CASE STUDY: LEAD SCORING

BY

1. Lakshmi Ramya
2. Hemendra Sai

Problem Statement & Analysis:

- An education company X sells online courses to various professionals and students through their website.
- They market these courses via different platforms. Based on the leads generated via these platforms, their marketing team pursues them individually.
- Our task at hand is to identify the specific factors that contributes significantly for a lead to get converted.
- We need to build a model and train it on the existing data to classify a lead as hot lead based on the conversion probability Their typical lead conversion has been around 30%.
- We need to build a logistic regression model wherein 80% of the leads that could get converted have to be classified as hot leads as per our model.

Approach:

- Our analysis and building of the model involved the following steps.
    - Data Cleaning
    - Data Preparation
    - Train_Test_Split
    - Correlation, Recursive feature Elimination, VIF

- **Data Cleaning:**

- Given dataset consists of 9240 rows and 37 columns.
- Initially all the ‘select' level items have been converted to null.
- Thereafter some of the columns which have more than 45% of the values missing have been dropped from the analysis.

- Apart from this, we have also identified certain columns which have missing values and are highly skewed. Since imputing them with the mode only increases their skewness further, we have dropped them as well.
- After the above process, we had 9240 rows and 28 columns.
- Some of the variables have less than 200 missing values and we have deleted such rows bringing the total rows to 9074.
- After this, we have 4 out of 28 rows that have missing values
- Since all of them are categorical variables, we have chosen to impute them with mode. Thus, all the missing values have been treated.
- After this, we ran a for loop to find out number of unique values in each column except the numerical and index columns.
- We dropped the columns which have only one unique value. We also removed some of the columns which are highly skewed.
- After the above process, the number of columns have come down to 15 with 9074 rows.

- **Data Preparation:**
- We have observed that some of the categorical variables have many different values with low number of data points for each category.
- Therefore, we have chosen to club some of the categories in some columns to reduce the skewness in different categories.
- Variables which have been treated that way are as follows.
- **Country**
- **City**
- **Specialization**
- **Tags**
- **Lead Source**
- **Last notable activity**
- After this we have converted the variables having 2 categories into binary variables.
- We have also created dummy variables for the other categorical variables.

- **Train-test split:**
- We have divided the dataset to dependent and independent variables by creating data sets X and y
- We have then divided th dataset to train and test with 70% of available data to be used for training the model.
- **Correlation, Recursive feature Elimination, VIF:**
- Using the RFE library from scikit learn, we have obtained the top 15 significant variables for building the model.
- Using these variables, we have performed logistic regression.
- Initially we have eliminated the variables with high P-value and performed regression again.
- In the 2nd regression, all the coefficients turned out to be significant. Therefore, we have proceeded to check for multi-collinearity. There is only one variable with VIF > 5
- After eliminating that variable, we performed regression again and finally obtained the model with significant coefficients and low multi-collinearity.

Final Model:

Final model consists of following variables that are significant

	Coefficient	Std err	z	P-value	2.5%'le	97.5%'le
const	-0.9347	0.124	-7.528	0.000	-1.178	-0.691
Do Not Email	-1.4107	0.169	-8.339	0.000	-1.742	-1.079
Total Time Spent on Website	1.1455	0.040	28.726	0.000	1.067	1.224
Lead Origin_Landing Page Submission	-0.2049	0.091	-2.241	0.025	-0.384	-0.026
Lead Origin_Lead Add Form	4.3319	0.226	19.204	0.000	3.890	4.774
Lead Origin_Lead Import	1.6907	0.448	3.777	0.000	0.813	2.568
Lead Source_Google	0.3644	0.080	4.539	0.000	0.207	0.522
Lead Source_Olark Chat	1.0229	0.127	8.082	0.000	0.775	1.271
Specialization_strong demand	-0.1785	0.083	-2.140	0.032	-0.342	-0.015
Tags_lost	1.0083	0.111	9.066	0.000	0.790	1.226
Tags_unresponsive	-1.1399	0.152	-7.509	0.000	-1.438	-0.842
Last Notable Activity_Modified	-0.8424	0.086	-9.792	0.000	-1.011	-0.674
Last Notable Activity_Other activity	-0.3013	0.129	-2.344	0.019	-0.553	-0.049
Last Notable Activity_SMS Sent	1.5179	0.088	17.245	0.000	1.345	1.690

- Above table suggests the variables which are significant in converting the lead.
- **Recommendation for CEO would be to concentrate on leads which have positive coefficients as increasing them would lead to exponential increase in the probability of conversion.**
- **For example, lets take an example of lead sourcing, model suggests that keeping everything else constant, a lead sourced from Olark chat increase the log odds of converting the lead by 1.0229.**
- **Same way, a lead originated from lead add form increases the log odds of converting the lead by 4.3319.**
- **This gives an initial indication to the marketing team on what leads to be pursued diligently.**

