Using Kernel Methods and Model Selection for Prediction of Preterm Birth

Ilia Vovsha, Ansaf Salleb-Aouissi IV2121, AS2933@COLUMBIA.EDU
Computer Science Department
Columbia University, NY, NY

Anita Raja, Thomas Koch, Alex Rybchuk ARAJA, KOCH, RYBCHUK@COOPER.EDU
Albert Nerken School of Engineering
The Cooper Union, NY, NY

Axinia Radeva, Ashwath Rajan, Yiwen Huang, Hatim Diab, Ashish Tomar
ARADEVA, ANR2121, YH2726, HDIAB, AST2124@COLUMBIA.EDU
Center for Computational Learning Systems
Columbia University, NY, NY

Ronald Wapner RW2191@CUMC.COLUMBIA.EDU
Department of Obstetrics and Gynecology
Columbia University Medical Center.

Abstract

We describe an application of machine learning to the problem of predicting preterm birth. We conduct a secondary analysis on a clinical trial dataset collected by the National Institute of Child Health and Human Development (NICHD) while focusing our attention on predicting different classes of preterm birth. We compare three approaches for deriving predictive models: a support vector machine (SVM) approach with linear and non-linear kernels, logistic regression with different model selection along with a model based on decision rules prescribed by physician experts for prediction of preterm birth. Our approach highlights the pre-processing methods applied to handle the inherent dynamics, noise and gaps in the data and describe techniques used to handle skewed class distributions. Empirical experiments demonstrate significant improvement in predicting preterm birth compared to past work.

1. Introduction

Premature or preterm birth (PTB) is a major long-lasting public health problem with heavy emotional and financial consequences to families and society (March of Dimes, 2012; Conova, 2016). PTB is the leading cause of neonatal mortality and, long-term disabilities. Furthermore, over 26 billion dollars are spent annually on the delivery and care of the 12-13% of infants who are born preterm in the United States (Behrman et al., 2007). A crucial challenge is to identify women who are at the highest risk for very early preterm birth and to develop interventions. Equally important, would be the ability to identify women at the lowest risk to avoid unnecessary and costly interventions. A particularly challenging population to determine PTB risk is first time mothers (nulliparous women) due to the lack of prior pregnancy history.

Prediction of preterm birth represents a compelling application from a machine learning perspective. It has been an exceedingly challenging problem, predominantly due to (1) the inherent complexity of its heterogeneous multifactorial etiology, (2) the temporal dynamics of pregnancy and (3) the lack of approaches capable of integrating and interpreting large multidimensional data.

Risk factors of PTB are heterogenous and include history of PTB, race, age, parity of the mother, bacterial vaginosis, urinary tract infection, smoking, bleeding, cervix length. Most studies to date have examined individual risk factors independently of each other through univariate analyses of their coincidence with PTB. While these studies led to many insights on the PTB problem, current models lack sufficiently good prediction to be used clinically (Mercer et al., 1996). Previous results on this dataset using a multivariate logistic regression model show a sensitivity of 24.2% and 18.2%, and specificity of 28.6% and 33.3%, for nulliparous and multiparous women respectively.

We describe our efforts towards developing multivariate linear and non-linear models that integrate all risk factors for predicting preterm birth.[1] We use the "Preterm Prediction Study," a clinical trial dataset collected by the National Institute of Child Health and Human Development (NICHD) – Maternal-Fetal Medicine Units Network (MFMU). We compare three approaches for deriving predictive models: a support vector machine (SVM) approach with linear and non-linear kernels, logistic regression with model selection along with a hand-picked model.

We also focus our attention (as recommended by (NICHD, 2005)) on predicting (1) any kind of preterm birth, (2) spontaneous preterm birth, and (3) predicting preterm birth for nulliparous women. Furthermore, etiologies of preterm birth are believed to be different as pregnancy progresses. Hence, we also derive models at different time points, which represent the three main prenatal visits in the preterm prediction study, that is at 24, 26 and 28 weeks gestation. Our results for the spontaneous preterm birth class at 28 weeks gestation show an improvement of 20% and 30% for sensitivity and specificity respectively as compared to (Mercer et al., 1996). In addition, we obtain approximately 50% sensitivity and specificity across other data classes and time points.

This paper is organized as follows: in Section 2, we provide an overview of the risk factors and state-of-the-art systems for predicting PTB. We describe the Preterm Prediction Study dataset in Section 3 along with our pre-processing methods. We present our approach in Section 4 and our empirical evaluation along with a discussion of our results in Section 5. Finally, we discuss the significance and impact of this study and outline future work in Section 6.

2. Background

In this section, we describe the known risk factors for PTB and review state-of-the-art approaches to devise a risk-scoring system for PTB.

Risk Factors for Preterm Birth: Approximately 30% of preterm deliveries are indicated based on maternal or fetal conditions such as mother's preeclampsia and intra uterine growth restriction. The remaining 70%, known as spontaneous PTB (SPTB), occur follow-

[1] A preliminary version of this work appeared in (Vovsha et al., 2014).

ing the onset of spontaneous preterm labor, prelabor Premature Rupture Of the Membranes (pPROM), or cervical insufficiency (Goldenberg et al., 2008). Spontaneous preterm labor is a heterogeneous condition, the final common product of numerous biologic pathways that include immune, inflammatory, neuroendocrine, and vascular processes (Behrman et al., 2007). Epidemiological investigations have largely associated single factors with PTB. Of the many risk factors for preterm labor, a prior history of preterm delivery is the most predictive with a recurrence risk as high as 50% depending on the number and gestational age of previous deliveries (Goldenberg et al., 2008). It has been shown (Goldenberg et al., 1998) that the odds ratio of SPTB was highest for a positive fetal fibronectin test, followed by short cervix (Crane and Hutchens, 2008) and history of prior PTB. However, in practice, prior history of preterm delivery is used as the most predictive indicator of PTB in most clinical settings. Risk factors include race (Goldenberg et al., 2008), low socioeconomic status, extremes in age, single marital status (Smith et al., 2007; Thompson et al., 2006), low pre-pregnancy body mass index, (Hendler et al., 2005), and high-risk behaviors during pregnancy (e.g. tobacco, cocaine and heroin use). Psychological factors (Grote et al., 2010), (Zhu et al., 2010), and obstetrical conditions (Goldenberg et al., 2000; Gotsch et al., 2007; Romero et al., 2006; Tita and Andrews, 2010) are also known to increase the risk of PTB. Additional risk factors include closely spaced gestations (Conde-Agudelo et al., 2006), multiple gestations (Goldenberg et al., 2008), assisted reproductive technologies (Allen et al., 2006), exposure to tobacco smoke (Kharrazi et al., 2004; Jaakkola et al., 2001), and genetic factors (Porter et al., 1997; Winkvist et al., 1998). Those judged at risk for PTB are typically treated by prenatal administration of progesterone 17 OHPC (IM progesterone) (Acog, 2008; Flood and D. Malone, 2012). However, pregnant nulliparous women are often not treated due to the lack of prior pregnancy history.

