Lead Scoring Case Study

Presented by
Viraj Malgi
&
Vishakha Kude

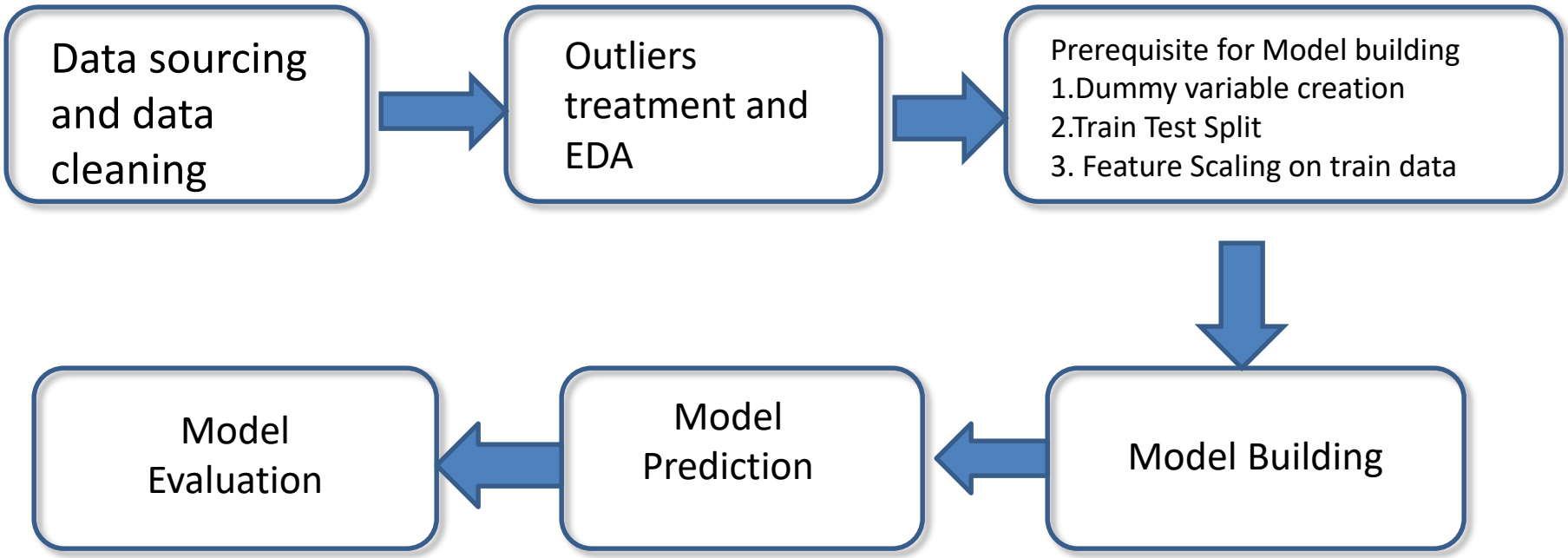
Problem Statement :

- An education company X Education sells online courses to industry professionals.
- The company markets its courses on several websites and search engines like Google.
- The leads are acquired when people land on the website and fill up forms providing phone number or email address, also from past referrals.
- After acquiring leads from various sources, marketing team starts lead conversion process by making calls and writing emails. The company gets a lot of leads but
- the current lead conversion rate at X education is poor, around 30%.
- X education wishes to make lead conversion process more efficient by finding potential leads or hot leads.
- The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%.

Objectives to be achieved :

- 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
- Model to be flexible enough to accommodate future requirements.

Analysis Approach



Data sourcing and Exploration :

Got some useful insights after data exploration.

1. Columns with high % missing values, columns with unique values, columns with ‘Select’ values which is nothing but null, rows with missing values, removed columns which were least influential in analysis ex. Country
2. Imputed null values for categorical variables.
3. Few features having outliers will impact the model, so treated outliers using IQR method
4. Data visualisation in univariate and bivariate analysis helped to visualise data spread across the features.

Prerequisites for Model Building :

Before we start building logistic regression model, we need to prepare dataset accordingly.

