



Problem Statements & Assignment Planing

DO HOANG HUONG LIEN
NGUYEN HUU LIEM

Problem Statement

Stakeholders: An education company named X Education.

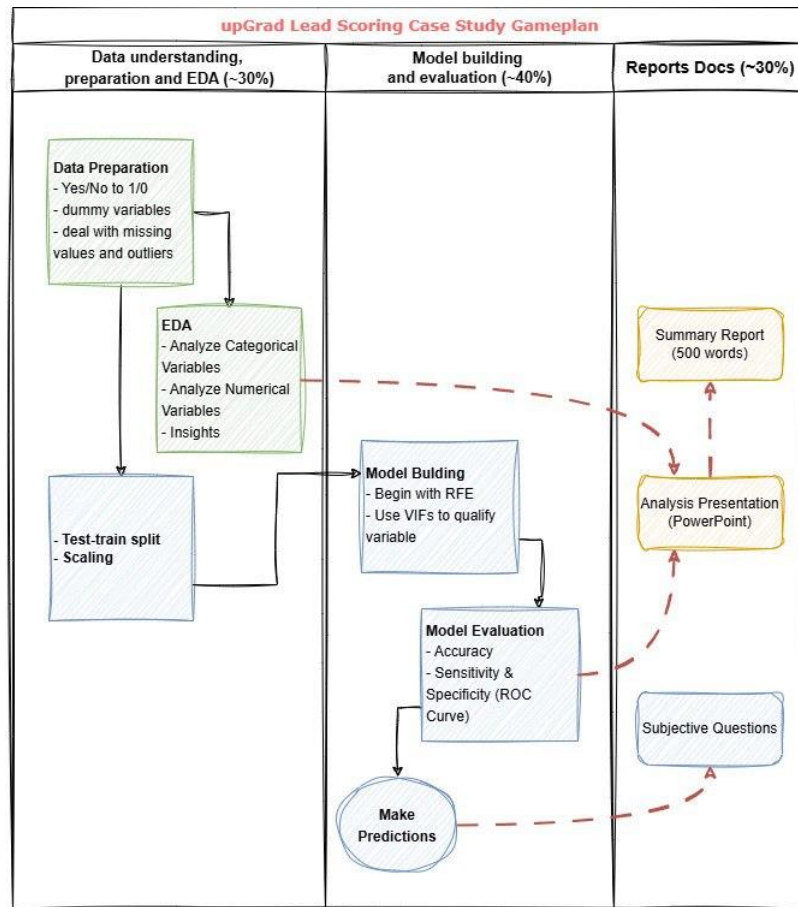
Business Requirements: Identify the most potential leads, also known as 'Hot Leads'.

Data Analyst Requirements: Able to identify the leads that are most likely to convert into paying customers, with a target lead conversion rate to be around 80%.

Model Requirements: A logistic regression model to assign a lead score between 0 and 100 to each of the leads which can be used by the company to target potential leads.



Assignment Planing



Huong Lien

Mamta S. Kamadi

Huu Liem

Actual Contributions:

Huong Lien:

As planed, plus the Presentation

Huu Liem:

As planed, plus the Presentation & Summary Report

Mamta S. Kamadi

Absolutely NONE.

All her works has been detected as plagiarism.



Data Preparation and EDA

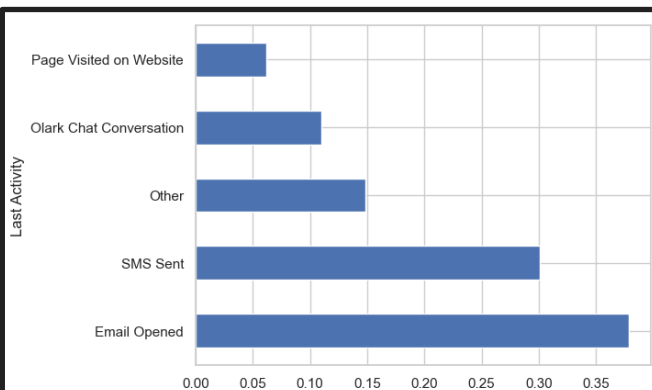
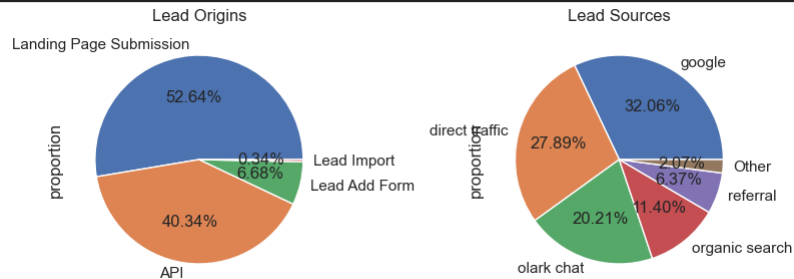
DO HOANG HUONG LIEN
NGUYEN HUU LIEM

