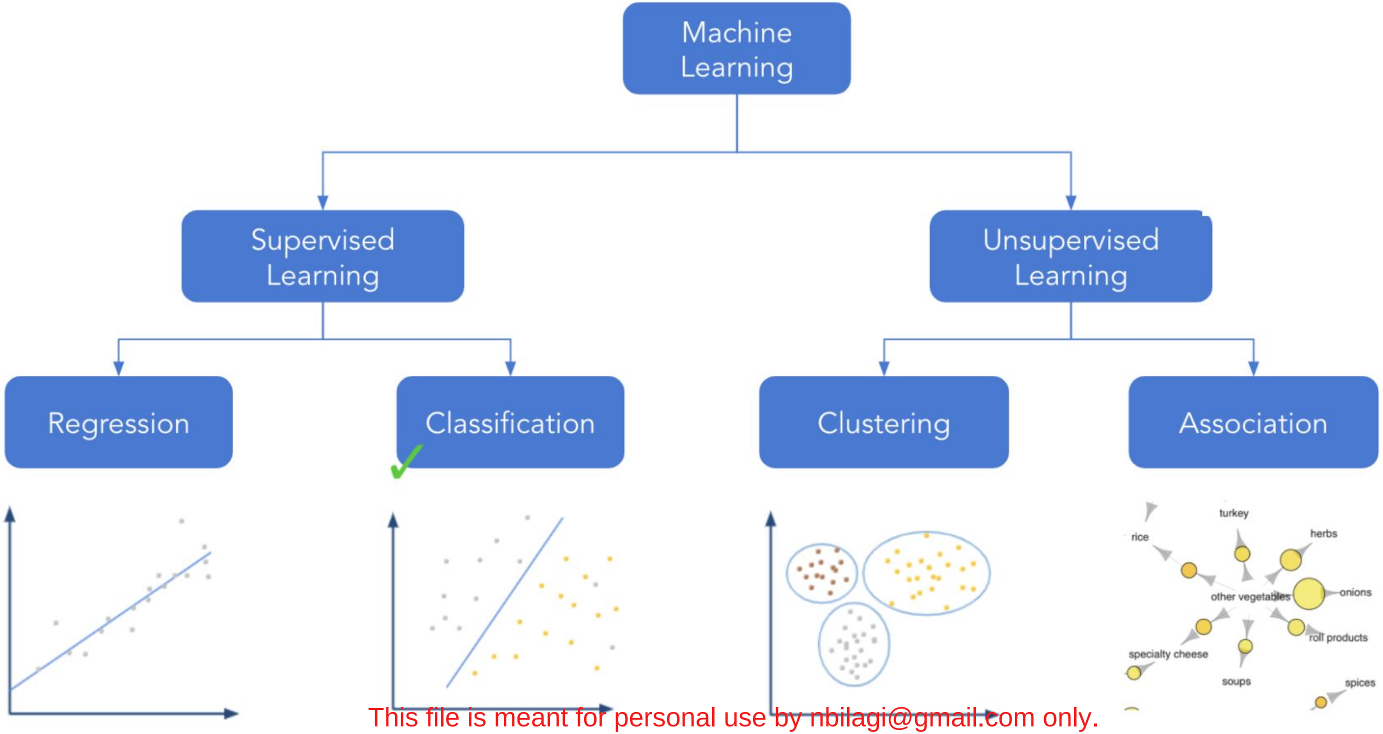
Supervised Learning Classification

Agenda

- Standard Process of DS projects
- Visiting Basics
- Odds/Probability
- Binomial logistic regression

Machine Learning

In this session, we shall cover classification



Machine Learning

Supervised Learning

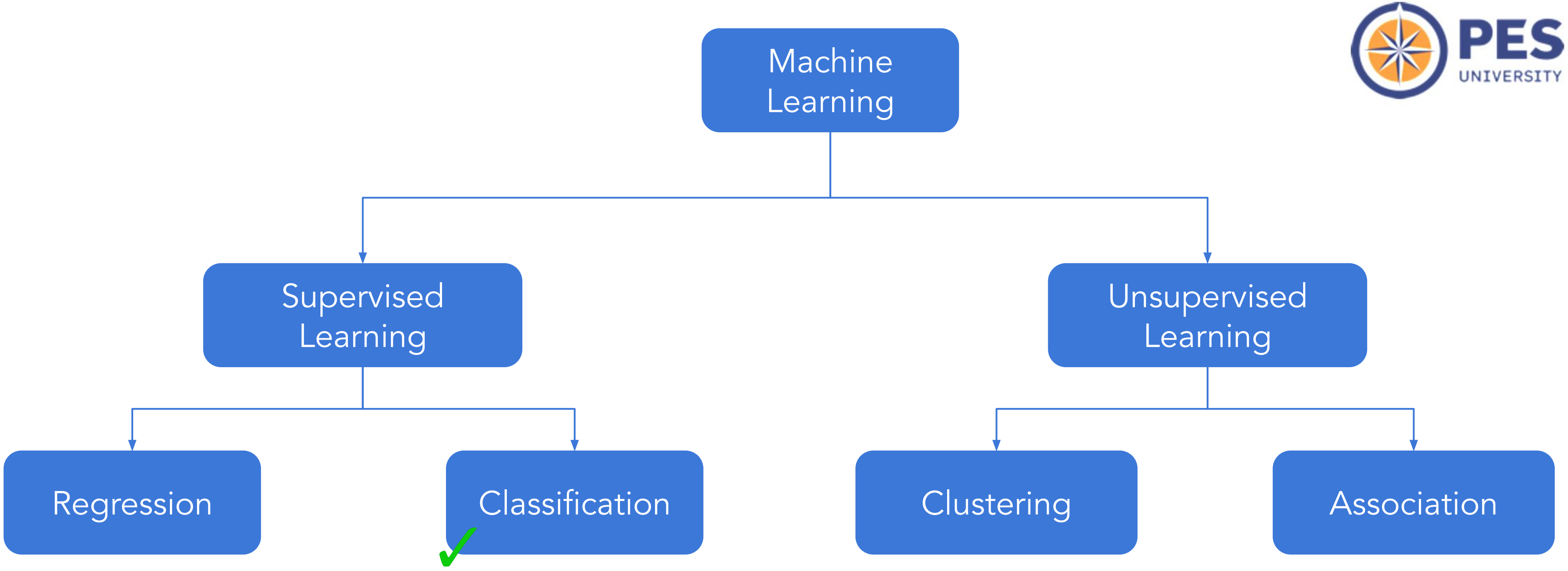
Unsupervised Learning

Regression

Classification ✓

Clustering

Association



Machine Learning

Supervised Learning

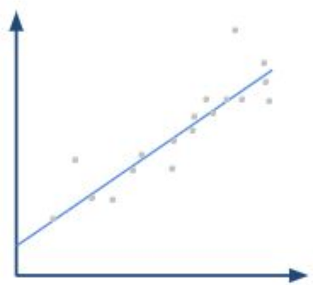
Unsupervised Learning

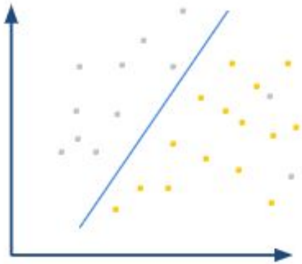
Regression

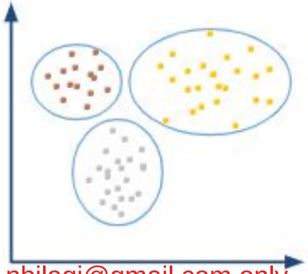
Classification ✓

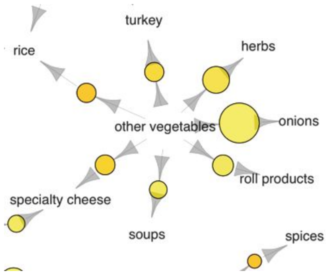
Clustering

Association









Supervised Learning

Supervised learning

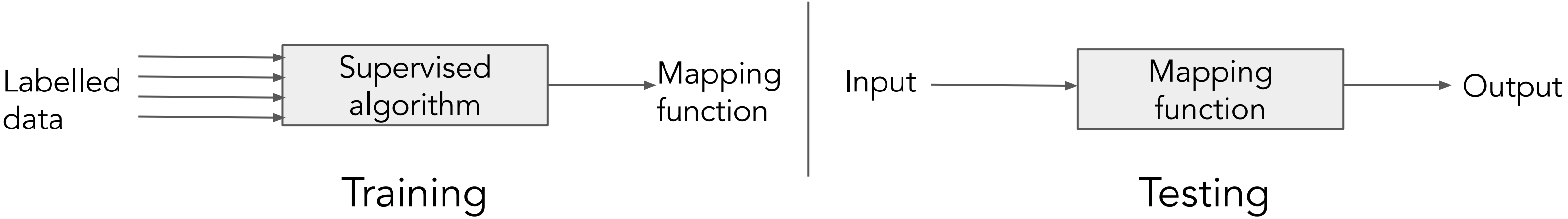
- Supervised learning aims at finding a model that maps the output (target) variable to the input (predictor) variables

- The data used for supervised learning is labelled data i.e. for each set of input data there is known output data, the aim is to find the mapping function

Example: Detection of phishing emails based on certain phrases like ‘You have won a million’. More such phrases are prespecified while training the model. So if a new email also contains a similar phrase such emails can directly be tagged as spam.

