9/7/2020

# Lead Scoring Case Study

## Summary Report

Srinivasaragavan V/ Vishal Yadav

# Summary

This analysis is done for X Education and to find ways to get more industry professionals to join their courses. The basic data provided gave us a lot of information about how the potential customers visit the site, the time they spend there, how they reached the site and the conversion rate.

The following are the steps used:

1. **Cleaning data:**
   The data was partially clean except for a few null values and the option select had to be replaced with a null value since it did not give us much information. Unwanted and columns with high skewness were dropped.

2. **EDA:**
   A quick EDA was done to check the condition of our data. It was found that a lot of elements in the categorical variables were irrelevant. The numeric values seem good but outliers were found and capped statistically.

3. **Dummy Variables:**
   The dummy variables were created. And the original columns were removed from the data frame.

4. **Train-Test split:**
   The split was done at 70% and 30% for train and test data respectively.

5. **Model Building:**
   Firstly, RFE was done to attain the top 15 relevant variables. Later the rest of the variables were removed manually depending on the VIF values and p-value (The variables with VIF < 5 and p-value < 0.05 were kept).

6. **Model Evaluation:**
   A confusion matrix was made. Later on the optimum cut off value (using ROC curve) was used to find the accuracy, sensitivity and specificity which came to be around 80% each.

7. **Prediction:**
   Prediction was done on the test data frame and with an optimum cut off as 0.35 with accuracy, sensitivity and specificity of 80%.

8. **Precision – Recall:**
   This method was also used to recheck and a cut off of 0.45 was found with Precision around 77% and recall around 71% on the test data frame.

# Final Observation:

### Comparison of the values obtained for Train & Test:

```
print("Train Data Accuracy    :{} %".format(round((trainaccuracy*100),2)))
print("Train Data Sensitivity :{} %".format(round((trainSensitivity*100),2)))
print("Train Data Specificity :{} %".format(round((trainSpecificity*100),2)))
print("Train Data F1 Score    :{}  ".format(round((trainF1_score_train),2)))
print("Test Data Accuracy     :{} %".format(round((testaccuracy*100),2)))
print("Test Data Sensitivity  :{} %".format(round((testsensitivity*100),2)))
print("Test Data Specificity  :{} %".format(round((testspecificity*100),2)))
print("Test Data F1 Score     :{}  ".format(round((testF1_score),2)))
```

```
Train Data Accuracy    :80.95 %
Train Data Sensitivity :80.73 %
Train Data Specificity :81.08 %
Train Data F1 Score    :0.72
Test Data Accuracy     :81.54 %
Test Data Sensitivity  :80.73 %
Test Data Specificity  :81.08 %
Test Data F1 Score     :0.74
```

### Lead scores for varying cut-off probability

```
prob = []
potential_leads = []
for i in np.arange(0.05,1,0.05):
    prob.append(i)
    potential_leads.append(sum(Y_train_pred_final.Converted_prob.map(lambda x: 1 if x > i else 0)) +
                           sum(Y_pred_final.Converted_prob.map(lambda x: 1 if x > i else 0)))
    projected_leads = pd.DataFrame([prob,potential_leads]).T.rename(columns={0:'Probability Cut-Off',1:'Projected Leads'})
```

```
projected_leads
```

| | Probability Cut-Off | Projected Leads |
|---|---|---|
| 0 | 0.05 | 7870.0 |
| 1 | 0.10 | 6739.0 |
| 2 | 0.15 | 5801.0 |
| 3 | 0.20 | 5381.0 |
| 4 | 0.25 | 4487.0 |
| 5 | 0.30 | 4117.0 |
| 6 | 0.35 | 3794.0 |
| 7 | 0.40 | 3537.0 |
| 8 | 0.45 | 3328.0 |
| 9 | 0.50 | 2891.0 |
| 10 | 0.55 | 2704.0 |
| 11 | 0.60 | 2472.0 |
| 12 | 0.65 | 2273.0 |
| 13 | 0.70 | 2103.0 |
| 14 | 0.75 | 1752.0 |
| 15 | 0.80 | 1585.0 |
| 16 | 0.85 | 1298.0 |
| 17 | 0.90 | 928.0 |
| 18 | 0.95 | 553.0 |

**Selecting the coefficients of the selected features from our final model excluding the intercept**

```
[7]: pd.options.display.float_format = '{:.2f}'.format
     new_params= round(res9.params.sort_values(ascending=False),2)
     new_params
```

