Case Study – Lead Score

A summary Report

Group Name	**Mind Benders**
Team Member 1	**Rahul Kumar**
Team Member 2	**Adithya Chadalawada**
Course	**PGDDS**
Batch	**Dec 2018**

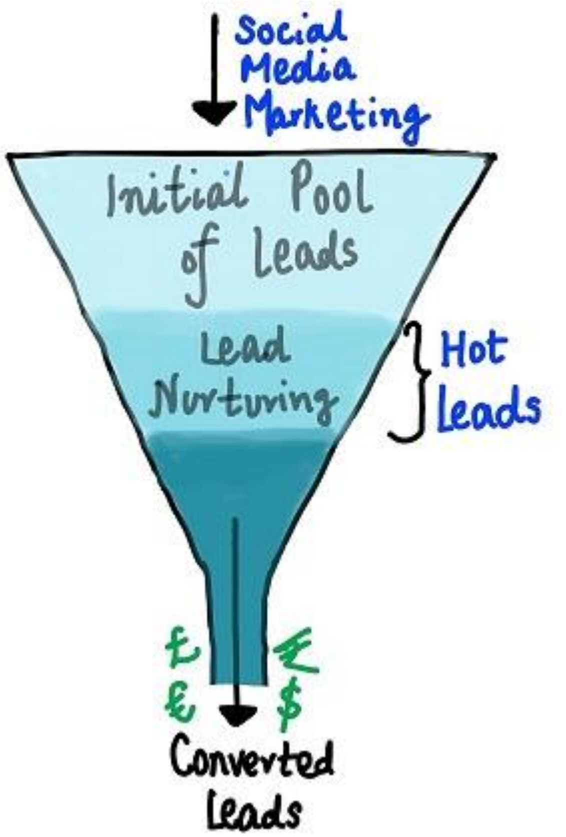
Contents

Introduction

An education company named X Education sells online courses to industry professionals. On any given day, many professionals who are interested in the courses land on their website and browse for courses.

The company markets its courses on several websites and search engines like Google. Once these people land on the website, they might browse the courses or fill up a form for the course or watch some videos. When these people fill up a form providing their email address or phone number, they are classified to be a lead. Moreover, the company also gets leads through past referrals. Once these leads are acquired, employees from the sales team start making calls, writing emails, etc. Through this process, some of the leads get converted while most do not. The typical lead conversion rate at X education is around 30%.

Now, although X Education gets a lot of leads, its lead conversion rate is very poor. For example, if, say, they acquire 100 leads in a day, only about 30 of them are converted. To make this process more efficient, the company wishes to identify the most potential leads, also known as ‘Hot Leads’. If they 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 we 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, we 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.

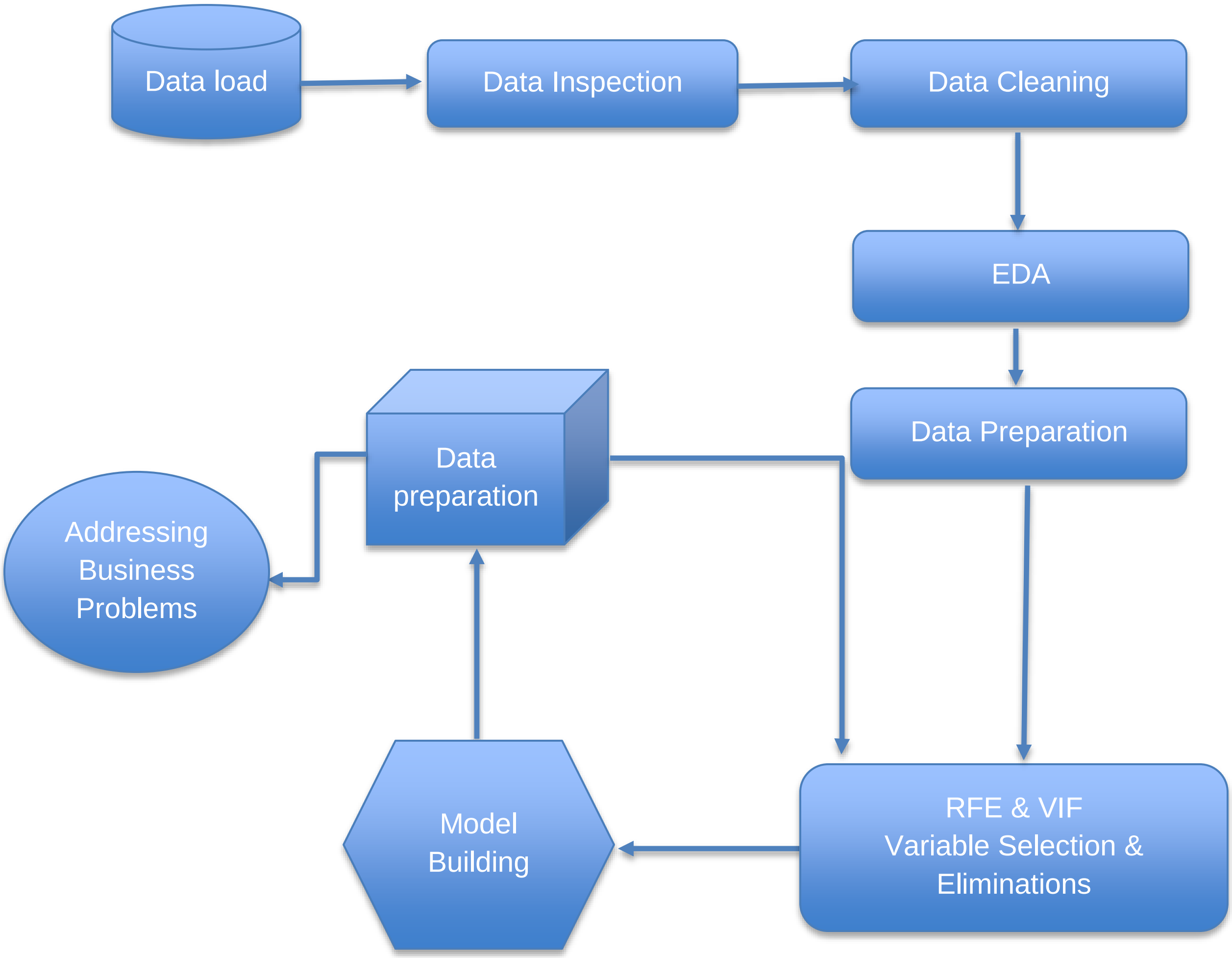
We need to work to select the most promising leads, i.e. the leads that are most likely to convert into paying customers. The company requires us to build a model wherein we need to assign a lead score to each of the leads such that the customers with higher lead score have a higher conversion chance and the customers with lower lead score have a lower conversion chance. The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%.

