upGrad

#LifeKoKaroLift

Data Science Certification Program

upGrad

Course : Machine Learning

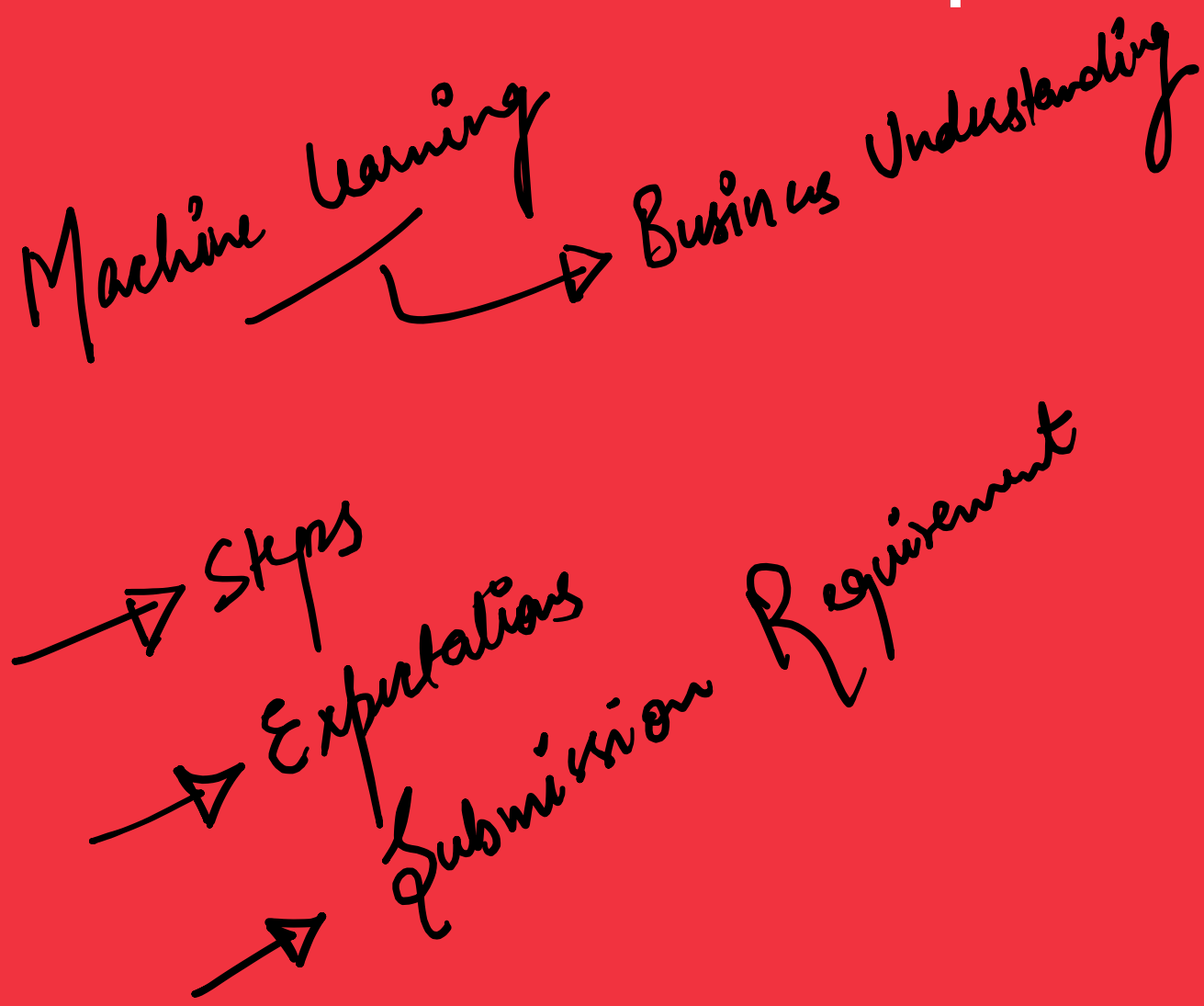
Lecture On : Lead Scoring Assignment

Instructor : Shivam Garg

upGrad

Today's Agenda

1. Problem Statement
2. Assignment Walkthrough
3. Doubt Session



Assignment: Problem Statement

[Ed-tech Industry]

