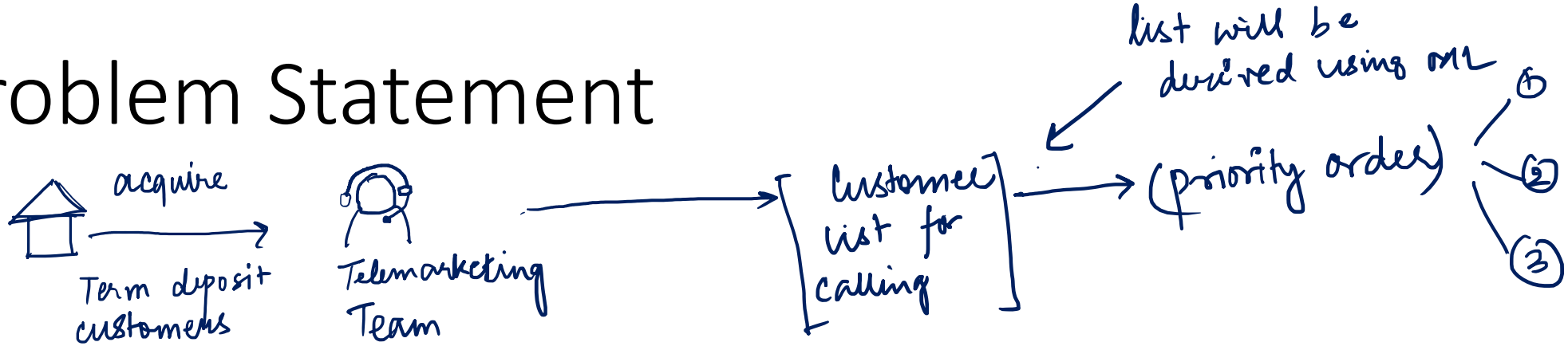
# Predicting the Success of Telemarketing Campaign

Live Session :  Project Discussion

# Problem Statement




**"The Telemarketing Team of a Bank runs campaigns to expand the term deposit portfolio. You are requested to enable prioritization for the Telemarketing team, so that overall responses and ROI of the campaign increases"**

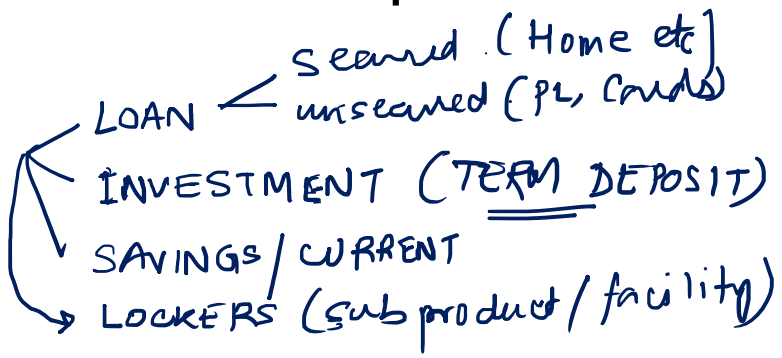
{ROI → Return on Investment
Profit made on the invested amount}

# Some Definitions First

**Marketing Channel :** Avenue of communication to customer for Business Purpose

**Telemarketing:** It is a Marketing Channel where customers are called up for Offers

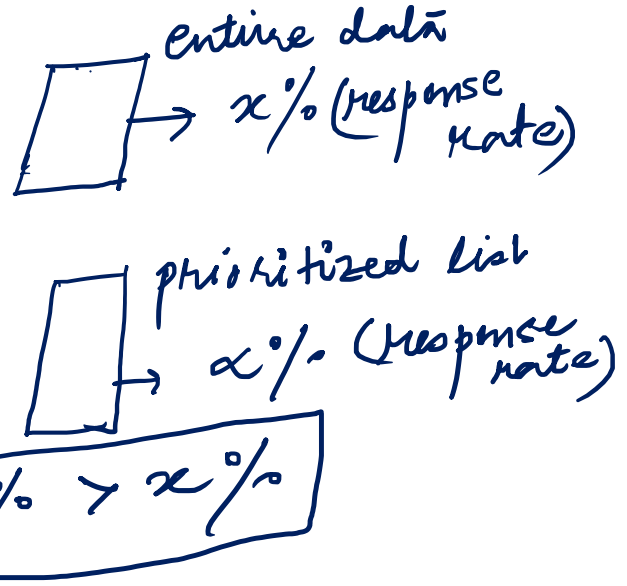
**Portfolio:** A group of customers under a product category




