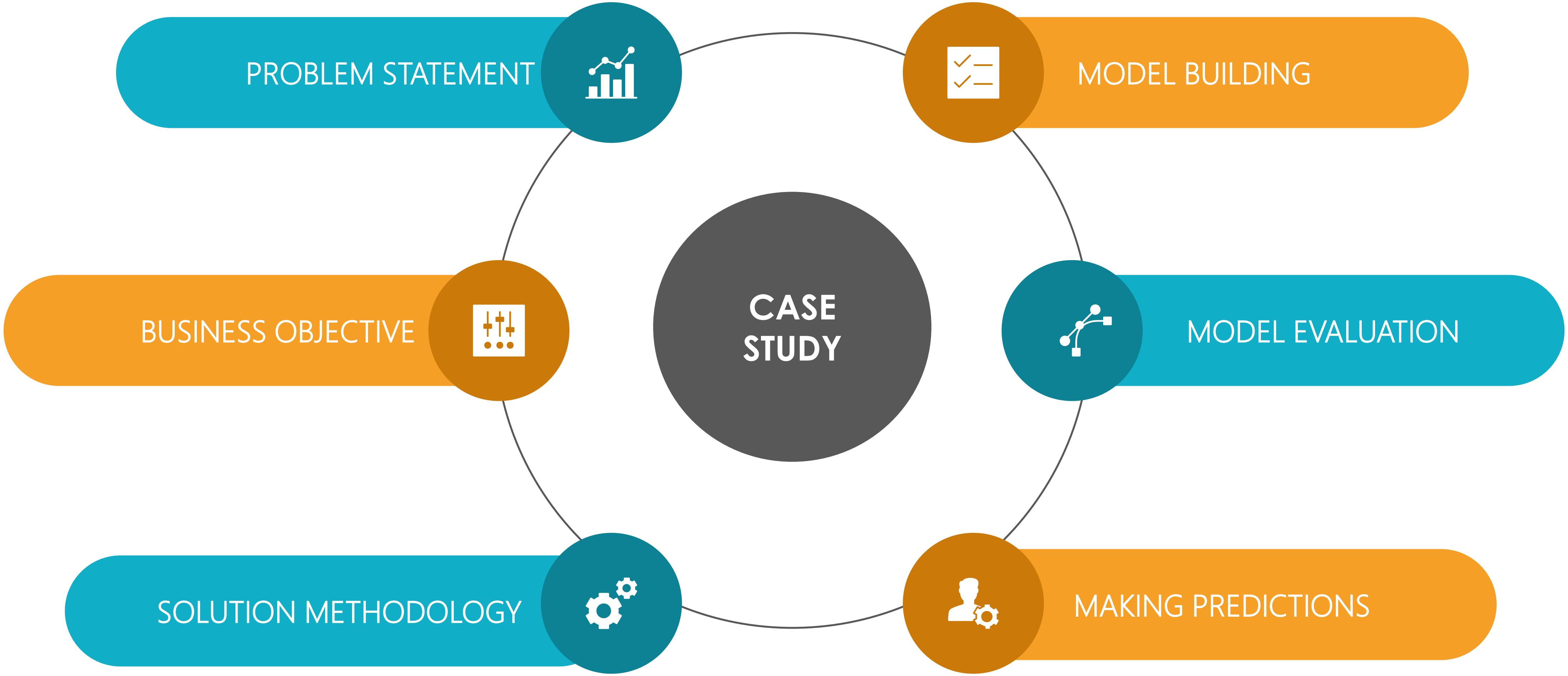
Lead Score Case study

Presentation

By
1. Arpitha V
2. Natasha Najeeb

Date : 03rd Jan 2023

Case study analysis



CASE STUDY

PROBLEM STATEMENT

BUSINESS OBJECTIVE

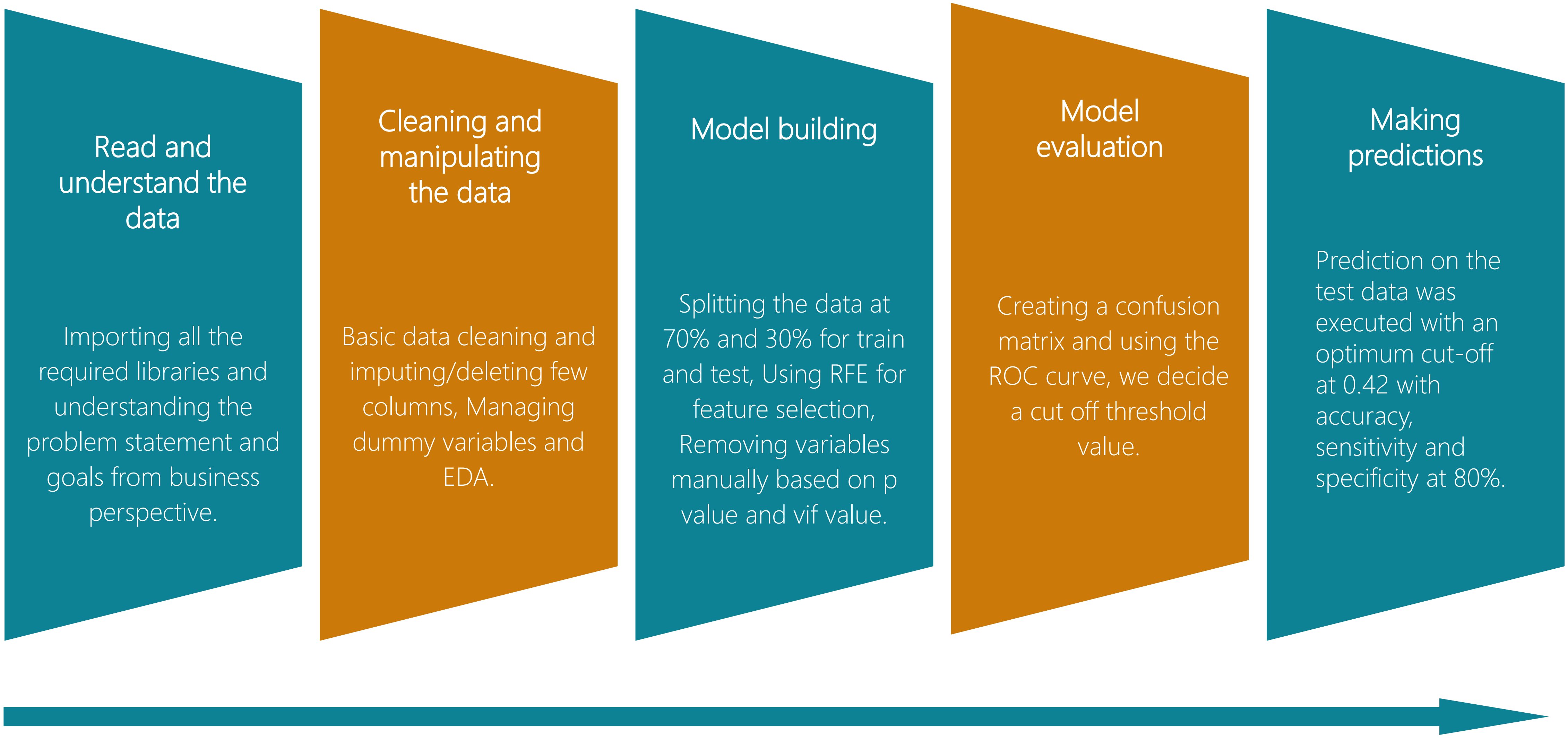
SOLUTION METHODOLOGY

MODEL BUILDING

MODEL EVALUATION

MAKING PREDICTIONS

Case study analysis



Read and understand the data

Importing all the required libraries and understanding the problem statement and goals from business perspective.

Cleaning and manipulating the data

Basic data cleaning and imputing/deleting few columns, Managing dummy variables and EDA.

Model building

Splitting the data at 70% and 30% for train and test, Using RFE for feature selection, Removing variables manually based on p value and vif value.