upGrad

Classification problem

→ upgrad

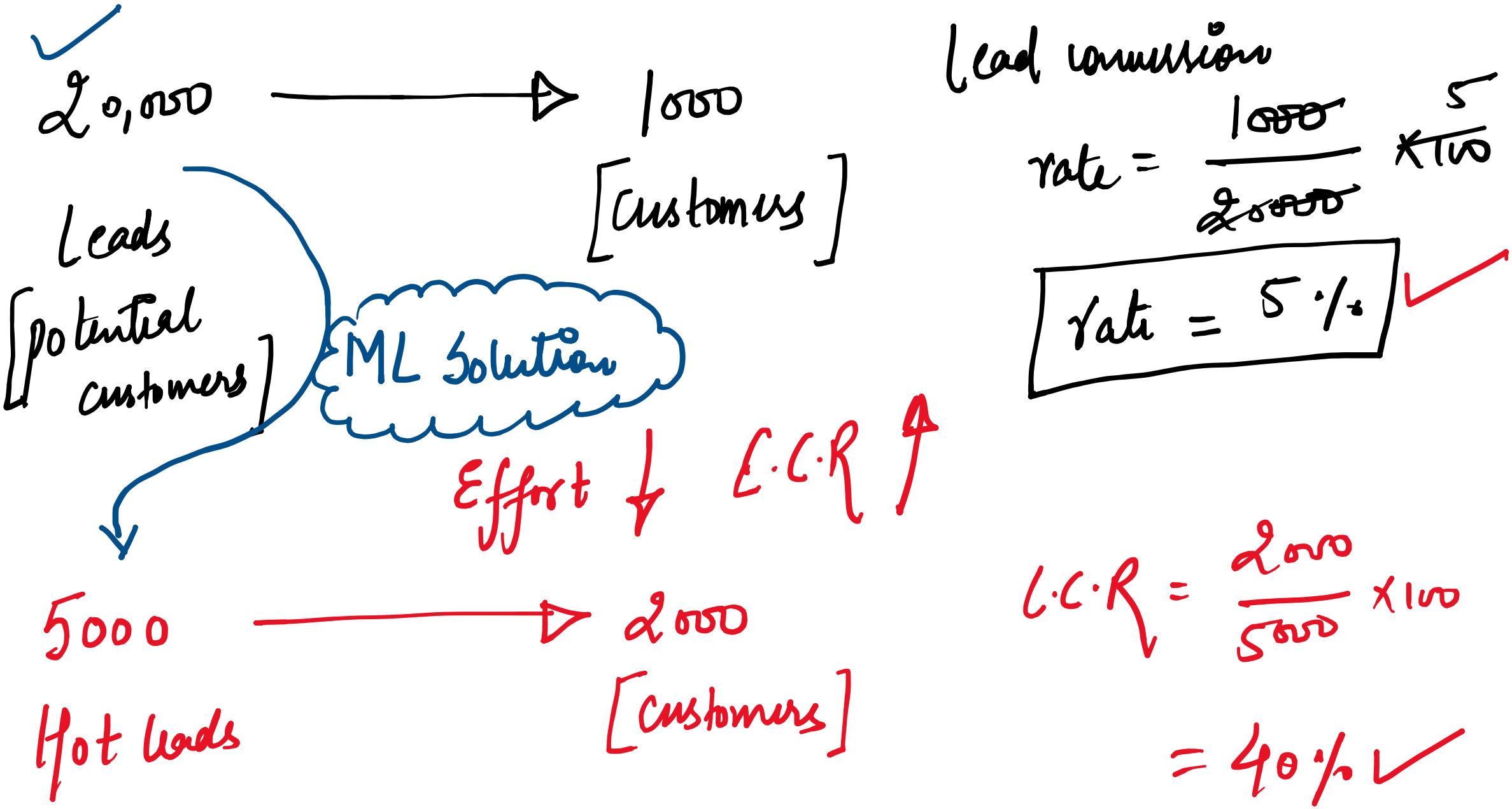
An education company named X Education sells online courses to industry professionals.The company markets its courses on several websites and search engines like Google. Once these people land on the website, they might browse the courses or fill up a form for the course or watch some videos. When these people fill up a form providing their email address or phone number, they are classified to be a lead. Now, although X Education gets a lot of leads, its lead conversion rate is very poor. For example, if, say, they acquire 100 leads in a day, only about 30 of them are converted.

