Leads Scoring Case Study

Problem Statement

- X Education is an education company which sells online courses to industry professionals.
- Once customers land on the website, they might browse the courses or fill up a form for the course or watch some videos.
- People who fill up the form providing their email address or phone number are classified as a lead.
- Some leads even come through referrals.
- The typical lead conversion rate at X education is around 30%.
- Now the firm want to identify the most potential leads, also known as ‘Hot Leads’.
- The aim here is to build a model wherein a lead score will be assigned to each of the leads.
- The customers with a higher lead score have a higher conversion chance and vice versa
- We are given leads dataset from the past with around 9000 data points and target variable 'Converted'.

Analysis Approach

Reading and Understanding Data:

- Read and understood the data. Used data dictionary whenever necessary
- The dataset had 9240 records and 37 fields.
- Upon analysing further, identified multiple columns with missing values

```
leads.isna().sum().sort_values(ascending=False)/leads.shape[0]*100
```

```
Lead Quality                                     51.590909
Asymmetrique Activity Index                      45.649351
Asymmetrique Profile Score                       45.649351
Asymmetrique Activity Score                      45.649351
Asymmetrique Profile Index                       45.649351
Tags                                             36.287879
Lead Profile                                     29.318182
What matters most to you in choosing a course    29.318182
What is your current occupation                  29.112554
Country                                          26.634199
How did you hear about X Education               23.885281
Specialization                                   15.562771
City                                             15.367965
Page Views Per Visit                              1.482684
TotalVisits                                       1.482684
Last Activity                                     1.114719
Lead Source                                       0.389610
```

Data Cleansing:

- Performed data cleansing to eliminate missing values and any other kind of irregularities
- Dropped columns with 35% or above missing values
- Evaluated the values in various columns and identified columns with data imbalance
- Dropped the columns with very high data imbalance
- As per the business understanding achieved, dropped non contributing features like ID, country etc.
- Columns with “Select” values were also handled like missing values
- Fields with very high percentage of “Select” values were dropped, except the column “Specialization”
- For columns with smaller percentage of null values, dropped the null value records
- At the end of the data cleansing activities we were able to retain 69% of the records and 12 attributes.

Data Preparation:

- Prepared the data for modelling by properly handling numerical and categorical variables
- Created dummy variables for categorical variables
- While creating dummy variable for “Specialization” special care was taken to drop one of the dummy records manually
- This was done to eliminate the dummy variable corresponding to “Select” value in “Specialization”

Train-Test Split

- To proceed with modelling, we divided our variables into feature variable set (X) and target variable (y)
- Performed train test split on the above datasets in the ratio 70:30

Feature Scaling

- Performed feature scaling on the numeric variables using MinMaxScaler
- Fit_transform() was applied on train and transform() was applied on test set

Feature Correlations

- Created correlation matrix for the feature variables
- Due to large number of variables was unable to make any conclusions

Feature Selection

- Used RFE to select features from the train dataset
- Selected a list of 15 features from a set of 74 using RFE

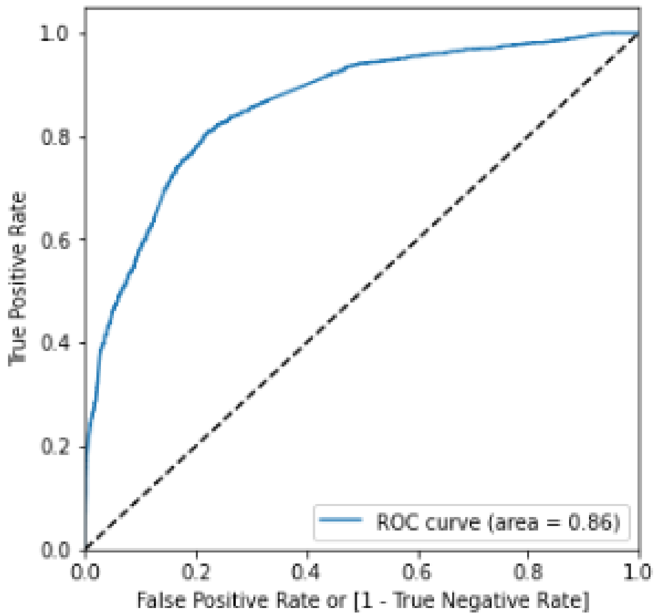
Model Building

- The first model built the model using the RFE features had high p value and vif values
- Eliminated the high-p high-vif record and repeated the model creation and elimination process
- The Model 5 was finalised due to optimal p and vif values

