Lead Scoring Case Study

Problem Statement

- The company markets its courses on several websites and search engines like Google.
- Once these people land on the website, they might browse the courses or fill up a form for the course or watch some videos.
- When these people fill up a form providing their email address or phone number, they are classified to be a lead. Moreover, the company also gets leads through past referrals.
- Once these leads are acquired, employees from the sales team start making calls, writing emails, etc. Through this process, some of the leads get converted while most do not.
- The typical lead conversion rate at X education is around 30%.
- Now the firm want to identify the most potential leads, also known as ‘Hot Leads’.
- The aim here is to build a model wherein a lead score will be assigned to each of the leads.
- The customers with a higher lead score have a higher conversion chance and vice versa
- We are given leads dataset from the past with around 9000 data points and target variable 'Converted'.

Analysis Approach

- Reading and Understanding data:
  - Read and understood the data. Used data dictionary to understand the fields.
  - The dataset had 9240 records and 37 fields.
  - Checked for data duplicity and found there were no duplicates.
  - Found 2 columns which are having ID information and are dropped.
  - Upon analyzing further, identified multiple columns with missing values

```
# Checking null values percentage in each column
round(lead.isnull().sum().sort_values(ascending=False)/lead.shape[0]*100,2)
```

```
Lead Quality                                   51.59
Asymmetrique Activity Index                    45.65
Asymmetrique Profile Score                     45.65
Asymmetrique Profile Index                     45.65
Asymmetrique Activity Score                    45.65
Tags                                           36.29
Lead Profile                                   29.32
What matters most to you in choosing a course  29.32
What is your current occupation                29.11
Country                                        26.63
How did you hear about X Education             23.89
Specialization                                 15.56
City                                           15.37
Page Views Per Visit                            1.48
TotalVisits                                     1.48
Last Activity                                   1.11
Lead Source                                     0.39
```

Analysis Approach

- Data Cleansing & Analysis:
  - Performed data cleansing to eliminate missing values and any other kind of irregularities
  - Dropped columns with 45% or above missing values
  - Evaluated the values in various columns and identified columns with data imbalance
  - Dropped the columns with very high data imbalance
  - Columns with “Select” values were also handled like missing values
  - Fields with very high percentage of “Select” values were dropped
  - Missing values are imputed with Mode of the data.
  - For columns having low frequency/count values, merged all these columns under the name of Others
  - For columns with smaller percentage of null values, dropped the null value records
  - Created Count plots to analyze categorical columns.
  - Created heatmap to analyze numerical attributes with target variables
  - At the end of the data cleansing activities, we were able to retain 8953 records and 12 attributes.

Analysis Approach

- Data Preparation
  - Prepared the data for modelling by properly handling numerical and categorical variables
  - Created dummy variables for categorical variables
  - Special care was taken for “Specialization” column to manually drop the dummy variable corresponding to “Specialization_Not Available” to handle the “Select” value case
  - Dropped Others category columns as these are merge of multiple small data values.
  - Total feature count was 56 at the end of this process.
- Train Test Split
  - To proceed with modelling, we divided our variables into feature variable set (X) and target variable (y)
  - Performed train test split on the above datasets in the ratio 70:30
- Feature Scaling
  - Performed feature scaling on the numeric variables using StandardScaler
  - Fit_transform() was applied on train and transform() was applied on test set
- Feature Correlations
  - Created correlation matrix for the feature variables
  - Due to large number of variables was unable to make any conclusions
- Feature Selection
  - Used RFE to select features from the train dataset
  - Selected a list of 15 features from a set of 56 using RFE.

