Leads Scoring Case Study Submission Report

To build a Logistic Regression Model to predict whether a lead online courses for an education company named X education would successfully converted or Not...

Group Members: Prajay Urkude, Atish Mistry & Achanta Pradeep Sandilya

Business objective

- To help X Education to select the most promising leads (Hot Leads), these are most likely to convert into paying customers

- To build a logistic regression model to assign a lead score value between 0 to 100 each of the values can be used by the company to target potential leads

The objective is classified into sub - goals observe below:

1. Create a logistic regression model to Predict the lead conversion probabilities for each lead.

2. Decide on a Probability threshold value above which a lead will be predicted as converted whereas not converted if it is below it..

3. Multiply the lead conversion probability to arrive at the lead score value of each lead.

Problem Solving Ideology

- This Project has divide the entire case study into various check points to meet each of the subgoals.

1. Understanding Data & Preparing the data

2. Applying Recursive feature elimination to identify the best performing subset of features for building the model.

3. Building the model with features selected by RFE. Eliminate all features with high p-values and VIF values and finalize the model

4. Perform model evaluation with various metrics like sensitivity, specificity, precision, recall, etc.

5. Decide on the probability threshold value based on Optimal cutoff point and predict the dependent variable for the training data.

6. Use the model for prediction on the test dataset and perform model evaluation for the test set.

Data Preparation and 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:

1. Remove columns which has only one unique value

2. Removing rows where a particular column has high missing values , here Lead Source’ is an important column for analysis. Hence all the rows that have null values for it were dropped.

3. Imputing Null values with Median, here , 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.

4. Imputing Null values with Mode, here , 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 occurring value). Then binned rest of category into 'Outside India’.

5. Handling 'Select' values in some columns, here there are some columns in dataset which have a level/value called ‘Select’, these select values columns can convert into Null Values by using NaN.

6. Assigning a Unique Category to NULL/SELECT values, all the nulls in the columns were binned into a separate column ‘Unknown’, these Unknown levels for each of these columns will be finally dropped during dummy encoding

7. Outlier Treatment , here the outlier presents in the “Total visits” & “Pages View Per Visit” these outlier columns were removed based on the IQR range

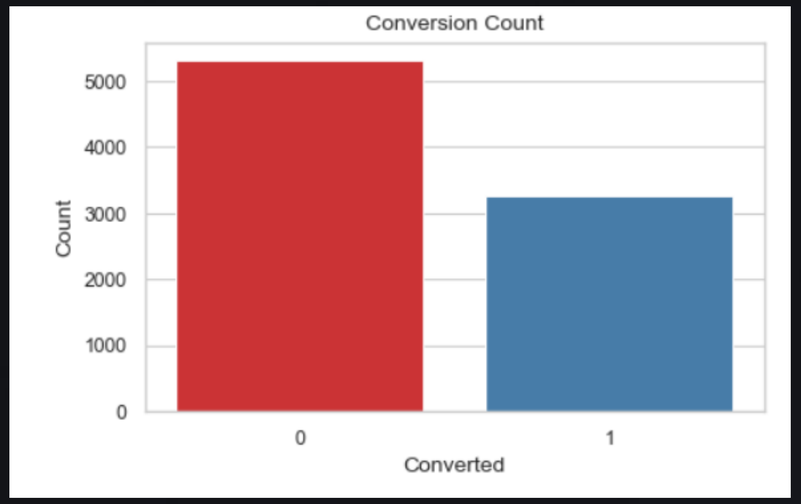
8. Binary Encoding, There are some variables which having Yes or No, True & False to 0 & 1

9. Dummy Encoding , For the following categorical variables with multiple levels, dummy features (one-hot encoded) were created, these are ’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

