LEAD SCORING CASE STUDY

SUBMISSION

Group Members:

1. Ayyappa P
2. Deepa Garg
3. Subarna Saha
4. Manasa Maiya

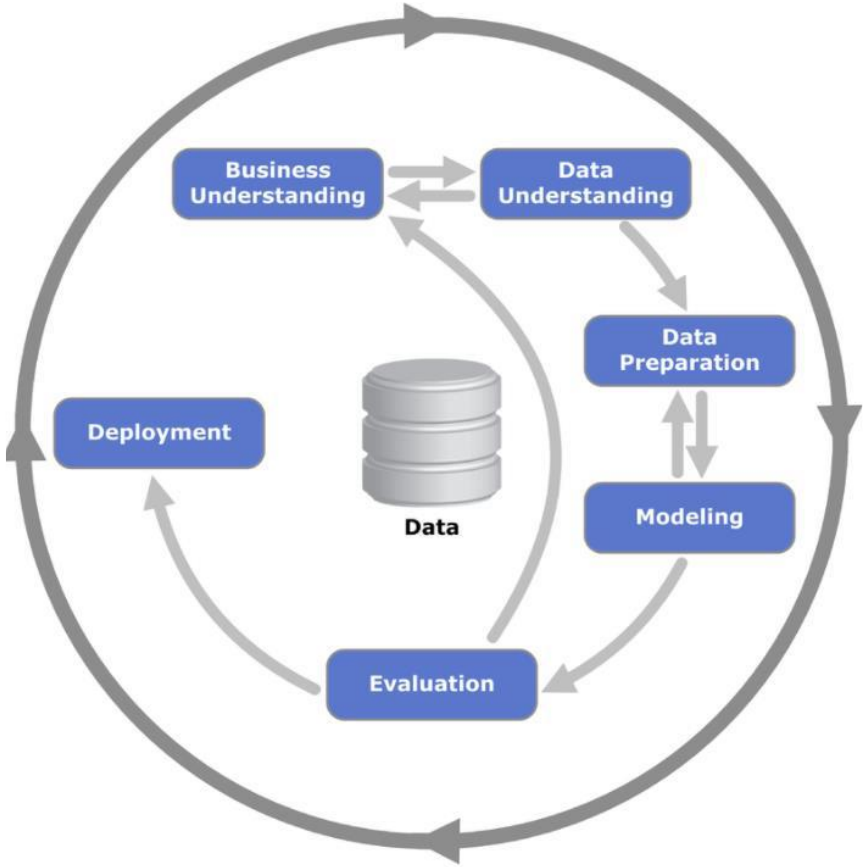
Case Study Objectives

An education company named X Education sells online courses to industry professionals.

- The objective of this analysis is to select the most promising leads, i.e. the leads that are most likely to convert into paying customers, also known as ‘Hot Leads’ for X Education.
  If we successfully identify this set of leads, the lead conversion rate would go up as the sales team will now be focusing more on communicating with the potential leads rather than making calls to everyone.
- Need to build a model by assigning a lead score to each of the leads such that the customers with higher lead score have a higher conversion chance and the customers with lower lead score have a lower conversion chance.
- Target is to get lead conversion rate to be around 80%.

Problem Solving Methodology

We follow the CRISP-DM methodology.



Phase 1 -Business Understanding

Phase 2 -Data Understanding

Phase 3 -Data Preparation

Phase 4 –Modeling

Phase 5 –Evaluation

Phase 6 –Deployment

Business & Data Understanding

Phase 1 -Business Understanding

Need to identify the most potential leads, also known as ‘Hot Leads’ to increase the lead conversion rate.

Phase 2 -Data Understanding

- The data set taken for the analysis is the leads dataset from the past with around 9000 data points. This dataset consists of various attributes such as Lead Source, Total Time Spent on Website, Total Visits, Last Activity, etc which can be used to decide whether the lead will be converted or not.
- The target variable, in this case, is the column ‘Converted’ which tells whether a past lead was converted or not wherein 1 means it was converted and 0 means it wasn’t converted.

Phase 3: Data Cleaning & Preparation

- Imported the data from Leads dataset & loaded in data frame.
- Many of the categorical variables had a level called 'Select' . This has been handled by replacing it by null value.
- Dropped the columns which had more than 3000 missing values.
- Dropped the columns that don’t change as there was no variance.
- Removed & corrected the dirty data by handling null /NaN values.
- Imputation: Identified all the NA values and replace them with appropriate value.
- Converted binary variables(Yes/No) to 1/0.
- Created dummy variables for categorical variables with multiple level.
- Checked for outlier's in continuous variables & performed outlier treatment.
- Identified the variables which needs to be scaled.

