IST 718 Group #2

Final Project Report

Professor: Willard Williamson

Student Names: Yifan Wang, Jing Sun, Zequn Che

Table of Contents

1. Abstract

Newborns are the hopes of the parents and also the future of the world. People all hope that newborn babies could be healthy. However, there are still some unfortunate infants which are not healthy when they are born. For helping enhance the health rate of newborn infants, we get the open data (Natality Birth Data) of infants' health situations and the information of their parents from the National Bureau of Economic Research. The dataset records most situations of newborn infants in detail. We believe analyzing the dataset will help us find the key factors of fetal viability and newborns' health situation. Furthermore, analyzing the dataset could build a model which can predict the health of newborn infants.

The main goal of this project is to perform data analytics on 2018 Natality Birth Data and find insights related to the newborn infants' health. By building models to predict natality and infants' health situation, we are able to interpret the models and capture key influencing factors.

2. Data

2.1. Data Collection

In this project, we use 2018 natality data from the National Vital Statistics System of the National Center for Health Statistics. The original dataset has 3801534 rows and 240 columns. Each row represents one birth record in 2018.

The predictors can be roughly be categorized into categories:

- Basic birth info: birth time, birth place, plurality, sex, parents' age, race and education, etc.
- Maternal Behavior: prenatal care began time, smoking habit, mother's height, weight, pregnancy history
- Pregnancy risk factors: pre-pregnancy diabetes, gestational diabetes, gestational hypertension
- Infections present: gonorrhea, syphilis, chlamydia, hepatitis B
- Characteristics of labor and delivery: induction of labor, augmentation of labor,

steroids, antibiotics, delivery method

- Maternal morbidity: maternal transfusion, perineal laceration, ruptured uterus, admit to intensive care

Link to the dataset: https://data.nber.org/data/vital-statistics-natality-data.html

There are many columns presenting the same feature but with different coding methods. For example, there are 3 columns representing "mothers' age" feature. One is single years of age, the other two bin ages with different ranges. Since the dataset is too large to load in Google Colab, we went over all the columns and dropped those repetitive columns. We ended up decreasing the dataset to 78 columns.

Interesting about the data: Out of 3801534 newborns records in the dataset, 9683 are dead, meaning the newborn mortality rate is 0.255%. 421343 infants have abnormal conditions, which represent 11.083% of the total. 13314 infants have Congenital anomalies, which is 0.350% of the total.

2.2. Data Cleaning and Exploration

First of all, we used the Pandas package to load the original dataset and perform some necessary dataset cleaning steps. The original dataset contains 240 columns. There are some empty columns and some columns presenting the same feature but with different coding methods. For example, there are 3 columns all representing mothers' age. One is single years of age, the other two bin ages with different ranges. We went over all the columns and dropped empty and repetitive columns. We ended up decreasing the dataset to 78 columns. The current dataset contains features like birthplaces, parents' ages, how many cigarettes the parents smoke every day before pregnancy, and if the parents have diabetes. Then we generated a new CSV file after the above steps for future convenience.

Next, we read the new CSV file via PySpark. After reading the dataset, we printed the dataset schema to check each column's data type. The result shows each column of our dataset is a string type. We casted them from the string into integers or floats.

Moreover, we performed a statistical analysis of our dataset. We first want to know if the data frame has the same rows or duplicate data records by count function and distinct function. The count of rows is 3801534, and the count of distinct rows is 3801534, which means there is no redundant information. Next, we checked if our data frame has any missing data, and