Supervised learning

Supervised learning aims at finding a model that maps the output (target) variable to the input (predictor) variables



Labelled data

Supervised algorithm

Mapping function

Training

Input

Mapping function

Output

Testing

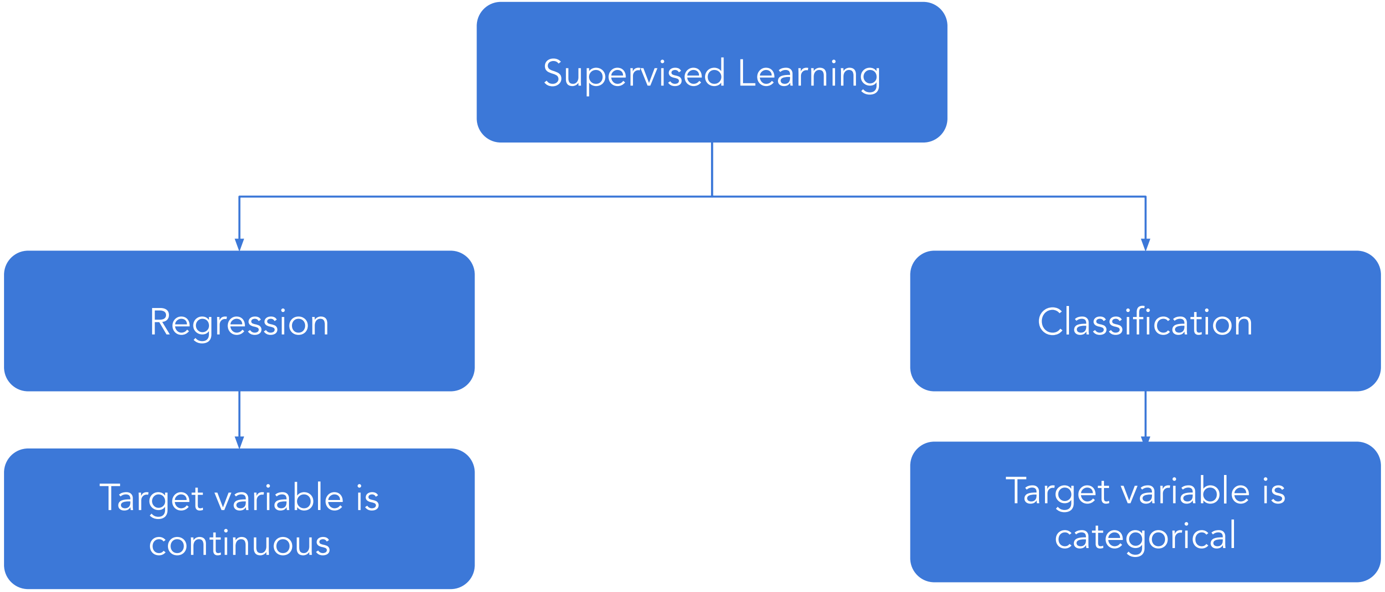
Example: Detection of phishing emails based on certain phrases like ‘You have won million’. More such phrases are prespecified while training the model. So if a new email also contains a similar phrase such emails can directly be tagged as spam.

Supervised learning problems

Supervised learning is mainly used for two types of problems:

- Regression problem
- Classification problem

Regression vs classification



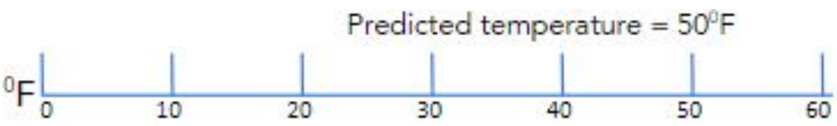
Supervised Learning

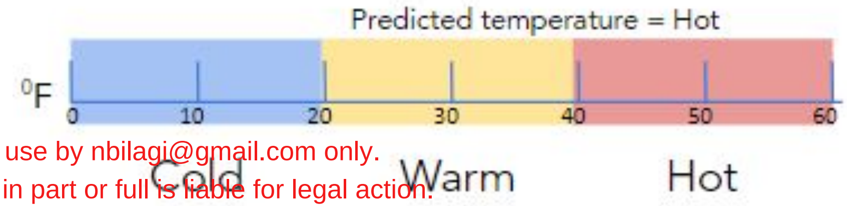
Regression

Target variable is continuous

Classification

Target variable is categorical





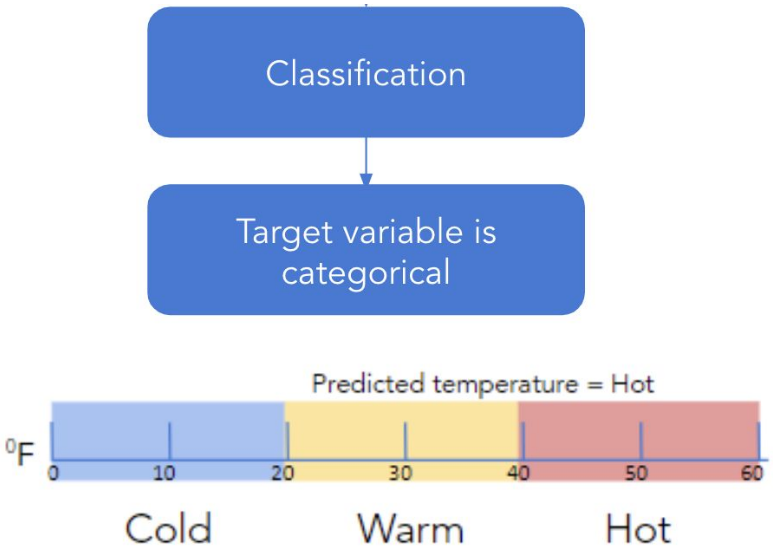
Classification

Class label

For a classification, the target variable has categories.

In the example, Cold, Warm and Hot are the categories of the target variable.

These categories are called the class labels.



Classification

Target variable is categorical

Classification

- An instance is mapped to one of many available labels
- Labels are the fixed number of values taken by the target variable
- The machine learns the pattern from train data where the labels are known for all instances. Then the learning can be used on new data where labels need to be predicted

Example of classification

Consider the example where we have inventory data for an online retailer which includes number of orders, type and demand area for each item. The aim is to classify the items based on if they have a high or low demand.

This process can be automated using machine learning algorithms for classification.

Types of classification

- Binary classification:

  Classification with only two class labels

  Example: Emails can be classified into spam and ham

- Multiclass classification:

  Classification with more than two distinct class labels

  Example: Classification of land based on types of soils

Unsupervised Learning

Unsupervised learning

- Unsupervised learning aims to learn more about given data

- The data used for unsupervised learning has no labels i.e. there is no desired outcome or correct answer given

Example: Consider a dataset with information about flowers. We just know the data has records of flowers and their different characteristics. Using unsupervised learning we can group flowers with similar characteristics and try to find if they belong to a certain species.