Model evaluation

Creating a confusion matrix and using the ROC curve, we decide a cut off threshold value.

Making predictions

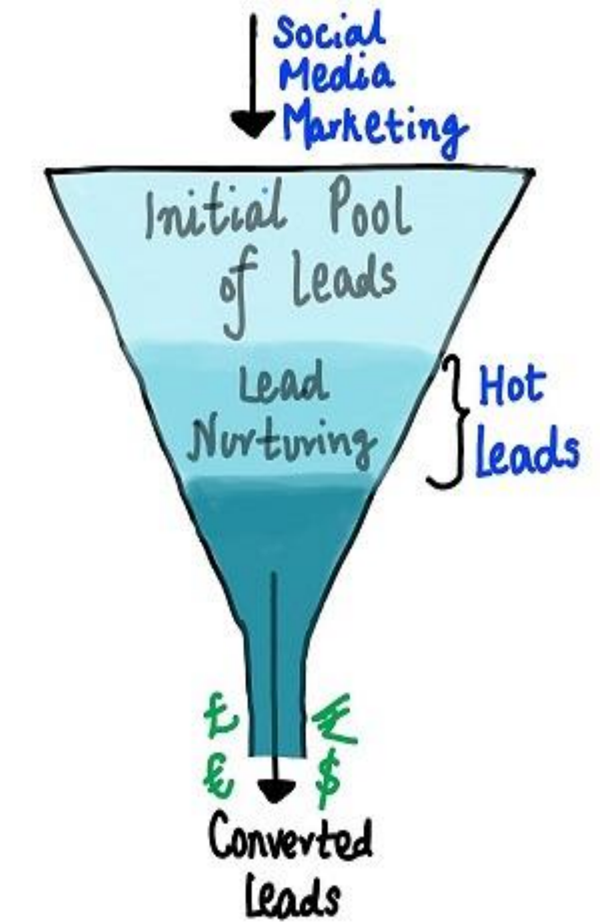
Prediction on the test data was executed with an optimum cut-off at 0.42 with accuracy, sensitivity and specificity at 80%.

Case study analysis

Problem Statement

An education company named X Education sells online courses to industry professionals. 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%.

If they successfully identify the most likely to convert to paying customers set of leads, the lead conversion rates should go up as the sales team will now be focusing more on communicating with the potential leads rather than making calls to everyone. The target lead conversion rate is around 80%.



