ANITA B.ORG 2024

GRACE HOPPER CELEBRATION

me +we

Data Science Storytelling:

Using ML to Predict a Baby's Birth Weight

Anubha Bhargava

Data Scientist at Virtualitics

ANITA B.ORG 2024
GRACE HOPPER CELEBRATION

Agenda

- About Me
- Problem Statement
- Exploratory Data Analysis
- Feature Engineering
- Model Selection
- Model Evaluation
- Case Studies
- Conclusion

ANITA B.ORG 2024 GRACE HOPPER CELEBRATION

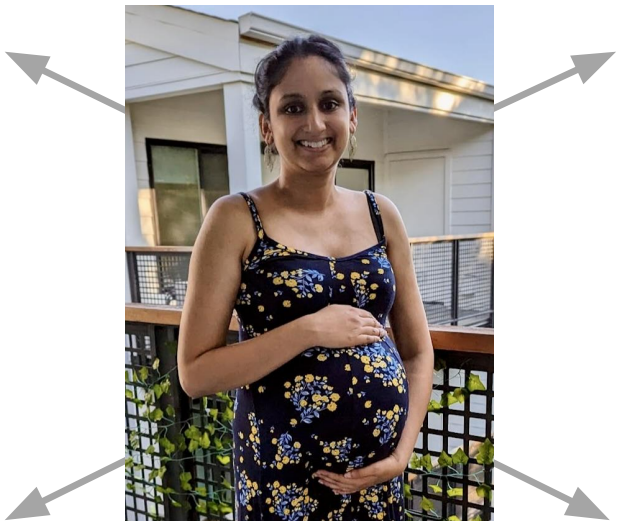
A Bit About Me

Personal:

- I am married and have a 1 year old son Ridhaan.
- I grew up in Boston and currently reside in New York.

Education:

- Bachelor's Degree in Electrical Engineering from Rensselaer Polytechnic Institute (RPI)
- Master's Degree in Electrical Engineering, Concentration: Data Science from Columbia University



Anubha Bhargava

Career:

- Currently, I work as a data scientist at Virtualitics.
- I have ~7 years of work experience as a data scientist, building models for a variety of industries, including health and wellness, marketing and defense.

Hobbies:

- Dance, Yoga, Crafting, Hiking, Traveling, Skiing

Problem Statement

- **Objective:** Building a model to predict a newborn's birth weight.
- **Why?** Complications can result if your baby is not a normal weight.

- **Underweight** (less than 5 lbs, 8 ounces at birth):
  - Usually caused by premature birth (before 37 weeks)
  - Short term / long term complications can result, like low oxygen levels at birth, breathing problems, nervous system problems, digestive problems and developmental delays.

- **Overweight** (over 8 lbs, 13 ounces at birth):
  - For the mother, increases the need of a C-section, type 2 diabetes, heart disease, asthma and obesity.

Pampers, https://www.pampers.com/en-us/baby/newborn/article/average-baby-weight

Data

Open Source Dataset from Kaggle - **Features Used** (26):

Information about Mother:

- Mother's Birthplace (US, Outside of US)
- Mother's Height
- Pre-pregnancy weight
- Married (Y/N)
- Body Mass Index (BMI)
- Mother's Age
- Mother's Education
- Time since last pregnancy
- Time since last birth
- Prior Number of Births
- Prior Number of Terminations

Information about Father:

- Father's Age
- Father's Education

Factors During Pregnancy:

- Risk Factors Reported
- Smoked (Y/N)
- Weight Gain
- Month Prenatal Care Began
- # of Prenatal Visits
- Length of Pregnancy (Derived)
- Length of Prenatal Care (Derived)

Information about Newborn:

- Sex (M/F)

Factors During Delivery:

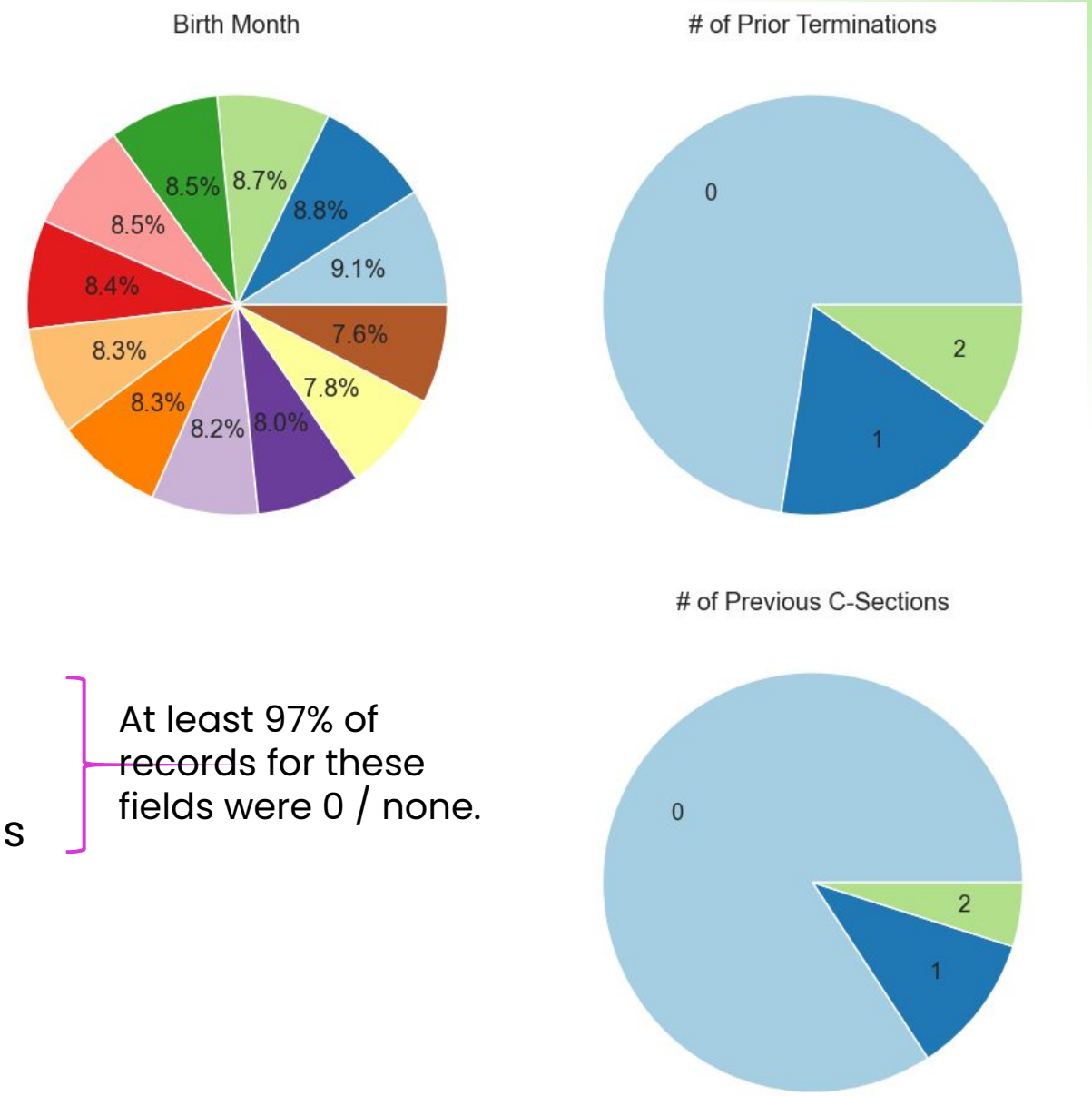
- Delivery Method (Normal, C-Section)
- Induction of Labor
- Payment Type (Medicaid, Private Insurance)
- Birthplace (Hospital, Home)
- Attendant at Birth (Doctor, Midwife)