Risk Scoring Systems for Predicting Preterm Birth: In the late 1960's, Papiernik proposed an empirical method for estimating the risk of premature delivery (Papiernik-Berkhauer, 1969). In this approach, maternal characteristics are grouped into four series of comparable variables (social status, obstetric history, work conditions, pregnancy characteristics) in a two-dimensional table. Point values varying from 1 to 5 according to the degree of their importance are assigned to all characteristics. The sum of the points gives the risk of Premature delivery. Papiernik's risk table was later modified by Creasy et al. and used in the risk of preterm delivery (RPD) system proposed in (Creasy et al., 1980) (Appendix, Table 4). Further assessment of the prediction performance of Creasy's table (Edenfield et al., 1995) on another population has shown low performance.

Another graded risk system was proposed (Mercer et al., 1996) in the context of the NICHD MFMU preterm prediction study. The results of a multivariate logistic regression were modest with sensitivity of 24.2% and 18.2%; specificity of 28.6% and 33.3%, respectively for nulliparous and multiparous women. This constitutes our baseline for comparison.

Goodwin et al. (2001) have explored the use of data mining techniques to predict preterm labor. They have identified seven demographic variables that predict preterm birth. While these results are interesting, there are concerns whether the sampling of a particular demographic would be representative of more general population. Furthermore, the experiment procedure is unclear – for example, the Area Under Curve (AUC) could have been obtained on a validation set or an unseen test set; consequently it is difficult to reproduce their results. Courtney et al. (2008) describe a secondary analysis showing that the demographic preterm prediction model generated in (Goodwin et al., 2001) generalizes to a broader population with a modest accuracy. Today, there is no widely tested risk scoring/prediction system that combines PTB factors (Davey et al., 2011).

Our models differ from the previous work as follows: (1) the dataset we study represents a diverse population from ten medical centers across the US, (2) we derive predictive models at different stages in pregnancy, (3) we derive models for specific classes of patients, namely nulliparous women and also for spontaneous PTB, and (4) the procedure we use to evaluate our models is robust and reproducible.

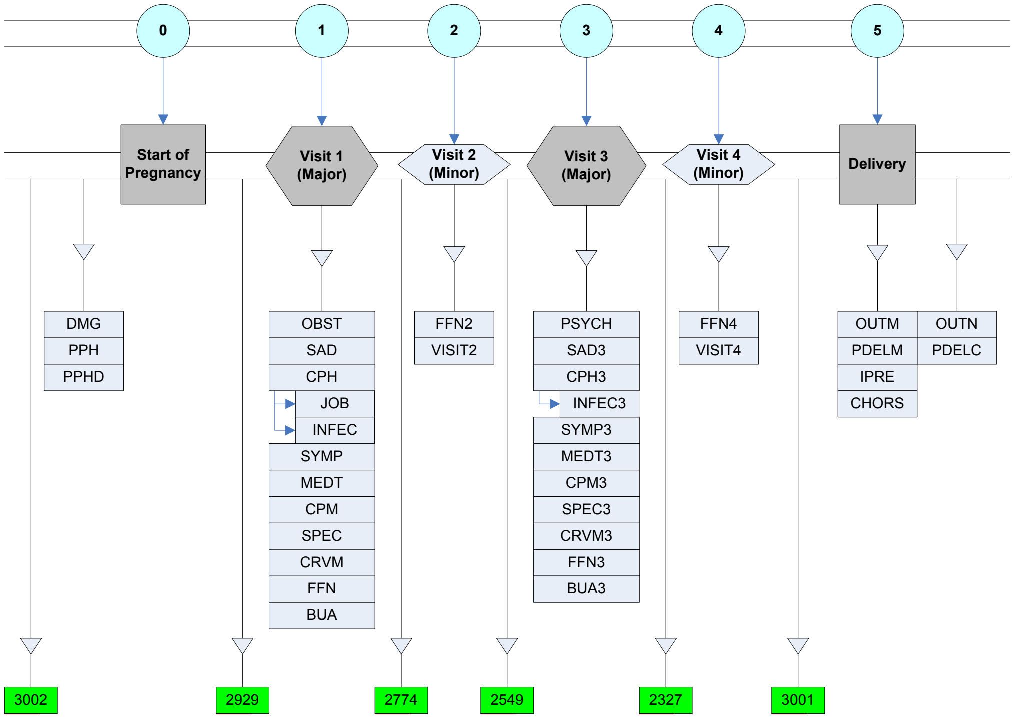
3. The Preterm Prediction Study Data

We have obtained the released data set for the *Preterm Prediction Study*, performed by the NICHD Maternal Fetal Medicine Units (MFMU) Network between 1992 to 1994. This study is an observational prospective study of 3,073 women with singleton pregnancies recruited at less than 24 weeks gestational age. Of the women enrolled, 2,929 participated in the study at the 10 participating MFMU centers across the United States between October 1992 and July 1994. There were 1,711 multiparous and 1,218 nulliparous women. The incidence of spontaneous preterm birth was 10.3% overall 8.2% for nulliparous and 11.9% for multiparous women (Mercer et al., 1996). Henceforth, we will refer to this data as the *MFMU data*. Participating women in this study were followed up by research nurses during four visits at 24 (time T0), 26 (time T1), 28 (time T3) and 30 (time T4) weeks gestation for screening tests. The MFMU data timeline is illustrated in Figure 1. The data collected from all visits Maternal Fetal Medicine Units Network (1994) has over 400 variables in all. These include demographic, behavioral, medical history, previous and current pregnancy history, digital cervical examination, vaginal ultrasound, cervical and vaginal fetal fibronectin, KOH prep for yeast tests, and a psychosocial questionnaire. The detailed outcome of the pregnancies is as follows for spontaneous PTB <32 weeks (2%), <35 weeks (4%), <37 weeks (10%); indicated PTB <37 weeks (4%). The Preterm prediction cohort singletons was released only in April 2007 under the study title: *Screening for Risk Factors for Spontaneous Preterm Delivery in Singletons and Twins* (MFM, 2007).

The MFMU data is a very rich and highly structured dataset. As a result, multiple processing steps are required. We face several challenges, including the complexity of data, missing data and skewed class distribution (addressed in Section 4). For reproducibility purposes, we describe our preprocessing steps in considerable detail in Vovsha et al. (2013).

Complexity of Data: We handle the complexity of the data by organizing features into groups (according to the original questionnaire) as depicted in Figure 1. At each visit, a set of feature groups is collected. We focus our study on the three major visits at time T0, T1 and T3.

