LEAD SCORING CASE STUDY

SUBMISSION

Group Members:

1. Arpit Joshi
2. Harshit Tiwari
3. Vinay Dharwadkar
4. Bhaswati Paul

Problem Statement

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%.

Business Objectives:

X Education needs a model to indentify the leads that are most likely to convert into paying customers by assigning 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.

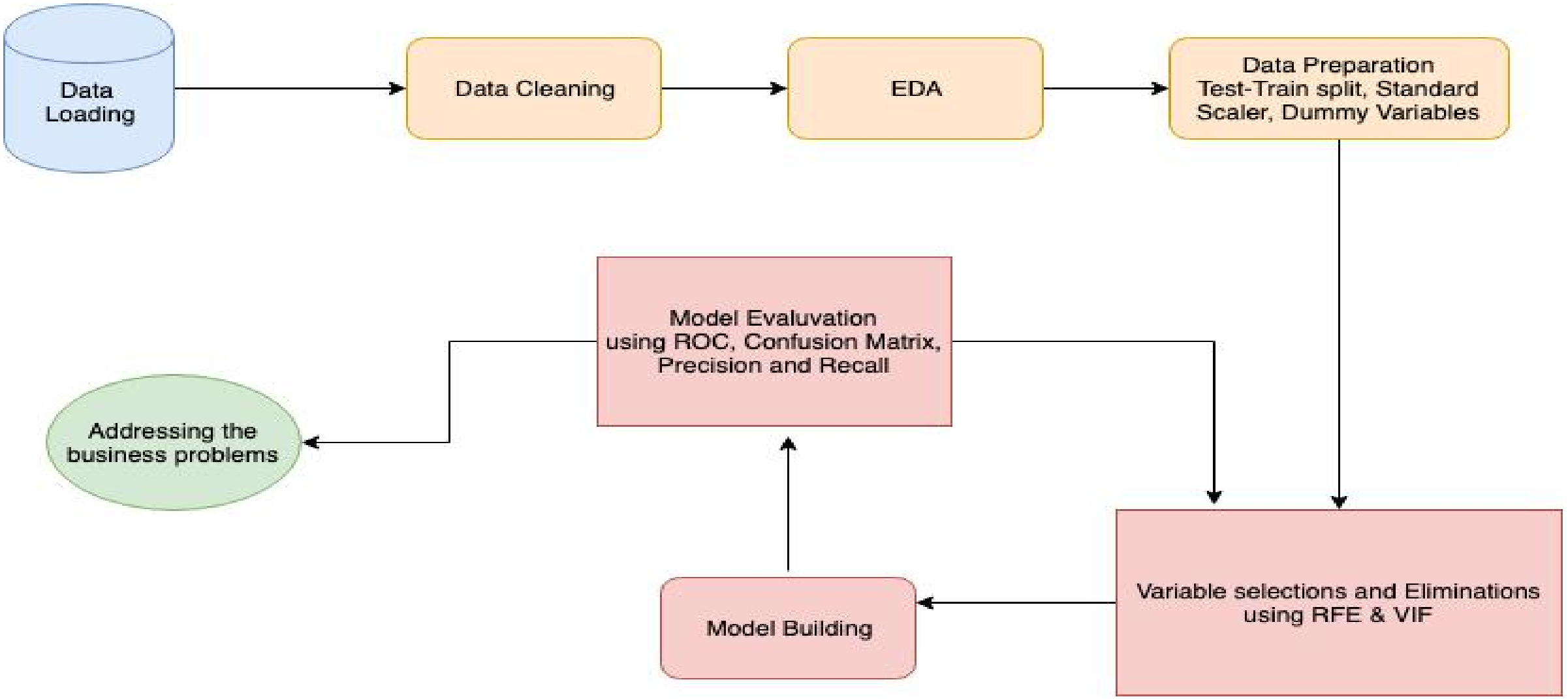
The CEO has ballpark of the target lead conversion rate to be around 80%.

Case-Study Goals:

To 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.

An adjustable model such that some more problems presented by the company (if the company's requirement changes in the future), the model is able to adapt to it.