Data

Features Not Used:

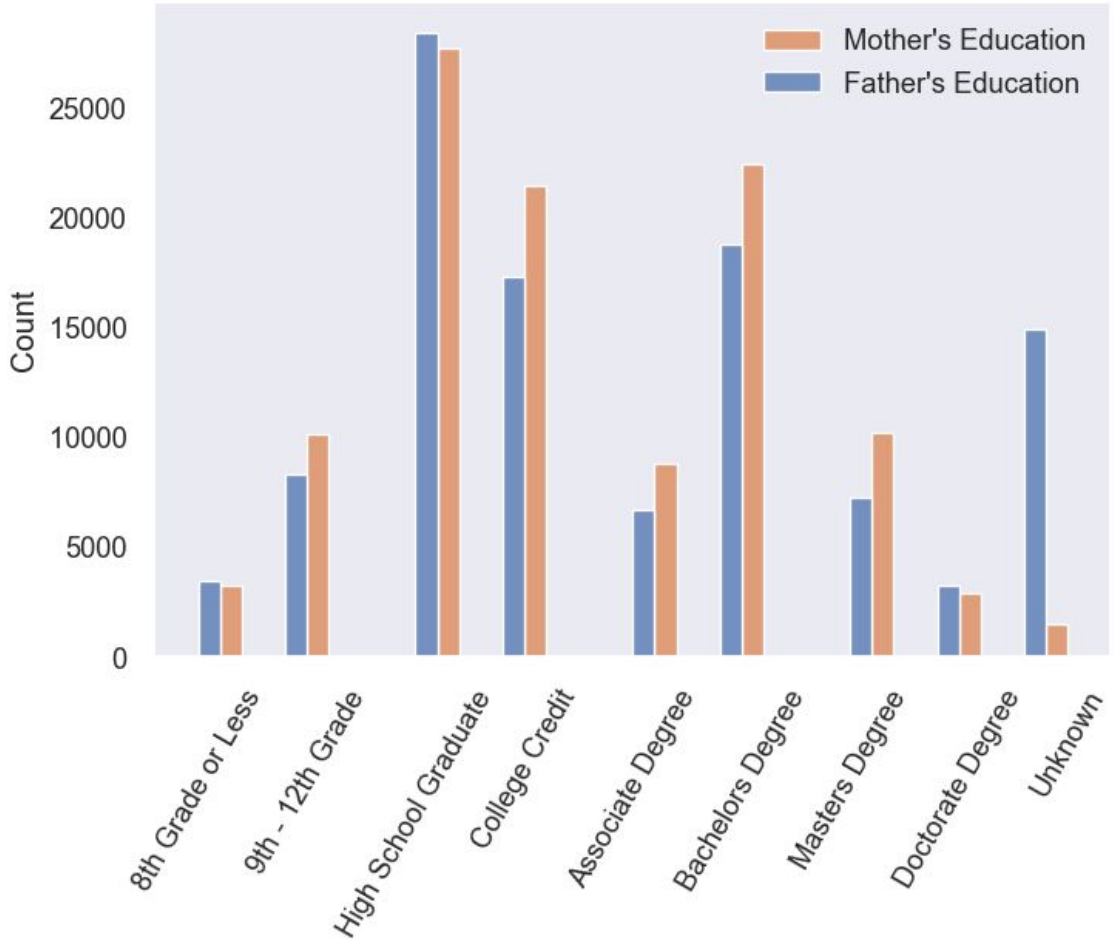
- Last Period Month
- Birth Month
- Time of Birth (Hour/Minutes)
- Day of Birth
- # Previous C-Sections
- # of Infections Reported
- Maternal Morbidity
- # Children Dead from Previous Live Births

At least 97% of records for these fields were 0 / none.

