1 Challenge

Our challenge is to predict the birth weight of Sammi Sternbach’s baby. Sammi, an obstetrics nurse at the Stony Brook Hospital was due on 8th December, 2014. Now, she is in her 41st week of pregnancy. The birth weight of a baby can be predicted based on the external and internal factors associated with the mother and her environment.

Every parent in the world desires for a healthy baby. The best measure of the health of a baby is its weight. Prediction of the birth weight in normal range is an indicator for the parents as well as the doctors that the baby is healthy. A significant deviation can be an indication of an abnormality. It may be due to a change in the mother’s habits or due to the change in the environment, knowingly or unknowingly. Another instance which excites people to predict the baby weight is the concept of Baby Pools. It also carries rewards for the closest prediction.

Factors that help to predict the birth weight are ethnicity of the parents, education of the parents, medical conditions of the mother etc. Location also plays an important role along with changes in food habits and environment. Predictions based upon a large data-set of these factors indicates the general weight distribution for a certain category of people. Ability to apply predictive models from a niche database with knowledge in Data Science helps us to predict the baby weight with minimal errors.

2 Background

Research in the field of birth weight began with discovery of increasing low birth weight among babies. Premature birth and low birth weight are the leading causes of increase in infant and neonatal mortality (death before 28 days of age) rates in the United States. The low birth weight makes up 60-80% of the infant
mortality rate. On an average, birth weight less than 5.5 pounds is termed low birth weight (LBW) whereas birth weight higher than 9 pounds is termed as high birth weight (HBW). A general healthy baby weighs 5.5 to 9 pounds.

Few interesting facts that we stumbled upon are, the worlds heaviest baby (born to a healthy mother) weighed 10.2kg (22.5 pounds) whereas the worlds lightest (healthy) baby weighed just 260g (0.6 pounds). The longest pregnancy on record was 375 days but surprisingly the baby weighed only 6 pounds 15 ounces which is quite a normal birth weight.

![Fig. 1. Effect of sex of the baby on birth weight](image)

Low birth weight infants are more likely to experience physical and developmental health problems or die during the first year of life than infants of normal weight. These babies tend to be independent of gestational age or the baby being born premature [5].

According to preliminary data, 8.2% of infants were born with low birth weight in 2009. In 2006, the rate of low birth weight was recorded highest in four decades (8.3%). The increase in multiple births, which are at high risk of low birth weight, strongly influenced this increase. The rates of low birth weight also rose for singleton births.[5]

In 2009, the rate of low birth weight was much higher among infants born to non-Hispanic Black women (13.6%) than infants born to mothers of other racial/ethnic groups. It was followed by a rate of 8.3% among Asian/Pacific
Islanders, and by a rate of 7.3% among American Indian/Alaska Natives. Low birth weight occurred among 7.2% of infants born to non-Hispanic White women, while infants of Hispanic women experienced the lowest rate (6.9%). Low birth weight levels in 2009 were not significantly different from 2008 for non-Hispanic White, non-Hispanic Black, and Hispanic infants.[5]

Low birth weight also varied by maternal age. In 2008, the rate of low birth weight was highest among babies born to women younger than 15 years of age (12.4%), followed by babies born to women aged 40 - 54 years (11.8%). The lowest rates occurred among babies born to mothers aged 25 - 29 years and 30 - 34 years (7.4% and 7.6%, respectively).[5]

Numerous barriers often stand between pregnant women much needed care. For example, the inability to pay for services causes many women to delay or even forgo prenatal care. Women tend to be not aware of the importance of prenatal or preventive care. In addition, women and their families are often overwhelmed by the stresses of poverty. Recent studies are now focusing on the relationship between stress and poor birth outcomes, especially in high-risk women. [4]

3 Literature review

We have reviewed research papers which discusses the effect of birth weight on mortality rate, prediction of baby weight, importance of birth weight, understanding the relationship between prenatal care and birth weight, and effect of maternal anaemia on birth weight.