Model Building:

- We have not used correlation metrics (heat map) due to high numbers of variables / columns. Due to this we dropped multi – correlated variables once we build our first model.
- After downloading required packages (Stats Model) for Logistic Regression, our first model:

Generalized Linear Model Regression Results

Dep. Variable:	Converted	**No. Observations:**	5958
Model:	GLM	**Df Residuals:**	5896
Model Family:	Binomial	**Df Model:**	61
Link Function:	logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	nan
Date:	Sat, 02 Mar 2019	**Deviance:**	nan
Time:	20:19:30	**Pearson chi2:**	6.35e+18
No. Iterations:	100	**Covariance Type:**	nonrobust

	coef	std err	z	P>|z|	[0.025	0.975]
const	5.478e+16	9.59e+07	5.71e+08	0.000	5.48e+16	5.48e+16
Do Not Email	-5.692e+14	4.9e+06	-1.16e+08	0.000	-5.69e+14	-5.69e+14
TotalVisits	-1.539e+14	1.44e+06	-1.07e+08	0.000	-1.54e+14	-1.54e+14
Total Time Spent on Website	7.971e+14	1.02e+06	7.82e+08	0.000	7.97e+14	7.97e+14
Page Views Per Visit	9.15e+10	1.6e+06	5.71e+04	0.000	9.15e+10	9.15e+10
Search	-1.038e+15	3.12e+07	-3.33e+07	0.000	-1.04e+15	-1.04e+15
Newspaper Article	29.1245	6.42e-07	4.54e+07	0.000	29.125	29.125
X Education Forums	134.8878	4.87e-07	2.77e+08	0.000	134.888	134.888
Newspaper	560.2821	8.07e-07	6.94e+08	0.000	560.282	560.282
Digital Advertisement	9.084e+13	4.76e+07	1.91e+06	0.000	9.08e+13	9.08e+13
Through Recommendations	3.698e+14	4.41e+07	8.39e+06	0.000	3.7e+14	3.7e+14

- We have multiple variables it is important to eliminate some variables, to make this model more sustainable and actionable by business. Though p – values for most of the variable was close to 0.

Model Building: Feature Elimination Using RFE & VIF

- Import Sklearn Model for RFE. **Number of features = 15.** Top 15 features, selected by RFE are:

```
Index(['Do Not Email', 'Total Time Spent on Website',
       'Lead Origin_Lead Add Form', 'Lead Source_Olark Chat',
       'Lead Source_Reference', 'Lead Source_Welingak Website',
       'Last Activity_Email Opened', 'Last Activity_Had a Phone Conversation',
       'Last Activity_SMS Sent', 'What is your current occupation_Housewife',
       'What is your current occupation_Unknown',
       'What is your current occupation_Working Professional',
       'Last Notable Activity_Modified', 'Last Notable Activity_Unreachable',
       'Last Notable Activity_Unsubscribed'],
      dtype='object')
```

- Re – run the model based on above variables, checked VIF and P-Values, based on these values removed some variables and run he model again. These step's have been repeated to get a sustainable model.
- Final model had 13 variables and coefficients are as follows:

Generalized Linear Model Regression Results

Dep. Variable:	Converted	**No. Observations:**	5958
Model:	GLM	**Df Residuals:**	5944
Model Family:	Binomial	**Df Model:**	13
Link Function:	logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-2427.2
Date:	Sat, 02 Mar 2019	**Deviance:**	4854.5
Time:	20:19:31	**Pearson chi2:**	5.94e+03
No. Iterations:	7	**Covariance Type:**	nonrobust