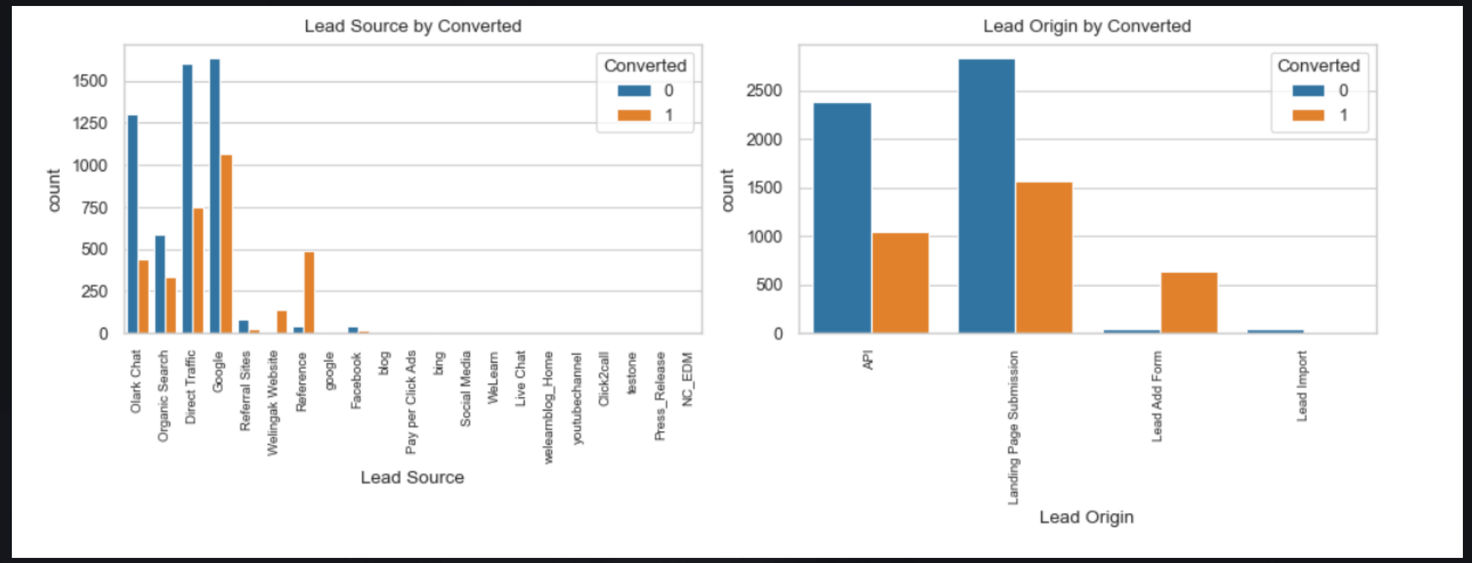
10. Train Test Split Model, The original dataframe was split into train and test dataset. The train dataset was used to train the model and test dataset was used to evaluate the model.

11. Feature Scaling, This 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. ‘Standardization’ was used to scale the data for modeling. It basically brings all of the data into a standard normal distribution with mean at zero and standard deviation one.

