WEEK 12

Introduction to Machine Learning

Lecture outline

- **Machine Learning Definition**

- **Evaluation of the Logit Models:**
  - *Train and Test Datasets*
  - *Confusion Matrix, Accuracy*
  - *Sensitivity*
  - *Specificity*
  - *Precision*

Machine Learning Definitions

- **Algorithm**:

  A Machine Learning algorithm is a set of rules and statistical techniques used to learn patterns from data and draw significant information from it. It is the logic behind a Machine Learning model.

  Ex: Linear Regression or Logistic Regression algorithm.

- **Model**:

  A model is the main component of Machine Learning.

  A model is trained by using a Machine Learning Algorithm.

  An algorithm maps all the decisions that a model is supposed to take based on the given input, in order to get the correct output.

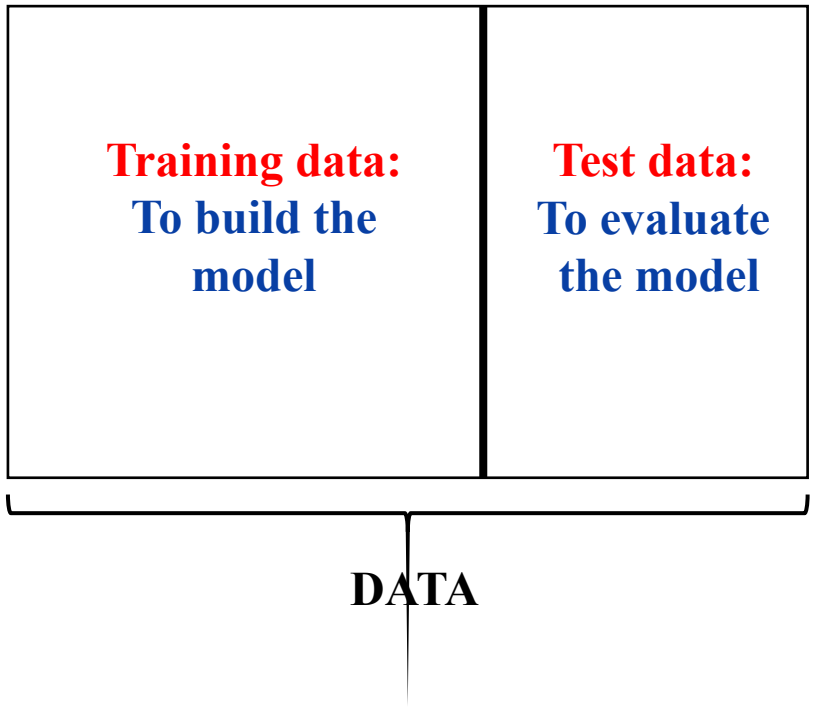
Data Partitioning:

