Data Preprocessing

Agenda

1. Data Preprocessing
2. Steps involved in model building

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?

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

What are the steps involved in Data Preprocessing?

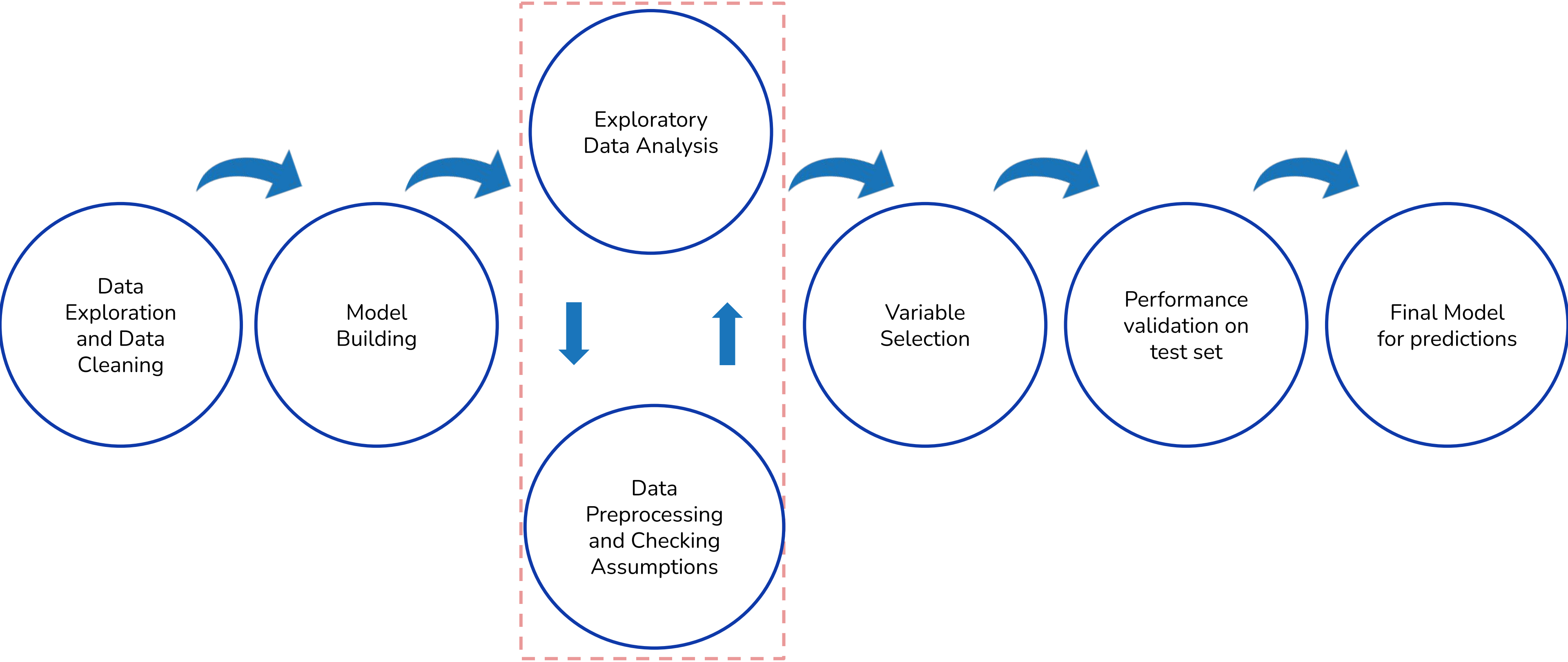
Broadly it is divided into the following steps:

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

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

What are the steps involved in model building?


Data Exploration and Data Cleaning
Model Building
Exploratory Data Analysis
Data Preprocessing and Checking Assumptions
Variable Selection
Performance validation on test set
Final Model for predictions

greatlearning
Power Ahead

Happy Learning !