X EDUCATION LOGISTIC REGRESSION CASE STUDY

SUBMISSION

Group Name:

1. **Divij Jawarani**
2. **Dinesh Challa**

X Education Logistic Regression

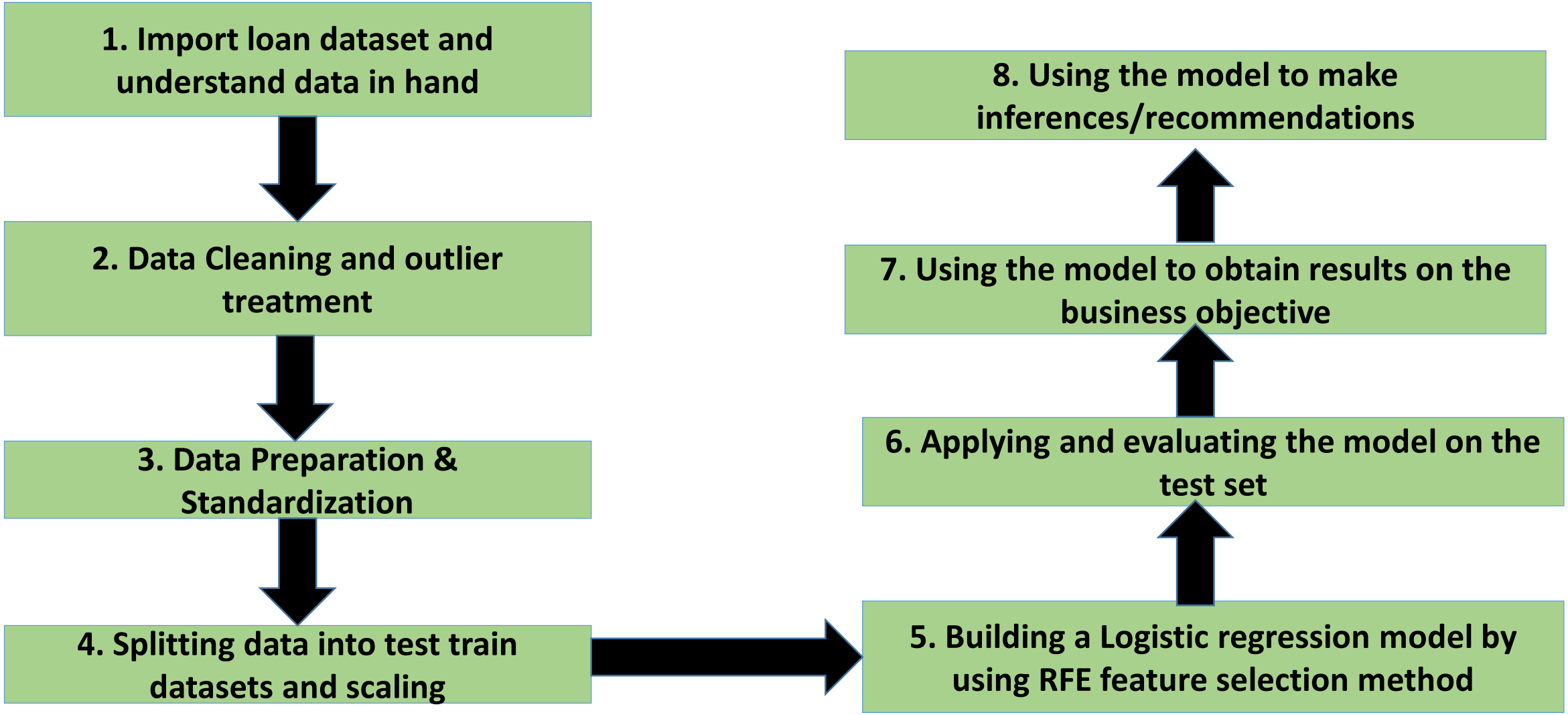
X Education is a company which sells online courses to industry professionals. The company markets its courses on several websites and search engines, people might browse or come across these courses and fill up forms with their details. These people are then identified as leads. X Education would like to increase their lead conversion rate which is currently at 30%.

