Lead Scoring Case Study
SUBMISSION REPORT

- *To build a Logistic Regression Model to predict whether a lead for online courses for an education company named X Education would be successfully converted or not.*

Group Members :

Anurag Thawait

Shaik Farook baba

Preeti Gupta

Business Objective

- *To help X Education to select the most promising leads(Hot Leads), i.e. the leads that are most likely to convert into paying customers.*
- *To build a logistic regression model to assign a lead score value between 0 and 100 to each of the leads which can be used by the companyto target potential leads.*

The objective is thus classified into the following sub-goals:

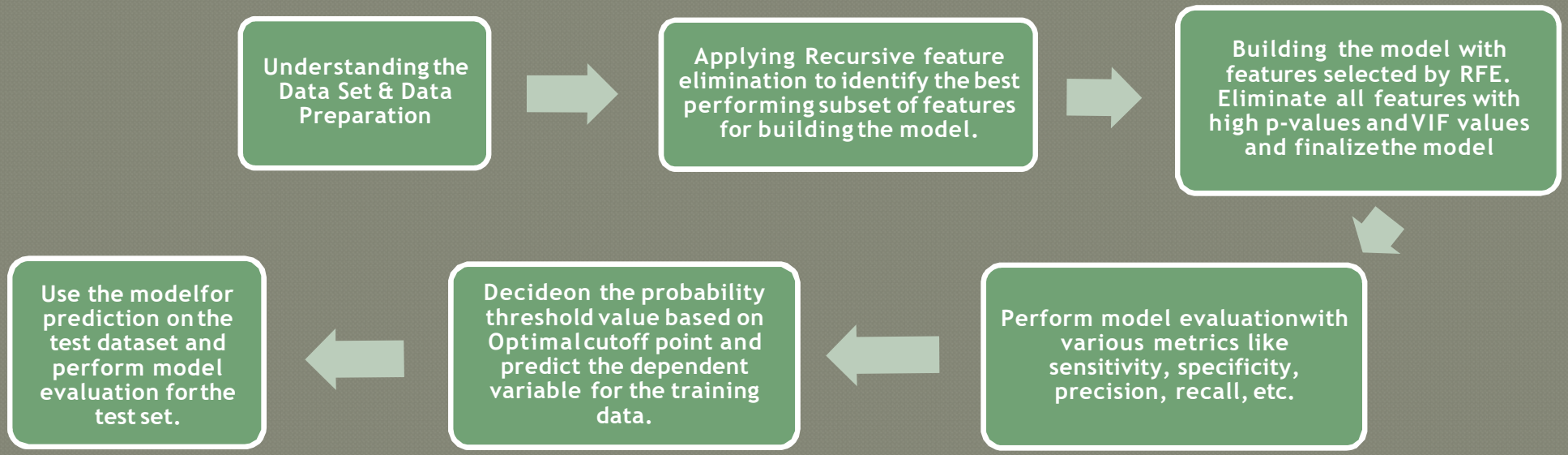
Create a Logistic Regression model to predict the Lead Conversion probabilities for each lead.

Decide on a probality threshold value above which a lead will be predicted as converted, whereas not converted if it is belowit.

Multiply the Lead Conversion probabilityto arrive at the Lead Score value for eachlead.

Problem Solving Methodology

- The approach for this project has been to divide the entire case study into various checkpoints to meet each of the sub-goals. The checkpoints are represented in a sequential flow as below:



Data Preparation & Feature Engineering

The following data preparation processes were applied to make the data dependable so that it can provide significant business value by improving Decision Making Process:

Remove columns which has only one unique value	• Deleting the following columns as they have only one unique value and hence cannot be responsible in predicting a successful lead case – **'Magazine', 'Receive More Updates About Our Courses' , 'Update me on Supply Chain Content' , 'Update me on Supply Chain Content'** and **'I agree to pay the amount through cheque'.**
Removing rows where a particular column has high missing values	• **'Lead Source'** is an important column for analysis. Hence all the rows that have null values for it were dropped.
Imputing NULL values with Median	• The columns **'TotalVisits'** and **'Page Views Per Visit'** are continuous variables with outliers. Hence the null values for these columns were imputed with the column median values.
Imputing NULL values with Mode	• The columns **'Country'** is a categorical variable with some null values. Also majority of the records belong to the Country 'India'. Thus imputed the null values for this with mode(most occuring value). Then binned rest of category into 'Outside India'.