Data Preparation

- Rename some columns to a shorter and more straightforward name. E.x: “What is your current occupation” to “occupation”
- Drop cols with just a single variable
- Clean categorical variables using RegEx, deal with low quality variable (E.x some columns contain “Select”)
- Impute missing value in numerical variables
- Variable conversion (Yes/No to 1/0)
- Identify and remove Outliers from “TotalVisits” and “Page Views Per Visit” using percentiles. Removing outliers will provide a better Scaler later on.

EDA

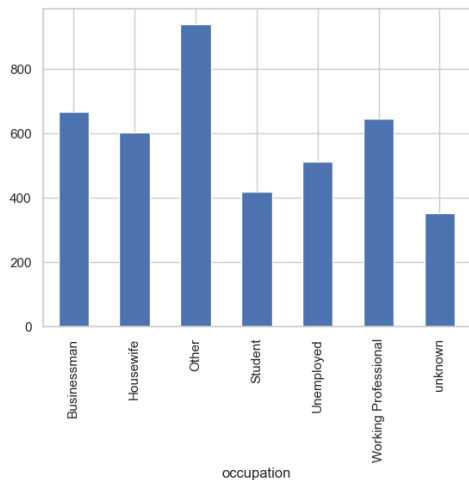
- The origin of leads mainly is from "Landing Page Submission" and "API"
- The source of leads mainly is from "google" and "direct traffic"



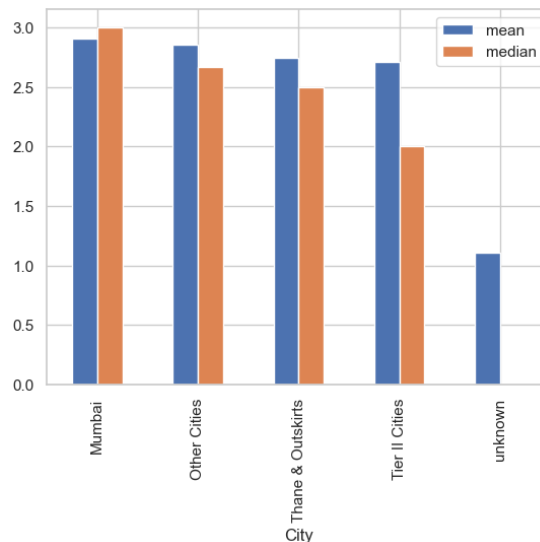
- Before the leads get converted, users were open their Email or sent SMS.

- Most of the customers are okay with receiving emails and calls about the course with over 90% say "No" with "Do not email" or "Do not call" even nearly 100%.
- However, almost advertisement platforms or measurements are not really effective, because almost 100% customers said "No" about seeing ad before. Only in "free_copy_mastering_interview" field, there is 31,8% customers having desire to receive a free copy of 'Mastering the Interview'.
- Only 1/6 Leads are marked as Potential

EDA

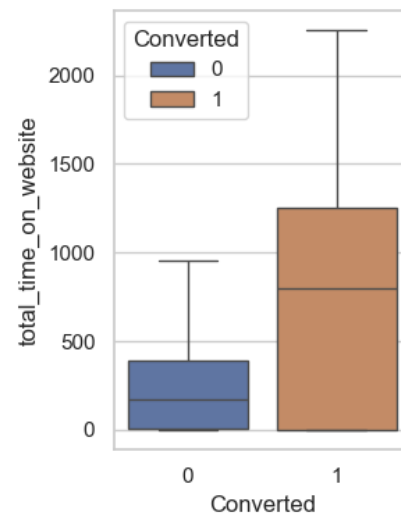


- People who spent the most time on website are Businessman or working Professional



- Mumbai, Thane & Outskirts, and other cities have equal mean and median values, indicating symmetric data distributions.

- Tier II Cities has a higher mean than median, suggesting a right-skewed distribution.



