Leads Case Study

By Sahil Thakare, Devavratha Bhat

Business Objective

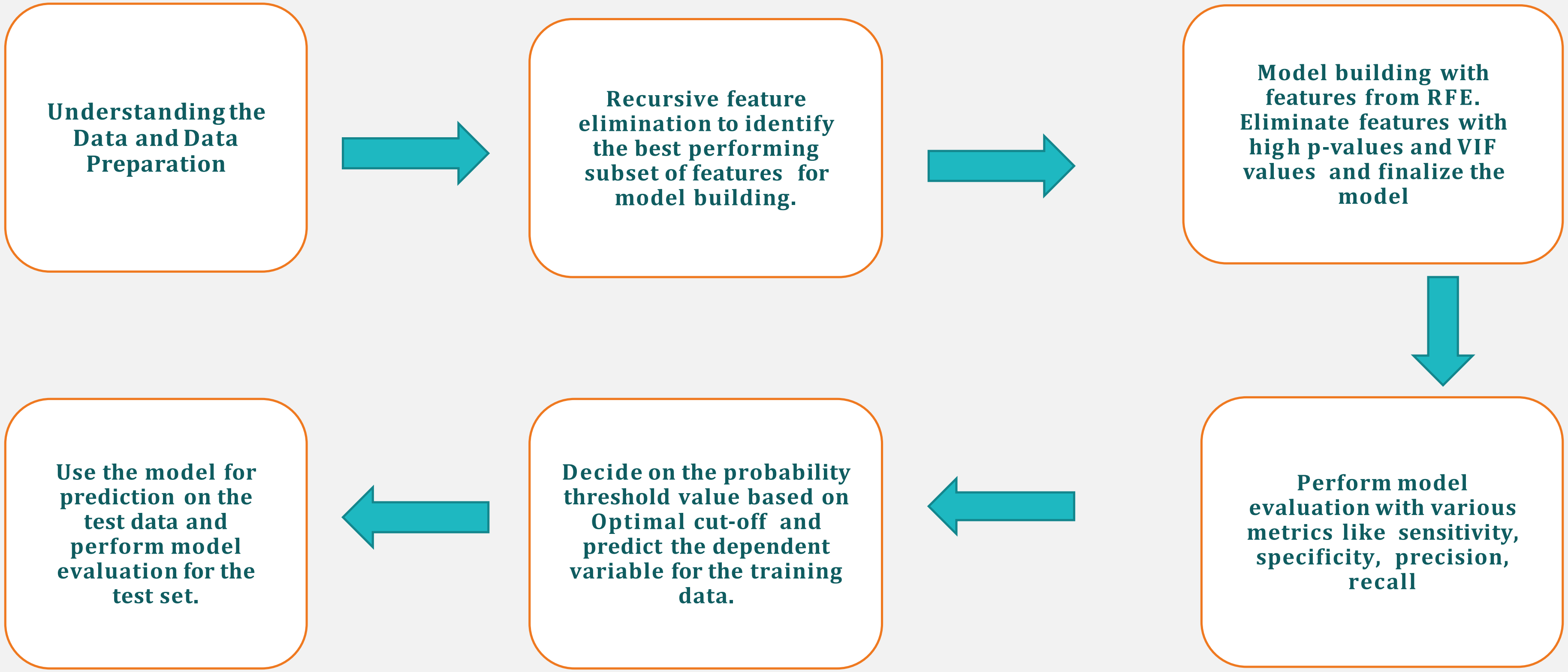
The objective of this case study is broadly broken down as below

Help X Education to select Hot Leads i.e. the leads that are most likely to convert into paying customers.

Build a logistic regression model to assign a lead score value to each of the leads which can be used target potential leads.

High Level Approach

Below is the high level approach towards the case study:



Understanding the Data and Data Preparation

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

Model building with features from RFE. Eliminate features with high p-values and VIF values and finalize the model

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

Decide on the probability threshold value based on Optimal cut-off and predict the dependent variable for the training data.

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

Data Preparation and Feature Engineering

The below data preparation process was applied for improving Decision Making:

Remove columns which has only one unique value	Few columns that were having only one unique values were dropped , few of them are as below **Magazine’, ‘Receive More Updates About Our Courses’ , ‘Update me on Supply Chain Content’ and ‘I agree to pay the amount through cheque’.**
Removing rows where a 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 Mode	The columns ‘**Country**’ is a categorical variable with null values. majority of the records belong to the ‘India’, so imputed the null values for this with mode

Data Preparation and Feature Engineering

The below data preparation process was applied for improving Decision Making:

Handling 'Select' values in some columns

There are some columns with value 'Select‘ which is the default option and many customer might 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 imputed using mode depending on the variables.
The Unknown levels for each of these columns were dropped during dummy encoding.