Analysis Approach

- Model Building
  - The first model built the model using the RFE features and stats model.
  - There were 2 features with high VIF value (>5) and 1 feature with high p value (>0.05)
  - Eliminated the high VIF record and repeated the model creation and elimination process
  - The Model 3 was finalized with optimal p and VIF value

Generalized Linear Model Regression Results

Dep. Variable:	Converted	No. Observations:	6267
Model:	GLM	Df Residuals:	6253
Model Family:	Binomial	Df Model:	13
Link Function:	Logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-1263.3
Date:	Mon, 23 Jan 2023	Deviance:	2526.6
Time:	13:02:18	Pearson chi2:	8.51e+03
No. Iterations:	8	Pseudo R-squ. (CS):	0.6037
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-1.1179	0.084	-13.382	0.000	-1.282	-0.954
Total Time Spent on Website	0.8896	0.053	16.907	0.000	0.786	0.993
Lead Origin_Lead Add Form	1.6630	0.455	3.657	0.000	0.772	2.554
Lead Source_Direct Traffic	-0.8212	0.127	-6.471	0.000	-1.070	-0.572
Lead Source_Welingak Website	3.8845	1.114	3.488	0.000	1.701	6.068
Last Activity_SMS Sent	1.9981	0.113	17.718	0.000	1.777	2.219
Tags_Closed by Horizzon	7.1955	1.020	7.053	0.000	5.196	9.195
Tags_Interested in other courses	-2.1318	0.406	-5.253	0.000	-2.927	-1.336
Tags_Lost to EINS	5.9177	0.611	9.689	0.000	4.721	7.115
Tags_Others	-2.3737	0.206	-11.507	0.000	-2.778	-1.969
Tags_Ringing	-3.4531	0.238	-14.532	0.000	-3.919	-2.987
Tags_Will revert after reading the email	4.5070	0.188	24.002	0.000	4.139	4.875
Last Notable Activity_Modified	-1.6525	0.124	-13.279	0.000	-1.896	-1.409
Last Notable Activity_Olark Chat Conversation	-1.8023	0.491	-3.669	0.000	-2.765	-0.839

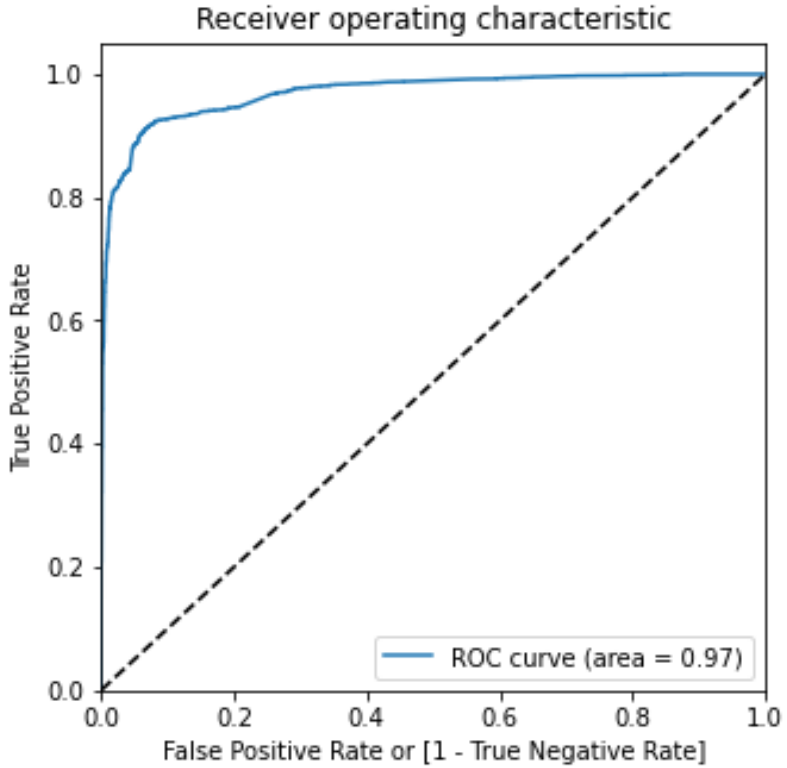
```
# Checking the parameters and their coefficient values
lr3.params
```

```
const                                          -1.117939
Total Time Spent on Website                     0.889556
Lead Origin_Lead Add Form                       1.662993
Lead Source_Direct Traffic                     -0.821189
Lead Source_Welingak Website                    3.884479
Last Activity_SMS Sent                          1.998134
Tags_Closed by Horizzon                         7.195456
Tags_Interested in other courses               -2.131806
Tags_Lost to EINS                               5.917665
Tags_Others                                    -2.373716
Tags_Ringing                                   -3.453095
Tags_Will revert after reading the email        4.506971
Last Notable Activity_Modified                 -1.652548
Last Notable Activity_Olark Chat Conversation  -1.802306
dtype: float64
```

From the lr3 model summary, it is evident that all our coefficients are not equal to zero for all 13 features
The VIF for all features/variables is now less than 5. This means there is very low multicollinearity between the features.
The p-values for all the features/variables is very low and hence are significant in the model.

