



Inspire...Educate...Transform.

Introduction to Data processing in R and Python

Jagan Basa

Lead Data Scientist, INSOF

We will learn

- What is Data science & pre-processing
- Data science Process
- Skills required and mapping the skills
- Key steps in the Data processing

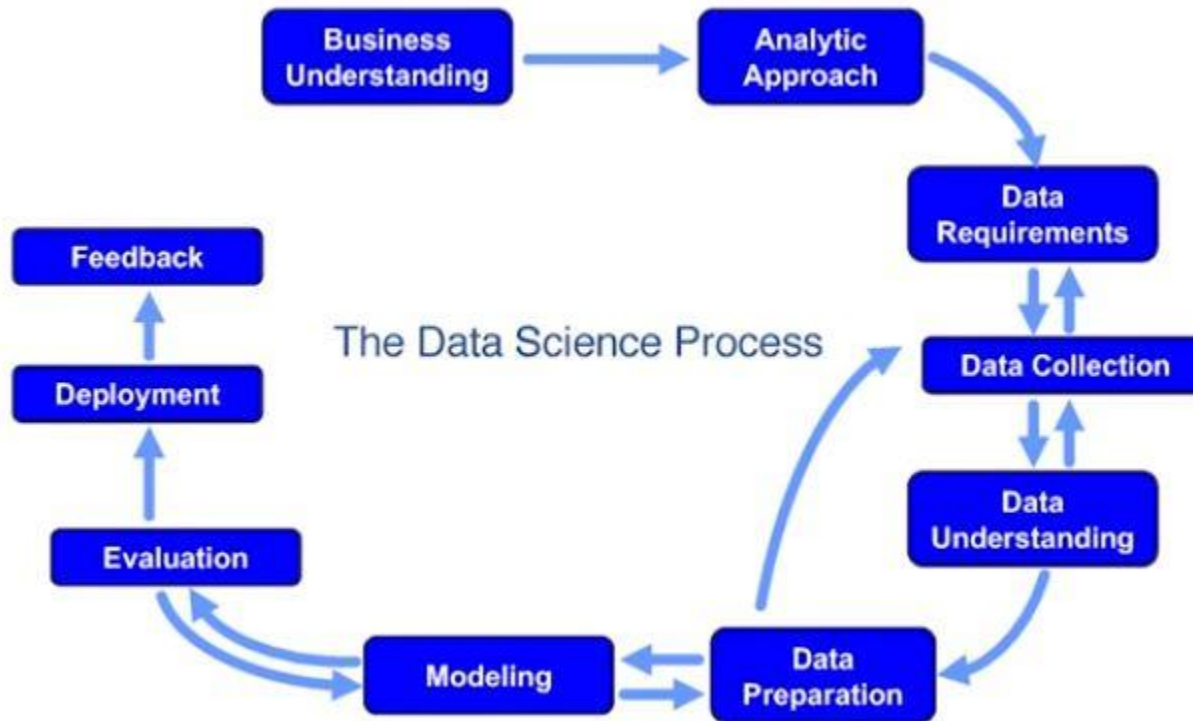


What is Data Science

Data Science in a Nutshell



The Data Science Process



4

The Data Science Skills

The “Perfect” Hire



Can't hire one person for each role?

- Subject matter expertise & industry experience
- Caught up with the latest tools & technologies
- Experience with big data & big data infrastructure
- Experience with machine learning and building pipelines
- Experience with app development
- Also happy to do simple ad-hoc reports



The Data Science Skill Map

Mapping Roles in a Data Science Team



Data Analyst (Jr. Data Scientist, BI analysts)

Spreadsheets, statistics, ad-hoc reports, visualizations
Small data



Data Scientist

Machine learning & statistics, big data
Can come up with new algorithms to solve a problem



App Developer

Writes code to build apps -- using data & models
Use existing blackbox APIs for machine learning/data science



Data Engineers

Architects how data is organized & ensures operability
Infrastructure around data storage, management, cleaning
ETL pipeline, data warehousing

5



Mapping Roles

The diagram illustrates the mapping of roles to the stages of the CRISP-DM process. The roles are represented by colored icons: DS (Data Scientist, teal), DE (Data Engineer, orange), App (Application Developer, blue), and Biz (Business Analyst, green). The process stages are represented by blue boxes: Business Understanding, Analytic Approach, Data Requirements, Data Collection, Data Understanding, Data Preparation, Modeling, Evaluation, Deployment, and Feedback. The flow of the process is indicated by blue arrows, with a red curved arrow highlighting the feedback loop from Evaluation back to Business Understanding. The roles are mapped to the stages as follows: Biz is involved in Business Understanding; DS is involved in Analytic Approach; DE and DS are involved in Data Requirements; DE is involved in Data Collection; DE and DS are involved in Data Understanding; DS, DE, and App are involved in Data Preparation; DS is involved in Modeling; DS is involved in Evaluation; DS and App are involved in Deployment; and DS and Biz are involved in Feedback.

