Lead score Case study Report

Amrita

Amitlasure

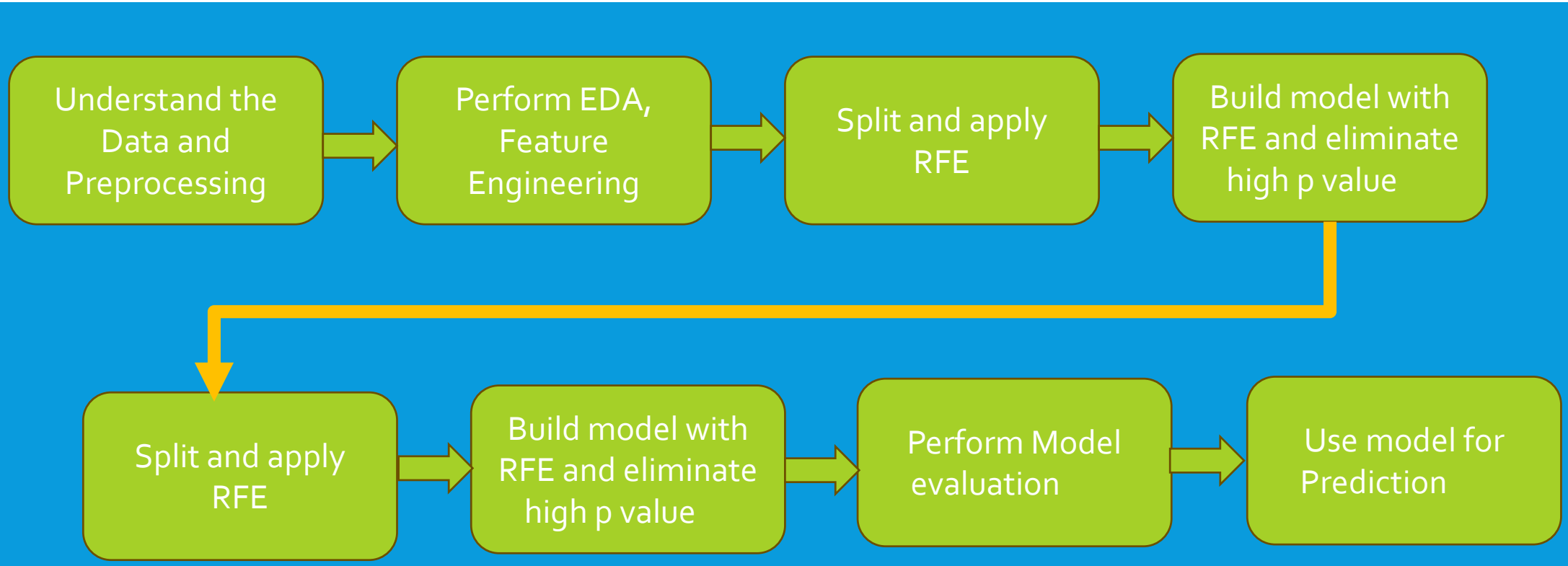
Amarnath

Business Objective

Support X Education to select the Potential leads(HotLeads), i.e. the leads that are most likely to convert into actual customers.

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 organization to target potential leads.