1. Creating dummy variables : For categorical variables , we have to create dummy variables using pd.get_dummies().
2. Train Test Split : We build model on training data set and evaluate on test dataset. Hence , we need to split original dataset in train and test sets with 70-30 ratio.
3. Feature scaling : If numerical feature contains high range values , it will create bias in model, so we have to standardize/ scale feature . We have used min-max scaler to scale data of all features between 0 to 1.

Logistic Regression Model Building :

This is basically process of finding most significant features in the dataset which defines probability of lead conversion into paying customers.

- Once we build initial model using Train data, we get coefficients of each variable present.
- We have used RFE method, to select most significant 15 features.
- After its iterative process to remove feature with high p value and VIF and build model again. We keep removing insignificant feature one by one until we get p value for all feature in model < 0.05 and VIF < 5
- Below is the final regression model we build

	coef	std err	z	P>|z|	[0.025	0.975]
const	-0.7914	0.086	-9.241	0.000	-0.959	-0.624
Total Time Spent on Website	4.8021	0.177	27.166	0.000	4.456	5.149
Lead Source_Olark Chat	1.2755	0.106	12.021	0.000	1.068	1.483
Lead Source_Reference	4.4953	0.271	16.579	0.000	3.964	5.027
Lead Source_Welingak Website	6.7074	1.018	6.590	0.000	4.713	8.702
Do Not Email_Yes	-1.8951	0.181	-10.461	0.000	-2.250	-1.540
Last Activity_Had a Phone Conversation	2.4275	1.082	2.243	0.025	0.306	4.549
Last Activity_Olark Chat Conversation	-1.1476	0.198	-5.784	0.000	-1.536	-0.759
What is your current occupation_Working Professional	2.7616	0.191	14.442	0.000	2.387	3.136
Last Notable Activity_Email Link Clicked	-2.1065	0.286	-7.364	0.000	-2.667	-1.546
Last Notable Activity_Email Opened	-1.5278	0.092	-16.651	0.000	-1.708	-1.348
Last Notable Activity_Modified	-2.0428	0.102	-20.039	0.000	-2.243	-1.843
Last Notable Activity_Olark Chat Conversation	-1.7627	0.370	-4.759	0.000	-2.489	-1.037
Last Notable Activity_Page Visited on Website	-1.6121	0.217	-7.425	0.000	-2.038	-1.187

Model Prediction:

Time to predict model !

- Basically its finding probability for each leads using const and coefficients we found in final model.
- Then we create a dataframe with the actual Converted column values and Predicted probabilities for each Lead Number.
- Now we decide cutoff of probability below which predicted conversion is 0 else 1.
- Using actual converted column values and predicted conversion column values we get confusion metrics, using that we can calculate statistics of the model.

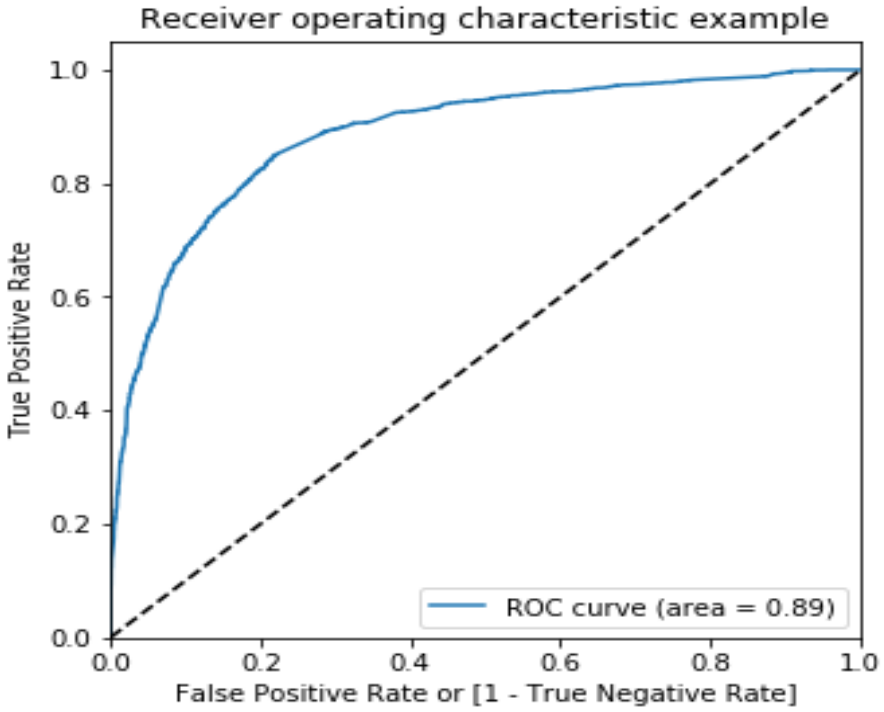
Statistics for model at cutoff value for Converted_prob > 0.51