Exploratory Data Analysis

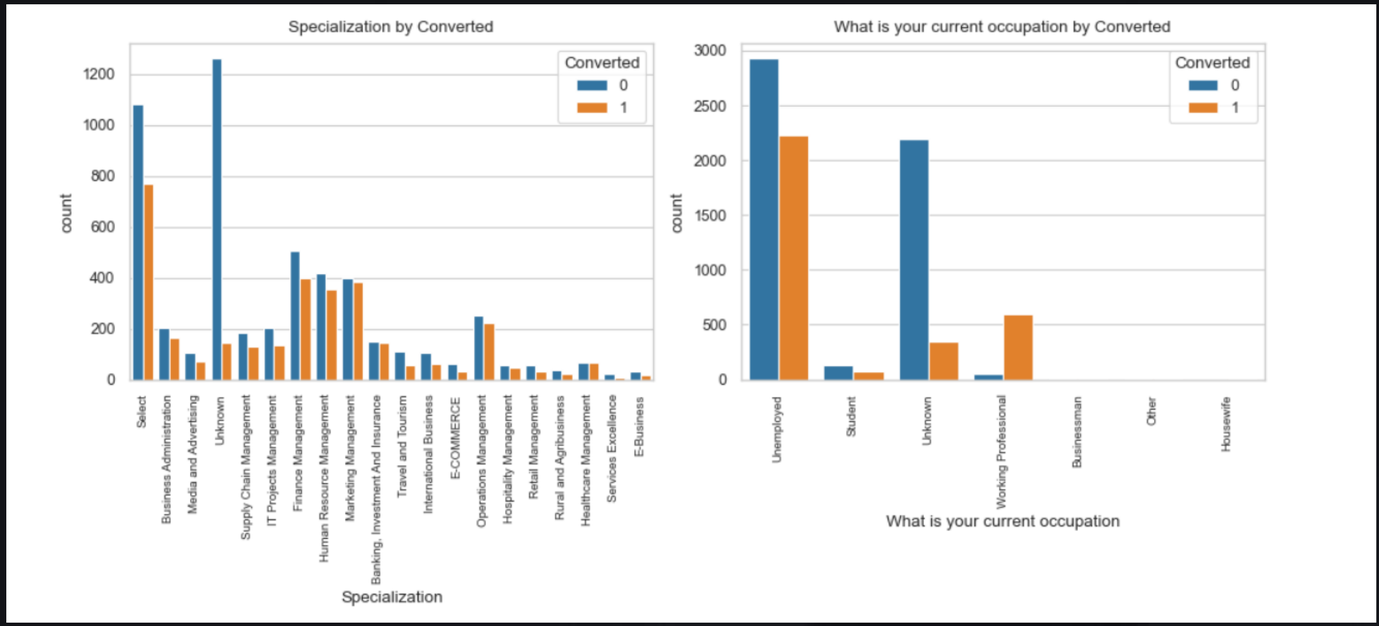


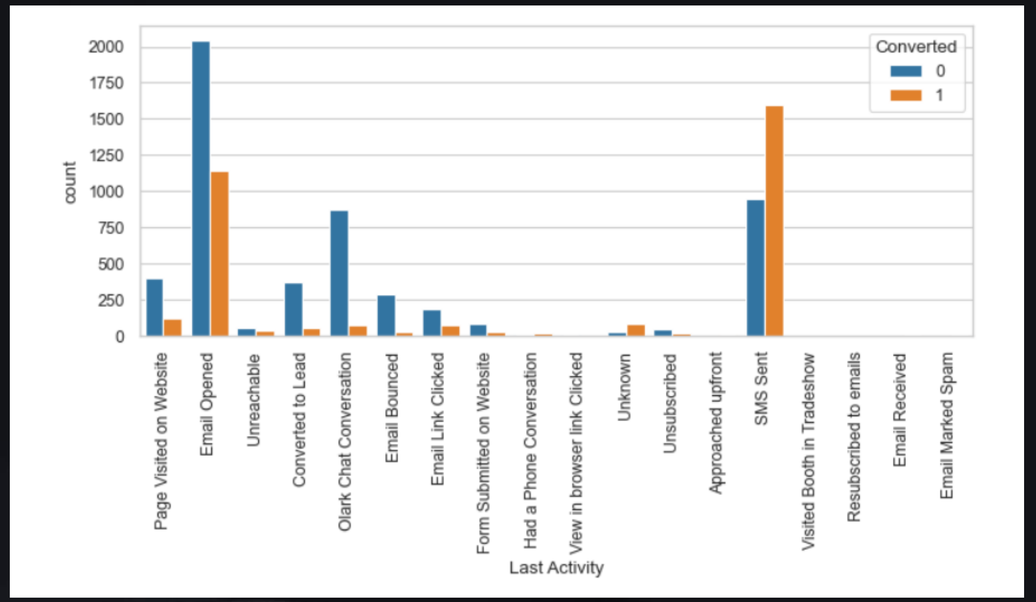
We have around **30%** of Conversion Rate



- The count of leads from the Google and Direct Traffic is maximum
- The conversion rate of the leads from Welingak Website is maximum than the number of leads
- API and Landing Page Submission has less conversion rate(~30%) but counts of the leads from them are considerable
- The count of leads from the Lead Add Form is pretty low but the conversion rate is very high

Exploratory Data Analysis





- Looking at above plot, we can say that working professionals have high conversion rate
- Number of Unemployed leads are more than any other occupation where ~50% leads are converted.
- Specialization "Finance Management", "HR Management", "Marketing Management" "Operation management" we are getting more leads and high conversion rate.