Business objective: The CEO of X Education would like to assign lead scores to all the leads and identify the hot leads such that the conversion rate is 80%

Goals of data analysis: The goals are divided into three main sub-goals

- Build a logistic regression model which gives a lead score to each lead and can identify as a hot lead
- Try to find optimal lead score to increase lead conversion to 80%
- Give recommendations to solve the given problems of X Education

Problem solving methodology



Logistic Regression Model

We used feature selection through RFE method to build our logistic regression model, after a few iterations, we got our final model

Final Model

Generalized Linear Model Regression Results

Dep. Variable:	converted	**No. Observations:**	5764
Model:	GLM	**Df Residuals:**	5751
Model Family:	Binomial	**Df Model:**	12
Link Function:	logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-2676.0
Date:	Mon, 10 Jun 2019	**Deviance:**	5352.0
Time:	12:11:46	**Pearson chi2:**	6.03e+03
No. Iterations:	6	**Covariance Type:**	nonrobust

	coef	std err	z	P>|z|	[0.025	0.975]
const	-1.1402	0.050	-22.641	0.000	-1.239	-1.042
do not email	-1.0919	0.198	-5.502	0.000	-1.481	-0.703
total time spent on website	0.9784	0.039	25.207	0.000	0.902	1.055
lead origin_Lead Import	1.1362	0.493	2.305	0.021	0.170	2.102
lead source_Olark Chat	0.9419	0.099	9.512	0.000	0.748	1.136
lead source_Reference	4.1427	0.244	16.947	0.000	3.664	4.622
last notable activity_SMS Sent	1.6387	0.080	20.587	0.000	1.483	1.795
last notable activity_Unreachable	2.0494	0.610	3.357	0.001	0.853	3.246
last notable activity_Unsubscribed	0.9737	0.461	2.113	0.035	0.070	1.877
last activity_Converted to Lead	-1.0987	0.195	-5.637	0.000	-1.481	-0.717
last activity_Email Bounced	-0.9980	0.371	-2.689	0.007	-1.725	-0.271
last activity_Had a Phone Conversation	1.9594	0.750	2.611	0.009	0.489	3.430
last activity_Olark Chat Conversation	-1.5499	0.161	-9.620	0.000	-1.866	-1.234

The p-values of each variable are close to 0 and thus all the variables are significant

VIFs

	Features	VIF
0	do not email	2.12
9	last activity_Email Bounced	1.93
3	lead source_Olark Chat	1.59
11	last activity_Olark Chat Conversation	1.36
1	total time spent on website	1.24
5	last notable activity_SMS Sent	1.16
7	last notable activity_Unsubscribed	1.16
4	lead source_Reference	1.10
2	lead origin_Lead Import	1.00
6	last notable activity_Unreachable	1.00
8	last activity_Converted to Lead	1.00
10	last activity_Had a Phone Conversation	1.00

The VIFs of all variables are less than 5 and are under control

Logistic Regression Model

Top 3 variables

The final variables and their corresponding coefficients are as follows

const	-1.140212
do not email	-1.091935
total time spent on website	0.978435
lead origin_Lead Import	1.136200
lead source_Olark Chat	0.941901
lead source_Reference	4.142660
last notable activity_SMS Sent	1.638688
last notable activity_Unreachable	2.049392
last notable activity_Unsubscribed	0.973723
last activity_Converted to Lead	-1.098694
last activity_Email Bounced	-0.997980
last activity_Had a Phone Conversation	1.959365
last activity_Olark Chat Conversation	-1.549923

Coefficients here can be assumed as weightage of each variable in determining the odds of conversion. Hence, the top 3 variables would be:

- lead source_Reference (4.14)
- last activity_Had a Phone Conversation (1.95)
- last notable activity_SMS Sent (1.63)

Initial results

At cut-off point 0.5

Results obtained by manually selecting a 0.5 cut-off point for predicting conversion

