Machine learning.

## Predicting Modelling Cases.

→ Movie prediction → Sales prediction
→ Viewing traffic of website → predicting stock market

**Predictive Modelling :**

↳ Making use of past data
↳ Predict the future using past data.

* Recommending movie
  ↳ Horror Movie ↑↑↑ likes much
  P.M task.
* Sales prediction
  ↳ Analysing past Data.
  No future outcome or forcast.
  Not a predictive Modelling task.
  ↳ Detective Analysis.
* traffic website : Not a P.M.
* Stock Market : P.M task

## types of predictive Modelling.

**House price prediction**

- Size
- Locality
- No. of rooms
- Price.

} data → Supervised

**Google News Clustre**

No data → Unsupervised

House Price Prediction  Energy Prediction

Price

↑ Continuous.

Yes
No
Co
→ Discrete

Un-supervised learning.

Clustering  Segmentation

Stages of Predictive Modelling

→ Problem Def
→ Hypothesis Generation
→ Data Extraction/Collection.
→ Data Exploration and Transformation
→ Predictive Modelling
→ Model Deployment/Implementation

Problem Statement-

* Identify the right problem statement, ideally formulate the problem Mathematically.

• Bad Problem Statement.

↳ Screenshot.

# Problem Definition

**Bad Problem Statement:** Want to improve the profitability of credit card customers

- Want to increase the APR of credit cards
- Want to deploy different APR for different segment of customers
- Want to identify the customers segments having lower default rate
- Want to have different APR and other benefits for different customers segments (On expected default rate) to maximize profit

## Hypothesis Generation

List down all possible variables, which might influence problem objective.

#Quality of Model depends on the Hypothesis

⇒

Should Hypothesis Gen be done before (&) after looking at the data.

Hypothesis gen to be done before looking at the data.

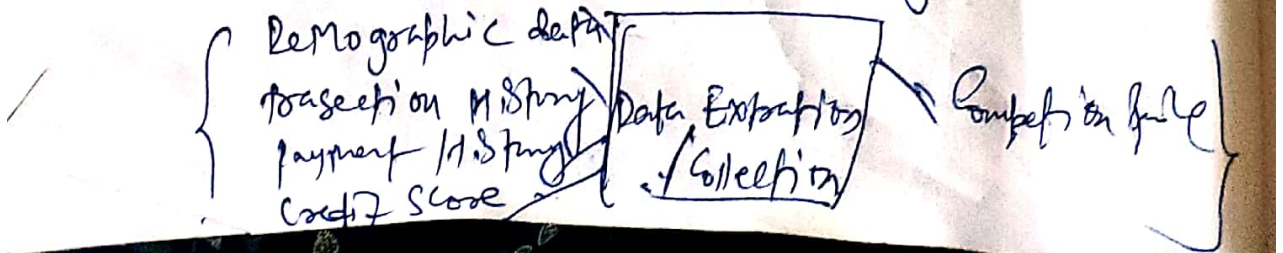
Why?

→ Lets you think of all the factors which might affect the problem without being biased.

→ Stop time wastage in analyzing all available data

## Data Extraction / Collection

# Extract / Collect data from diff sources and combine those for exploration and Model building



# Data Exploration

→ Making sense out of data.

→ Good Analyst → knows his her data.

- → Reading the data.
- → Variable Identification
- → Univariate Analysis (visualization)
- → Bivariate analysis
- → missing value treatment
  - → mean → mode
- → Outlier treatment
- → Variable ~~data~~ transformation

## Reading the data

- → csv file .csv
- → excel file xlsx

→ Read_csv()
→ Shape
→ Columns
→ .head()

## Variable Identification

```
import pandas as pd
file = pd.read_csv('example')
file.shape
file.Head()
```

→ file.dtypes.

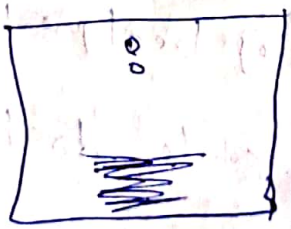
## Univariate Analysis.

1. What is Univariate Analysis.

↳ focus single variable at time
↳ summarize the variable
↳ make sence out of that summary to discover insights, anomolis etc.

2. Why Univariate Analysis.

Box plot (SS)



Univariate helps to detect anomolies in data.