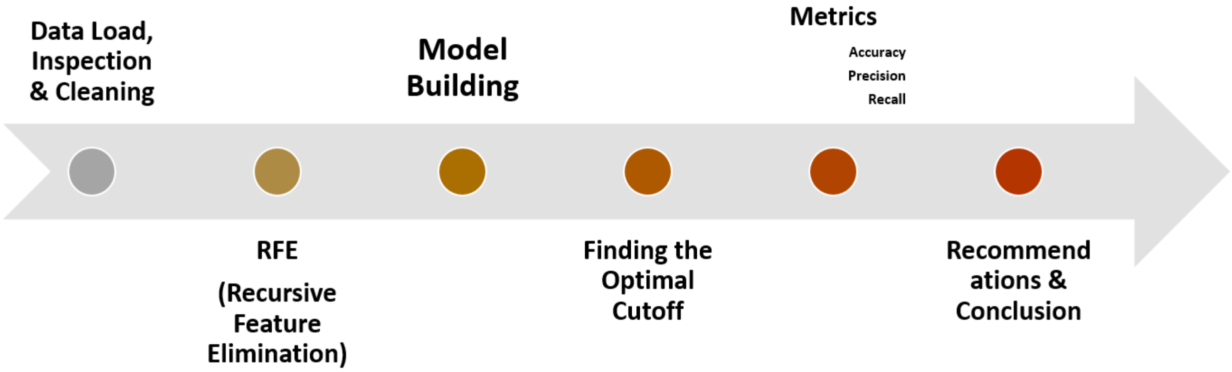
Goals of the Case Study

There are quite a few goals for this case study.
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. 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.

There are some more problems presented by the company which our 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.

Solution Approach




Data Load, Inspection & Cleaning
RFE (Recursive Feature Elimination)
Model Building
Finding the Optimal Cutoff
Metrics
Accuracy
Precision
Recall
Recommendations & Conclusion

Data Load

The leads data was provided in a CSV file, we loaded into a dataframe and named as 'leads_data'. In original dataset, there are 37 features or fields and 9240 rows

Data Cleaning

While inspecting the dataset, we found lots of data inconsistency like missing values, invalid text, feature with only 1 value etc.

Invalid text

- All the "Select" values in the data were replaced with Null values.

NaN values

- Columns having high percentage (more than 30%) of null values were dropped, since they did not contain adequate information for Analysis.

Single unique value

- Columns having a single unique value were also dropped.

Categorical variables

- Categorical variables having two unique values were mapped to 1 or 0.

Data inconsistency

- Some columns, having two unique values, had around 99% of same values. They were also dropped.

Impute

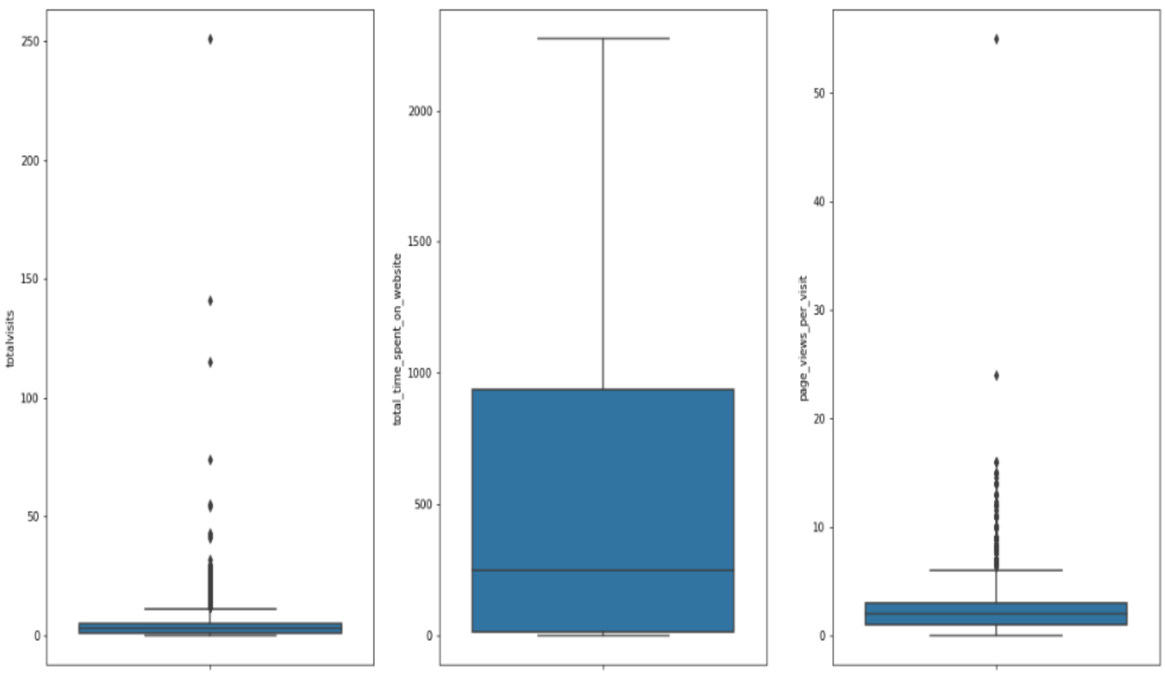
- Missing values of a few columns, which were important from the business perspective, were imputed using statistical measures such as Mean, Median and Mode.

