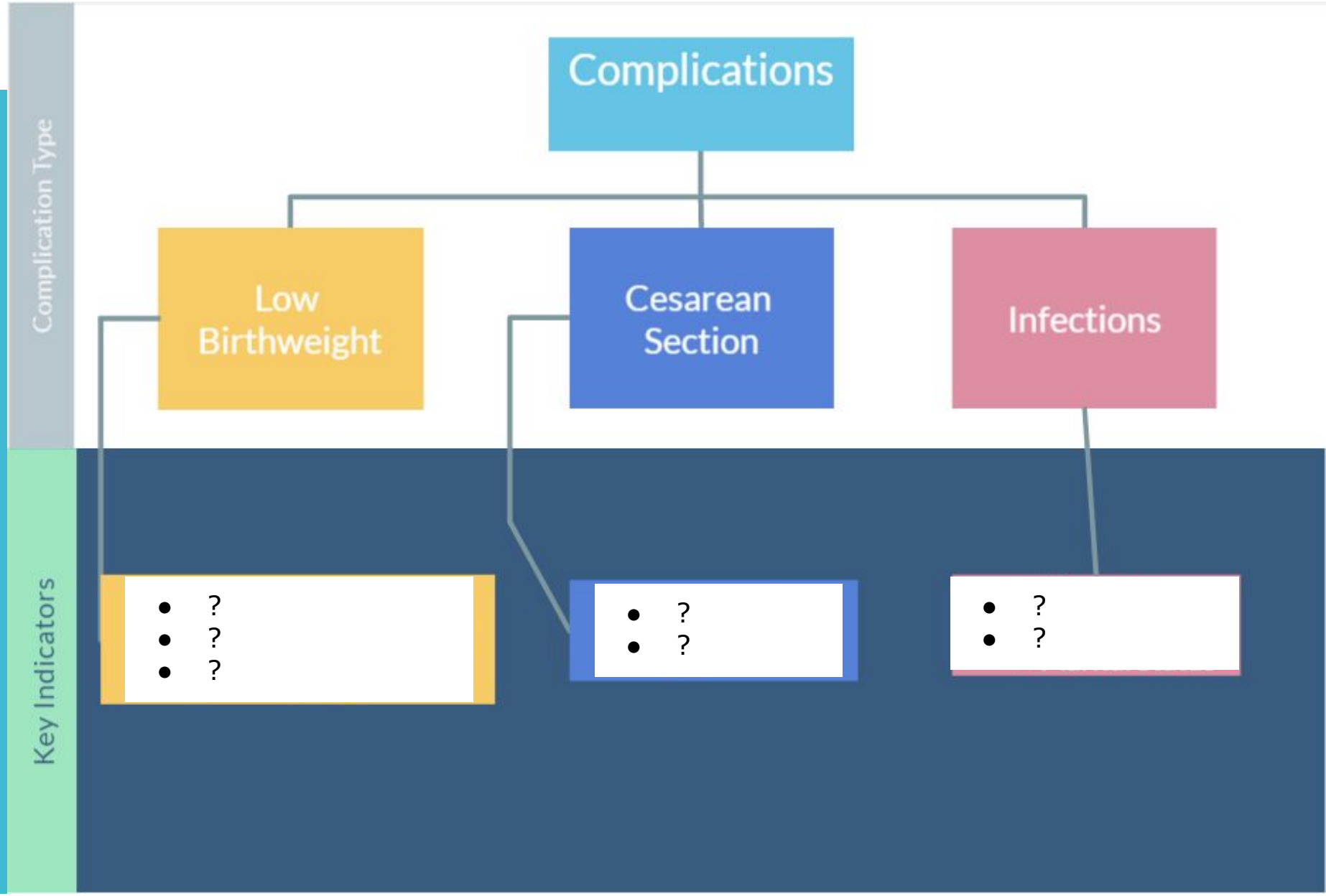
Predicting complications at Birth

Dataset: **US births (2018) - Kaggle**

Problem: Birth Complications



Introduction : Low birthweight as a Proxy for Complications

- A baby with low birthweight may be at increased risk for complications.
- The baby's tiny body is not as strong and he or she may have a harder time eating, gaining weight, and fighting infection.
- Because they have so little body fat, low birthweight babies often have difficulty staying warm in normal temperatures.
- A birthweight less than 2,500 grams (5 pounds, 8 ounces) is diagnosed as low birthweight.
- It becomes important for the parents and doctors as well to know whether the child will be underweight at birth or not as it gives a direct indication of the baby’s health.
- Hence our goal is to predict given a set of predictors in the dataset, whether the child will be underweight at birth or not.

EDA

Dataset: **US births (2018) - Kaggle**

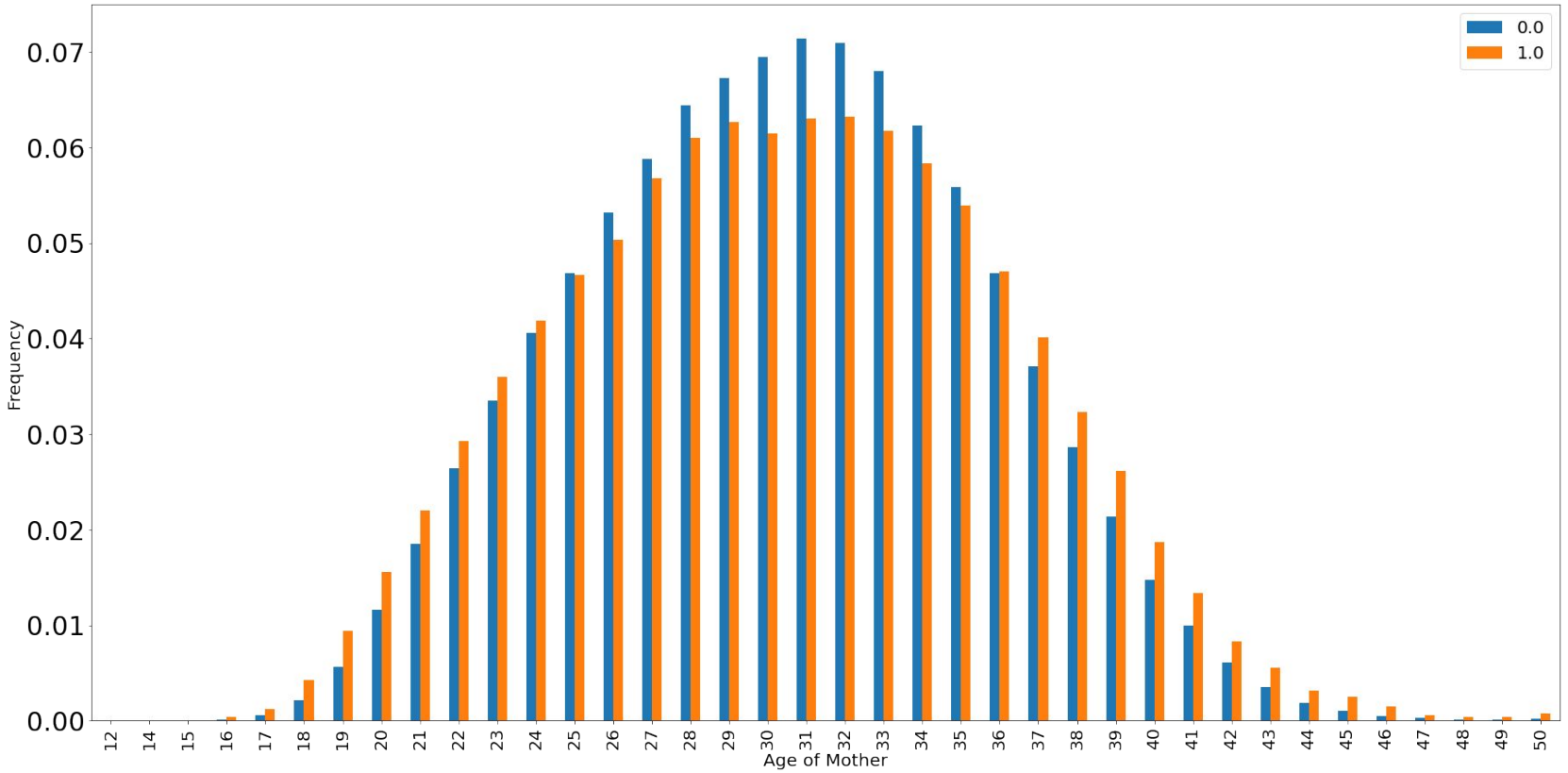
EDA

- Missing or unreported values in the dataset stored as 9, 99, 99.9 etc., replaced such values with NaNs.
- Cleaned the dataset of the missing values and kept relevant columns. Dropped columns such as payment method, place of birth etc.
- Correlational matrix shows very low correlation between predictors and response variable, hence further analysis needed with respect to the target variable.
- Approx. 92% of the dataset is normal weight at birth (negative class in binary classification) → Baseline accuracy for this dataset of 92%

EDA continued

Plot of mother's age for both positive and negative class:

According to our dataset, for mothers under 24 years old, there is a higher likelihood for delivering an underweight child than for the age group 25-34. This pattern can also be observed for mothers 35 and older ("Advanced Maternal Age").