Outlier Treatment

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

Binary Encoding

Converting the binary variables (Yes/No) to 0/1:
‘**Do Not Email**', '**Do Not Call**‘ were the variables converted.

Data Preparation and Feature Engineering

The below data preparation process was applied for improving Decision Making:

One Hot Encoding

For the following categorical variables with multiple levels, dummy features (one-hot encoded) were created:

'Lead Quality', '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 data frame 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

‘**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.

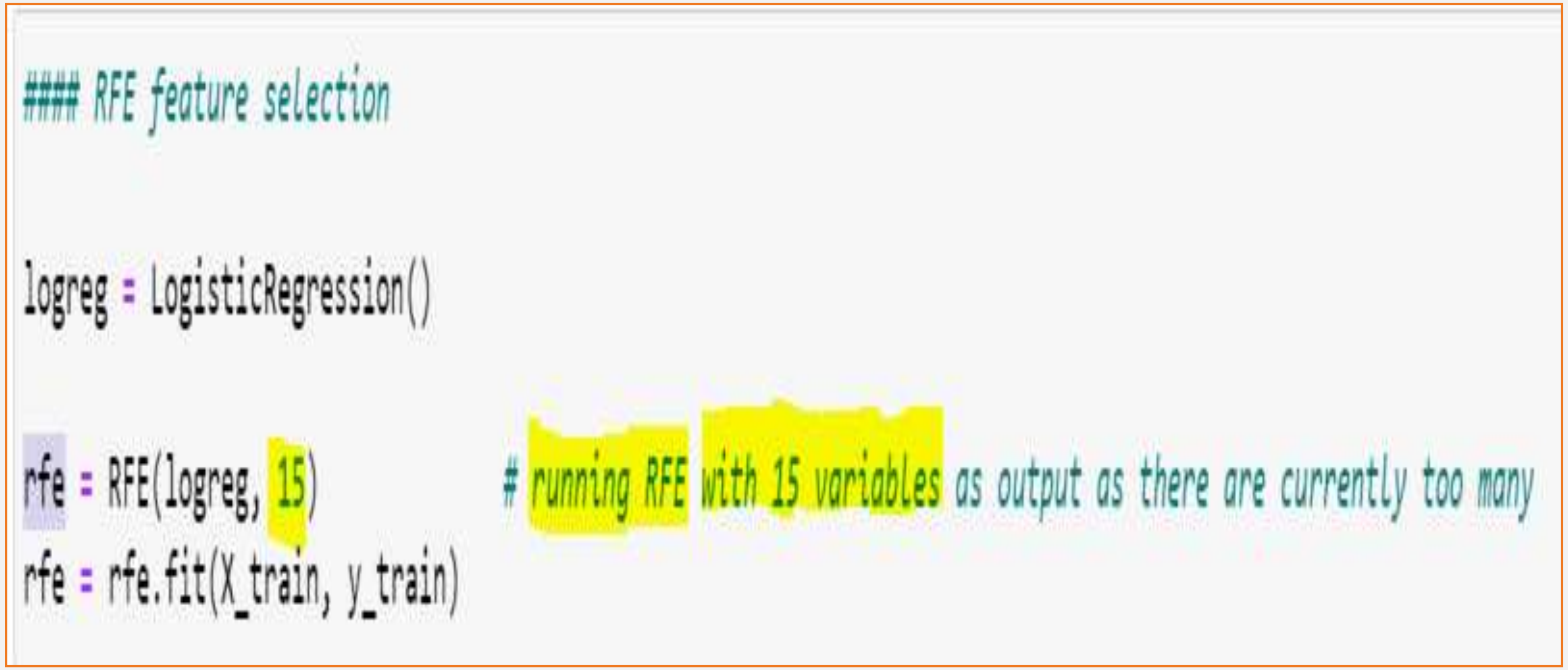
Feature Selection using RFE

Recursive feature elimination is an optimization technique for finding the best performing subset of features.

Repeated construction of 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.

In this case study we used the output number of variables to 15 ,i.e. top 15 best variables used for creating the model.