Standard Process for Data Science Project

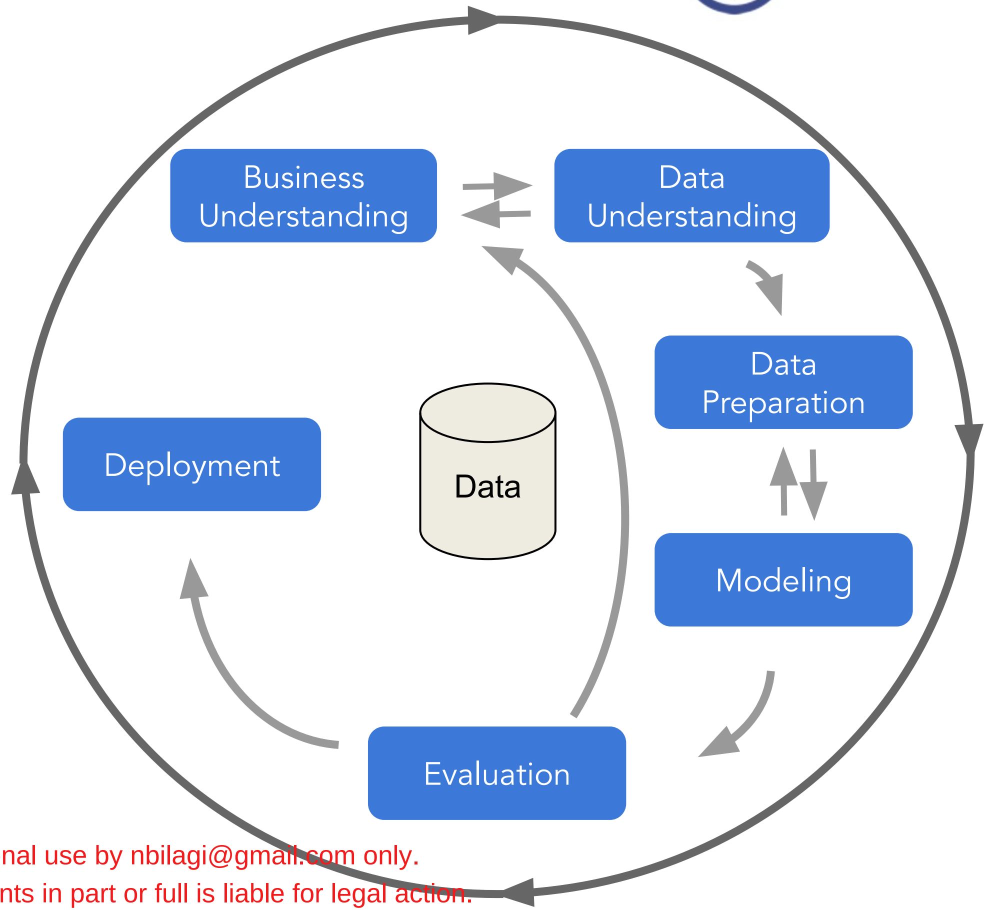
CRISP-DM

Cross Industry Standard Process for Data Mining (CRISP-DM) is a standard process used for data mining

CRISP-DM phases

CRISP-DM breaks data mining into six phases:

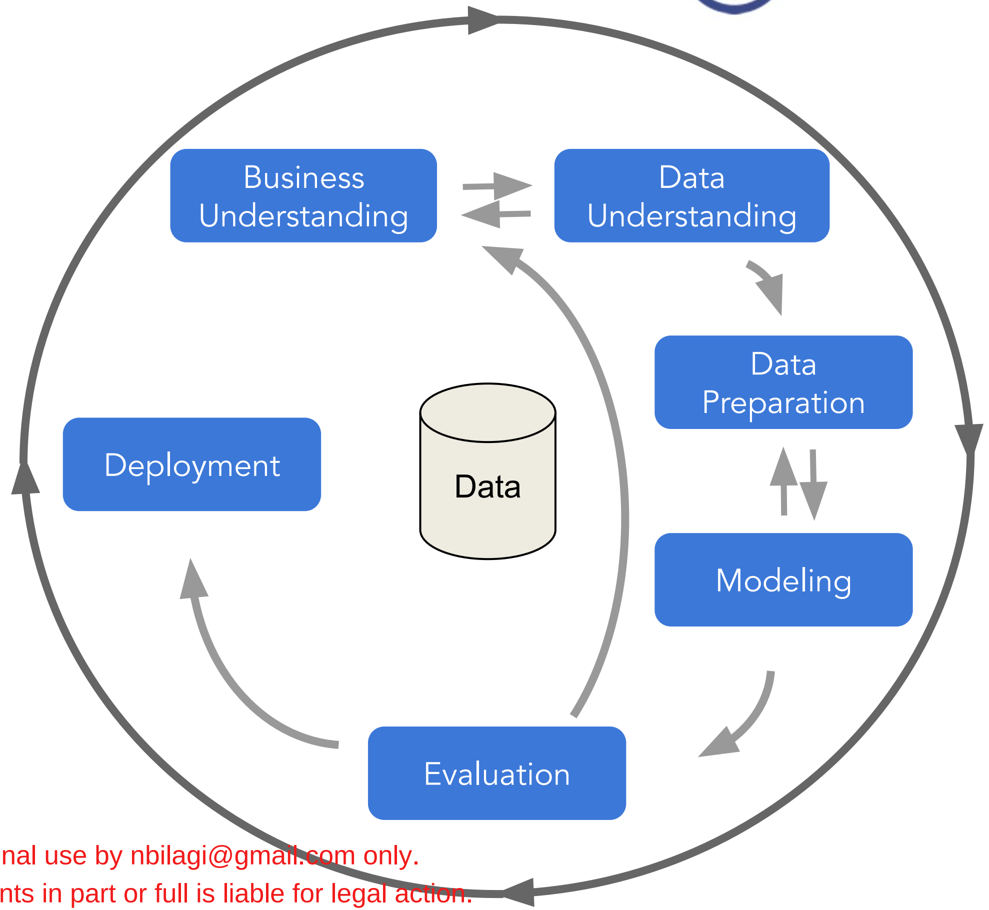
- Business Understanding
- Data Understanding
- Data Preparation
- Modeling
- Evaluation
- Deployment



CRISP-DM phases

CRISP-DM breaks data mining into six phases:

- **Business Understanding**
- Data Understanding
- Data Preparation
- Modeling
- Evaluation
- Deployment



Business understanding

In this phase we define what problem we are trying to solve.

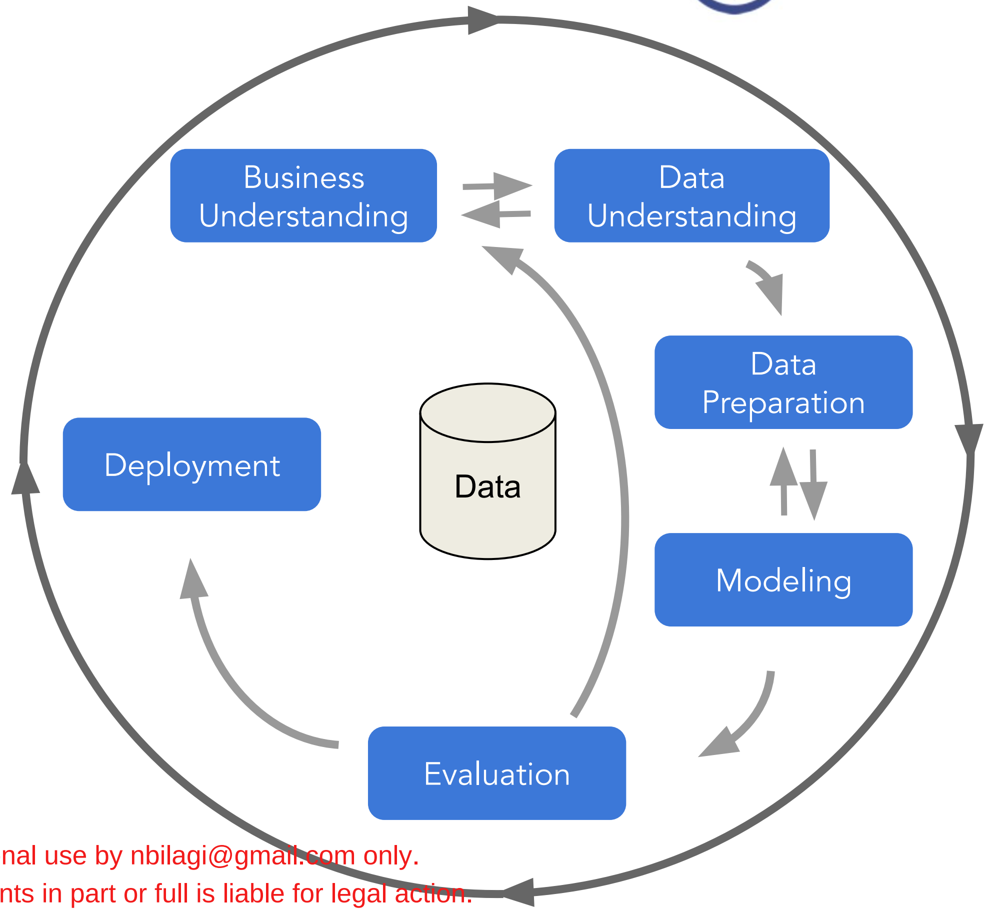
Example: Consider the example where we have inventory data for an online retailer which includes number of orders, type and demand area for each item. The aim is to classify the items based on if they have a high or low demand. So we can define clear business problems like:

- Is type of item related to the demand for the item?
- Can attributes in the considered data be used to classify the entire inventory list with reasonable accuracy?

