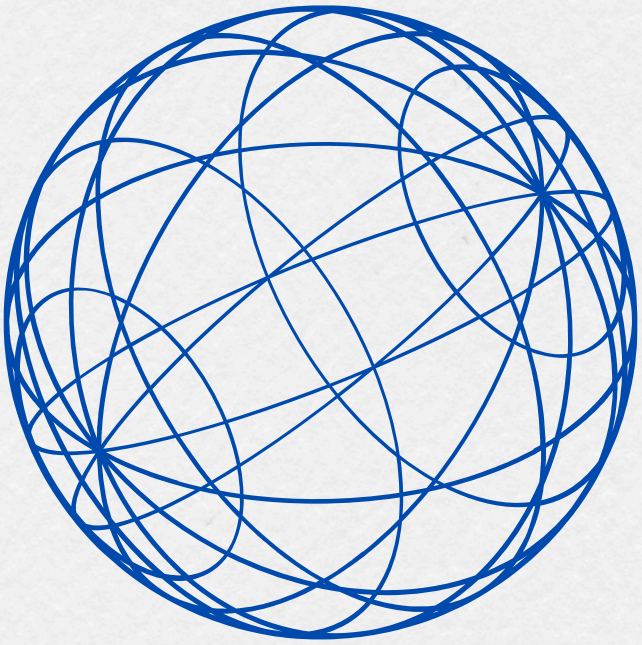
Upen

work of

FEATURE ENGINEERING REFLECTION ASSIGNMENT

Angel gabriel zuñiga



Presentado por: zuñiga

Objective

Develop an intuitive understanding of how to prepare and improve data (features) for machine learning, using the Titanic dataset as a real-world example.

Context

The Titanic dataset contains information about passengers and whether they survived. Your task is to think critically about the features (columns), how they affect prediction, and how they could be transformed or improved.

Instructions

Answer all reflection questions using your own reasoning. Your answers must be based on personal analysis.

Part 1: Understanding the Problem

1. What is the goal of the Titanic dataset?
    a. The goal is to have a collection of data on the Titanic's passengers to predict whether a passenger survived or not based on various attributes (features) such as age, class, gender, and fare.
2. Why is this prediction (survival) important or interesting?
    a. It is interesting because it combines human and societal factors (like class and gender) with data to predict life-or-death outcomes. It's also historically significant and a classic example to learn how machine learning can extract patterns from real-world scenarios.
3. What kind of real-life decisions could a similar model support?
    a. A similar model could prioritize rescue operations in emergencies, assess risk in transportation systems, or assist in health care triage decisions based on patient data.

Part 2: Thinking About the Features

4. Five selected features: Age, Sex, Pclass, Fare, Embarked

5. For each one:

- · Age
  - · Represents: The age of the passenger.
  - · Type: Numerical (continuous).
- · Sex
  - · Represents: Gender of the passenger.
  - · Type: Categorical.
- · Pclass
  - · Represents: Passenger class (1st, 2nd, 3rd).
  - · Type: Categorical (ordinal).
- · Fare
  - · Represents: How much the passenger paid.
  - · Type: Numerical (continuous).
- · Embarked
  - · Represents: Port of embarkation
  - · Type: Categorical.

6. Which feature do you think is most useful for predicting survival? Why?
Feature: Sex
Reason: The Sex feature is highly predictive because historical data from the Titanic shows that women were much more likely to survive than men, due to the "women and children first" evacuation policy. When analyzing the dataset, survival rates for females are significantly higher than for males, making it a strong indicator of survival.

7. Which feature might be least useful or confusing? Why?
Feature: Ticket
Reason: The Ticket feature contains ticket numbers, which are a mix of letters and digits with no clear pattern or consistent structure. It’s hard to extract meaningful information from it directly.

8. New feature idea: PortAndClass = Embarked + Pclass. Combines embarkation port with class, e.g., "S3", "C1", which may reflect socioeconomic patterns or groupings.

Part 3: Improving the Dataset

- 9. Feature selected: Age

How would you improve or transform it?

- Create categorical bins like: child (0–12), teen (13–19), adult (20–59), senior (60+).

10. What might happen if we keep irrelevant or misleading features?

- They may introduce noise, reduce model accuracy, and lead to incorrect conclusions. The model could overfit or learn false correlations.

11. How could missing or inaccurate data affect the model's predictions?

- Missing or inaccurate data can bias the model, cause it to ignore important patterns, or give disproportionate weight to unreliable features.

Part 4: Ethical Thinking

Could any feature in the Titanic dataset introduce bias or unfairness in the model? Explain your reasoning.

Yes. Features like Sex and Pclass can reflect societal biases. For example, women and higher-class passengers had a higher chance of survival because of social norms and access to lifeboats. A model trained on such data might incorrectly assume that these groups are more deserving or more likely to survive in general, which could be unfair if used in other contexts.