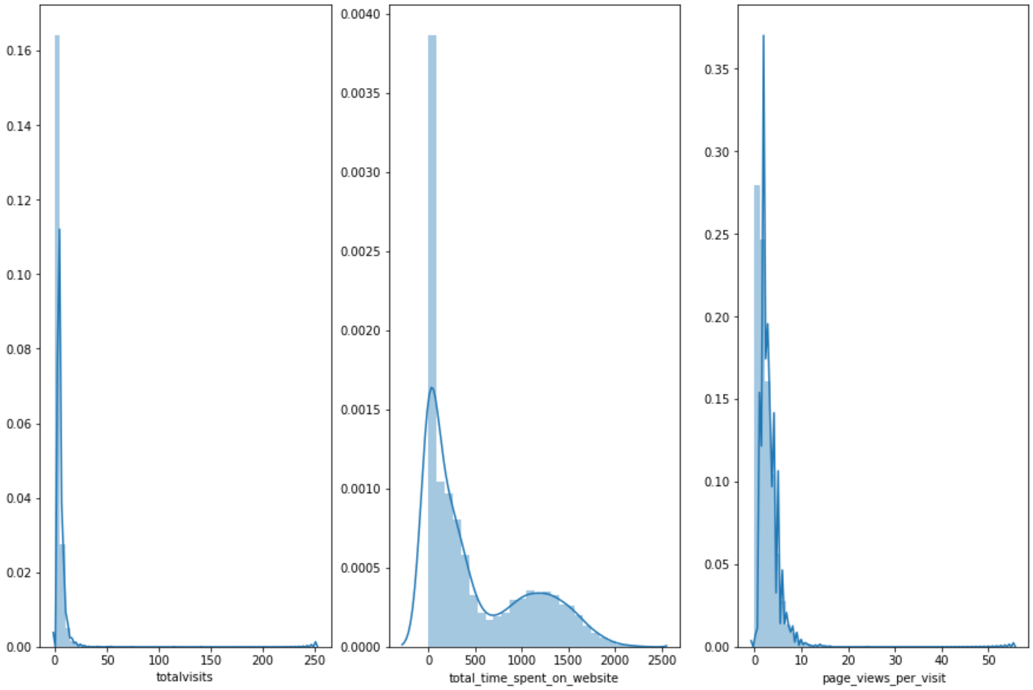
Duplicate values

- Columns 'last_activity' and 'last_notable_activity' had duplicated values. Hence one of them was dropped

Outlier Treatment

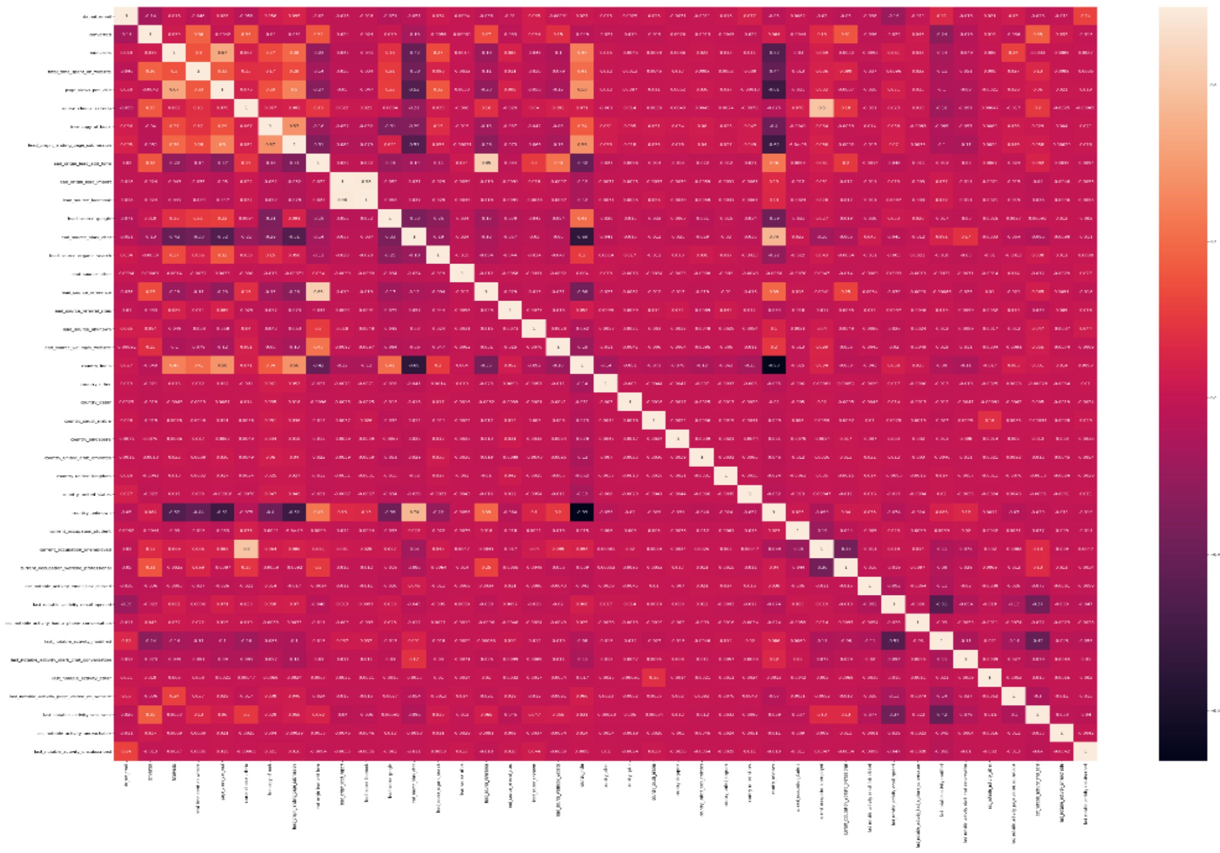
Some variables (such as TotalVisits, Total Time Spent on Website, Page Views Per Visit) had outliers (above fig) After doing some analysis, these outliers were treated by capping them to a certain value such as:
page_views_per_visit: cap to 16, meaning if the value is 16 it would mean that the page views per visit are 16 or more
totalvisits: cap to 30, meaning if the value is 30 it would mean that the total visits are 30 or more