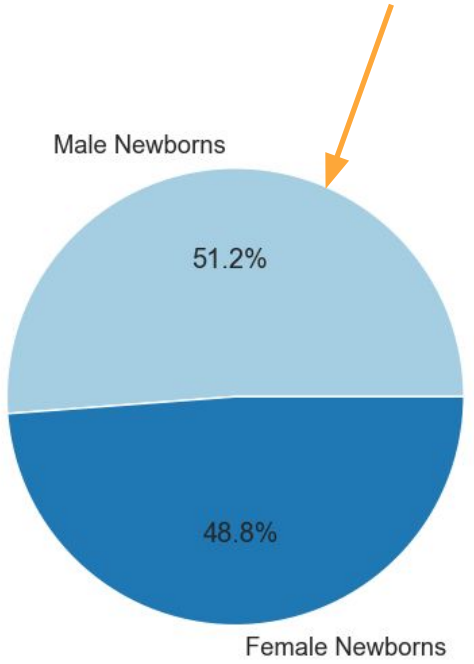


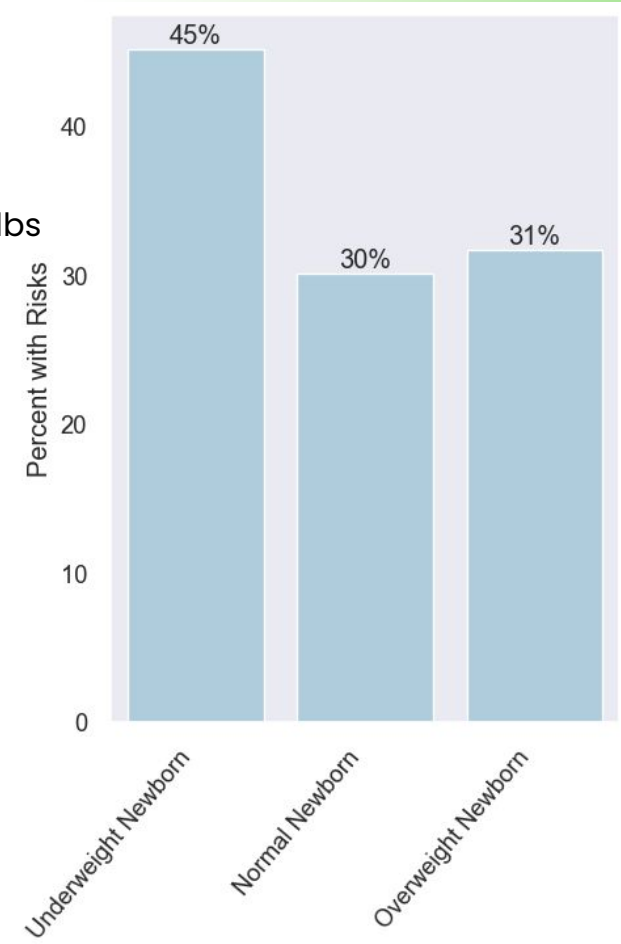
ANITA B.ORG 2024 GRACE HOPPER CELEBRATION

Exploratory Data Analysis



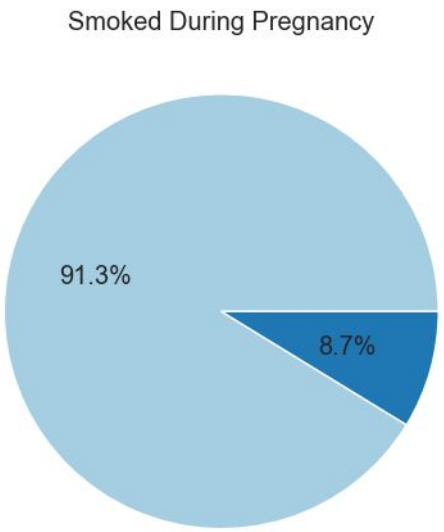
Average weight:
Male newborns: 7.3 lbs
Female newborns: 7.0 lbs

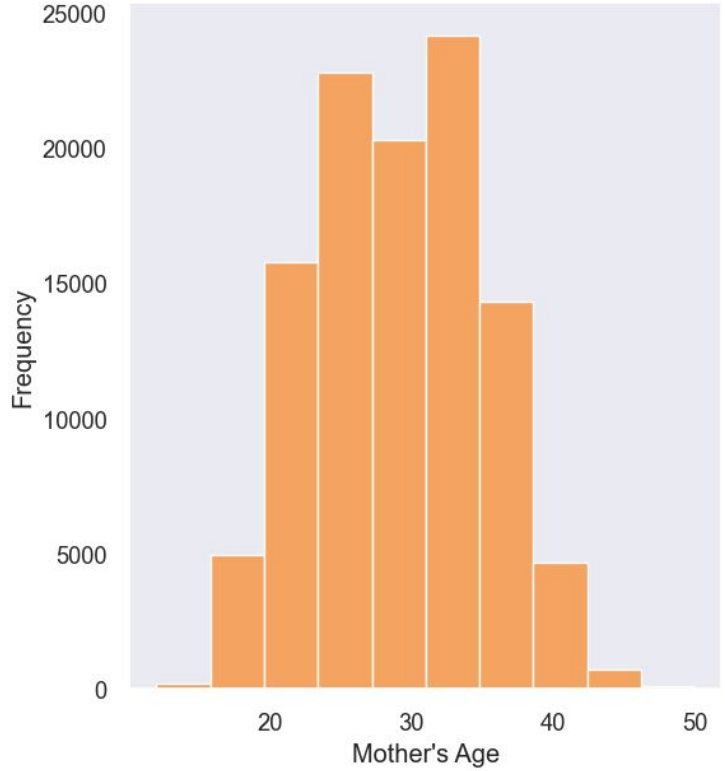


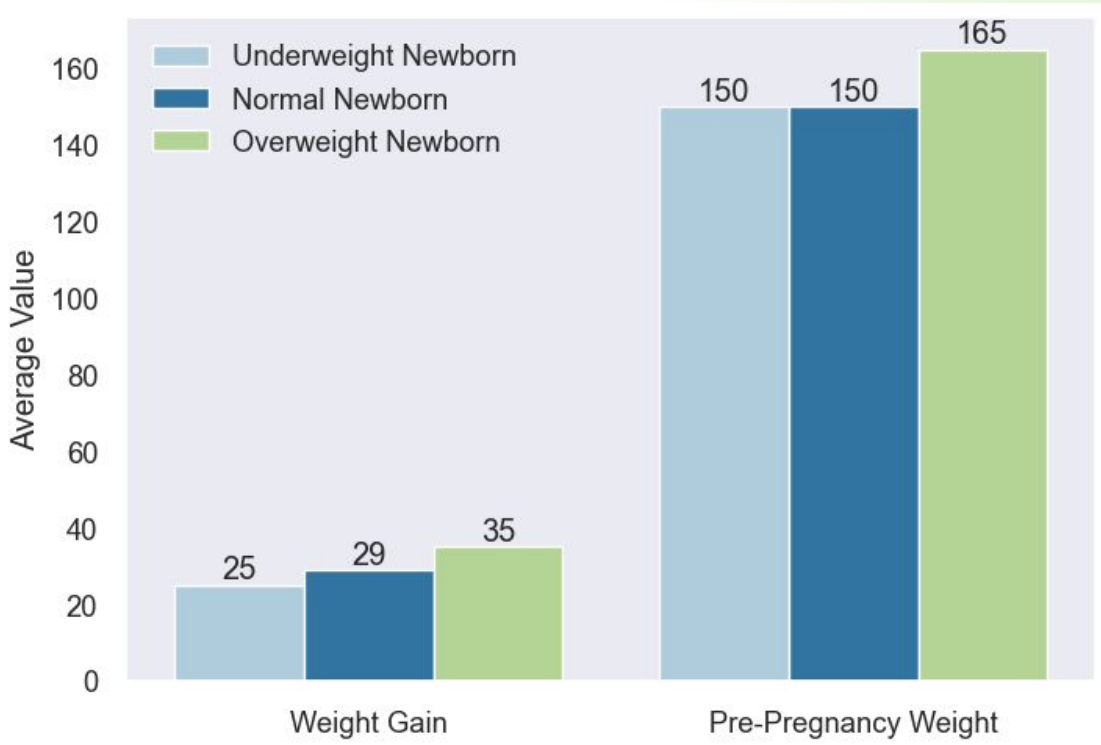


- 13% of father's education are unknown, and 1% for mothers.
- 45% of mothers that delivered an underweight newborn had risks during pregnancy, ~15% more than those with normal/overweight babies.