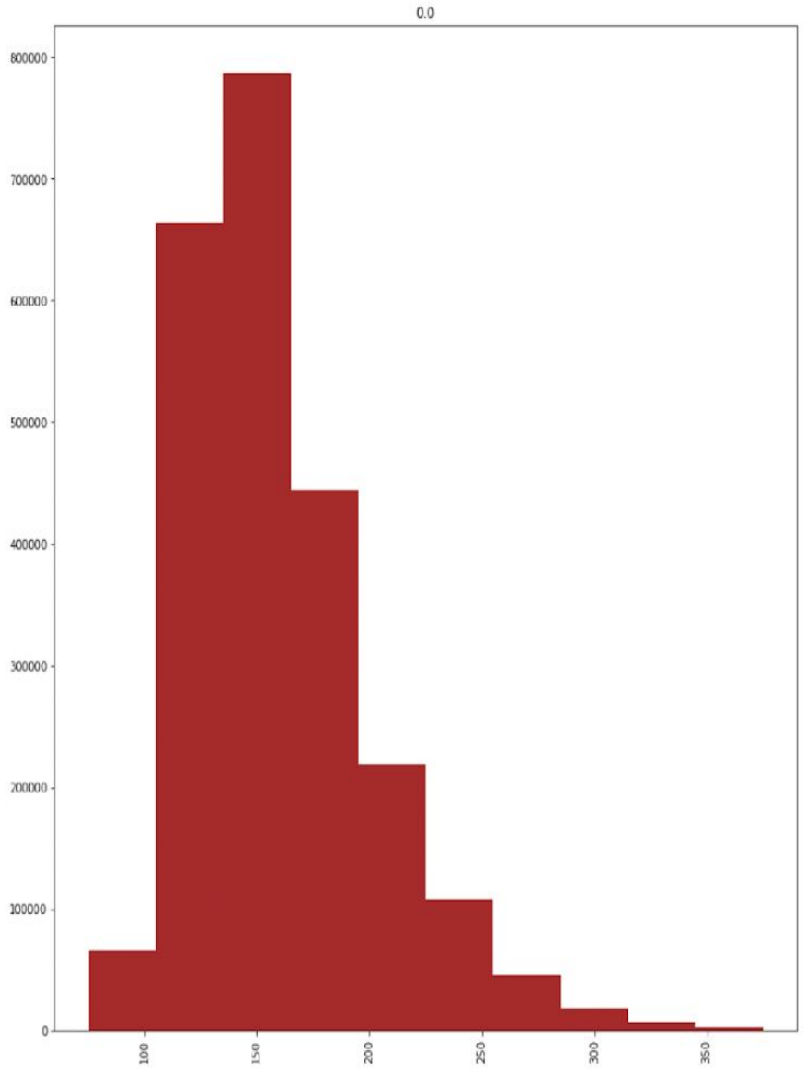


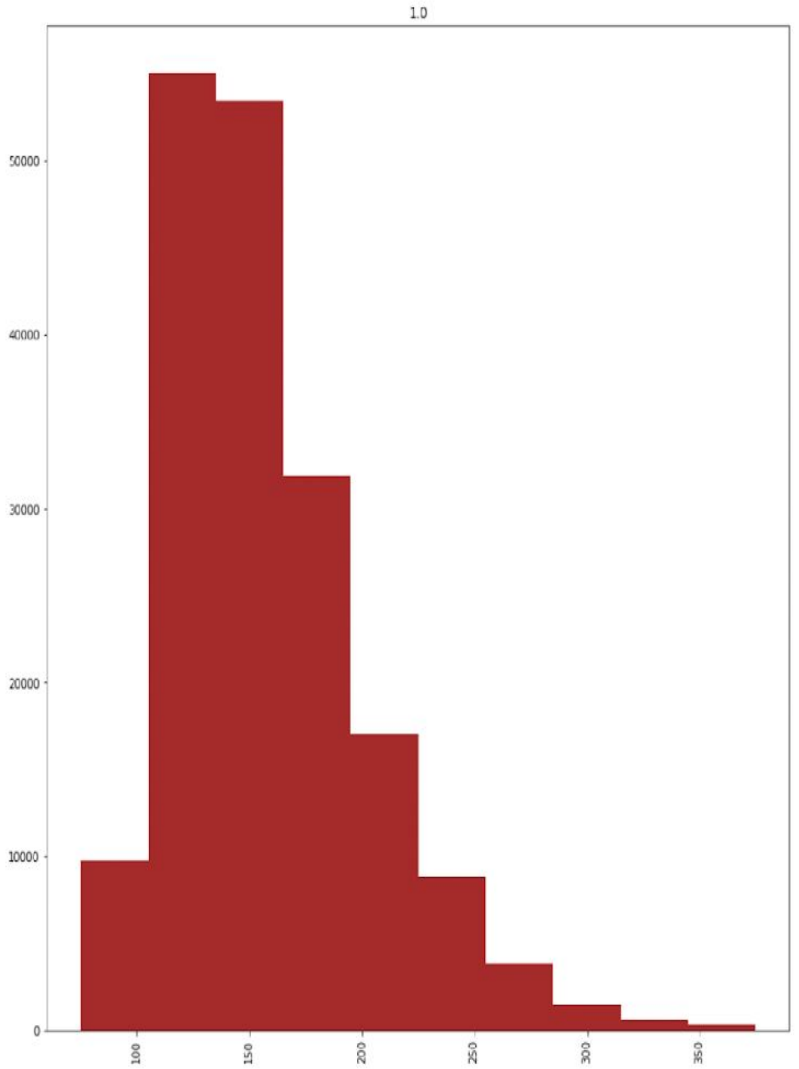
EDA continued

Plot of mother's pre-pregnancy weights for both classes.

On average, the pre-pregnancy weight reported by the mothers in the negative class is less than that of the mothers in the positive class.

The graphs confirm the same as there are more women below 100 in the positive class than in the negative class, suggesting pre-pregnancy weight of the mother is related to the birth weight of the baby.





EDA continued

On average the number of pre-natal visits for the negative class in more than that for the positive class.

