Machine Learning Engineer Nanodegree

Capstone Proposal

Sabrina Aliyeva
August 30th, 2017

Predicting Likelihood of Natural Birth Intervention for a C-section

Domain Background

Many believe the C-section vs. vaginal birth debate is a matter of choice. However, there is an evidence that proves natural birth is safest for most moms and babies [2]. There can be situations when C-section is the best option for mother and child, sometimes lifesaving. However, there are also major risks involved for mother and baby. C-section is a major abdominal surgery. There is a higher risk of blot clots following a C-section than a vaginal birth [1]. Hysterectomy is more common after cesarean. The risk of maternal death is higher. Babies born by C-section are much more likely to have respiratory problems than babies born by vaginal birth. They are also at much higher risk for developing asthma later in life [3].

Around the world C-Section based delivery is approximately 19%, whereas in USA it was estimated 32.7% in 2013. It's clear that a cesarean holds many risks. Vaginal birth is much safer for both mother and baby in most situations [4][5].

When I was giving a birth in January 2017 I was terrified that I could end up in C-section, thankfully I didn't. Hence, for my capstone projected I decided to work on predicting a likelihood of intervention of a vaginal delivery for a C-section based on women health history in USA.

Problem Statement

This capstone project will be an attempt to train and tune a classifier which should correctly identify if a woman had vaginal or cesarean delivery. I will decide on two models that will be trained, tested and validated against provided data. The result of how good the classifier is will be measured using Scikit-Learn metric's module metrics.f1_score and metrics.confusion_matrix functions. Any classification model can point at what features from dataset make the most impact on the accuracy of prediction. I will determine the most important features, then, use those on the dataset for the next year and compare the accuracy.

Datasets and Inputs

CDC birth data represents all births registered in the 50 States, the District of Columbia, and New York City. The Centers for Disease Control and Prevention's National Center for Health Statistics (NCHS) receives these data as electronic files, prepared from individual records processed by each registration area, through the Vital Statistics Cooperative Program. Birth data for the U.S. are limited to births occurring within the United States to U.S. residents and nonresidents [6].

For this project, I chose to work with CDC birth data from 2014. The data has 241 features and about 4 000 000 patients. It also has all the information about delivery and history of each woman who tried for vaginal and C-section delivery. The data is a mix of categorical and binary data. So I have to transfer all the categorical features into numerical/binary. The target column is binary column, where "1" is Vaginal and "2" is C-section delivery.

Before I get to the choosing a classifier I have to clean the data. I have to only select the features, which have information prior to the delivery. The features which are going into the model is a medical history of women such as age, weight, prior C-section history, whether a woman had high blood pressure during pregnancy, if the baby is breached, education of mother and father, race of mother and father, and etc. The Table 1 describes all available features from 2014 CDC Birth data.

There are also a lot of missing values; hence, I will remove the columns that contain 40% of "NaN" values. For random missing values, I will replace them with the appropriate values from the CDC documents. For example, "mother's education" is divided into buckets ranging from 1 to 9, where value 9 is for "Unknown", hence if there are "NaN" values in "mother's education" columns, they will be replaced with value 9. Another example is column "Prior birth, now living" where the range is 00-30 and 99 is if information is unknown or not available. Therefore all "NaN" values in this column will be replaced with value 99. After cleaning the data there will 101 features and 3 998 175 data points left the dataset.

Table 1: List of features.

dob_yy	fagerec11	cig_2	f_rf_gdiab	f_ob_fail	f_mm_rupt	oegest_r3	ca_cleft	uca_downs
dob_mm	frace31	cig_3	f_rf_phyper	ld_indl	f_mm_uhyst	dbwt	ca_clpal	
dob_tt	frace6	cig0_r	f_rf_ghyper	ld_augm	f_mm_aicu	bwtr12	ca_downs	
dob_wk	frace15	cig1_r	f_rf_eclamp	ld_ster	no_mmorb	bwtr4	ca_disor	
bfacil	fbrace	cig2_r	f_rf_ppb	ld_antb	attend	ab_aven1	ca_hypo	
f_facility	fhisp_r	cig3_r	rf_inftr	ld_chor	mtran	ab_aven6	f_ca_limb	
bfacil3	f_fhisp	f_cigs_0	rf_fedrg	ld_anes	pay	ab_nicu	f_ca_cleftlp	
mage_impflg	fracehisp	f_cigs_1	rf_artec	f_ld_indl	pay_rec	ab_surf	f_ca_cleft	
mage_repflg	feduc	f_cigs_2	f_rf_inft	f_ld_augm	f_pay	ab_anti	f_ca_downs	