- Converted Leads have doubled total time on website compared to non-converted leads

EDA

- *Some strong indicators of a promising lead:*

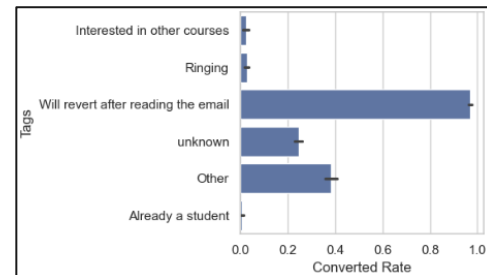
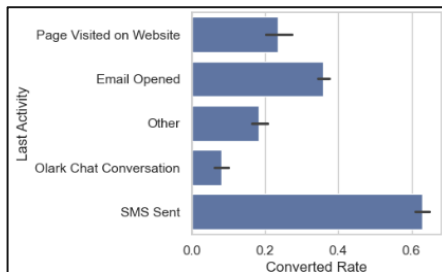
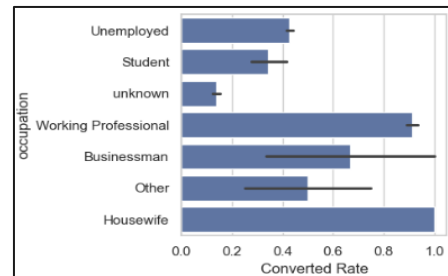
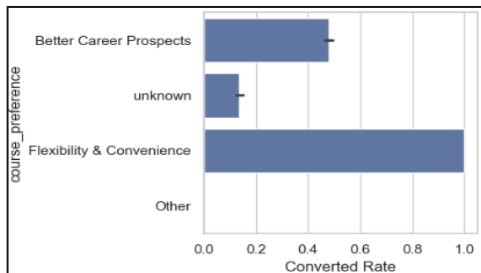
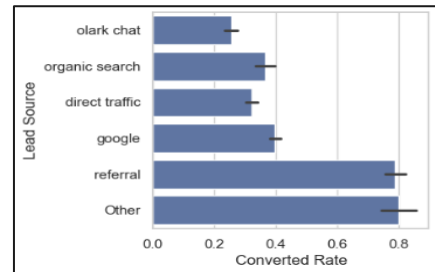
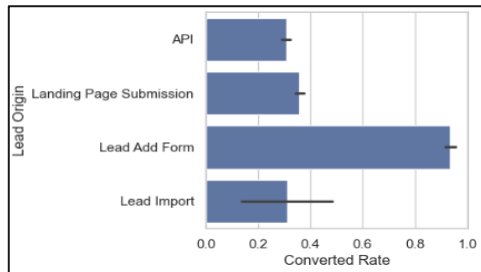
+ **Lead Origin:** Lead Ad Form. This form initiate good leads

+ **Lead Source:** referral. Referral Source brings quality leads

+ **Last Activity:** SMS Sent. This will be a good signal

+ **Occupation:** Housewife, Professional
+ **Course Prefers:** Flexibility & Convenience. Seems like the company is delivering this aspect.

+ **Tags:** Will revert after reading the email. They definitely will!

