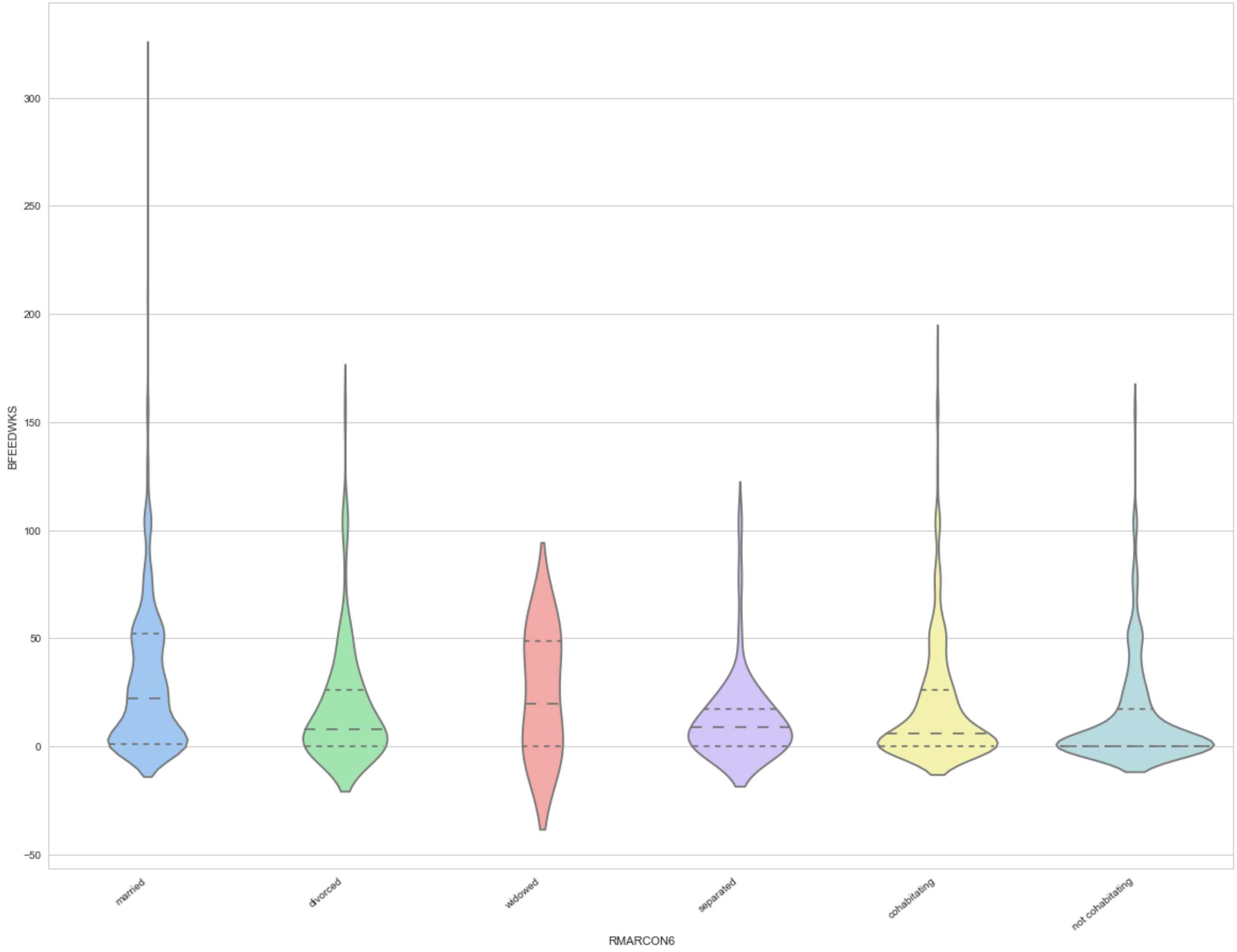
The graph below is a scatterplot showing respondents on the x axis and BFEEDWKS on the y axis. Most women breastfeed for less than a year and very few breastfeed past 3 years. From the data in the chart below, the mean is 22 weeks and the median is 9 weeks.



