LEAD SCORE CASE STUDY

LOGESTIC REGRESSION

THULASI KRISHNA

Problem statement

- X-Education is an education company sells online Education courses to professionals and does its marketing through many online advertisements. Company gets information through different channels and if candidates enquiring with certain education level it calls lead. Typically lead conversion is 30% of certain education. Company identifying Hot Leads on certain criteria also. Lead conversion ratio is lesser than number of enrollment. company given Target to achieve 80% of total enrollment

Business Goal

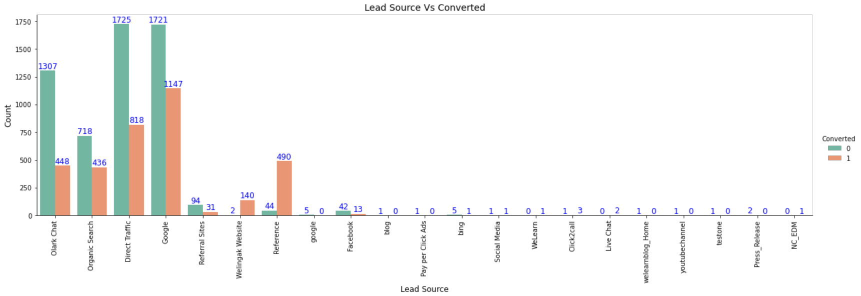
- Build a logistic regression model to assign a lead score between 0 and 100 to each of the leads which can be used by the company to target potential leads. A higher score would mean that the lead is hot, i.e. is most likely to convert whereas a lower score would mean that the lead is cold and will mostly not get converted. There are some more problems presented by the company which your model should be able to adjust to if the company's requirement changes in the future so you will need to handle these as well. These problems are provided in a separate doc file. Please fill it based on the logistic regression model you got in the first step. Also, make sure you include this in your final PPT where you'll make recommendations.

Strategy

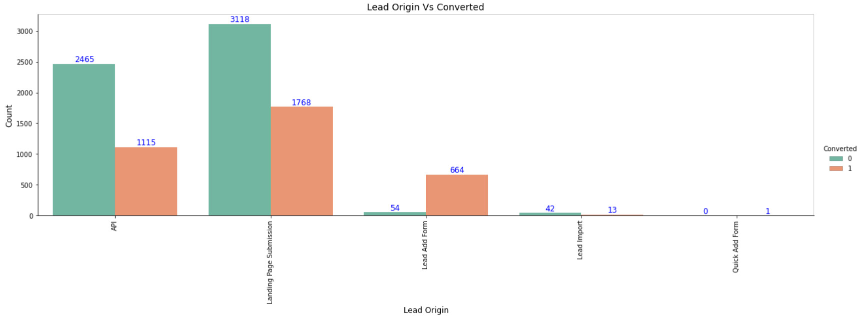
- Import Data
- Clean and prepare the required data for future analysis
- Perform EDA to list out the significant features for lead conversion
- Select the features using RFE
- Scale the features using Standard scaling
- Create the dummy values for categorical variables and prepare the data for creating the model
- Split the train and test data
- Build a fine logistic model by dropping the columns with high p-value as they are insignificant for the predictions
- Assign the lead score for each model based on the final prediction evaluated using optimal cut off value
- Calculate the metrics like confusion matrix,Accuracy,sensitivity,specificity
- Plot ROC curve using the precision and recall values
- Test the model on test data
- Repeat the metrics calculations and validate these values with train data metrics