- Continuous variable.
  - ↳ Central tendency and dispersion
    - ↳ Mean/Median/Mode/SD
  - ↳ Distribution of variable
    - ↳ Symmetric/Right skewed/Left skewed.
  - ↳ Presence of missing vale
  - ↳ outliers.

method (Continuous Variable method)

↳ Tabular Method

↳ mean, med, mode, std, and missing value

↳ Graphical method

↳ Distribution of variable presence of outliers.

df.describe() (pandas method)

↳ count, mean, std, min etc)

↳ Histogram → : Distribution of

Boxplot → : outliers Ref. Variable.

pd.DataFrame.hist()

example :-

df.describe() :- only Continuous Variable.

Plotting Histogram

df['Age'].plot.hist() : Distribution of variable

box plot:

df['Age'].plot.box() : outliers

Univariate

↳ Categorical Value.

↳ Sex

↳ Class

↳ Cancer.

- Count
- Count %.

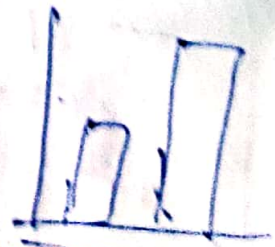
- Tabular Method → Frequency Table
- Graphical Method → Bar Plot.

`df['sex'].value_count()`

Male 577
Female 314 } Count

Male 0.64
Female 0.35 } Count % =

Bar graph



`df['sex'].value_count()/len(df['sex'])`

`df['sex'].value_counts().plot.bar()`

`(df['sex'].value_counts()/len(df['sex'])).plot.bar()`

# Bivariate Analysis

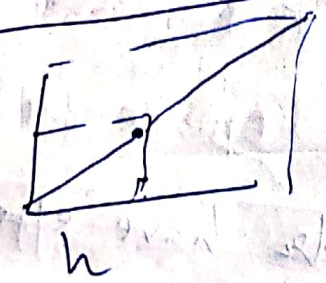
1. What is Bivariate Analysis

- When two variables are studied together for their empirical relationship
- When you want to see whether two variables are associated with each other.

Ex: [ ↑ Weight ∝ ↑ Weight ]

2. Why Bivariate

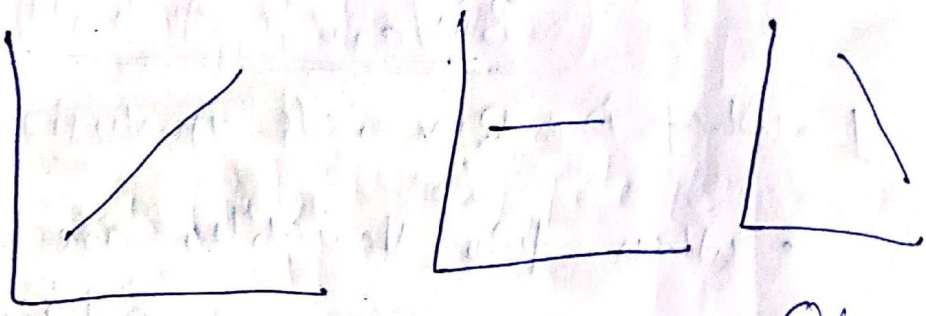
↳ It helps in prediction




3. Type of Bi-variate Analysis

↳ Age - Fare Price (Scatter Plot)
↳ Gender - Mean Age (Bar Plot)
↳ Gender - Survived (Two-way table

Analysis test [ $\rho = \frac{Cov(x,y)}{\sqrt{S_x^2 * S_y^2}}$ ]



Positive - Correlation

No - Correlation

-ve Correlation

* ~~Gender~~

Gender Mean

Analysis test $t = \dfrac{\bar{y}_1 - \bar{y}_2}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}}$

$t < 0.5$

* Does gender have effect on the several suicides.

$$\chi^2 = \frac{\sum (O-E)^2}{E}$$

chi-sq test.

| Suicide | 0 | 1 |
|---|---|---|
| Sex | | |
| f | 81 | 233 |
| m | 468 | 109 |

# Jupyter Notebook

→ import Pandas as pd

→ file = pd.read_csv('titanic.csv')

→ file.head()

→ file.dtypes

# Confi - Continuous (Bivariate Analysis)
(Age) (Fare)

→ file.plot.scatter('Age', 'Fare')

→ df.corr() (or) file.corr()