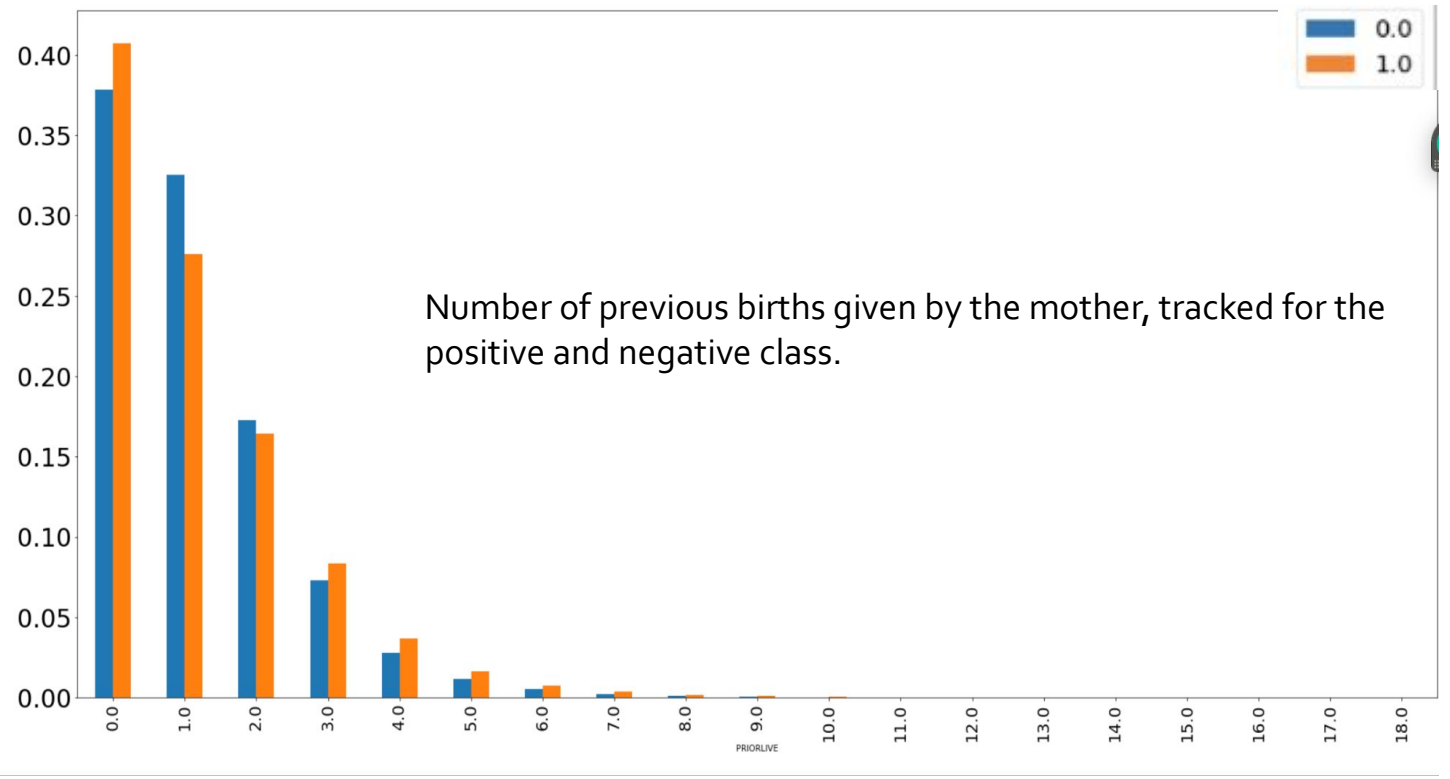
The number of births previously given by the mother shows a weak relationship to the birth weight of the child.

On average, there's a higher ratio for the first birth to be underweight and it's comparatively less likely to deliver an underweight child after 1 or 2 previous births. However, the likeliness of the positive class increases after 3 previous births.

target	PREVIS
0.0	11.432553
1.0	9.680133

Average negative class prenatality visits to the doctor

Average positive class pre natality visits to the doctor



Number of previous births given by the mother, tracked for the positive and negative class.

EDA continued

Feature engineering:

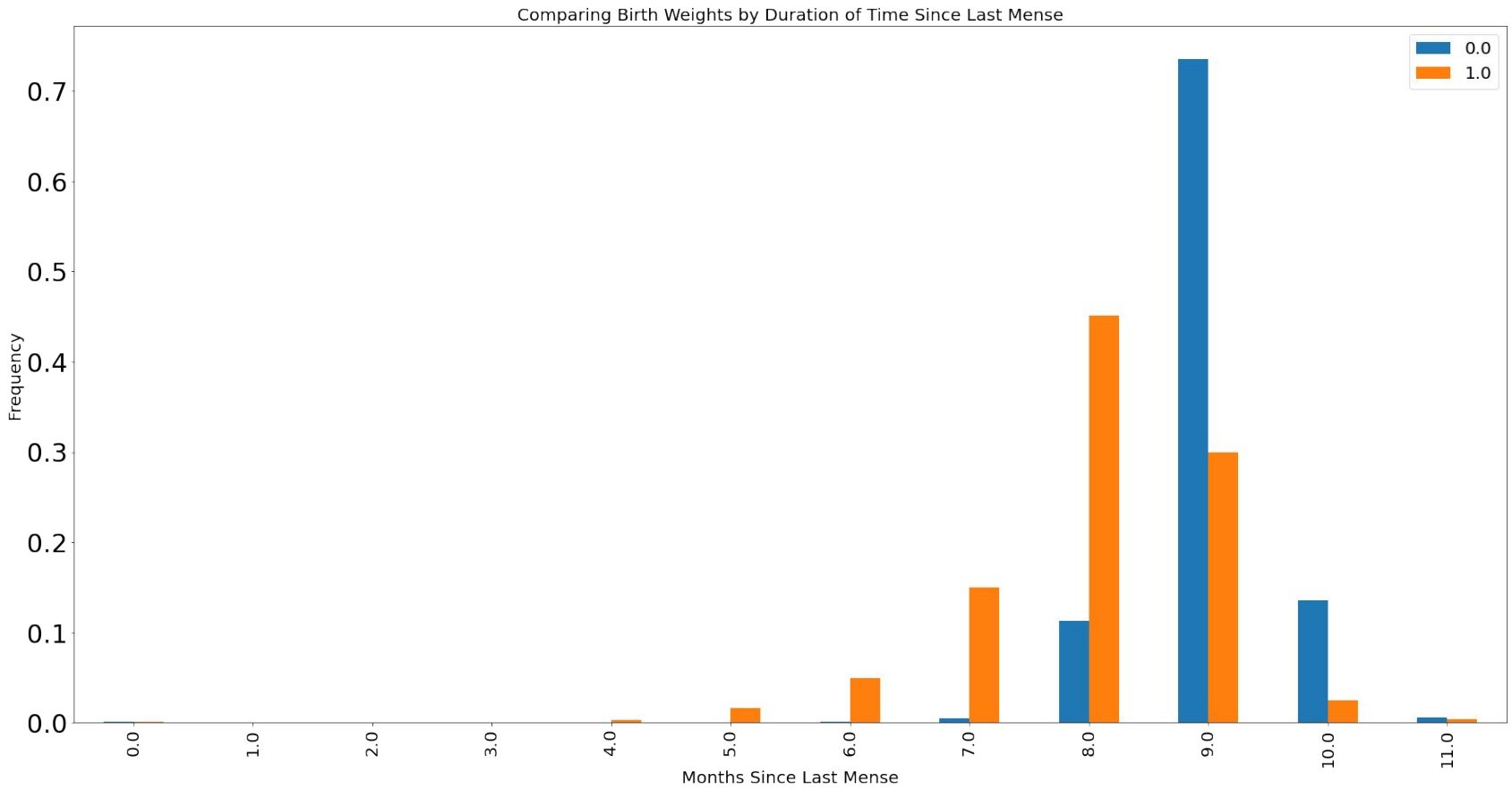
Created No menses duration (NMD)

NMD = Birth-Month - Last normal Menses Month

Expected value of this feature: 9 months.

Compared this feature for both classes:

For the mothers who delivered underweight babies, i.e., in the positive class, the instances of NMD value being 6, 7, 8 is more than the negative class, showing that women whose normal menstrual cycles stops 6,7, or 8 months before the date of birth have a higher possibility to deliver an underweight baby.



Birth Weight Modeling

Dataset: **US births (2018) - Kaggle**

Models - Low Birthweight

1. Logistic Regression
2. Decision Tree
3. Bagging
4. Random forests.
5. Gradient boosting

The accuracy with logistic regression (93.27%) is better than the baseline(92%) but the recall is only 6%.

To improve the recall, we upsampled the positive class as they are comparatively lesser in count.

Logistic regression accuracy on the test data with upsampling is: 80%, and recall improved significantly to 70%.
However accuracy takes a hit.

Tested other models without the upsampling as it is affecting the accuracy.