Exploratory Data Analysis

Major conversion in the lead source is from google

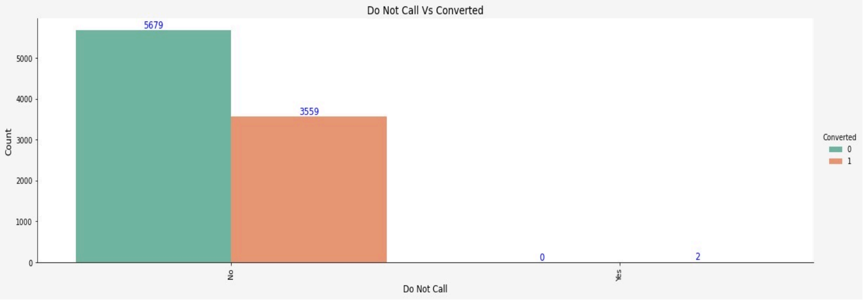


Maximum conversion happened from Landing Page Submission

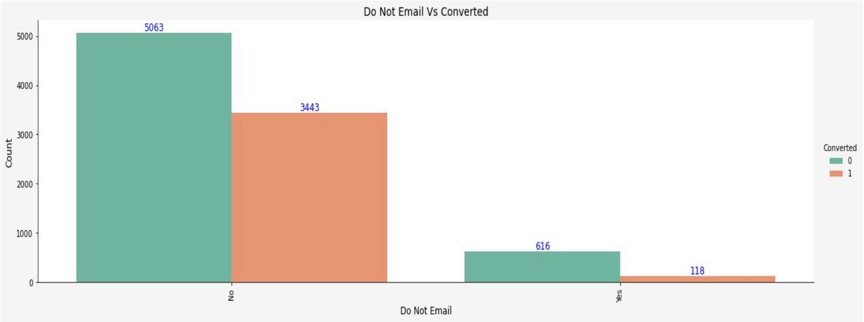


EDA Continues…

Major conversions happened when calls were made

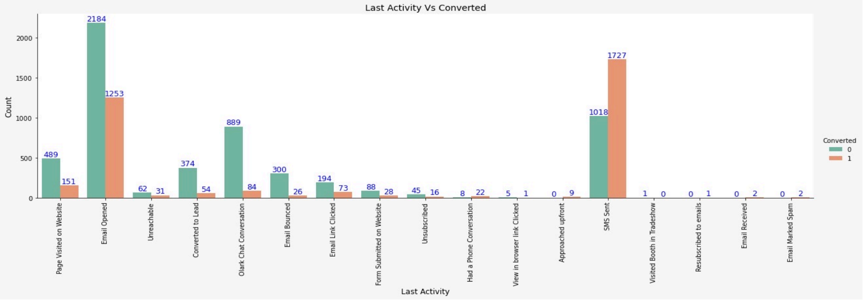


Major conversion has happend from the emails that have been sent

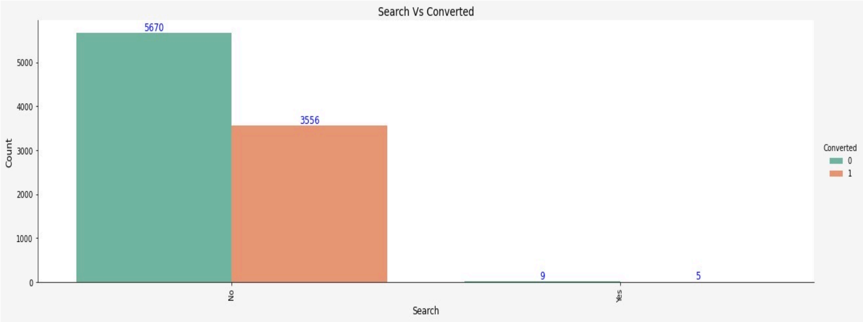


EDA Continues…

Huge number of convertions happened with Email Opened last activity, but the conversion rate is high for SMS Sent

