LEAD SCORING CASE STUDY

PROBLEM SOLVING USING LOGISTIC REGRESSION

By- Harshit Gosain

BUSINESS OBJECTIVE

- 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.

- To make this process more efficient, the company wishes to identify the most potential leads, also known as 'Hot Leads'. If they successfully identify this set of leads, the lead conversion rate should go up as the sales team will now be focusing more on communicating with the potential leads rather than making calls to everyone.

- Our objective is to help select promising leads using logistic regression.

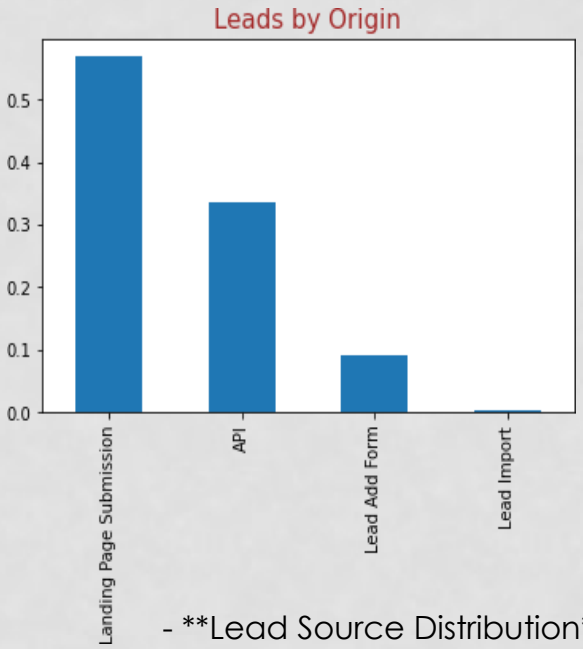
SOLUTION METHODOLOGY

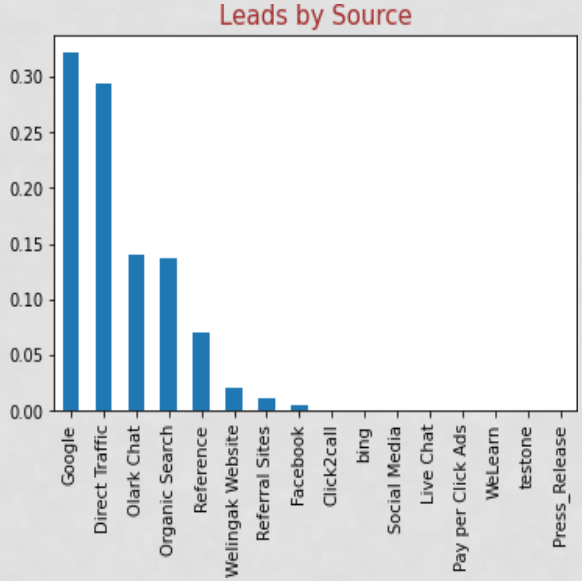
- Data Cleaning and manipulation
- Exploratory Data Analysis
- Model Building
- Model Evaluation
- Model Prediction on Testset
- Inferences
- Recommendation