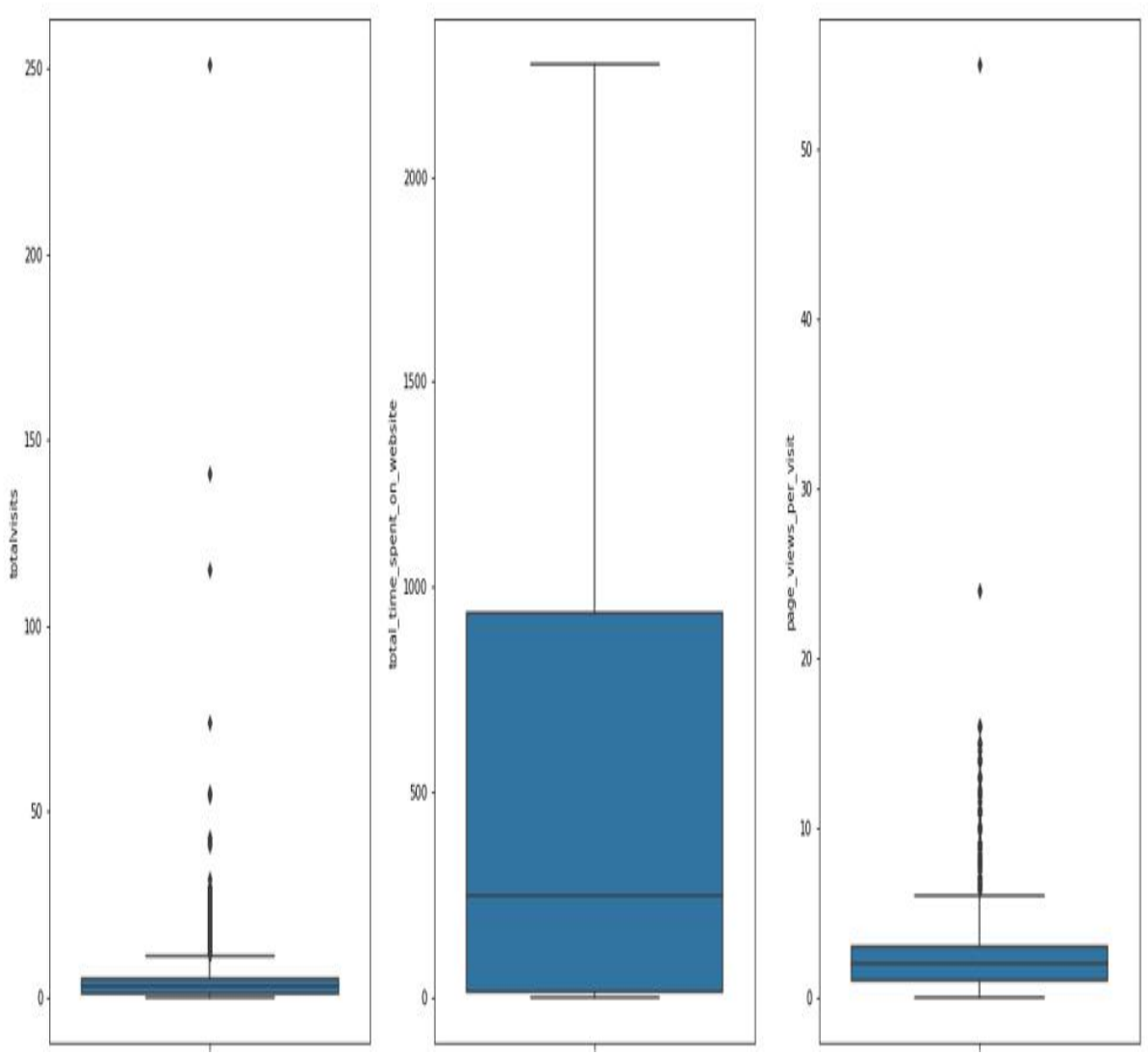
Flow chart of the methodology followed

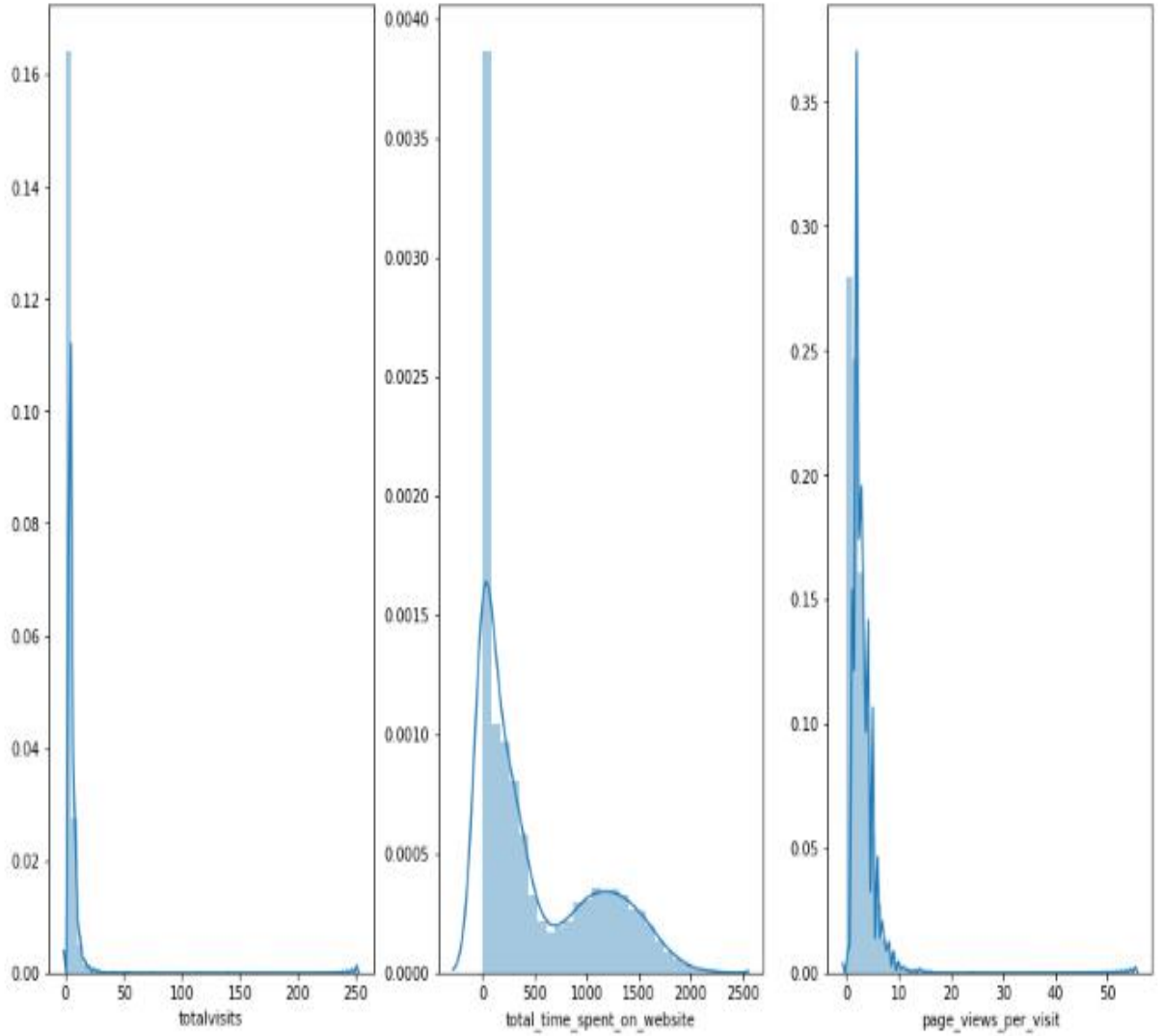


Data Cleaning

- All the “Select” values in the data were replaced with Null values.
- All columns having high percentage (more than 30%) of null values were dropped, since they did not contain adequate information for Analysis.
- Columns having a single unique value were also dropped.
- Categorical variables having two unqiue values were mapped to 1 or 0.
- Some columns, having two unique values, had around 99% of same values. They were also dropped.
- Missing values of a few columns, which were important from the business perspective, were imputed using statistical measures such as Mean, Median and Mode.
- Columns 'last_activity' and 'last_notable_activity' had duplicated values. Hence one of them was dropped.

