Prepare for Class 5
09 / 27 /22

Big Data Pitfalls

There are many big data pitfalls. Below is a listing of some of those pitfalls:

1. Lack of a clear goal for the analysis. Many times data scientist begin collecting mass data but lack an clear research question.
2. Focusing of own data. Analyst often to do not explore outside data such as from Social media, web content and other supplemental data.
3. Not complying with privacy regulations. Data must be treated with respect. Voliation of privacy regulations can lead to serious consequences.
4. Failure to consider confirmation bias. Often there is a tendency to find the results that support the answers the researcher is looking for rather than trying to reject a hypothesis.
5. Analyzing too small of a sample size. Small sample sizes make it difficult to determine whether a variable is a statistical outlier or not.
6. Failure to verify to external validation. Model success can only be concluded when the model is compared to data not used during model building.
7. Not accounting fo “Garbage in, garbage out” (GIGO). The validity of a finding is strongly dependenbt on the quality of the data learned from. Confounding biases may lead to surprising associations but those results might be spurius.

Overfitting and Overparameterization

Most often analyst are attempting to obtain a low enough “p-value” so as to “prove” their model correct. Often, however, analyst fail to ask whether their model is “too good.” Overfiting and overparameterization can result in models that are “too good.” Overfitting is where a model uses too many parameters relative to the sample size leading to a good fit with the sample data but a poor fit to new data. Overparameterization is simply using an excessive number of parameters. Overparameterization leads to the “dimensionality curse.” That is the more variables an analyst combine in their model, the higher the chances of a spurious positive finding.

References

“Overfit vs Overparameterize - What's the Difference?” 2017. WikiDiff. https://wikidiff.com/overparameterize/overfit (September 22, 2022).

Lamata, Pablo. 2020. “Avoiding Big Data Pitfalls.” Heart and metabolism : management of the coronary patient. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7610672/ (September 22, 2022).

Linden, Rick Van der. 2022. “Big Data: 7 Pitfalls: Strategies for Big Data Success.” Big Data | 7 Pitfalls | Strategies for Big Data Success. https://www.passionned.com/the-7-biggest-big-data-pitfalls/ (September 22, 2022).