Importance of birth weight related to mortality

This paper [1] discusses the importance of birth weight and evaluates if low birth weight can be a cause for mortality. It introduces the concept of LBW paradox. This paradox states that, LBW babies with high rates of infant mortality rate have low mortality than those with low rates of infant mortality rate. The history of this is entwined with one of the most famous controversies in the history of epidemiology: the debate over role of casual cigarette smoking.

In 1950s, researchers found that mothers who smoked delivered smaller babies and there was a greater chance of high mortality. But this came with an interesting observation: LBW babies born to mothers who smoked had lower mortality than LBW babies of mothers who did not smoke. This leads us to a rather strange conclusion that If a baby was born LBW, it seemed like an advantage to a mother who smoked.
This paradox is also associated with factors such as altitude at which a baby is born, race of the baby, twins etc. It was shown that babies in Colorado are smaller when compared to those born in places at lesser altitude.

**Accurate prediction of term birth weight from prospectively measurable maternal characteristics**

In this paper[2], the authors propose an equation for the prediction of birth weight based on just the maternal characteristics. Data was collected about features such as age, parity, height, weight, level of obesity, rate of pregnancy, weight gain, fetal gender and length of gestation period. Using a cross-validated split sample multiple regression analysis, the combination of the features that have the best predictive power on birth weight were determined. The final equation that predicts the birth weight is as follows:

\[
\text{birth weight (g)} = \text{gestational age (days)} \times [9.40 + 0.255 \times \text{gender} + \\
0.000232 \times \text{height (cm)} \times \text{maternal weight at 26 wk (kg)} + \\
4.89 \times \text{3rd trimester weight gain rate (kg/d)} \times (\text{parity} + 1)]
\]

where: gender = -1 for females;
+1 for males;
0 for unknown gender
and gestational age = conceptual age (days) + 14

Predictive accuracy was assessed using a jackknifing procedure. The results were compared to similar types of birth weight predictions obtained via previous algorithms. These variables explained 33% of variance in birth weight and 10.8% variance in predicted birth weight. One of the major advantages of this equation is that values of all variables can be obtained by the beginning of third trimester. This is helpful in making an early prediction of the birth weight and thereby can be used by doctors to advise preventive measures in case of abnormalities. One limitation of this equation is that it doesn’t consider smoking as a factor affecting birth weight.

**Accuracy of Ultrasonographic prediction of term birth weight**

In this paper[3], accuracy of 25 different ultrasonic algorithms for prediction of the birth weight was evaluated and the results were compared to the equation above.

Three main parameters obtained from ultrasonography which are used in the prediction of birth weight are: fetal bi-parietal diameter (BPD), abdominal circumference (AC), and femur length (FL). The 25 algorithms were divided into 5 categories based on fetal biometric measurements. Paired t tests were used to make comparisons within and between algorithms that used the same sets of
fetal biometric measurements. Two-tailed t tests were used to identify the algorithm or group of algorithms within each category. This significantly gave more accurate birth weight predictions than other algorithms. Finally, likelihood analyses was performed to assess the differences in the percentage of birth weights that were predicted accurately to within 10% and 15% of the actual birth weight for the best performing ultrasonography algorithms which contained maternal characteristic equation.

Comparisons among different categories of ultrasonic prediction algorithms revealed that algorithms that were solely based on the fetal AC or the AC in combination with the BPD (with or without FL) were significantly more accurate than the algorithm of Warsof et al [3] that is based purely on the FL. The results of this study are interesting, particularly because the ultrasonography algorithms that were based exclusively on the measurement of the fetal AC proved to be as accurate as the other classes of equations that are based on multiple standard ultrasonography fetal measurements.

Based on updates from Sammi, we have studied a few research papers to analyze their findings and applied them on our data-set.