Data Cleaning - Outlier Treatment





Data Cleaning - Outlier Treatment

Some variables (such as TotalVisits, Total Time Spent on Website, Page Views Per Visit) had outliers (as depicted in the plots in the previous slide)

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

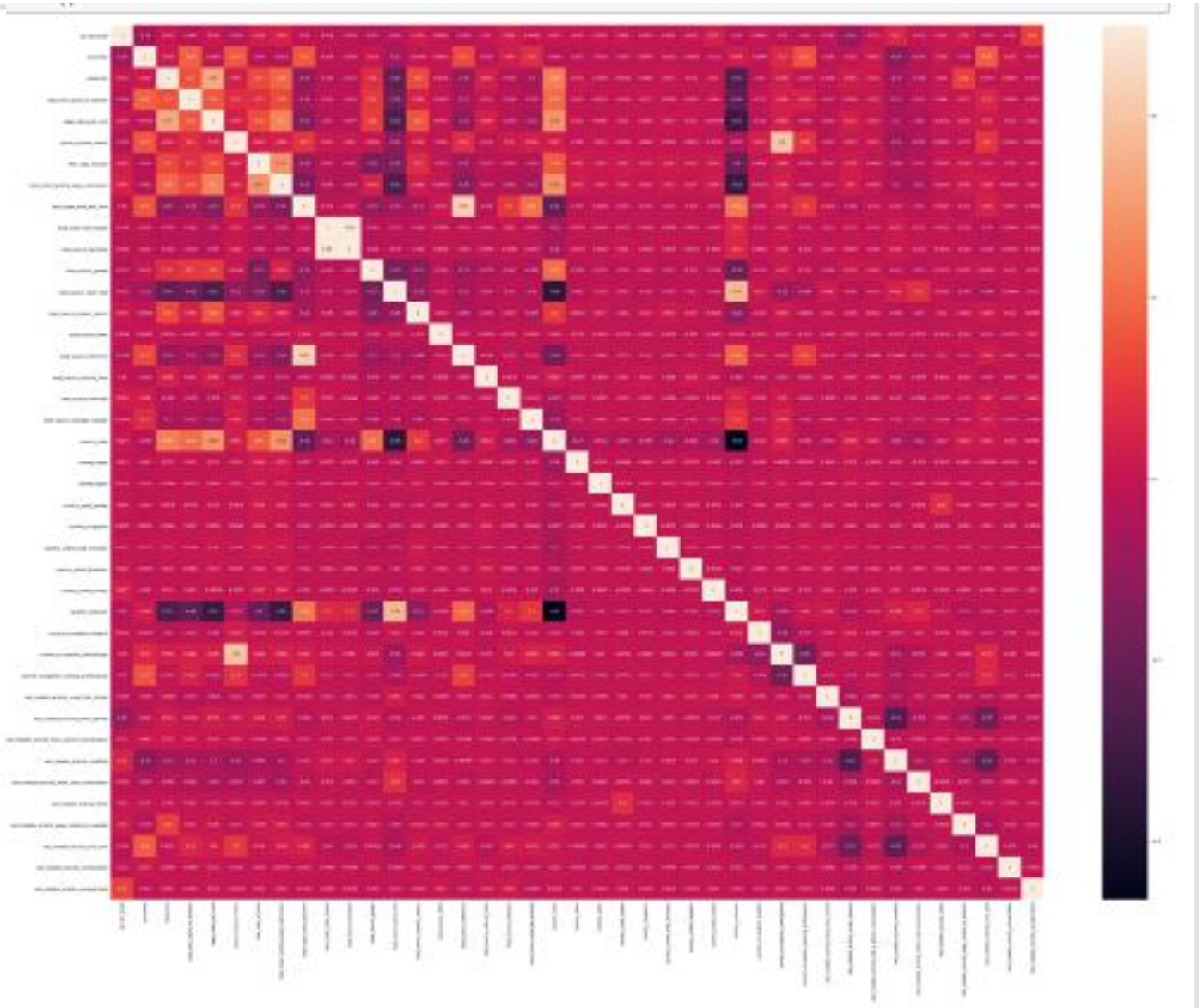
totalvisits: cap to 30, meaning if the value is 30 it would mean that the total visits are 30 or more.

Data Cleaning

	Var 1	Var 2	coeff
3	lead_origin_lead_import	lead_source_facebook	0.981709
5	lead_origin_lead_add_form	lead_source_reference	0.852594
6	country_unknown	lead_source_olark_chat	0.741415
7	course_choose_criteria	current_occupation_unemployed	0.798003