- **Training Data:**

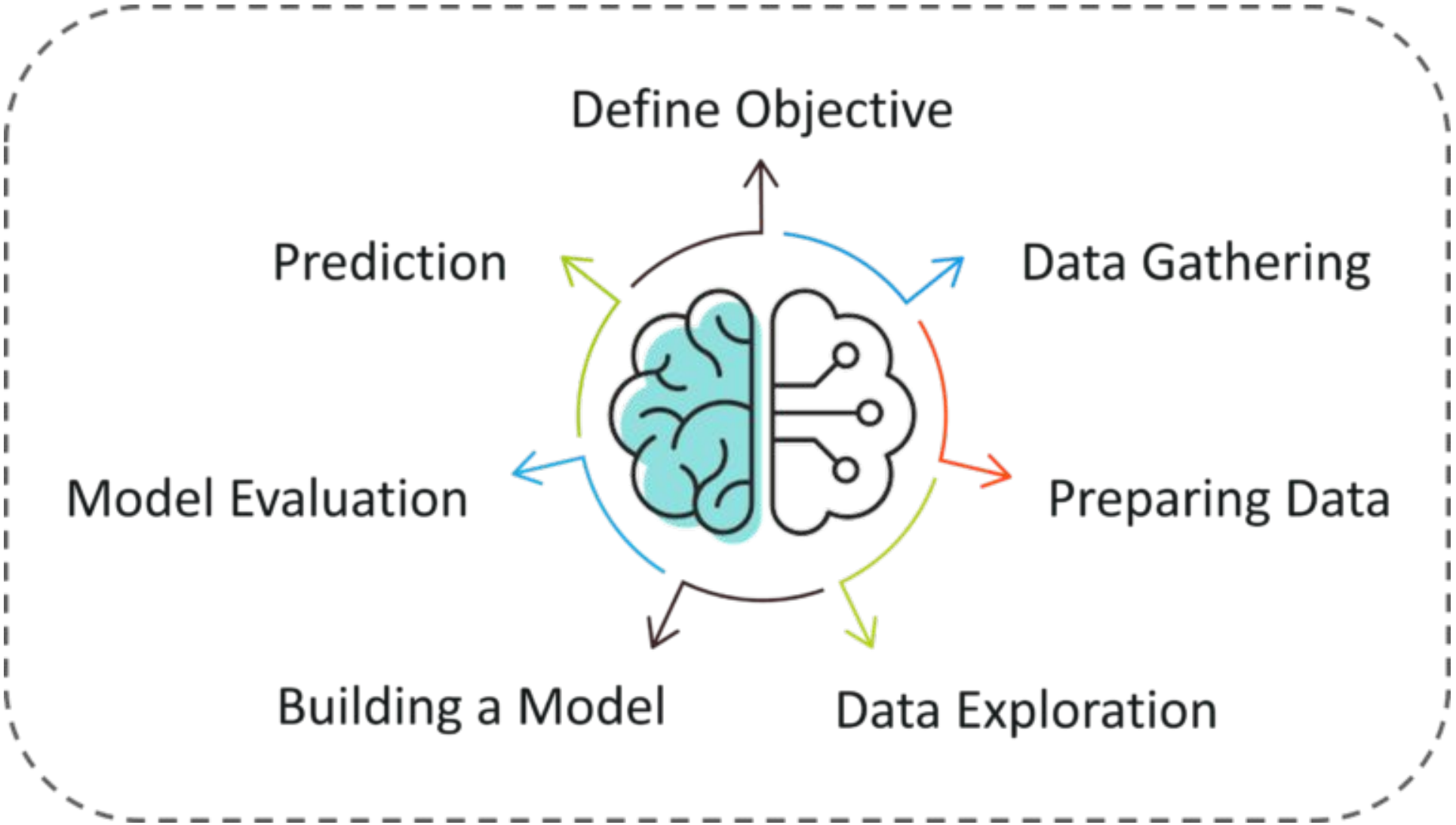
The Machine Learning model is built using the training data. The training data helps the model to identify key trends and patterns essential to predict the output.

- **Testing Data:**

After the model is trained, it must be tested to evaluate how accurately it can predict an outcome. This is done by the testing data set.



Machine Learning Process



Model Evaluation: Titanic Dataset

Fit a model to predict survival based on Pclass, Gender, Age and Fare. Use first 700 observations to fit your model (train dataset). Use the remaining observations (187 obs) as test dataset and evaluate your model using Confusion Matrix*.

Actual Class \ Predicted Class	Positive 1	Negative 0	
Positive 1	True Positive (TP)	False Negative (FN) **Type II Error**	**Sensitivity** $\frac{TP}{(TP+FN)}$
Negative 0	False Positive (FP) **Type I Error**	True Negative (TN)	**Specificity** $\frac{TN}{(TN+FP)}$
	Precision $\frac{TP}{(TP+FP)}$	**Negative Predictive Value** $\frac{TN}{(TN+FN)}$	**Accuracy** $\frac{TP+TN}{(TP+TN+FP+FN)}$

Confusion Matrix:

A tabular display (2X2 in the binary case) of the record counts by their predicted and actual classification status

Model Evaluation: Titanic Dataset

Now we apply the model to predict unseen data (test data) and compute the accuracy.

We use 0.5 as the **cutoff point**:

if $p > 0.5$, then we classify the passenger as survived (1).

if $p \leq 0.5$, then we classify the passenger as perished (0).

Confusion Matrix:

