Data Exploration, Visualization and Feature Engineering

datasciencedojo
unleash the data scientist in you

Data Beats Algorithm but…

- More data will yield good generalization performance – even with a simple algorithm
- But there are caveats
  - Amount of data may have diminishing returns
  - Data quality and variety matters
  - A decent performing learning algorithm is still needed
  - Most importantly, extracting useful features out of data is important

Dispelling a Common Myth

There is no algorithm that would take raw data and give you actionable insights

Janitorial Work is Important

Not spending time on understanding your data is a common source all sorts of problems!

Objectives of The Session

- Training you to be a good data science janitor
- High level thinking process of exploring and visualizing a data set before building a model
- How to summarize your findings

Agenda

Data exploration and visualization using R

Some graphics packages

Azure ML studio visualization and exploration capabilities

- Building and evaluating predictive models in Azure ML Studio

A Lot of Material to Cover…

Don’t worry about syntax, just try to understand the process. You can look up syntax any time