**Grade: 100%**

IST 718 Group 2

Project Proposal

Professor: Willard Williamson

Student Names: Yifan Wang, Jing Sun, Zequn Che

**Objective**

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.

**Data Set Description**

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, birthplace, 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.

**Preliminary Data 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.

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:

表格

描述已自动生成

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:

图表

描述已自动生成

**Predictions**

We are interested in three target variables:

- Infant living at time of report

- Abnormal conditions of the newborn

- Congenital anomalies of the newborn

**Inference**

- Determine which predictors are important for a live birth

- Determine which features are important to predict if newborns have abnormal conditions

- Determine which features are important to predict if newborns have congenital anomalies

These important predictors will have implications for medical research and treatment to increase birth rate and newborn health.

**Non-Spark Packages**

* Handy\_Spark