X-EDUCATION ONLINE SERVICES

CASE STUDY REPORT FOR IDENTIFICATION OF LEADS

TABLE OF CONTENTS

INTRODUCTION

X Education is an online education company that sells online courses to industry professionals in which many professionals interested in the courses land on their website and browse for courses.

The company targets on identifying the most potential leads form its customers, also known as ‘Hot Leads’. If we are able to successfully identify this set of leads, the lead conversion rate should go up as the sales team will now be focusing more on communicating with the potential leads rather than making calls to everyone. A typical lead conversion process can be represented using the following funnel:

As you can see, there are a lot of leads generated in the initial stage (top) but only a few of them come out as paying customers from the bottom. In the middle stage, you need to nurture the potential leads well (i.e. educating the leads about the product, constantly communicating etc.) in order to get a higher lead conversion.

FOCUS

The main focus is to select the most promising leads, i.e. the leads that are most likely to convert into paying customers. The company requires to build a model wherein we assign 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. The CEO, in particular is actually focusing on ballpark of the target lead conversion rate to be around 80%.

DATA ACQUIRED

Main data consists of 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.. 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.

MAIN GOALS OF STUDY

1. To build a logistic regression model to assign a lead score between 0 and 100 to each of the leads which can be used 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.
2. There are some more problems presented by the company by which the 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. This need to be filled based on the logistic regression model you got in the first step along with highlighted recommendations.

LINEAR REGRESSION ANALYSIS

PRIMARY ANALYSIS

From the actual data,

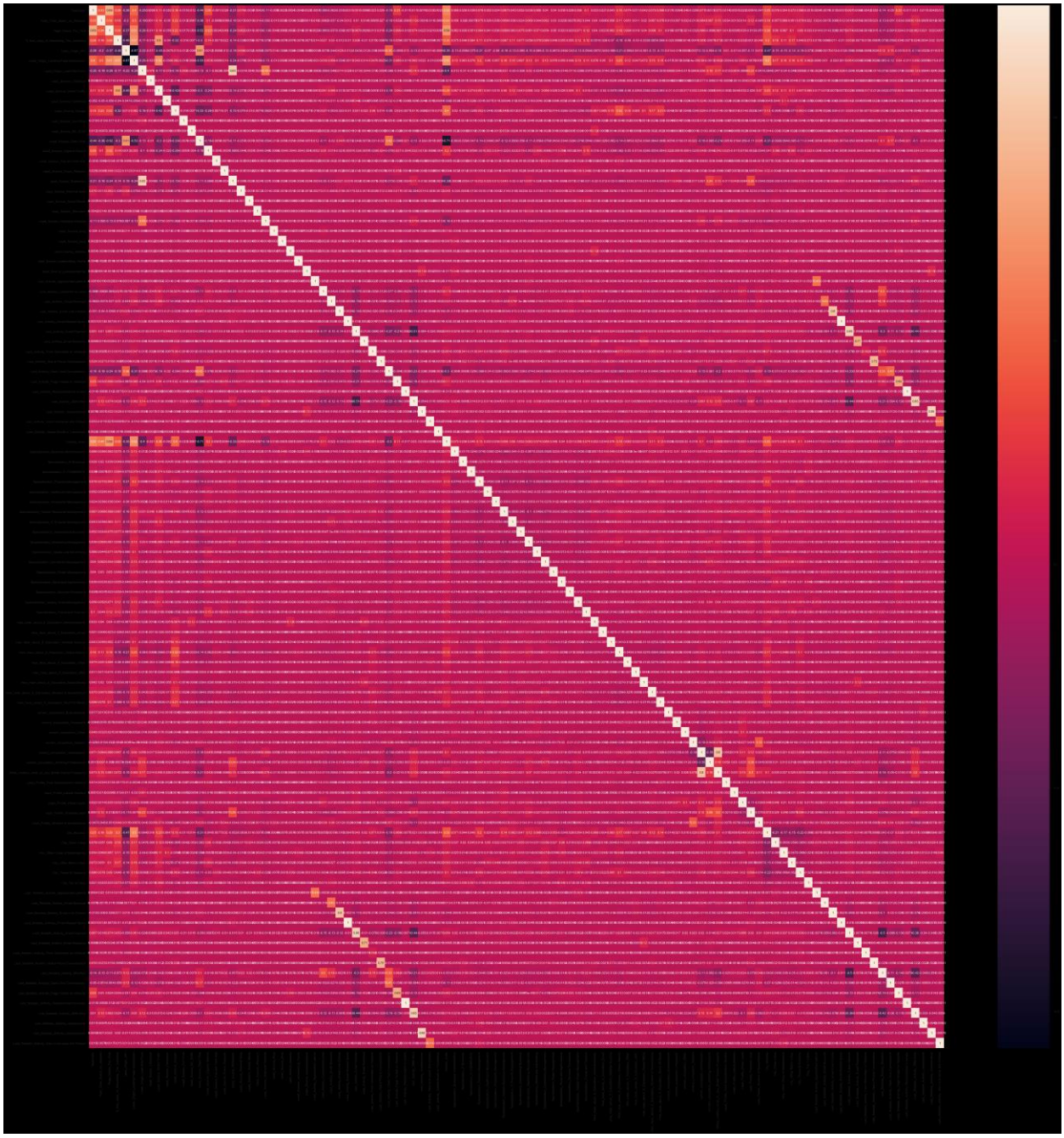
We have taken all the important leads that are needed to extract the understandings about the leads and the following primary findings are shown:

