Introduction to Data Science

Introduction to Data Science Workflow

Session 1

From Introduction to Analysis

- Define the Business Goal
- Collect and manage data
  - Read the data
  - Pre-Processing
  - Data Visualization
- Demo – Churn analysis (part 1)
- Build the model – Introduction
  - Machine Learning;
  - Supervised and Unsupervised Learning;
  - Introduction to models (Classification, Regression, Cluster and Association)

Session 2

Deep dive into Analysis

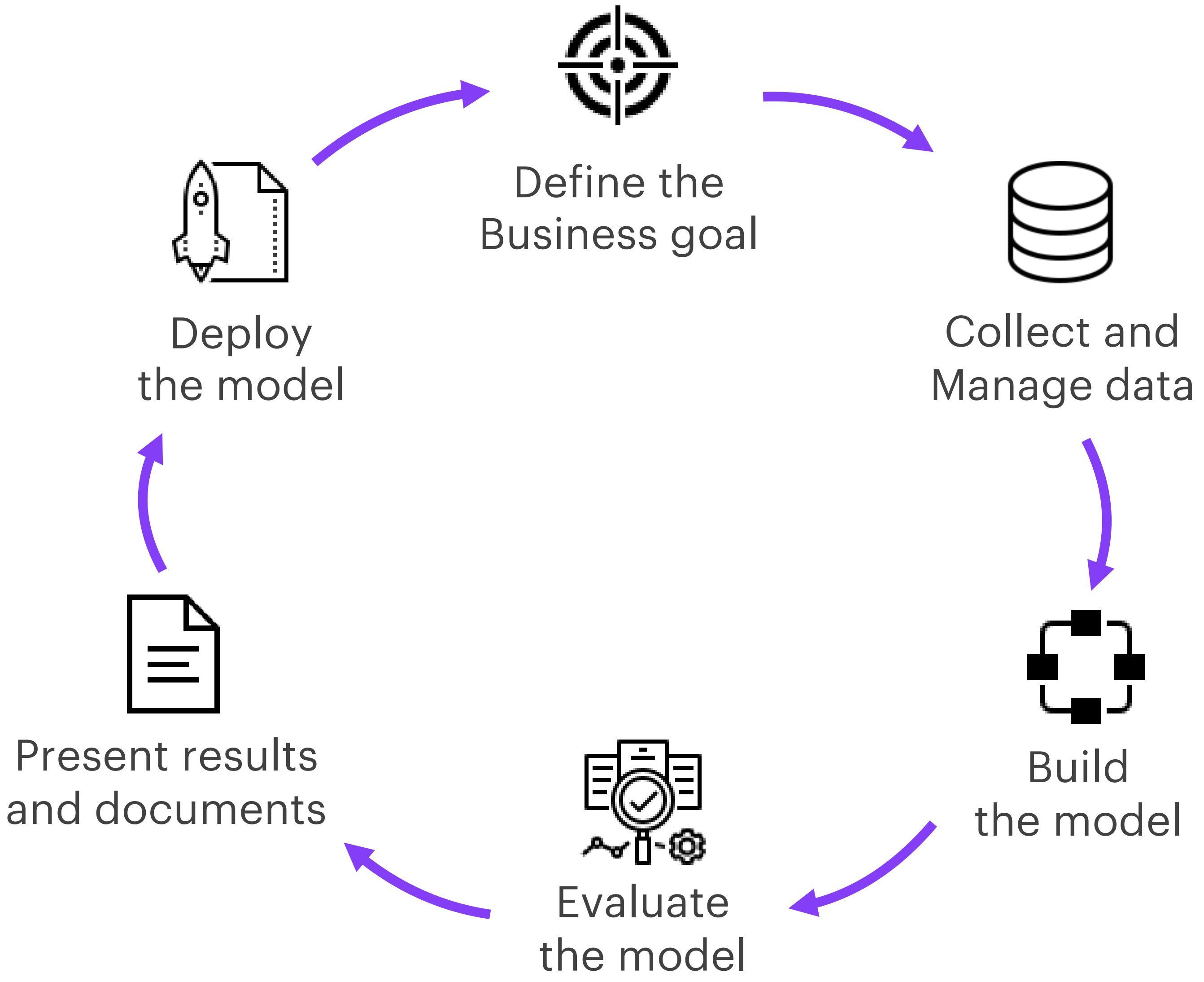
- Build the model – Deep dive
  - Training and test data
  - Linear Regression model
  - Classification models (Logistic Regression, Decision Tree and Random Forest)
  - Clustering models (k-Means, hierarchical clustering)
- Evaluate the model
- Present results and documents
- Demo – Churn analysis (part 2)
- Demo – Customer segmentation

Session 2

Introduction to Data Science Workflow -

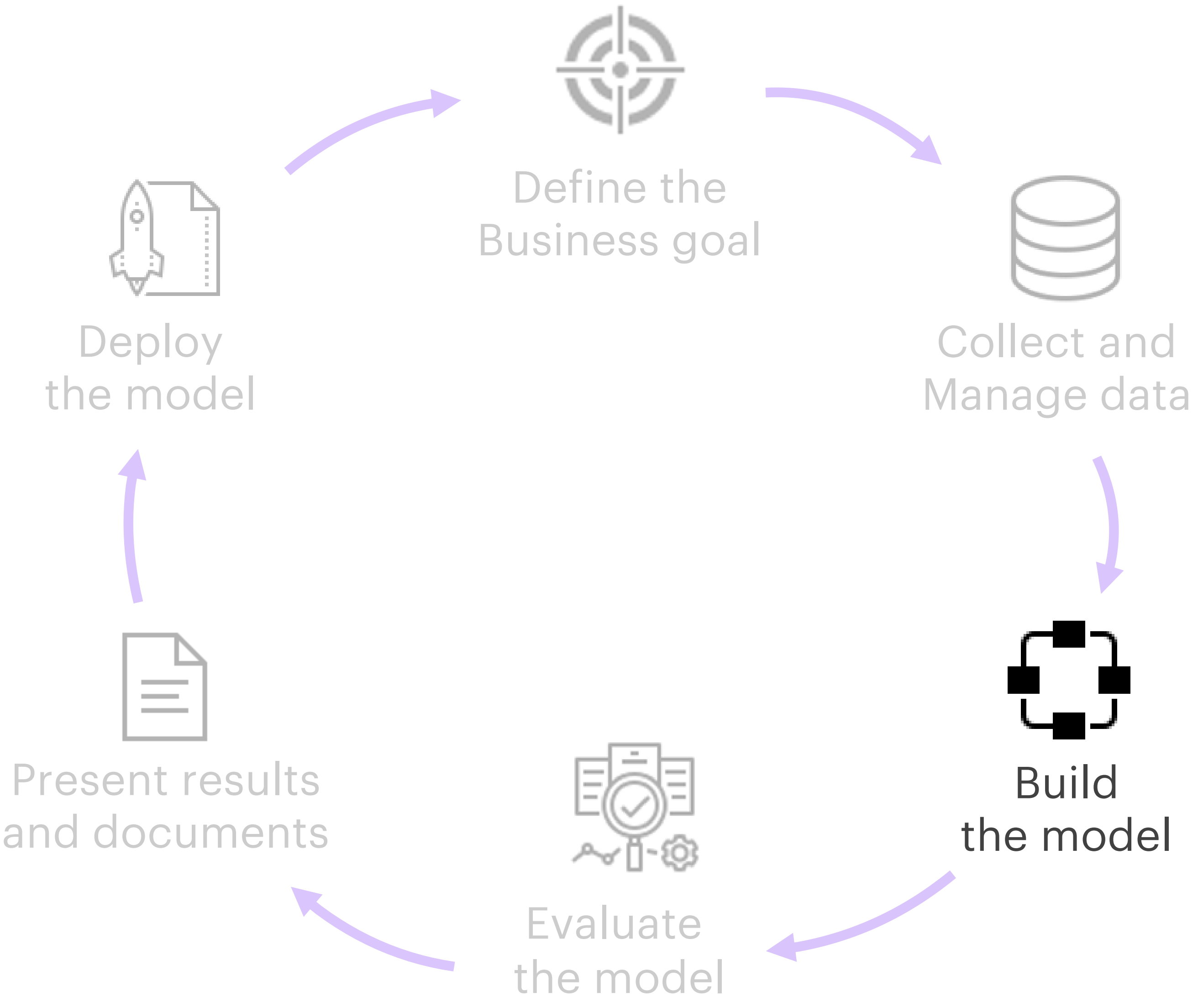
Deep Dive into Analysis

The Data Science workflow



Session 2

Build the model – Deep dive



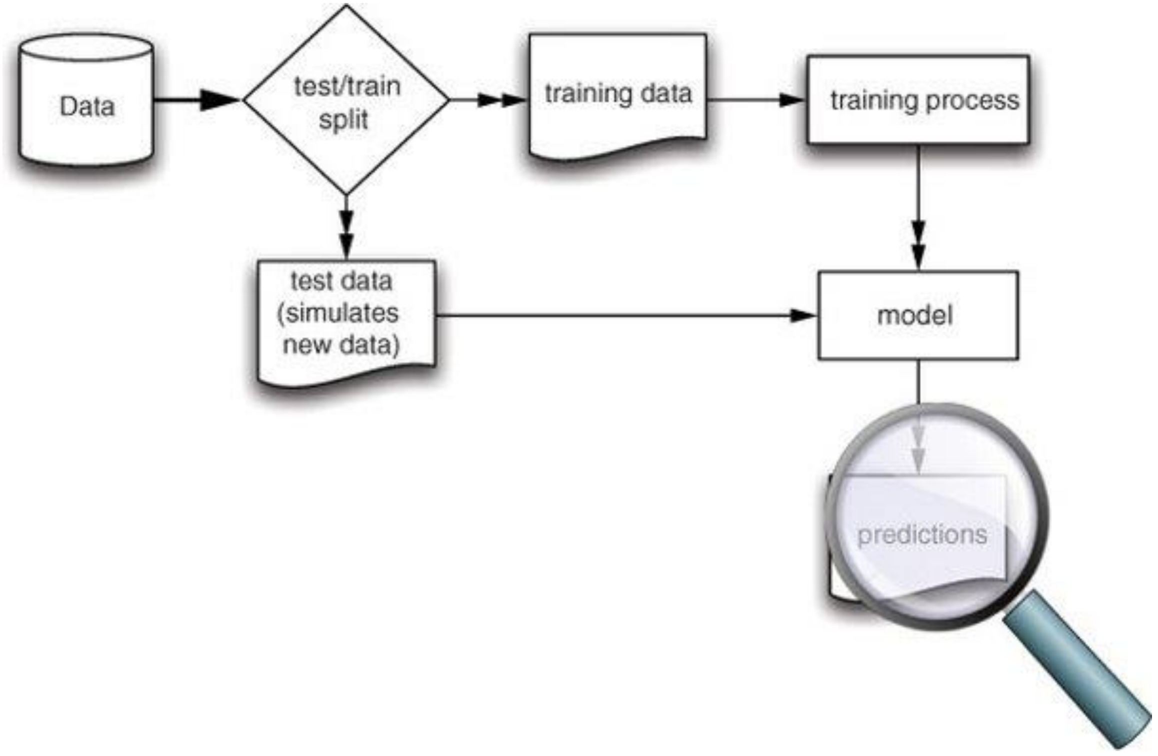
Back to slide

Organizing data for Modeling Process

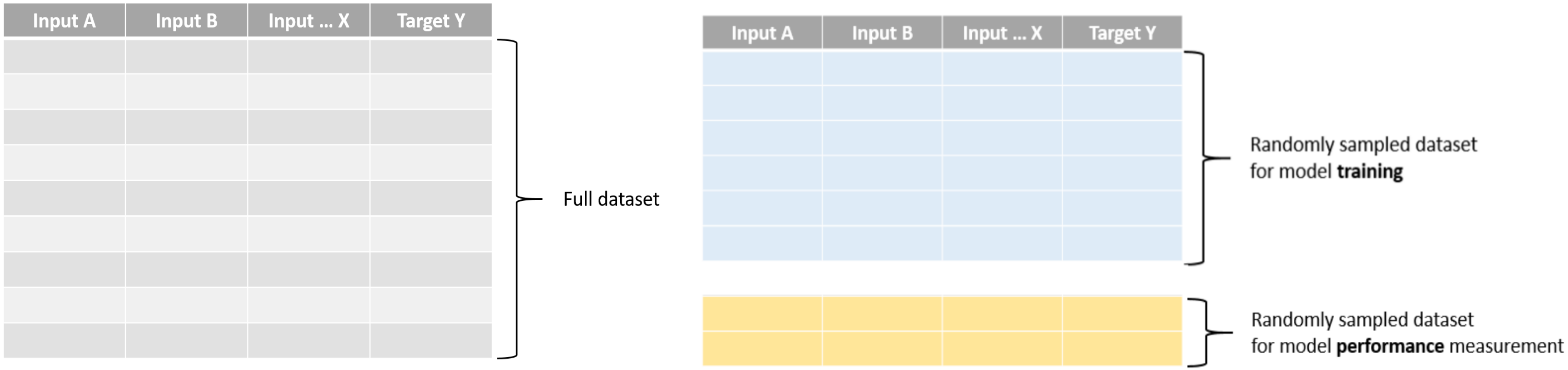
Training & Test data

In order to evaluate the future performance of the model, the data is divided into:

- **Training set** — a subset to build the model.
- **Test set** — a subset to test the trained model and gives an indication of the performance with new data.