Methodology



Data Preparation and EDA

As part of the cleansing Categorical values like

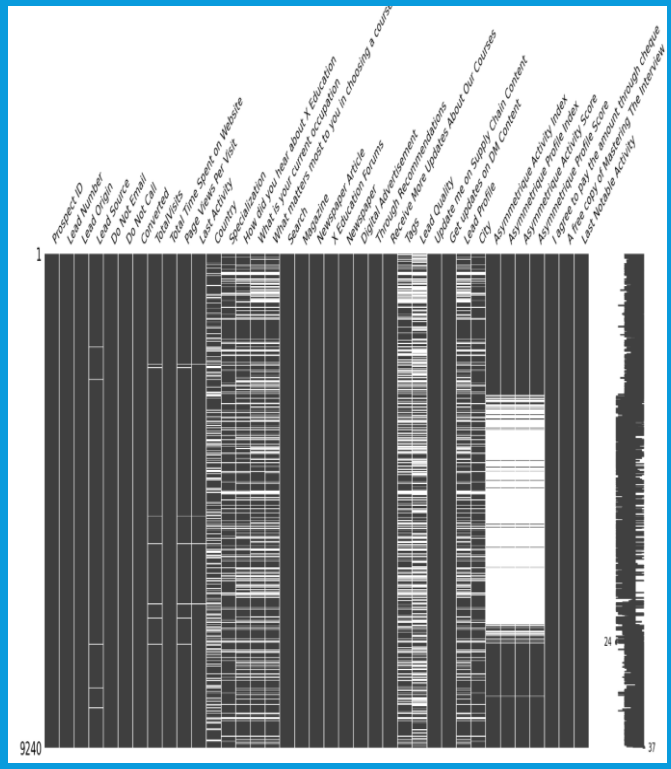
Country → Imputed with Unknown

Specialization, How did you hear about X Education, What is your current occupation, Lead Quality are imputed with "Other"

Tags, City and What matters most to you in choosing a course imputed with Mode

The Select value is converted to null

The columns Don't call, Don't Email are being converted to Binary values 0 and 1



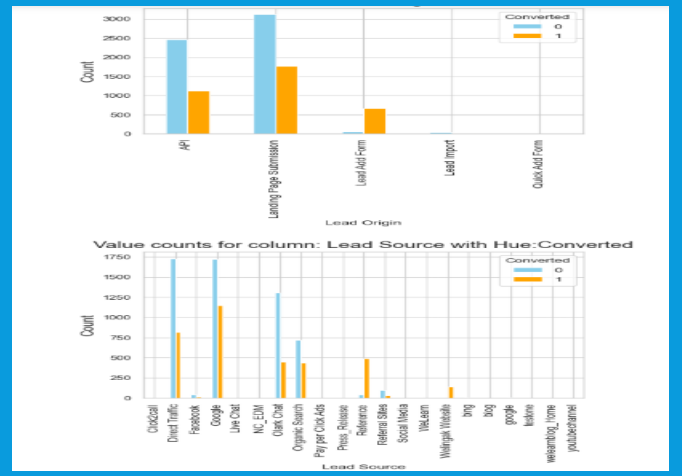
....cont

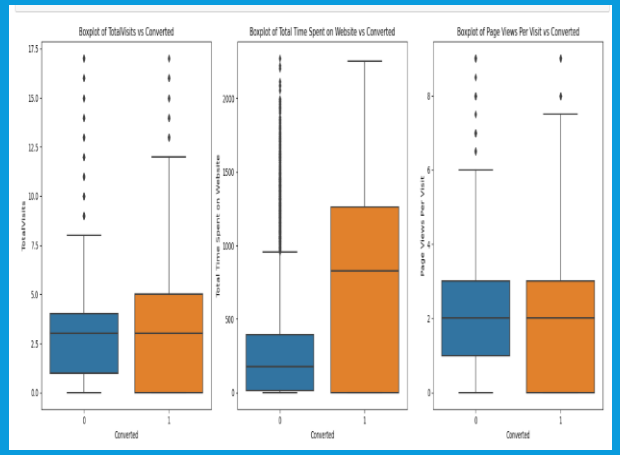
Data Preparation and EDA

Performed Outlier treatment

Dropped unnecessary columns

Introduced Dummies for improving the features: 'Lead Origin', 'Lead Source', 'Last Activity', 'Specialization', 'What is your current occupation', 'Tags','Lead Quality','City', 'How did you hear about X Education']





Split and apply RFE

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.