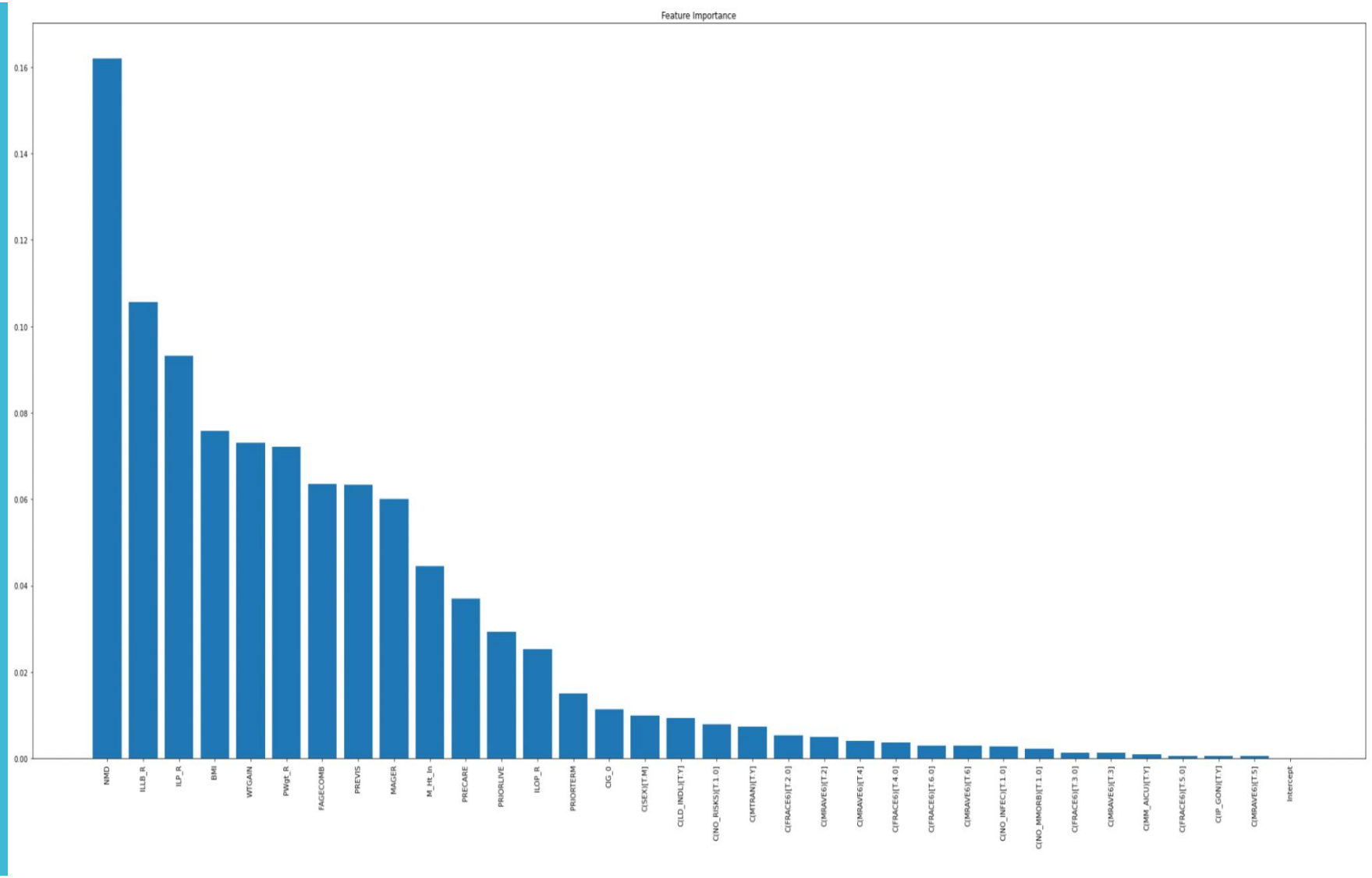
Summary of other models accuracies and recall:

	Training Accuracy	Test Accuracy	Train Recall	Test Recall
Decision Tree	1.000000	0.897016	1.000000	0.647006
Bagging	0.996931	0.933126	0.996931	0.632044
Random Forest	0.999461	0.931447	0.999461	0.626168
Gradient Boosting	0.986430	0.935591	0.986430	0.660284

Feature Importance

NMD turns out to be the best indicator to predict the birth weight

Other important factors include Mother's age, BMI, pre-pregnancy weight



Insights and findings

Mother's current physical factors affecting the birth weight of the child:

1. Mothers height in inches
2. BMI
3. Prepregnancy Weight
4. Weight gained during pregnancy.

Mother's medical related factors affecting birth weight of the child:

1. NMD - most important factor based on our data
2. Interval since last pregnancy/last birth/other pregnancy.
3. Mother’s age.

Other miscellaneous factors:

1. Father’s age
2. Number of prenatal visits
3. When did the prenatal care begin
4. Prior births which are now living
5. Number of Cigarettes smoked during pregnancy.

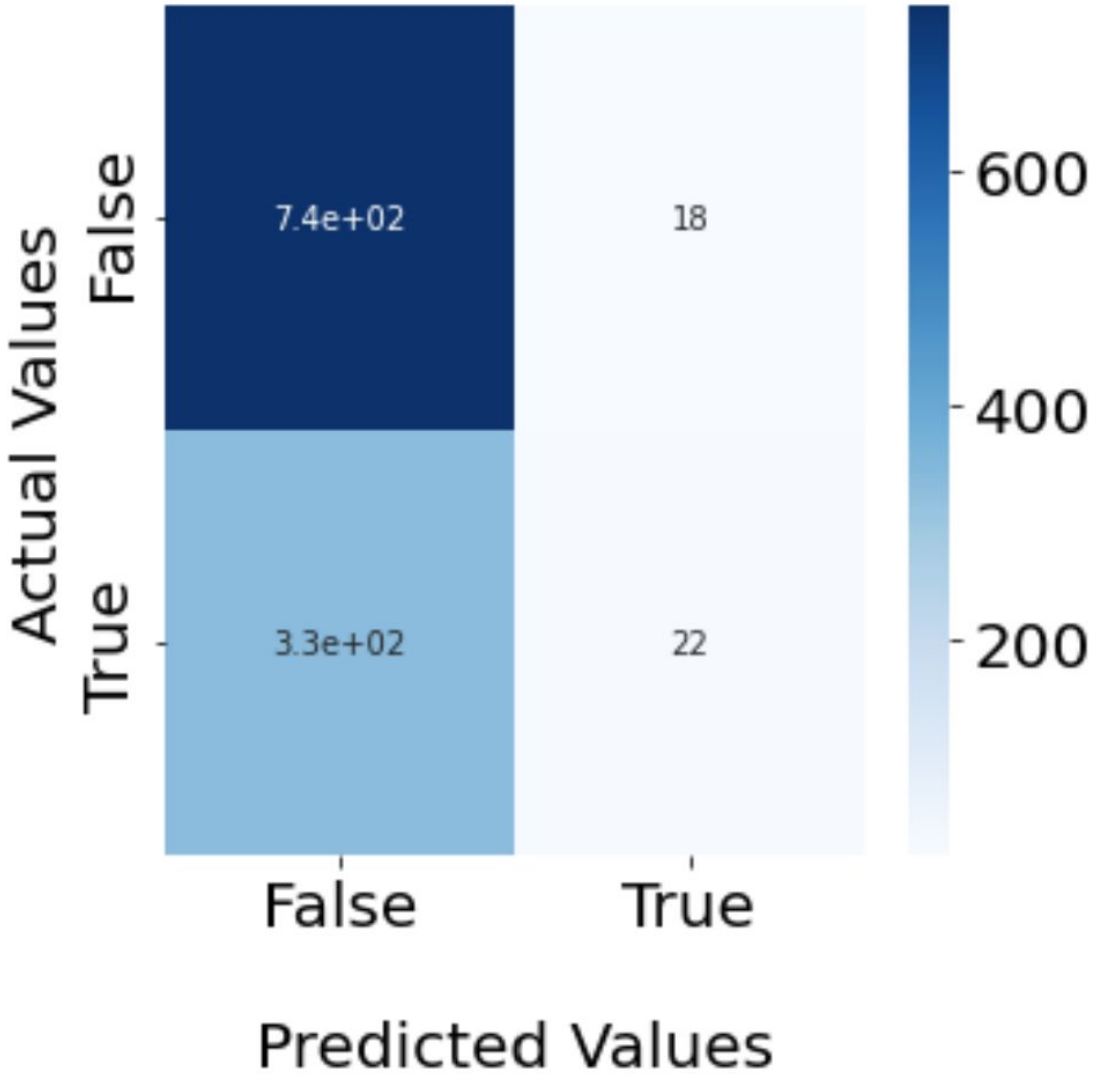
All other factors play comparatively lesser important role in determining the weight of the child.