Data Preparation & Feature Engineering contd…

Handling 'Select' values in some columns

- There are some columns in dataset which have a level/value called 'Select'. This might have happened because these fields in the website might be non mandatory fields with drop downs options for the customer to choose from. Amongst the dropdown values, the default option is probably 'Select' and since these aren't mandatory fields, many customer might have have chosen to leave it as the default value 'Select'.
- The Select values in columns were **converted to Nulls**.

Assigning a Unique Category to NULL/SELECT values

- All the nulls in the columns were binned into a separate column **'Unknown'**.
- Instead of deleting columns with huge null value percentage(which results in loss of data), this strategy adds more information into the dataset and results in the change of variance.
- The Unknown levels for each of these columns will be finally dropped during dummy encoding.

Outlier Treatment

- The outliers present in the columns '**TotalVisits**' & '**Page Views Per Visit**' were finally removed based on iterquatile range analysis.

Binary Encoding

- Converting the following binary variables (Yes/No) to 0/1:
- '**Search**','**Do Not Email**', '**Do Not Call**', '**Newspaper Article**', '**X Education Forums**', '**Newspaper**', '**Digital Advertisement**','**Through Recommendations**' and '**A free copy of Mastering The Interview**'

Data Preparation & Feature Engineering contd…

Dummy Encoding

- For the following categorical variables with multiple levels, dummy features (one-hot encoded) were created:
- **'Lead Quality','Asymmetrique Profile Index','Asymmetrique Activity Index','Tags','Lead Profile', 'Lead Origin','What is your current occupation', 'Specialization', 'City','Last Activity', 'Country'** and **'Lead Source','Last Notable Activity'**

Test-Train Split

- The original dataframe was split into **train and test** dataset. The train dataset was used to train the model and test dataset was used to evaluate the model.

Feature Scaling

- Scaling helps in interpretation. It is important to have all variables(specially categorical ones which has values 0 and 1) on the same scale for the model to be easily interpretable.
- **'Standardisation'** was used to scale the data for modelling. It basically brings all of the data into a standard normal distribution with mean at zero and standard deviation one.

```
Data columns (total 37 columns):
 #   Column                                         Non-Null Count  Dtype
---  ------                                         --------------  -----
 0   Prospect ID                                    9240 non-null   object
 1   Lead Number                                    9240 non-null   int64
 2   Lead Origin                                    9240 non-null   object
 3   Lead Source                                    9204 non-null   object
 4   Do Not Email                                   9240 non-null   object
 5   Do Not Call                                    9240 non-null   object
 6   Converted                                      9240 non-null   int64
 7   TotalVisits                                    9103 non-null   float64
 8   Total Time Spent on Website                    9240 non-null   int64
 9   Page Views Per Visit                           9103 non-null   float64
 10  Last Activity                                  9137 non-null   object
 11  Country                                        6779 non-null   object
 12  Specialization                                 7802 non-null   object
 13  How did you hear about X Education             7033 non-null   object
 14  What is your current occupation                6550 non-null   object
 15  What matters most to you in choosing a course  6531 non-null   object
 16  Search                                         9240 non-null   object
 17  Magazine                                       9240 non-null   object
 18  Newspaper Article                              9240 non-null   object
 19  X Education Forums                             9240 non-null   object
 20  Newspaper                                      9240 non-null   object
 21  Digital Advertisement                          9240 non-null   object
 22  Through Recommendations                        9240 non-null   object
 23  Receive More Updates About Our Courses         9240 non-null   object
 24  Tags                                           5887 non-null   object
 25  Lead Quality                                   4473 non-null   object
 26  Update me on Supply Chain Content              9240 non-null   object
 27  Get updates on DM Content                      9240 non-null   object
 28  Lead Profile                                   6531 non-null   object
 29  City                                           7820 non-null   object
 30  Asymmetrique Activity Index                    5022 non-null   object
 31  Asymmetrique Profile Index                     5022 non-null   object
 32  Asymmetrique Activity Score                    5022 non-null   float64
 33  Asymmetrique Profile Score                     5022 non-null   float64
 34  I agree to pay the amount through cheque       9240 non-null   object
 35  A free copy of Mastering The Interview         9240 non-null   object
 36  Last Notable Activity                          9240 non-null   object
```

Distribution of NULL values in the different columns of the original dataset before null-handling.

