Leads scoring - Case study

Group members :
1) Pranjal Satish Tikande
2) Chaman Jha
3) Alistair Dsilva

Problem Statement

- X Education is an online education company who sells courses to Industry professionals
- They get their leads from multiple sources such as Google, past referrals, calls, emails. However the lead conversion rate at X education is around 30% which is very poor
- To make this process more efficient, the company wishes to identify the most potential leads, also known as ‘Hot Leads’.
- Our goal here is the identify Hot Leads by building a Logistic Regression model to help achieve the target of 80% conversion rate

Solution Methodology

- Data cleaning
  - Data imputing
  - missing value treatment
- EDA
  - EDA on Numerical Variables
  - EDA on Categorical Variables
  - Converting categorical variables to numerical variables by using preprocessing(Label encoder)
- Building a logistic regression model
  - Splitting the dataset in train and test set
  - Data scaling
  - Correlation matrix
  - Recursive Feature elimination - Course tuning
  - Feature elimination based on P-value and VIF - Fine tuning
  - Finding the optimal cutoff using the ROC curve
  - Prediction on the test set
- Prediction on test set
  - Precision and Recall tradeoff
  - Top feature analysis

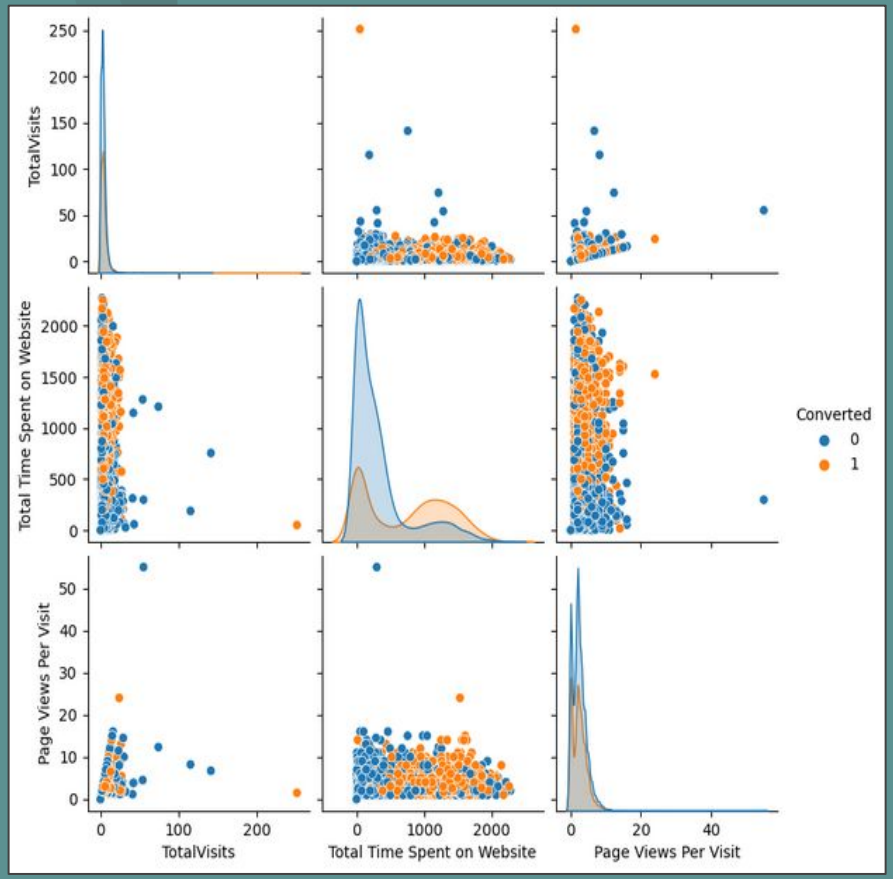
Data imputation/manipulation

- Replacing the missing values in ‘Lead Quality’, ‘Country’, ‘Specialization’, 'How did you hear about X Education', 'Tags', 'What is your current occupation' as ‘unknown’
- Replacing missing values in ‘City’, ‘Lead Profile’, 'What matters most to you in choosing a course' as ‘Other’
- Replacing ‘'Asymmetrique Activity Index' values with numerical values(1,2,3 for High, Medium and Low values)
- Created a list of all categorical columns and stored it in object_columns - used the LabelEncoder preprocessor to replace all feature values with dummy variables