Caesarean Section Modeling

Dataset: **US births (2018) - Kaggle**

Models - Caesarean Section

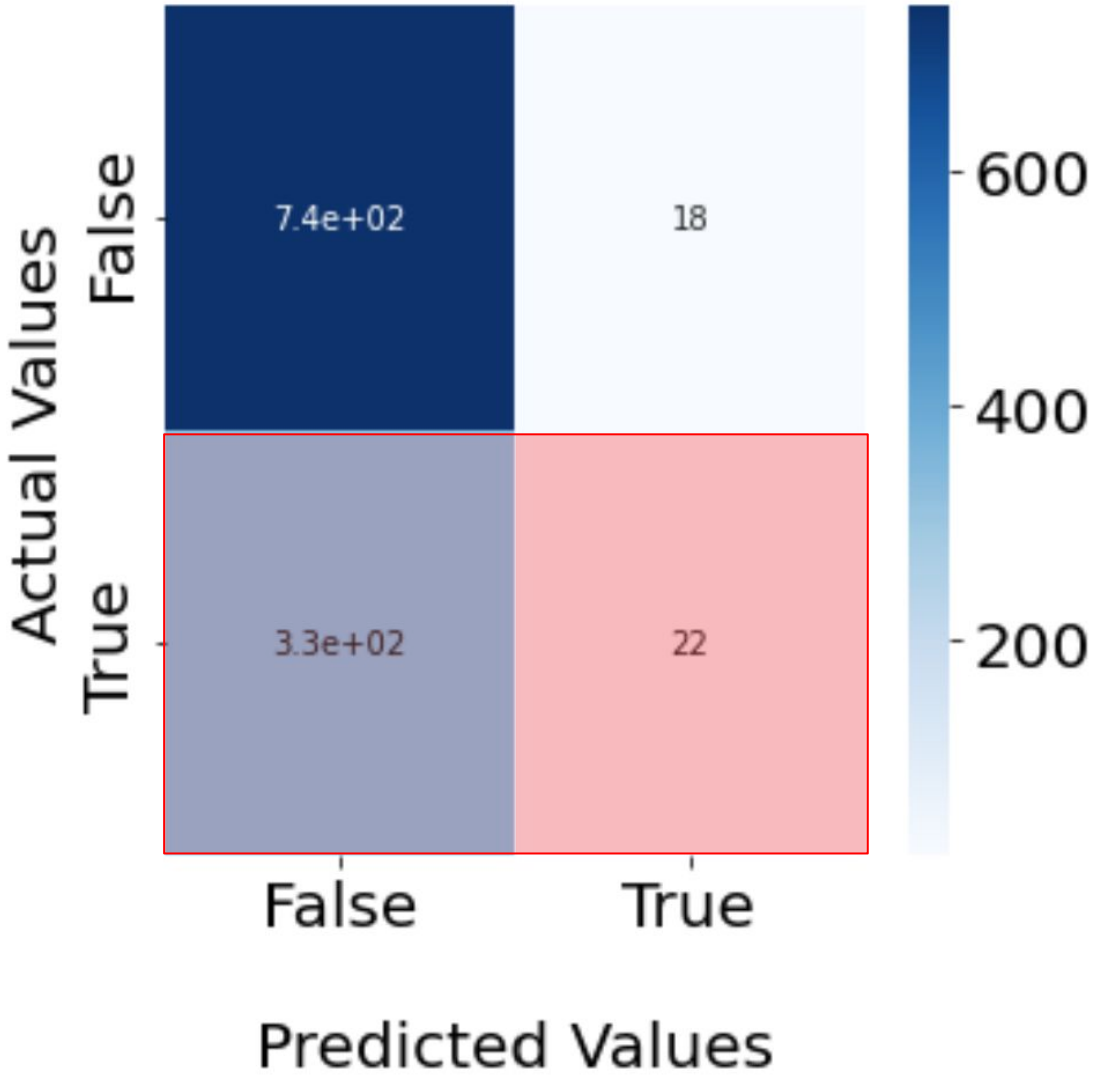
Univariate Confusion Matrix (1000s)



Models - Caesarean Section

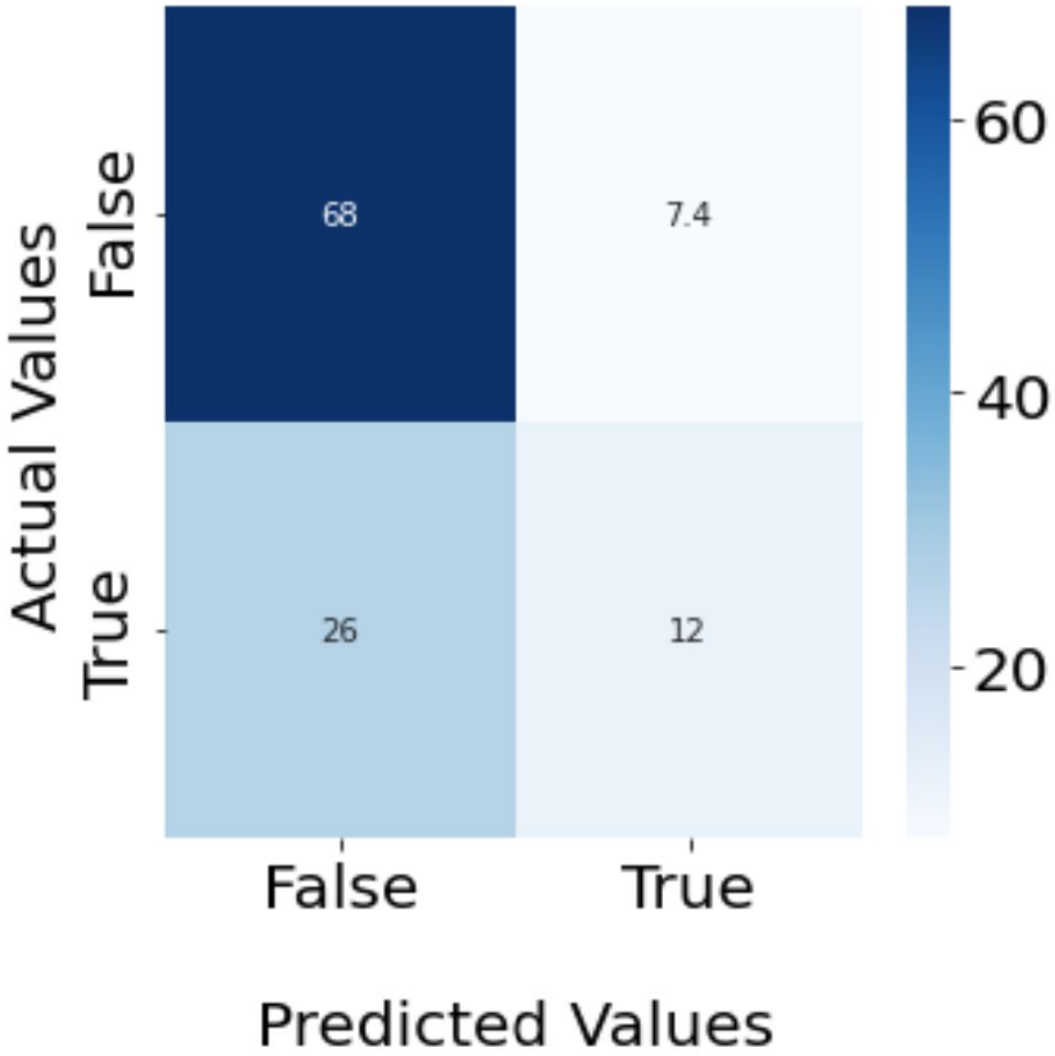
Precision	**55%**
Recall	**6%**

Univariate Confusion Matrix (1000s)