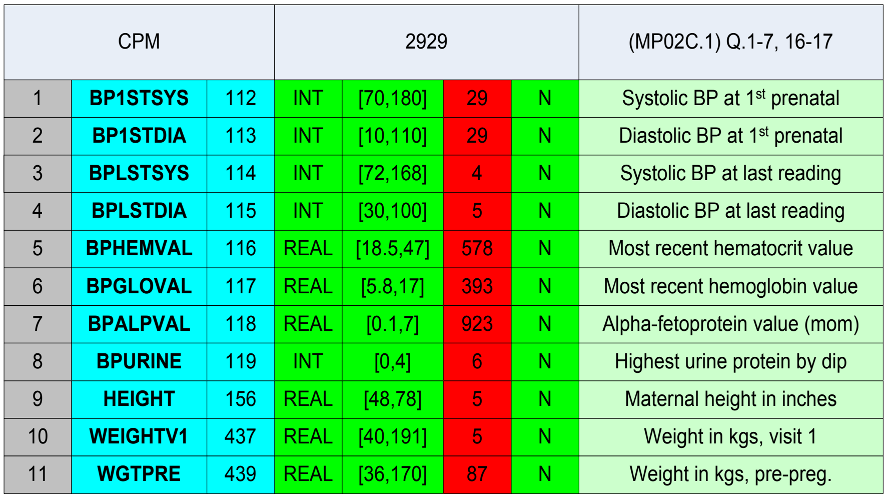
Since features are obtained from various sources, they are not always uniform. We undertake several processing steps to convert the data into a standard numerical format suitable for off-the-shelf machine learning algorithms. In particular, Yes/No features are converted to binary (1/0) values, categorical features are converted to a set of binary features, unusual values (e.g., "> 3", "2-3") are replaced with reasonable approximations (4, 2.5 respectively), and features with arbitrary ranges are normalized to the [0,1] interval.



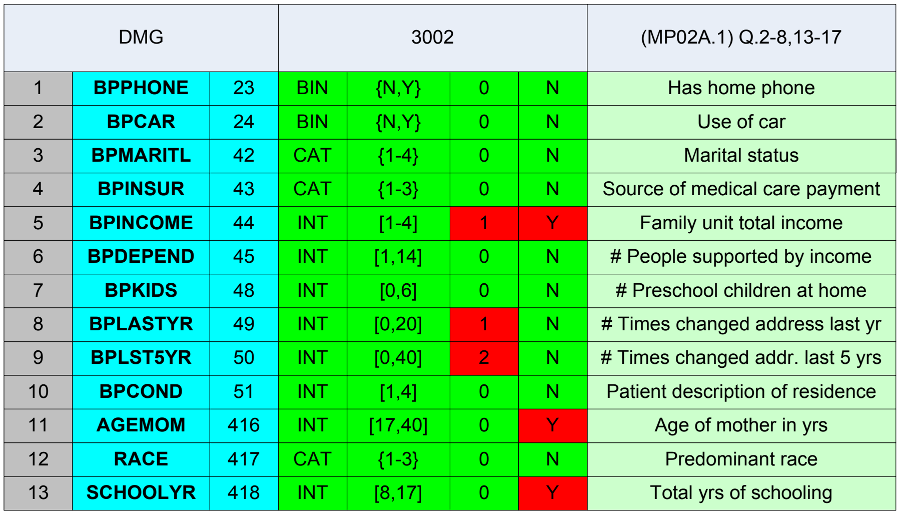
Acronym	Description	Acronym	Description
DMG	Demographics and Home Life	CRVM	Cervical Measurements V1
PPH	Previous Pregnancy History	CRVM3	Cervical Measurements V3
PPHD	Previous Pregnancy History Detail	FFN	Fetal Fibronectin Analysis V1
OBST	Obstetrical and Medical Complications	FFN2	Fetal Fibronectin Analysis V2
SAD	Substance Use V1	FFN3	Fetal Fibronectin Analysis V3
SAD3	Substance Use V3	FFN4	Fetal Fibronectin Analysis V4
CPH	Current Pregnancy History V1	BUA	Blood and Urine Analysis V1
CPH3	Current Pregnancy History V3	BUA3	Blood and Urine Analysis V3
JOB	Current or Last Job	PSYCH	Psychological Questionnaire
INFEC	Infections During This Pregnancy V1	VISIT2	Yeast and Intercourse Variables V2
INFEC3	Infections During This Pregnancy V3	VISIT4	Yeast and Intercourse Variables V4
MEDT	Medications and Treatments V1	OUTM	Pregnancy Outcome, Maternal Data
MEDT3	Medications and Treatments V3	OUTN	Pregnancy Outcome, Neonatal Data
SYMP	Symptoms During Previous Week V1	OUTS	Pregnancy Outcome Status
SYMP3	Symptoms During Previous Week V3	IPRE	Indicated Preterm Birth Reasons
CPM	Current Pregnancy Measurements V1	CHORS	Chorioamnionitis Suspected
CPM3	Current Pregnancy Measurements V3	PDELM	Preterm Delivery, Maternal Data
SPEC	Specimen Collection V1	PDELC	Preterm Delivery, Clinical Data
SPEC3	Specimen Collection V3		

Figure 1: Illustration of MFMU data timeline and description of the set of feature groups. The numbers at the bottom of the diagram indicate the number of patients that reached that point in time of the study. These numbers decrease with time for several possible reasons including: patients withdrawing from study/delivered/lost to follow up/skipped major visit etc. The last number (3001) indicates the total number of patients with known pregnancy outcomes.

Table 1: (Left) CPM group features; (Bottom) Right group features. From left to right: feature number in the group, name, feature number in the raw data, type, range, number of missing values, processing flag, description.



CPM			2929				(MP02C.1) Q.1-7, 16-17
1	**BP1STSYS**	112	INT	[70,180]	29	N	Systolic BP at 1st prenatal
2	**BP1STDIA**	113	INT	[10,110]	29	N	Diastolic BP at 1st prenatal
3	**BPLSTSYS**	114	INT	[72,168]	4	N	Systolic BP at last reading
4	**BPLSTDIA**	115	INT	[30,100]	5	N	Diastolic BP at last reading
5	**BPHEMVAL**	116	REAL	[18.5,47]	578	N	Most recent hematocrit value
6	**BPGLOVAL**	117	REAL	[5.8,17]	393	N	Most recent hemoglobin value
7	**BPALPVAL**	118	REAL	[0.1,7]	923	N	Alpha-fetoprotein value (mom)
8	**BPURINE**	119	INT	[0,4]	6	N	Highest urine protein by dip
9	**HEIGHT**	156	REAL	[48,78]	5	N	Maternal height in inches
10	**WEIGHTV1**	437	REAL	[40,191]	5	N	Weight in kgs, visit 1
11	**WGTPRE**	439	REAL	[36,170]	87	N	Weight in kgs, pre-preg.