	coef	std err	z	P>|z|	[0.025	0.975]
const	-1.5889	0.111	-14.339	0.000	-1.806	-1.372
Do Not Email	-1.2796	0.202	-6.331	0.000	-1.676	-0.883
Total Time Spent on Website	1.0631	0.042	25.506	0.000	0.981	1.145
Lead Source_Olark Chat	1.2407	0.104	11.901	0.000	1.036	1.445
Lead Source_Reference	3.6413	0.231	15.783	0.000	3.189	4.093
Lead Source_Welingak Website	6.1474	1.014	6.063	0.000	4.160	8.135
Last Activity_Email Opened	0.7112	0.110	6.481	0.000	0.496	0.926
Last Activity_Had a Phone Conversation	3.8169	1.270	3.006	0.003	1.328	6.306
Last Activity_SMS Sent	1.8428	0.111	16.552	0.000	1.625	2.061
What is your current occupation_Unknown	-1.2007	0.090	-13.403	0.000	-1.376	-1.025
What is your current occupation_Working Professional	2.4504	0.197	12.446	0.000	2.065	2.836
Last Notable Activity_Modified	-0.5572	0.091	-6.134	0.000	-0.735	-0.379
Last Notable Activity_Unreachable	2.5806	0.626	4.120	0.000	1.353	3.808
Last Notable Activity_Unsubscribed	1.5263	0.539	2.829	0.005	0.469	2.584

Model Building: Confusion Matrix and Accuracy%

- Predicted probability of getting 1 / converted on our train dataset, called “Lead_Prob”, used predicted state i.e. probability cut off to identify if customer converted or not.
- **Cut off we used = 0.5**

	Converted	Lead_Prob	CustID
0	0	0.078195	6227
1	0	0.116674	6322
2	0	0.045781	3644
3	0	0.129082	3011
4	1	0.869345	8267

	Converted	Lead_Prob	CustID	predicted
0	0	0.078195	6227	0
1	0	0.116674	6322	0
2	0	0.045781	3644	0
3	0	0.129082	3011	0
4	1	0.869345	8267	1

- To check accuracy of model, created confusion matrix and accuracy%.
- So we can see few misclassification as well such as 41 and 709.
- However, accuracy% for this model is 81.06%

```
# Predicted     not_lead    lead
# Actual
# not_lead        3301      419
# lead             709      1529
```

```
# Let's check the overall accuracy.
print(metrics.accuracy_score(y_train_pred_final.Converted, y_train_pred_final.predicted))
```

```
0.81067472306143
```

Model Building: Metrics Beyond Accuracy & ROC Curve

- As accuracy tells model performance based on both class, while in our case only single class (Not Converted) is important. So we also checked sensitivity, specificity, positive / negative predictive value & precision.

```
# Let's see the sensitivity of our logistic regression model
TP / float(TP+FN)
```

```
0.6831992850759607
```

```
# Let us calculate specificity
TN / float(TN+FP)
```

```
0.8873655913978494
```

```
# Calculate false postive rate - predicting as hot leads when customer did not convert
print(FP/ float(TN+FP))
```

```
0.11263440860215054
```

```
# positive predictive value
print (TP / float(TP+FP))
```

```
0.7849075975359343
```

```
# Negative predictive value
print (TN / float(TN+ FN))
```

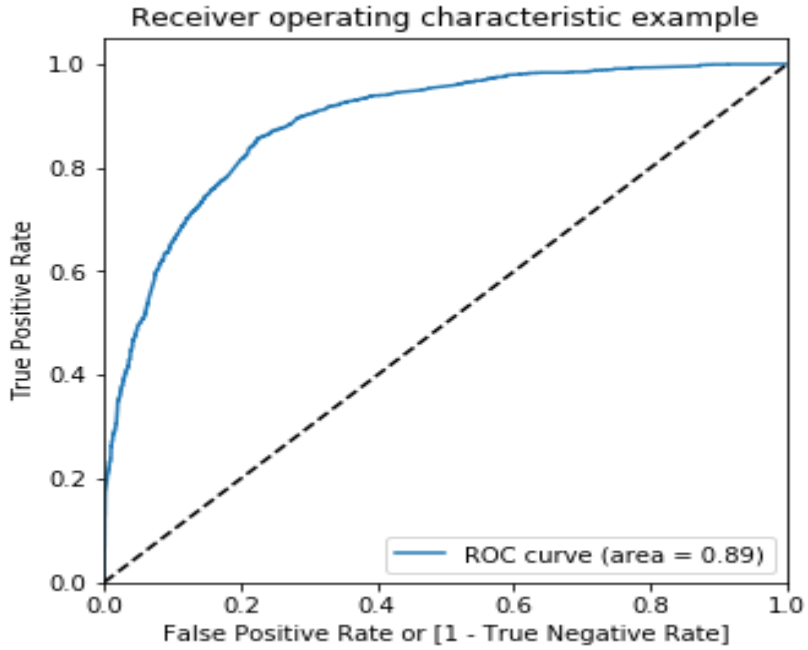
```
0.8231920199501247
```

```
#Precision
#TP / TP + FP

confusion[1,1]/(confusion[0,1]+confusion[1,1])
```