Models - Caesarean Section

Multivariate Confusion Matrix (1000s)*

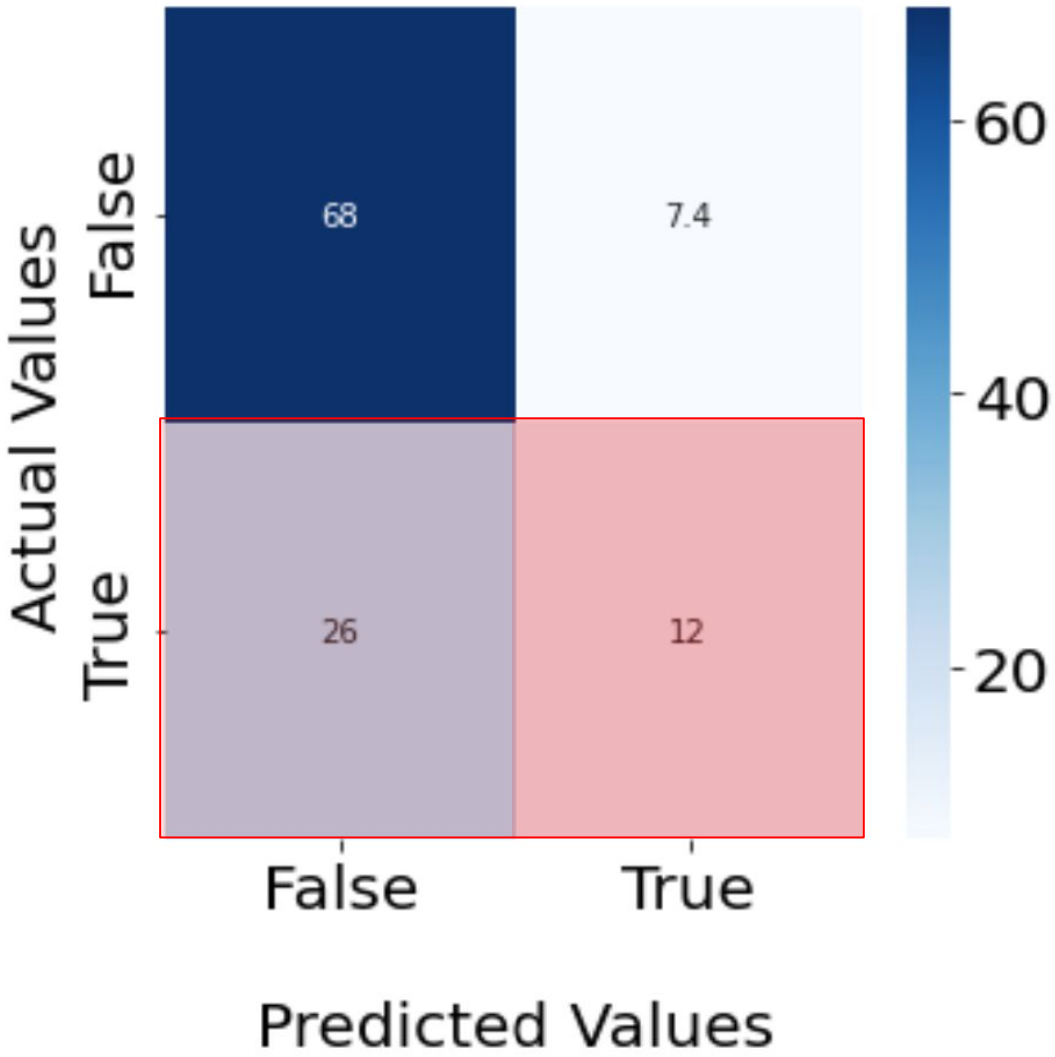


*Values are smaller as multivariate removes more missing records with missing values.