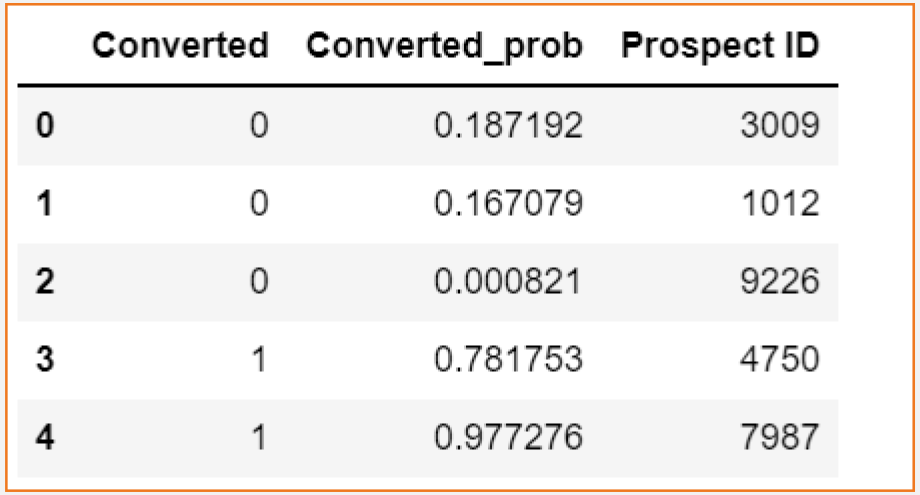


```
#### RFE feature selection

logreg = LogisticRegression()

rfe = RFE(logreg, 15)             # running RFE with 15 variables as output as there are currently too many
rfe = rfe.fit(X_train, y_train)
```

Predicting conversion probability and predicted column

Creating a data frame with the actual Converted flag and the predicted probabilities.

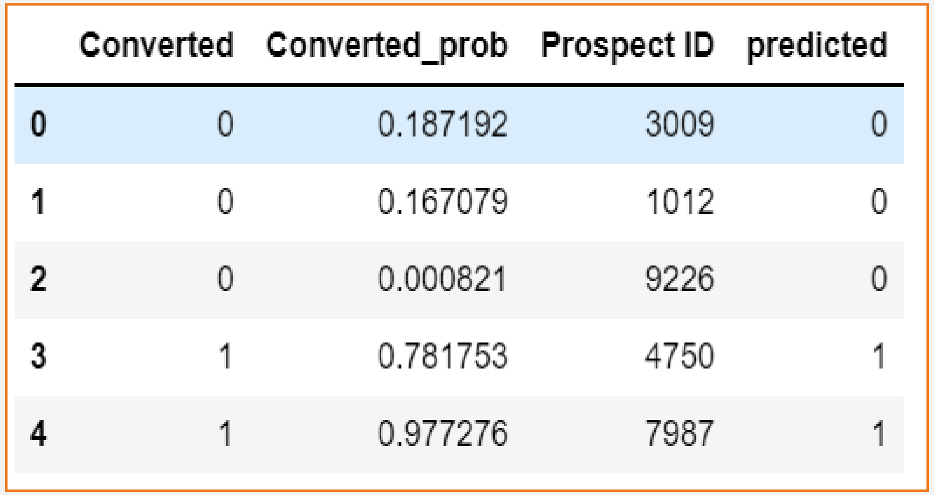
Showing head of the data on the right



	Converted	Converted_prob	Prospect ID
0	0	0.187192	3009
1	0	0.167079	1012
2	0	0.000821	9226
3	1	0.781753	4750
4	1	0.977276	7987

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

Showing head of the data frame on the right



	Converted	Converted_prob	Prospect ID	predicted
0	0	0.187192	3009	0
1	0	0.167079	1012	0
2	0	0.000821	9226	0
3	1	0.781753	4750	1
4	1	0.977276	7987	1