DMG			3002				(MP02A.1) Q.2-8,13-17
1	**BPPHONE**	23	BIN	{N,Y}	0	N	Has home phone
2	**BPCAR**	24	BIN	{N,Y}	0	N	Use of car
3	**BPMARITL**	42	CAT	{1-4}	0	N	Marital status
4	**BPINSUR**	43	CAT	{1-3}	0	N	Source of medical care payment
5	**BPINCOME**	44	INT	[1-4]	1	Y	Family unit total income
6	**BPDEPEND**	45	INT	[1,14]	0	N	# People supported by income
7	**BPKIDS**	48	INT	[0,6]	0	N	# Preschool children at home
8	**BPLASTYR**	49	INT	[0,20]	1	N	# Times changed address last yr
9	**BPLST5YR**	50	INT	[0,40]	2	N	# Times changed addr. last 5 yrs
10	**BPCOND**	51	INT	[1,4]	0	N	Patient description of residence
11	**AGEMOM**	416	INT	[17,40]	0	Y	Age of mother in yrs
12	**RACE**	417	CAT	{1-3}	0	N	Predominant race
13	**SCHOOLYR**	418	INT	[8,17]	0	Y	Total yrs of schooling

We review each feature with non-standard values manually and decide what is the most appropriate processing step.

For example, consider some of the features from the "DMG" group (Table 1). The BPPHONE (has home phone) and BPCAR (use of car) are Yes/No features, BPMARITL (marital status) is a categorical feature with four categories, and the AGEMOM (age of mother) and SCHOOLYR (total years of schooling) features both have unusual values and different integer ranges. We replace unusual values in the the last two features (AGEMOM, SCHOOLYR) by rounding off from below and from above, e.g., for AGEMOM, "≤ 17" is replaced with 17 and "≥ 40" is replaced with 40.

Furthermore, PTB data is characterized by complex interdependencies among its features (physiological as well as socio-economic) which contributes to the difficulty of accurate prediction of PTB. We propose that use of non-linear methods like using the RBF kernel would pick up on these complex non-linear interdependencies and improve the prediction accuracy. We also propose to use logistic regression with model selection to automatically include whole groups of co-advisorlinear predictors and hence take into account this aspect of the data.

Missing Data: Our main objective in this work is to retain as many features and examples as possible. Hence, we prefer to fill in (complete) values rather than delete features. Since a substantial number of features is missing, we follow a simplified approach and treat features equally whether they are randomly or structurally absent. Most missing values can be reasonably completed by inserting a default value (e.g., 'No' or 0), the most common value (for categorical features), or the mean value (for numeric features). However, some features require non-trivial processing steps, and for those we sometimes include the range, mean or median, and other features in the computation as well.

As a concrete example, consider the features from the "`INFEC`" group. All 10 features in this group are Yes/No features that can be structurally absent if the patient did not report any infections during pregnancy. As such, we complete any missing values with the default 'No' value. On the other hand, the features from the "`CPM`" group are all numbers from some range of values (Table 1). All 11 of these features can be randomly missing due to the patient not undertaking a test or measurement not being reported after a visit. For several of these features (1-4), we prefer to complete the missing values with the mean of the actual responses. However, other features (5-7) have too many missing values (as shown in Table 1, column 6), and so we (reluctantly) remove these features from the dataset. Finally, some features (9-11) have particular meaning (weight, height at different points in the pregnancy), and hence we apply non-trivial processing steps to impute the value from the available information. For example, if the weight is not measured during visit 1, then we set feature 10 (`WEIGHTV1`) to the weight before pregnancy (feature 11) plus the average difference between the weights of all mothers at visit 1 and their weights before pregnancy.

4. Method for Prediction of PTB

In this section, we consider support vector machines (SVMs) and regression methods with model selection. We frame PTB as a binary classification problem, where patients who deliver a baby preterm (full-term) are assigned the positive (negative) class respectively. At every tick $(0, 1, 3)$, each patient (example) is described by a complete feature vector (see Section 3) and a label (x_i, y_i), $y_i \in \{+1, -1\}$. To validate our results, we repeat the following procedure throughout: each dataset is randomly divided into train and test sets with an 80/20 ratio, and each class is split proportionally between the sets. We then apply 5-fold cross-validation (CV) to the train set to determine the best model and optimal parameters (if any). The best model is tested on the (unseen) test set, and confusion matrices for various subsets of the data are recorded.

The metrics used are sensitivity (the percent of positive instances that are correctly predicted as positive), specificity (the percent of negative instances that are correctly predicted as negative) and the geometric mean (the square root of the product of the specificity and sensitivity).

Support Vector Machines We use support vector machines (SVMs) that belong to the family of maximum margin classifiers (Vapnik, 1995). The standard approach is to solve the soft-margin formulation (Boser et al., 1992; Cortes and Vapnik, 1995):

$$\begin{aligned} \min_{w,\xi} \quad & \frac{1}{2}||w||^2 + C \sum_{y_i} \xi_i \\ \text{s.t.} \quad & y_k[w^\top x_k + b] \geq 1 - \xi_k, \quad \xi_k \geq 0 \quad \forall k \in 1, ..., n \end{aligned}$$

where C is a positive constant determining the tradeoff between maximizing the margin and minimizing the misclassifications. The ξ's are slack variables that allow to calculate the misclassifications. An example x_i is misclassified if its corresponding slack variable $\xi_i \geq 1$. A margin error occurs if $0 \leq \xi_i \leq 1$. A large C corresponds to assigning a higher penalty to errors. This is useful, since in practice, data is rarely linearly separable. Typically, the SVM produces a classifier that labels examples x with $y = sign(w^T.x + b)$.

To account for the large discrepancy between the number of examples in each class, we scale the hinge loss penalty from the cost function proportionally to the size of each class. The cost function is thus a slightly modified version of the soft-margin SVM formulation:

$$\begin{aligned} \min_{w,\xi} \quad & \frac{1}{2}||w||^2 + C_- \sum_{y_i=-1} \xi_i + C_+ \sum_{y_j=+1} \xi_j \\ \text{s.t.} \quad & y_k[w^\top x_k + b] \geq 1 - \xi_k, \quad \xi_k \geq 0 \quad \forall k \in 1, ..., n \end{aligned}$$

By assigning different misclassification costs, we can give equal overall weight to each class in measuring performance. In order to avoid tuning two cost parameters, we set:
$C_+ \times n_+ = C_- \times n_-$ where $n_+(n_-)$ is the number of positive (negative) examples (Ben-Hur and Weston, 2010). In our experiments, we use an SVM with linear, polynomial of degree 2 and 3 along with a radial basis function (RBF) kernels.

