**Machine Learning Engineer Nanodegree**

**Capstone Project**

Elissa Backas   
August 26th, 2017

1. 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](https://www.aap.org/en-us/about-the-aap/aap-press-room/pages/aap-reaffirms-breastfeeding-guidelines.aspx)
2. <https://www.cdc.gov/breastfeeding/data/reportcard.htm>
3. [https://www.ncbi.nlm.nih.gov/pmc/articles/PMC 1595282/](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1595282/)

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.

# Evaluation Metrics

I expect age, born outside the US, marital status, working status, poverty level, education level 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 R2 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 = 1 - residual sum of squares / total sum of squares. [https://en.wikipedia.org/wiki/Coefficient\_ of\_ determination](https://en.wikipedia.org/wiki/Coefficient_of_determination) |  |

II. Analysis

Data Exploration

I'm using the National C enter for Health Statistics (NC HS ). (2016). 2013-2015 National Survey of

Family Growth Public Use Data and Documentation. Hyattsville, MD: CDC National C enter for

Health Statistics. Retrieved from[http://www.cdc.gov/nchs/nsfg/nsfg\_ 2013\_ 2015\_ puf.htm](http://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 may also use some data from the main female respondent 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.

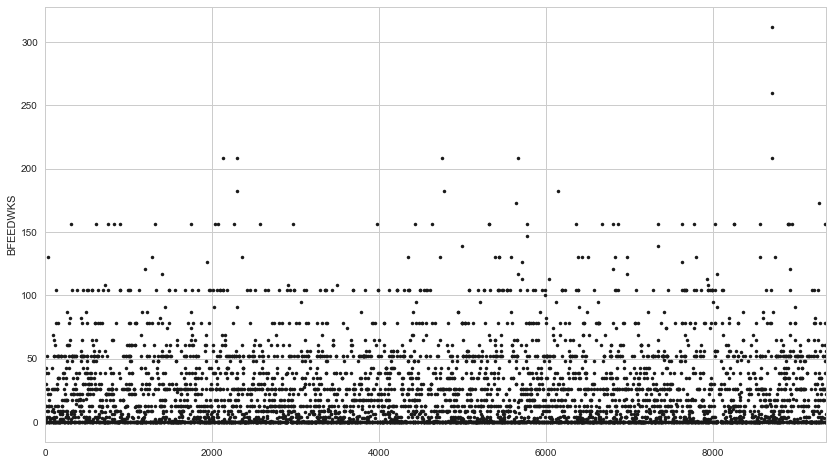
I also need to drop the women who did not have a pregnancy end in a live birth or who are still breastfeeding. For multiples, I'm going to assume breastfeeding duration was equal and will take other features as needed for the first child only.

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

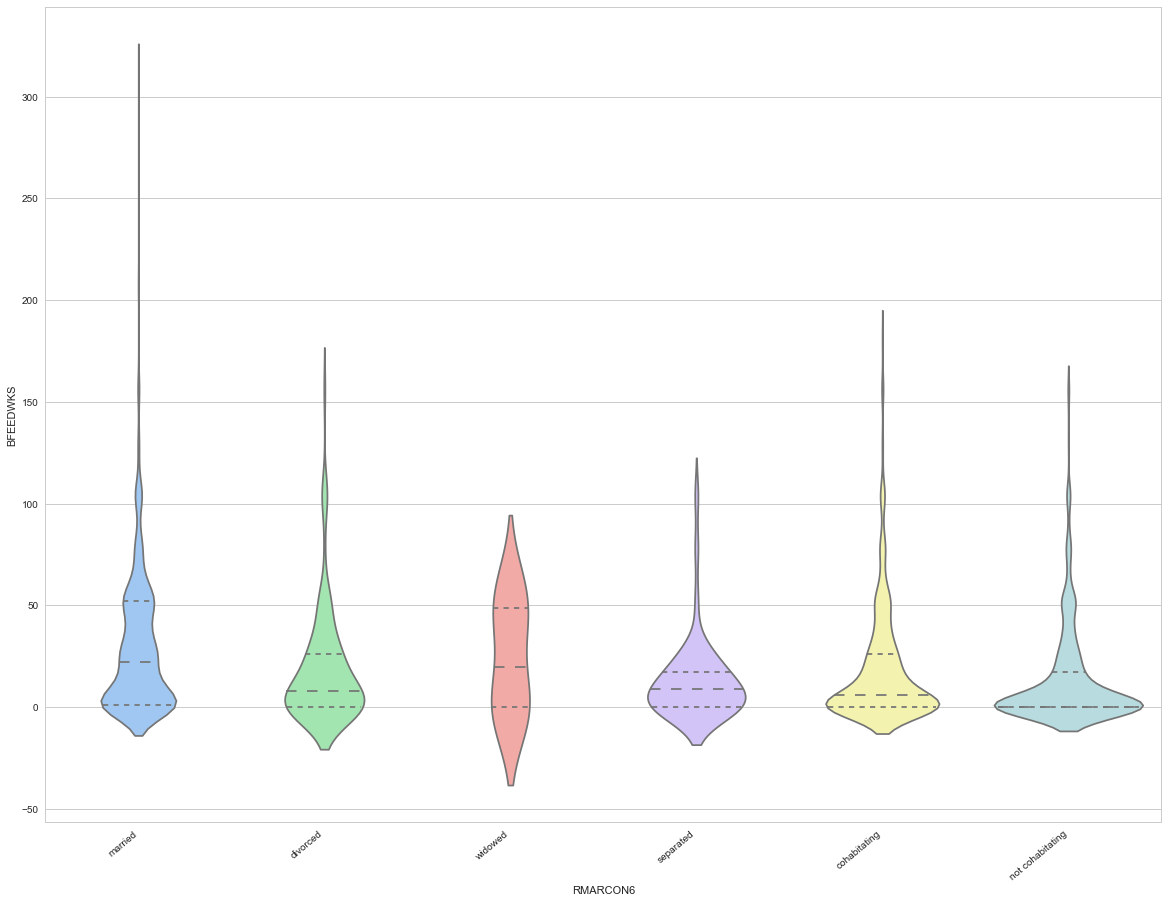
The graphs below is a scatterplot showing respondents on the x axis and BFEEDWKS on the y axis. Most women breastfeed for less than a year and very few breastfeed past 3 years. From the data in the chart below, the mean is 22 weeks and the median is 9 weeks.