**ROI :** Acronym for Return on Investment

# Why it Matters

- Telemarketing Team generally has fixed resources
- The team will have a fixed call handling capacity in a given time frame (suppose x calls/day)
- If the Telemarketing team receives a pre-selected list of customers to call, then they can focus on them only
- ROI will increase in 2 ways
  - Reduction of investment by not calling up everyone
  - Increase in rate of response among the (prioritized customer list)

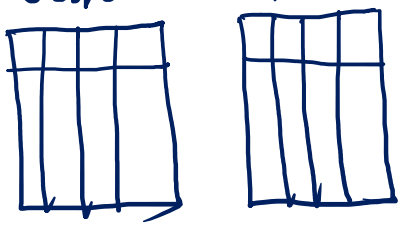


# Data Source

- **This dataset is based on "Bank Marketing" UCI dataset**
- **The full description along with dataset is available here : http://archive.ics.uci.edu/ml/datasets/Bank+Marketing**
- **This dataset is enriched with a few social and economic attributes**
- **Due to confidentiality clauses all attributes are not mentioned**
- **The binary classification goal is to predict if the client will subscribe a bank term deposit**

# Data Capture Process

Jan Feb Mar Apr May June . . . Dec (Monthly data)



"Suppose that the Telemarketing team has called up per customers in Jan and Feb"

| Prospect id | Called (Y/N) | Age | education |
|---|---|---|---|
| abc 123 | 1 | - | - |
| def 234 | 0 | - | - |

→ every month · (Jan & Feb)

① (Jan & Feb) who have "called = Y" —— (A)

② A → Left join (Mar, Apr, May) . .
if customer responded in Mar, Apr
May → Target = 1 or Target = 0

③ · df['Target'].mean() → RESPONSE RATE

→ Purpose for any ML project → Utilize the history and predict the future

[Jan – Feb] ⟶ Mar, Apr, May

↓ ↓

Historical period — Future period

{This is where independent variables come from} {This is where target comes from}

"Utilize last 2 month's of data to predict chance of success in the next 3 months" → Sample/example prob statement.

# Data Description

| Variable | Description |
|---|---|
| Age | Age of Customer |
| Job | Type of Job (Categorical : “admin”, ”blue-collar”, “entrepreneur”, “housemaid”, ”management”, "retired","self-employed","services","student","technician","unemployed","unknown") |
| Marital | marital status(categorical:"divorced","married","single","unknown") |
| education | (categorical: “basic.4y”, “basic.6y”, “basic.9y”, “high.school”, “illiterate”, “professional.course”, “university.degree”, “unknown”) |
| default | default: has credit in default? (categorical: “no”, “yes”, “unknown”) |
| housing | housing: has housing loan? (categorical: “no”, “yes”, “unknown”) |
| loan | loan: has personal loan? (categorical: “no”, “yes”, “unknown”) |
| contact | contact: contact communication type (categorical: “cellular”, “telephone”) |
| month | month: last contact month of year (categorical: “jan”, “feb”, “mar”, …, “nov”, “dec”) |
| day_of_week | day_of_week: last contact day of the week (categorical: “mon”, “tue”, “wed”, “thu”, “fri”) |
| duration | duration: last contact duration, in seconds (numeric). |
| campaign | campaign: number of contacts performed during this campaign and for this client (numeric, includes last contact) |
| pdays | pdays: number of days that passed by after the client was last contacted from a previous campaign (numeric; 999 means client was not previously contacted) |
| previous | previous: number of contacts performed before this campaign and for this client (numeric) |
| poutcome | poutcome: outcome of the previous marketing campaign (categorical: “failure”, “nonexistent”,“success”) |
| emp.var.rate | emp.var.rate: employment variation rate — (numeric) |
| cons.price.idx | cons.price.idx: consumer price index — (numeric) |
| cons.conf.idx | cons.conf.idx: consumer confidence index — (numeric) |
| euribor3m | euribor3m: euribor 3 month rate — (numeric) |
| nr.employed | nr.employed: number of employees — (numeric) |
| y | target variable - has the client subscribed to term deposit (1/0) |