- The conversion rate for "SMS sent" is higher than the "Emailed open" even if it has less count than the "Email Opened“
- The count for last activity "Emailed Open" is high where conversion is low.

Feature Selection Using RFE

Recursive feature elimination (RFE) is a feature selection method that fits a model and removes the weakest feature (or features) until the specified number of features is reached.

Running RFE with 20 variables

```
col = X_train.columns[rfe.support_]
col
```

```
Index(['Lead Source_Welingak Website', 'Lead Quality_Worst',
       'Asymmetrique Activity Index_03.Low', 'Tags_Already a student',
       'Tags_Closed by Horizzon', 'Tags_Diploma holder (Not Eligible)',
       'Tags_Interested  in full time MBA', 'Tags_Interested in other courses',
       'Tags_Lost to EINS', 'Tags_Not doing further education', 'Tags_Ringing',
       'Tags_Will revert after reading the email', 'Tags_invalid number',
       'Tags_number not provided', 'Tags_opp hangup', 'Tags_switched off',
       'Tags_wrong number given', 'What is your current occupation_Unemployed',
       'What is your current occupation_Working Professional',
       'Last Activity_SMS Sent'],
      dtype='object')
```

These are the Train columns

```
from sklearn.linear_model import LogisticRegression
logreg = LogisticRegression()
```

```
from sklearn.feature_selection import RFE
rfe = RFE(estimator=logreg, n_features_to_select=20)             # running RFE with 20 variables as output
rfe = rfe.fit(X_train, y_train)
```

```
# Fit RFE to the training data
rfe.fit(X_train, y_train)
```

RFE

RFE(estimator=LogisticRegression(), n_features_to_select=20)

estimator: LogisticRegression

LogisticRegression()