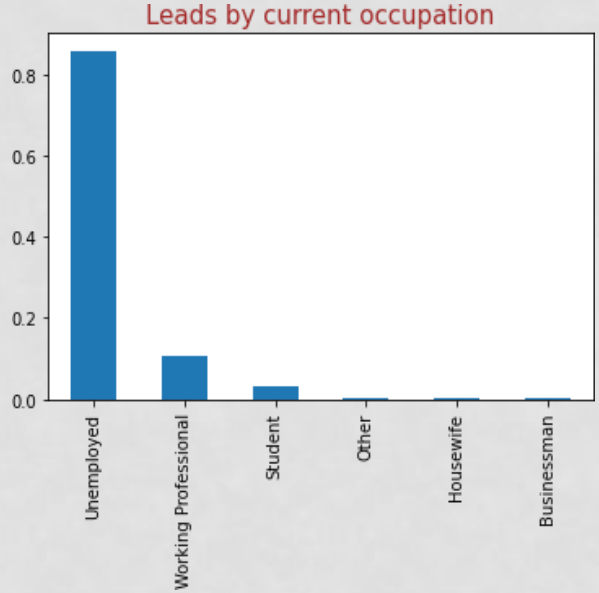
DATA CLEANING

- - **Handling Missing Values**: Given the abundance of data points (over 9000), columns with more than 30% missing values were eliminated to maintain data integrity and analysis efficiency.
- - **Omitting Irrelevant Variables**: Since X Education offers online courses, variables related to city and country were deemed irrelevant and thus removed from the dataset.
- - **Eliminating Record Identifiers**: Columns such as Prospect ID and Lead Number, serving merely as record identifiers, were dropped as they do not contribute to predictive modeling.
- - **Addressing Skewed Data**: Columns exhibiting skewed data points were removed as they lack predictability value, ensuring that only relevant and informative features are retained for analysis.
- - **Conversion Rate**: After data cleaning and preprocessing, a conversion rate of 48% was determined, signifying the proportion of leads successfully converted. This serves as a benchmark for evaluating the effectiveness of subsequent predictive models and strategies.

UNIVARIATE ANALYSIS





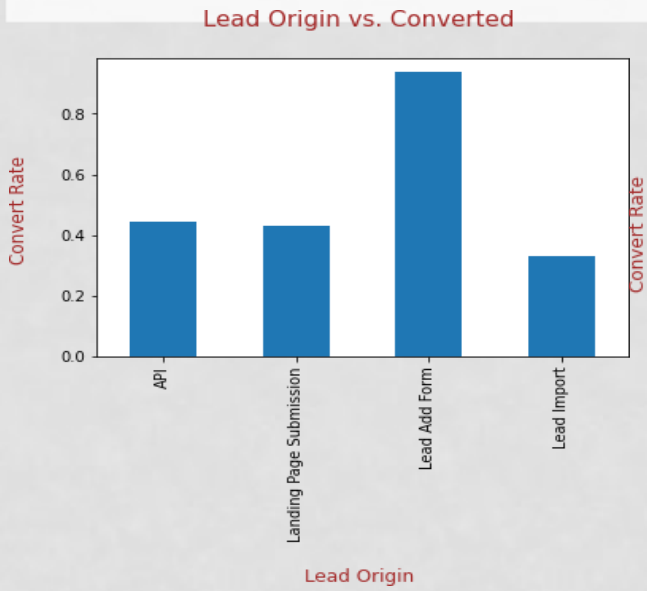


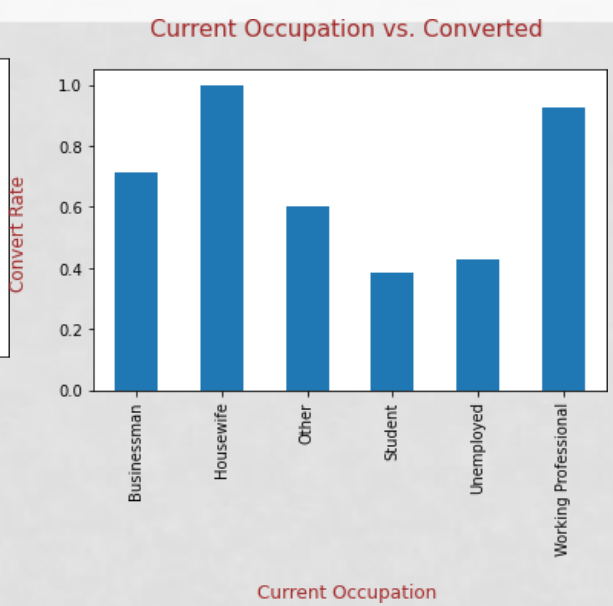
- **Lead Source Distribution**: The primary sources generating leads for X Education are Landing Page Submissions and API, with the former being the most dominant channel.