# Exploratory Data Analysis

## Data Understanding – Univariate Analysis

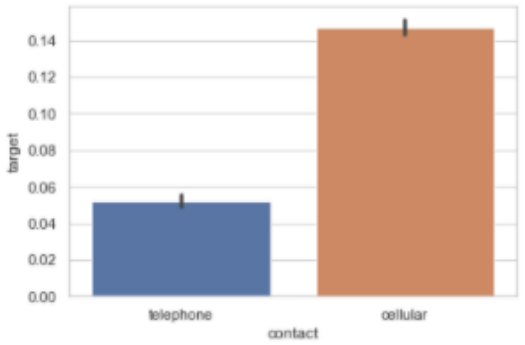
- How well populated is the data?
- How much variation is there in the variables given to you?
- What are the unique levels for the categorical variables
- What is the proportion of missing data for the given raw variables? Discard variables that are more than 25% missing in values
- Missing Value Imputation Methods : Mean for Numeric and Mode for Categorical

# Bi-Variate Plots

**Visualizations to reveal Bi-Variate data patterns and relationships**

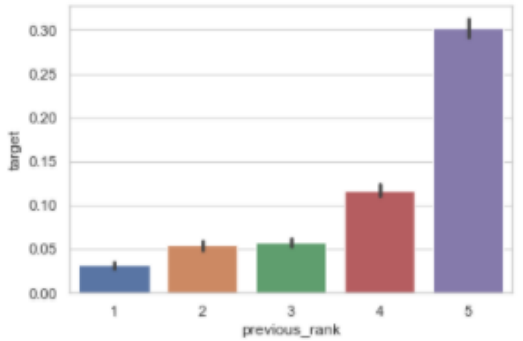
- Plot Categorical Independent Variable Levels on X- Axis
- Plot Average Value of Dependent Variable on Y-Axis
- In case of Continuous Independent variable, break them into ranks and then plot them on the X-Axis

Avg Value of Y




Categorical Independent Variable

Avg Value of Y




Ranks of Continuous Variable

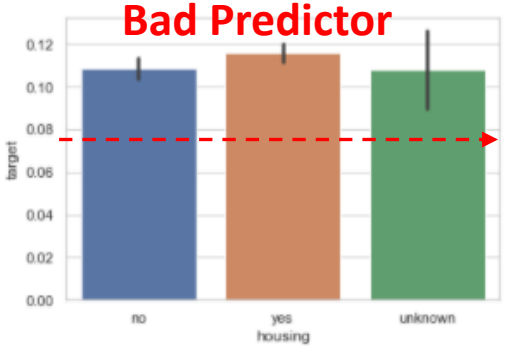
# Insights : Bi-Variate Plots

**We are looking for variables/features that differentiates the average value of Y**

- Based on plots we can determine **which variables can predict Y**
- We can even group up some levels or club variable to achieve differentiation
- Examples of good and bad outcome is given below to clearly explain the expectation