Dropped columns: 'Prospect ID', 'Lead Number', 'Asymmetrique Profile Index','Asymmetrique Activity Score','Asymmetrique Profile Score' - as they did not all value

Dropped rows of columns with less than 2% missing values

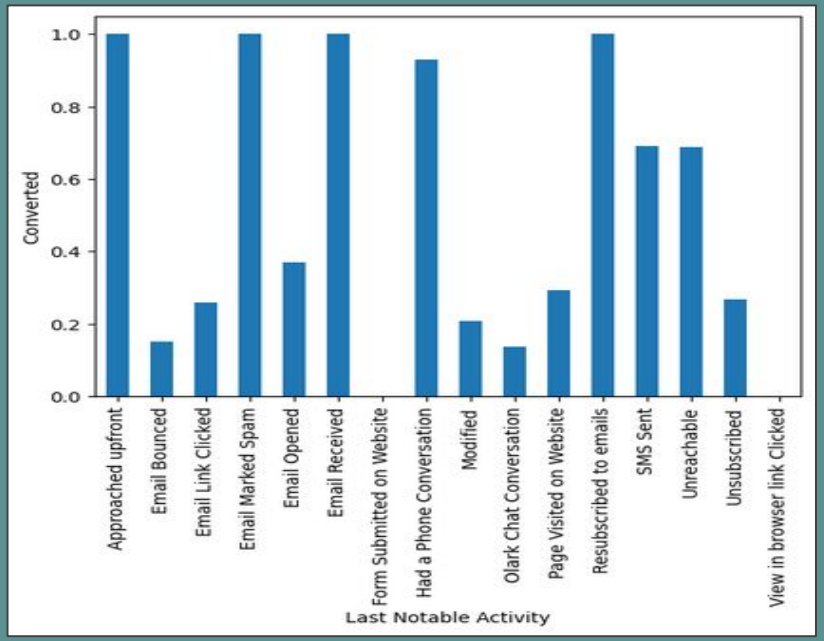
EDA