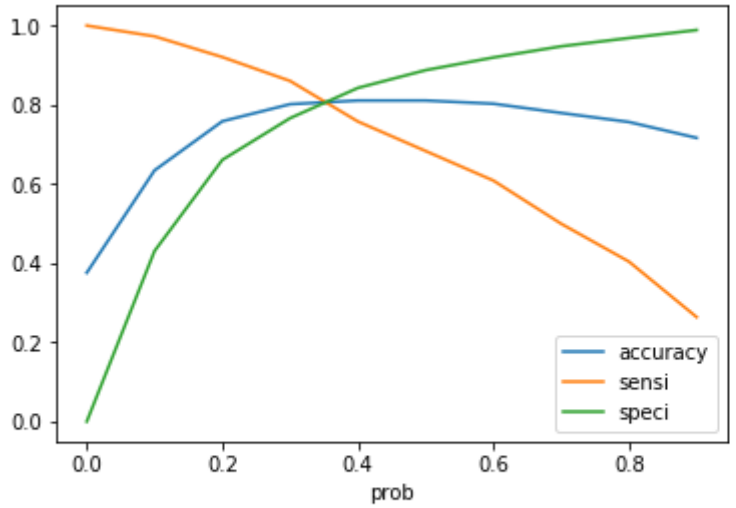
```
0.7849075975359343
```

- To improve the sensitivity of our model, we need to redefine our cut off point.
- ROC Curve: Trade off b/w TPR and FPR. As we know higher the area under the curve of an ROC, the better is model

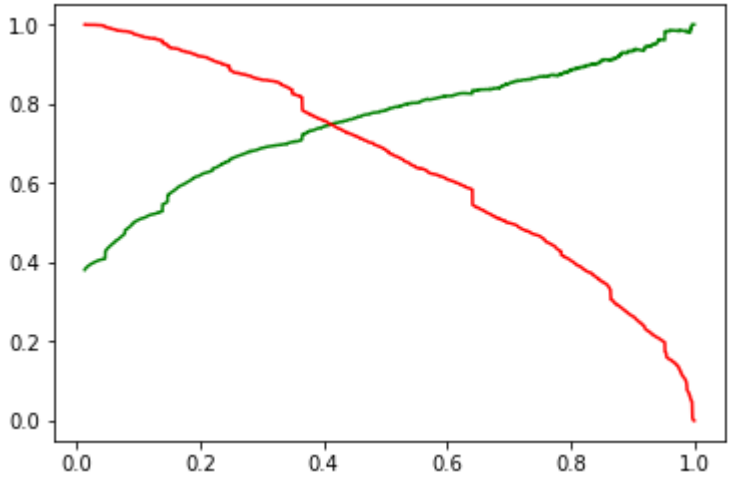


Model Building: Finding the Optimal Threshold

- We calculated values of Accuracy, Sensitivity & Specificity at different cut off values and stored them in a data frame.
- As we can see from below chart, 0.35 is optimum probability cut off for our dataset.
- Accuracy with 0.35 cut off is = 80.3%



- Precision and Recall Tradeoff with 0.35 cut off value
- Precision = 0.784
- Recall = 0.683
- Appended this model on Test dataset with accuracy = 81.04% & Precision = 0.713



From the curve above, 0.35 is the optimum point to take it as a cutoff probability.

Business Requirement & Recommendations:

- As we had business requirement **- Target lead Conversion Rate at ~80%**
- We re-run the model with probability cut off at 0.5 to achieve precision of ~80%.
- With Accuracy Score = 83% & Precision = 0.815

Business Recommendations:

- As we have ~81% of in both train and test database, with precision of 80%, it means we have identified our most of the converted customers correctly.
- However, **by changing cut off limit we can achieve business target**. As by reducing cut off, education group can have more target's while increasing cut off they can limit out targets and will focus only on those customer's those have very high probability of conversion or "Hot Leads".
- Important variables based on our model: **Lead source_Wellingak Website, Lead Activity_Had a Phone Conversation and Lead source_Reference**
- Important Dummy Variables are: **Lead source, Last Activity and Last Notable Activity**
- As we can see from our important variables, it is quite important to understand lead source, in our case it is website link and reference. Which suggest education group to keep **content up to date on online sources** also, have **high touch base with potential customers**.
- Education group can improve customer can include **telephonic discussion as important touch point** with potential customers.