Exploratory Data Analysis







- The average mother's age is 29 years old.
- Those with more overweight babies had a higher pre-pregnancy weight and weight gain; those with underweight babies gained the least.

Feature Engineering

Length of Pregnancy:

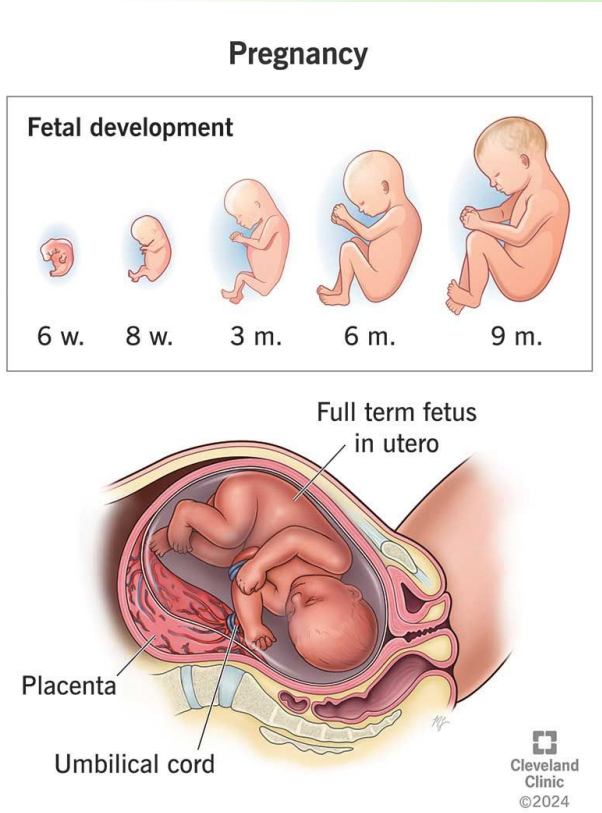
$$Pregnancy\ Length = (DOB\ Month\ - Last\ Period\ Month) * Average\ Days\ Per\ Month$$

Length of Prenatal Care:

$$Care\ Length = Pregnancy\ Length\ - (\ Prenatal\ Care\ Month\ * Average\ Days\ Per\ Month)$$

Example:

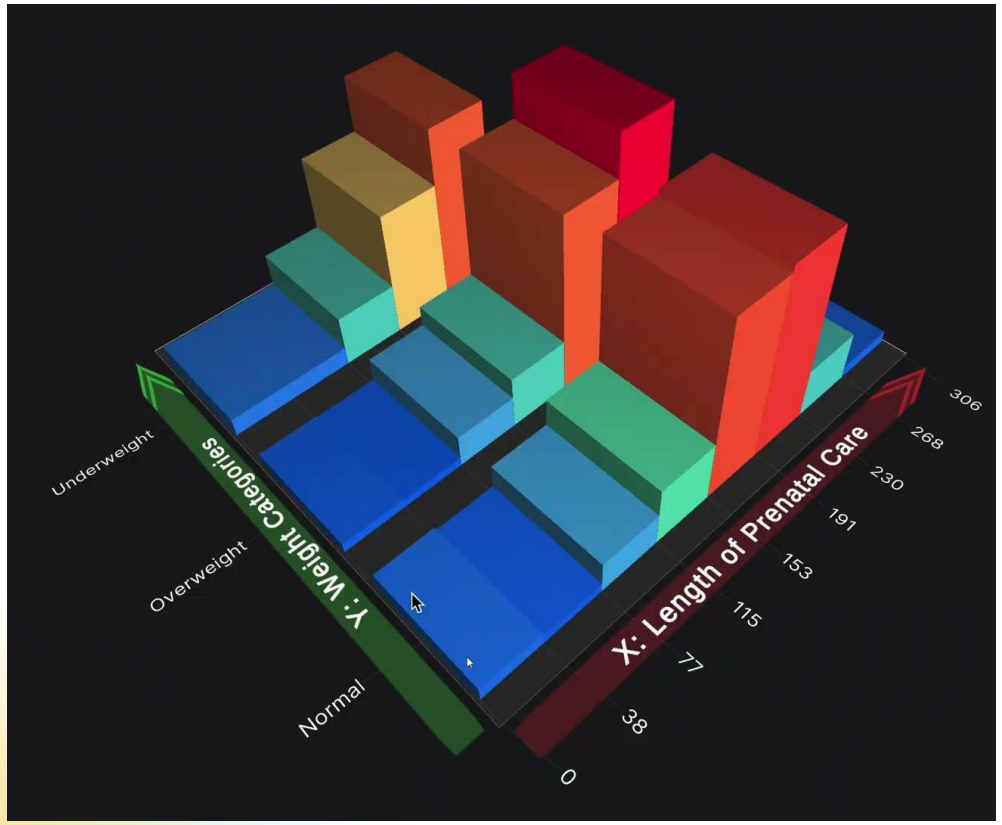
$$Pregnancy\ Length = (8\ - 2) * 30.4 = 182.4\ days$$
$$Care\ Length = 182.4\ days\ - (\ 3 * 30.4) = 91.2\ days$$