Training & Test Data

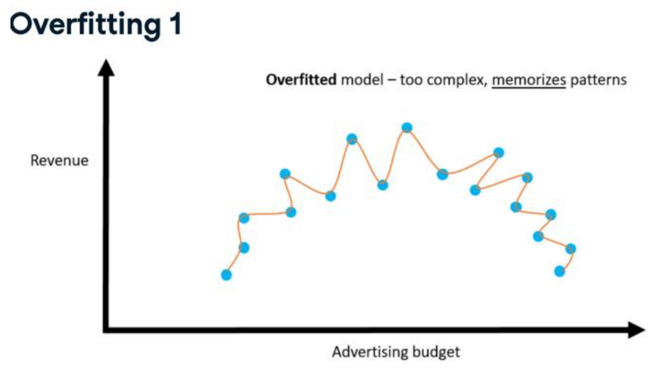


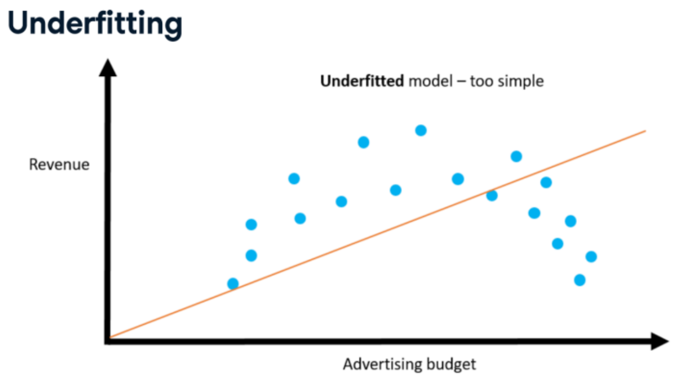
Assumption on Test set:

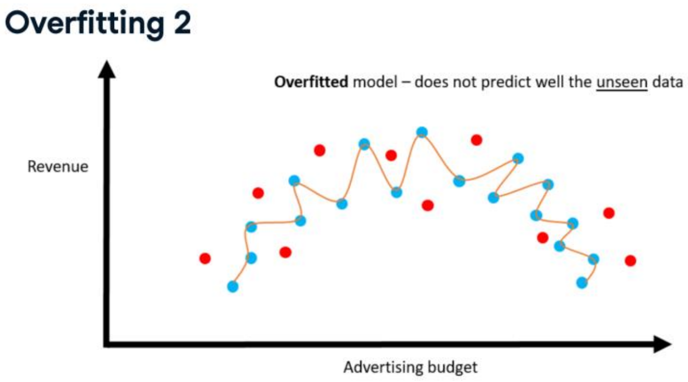
- Is large enough to have statistically meaningful results.
- Is representative of the data – No Test set with different characteristics than the training set.

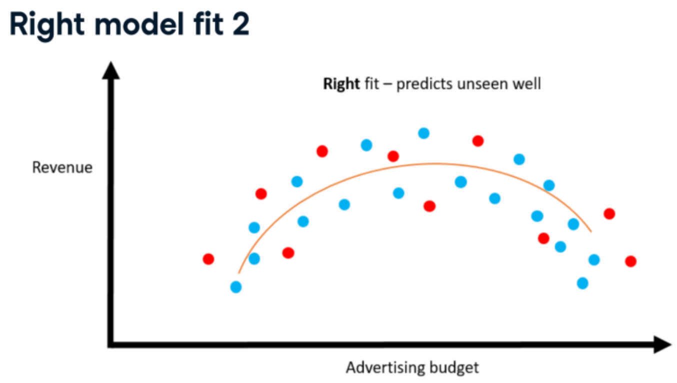
Training & Test Data

- **Overfitting**: when the model fits exactly against its training data. The algorithm cannot perform accurately against unseen data.
- **Underfitting:** when the model is unable to capture the relationship between the input and output variables, generating a high error rate on both the training set and unseen data. It occurs when a model is too simple.









Mapping Business Problems to a good Machine Learning

As a Data Scientist there are a lot of business problems that your team might be called on to address. For example:

- Predicting what customers might buy, based on past transactions;
- Identifying fraudulent transactions;
- Grouping customers with similar behaviour (segmentation);
- Evaluation of campaigns;
- How much the company should spend to buy certain Adwords on search engines;

Mapping Business Problems to a good machine Learning

All these different kinds of suggest a different statistical approach to try and they are generally grouped in three categories

CLASSIFICATION

- **GOAL: Assigning a label** to the data.
- **EXAMPLE**: **classification of products,** based on attributes and/or text descriptions of the products.

CLASSIFICATION MODELS: SUPERVISED LEARNING MODELS

CLUSTERING

- **GOAL: Discovering patterns** in the data.
- **EXAMPLE**: **identifying groups of customers** with the same buying patterns.

CLUSTERING MODELS: UNSUPERVISED LEARNING MODELS

SCORING

- **GOAL: Assigning numerical values**.
- **EXAMPLE**: **predicting the increase in sales after a marketing campaign**.

SCORING MODELS: SUPERVISED LEARNING MODELS

Build the model – Deep Dive

DEMO TOPICS

Scoring	Classification			Clustering	
Linear Regression	Logistic Regression	Decision Tree	Random Forest	K-Means	Hierarchical

Build the model – Deep Dive

DEMO TOPICS

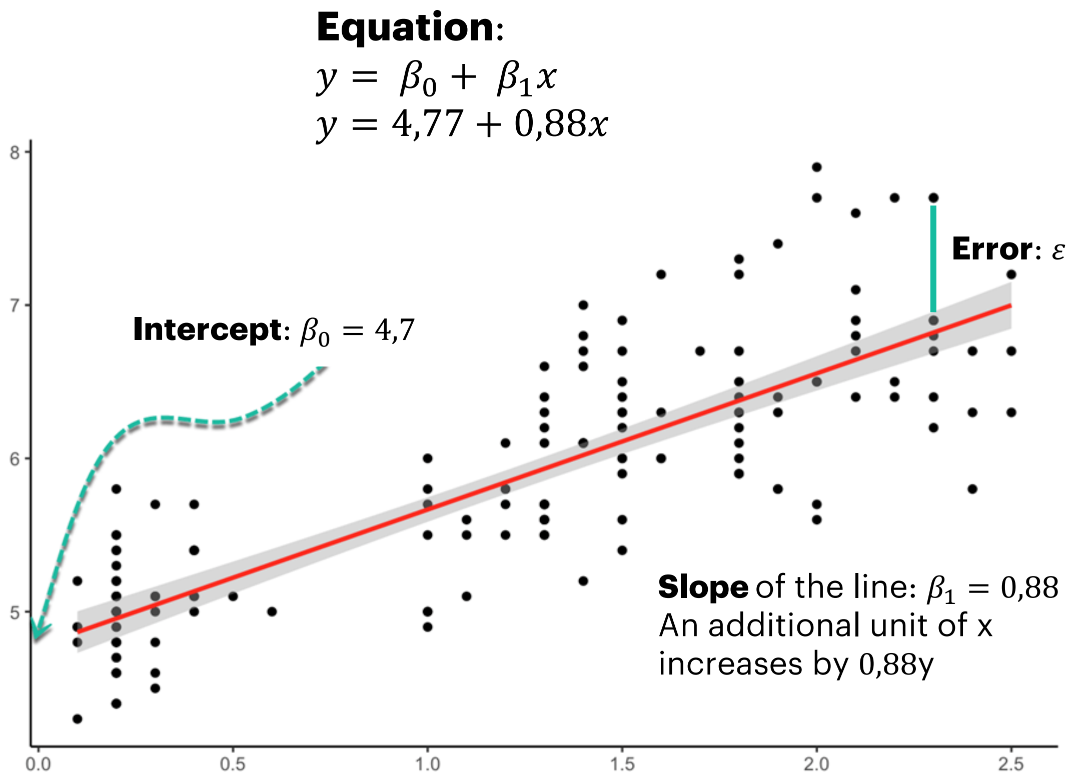
Scoring	Classification			Clustering	
Linear Regression	Logistic Regression	Decision Tree	Random Forest	K-Means	Hierarchical

Linear Regression

Linear Regression describes the relationship between quantitative variables by fitting a **line** to the observed data.

$$y = \beta_0 + \beta_1 X + \varepsilon$$

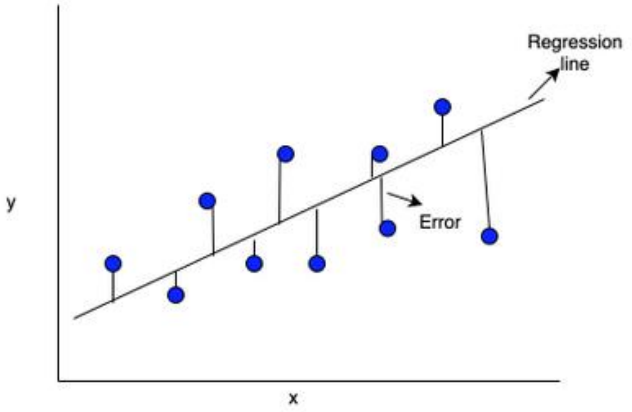
- y is the predicted value of the dependent variable (y).
- β_0 is the intercept, the predicted value of y when the x is 0.
- β_1 is the regression coefficient.
- ε is the error of the estimate.