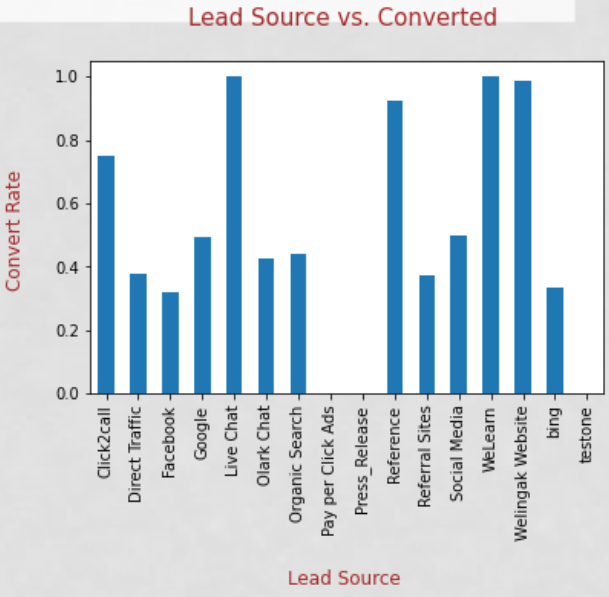
- **Traffic Origins**: The majority of leads originate from 'Google' and 'Direct Traffic', with 'Olark Chat' and 'Organic Search' following closely behind in terms of lead generation.

- **Customer Employment Status**: A significant portion of leads comes from individuals categorized as 'Unemployed', indicating a potential demographic segment that engages with X Education's offerings.

BI-VARIATE ANALYSIS







- **Conversion Probability by Lead Origin**: Leads originating from Add Forms exhibit a higher likelihood of conversion compared to other lead sources.

- **Conversion Trends by Occupation**: Individuals classified as 'Working Professionals' and 'Housewives' demonstrate a higher propensity for conversion compared to other occupational categories.

- **Influential Lead Sources**: Lead sources such as Live Chat, Reference, WeLearn, and the Welingak Website stand out for their higher conversion rates, suggesting their effectiveness in attracting leads likely to convert into customers.

MODEL BUILDING

- - **Data Splitting**: The dataset was divided into training and testing sets using a 70:30 ratio to facilitate model training and evaluation.
- - **Numerical Feature Scaling**: Numerical features were scaled using the MinMax scaler to ensure that all features are on a similar scale, preventing any particular feature from dominating the model due to its larger magnitude.
- - **Feature Selection with Recursive Feature Elimination (RFE)**: RFE was employed to identify the 15 most important features contributing to lead conversion prediction, thereby reducing dimensionality and enhancing model interpretability.
- - **Statistical Significance Analysis**: Features were further refined by evaluating their p-values and Variance Inflation Factors (VIF) to eliminate statistically insignificant predictors, ensuring that the final model comprises only relevant and impactful features.
- - **Final Feature Set**: After feature selection and refinement, the model was narrowed down to 11 features deemed crucial for predicting lead conversion.
- - **Lead Scoring**: A lead score was created by multiplying the predicted conversion probability by 100, yielding a score ranging between 0 and 100. Higher scores indicate 'hot leads' with a greater likelihood of converting, enabling the prioritization of leads based on their conversion potential.

MODEL EVALUATION

Generalized Linear Model Regression Results

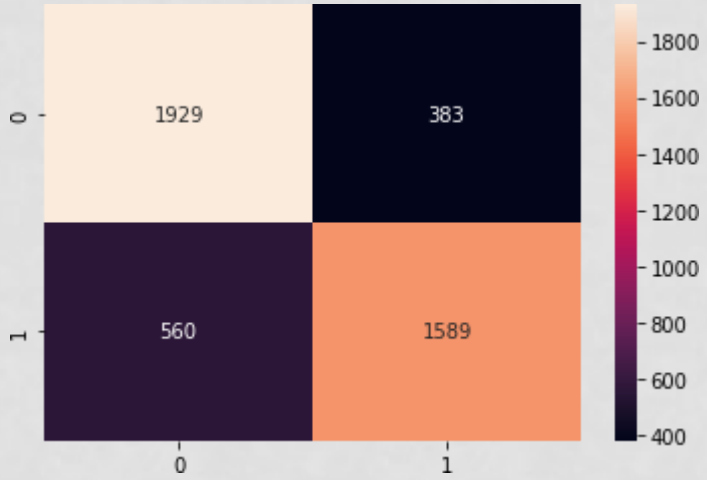
Dep. Variable:	Converted	**No. Observations:**	4461
Model:	GLM	**Df Residuals:**	4449
Model Family:	Binomial	**Df Model:**	11
Link Function:	logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-2079.1
Date:	Mon, 14 Nov 2022	**Deviance:**	4158.1
Time:	15:08:31	**Pearson chi2:**	4.80e+03
No. Iterations:	7		
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	0.2040	0.196	1.043	0.297	-0.179	0.587
TotalVisits	11.1489	2.665	4.184	0.000	5.926	16.371
Total Time Spent on Website	4.4223	0.185	23.899	0.000	4.060	4.785
Lead Origin_Lead Add Form	4.2051	0.258	16.275	0.000	3.699	4.712
Lead Source_Olark Chat	1.4526	0.122	11.934	0.000	1.214	1.691
Lead Source_Welingak Website	2.1526	1.037	2.076	0.038	0.121	4.185
Do Not Email_Yes	-1.5037	0.193	-7.774	0.000	-1.883	-1.125
Last Activity_Had a Phone Conversation	2.7552	0.802	3.438	0.001	1.184	4.326
Last Activity_SMS Sent	1.1856	0.082	14.421	0.000	1.024	1.347
What is your current occupation_Student	-2.3578	0.281	-8.392	0.000	-2.908	-1.807
What is your current occupation_Unemployed	-2.5445	0.186	-13.699	0.000	-2.908	-2.180
Last Notable Activity_Unreachable	2.7846	0.807	3.449	0.001	1.202	4.367