mager	priorlive	f_cigs_3	f_rf_inf_drg	f_ld_ster	f_pay_rec	ab_seiz	f_ca_chrom	
mager14	priordead	cig_rec	f_rf_inf_art	f_ld_antb	apgar5	f_ab_vent	f_ca_hypos	
mager9	priorterm	f_tobaco	rf_cesar	f_ld_chor	apgar5r	f_ab_vent6	no_congen	
mbstate_rec	lbo_rec	m_ht_in	rf_cesarn	f_ld_anes	f_apgar5	f_ab_nicu	itran	
restatus	tbo_rec	f_m_ht	f_rf_cesar	no_lbrdlv	apgar10	f_ab_surfac	ilive	
mrace31	illb_r	bmi	f_rf_ncesar	me_pres	apgar10r	f_ab_antibio	bfed	
mrace6	illb_r11	bmi_r	no_risks	me_rout	dplural	f_ab_seiz	f_bfed	
mrace15	ilop_r	pwgt_r	ip_gon	me_trial	imp_plur	no_abnorm	ubfacil	
mbrace	ilop_r11	f_pwgt	ip_syph	f_me_pres	setorder_r	ca_anen	urf_diab	
mraceimp	ilp_r	dwgt_r	ip_chlam	f_me_rout	sex	ca_mnsb	urf_chyper	
mhisp_r	ilp_r11	f_dwgt	ip_hepatb	f_me_trial	imp_sex	ca_cchd	urf_phyper	
f_mhisp	precare	wtgain	ip_hepatc	rdmeth_rec	dlmp_mm	ca_cdh	urf_eclam	
mracehisp	f_mpcb	wtgain_rec	f_ip_gonor	dmeth_rec	dlmp_yy	ca_omph	ume_forcep	
mar_p	precare5	f_wtgain	f_ip_syph	f_dmeth_rec	compgst_imp	ca_gast	ume_vac	
dmar	previs	rf_pdiab	f_ip_chlam	mm_mtr	obgest_flg	f_ca_anen	uop_induc	
mar_imp	previs_rec	rf_gdiab	f_ip_hepatb	mm_plac	combgest	f_ca_menin	uld_breech	
f_mar_p	f_tpcv	rf_phype	f_ip_hepatc	mm_rupt	gestrec10	f_ca_heart	uca_anen	
meduc	wic	rf_ghype	no_infec	mm_uhyst	gestrec3	f_ca_hernia	uca_spina	
f_meduc	f_wic	rf_ehype	ob_ecvs	mm_aicu	lmpused	f_ca_ompha	uca_ompha	
fagerpt_flg	cig_0	rf_ppterm	ob_ecvf	f_mm_mtr	oegest_comb	f_ca_gastro	uca_cleftlp	
fagecomb	cig_1	f_rf_pdiab	f_ob_succ	f_mm_	oegest_r10	ca_limb	uca_hernia	

Solution Statement

Using the input features, I just mentioned above I am going to use classification algorithms. I will firstly, split input data into training and testing set. Since I have binary classification problem I will first try Logistic Regression on training set, due to so many features there might be some sort of multicollinearity. Hence, I will try Regularization (L2) and Logistic Regression which should remove collinear predictors by lowering their weights closer to zero. I will, then, check the performance of the trained Logistic Regression of testing set by looking at F1 score and confusion matrix.

Since, it is classification problem I suspect there might be imbalanced classes, such that one class will be always misclassified. Hence, I will probably have to undersample my data.

Next, I will try Random Forest Classifier. I will, then, use F1 score to see how accurate my prediction is. I will also consider tuning Random Forest Classifier to achieve a better F1 score. I would also like to look at confusion matrix because I am more interested in predicting two classes correct and not just overall F1 score. After I choose a predictor with a relatively good F1 test score I will look at most important features to understand based on what that particular algorithm made predictions. Next, I will extract probabilities of that predictor. I would also like to use the same important features and model on CDC birth data 2015 and see if I can get approximately the same F1 score.

Benchmark Model

Unfortunately, not much work has been done on prediction of vaginal delivery intervention for C-section. The only paper I found somehow close to this capstone project is about Predicting Caesarean Delivery with Decision Tree Models. There, authors explored Logistic Regression and Decision Tree methods to determine if a woman is going to have C-section based on 78 features. They found that accuracy of both methods is ~ 90%. Given all these results and the availability of the data set, a direct comparison with this paper's results is not possible. However, I would like to continue my study.