→ file['Age'].corr(File['Fare'])

→ # Categorical - Continuous Bivariate Analysis

→ file.groupby('sex')['Age'].mean()

→ file.groupby('sex')['Age'].mean().plot.bar()

→ from scipy.stats import ttest_ind

males = file[File['Sex'] == 'male']

female = file[File['Sex'] == 'female']

ttest_ind(males['Age'], female['Age'], nan_policy

## Categorical - Categorical: Bivaite Analysis

→ pd.crosstab(file['sex'], file['survived'])

→ from scipy.stats import chi2_contingency

chi2_contingency(pd.crosstab(file['sex'], file['survived']))

---

## Missing value.

→ Reasons of missing value:

- → Non-Response
- → Error in Data Collection
- → Error in Reading Data.

→ Types of missing Vale

1. missing completing at random (MCAR)
2. missing at random (MAR)
3. missing not at random (MNAR)

---

MCAR.

↳ Missing value have no relation to
1. the variable in which missing value exist
2. Other variable in the dataset.

MAR.

. Missing value have no relation to
1. the variable in which M.V exist
. Missing values have relation to
1. the variable other than the variable in which missing value exists.

MANR.

missing value have relation
1. the variable in which
missing value exists.

# Identifying M.V

. describe()
. isnull()

# Deal with Missing value

1 Imputing 2. Deleting

Imputation
- Conti ← Mean / Median / Regression Model
- Cato ← Mode / Classification Model

# Deletion

- Row-wise Deletion
- Column-wise Deletion

## Jupyter Notebook

```
-> import pandas as pd
-> file=pd.read_csv('data.csv')
-> file.shape
-> file.describe() - count   Pass Sur Ser Age Price fare
                            10  10  9  8  7  6
-> file.isnull().sum() # Identify M.V
-> # drop all row
-> file.dropna().isnull().sum()
                          -> gives sum
-> file.dropna(how='all').shape # remove all missing value
   # drop columns with missing value
->  file.dropna(axis=1).shape
->  file.dropna(axis=1,how='all').shape
```

#How to fill Missing value.

→ file.fillna(0) # Temporary

→ file.fillna(0, inplace=True) # Permanently

→ file['Age'].fillna(0) # Particular Column

→ # mean of column.

→ file['Age'].fillna(file['Age'].mean())

## outliers treatment

→ Reasons for outliers.

1. Data - Entry Errors
2. Measurment Errors.
3. processing Errors
4. Change in underlying population.

→ Types outliers

↳ univariate

↳ Bivariate

→ Identifying outliers

Graphical Method (Boxplot, Scatter Plot

## Treating outliers-

1. Del obs 2. Trasformation and Binning Val.
3. Imputing outlins like miss. Value
4. treat them Seperately

## Jupyter note book:

# Univariate outlier Det.

→ df['Age'].plot.box()

→ df.plot.scatter('Age', 'Fare') # Bivariate outlins

## # Removing outliers-

df=df[df['Fare']<300]

# Replacing outlins in age with mean age value.

df.loc[df['Age']>65, 'Age'] = np.mean(df['Age'])

# formula for predicting outliers

Q1 - 1.5 * IQR or > Q3 + 1.5 * IQ

# Variable transformation

## What is Variable transformation

Variable transformation is a process by which we replace a variable with some fxn of that variable.

Example: replacing a variable x with logarithm.

2. we change the distribution (or) relationship of a variable with others.

## Why?

Variable transformation is used to

1. change the scale of variable.
   if 10 variables are measured in Km and 1 in miles.

2. transformation Non-linear relationships into linear relationships.

3. create symmetric distribution from skewed distribution

Common method of variable Transformation

1. Logarithm - taking log of the value
   → reduces right skewedness of the variable
2. Squared root -
3. ~~Sq~~ Cube root
4. Binning - ~~Using~~
   ↳ used to convert continuous into categorical

Jupyter Notebook.

