LEAD SCORING CASE STUDY

SUBMISSION

Group Name:

1. Prateek Gharat
2. S S K M Chaitanya Pusuluri

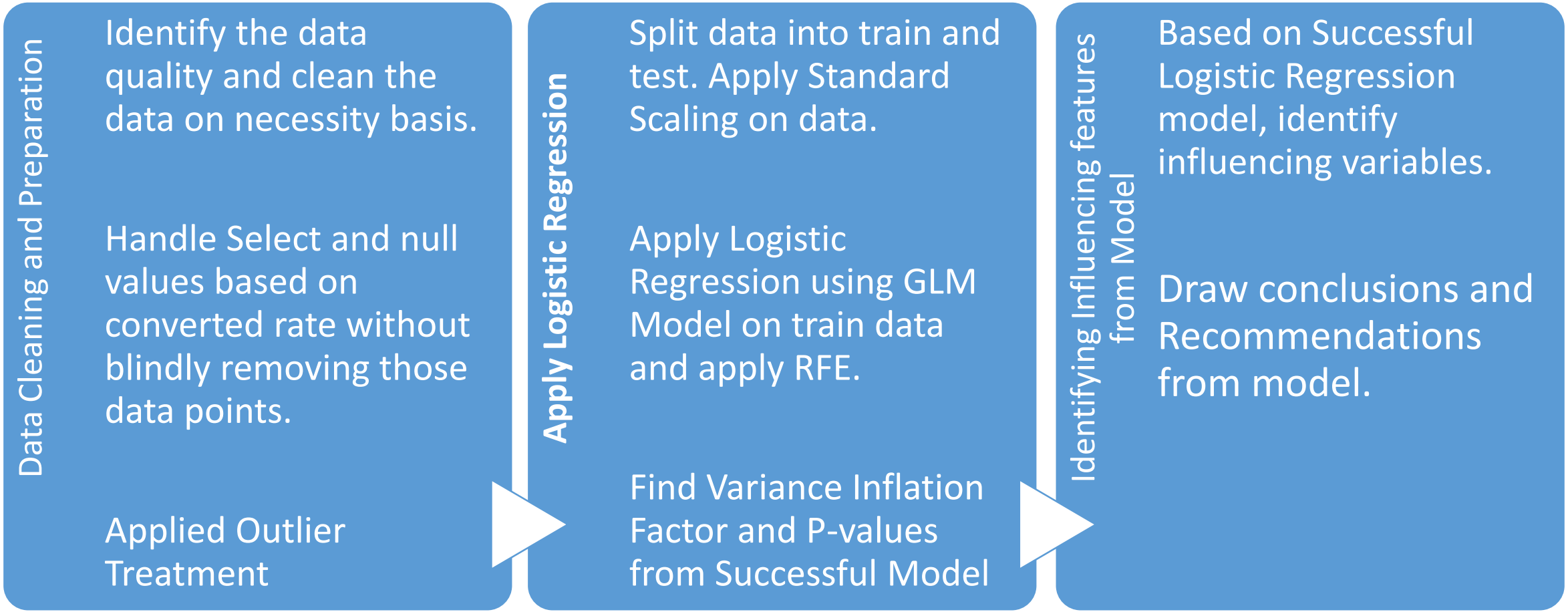
Abstract

Objective:- Identify the most potential leads/Hot Leads, which in turn increases the lead conversion rate.

Approach:-

- Data Cleaning and Preparing data for analysis.
- Split data into train and test. Apply Logistic Regression with GLM model.
- Identify Variables which are influencing model.