CRISP-DM phases

CRISP-DM breaks data mining into six phases:

- ○ Business Understanding
- ○ **Data Understanding**
- ○ Data Preparation
- ○ Modeling
- ○ Evaluation
- ○ Deployment



Data understanding

This phase involves understanding the data considered for finding the solution.

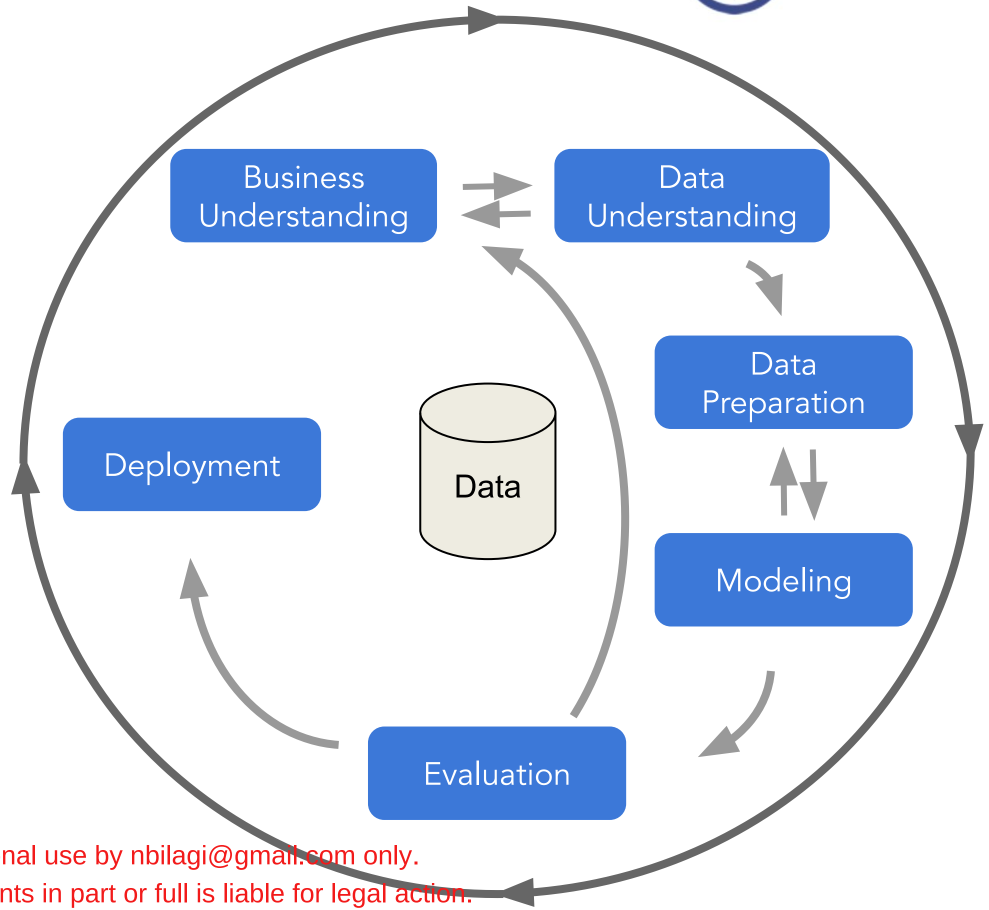
Example: Consider the example where we have inventory data for an online retailer which includes number of orders, type and demand area for each item. The aim is to classify the items based on if they have a high or low demand.

It is important to know if the items are perishable or non-perishable. For instance, items like dairy, cosmetics, and so on, can not be stocked, and hence adequate inventory should be available to meet the demand.

CRISP-DM phases

CRISP-DM breaks data mining into six phases:

- ○ Business Understanding
- ○ Data Understanding
- ○ **Data Preparation**
- ○ Modeling
- ○ Evaluation
- ○ Deployment



Data preparation

This phase involves cleaning and processing the data to be in a format suitable for the model used to solve the problem.

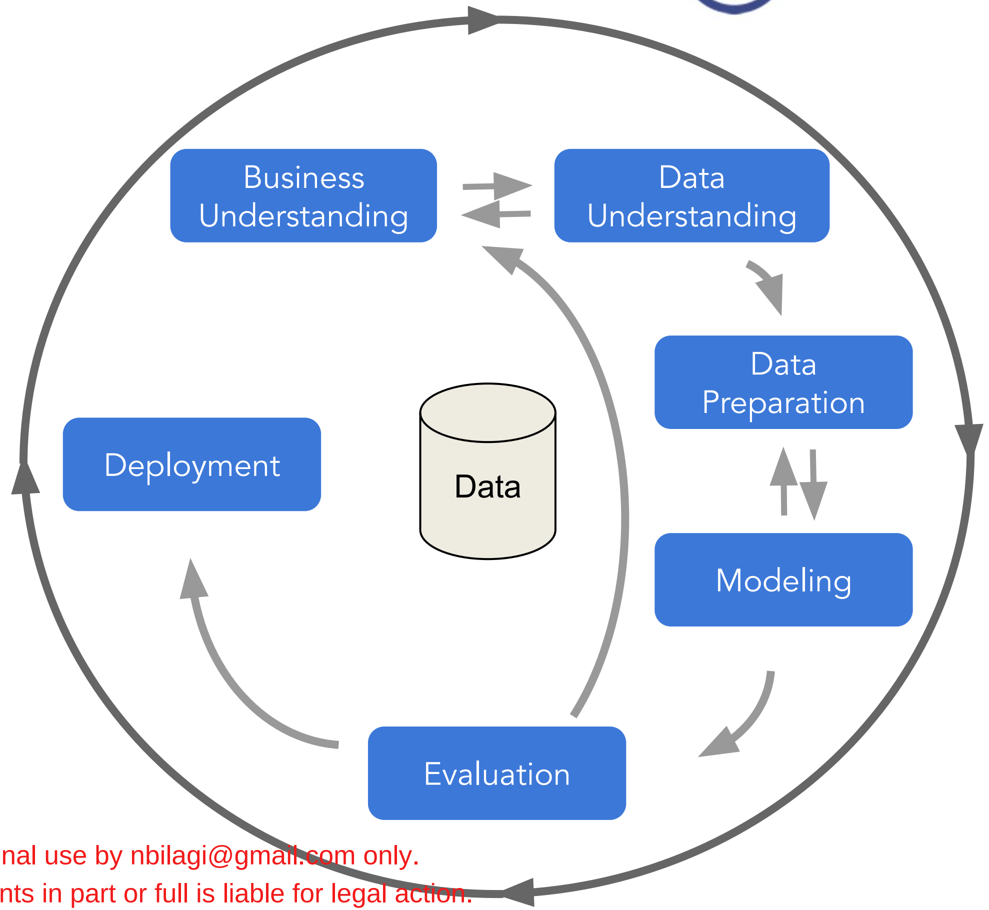
Example: Consider the example to classify the items based on if they have a high or low demand. We can prepare the data as follows:

- Treat the missing values
- The categorical variables need to be dummy encoded
- Check for correlation among variables

CRISP-DM phases

CRISP-DM breaks data mining into six phases:

- Business Understanding
- Data Understanding
- Data Preparation
- **Modeling**
- Evaluation
- Deployment



Modeling

- This phase involves finding the model that captures the solution to the business problem using available data