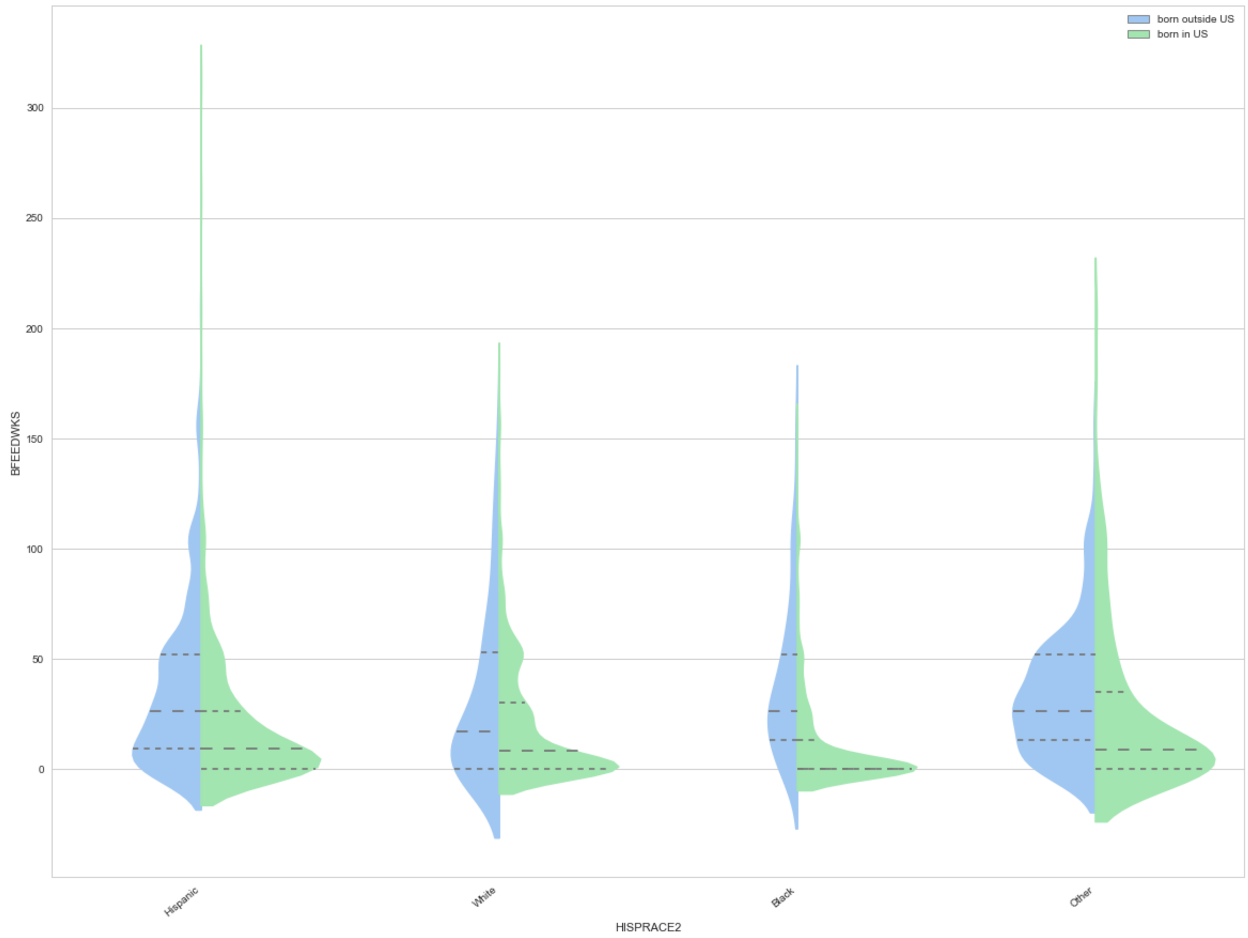
Statistics for some of the continuous features:

| | | | | |
|---|---|---|---|---|
| Name: **WKSGEST**<br>5376.000000<br>mean 38.538690<br>std 2.423987 | Name: **AGEPREG**,<br>5376.000000<br>mean 2563.488467<br>std 550.506578 | **Name: BFEEDWKS,**<br>**5376.000000**<br>**mean 22.109375**<br>**std 30.210073** | Name: **EDUCAT**,<br>5376.000000<br>mean 12.916667<br>std 2.703268 | Name: **POVERTY**<br>5376.000000<br>mean 174.792783<br>std 150.195022 |
| min 23.000000<br>25% 38.000000<br>50% 39.000000<br>75% 40.000000<br>max 48.000000 | min 1325.000000<br>25% 2125.000000<br>50% 2512.000000<br>75% 2958.000000<br>max 4283.000000 | **min 0.000000**<br>**25% 0.000000**<br>**50% 9.000000**<br>**75% 35.000000**<br>**max 312.000000** | min 9.000000<br>25% 11.000000<br>50% 12.000000<br>75% 14.000000<br>max 19.000000 | min 5.000000<br>25% 59.000000<br>50% 117.000000<br>75% 243.000000<br>max 500.000000 |

Exploratory Visualizations



This is a violin plot with the respondent's informal marital status on the x-axis and the number of weeks the respondent breastfed for along the y-axis. From the plot it appears that married and widowed women breastfeed longer than the other groups. They have the highest medians and more women at about the 50 weeks line. Most of the women in the never married, not cohabitating group didn't breastfeed at all, the median is at 0 and 75% below the median for the married group. Though several made it past 50 weeks. I would have thought the cohabitating group would have breastfeed for almost as long as the married group, but the median and 75% lines are well below that group, though it has the second longest point, so a few cohabitating women breastfed for an extended amount of time. This data is also contrary to what the Utah study found, that divorced and separated women were more likely to breastfeed than married women, but that study only looked at women in Utah.



This is a similar plot comparing race. At first I noticed Hispanic women have the highest number of weeks in this study, which is contrary to the rates I have read elsewhere, so I split by BRNOUT comparing those born outside the US and those born in the US. Hispanic women only have higher rates if born outside the US which makes more sense. In general, women who were born outside the US breastfed longer. Black women born in the US have the lowest median and 75%, which is in line with my research.

## Algorithms and Techniques

I will use PCA to reduce the number of features I have. Transforms the features by picking out principle components of correlated features that explain the maximum variance of the data.

I will then try a few different regressors: DecisionTreeRegressor, RandomForestRegressor, and MLP regressor, to see which has the best R^2 score. Decision tree algorithms seem a good fit here because I'm just as interested in interpreting the model to see how important the features are as I am in the model itself.  To try something different and see if neural network algorithm can get better accuracy than the decision trees, I will try MLP Regressor.

I'll start with DecisionTreeRegressor because it's the simplest decision tree algorithm. It creates a tree by splitting on some criteria that will offer the most information gain and then another and another until there are branches to determine where a data point fits. It's simple, but prone to overfitting.

RandomForestRegressor is a boosting algorithm which improves on regular decision trees by creating a forest of decision trees based on subgroups of the data and creates a model by averaging the results of these trees together. Since the grouping of features is random, the splits will differ between trees.

MLP regression does not use trees, but a multilayer perceptron neural network. It consists of neurons with weights that fire based on an activation function. The algorithm works by minimizing the error of the output by first picking random weights, seeing what the error is and determining the direction the weights should be adjusted to minimize the error and continuing to make small adjustments until the error is at a minimum.

I'll then use grid search with cross validation to fine tune the algorithm with different parameters. Grid search is a technique that optimizes parameters by going through a grid of combinations and determining which is best by validating the results. It tries each parameter in a given list in a separate experiment and cross validates each on a separate set of test data. It compares scores for each experiment to determine which parameters were most successful at giving an optimized result. I'll use R^2 for the scoring function to compare models.

Once I have an optimized model, I will compare the feature importance to the benchmark.

# Benchmark Model

This Australian study found the following demographic factors strongly correlated with a longer duration for breastfeeding: https://internationalbreastfeedingjournal.biomedcentral.com/articles/10.1186/1746-4358-1-18 being born in an Asian country and older maternal age. Negatively correlated factors included: the mother smoking 20 or more cigarettes per day pre-pregnancy and maternal obesity.

