upGrad

#LifeKoKaroLift

Lead Scoring: Case Study

upGrad

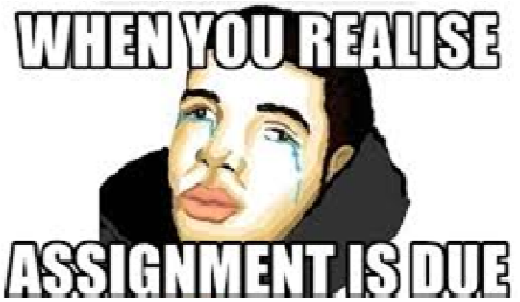
What we will cover in this session?

1. Problem Statement
2. Assignment walkthrough
3. QnA

upGrad

#LifeKoKaroLift

Lead Scoring: Assignment Walkthrough



Assignment

upGrad

Assignment Problem Statement

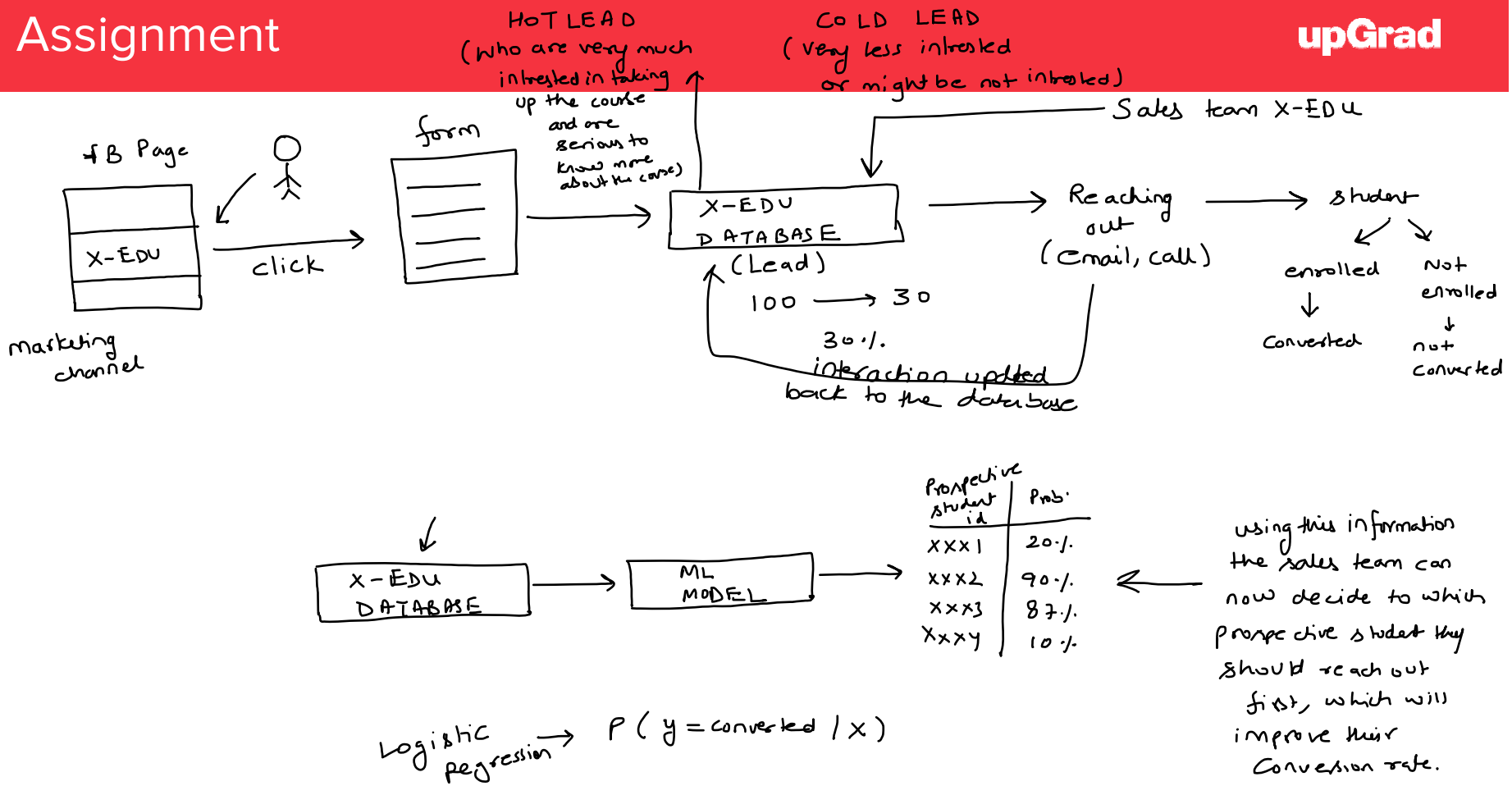
An education company named X Education sells online courses to industry professionals.The company markets its courses on several websites and search engines like Google. Once these people land on the website, they might browse the courses or fill up a form for the course or watch some videos. When these 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?