Problem solving methodology

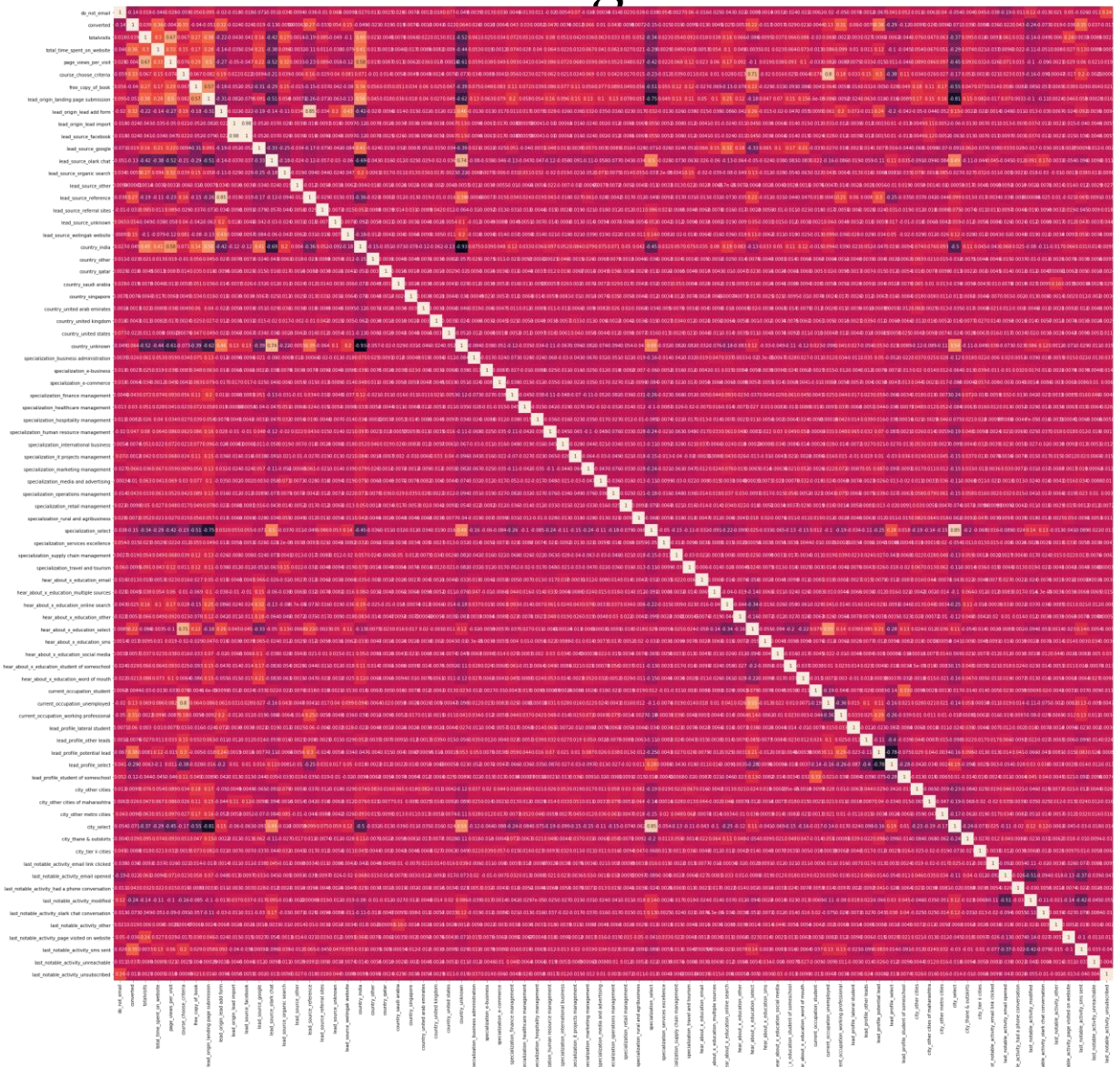


Analysis

Data Cleaning and Preparation:-

- Checking for percentage of leads converted and not converted in the given data.
- Drop numerical Columns having high null percentage(>30 %).
- As per the data 'select' which means the customer had not selected this option while filling the form is as good as NaN. So we replaced NaN values with Select for “lead_profile”, “specialization” and “city” without dropping these columns. If we drop these columns we will loose 40% of data.
- Dropped Columns which are having least variance of data.
- Applied Outlier treatment.