What you need to do?

→ hot leads

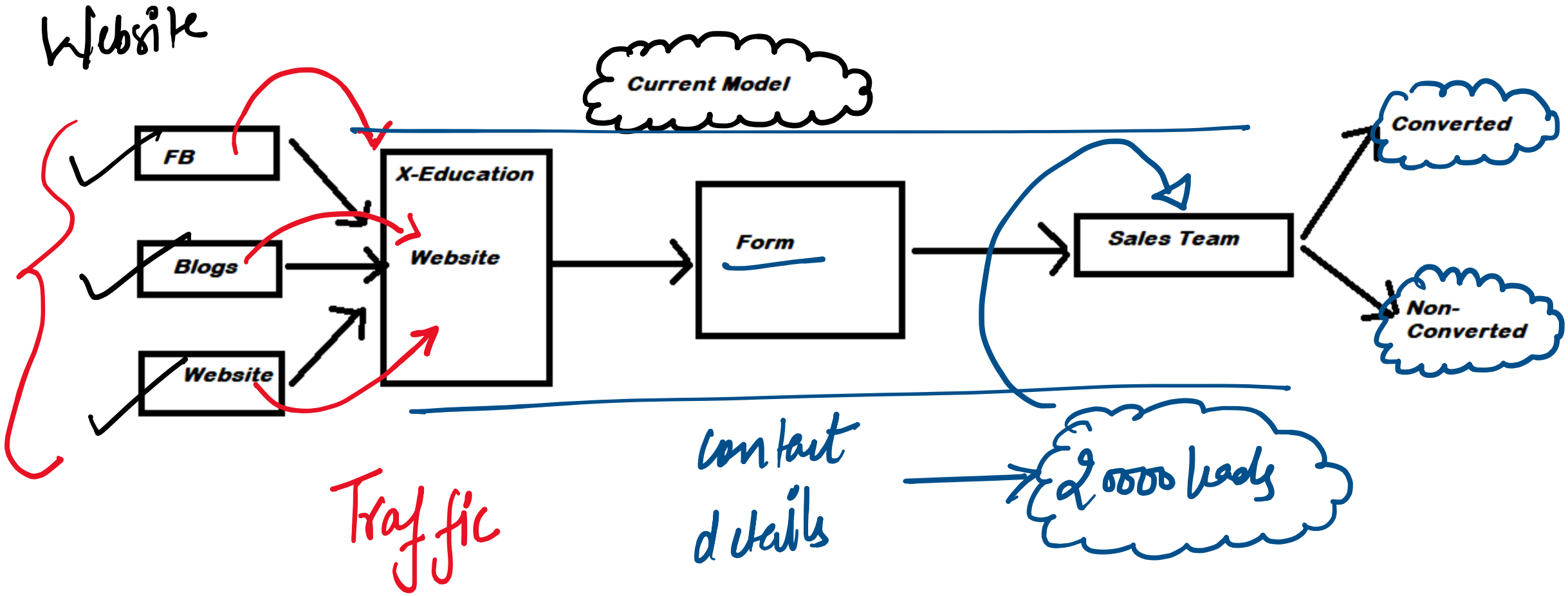
- X Education has appointed you to help them select the most promising leads, i.e. the leads that are most likely to convert into paying customers.
- The company requires you to build a model wherein you need to assign a lead score to each of the leads such that the customers with higher lead score have a higher conversion chance and the customers with lower lead score have a lower conversion chance.

Lead Score = 100 x prob.