**Bad Predictor**

Avg Value of Y

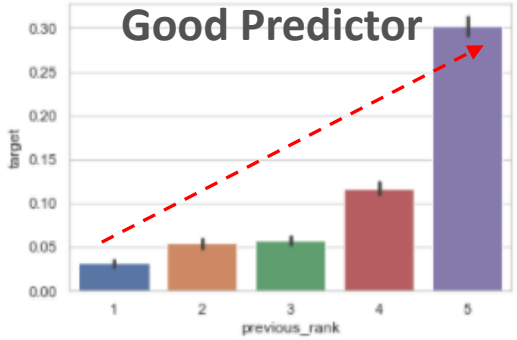



No Slope

Categorical Independent Variable

**Good Predictor**

Avg Value of Y



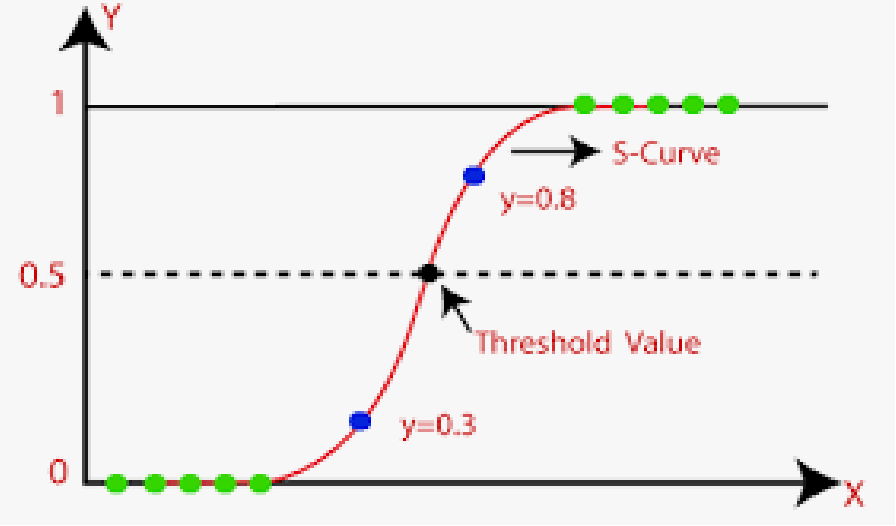

Good Slope

Ranks of Continuous Variable

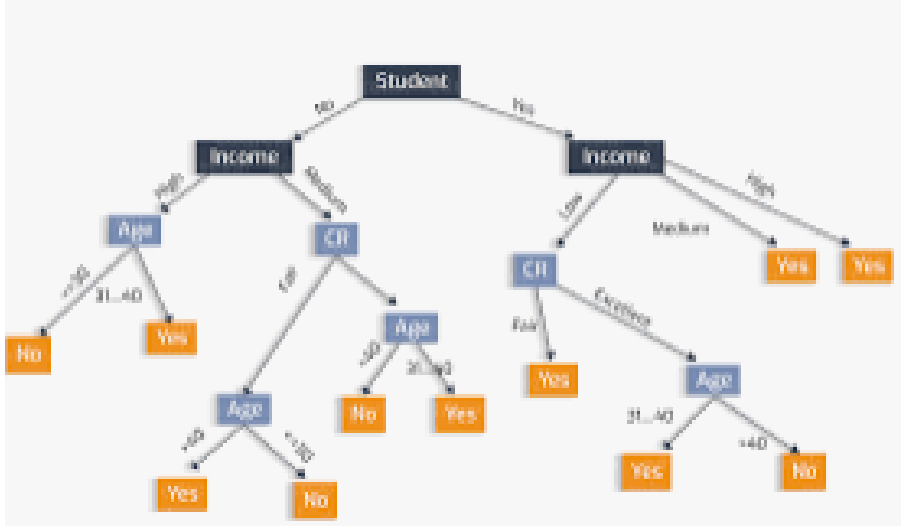
# Classification Algorithms

**Bi-directional approach : Both classifier algorithms (Statistical and ML Based) can be tried out and results compared for final deployment**