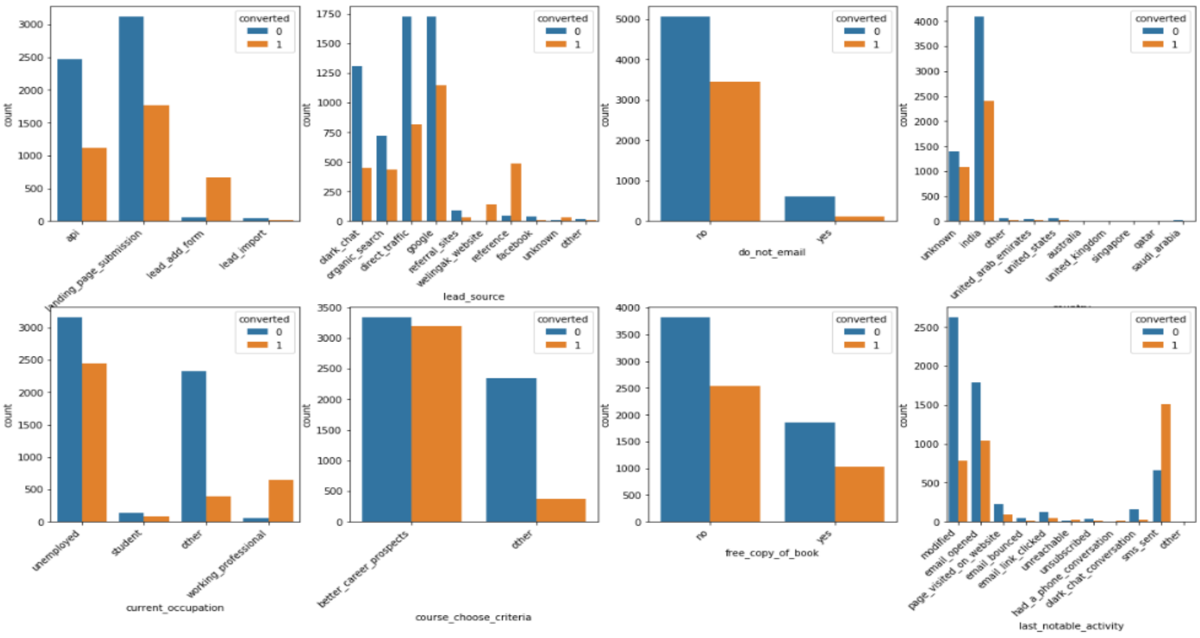
High correlation

- Variables with High correlation were observed using a Heat Map and the correlation matrix and dropped



Exploratory Data Analytics (EDA)

EDA is used to get a better understanding of the variables and to prepare the data so that it could be used for Logistic Regression Modelling.



Recursive Feature Elimination (RFE)

RFE was used to eliminate features or variables and to select the most important features for model building.

List of the features Selected

do_not_email
total_time_spent_on_website
course_choose_criteria
lead_origin_lead_add_form
lead_source_welingak_website
country_Qatar
country_unknown
current_occupation_working_professional
last_notable_activity_had_a_phone_conversation
last_notable_activity_other
last_notable_activity_sms_sent last_notable_activity_unreachable