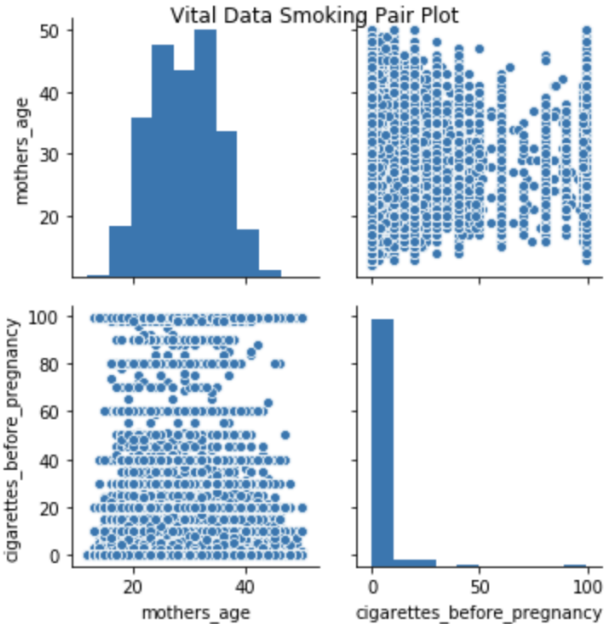
we found no missing values in the dataset. Then we grouped our data by our target variable to count the frequency distribution of living cases and dropped any unknown records. We noticed that the live number is 3784682, and the death number is 9683, which means our dataset has an imbalanced phenomenon. Therefore, some necessary data processing steps might apply to future development. After the above steps, we created dummy variables to encode our target variable as 1 or 0, renamed the dummy column Y, and dropped the original target column and dummy column N.

Target Variable	Yes	No
No abnormal condition	3376738	421343
No congenital anomalies	3782113	13314
infant alive	3224846	2213

We extracted the numerical data from our data frame and performed the describe function on them to get statistical summaries. Part of the results as the following figure:

	summary	mothers_age	fathers_age	cigarettes_before_pregnancy	cigarettes_1_trimester	cigarettes_2_trimeste
0	count	3794365	3794365	3794365	3794365	379436
1	mean	29.009387868589343	39.596097370706296	1.550107330212038	1.131522402299199	0.969778342357680
2	stddev	5.8052783803689225	22.449944281961567	8.107888249657071	7.4407016553343075	7.237194118143928
3	min	12	11	0	0	
4	max	50	99	99	99	9

One more interesting thing we found from our data points was that smoking before pregnancy is a common phenomenon in different ages. The following plot is our visualization result:



We are interested in three target variables:

- **Infant living** at time of report
- **Congenital anomalies** of the newborn
- **Abnormal conditions** of the newborn

After the previously dataset exploration and dimension reduction, our dataset contains 75 columns that served for our three inferences. For our first inference, we performed a deeper data cleaning based on the existing columns. First, we created a subset containing 21 columns after filtering, which contains what we need for deciding which predictors are important for a live birth.