Data Cleaning

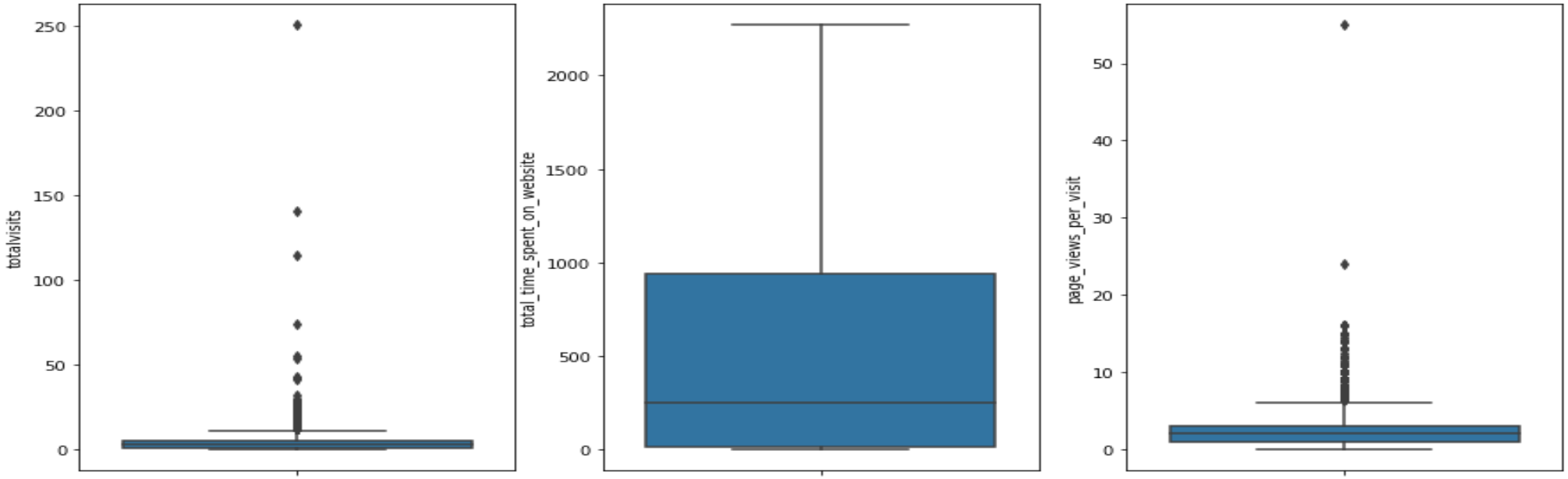


Variables with High correlation were found from the heatmap and dropped accordingly.

	Variable 1	Variable 2	coefficient
1	course_choose_criteria	current_occupation_unemployed	0.798003
4	lead_origin_lead import	lead_source_facebook	0.981709
6	lead_origin_lead add form	lead_source_reference	0.852594
7	country_unknown	lead_source_olark chat	0.742487
9	course_choose_criteria	hear_about_x_education_select	0.705097
10	course_choose_criteria	current_occupation_unemployed	0.798003
11	city_select	specialization_select	0.845374

Outlier Treatment

Outlier Treatment



Column Name	no	yes
do_not_call	9238	2
search	9226	14
newspaper_article	9238	2
x_education_forums	9239	1
newspaper	9239	1
digital_advertisement	9236	4
through_recommendations	9233	7