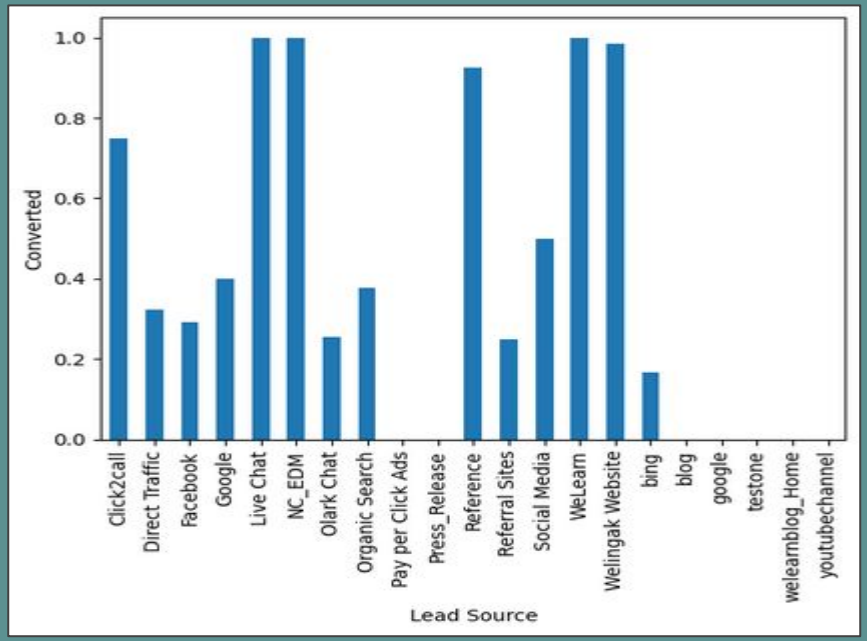


EDA on numerical columns

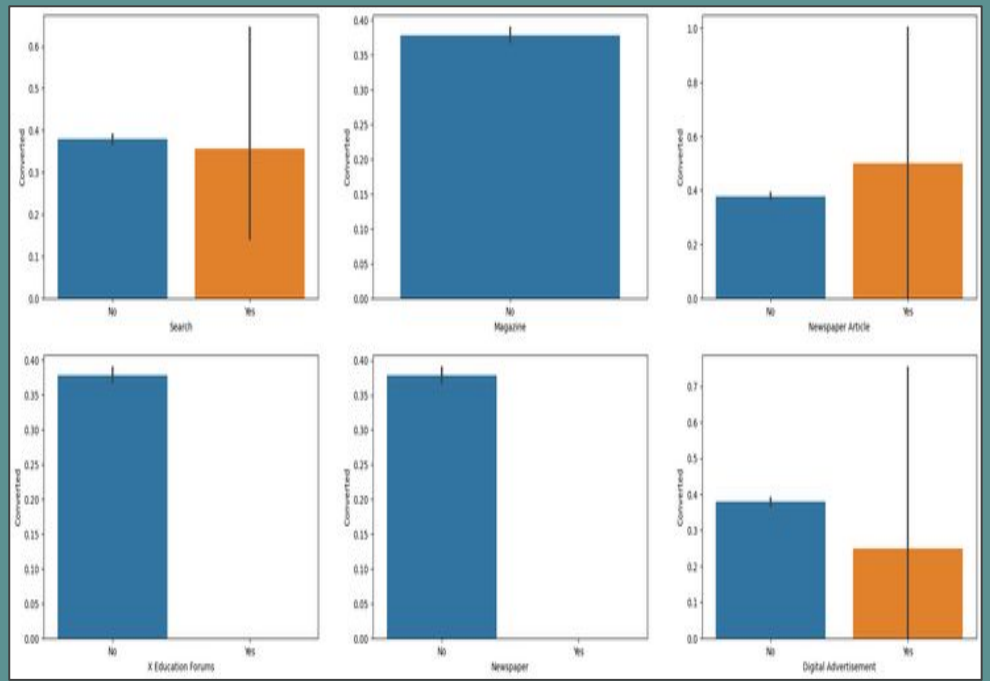
Here we can clearly notice that the 'page views per visit', 'Total time spent on website' and 'total visits' are directly proportionate with the target variable('Converted')