Logistic and Lasso Regression: The regression study in this paper is motivated by the desire to create a meaningful baseline model to evaluate the performance of linear models in this problem space and assess the benefit of model selection methodologies. We consider two logistic regression model selection methodologies: l_1 lasso regression and elastic net regression (Zou and Hastie, 2005; Tibshirani, 1996). Lasso regression uses an (l_1 norm) penalty to encourage sparse solutions and perform a level of feature selection. Elastic net regression combines the sparsity induction of the l_1 norm to eliminate the trivial covariates, while using the ridge regression l_2 norm to automatically include whole groups of collinear predictors once a single covariate is added (Zou and Hastie, 2005).

Since the class distribution is skewed towards the negative (full term) class (skewed distribution challenge from the previous section), we use oversampling techniques to achieve 1:1 levels of negative to positive examples. Specifically, we use the adaptive synthetic (ADASYN) sampling approach (He et al., 2008) to adaptively generating minority data to balance the dataset.

The Creasy Baseline: We implemented the risk of preterm delivery score proposed by Creasy, as discussed in Section 2, in order to provide a baseline for comparison. In Creasy's system, patients were initially screened, and then given a follow up screening at

26 to 28 weeks' gestation. This follow up screening provided additional information on the pregnancy and improved the accuracy of the risk of preterm delivery (RPD) system (Creasy et al., 1980). Similarly, our implementation had an initial screening at T0 and then added information at the T1 and T3 time points. Our implementation had two key differences to the RPD system.

In RPD, patients are given a risk score based on factors involving socioeconomic status, past history, daily habits and status of the pregnancy. Consider for instance a mother who encountered DES (diethylstilbestrol) exposure, has a very low socioeconomic status, and is diagnosed with hypertension. This patient is assigned a score of 5 in our implementation. We manually classified these factors into three feature categories. In the first category, features did not require any modification and are used as is, such as having two children at home or hypertension. The second category of factors were simply discounted because they were either no longer prevalent (DES exposure) or the dataset lacked information on that feature (head being engaged). In the third category, features were ambiguous. For example, Creasy does not provide a clear distinction between low socioeconomic status and very low socioeconomic status. In order to compensate, we assigned a reasonable definition so that their score could be taken into account. The RPD scoring system as well as the mapping we developed of RPD risk score factors to MFMU features are available in the Appendix (Tables 4, and 5).

In RPD, the numerical risk factor is then translated into low, medium, or high risk of PTB, whereas SVM and regression methodologies are marked as only low or high risk. We modified RPD by eliminating the medium category. We ran the analysis twice, using seven and thirteen points as our boundary between low and high risk (respectively marked Creasy-7 and Creasy-13 in Table 3). A cut-off of seven points agreed with Creasy's definition of medium risk, but did not produce a similar distribution of PTB patients found in (Creasy et al., 1980), whereas a cut-off of thirteen produced a similar distribution.

5. Empirical Evaluation

Results: For each of the three problems above, we derive prediction models at different time points (ticks). Each tick (T0, T1, T3) represents a critical point (major visit) at which information is collected. In Table 2, we list the ratio of positive to negative examples in each dataset and the number of features at each tick.

We use the glmnet package to run our regression experiments. The package implements coordinate descent to train the elastic net and lasso models (Friedman et al., 2010). We display results for models with weighting factor $r = 1$ only, as there is little difference in performance between using $r = 2$ and $r = 1$. We use the sklearn package for random forest and used the "class weight" parameter to handle the data imbalance. We obtain our SVM models with modified code based on the LIBSVM package (Chang and Lin, 2011).

Table 2: Class size and feature count

	T0	**T1**	**T3**
Feature count	50	205	316
All data	434 / 2,568	423 / 2,506	334 / 2,215
Spontaneous only	309 / 2,568	302 / 2,506	240 / 2,215
Nulliparous only	156 / 1,087	153 / 1,065	112 / 951

We present our results in Table 3. For each algorithm, we show the sensitivity, specificity, and geometric mean (g-mean) performance measures (rounded off to two decimal places) on the unseen test set, averaged over five runs plus-minus the standard deviation.

Table 3: Average test rates for all algorithms at each tick

	Sensitivity			**Specificity**			**g-mean**		
	T0	**T1**	**T3**	**T0**	**T1**	**T3**	**T0**	**T1**	**T3**
Preterm vs. Fullterm, All data									
Lasso	0.59	0.52	0.50	0.59	0.67	0.73	0.59 ± 0.02	0.59 ± 0.02	0.60 ± 0.03
Elastic Net	0.59	0.51	0.50	0.59	0.67	0.73	0.59 ± 0.02	0.59 ± 0.02	0.60 ± 0.03
Linear SVM	0.40	0.43	0.45	0.83	0.82	0.84	0.58 ± 0.04	0.59 ± 0.02	0.62 ± 0.03
Poly. SVM d2	0.56	0.62	0.67	0.63	0.65	0.66	0.59 ± 0.03	0.64 ± 0.01	0.6 ± 0.04
Poly. SVM d3	0.55	0.27	0.23	0.62	0.87	0.93	0.57 ± 0.03	0.49 ± 0.02	0.46 ± 0.07
RBF SVM	0.58	0.55	0.59	0.62	0.72	0.72	0.60 ± 0.04	0.63 ± 0.03	0.65 ± 0.03
Random Forest	0.3	0.3	0.3	0.86	0.88	0.93	0.51 ± 0.07	0.48 ± 0.09	0.5 ± 0.1
Creasy-7	0.30	0.22	0.21	0.88	0.91	0.93	0.52 ± 0.04	0.45 ± 0.03	0.44 ± 0.03
Creasy-13	0.29	0.31	0.31	0.86	0.89	0.91	0.50 ± 0.04	0.52 ± 0.06	0.53 ± 0.03
Preterm vs. Fullterm, Spontaneous only									
Lasso	0.53	0.35	0.36	0.54	0.66	0.67	0.53 ± 0.02	0.48 ± 0.04	0.48 ± 0.05
Elastic Net	0.52	0.36	0.36	0.55	0.65	0.67	0.53 ± 0.03	0.48 ± 0.04	0.49 ± 0.05
Linear SVM	0.50	0.53	0.47	0.50	0.51	0.57	0.49 ± 0.02	0.52 ± 0.03	0.52 ± 0.02
Poly. SVM d2	0.56	0.44	0.41	0.48	0.58	0.6	0.51 ± 0.03	0.5 ± 0.02	0.49 ± 0.03
Poly. SVM d3	0.42	0.17	0.02	0.62	0.86	0.93	0.51 ± 0.01	0.38 ± 0.08	0.11 ± 0.1
RBF SVM	0.40	0.40	0.43	0.59	0.60	0.58	0.49 ± 0.04	0.49 ± 0.04	0.50 ± 0.02
Random Forest	0.08	0.03	0.03	0.95	0.97	0.98	0.2 ± 0.1	0.13 ± 0.1	0.14 ± 0.1
Creasy-7	0.09	0.10	0.10	0.88	0.89	0.92	0.28 ± 0.03	0.30 ± 0.02	0.30 ± 0.03
Creasy-13	0.07	0.11	0.08	0.88	0.90	0.91	0.25 ± 0.08	0.30 ± 0.05	0.26 ± 0.07
Preterm vs. Fullterm, Nulliparous only									
Lasso	0.36	0.35	0.31	0.58	0.68	0.75	0.46 ± 0.06	0.48 ± 0.05	0.47 ± 0.11
Elastic Net	0.35	0.35	0.30	0.58	0.69	0.76	0.45 ± 0.07	0.49 ± 0.06	0.47 ± 0.06
Linear SVM	0.40	0.40	0.42	0.59	0.60	0.66	0.48 ± 0.03	0.49 ± 0.06	0.52 ± 0.07
Poly. SVM d2	0.49	0.38	0.38	0.46	0.63	0.73	0.46 ± 0.04	0.48 ± 0.04	0.52 ± 0.05
Poly. SVM d3	0.38	0.15	0.18	0.61	0.88	0.93	0.48 ± 0.05	0.36 ± 0.05	0.4 ± 0.06
RBF SVM	0.41	0.34	0.42	0.64	0.64	0.68	0.50 ± 0.05	0.46 ± 0.06	0.53 ± 0.08
Random Forest	0.12	0.09	0.03	0.92	0.93	0.98	0.31 ± 0.09	0.25 ± 0.14	0.14 ± 0.1
Creasy-7	0.02	0.13	0.13	0.87	0.88	0.92	0.06 ± 0.13	0.34 ± 0.05	0.35 ± 0.05
Creasy-13	N/A	0.10	0.14	N/A	0.88	0.90	N/A ± N/A	0.22 ± 0.21	0.27 ± 0.25

