Gender Prediction: High or Low Pregnancy?

David D Berberena

Bellevue University

DSC 500 Introduction to Data Science

Nasheb Ismaily

September 10, 2023

Gender Prediction: High or Low Pregnancy?

One of mankind’s most popular “old wives’ tales” surrounding the gender of a baby is that the way a woman’s body carries the fetus determines the sex of the baby. Now this can be interpreted in many ways, whether the woman in question has a wide belly, a “torpedo” belly, a small belly, or a large belly. But the most common indicator of the baby’s sex according to the “old wives’ tale” is whether the woman carries high or low in her pregnancy. WebMD states that mothers who carry high are thought to have girls, whereas mothers who carry low are thought to have boys. (Johnson, 2023) This sentiment is mimicked by society and prompts further investigation as to whether this holds true or not according to the data.

# CRISP-DM Methodology

The CRISP-DM, or Cross Industry Standard Process for Data Mining methodology is one of the most popular processes in data science used for data collection, analysis, and implementation. The six steps involved in this process are as follows: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment. (Hotz, 2023) I will be delving further into this model's second and third steps, which are data understanding and data preparation. These steps consider what it takes to gather data concerning the prompt at hand, which as stated above is the question of whether carrying high or low determines the gender of the baby a woman is carrying.

## Data Understanding

To answer the question of a high pregnancy being a girl and a low pregnancy being a boy, we first need to identify the variables with which we can properly go about gathering data. Establishing controls within this scenario is important, as there is a myriad of ways in which pregnancies can be affected. Normal pregnancies only are to be considered to attain clean data for the question at hand. This means that the babies born are deemed healthy and were born on time (no premature or late births). This also means that the women studied are also healthy and report normal pregnancy symptoms. A substantial number of women who meet these standards are to be considered. Pure speculation leads me to believe that studies with at least 100 women are to be considered credible, as any number below this could lead to insufficient data. The studies needed should include the diagnosis of the pregnancy (high or low) at the same point in every woman’s pregnancy (middle to end of the second trimester). The last indicator of a clean data source is that there is an actual result that stems from the study, meaning that the number of girls and boys being born from these diagnoses of high and low pregnancies is quantified at the end of the study. With all this in mind, data preparation is the next step in the CRISP-DM methodology to conduct data collection and analyze the findings.

### Data Preparation

Once the variables and controls are identified, it is time to consider where and how to collect the data necessary to begin answering this question. Hospital medical data is the most relevant data to begin with, as many pregnancies are documented through doctors via regular check-ups. I would have access to thousands of thoroughly documented pregnancies of women around the world and could begin the cleaning process by applying the previously stated controls and criteria. Google data also could prove useful as there are many scientific studies done on pregnant women that also focus on what elements of pregnancy could potentially affect the gender outcome. Once data is gathered from these two main sources, the data would need to be cleaned by ridding ourselves of irrelevant figures or data obtained without keeping the outlined parameters at the forefront of the research. After data cleaning takes place, we would be able to compile the data so it can be easily read, and the data can be assessed for the answer to our initial question.

Moving Forward: Carrying high or low being a deciding factor in the sex of a woman’s baby prompts many questions, such as the health of the woman, the health of the baby, the weight of the baby once born, and more. Figuring out how to start your research is key to reaching a plausible and credible answer to the query, and the CRISP-DM methodology helps data scientists significantly in this journey. From understanding that controls need to exist in a study such as this to interpreting the data gleaned from hospital records of thousands of normal pregnancies with a corresponding “high” or “low” designation, the thought process behind data collection, analysis, presentation, and implementation is paramount to unlocking the most accurate answer the data can provide.

References

Johnson, T. C., MD (2023, April 23). Pregnancy: What to Know About How You're Carrying. WebMD. Retrieved September 5, 2023, from <https://www.webmd.com/baby/pregnancy-know-how-you-are-carrying>

Hotz, N. (2023, January 19). What is CRISP DM? Data Science Process Alliance. Retrieved September 5, 2023, from https://www.datascience-pm.com/crisp-dm-2/