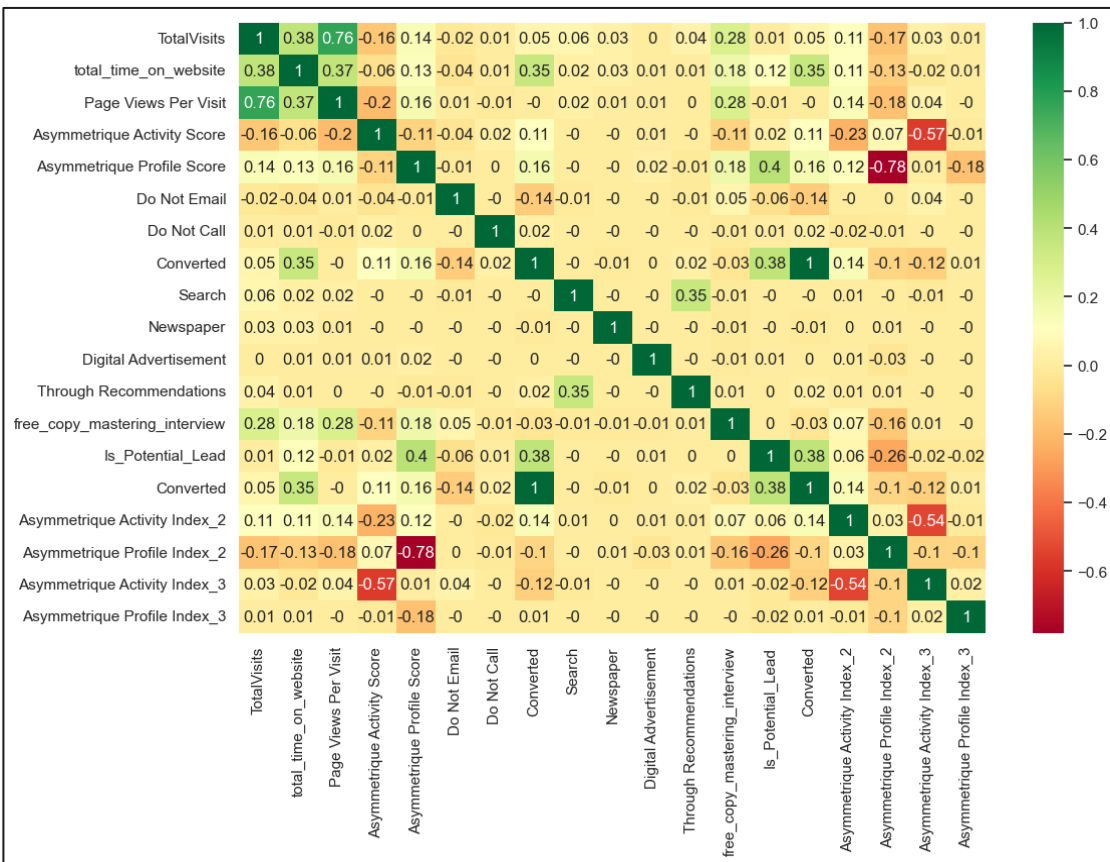




Model building & Evaluation

DO HOANG HUONG LIEN
NGUYEN HUU LIEM

Test-train split and scaling



Some key observations:

- 'TotalVisits' and 'total_time_on_website' have a strong positive correlation (0.38).
- 'Page Views Per Visit' shows a strong correlation with 'TotalVisits' (0.76).
- The 'Profile Index' and 'Activity Index' dummy variables have high correlation with the 'Asymmetrique Score' dummy variables. We should drop those dummy columns to avoid collinearity.

Model building

Pseudo R-squared (CS) of 0.5487

- This is a measure of how well the model fits the data. In logistic regression, Pseudo R-squared values can vary between 0 and 1, with higher values indicating better fit.

- A value of 0.5487 suggests a reasonably good fit, meaning the model explains about 54.87% of the variance in the dependent variable 'Converted'

Generalized Linear Model Regression Results			
Dep. Variable:	Converted	No. Observations:	6046
Model:	GLM	Df Residuals:	6031
Model Family:	Binomial	Df Model:	14
Link Function:	Logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-1586.0
Date:	Sat, 19 Oct 2024	Deviance:	3172.0
Time:	14:10:26	Pearson chi2:	1.31e+04
No. Iterations:	7	Pseudo R-squ. (CS):	0.5487
Covariance Type:	nonrobust		

Model building

Significance of Variables

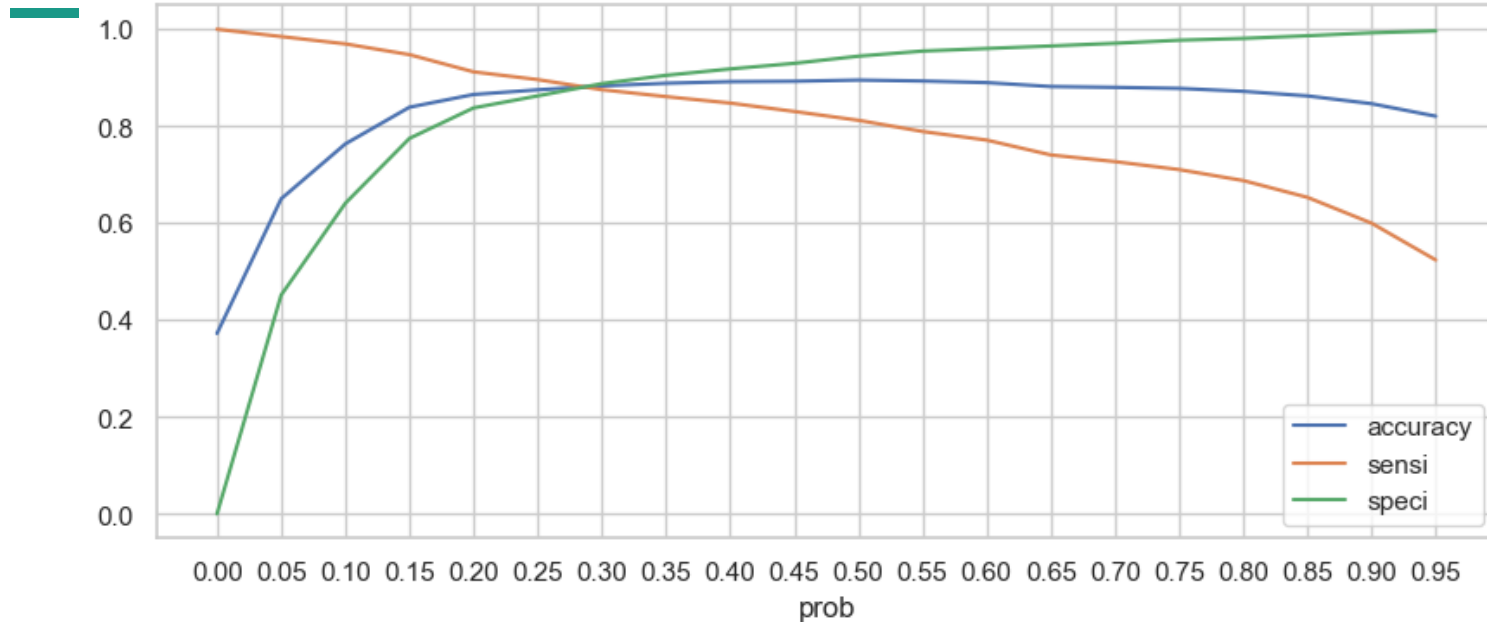
Almost all the variables have p-values close to 0, which indicates that they are statistically significant in predicting the probability of conversion. The most significant features include:

- Do Not Email
- total_time_on_website
- Lead Origin_Lead Add Form
- Tags_Will revert after reading the email
- Last Notable Activity_SMS Sent


	coef	std err	z	P> z	[0.025	0.975]
const	-0.4979	0.126	-3.955	0.000	-0.745	-0.251
Do Not Email	-1.7676	0.233	-7.592	0.000	-2.224	-1.311
total_time_on_website	1.0307	0.052	19.874	0.000	0.929	1.132
Search	2.6264	1.224	2.145	0.032	0.227	5.026
Is_Potential_Lead	0.6390	0.149	4.286	0.000	0.347	0.931
Lead Origin_Lead Add Form	2.9709	0.271	10.974	0.000	2.440	3.502
Lead Source_direct traffic	-1.6411	0.152	-10.779	0.000	-1.939	-1.343
Lead Source_google	-1.1968	0.142	-8.437	0.000	-1.475	-0.919
Lead Source_organic search	-1.3144	0.180	-7.317	0.000	-1.667	-0.962
Last Activity_Olark Chat Conversation	-1.4674	0.195	-7.510	0.000	-1.850	-1.084
course_preference_unknown	-0.6148	0.118	-5.229	0.000	-0.845	-0.384
Tags_Other	0.4350	0.122	3.569	0.000	0.196	0.674
Tags_Ringing	-3.7740	0.280	-13.466	0.000	-4.323	-3.225
Tags_Will revert after reading the email	4.0597	0.202	20.090	0.000	3.664	4.456
Last Notable Activity_SMS Sent	2.0124	0.114	17.688	0.000	1.789	2.235

Model evaluation

The the variation in model performance metrics—accuracy, sensitivity (sensi), and specificity (speci)—across different probability thresholds.



- For the best performance in both statistics side and business side, we pick the Cutoff point of 0.28, to achieve Accuracy & Sensitivity & Specificity of 0.88. This will allow lead conversion rate to be around 80%.

- 
1. **Accuracy:** The blue line represents the overall accuracy of the model, showing how well the model predicts conversions at different probability thresholds. It peaks around the 0.3 threshold.
 1. **Sensitivity (Sensi):** The orange line indicates sensitivity, or the true positive rate. It's highest at lower thresholds, meaning the model is better at identifying actual conversions when it allows more predictions to be positive.
 1. **Specificity (Speci):** The green line shows specificity, or the true negative rate. It's highest at higher probability thresholds, indicating that the model is better at correctly identifying non-conversions when it requires stronger evidence to predict conversion.
 1. **Optimal Threshold:** The intersection points and relative heights can help determine an optimal threshold where the model maintains a balance between sensitivity and specificity, often aligned with high accuracy.

Evaluation test-set

Confusion Matrix

	Predicted Not Converted (0)	Predicted Converted (1)
Actual Not Converted (0)	1388	204
Actual Converted (1)	128	872

- So we end up with a model with 0.88 accuracy & sensitivity & specificity on the train test, and 0.87 accuracy & sensitivity on the test set.
- A drop of only 0.01 indicate a reliable model had been build!



Thank You!