Observations: As we stated in Section 2, the test error of a multivariate logistic regression model was modest (Mercer et al., 1996) with sensitivity of 24.2% and 18.2%; specificity of 28.6% and 33.3%, respectively for nulliparous and multiparous women. This constitutes our baseline for comparison. Our results for the spontaneous preterm birth class at 28 weeks gestation using linear SVMs are 47% sensitivity and 57% specificity showing an improvement of 20% and 30% respectively as compared to (Mercer et al., 1996). In addition, we obtain approximately 50% sensitivity and specificity across other data classes and time points.

We observe that SVM with a non-linear (RBF) kernel performs slightly better than linear SVM for the full data. We believe that a larger data set would highlight the advantage of the non-linear method more prominently. When we consider the entire (full) dataset, the linear/RBF SVM performs better with increasing ticks (T0 to T3 i.e., as the pregnancy progresses). This reflects our intuition that as we increasingly obtain more information (features) about each patient (example), we expect to better discriminate between them. SVMs for spontaneous and nulliparous data can sometimes lead to a poor performance. For instance the g-mean for polynomial SVMs with degree 3 for T3 is 0.11 ± 0.1. We consider the nulliparous data only to be the most difficult of the three datasets. This is especially clear at T0 when most of the critical features are derived from previous pregnancy history which is not available for nulliparous women. The high number of support vectors (not shown in the tables) required for the SVMs solution throughout the SVMs runs (across ticks, kernels, data) indicates that the preferable decision rule is approximately linear. In other words, under-fitting the data (small C value) generalizes better to unseen examples. We observe a poor performance of the Random Forest method, probably due to overfitting of decision trees on this kind of data. We have also shown that the *machine-picked* linear model presented in this paper outperforms Creasy table (Creasy et al., 1980) *hand-picked* model. In summary, our study demonstrates that model selection and non linear kernels are promising approaches for prediction of PTB.

6. Significance and Impact

Preterm birth is a challenging and complex real world problem that pushes the boundary of machine learning state-of-the-art methodologies. Today, there does not exist an effective prediction system to identify women at risk of PTB to prevent this adverse pregnancy outcome. Specifically, nulliparous women (first time mothers-to-be) remain the most vulnerable population.We present a comprehensible and reproducible study that demonstrates that more accurate prediction of preterm birth is not an elusive task. Our best performing algorithms attained (balanced) accuracy rates of 60%. We have demonstrated significant improvement compared to previous prediction performance on the same type of data and developed models that integrate heterogenous risk factors.

For future work, we plan to conduct larger scale experiments on other sources of data to study preterm birth including existing datasets and other data collected from electronic health records at a large urban hospital.

Acknowledgment

We thank the reviewers for their helpful comments. This work is partially funded by the National Science Foundation (NSF) under agreement number IIS-1454855 and IIS-1454814. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect those of the National Science Foundation.

We also acknowledge the assistance of NICHD, the MFMU Network, and the Protocol Subcommittee in making the database available on behalf of the project. The contents of this paper represent the views of the authors and do not represent the views of the Eunice Kennedy Shriver National Institute of Child Health and Human Development Maternal-Fetal Medicine Units Network or the National Institutes of Health.The authors are very thankful to Dr. Tara Randis and Dr. Mary McCord for their invaluable contributions on preterm birth background and insightful discussions in an early stage of this project.

References

MFMU dataset. https://mfmu.bsc.gwu.edu/publicly-available-data-sets, 2007.

Acog. ACOG Committee Opinion. Use of progesterone to reduce preterm birth. *Obstetrics and Gynecology*, 112(4):963–965, 2008.

V. M. Allen, R. D. Wilson, and A. Cheung. Pregnancy outcomes after assisted reproductive technology. *Journal of obstetrics and gynaecology Canada*, 28(3):220 – 50, 2006. ISSN 1701-2163.

R.E. Behrman, A.S. Butler, Institute of Medicine (U.S.). Committee on Understanding Premature Birth, and Assuring Healthy Outcomes. *Preterm birth: causes, consequences, and prevention.* National Academies Press, 2007. ISBN 9780309101592. URL http://books.google.com/books?id=9c_7kxBsKzIC.

A. Ben-Hur and J. Weston. A user's guide to support vector machines. 609:223–239, 2010.

B. Boser, I. Guyon, and V. Vapnik. A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on Computational learning theory*, COLT '92, pages 144–152, New York, NY, USA, 1992. ACM.

C-C. Chang and C-J. Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011. Software available at urlhttp://www.csie.ntu.edu.tw/ cjlin/libsvm.

A. Conde-Agudelo, A. Rosas-Bermudez, and A. C. Kafury-Goeta. Birth spacing and risk of adverse perinatal outcomes: a meta-analysis. *JAMA : the journal of the American Medical Association*, 295(15):1809 – 23, 2006. ISSN 0098-7484.

Susan Conova. Why Mothers Deliver Early - And How To Stop It. *Columbia Medicine*, 35 No. 2:16–21, 2016.

C. Cortes and V. Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, 1995.

Karen L. Courtney, Sara Stewart, Mihail Popescu, and Linda K. Goodwin. Predictors of preterm birth in birth certificate data. In *MIE*, volume 136 of *Studies in Health Technology and Informatics*, pages 555–560. IOS Press, 2008.

