upGrad

#LifeKoKaroLift

Lead Scoring: Case Study

upGrad

What we will cover in this session?

1 Problem Statement

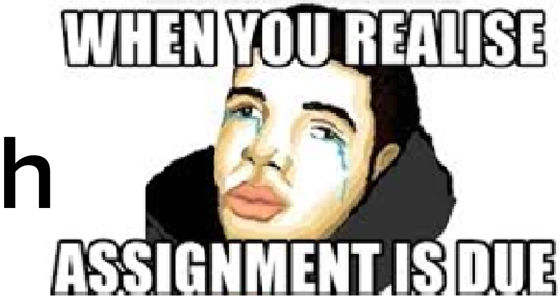
2 Assignment walkthrough

3 QnA

upGrad

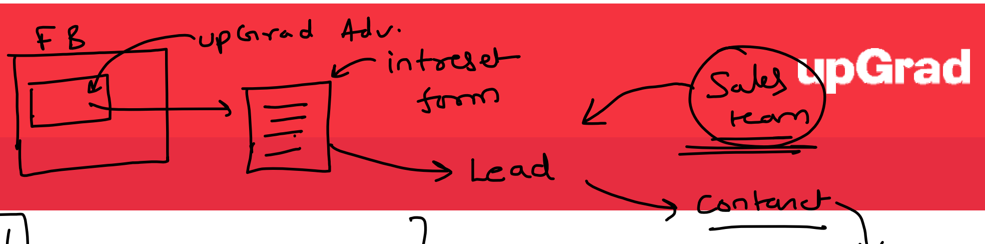
#LifeKoKaroLift

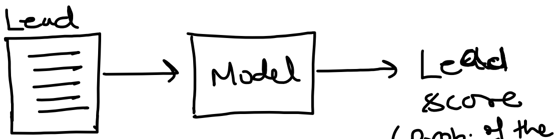
Lead Scoring: Assignment Walkthrough



Assignment

Problem: The system is not optimised





(prob. of the student getting converted into full time student)

20% → 80%

Assignment Problem Statement

100% → 30%

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.

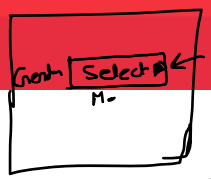
What you need to do?

- 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

{ EDA
outliers



→ current exp . Select] Not applicable

i) check for columns having "select" field
↓
missing data

ii) %age of missings values

iii) Drop those columns with high %age of missing value

iv) check for # of unique categories for all categorical columns (identify highly skewed columns and then drop them)

Assignment Steps

np.nan } missing fields

Data Cleaning

- 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.
- Check the number of unique categories in each categorical column. Here you may need to do something.
- For the columns with less percentage of missing, use some imputation technique.
- Finally check the percentage of rows retained in data cleaning process.

Cx	
A	90%
B	2%
C	1
D	2%

Cx	
A	25%
B	25%
C	15%
D	10%
E	10%
F	10%
G	5%
H	1%
	1%
	%

combine them "other"

vii) you can drop rows with high missing %age

viii) check the %age of rows that are retained.

v) Identify such categorical columns having so many categories but with very less %age of rows := combine categories and name it to "other"

vi) for columns with less %age of missing → impute them

Assignment

(0 15 85 100)

85 → hot lead

upGrad

model → 1000

900 → 1

↓

90%] Conversion

Assignment Steps

Data Preparation

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

problem → row-1

	C1	C2	C3	C4	C5	C6	C7	C8
row-1	NA	0	NA	1	NA	NA	NA	NA →

Assignment

Predict ( ) → class

Predict_Prob ( ) → Prob × 100

↓

y = 1

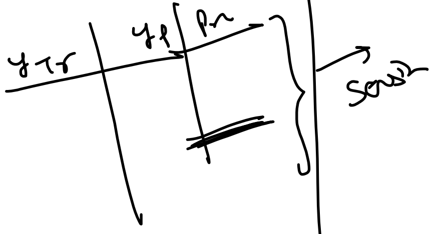
Assignment Steps

Modelling

RFE + Manual approach

- 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. (0 — 100)

$P(y=1|x) \times 100 = \text{Score}.$



Independent variables

- user input variables
- Score variables
- Activity variables

Score variables:
- Lead quality
- Tags
- Asymmetrique activity index

↓

These are those variables that are generated by the sales team after the discussion with student

(Drop them)

Assignment

Assignment-Submission

Zip → Jupyter notebook
→ PDF (subjective answer)
→ PDF (summary report)
→ PDF (PPT)

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

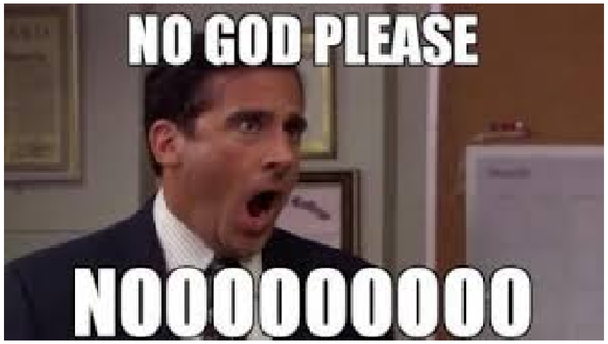
What to keep in mind

- 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.

upGrad
#LifeKoKaroLift

Generate model with best sensitivity → use the model to predict prob. → prob × 100 = score

Quiz Time



upGrad
#LifeKoKaroLift

Thank You!

References:
towardsdatascience