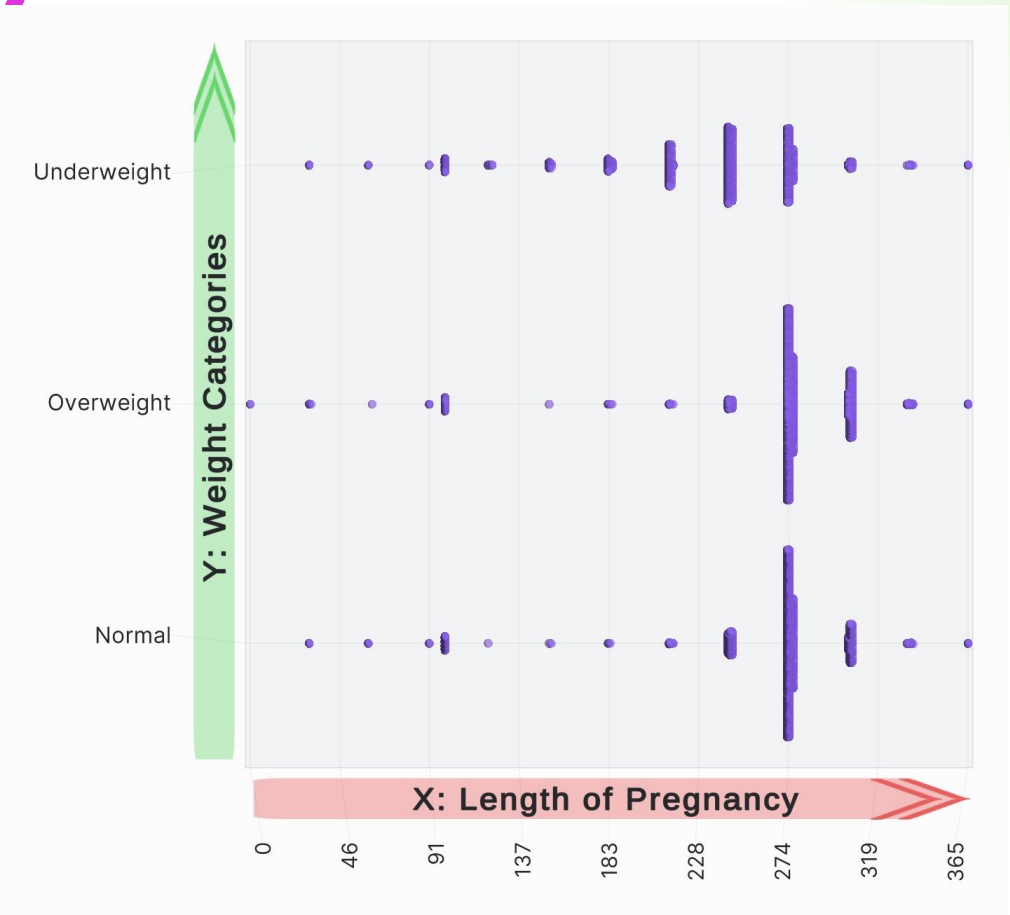


ClevelandClinic, ,https://my.clevelandclinic.org/health/articles/pregnancy

Exploratory Data Analysis

Visualizations using VIRTUALITICS

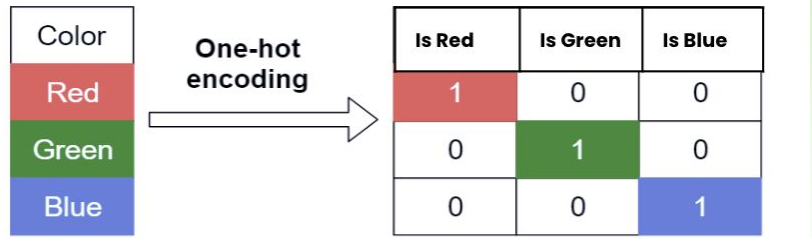


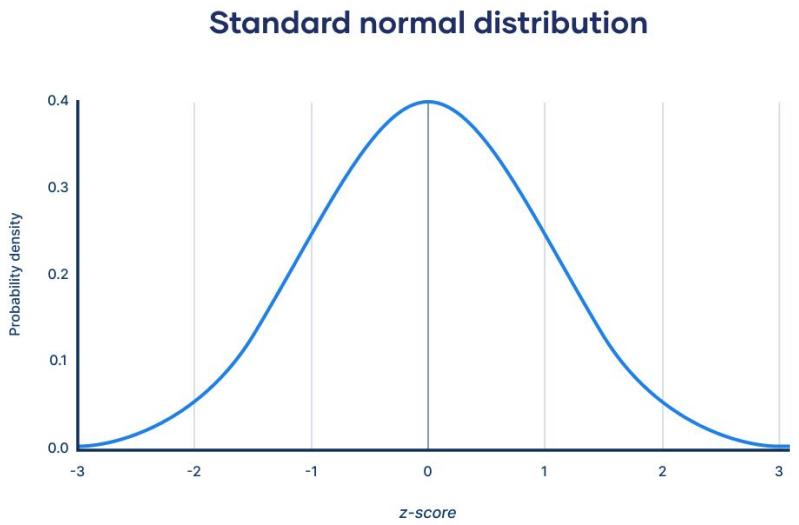


The length of prenatal care and pregnancy is on average longer for normal / overweight babies.

Preprocessing

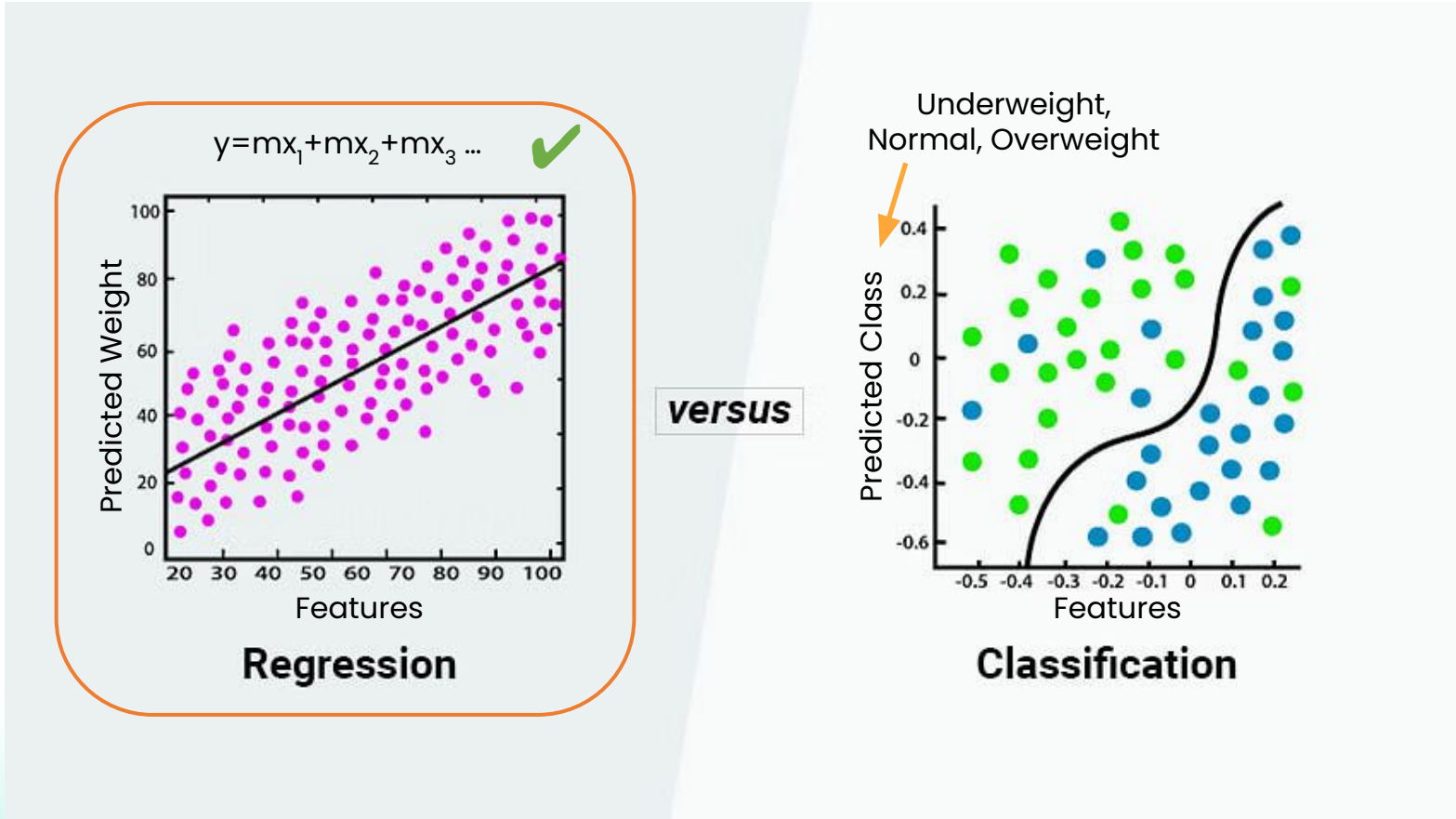
- **One Hot Encoding:**
  - Converts categorical features into numeric ones
- **Outlier Removal:**
  - Removing rows with a z-score greater than 3.
  - A z-score measures how many standard deviations above/below the mean a data point is.
- **Tested Methods of Imputation:**
  - Removing rows with null value (this method yielded better results).
  - Imputing values (99, etc.) for null values
- **Split Data Into Training and Testing:**