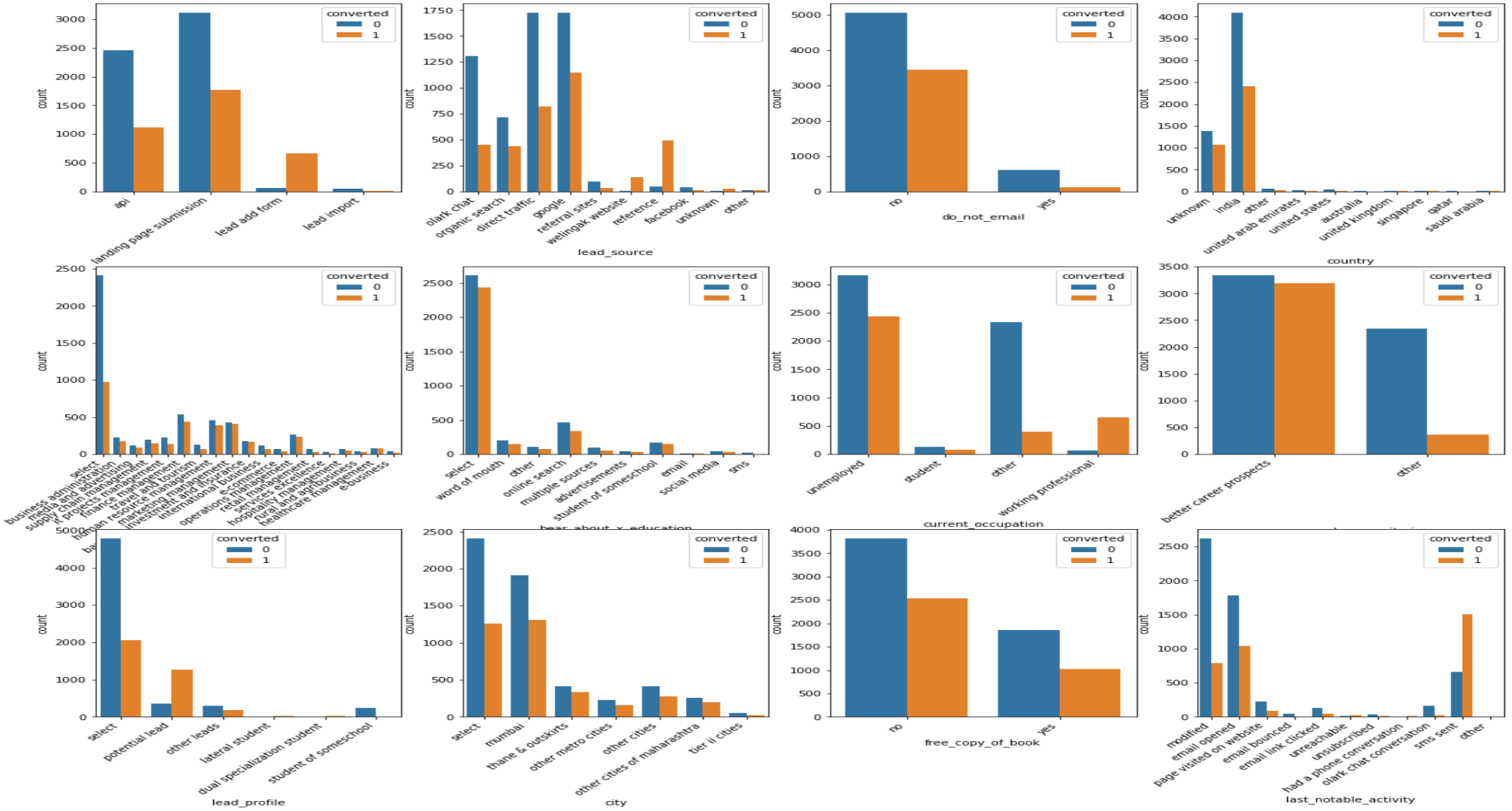
Less Data Variance columns

Analysis

Apply Logistic Regression

- Split data into train and test by taking train size as 70% and test size as 30%.
- Applied StandardScaler on train data.
- Run GLM Model on train data and verify the p-values and Variance Inflation Factor.
- Also find Confusion matrix, Accuracy Score.
- Apply RFE on this train data. Using this RFE columns we will re-run GLM model till we found the best VIF values(5<vif<7 acceptable range) and p-value(<0.05).

Exploratory Data Analytics (EDA)



Checking for converted and non converted leads for categorical variables

Recursive Feature Elimination (RFE)

```
col = X_train.columns[rfe.support_]
```

```
list(col)
```

```
['do_not_email',
 'total_time_spent_on_website',
 'lead_origin_lead add form',
 'lead_source_welingak website',
 'country_qatar',
 'current_occupation_working professional',
 'lead_profile_other leads',
 'lead_profile_select',
 'lead_profile_student of someschool',
 'last_notable_activity_had a phone conversation',
 'last_notable_activity_sms sent',
 'last_notable_activity_unreachable']
```

```
X_train.columns[~rfe.support_]
```

```
Index(['totalvisits', 'page_views_per_visit', 'course_choose_criteria',
       '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', 'country_unknown',
       'specialization_business administration', 'specialization_e-business',
       'specialization_e-commerce', 'specialization_finance management',
       'specialization_healthcare management',
       'specialization_hospitality management',
       'specialization_human resource management',
       'specialization_international business',
       'specialization_it projects management',
       'specialization_marketing management',
       'specialization_media and advertising',
       'specialization_operations management',
       'specialization_retail management',
       'specialization_rural and agribusiness',
       'specialization_services excellence',
       'specialization_supply chain management',
       'specialization_travel and tourism', 'hear_about_x_education_email',
       'hear_about_x_education_multiple sources',
       'hear_about_x_education_online search', 'hear_about_x_education_other',
       'hear_about_x_education_sms', 'hear_about_x_education_social media',
       'hear_about_x_education_student of someschool',
       'hear_about_x_education_word of mouth', 'current_occupation_student',
       'lead_profile_lateral student', 'lead_profile_potential lead',
       'city_other cities', 'city_other cities of maharashtra',
       'city_other metro cities', 'city_select', 'city_thane & outskirts',
       'city_tier ii cities', 'last_notable_activity_email link clicked',
       'last_notable_activity_email opened', 'last_notable_activity_modified',
       'last_notable_activity_olark chat conversation',
       'last_notable_activity_other',
       'last_notable_activity_page visited on website',
       'last_notable_activity_unsubscribed'],
      dtype='object')
```

Recursive Feature Elimination (RFE) was used to eliminate the columns or variables and select the relevant once for model building.

