Machine Learning Engineer Nanodegree

Capstone Proposal

Elissa Backas August 15, 2017

Proposal

Domain Background

"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 CDC 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-reaffirms-breastfeeding-guidelines.aspx
2. https://www.cdc.gov/breastfeeding/data/reportcard.htm
3. 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.

Datasets and Inputs

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 Center 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 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.

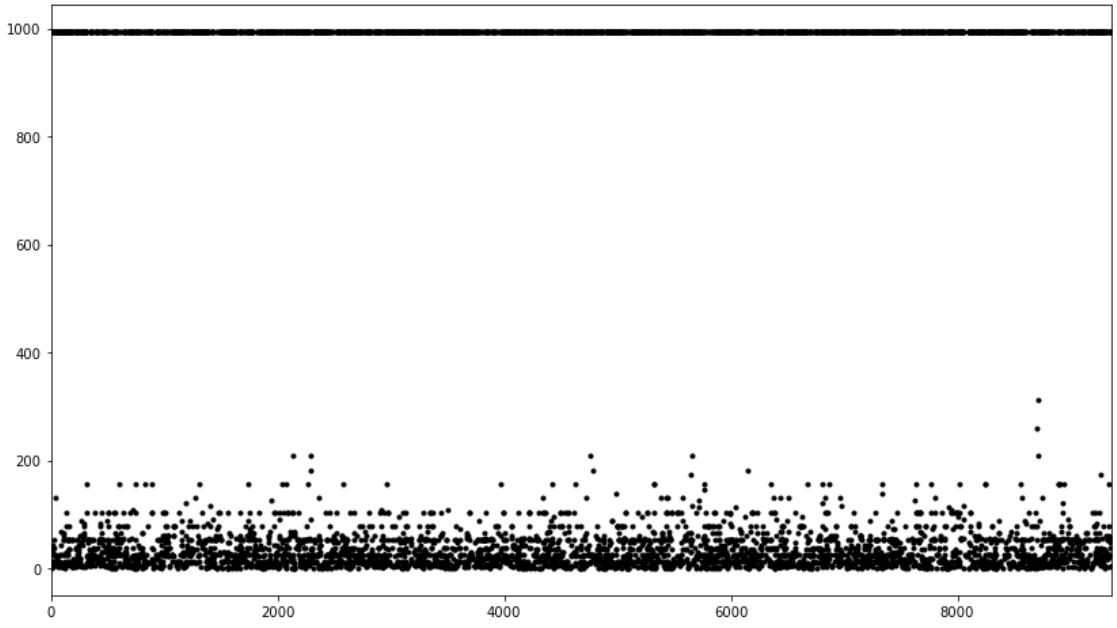
Data in the survey was collected by female interviewers, in person, taking down responses on laptops, averaging 74 minutes. Interviewees were compensated. Respondents were given the opportunity to revise answers if they seemed inconsistent, but there may still be errors in the data due to human error.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

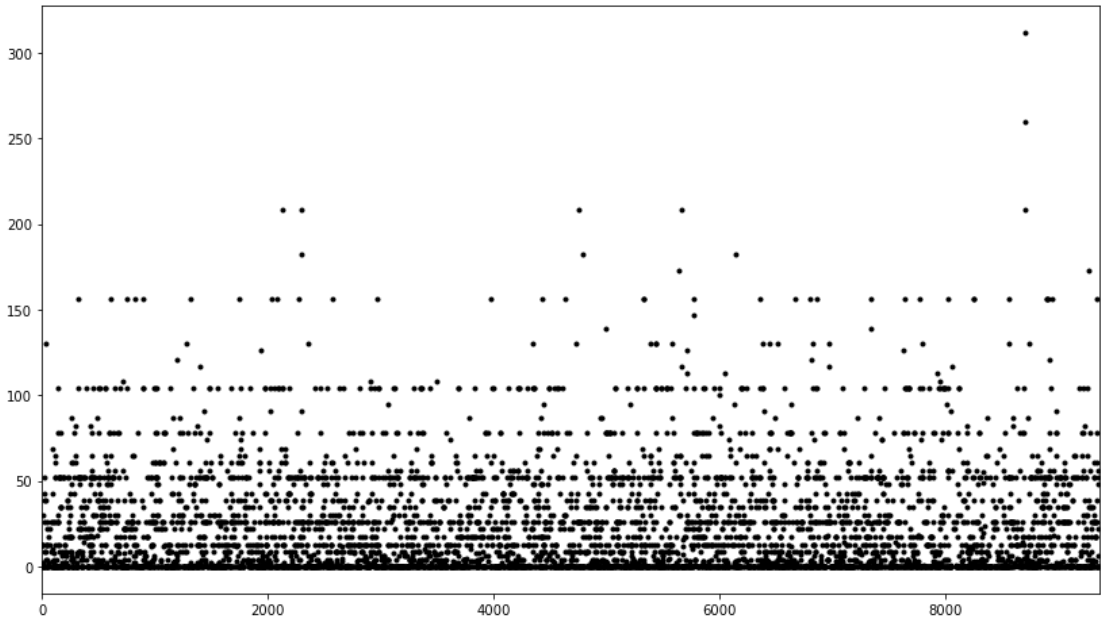
These are the fields in the data I'm keeping:

- 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
- HPAGELB "BD-6 Father's age at time of child(ren) s birth" #continuous
- 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
- CMKIDIED,2,3 "CM for child's date of death - 1st from this pregnancy" #only multiples died, will remove those rows from the data
- OUTCOM_S "Outcome of pregnancy (based on corrected/chron sorted data)" #discrete
- 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" #discrete

- 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" #discrete
- 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
- RACE "Race" #discrete
- HISPANIC "Hispanic origin" #discrete
- HISPRACE "Race & Hispanic origin of respondent - 1977 OMB standards (respondent recode)" #discrete
- HISPRACE2 "Race & Hispanic origin of respondent - 1997 OMB standards (respondent recode)" #discrete
- RCURPREG "Pregnant at time of interview" #discrete
- 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" #I'd like to use this, but not enough data

The graphs below are scatterplots showing respondents on the x axis and BFEEDWKS on the y axis. The plot on the top was before I removed still breastfeeding respondents and set never breastfed response to 0 weeks.
The plot on the bottom was after these changes. From the plot, we can see 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,** dtype: int64 count 5376.000000 mean 38.538690 std 2.423987	Name: **AGEPREG,** dtype: float64 count 5376.000000 mean 2563.488467 std 550.506578	**Name: BFEEDWKS, dtype: int64 count 5376.000000 mean 22.109375 std 30.210073**	Name: **EDUCAT,** dtype: int64 count 5376.000000 mean 12.916667 std 2.703268	Name: **POVERTY,** dtype: int64 count 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

Solution Statement

Use supervised learning to create a regression model, based on demographic information, that can predict how long a woman will breastfeed. Use feature_importance to choose which factors best predict breastfeeding duration. Input would be information such as age, race, education level etc and output would be the number of weeks she is likely to breastfeed for.

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=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. 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.

Evaluation Metrics

I expect age, weight (if I can pull it out of the other survey), smoking habits (if there's enough data), country of origin, 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 NSFG 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

Project Design

First I had to make sure the data was in a usable format, so I already 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 I will drop. 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. I need to figure out if the breastfeeding duration differs at all and if I can drop the statistic for the other children.

After dropping unnecessary data, I am going to scale my continuous features such as age and one hot encode my discrete features. I will use PCA to reduce the number of features I have.

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