SNS COLLEGE OF TECHNOLOGY

An Autonomous Institution
Coimbatore-35

DEPARTMENT OF ARTIFICIAL INTELLIGENCE & DATA SCIENCE

23ADT202 – FUNDAMENTALS OF DATA SCIENCE AND ANALYTICS

II YEAR IV SEM

UNIT I – INTRODUCTION TO DATA SCIENCE

BUILD THE MODEL

BUILD THE MODEL

To build the model, data should be clean and understand the content properly.

BUILD THE MODEL

EMPATHY:

• To build the model, data should be clean and understand the content properly. The components of model building are as follows:

a) Selection of model and variable

b) Execution of model

c) Model diagnostic and model comparison

• Building a model is an iterative process. Most models consist of the following main steps:

1. Selection of a modeling technique and variables to enter in the model

2. Execution of the model

3. Diagnosis and model comparison

Model and Variable Selection

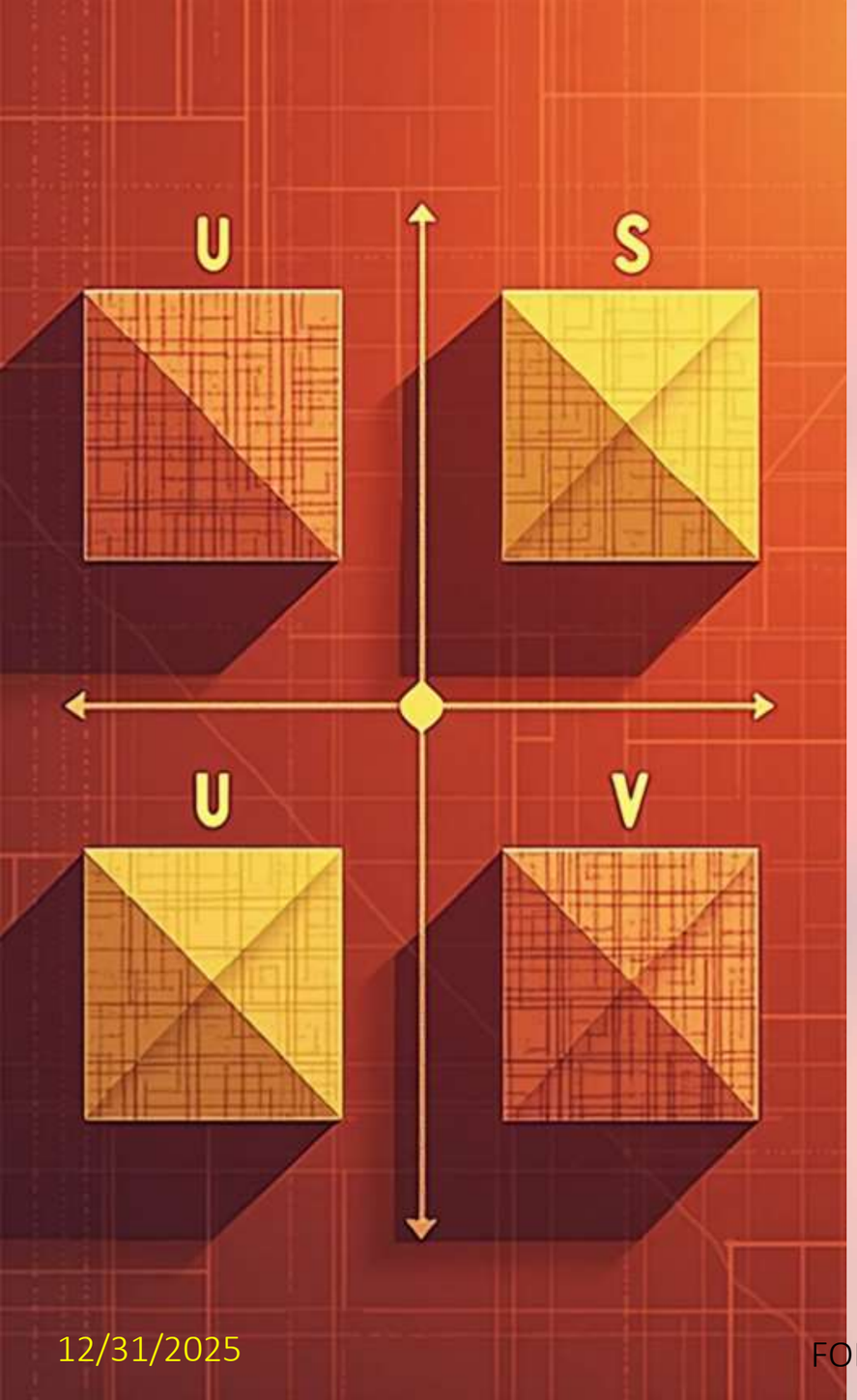
DEFINE:

1. Must the model be moved to a production environment and, if so, would it be easy to implement?
2. How difficult is the maintenance on the model: how long will it remain relevantif left untouched?
3. Does the model need to be easy to explain?

Model Execution

• Various programming language is used for implementing the model. For model execution, Python provides libraries like StatsModels or Scikit-learn. These packages use several of the most popular techniques.

• Coding a model is a nontrivial task in most cases, so having these libraries available can speed up the process. Following are the remarks on output:



a) Model fit: R-squared or adjusted R-squared is used.

b) Predictor variables have a coefficient: For a linear model this is easy to interpret.

c) Predictor significance: Coefficients are great, but sometimes not enough evidence exists to show that the influence is there.

- Linear regression works if we want to predict a value, but for classify something, classification models are used.

- The k-nearest neighbors method is one of the best method.

IDEATE:

Following commercial tools are used :

1. SAS enterprise miner: This tool allows users to run predictive and descriptive models based on large volumes of data from across the enterprise.

2. SPSS modeler: It offers methods to explore and analyze data through a GUI.

3. Matlab: Provides a high-level language for performing a variety of data analytics, algorithms and data exploration.

4. Alpine miner: This tool provides a GUI front end for users to develop analytic workflows and interact with Big Data tools and platforms on the back end.

Open Source tools

TESTING:

1. R and PL/R: PL/R is a procedural language for PostgreSQL with R.

2. Octave: A free software programming language for computational modeling, has some of the functionality of Matlab.

3. WEKA: It is a free data mining software package with an analytic workbench. The functions created in WEKA can be executed within Java code.

4. Python is a programming language that provides toolkits for machine learning and analysis.

5. SQL in-database implementations, such as MADlib provide an alterative to in memory desktop analytical tools.

Model Diagnostics and Model Comparison

Try to build multiple model and then select best one based on multiple criteria. Working with a holdout sample helps user pick the best-performing model.

- In Holdout Method, the data is split into two different datasets labeled as a training and a testing dataset.
- This can be a 60/40 or 70/30 or 80/20 split. This technique is called the hold-out validation technique.
- Suppose we have a database with house prices as the dependent variable and two independent variables
- showing the square footage of the house and the number of rooms. Now, imagine this dataset has 30 rows.
- The whole idea is that you build a model that can predict house prices accurately.

PROTOTYPING:

• To 'train' our model or see how well it performs, we randomly subset 20 of those rows and fit the model. The second step is to predict the values of those 10 rows that we excluded and measure how well our predictions were.

• As a rule of thumb, experts suggest randomly sample 80% of the data into the training set and 20% into the test set.

• The holdout method has two basic drawbacks :

1. It requires an extra dataset.

2. It is a single train-and-test experiment; the holdout estimate of error rate will be misleading if we happen to get an "unfortunate" split.