Model Building

Generalized Linear Model Regression Results

Dep. Variable:	converted	No. Observations:	6468
Model:	GLM	Df Residuals:	6456
Model Family:	Binomial	Df Model:	11
Link Function:	logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-2613.1
Date:	Sun, 09 Jun 2019	Deviance:	5226.1
Time:	15:01:53	Pearson chi2:	7.54e+03
No. Iterations:	7	Covariance Type:	nonrobust

	coef	std err	z	P>|z|	[0.025	0.975]
const	0.1567	0.088	1.774	0.076	-0.016	0.330
do_not_email	-1.2971	0.165	-7.855	0.000	-1.621	-0.973
total_time_spent_on_website	0.9151	0.035	26.024	0.000	0.846	0.984
lead_origin_lead add form	2.9227	0.190	15.395	0.000	2.551	3.295
lead_source_welingak website	2.6760	0.747	3.582	0.000	1.212	4.140
current_occupation_working professional	2.3955	0.192	12.457	0.000	2.019	2.772
lead_profile_other leads	-1.3606	0.165	-8.234	0.000	-1.684	-1.037
lead_profile_select	-1.7808	0.096	-18.634	0.000	-1.968	-1.593
lead_profile_student of someschool	-3.3215	0.428	-7.765	0.000	-4.160	-2.483
last_notable_activity_had a phone conversation	2.8404	1.152	2.465	0.014	0.582	5.099
last_notable_activity_sms sent	1.6838	0.080	21.169	0.000	1.528	1.840
last_notable_activity_unreachable	1.5245	0.546	2.790	0.005	0.453	2.596

	Features	VIF
2	lead_origin_lead add form	1.38
9	last_notable_activity_sms sent	1.34
6	lead_profile_select	1.33
3	lead_source_welingak website	1.24
4	current_occupation_working professional	1.16
0	do_not_email	1.09
1	total_time_spent_on_website	1.07
5	lead_profile_other leads	1.01
7	lead_profile_student of someschool	1.01
8	last_notable_activity_had a phone conversation	1.00
10	last_notable_activity_unreachable	1.00

Using RFE and manual feature elimination for features having P-Value more than 0.05 and VIF more than 5. We reached a final model with P-Value less than 0.05 and VIF less than 5.

Analysis

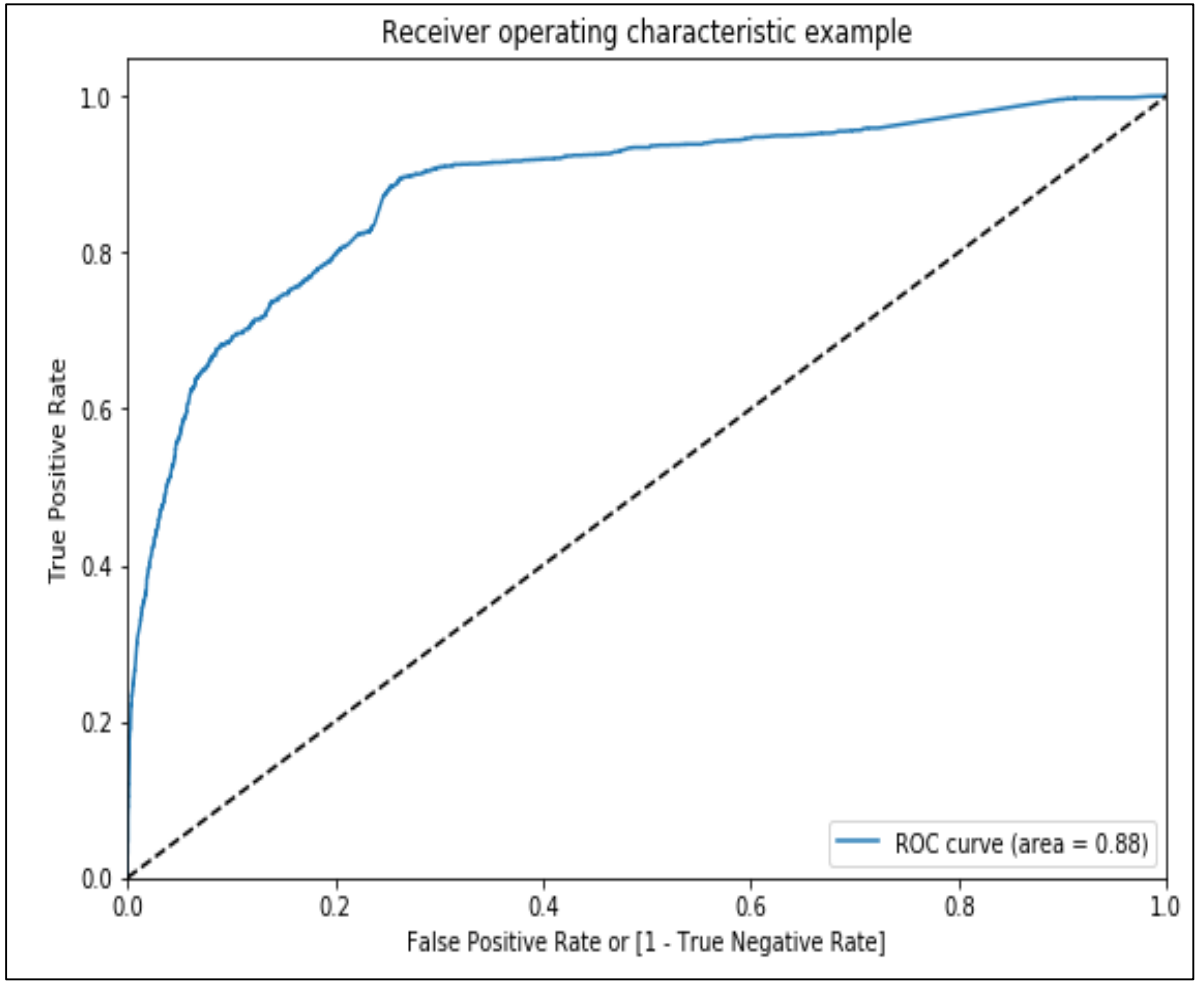
Identifying Influencing features from Model

