Lead Scoring Case Study

Vijay Kumar

Sunil Shaw

Vignesh

Problem Statement

Build an efficient model to identify potential leads who will enroll for an online course in an education company named X Education

Analysis Approach

1. Data Cleaning: Loading data set, understanding and cleaning data
2. EDA: Check imbalance, univariate and bivariate analysis
3. Data Preparation: Dummy variables, test-train split and feature scaling
4. Model building: RFE for top features, manual feature selection and finalizing model
5. Predictions on test data: Compare train vs test metrics, assign lead score
6. Recommendations: Suggest top features to focus for higher conversion

Data Cleaning

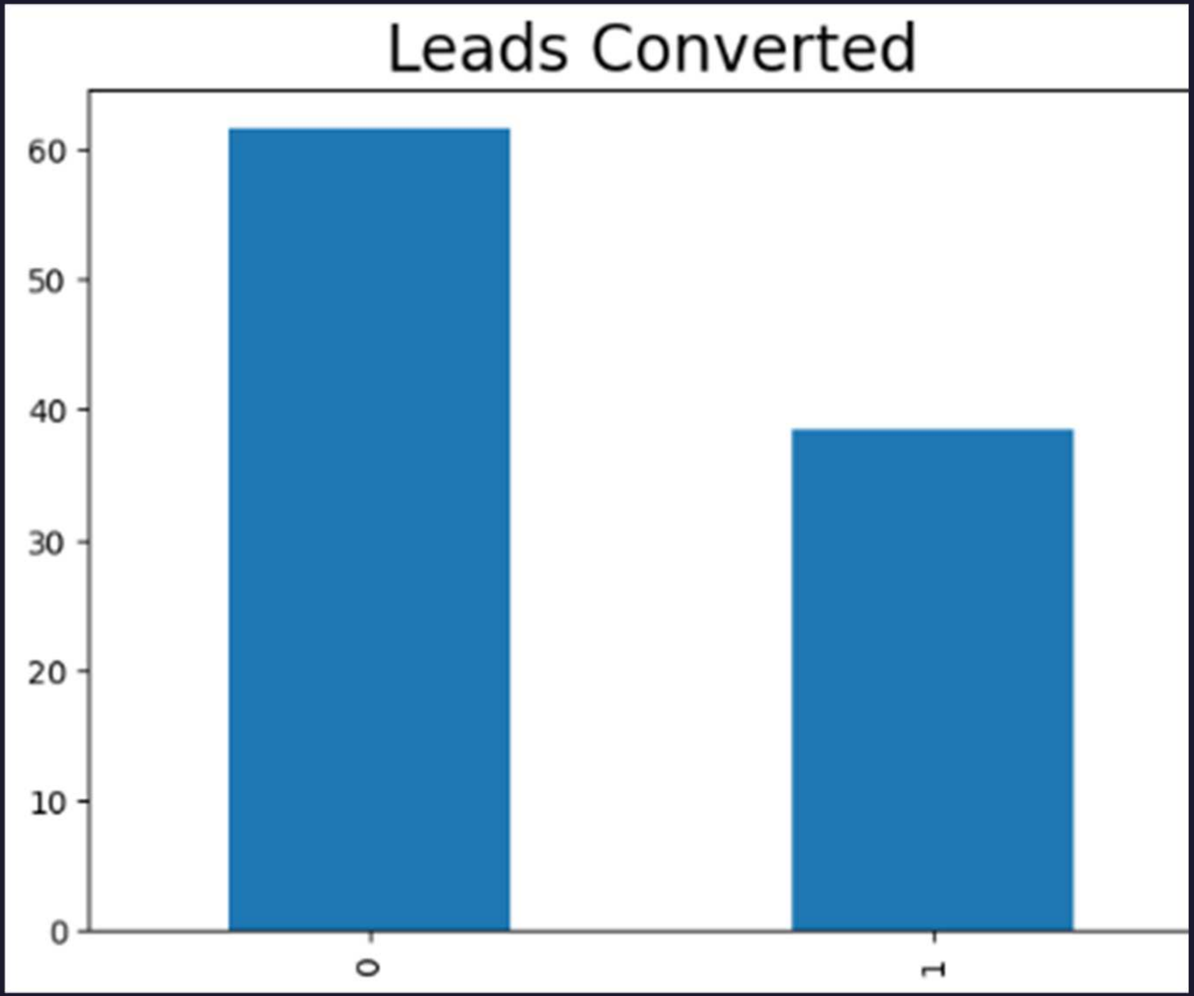
1. "Select" level represents null values for some categorical variables, as customers did not choose any option from the list.
2. Columns with over 35% null values were dropped.
3. Features which do not add any insights are dropped (country and what matters most to you in choosing a course)
4. Columns with no use for modeling (Prospect ID, Lead Number) or only one category of response were dropped.
5. Null values in categorical columns are replaced with most commonly occurring values
6. Null values in numerical columns are replaced with median value
7. There are outliers in numerical columns which will be handled using scaling technique
8. Binary categorical variables were mapped

