Lead Scoring Case Study

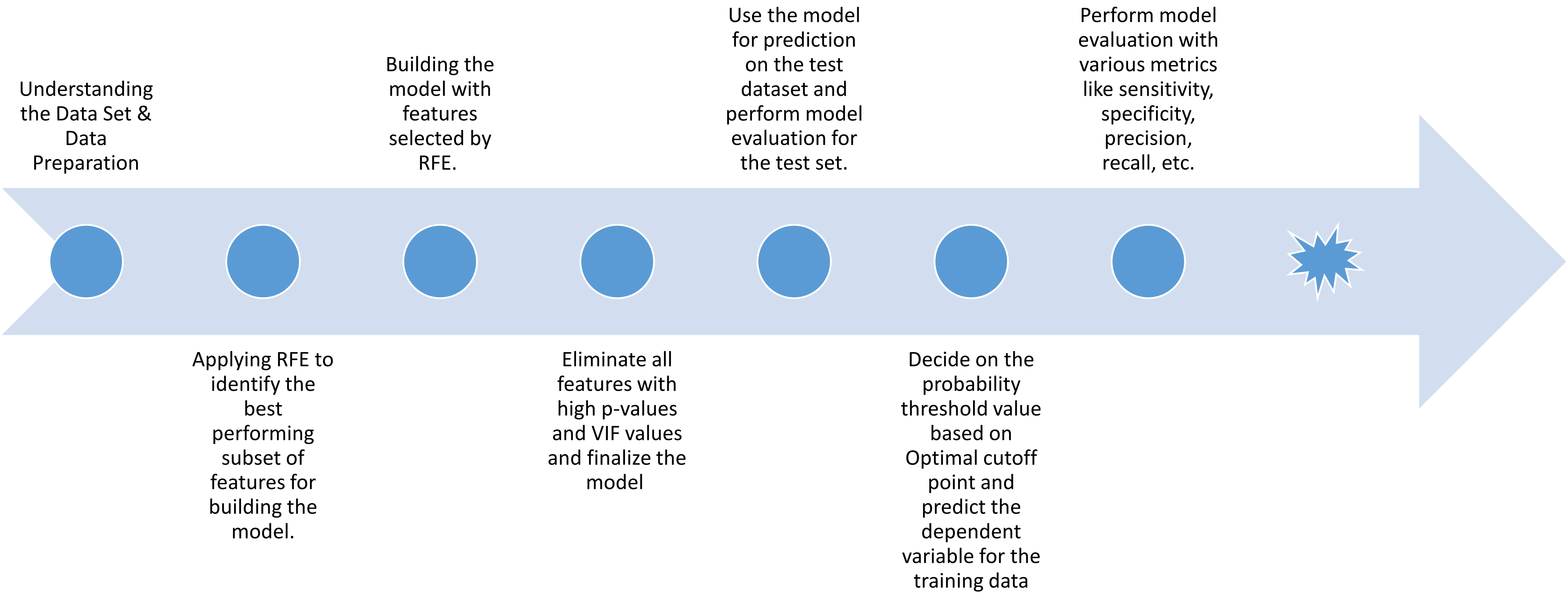
Group Members
Piyush, Kiran and Meenakshi

Business Objective

- **Problem Statement**
- 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 t**o assign a lead score value between 0 and 100 to each of the leads which can be used by the company to target potential leads

- **Subcategories of Objectives :**
- Logistic Regression model to predict the Lead Conversion probabilities for each lead
- Decide on a probability threshold value based on which the lead will be predicted as converted and vice versa.
- Multiply the Lead Conversion probability to arrive at the Lead Score value for each lead.