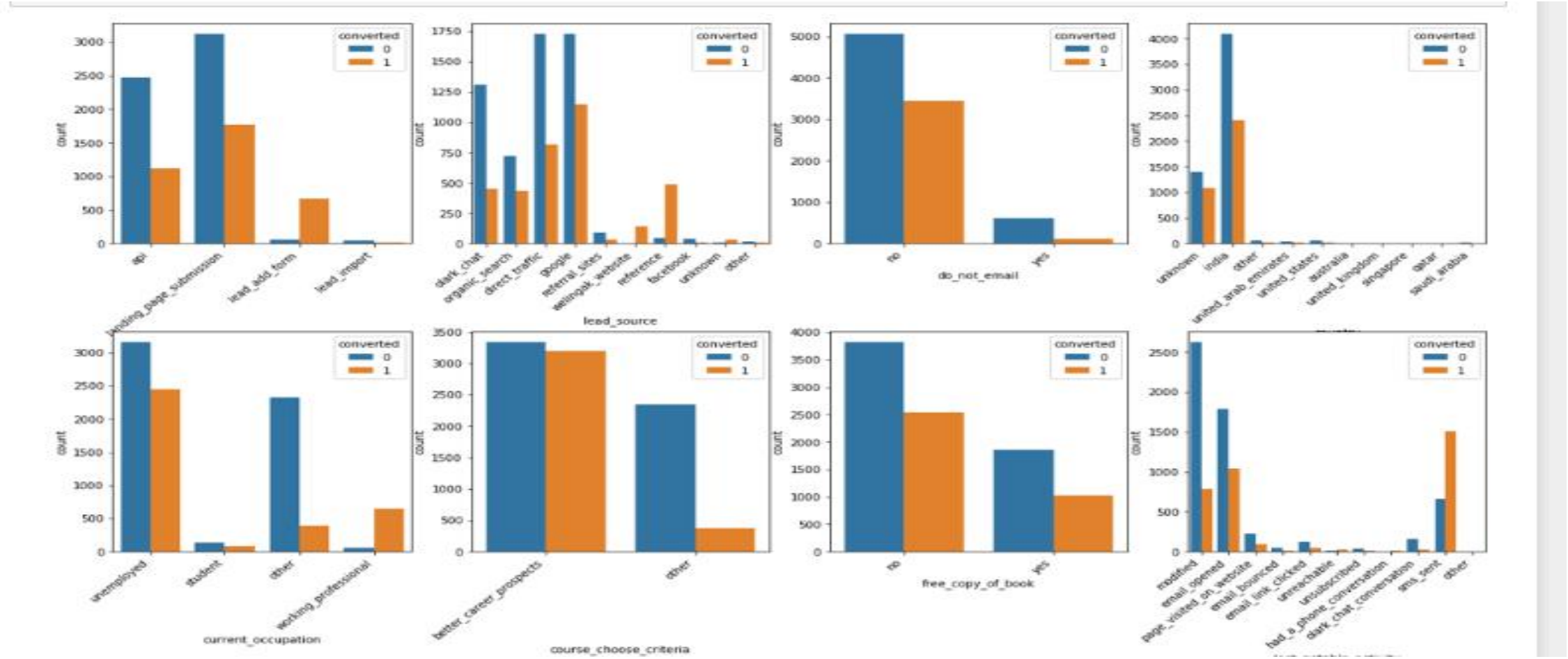
** 4 out the variables have very high correlation. Lets drop these. **

Variables with High correlation were observed using a Heat Map (as depicted on left) and the correlation matrix (as depicted above) and dropped.



Exploratory Data Analytics (EDA)

EDA (as depicted below an in graphs of previous slides) was 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. The list of the features Selected is depicted below, followed by the list of columns Eliminated by RFE.

```
In [135]: col = X_train.columns[rfe.support_]

In [136]: list(col)

Out[136]: ['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']

In [137]: X_train.columns[~rfe.support_]

Out[137]: Index(['totalvisits', 'page_views_per_visit', 'free_copy_of_book', 'lead_origin_landing_page_submission', 'lead_origin_lead_imp
          ort', 'lead_source_google', 'lead_source_organic_search', 'lead_source_other', 'lead_source_referral_sites', 'lead_source_unkno
          wn', 'country_india', 'country_other', 'country_saudi_arabia', 'country_singapore', 'country_united_arab_emirates', 'country_un
          ited_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_activ
          ity_page_visited_on_website', 'last_notable_activity_unsubscribed'], dtype='object')
```

Model Building

Manual Feature Elimination

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 (left) and VIFs (right) less than 5. (as depicted below). This was done to eliminate multi-collinearity amongst the features. The final list of 10 features in the model is as shown below (right).

