Weight at Birth Prediction

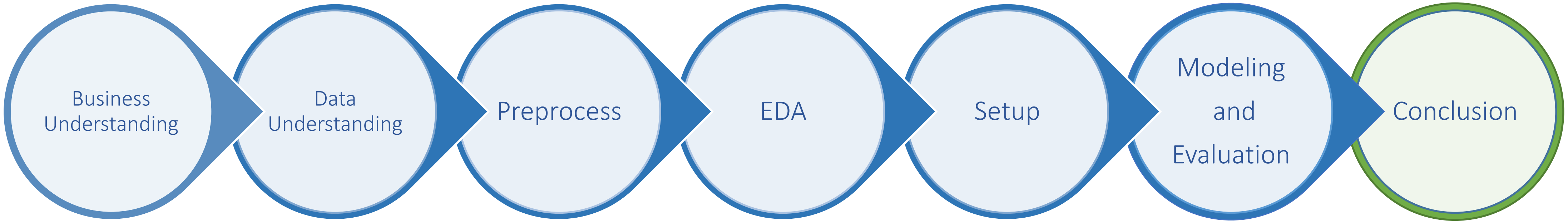
Mohammad Mirsafaei

Elaheh Toulabi Nejad

Dataset

- Name: US_births(2018)
- Columns: 55
- Rows: 3.8M
- Context: This data contains Information about all of the child birth in the United States in the year of 2018).
- Task: This data could be used for predicting the weight of a baby.

Progress



Business Understanding

Data Understanding

Preprocess

EDA

Setup

Modeling and Evaluation

Conclusion

Challenges

Large amount of missing values

- Analyzing data and replacing with mean or median whenever needed.

Null values in BMI column

- Reconstructing part of the dataset based on available columns(Height, Weight)
- Replacing rest of the dataset with mean

Statistical columns with no effect on weight

- Analyzing data and removing some columns based on their usage and meaning that had no effect on prediction

High volume of data and hardware limitations

- Sampling over 600,000 records of data
- Running statistical tests to ensuring sample validity.

Columns with little effect

- Using OLS model and P-value to determine what columns should be removed

High number of columns

- Using Dimension Reduction methods (PCA) for reducing number of column up to 75%

No fixed number of groups for outcome variable

- Analyzing each potential group of data.
- Searching through articles and reliable sources such as WHO
- Using experts knowledge
- Running most of the models on each assumption and analyzing and comparing results

Selecting between evaluation parameters

- Deciding to concentrate on Recall and FScore more than accuracy

Unbalanced Dataset

- Balancing dataset based on idea of improving “Low” category

Brief Checklist Explanation

P1

- Columns: 55
- Rows: 3.8M

P2

- Distribution of Weight
- Correlation of Columns with Weight
- Pair Plot
- Impact of Categorical Features On Weigh
    - Sex
    - Smoked
    - Prior Dead
    - First Birth
    - Previous Cesarian

P3

- Dropping null columns
- Dropping columns based on meaning
- Constructing null values whenever was possible with other columns
- Adding new features
- Standardization
- Removing outliers
- Sampling

P4

- Linear Models: OLS, Linear Regression, Ridge, Lasso, Decision Tree Regressor, Random Forest Regressor
- Logistic Models: Logistic Regression, Random Forest Classifier, Decision Tree Classifier, Neural Network, GaussianNB

P5&P6

- Linear Models:

Model	R-squared	MAE	MSE	RMSE
OLS	0.256	363.4576	222061.6544	471.2341
LinearRegression	0.2528	363.4576	222061.6544	471.2341
Ridge	0.2528	363.4576	222061.6494	471.2341
Lasso	0.2527	363.4378	222056.3011	471.2285
DecisionTreeRegressor	-0.4930	517.7256	443736.9522	666.1358
RandomForestRegressor	0.2833	356.4205	213002.4881	461.5219

P5&P6

- Logistic Models:

#class	Model	Avg Precision	Avg Recall	Avg F1-score
11	LogisticRegression	0.36	0.22	0.25
	Random Forest	-	-	-
	Decision Tree	0.17	0.23	0.20
	Naive Bayes	0.33	0.16	0.19
	Neural Network	0.36	0.27	0.27
2	LogisticRegression	0.79	0.71	0.73
	Random Forest	0.79	0.82	0.79
	Decision Tree	0.75	0.67	0.70
	Naive Bayes	0.77	0.70	0.72
	Neural Network	0.78	0.70	0.73

P5&P6

- Logistic Models:

#class	Model	Low Precision	Low Recall	Low F1-score
2	LogisticRegression	0.38	0.67	0.48
	Random Forest	0.6	0.28	0.38
	Decision Tree	0.32	0.55	0.40
	Naive Bayes	0.35	0.59	0.44
	Neural Network	0.37	0.63	0.46
11	LogisticRegression	-	-	0.2506
	Random Forest	-	-	-
	Decision Tree	-	-	0.0225
	Naive Bayes	-	-	0.1928
	Neural Network	-	-	0.1731

P7

- Improving Recall and F1-score of “Low” category by balancing dataset.
- Using PCA and analyzing best number of columns

#class	Model	Low Precision		Low Recall		Low F1-score	
		Before	After	Before	After	Before	After
2	LogisticRegression	0.66	0.38	0.23	0.67	0.35	0.48
	Random Forest	0.67	0.60	0.20	0.28	0.31	0.38
	Decision Tree	0.57	0.32	0.17	0.55	0.28	0.40
	Naive Bayes	0.44	0.35	0.30	0.59	0.36	0.44
	Neural Network	0.58	0.37	0.28	0.63	0.38	0.46

P8

- Deciding to define new logistic problem due to flaw in data.
- Achieving 356.63 MAE by using “Random Forest Regressor” in linear part that is acceptable value.
- Using 11 categories due to ISCD standard and also 2 categories by consulting with experts in field.

#Class	Model	Precision	Recall	F1-score	Precision for low class	Recall For low class	F1-score For low class
2	Logistic Regression	0.79	0.71	0.73	0.38	0.67	**0.48**
11	Logistic Regression	0.36	0.22	0.25	-	-	**0.2506**