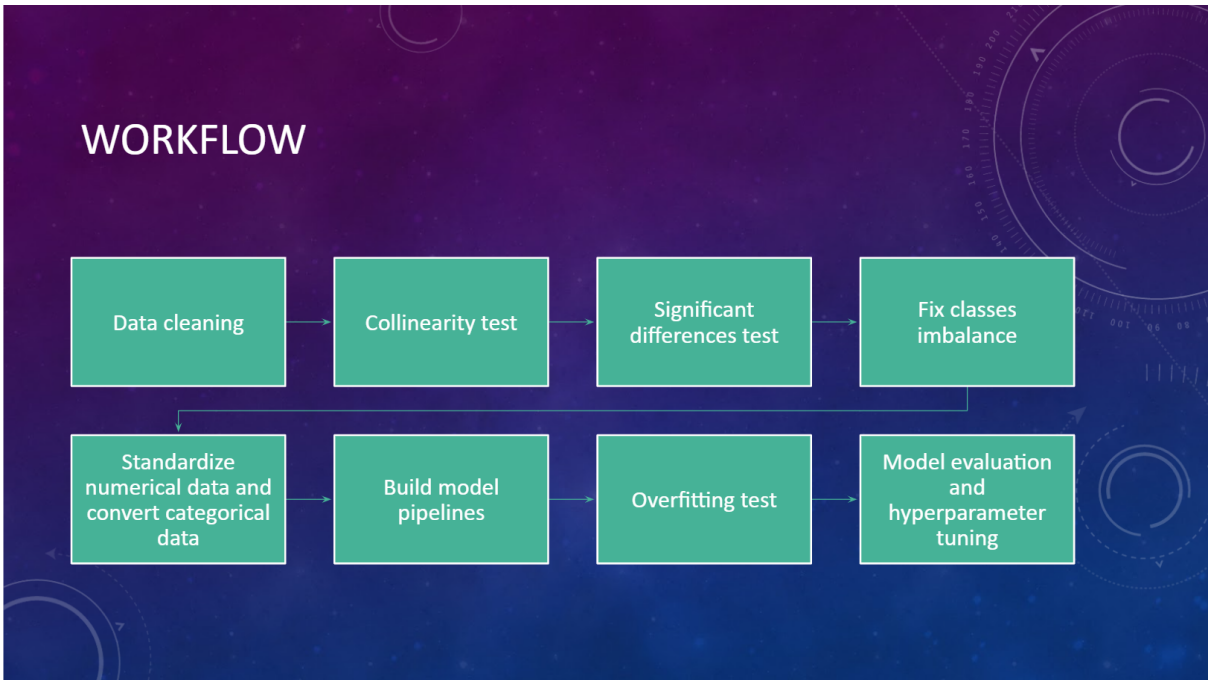
Next, we read the new subset file via our customized struct field function and schema. We pre-defined the columns' data type based on their values. After the above steps, we checked if there were any missing values, then we filtered out all unknown cases, because we could not decide whether those unknown cases were alive or not. Moreover, we calculated the correlations between all the numeric data to see if any collinearity exists. We dropped three columns since their values are extremely close to 1, which means they are also predicting other features except our target.

In addition, we casted our target label from categorical values into 1 and 0 to make it in the right format for our models. Next we calculated the Pearson correlation coefficients to if

there are any significant differences within our categorical features. We dropped any column that has the P-Value greater than 0.05.

3. Methodology

Data science workflow



We first performed several data cleaning steps, such as check duplicate values, check missing values, cast data types, create dummy variables; Then we performed a collinearity analysis on all numerical data, and remove any features that scored close to 1; Next we performed significant difference analysis on categorical variables and remove any features with P-value greater than 0.05; we tried some strategie to solve the class imbalance; then we build pipelines to train our models and check if there existing overfitting; last, we selected the best model based on the performance and applied hyperparameter tuning.

4. Models

4.1. Models:

During our first inference task, we used two common supervised classification models:

Logistic Regression and **Random Forest**.

For **Logistic Regression**: This model is used to predict the class or category of individuals based on one or multiple predictor variables(X). It is used to model a binary outcome, that is a variable, which can have only two possible values. Logistic regression does not return directly the class of observations. It used to estimate the probability of class membership. The probability will range between 0 and 1.

For **Random Forest**, it consists of a large number of individual decision trees that operate as an ensemble. Each individual tree in the random forest spits out a class prediction and the class with the most votes becomes our model's prediction.

4.2. Problems and Solutions:

For qualifying the above two models' requirements, we performed necessary data transformations. For example: We built pipelines to transfer categorical ordinal data into dummy variables via combining the StringIndexer and OneHotEnoder. Moreover, we applied under sampling on our original dataset to handle the imbalanced classes issue and trained the smaller dataset with Logistic Regression to see if the under sampling helped us or not. Last, we extracted the pipeline information after fitting it with the training set and compared its results with the transformed model to see if there exists overfitting. We found the ROC score did have overfit, but the PR score did not. So, we added a regularization parameter during the hyperparameters tuning phase to solve the overfit.

4.3. Evaluation metrics:

Since our dataset exists the imbalanced classes phenomenon, so accuracy rate is not the proper evaluation indicator for this project. So, we decided to use the PR score, which summarize the trade-off between the true positive rate and the positive predictive value for our model using different probability thresholds, as our primary choice; Then we also decided use the ROC score as the second choice, since the area under ROC shows how well the performance of our classifiers during the inference task one, we would know if the our model selections were proper.