Scribbr, https://www.scribbr.com/statistics/standard-normal-distribution/

Model Selection

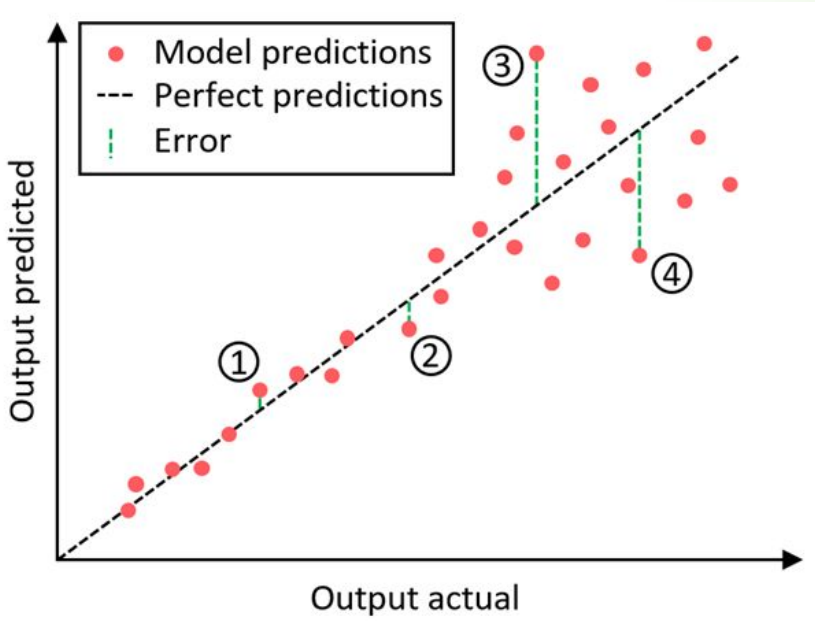


Simplilearn, https://www.simplilearn.com/regression-vs-classification-in-machine-learning-article

Model Selection

- **Lasso Regression** (L1 regularization):
  - Adds the absolute value of the coefficient in regression as a penalty term.

- **Ridge Regression** (L2 regularization):
  - Adds the squared magnitude of the coefficient as a penalty term.
  - L2 is not as robust to outliers (the squared magnitude will amplify the differences in the error of the outliers).

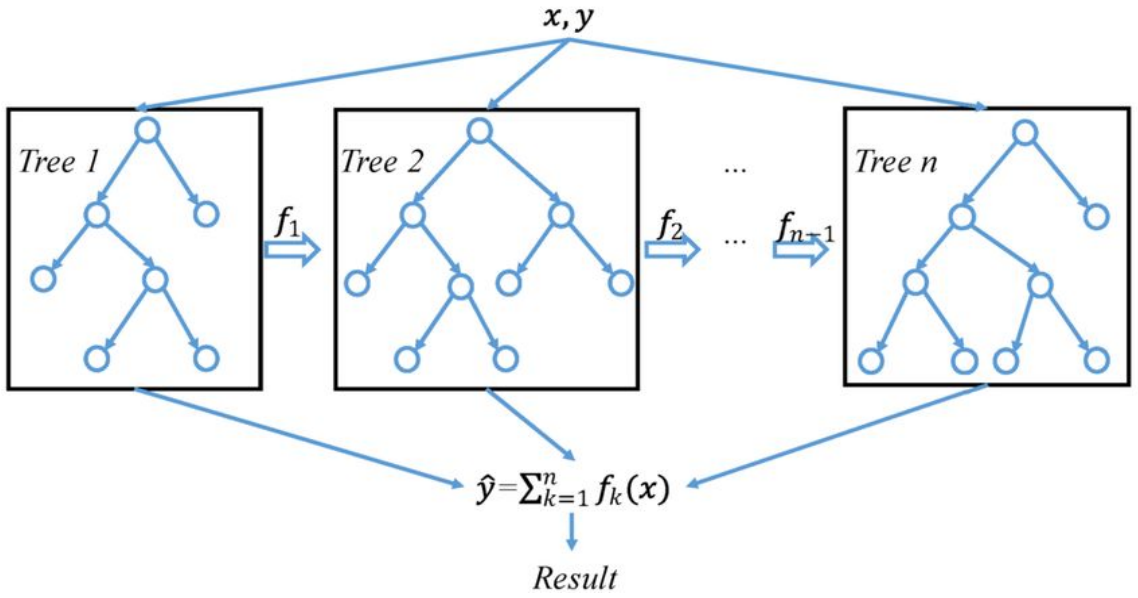


Monolith Support, https://support.monolithai.com/support/solutions/articles/80000953338-predicted-vs-actual

L1 and L2 Regularization are used to reduce the error and prevent overfitting.