Models - Caesarean Section

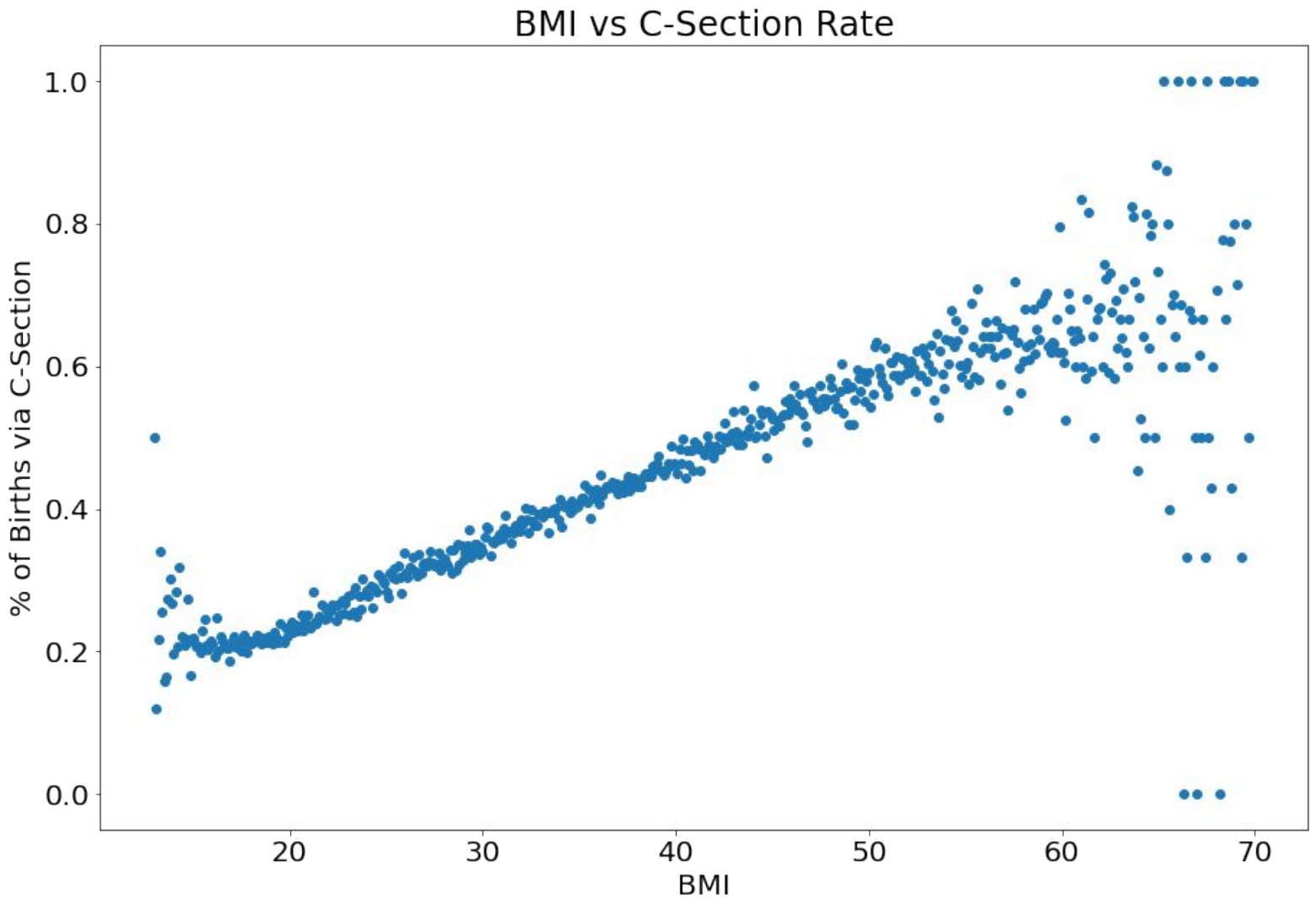
Precision	**62%**
Recall	**31%**

Multivariate Confusion Matrix (1000s)*



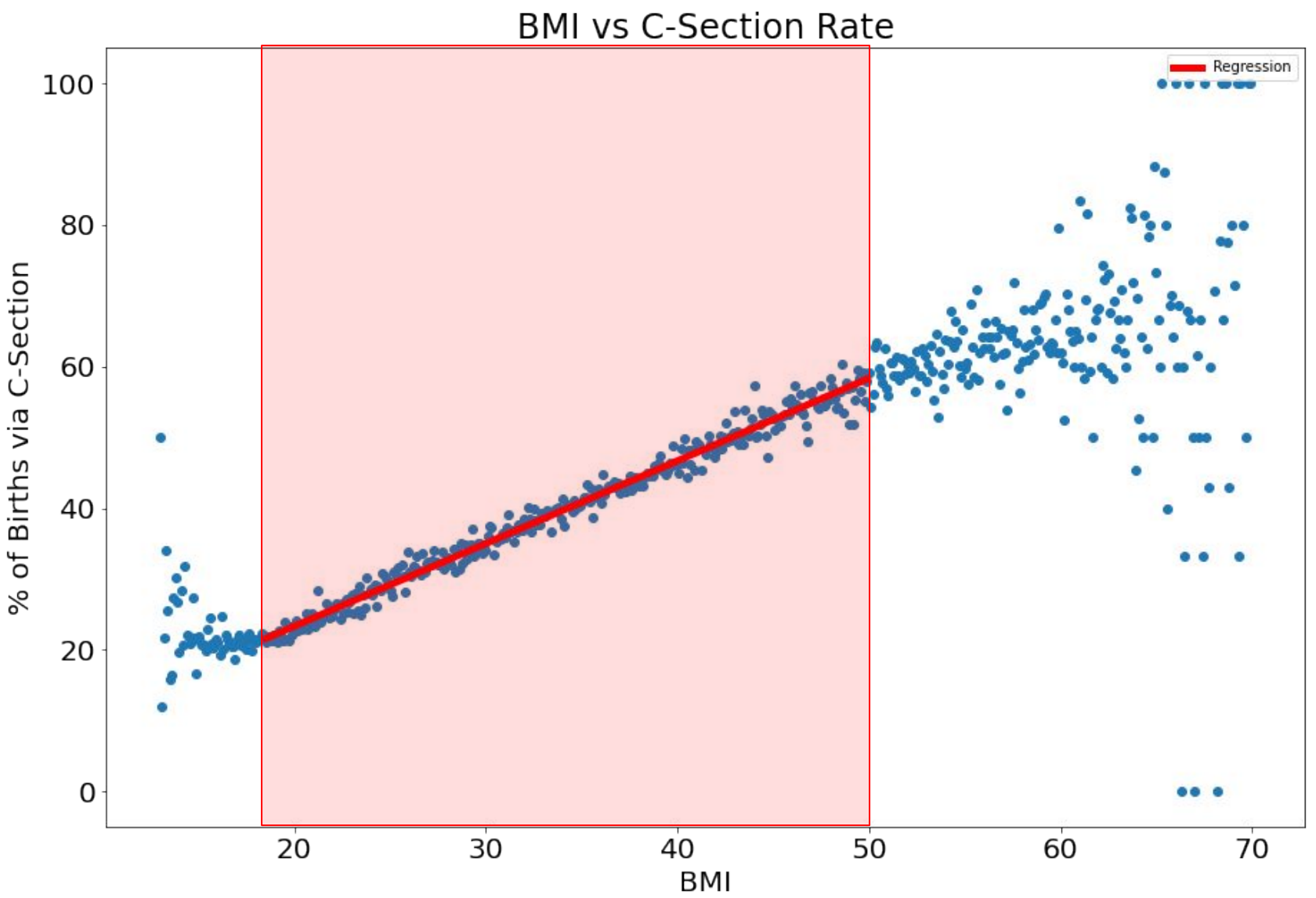
*Values are smaller as multivariate removes more missing records with missing values.

Models - Caesarean Section



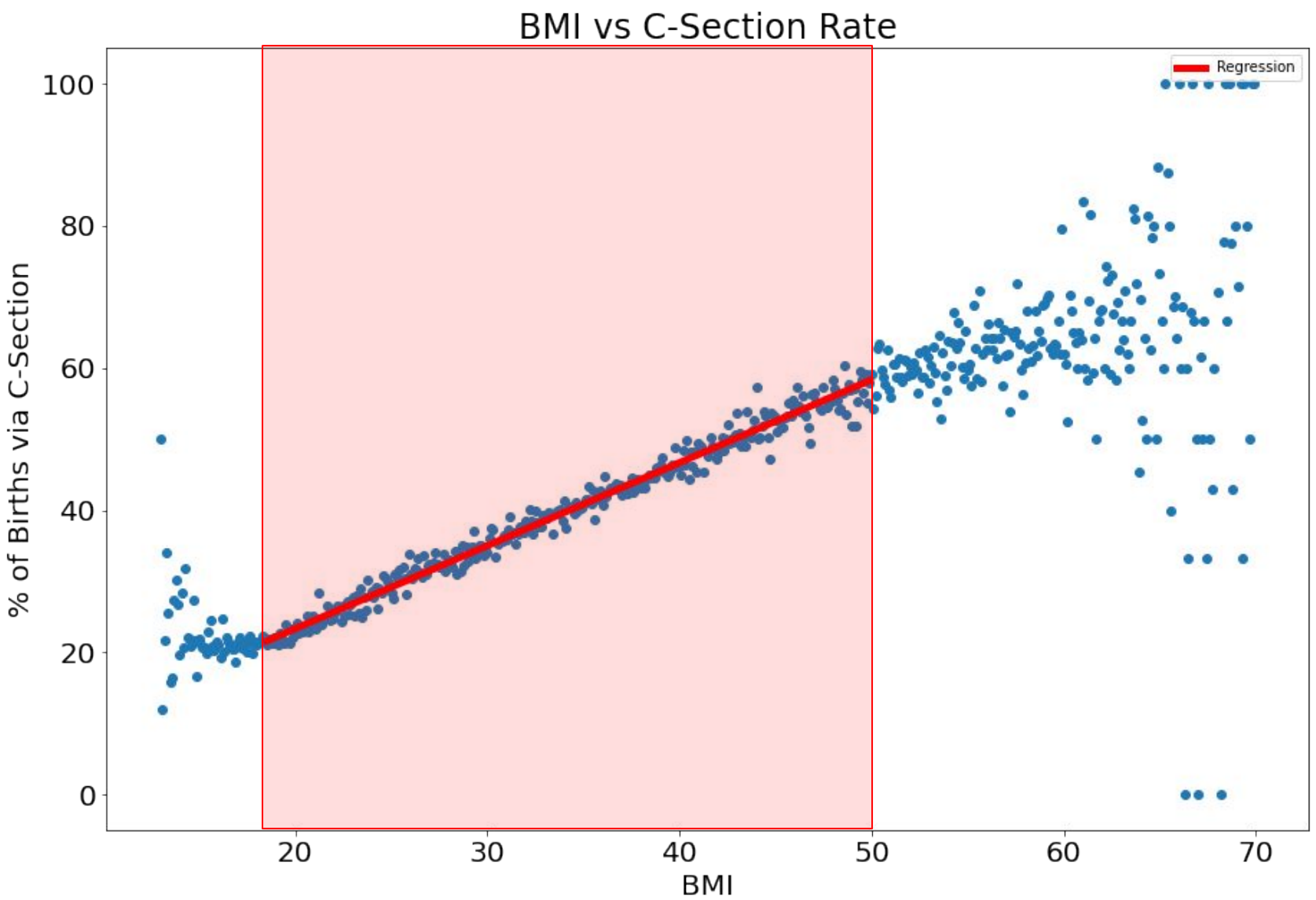
Models - Caesarean Section

BMI Range	Category
< 18.5	Underweight
18.5 - 24.9	Normal Weight
25 - 29.9	Overweight
30 - 39.9	Obese
> 40	Morbidly Obese