J.M.G. Crane and D. Hutchens. Transvaginal sonographic measurement of cervical length to predict preterm birth in asymptomatic women at increased risk: a systematic review. *Ultrasound Obstet Gynecol*, 31(5):579–87, 2008.

RK Creasy, BA Gummer, and GC. Liggins. System for predicting spontaneous preterm birth. *Obstet Gynecol*, 55(6):692–695, 1980. URL http://www.ncbi.nlm.nih.gov/pubmed/20813278.

M. Davey, L. Watson, J.A. Rayner, and S. Rowlands. Risk scoring systems for predicting preterm birth with the aim of reducing associated adverse outcomes. *Cochrane Database Syst Rev*, 11, 2011.

SM Edenfield, SD Thomas, WO Thompson, and JJ Marcotte. Validity of the creasy risk appraisal instrument for prediction of preterm labor. *Nursing research*, 44(2), 1995. URL `http://europepmc.org/abstract/MED/7892143`.

Karen Flood and Fergal D. Malone. Prevention of preterm birth. *Seminars in Fetal and Neonatal Medicine*, 17(1):58 – 63, 2012.

Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22, 2010. URL `http://www.jstatsoft.org/v33/i01/`.

R.L. Goldenberg, J.D. Iams, B.M. Mercer, P.J. Meis, A.H. Moawad, R.L. Copper, A. Das, E. Thom, F. Johnson, D. McNellis, M. Miodovnik, J.P. Van Dorsten, S.N. Caritis, G.R. Thurnau, and S.F. Bottoms. The preterm prediction study: the value of new vs standard risk factors in predicting early and all spontaneous preterm births. nichd mfmu network. *Am J Public Health*, 88(2):233–8, 1998.

Robert L. Goldenberg, John C. Hauth, and William W. Andrews. Intrauterine infection and preterm delivery. *New England Journal of Medicine*, 342(20):1500–1507, 2000. URL `http://www.nejm.org/doi/full/10.1056/NEJM200005183422007`.

Robert L Goldenberg, Jennifer F Culhane, Jay D Iams, and Roberto Romero. Epidemiology and causes of preterm birth. *The Lancet*, 371(9606):75 – 84, 2008. ISSN 0140-6736. URL `http://www.sciencedirect.com/science/article/pii/S0140673608600744`.

L.K. Goodwin, M.A. Iannacchione, W.E. Hammond, P. Crockett, S. Maher, and K. Schlitz. Data mining methods find demographic predictors of preterm birth. *Nursing Research*, 50(6):340–5, 2001.

F. Gotsch, R. Romero, J. P. Kusanovic, S. Mazaki-Tovi, B. L. Pineles, O. Erez, J. Espinoza, and S. S. Hassan. The fetal inflammatory response syndrome. *Clinical Obstetrics and Gynecology*, 50(3):652 – 83, 2007. ISSN 0009-9201. URL `http://journals.lww.com/clinicalobgyn/Fulltext/2007/09000/The_Fetal_Inflammatory_Response_Syndrome.11.aspx`.

Nancy K. Grote, Jeffrey A. Bridge, Amelia R. Gavin, Jennifer L. Melville, Satish Iyengar, and Wayne J. Katon. A meta-analysis of depression during pregnancy and the risk of preterm birth, low birth weight, and intrauterine growth restriction. *Arch Gen Psychiatry*, 67(10):1012–1024, 2010. URL `http://archpsyc.ama-assn.org/cgi/content/abstract/67/10/1012`.

Haibo He, Yang Bai, Edwardo A. Garcia, and Shutao Li. Adasyn: Adaptive synthetic sampling approach for imbalanced learning. In *IJCNN*, pages 1322–1328, 2008.

Israel Hendler, Robert L. Goldenberg, Brian M. Mercer, Jay D. Iams, Paul J. Meis, Atef H. Moawad, Cora A. MacPherson, Steve N. Caritis, Menachem Miodovnik, Kate M. Menard, Gary R. Thurnau, and Yoram Sorokin. The preterm prediction study: Association between maternal body mass index and spontaneous and indicated preterm birth. *American*

Journal of Obstetrics and Gynecology, 192(3):882 – 886, 2005. ISSN 0002-9378. URL http://www.sciencedirect.com/science/article/pii/S0002937804010397.

Jouni J. K. Jaakkola, Niina Jaakkola, and Kolbjrn Zahlsen. Fetal growth and length of gestation in relation to prenatal exposure to environmental tobacco smoke assessed by hair nicotine concentration. *Environmental Health Perspectives*, 109(6):pp. 557–561, 2001. ISSN 00916765. URL http://www.jstor.org/stable/3455027.

M. Kharrazi, G. N. DeLorenze, F. L. Kaufman, B. Eskenazi, Jr. Bernert, J. T., S. Graham, M. Pearl, and J. Pirkle. Environmental tobacco smoke and pregnancy outcome. *Epidemiology*, 15(6):660 – 70, 2004. ISSN 1044-3983. URL http://journals.lww.com/epidem/Fulltext/2004/11000/Environmental_Tobacco_Smoke_and_Pregnancy_Outcome.4.aspx.

March of Dimes. Born too soon report. http://www.who.int/pmnch/media/news/2012/201204_borntoosoon-report.pdf. 2012.

Maternal Fetal Medicine Units Network. Screening for Risk Factors for Spontaneous Preterm Birth – manual of operations, 1994.

B.M. Mercer, R.L. Goldenberg, A. Das, A.H. Moawad, J.D. Iams, P.J. Meis, R.L. Copper, F. Johnson, E. Thom, D. McNellis, M. Miodovnik, M.K. Menard, S.N. Caritis, G.R. Thurnau, S.F. Bottoms, and J. Roberts. The preterm prediction study: A clinical risk assessment system. *American Journal of Obstetrics and Gynecology*, 174(6):1885 – 1895, 1996. ISSN 0002-9378. URL http://www.sciencedirect.com/science/article/pii/S0002937896702259.

NICHD. Pregnancy and Perinatology Branch – Strategic Plan 2005-2010, 2005.

E. Papiernik-Berkhauer. [coefficient of premature delivery risk (c.p.d.r)]. *Presse Med*, 77 (21):793–4, 1969.

T. F. Porter, A. M. Fraser, C. Y. Hunter, R. H. Ward, and M. W. Varner. The risk of preterm birth across generations. *Obstetrics and gynecology*, 90(1):63–7, 1997. ISSN 0029-7844.

R Romero, J Espinoza, JP Kusanovic, F Gotsch, S Hassan, O Erez, T Chaiworapongsa, and M Mazor. The preterm parturition syndrome. *BJOG: An International Journal of Obstetrics and Gynaecology*, 113:17–42, 2006. ISSN 1471-0528. URL http://dx.doi.org/10.1111/j.1471-0528.2006.01120.x.