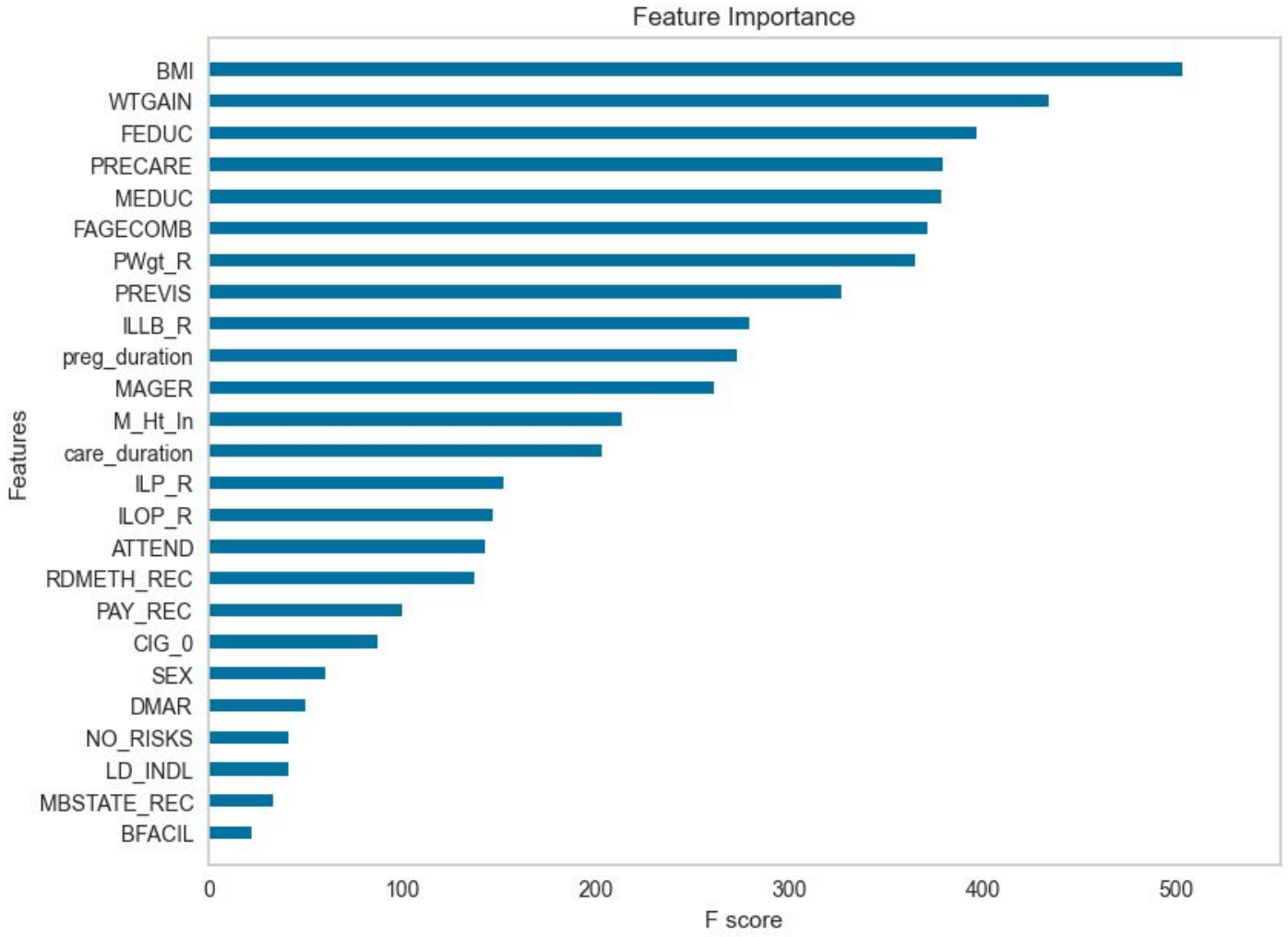
Model Selection

- **XGBoost Regression:**
  - Uses gradient boosting decision trees for regression.
  - Ensemble model constructed from decision tree models and fit to correct the predicted errors made by prior models.



ResearchGate, https://www.researchgate.net/figure/A-general-architecture-of-XGBoost_fig3_335483097

Model Feature Importance



Features with the highest importance:

- **BMI:** Body Mass Index
- **WTGAIN:** Weight gain in pounds
- **FEDUC:** Father's Education
- **PRECARE:** Month Prenatal Care Began
- **MEDUC:** Mother's Education
- **FAGECOMB:** Father's Age
- **PWgt_R:** Pre-pregnancy Weight
- **PREVIS:** Number of Prenatal Visits

Features with the lowest importance:

- **BFACIL:** Birth Place - Hospital, Home,
- **MBSTATE_REC:** Mother's Birthplace - born in the US, born outside the US, etc.
- **LD_INDL:** Induction of Labor
- **NO_RISKS:** Risks during pregnancy
- **DMAR:** Marital Status

Model Evaluation

- Using XGBoost using all the features has the lowest MSE and RMSE.
- The features with the lowest importance can be removed from the model, since they don't affect MSE or RMSE.
- Splitting the models yields slightly better results for female.

Model	MSE	RMSE
XGBoost – All Features	1.06	1.03
Ridge Regression	1.53	1.24
Lasso Regression	1.59	1.26
XGBoost - Upsampled	1.43	1.19
XGBoost – Selected Features	1.06	1.03
XGBoost - Female	1.03	1.03
XGBoost - Male	1.06	1.06

