Calanit Kamala

Springboard 2021

CAPSTONE II FINAL REPORT

Birth outcome classification model

For some, childbirth is perceived as a natural process full of magic and mystery. For others, the miracle of modern medicine is the magic they seek to have a safe birth that results in the best outcome. An on-going issue of concern is whether medical necessity determines the extent of intervention during child-birth.

The pinnacle of medical intervention during childbirth is the Cesarean operation, more commonly known as a c-section. The term for Cesarean birth comes from the Latin verb *Caedere*, which means 'to cut'. In ancient times a c-section was performed when the mother passed away in labor, in an attempt to save the baby. Only with the development of modern medicine and anesthesia, mothers were able to survive the operation.

Over the last few decades c-section rates have increased significantly. Today, nearly a third of U.S. births result in a c-section, a significant increase from the 20% rate in 1995. 80% of those c-sections were performed in first birth, low risk pregnancy, which may indicate that c-sections are sometimes preformed not for medical reasons. In fact, the WHO estimates that 15% of c-sections performed world-wide are not medically necessary.

C-section is undoubtedly a risky operation that prolongs mother's recovery and potentially affects breastfeeding and baby's immune system. Unfortunately, there are strong incentives to resort to a c-section even when one is not medically necessary.

First, a c-section has very established process and time-line. The baby is born within 10 minutes and the surgery is complete in less than an hour. For busy doctors with schedule constraints, that level of certainty may be appealing. Second, doctors in the U.S are paid significantly more for a c-section than a vaginal birth.

When incentives between doctors and their patients are misaligned, data-driven analysis can empower individuals to have the appropriate information to make informed health choices.

By employing a ML model to classify birth outcomes, the understanding of the variables that affect the prevalence of c-section can be further understood. Understanding the variation in birth outcomes can empower individuals as well as health practitioners to reduce medically unnecessary c-section rates.

Data Collection & Date Cleaning

The CDC provides detailed birth information on all U.S. births that includes extensive details on mother's health and socio-economic characteristics, as well as birth outcomes. Annual data is available to download at https://www.cdc.gov/nchs/data_access/vitalstatsonline.htm. The data is in fixed format, with each record being comprised of a collection of letters, numbers and spaces to create a length of 1330 characters per record. Here is an example of a single record:

```
201901    11353        11               1                         29094    1
1033033 0  013 X1  14 1                   310503303   00134 1     010001  2
3               09708   01803   01803     0211          28  111
N10000000000001111N1         62129.83    1631   191 12831
NYNYNN111111NXX111N00 110     NNNNN111111      NN 11
NNNNNY1111110     14U111321     NNNNN 11111 1     3N2211     0831885   1    9
F 03  2018     40082    330411537 042     NNYNNN 111111 0
NNNNNN111111NNNNNN1111111     NYY1
```

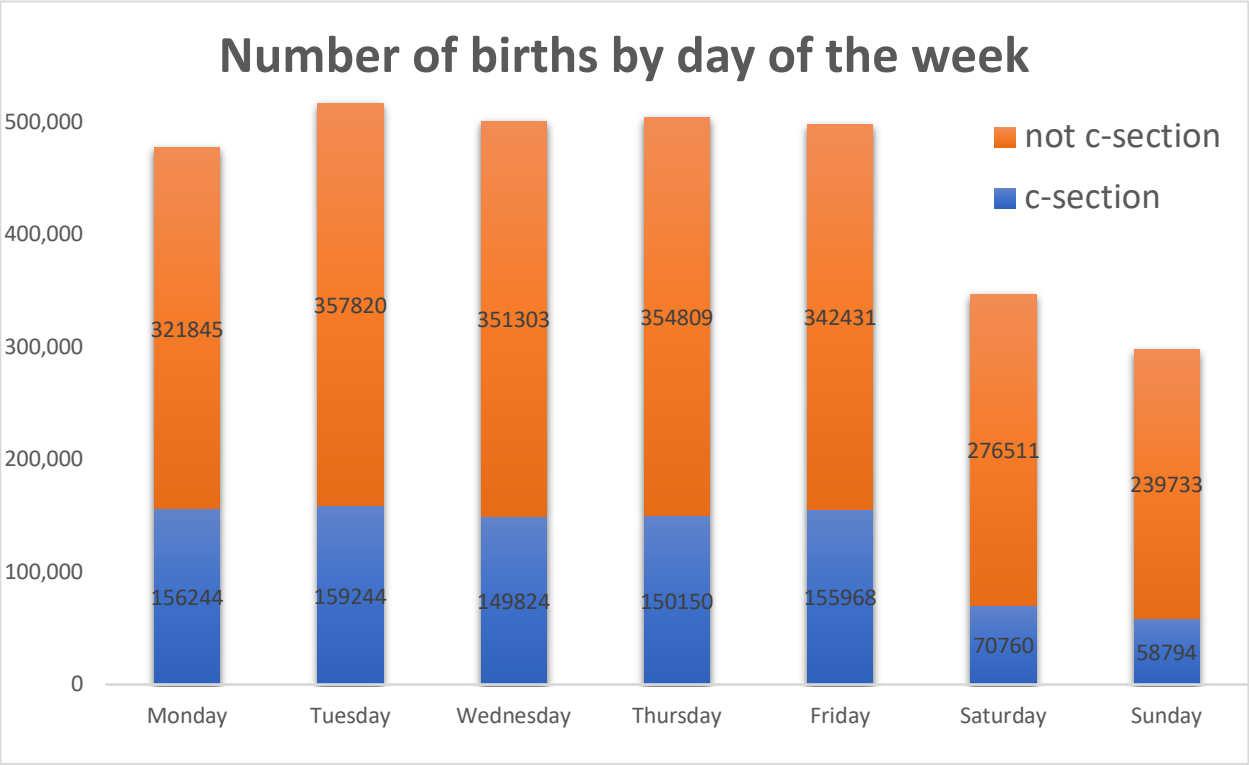
Given the format of data, the data cleaning process was extensive, as each column of data had to be identified and labeled. After an initial ingestion of the data, 57 features were identified with a total of 3,757,582 observations that represent 2019 births in the 50 U.S. states. The sample was then limited to include only hospital, single births that had no mortality or abnormal conditions associated with them (for either mother or newborn) for a total of 3,146,450 observations and 51 features.