For MSE, it is not an appropriate cost function for our project. There are two main reasons that MSE is a bad choice for binary classification problems: first, using MSE means that we

assume that the underlying data has been generated from a normal distribution, but the reality is our dataset is not following normal distribution; Secondly, the MSE function is non-convex for binary classification. In simple terms, if a binary classification model is trained with MSE Cost function, it is not guaranteed to minimize the Cost function.

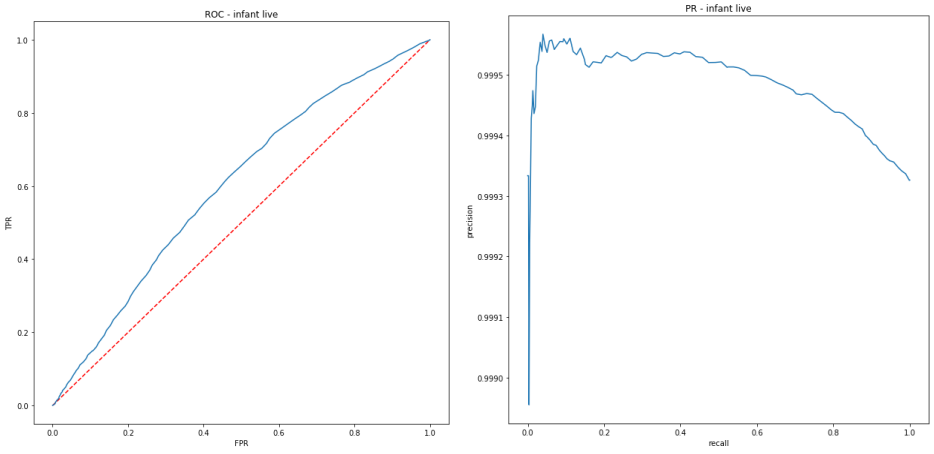
4.4. Inference study:

4.4.1. Infant living:

According to our inferences' results, the logistic regression model suggests that the following features has positive impact of infant live birth: weight gain, gestational hypertension, pre-pregnancy hypertension, previous preterm birth, cigarettes smoke numbers during the first period of trimester, and mothers' height.

Our random forest model suggests that the importance of features for infant live birth are previous preterm birth, hypertension eclampsia, pre-pregnancy weight, and birth places. Especially for the previous preterm birth, it scored almost 0.90, which means this feature has extremely influence on infant living.

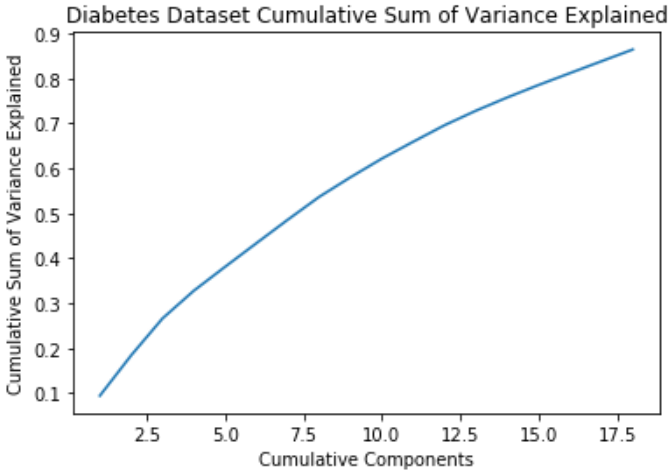
Last, we used the train-validation split model's result to decide the top seven features that impact newborns live, we noticed that if the parents smoke during the first trimester has a huge negative influence on our predictions.



As a summarize, we believe the feature: number of cigarettes smoked during the first trimester, is what we should pay attention to when predicting if the infant will live or not. We strongly recommend the parents, especially mothers, stop smoking during the first trimester.