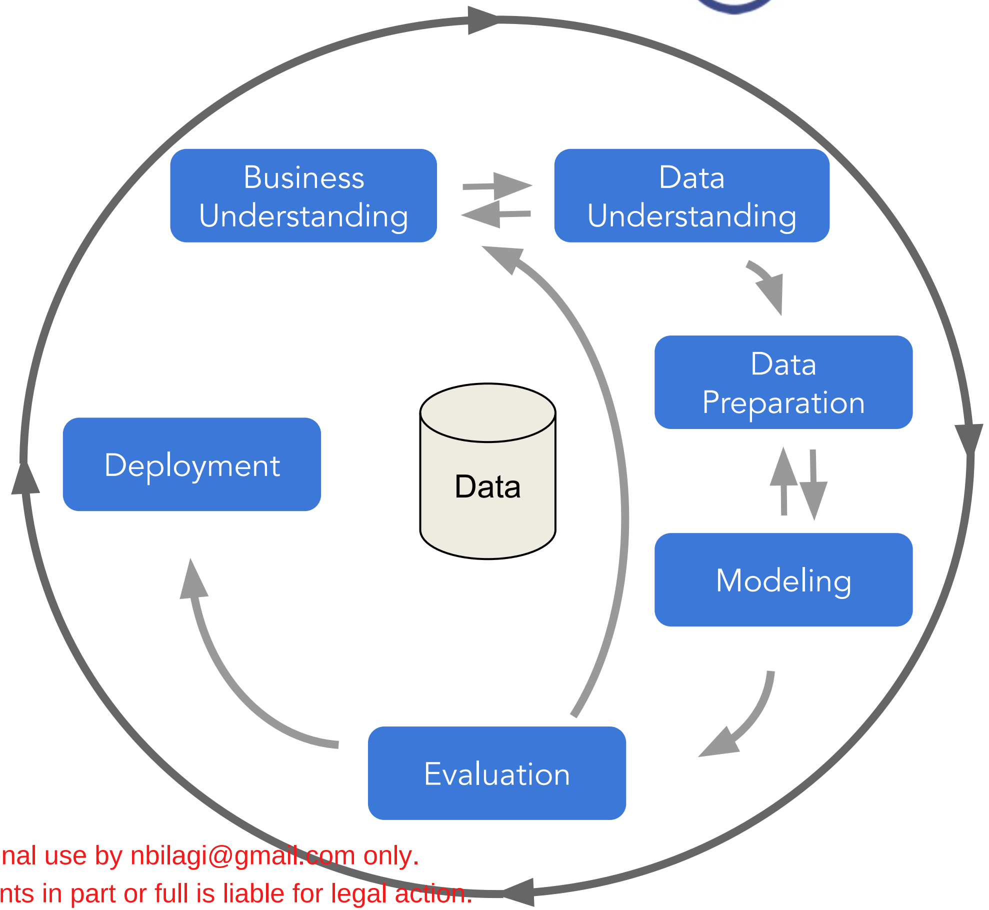
- We may have to try multiple models and go back and forth between data preparation and modelling to choose the correct model

Example: In the modelling phase we try to find a function that maps the attributes like number of orders, type, etc from the data to demand for the item.

CRISP-DM phases

CRISP-DM breaks data mining into six phases:

- ○ Business Understanding
- ○ Data Understanding
- ○ Data Preparation
- ○ Modeling
- ○ **Evaluation**
- ○ Deployment



Evaluation

Once the model is built we need to check how good the model performs on unseen data. This process is done during the evaluation phase.

Example: We can check the model performance on data for which we know the actual demand. Using that data we can compare the predicted and actual values and evaluate.

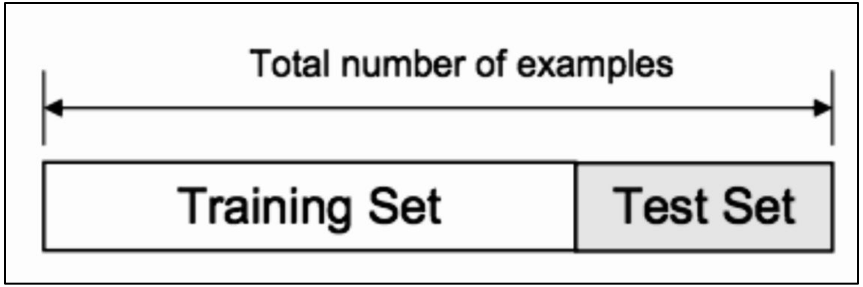
Train-Test Split

Train-test split

- The most straightforward technique that is used to evaluate the performance of a machine learning algorithm is to use different subsets of a dataset
- We can split our original dataset into two parts (train and test)
- Build the model on the training dataset, make predictions on the test dataset and evaluate the forecasts against the expected results

Train-test split

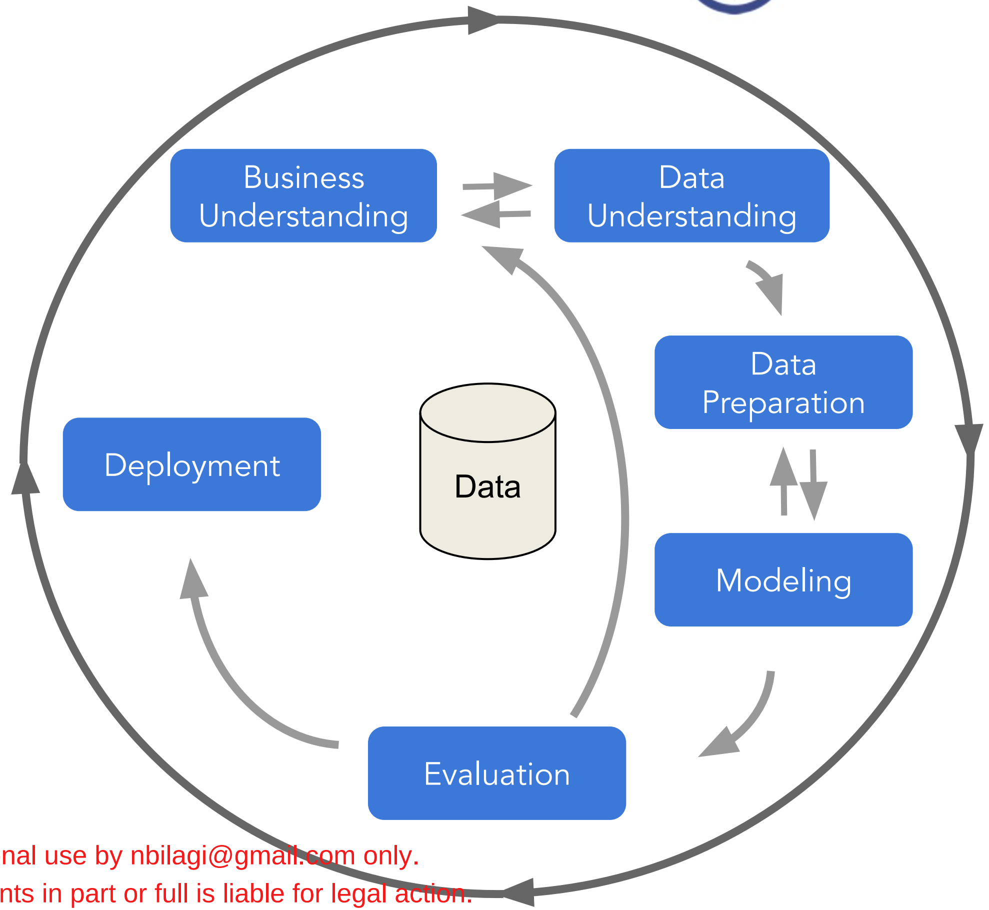
The size of the split can depend on the size of the dataset, although it is common to use 70% of the data for training and the remaining 30% for testing.



CRISP-DM phases

CRISP-DM breaks data mining into six phases:

- Business Understanding
- Data Understanding
- Data Preparation
- Modeling
- Evaluation
- **Deployment**



Deployment

If we are satisfied with the performance of the model from the previous phase we deploy it in the deployment phase

Example: For the considered example of predicting demand for an item, perhaps we could develop an app that takes input as the attribute values for an item and returns the demand for that item to the retailer.