Case study analysis

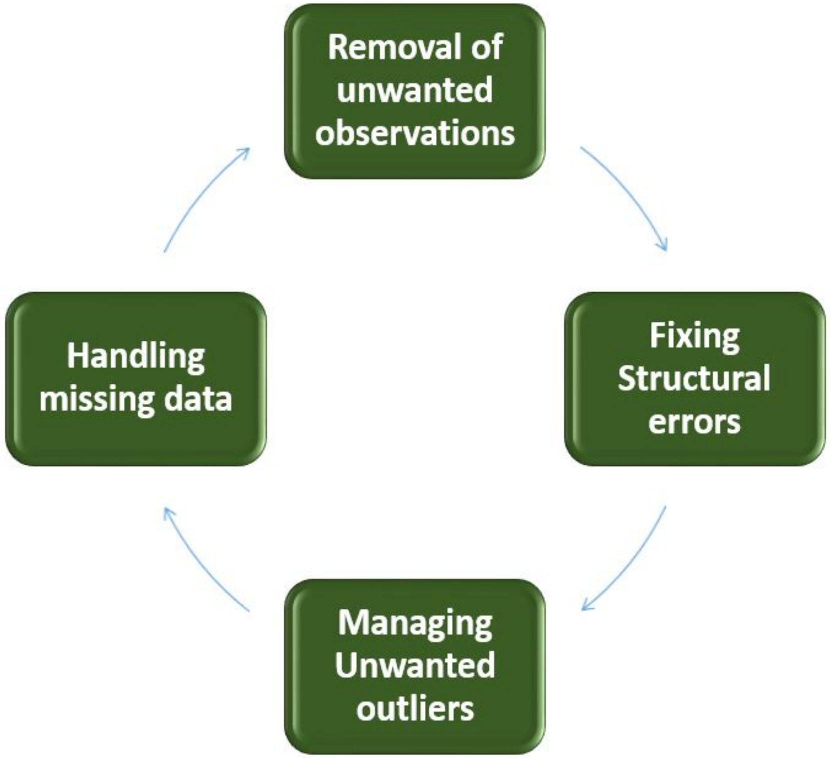
Goal of the case study

- Build a logistic regression model to assign a lead score between 0 and 100 to each of the leads which can be used by the company to target potential leads. A higher score would mean that the lead is hot.
- To build a model which identifies the ‘hot leads’.
- There are some more problems presented by the company which the model should be able to adjust to if the company's requirement changes in the future so we will need to handle these as well.

Case study analysis

Data Cleaning and EDA

- Specifications of the given dataset: (9240, 37)
- Dropping columns with null values >3000 records
- Columns like 'Magazine', 'Newspaper Article', 'X Education Forums', 'Newspaper', 'Digital Advertisement', 'Through Recommendations', 'Receive More Updates About Our Courses', 'Update me on Supply Chain Content’, have been dropped.
- Imputing null values in some columns
- Removing redundant columns in the dataset