Feature Selection Using RFE

•**Recursive feature elimination** is an optimization technique for finding the best performing subset of features. It is based on the idea of repeatedly constructing a model and choosing either the best (based on coefficients), setting the feature aside and then repeating the process with the rest of the features. This process is applied until all the features in the dataset are exhausted. Features are then ranked according to when they were eliminated.

Model Building using Stats Model & RFE:

In [108]:
```
import statsmodels.api as sm
from sklearn.linear_model import LogisticRegression
logreg = LogisticRegression()
```

In [109]:
```
from sklearn.feature_selection import RFE
rfe = RFE(logreg, 20)             # running RFE with 20 variables as output
rfe = rfe.fit(X_train, y_train)
```

•**Running RFE with the output number of the variable equal to 20.**

Building the model

- Generalized Linear Models from StatsModels is used to build the Logistic Regression model.
- The model is built initially with the 20 variables selected by RFE.
- Unwanted features are dropped serially after checking p values (< 0.5) and VIF (< 5) and model is built multiple times.
- The final model with 16 features, passes both the significance test and the multi-collinearity test.

Features	VIF
Tags_Closed by Horizzon	1.26
Tags_Not doing further education	1.23
Tags_switched off	1.17
Tags_Interested in full time MBA	1.10
Lead Source_Welingak Website	1.08
Asymmetrique Activity Index_03.Low	1.07
Tags_Lost to EINS	1.06
Tags_opp hangup	1.02
What is your current occupation_Working Profes...	0.77
Lead Quality_Worst	0.67
Tags_Ringing	0.58
Tags_Interested in other courses	0.38
Tags_Already a student	0.36
Tags_Will revert after reading the email	0.09
What is your current occupation_Unemployed	0.01
Last Activity_SMS Sent	0.00

Features	p-Value
const	0.00
Lead Source_Welingak Website	0.00
Lead Quality_Worst	0.00
Asymmetrique Activity Index_03.Low	0.00
Tags_Already a student	0.00
Tags_Closed by Horizzon	0.00
Tags_Interested in full time MBA	0.00
Tags_Interested in other courses	0.00
Tags_Lost to EINS	0.00
Tags_Not doing further education	0.00
Tags_Ringing	0.00
Tags_Will revert after reading the email	0.00
Tags_opp hangup	0.00
Tags_switched off	0.00
What is your current occupation_Unemployed	0.00
What is your current occupation_Working Profes...	0.00
Last Activity_SMS Sent	0.00

Predicting the Conversion Probability and Predicted column

Creating a dataframe with the actual Converted flag and the predicted probabilities.

Showing top 5 records of the dataframe in the picture on the right.

	Converted	Conversion_Prob	LeadID
0	0	0.064688	8529
1	0	0.009566	7331
2	1	0.762190	7688
3	0	0.077626	92
4	0	0.077626	4908

Creating new column 'predicted' with 1 if Conversion_Prob > 0.5 else 0

Showing top 5 records of the dataframe in the picture on the left.