- Overall_Accuaracy :0.82
- Sensitivity :0.69
- Specificity : 0.89
- False positive rate : 0.11
- **Precision(Positive predictive Value) : 0.8**
- Negative predictive Value : 0.83
- Sensitivity - Specificity : 0.69 , 0.89
- Precision - Recall : 0.8 , 0.69
- F1 score is : 0.7409395973154361

- For finding optimal cut-offs we have sensitivity-specificity view and precision- recall view, covered in upcoming slides
- We use one more technique to see if model is good or not which is ROC curve . It depicts relative trade offs between true positive(benefits) and false positive(costs). covered in upcoming slides

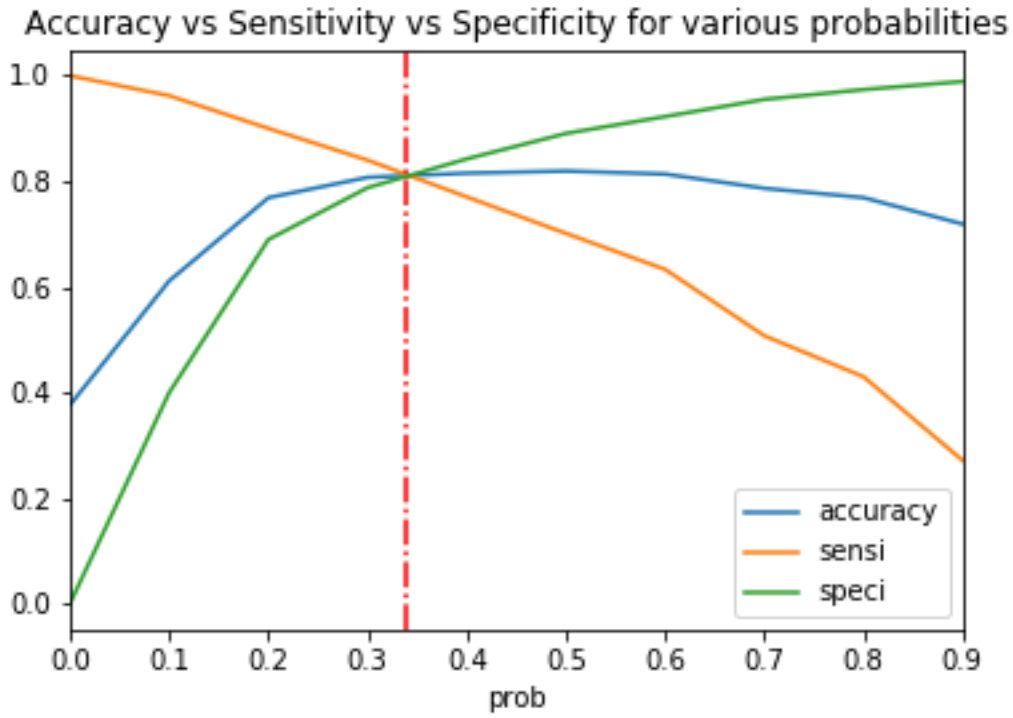
ROC Curve :

Receiver operating characteristic curve is a tool to select optimal model.
It depicts relative trade offs between true positive(benefits) and false positive(costs).
Greater the area under curve, better is the model.