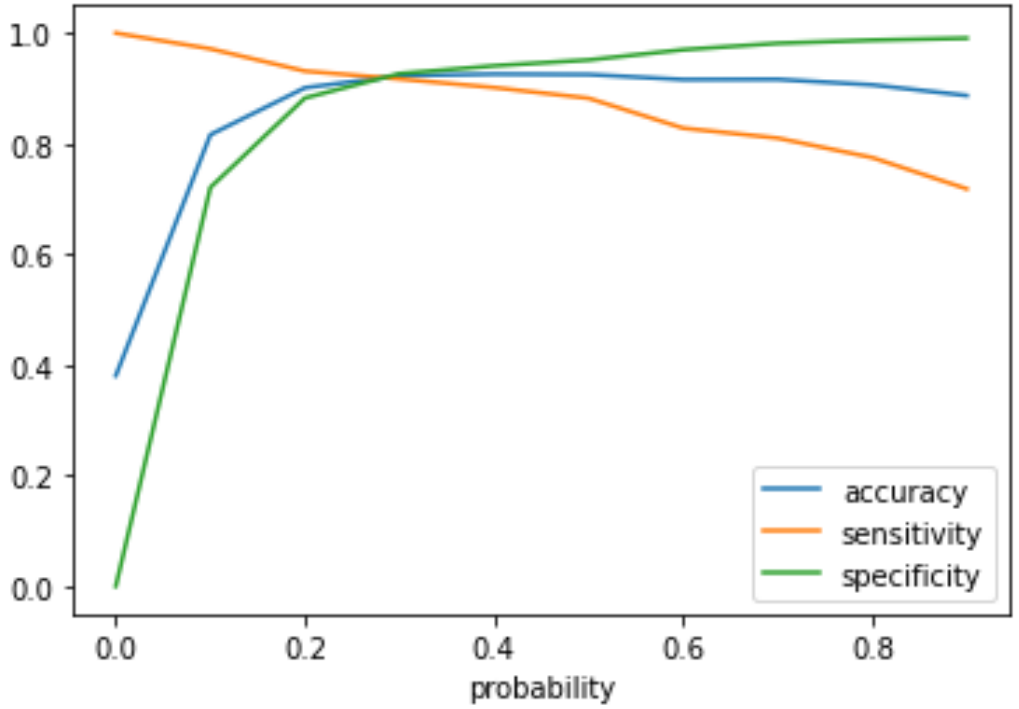
Analysis Approach

- Model Evaluation
  - Model evaluation was performed by making predictions on train with an arbitrary cut off of 0.5
  - Metric Values are:
    - Accuracy – 92.50%
    - Sensitivity – 88.21%
    - Specificity – 95.13%
    - False Positive Rate – 4.86%
    - Positive Predictive Value – 91.75%
    - Negative Predictive Value – 92.92%
  - To evaluate the model further ROC curve was plotted
  - The ROC curve is better if value is close to 1 (i.e., area of the curve). a very good predictive model.
  - The ROC value is 0.97 indicative of a very good predictive model