- 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.

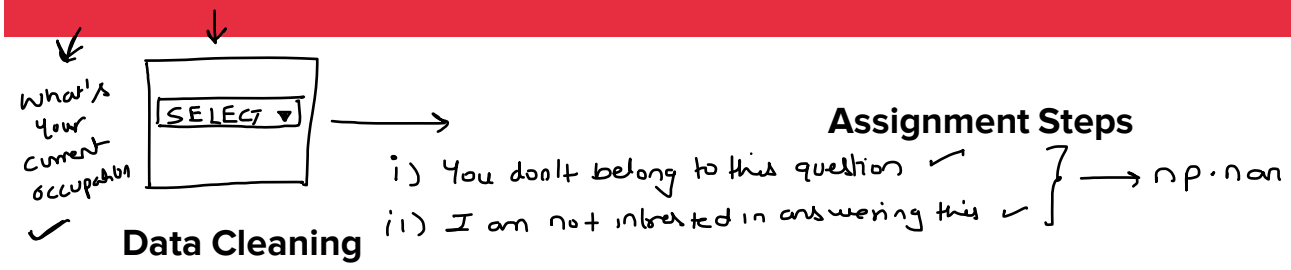
Assignment

upGrad

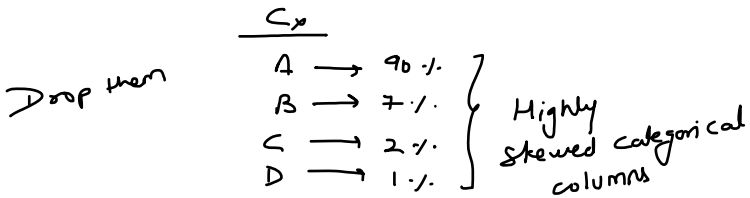


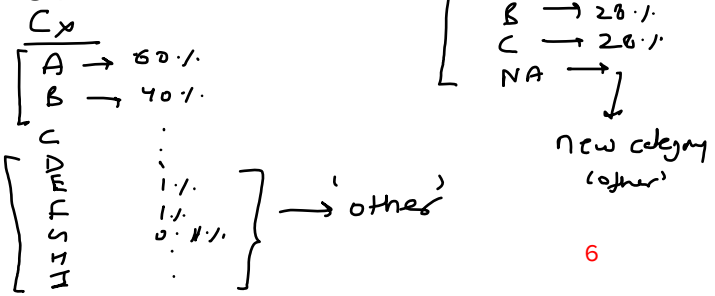
Assignment

upGrad



- Handle the "Select" level that is present in many of the categorical variables.
- Drop columns that are having high percentage of missing values. Check all the columns before dropping them. ≥ 40%.
- Check the number of unique categories in each categorical column. Here you may need to do something. → Row-wise cleaning
- For the columns with less percentage of missing, use some imputation technique. mode, mean, median
- Finally check the percentage of rows retained in data cleaning process.





