Lead Score Case Study

PGDDS – Cohort 14

Vidhu Jain

Problem Statement

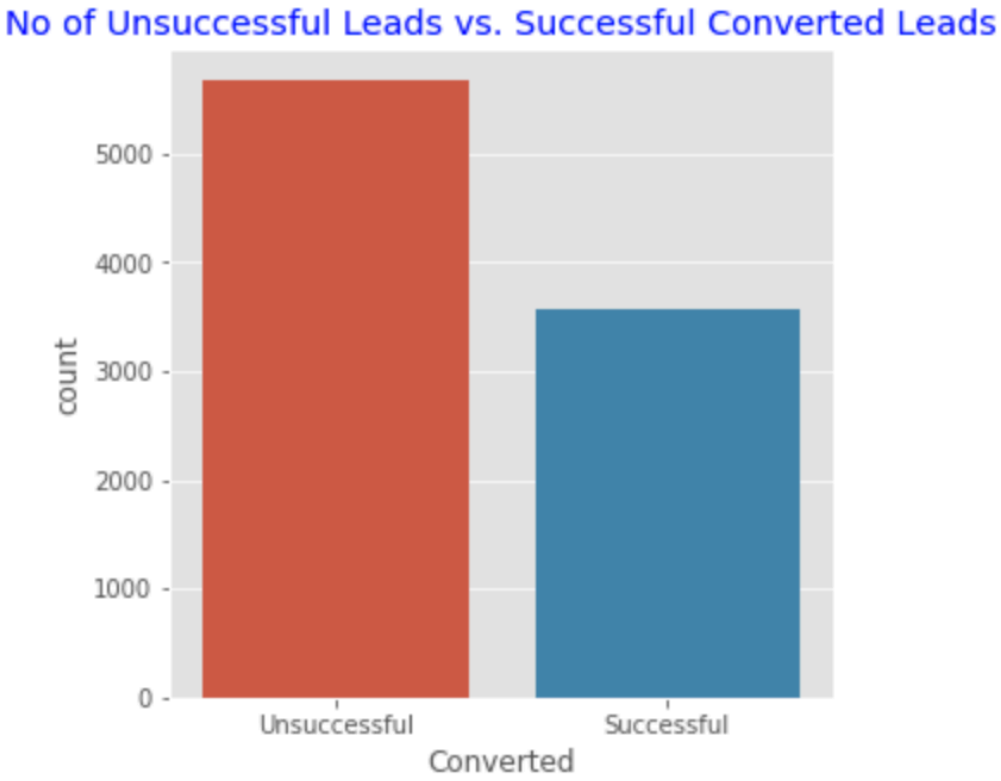
An education company named X Education sells online courses to industry professionals. The company markets its courses on several websites and search engines like Google. Once these people land on the website, they might browse the courses or fill up a form for the course or watch some videos. When these people fill up a form providing their email address or phone number, they are classified to be a lead.

Now, although X Education gets a lot of leads, its lead conversion rate is very poor. To make this process more efficient, the company wishes to identify the most potential leads, also known as 'Hot Leads'.

X Education has appointed you to help them select the most promising leads, i.e. the leads that are most likely to convert into paying customers. The company requires you to build a model wherein you 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%.