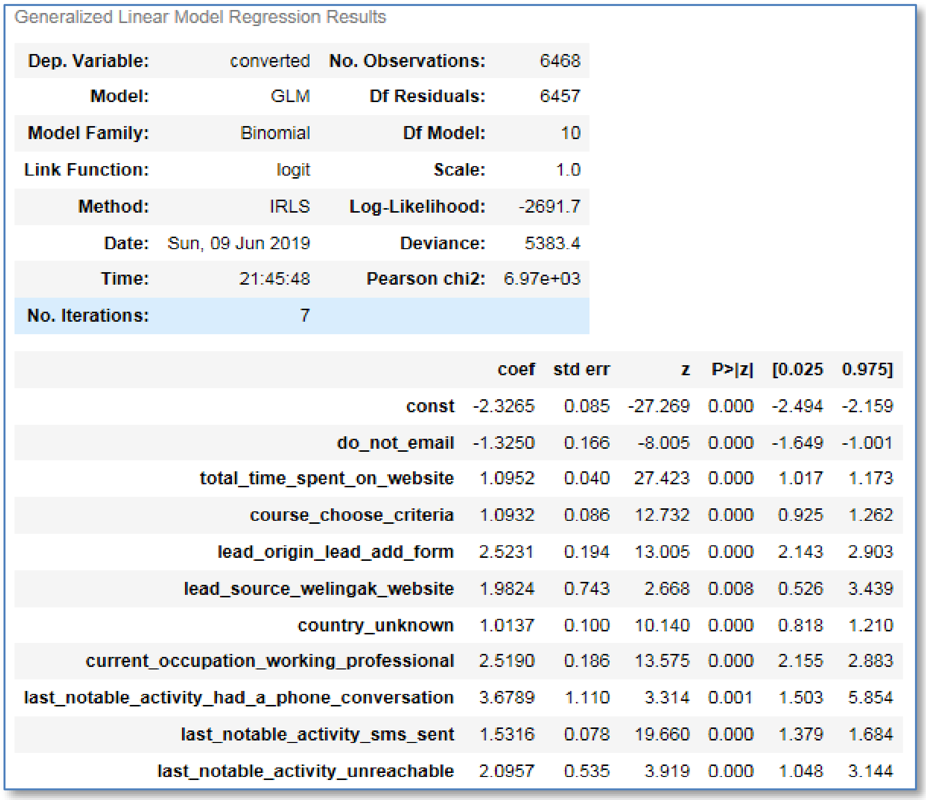
List of columns Eliminated by RFE

Totalvisits

page_views_per_visit
free_copy_of_book
lead_origin_landing_page_submission
lead_origin_lead_import
lead_source_google
lead_source_organic_search
lead_source_other
lead_source_referral_sites
lead_source_unknown
country_india
country_other
country_saudi_arabia
country_Singapore
country_united_arab_emirates
country_united_kingdom
country_united_states
current_occupation_student
last_notable_activity_email_link_clicked
last_notable_activity_email_opened
last_notable_activity_modified
last_notable_activity_olark_chat_conversation
last_notable_activity_page_visited_on_website
last_notable_activity_unsubscribed

Model Building

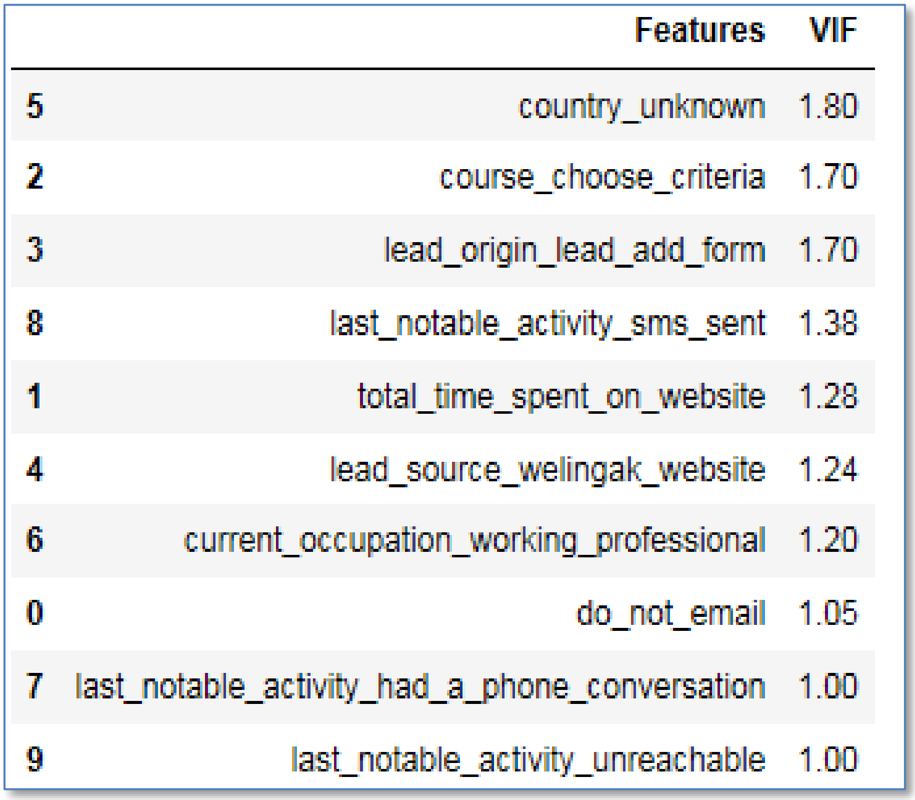
After RFE, we manually eliminated features which had High P-Value (greater than 0.05) and High VIF values (greater than 5), until we reached a model which had all features with P-value less than 0.05 and VIFs less than 5.
This was done to eliminate multi-collinearity amongst the features. The final list of 10 features in the model.