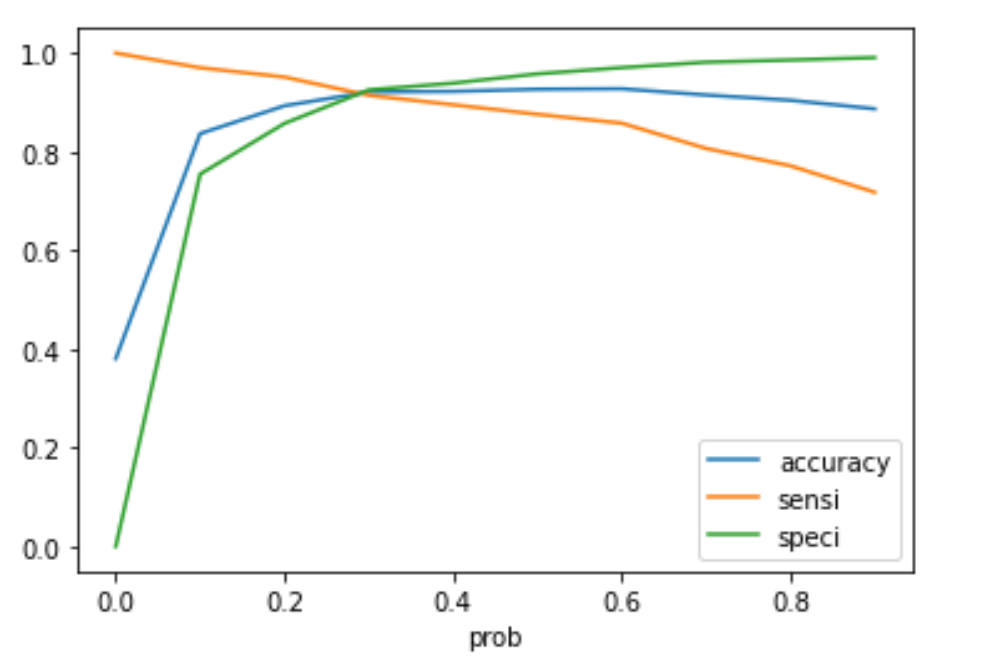
	Converted	Conversion_Prob	LeadID	predicted
0	0	0.064688	8529	0
1	0	0.009566	7331	0
2	1	0.762190	7688	1
3	0	0.077626	92	0
4	0	0.077626	4908	0

Finding Optimal Probability Threshold

Optimal cutoff probability is that prob where we get balanced sensitivity and specificity.

Optimal Probability Threshold

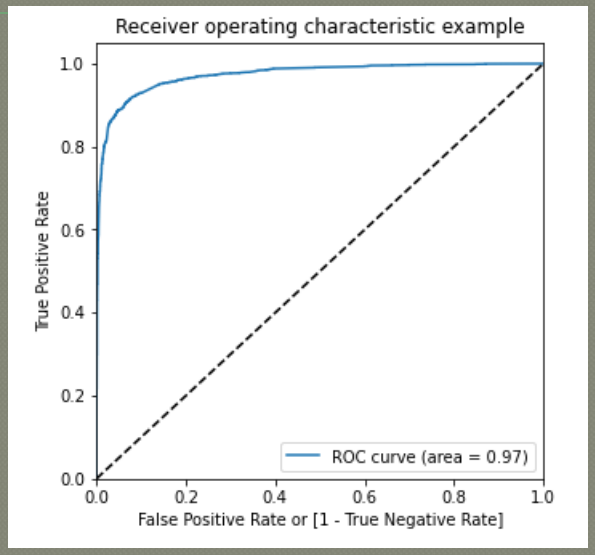
- The accuracy sensitivity and specificity was calculated for various values of probability threshold and plotted in the graph to the right.
- From the curve above, **0.3** is found to be the optimum point for cutoff probability.
- At this threshold value, all the 3 metrics - accuracy sensitivity and specificity was found to be well above 80% which is a well acceptable value.



Plotting the ROC Curve & Calculating AUC

Receiver Operating Characteristics (ROC) Curve

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



Area under the Curve (GINI)

- By determining the Area under the curve (AUC) of the ROC curve, the goodness of the model is determined. Since the ROC curve is more towards the upper-left corner of the graph, it means that the model is very good. The larger the AUC, the better will is the model.
- The value of AUC for our model is 0.97.

As a rule of thumb, an AUC can be classed as follows,

- 0.90 - 1.00 = excellent
- 0.80 - 0.90 = good
- 0.70 - 0.80 = fair
- 0.60 - 0.70 = poor
- 0.50 - 0.60 = fail

Since we got a value of 0.9678, our model seems to be doing well on the test dataset.