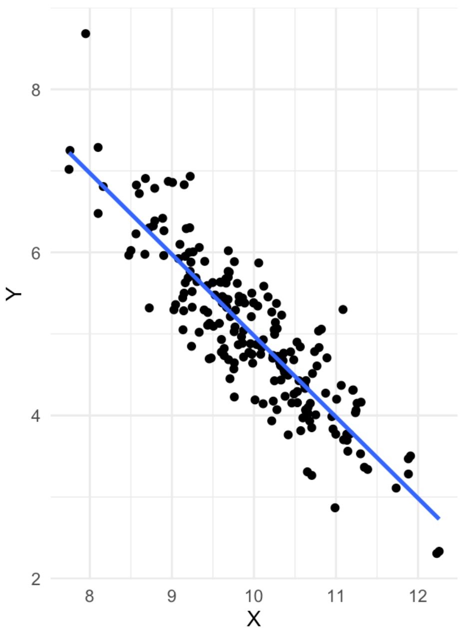


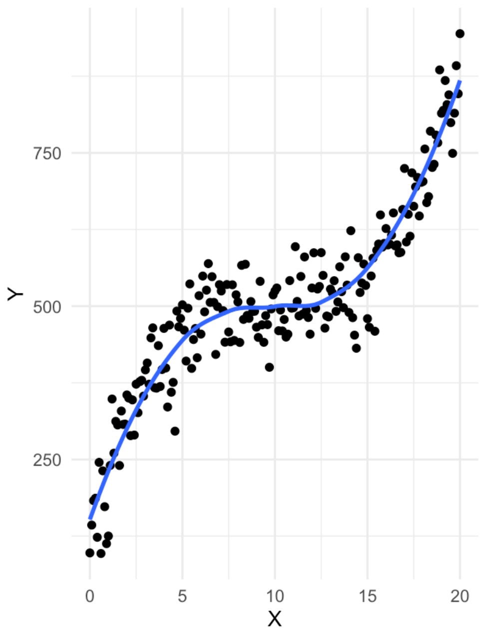
Linear Regression

Assumptions:

- **Homogeneity of variance** (homoscedasticity): the size of the error doesn't change significantly across independent variable.
- **Independence of observations**: there are no hidden relationships among observations (random sample).
- **Normality**: The data follows a normal distribution.
- The relationship between the independent and dependent variable is **linear**.







Linear Regression - Pros and Cons

ADVANTAGES

- Simple model
- Computationally efficient – no complicated calculations and is fast with large amount of data
- Interpretability of the Output – allows to determine the influence of variables looking to the coefficients

DISADVANTAGES

- Overly-Simplistic – to simple to capture reality
- Based on assumptions
- Affected by Outliers

Build the model – Deep Dive

DEMO TOPICS

Scoring

Classification

Clustering

Linear Regression

Logistic Regression

Decision Tree

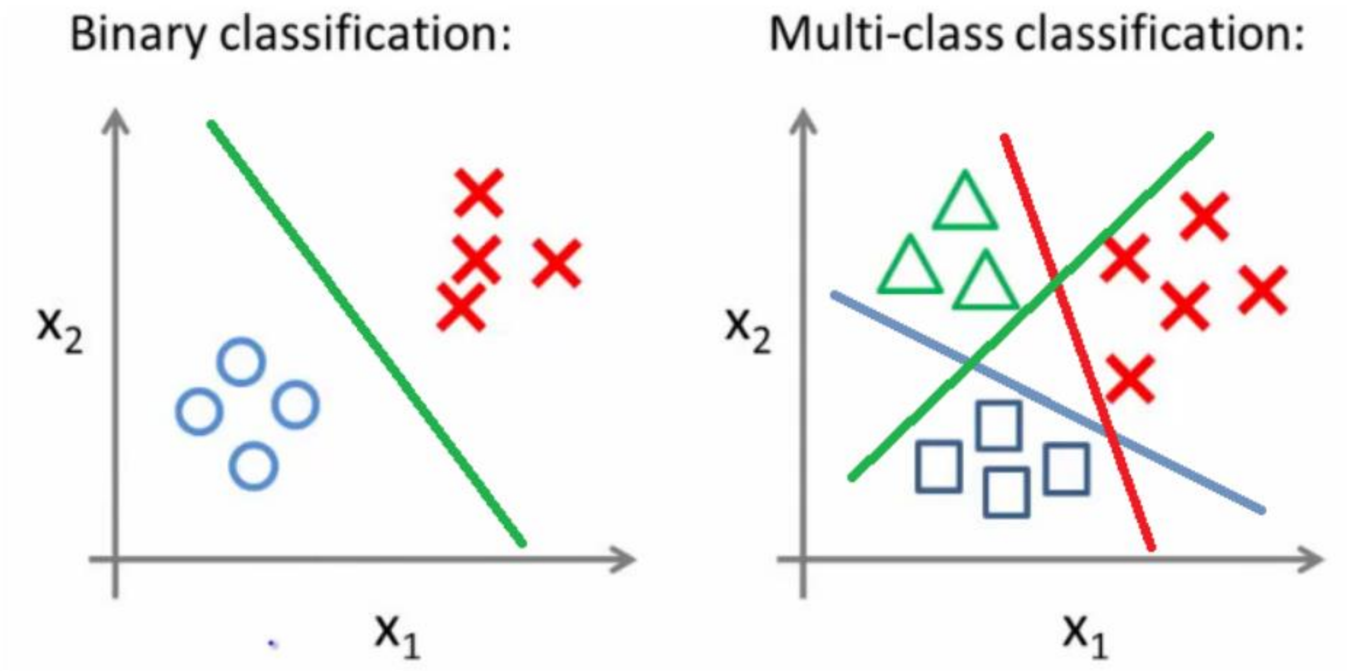
Random Forest

K-Means

Hierarchical

Classification

The Classification is a **supervised machine learning algorithm** that sorts the input data into different categories.



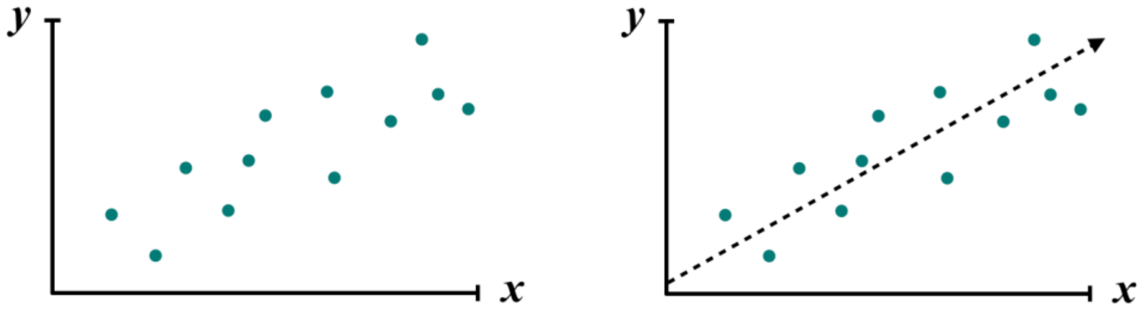
Build the model – Deep Dive

DEMO TOPICS

Scoring	Classification			Clustering	
Linear Regression	Logistic Regression	Decision Tree	Random Forest	K-Means	Hierarchical

Logistic Regression - Intro

The logistic regression involves **predicting an outcome Y** (dependent variable) **using one or more predictors**, labeled as **X** variables.



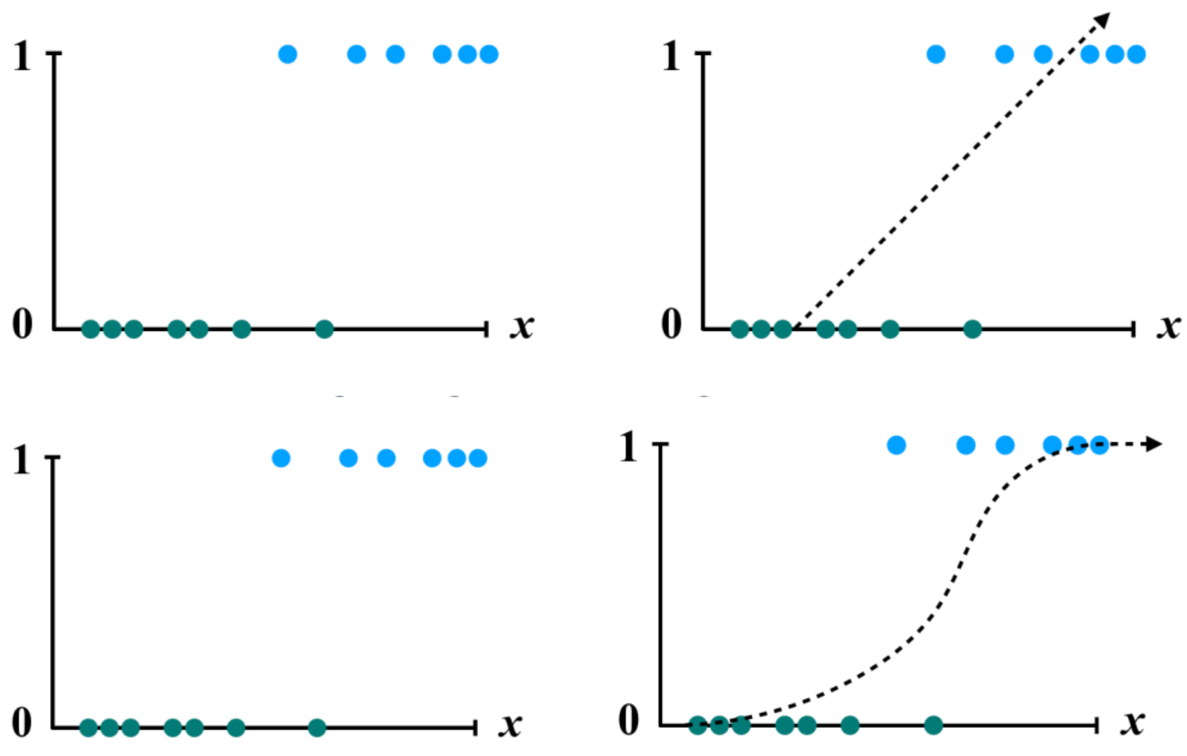
The Y might reflect something like income of life expectancy, while the X could represent age or education.

Linear Regression involves fitting the straight line to this data that best captures the relationship between x and y terms.

Logistic Regression - Intro

The Logistic Regression instead of trying to model data with a straight line, uses a curve.

A type of S-shaped curve called a **logistic function** has the property that for any input value of x, **the output is always between 0 and 1 just like a probability**. The greater this probability, the more likely the outcome is to be the one labeled '1'.



Logistic Regression

The outputs of Logistic regression rappresent the probability (p) of events.

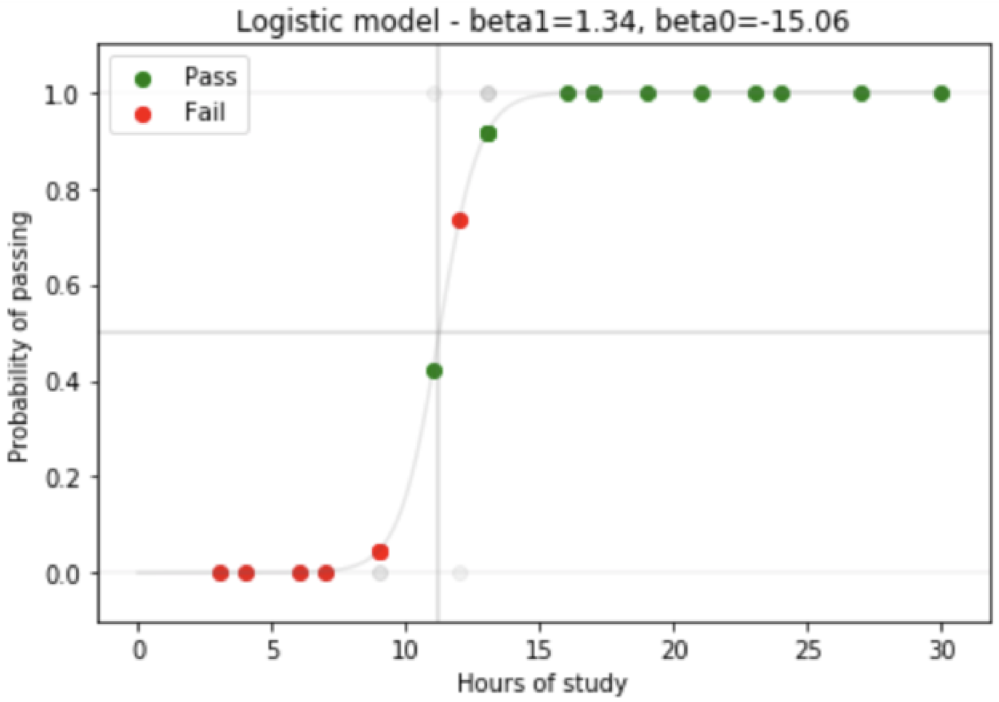
If **p > 0.5** then data is labeled as 1 and if **p < 0.5** the data is labeled as 0.