	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

Model Evaluation

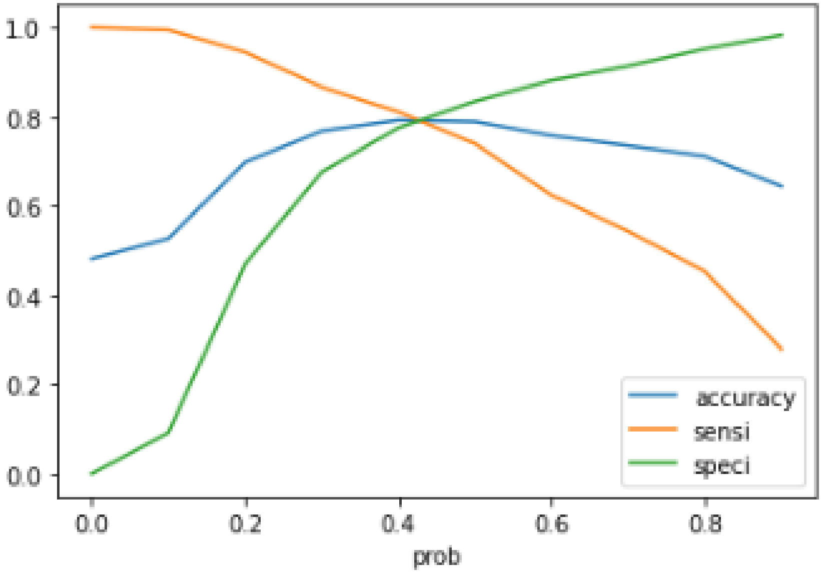
- Model evaluation was performed by making predictions on train with an arbitrary cut off of 0.5
- Accuracy, sensitivity, specificity values were 78.86%, 73.94% and 83.43% respectively
- To evaluate the model further ROC curve was plotted



- The curve is as expected denoting the model is effective.
- The area under the curve is 0.86, which is a good value for a model

Optimal cut-off

- To get an optimal cut-off value, plotted trade off between accuracy, sensitivity and specificity
- The optimal value was identified 0.42

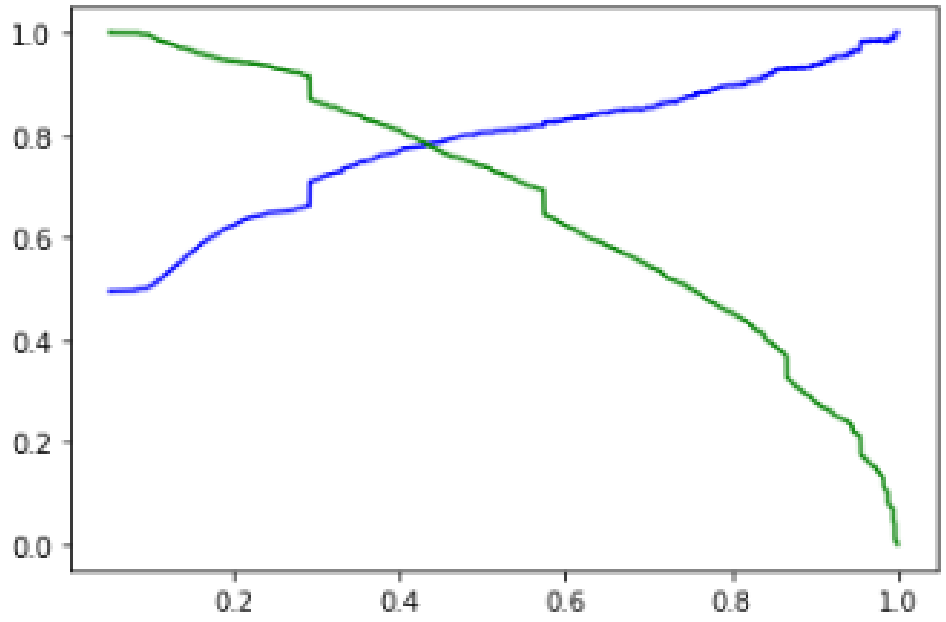


Evaluation of new threshold

- The model was evaluated with new threshold
- Accuracy, sensitivity, specificity values were 79.08%, 79.33% and 78.84% respectively
- These matrix values are all in optimal range

Precision and Recall

- In addition to other metrices, calculated precision and recall
- The precision and recall values 77.71% and 79.33% with cut-off=0.42
- Created the precision-recall trade off plot
- Even this plot gave the optimum threshold value as 0.42



Predictions on Test Set

- The model was tested on the test set with threshold=0.42
- The effectiveness was evaluated using the metrices
- Confusion matrix was created and evaluation was performed
- Accuracy, sensitivity, specificity values were 78.45%, 77.94% and 78.91% respectively
- The precision and recall values 77.27% and 77.94%
- The matrix values of test is very comparable to values received for train
- This makes sures that the model is effective and not overfitting

Suggestions:

As per the model, here are few suggestions to improve the conversion rate

- Focus on those customers who had the highest number of visits to the website
- Customers who has spent maximum amount of time on website could be prioritized
- There is a very high chance of these customers getting converted.
- People who are unemployed and students have lesser chance of enrolling
- Better move such customers to lesser priority list
- Telephonic communication could be focussed since that seems to have more impact
- Customer who has marked Do Not Email as Yes has very less chance of conversion
- Customers who was marked as lead by adding form has higher chance of converting to lead
- Focusing on the above parameters could heavily impact the conversion rate

Thank you!