	Features	VIF
9	What is your current occupation_Unemployed	2.82
1	Total Time Spent on Website	2.00
0	TotalVisits	1.54
7	Last Activity_SMS Sent	1.51
2	Lead Origin_Lead Add Form	1.45
3	Lead Source_Olark Chat	1.33
4	Lead Source_Welingak Website	1.30
5	Do Not Email_Yes	1.08
8	What is your current occupation_Student	1.06
6	Last Activity_Had a Phone Conversation	1.01
10	Last Notable Activity_Unreachable	1.01

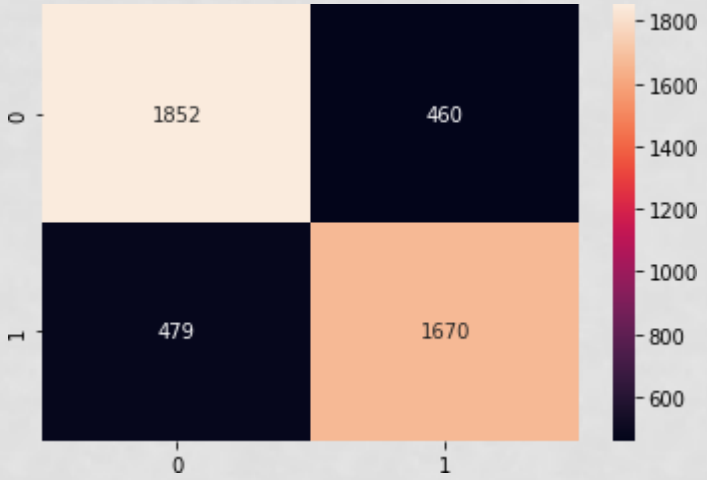
MODEL EVALUATION

- Training Set



Accuracy	78.86%
Sensitivity	73.94%
Specificity	83.43%

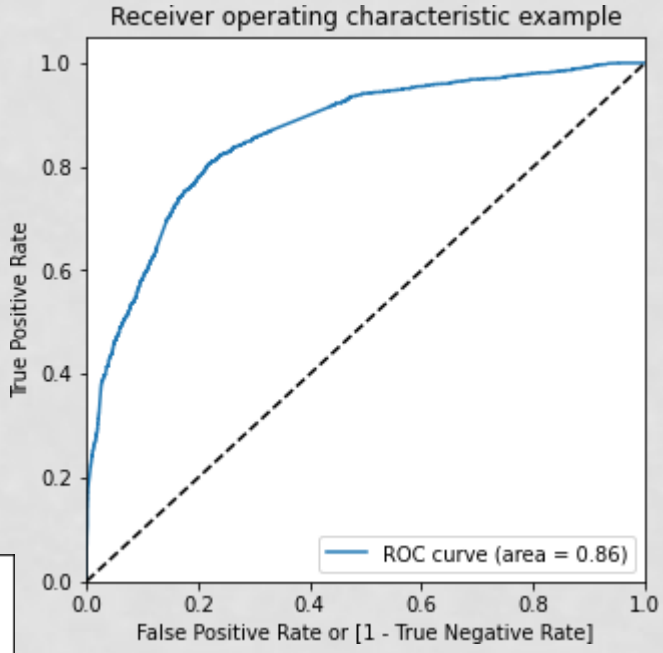
Test Set

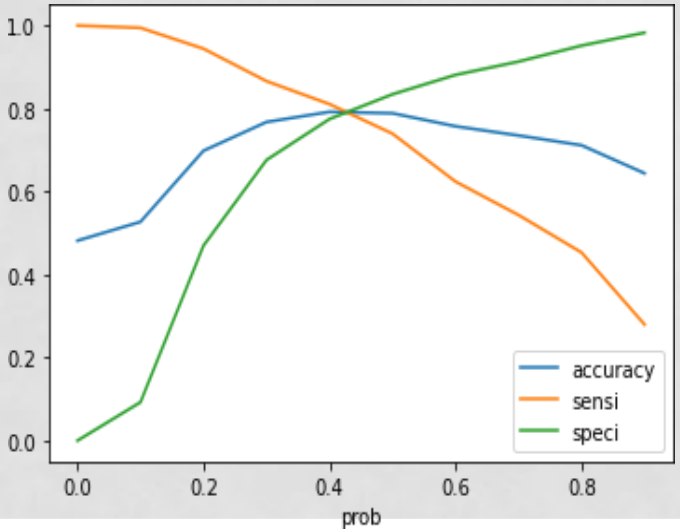


Accuracy	78.95%
Sensitivity	77.71%
Specificity	80.10%
Precision	78.40%
Recall	77.71%

MODEL EVALUATION - ROC/CUTOFF

	prob	accuracy	sensi	speci
0	0	48.17%	100.00%	0.00%
0.1	0.1	52.70%	99.44%	9.26%
0.2	0.2	69.83%	94.42%	46.97%
0.3	0.3	76.75%	86.60%	67.60%
0.4	0.4	79.20%	81.06%	77.47%
0.5	0.5	78.86%	73.94%	83.43%
0.6	0.6	75.72%	62.40%	88.11%
0.7	0.7	73.50%	54.35%	91.31%
0.8	0.8	71.15%	45.32%	95.16%
0.9	0.9	64.40%	27.97%	98.27%





INFERENCES

- **Top Three Numerical Variables Contributing to Conversion Probability:**
- 1. TotalVisits: The number of visits made by the lead to X Education's website.
- 2. Total Time Spent on Website: The cumulative duration spent by the lead navigating X Education's website.
- 3. Lead Origin_Lead Add Form: Origin of the lead, particularly those generated through the 'Lead Add Form', indicating a higher probability of conversion.

- **Top Three Categorical/Dummy Variables for Focused Attention on Conversion Probability:**