- To evaluate Model, we need to calculate Sensitivity, Specificity.
- Also Plot ROC Curve and Optimal Cutoff Point.
- As per Precision-Recall Tradeoff, the cutoff is around 0.425 (between 0.4 and 0.45) . We can choose the cut-off as 0.47 and use the Precision-Recall-Accuracy metrics to evaluate the model.
- Identified list of features that have influence on Lead Score

'do_not_email', 'total_time_spent_on_website', 'lead_origin_lead add form', 'lead_source_welingak website', 'current_occupation_working professional', 'lead_profile_other leads', 'lead_profile_select', 'lead_profile_student of someschool', 'last_notable_activity_had a phone conversation', 'last_notable_activity_sms sent', 'last_notable_activity_unreachable'

ROC Curve

Using RFE and manual feature elimination we arrived at final model with the ROC curve and important metrics.




```
#True positive
TP = Confusion_Mat[1,1]

#True negatives
TN = Confusion_Mat[0,0]

#False positives
FP = Confusion_Mat[0,1]

#False negatives
FN = Confusion_Mat[1,0]
```

```
#Sensitivity of our logistic regression model
TP / float(TP+FN)
```

```
0.6824817518248175
```

```
#Specificity of the model
TN / float(TN+FP)
```

```
0.9067966016991504
```

```
#Calculate false postive rate - predicting converted when lead has not converted
print(FP/ float(TN+FP))
```

```
0.09320339830084957
```

```
# positive predictive value
print (TP / float(TP+FP))
```