It measures the relationship between the “**Label**” on the Y-axis and "**Features**" on the X-axis using a logistic function as shown in this figure.

e.g.: Relation between hours of Study and probability of passing

Logistic regression equation:

$$\ln(\mathrm{Odds}) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k$$



Logistic Regression - Pros and Cons

ADVANTAGES

- Simple to implement and interpret in terms of data classification
- It can be easily extended to multiple classes (multinomial regression)
- Interpretability of the Output – allows to determine the influence and importance of variables

DISADVANTAGES

- Does not capture complex relationships
- In high-dimensional data, it may lead to overfitting
- Independent variables are linearly related to the log odds

Build the model – Deep Dive

DEMO TOPICS

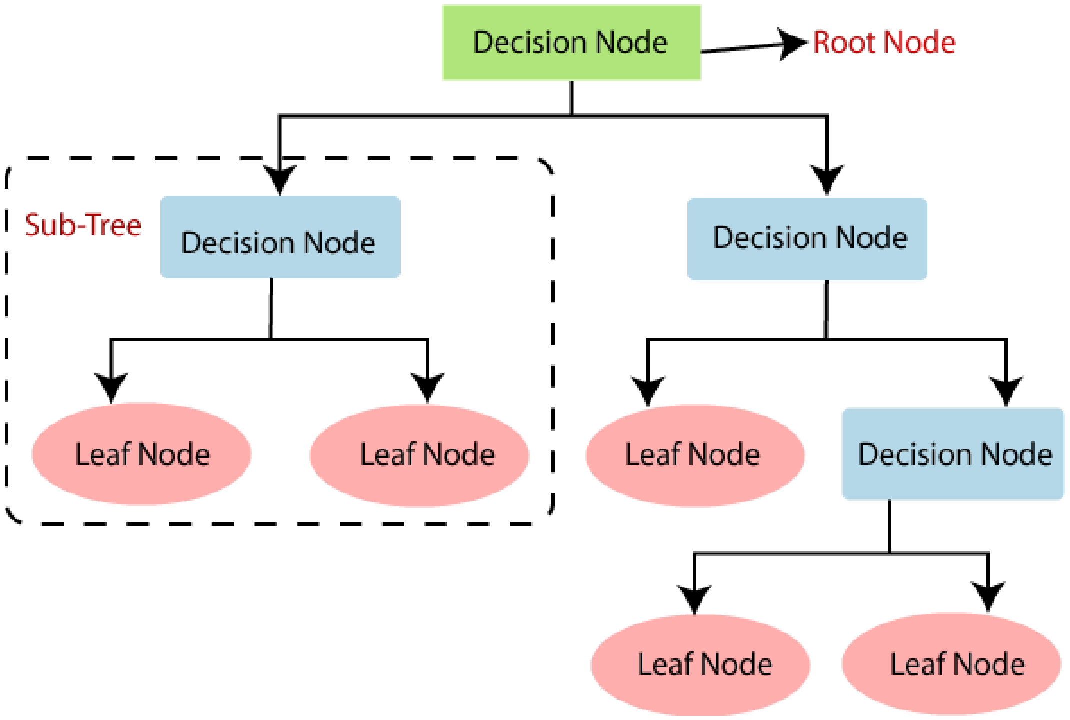
Scoring	Classification			Clustering	
Linear Regression	Logistic Regression	Decision Tree	Random Forest	K-Means	Hierarchical

Decision Trees

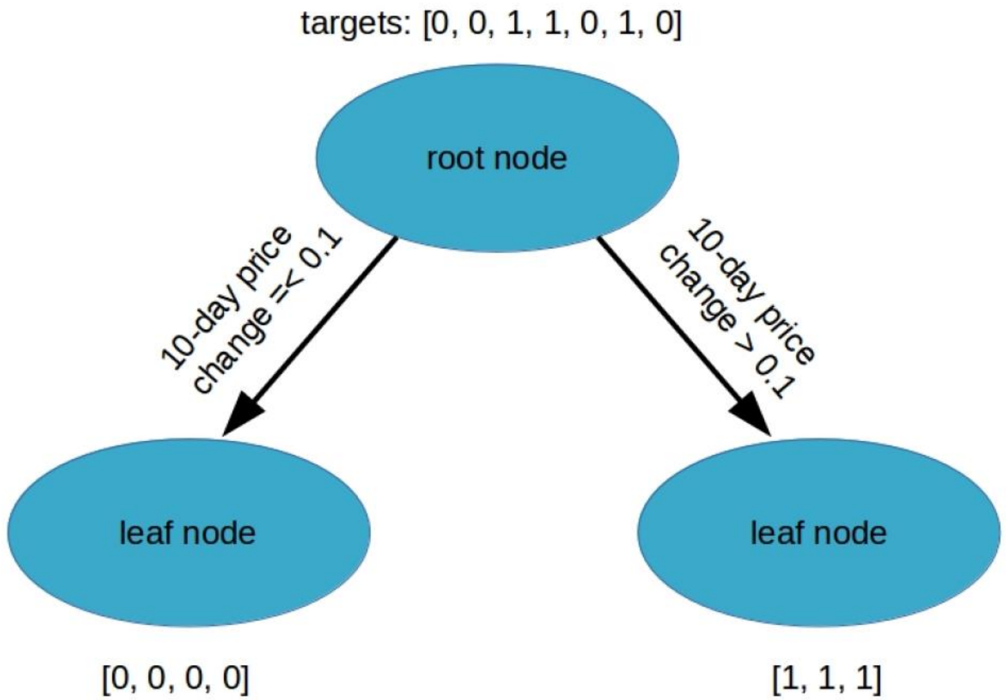
Decision Trees are a supervised learning method used for **classification** and **regression**.

It's very helpful due to its easy interpretability.

The general structure is hierarchical; starting from a root node (a starting point) the tree is split in other nodes through branches.



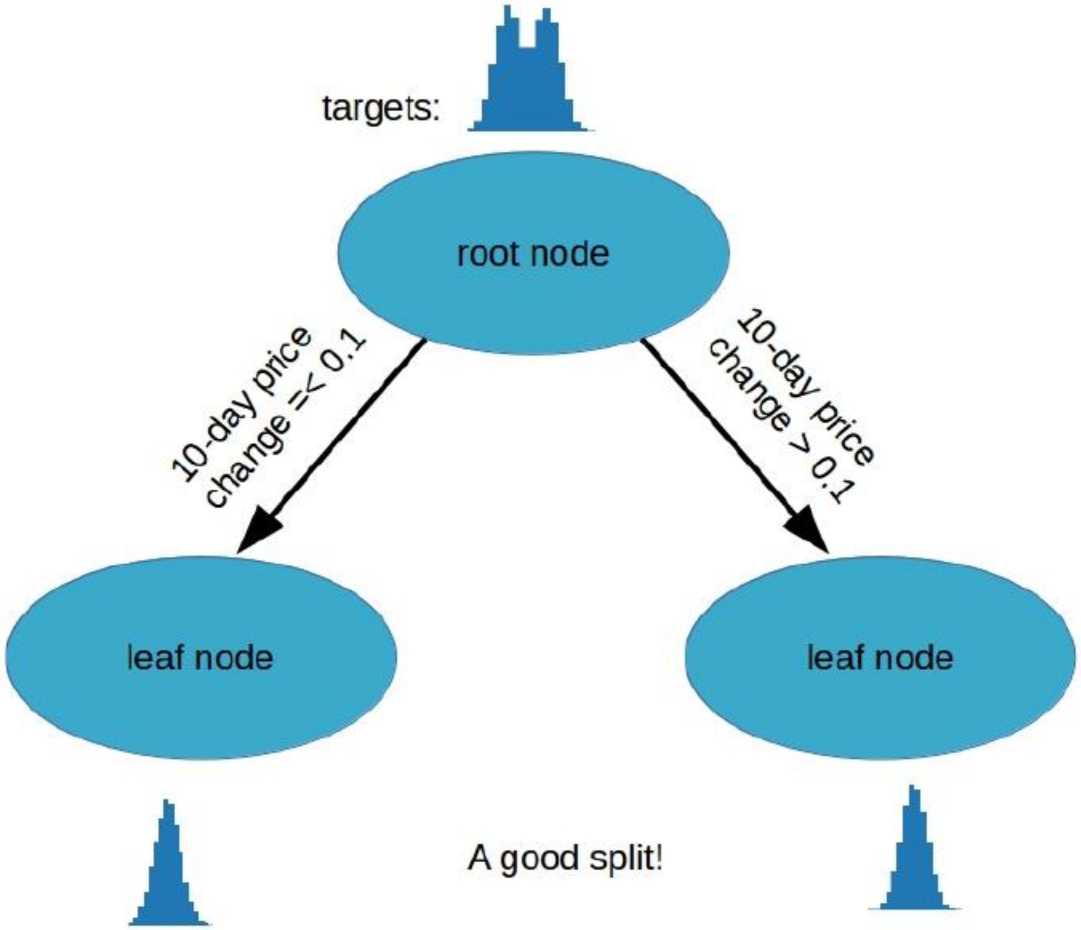
Decision Trees



The decision tree is a method used to split observations into different sub-groups, that determines a parent-child relationship.

Goal: create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features.

Decision Trees



Trees split data based on features to get the **best possible predictions.**

In the case of binary classification, we would try to group all the 0s on one side, and the 1s on another side.

The tree uses "**purity**“* of the leaf nodes to choose the best feature for making splits at each node.

Purity is a measurement of **homogeneity** of targets in a leaf node.