Visiting Basics

Odds vs probability

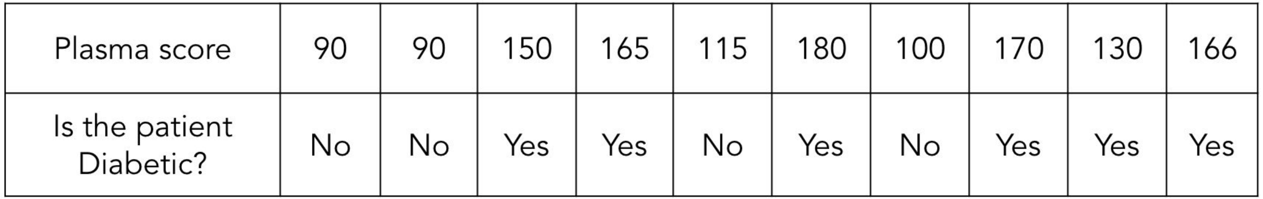
Odds of an event are the ratio of number of observations in favour of an event to number of observations not in favour of the event

$$\text{odds} = \frac{\text{number of observations in favour of the event}}{\text{number of observations not in favour of the event}}$$

Probability of an event is the ratio of number of observations in favour of an event to all possible observations

$$\text{probability} = \frac{\text{number of observations in favour of the event}}{\text{number of observations}}$$

Odds vs probability



Plasma score	90	90	150	165	115	180	100	170	130	166
Is the patient Diabetic?	No	No	Yes	Yes	No	Yes	No	Yes	Yes	Yes

For the above data, the odds of a patient having diabetes is given by,

$$\text{odds} = \frac{\text{number of patients having diabetes}}{\text{number of patients not having diabetes}} = \frac{6}{4}$$

For the above data,, the probability of a patient having diabetes is given by,

$$\text{probability} = \frac{\text{number of patients having diabetes}}{\text{Total number of patients}} = \frac{6}{10}$$

Log of odds

$$\text{odds of having diabetes} = \frac{6}{4}$$

$$\text{log(odds of having diabetes)} = \ln(1.5) = 0.405$$

$$\text{odds of not having diabetes} = \frac{4}{6}$$

$$\text{log(odds of not having diabetes)} = \ln(0.667) = -0.405$$

- As we can see if we only consider odds value, the magnitude for each class value taken by variable is very different
- Hence the log(odds) value is consider so that no matter which class the magnitude is same
- Log of odds is the logit function used in logistic regression

$\text{odds of having diabetes} = \frac{6}{4}$	$\text{odds of not having diabetes} = \frac{4}{6}$
log(odds of having diabetes) = ln (1.5) = 0.405	$\log(\text{odds of not having diabetes}) = \ln(0.667) = -0.405$

log(odds of having diabetes) = ln (1.5) = 0.405

Relation between odds and probability

If P(A) is probability of event A

$$\text{Odds} = \frac{\text{P(A)}}{\text{1-P(A)}}$$

$$\text{Probability} = \frac{\text{odds}}{\text{1+odds}}$$

$$\log(\text{Odds}) = \ln\left(\frac{\text{P(A)}}{\text{1-P(A)}}\right)$$

Odds ratio

- Odds ratio refers to the ratio of odds
- Odds ratio can be used to determine the impact of a feature on target variable
- For our considered example the odds ratio can be calculated as,

$$\text{odds ratio} = \frac{\text{odds of patient having diabetes}}{\text{odds of patient not having diabetes}} = \frac{\frac{6}{4}}{\frac{4}{6}} = \frac{9}{4}$$

Question:

Patients with high sugar diet are considered more susceptible to diabetes. How can we determine whether sugar content in diet has an impact on possibility of a patient getting diagnosed with diabetes? Consider the following sample data.

Sugar content in diet	High	High	Low	High	Low	High	High	Low	High	Low
Is the patient Diabetic?	Yes	No	Yes	Yes	No	Yes	Yes	No	No	No

Solution:

From the given sample data we can calculate:

1. Odds of a patient having diabetes given he has high sugar diet

$$\frac{\text{number of patients having diabetes given he has high sugar diet}}{\text{number of patients not having diabetes given he has high sugar diet}} = \frac{4}{2}$$

2. Odds of a patient having diabetes given he has low sugar diet

$$\frac{\text{number of patients having diabetes given he has low sugar diet}}{\text{number of patients not having diabetes given he has low sugar diet}} = \frac{1}{3}$$

Solution continued:

From 1 and 2 we can calculate odds ratio:

$$\text{odds ratio} = \frac{\text{odds of a patient having diabetes given he has high sugar diet}}{\text{odds of a patient having diabetes given he has low sugar diet}} = \frac{\frac{4}{2}}{\frac{1}{3}} = 6$$

Thus from the odds ratio we can see that patients with a high sugar diet are 6 times more susceptible to diabetes compared to patients who have a low sugar diet.

Binomial Logistic Regression

Question:

Consider the example below about whether or not a patient has diabetes based on plasma score. Can we use linear regression line to predict the whether the patient is diabetic?

Plasma score	90	90	150	165	115	180	100	170	130	166
Is the patient Diabetic?	No	No	Yes	Yes	No	Yes	No	Yes	Yes	Yes

Answer:

The example about whether or not a patient has diabetes based on plasma score is a classification problem. Moreover, the target variable is categorical. Hence, we can not use linear regression line to predict the whether the patient is diabetic. We classify them as diabetic and non-diabetic.

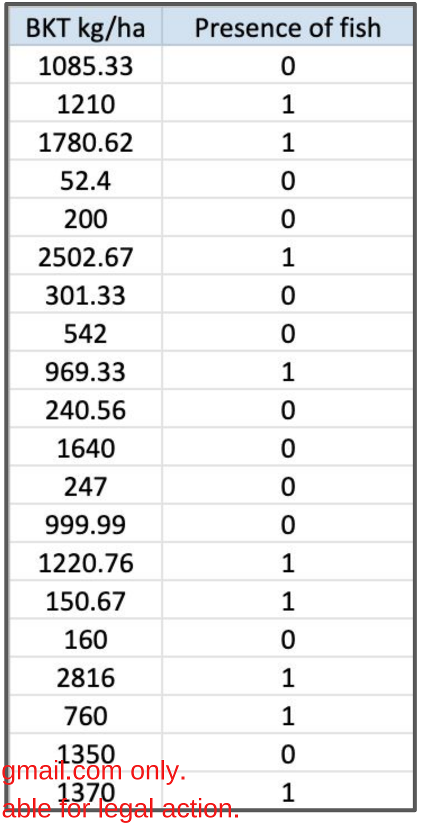
Plasma score	90	90	150	165	115	180	100	170	130	166
Is the patient Diabetic?	No	No	Yes	Yes	No	Yes	No	Yes	Yes	Yes

Logistic regression

- Here on we shall consider the adjacent data

- The data is tell us the presence of one fish depending on the density of other fish in a lake

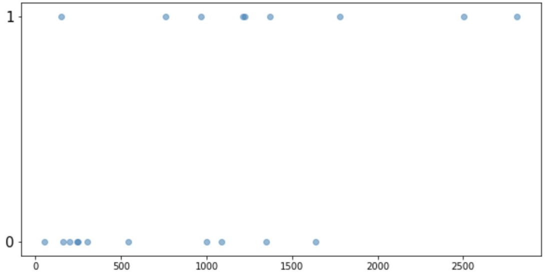
- If they compete with each other, then higher density of one may suggest absence of the other whereas if they are symbiotic, high density of one may promote the other



BKT kg/ha	Presence of fish
1085.33	0
1210	1
1780.62	1
52.4	0
200	0
2502.67	1
301.33	0
542	0
969.33	1
240.56	0
1640	0
247	0
999.99	0
1220.76	1
150.67	1
160	0
2816	1
760	1
1350	0
1370	1

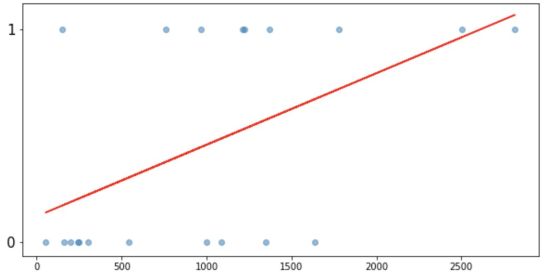
Logistic regression

- Consider the scatter plot of the previous data (data on slide 44)