It is important to note that the rich data set precludes geographic and specific income information in order to avoid client identification. Although income can be imputed with other variables such as payment provider and WIC status, geographical information is completely unavailable.

After features were created, null values also needed to be properly labeled since different variables have different null values such as 9, 99, 999 and U for unknow and X for irrelevant. After analyzing null variables and removing data with nulls, as well as limiting sample age to women ages 15 to 45, the final shape of the data is 2,874,472 with 47 features.

A quick exploration of the data immediately reveals that human preferences have a significant impact on delivery timing. A histogram of births by day and method (c-section versus not c-section) reveals an interesting trend: Saturday and Sunday have a significantly lower birth rate. This is especially true for c-sections, where birth rates decrease proportionally more on the weekend, with 7% of c-section occurring on Sunday, relative to 11% of non c-section births. There is no biological reason for this trend to occur – newborns do not choose the timing of their birthday by the day of the week. The relatively similar distribution of the number of births

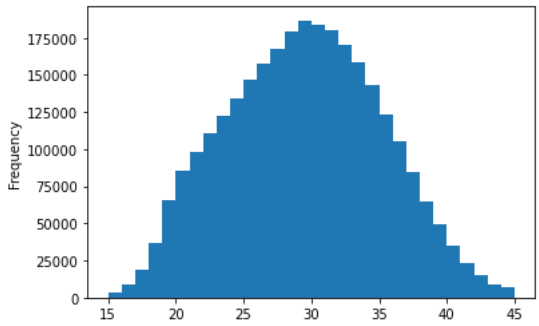
on weekdays further demonstrates that the low birth rates on the weekend is engineered by humans.



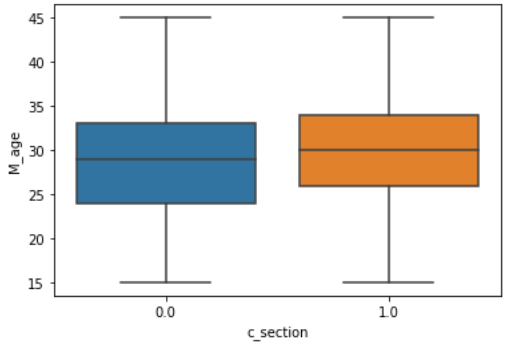
Exploratory Data Analysis

During the EDA phase, I examined the distribution of each feature, by delivery type. More specifically I looked at variables such as mother's age, baby's weight, number of prenatal visits, to explore whether these variables change with type of delivery. Median age of mothers who had a c-section is slightly higher than other birth outcomes. Median birth weight is very similar for both birth types, though variation of birth weight is higher for c-section babies. Similar trends can be observed with the number of prenatal visits.