The study: "Demographic Factors that Predict Breastfeeding in the Early Postpartum Period in Utah Women": https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=1&ved=0ahUKEwiG5mmmd3VAhUS 3YMKHW7uBusQFggoMAA&url=http%3A%2F%2Fdigitalcommons.usu.edu%2Fcgi%2Fviewcontent.cgi%3Farticle%3D1029%26context%3Detd&usg=AFQjCNF9xJ 1wXWjYC VnZ4WzTPe4yuhIPOg) found the following factors correlated with breastfeeding duration: age, marital status, WIC participation, maternal education level, and maternal employment. This study found older mothers were more likely to continue breastfeeding longer. Single women were less likely to breastfeed while divorced and separated women were more likely compared to married women. Enrollment in WIC correlated negatively with breastfeeding. More education was positively correlated with breastfeeding.

This paper also mentions that other studies have found higher parity to be positively correlated with breastfeeding success as well as race as factors.

Coincidently, I've been working on this project during Black Breastfeeding week and learned that black women are far less likely to breastfeed than other racial groups. "The most recent CDC data show that 75% of white women have ever breastfed versus 58.9% of black women."
http://blackbreastfeedingweek.org/

For my benchmark model, I will use the mean for all training data points. The benchmark gave an R^2 score of -0.0245 on the test data.

# III. Methodology

## Data Preprocessing

To see which columns had enough data and what rows I needed to drop, I first printed the counts for all of the columns I thought were relevant. I found out I couldn't use the smoking data and the years in US as most rows did not have values for these columns. I dropped women still breastfeeding and women with no breastfeeding data. If a woman never breastfed, I set her number of weeks equal to 0. I dropped women who didn't have any live babies born from this pregnancy and I dropped babies who died soon after birth or were given away for adoption. I noticed that the baby's date of birth and father's age columns had a few "not ascertained", "refused", and "don't know answers" which I filled in with the mean because I didn't want to drop those rows all together.

I applied a logarithmic function to the continuous features so the extremes in the data didn't skew the results as much. I tested with and without the logarithmic function applied and the model did better with it. I then used MinMaxScaler to scale my continuous features to be between 0 and 1 and get_dummies to one hot encode my discrete features.

## Implementation

I tried using PCA to reduce the number of features, but the results were disappointing. The first 15 dimensions only cover about 65% of the variance in the data. I tried using the reduced features with the Regression models, but I got worse results than with the original data.

Reduced R^2 results:
DecisionTreeRegressor r^2 train score 1.0000
DecisionTreeRegressor r^2 test score -0.7685
RandomForestRegressor r^2 train score 0.8539

RandomForestRegressor r^2 test score 0.2169
MLPRegressor r^2 train score 0.1101
MLPRegressor r^2 test score 0.0186

Since training time isn't bad anyway with so little data, I chose to not use the reduced features. However, I added back in BFEEDWKS and use the PCA components to see if features were positively or negatively correlated with BFEEDWKS.

I dropped BFEEDWKS and used that as my label. Using the chosen features, I split the data into training and test sets, reserving 10% for testing. I then trained the data using DecisionTreeRegressor, RandomForestRegressor and MLPRegressor.

I compared the R^2 score for each:
DecisionTreeRegressor r^2 train score 1.0000
DecisionTreeRegressor r^2 test score -0.2766
RandomFoestRegressor r^2 train score 0.8688
RandomFoestRegressor r^2 test score 0.3423
MLPRegressor r^2 train score 0.5220
MLPRegressor r^2 test score 0.2176

## Refinement

I took the RandomForestRegressor and played with the data and the parameters. As I mentioned above, I applied a logarithmic function to my continuous features which bumped the r^2 score up by .1. With a high train score and low test score, the model seems to be overfitting. I used GridSearchCV to tune all of the parameters which includes n_estimators, min_samples_split, min_samples_leaf, and max_leaf_nodes. I first looked at two parameters at a time and played with many different values, honing in on the best parameters and then set those and looked at the next two. I found the rest of the parameters were best left as default. This improved the model from 3423 to .3621.

I then used the improved model and looked at the feature importance. I took the top 20 features and studied them by looking at their correlations via PCA and through a heatmap. I then took these 20 features and used only them in my model.

# IV. Results

## Model Evaluation and Validation

The final model is the refined RandomForestRegressor using the 20 most important feature as ranked by the feature_importances attribute. The R^2 score of the reserved test data is close to the model using all features, but 20 is a more reasonable number to work with if you are taking down data on a particular woman.

The model predicted I would breastfeed for 34 weeks which is far less than I actually breastfed for, but that's a fair enough estimate in this case to say that I probably don't need extra support.