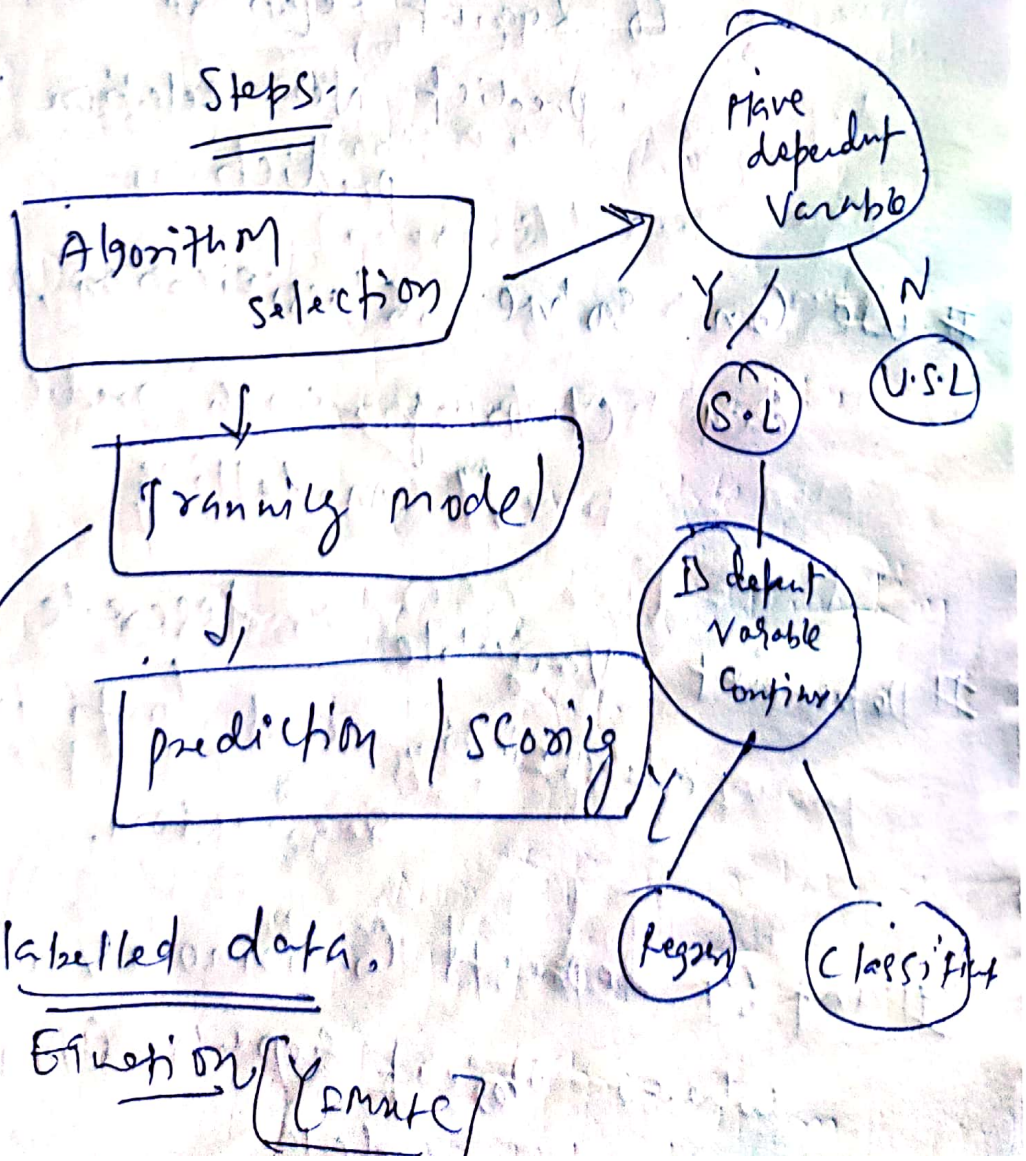
→ df['Age'].plot.hist()
   np.log(df['Age'])
→ np.sqrt(df['Age']).plot.hist()
→ df['Age'].plot.hist()
→ bins = [0, 15, 80]
→ group = ['child', 'Adult']
→ df['type'] = pd.cut(df['Age'], bins, labels=groups)

df.head()

df['type'].value_count()

# Basics of Model

→ Model Building is a process to create a mathematical model for estimating/predicting the future behavior based on past data




- past data (known dependent variable)
- Used to train model

- future data (Unknown dependent variable)
- Used to score

## Making Predictions

↳ Learnt Relations will predict New Relation i.e prediction

# we can solve Classification, Regression and Clustering with predictive Modelling.

# Dependent Variable is discrete in classification and conti in Regression.

# int 64 Represents Continuous Variable whereas Object Represents Categorical data