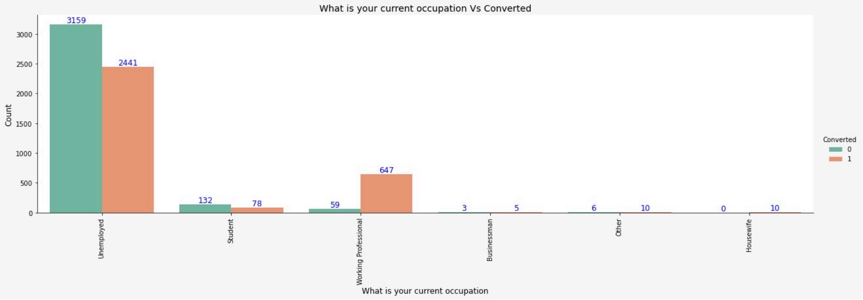


Conversion rate is high on leads who are not through search

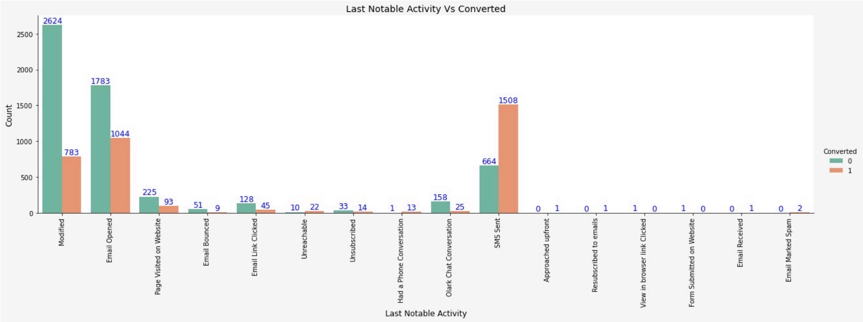


EDA Continues…

More conversion happened with people who are unemployed

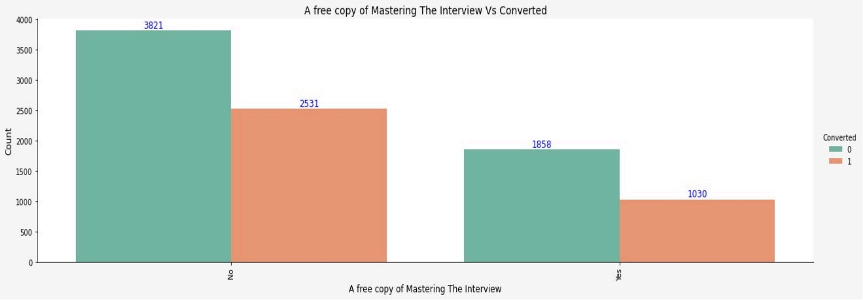


Most Leads are converted with messages , even Emails also include leads



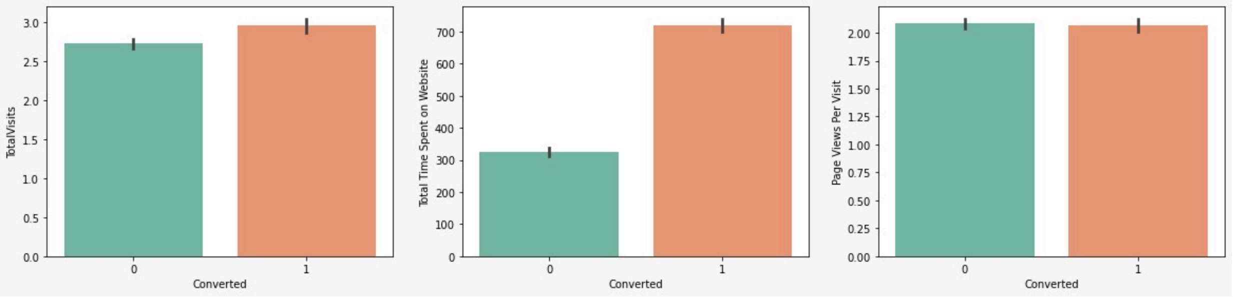
EDA Continues…

Leads prefer less copy of interviews

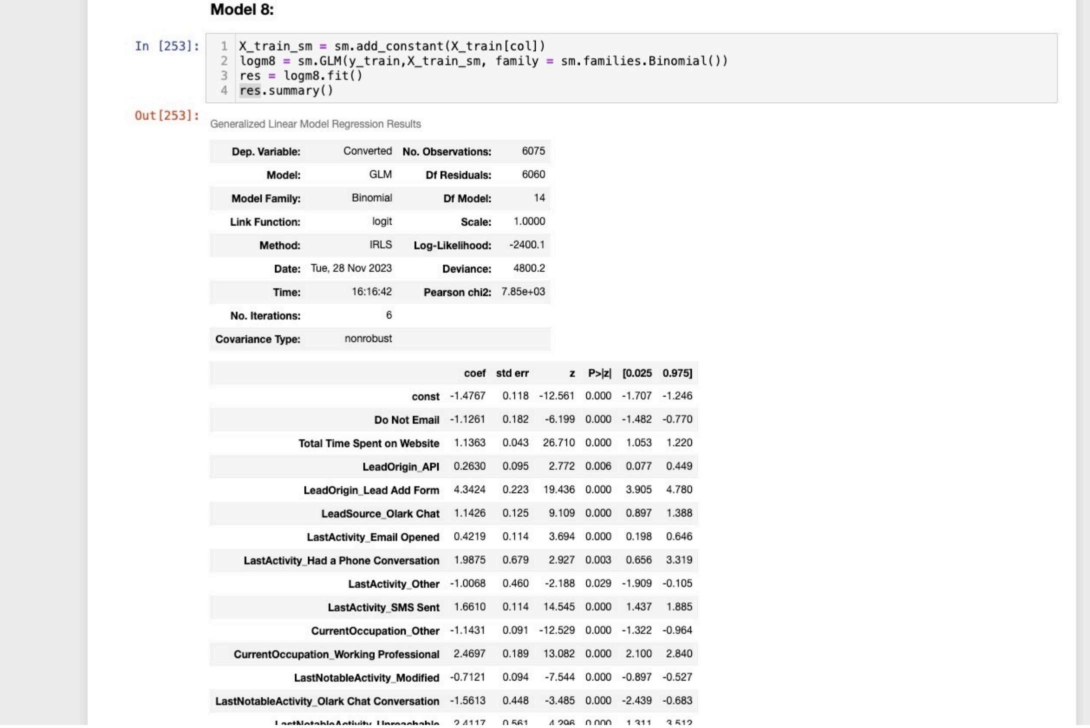


EDA Continues

- People spending more time on website are converting to leads
- More number of visits have slightly more chances to get converted to lead