LogisticRegression

Caption

Model Building

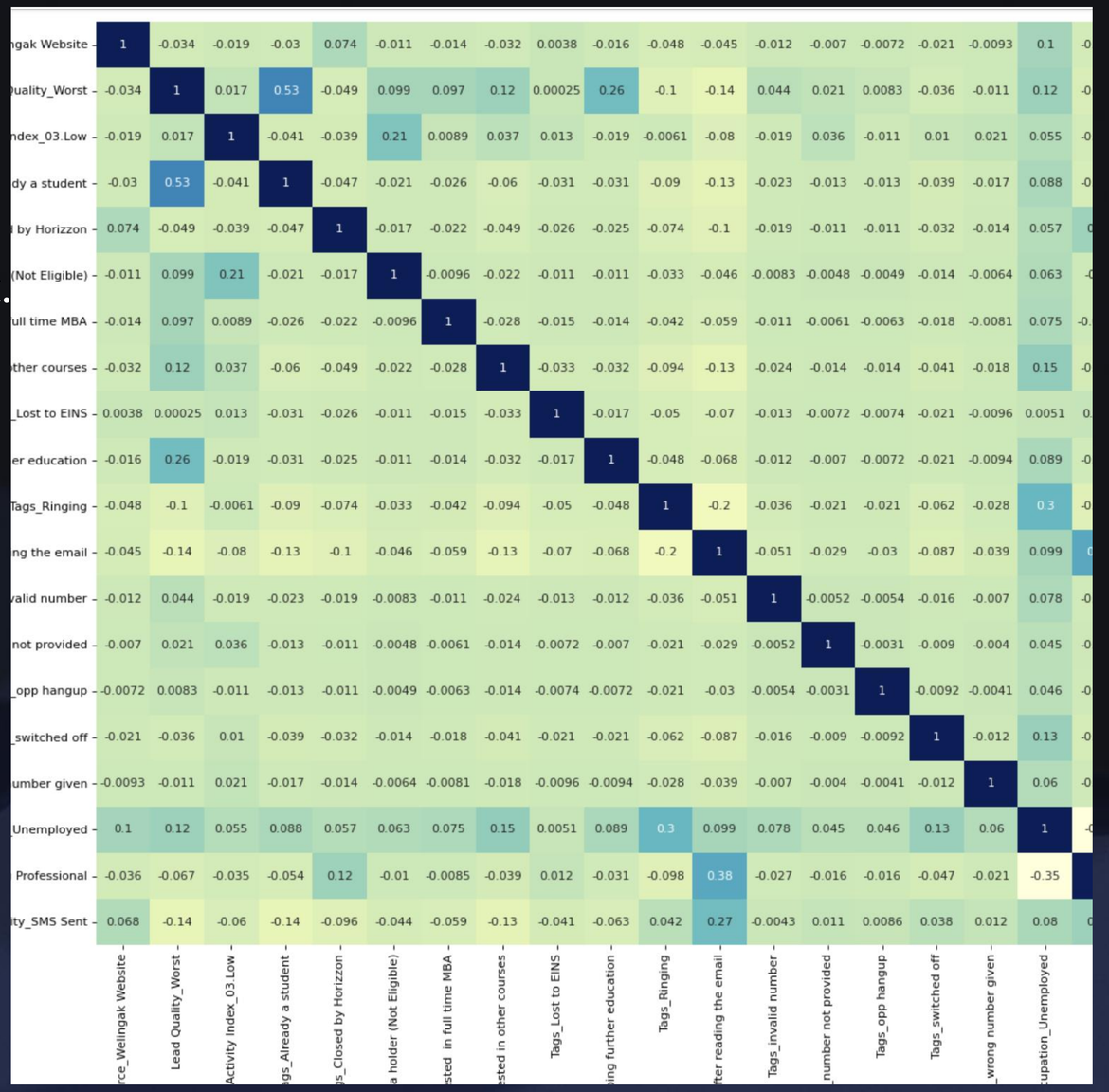
1. The model is built initially with the 20 variables selected by RFE.
Unwanted features are dropped serially,

2. after checking p values (< 0.5) and VIF (< 5) and
model is built multiple times, the final model with 16 features.

3.It passes both the significance test and the multi-collinearity test.

From given A heat map consisting of the final 16 features proves that

there is no significant correlation between the independent variables



Caption

Predicting conversion probability and Prediction columns

- Here we created a data frame with the actual Converted flag and the predicted probabilities.

99]:

	Converted	Conversion_Prob	LeadID
0	0	0.065692	8529
1	0	0.009069	7331
2	1	0.833555	7688
3	0	0.076360	92
4	0	0.076360	4908

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

00]:

	Converted	Conversion_Prob	LeadID	predicted
0	0	0.065692	8529	0
1	0	0.009069	7331	0
2	1	0.833555	7688	1
3	0	0.076360	92	0
4	0	0.076360	4908	0

Caption

Finding optimal probability Threshold

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

1. The accuracy sensitivity and specificity was calculated for various values of probability threshold and plotted in the graph to the right.