Decision Tree- Pros and Cons

ADVANTAGES

- Simple to understand and interpret. Trees can be visualized
- Flexible – used for Classification or Regression
- Requires minimal data preparation and can handle missing values
- Capture of non-linear patterns

DISADVANTAGES

- Creation of over-complex trees – overfitting
- Unstable – small variations leads to completely different tree
- Biased trees if some classes dominate

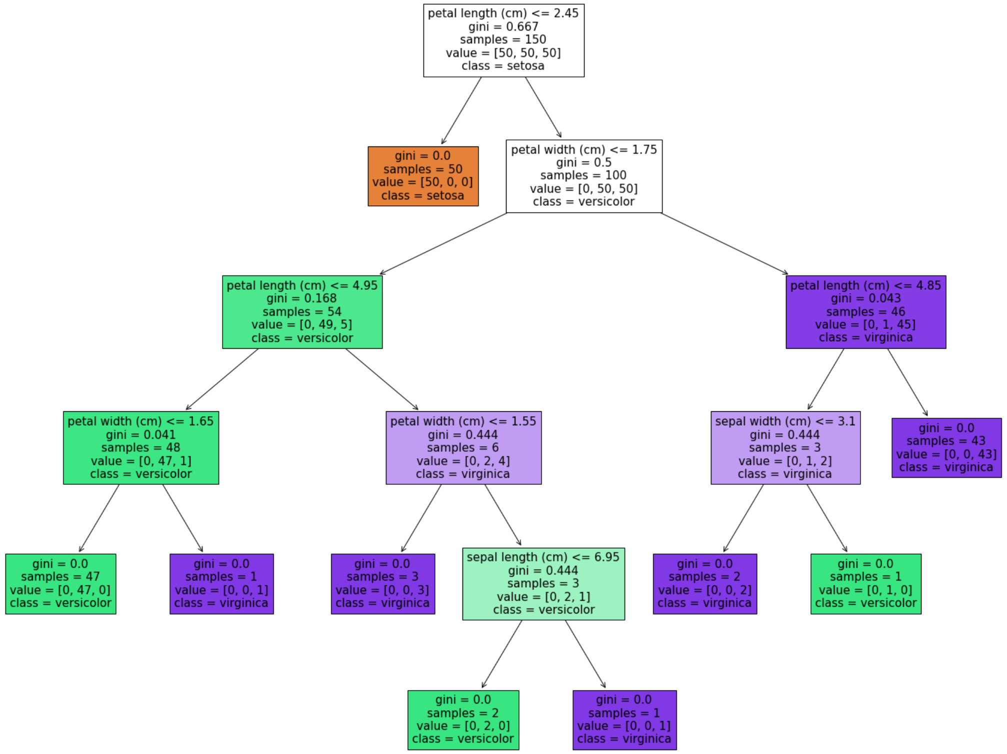
How to visualize Decision Tree

Here below an example of Decision Tree visualizations using Python

Each node in the decision tree has the following characteristics:

- Variable used for the split and for which value.
- Gini index[1] - measuring the disparity of a distribution
- Sample size
- Value – split of the sample between classes
- Assigned class

Visualization example



[1]Gini Index for Decision Tree

Build the model – Deep Dive

DEMO TOPICS

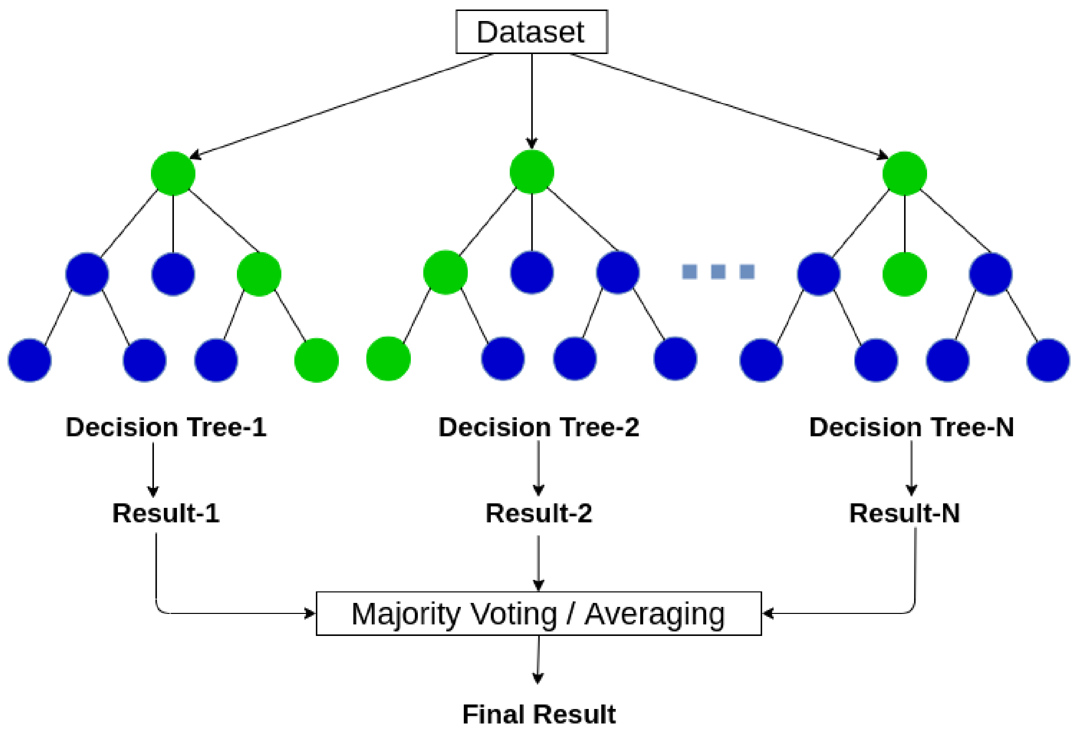
Scoring	Classification			Clustering	
Linear Regression	Logistic Regression	Decision Tree	Random Forest	K-Means	Hierarchical

Random Forest

- Random Forest is a supervised machine learning algorithm that **combines** multiple decision trees to create a "forest."

- The idea is to create a large number of **uncorrelated** decision trees by sampling with replacement several random samples from the training set to get a more accurate prediction

- The output chosen by the majority voting of the decision trees becomes the final result (while in case of regression the average between all decision trees is computed)



Random Forest - Pros and Cons

ADVANTAGES

- Accuracy: more accurate outcomes (also with missing value) and resolves the problem of overfitting
- Efficiency on a large database
- Versatility – can be used for Classification or Regression

DISADVANTAGES

- Require a lot of memory on larger projects
- Slower than other algorithms

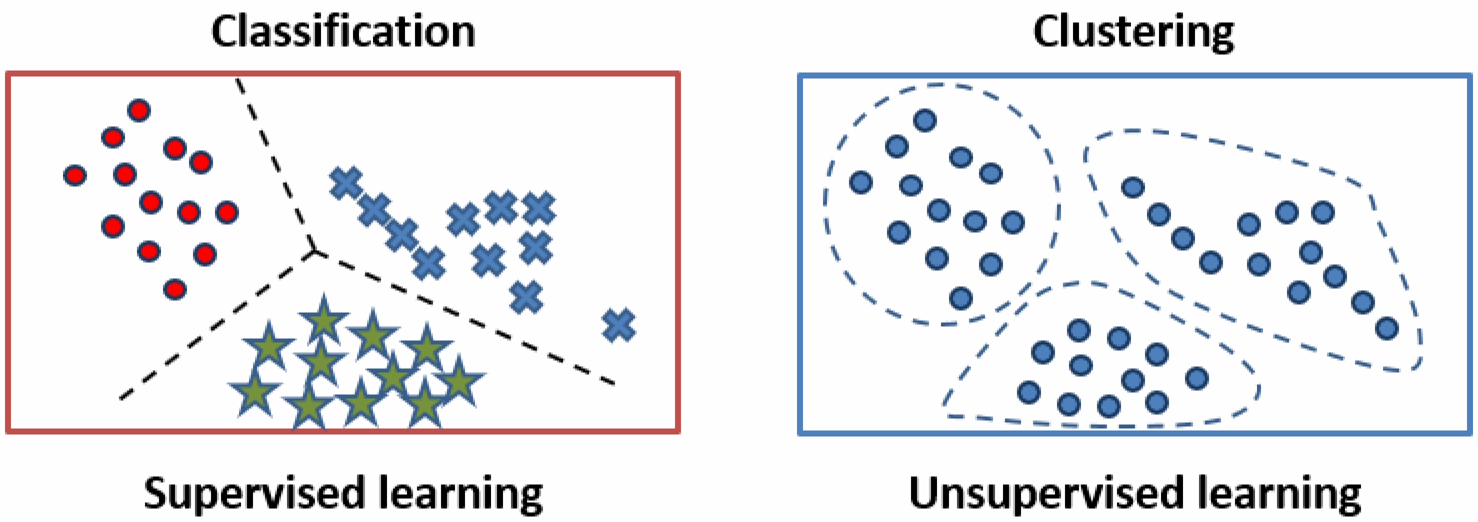
Build the model – Deep Dive

DEMO TOPICS

Scoring	Classification			Clustering	
Linear Regression	Logistic Regression	Decision Tree	Random Forest	K-Means	Hierarchical

Clustering

Clustering is **unsupervised machine learning algorithm** that divide the data points into several group.
Each cluster is distinct from each others and the data within each cluster are broadly similar to each other.



Build the model – Deep Dive

DEMO TOPICS

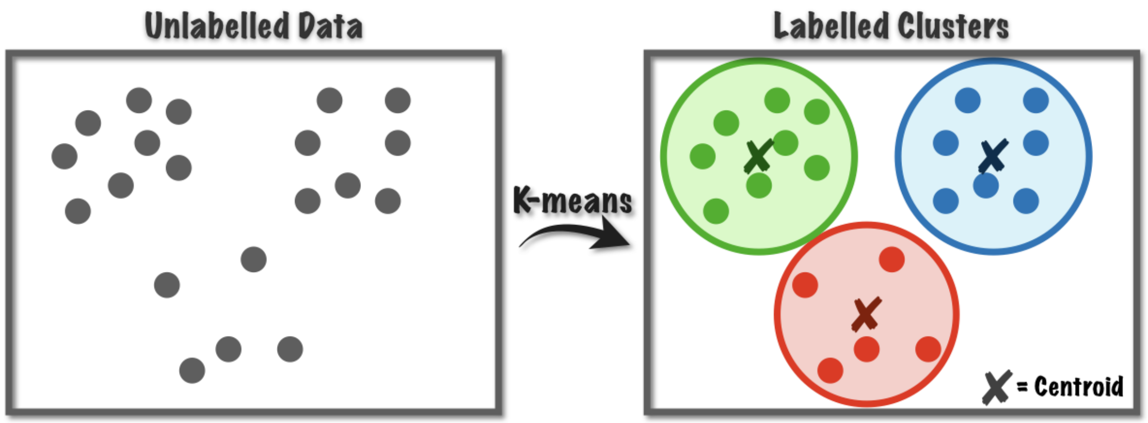
Scoring	Classification			Clustering	
Linear Regression	Logistic Regression	Decision Tree	Random Forest	K-Means	Hierarchical

K-Means

- K-Means Clustering is Unsupervised Learning algorithm, which groups the unlabeled data into different cluster.

- It allows to discover the categories of groups without the need of any training.

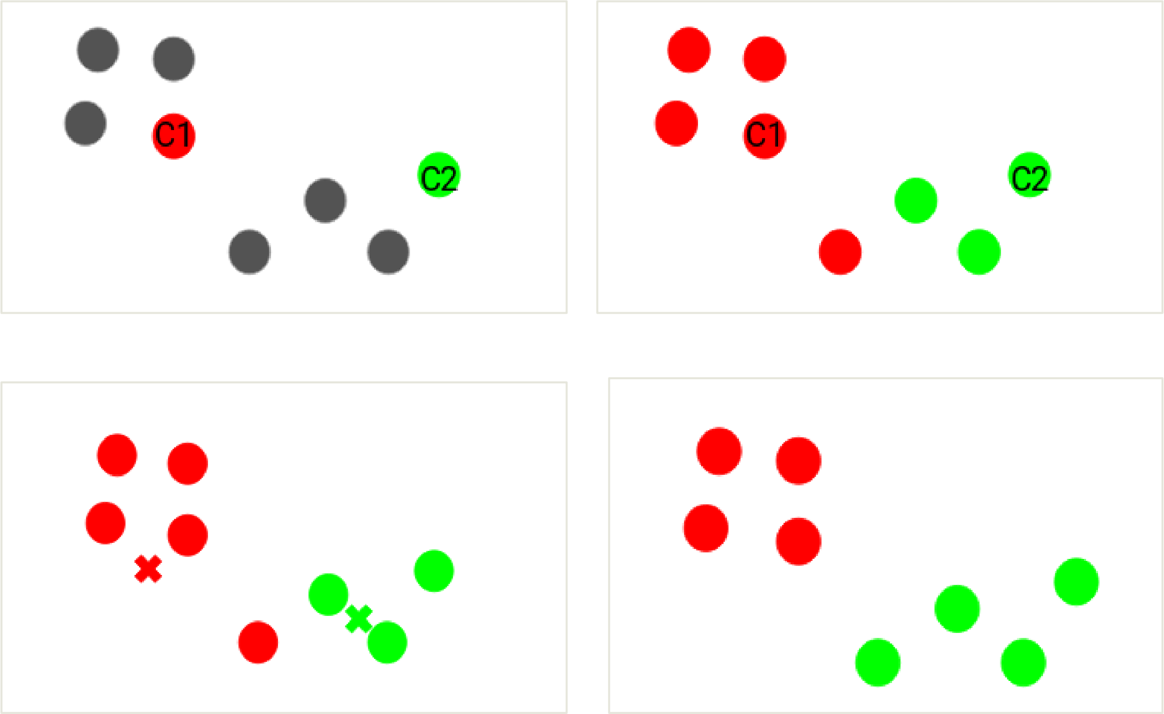
- The parameter “K” represents the number of clusters to be formed, and to determine the optimal value the **elbow method** can be used by iterative process.