Removal of unwanted observations

Fixing Structural errors

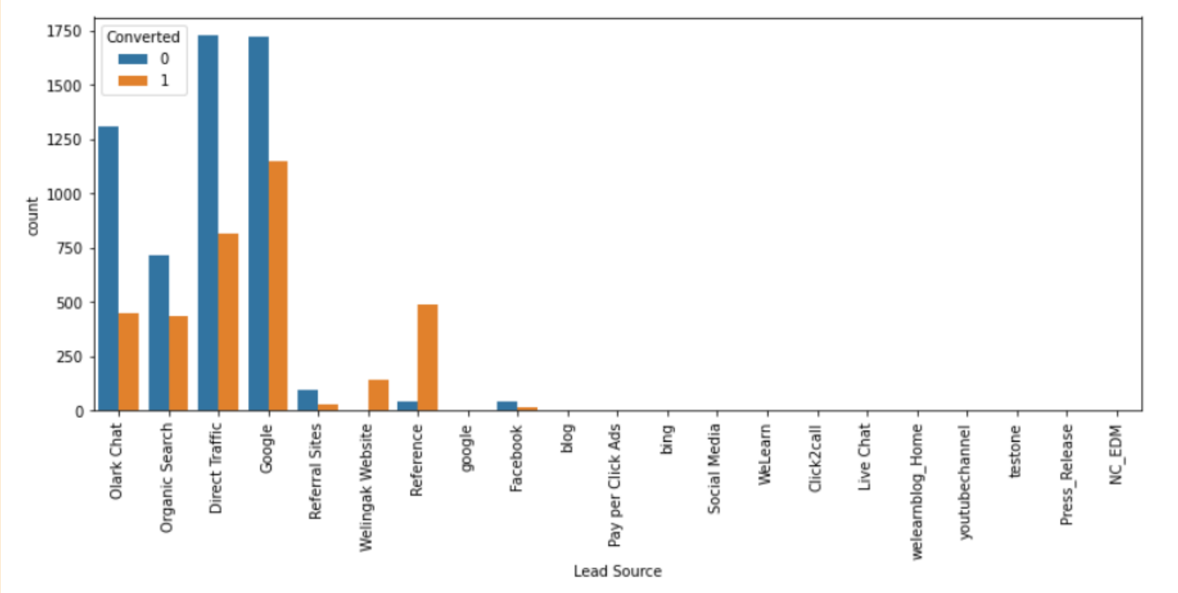
Managing Unwanted outliers

Handling missing data

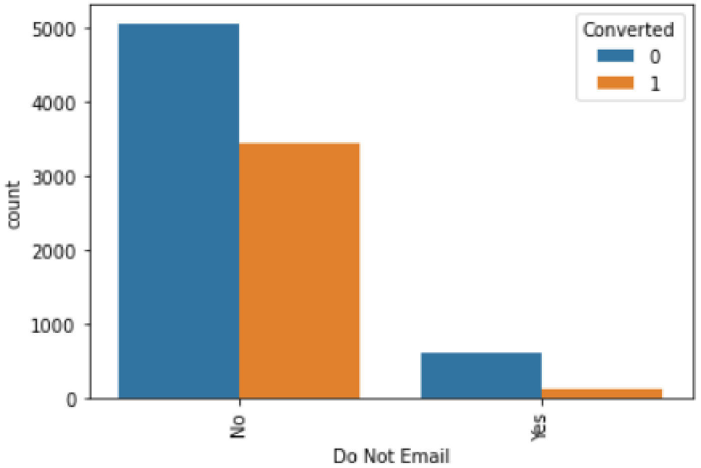
Case study analysis

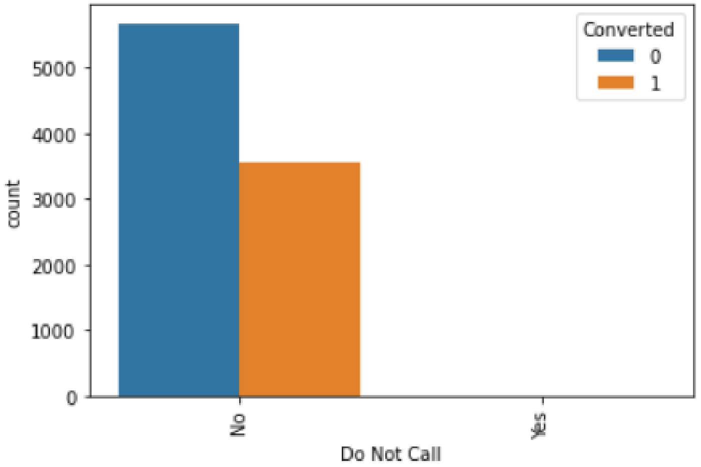
EDA

1. Lead Source

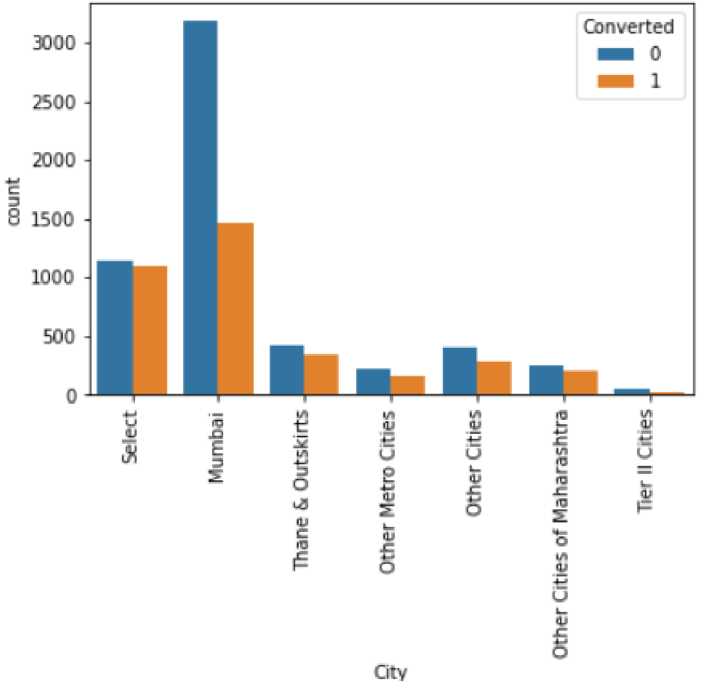


2. Do Not Email