Finding Optimal probability threshold

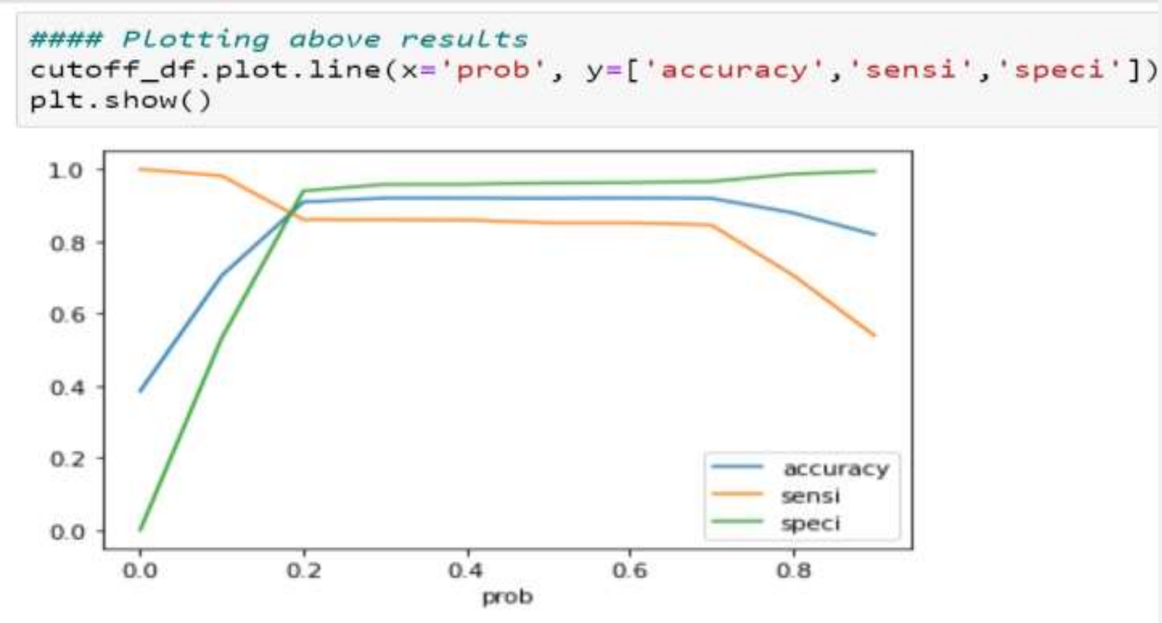
Optimal cut-off probability is that probability of balanced sensitivity and specificity.

Optimal Threshold Probability

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, **0.20 is** optimum point for cut-off probability.

At this threshold value, accuracy sensitivity and specificity was above 85% which is acceptable value.

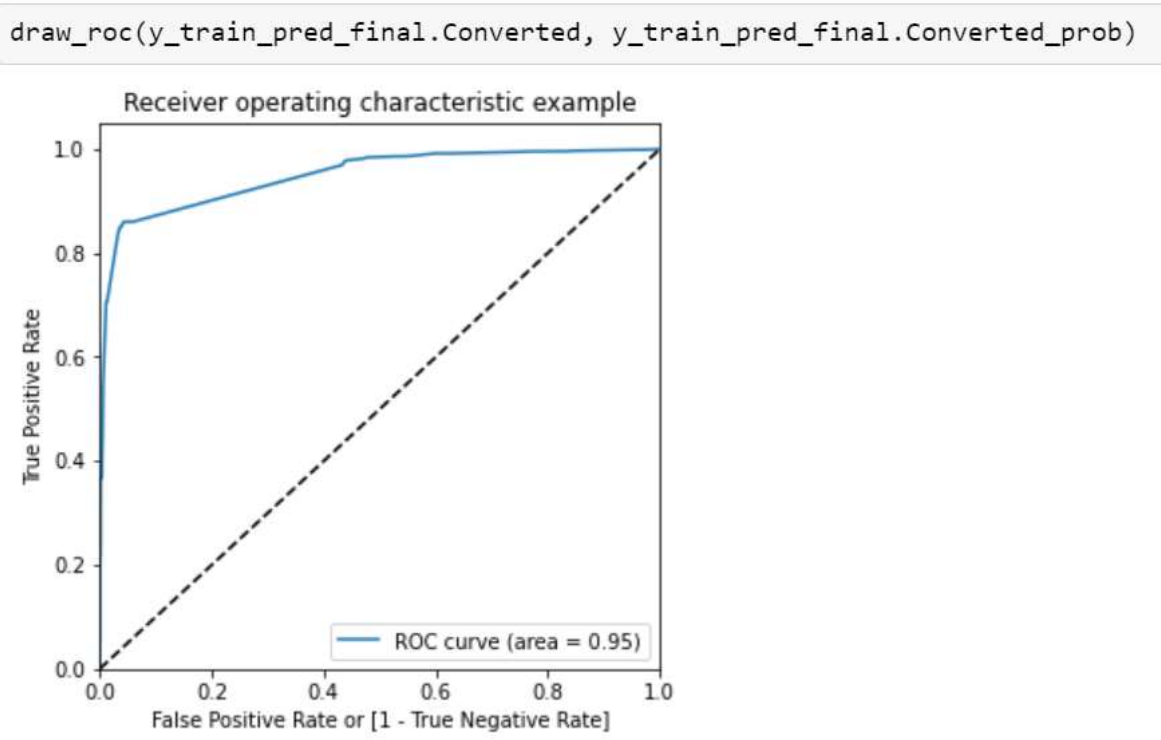


```
#### Plotting above results
cutoff_df.plot.line(x='prob', y=['accuracy','sensi','speci'])
plt.show()
```