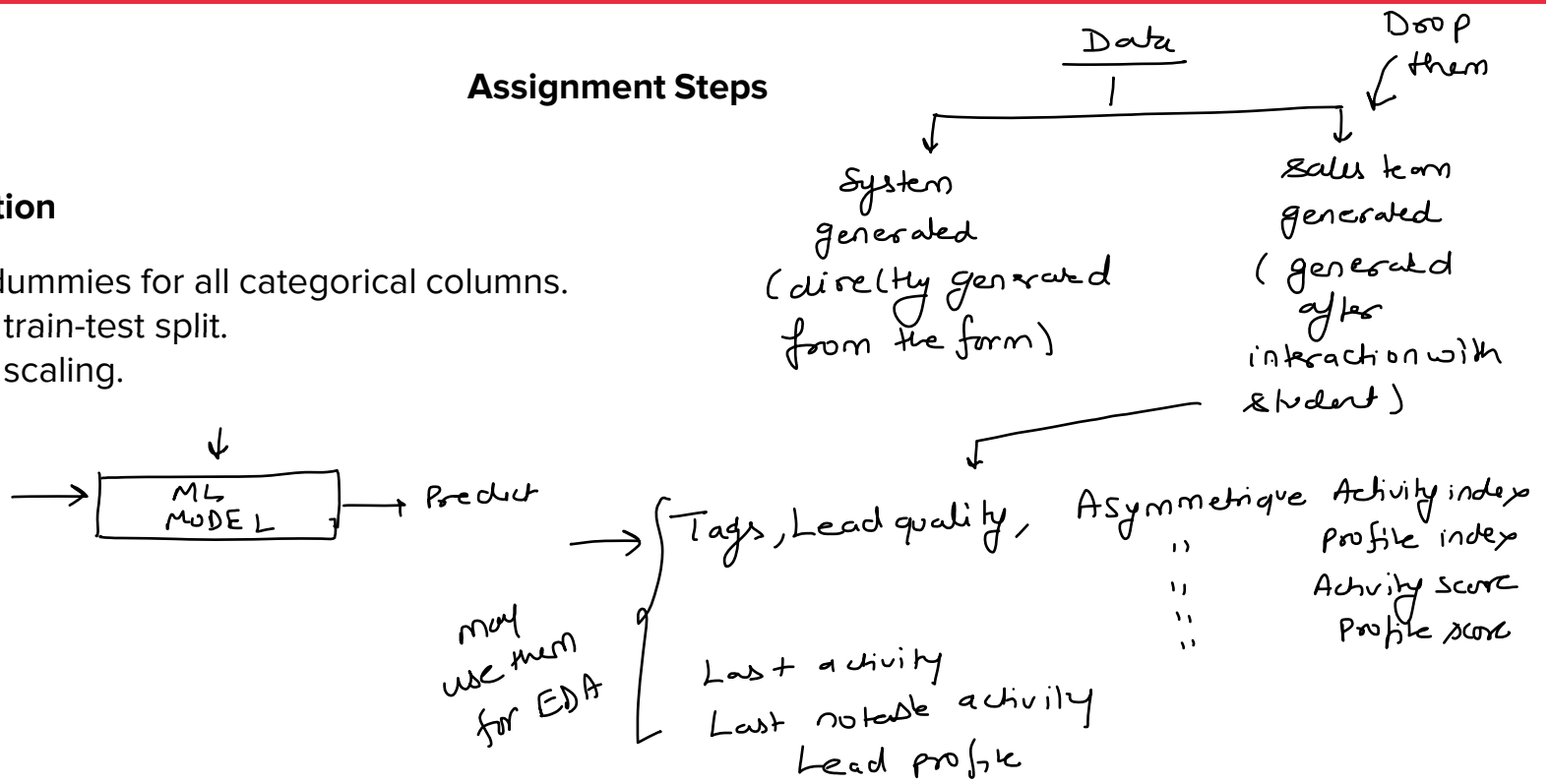
Assignment

upGrad

Assignment Steps

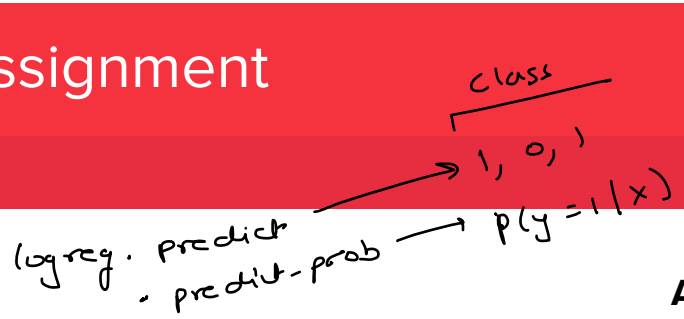
Data Preparation

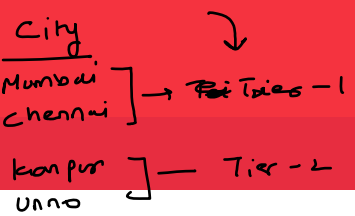
- Create dummies for all categorical columns.
- Perform train-test split.
- Perform scaling.