And do not call have
 similar inferences.

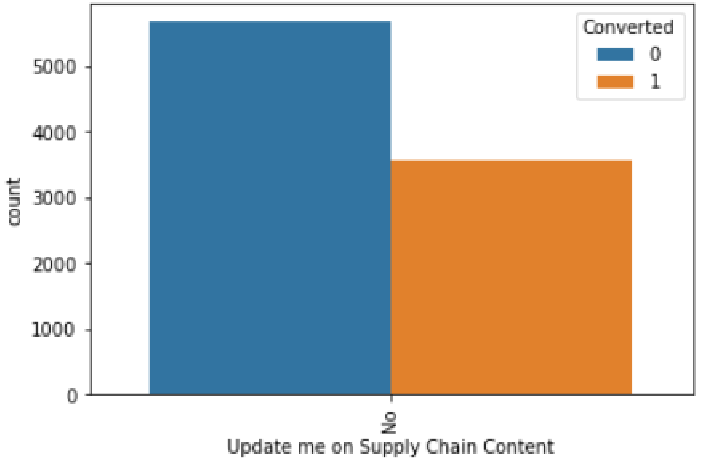


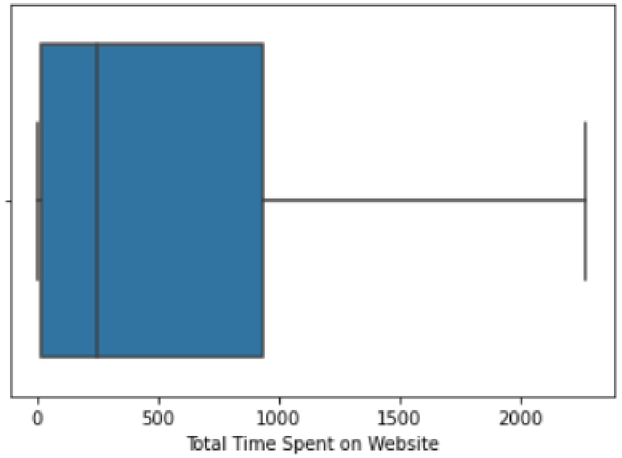


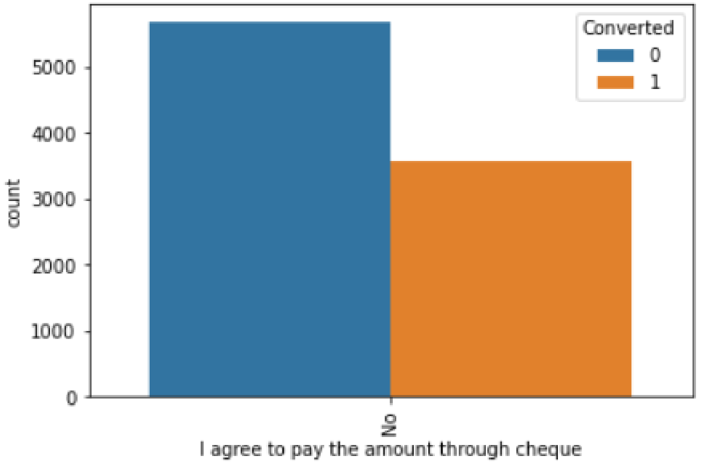
Case study analysis

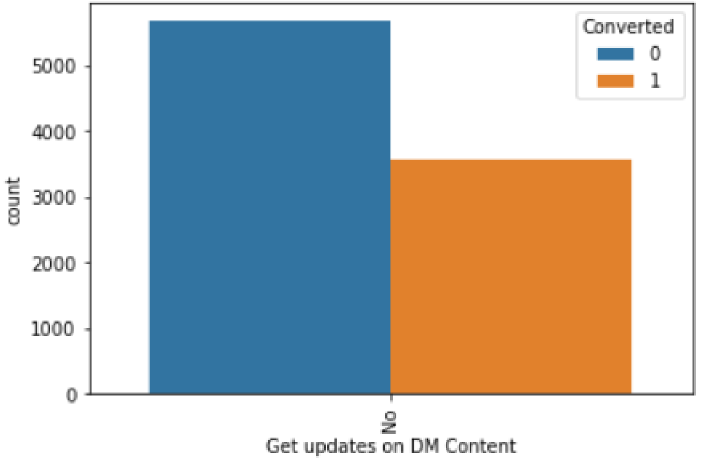


Most variables have only “NO” as the value.
Hence no inference can be drawn from these.