- 1. Lead Origin_Lead Add Form: Leads originating from the 'Lead Add Form' source exhibit a significant potential for conversion.
- 2. Last Activity_Had a Phone Conversation: Leads engaging in phone conversations as their last activity demonstrate a higher likelihood of conversion, suggesting the importance of personalized communication.
- 3. Lead Source_Welingak Website: Leads sourced from the 'Welingak Website' show a notable propensity for conversion, emphasizing the effectiveness of this particular lead source in driving conversions.

RECOMMENDATION

- Depending on the requirements the model needs to be tweaked such that
- **Scenario 1**: So when the company has more interns we need have lower cutoff threshold so that our model can predict almost all leads. The flip side to this decrease in threshold will be that we will misclassify some non-conversions as conversions but this is a good tradeoff given we have mode manpower to deal with it.
- **Scenario 2**: Typically, when the company has less people to call potential customers so its good to have more accurate predictions in which case the model specificity should be much more higher. This would mean form the above graph the we would have to choose a cutoff point which is much higher. The tradeoff of this is that we are going to miss some leads but given that the company has less manpower who can focus more on correctly predicted leads.
- **Scenario 3**: The company should focus on sending automated SMS and emails to potential leads during the time they have less manpower which allows for cost effective lead conversion without manual intervention.

Thank you