20,000
1000
[customers]
Lead conversion
rate = 1000/20000 x 100
5
rate = 5%
Leads
[potential customers]
ML Solution
Effort ↓ L.C.R ↑
5000
2000
[customers]
Hot leads
L.C.R = 2000/5000 x 100
= 40%

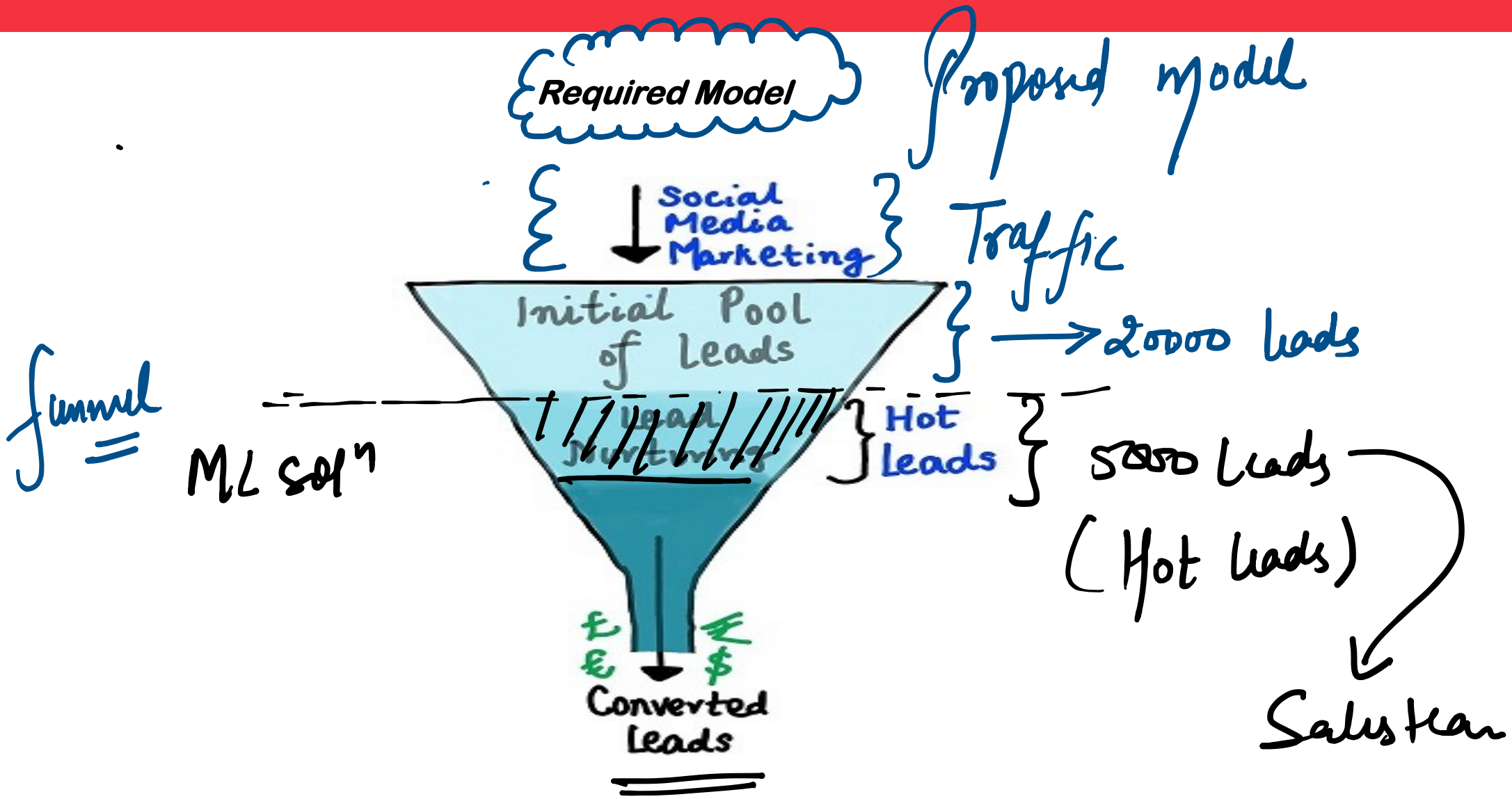
Assignment: Problem Statement

upGrad



Assignment: Problem Statement

upGrad



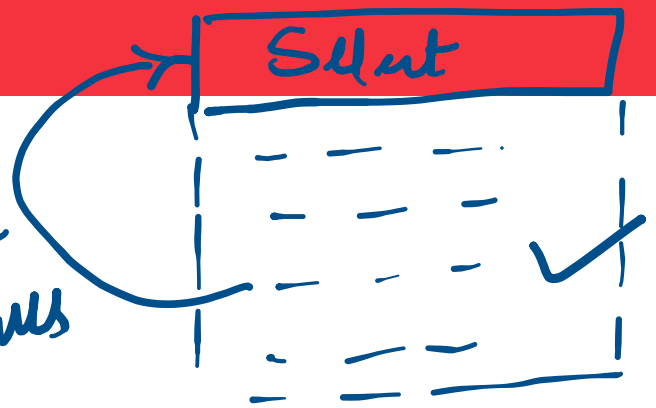
→ Assignment Steps:-

1- Data Cleaning / Data preprocessing.