2. From the curve above, 0.33 is found to be the optimum point for cutoff probability.

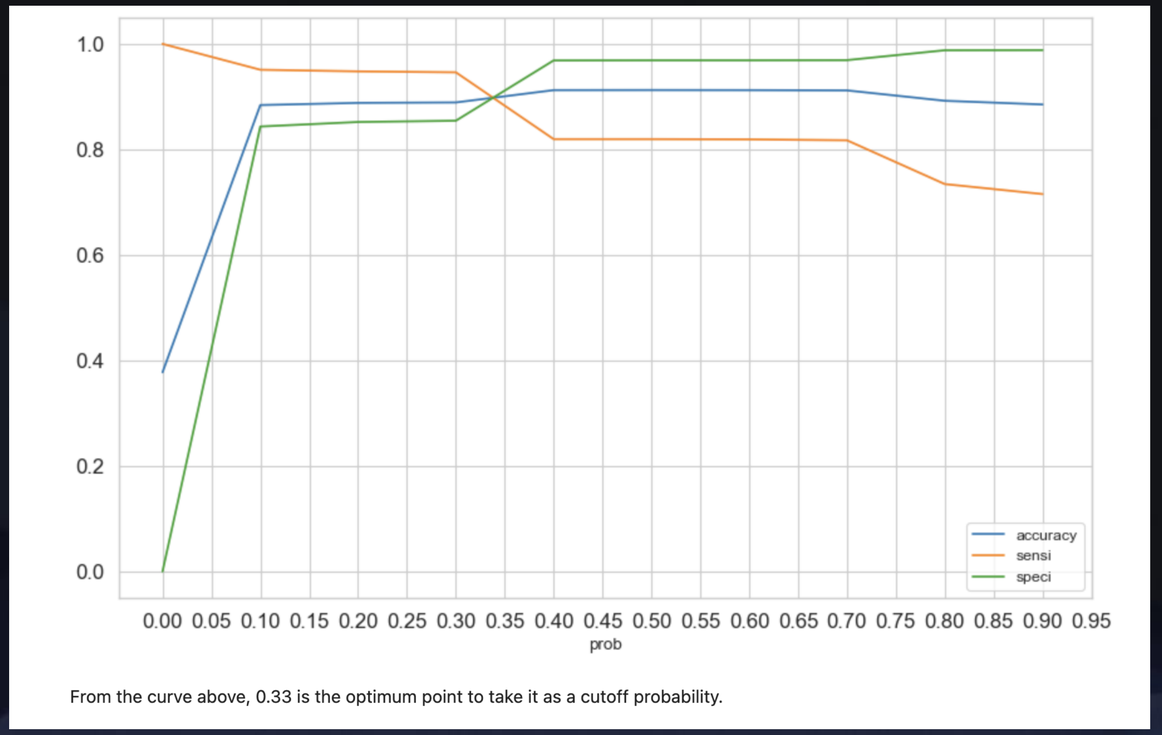
3. At this threshold value, all the 3 metrics - accuracy sensitivity and specificity was found to be well above 80% which is a well acceptable value.