## Binomial LR Algorithm




## Tree Based ML Algorithm



# Precision , Recall and F1 Score

↓ out of everything that actually happened how many are predicted

**Precision**

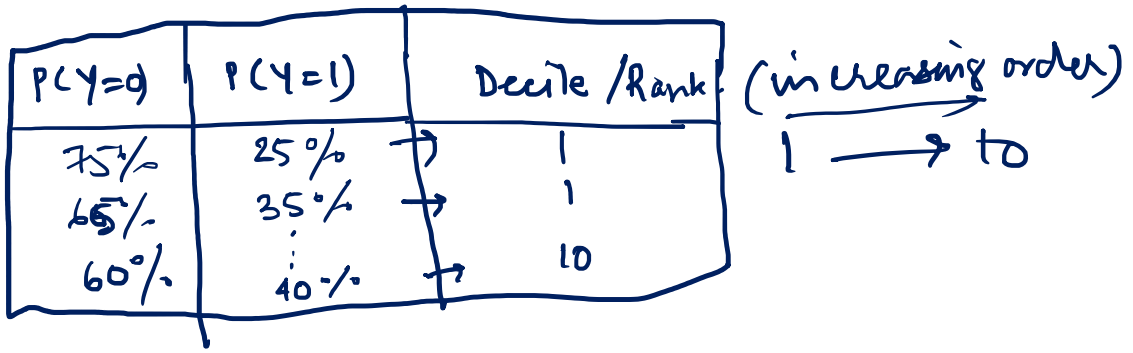
| | | Actuals | | |
|---|---|---|---|---|
| | | Positive | Negative | |
| Model Predicted | Positive | A | B | Predicted Positive Rate: A/(A+B) |
| | Negative | C | D | Predicted Negative Rate: D/(C+D) |
| | | Sensitivity: A/(A+C) | Specificity: D/(B+D) | Accuracy = (A+D) / (A+B+C+D) |

Out of everything the model is saying how many are true →

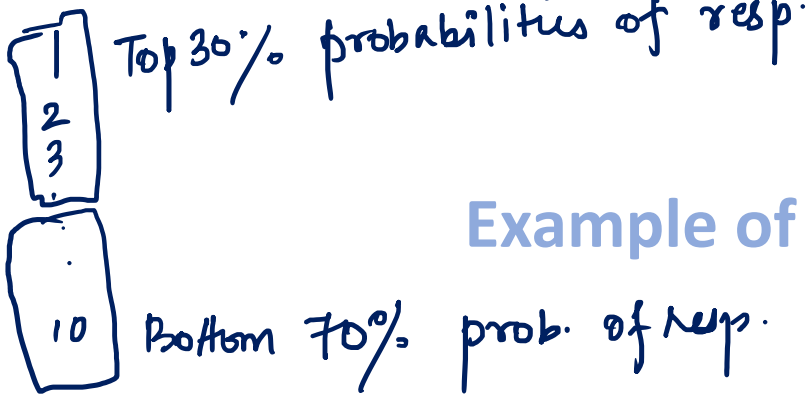
**Recall** (True Positive Rate)

$$\text{F1 Score} = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}$$

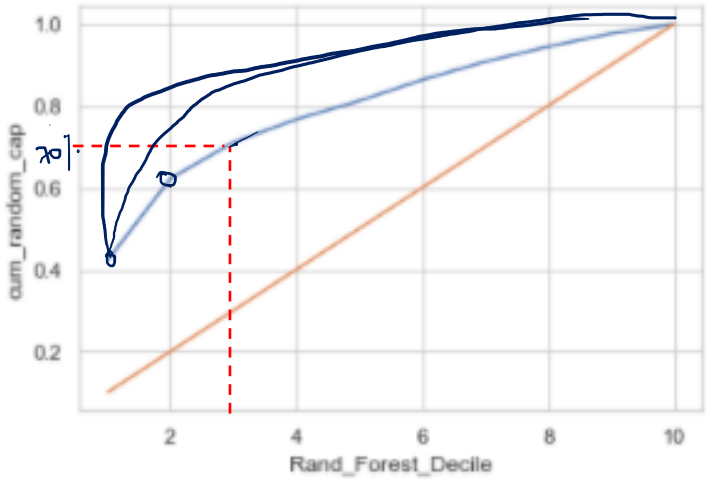
# Gains Chart




- Build deciles based on Predicted Probability
- Summarize Data across the deciles to calculate the Cumulative Actual Event Capture Rate
  - *Target=1 is Event and Target=0 is Non-Event*
- Compare Models based on the Cumulative Event Capture Rate till Decile 3 ✓




**Example of Gains Chart**



A ↓ Audience

P ↓ Priority Order

T ↓ Treatment

M ↓ Measurement → (HYPOTHESIS TESTING)

- i) Whether the model is effective?
- ii) Whether the overall campaign was effective?