Exploratory Data Analysis

No. of successful vs. unsuccessful leads given in data set

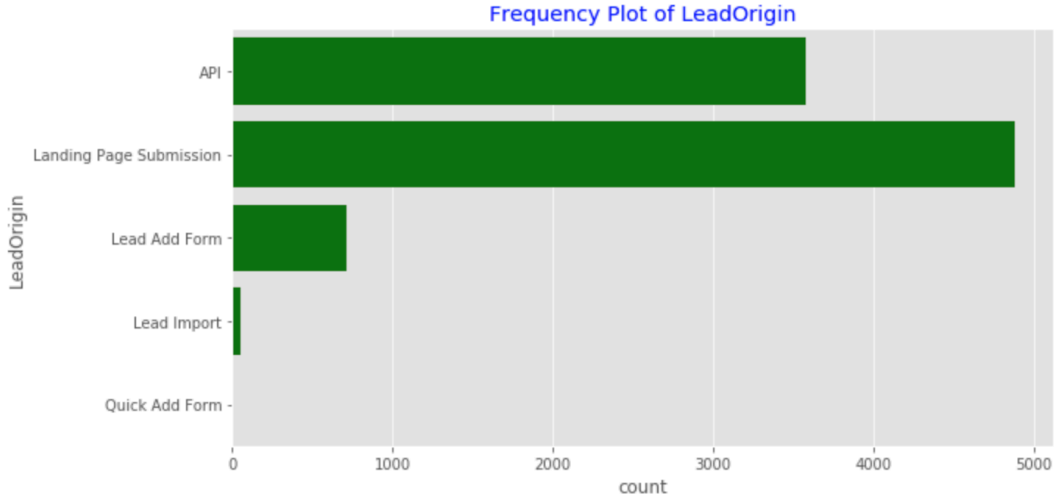


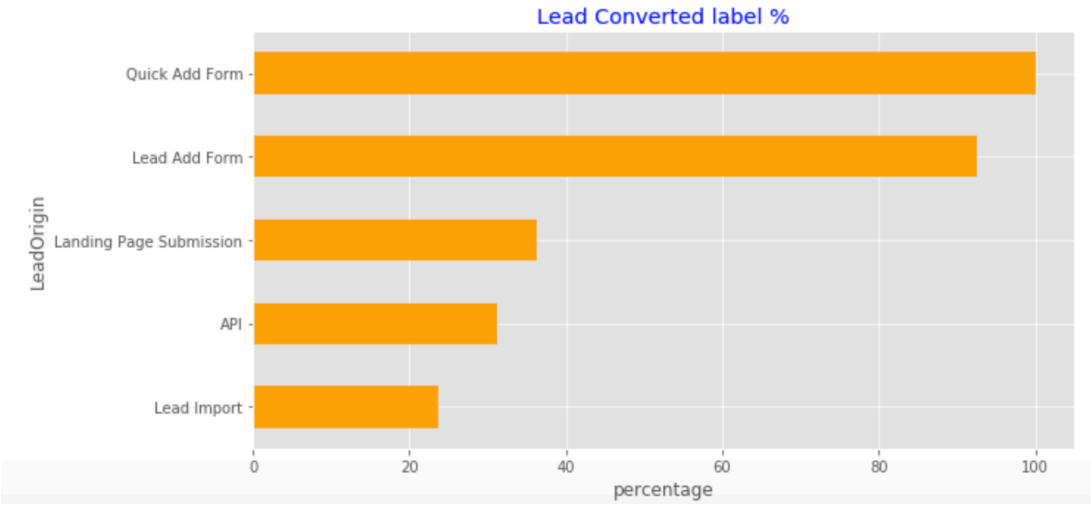
Observations

- The Target class is not highly imbalanced with
- 61.46% of observations as "0" - labeled as non-Converted or unsuccessful leads
- 38.54% of observations as "1" - labeled as Converted or successful leads