K-Means

Steps

1. Select K to decide the number of cluster. E.g K=2.
2. Select random K points or centroids.
3. Assign each data point to their **closet** centroid, which will form the predefined K clusters
4. Recompute the centroids of newly formed clusters and repeat the 3rd step.



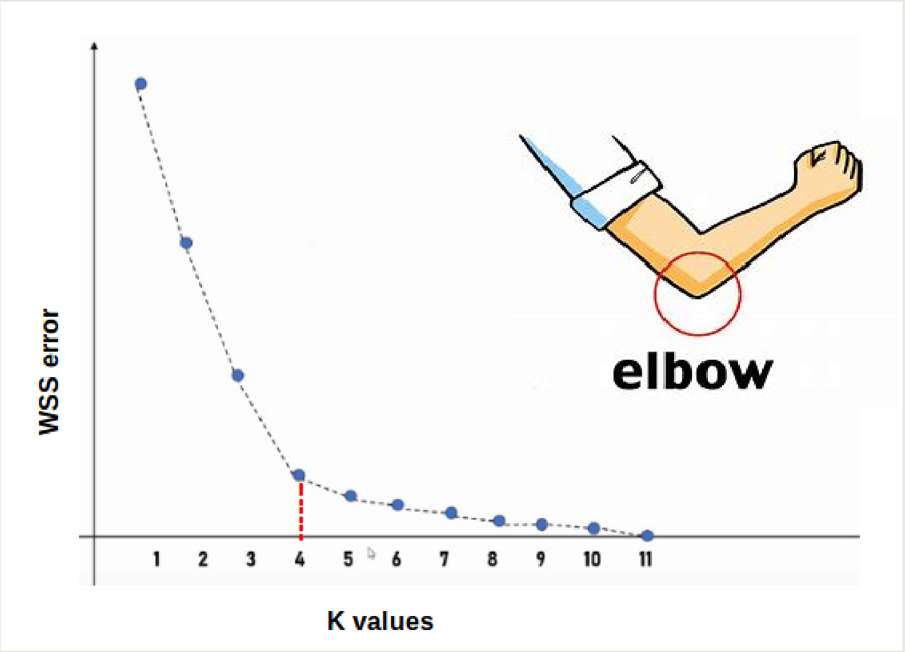
K-Means

Elbow Method

The Elbow Method runs K-means clustering on dataset for a range of values of K (e.g. 1 to 10).

1. Perform K-Means Clustering with different values of K and calculate average distances to the centroid for each K.

2. Plot the average distances and find the point where the line «falls»



K-Means - Pros and Cons

ADVANTAGES

- Simple to implement
- Generalizes to clusters of different shapes and sizes

DISADVANTAGES

- Choosing k manually
- Being dependent on initial values
- Affected by Outlier

Build the model – Deep Dive

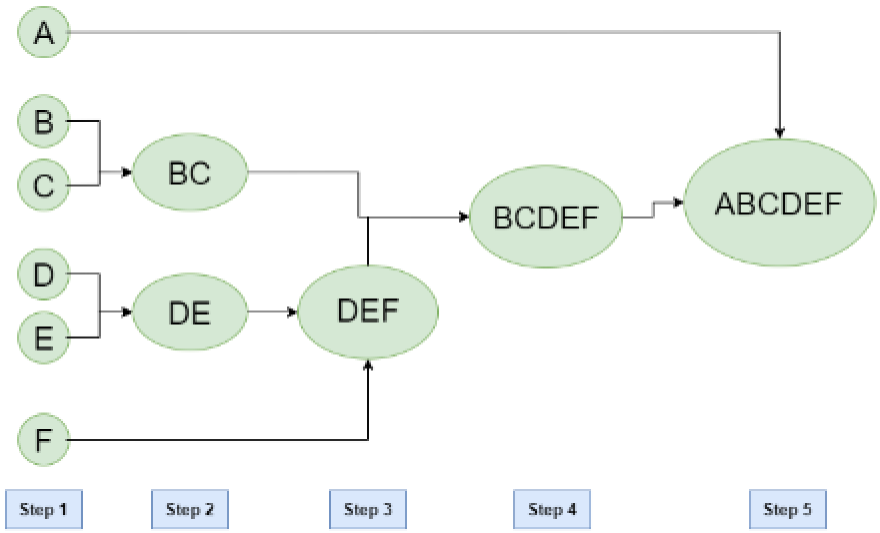
DEMO TOPICS

Scoring	Classification			Clustering	
Linear Regression	Logistic Regression	Decision Tree	Random Forest	K-Means	Hierarchical

Hierarchical clustering

- Hierarchical is unsupervised machine learning algorithm that groups similar objects into groups.

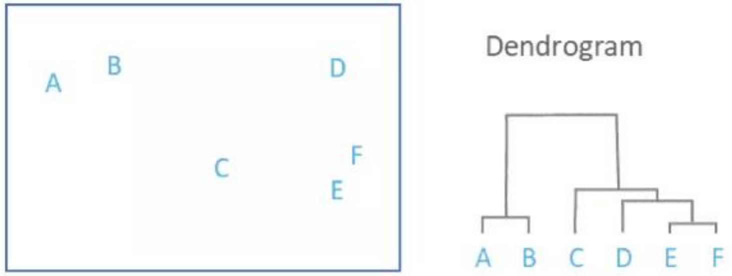
- The main output is the **dendrogram**, which shows the hierarchical relationship between the clusters.
  This allows to decide the level or scale of clustering that is most appropriate for application.

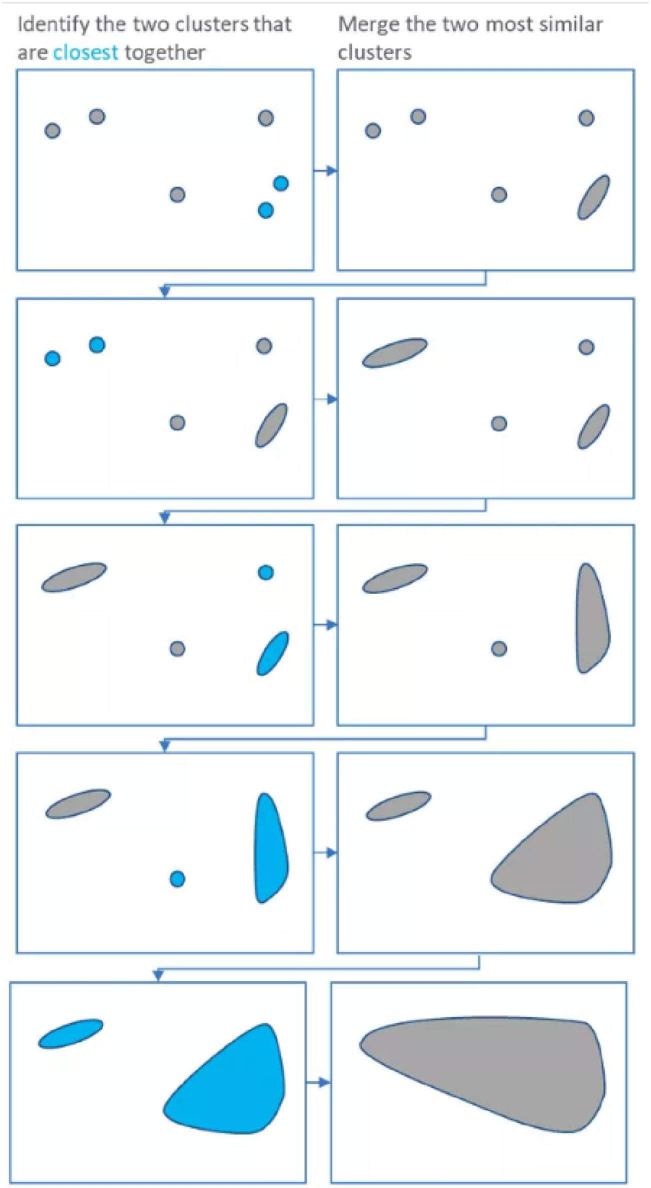


Hierarchical clustering

Steps

1. Treating each observation as a separate cluster
2. Identify the two clusters that are closest together by measures of distance (similarity)
3. Merge the two most similar clusters.
4. This iterative process continues until all the clusters are merged.





Hierarchical - Pros and Cons

ADVANTAGES

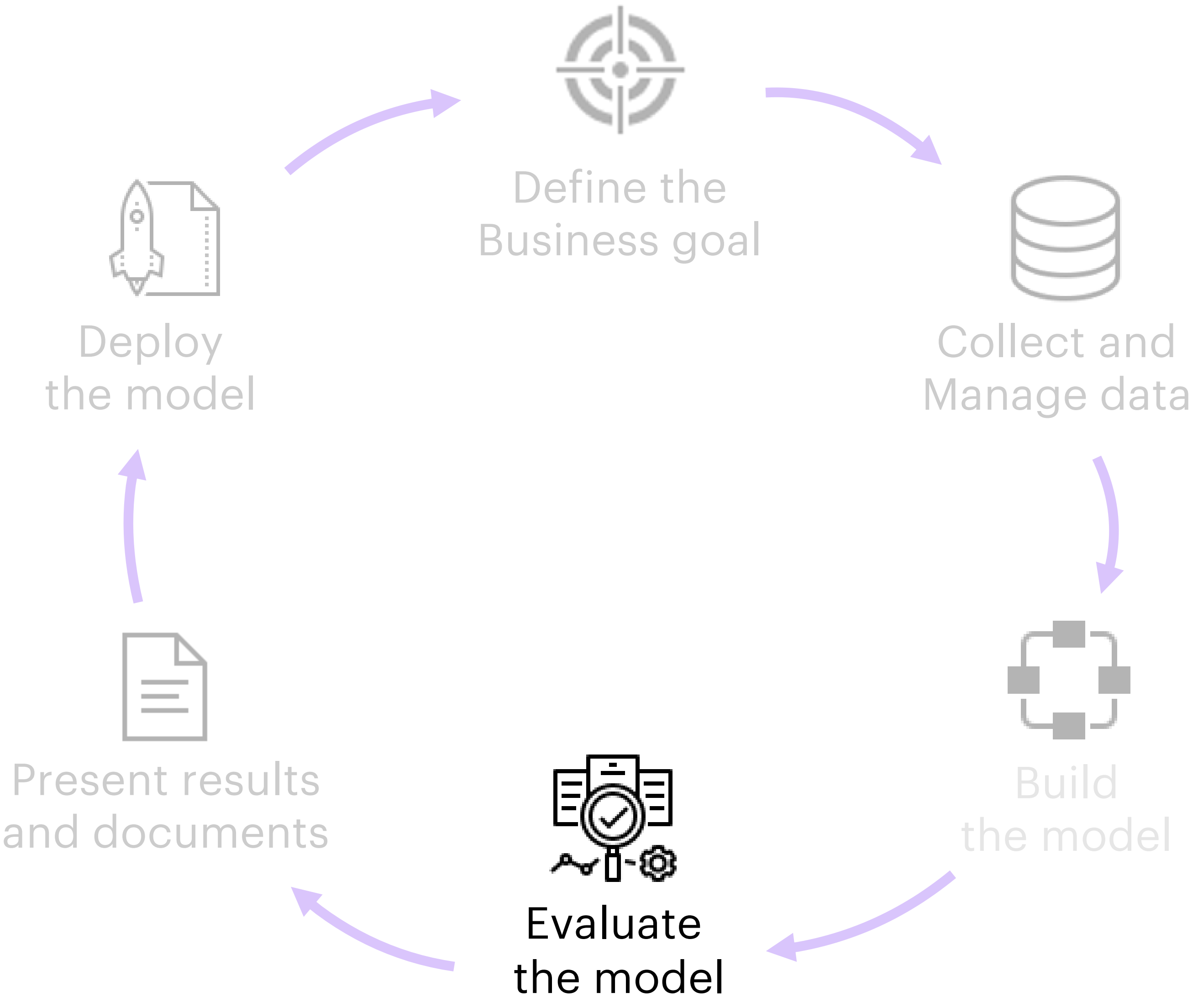
- Identifies the optimal number of clusters itself
- Dendrograms - visualization simple to understand
- Good for small data sets

DISADVANTAGES

- Computationally demanding
- Fails on larger sets
- Other disadvantages due to the similarity index used

Session 2

Evaluate the model



[Back to slide](#)

Performance

Confusion Matrix

Confusion Matrix is a **performance measurement** for machine learning classification problem.
It is a table with 4 different combinations of predicted and actual values:

		Actual Values	
		Positive (1)	Negative (0)
Predict Values	Positive (1)	TRUE POSITIVE (TP)	FALSE POSITIVE (FP)
	Negative (0)	FALSE NEGATIVE (FN)	TRUE NEGATIVE (TN)

→ Allows to evaluate the performance of a classification model using: **Accuracy, Precision and Recall**.