**Reference range of birth weight with gestation and first-trimester prediction of small-for-gestation neonates**

In this paper [2], there was a prospective screening study for adverse obstetric outcomes in women attending their routine first hospital visit in pregnancy.
In this study, which is held at 11 weeks to 13 weeks of gestation, they recorded maternal characteristics and performed a trans-abdominal ultrasound scan to confirm gestational age from the measurement of the fetal crown-rump length (CRL) and to diagnose any major fetal abnormalities.

Data on pregnancy outcome were collected from the hospital maternity records or their general medical practitioners. Patients were asked to complete a questionnaire on maternal age, racial origin (Caucasian, African, South Asian, East Asian, and mixed), cigarette smoking during pregnancy (yes or no), parity (nulliparous if there were no previous pregnancies beyond 23 completed weeks or parous), birth weight of previous neonates (only SGA, only non-SGA, or mixture of SGA and non-SGA), method of conception (spontaneous or assisted), and medical history of chronic hypertension and pre-pregnancy diabetes mellitus.

According to this paper, in the total population of 33,602 pregnancies there was a significant association between birth weight and gestational age at delivery:

\[
\text{Expected log}_{10} \text{birth weight} = -0.6329 + 0.1873 \times (GA) - 0.0021 \times (GA)^2;
\]

\[R^2 = 0.574, \ SD = 0.0581, \ p < 0.0001\]

**Understanding the relationship between prenatal care and birth weight**

This paper[3] was used to study the potential effect of prenatal care on birth weight. Several methodological issues present serious challenges, including the identification of ways to measure prenatal care. The team investigated effectiveness of early prenatal care initiation in producing positive birth weight outcomes. The study shows that prenatal care is a significant determinant of birth weight, with effects ranging from decrease in birth weight by a margin of 36 to 107 grams for each month of delay in care initiation.
Maternal anaemia and its associated risks for preterm delivery and low birth weight

This paper[6] estimates the effect of maternal anaemia on various perinatal outcomes. A total of 1,174 anaemic and 547 non-anaemic women were enrolled. Their median age was 24 years (range 14 - 46 years) and median parity was 2 (range 0 - 17). The prevalence of anaemia and severe anaemia was 68% and 5.8%, respectively. The risk of pre-term delivery increased significantly with the severity of anaemia, with odds ratios of 1.4, 1.4 and 4.1 respectively for mild, moderate and severe anaemia. The corresponding risks for LBW and VLBW were 1.2 and 1.7, 3.8 and 1.5, and 1.9 and 4.2 respectively. This paper concludes that the risks of preterm delivery and LBW increased in proportion to the severity of maternal anaemia.
Fig. 4. Effect of Anaemia on birth weight

4 Data Matrices

The Dataset is obtained from Vital Statistics Data on Births and related parameters from the UNC Odum Institute. We acknowledge the State Center for Health Statistics (SCHS) and the Howard W. Odum Institute for Research in Social Science at UNC at Chapel Hill as the source of data. The current dataset is the part of a huge dataset of vital statistics details across years. The dataset has majority of records for the state of North Carolina and a significant number of outer state cases. It contains 133422 cases for the year 2008 and similar data is available for the past 20 years across 125 variables. We used the following dataset for initial analysis and prediction. The following are some interesting variables observed in the dataset.

- Race of Mother and Father
- Sex of the Baby
- Age of Mother
- Total Pregnancies
- Completed Weeks of Gestation
- Month Prenatal Care Began
- Locality
- Plurality
- Average # Cigarettes Used Daily
- Average # Drinks Consumed Weekly
- Medical History
- Education of Parents
- Apgar Score
Race of Mother And Father
We observed that Race plays an important role in the baby weight. There are 8 categories of races observed. They are Other Nonwhite, White, Black, American Indian, Chinese, Japanese, Hawaiian, Filipino and Other Asian or Pacific Islander.

Sex of the Baby
The birth weight of the baby is also co-related to the sex. Male babies are found to be a little heavier than the female babies as shown in the further section.