Generalized Linear Model Regression Results

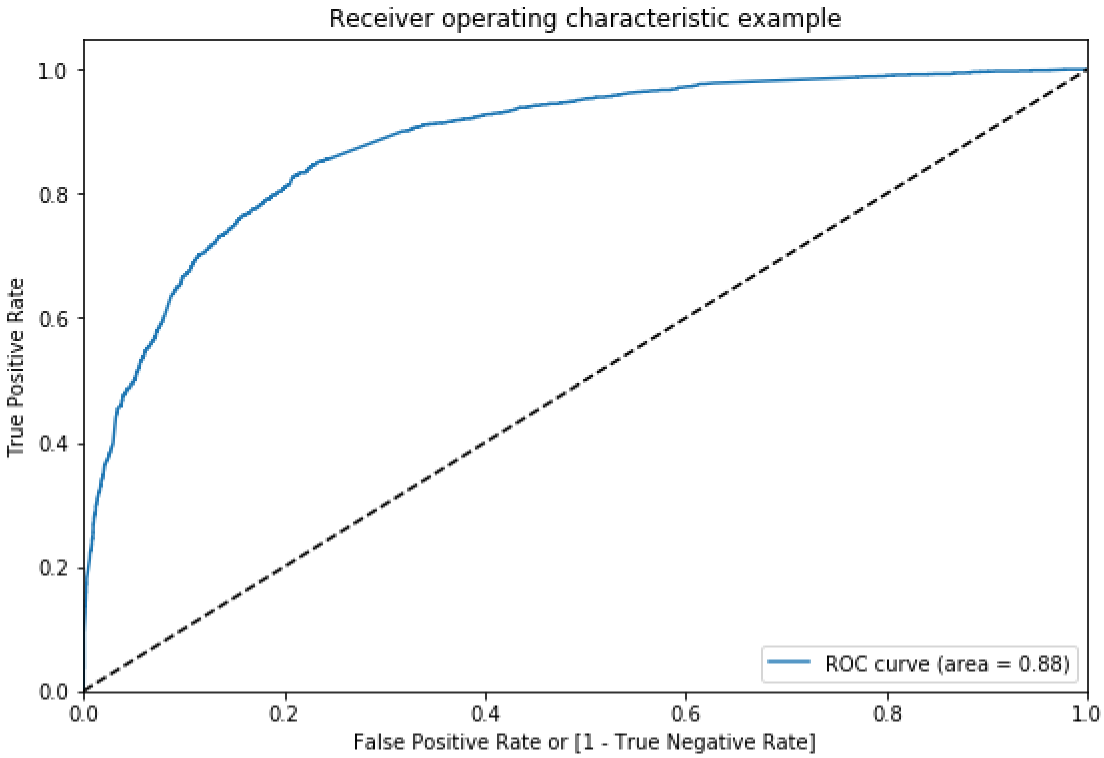
Dep. Variable:	converted	No. Observations:	6468
Model:	GLM	Df Residuals:	6457
Model Family:	Binomial	Df Model:	10
Link Function:	logit	Scale:	1.0
Method:	IRLS	Log-Likelihood:	-2691.7
Date:	Sun, 09 Jun 2019	Deviance:	5383.4
Time:	21:45:48	Pearson chi2:	6.97e+03
No. Iterations:	7		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-2.3265	0.085	-27.269	0.000	-2.494	-2.159
do_not_email	-1.3250	0.166	-8.005	0.000	-1.649	-1.001
total_time_spent_on_website	1.0952	0.040	27.423	0.000	1.017	1.173
course_choose_criteria	1.0932	0.086	12.732	0.000	0.925	1.262
lead_origin_lead_add_form	2.5231	0.194	13.005	0.000	2.143	2.903
lead_source_welingak_website	1.9824	0.743	2.668	0.008	0.526	3.439
country_unknown	1.0137	0.100	10.140	0.000	0.818	1.210
current_occupation_working_professional	2.5190	0.186	13.575	0.000	2.155	2.883
last_notable_activity_had_a_phone_conversation	3.6789	1.110	3.314	0.001	1.503	5.854
last_notable_activity_sms_sent	1.5316	0.078	19.660	0.000	1.379	1.684
last_notable_activity_unreachable	2.0957	0.535	3.919	0.000	1.048	3.144



	Features	VIF
5	country_unknown	1.80
2	course_choose_criteria	1.70
3	lead_origin_lead_add_form	1.70
8	last_notable_activity_sms_sent	1.38
1	total_time_spent_on_website	1.28
4	lead_source_welingak_website	1.24
6	current_occupation_working_professional	1.20
0	do_not_email	1.05
7	last_notable_activity_had_a_phone_conversation	1.00
9	last_notable_activity_unreachable	1.00