We can see below graph figure in the next slide

Finding optimal probability Threshold

- From Above Curve Graph diagram

- 0.33 is the optimum point to take it as a cutoff probability



Caption

ROC Curve & Finding AUC

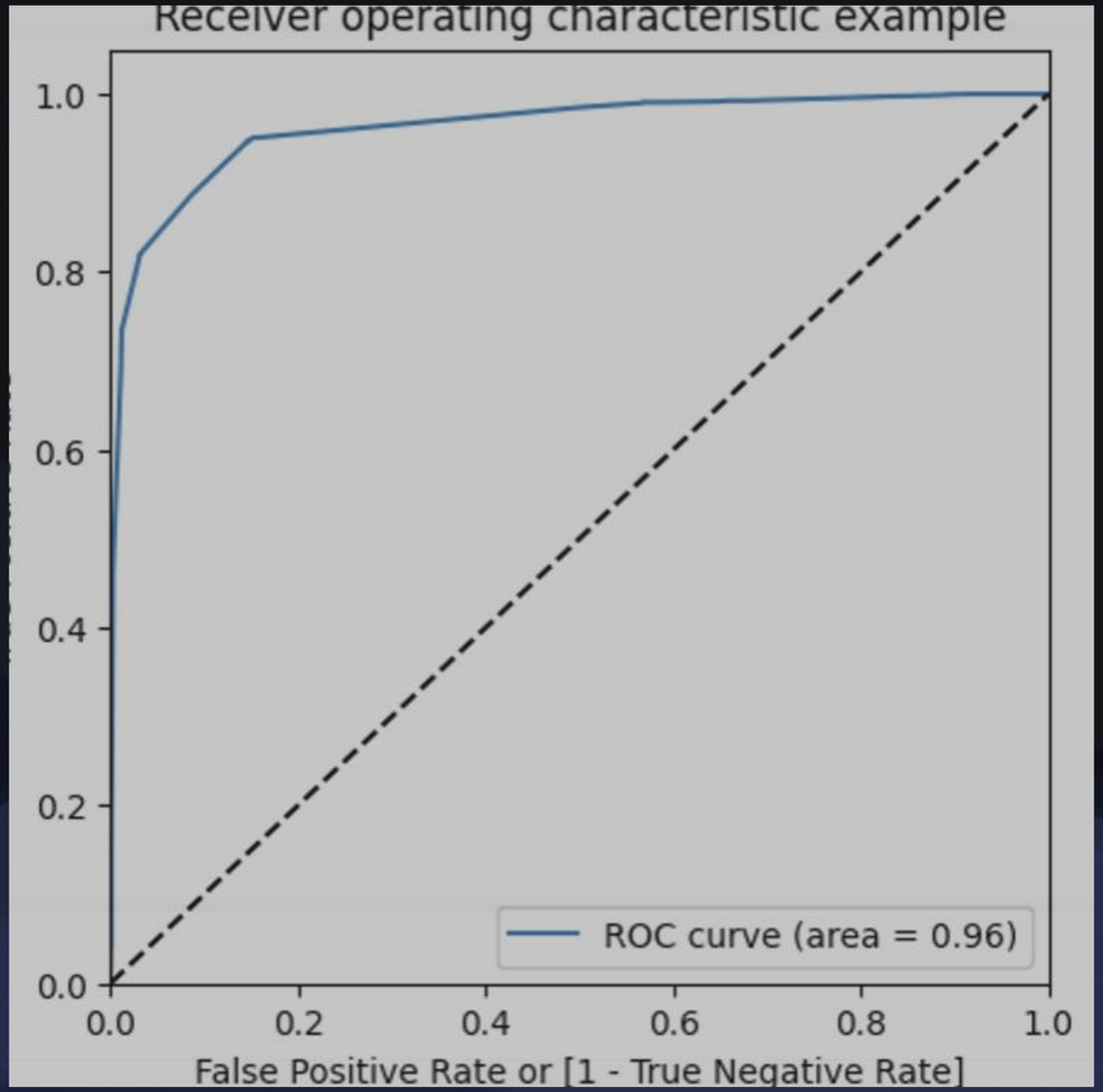
Receiver Operating Characteristics (ROC) Curve:

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 0.9678.



Caption

Model Evaluation on Train Data

Below values are build from Model evaluation of the lead score on Trained Data

Probability Threshold Value is 0.33

Accuracy of this model is 0.903

Sensitivity of this model is 0.887

Specificity = 0.913

False Positive Rate (FPR) = 0.087

Positive Predictive Value = 0.86

Negative Predictive value = 0.93

Precision value of this model = 0.861

Recall value of this Model = 0.887

F1 score of this model = 0.874

Area Under curve = 0.962

Making prediction the test data

- The final model on the train dataset is used to make predictions for the test 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 0.33, the leads from the test dataset were predicted if they will convert or not.

Model Evaluation on the Test Data

Below values are build from Model evaluation of the lead score on TestData

Probability Threshold Value is 0.33

Accuracy of this model is 0.906

Sensitivity of this model is 0.889

Specificity = 0.916

False Positive Rate (FPR) = 0.084

Positive Predictive Value = 0.87

Negative Predictive value = 0.928

Precision value of this model = 0.87

Recall value of this Model = 0.889

F1 score of this model = 0.879

Area Under curve = 0.968

Cross Validation Score = 0.913

	precision	recall	f1-score	support
0	0.93	0.92	0.92	1577
1	0.87	0.89	0.88	996
accuracy			0.91	2573
macro avg	0.90	0.90	0.90	2573
weighted avg	0.91	0.91	0.91	2573

Caption

Lead Score Calculation

Lead Score is calculated for all the leads in the original data frame.