Age of Mother
This is of not much significance but upon analysis was observed that the probability of lower birth weight is slightly lower among older women.

Parity
It has been shown that mothers with first pregnancy give birth to babies who weigh less.

Completed Weeks of Gestation
This parameter holds a direct relation to the weight of the baby.

Month Prenatal Care Began
This parameter is to help establish the growth of the baby from the 1st trimester to delivery. Generally it has been shown that the maximum growth of the baby is in the 3rd trimester of pregnancy.

Location
Apart from race, location plays an important role in the determination of the baby weight. People from different locations are subjected to different variety of food and environment. It contains all the states of the United States with a majority from North Carolina.

Plurality
This tells whether it is a singleton or twins or multiple baby births. It has been shown that as the multiplicity of the baby number increases, the weight of the babies decreases. We will verify this claim from the data available.

Other Interesting Variables

Average # Cigarettes Used Daily
This data is available and has been indicated that cigarettes tends to affect the weight of the baby. Our next goal is to uncover a relation between these parameters.
**Average # Drinks Consumed Weekly**

The effect of drinks on baby weight has been mentioned previously. We will try and establish a relation from the dataset.

**Medical History**

This is one of the major outlier cases in prediction of birth weight of the babies. Medical conditions irrespective of the seriousness can affect the weight of the baby. For example, a mother with diabetes tends to have a heavier baby.

**Education of Parents**

This parameter looks esoteric, but it might help to indicate any correlation between the education of Parents and baby weight. The educational background of the parents tends to affect their lifestyle and habits.

**Apgar score**

The Apgar score indicates the health of the baby. We are looking at the possibility of establishing a relationship between the health of the baby and the Apgar score.

This dataset has a lot of additional data (such as Hospital Type Code) which is not relevant to our project. For the year 2008, it had 133422 rows with 125 columns. We trimmed it down by removing those features(columns) which are not relevant. We have also refined the dataset by splitting the categorical variables into several binary variables.

The below parameters were removed. These are either unwanted features or post birth parameters that cannot be considered for prediction.

- Hospital Type Code
- Attendant
- Pounds of Birth weight
- Ounces of Birth weight
- Events of Labor
- Conditions of the Newborn

The pounds and ounces are combined to give the birth weight in decimal.

The nominal variables such as ethnicity of parents are converted into binary. For instance, race of mother was converted to different binary variables such as White_Mom, White_Dad, Black_Mom, Black_Dad, Hisp_mom, Hisp_dad, etc.
After cleansing, the final dataset has 69051 rows and 52 columns. The dataset was divided into training, test and evaluation data in a random manner. The training set comprises of 80% of the data i.e 55240 rows. The remaining 20% (i.e., 13810 rows) are used as testing data set.

4.1 Sammi Sternbach & Updates

The main aim of our project is to determine the weight of Sammi Sternbach’s baby. She is currently pregnant and her baby is due in December 2014.

We got to know that Sammi’s sister is pregnant and we were informed by Sammi that her sister had a baby boy who weighed 8lb 9oz. It was her third baby. The first baby weighed over 8lbs and second baby weighed 7 lbs. 4 oz. Also, we were informed that Sammi’s sister is a gestational diabetic.

Also, we got a few updates about Sammi. Her current gestational age is 41 weeks. Her Prenatal care began in May and she has had 10-11 visits so far. She is an anaemic and her weight gain at 41 Weeks is 20 lbs.

![Fig. 5. Sammi(Right) and her sister(Left) : At 27 weeks](image)
Fig. 6. Sammi’s sister gave birth to a baby boy weighing 8lb 9oz

5 Observations

Fig. 7. Effect of Smoking on birth weight
In first figure, the red line represents smoker mother and the blue line is for non-smoker mother. LBW paradox can be clearly observed in the figure above as birth weight of the baby of a non-smoker mother tend to be less than that of a smoker mother.