2- Data preparation } →

3- Model Building & Evaluation }

upGrad



Data Cleaning:

1. Handle the "Select" level that is present in many of the categorical variables.
2. Drop columns that are having high percentage of missing values. Check all the columns before dropping them.
3. Check the number of unique categories in each categorical column. Here you may need to do something.
4. For the columns with less percentage of missing, use some imputation technique.
5. Finally check the percentage of rows retained in data cleaning process.

} Data Sanity Checks

6. Outlier Treatment. & Columns

A

% of missing values

15% ✗ ⟶ 75% ✓

Select ⟶ np.nan } Null

% of missing values ⟶ Threshold (30-50%) ⟶ Drop columns (> threshold)

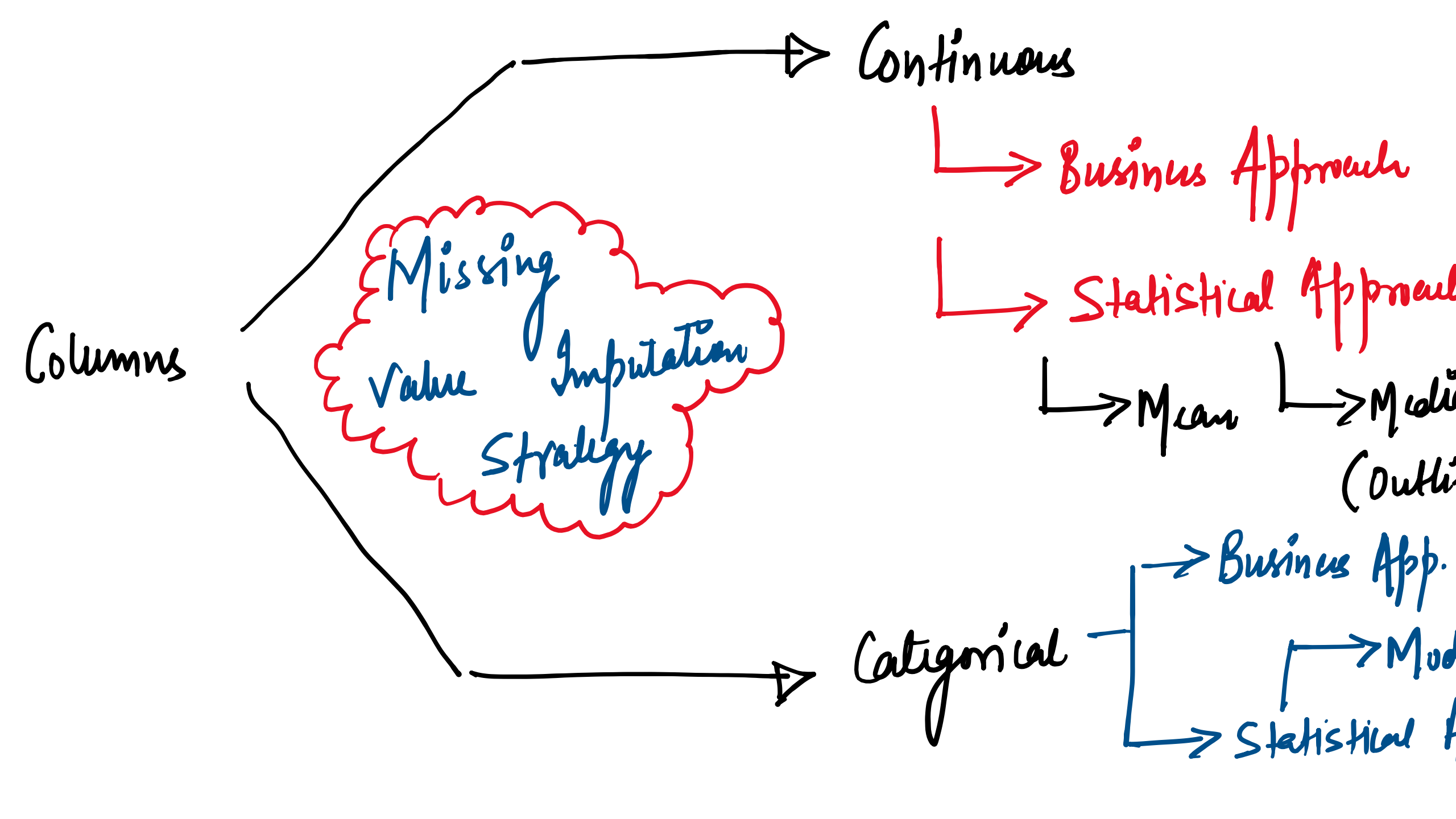
Country (100 countries)