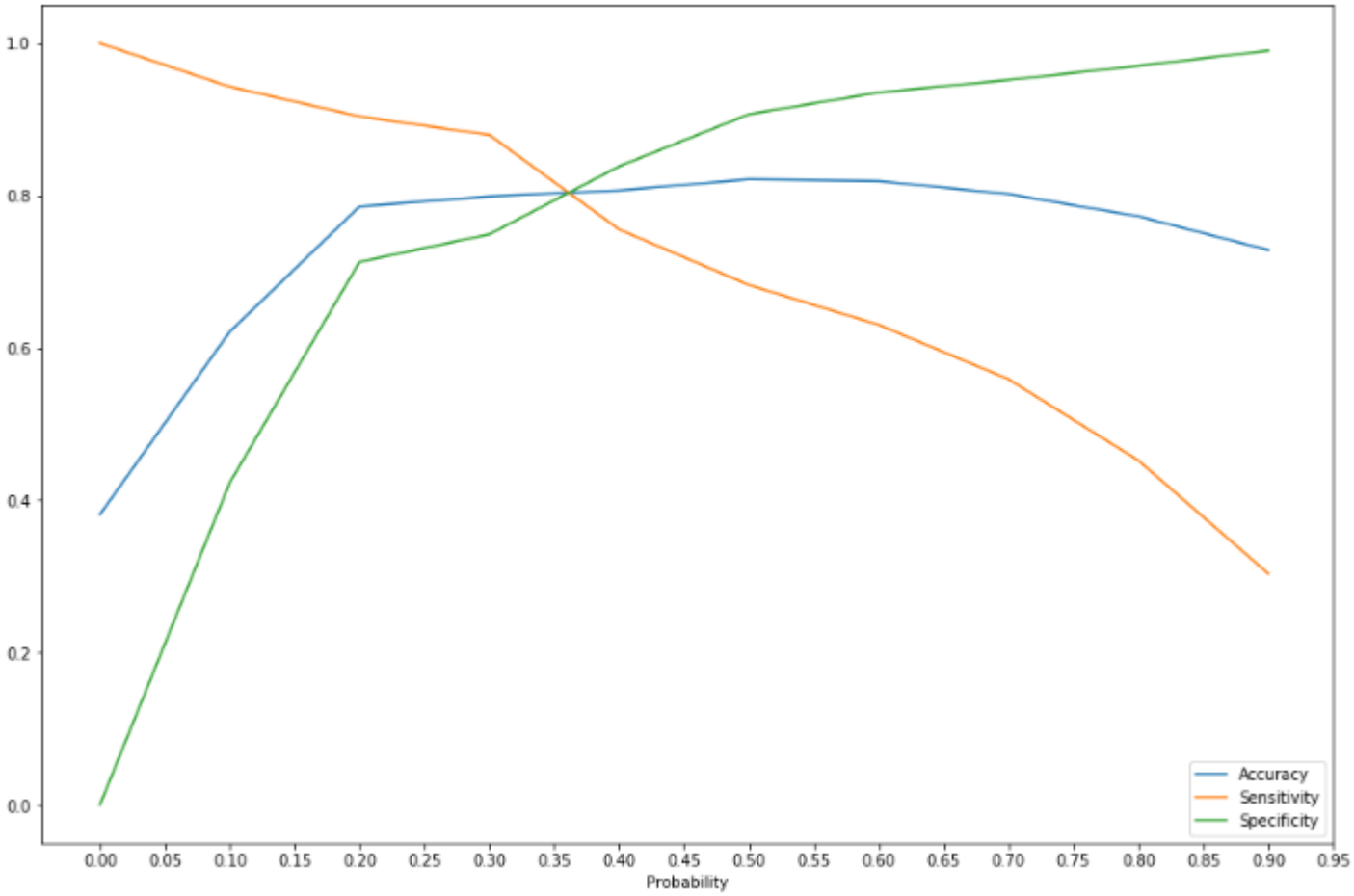
```
0.818579766536965
```

```
# Negative predictive value
print (TN / float(TN+ FN))
```

```
0.822529465095195
```