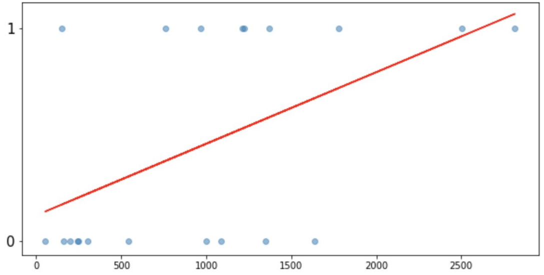
- Fit a linear regression line to it

- Note the line is not a true representative of the data

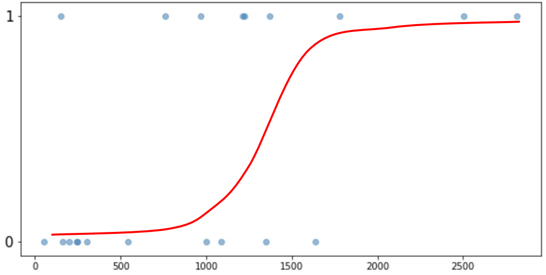


Logistic regression

- A S-shaped curve as in the figure below gives the true relationship



- Such a curve is given be the sigmoid function

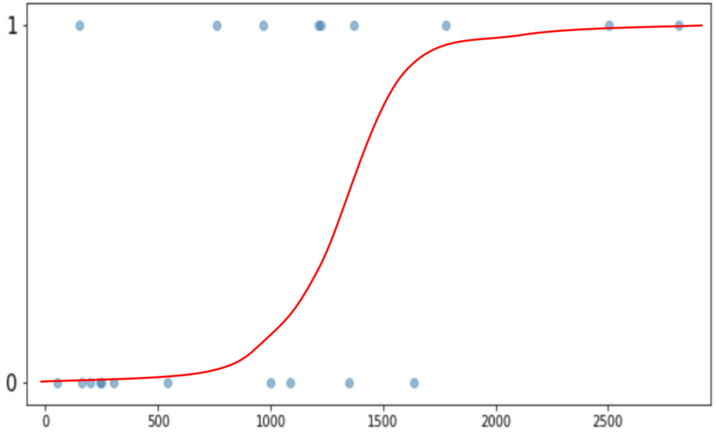


What is a sigmoid function?

- The sigmoid function is a mathematical function which is S-shaped and is given by

$$f(x) = \frac{1}{1 - \exp^{-z}}$$

- It exists between 0 to 1



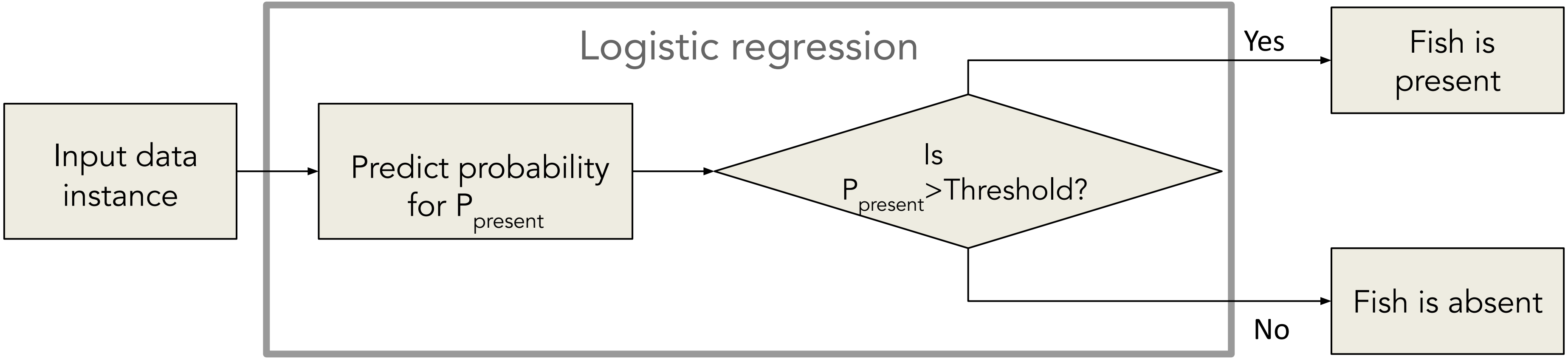
Logistic regression

- Logistic regression is a binary classification algorithm
- It predicts the probability of occurrence of a class label. Based on these probabilities the data points are labelled
- A threshold (or cut-off; commonly a threshold of 0.5 is used) is fixed, then

	Classify as
threshold < probability	Presence of fish
Probability < threshold	Absence of fish

Main steps in logistic regression

Consider that logistic regression is used to identify whether or not a patient is suffering from diabetes



Logistic regression

- The logistics regression is given by

$$\pi(x) = \frac{e^{\beta_0 + \sum_{i=1}^{n} \beta_i x_i}}{1 + e^{\beta_0 + \sum_{i=1}^{n} \beta_i x_i}}$$

- Here, π(x) is the conditional expectation of the outcome given the values for independent variables , i.e. E(Y|X)

- It predicts the probability of occurrence of a class label by fitting the data to a function called logit function, hence called logit regression

Probability as output of logistic regression

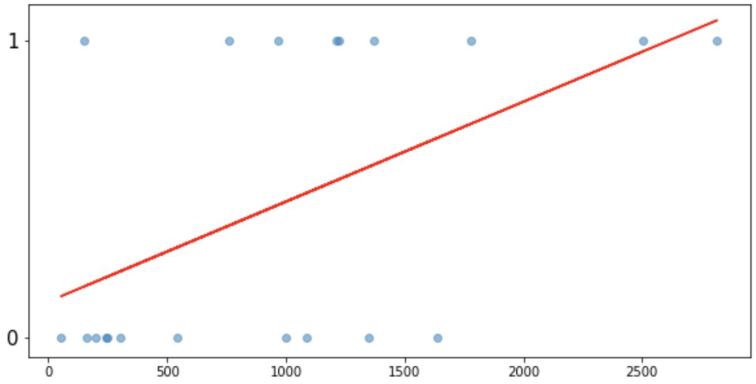
- The logistic regression model is given by $\pi(x) = \frac{e^{\beta_0 + \sum_{i=1}^{n} \beta_i x_i}}{1+e^{\beta_0 + \sum_{i=1}^{n} \beta_i x_i}}$

- Taking limits tending to -∞ on both side, $\lim_{x \to -\infty} \frac{e^x}{1+e^x} = 0$

- Taking limits tending to ∞ on both side, $\lim_{x \to \infty} \frac{e^x}{1+e^x} = 1$

- Thus π(x) lies in-between 0 and 1, i.e. π(x) ∈ [0,1] and can be viewed as probability

Logistic regression



- Since we are predicting probabilities it is important for values to be between 0 and 1
- Consider the points A and B for which values for plasma score are 80 and 177 respectively, the probability values are out of the expected range 0 to 1
- Hence linear regression cannot be directly used to predict probabilities

Usage

- Classification:

  The ICU in a hospital is assigned on priority to high risk patients. Logistic regression can be used to classify the list of patients into high risk and low risk records.

- Profiling:

  Nuclear fuel companies moderate various factors like pressure, temperature, etc. to produce high yielding and low yielding fuels. Based on different parameters for the current day it can be predetermined whether the fuel produced will be high yielding or not. The company can then alter the parameters to produce high yielding fuel everyday.

Logistic regression

- Thus for a binary logistic classification where the target variable takes two values names 0 and 1, we have

Class labels	0	1
P[Y=y | X]	1-π(x)	π(x)

- π(x) denotes the probability that the response is present for the records for some combination of values that the independent variables take, i.e. for X=x

- 1-π(x) denotes the probability that the response is absent for the records for some combination of values that the independent variables take, i.e. for X=x

Linearization

- To estimate the parameter we need to linearize the function. We use the logit transformation

$$\eta = \ln \frac{\pi}{1-\pi}$$

$$\pi(x) = \frac{e^{\beta_0 + \sum_{i=1}^{n} \beta_i x_i}}{1+e^{\beta_0 + \sum_{i=1}^{n} \beta_i x_i}}$$

- The ratio π/(1-π) is called the odds

- Hence the logit transformation is also known as the log-odds

Linearization

We have,

$$\frac{\pi(x)}{1-\pi(x)} = \frac{\dfrac{\exp^{\beta_0+\sum_{i=1}^{n} \beta_i x_i}}{1+\exp^{\beta_0+\sum_{i=1}^{n} \beta_i x_i}}}{1-\dfrac{\exp^{\beta_0+\sum_{i=1}^{n} \beta_i x_i}}{1+\exp^{\beta_0+\sum_{i=1}^{n} \beta_i x_i}}} = \frac{\dfrac{\exp^{\beta_0+\sum_{i=1}^{n} \beta_i x_i}}{\cancel{1+\exp^{\beta_0+\sum_{i=1}^{n} \beta_i x_i}}}}{\dfrac{1}{\cancel{1+\exp^{\beta_0+\sum_{i=1}^{n} \beta_i x_i}}}} = \frac{\exp^{\beta_0+\sum_{i=1}^{n} \beta_i x_i}}{1}$$

That is,

$$\frac{\pi(x)}{1-\pi(x)} = \exp^{\beta_0+\sum_{i=1}^{n} \beta_i x_i}$$

Linearization

Taking natural log on both sides we have,

$$\ln \frac{\pi(x)}{1-\pi(x)} = \ln \exp^{\beta_0 + \sum_{i=1}^{n} \beta_i x_i}$$

$$\ln \frac{\pi(x)}{1-\pi(x)} = \beta_0 + \sum_{i=1}^{n} \beta_i x_i$$

Thus, we have a linear relationship.