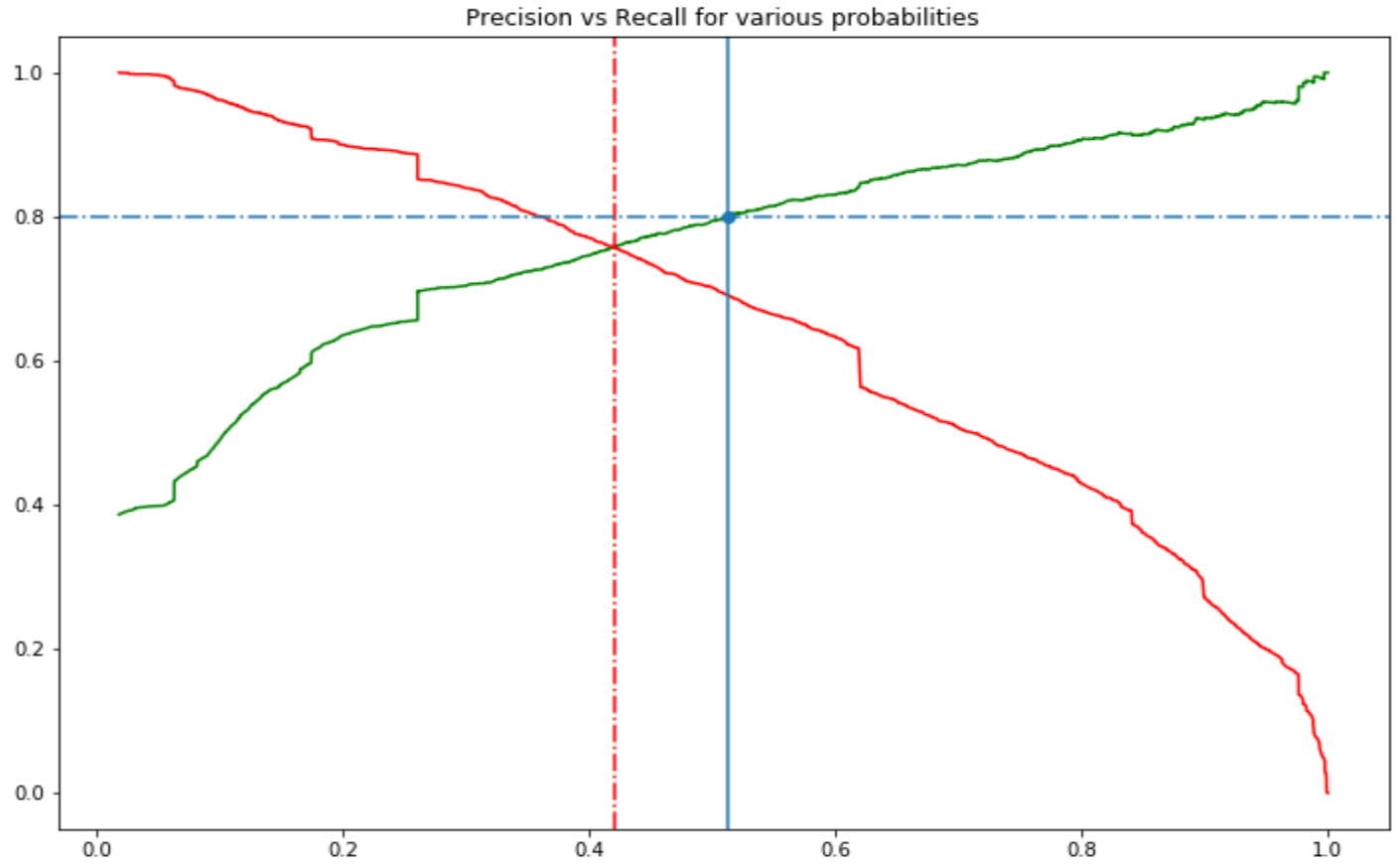
Sensitivity-Specificity View:

To find balanced statistics we have to find tradeoff between accuracy, sensitivity and specificity



Precision – Recall View:

We need positive predictive value or precision at 0.8 which will ensure conversion ratio of 80% and with cutoff for converted prob value as 0.51 is giving us 0.80 precision value. Hence we will go with cutoff of 0.51



Model Evaluation on Test data:

- We predict the probabilities for each lead using build model on train data.
- We predict conversion column using cutoff we decided using Precision- Recall view which is converted_prob> 0.51 results in 1 else 0.
- Using actual converted column and predicted conversion column we build confusion matrix and find statistics

Statistics for test data at cutoff value for Converted_prob > 0.51

- Overall_Accuaracy :0.79
- Sensitivity :0.66
- Specificity : 0.87
- False positive rate : 0.13
- **Precision(Positive predictive Value) : 0.75**
- Negative predictive Value : 0.81
- Sensitivity - Specificity : 0.66 , 0.87
- Precision - Recall : 0.75 , 0.66
- F1 score is : 0.702127659574468

Assigning Lead score :

Using probabilities for converted columns, each Lead Number has been assiged with Lead score between 0 to 1, to which company can follow to find potential leads or hot leads. Higher the lead score higher the chances of getting converted to paying costomer.

```
# Assigning lead score for test set

y_pred_final['Lead_Score'] = y_pred_final['Converted_Prob']*100

y_pred_final.sort_values(by = 'Lead_Score',ascending = False).head()
```

	Lead Number	Converted	Converted_Prob	final_predicted	Lead_Score
2060	7187	1	0.999013	1	99.901340
2178	8120	1	0.999001	1	99.900108
347	1614	1	0.998521	1	99.852070
0	4775	1	0.998446	1	99.844622
1271	4771	1	0.998446	1	99.844622

- **Top three features contributing most towards probability**

1. Lead Source
2. Total Time Spent on Website
3. What is your current occupation

- **Top 3 categorical variables in the model, contributing most**

1. Lead Source_Welingak Website (coeff value - 6.7074)
2. Lead Source_Reference (coeff value - 4.4953)
3. What is your current occupation_Working Professional (coeff value - 2.7616)

- **X Education has a period of 2 months every year during which they hire some interns. The sales team, in particular, has around 10 interns allotted to them. So during this phase, they wish to make the lead conversion more aggressive. So they want almost all of the potential leads (i.e. the customers who have been predicted as 1 by the model) to be converted and hence, want to make phone calls to as much of such people as possible. Suggest a good strategy they should employ at this stage.**

➢ The cutoff value should be minimized to include more number of leads which are predicted as yes. So , how much we should minimize ? For that we should consider sensitivity and specificity view , the trade off between accuracy , sensitivity and specificity will give us balanced statistics for the model with increased number of leads predicted as yes.

- **Similarly, at times, the company reaches its target for a quarter before the deadline. During this time, the company wants the sales team to focus on some new work as well. So during this time, the company's aim is to not make phone calls unless it's extremely necessary, i.e. they want to minimize the rate of useless phone calls. Suggest a strategy they should employ at this stage.**

➢ As company want to minimize useless phone calls, false positive rate should be minimized. We need maximum True predictive value i.e. precision. Hence, we should be looking at maximizing the precision and can choose the threshold from Precision-Recall curve.

Conclusion :

We build Logistic regression model giving 80 % conversion rate and lead scores are applied to each leads which helps in finding potential or hot leads (with higher lead score).

Top features to be focused on are :

1. Lead Source_Welingak Website (coeff value - 6.7074)
2. Total Time Spent on Website (coeff value - 4.8021)
3. Lead Source_Reference (coeff value - 4.4953)
4. What is your current occupation_Working Professional (coeff value - 2.7616)

Recommendations :

1. Provide referral bonus to attract more leads through Reference.
2. Advertisement department should broadcast adds about courses on Job search websites like naukri.com,Linkedin etc to attract more working professionals who tend to search jobs on these websites.
3. Keep an eye on Welingak Website to find new leads.