Summary

To build the prediction model for an Education X company we have used the below steps

EDA:

The data check is done and a few columns were converted into more readable for and below observations found

- Prospect ID and Lead Number are both unique identifiers. which don't serve any purpose in data analysis, hence we can drop these columns.
- The Column 'Tags' is the one filled by the Sales team for their reference, as per business it has no relevance to the outcome i.e lead conversion, hence dropping this as well.
- Also, the Columns filled by the sales team are 'Last Activity', 'Last notable Activity', and 'lead quality' are also okay to drop.
- There few columns like 'Specialization', and 'source' that contains values as 'Select', which looks like they drop down to collect data and not mandatory fields, so we can convert these 'Select' words to NULL
- Columns with more than 45% NULL values have been dropped.

The country column's NULL value has been derived where City information was available.

Data Imputation:

- The Country, City, and Lead Source column's data have been imputed with the most common value using the mode function.
- Specialization, course selection, and occupation column data have been imputed with a random selection method.
- Total Visits and Page Views per visits missing value has been imputed with their median value.

There are few columns with only one unique value

- Magazine
- courses_updates
- supply_chain_content_updates
- dm_content_updates
- cheque_payment

Since the above-mentioned columns have only ONE value, hence these don't contain any relevance. So dropping them from the dataset

These are binary data columns, having only 'Yes' and 'No' as values
These columns show DATA IMBALANCE, as almost all records have the same value
Because of heavy data imbalance, we can drop the following columns as well

* We are not dropping mastering_interview column as doesn't show data imbalance
* We are not dropping 'Do Not Email' column as well as its has some data

Uni variate Analysis

Observations:

1) Landing Page Submission is a good lead Origin, it has a better chance of Conversion in comparison to others
2) Online Search engines are a good lead source, it has 90% of the traffic
3) Data mostly belongs to India and in that predominately from Mumbai City
4) The 'Management' category of course is the most popular, it has more than 70% of leads
5) The 'Unemployed' section has most of the inquiries
6) People going to course for 'Better carrier choice'

Numerical data analysis

Observations:

1) Time spent on the website has a strong and positive relation with the target variable 'Converted'
2) Page views per visit has a week and negative relation with the target variable 'Converted'

Outlier Analysis

Observations:

1) Look at 1st and 3rd box plots and the statistics, there are upper bound outliers in both total_visits and page_views_per_visit columns.
2) We can also see that the data can be capped at the 99 percentile.
3) We are not going to treat Outliers as of yet.

Data Preparation and Model Building:

Dummy value creation

'Do Not Email', 'mastering_interview' are binary data columns

Creating dummy variables for categorical columns

'Lead Origin', 'Lead Source', 'Country',
'Specialization',
'occupation', 'course_selection_reason', 'City'

Test and Train data split

Split the data-set into 70% and 30% for train and test respectively with the random state as 10.

Model Building

Since a number of feature is more let go with the
RFE method first and then we will use the manual method to further fine tune the model
Removed feature with more than 5 VIF values and more than 0.05 as p-value in multiple iterations.

With the cut-off as 0.5, we have around

Accuracy of 79.1%

sensitivity of 64.2%

Specificity of 88.52%.

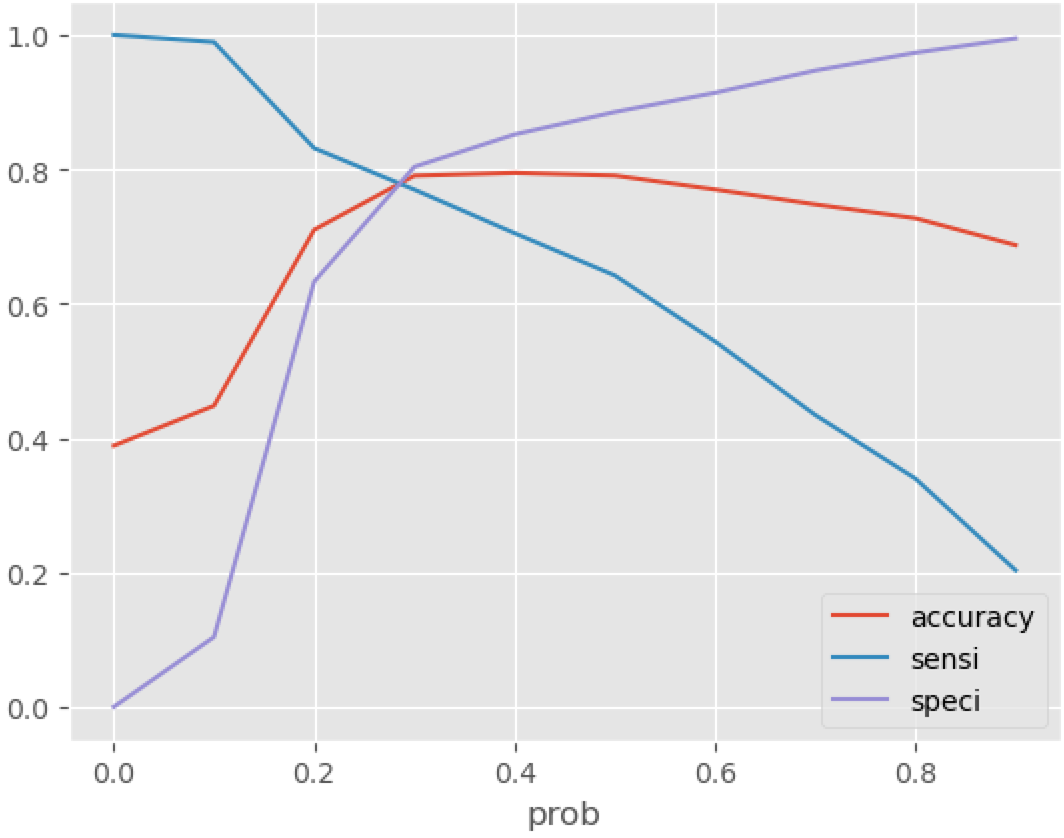
Optimise Cut off (ROC Curve)

