Brief Summary of Lead Conversion Case Study

The lead conversion case study is about creating a logistic regression model, so that the education company XEducation can focus on their hot leads (i.e. high probability for a lead to get converted) which can buy the online courses.

The file Lead.CSV has the leads details i.e. customers who has fill up the forms after browsing the courses.

We have created a logistic regression model to predict the reference variables which has the most impact on conversion variable. So that company can focus on those variables and those leads to increase the lead conversion rate.

As per the current model:

- Accuracy is approx. 90% ,
- Sensitivity : Predicted Converted Out of Actual Converted : 0.82
- Specificity : Predicted Not Convertedd out of Actual Not Converted : 0.95
- False Positive Rate : Predicted Converted when Lead Has Not converted : 0.05
- False Negative Rate: Predicted Not converted When The Lead Has Actually converted : 0.18

Below are the steps in brief that were followed:

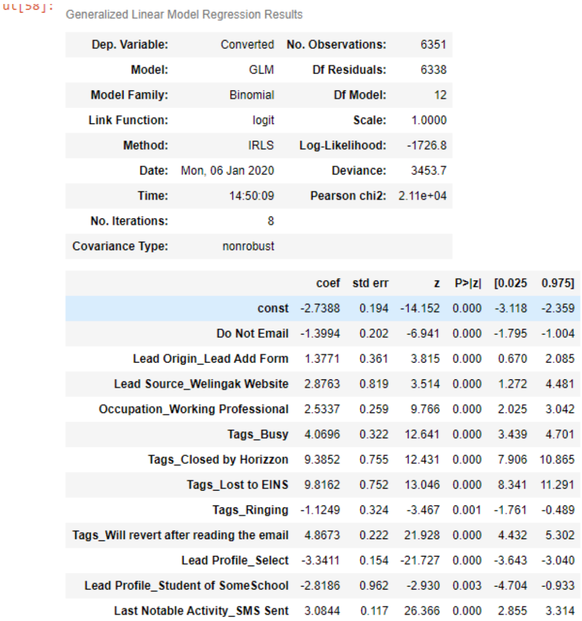
1. Importing Data
    a. The file has 9240 data rows and 37 attributes.
2. Inspecting Data
    a. 4 columns has more than 46% of nulls
    b. 4 columns has Select as Data Value
3. Cleaning and preparing the data for building model:
    a. Replacing the "Select" value with Null
    b. Dropping the columns having more than 46% of nulls
    c. Dropping the Data which has nulls.
    d. Converting the Yes/No columns to (1/0)
    e. Identifying and imputing outliers
4. Splitting the data in train and test
5. Feature Scaling to bring all the variables in common scale
6. Looking at data correlation to eliminate multicollinearity
7. Feature Selection using RFE
    a. Fitting the model in RFE for 13 output variables
    b. Model assessment with Stats Model
    c. Dropped the column having high P values and recreating model
    d. Validation of VIF values to be approx. 3 and less.
    e. Merging the predicted probabilities with Predicted Flag
    f. Decide for a cutoff (0.5) to derive the Predicted Converted Flag

    g. Derive the confusion matrix to assess the model for Train data
    h. Check for model accuracy, sensitivity , specificity , False +ve Rate , False -ve Rate, Precision and Recall.
    i. Plot the ROC Curve to Derive the Tradeoff between Sensitivity and Specificity

8. Find Actual Cutoff point to predict Converted Flag as per the ROC
9. Apply the model in Test Data to validate the model
10. Derive the confusion matrix to identify:
    a. Rate of correctly Predicted Converted lead out of Actual Total Converted lead – Sensitivity
    b. Rate of Correctly Predicted non converted lead out of Total Actual Non-Converted lead – specificity
    c. Rate of Incorrectly Predicted Converted Leads out of Total Actually Non Converted Leads – False +ve Rate
    d. Rate of incorrectly Predicted Not converted Lead out Of Total Actually Converted Leads – False -ve Rate.

Conclusion

As per the model after dropping null columns, null values, Outlier Treatment , Scaling below are the 12 major features that contribute to lead conversion prediction :



ut[58]:

Generalized Linear Model Regression Results

Dep. Variable:	Converted	**No. Observations:**	6351
Model:	GLM	**Df Residuals:**	6338
Model Family:	Binomial	**Df Model:**	12
Link Function:	logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-1726.8
Date:	Mon, 06 Jan 2020	**Deviance:**	3453.7
Time:	14:50:09	**Pearson chi2:**	2.11e+04
No. Iterations:	8		
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-2.7388	0.194	-14.152	0.000	-3.118	-2.359
Do Not Email	-1.3994	0.202	-6.941	0.000	-1.795	-1.004
Lead Origin_Lead Add Form	1.3771	0.361	3.815	0.000	0.670	2.085
Lead Source_Welingak Website	2.8763	0.819	3.514	0.000	1.272	4.481
Occupation_Working Professional	2.5337	0.259	9.766	0.000	2.025	3.042
Tags_Busy	4.0696	0.322	12.641	0.000	3.439	4.701
Tags_Closed by Horizzon	9.3852	0.755	12.431	0.000	7.906	10.865
Tags_Lost to EINS	9.8162	0.752	13.046	0.000	8.341	11.291
Tags_Ringing	-1.1249	0.324	-3.467	0.001	-1.761	-0.489
Tags_Will revert after reading the email	4.8673	0.222	21.928	0.000	4.432	5.302
Lead Profile_Select	-3.3411	0.154	-21.727	0.000	-3.643	-3.040
Lead Profile_Student of SomeSchool	-2.8186	0.962	-2.930	0.003	-4.704	-0.933
Last Notable Activity_SMS Sent	3.0844	0.117	26.366	0.000	2.855	3.314

Below are the Accessibility Numbers:

Train Data Set:

- Accuracy : 89.9%
- Sensitivity : Predicted Converted Out of Actual Converted : 0.82
- Specificity : Predicted Not Converted out of Actual Not Converted : 0.95
- False Positive Rate : Predicted Converted when Lead Has Not converted : 0.05
- False Negative Rate : Predicted Not converted When The Lead Has Actually converted : 0.18

Test Data Set:

- **Accuracy: 89.3**
- **Sensitivity : Predicted Converted Out of Actual Converted : 0.8**
- **Specificity : Predicted Not Converted out of Actual Not Converted : 0.95**
- **False Positive Rate : Predicted Converted when Lead Has Not converted : 0.05**
- **False Negative Rate : Predicted Not converted When The Lead Has Actually converted : 0.2**

Learnings:

As per the current cutoff of 0.3 %

The model has approx. 89 % accuracy with 80% sensitivity and 95% Specificity, hence we can apply the model to predict approx. 80% chances of a lead to get converted. Salesperson can focus on the leads to increase the Lead Conversion Rate of XEducation.

We can decrease the cutoff to 0.2, to predict more leads that has chances to get converted. In this case with more staff sales staff, we can focus on more leads.

Similarly, in we can increase the cutoff to be conservative about the cases which to get converted so that sales team can focus on the strategy for the next year.