Model Evaluation

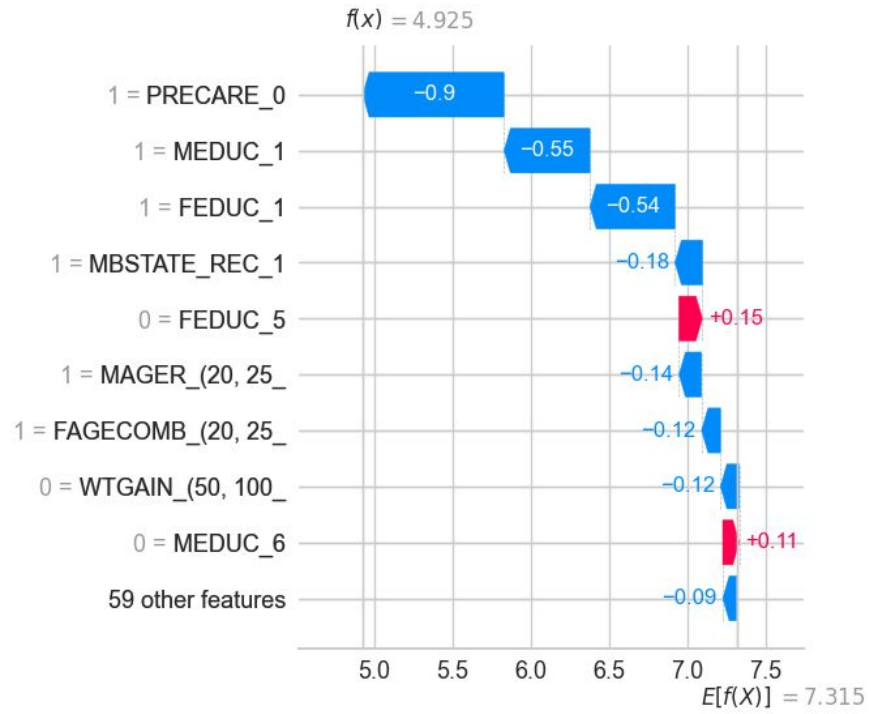
- A prediction interval is an estimated range of values
- Adding a prediction interval will provide a measure of reliability to the predictions
- Using the MAPIE regressor to calculate intervals.

ANITA B.ORG 2024 GRACE HOPPER CELEBRATION

Case Studies: Underweight & Normal Newborn

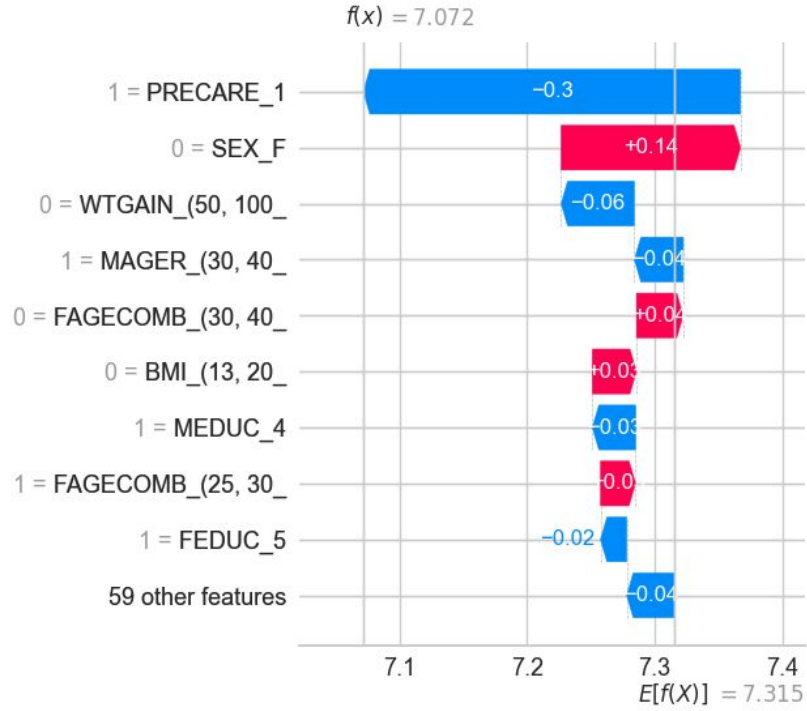
Underweight (Actual: 4.4 lbs, Predicted: 4.9 lbs):

- Did not receive prenatal care
- Both mother and father's education was less than 8th grade
- Mother's age was 20-25



Normal (Actual: 7.2 lbs, Predicted: 7.1 lbs):

- Started prenatal care the first month into pregnancy
- Mother received some college credit, father has an associate's degree
- Mother's age was between 30 – 40



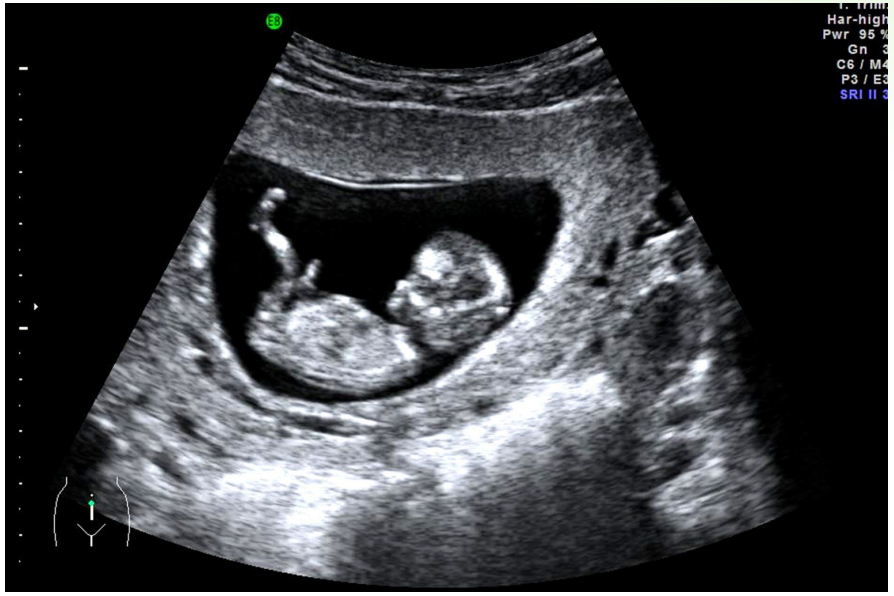
Conclusion

- It is important to think about the problem at a high-level by assessing:
  - if the features are useful to the prediction
  - if the predictions are useful to an end user

- Future improvements: Improving the model to accurately predict a value rather than an interval.

- Current method for determining weight is an ultrasound scan by gynecologist through the pregnancy

- ML predictions could be used concurrently for better diagnosis support



NHS, https://www.nhs.uk/pregnancy/your-pregnancy-care/12-week-scan/

Findings & Recommendations from the March of Dimes

Reasons a baby could be underweight:

- Preterm labor (delivering before 37 weeks of pregnancy)
- Chronic health conditions (high blood pressure, diabetes, heart, lung and kidney problems)
- Taking certain medicines to treat certain health conditions
- Infections (rubella, chickenpox, toxoplasmosis)
- Not gaining enough weight during pregnancy
- Being pregnant with multiples (twins, triplets or more)
- Smoking, drinking alcohol, using drugs
- Age: Being a teen (especially younger than 15) or being older than 36

To have a normal baby:

- Get regular prenatal care
- Weight before pregnancy matters, as does how much weight is gained during pregnancy
- Take appropriate steps to manage chronic health conditions

ANITA B.ORG 2024
GRACE HOPPER CELEBRATION

THANK YOU

me+we