Lead Score = 100 * Conversion probability

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 and vice 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.

	Lead Number	Converted	Conversion_Prob	final_predicted	Lead_Score
0	660737	0	0.031109	0	3
1	660728	0	0.009566	0	1
2	660727	1	0.801308	1	80
3	660719	0	0.009566	0	1
4	660681	1	0.955452	1	96
5	660680	0	0.077626	0	8
6	660673	1	0.955452	1	96
7	660664	0	0.077626	0	8
8	660624	0	0.077626	0	8
9	660616	0	0.077626	0	8

Lead Score Calculation

Feature Importance

16 features have been used by our model to successfully predict if a lead will get converted or not.

The Coefficient (beta) values for each of these features from the model parameters are used to determine the order of importance of these features.

Features with high positive beta values are the ones that contribute most towards the probability of a lead getting converted.

Similarly, features with high negative beta values contribute the least.

```
Lead Source_Welingak Website                              3.61
Lead Quality_Worst                                       -3.18
Asymmetrique Activity Index_03.Low                       -2.34
Tags_Already a student                                   -3.45
Tags_Closed by Horizzon                                   5.44
Tags_Interested  in full time MBA                        -2.66
Tags_Interested in other courses                         -2.63
Tags_Lost to EINS                                         6.71
Tags_Not doing further education                         -3.35
Tags_Ringing                                             -3.84
Tags_Will revert after reading the email                  3.87
Tags_opp hangup                                          -3.08
Tags_switched off                                        -4.73
What is your current occupation_Unemployed                1.67
What is your current occupation_Working Professional      1.89
Last Activity_SMS Sent                                    1.97
dtype: float64
```

```
Lead Source_Welingak Website                              12
Lead Quality_Worst                                         9
Asymmetrique Activity Index_03.Low                         3
Tags_Already a student                                     8
Tags_Closed by Horizzon                                    1
Tags_Interested  in full time MBA                         11
Tags_Interested in other courses                           5
Tags_Lost to EINS                                          6
Tags_Not doing further education                           2
Tags_Ringing                                              13
Tags_Will revert after reading the email                  14
Tags_opp hangup                                           15
Tags_switched off                                          0
What is your current occupation_Unemployed                10
What is your current occupation_Working Professional       4
Last Activity_SMS Sent                                     7
dtype: int64
```

Using Quick Sort Algorithm

Inference

- We finally choose a model with the following Lists
- All variables have p-value < 0.05.
- All the features have very low VIF values, meaning, there is hardly any multicollinearity among the features. This is also evident from the heat map.The overall accuracy of 0.9056 at a probability threshold of 0.33 on the test dataset is also very acceptable. The conversion probability of a lead increases with increase in values of the following features in descending order Features with Positive Coefficient Values

  Tags_Lost to EINS

  Tags_Closed by Horizzon

  Tags_Will revert after reading the email

  Lead Source_Welingak Website

  Last Activity_SMS Sent

  What is your current occupation_Working Professional

  What is your current occupation_Unemployed

Inference

The conversion probability of a lead increases with decrease in values of the following features in descending order:

Features with Negative Coefficient Values

Tags_switched off

Tags_Ringing

Tags_Already a student

Tags_Not doing further education

Lead Quality_Worst

Tags_opp hangup

Tags_Interested in other courses

Asymmetrique Activity Index_03.Low

Recommendations

- Another point to note here is that, depending on the business requirement, we can increase or decrease the probability threshold value with in turn will decrease or increase the Sensitivity and increase or decrease the Specificity of the model.

- High Sensitivity will ensure that almost all leads who are likely to Convert are correctly predicted where as high Specificity will ensure that leads that are on the brink of the probability of getting Converted or not are not selected.