Model Family:	Binomial	Df Model:	10
Link Function:	logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-2691.7
Date:	Sun, 03 Mar 2019	Deviance:	5383.4
Time:	17:31:43	Pearson chi2:	6.97e+03
No. Iterations:	7	Covariance Type:	nonrobust

	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

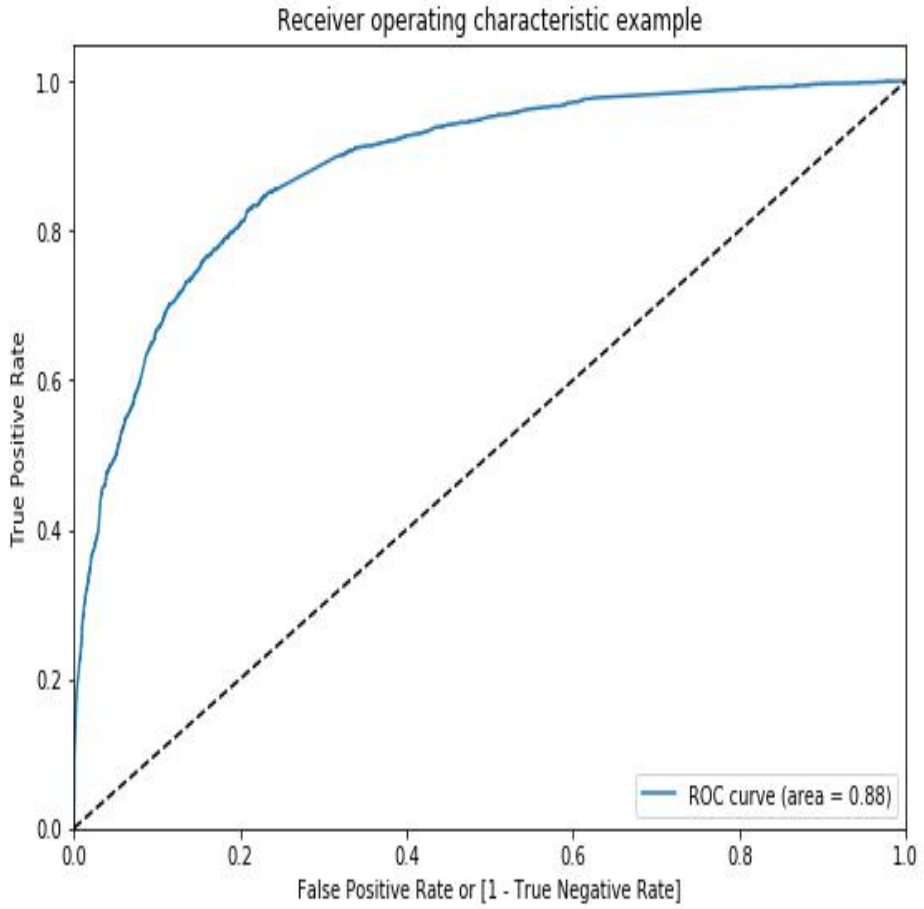
Out[158]:

	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

Model Building

Final Model

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 (left) and metrics (right) are depicted below.