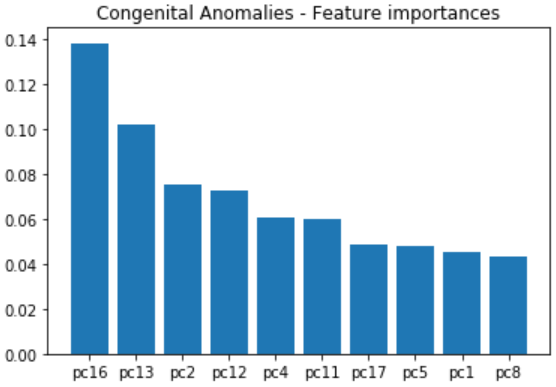
4.4.2. Congenital anomalies

In the analysis of the congenital anomalies situations, we used 23 features to analyze the results. We first filtered out the unknown target features. By PCA analysis, we kept 18 features. The 18 principle components achieve about 86.45% of the total variance in the dataset. Since the dataset is very imbalanced, we created both train and test data which have a fair number of instances from minority class. We split the dataset by class and resample respectively and combine them.



For the sake of running time, we use TrainValidationSplit for hyper-parameter tuning instead of cross validation. We have 3 methods to solve the problem. Method #1, we undersampled the data. Method #2, we assign higher weight to instances from minority class. The last method, we SMOTE and undersample the data and then do the Random Forest. By using these three methods, we could see there is a huge improvement in predicting minority class. The three methods do not have distinctive differences in regards to PR score. We used the undersampling & random forest method which has the highest PR score.

	mode	PR
0	RandomForest-undersampling	0.998312
1	LogisticRegression-weighted	0.998302
2	RandomForest-SMOTE	0.998323



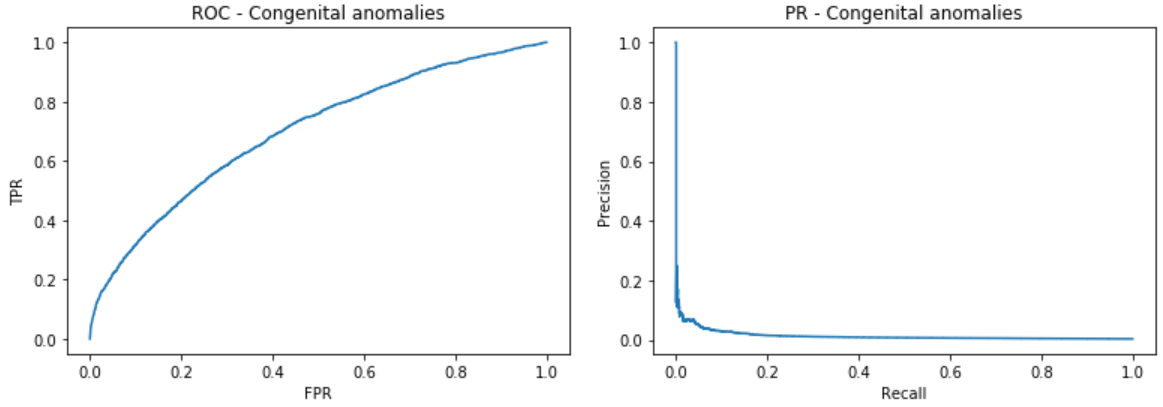
Highest loadings of pc16

	features	loading	abs_loading
14	dad_raceclassVec_5	0.917674	0.917674
33	prior_dead	0.296104	0.296104
42	weight_gain	0.181359	0.181359
11	dad_raceclassVec_4	-0.138303	0.138303
12	dad_raceclassVec_6	-0.0912606	0.0912606
45	gestation	0.0701347	0.0701347
31	mothers_age	0.0339018	0.0339018
32	fathers_age	0.0311349	0.0311349
35	prenatal_visit	0.02992	0.02992
22	delivery_methodclassVec_1	0.0180129	0.0180129