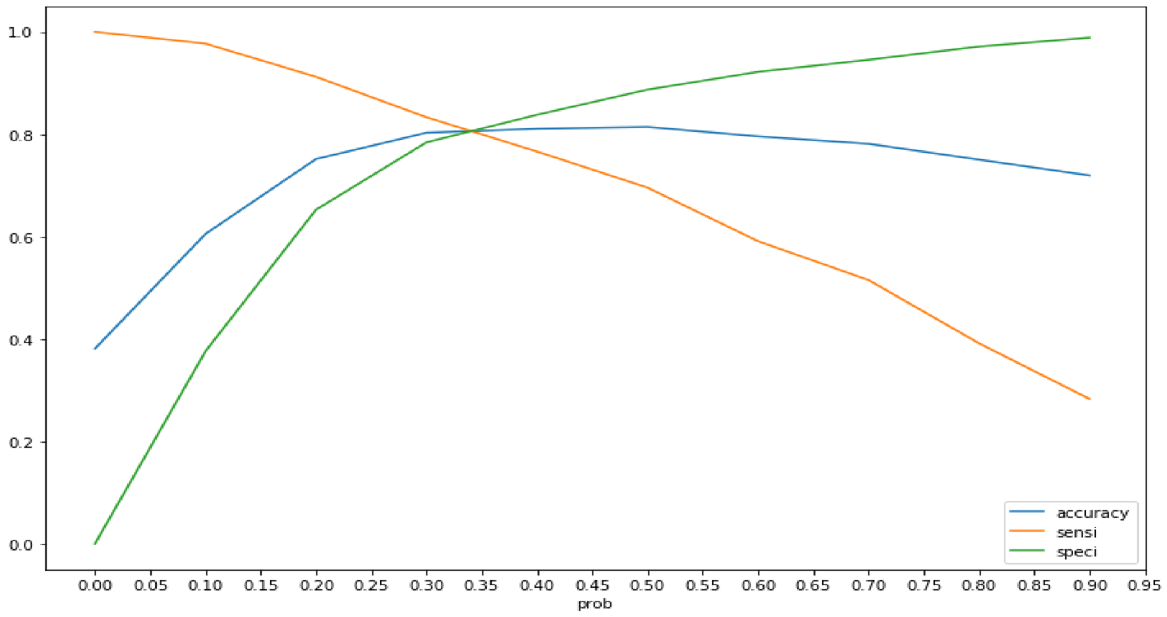
After removing all features having P-value more than 0.05 and VIF values more than 5, and re-building the model few times, we arrived at a decent Logistic Regression model, who's ROC curve and metrics are as below



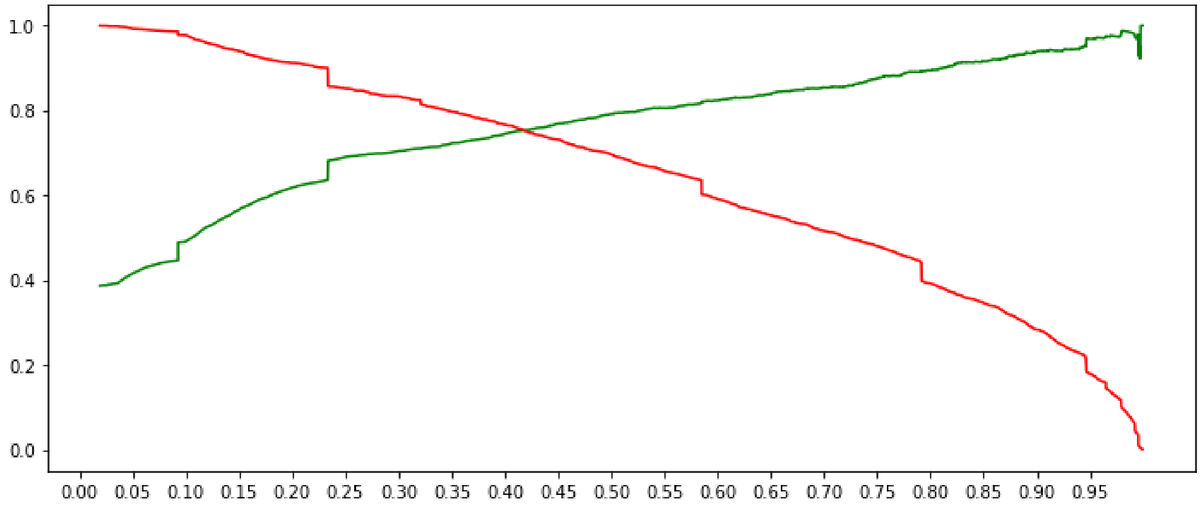
Sensitivity of Model = 0.6958637469586375
Specificity = 0.8873063468265867
False Positive Rate = 0.11269365317341329
Positive Predictive Value = 0.7918781725888325
Negative Predictive Value = 0.8256219483840967

Finding the Optimal Cutoff

From the accuracy-sensitivity-specificity plot, we observed that: The accuracy is at peak and remains constant between 0.2 and 0.53
All the three metrics converge at **0.35**



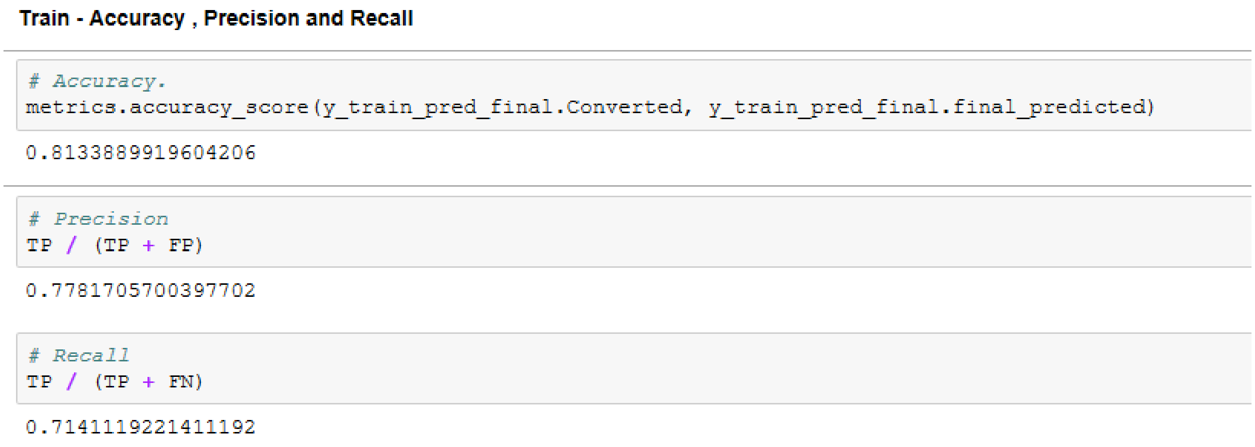
From the Precision vs Recall plot, we observed that: the cutoff is around approx. **0.425** (between 0.4 and 0.45)
Considering both the aspects, we chose the cut-off as 0.47 and use the Precision-Recall-Accuracy metrics to evaluate our model.