Interpreting the parameter

- The logistic regression is given by

$$\ln \frac{\pi(x)}{1-\pi(x)} = \beta_0 + \sum_{i=1}^{n} \beta_i x_i$$

- In a linear regression model, β$_1$ gives the average change in Y associated with a one-unit increase in X

- Whereas, in a logistic regression model, increasing X by one unit changes the log odds by β$_1$, or equivalently it multiplies the odds by exp{β$_1$}

Interpreting for our example

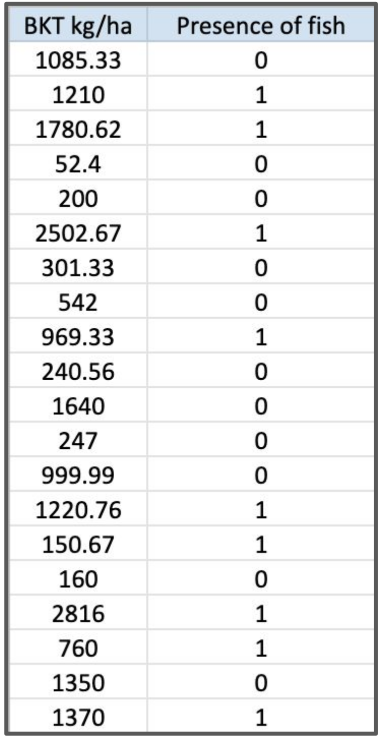
- **The logistic regression is given by**

$$\ln \frac{\pi(x)}{1-\pi(x)} = 0.142 + 1.0018 \ \textbf{BKT kg/ha}$$

- Increasing BKT kg/ha by one unit changes the log odds by 1.0018

OR

- Increasing BKT kg/ha by one unit multiplies the odds by exp{1.0018}

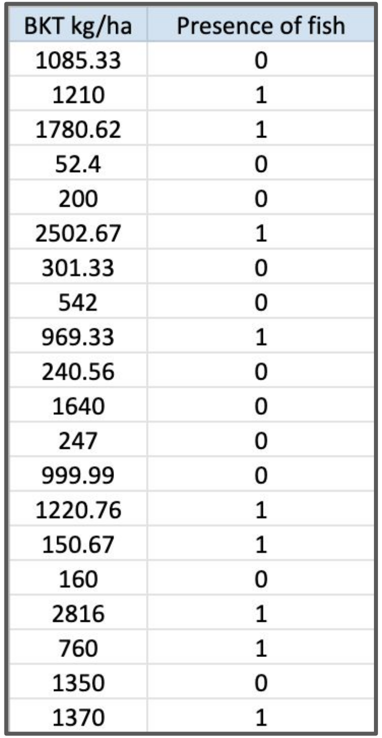


BKT kg/ha	Presence of fish
1085.33	0
1210	1
1780.62	1
52.4	0
200	0
2502.67	1
301.33	0
542	0
969.33	1
240.56	0
1640	0
247	0
999.99	0
1220.76	1
150.67	1
160	0
2816	1
760	1
1350	0
1370	1

Interpreting for our example

	Coefficient	Odds = $e^{\text{coefficient}}$
Intercept	0.142	1.525
BKT kg/ha	1.0015	2.722

- The odds of the fish being present irrespective of BKT is 1.525
- The odds for BKT is 2.722, i.e., for unit increase in BKT the chances of presence of the fish increases by 2.722 times than the chance if it not being present



BKT kg/ha	Presence of fish
1085.33	0
1210	1
1780.62	1
52.4	0
200	0
2502.67	1
301.33	0
542	0
969.33	1
240.56	0
1640	0
247	0
999.99	0
1220.76	1
150.67	1
160	0
2816	1
760	1
1350	0
1370	1

Parameter estimation

- The general form of logistic regression model is

$$y_i = E(y_i) + \epsilon_i$$

where the observations y_i are independent bernoulli random variables

- The expected value of y_i's is

$$E(y_i) = \pi_i = \frac{e^{\beta_0 + \sum_{i=1}^{n} \beta_i x_i}}{1 + e^{\beta_0 + \sum_{i=1}^{n} \beta_i x_i}}$$

Parameter estimation

- The method of least squares fails

- Since, unlike the linear regression, the closed form solution does not exist

where a closed form solution is an exact solution evaluated with a fixed number of operations

- So we use the method of maximum likelihood estimation

Question:

Consider the example below about whether or not a patient has diabetes based on plasma score. Since the logistic regression equation for this example will represent a straight line can we determine the residual values?

Plasma score	90	90	150	165	115	180	100	170
Is the patient Diabetic?	No	No	Yes	Yes	No	Yes	No	Yes

Solution:

For the considered example the equation of logistic regression line is given by,

$$\ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 \text{Plasma Score}$$

p = probability of a patient having diabetes

$\ln\left(\frac{p}{1-p}\right)$ = logit function

β_0 = Intercept value

β_1 = Coefficient for variable Plasma Score

Plasma Score = Values taken by variable Plasma Score

Solution continued:

The value of logit function i.e the LHS for a patient who has diabetes is given by

$$\ln\left(\frac{p}{1-p}\right) = \ln\left(\frac{1}{0}\right) = \ln(1) - \ln(0) = 0 - (-\infty) = +\infty \qquad \ldots\text{(i)}$$

The value of logit function i.e the LHS for a patient who has diabetes is given by

$$\ln\left(\frac{p}{1-p}\right) = \ln\left(\frac{0}{1}\right) = \ln(0) - \ln(1) = -\infty - 0 = -\infty \qquad \ldots\text{(ii)}$$

Solution continued:

From i and ii we can infer that the logistic regression line will approach $+\infty$ and $-\infty$ values.

So the residuals will also take infinite values and hence cannot be determined.

Why we cannot use least squares for logistic regression?

- The residual values for the logistic regression line cannot be determined
- Least squares optimization cannot be performed without determining residuals for the line
- So we cannot use least squares optimization for logistic regression and instead use an optimization approach called maximum likelihood estimation
- Note: least squares for linear regression is a special case of maximum likelihood estimation

Maximum likelihood estimation

- The method of Maximum Likelihood Estimation (MLE) is a method of estimating the parameter of a function by maximizing the likelihood function

- The likelihood function is the joint probability density function of the sample

- For binomial logistic regression the data follows the bernoulli distribution. So the probability distribution is given by

$$f(y_i) = \pi_i^{y_i}(1 - \pi_i)^{1-y_i}$$

where y_i takes value 0 or 1 and π_i is the probability

Maximum likelihood estimation

- The likelihood function (L) is given as

$$L = \prod_{i=1}^{n} f(y_i) = \prod_{i=1}^{n} \pi_i^{y_i} (1 - \pi_i)^{1-y_i}$$

since we have an independent sample, L is the product of all probabilities

- Taking log on both sides, we have

$$\ln L = \ln \prod_{i=1}^{n} \pi_i^{y_i} (1 - \pi_i)^{1-y_i}$$

$$\ln L = \sum_{i=1}^{n} y_i \ln \pi_i + \sum_{i=1}^{n} (1 - y_i) \ln(1 - \pi_i)$$

Maximum likelihood estimation

- Taking log on both sides, we have

$$\ln L = \sum_{i=1}^{n} y_i \pi_i + \sum_{i=1}^{n} (1 - y_i)(1 - \pi_i)$$

- So solving, we have

$$\ln L = \sum_{i=1}^{n} [y_i \ln \frac{\pi_i}{1-\pi_1}] + \sum_{i=1}^{n} \ln(1 - \pi_i)$$

- This equation is further solved by numerical method such as the Newton-Raphson method in order to get the estimates

Thank You