Exploratory Data Analysis

Lead Origin



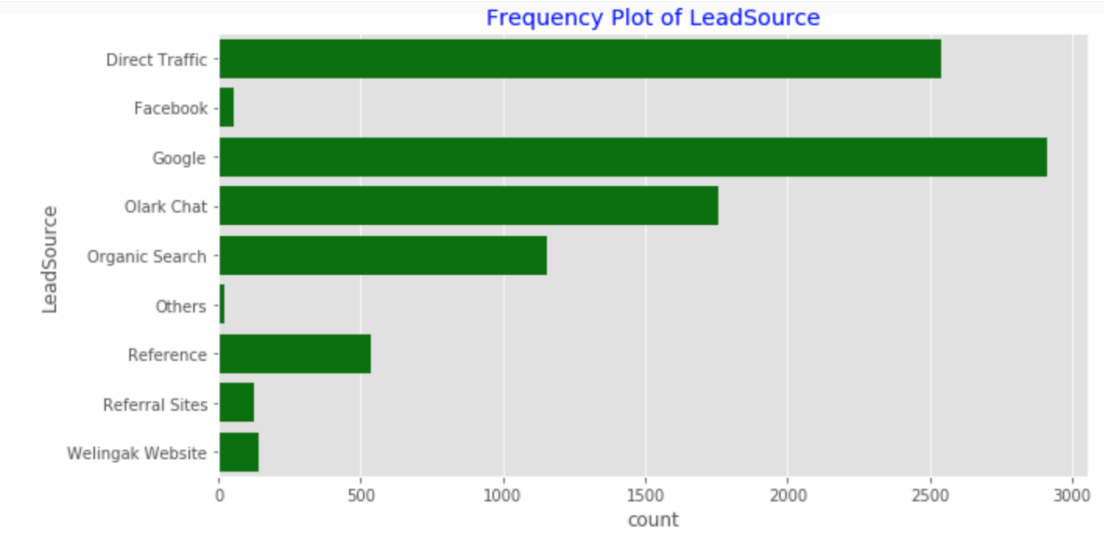


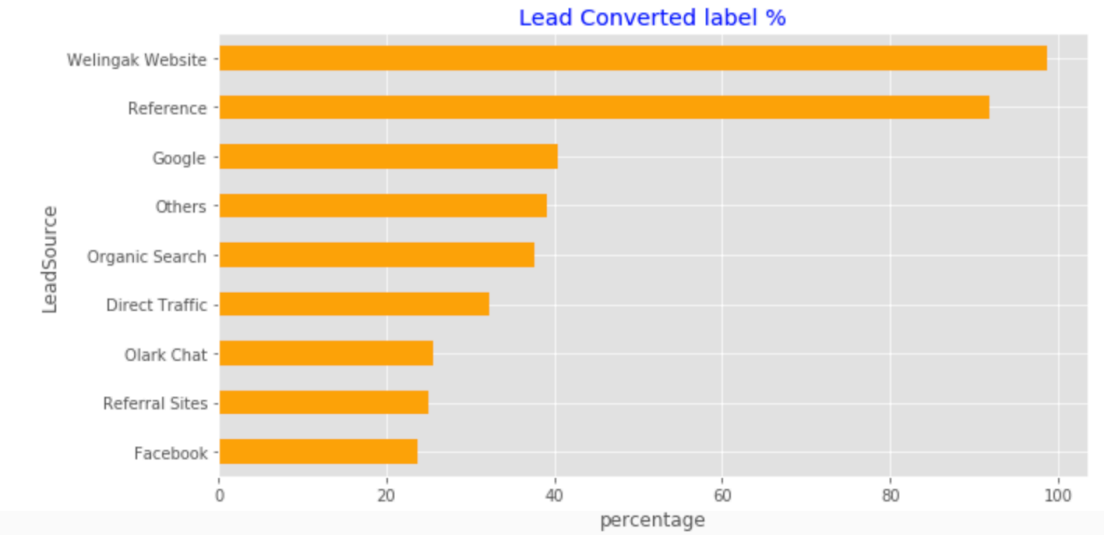
Observations

- Maximum observations have Lead Origin option as 'Landing Page Submission’ (4881 in total)
- Lead Origin with 'Quick Add Form' is successfully converted and there is only one observation with such Lead Origin
- 'Lead Add Form' comparatively has the highest lead conversion rate (92.48%)