Actual \ Predicted	0	1
0	104	14
1	20	49

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  3.4137743  0.4786357   7.132 9.87e-13 ***
Pclass2     -1.0557724  0.3303871  -3.196 0.001396 **
Pclass3     -2.3438225  0.3337248  -7.023 2.17e-12 ***
Sexmale     -2.5773356  0.2100574 -12.270  < 2e-16 ***
Age         -0.0284005  0.0079198  -3.586 0.000336 ***
Fare        -0.0009845  0.0024805  -0.397 0.691433
```

We incorrectly predicted survival status of 20 + 14 = 34 out 187 passengers from the test data.

That leaves us with an **accuracy level** of $\frac{104+49}{187} = 0.818$ or $\approx 82\%$

Model Evaluation: Titanic Dataset

Sensitivity: The percent (or proportion) of all 1s that are correctly classified as 1s.

$$\text{Sensitivity} = \frac{49}{49+20} = 0.71$$

Specificity: The percent (or proportion) of all 0s that are correctly classified as 0s.

$$\text{Specificity} = \frac{104}{104+14} = 0.88$$

Precision: The percent (or proportion) of predicted 1s that are actually 1s.

$$\text{Precision} = \frac{49}{49+14} = 0.78$$

Confusion Matrix:

	Predicted	
Actual	0	1
0	104	14
1	20	49

Exercises

Assign first 800 of the Global Findex data to train data and the remaining 200 of them to test data.

Fit a multiple logit model to predict whether the person saved in the past 12 months (1) or not saved (0) based on age, gender, education level, and employment status.

a) Compute the odds ratio and probability for a female, 30 years old, with secondary education, and employed. Would you classify this person as saved or not saved? Use $p > 0.5$ as cutoff point to classify as saved (if $p > 0.5$, then Saved = 1).

b) Evaluate your model performance on test data using accuracy, sensitivity, specificity and precision ratios. Use $p > 0.5$ as cutoff point to classify as saved (if $p > 0.5$, then Saved = 1).

Fit another multiple logit model to predict whether the person saved (1) or not saved (0) using all other variables as predictor variables.

c) Evaluate your model performance on test data using accuracy, sensitivity, specificity and precision ratios. Use $p > 0.5$ as cutoff point to classify as saved (if $p > 0.5$, then Saved = 1).

Exercise solutions

The fitted model is given here:

```
Coefficients:
                                  Estimate Std. Error z value Pr(>|z|)
(Intercept)                     -1.329265   0.283558  -4.688 2.76e-06 ***
GenderMale                       0.418525   0.156715   2.671  0.00757 **
Age                              0.004852   0.004570   1.062  0.28835
Educationcompleted tertiary or more 1.329543   0.300350   4.427 9.57e-06 ***
Educationsecondary               0.372900   0.216837   1.720  0.08548 .
Employment                       0.252890   0.165266   1.530  0.12597
```

b. $\ln\left(\frac{p}{1-p}\right) = -1.3293 + 0.0049 * 30 + 0.3729 + 0.2529 * 1$

= -0.5565

Odd ratio: $\frac{p}{1-p} = e^{-0.5565} => \frac{p}{1-p} = 0.5732$

$p = \frac{0.5732}{1+0.5732} = 0.36$. Since $p < 0.50$, we classify this female as Not Saved in the past 12 months.

Confusion Matrix:

Actual \ Predicted	0	1
0	112	16
1	58	14

c.

$$\boldsymbol{Accuracy = \frac{112+14}{200} = 0.63}$$

$$\textbf{Specificity} = \frac{112}{112+16} = 0.88$$

$$\textbf{Sensitivity} = \frac{14}{14+58} = 0.19$$

$$\textbf{Precision} = \frac{14}{14+16} = 0.47$$

Exercise solutions


```
Coefficients:
                                 Estimate Std. Error z value Pr(>|z|)
(Intercept)                    -1.094863   0.356990  -3.067  0.00216 **
GenderMale                      0.331404   0.165698   2.000  0.04549 *
Age                             0.002609   0.004929   0.529  0.59661
Educationcompleted tertiary or more  0.629259   0.329735   1.908  0.05634 .
Educationsecondary              0.285304   0.226779   1.258  0.20837
Income_statusMiddle 20%        -0.671775   0.241542  -2.781  0.00542 **
Income_statusPoorest 20%       -0.996978   0.252446  -3.949 7.84e-05 ***
Income_statusRichest 20%        0.195545   0.229471   0.852  0.39413
Income_statusSecond 20%        -0.586627   0.249244  -2.354  0.01859 *
Employment                      0.233531   0.173441   1.346  0.17816
Has.a.debit.card                0.513292   0.232207   2.210  0.02707 *
Paid.bill.online                0.713257   0.362500   1.968  0.04911 *
Owns.a.mobile.phone             0.337022   0.202921   1.661  0.09674 .
Has.a.fin.account              -0.087602   0.214503  -0.408  0.68298
```

Confusion Matrix:

Actual	Predicted 0	Predicted 1
0	98	30
1	37	35

$$\text{Accuracy} = \frac{98+35}{200} = 0.67 \qquad \text{Precision} = \frac{35}{35+30} = 0.54$$

$$\text{Sensitivity} = \frac{35}{35+37} = 0.49 \qquad \text{Specificity} = \frac{98}{98+30} = 0.77$$

CONGRATS!!!

Thank you for your attention!

Congratulations!!!

You mastered

the “Introduction to Statistics and Data Science”

References

1. James G. et al. An Introduction to Statistical Learning (with Applications in R). ISBN: 978-1-4614-7137-0. p.p. 156 – 160.

2. https://www.datacamp.com/community/tutorials/logistic-regression-R

3. https://machinelearningmastery.com/logistic-regression-for-machine-learning/