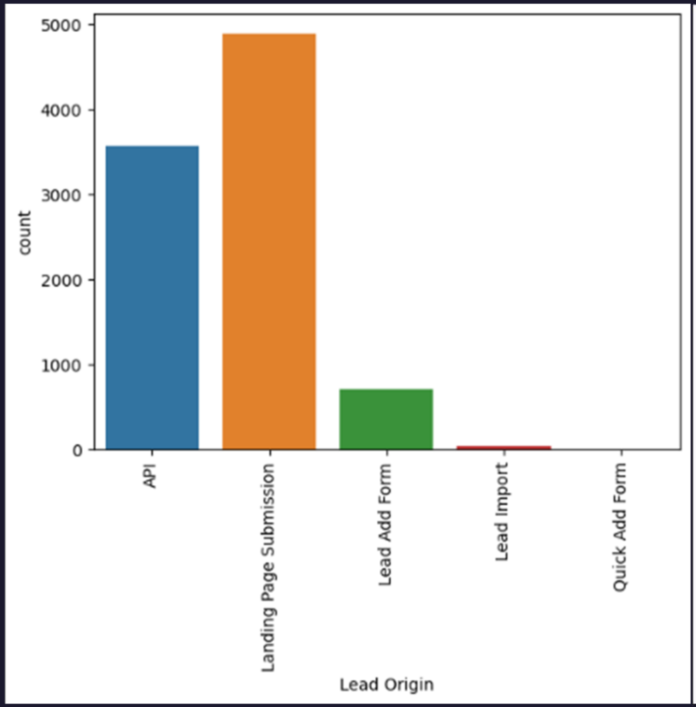
EDA

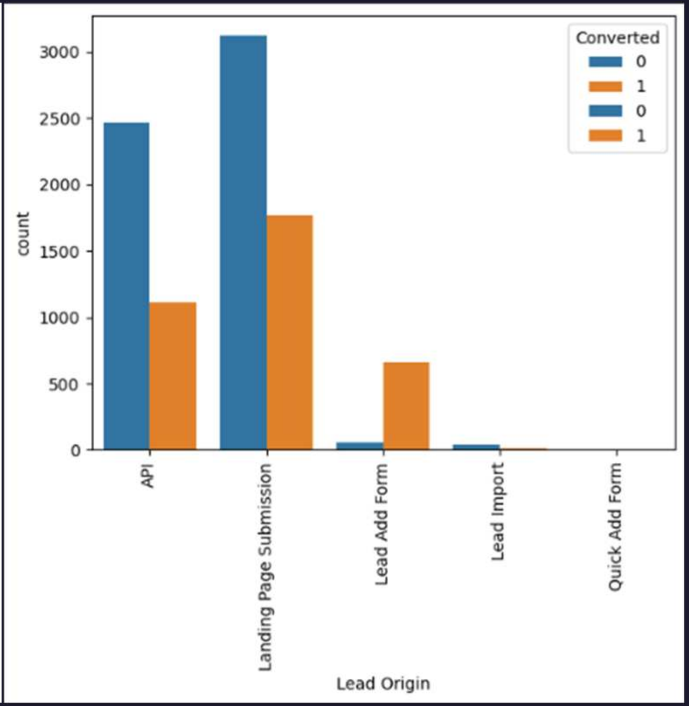
1. Target variable data is imbalanced

2. Conversion rate is of 38.5%, meaning only 38.5% of the people have converted to leads.(Minority)