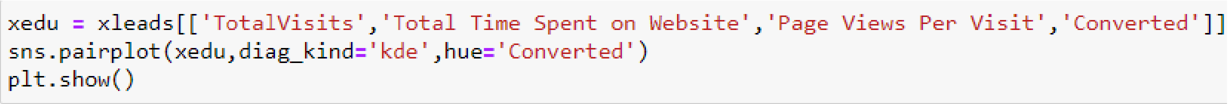




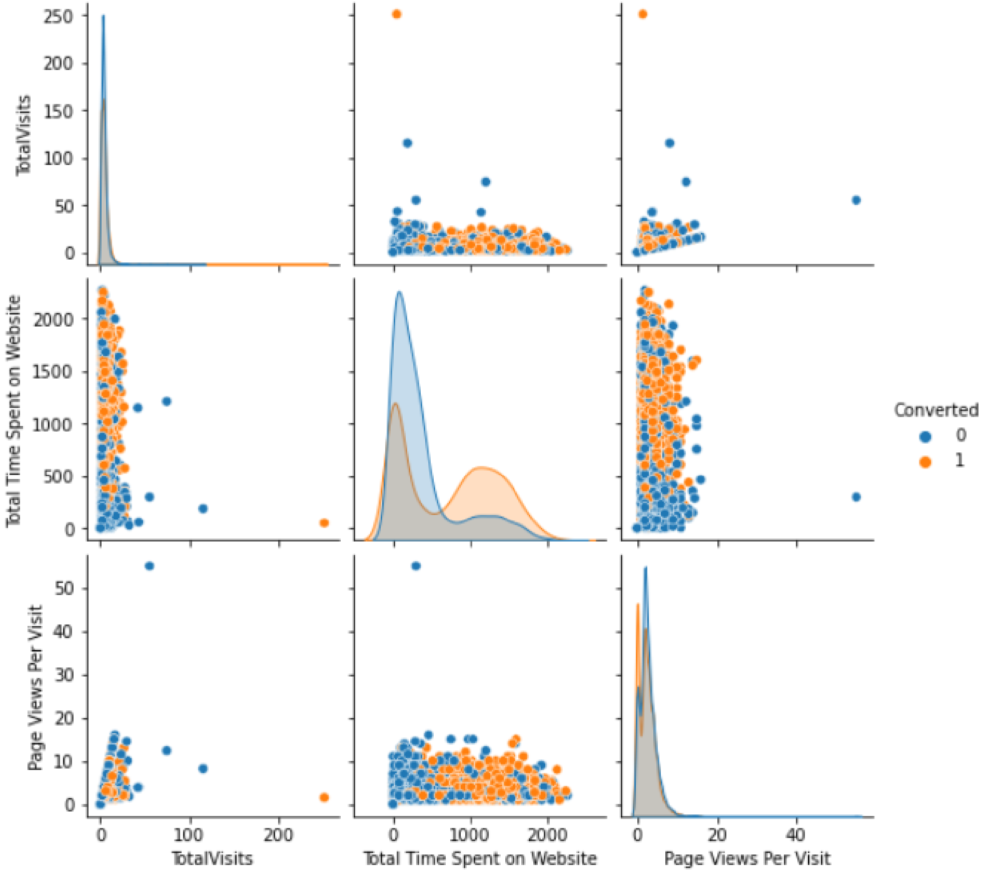
Case study analysis

Data Preparation

- Converting some binary categorical variables (Yes/No) to 1/0
- Adding dummy variables for the categorical features 'Lead Origin', 'Lead Source', 'Last Activity', 'Specialization', 'What is your current occupation', 'City', 'Last Notable Activity'
- Splitting the Data into Training and Testing Sets.we have chosen 70:30 ratio.
- For Scaling, we use MinMaxScaler.
- Use RFE for Feature Selection.
- Running RFE with 15 variables as output.



```
xedu = xleads[['TotalVisits','Total Time Spent on Website','Page Views Per Visit','Converted']]
sns.pairplot(xedu,diag_kind='kde',hue='Converted')
plt.show()
```



Case study analysis

Model Building

- We assess the model using StatsModels.
- After building each model. we remove the variable whose p-value is high (greater than 0.05) and VIF value is also high (greater than 5). We rebuild each model again and check the values to be dropped again.
- We have 11 columns in the final model.

Making predictions on the Training dataset

- Choosing an arbitrary cut-off probability point of **0.5** to find the predicted labels. We create a new column 'predicted' with 1 if Converted_Prob > 0.5 else 0.
- Making the confusion matrix which indicates:

Predicted	not_converted	converted
Actual		
not_converted	1929	383
converted	560	1589

Case study analysis

Model Evaluation

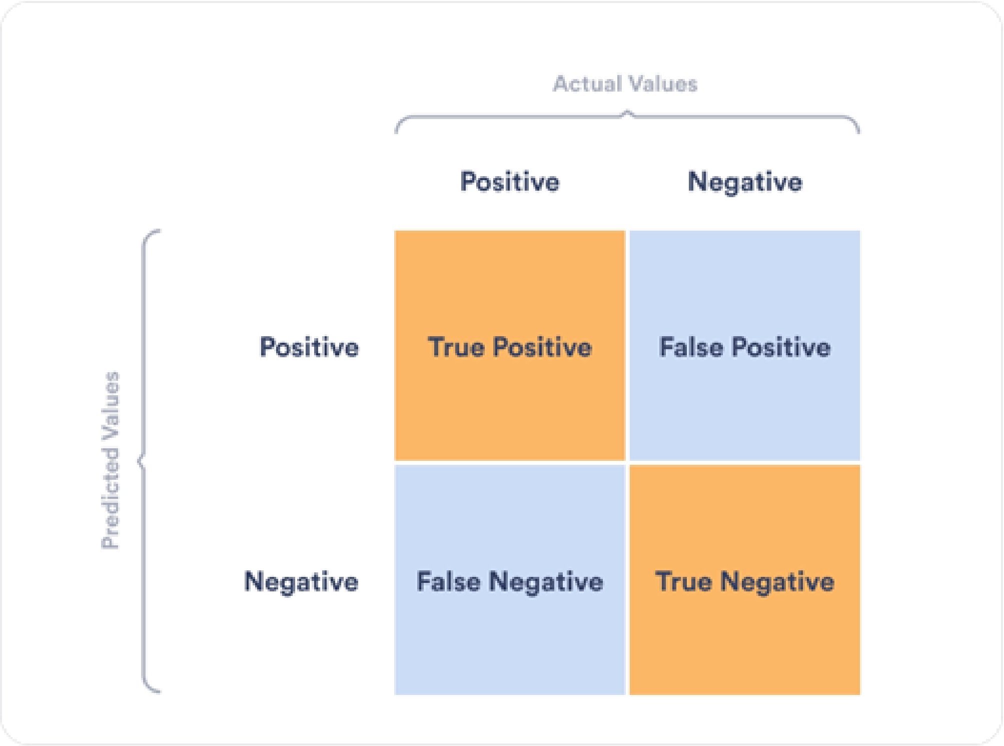
- Finding the Accuracy, Sensitivity and Specificity on train data set using the parameters of the confusion matrix.

  TP = confusion[1,1] -true positive
  TN = confusion[0,0] - true negatives
  FP = confusion[0,1] - false positives
  FN = confusion[1,0] - false negatives

- Accuracy – 78.86%
- Sensitivity – 73.94%
- Specificity – 83.43%

- Defining a threshold value: To optimise the arbitrary value of 0.5 which was chosen previously. For this purpose we use the ROC curve (receiver operating characteristic curve).

- From our model, Area under the ROC curve is 0.86.