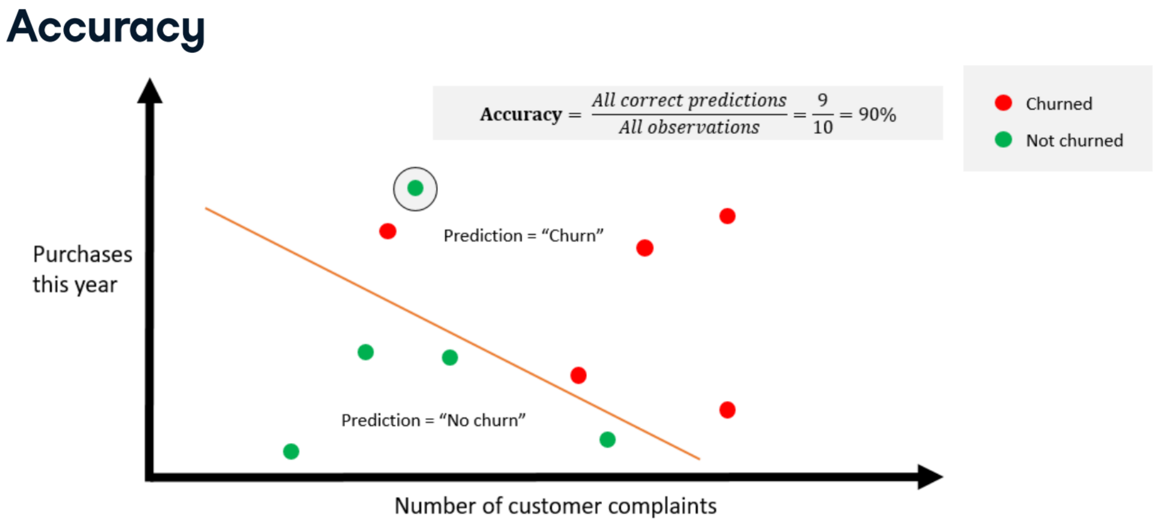
Performace

Accuracy

- Accuracy is the ratio of correctly predicted observation to the total observations:

$$Accuracy = \frac{TP + TN}{Total}$$

- Is preferred to use only with **symmetric datasets** where values of false positive and false negatives are almost same.



Performance

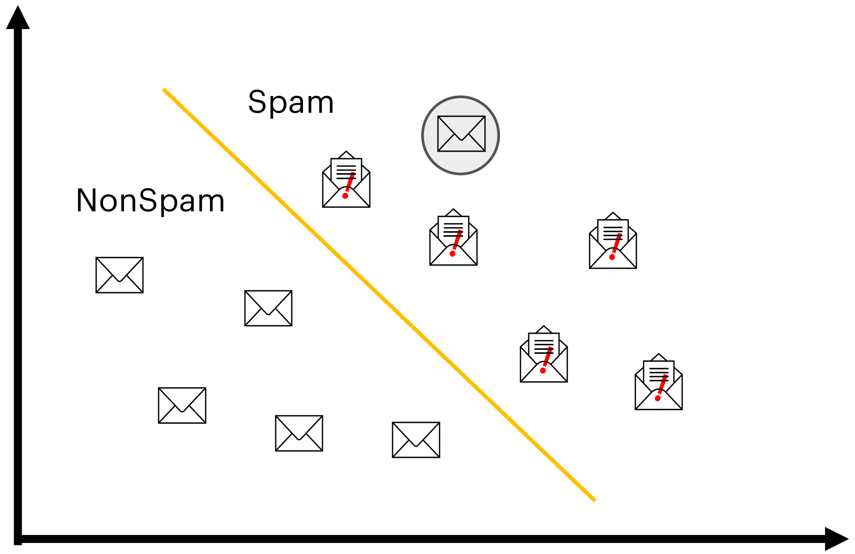
Precision

- Precision is the ratio of correctly predicted positive observations to the total predicted positive observations. It is a good measure to determine, when the costs of False Positive is high.

$$Precision = \frac{TP}{TP + FP}$$

- e.g. In **email spam** detection, an email non-spam has been identified as spam (False Positive) can cause the loss of important information.

$$Precision = \frac{\textbf{Correct}\ Spam\ prediction}{Observation\ \textbf{predicted}\ as\ Spam} = \frac{5}{6} = \sim 83\%$$



Performace

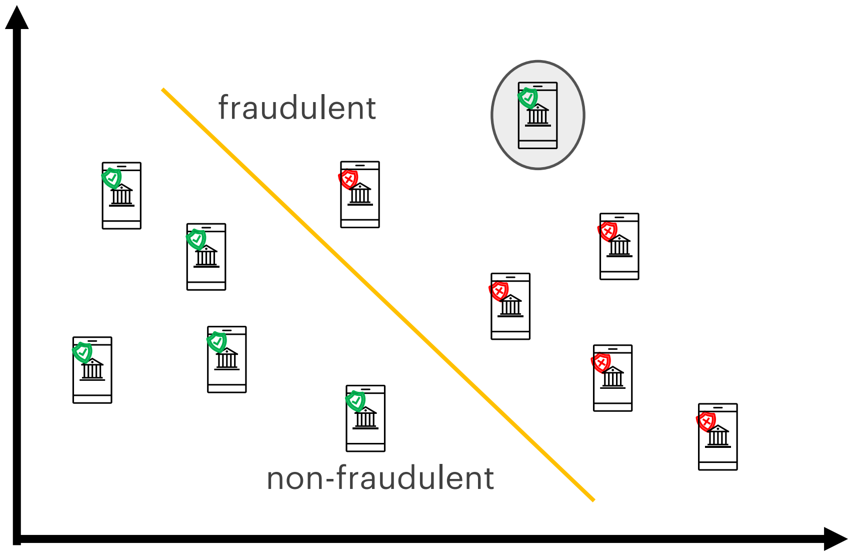
Recall

➢ Recall is the ratio of correctly predicted positive observations to the all observations in actual positive class

$$Recall = \frac{TP}{TP + FN}$$

➢ e.g. In **fraud detection**, if a fraudulent transaction is predicted as non-fraudulent (False Negative), the consequence can be very bad for the bank.

$$Recall = \frac{\textbf{Correct}\ fraudulent\ prediction}{All\ \textbf{actual fraudulent}observation} = \frac{5}{5} = 100\%$$

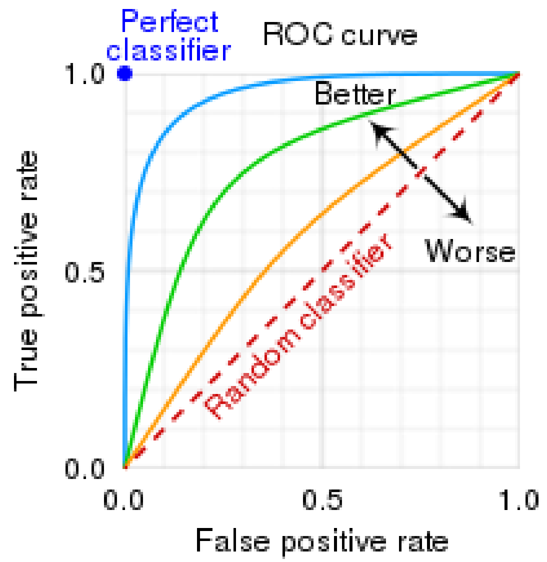


Performance

ROC/AUC

- **ROC Curve** (Receiver Operating Characteristic) is graph that shows the performance of a classification model at all classification thresholds.
- **AUC Curve** (Area under the ROC) provides an aggregate measure of performance across all possible classification thresholds.

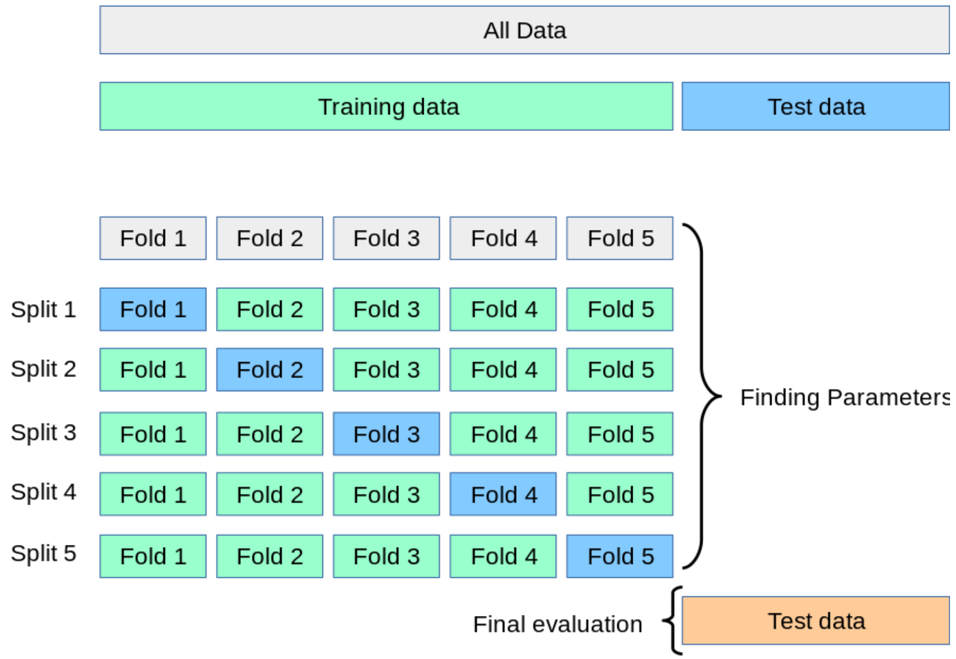
- The higher the area under the ROC curve (AUC), the better the classifier. A perfect classifier would have an AUC of 1.



Cross Validation – K fold CV

Following on Performance topics, here's a focus on performance metrics related to Cross Validation.

The K-fold Cross Validation allows to run a single model on different combinations of training/test sets and provide a more robust metric of performance.