3. While 61.5% of the people didn't convert to leads. (Majority)

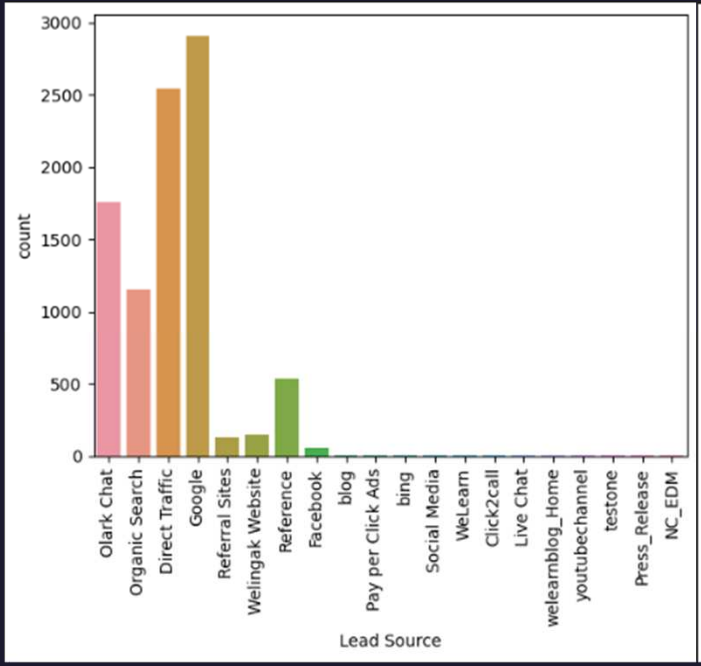


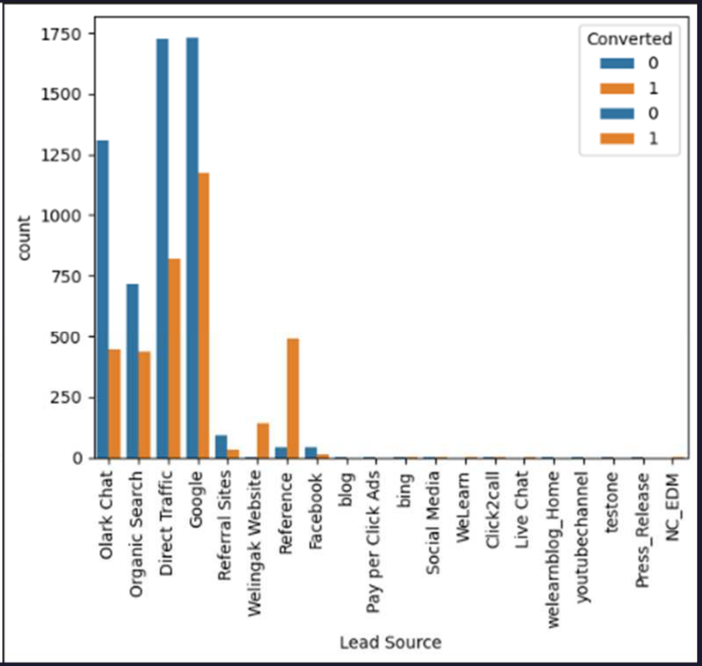




Lead Origin

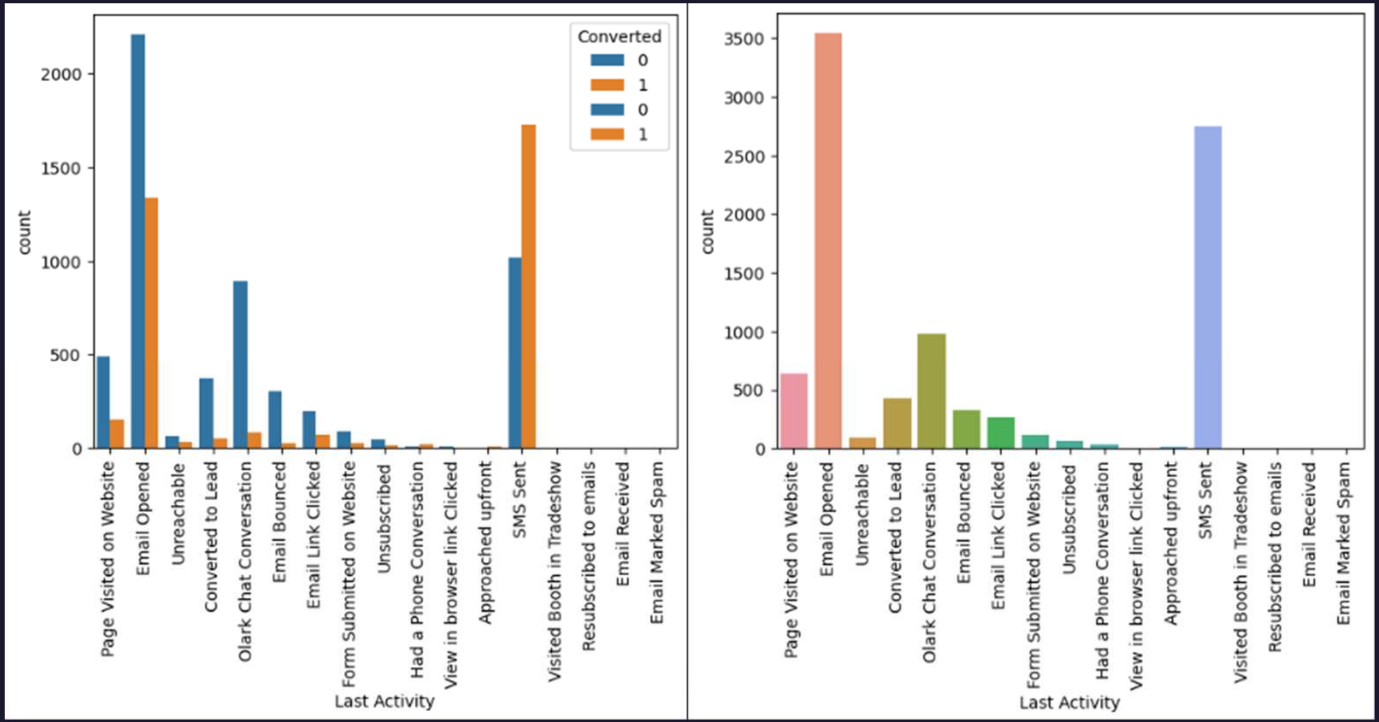
- Most number of customers were identified by "Landing Page Submission"