Problem Solving Methodology



Data Preparation and Feature Engineering

The following data preparation processes were applied to make the data dependable and significant business value by improving Decision Making Process:

Steps	Data Points
Remove columns with single unique value	“Magazine”, “Receive More Updates About our Course”, “Update me on Supply Chain Content”, “I agree to pay the amount through cheque”.
Remove rows where a column has high missing value	“Lead Source”
Imputing null values with Median	“Total Visits”, “Page Views Per Visit” [continuous variable]
Imputing null values with Mode	“Country” [Categorical Variable]
Handling 'Select' values in the columns	Default Option “Select” with only Single Value. Converted to Null
Assigning a Unique Category to NULL/SELECT values	All the nulls in the columns were binned into a separate column ‘Unknown’
Outlier Treatment	“TotalVisits” & “Page Views Per Visit” based on interquartile range analysis.
Binary Encoding	Binary variables (Yes/No) to 0/1: 'Search', 'Do Not Email', 'Do Not Call', 'Newspaper Article', 'X Education Forums', 'Newspaper', 'Digital, ‘Advertisement`, ‘Through Recommendation’, and ‘A free copy of Master the Interview’
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', 'Lead Source', 'Last Notable Activity`
Test-Train Split	Split the dataset to train and evaluate the model
Feature Scaling	‘Standardisation’

Feature Selection via RFE

Running RFE with the output number of the variable equal to 20

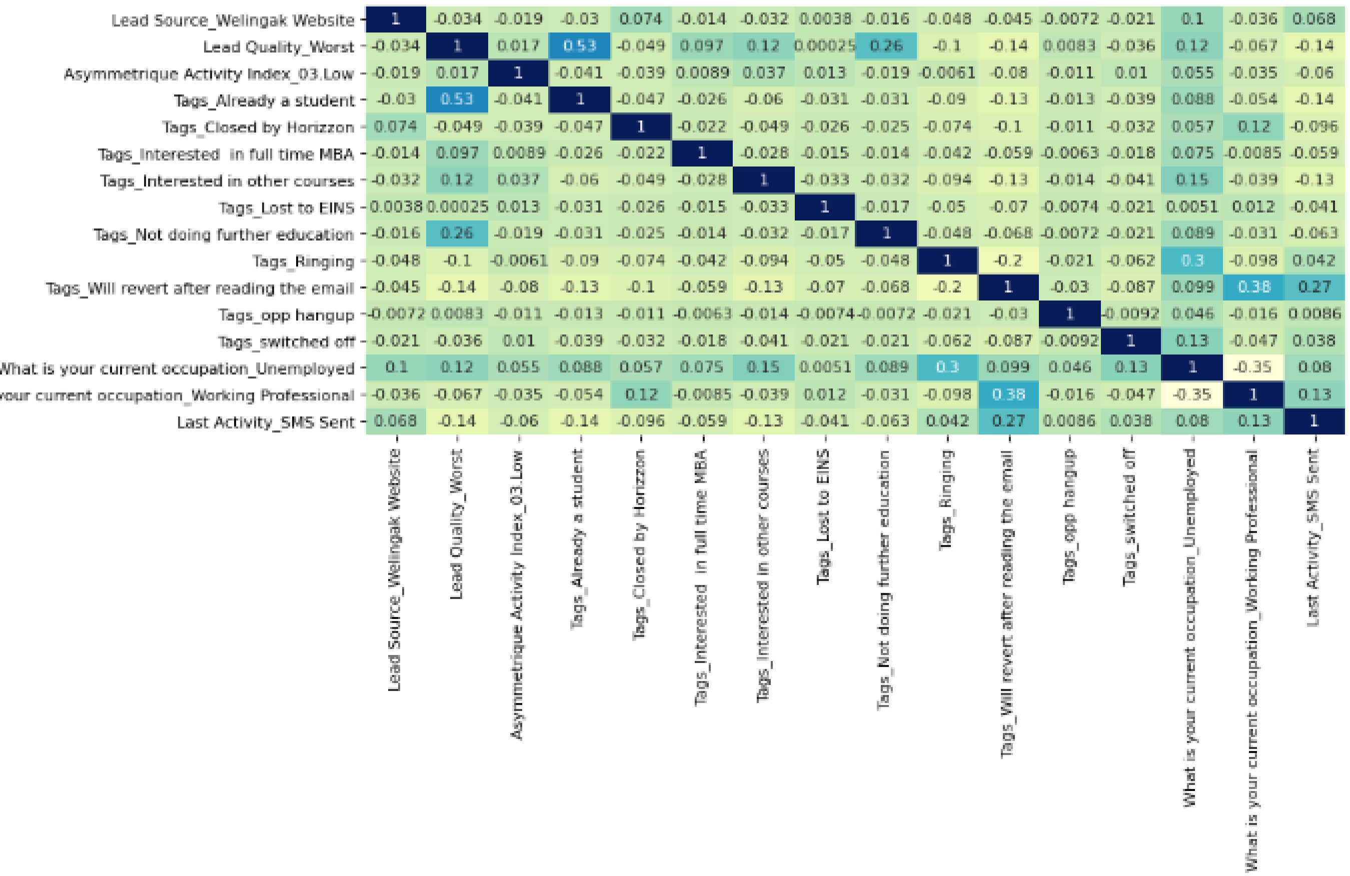
```
from sklearn.linear_model import LogisticRegression
logreg = LogisticRegression()
```

```
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)
```

```
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')
```