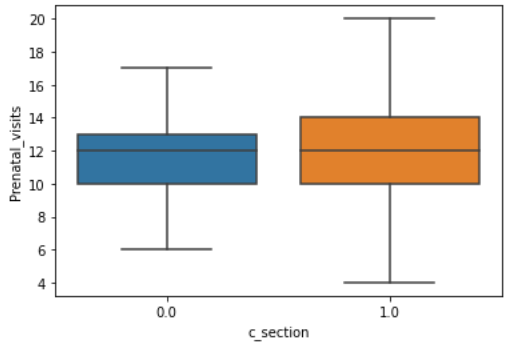
Distribution of mother's age



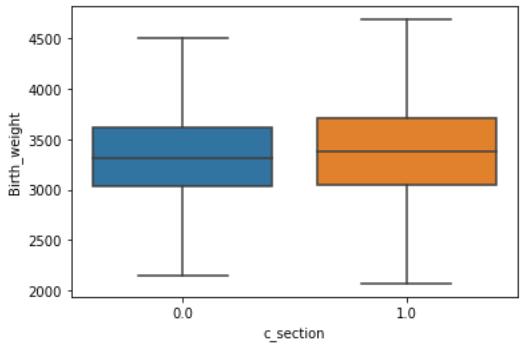
Boxplot of mother's age by c-section outcome



Boxplot of number of prenatal visits by c-section outcome



Boxplot of number of baby's birth weight by c-section outcome

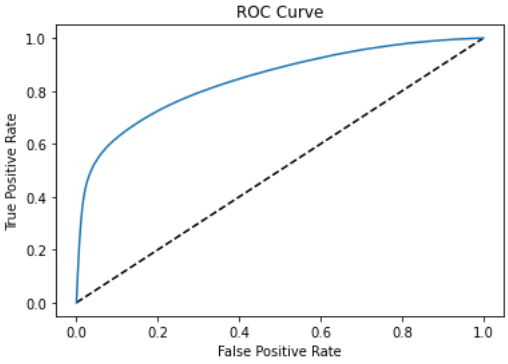


DATA PRE-PROCESSING

Categorical features were transformed to dummy variables. In an addition, the data was scaled using standard scaler, in order to make sure the data is consistent across types of data. The final number of features before modeling was 51.

MODELING

Given the large number of observations, and the nature of the research question, logistic regression was the first model explored for the classification. The dependent variable is the dummy variable c-section, which equals 1 is a birth ended in a c-section and 0 if it did not. Accuracy of the initial logistic regression model was 83% which is fairly high. Since the data is somewhat imbalanced, it is important to examine other classification metrics as well. Precision is 0.79 and recall is 0.56, with F-1 score of 0.65. The first model ROC curve:



Since classes of the dependent variable are somewhat imbalanced (28.4% of the sample had a c-section), a weighted logistic regression was also employed. While accuracy did not change much (0.79), precision declined to 0.61 but recall increased to 0.71.

The next step in model optimization was running a recursive feature extraction algorithm, to optimize the number of features. 28 features were found to be optimal, though the change in accuracy was small.

Since there is a large number of dummy variables and some of the categorical data can be highly correlated, a logistic regression can suffer from multicollinearity problems, skewing the results of the model. Therefore, a classification model that is based on a decision tree learner

may provide a more robust modeling approach, since multicollinearity among features does not pose a problem with decision trees.

As part of the optimization process, a few decision tree models were tested, all taking class imbalance into account. Table 1 provides a summary of most important metrics.

Table 1. Performance metrics across 5 models

	Accuracy	Precision	Recall	F-1	AUC
Logistic regression - baseline	0.83	0.79	0.56	0.65	0.84
Weighted logistic regression	0.79	0.61	0.71	0.66	0.84
Weighted Decision tree classifier	0.77	0.6	0.6	0.6	0.72
Weighted Random forest classifier	0.85	0.84	0.56	0.68	0.86
Weighted XGBoost classifier	0.84	0.92	0.5	0.64	0.88

The weighted XGBoost classifier results in best performance. Precision is the highest at 92%, which means the number of false positive (classified as c-section when it is not a c-section) identified by the model is very low. Recall is the lowest for this model, which provides a clear example of the precision-recall trade-off. However, for the purpose of this model, having higher precision is more desirable because it means that the model very accurately identifies the true positive cases. AUC is also highest for this model.

Table 2 provides the top 10 most important features for the weighted XGBoost model. One interesting observation is that the weekend dummy variable is in 9th position

(out of 51 features). The weekend dummy has more explanatory power than mother's height, pre-labor weight or baby's weight and all of the mother's socio-economic variables.

Table 2. Top 10 features, by feature importance

Feature	Feature importance
No_risk_factors	0.23
Attendant_type	0.12
Fetal_pres	0.11
Augment_N	0.11
Induction_N	0.08
Prev_premie_N	0.06
Gest_diabetese_N	0.04
Gest_hypertension_N	0.04
Weekend	0.03
Antibiotics_N	0.03

SUMMARY

Using detailed birth data, I was able to train a classification model that correctly estimated whether a birth was a c-section with 92% precision, using a XGBoost tree-based learner. The model confirmed that practitioner preferences can be a significant factor in determining whether a birth ends up in a c-section, since c-sections happen significantly less on the weekend. If the number of c-sections preformed on the weekend provides a bench mark to represent necessary c-sections, this means that parents with births during weekdays should be particularly aware of the circumstance of their birth, if they want to avoid a medically unnecessary c-section.