		Actual Values	
		Positive	Negative
Predicted Values	Positive	True Positive	False Positive
	Negative	False Negative	True Negative

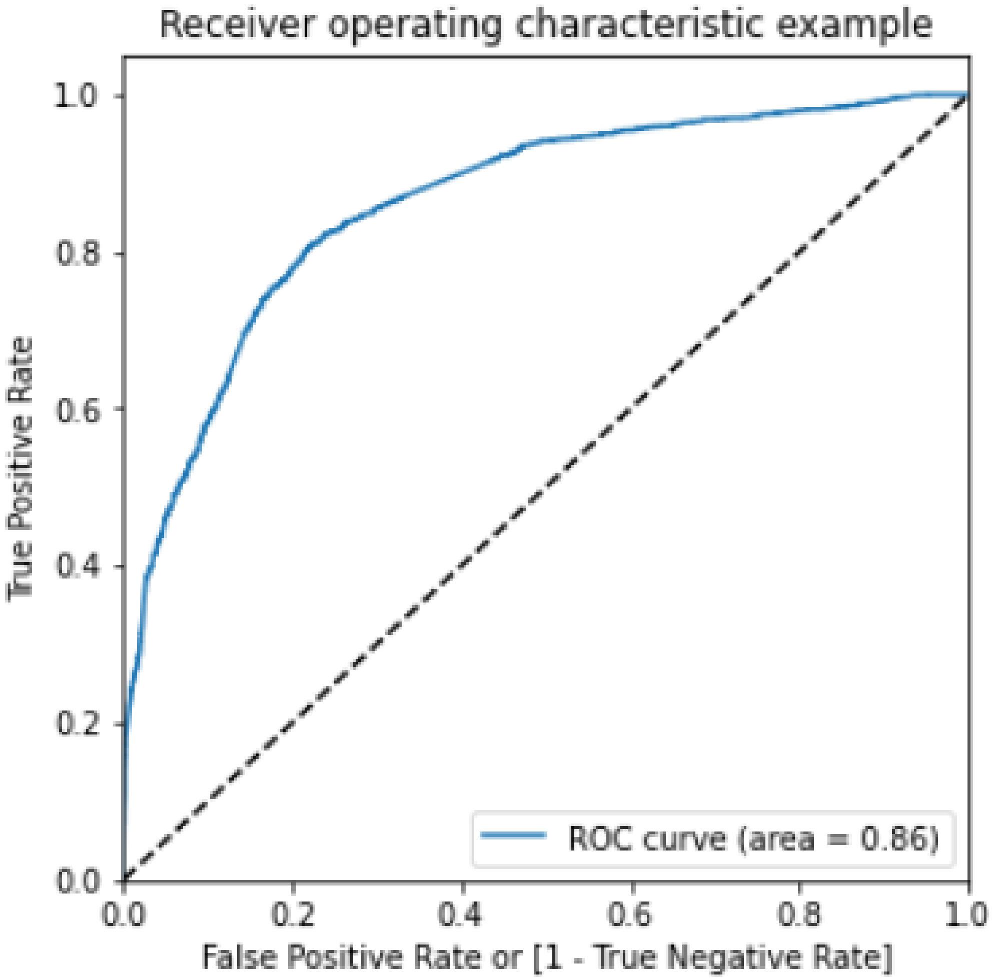
Confusion Matrix

Case study analysis

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



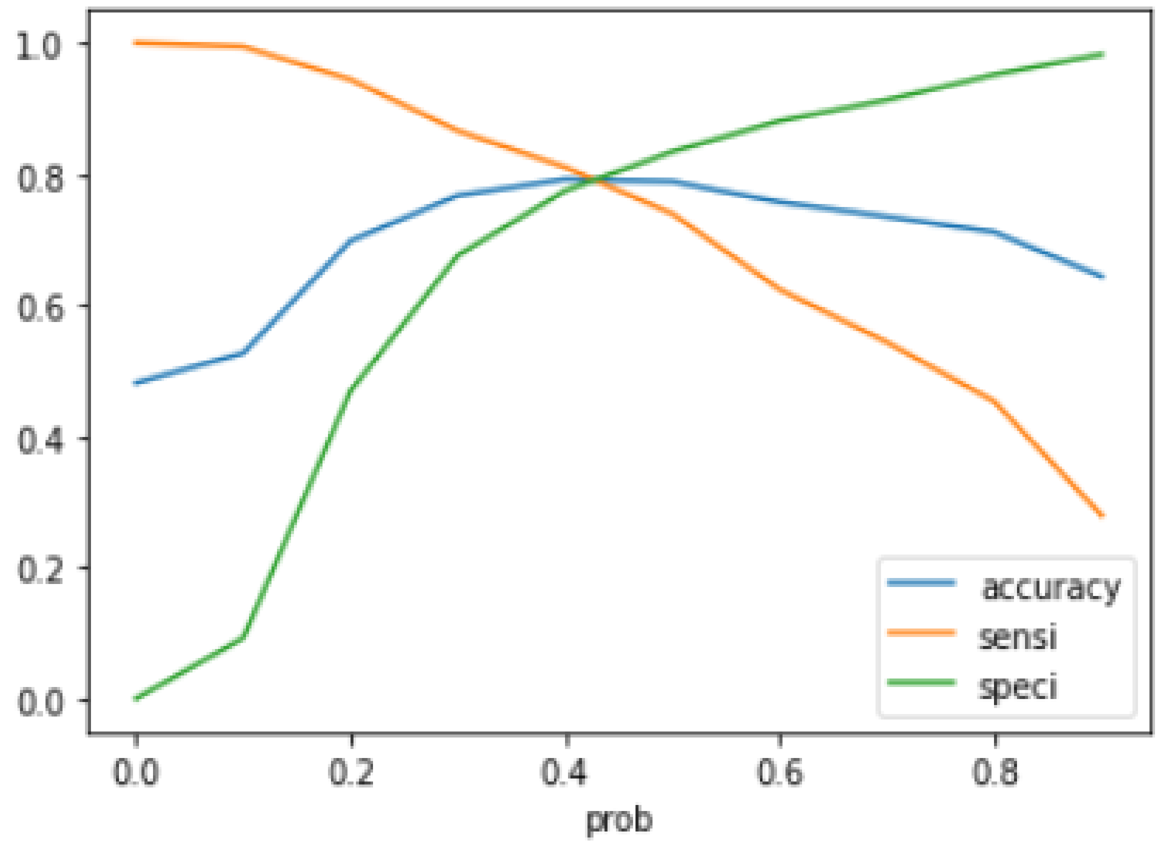
Case study analysis

Finding Optimal Cut off Point

- Optimal cut off probability is that probability where we get balanced sensitivity and specificity, this is the threshold value.

- From the graph it is visible that the optimal cut off is approximately at 0.42 and from calculations all the three values of accuracy, sensitivity and specificity seem to be quite close at this point.

- Based on this value, we make another confusion matrix and take the values of accuracy, sensitivity and specificity.



Case study analysis

Model Evaluation of training dataset

Accuracy – 79.08%
Sensitivity – 79.39%
Specificity – 78.84%

Model Evaluation of test dataset

- We repeat all the operations performed on training dataset on test dataset as well.

Accuracy, Sensitivity and Specificity on test data set

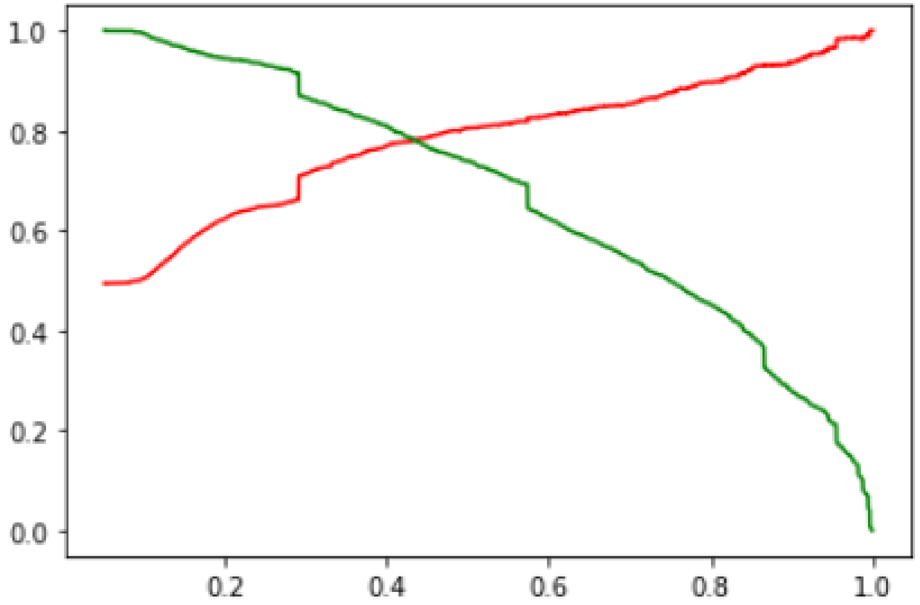
Accuracy – 78.45%
Sensitivity – 77.27%
Specificity – 77.94%

Case study analysis

Precision and Recall Tradeoff

For our Business Problem, The appropriate Metrics to be use will be Recall and Precision.

- Because this will help us to identify the predicted CONVERTED is actual CONVERTED and
- Probability that an actual CONVERTED case is predicted correctly. So we will use the Precision-Recall trade off curve to identify the most optimised threshold.
- The precision and Recall seem to have trade-off at .44, hence .44 will be used as threshold on test data