Results: Training Set

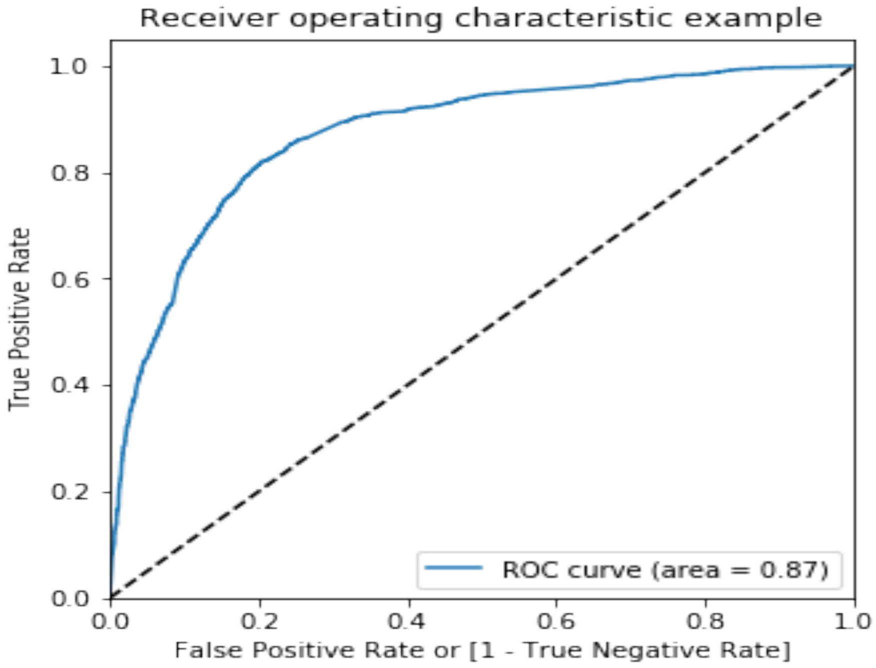
- **Before ROC curve:**
- After obtaining predicted y- values i.e probability of conversion for each data point, we have chosen an initial cutoff point for probability of conversion as 0.5 for the lead to get converted.
- Based on this, we have obtained the following confusion matrix

		Predicted	
		Not converted	Converted
Actual	Not converted	3415	490
	Converted	770	1676

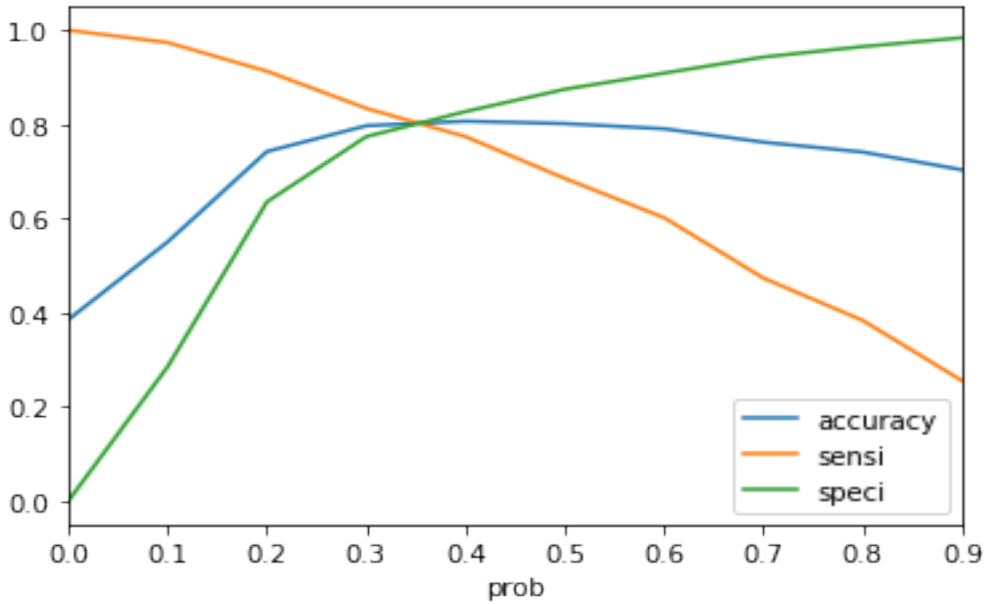
Based on the above matrix, accuracy and sensitivity of the model is as follows.

Accuracy	0.80
Sensitivity	0.68
Specificity	0.87

- **ROC Curve:**
- We have created additional columns with probability cutoffs at 0,0.1,0.2,0.3,0.4,0.5,0.6,0.7,0.8,0.9,1.0
- Based on these observations, we have plotted ROC curve and obtained the following curve.



We have also plotted accuracy, sensitivity, specificity for the above observations and plotted them against the various cutoff points.



- From the above curve, it is observed that optimal cutoff point is between 03 and 0.35.
- We have analyzed the model at a cutoff of 0.35 and obtained the following results.

		Predicted	
		Not converted	Converted
Actual	Not converted	3129	776
	Converted	458	1988

Accuracy	0.80
Sensitivity	0.81
Specificity	0.80

Result: Test set:

- Using the model that we have built, we have applied it to test set and obtained the following results. We have taken the cutoff probability at 0.30 for the test set.

		Predicted	
		Not converted	Converted
Actual	Not converted	1349	385
	Converted	194	795

Accuracy	0.79
Sensitivity	0.80
Specificity	0.78

Recommendations to CEO:

- As it can be observed from the plot between accuracy, sensitivity and specificity, choosing appropriate cutoff is a tradeoff between sensitivity and specificity
- Higher sensitivity will increase the predictability of the model in detecting the leads that could get converted.
- But high sensitivity also leads to a lower specificity which reduces the predictability of the model in detecting the leads that will not get converted. This may lead to making too many unnecessary calls thereby decreasing the efficiency.
- As per the curve, our optimal cutoff is at 0.3-0.35 which has produced a sensitivity of 0.81 for training set and 0.80 for test set with specificity also around the same value.
- We recommend the CEO to decrease the cutoff only when he wants the model to predict higher percentage of leads that could actually get converted although this may reduce specificity