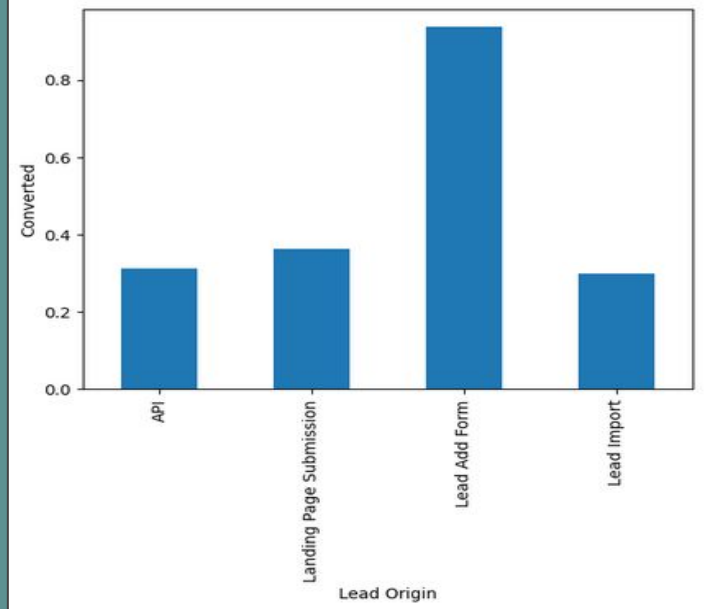
EDA on categorical variables #1

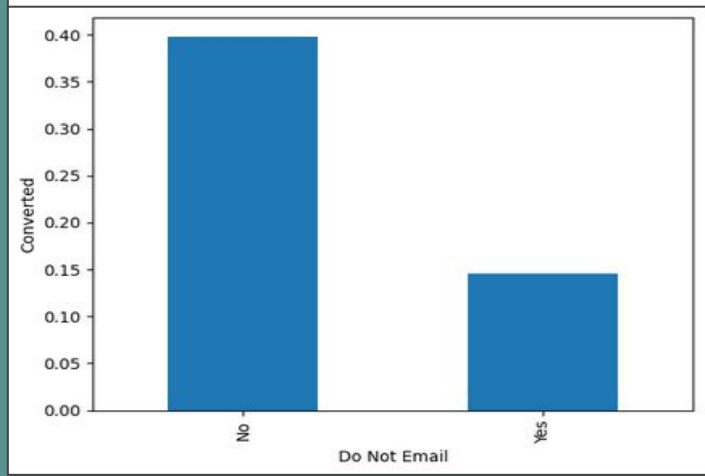




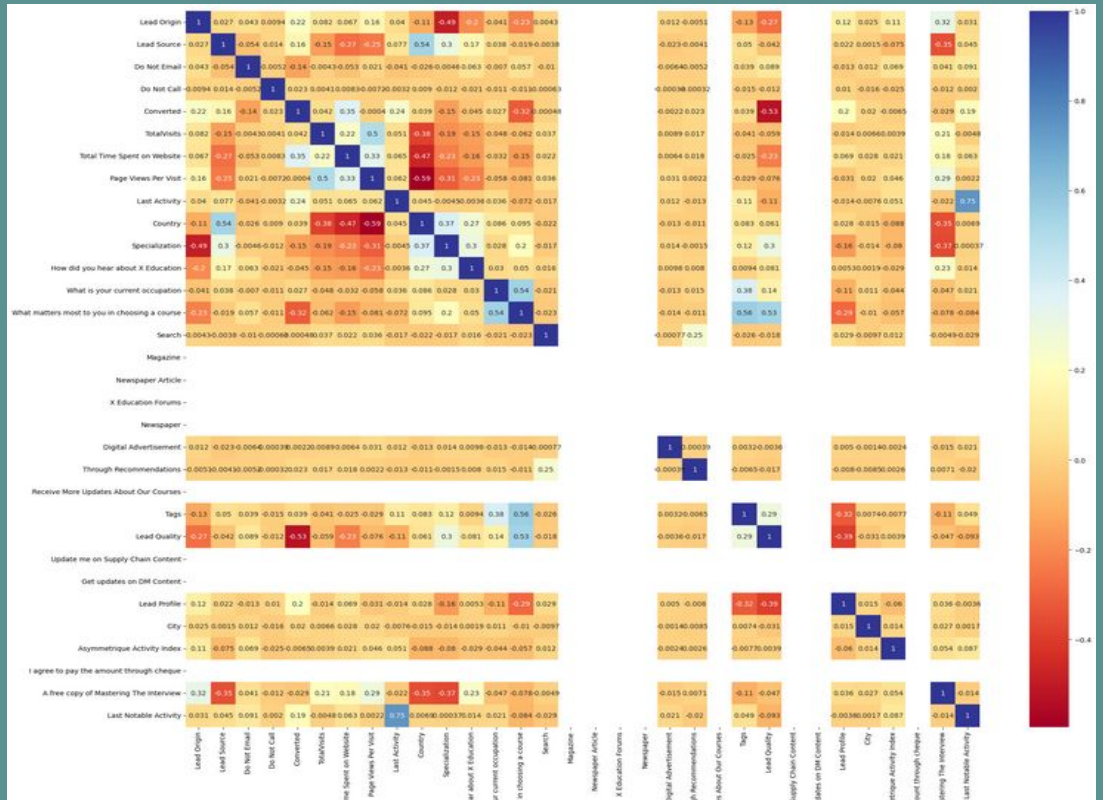
EDA Categorical variable - #2







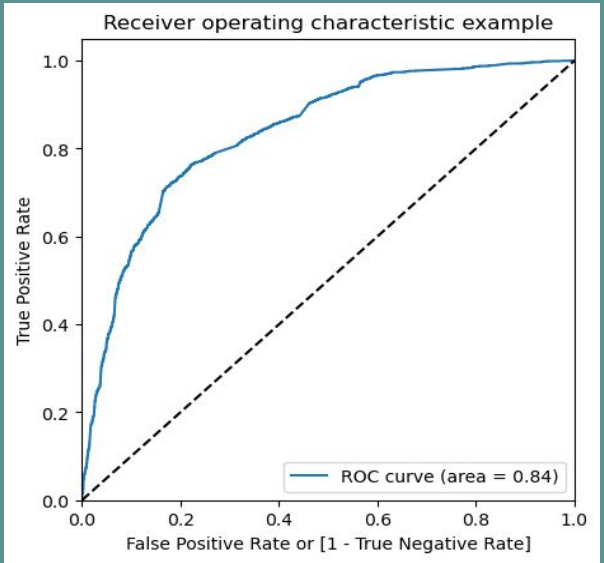
Correlation Matrix

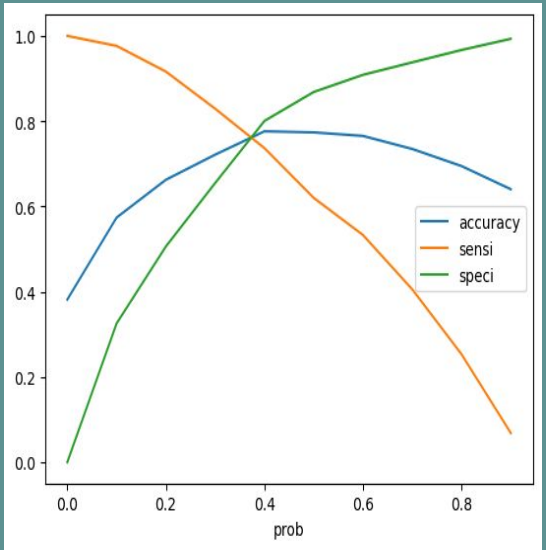


Model building

- Splitting the dataset into train and test sets
- Splitting the train data into X and y variables, where x contains all variables and y contains the target variable
- Conducting RFE on the X_train and y_train dataframes - RFE has selected 15 columns
- Feature elimination based on high VIFs and P-values - Cutoff 0.5
- With the current cut off of 0.5 VIF and 77% Accuracy. The sensitivity stands at 62% and specificity metric at 87%

