

Data Preprocessing

This file is meant for personal use by gelson.diaz@austin.utexas.edu only.

Sharing or publishing the contents in part or full is liable for legal action.

Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.

Agenda

1. Data Preprocessing
2. Steps involved in model building

This file is meant for personal use by gelson.diaz@austin.utexas.edu only.

Sharing or publishing the contents in part or full is liable for legal action.

Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.

Questions to discuss

1. What is Data Preprocessing and what are the steps involved in it?
2. Why we need data preprocessing?
3. What are the steps involved in model building?

This file is meant for personal use by gelson.diaz@austin.utexas.edu only.

Sharing or publishing the contents in part or full is liable for legal action.

Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.

What is Data Preprocessing?

- Data preprocessing is a collective term used to describe a collection of approaches that help get the data ready for analysis
- It helps us clean and transform the raw data to an efficient format for analysis and modeling
- It is generally very contextual and requires good domain understanding to do this well
- Different datasets need different kinds of preprocessing

This file is meant for personal use by gelson.diaz@austin.utexas.edu only.

Sharing or publishing the contents in part or full is liable for legal action.

Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.

What are the steps involved in Data Preprocessing?

Broadly it is divided into the following steps:

1. Data format checks
 - a. Data dimension
 - b. Data types
2. Data Consistency
 - a. Missing, inconsistent, duplicate values
 - b. Outliers
 - c. Data distribution and skewness
3. Feature Engineering
 - a. Variable transformations
 - b. Feature extraction

This file is meant for personal use by gelson.diaz@austin.utexas.edu only.

Sharing or publishing the contents in part or full is liable for legal action.

Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.

Why do we need Data Preprocessing?

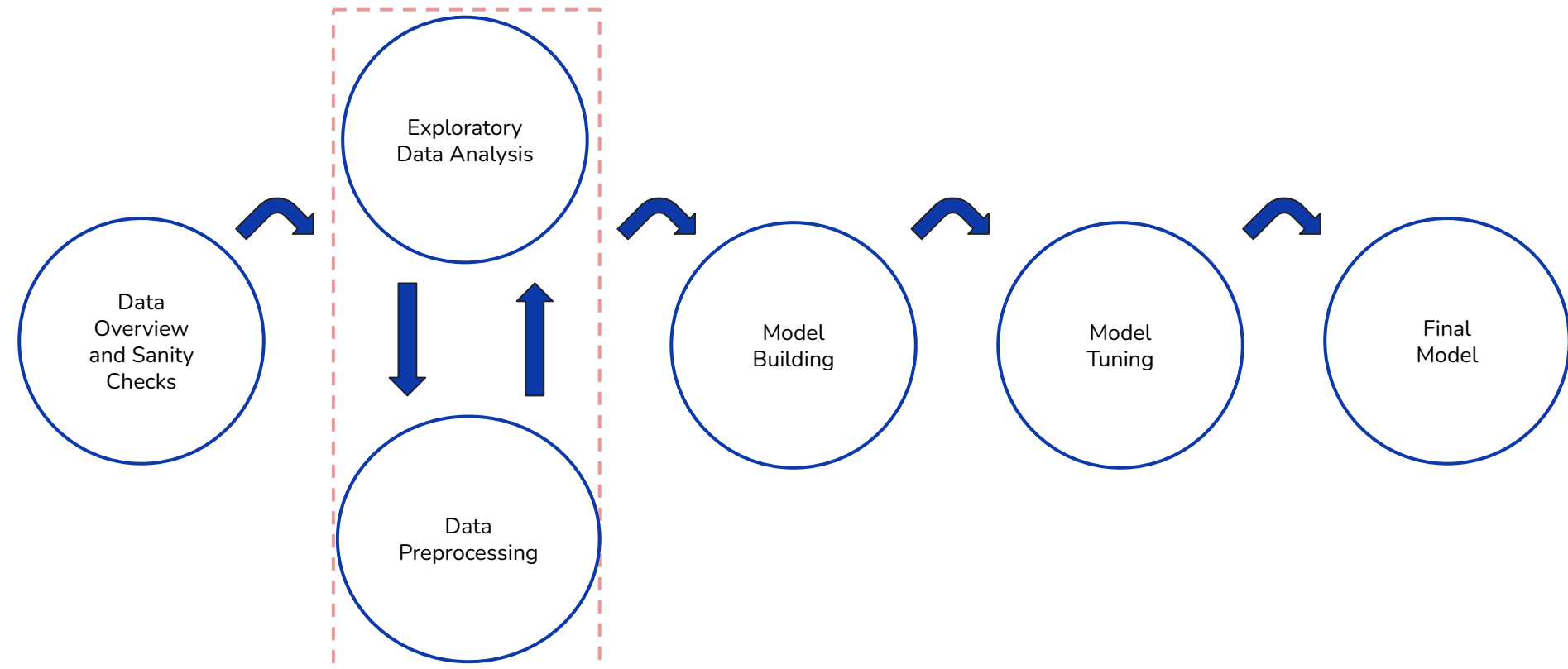
- Data preprocessing is a crucial step in the cycle of building a model from raw data
- Data preprocessing takes up approximately 60-80% of the time in a modeling project
- We often need to iterate between exploratory data analysis and data preprocessing to obtain the optimal data to get the desired model performance
- Data preprocessing also helps us in case our model does not satisfy any underlying statistical assumptions

This file is meant for personal use by gelson.diaz@austin.utexas.edu only.

Sharing or publishing the contents in part or full is liable for legal action.

Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.

What are the steps involved in model building?



This file is meant for personal use by gelson.diaz@austin.utexas.edu only.

Sharing or publishing the contents in part or full is liable for legal action.

Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.



Happy Learning !

