Lead Scoring Case Study

Ashik Mohammed S

Gopi Jagini

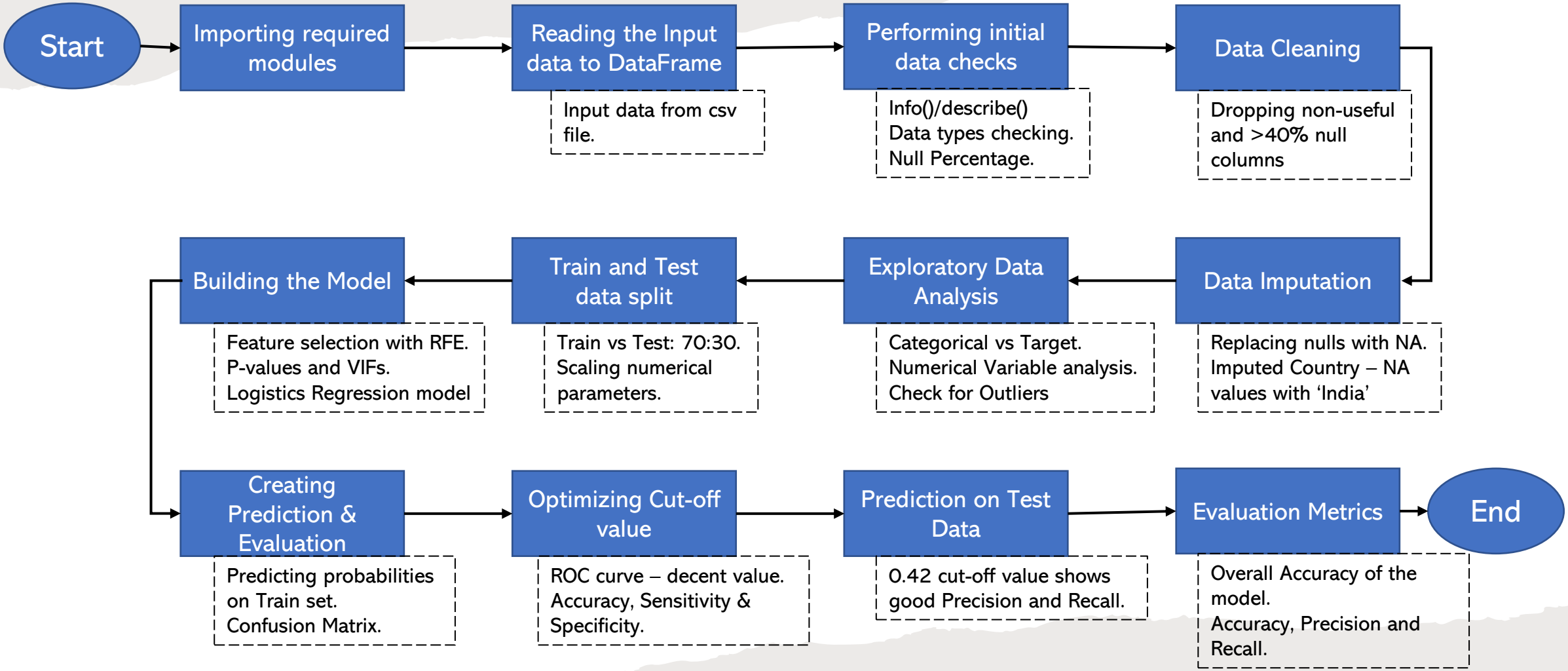
upGrad & IIITB | Data Science Program

Problem Statement:

- X Education sells online courses to industry professionals.
- People who are interested in the courses land on their website and browse for courses.
- When these people fill up a form providing their email address or phone number, they are classified to be a lead.
- The typical lead conversion rate at X education is around 30%.
- To make this process more efficient, the company wishes to identify the most potential leads, also known as 'Hot Leads'.
- The company wishes to build a model to assign a lead score to each of the leads to identify the hot leads (leads with high lead score).