Metrics on the Train Set

The results of our Logistic Regression Model on the **Train Set** is:
About 81% Accurate (Accuracy)

About 78% Precise (Precision)
About 72% Recall Rate



Train - Accuracy , Precision and Recall

```
# Accuracy.
metrics.accuracy_score(y_train_pred_final.Converted, y_train_pred_final.final_predicted)
```

```
0.8133889919604206
```

```
# Precision
TP / (TP + FP)
```

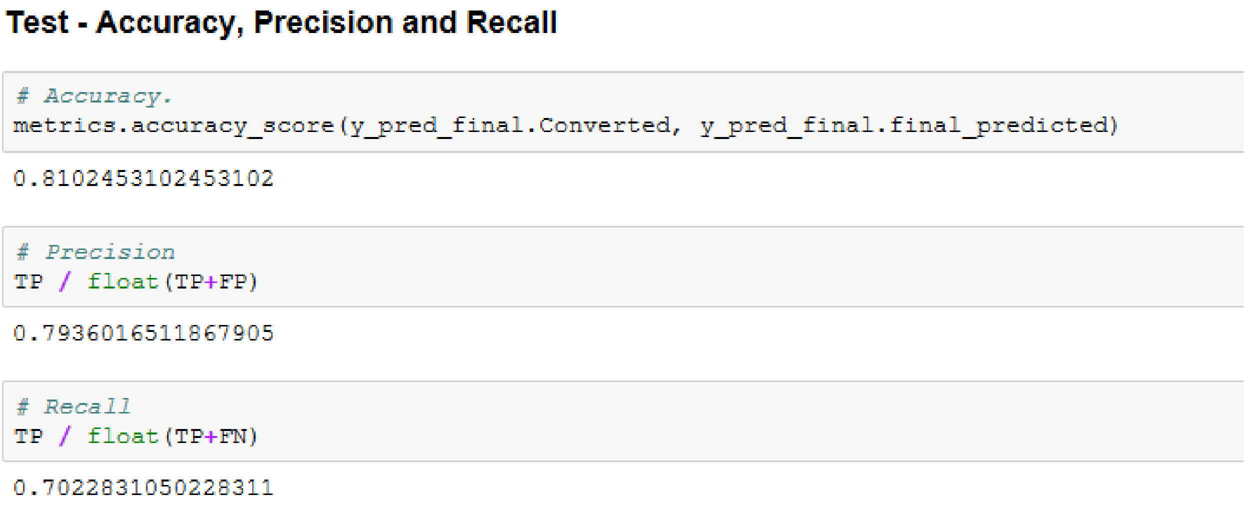
```
0.7781705700397702
```

```
# Recall
TP / (TP + FN)
```

```
0.7141119221411192
```

Metrics on the Test Set

The results of our Logistic Regression Model on the **Test Set** is:
About 81% Accurate (Accuracy)
About 79% Precise (Precision)
About 70% Recall Rate



Test - Accuracy, Precision and Recall

```
# Accuracy.
metrics.accuracy_score(y_pred_final.Converted, y_pred_final.final_predicted)
```

```
0.8102453102453102
```

```
# Precision
TP / float(TP+FP)
```

```
0.7936016511867905
```

```
# Recall
TP / float(TP+FN)
```

```
0.7022831050228311
```

Recommendations & Conclusions

The organization should focus on Leads having a Lead-Score of 47 and above, they should be considered as Hot-Leads, since they are very likely to convert.
Resources, Time and Effort should not be wasted by focusing on Leads having a Lead-Score of below 47. They can be considered as Cold-Leads and should be avoided, since they are very less likely to convert.
Leads having "Last Notable Activity" as "Had a phone Conversation" **OR** "Current Occupation" as "Working Professional" **OR** "Lead Origin" as "Lead Add Form" (and LeadScore of more than 47) should have the most focus on and pursued extensively, since these can be categorized as "Very Hot Leads" and have very high chances of conversion. To prove the statement is correct, these are the Top-3 Predictor variables for our Model.
Also, Leads that have "Yes" for "Do Not Email" (and LeadScore less than 47) should NOT be pursued or resources should not be wasted on them since they can be categorized as "Very Cold Leads" and are least likely to convert.

After evaluating our Model based on the Accuracy, Precision and Recall values, we can safely conclude that the model would help X Education to identify the leads that are most likely to convert into paying customers.
Since the model has an Accuracy and Precision of about 80%, it would also help meet the CEO's ballpark target of lead conversion rate to be around 80%.
Also, the model built is adjustable and if the company's requirement changes in the future,
We can do the following:-
1. When there are more people to contact the leads and try to convert then we can lower the cutoff to get more projected leads.
2. When the target has been met, we can increase the cut-off to ensure that we get only few projected leads which are having a very high probability of conversion

Learnings / Take Away

Data Cleaning was very major part in this case study
Identification of variable with having inconsistent values and check the possible solution for impute with mean, median or mode
Which one to use RFE or PCA?
We decided to use RFE as it will be easily interpretable & also can answer what business education firm wants.
Feature scaling and Feature elimination
Model Building

Training & testing data set
Finally saw how the result set varies, if company needs to parametrized the input.
And so on…

-------------------------------- End of File ---------------------------------