L K Smith, E S Draper, B N Manktelow, J S Dorling, and D J Field. Socioeconomic inequalities in very preterm birth rates. *Archives of Disease in Childhood - Fetal and Neonatal Edition*, 92(1):F11–F14, 2007. URL http://fn.bmj.com/content/92/1/F11.abstract.

John M. D. Thompson, Lorentz M. Irgens, Svein Rasmussen, and Anne Kjersti Daltveit. Secular trends in socio-economic status and the implications for preterm birth. *Paediatric and Perinatal Epidemiology*, 20(3):182–187, 2006. ISSN 1365-3016. URL http://dx.doi.org/10.1111/j.1365-3016.2006.00711.x.

R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society (Series B)*, 58:267–288, 1996.

Alan T.N. Tita and William W. Andrews. Diagnosis and management of clinical chorioamnionitis. *Clinics in Perinatology*, 37(2):339 – 354, 2010. ISSN 0095-5108. URL http://www.sciencedirect.com/science/article/pii/S0095510810000217. Early Onset Neonatal Sepsis.

V. N. Vapnik. *The nature of statistical learning theory.* Springer-Verlag New York, Inc., New York, NY, USA, 1995. ISBN 0-387-94559-8.

Ilia Vovsha, Ansaf Salleb-Aouissi, Axinia Radeva, Anita Raja, Ashish Tomar, Ronald Wapner, , Hatim Diab, Ashish Tomar, and Ashwath Rajan. Data pre-processing for the preterm prediction study MFMU dataset. http://www1.ccls.columbia.edu/~ansaf/CING/CCLS-13-04.pdf. Technical report, CCLS Columbia University, 2013.

Ilia Vovsha, Ashwath Rajan, Ansaf Salleb-Aouissi, Anita Raja, Axinia Radeva, Hatim Diab, Ashish Tomar, and Ronald Wapner. Predicting preterm birth is not elusive: Machine learning paves the way to individual wellness. AAAI Spring Symposium Series. 2014. URL https://www.aaai.org/ocs/index.php/SSS/SSS14/paper/view/7694/7788.

Anna Winkvist, Ingrid Mogren, and Ulf Hgberg. Familial patterns in birth characteristics: impact on individual and population risks. *International Journal of Epidemiology*, 27(2): 248–254, 1998. URL http://ije.oxfordjournals.org/content/27/2/248.abstract.

Peng Zhu, Fangbiao Tao, Jiahu Hao, Ying Sun, and Xiaomin Jiang. Prenatal life events stress: implications for preterm birth and infant birthweight. *American Journal of Obstetrics and Gynecology*, 203(1):34.e1 – 34.e8, 2010. ISSN 0002-9378. URL http://www.sciencedirect.com/science/article/pii/S000293781000236X.

Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, Series B*, 67:301–320, 2005.

Appendix

Table 4 shows the Risk of Preterm Delivery system (RPD) Creasy et al. (1980). This scoring system is a modification of the CPDR system proposed in Papiernik-Berkhauer (1969). The final score is computed by addition of the number of points given any item. A final score between 0 and 5 is classified as *low risk*; a score between 6 and 9 as *medium risk* and any score higher or equal to 10 deemed as *high risk* score for preterm birth.

Table 4: Risk of Preterm Delivery (RPD) Creasy et al. (1980)

Points	**Socioeconomic status**	**Past history**	**Daily habits**	**Current pregnancy**
1	2 children at home	1 abortion	Work outside home	Unusual fatigue
	Low socioeconomic status	Less than 1 year since last birth		
2	Younger than 20 years	2 abortions	More than 10 cigarettes per day	Less than 13 kg gain by 32 weeks' gestation
	Older than 40 years			
	Single parent			Albuminuria
				Hypertension
				Bacteriuria
3	Very low socioeconomic status	3 abortions	Heavy work	Breech at 32 weeks
			Long tiring trip	Weight loss of 2 kg
	Shorter than 150 cm			Head engaged
	Lighter than 45 kg			Febrile illness
4	Younger than 18 years	Pyelonephritis		Metrorrhagia after 12 weeks' gestation
				Effacement
				Dilatation
				Uterine irritability
5		Uterine anomaly		Placenta previa
		Second trimester abortion		Hydramnios
		DES exposure		
10		Premature delivery		Twins
		Repeated second-trimester abortion		Abdominal surgery

Table 5 shows the mapping we have developed of RPD factors to MFMU features.

Table 5: Mapping of RPD factors to MFMU dataset

RPD Factor	**MFMU Feature**
2 children at home	BPKIDS ≥ 2
Low socioeconomic status	BPINCOME == 1 and SCHOOLYR == (13 or 14)
Younger than 20 years	AGEMOM < 20 and $\neq 18$
Older than 40 years	AGEMOM == 40
Single parent	BPMARITL_2 == 1 or BPMARITL_3 == 1
Very low socioeconomic status	BPPHONE == 0 and BPCAR == 0 and BPINCOME == 1 and BPWORK == 0 and SCHOOLYR < 13
Shorter than 150 cm	HEIGHT < 59
Lighter than 45 kg	WGTPRE < 45
Younger than 18 years	AGEMOM = 18
1 abortion	BPINDUCE == 1
Less than 1 year since last birth	LASTPREG == 0
2 abortions	BPINDUCE == 2
3 abortions	BPINDUCE > 2
Pyelonephritis	BPINFEC == 1 and PYELO == 1
Uterine anomaly	BPFIBR == 1 or BPLOWER_2 == 1
Second trimester abortion	SECAB > 0
DES exposure	N/A
Premature delivery	PRETERM == 1
Repeated second-trimester abortion	N/A
Work outside home	BPJOB == 1
More than 10 cigarettes per day	BPSMOKE == 1 and CIGSPRE >= 10
Heavy work	N/A
Long tiring trip	N/A
Unusual fatigue	BPSTAND == 1 or BPBREAK == 0 or BPVIBES == 1 or BPHRS > 50
Less than 13 kg gain by 32 weeks gestation	WEIGHTV3 - WEIGHTV1 < 13
Albuminuria	BPURINE > 0
Hypertension	BPHYPER == 1
Bacteriuria	BACTER == 1
Breech at 32 weeks	N/A
Weight loss of 2 kg	WEIGHTV3 - WEIGHTV1 < -2
Head engaged	N/A
Febrile illness	HERPES == 1 or VHERPES == 1 or CYS == 1 or VCYS == 1
Metrorrhagia after 12 weeks' gestation	BPVAG2ND == 1 or PERBLD == 1
Effacement	N/A
Dilatation	BPCRVLT < 25
Uterine irritability	N/A
Placenta previa	N/A
Hydramnios	OLIGO == 1
Twins	N/A
Abdominal surgery	BPABD == 1