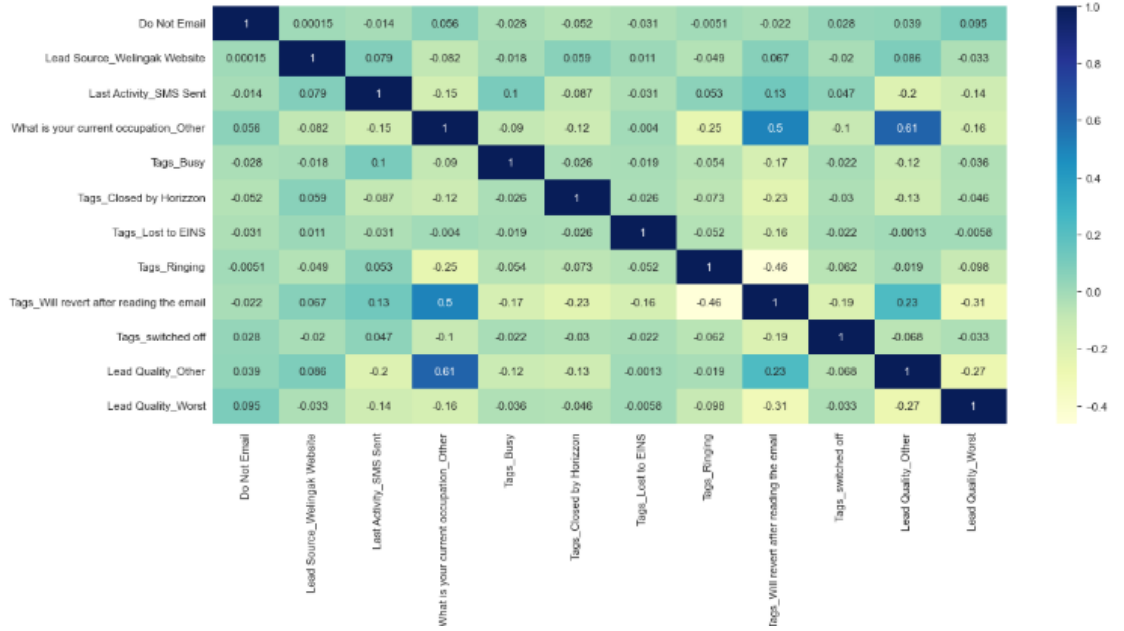
Scaling helps in interpretation. It is important to have all variables(specially categorical ones which has values 0 and 1) on the same scale for the model to be easily interpretable.

'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.

'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.

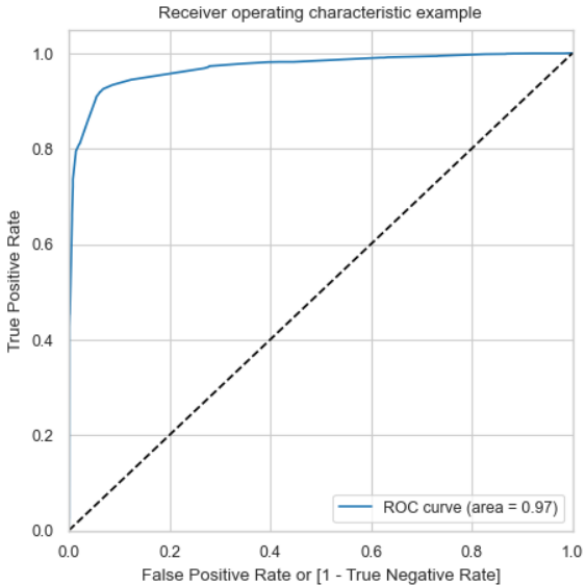
Model building

- •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.



Plotting ROC and Calculating AUC

- ROC (Receiver Operating Characteristic) and AUC (Area Under the Curve) are used to measure a model's performance, while the Gini coefficient is a metric derived from ROC_AUC. The Gini coefficient is a popular choice for non-technical audiences because it provides a single-figure snapshot of model effectiveness



Calculating the Area Under Curve (AUC) GINI

```
def auc_val(fpr,tpr):
    AreaUnderCurve = 0.
    for i in range(len(fpr)-1):
        AreaUnderCurve += (fpr[i+1]-fpr[i])
    AreaUnderCurve *= 0.5
    return AreaUnderCurve
```

```
auc = auc_val(fpr,tpr)
auc
```