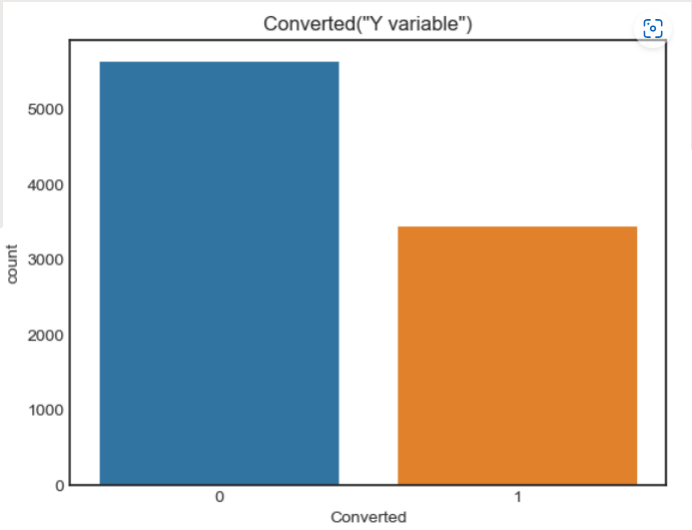
Goals	
	To build a logistic regression model to assign a lead score between 0 and 100 to each lead
	To provide adjustments in the model based on the company's requirements in the future

Analysis Approach:

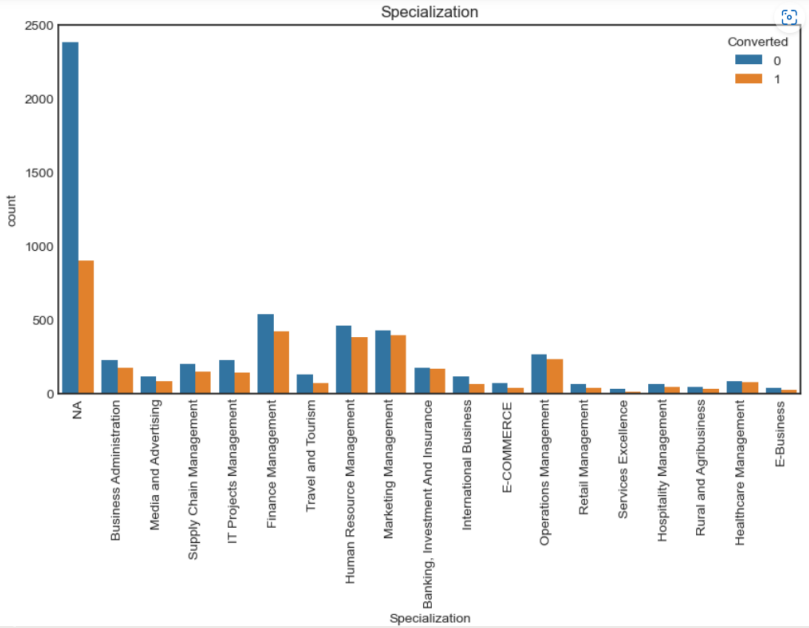


Note: Showing some for reference

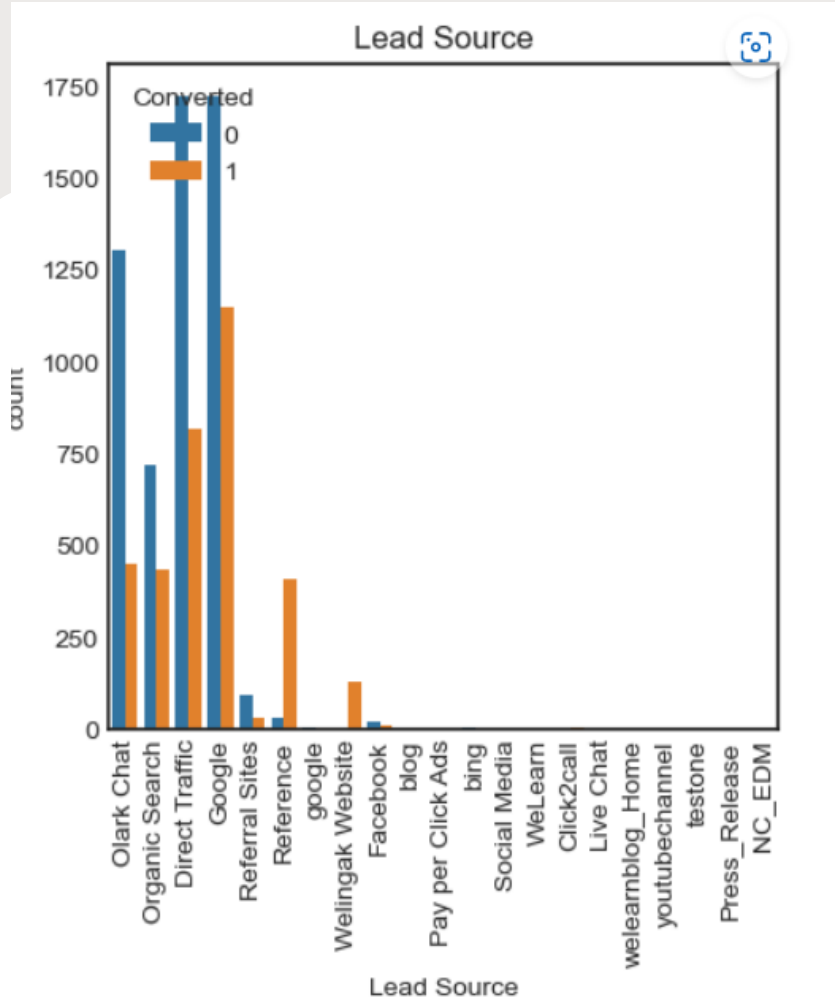
Exploratory Data Analysis:



- Count of Converted vs Non-Converted.
- Converts are low in count.



- Specialization vs Converted
- All Management related profiles have more entries
- NA category has more percentage of data
- Conversion percentage is almost same for all the categories



- Lead Score Vs Lead Source
- Direct Traffic and Google contribute to more percentage of leads
- The conversion rate through Reference is very high
- The conversion rate through Google search is also high

Model Building: Logistic Regression

```
# Doing LogisticRegression and RFE
lr = LogisticRegression()
rfe = RFE(lr,n_features_to_select=20)
rfe = rfe.fit(X_train,y_train)
```

- Logistic Regression Model
- Parametric selection using RFE

Model Evaluation:

```
array([[3492,  423],
       [ 717, 1661]], dtype=int64)
```

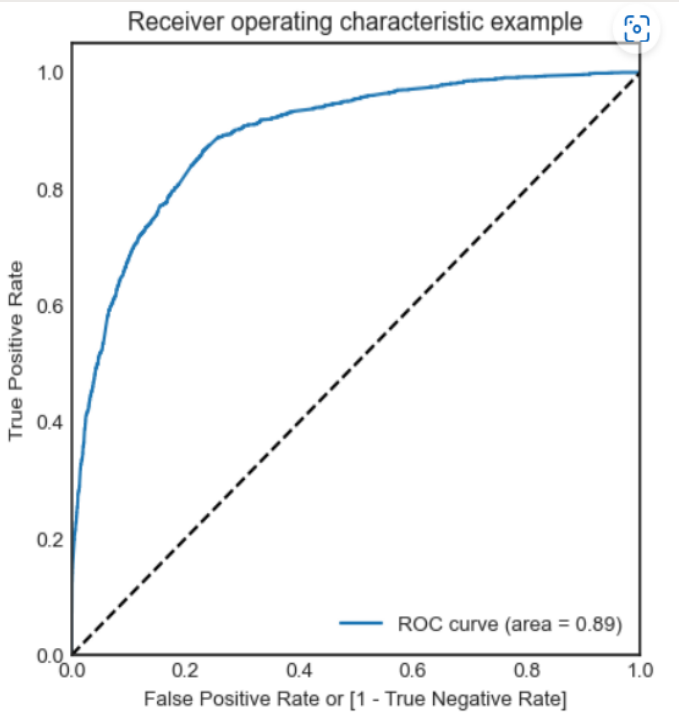
Confusion Matrix

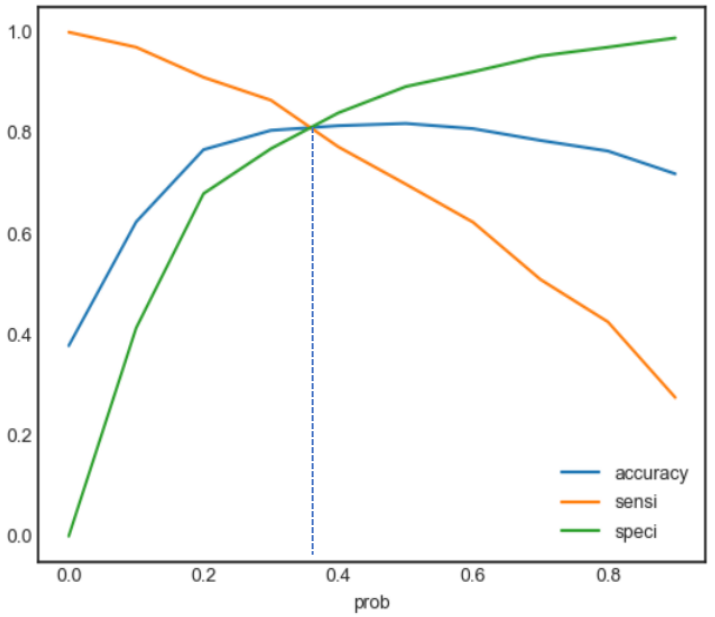
Accuracy: 81.3%
Sensitivity: 68.7%
Specificity: 88.9%