ROC curve

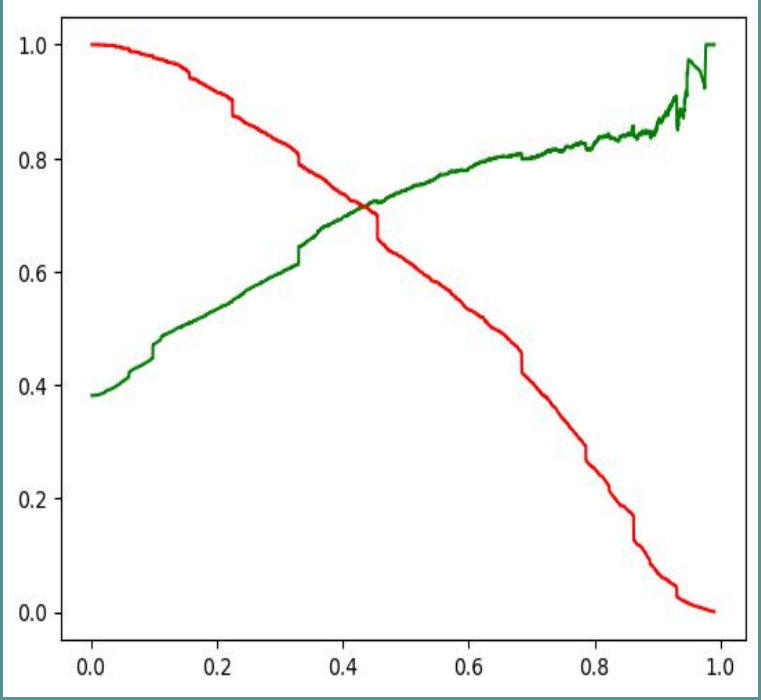




Finding the optimal cutoff

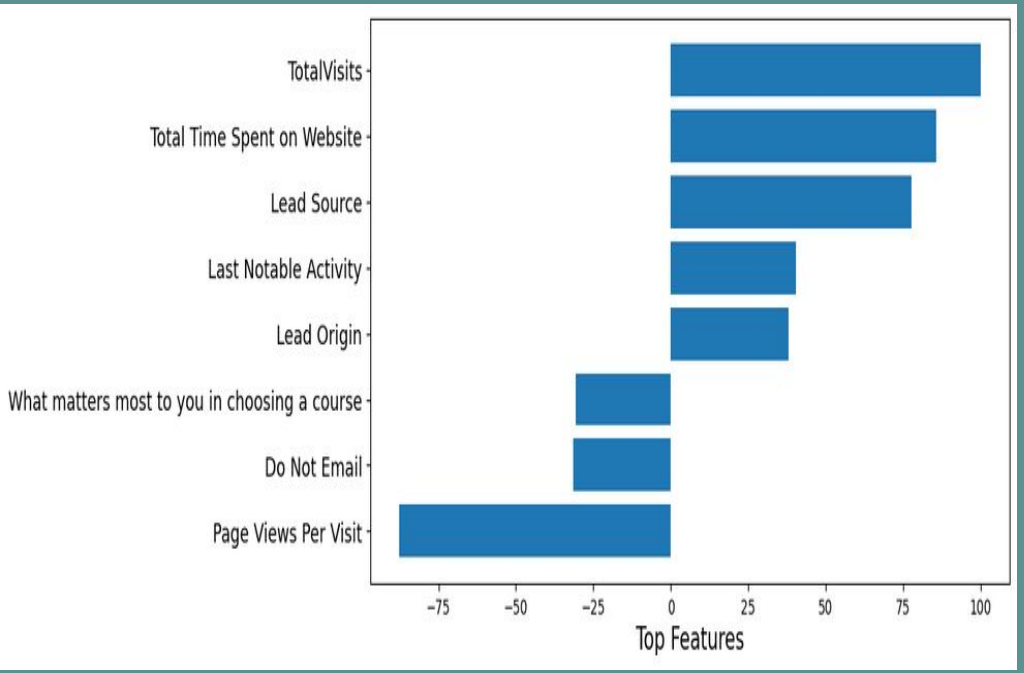
- The area under the ROC curve is 0.84, which is a decent figure
- From the graph it is visible that the optimal cut off is at 0.35

Precision-Recall Tradeoff



- Looking at the Precision and Recall curve - we conclude that the optimal cut off value is 0.41
- With the cut off at 0.41, the Accuracy stands at 78%, Precision at 67% and Recall at 76%

Conclusion



- In conclusion the top features that affect the ‘Conversion’ rate are :
  - Lead Origin
  - Lead Source
  - Do Not Email
  - TotalVisits
  - Total Time Spent on Website
  - Page Views Per Visit
  - What matters most to you in choosing a course
  - Last Notable Activity