For benchmark model I will construct Random Forest from scratch instead of using the one provided by sklearn. I will use Bagging method on out-of-box Decision Tresses where I will split each node on Impurity using Gini Index. I will, then, use my CDC data on the out-of-box Random Forest and later will compare test F1 score with Random Forest provided by sklearn.

Evaluation Metrics

The evaluation metrics for classification model will be the confusion matrix and F_1 score. Since a supervised classifier has known labelled data, we can determine the number of true positives, false positives, and false negatives these are ultimately derived from the confusion matrix.

Where each of the terms in confusion matrix can be defined as following:
- True positives: Entries that are correctly labelled
- False positives: Entries that a wrongly identified with a given label
- False negatives: Entries for a given label that are wrongly identified with other labels

In the general case, a confusion matrix is simply a matrix illustrating the mapping from the true labels to the predicted labels. Elements along the diagonal represent a correct classification, whereas the off-diagonal represent a misclassification. A confusion matrix can be a useful check to see what type of child delivery in particular are most likely confused for one another.

F_1 score is a measure of an accuracy. It considers both the precision and the recall of the test to compute the score: where precision is the result of the number of true positives divided by the sum of true positives and false negatives. This can be given by the following equation,

precision = tp / (tp + fp)

where tp and fp stand for true positive and false positive respectively.

Recall is the result of dividing the true positives by the sum of true positives with false negatives. This can be given as follows,

recall = tp / (tp + fn)

where tp and fn stand for true positive and false negative respectively

$F_1 = 2 \bullet precision \bullet recall / (precision + recall)$

The F_1 score can be interpreted as a weighted average of the precision and recall, where an F_1 score reaches its best value at 1 and worst at 0. These metrics altogether will give us an insight to determine how well the classifier predicts each type of child delivery.

Project Design

The first step of this capstone project is the collection and cleaning data. As I mention before I am using CDC Birth data from year 2014. This is a large data set and hence there are a lot of missing values. The idea to tackle this problem is to use CDC documentation about Birth data, where it is specified about each feature as well as values (digits) which were used instead of missing values. I assume that there still is a lot of NaN values and hence I am planning to replace them with corresponding values for each column. If there are more that 40% on missing values in a column, then I am planning to simply drop the column. In next step in cleaning data section, I drop redundant features as well as features which can carry bias information. This kind of features that give information about patients after the delivery. I, then, separate the data between women who actually had an attempt in vaginal delivery and women who chose or had to have directly C-section. For my capstone project, I only work with women who attempted a vaginal delivery which is 3 087 737 out of ~ 4 000 000.

The next part of this project will be prediction. I will consider essentially two machine learning methods: Logistic Regression and Random Forest. However, firstly I will have to transform the features into dummy variables, because there might be a mix of discrete and continuous data. The target column is binary column, where “1” is Vaginal and “2” is C-section delivery. Common problem in classification is that there might be imbalanced classes, such that one class will be always misclassified. Hence, I will have to undersample my data. Then, I will randomly split that entire data set into training and testing set, where about 30% of the data will be left for testing. Logistic Regression will be trained and tested first. I will derive F1 score and the confusion matrix to determine the accuracy of the classifier and I will repeat the same procedure for Random Forest. I will compare the accuracy of F1 score, confusion matrix and performance time of both methods and decide on the optimal model.

Next, I will probably tune the chosen method in order to improve the accuracy using GridSearch. In this capstone project I, also, want to extract important features that make the most impact on the prediction. Using information of the most important features as an output of Random Forest, I would like to apply the same model on 2015 CDC birth data. I am hoping to get approximately the same accuracy of F1 score as I estimate on 2014 CDC birth data.

References

[1] Am J Obstet Gynecol. Predicting cesarean delivery with decision tree models. 2000 Nov;183(5):1198-206

[2] Best Evidence C-Section. Childbirth Connection. childbirthconnection.org/article.asp?ck=10166.Childbirth Connection (2012).

[3] Vaginal Birth or Cesarean Birth: What Is at Stake for Women and Babies? New York: Childbirth Connection.

[4] What Every Woman Needs to Know About Cesarean Section. Childbirth Connection. childbirthconnection.org/pdfs/cesareanbooklet.pdf

[5] Martin JA, Hamilton BE, Osterman MJK, Curtin SC, Mathews TJ. Births: Final Data for 2014. National vital statistics reports; Hyattsville, MD: National Center for Health Statistics.

[6] National Center for Health Statistics. User Guide to the 2013 Natality Public Use File. Hyattsville, Maryland: National Center for Health Statistics. Annual product 2014. Available for downloading at: http://www.cdc.gov/nchs/data_access/VitalStatsOnline.htm