Why pre-process

- Poor model on good data is likely to be better than great model on poor data



Raw data

- Normally it is available in multiple tables
 - Merge them
 - Fill missing values



More on attributes

- Type
 - Numeric, Categorical and Ordinal
- Actionable
 - Focus, changeable



Process

- Take a subset from a table if needed
- Typeset the attributes correctly
- Do descriptive statistics and understand the data



Missing values

- Ignore
- Fill with a central statistic



KNN Imputation

- Find nearest neighbors based on existing attributes

[illegible]

Euclidian for numeric
and 1 or zero for
categorical

- Take an average of only the nearest neighbors
 - Mean for numeric
 - Mode for categorical



Data standardization

- Let us say, we are measuring distance between records for some purpose

Employee	Age	Income
1	24	50000
2	25	55000
3	60	51000

This dominates completely



Bring to same range

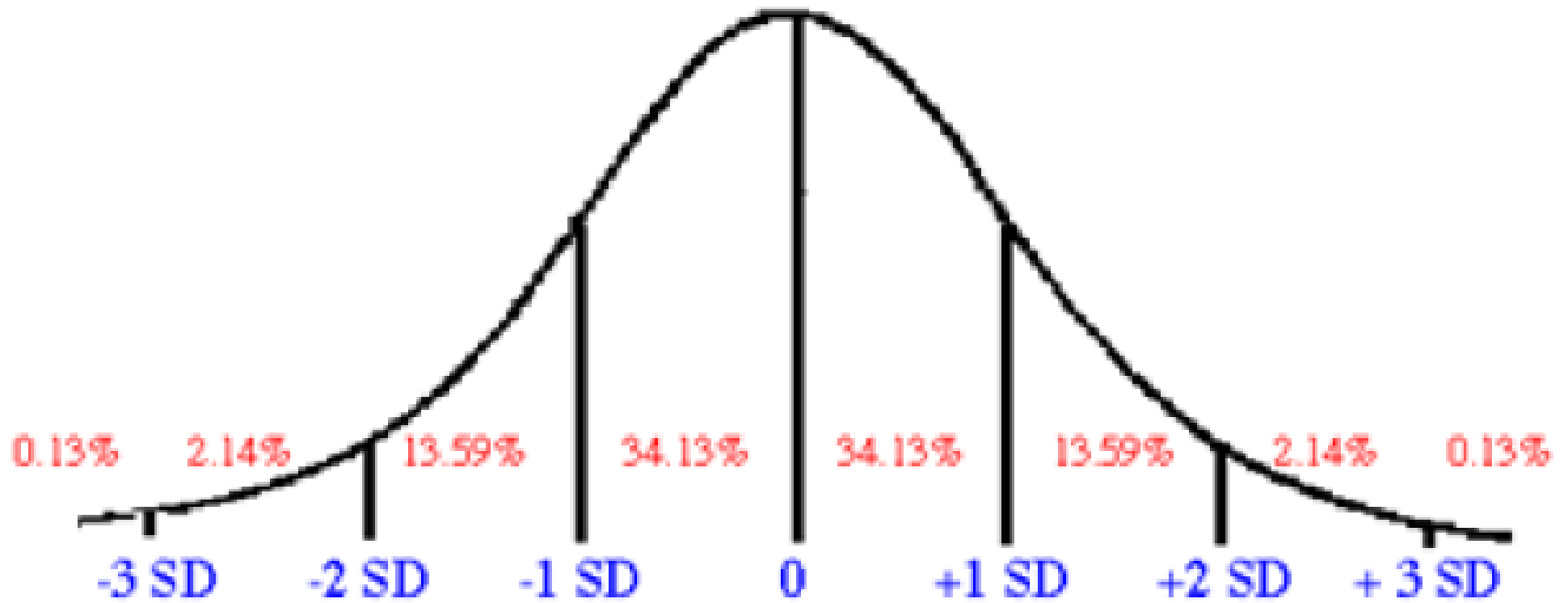
$$Value_{new} = \frac{Value - minValue}{maxValue - minValue} \quad \text{Range is 0 to 1}$$

$$\text{Min max for 25: } \frac{25 - 24}{60 - 24}$$



- Min-Max is extremely sensitive to the outliers.
- Min-max of : (1, 2, 1001) is (0, 0.001, 1)





$$\text{newValue} = \frac{\text{Value} - \text{mean}}{\text{standard deviation}}$$



Changing the type

- Neural networks
 - All attributes must be numeric
- Naive Bayes
 - All attributes must be categorical



Numeric to categorical

- Manual
- Equal frequency
 - Number of samples in each bin
- Equal width
 - Interval is same (good for uniform distributions)



Ordered and categorical

- Merging multiple bins
 - Verify the frequencies
 - Convert them to numeric and recode



Categorical to numeric

- How do we set up categorical variables in distance metrics
 - Create as many dummy variables as there are options
 - Code as 100, 010,...







HYDERABAD

Office and Classrooms

Plot 63/A, Floors 1&2, Road # 13, Film Nagar,
Jubilee Hills, Hyderabad - 500 033
+91-9701685511 (Individuals)
+91-9618483483 (Corporates)

Social Media

Web: <http://www.insofe.edu.in>
Facebook: <https://www.facebook.com/insofe>
Twitter: <https://twitter.com/Insofeedu>
YouTube: <http://www.youtube.com/InsofeVideos>
SlideShare: <http://www.slideshare.net/INSOFE>
LinkedIn: <http://www.linkedin.com/company/international-school-of-engineering>

This presentation may contain references to findings of various reports available in the public domain. INSOFE makes no representation as to their accuracy or that the organization subscribes to those findings.

BENGALURU

Office

Incubex, #728, Grace Platina, 4th Floor, CMH Road,
Indira Nagar, 1st Stage, Bengaluru – 560038
+91-9502334561 (Individuals)
+91-9502799088 (Corporates)

Classroom

KnowledgeHut Solutions Pvt. Ltd., Reliable Plaza,
Jakkasandra Main Road, Teacher's Colony, 14th Main
Road, Sector – 5, HSR Layout, Bengaluru - 560102