Precision was around 78% and recall around 77% on the test data frame.

Case study analysis

Based on the value of 0.44, we find the below parameters for the training dataset:

Accuracy – 78.95%
Sensitivity – 78.40%
Specificity – 77.71%

Variables impacting the lead conversion rate

- TotalVisits
- Total Time Spent on Website
- Lead Origin_Lead Add Form
- Last Notable Activity_Unreachable
- Last Activity_Had a Phone Conversation
- Lead Source_Welingak Website
- Lead Source_Olark Chat
- Last Activity_SMS Sent
- const
- Do Not Email_Yes
- What is your current occupation_Student
- What is your current occupation_Unemployed

Case study analysis

Recommendations

- The company should make calls to the leads coming from the lead sources "Welingak Websites" and “Olark chat" as these are more likely to get converted.
- The company should make calls to the leads who spent "more time on the websites" and leads who have greater number of total “number of visits” to the website as these are more likely to get converted.
- The company should make calls to the leads whose current occupation is “Working Professional” as they are not likely to get converted.
- The company should make calls to the leads whose last activity was “SMS Sent or a phone conversation” as they are more likely to get converted.
- The company should not make calls to the leads whose lead origin is “Add form" as they are likely to get converted.
- The company should not make calls to the leads whose current occupation is “Unemployed” as they are not likely to get converted.
- The company should not make calls to the leads who chose the option of "Do not Email" as "yes" as they are not likely to get converted.

Thank You