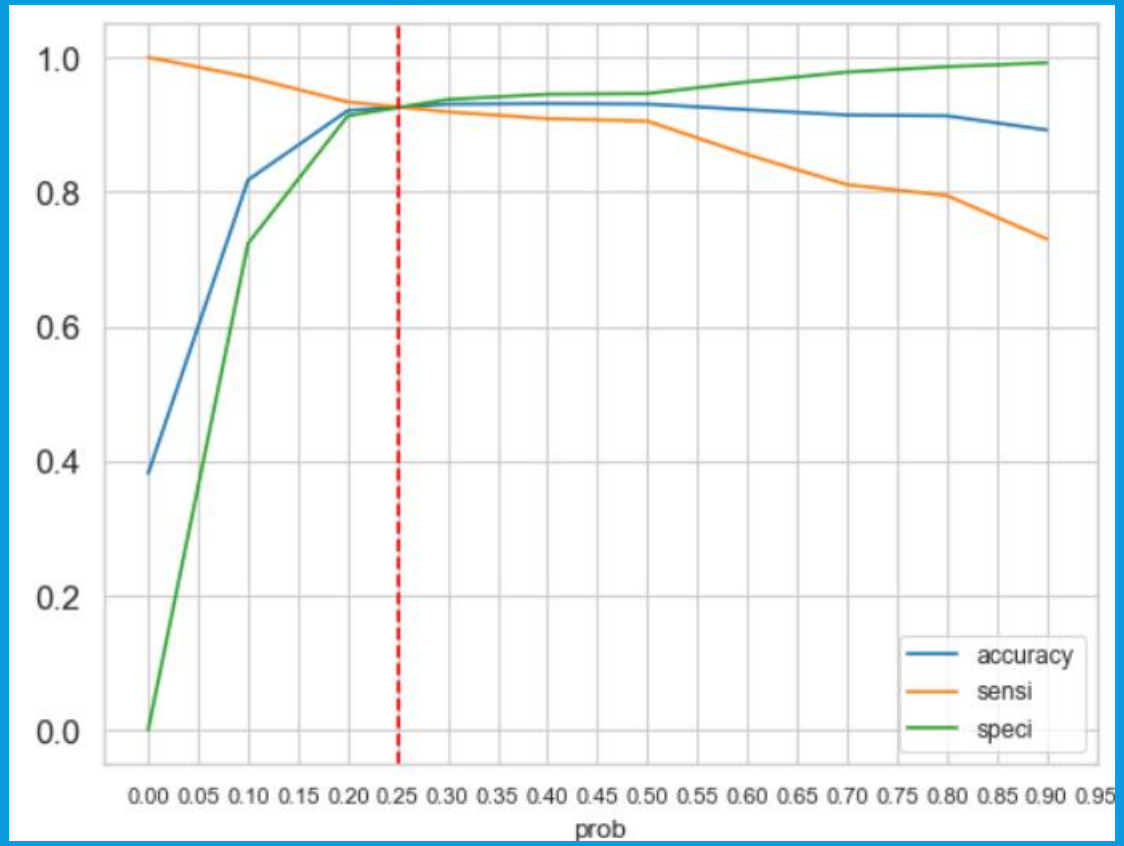
```
0.9697555419370606
```

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
```

Find Optimal prob threshold

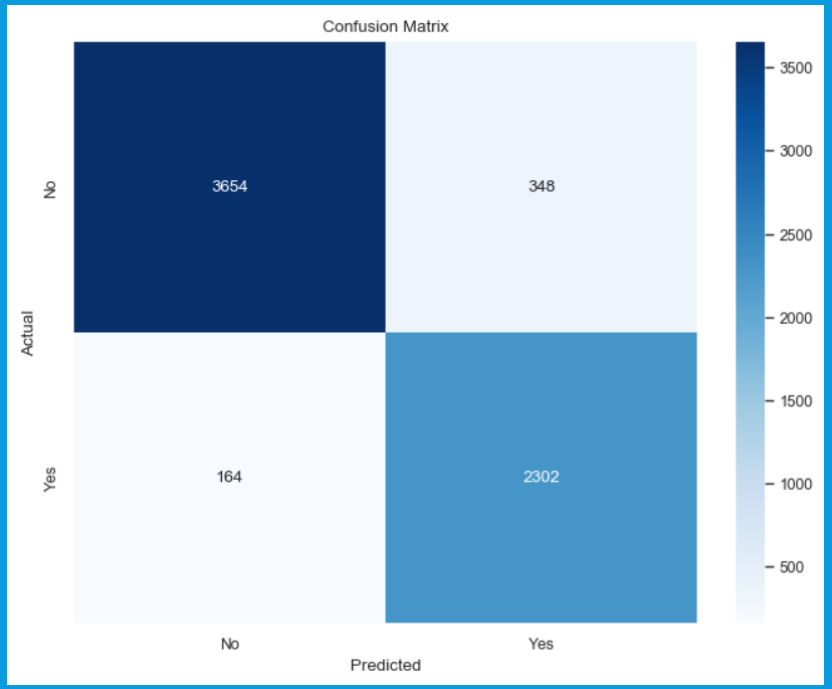
- The accuracy, Sensitivity and Specifity is calculated as shown in graph.
- From the curve 0.25 is derived to be the optimum point for cutoff probability.
- At this threshold value all the 3 metrics : Accuracy, Sensitivity and specificity were above 86%.



Evaluation of Model on Train Data

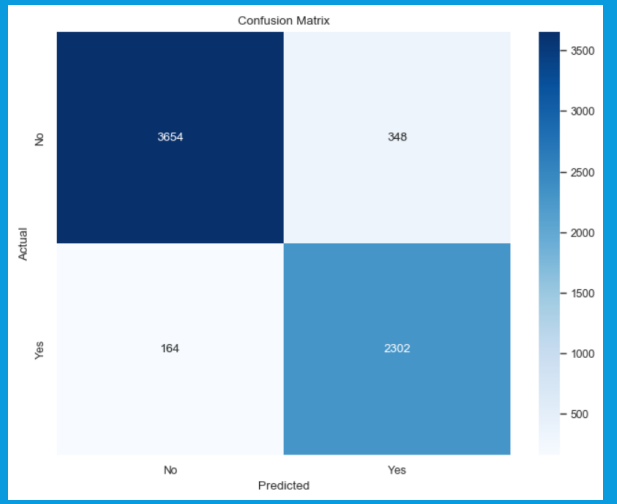
The below calculated the Metrics with the Probability Threshold 0.25

Sensitivity: 0.933495539349554
Specificity: 0.9130434782608695
Accuracy: 0.9208410636982065
FPR: 0.08695652173913043
TPR: 0.8686792452830189
TNR: 0.957045735987428
F1 score: 0.899