Lead Source

- Most number of lead sources are from Google and Direct Traffic



Last Activity

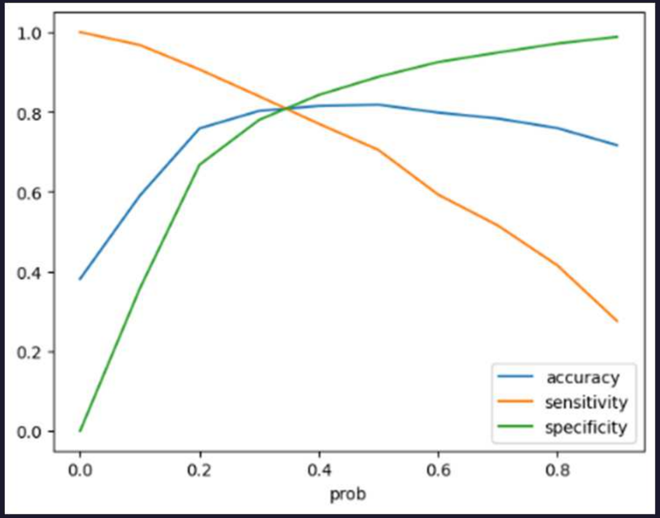
- Most of the customers last activities are SMS sent and Email Opened

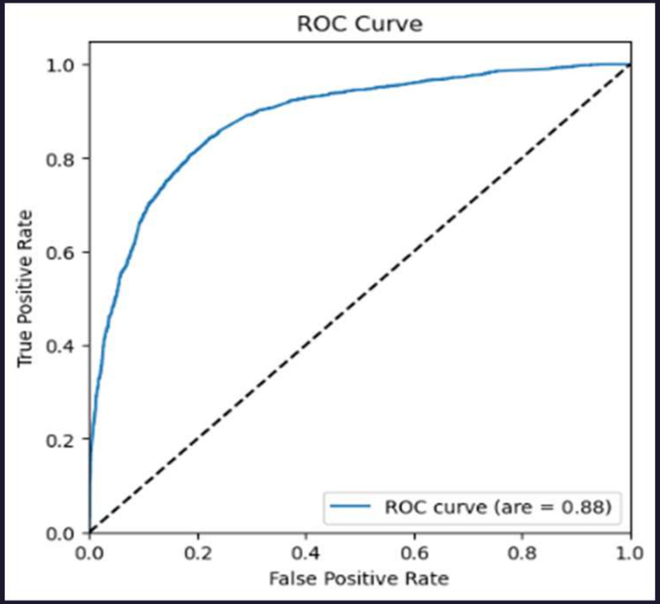
Data Preparation

1. Binary level categorical columns were already mapped to 1 / 0 in previous steps
2. Created dummy features for categorical variables which has more than 2 levels
3. Splitting train and test sets
4. Feature scaling: min-max scaling is used to scale features

Model Building

1. **Recursive Feature Elimination** is used to select top 20 features
2. Then the model is manually tuned to end up with 16 features
3. Features with p-value higher than 0.05 are dropped one by one and model is built again
4. VIF is withing range (<5) for all the features
5. Model 6 is the final model