Final Model Summary



Model 8:

In [253]:

```
X_train_sm = sm.add_constant(X_train[col])
logm8 = sm.GLM(y_train,X_train_sm, family = sm.families.Binomial())
res = logm8.fit()
res.summary()
```

Out[253]:

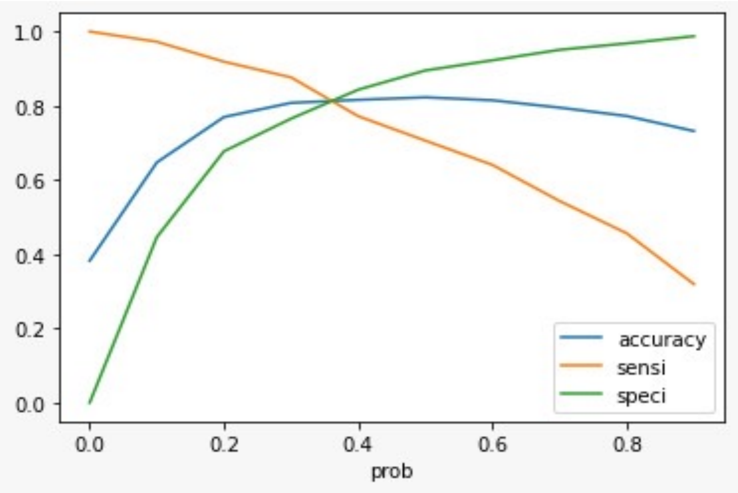
Generalized Linear Model Regression Results

Dep. Variable:	Converted	**No. Observations:**	6075
Model:	GLM	**Df Residuals:**	6060
Model Family:	Binomial	**Df Model:**	14
Link Function:	logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-2400.1
Date:	Tue, 28 Nov 2023	**Deviance:**	4800.2
Time:	16:16:42	**Pearson chi2:**	7.85e+03
No. Iterations:	6		
Covariance Type:	nonrobust		

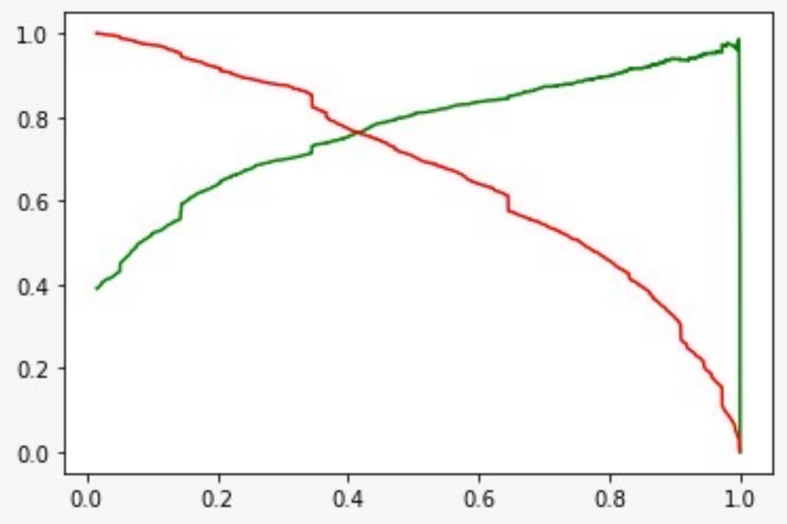
	coef	std err	z	P>|z|	[0.025	0.975]
const	-1.4767	0.118	-12.561	0.000	-1.707	-1.246
Do Not Email	-1.1261	0.182	-6.199	0.000	-1.482	-0.770
Total Time Spent on Website	1.1363	0.043	26.710	0.000	1.053	1.220
LeadOrigin_API	0.2630	0.095	2.772	0.006	0.077	0.449
LeadOrigin_Lead Add Form	4.3424	0.223	19.436	0.000	3.905	4.780
LeadSource_Olark Chat	1.1426	0.125	9.109	0.000	0.897	1.388
LastActivity_Email Opened	0.4219	0.114	3.694	0.000	0.198	0.646
LastActivity_Had a Phone Conversation	1.9875	0.679	2.927	0.003	0.656	3.319
LastActivity_Other	-1.0068	0.460	-2.188	0.029	-1.909	-0.105
LastActivity_SMS Sent	1.6610	0.114	14.545	0.000	1.437	1.885
CurrentOccupation_Other	-1.1431	0.091	-12.529	0.000	-1.322	-0.964
CurrentOccupation_Working Professional	2.4697	0.189	13.082	0.000	2.100	2.840
LastNotableActivity_Modified	-0.7121	0.094	-7.544	0.000	-0.897	-0.527
LastNotableActivity_Olark Chat Conversation	-1.5613	0.448	-3.485	0.000	-2.439	-0.683
LastNotableActivity_Unreachable	2.4117	0.561	4.296	0.000	1.311	3.512