I wouldn't use this as the only predictor of breastfeeding success, but with some additional data, I think it could be a helpful tool.

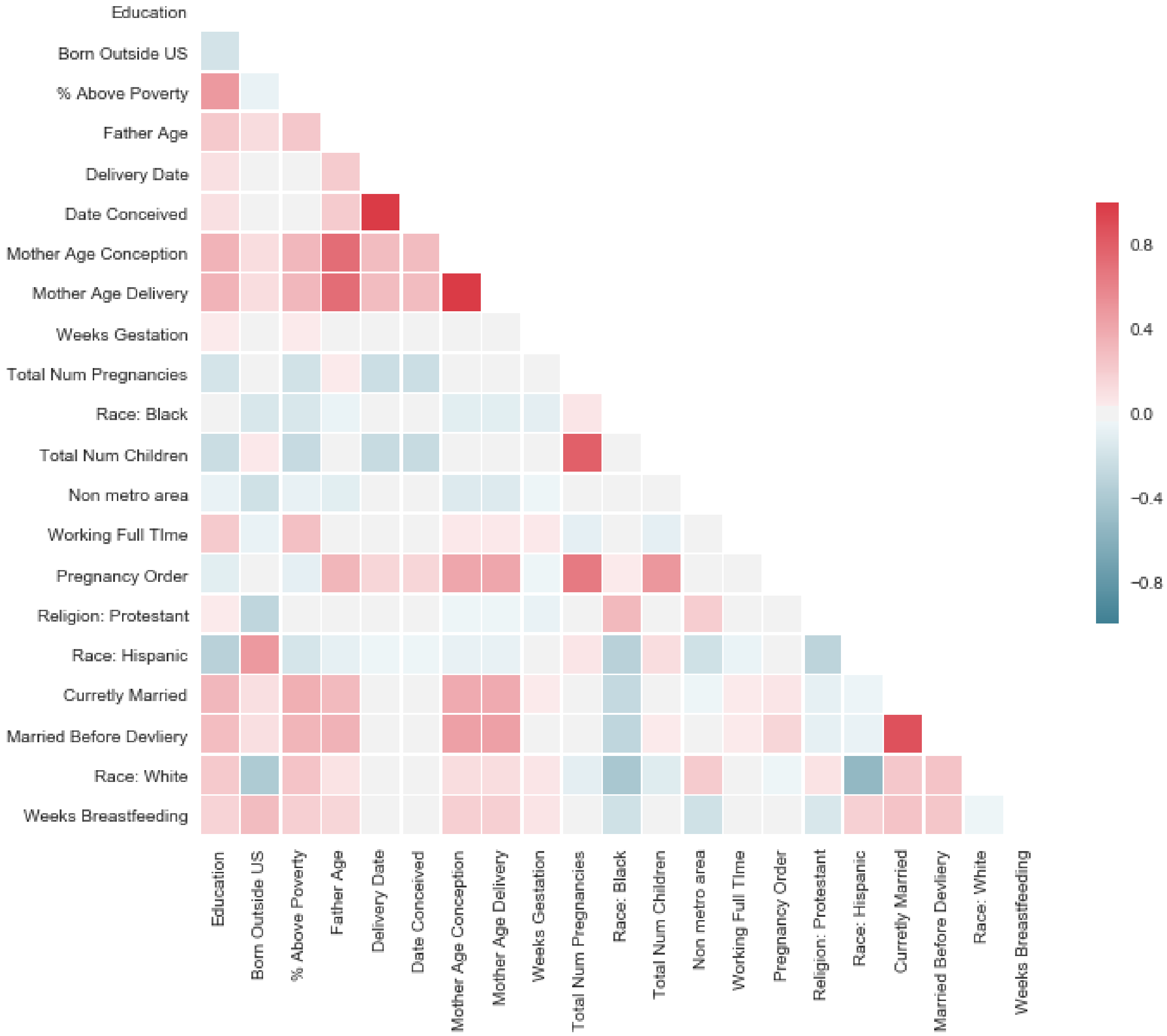
## Justification

The model is still not generalizing very well, an R^2 score of .34 on the reserved test data. It's better than just using the benchmark model, the average of the training data, which gave a score of -0.0245, but the score is still far from a perfect 1. I believe it is significant enough to give health care providers an idea of how long a woman may breastfeed for.