1. There are totally 37 variables, with 6 numeric, 15 categorial and 10 boolean variables
2. About 12% of cells are missing (has no values)

3. There is no duplicate rows
4. There are many variables with very high percentage of missing values that need to cleaned
5. There are a number of variables that has only single value which has no effect in the final result. These will be removed.
6. Around 30% values have missing values. So, we need to remove higher than 30% missing values.

SCALING OF DATA

We use StandardScaler for scaling data being prescribed. Now, the data is being taken for unit conversion in which the conversion rate tends to be at 38.05. Then, it is checked for correlation matrix against each of the variables for clarification regarding the leads.



The correlation matrix shows some degree of correlation (insignificance) for some variables but most are not or less correlated.

MODEL BUILDING

The leads need to be analysed for train – test analysis which is the route to model building. Here, different models are being built in order to reach optimal level.

Dep. Variable:	Converted	No. Observations:	6352
Model:	GLM	Df Residuals:	6253
Model Family:	Binomial	Df Model:	98
Link Function:	logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	nan
Date:	Sun, 01 Mar 2020	Deviance:	nan
Time:	05:46:20	Pearson chi2:	7.18e+18
No. Iterations:	100		
Covariance Type:	nonrobust		

Then, the model is created using RFE (Recursive Feature Elimination) and also slightly deviated results are obtained.

Dep. Variable:	Converted	No. Observations:	6352
Model:	GLM	Df Residuals:	6331
Model Family:	Binomial	Df Model:	20
Link Function:	logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-2421.4
Date:	Sun, 01 Mar 2020	Deviance:	4842.8
Time:	05:46:23	Pearson chi2:	6.35e+03
No. Iterations:	23		
Covariance Type:	nonrobust		

The third model is the application of VIFs.