Exploratory Data Analysis

Lead Source



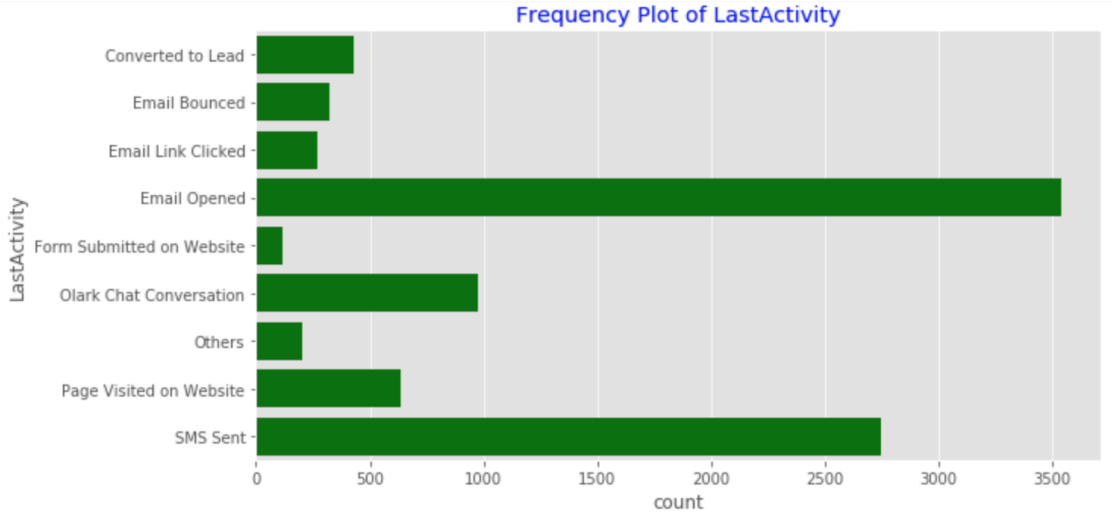


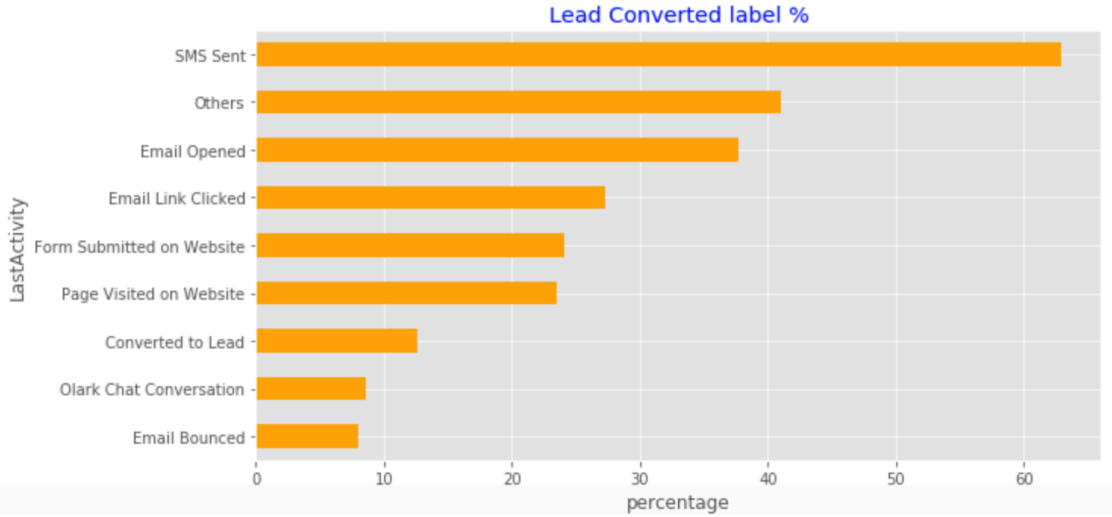
Observations

- Maximum leads are generated from 'Google' (2909) followed by 'Direct Traffic' (2539)
- 'Welingak Website' comparatively has the highest successful lead conversion rate (98.59%)
- Focus should be more on 'Facebook', 'Referral Sites', 'Olark Chat' to improve lead conversion rate