India	75%
U.S	15%
U.K	5%
Australia	2%
Pak	0.1%
Srilank	0.01%
⋮	⋮

97% ⟹ Others

Country (5 categories)

India	75%
U.S.	15%
U.K	5%
Aus	2%
Others	3%


Columns
Missing Value Imputation Strategy
Continuous
Business Approach
Statistical Approach
Mean
Categorical
Business App.
Mod
Statistical

Assignment Steps :

logistic model → linear model

upGrad

Data Preparation:

1. Create dummies for all categorical variables.
2. Perform train-test split.
3. Perform scaling.

[Mandatory]

Data → Train (70%) — Model Building (training)

Data → Test (30%) — Evaluation (Unseen)

Categorical column

No. of categories (n) = 3

No. of dummies = n - 1 = 2

A
B
A
C
B
A

A	B	C
1	0	0
0	1	0
1	0	0
0	0	1
0	1	0
1	0	0

pd.dummies(df[categorical], drop_first=True)

drop_first = True → first column get dropped

Assignment Steps :

p-value < 0.05 VIF < 5

upGrad

30 → (RFE) → 15 → (p-value) → 12 → (VIF) → 10 → final features → final model

Modelling:

1. Use techniques like RFE to perform variable selection.
2. Build a Logistic Regression model with good sensitivity.
3. Check p-value and VIF.
4. Find the optimal probability cutoff. [ROC-curve]
5. Check the model performance over the test data.(Confusion Matrix,Sensitivity,F1-score etc.)
6. Generate the score variable.