Makind Predictions

The conversion Matrix calculated based on Actual and Predicted values

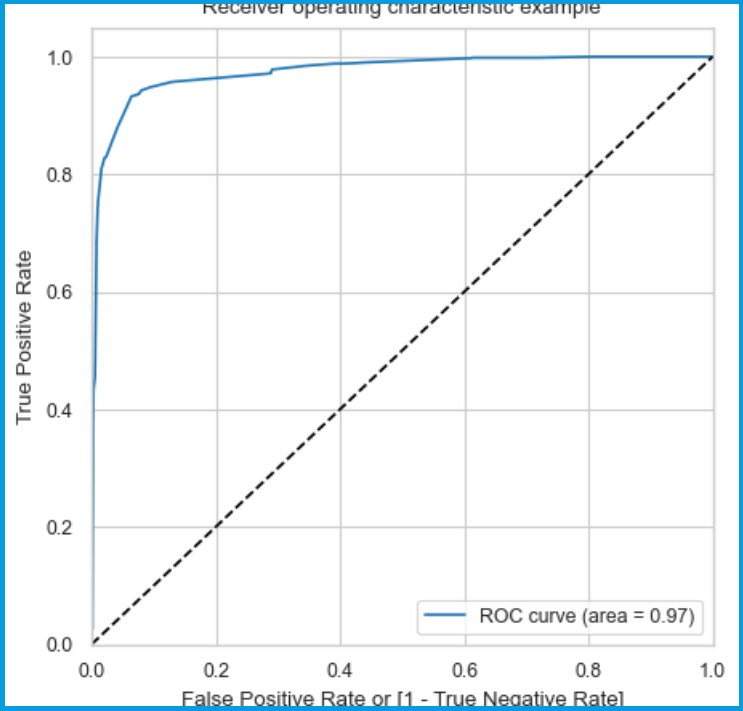


The Train dataset scaled using scaler.
The predicted probability was added with the probability threshold 0.25

	LeadID	Converted	Conversion_Prob	final_predicted
0	1871	0	0.01	0
1	6795	0	0.63	1
2	3516	0	0.01	0
3	8105	0	0.07	0
4	3934	0	0.26	1

Evaluation of model on test data

Area under the curve



	precision	recall	f1-score	support
0	0.96	0.91	0.93	1677
1	0.87	0.95	0.91	1095
accuracy			0.92	2772
macro avg	0.92	0.93	0.92	2772
weighted avg	0.93	0.92	0.92	2772

Lead Score calculation

	LeadID	Converted	Conversion_Prob	final_predicted	lead_score
157	4830	1	1.00	1	99
915	8412	1	1.00	1	99
1329	4812	1	1.00	1	99
2162	3736	1	1.00	1	99
939	3339	1	1.00	1	99