Assignment

upGrad



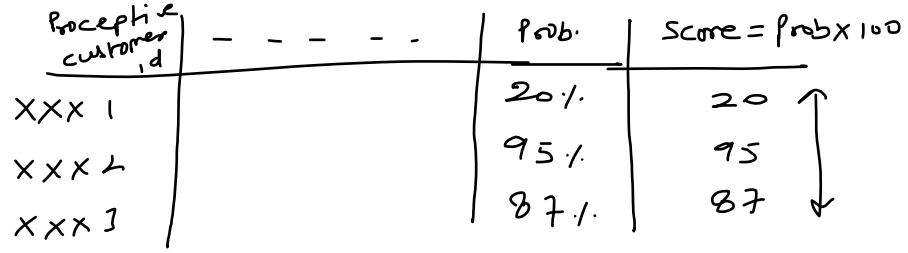


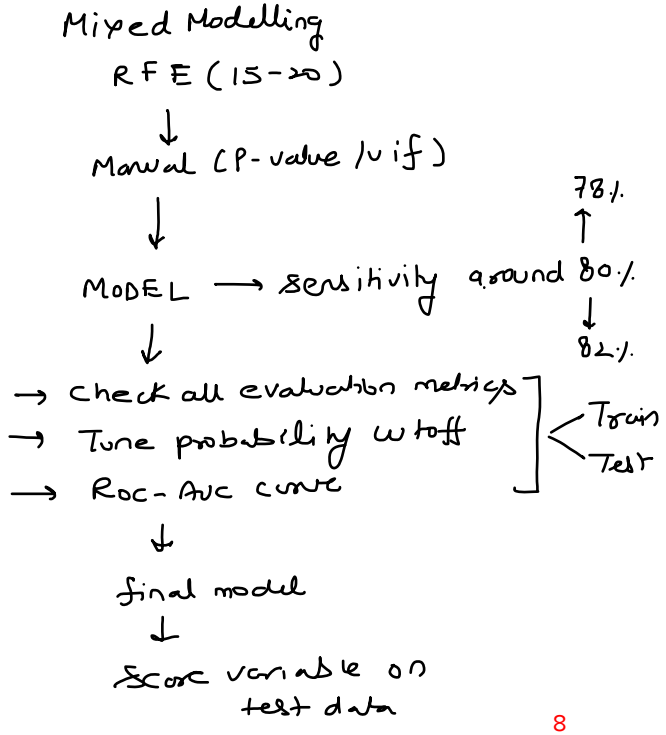
Assignment Steps

logreg. predict_prob (x-test)

Modelling

- Use techniques like RFE to perform variable selection.
- Build a Logistic Regression model with good sensitivity.
- Check p-value and VIF.
- Find the optimal probability cutoff.
- Check the model performance over the test data.
- Generate the score variable.





Assignment

upGrad

Assignment-Submission

Submission

- **Jupyter Notebook:** A well-commented Jupyter note with at least the logistic regression model, the conversion predictions and evaluation metrics.
- **Subjective Answers:** The word document filled with solutions to all the problems.
- The overall approach of the analysis in a presentation
  - Mention the problem statement and the analysis approach briefly
  - Explain the results in business terms
  - Include visualisations and summarise the most important results in the presentation
- **Summary Report:** A brief summary report in 500 words explaining how you proceeded with the assignment and the learnings that you gathered.
- **Presentation:** Make a presentation to present your analysis to the chief data scientist of your company (and thus you should include both technical and business aspects).

Assignment-Endnote

What to keep in mind

- Add comments after every cell of code. So that we can understand your approach and method.
- Describe the results.
- Use StackOverflow for dealing with syntax errors. Rather than being stuck at one place or waiting for someone to resolve your doubts, take action and use the resources available on the internet to save time.
- Post on the discussion forums for resolving any doubts you have
- Finally, write code manually instead of copy-pasting from the in-content notebooks provided. Builds a habit of writing code. It's okay to look and write, but don't just copy-paste under any circumstance. Because of just copy-pasting, a lot of our students have faced difficulties in the past when they had to write some code on their interview.