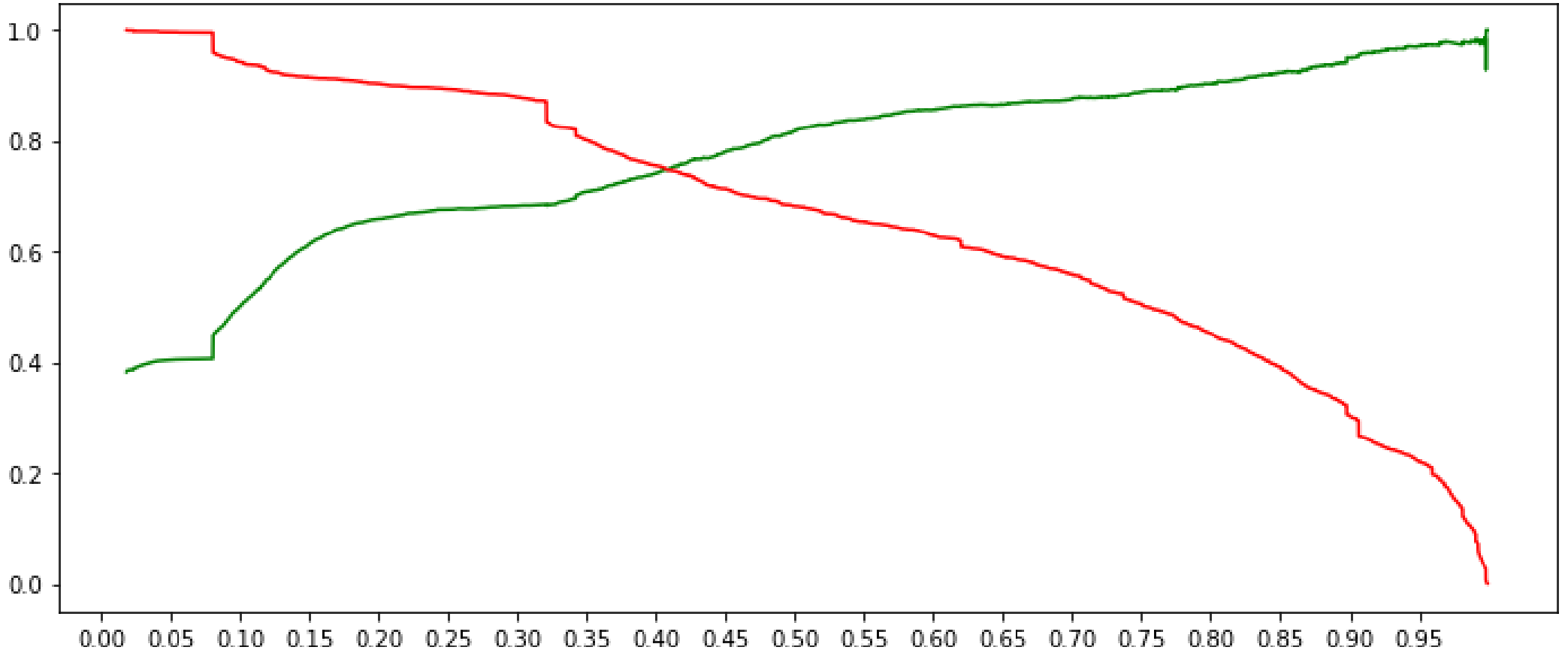
Accuracy vs Sensitivity vs Specificity

From the accuracy-sensitivity-specificity plot, we find that 0.35 is the optimum point where all 3 metrics merge.



Precision and recall tradeoff

As per Precision-Recall Tradeoff, the cutoff is around 0.425 (between 0.4 and 0.45) . We can choose the cut-off as 0.47 and use the Precision-Recall-Accuracy metrics to evaluate the model.



Result – Train and Test

Below is the result of train and test model.

Train - Accuracy , Precision and Recall

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

```
0.8161719233147805
```

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

```
0.7941040994933211
```

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

```
0.6991078669910786
```

Test - Accuracy, Precision and Recall

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

```
0.814935064935065
```

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

```
0.8108974358974359
```

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

```
0.6931506849315069
```

Recommendations

- Leads with cut-off value of 0.47 and above must be considered as Hot Leads and sales team should focus on calling these leads to achieve maximum conversions.

- As the cut-off range as per the model is 0.47, sales team should use there resources and time effectively to convert the hot leads (cut-off above 0.47) rather than focusing on cold once.

- The Top three variables in our model which contribute the most towards the probability of a lead getting converted are **Lead Origin, Last Notable Activity** and **Lead Source.**

- The Top 3 Categorical/Dummy variables in the model which should be focused the most on in order to increase the probability of lead conversion are **lead_origin_lead add form** (coefficient value of 2.9227), **last_notable_activity_had a phone conversation** (coefficient value of 2.8404) and **lead_source_welingak website** (coefficient value of 2.6760)

Conclusion

- As per our Logistic Regression Model, we can conclude that the model would help X Education to identify the leads that are most likely to convert into paying customers.
- 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.
- 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%.
- The model can be adjusted if X Education requirement changes in future like:
  - A period of 2 months every year during which X Education hire some interns and during this phase, they wish to make the lead conversion more aggressive also X Education want almost all the potential leads to be converted and hence, want to make phone calls to as much of such people as possible we can reduce the cut-off range to increase the projected leads. As per the model cut-off is 0.47, we can lower it up to 0.30 to 0.35 where the projected lead can be 4522 to 3972. An important thing to note is that as we lower the cut-off value it would also decrease the precision.
  - At times, the company reaches its target for a quarter before the deadline and during this time, the company wants the sales team to focus on some new work as well. Also 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, we can raise the cut-off value higher than 0.47 to target less customer but with high conversion value. We can set the cut-off range around 0.80 to 0.85 where the projected lead can be 1743 to 1469. Here the precision would be much high as we are contacting the hot leads which has high chances of conversion and will achieve the aim to avoid useless phone calls.