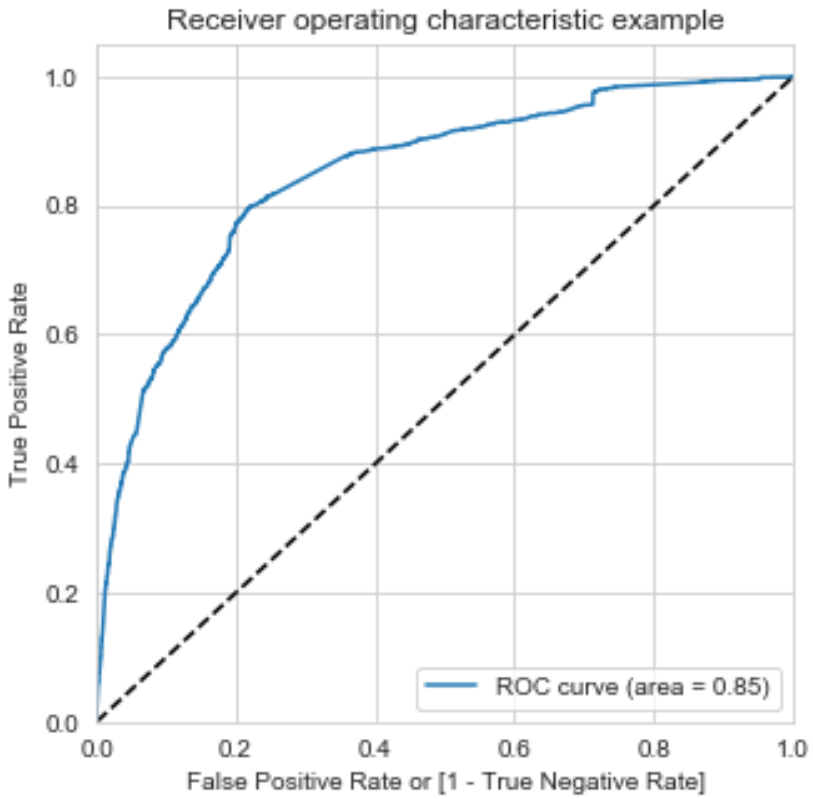
Confusion Matrix

Actual/Predicted	Not Converted	Converted
Not Converted	3089	498
Converted	772	1405

Metric results

Metric	Result
Accuracy	78%
Sensitivity	65%
Specificity	86%
False positive rate	14%

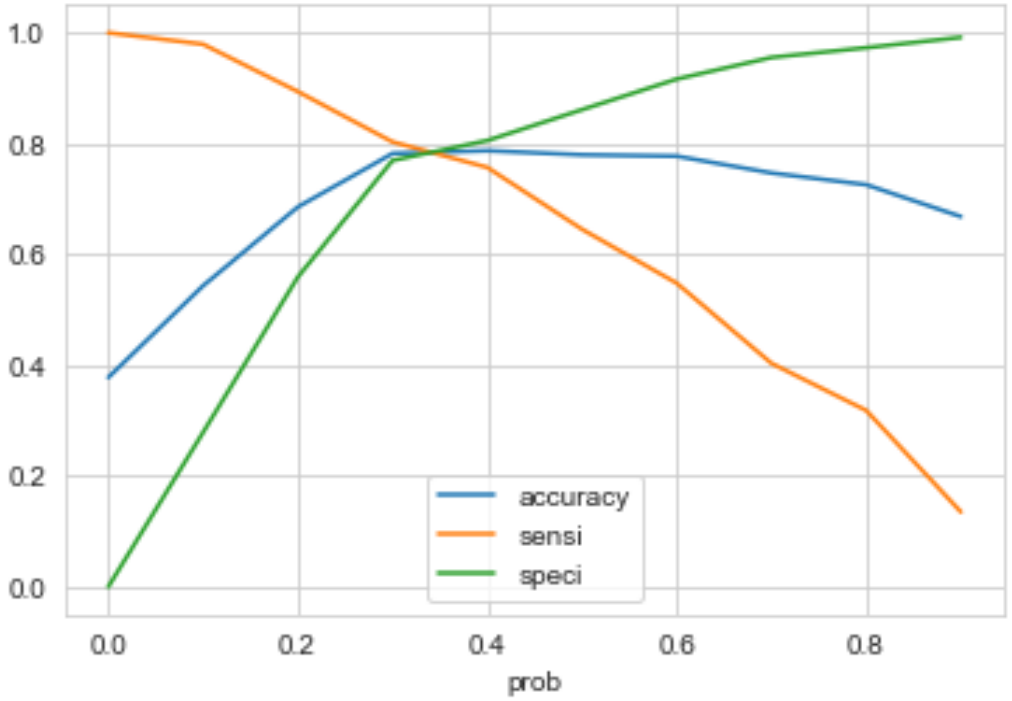
ROC Curve



Finding Optimal Cut-off Point

Optimal Cut-off point for predicting conversion

Optimal Cut-off point graph



Optimal Cut-off point is around 0.35

Confusion Matrix

Actual/Predicted	Not Converted	Converted
Not Converted	2832	755
Converted	471	1706