Model Evaluation

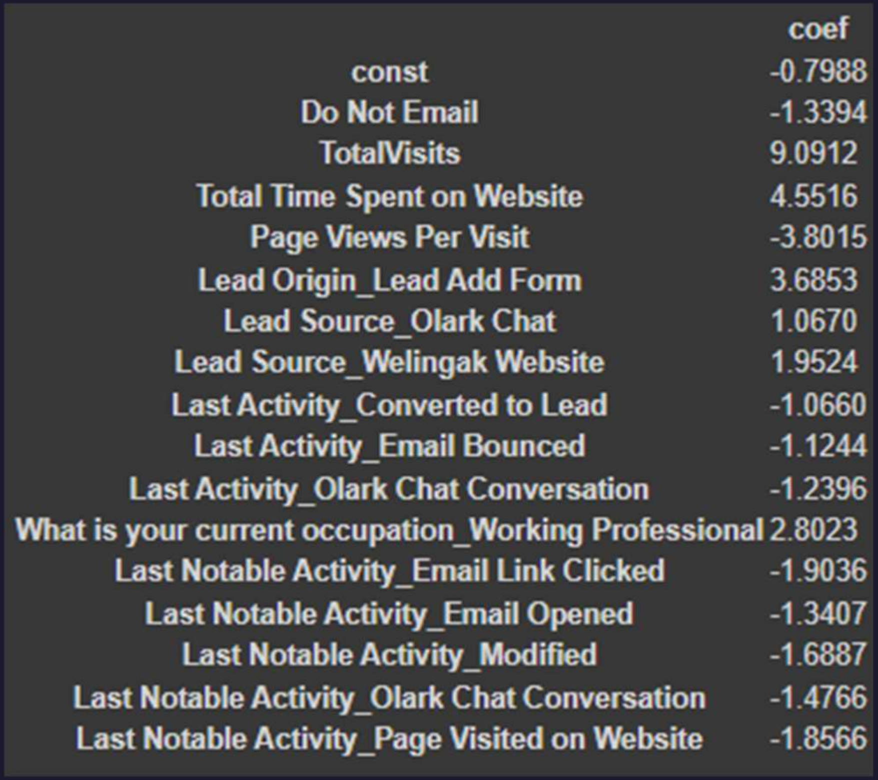
- CEO of X education has given a target lead conversion to be around 80%
- To identify 80% of potential leads sensitivity is the metric which should satisfy this target value
- From accuracy sensitivity and specificity curve 0.35 is the probability cutoff chosen to distinguish between leads who is going to convert and who does not
- With this value of cutoff, sensitivity of 81% is achieved
- Area under curve (AUC) for this model comes around 0.88 which is a good number
- This evaluation values are for train data, when model is used to predict for test data set similar value of AUC and sensitivity is obtained which indicates that model is reliable

Recommendations

- As per the problem statement, increasing lead conversion is crucial for the growth and success of X Education. To achieve this, we have developed a regression model that can help us identify the most significant factors that impact lead conversion.

- Features that have the highest positive coefficients should be given priority in our marketing and sales efforts to increase lead conversion.

- We have also identified features with negative coefficients that may indicate potential areas for improvement.



	coef
const	-0.7988
Do Not Email	-1.3394
TotalVisits	9.0912
Total Time Spent on Website	4.5516
Page Views Per Visit	-3.8015
Lead Origin_Lead Add Form	3.6853
Lead Source_Olark Chat	1.0670
Lead Source_Welingak Website	1.9524
Last Activity_Converted to Lead	-1.0660
Last Activity_Email Bounced	-1.1244
Last Activity_Olark Chat Conversation	-1.2396
What is your current occupation_Working Professional	2.8023
Last Notable Activity_Email Link Clicked	-1.9036
Last Notable Activity_Email Opened	-1.3407
Last Notable Activity_Modified	-1.6887
Last Notable Activity_Olark Chat Conversation	-1.4766
Last Notable Activity_Page Visited on Website	-1.8566

To Increase Lead Conversion Rates

- Features with positive coefficients should be targeted for marketing strategies.
- Develop strategies to attract high-quality leads from top-performing lead sources.
- Total visits and total time spent on website has very high positive co-efficients. So, increasing the number of ads in websites like welingak would help in landing lot of leads
- Incentives for providing reference that convert to lead, encourage providing more references.
- Working professionals should be targeted as they have high conversion rate, and they can also afford to pay higher fees too.
- Working professionals are also looking at a way to upskill themselves through online courses which is good opportunity for X education company

To identify areas of improvement

- Analyze negative coefficients for the features.
- Review those features to see what kind of improvement can be done.

Thank You