Models - Caesarean Section

Adj. R-Squared	**99.9%**
Coef	**0.017**



Infection Modeling

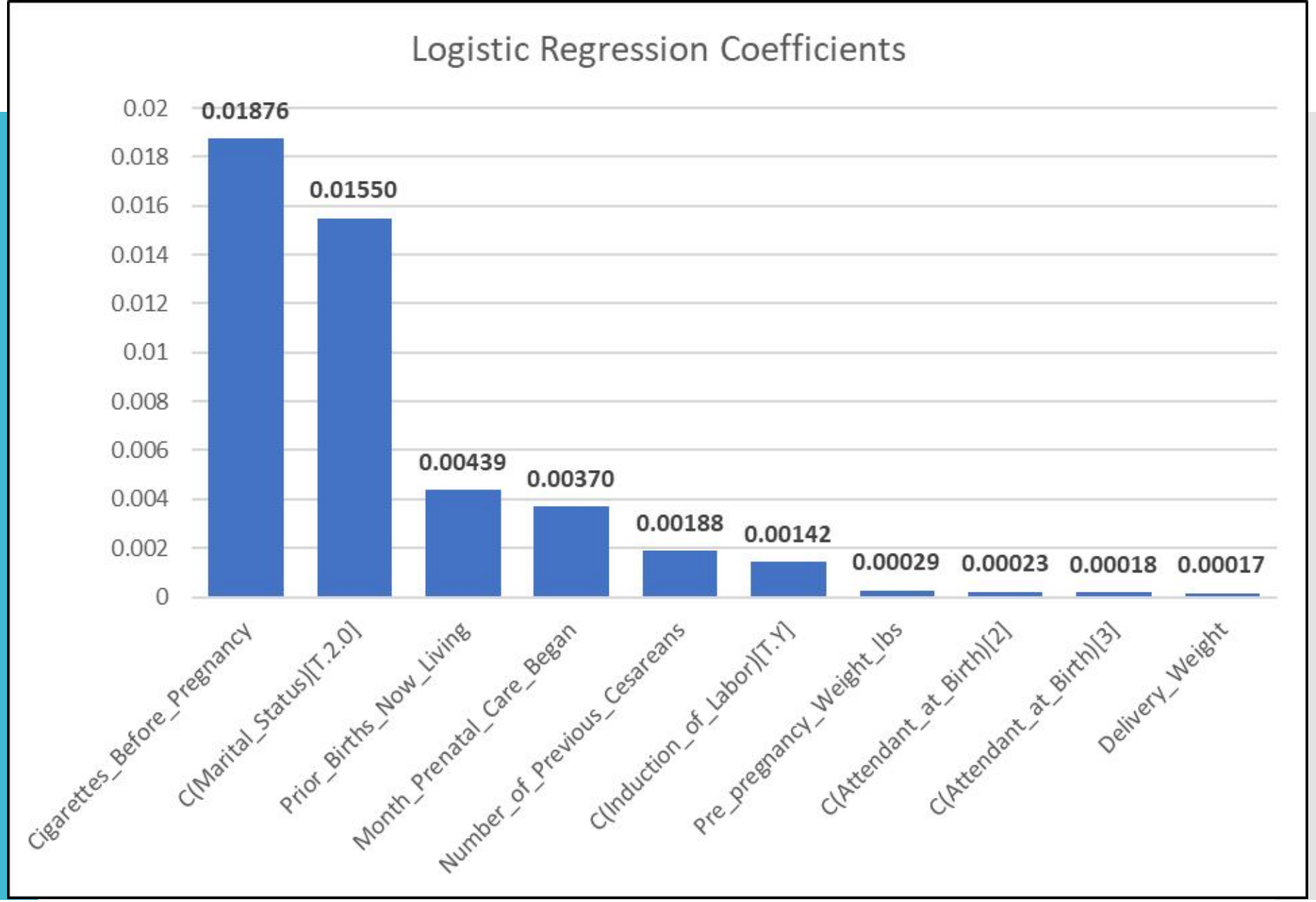
Dataset: **US births (2018) - Kaggle**

Infections - Models

Counts		
Births	No Infection	Infection
3,801,534	3,696,802	104,732

Accuracy	Infection	No Infection
Baseline	2.76%	97.24%
Decision Tree	2.99%	97.02%
Logistic Regression	2.98%	97.02%

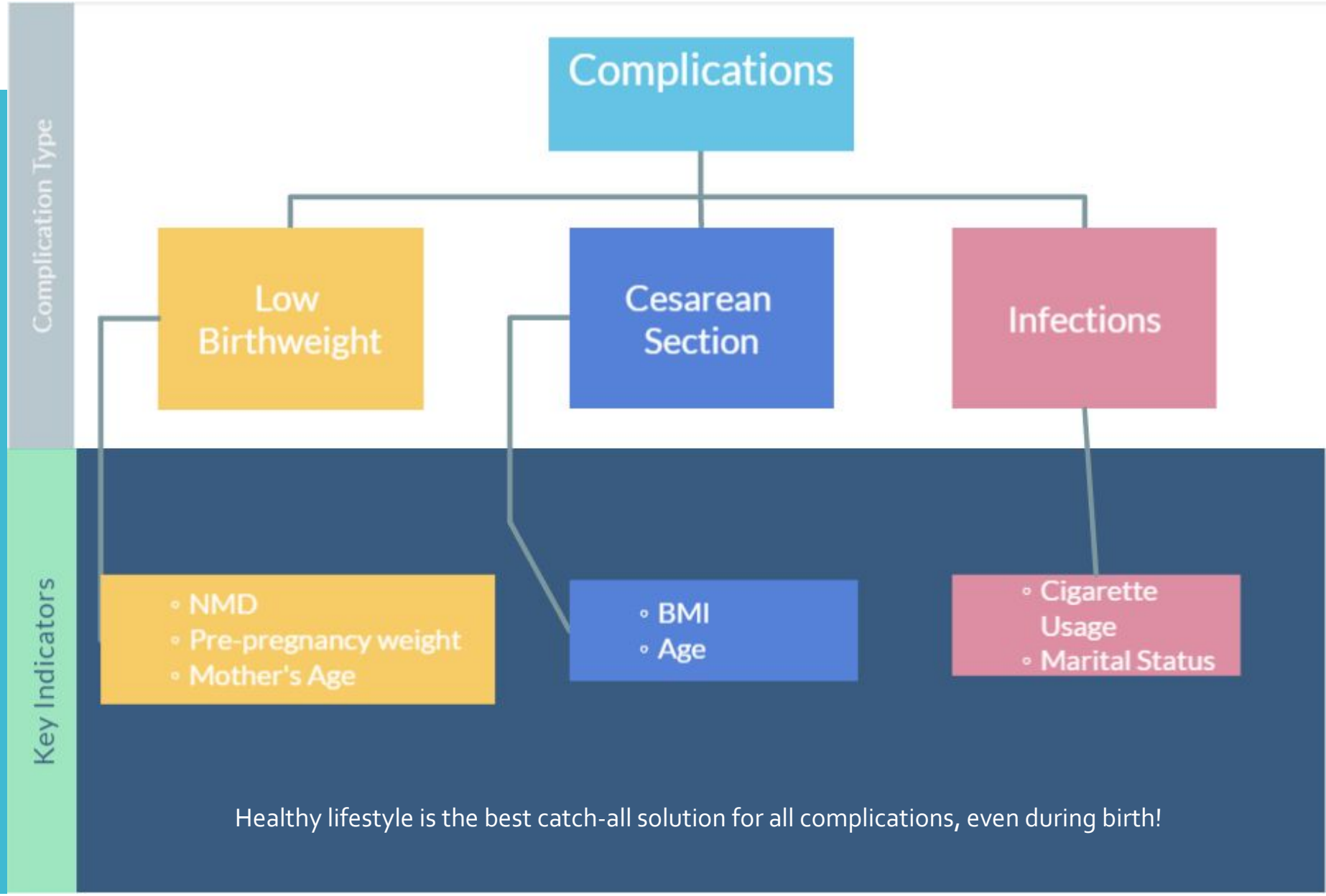
Infections - Variable Importance



Conclusions

Dataset: **US births (2018) - Kaggle**

Conclusions



Thank You

Questions?