Exploratory Data Analysis

Last Activity



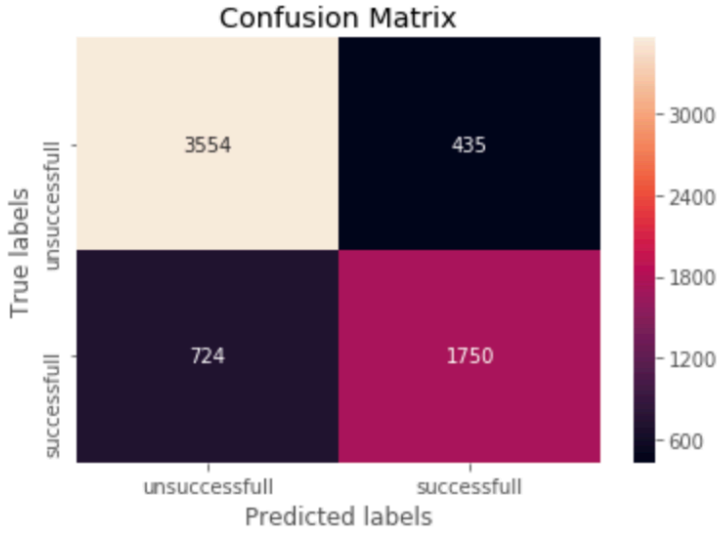


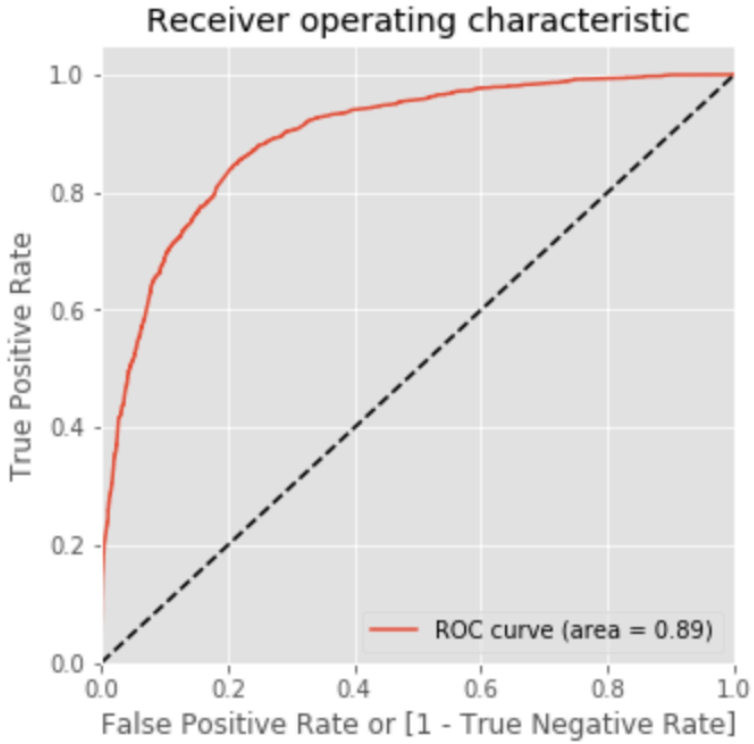
Observations

- Maximum leads have last activity as 'Email Opened’ (3540)
- Leads with last activity as 'SMS Sent' have the highest lead conversion rate (62.91%)

Build Logistic Regression Model

With Cutoff probability as 0.5






```
Accuracy score: 0.8206715147764196
Sensitivity score: 0.7073565076798707
f1-score: 0.7512341704228375
Precision score: 0.8009153318077803
Recall score: 0.7073565076798707
```

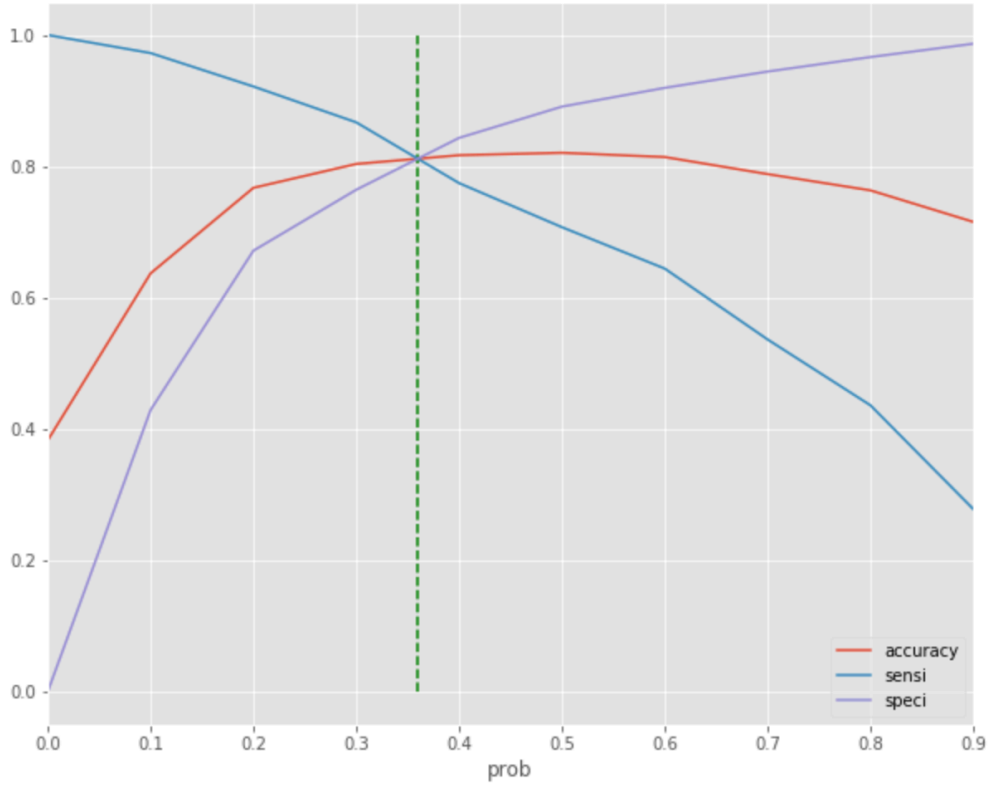
Observations

- Accuracy is 82%, Sensitivity is 70.7% and precision is 80%
- ROC Curve has area = 0.89