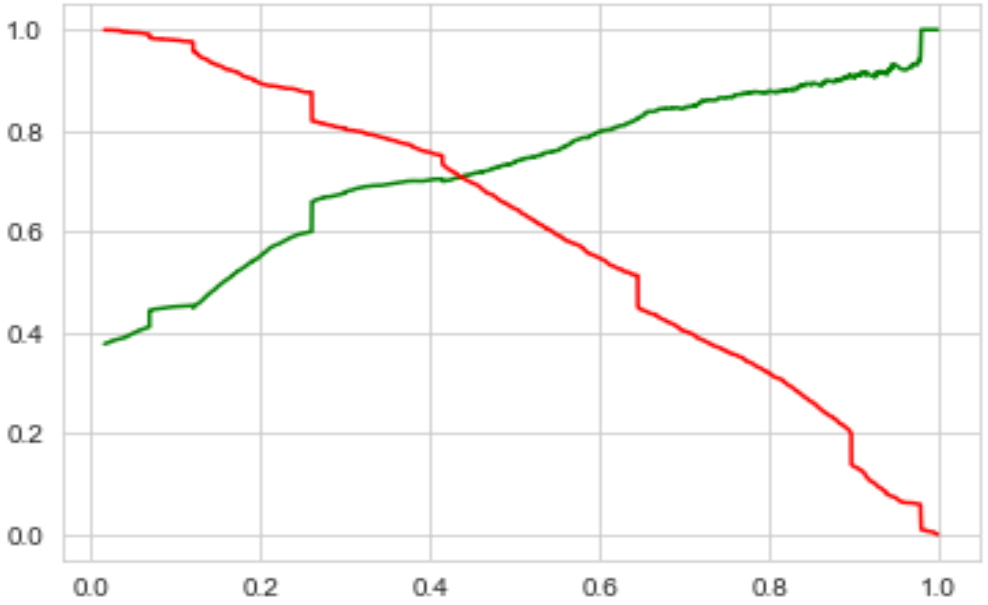
Metric results

Metric	Result
Accuracy	78%
Sensitivity	78%
Specificity	80%
False positive rate	21%

Precision and Recall

Finding out Precision and Recall values and cut-off point

Precision and Recall cut-off point graph



Precision and Recall Cut-off point is around 0.44

Metric results at cut-off point 0.44

Metric	Result
Precision	71%
Recall	70%

Solution to Business Objective

- The CEO of X Education would like to assign lead scores to all the leads and identify the hot leads such that the conversion rate is 80% i.e. the lead score cutoff needs to be adjusted in such a way that, of all the hot leads identified, 80% of them should convert.

- Since, our objective is to have 80% conversion, our model evaluation parameters should be precision and recall. Also, we need to make sure that our model precision is 80%.

Train Dataset Result

Cut-off Point	Precision	Recall
0.44	71%	70%
0.5	74%	65%
0.6	79%	55%
0.61	80%	54%

- Our conversion probability cut-off point is 0.61.
- Hence for a lead score >= 61, we are getting a precision of 80% i.e. of all the hot leads detected, 80% of them converted.

Test Dataset Result

Metric results at cut-off point 0.61	Result
Precision	78%
Recall	52%

- We are at 77.5% precision for the test set i.e. of all the hot leads that we detected 77.5% of them converted.

- So, we are very close to the results obtained in the training set. Hence, for 80% conversion we need to target all the leads >= 61 lead score / 0.61 conversion probability.

Conclusions

We would like to make the following suggestions to X Education:

- **Focus on the following variables the most:**
  i. **Lead Source_Reference** – The leads acquired through a reference have a higher chance of getting converted and they will be more interested to buy a course
  ii. **Last Activity_Had a phone conversation** – The leads which the employees have spoken on the phone to are more likely to get converted as the customer might want more details on phone which are not available on the portal
  iii. **Last Notable Activity_SMS sent** – The leads which send an SMS to the X Education employees are interested in the course and may have a higher chance of conversion

- **Focus on leads with lead score 61 or higher for 80% conversion:** We would recommend X Education to focus on hot leads i.e. leads with a lead score of 61 or higher, this will lead to 80% conversion rate

- **Reduce cut-off point to 0.19 to maximize leads during aggressive lead conversion:** On comparison to the optimum cut-off of 0.35 / 35 lead score (using accuracy, sensitivity & specificity), a cut-off of 0.19/19 lead score would increase the sensitivity by 12% to 90% but also increases the false positive rate from 20% to 46%. But since, X Education has got more interns, contacting those customers with lead score>= 19 would enable them in identifying around 90% of the potential leads.

- **Target leads with lead score 95 or higher to minimize rate of useless phone calls:** If X-education targets customers with lead score>= 95, there is a 0.928 probability of conversion. Since X-education has already met their target, they can now aim for those leads which have the highest conversion probability i.e. the best leads.