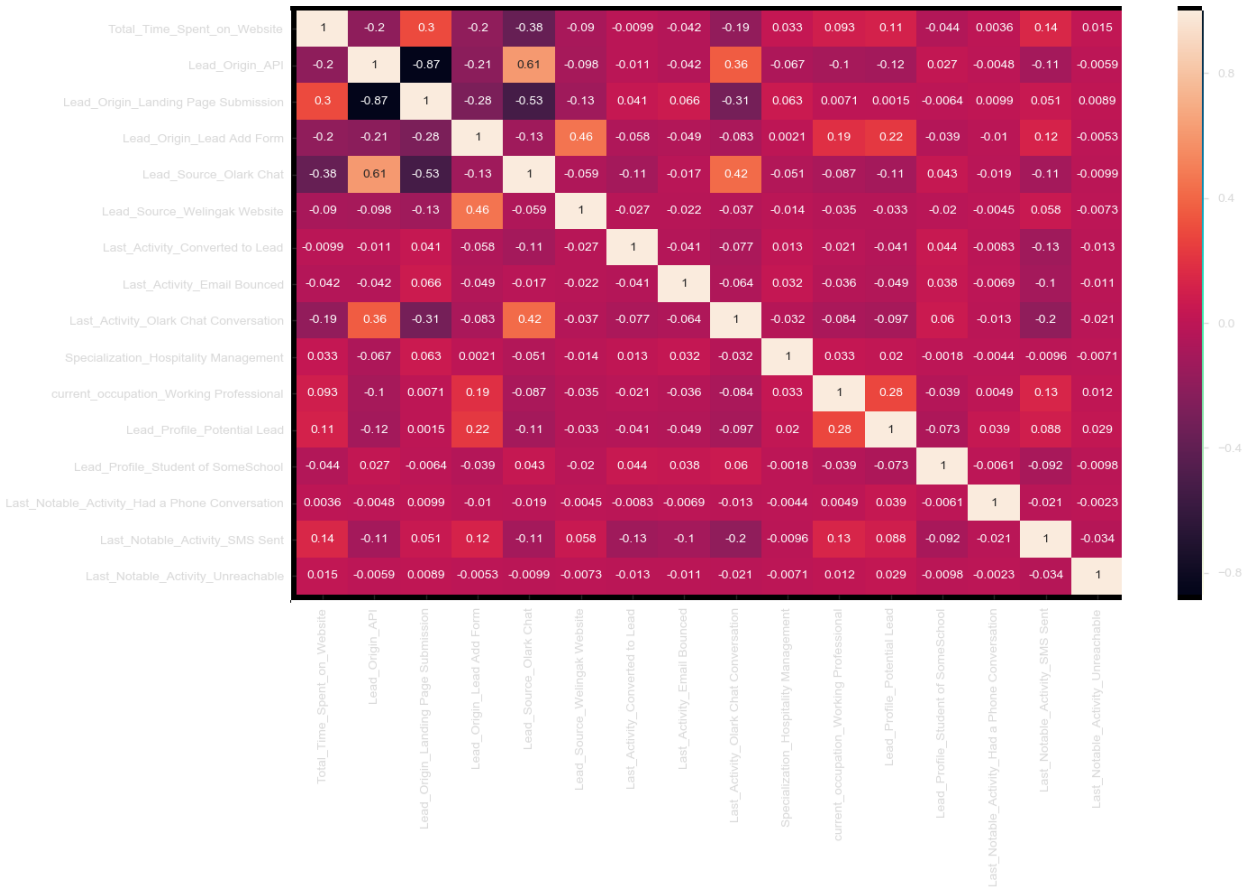
Dep. Variable:	Converted	No. Observations:	6352
Model:	GLM	Df Residuals:	6332
Model Family:	Binomial	Df Model:	19
Link Function:	logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-2467.1
Date:	Sun, 01 Mar 2020	Deviance:	4934.3
Time:	05:46:24	Pearson chi2:	6.60e+03
No. Iterations:	23		
Covariance Type:	nonrobust		

The fourth model is of extending VIFs.

Dep. Variable:	Converted	No. Observations:	6352
Model:	GLM	Df Residuals:	6335
Model Family:	Binomial	Df Model:	16
Link Function:	logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-2495.2
Date:	Sun, 01 Mar 2020	Deviance:	4990.5
Time:	05:46:24	Pearson chi2:	6.67e+03
No. Iterations:	7		
Covariance Type:	nonrobust		

Finally, the VIFs are assessed for proper conclusion and clearance.

Hence, we can now analyse the correlation matrix for the corresponding variables.

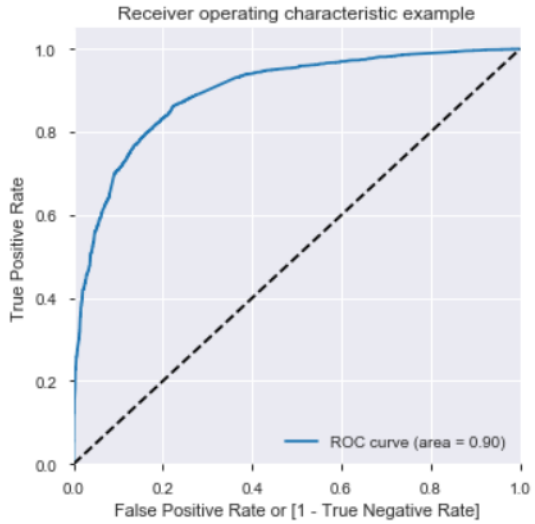


The following predictions are made from the above results:

- All variables have p-value < 0.05.
- All features are with very low values, meaning, there is hardly any multi-collinearity among the features.
- The overall accuracy of 0.8013 at a probability threshold of 0.05 is also very acceptable.