Model Evaluation on Train Data

- Optimal CutOff

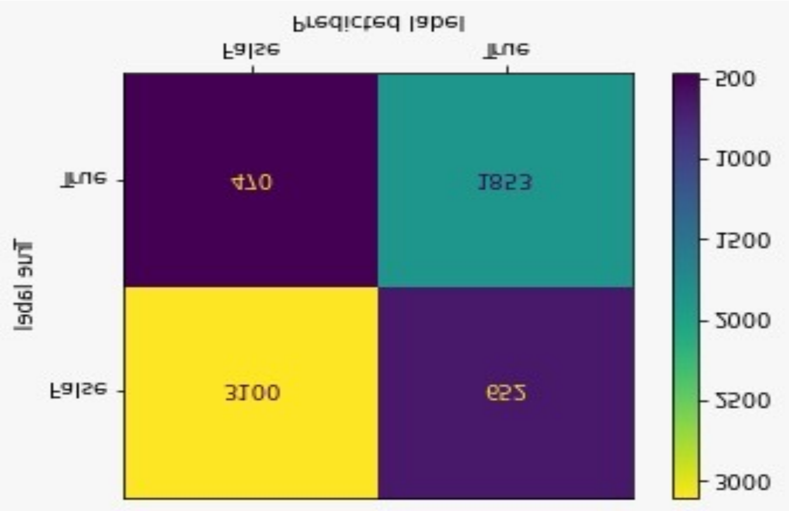


- Precision Recall Tradeoff



Model Metrics

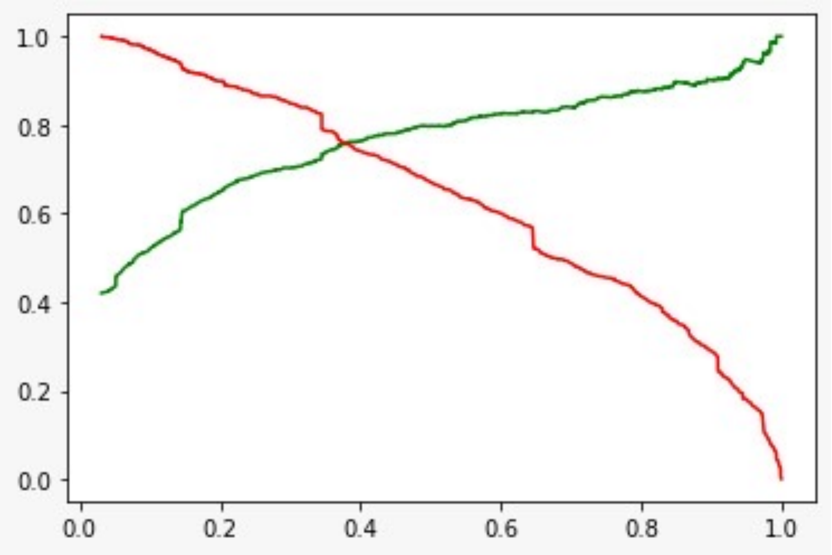
- Confusion Matrix



Accuracy:81.5%
Sensitivity: 79.8%
Specificity:82.6%
Precision:80.1%
Recall:&1%

Model evaluation test

- Precision recall trade off



Accuracy: 81.2%
Sensitivity: 77%
Specificity : 84%
Precision:75.4
Recall:76.6

Conclusion

- -We have achieved final prediction using Sensitivity and Specificity based on optimal cut off value

- Accuracy, Sensitivity and Specificity values of test set are around 81%, 77% and 84% which are approximately closer to the respective values calculated using trained set.
- Also the lead score calculated in the trained set of data shows the conversion rate on the final predicted model is around 80%
- Hence overall this model seems to be good.