It did pick out similar features as important as what the researched data predicted. Including age, born outside the US, marital status, working status, poverty level, education level and race. It also included parity as a predictive feature which some studies suggested was significant, even though the heatmap below shows it doesn't correlate with breastfeeding.

# V. Conclusion

## Free-Form Visualization



This is a heat map for correlations between our 20 most important features, listed in order of importance and BFEEDWKS. As expected, education level, percent above poverty level, being born outside the US, mother's age and being married, were positively correlated with the number of weeks breastfeeding while being black, negatively correlated with weeks breastfeeding.

The age of the father and being Hispanic also positively correlated with breastfeeding while being white, Protestant and from a non metro area negatively correlated with breastfeeding. There were several features the model considered more important but didn't have very high correlation to breastfeeding such as delivery date, date conceived, and birth order, number of pregnancies/children, pregnancy order and full time working status.

Looking at dimension 1 of the PCA reduced data, we see similar correlations:

```
EDUCAT          0.099311
BRNOUT_1        0.034671
POVERTY         0.103421
HPAGELB         0.037017
```

```
CMBABDOB       -0.001000
DATECON        -0.001159
AGECON          0.074027
AGEPREG         0.074417
WKSGEST         0.005539
PREGNUM        -0.009535
HISPRACE2_3    -0.113168
PARITY         -0.011101
METRO_3        -0.027812
LABORFOR_1      0.059916
PREGORDR        0.009227
RELIGION_3     -0.043474
HISPRACE2_1    -0.041036
RMAROUT6_1.0    0.354348
PMARPREG_2.0    0.351803
HISPRACE2_2     0.129877
BFEEDWKS        0.070969
```

Except being white positively correlated with breastfeeding and being Hispanic correlated negatively which is more accurate when combined with other factors, the other factors in the case of dimension 1 mostly being marital status, race and poverty level.

There are also obvious strong correlations between some pairs of features, that looking back, seems unnecessary to include both, such as age at conception/age at delivery, child's conception date/delivery date, if the mother was married before delivery/if she is currently married, and number of pregnancies/number of births.

# Reflection

Once I had decided to do a project on breastfeeding, I spent a lot of time looking for and researching what data I could use. Even the dataset I ended up with wasn't ideal and the number of women with breastfeeding data is small, skewed and missing some key features I would have liked to look at. I also struggled with importing the data into python and I would not have been able to if I didn't find existing code that already did this; I didn't understand the original data format well enough to do it myself. I also had to use several different files to interpret the coding for the questions and answers. This data gathering, importing and analysis was the hardest part of the project for me and took the most amount of time.

After studying the data, I manually picked out which features I thought were relevant to the problem. I studied visualizations to understand the data better. I scaled and encoded my features so they were all between 0 and 1. I tried several different regression algorithms and RandomForestRegressor produced the best initial results. I fine tuned the model with GridSearch. I tried the algorithm with reduced data from PCA, but this did not improve the R^2 score.

I instead looked at the feature importances produced by the RandomForest model and picked the top 20 features. I trained the model again with just those features to produce my final model. I then studied those 20 features using the PCA results and a heatmap to view the correlations with breastfeeding.

I had expected to be able to get a better R^2 score, but I think it's good enough to be used to give a general estimate for how long a woman may breastfeed for. I'm, glad the results for feature importance I ended up with were consistent with other research. I was surprised that education was the most important feature the model picked out and that working status was not more important.

## Improvement

One problem I had was the data itself. It is unbalanced and highly skewed towards 0. There are very few women in this study who breastfed for an extended period of time. I think having more data on breastfeeding women would greatly help. I think I'm also missing several features, such as if the mother was a smoker, that would have help predict breastfeeding success. Some of the respondents also were interviewed on pregnancies that happened years ago, which is why I left the date of birth as a feature, but having data for recent births only would probably help as more and more women are breastfeeding every year.

I could turn this into a classification problem instead of a regression problem, which might improve the results. I don't think there would be enough data to split on the 12 months or more mark, but maybe 6 months or more. I could also use oversampling to help with the unbalanced data. I was curious though to see how the regression model worked out as women have varying breastfeeding goals.