Data Quality Issues

Session 4

09/07/2024

Data Quality Issues

- Success of ML largely depends on the quality of the data.
- A data which has the right quality helps to achieve better prediction accuracy in case of supervised learning.
- There are multiple factors which lead to the data quality issues.

Data Quality Issues

- Incorrect sample set selection
- Random responding and motivated mis- responding
- Who are the true respondents?
- Errors in data collection
- Missing values
- Extreme values

Incorrect sample set selection

- The data may not reflect normal or regular quality due to incorrect selection of sample set.
- Eg: sales transaction from a festive period to predict for very next future period.
- If you are interested in studying the effects of smoking on a particular outcome, you must define what it means to be a smoker which helps make the research more precise.
- For example do you include people who just smoke occasionally? What about those smokers who previously used tobacco but have stopped?

Can data cleaning fix sampling problems?

- Unfortunately in many cases, poor sampling cannot be corrected by data cleaning.
- The goal of sampling is to gather data on whatever phenomenon is being studied in such a way as to make the best case possible for withdrawing inferences about the population of interest.

Response Set

- A response set is a strategy people use (consciously or otherwise)when responding to educational tests, questionnaires, or things like psychological tests (or even questions paused in casual conversation in the above example).
- Researchers should pay more attention to response sets and effects of random responding which can substantially increase probability of Type II errors.

Common types of Response Sets

- Random Responding
- Malingering and Dissimulation
- Social desirability
- Other response styles

Random Responding, Motivated Misresponding

- Let me ask you “How are you doing today”.
- Ever notice that some people tend to flip between extremes(e.g. wonderful or horrible) while others seem to be more stable (e.g. OK or fine) no matter what is going on. This is an example of one sort of Response set, where individual tend to vary in a narrow band around the average or vary around extremes.

Common types of Response Sets

- Random responding : Random responding is a response set in which individuals respond with little pattern or thought. The behaviour adds substantial error variance to analysis which completely make ineffective the usefulness of responses.
- This may be motivated by lack of preparation, reactivity to observation, lack of motivation to cooperate with testing or disinterest.
- If we are not careful, participants with lower motivation to perform at their maximum level may increase the odds of Type II errors masking real effects of our research through response sets such as random responding.

- Malingering is a response set where individuals falsify and exaggerate answers to appear weaker or more medically or psychologically symptomatic, often motivated by the goal of receiving services they would not otherwise be entitled to.
- Someone might pretend to be injured so they can collect an insurance settlement or obtain prescription medication.
- Dissimulation refers to a response set in which respondents falsify answers in an attempt to be seen in more negative or more positive side than honest answers can provide.
- These response sets are common in psychological test like “Do you have suicidal thought?”

- Social Desirability : Social desirability is related to malingering and dissimulation, in that it involves altering responses in systematic ways to achieve a desired goal. In this case to behave in the same way as most other people in a group or society, the respondent want to “look good” to the examiner.
- Socioeconomic status, religion, charitable acts, personal habits around smoking, drinking, drug
- For example, if a survey is sent from an organization that supports animal rights (such as PETA: People for the Ethical Treatment of Animals), people may underestimate the amount of meat they eat or express an overly negative attitude toward buying fur coats.
- Other response styles : Other response styles such as acquiescence and criticality are response patterns wherein individuals are more likely to agree with (acquiescence) or disagree with (criticality) questionnaire items in general, regardless of the nature of the item.

Detecting random responding in your research

- An important issue is whether we can be confident that what we call random responding truly is random?
- There is a large and well developed literature on how to detect many different types of response sets.
- Examples include addition of particular types of items to detect social desirability , altering instructions to respondents in particular ways, creating equally desirable items worded positively and negatively, use item response theory (ITR) to explicitly estimate a guessing parameter.

Detecting random responding in your research

- For example, I unknowingly say things which disturb my students.
- 4:Strongly Agree 3: Agree 2:Neutral 1:disagree 0: Strongly disagree
- Suppose we choose 4 and the respondent want to “look good” to the examiner.
- When the respondent score this question, he reverse the option.

Original Response	0	1	2	3	4
Reversed Response	4	3	2	1	0

Reference

Random Responding from Participants is a Threat to the Validity of Social Science Research Results

Jason W. Osborne and Margaret R. Blanchard