In second figure, the red line represents drinker mother and the blue line is for non-drinker mother. The above plot confirms that birth weight of the baby of a drinker mother tend to be less than that of a non-drinker mother.

6 Baseline Model

For our baseline model, we just considered the average of all birth weights in our data set.

The prediction for sammi’s child according to the baseline model is 6.834 lbs (3.1 Kgs).

7 Advanced Models

7.1 Linear Regression and Ridge Regression

We evaluated by fitting data to a multivariate regression model. We considered both Ordinary Least Squares (OLS) regression and Ridge regression. However,
there is not much of a difference in the final prediction of Ridge regression when compared to OLS. But one can notice the co-efficients and intercept values generated by Ridge regression seem more interpretable compared to the unusually large values generated by OLS regression. This is because of the penalty imposed by Ridge regression on variables having very high co-efficient values.

7.2 K Nearest Neighbours

K-nearest neighbours model identifies those specific instances which are most similar to the desired instance based on its features. However, care needs to be taken in-order to make sure that all features get equal consideration while measuring the similarity. For example, a feature based on a scale of 1-100 would have a more effect on the similarity than a feature based on scale of 0-1. To account for this, we normalized the data using z-scores. To measure the effect of k on the model performance, we have evaluated the model using different values of k ranging from 1 to 200. We have observed that error in prediction (in terms of RMSE value) is highest when k is 11 and lowest when k is in the range of 100. Starting from k =1, RMSE increases with k till k=11 and then we observe a sharp spike and RMSE starts to decline. This can better be visualized in the plot below.

![Plot of RMSE value for KNN Model at different values of K](image)

**Fig. 9.** Plot of RMSE value for KNN Model at different values of K
7.3 Decision Trees

We also tested our data against a decision tree model and evaluated the results at different depths. A plot of the Root Mean Squared Error at varying depths is shown below:

Fig. 10. Root Mean Squared Error at different depths in the Decision tree
Evaluation

All the above models are evaluated using the standard regression error metrics and by plotting the probability distribution of error. Results are as follows:

Fig. 11. Error probability distribution for Baseline, OLS, Ridge and KNN Models.
Lines representing mean and standard deviation of the error were also plotted in the above figure.

**Inference from the plot**

From the above plots, it can be inferred that OLS, Ridge and KNN have a narrower standard deviation of error than the baseline model. However, there isn’t a noticeable difference (in terms of model prediction accuracy) in the plots of OLS and Ridge models. KNN, however, has a slightly higher standard deviation for error and the average error is on the higher side too.

**Results of the evaluation metrics**

<table>
<thead>
<tr>
<th>MODEL</th>
<th>Root Mean Square Error</th>
<th>Mean Absolute Error</th>
<th>Explained Variance Score</th>
<th>R2 Score</th>
<th>Predicted Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline Model</td>
<td>1.344</td>
<td>9.86</td>
<td>1.11*e-16</td>
<td>0</td>
<td>6.834</td>
</tr>
<tr>
<td>Linear and Ridge Regression</td>
<td>1.056</td>
<td>0.830</td>
<td>0.3687</td>
<td>0.3685</td>
<td>7.4877</td>
</tr>
<tr>
<td>K-Nearest Neighbor(k=30)</td>
<td>1.2988</td>
<td>1.013725</td>
<td>0.132261</td>
<td>0.096206</td>
<td>7.0625</td>
</tr>
</tbody>
</table>

**Fig. 12.** Error Metrics
7.4 Ensemble Models

We observed that the decision tree-regression gave us a better result when compared to our baseline model, Linear regression, Ridge regression and K Nearest Neighbors. To improve this model further we are making use of ensemble techniques to boost the algorithm. By making use of the contributions of many decision trees we wanted to test if we could build a stronger model. We have subjected our data set to two ensemble models, the AdaBoost and Random Forest.