Highest loadings of pc13

	features	loading	abs_loading
16	no_riskclassVec_1	-0.499036	0.499036
17	no_riskclassVec_0	0.499036	0.499036
44	plurality	-0.395966	0.395966
23	delivery_methodclassVec_2	-0.307826	0.307826
22	delivery_methodclassVec_1	0.307826	0.307826
43	apgar	0.216733	0.216733
40	bmi	-0.164026	0.164026
41	prepregnancy_weight	-0.161004	0.161004
45	gestation	0.11448	0.11448
32	fathers_age	-0.10741	0.10741

From the inference analysis, we could see that the Component 16 and Component 13 are the two features with the highest contribution to the prediction. For the pc16, the key features are fathers' race, prior dead, and weight gain. For the pc13, the key features are 'no_riskclassVec' which is about the normal risky actions like pre-pregnancy diabetes or Hypertension. The plurality and delivery methods are also important to the congenital anomalies.

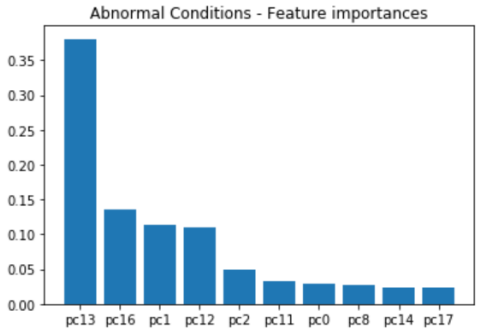


The PR plot is not ideal. It is because we can not improve precision by a lot. Everytime a true positive instance is predicted, it comes with false positive cases.

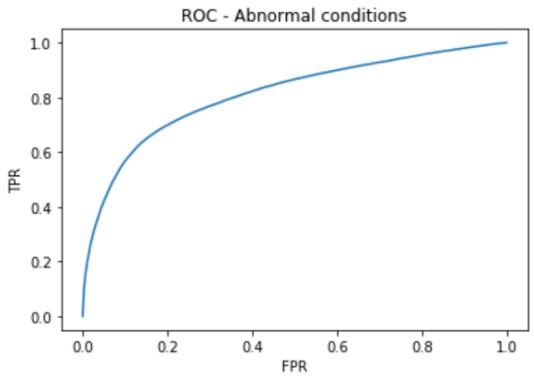
4.4.3. Abnormal conditions

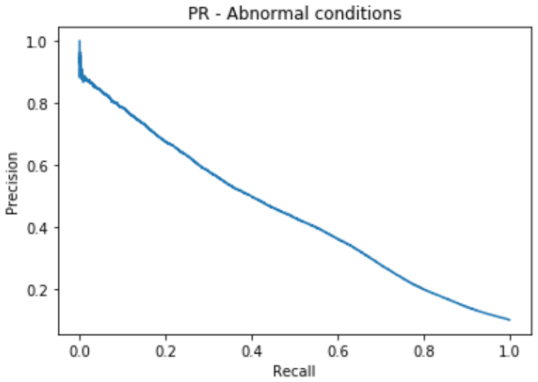
In the analysis of abnormal conditions, we used a similar process as the analysis of congenital anomalies part. We tested three different methods to predict minority class well. The three methods are undersampling, weighting, and SMOTE & undersampling. We chose the undersampling method which has the best performance. For the sake of running time, we directly go with the model in predicting congenital anomalies.

```
PR: 0.9669776119460161
confusion matrix:
[[ 40007  18486]
 [ 96431 431024]]
```



In the inference analysis, the outcome shows pc13 and pc16 are the key features which is similar to the analysis of congenital anomalies in 4.4.2.





The ROC and PR of the analysis are good.

5. Conclusion

5.1. Infant living

For improving the infant living rate. Puerperas should strictly stop smoking during the first trimester.

5.2. Congenital anomalies

For reducing the risks of congenital anomalies, we should pay attention to the prior death of infants, weight gain of the puerperas. The normal risky conditions like diabetes or hypertension are important features which will cause congenital anomalies.

5.3. Abnormal conditions

For avoiding any abnormal conditions, the health conditions of puerperas are

the key features which will influence the abnormal conditions of infants.