An ROC curve demonstrates several things:

It shows the tradeoff between sensitivity and specificity (any increase in sensitivity will be accompanied by a decrease in specificity).

The closer the curve follows the left-hand border and then the top border of the ROC space, the more accurate the test is.

The closer the curve comes to the 45-degree diagonal of the ROC space, the less accurate the test is.



The area under the ROC curve is 0.83.
ROC Curve gave optimal cut-off value as 0.3
With the cut-off value as 0.3 the model parameters are

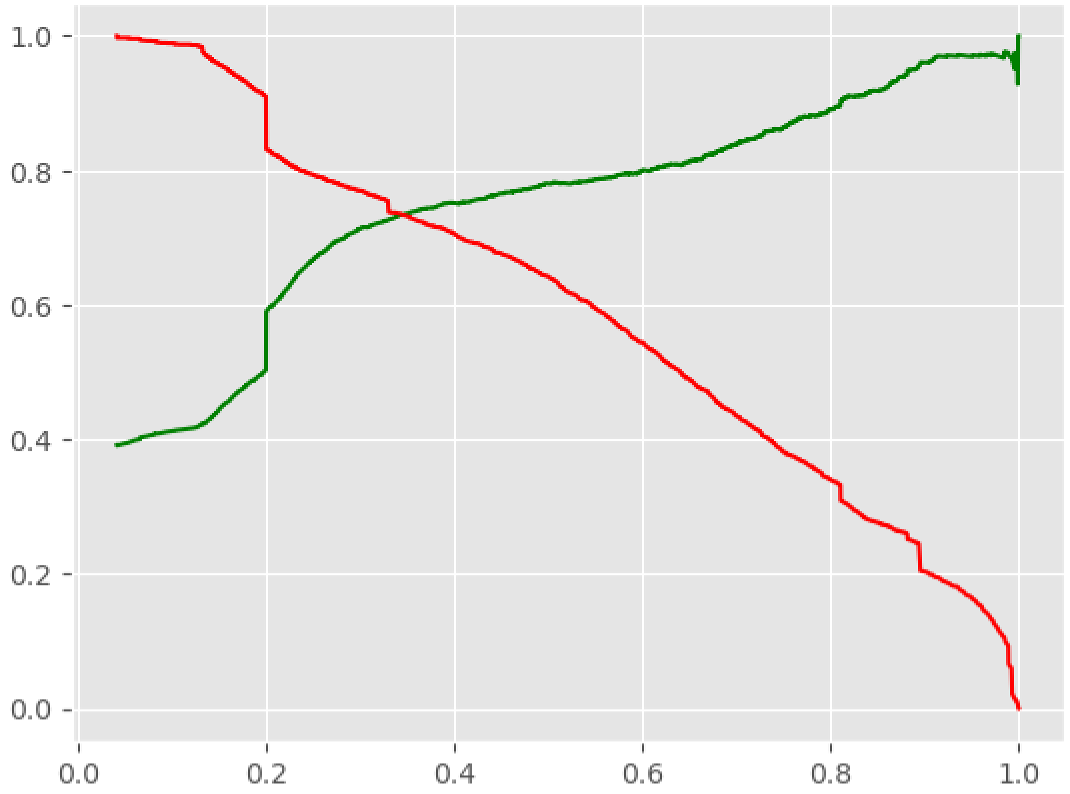
Accuracy= 79.07 %
sensitivity= 64.2 %
specificity= 88.52 %

Prediction on Test set

Model parameters on the test set is as followed

Accuracy= 78.93 %
sensitivity= 79.85 %
specificity= 78.38 %

Precision-Recall



The Precision-Recall curve shows that the cut-off is at 0.35

With the cut-off value at 0.35 the model parameters for the train data-set is
Accuracy= 79.07 %
sensitivity= 64.2 %
specificity= 88.52 %

With the same cut-off value the model parameters for test data-set is
Accuracy= 79.11 %
sensitivity= 74.88 %
specificity= 81.68 %

Overall: With the current cut off as 0.35 we have accuracy=79.22%, sensitivity=74.88% and specificity of around 81.68% for test dataset

Observation: There is a slight improvement in Accuracy and Specificity values whereas the significant drop in Sensitivity value with this new cutoff of 0.35 on test data set

Conclusion

Below is the list of parameters/feature which helps the predicting the Course sales

- Do Not Email
- TotalVisits
- time_on_website
- Page Views Per Visit
- Lead Origin_Landing Page Submission
- Lead Source_Olark Chat
- Lead Source_Reference
- Lead Source_Welingak Website

Specialization_Operations Management
occupation_Working Professional