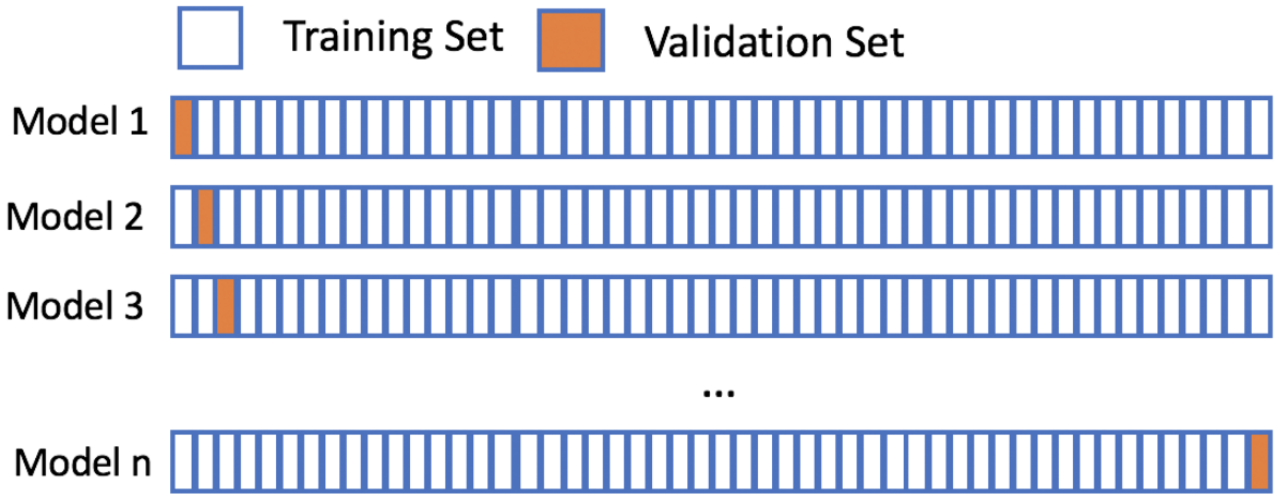
Steps

1. Split the dataset into k groups;
2. For each split: take one group as test data set and the remaining as training data set;
3. Fit the model on the training set and evaluate it on the test set;
4. Retain the evaluation score and get the mean value in order to determine the overall accuracy of the model;

Cross Validation - LOOCV

The Leave-One-Out-Cross-Validation is a technique used when data are limited and is an extension of **K-fold cross-validation**, where K is equal to n, the number of observations in the data.

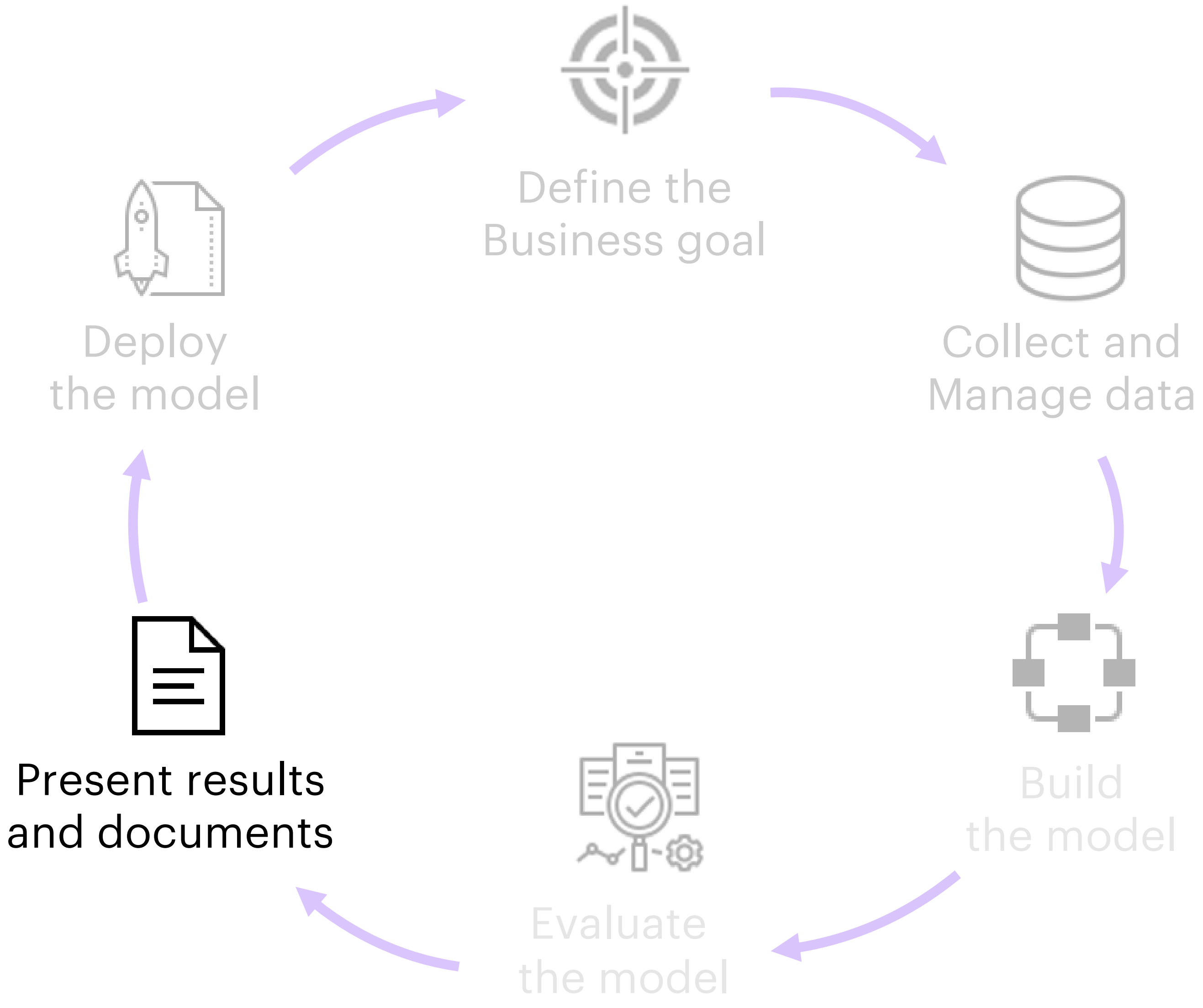
Each observation will be used as validation set, completely on its own.
For Model 1, all the data will be used for training except the first point, which will be used for validation. In Model 2, the second point is left out, in Model 3 the third, and so on.
The process is iterative until the last observation is used as validation set.



Session 2

Present results

and documents



[Back to slide]

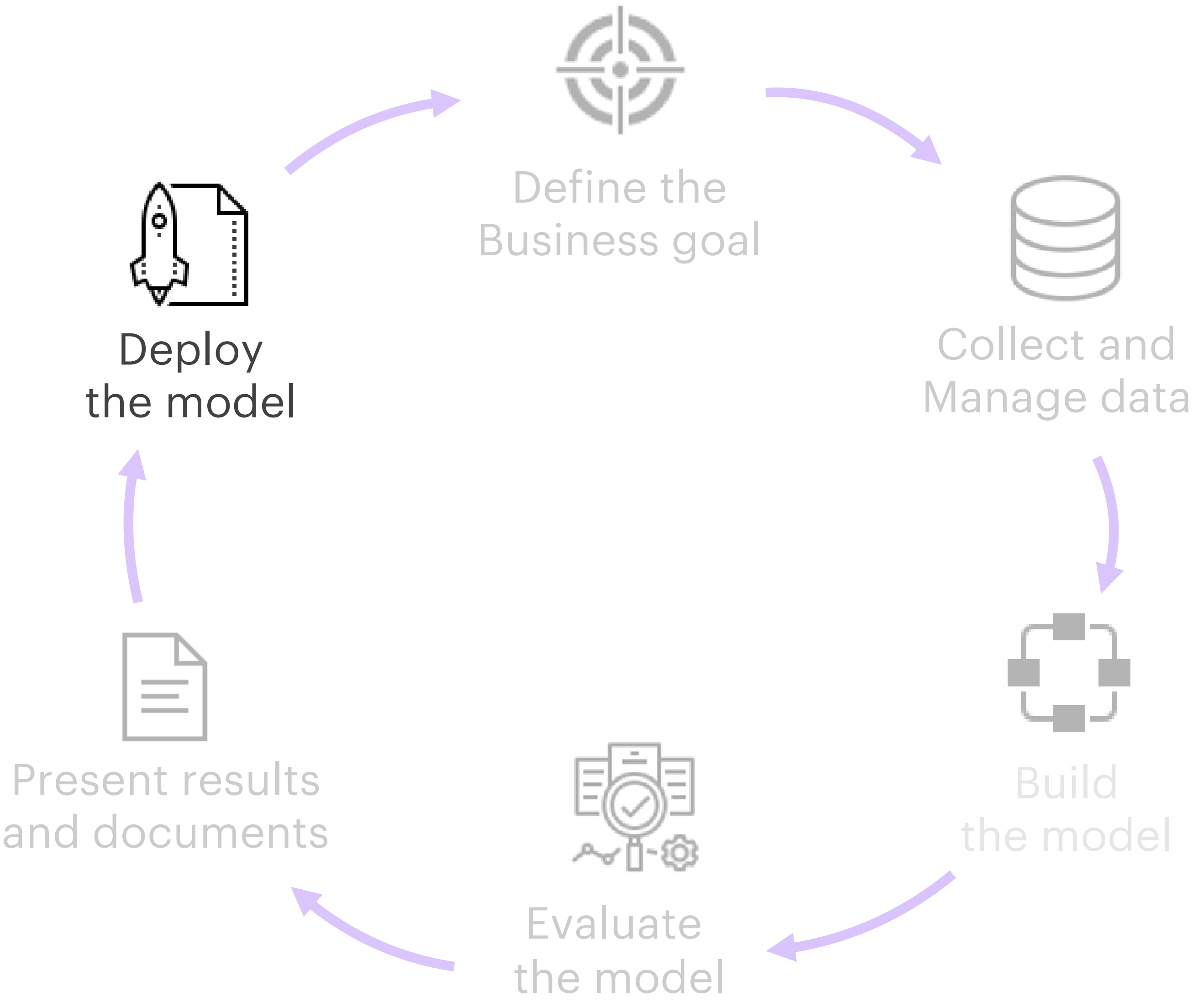
Present results & documents

In any communication strategy, there are several pieces we have to put together to create an effective story.

DATA STORY	TARGET AUDIENCE	DATA	VIZ	PRESENT
Is the practice of **building a narrative around a set of data** and its accompanying visualizations to **help convey the meaning of that data in a powerful** and compelling fashion	**Data professionals** like to talk about their methods. Less technical audiences however typically just care about what the result is and what the implications are.	Select the "**right data**": it implies including sufficient contextual insights in our story to support our main point without overloading our report with information.	Choose and adjust a **visualization to the message we want to convey**. Consider our audience expertise in the topic and familiarity with the concepts, to select what graphs easily interpretable.	To help structure the presentation, we should consider: its **purpose,** who is the **audience and** the **message** we want to get across.

Session 2

Deploy the model



Back to slide

Deploy the model

Make your models available to your partners for experimentation, testing, and production deployment.

DEPLOY THE MODEL	Embed the model you chose in dashboards, application.
MONITOR MODEL PERFORMANCE	Regularly test the performance of your model as your data changes to avoid model drift
IMPROVE YOUR MODEL	Continuously iterate and improve the model post deployment. Replace your model with an updated version to improve performance.

Session 2

DEMO – Churn Analysis

Back to slide

Session 2

DEMO – Customer Segmentation

Back to slide

References

- Data Science Workflow/Lifecycle
- Linear & Logistic Regression
- Decision Trees
- Random Forest
- K-Means
- Hierarchical Clustering
- Model Evaluation
- Presenting Results

Q & A

Thank You