Statistics for some of the continuous features:

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

Exploratory Visualization



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 abut the 50 weeks line. Most of the women in the never married, not cohabitating group didn't breastfeed at all, the mean 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.

Algorithms and Techniques

I will use PCA to reduce the number of features I have.

I will then try a few different regressors, DecisionTreeRegressor, RandomForestRegressor,

AdaBoostRegressor, MLP regressor, and 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.

I'll then use grid search with cross validation to fine tune the algorithm with different parameters.

Once I have my 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 VnZ4WzT](https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=1&ved=0ahUKEwiG5-mmmd3VAhUS3YMKHW7uBusQFggoMAA&url=http%3A%2F%2Fdigitalcommons.usu.edu%2Fcgi%2Fviewcontent.cgi%3Farticle%3D1029%26context%3Detd&usg=AFQjCNF9xJ1wXWjYCVnZ4WzTPe4yuhIPOg)

[Pe4yuhIPOg)](https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=1&ved=0ahUKEwiG5-mmmd3VAhUS3YMKHW7uBusQFggoMAA&url=http%3A%2F%2Fdigitalcommons.usu.edu%2Fcgi%2Fviewcontent.cgi%3Farticle%3D1029%26context%3Detd&usg=AFQjCNF9xJ1wXWjYCVnZ4WzTPe4yuhIPOg) 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. S ingle 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.

For my benchmark model, I will use the mean for all data points.

## 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 checked for babies who died soon after birth or 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 it was a big disappointment. 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:

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 is fast 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 -0.2486  
RandomForestRegressor r^2 train score 0.8690  
RandomForestRegressor r^2 test score 0.3261  
MLPRegressor r^2 train score 0.5709  
MLPRegressor r^2 test score 0.2287

### 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. I used GridSearchCV to tune all of the parameters. I first looked at two parameters at a time and played with several different values, honing in on the best parameters and then set those and look at the next two. This improved the model from .3261 to

I then used the improved model and looked at the feature importance. I took the top 15 features and studied them by looking at their correlations via PCA and through a heatmap. I then took these 15 features and used only them in my model.

## IV. Results

\_(approx. 2-3 pages)\_

### Model Evaluation and Validation

In this section, the final model and any supporting qualities should be evaluated in detail. It should be clear how the final model was derived and why this model was chosen. In addition, some type of analysis should be used to validate the robustness of this model and its solution, such as manipulating the input data or environment to see how the model’s solution is affected (this is called sensitivity analysis). Questions to ask yourself when writing this section:

- \_Is the final model reasonable and aligning with solution expectations? Are the final parameters of the model appropriate?\_

- \_Has the final model been tested with various inputs to evaluate whether the model generalizes well to unseen data?\_

- \_Is the model robust enough for the problem? Do small perturbations (changes) in training data or the input space greatly affect the results?\_

- \_Can results found from the model be trusted?\_

### Justification

In this section, your model’s final solution and its results should be compared to the benchmark you established earlier in the project using some type of statistical analysis. You should also justify whether these results and the solution are significant enough to have solved the problem posed in the project. Questions to ask yourself when writing this section:

- \_Are the final results found stronger than the benchmark result reported earlier?\_

- \_Have you thoroughly analyzed and discussed the final solution?\_

- \_Is the final solution significant enough to have solved the problem?\_

## V. Conclusion

\_(approx. 1-2 pages)\_

### Free-Form Visualization

In this section, you will need to provide some form of visualization that emphasizes an important quality about the project. It is much more free-form, but should reasonably support a significant result or characteristic about the problem that you want to discuss. Questions to ask yourself when writing this section:

- \_Have you visualized a relevant or important quality about the problem, dataset, input data, or results?\_

- \_Is the visualization thoroughly analyzed and discussed?\_

- \_If a plot is provided, are the axes, title, and datum clearly defined?\_

### Reflection

In this section, you will summarize the entire end-to-end problem solution and discuss one or two particular aspects of the project you found interesting or difficult. You are expected to reflect on the project as a whole to show that you have a firm understanding of the entire process employed in your work. Questions to ask yourself when writing this section:

- \_Have you thoroughly summarized the entire process you used for this project?\_

- \_Were there any interesting aspects of the project?\_

- \_Were there any difficult aspects of the project?\_

- \_Does the final model and solution fit your expectations for the problem, and should it be used in a general setting to solve these types of problems?\_

### Improvement

One problem I had was the data. 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 help predict breastfeeding success.

The reviewer of my proposal suggested I turn this into a classification 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.

\*\*Before submitting, ask yourself. . .\*\*

- Does the project report you’ve written follow a well-organized structure similar to that of the project template?

- Is each section (particularly \*\*Analysis\*\* and \*\*Methodology\*\*) written in a clear, concise and specific fashion? Are there any ambiguous terms or phrases that need clarification?

- Would the intended audience of your project be able to understand your analysis, methods, and results?

- Have you properly proof-read your project report to assure there are minimal grammatical and spelling mistakes?

- Are all the resources used for this project correctly cited and referenced?

- Is the code that implements your solution easily readable and properly commented?

- Does the code execute without error and produce results similar to those reported?