Assignment

upGrad

0 → -ve
1 → tve

y_{True}	y_{pred}	
0	0	TN
1	1	TP
0	1	fP
1	0	FN

✓ **Case-1 :** Cancer detection model.

→ Some non-cancerous case as cancerous → fp

→ cancerous case as non-cancerous → fN

$$\uparrow \text{Sensitivity} = \frac{TP}{TP + fN \downarrow}$$

✓ **Case-2 :** Email spam-ham classifier (Spam – 1, Ham – 0)

→ Some Ham email as spam → fp

→ Some spam email as ham → fN

$$\uparrow \text{Precision} = \frac{TP}{TP + fP \downarrow}$$

	Predicted		
		-ve	tve
Actual	-ve	TN	fp
	tve	fN	Tp

→ Converted → converted

Type-I → Converted → non-converted ✓ FN

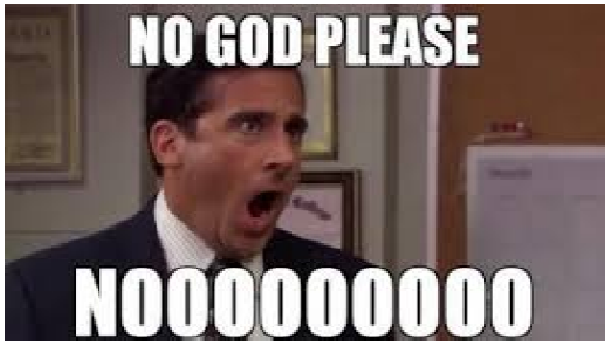
Type-II → non-converted → converted

$$\text{Sensitivity} = \frac{TP}{TP + fN}$$

upGrad

#LifeKoKaroLift

Quiz Time



Assignment

upGrad

Question-1: A study was designed to compare Red Bull energy drink commercials. Each participant was shown the commercials, A and B, in random order and asked to select the better one. There were 140 women and 130 men who participated in the study. Commercial A was selected by 65 women and by 67 men. Find the odds of selecting Commercial A for the men.

A • 0.51
B • 0.49
C • 1.04
D • 2.54

$$\text{Odds} = \frac{P(\text{Yes})}{P(\text{No})}$$

$$= \frac{0.51}{0.49} = 1.04$$

$$P(\text{people selecting A}) = \frac{67}{130} = 0.57$$

$$P(\text{people not selecting A}) = 1 - 0.57 = 0.43$$

Question-2: A survey on 250 customers was conducted for an automobile dealership. The customers were asked if they would recommend the service department to a friend. The number who responded Yes was 210. Find the odds of person responding yes.

A • 4.05
B • 5.25
C • 3.52

$$\text{odds} = \frac{210}{250} \times \frac{250}{70}$$

$$= \frac{21}{7} = 5.25$$

$$P(\text{Yes}) = \frac{210}{250}$$

$$P(\text{No}) = \frac{40}{250}$$

Assignment

upGrad

Question-3: Consider the confusion matrix given below. What is the accuracy of the model?

- 84%
- 82%
- 91%
- 92%

$$acc = \frac{TP + TN}{TP + TN + FP + FN} = \frac{50 + 100}{50 + 10 + 5 + 100} = \frac{150}{165} = 91\%$$

Question-4: What is the precision of the model?

- 84%
- 82%
- 91%
- 92%

$$\frac{TP}{TP + FP} = \frac{100}{100 + 10} = \frac{100}{110} = 91\%$$

n=165	**Predicted: NO**	**Predicted: YES**
Actual: NO	TN: 50	FP: 10
Actual: YES	FN: 5	TP: 100

Assignment

upGrad

Question-5: What is the recall of the model?

- 92%
- 95% ✓
- 91%

$$\frac{TP}{TP + FN} = \frac{100}{100+5} = \frac{100}{105} = 95\%$$

Question-6: What is the F1-Score of the model?

- 92% ✓
- 95%
- 91%

$$f1\text{-}s = \frac{2 \times P \times R}{P + R}$$

$$= \frac{2 \times 0.91 \times 0.95}{0.91 + 0.95}$$

$$= 92.1\%$$

n=165	Predicted: NO	Predicted: YES
Actual: NO	TN 50	FP 10
Actual: YES	FN 5	TP 100

upGrad
#LifeKoKaroLift

Thank You!

References:
towardsdatascience