Building the Model



A heat map consisting of the final 16 features proves that there is no significant correlation between the independent variables

Our latest model have the following features:

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 0.9125 at a probability threshold of 0.05 is also very acceptable.

So we need not drop any more variables and we can proceed with making predictions using this model only

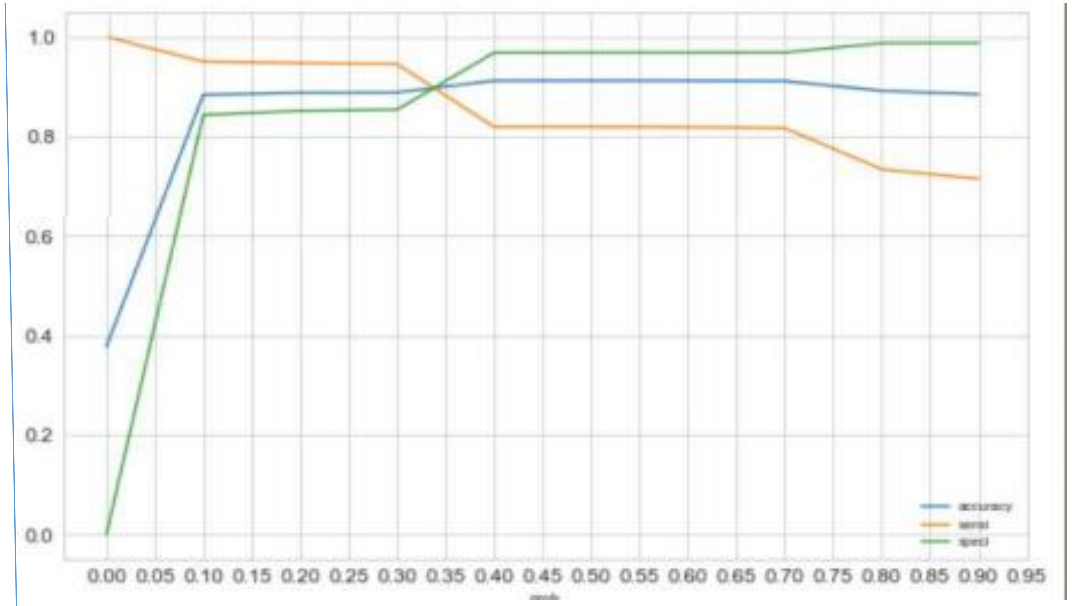
Conversion probability| Probability Threshold

- Creating a data frame with the actual converted flag and predicted probabilities.
- Showing top 5 records of the data frame.

	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 data frame.

	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

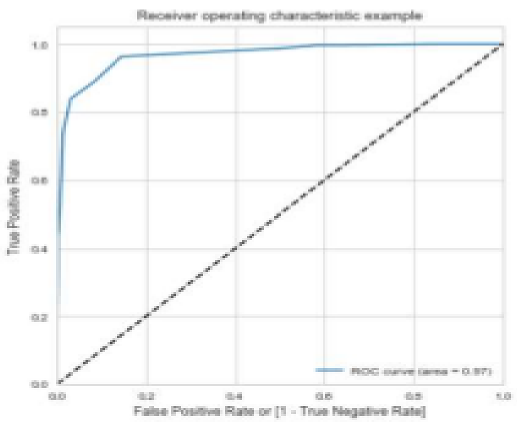


Optimal Probability Threshold

- From the curve above, 0.33 is the optimum point to take it as a cutoff probability.
    - At this threshold value, all the 3 metrics - accuracy sensitivity and specificity is above 80% which is a an acceptable value.