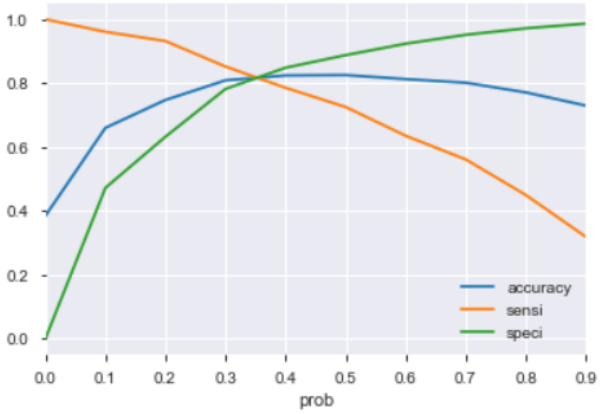
PLOTTING THE ROC CURVE

The ROC curve is being plotted for the following data and we get the following depictions:



The ROC is above 90%, which indicates a good response.

To find the optimal cutoff point, we plot the following data and obtain the following:



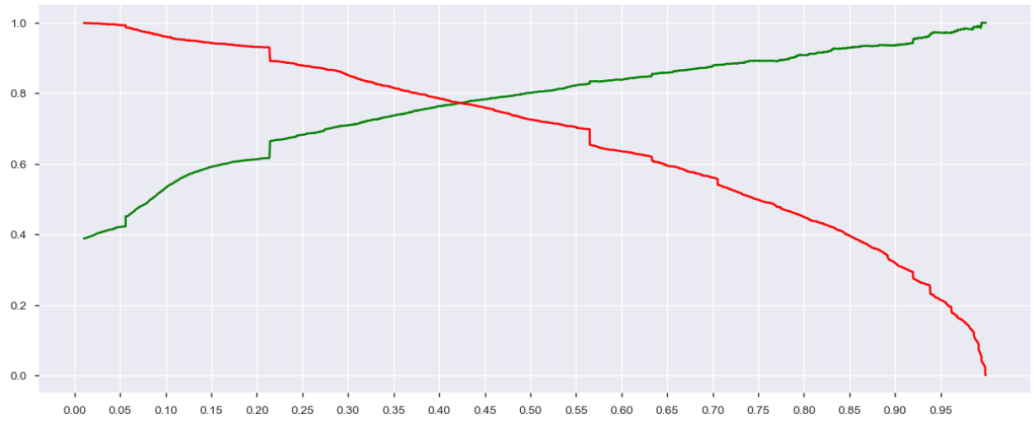
From the curve above, 0.35 is the optimum point to take it as a cutoff probability

The area under the curve is found to be 90. Hence, we can conclude that:

- Sensivity is around 82 while specifiy is around 82 that looks good.
- False positive is 18 and false negative is around 12.
- Positive predective is around 74 and negative predictive is around 88.

PRECISION AND RECALL

Now, we plot the precision and recall graph,

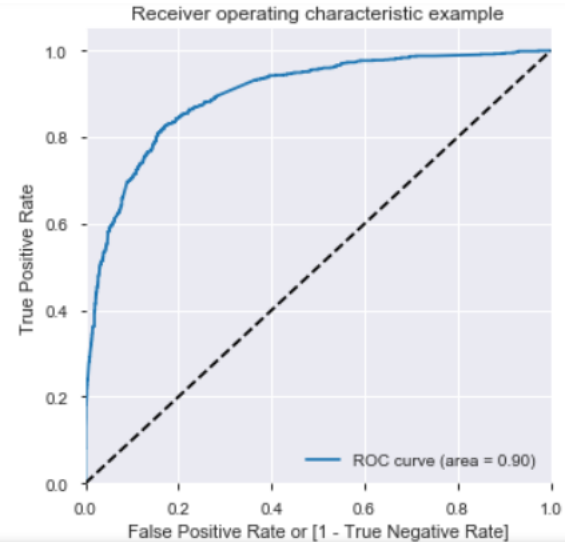


From the precision-recall graph above, we get the optical threshold value as close to .45. The X-Education requirement here is to have Lead Conversion Rate around 80%.