7.5 AdaBoost

This is an ensemble model which tends to vary the weights assigned to a point in classification. It is based on the principle of increasing the weight of points which are classified wrongly and decreasing the weight of points which are classified correctly. Thus it tends to set the decision boundary bit by bit to the most accurate position yielding a better R-square value and a minimal root mean square error compared to the decision tree regression. We are adding regression to determine the required numerical value.

7.6 Random Forest

This is another ensemble model which follows a different approach to AdaBoost. It takes the voting results of many decision trees to decide the best result. It is based on the principle of combining results of many weak decision trees and produce a single and accurate classifier. We are adding regression to determine the required numerical value. Although on an average case studies show ADA boost to be slightly better, random forest tends to perform best for Data-set. It has slightly better R-square and a minimal root mean square error comparatively.

Evaluation

Similar to the previous models, we evaluated these models using the error metrics and plot of probability distribution of error. The results are as below:
Error probability distribution

**Fig. 13.** Error Distributions of Classifier-Regression Models

**Inference from the plot**
We can infer from the plot that all the models are showcasing similar error distributions. The numerical values are evidence of this, where the root mean
square value is very near to 1 in all the models. We will have to consider other parameters to determine the better model.

The summary of evaluation metrics of all the models we have built from the beginning is shown in the table below:

### 7.7 Model Selection

<table>
<thead>
<tr>
<th>MODEL</th>
<th>Root Mean Square Error</th>
<th>Mean Absolute Error</th>
<th>Explained Variance Score</th>
<th>R2 Score</th>
<th>Predicted Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline Model</td>
<td>1.344</td>
<td>9.86</td>
<td>1.11*e-16</td>
<td>0</td>
<td>6.834</td>
</tr>
<tr>
<td>Linear and Ridge Regression</td>
<td>1.056</td>
<td>0.830</td>
<td>0.3687</td>
<td>0.3685</td>
<td>7.4877</td>
</tr>
<tr>
<td>K-Nearest Neighbor (k=30)</td>
<td>1.2988</td>
<td>1.013725</td>
<td>0.132261</td>
<td>0.096206</td>
<td>7.0625</td>
</tr>
<tr>
<td>Decision Tree (max depth=7)</td>
<td>1.0228</td>
<td>0.795126</td>
<td>0.406359</td>
<td>0.406292</td>
<td>7.5971</td>
</tr>
<tr>
<td>Ada Boost (estimators=30)</td>
<td>1.0338</td>
<td>0.805893</td>
<td>0.431822</td>
<td>0.429056</td>
<td>7.7089</td>
</tr>
<tr>
<td>Random Forest (depth=10)</td>
<td>0.9922</td>
<td>0.768459</td>
<td>0.451359</td>
<td>0.451355</td>
<td>7.6648</td>
</tr>
</tbody>
</table>

![Fig. 14. Model Selection Table](image)

Using the results from the error metrics and the error distribution plots, we have decided to choose Random Forest model to make our final prediction.
8 Final Prediction and Conclusions

Our prediction is made using the following details of Sammi. When we input these values into our trained Random forest regressor we get the following prediction:

- Gender of Baby: Female
- Fathers Age: 27
- Mothers Age: 27
- Parity of Mother: 1
- Gestation period: 41 Weeks
- Weight Gained: 20 lbs.
- Drinks per week: 0.25
- Anaemic: True
- Ethnicity of Father: WHITE
- Ethnicity of Mother: WHITE
- Marital Status: True
- Education of Father: 12 years
- Education of Mother: 16 years
- Pre Natal Care: True
- Number of Prenatal Visits: 11

The birth weight of Sammis baby is predicted to be 7 lb 10 oz at 41 Weeks.

9 Bibliography

References

6. Risks for preterm delivery and low birth weight are independently increased by severity of maternal anaemia Hussein L Kidanto, Ingrid Mogren, Gumilla Lindmark, Siriel Massawe, Lennarth Nystro