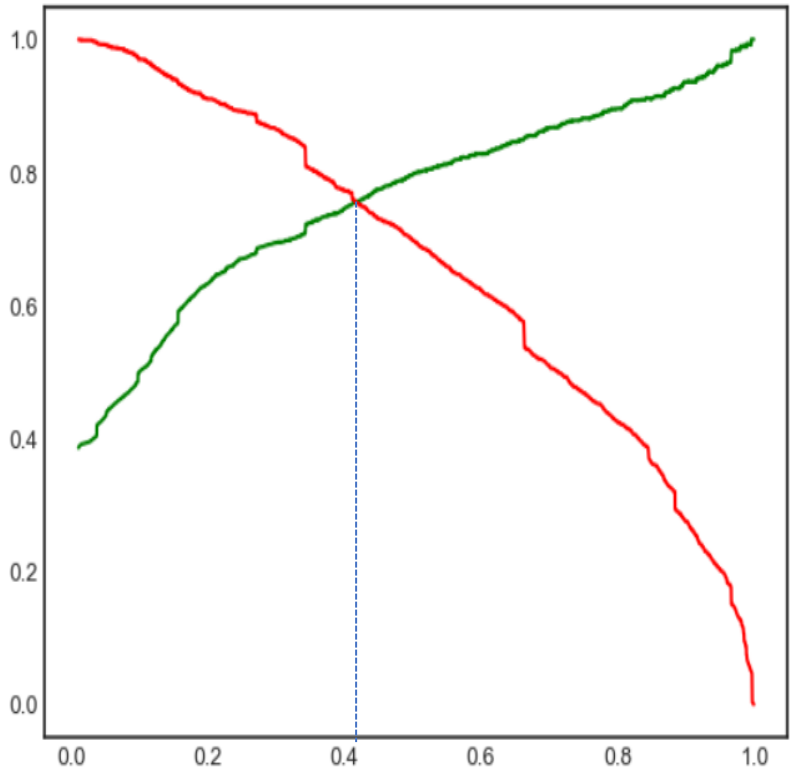
	Features	VIF
0	TotalVisits	2.51
11	Last Notable Activity_Modified	1.99
7	Last Activity_Olark Chat Conversation	1.97
1	Total Time Spent on Website	1.92
5	Do Not Email_Yes	1.88
6	Last Activity_Email Bounced	1.85
10	Last Notable Activity_Email Opened	1.68
3	Lead Source_Olark Chat	1.67
2	Lead Origin_Lead Add Form	1.51
12	Last Notable Activity_Olark Chat Conversation	1.35
4	Lead Source_Welingak Website	1.34
8	What is your current occupation_Working Profes...	1.17
13	Last Notable Activity_Page Visited on Website	1.15
9	Last Notable Activity_Email Link Clicked	1.06

For Feature Selection:
- P-value less than 0.05
- Decent VIF values

Model Optimization:







Area under ROC curve is more (near to 1). It indicates that, the model performance is good.

The above graph shows the optimized probability cut-off for better Accuracy, Sensitivity and Specificity.

Precision vs Recall Tradeoff: Indicates Optimized probability cut-off value of **0.42**

The final cut-off value chosen was 0.42. Which resulted in overall Accuracy of 80.83%, Precision of 75% and Recall of 81%

Recommendations:

- Top three variables which have impact on the conversion percentage are:

  Total Visits, Total Time Spent on Website, Lead Add Form.

- Top 3 categorical/dummy variables in the model which should be focused the most, for high lead conversion, are:

  Total Time Spent on Website, Lead Origin_Lead Add Form and What is your current occupation_Working Professional

- The emphasis should be more on the Lead Sources.

  For Eg: Through Reference, the conversion rate/percentage is very high

  Through Google Search, quantity of conversions is very high

- Management related professionals should be targeted more for good conversion.
- New age social media platforms like Instagram, Twitter should be utilized for promotions (rather than Facebook).


THANK YOU