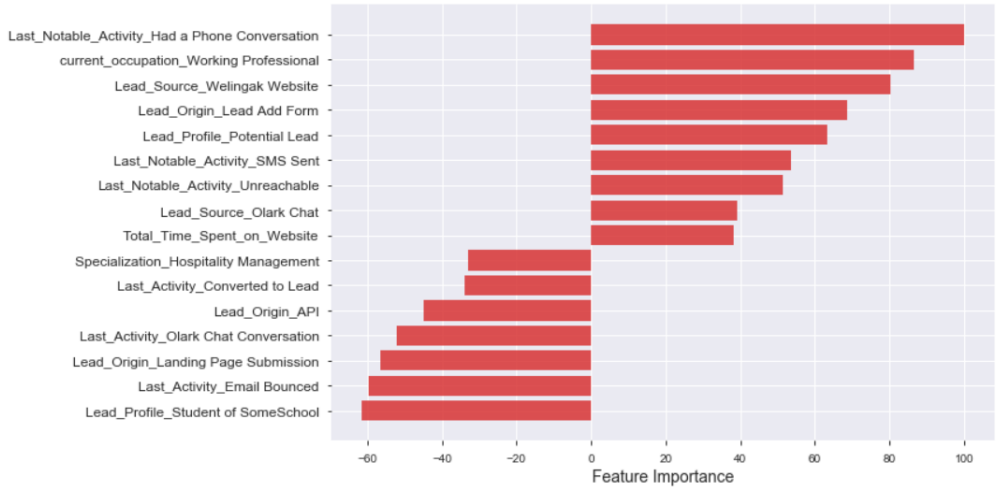
MAKING PREDICTIONS ON THE TEST SET

Since the data is already standardized and normalized, we will directly make predications on the test test.

Now, we plot the ROC curve for the following:



As seen above the GINI ratio is 90 which is imilar to the training dataset, so our model seems to be doing good on the test dataset.



CONCLUSION

After all models, we finally chose a model with the following characteristics:

- Sensivity is around 82 while specifiy is around 82 that looks good
- False positive is 18 and false negative is around 12
- Positive predective is around 74 and negative predictive is around 88

And the following statistics were calculated to predict the dependent variable

The following shows the different cutoff points with related accuracy, sensitivity and specificity

Following features contribute most to the model

- 'Total_Time_Spent_on_Website',
- 'Lead_Origin_API',
- 'Lead_Origin_Landing Page Submission',
- 'Lead_Origin_Lead Add Form',
- 'Lead_Source_Olark Chat',
- 'Lead_Source_Welingak Website',
- 'Last_Activity_Converted to Lead',
- 'Last_Activity_Email Bounced',
- 'Last_Activity_Olark Chat Conversation',
- 'Specialization_Hospitality Management',
- 'current_occupation_Working Professional',
- 'Lead_Profile_Potential Lead',
- 'Lead_Profile_Student of SomeSchool',
- 'Last_Notable_Activity_Had a Phone Conversation',
- 'Last_Notable_Activity_SMS Sent',
- 'Last_Notable_Activity_Unreachable']

When following features value increases, the conversion probablity of a lead will also increase

1. Last_Notable_Activity_Had a Phone Conversation
2. current_occupation_Working Professional

3. Lead_Source_Welingak Website
4. Lead_Origin_Lead Add Form
5. Lead_Profile_Potential Lead
6. Last_Notable_Activity_SMS Sent
7. Last_Notable_Activity_Unreachable
8. Lead_Source_Olark Chat
9. Total_Time_Spent_on_Website

When following features values decreases the conversion probablity of a lead will also increase

1. Lead_Profile_Student of SomeSchool
2. Last_Activity_Email Bounced
3. Lead_Origin_Landing Page Submission
4. Last_Activity_Olark Chat Conversation
5. Lead_Origin_API
6. Last_Activity_Converted to Lead
7. Specialization_Hospitality Management

Note that some of the above is not meanigful and contradictory. Since the time was limited we stopped the above effort.

Also note that depending on the business requirement, decreasing or increasing the Sensitivity and increase or decrease the Specificity of the model is dependent on the increase or decrease the probability threshold value.

High sensitivity will result more accurate lead while low sensitivity will result in less accurate lead.