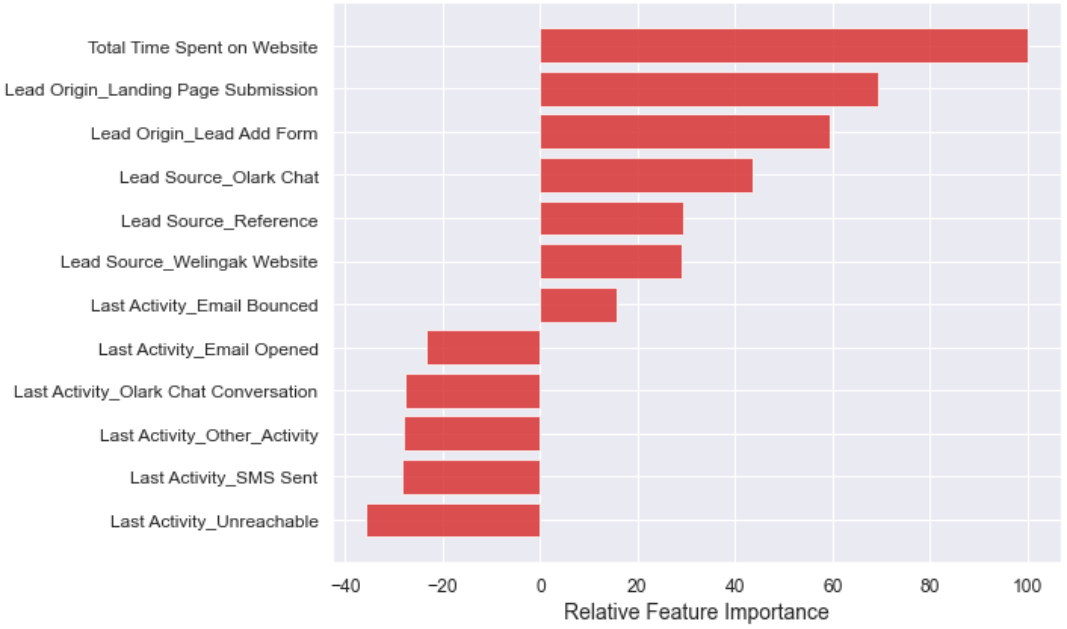
```
[7]: Lead Origin_Lead Add Form                              3.91
     What is your current occupation_Working Professional  2.71
     Last Activity_Other_Activity                          2.33
     Last Activity_SMS Sent                                1.71
     Lead Source_Olark Chat                                1.15
     Total Time Spent on Website                           1.13
     Last Activity_Email Opened                            0.61
     const                                                -0.91
     Lead Origin_Landing Page Submission                  -1.09
     Last Activity_Olark Chat Conversation                -1.10
     Specialization_Others                                -1.11
     Last Activity_Email Bounced                          -1.40
     dtype: float64
```

```
[9]: #Getting a relative coeffient value for all the features wrt the feature with the highest coefficient
     feature_importance = new_params
     feature_importance = 100.0 * (feature_importance / feature_importance.max())
     feature_importance.sort_values(ascending=False)
```

```
[9]: Lead Origin_Lead Add Form                             100.00
     What is your current occupation_Working Professional  69.31
     Last Activity_Other_Activity                          59.59
     Last Activity_SMS Sent                                43.73
     Lead Source_Olark Chat                                29.41
     Total Time Spent on Website                           28.90
     Last Activity_Email Opened                            15.60
     const                                                -23.27
     Lead Origin_Landing Page Submission                  -27.88
     Last Activity_Olark Chat Conversation                -28.13
     Specialization_Others                                -28.39
     Last Activity_Email Bounced                          -35.81
     dtype: float64
```

# Recommendations

X Education Company needs to focus on following key aspects to improve the overall conversion rate:




- Increase user engagement on Welingak website since this helps in higher conversion
- Focus on Working Professional which has high conversion certainty.
- Get Total Time Spent on Website increased by advertising and user experience which makes the customer engaging in the website. since this helps in higher conversion
- Improve the Olark Chat service since this is affecting the conversion negatively
- Improving Lead add form also improves the lead conversion with high certainty

Keeping these in mind the X Education can flourish as they have a very high chance to get almost all the potential buyers to change their mind and buy their courses.