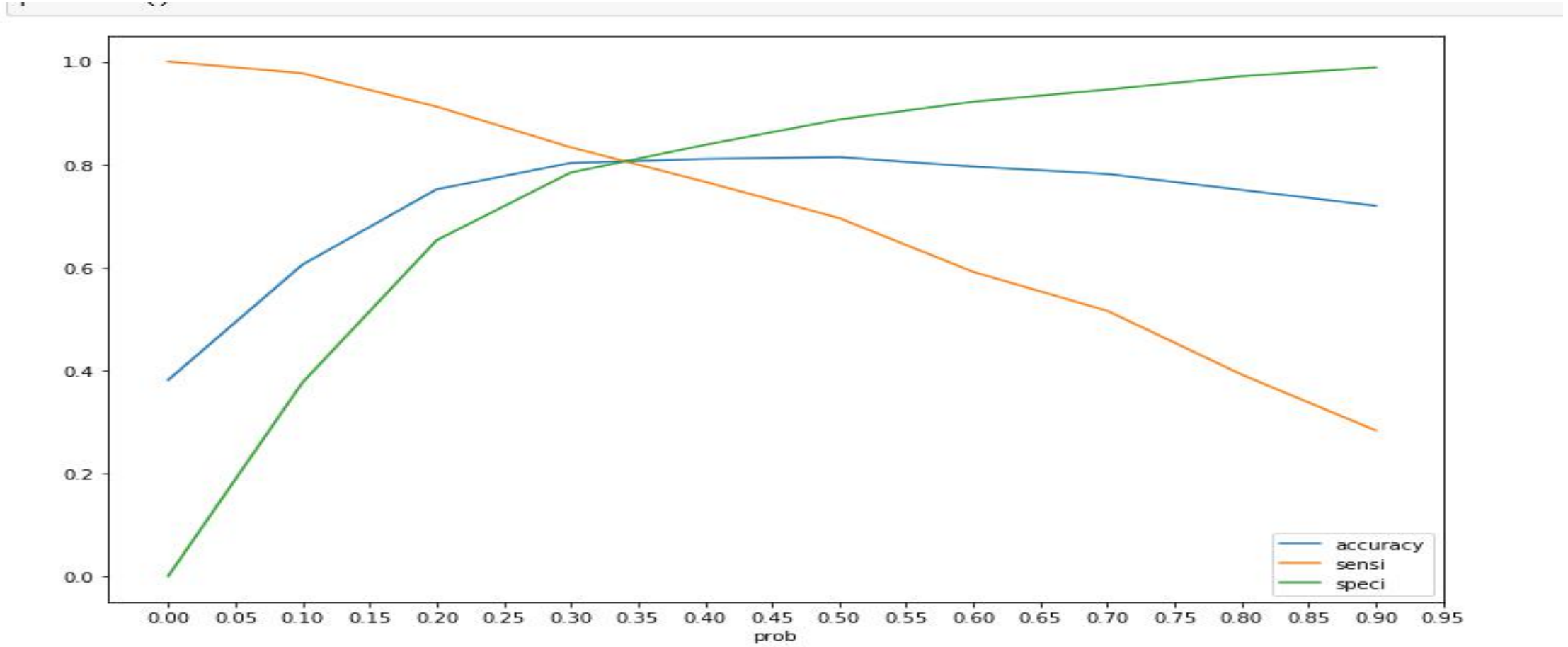
Metrics

```
In [168]: TP = confusion[1,1] # true positive
          TN = confusion[0,0] # true negatives
          FP = confusion[0,1] # false positives
          FN = confusion[1,0] # false negatives

In [169]: # Let's see the sensitivity of our logistic regression model
          TP / float(TP+FN)

Out[169]: 0.6958637469586375

In [170]: # Let us calculate specificity
          TN / float(TN+FP)

Out[170]: 0.8873063468265867

In [171]: # Calculate false postive rate - predicting converted when lead has not converted
          print(FP/ float(TN+FP))

          0.11269365317341329

In [172]: # positive predictive value
          print (TP / float(TP+FP))

          0.7918781725888325

In [173]: # Negative predictive value
          print (TN / float(TN+ FN))
```