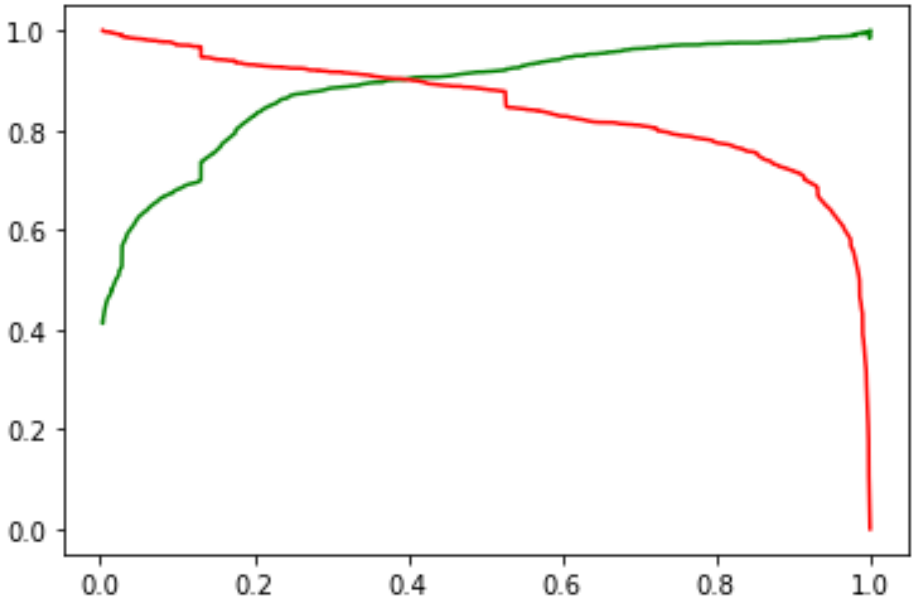
Analysis Approach

- Optimal Cutoff
  - To get an optimal cut-off value, plotted trade off between accuracy, sensitivity and specificity
  - The optimal value was identified 0.30
- Evaluation of new threshold
  - The model was evaluated with new threshold
  - Metric Values are:
    - Accuracy – 92.29%
    - Sensitivity – 91.70%
    - Specificity – 92.66%
    - False Positive Rate – 7.34%
    - Positive Predictive Value – 88.47%
    - Negative Predictive Value – 94.78%
  - These matrix values are all in optimal range



Analysis Approach

- Precision & Recall
  - In addition to other metrices, calculated precision and recall
  - The precision and recall values 88.48% and 91.70% with cut-off of 0.3
  - Created the precision-recall trade off plot
  - Even this plot gave the optimum threshold value between 03 and 0.4



Analysis Approach

- Predictions on Test Set
  - The model was tested on the test set with threshold of 0.3
  - The effectiveness was evaluated using the metrices
  - Confusion matrix was created and evaluation was performed
  - Metric Values are:
    - Accuracy – 92.70%
    - Sensitivity – 91.68%
    - Specificity – 93.32%
    - False Positive Rate – 6.68%
    - Positive Predictive Value – 89.21%
    - Negative Predictive Value – 94.90%
  - The precision and recall values are 89.21% and 91.68%
  - The matrix values of test is very comparable to values received for train
  - This makes sures that the model is effective and not overfitting
  - The Model seems to predict the Conversion/Success Rate very well and we should be able to give the CEO confidence in making good calls based on this model

Final Observation

Let us compare the values obtained for Train & Test:

Train Data:

1. Accuracy : 92.29%
2. Sensitivity : 91.70%
3. Specificity : 92.66%
4. Precision : 88.47%
5. Recall : 91.70%

Test Data:

1. Accuracy : 92.70%
2. Sensitivity : 91.68%
3. Specificity : 93.32%
4. Precision : 89.21%
5. Recall : 91.68%

Suggestions

As per the model, here are few suggestions to improve the conversion rate

- Focus on customers where lead tags are closed by Horizzon as they have high chances of conversion. Look for all the leads which are at status/tag just before the closed by Horizzon tag. Reach out to these stakeholders to move leads to closure state for more lead conversion.
- Focus on those customers where the tags/status is will revert after reading the email. Regularly connect with them for more lead conversion
- Focus on customers who found about the program from Welingak Website and who were marked as lead by adding form. There is a very high chance of these customers leads getting converted.
- Prioritize customers who spend the maximum/most time on the website.
- Give Less priority to customers where tag/status is interested in other courses or are in ringing as these are the least probable to give us a successful lead.
- Give Less priority to people with whom we have last activity as Olark Chat conversation or if lead source is Direct Traffic.
- Customers who was marked as lead by adding form has higher chance of converting to lead

Focusing on the above parameters could heavily impact the conversion rate