Evaluating the model on train datatset

Confusion Matrix

# Predicted / # Actual	Not Converted	Converted
Not Converted	3592	290
Converted	204	2181

Probability Threshold = 0.3

Accuracy	Sensitivity	Specificity	False Positive Rate	Positive Predictive Value
TP +TN/ (TP+TN+FN+FP)	TP / (TP+FN)	TN / (TN+FP)	FP/ (TN+FP)	TP / (TP+FP)
0.92	0.91	0.92	0.07	0.88

Negative Predictive Value	Precision	Recall	*F1 score = 2×(Precision*Recall)/(Precision+Recall)*	Area under the cuve
TN / (TN+ FN)	TP / TP + FP	TP / TP + FN		
0.94	0.88	0.91	0.88	0.96

Making predictions on the test set

- The final model on the train dataset is used to make predictions for thetest dataset
- The train data set was scaled using the scaler.transform function that was used to scale the train dataset.
- The Predicted probabilities were added to the leads in the test dataframe.
- Using the probability threshold value of 0.3, the leads from the test dataset were predicted if they will convert or not.

- The Conversion Matrix was calculated based on the Actual and Predicted 'Converted' columns.

Evaluating the model on test datatset

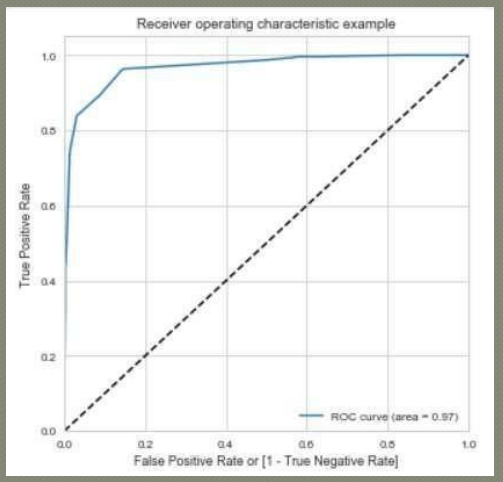
The following evaluation metrices were recorded for the test dataset.

Accuracy TP +TN/ (TP+TN+FN+FP)	Sensitivity TP / (TP+FN)	Specificity TN / (TN+FP)	Area under the cuve
0.92	0.91	0.93	0.96

Negative PredictiveValue TN / (TN+ FN)	Precision TP / TP + FP	Recall TP / TP + FN	*F1 score = 2×(Precision*Recall)/(Precision+Recall)*
0.93	0.89	0.90	0.88

False Positive Rate FP/ (TN+FP)	Positive PredictiveValue TP / (TP+FP)	Cross Validation Score
0.08	0.87	0.92

Area under the Curve



Lead Score Calculation

Lead Score is calculated for all the leads in the original dataframe.

Formula for Lead Score calculation is:

Lead Score = 100 * Conversion Probability

	Prospect ID	Converted	Converted_prob	Lead_Score
0	7681	0	0.022110	2
1	984	0	0.018112	2
2	8135	0	0.697404	70
3	6915	0	0.003927	0
4	2712	1	0.938270	94

The figure showing Lead Score for few records from the data set.

- The train and test dataset is concatenated to get the entire list of leads available.
- The Conversion Probability is multiplied by 100 to obtain the Lead Score for each lead.
- Higher the lead score, higher is the probability of a lead getting converted andvice versa,
- Since, we had used 0.33 as our final Probability threshold for deciding if a lead will convert or not, any lead with a lead score of 34 or above will have a value of '1' in the final_predicted column.

Conclusions

It was found that the variables that mattered the most in the potential buyers are :

- The total time spend on the Website.
- Total number of visits.
- When the lead source was:

a) Google
b) Direct traffic
c) Organic search
d) Welingak website

- When the last activity was:

a) SMS
b) Olark chat conversation

- When the lead origin is Lead add format.
- When their current occupation is as a working professional. Keeping these in mind the X Education can flourish as they have a very high chance to get almost all the potential buyers to change their mind and buy their courses.

UpGrad

THANK YOU