Plotting ROC curve

Receiver Operating Characteristics (ROC) Curve

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



```
draw_roc(y_train_pred_final.Converted, y_train_pred_final.Converted_prob)
```

Receiver operating characteristic example
True Positive Rate
False Positive Rate or [1 - True Negative Rate]
ROC curve (area = 0.95)
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.2
0.4
0.6
0.8
1.0

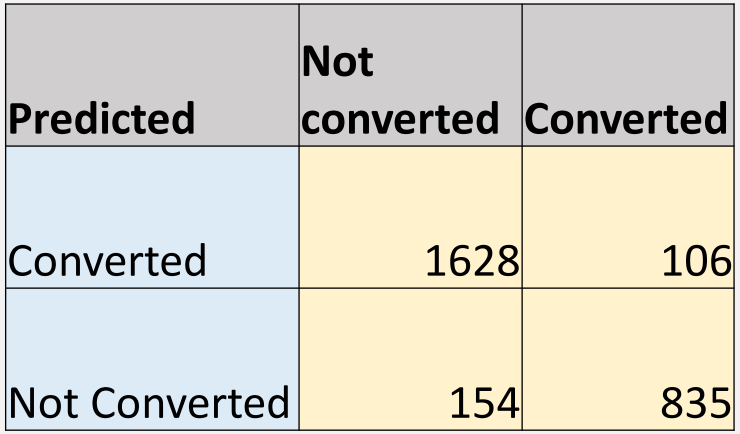
Evaluating model on train data set

Predicted	Not converted	Converted
Converted	3669	236
Not Converted	341	2105

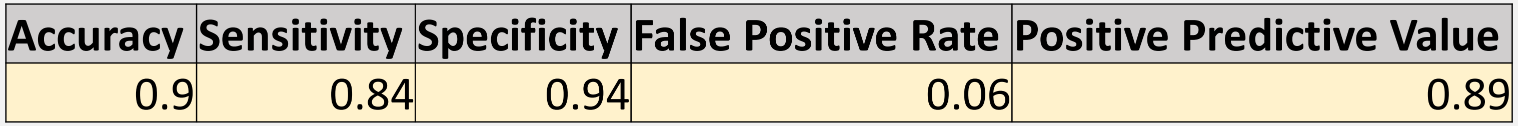
Accuracy	Sensitivity	Specificity	False Positive Rate	Positive Predictive Value
0.91	0.86	0.94	0.06	0.9

Precision	Recall
0.93	0.85

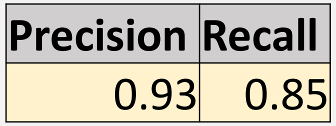
Evaluating model on test data set



Predicted	Not converted	Converted
Converted	1628	106
Not Converted	154	835



Accuracy	Sensitivity	Specificity	False Positive Rate	Positive Predictive Value
0.9	0.84	0.94	0.06	0.89



Precision	Recall
0.93	0.85

Recommendations

Which are the top three variables in your model that contribute most towards the probability of a lead getting converted?

Tags_Lost to EINS
Tags_Closed by Horizon
Tags_Will revert after reading the email

What are the top 3 categorical/dummy variables in the model which get maximum focus in order to increase the probability of lead conversion?

Tags_Lost to EINS
Tags_Closed by Horizzon
Tags_Will revert after reading the email

What would be the focus for an aggressive approach at conversion of leads to buyers

- **We will choose a lower threshold value for Conversion Probability. This will ensure the Sensitivity rating is very high which in turn will make sure almost all leads that are likely to Convert are identified**

What are the top 3 categorical/dummy variables in the model which get maximum focus in order to increase the probability of lead conversion?

We will choose a higher threshold value for Conversion Probability. Specificity rating is very high, ensuring all leads that are on the brink of the probability of getting Converted or not are not selected.