Finding the Optimal Cutoff

Accuracy vs Sensitivity vs Specificity graph

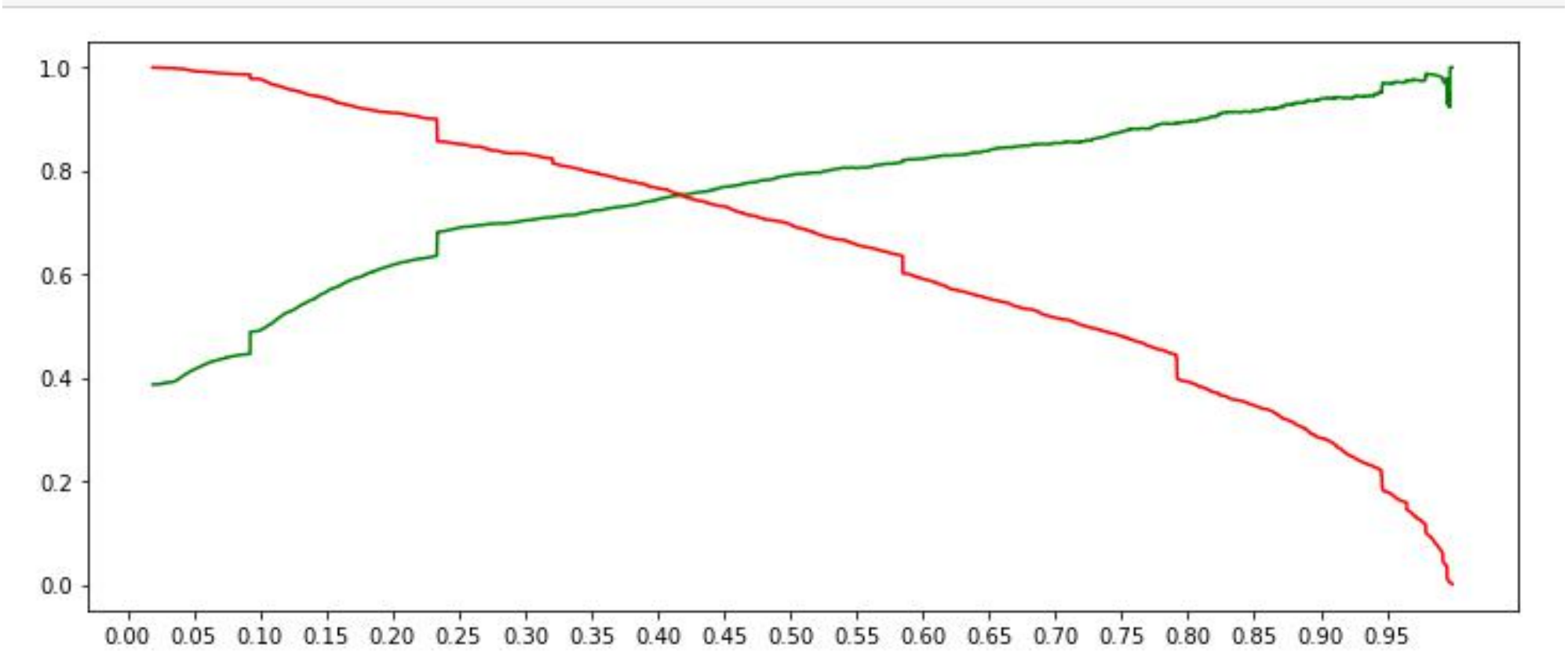
From the accuracy-sensitivity-specificity plot (depicted below), we observed that:
The accuracy is at peak and remains constant between 0.2 and 0.53
All the three metrics converge at 0.35.



Finding the Optimal Cutoff

Precision vs Recall graph

From the Precision vs Recall plot (depicted below), we observed that:
The cutoff is around 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.



Results

Metrics on the Train Set

The results (as depicted below) of our Logistic Regression Model on the **Train Set** was as follows:

- About 81% Accurate. (Accuracy)
- About 78% Precise. (Precision)
- About 72% Recall Rate.

Accuracy , Precision and Recall ¶

```
In [192]: # Accuracy.
          metrics.accuracy_score(y_train_pred_final.Converted, y_train_pred_final.final_predicted)

Out[192]: 0.8133889919604206

In [193]: # Precision
          TP / (TP + FP)

Out[193]: 0.7781705700397702

In [194]: # Recall
          TP / (TP + FN)

Out[194]: 0.7141119221411192
```

Results

Metrics on the Test Set

The results (as depicted below) of our Logistic Regression Model on the **Test Set (Unseen Data)** was as follows:

- About 81% Accurate. (Accuracy)
- About 79% Precise. (Precision)
- About 70% Recall Rate.

Test - Accuracy, Precision and Recall

```
In [213]: # Accuracy.
          metrics.accuracy_score(y_pred_final.Converted, y_pred_final.final_predicted)

Out[213]: 0.8102453102453102

In [214]: # Precision
          TP / float(TP+FP)

Out[214]: 0.7936016511867905

In [215]: # Recall
          TP / float(TP+FN)

Out[215]: 0.7022831050228311
```

Recommendations

- Leads having a Lead-Score of 47 and above should be considered as Hot-Leads and the organisations resources should be focused on those 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.

- Furthermore, 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 categorised as “Very Hot Leads” and have very high chances of conversion. (As these were 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 categorised as “Very Cold Leads” and are least likely to convert.

Conclusion

After evaluating our Logistic Regression Model (based on the Accuracy, Precision and Recall values as depicted in the previous slide), we can safely conlcude that the model would help X Education to indentify the leads that are most likely to convert into paying customers.
Since the model has an Accuracy and Precision of about 80% (as depicted in previous slide), it would also help meet the CEO's ballpark target of lead conversion rate to be around 80%.

Furthermore, 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 cut-off 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.

THE END