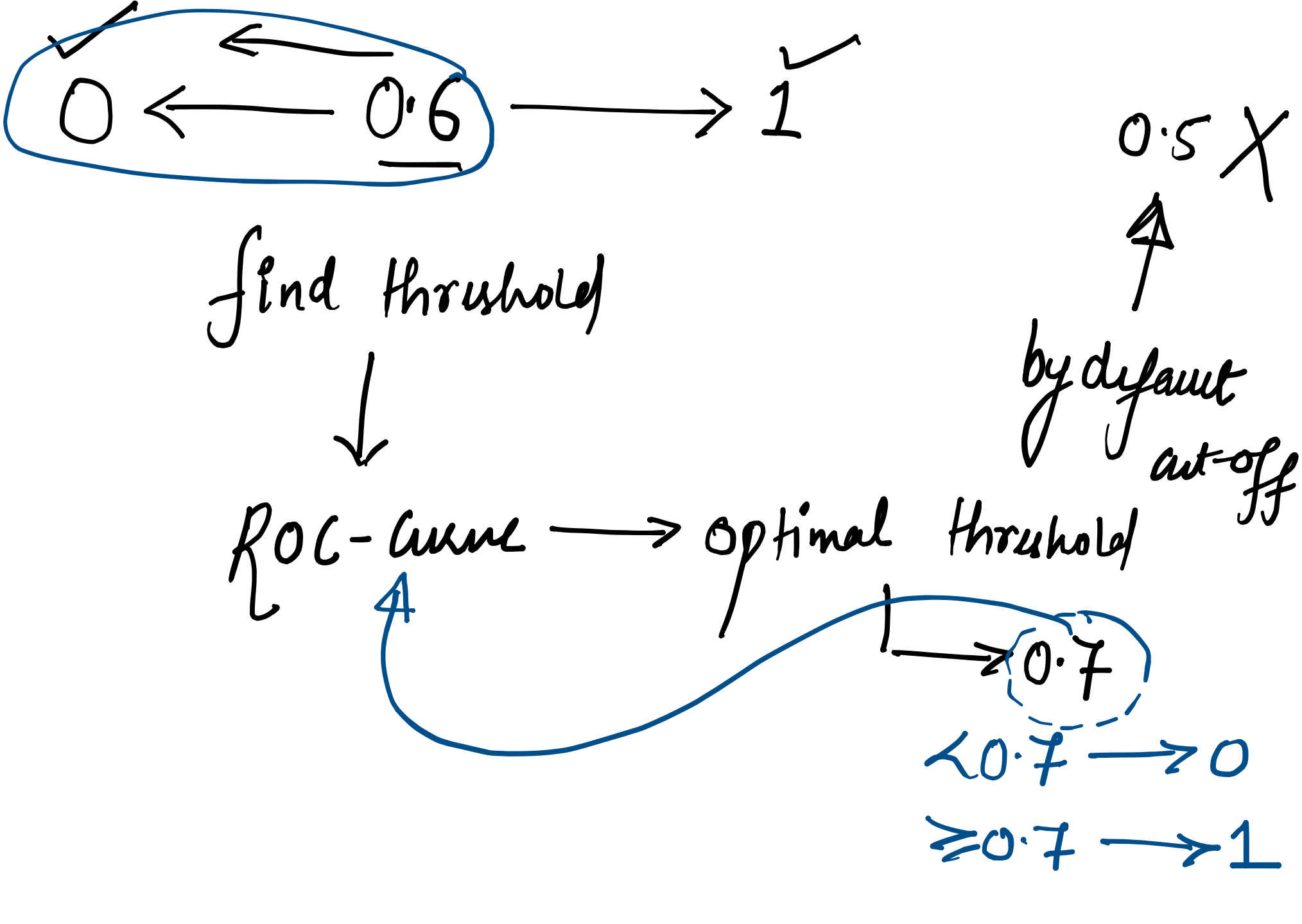
Iterative

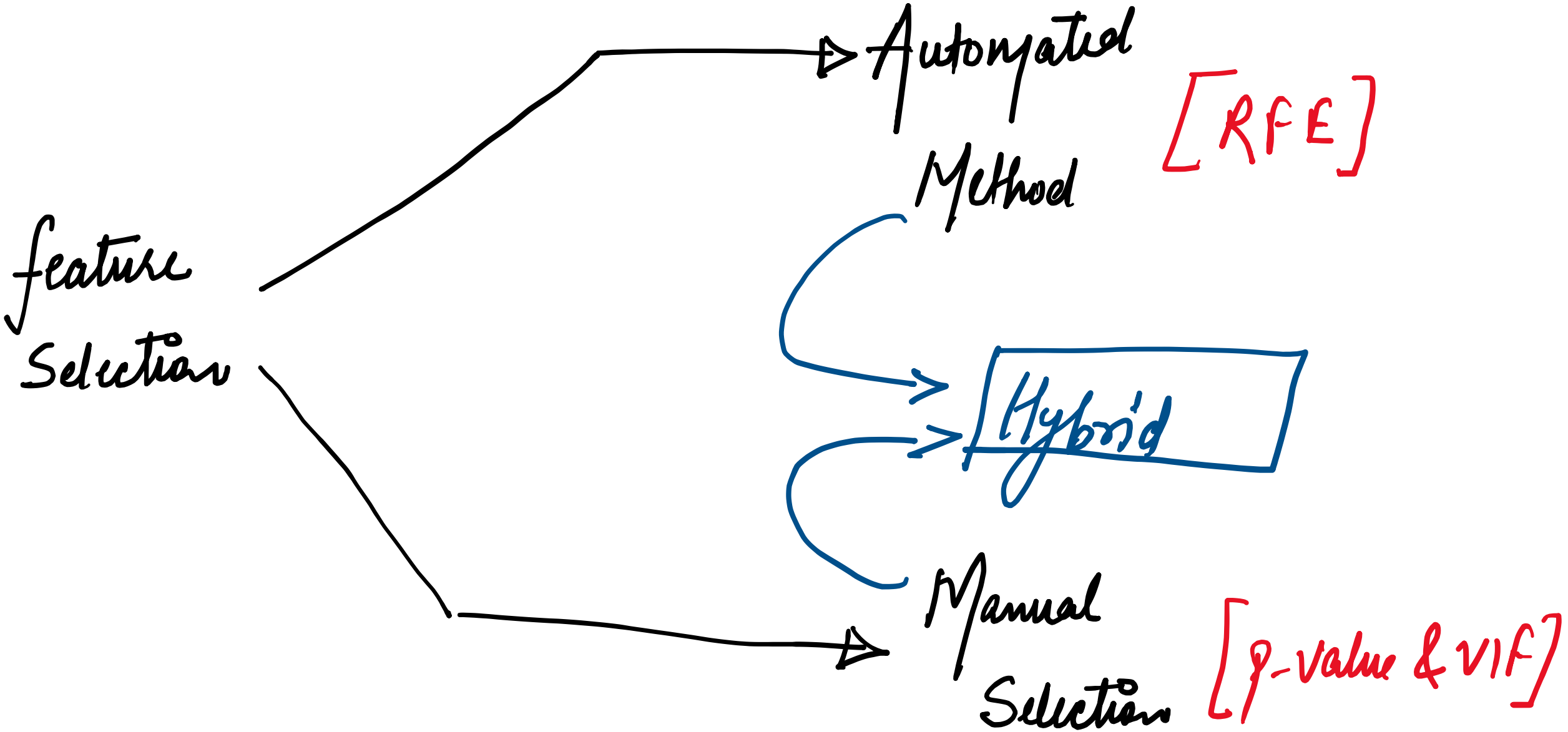
Recall

Score =	prob. × 100

prob. → Score

0.6 → 60


0
0.6
1
0.5
find threshold
by default cut-off
ROC-curve
optimal threshold
0.7
< 0.7 → 0
≥ 0.7 → 1


Feature Selection
Automated Method
[RFE]
Hybrid
Manual Selection
[p-value & VIF]

Assignment Steps :

(EDA + Data preprocessing + Data prep + modeling + evaluation + ROC + Score)

upGrad

Submission:

1. Jupyter Notebook: A well-commented Jupyter notebook with at least Logistic Regression model, the conversion predictions and evaluation metrices.
2. Subjective Answers: The word document filled with solutions to all the problems.
3. The overall approach of the analysis in the presentation:
   - Mention the problem statement and the analysis approach briefly.
   - Explain the results in business terms.
   - Include visualizations and summarize the most important results in the presentation.
4. Summary Report : A brief summary report in 500 words explaining how you proceed with the assignment and the learning that you gathered.

→ Jupyter Notebook → .ipynb format

→ Subjective Quest^{n}/Ans → Word (PDF)

→ Presentation → PPT → (PDF)

→ Summary Report (Word) → PDF

Zip file

↓

Upgrad portal

PPT (12-15 Slides)

→ Problem Statement + Assumptions (if any)

→ Approach

→ Graphs (EDA Results)

→ Conclusion or Result (Model performance, conf^n matrix)

→ Recommendation (Insights, impact on business, Recommendations)

Points to keep in mind

Short
Crisp
Clear

1- Do not mention code in your ppt.

2- Only put relevant results (important)

3- Graphs should be readable (not blur) & well-labelled
   → x
   → y
   → title

4- Do not write too much theory in your ppt. Always write in bullets.

What to keep in mind:

1. Add comments after every cell of code. So that we can understand your approach and method.
2. Describe the results.
3. Use Stack Overflow for dealing with syntax error, rather than being stuck at one place or waiting for someone to resolve your doubts, act and use the resources available on the internet to save time.
4. Post on the discussion forums for resolving any doubts you have.
5. Finally, write code manually instead of copy pasting from the in-content notebooks provided. Builds a habit of writing code. Its okay to look and write, but do not just copy-paste under any circumstance. Because of just copy pasting a lot of students have faced difficulties in the past when they had to write same code in their interview.