Plotting the ROC Curve and 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)
- The value of AUC for our model is **0.9678.**
- 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 the model is.

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 and Test Dataset

Train Data Set

- Probability Threshold 0.33
- Accuracy 0.903
- Sensitivity 0.887
- Specificity 0.913
- False Positive Rate 0.087
- Positive Predictive Value 0.860
- Negative Predictive Value 0.930
- Precision 0.861
- Recall 0.887
- F1 Score 0.874
- Area under the curve 0.962

Test Data Set

- Accuracy 0.906
- Sensitivity 0.889
- Specificity 0.916
- False Positive Rate 0.084
- Positive Predictive Value 0.870
- Negative Predictive Value 0.928
- Precision 0.870
- Recall 0.889
- F1 Score 0.879
- Area under the curve 0.968
- Cross Validation Score 0.913

Lead Score Calculation | Feature Importance

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,
- 0.33 is 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

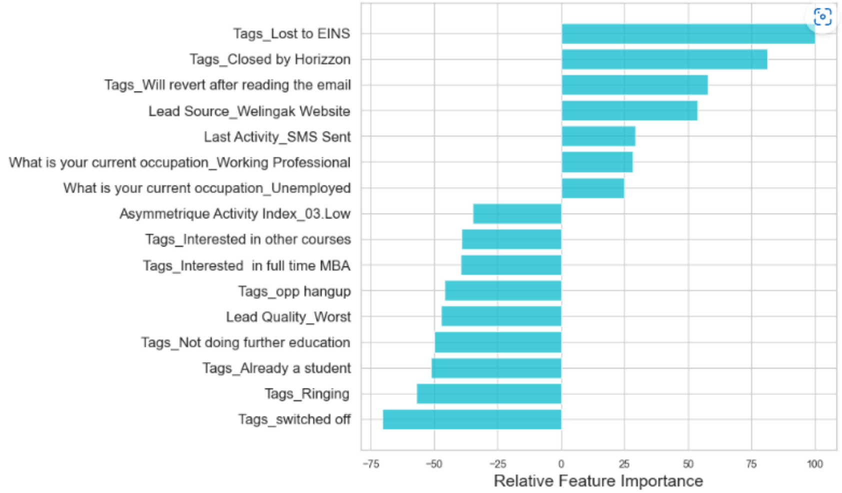
	Lead Number	Converted	Conversion_Prob	final_predicted	Lead_Score
0	660737	0	0.031109	0	3
1	660728	0	0.009566	0	1
2	660727	1	0.801308	1	80
3	660719	0	0.009566	0	1
4	660681	1	0.955452	1	96
5	660680	0	0.077626	0	8
6	660673	1	0.955452	1	96
7	660664	0	0.077626	0	8
8	660624	0	0.077626	0	8
9	660616	0	0.077626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

The Relative Importance of each feature is determined on a scale of 100 with the feature with highest importance having a score of 100.

feature_importance = 100.0 * (feature_importance / feature_importance.max())

The features are then sorted using Quick Sort algorithm.

Finally the sorted features are plotted in a bar graph in descending order of their relative importance.



Inference

- After trying several models, we finally chose a model with the following characteristics:
- 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.

Based on our model, some features are identified which contribute most to a Lead getting converted successfully.

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

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

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

Asymmetrique Activity Index_03.Low
Tags_Interested in other courses
Tags_Interested in full time MBA
Tags_opp hangup
Lead Quality_Worst
Tags_Not doing further education
Tags_Already a student
Tags_Ringing
Tags_switched off

The End

Group Members
Piyush, Kiran and Meenakshi