Build Logistic Regression Model

The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%. This means Sensitivity must be greater than 80%

For this, Check for optimal cutoff point



	cutoff	train_acc	train_sen	train_spec	train_prec	test_acc	test_sen	test_spec	test_prec
0	0.36	0.81	0.81	0.82	0.73	0.81	0.8	0.82	0.74
1	0.35	0.81	0.84	0.79	0.72	0.81	0.83	0.79	0.72
2	0.37	0.81	0.79	0.83	0.74	0.81	0.79	0.83	0.75

Observations

- Train and test accuracy, sensitivity and precision scores are almost similar for cutoff at 0.35
- Also, So we will choose cut-off as 0.35 and generate lead scores

Build Logistic Regression Model

Generate Lead score on train and Test Data with cutoff probability as 0.35

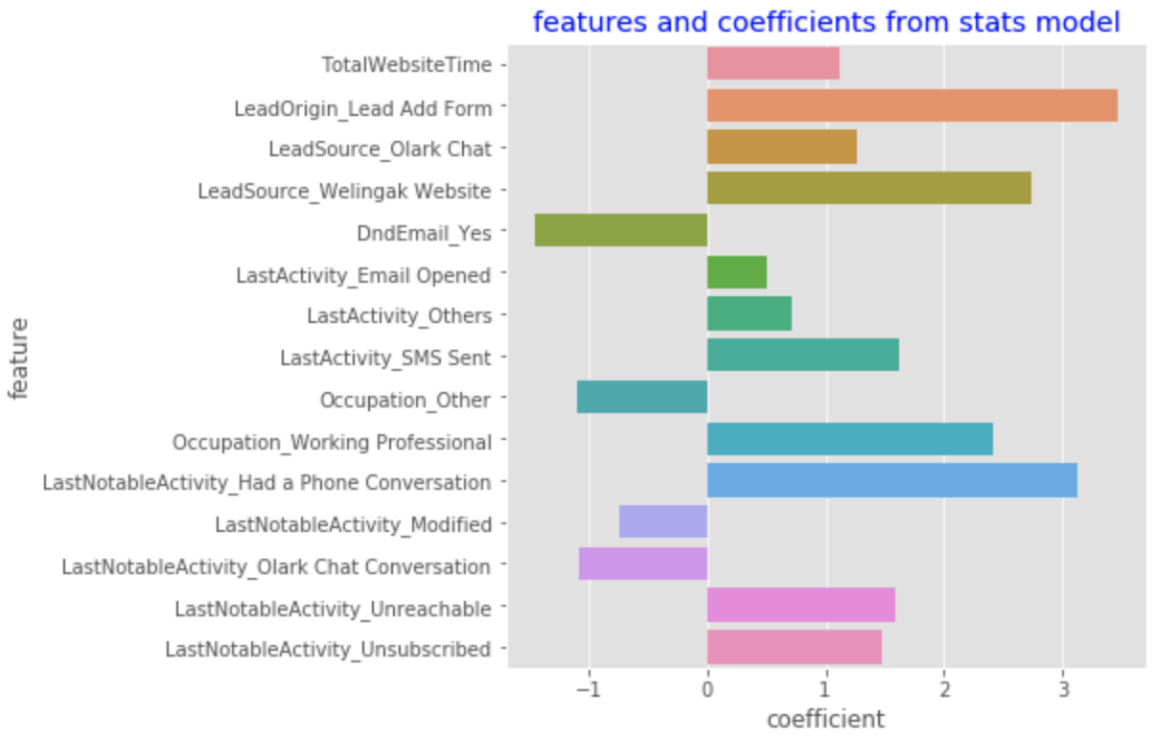
Snapshot of combined data set – train and test data with Lead Score

```
shape of combined dataset: (9234, 6)
```

	LeadId	Converted	Converted_Prob	predicted	final_predicted	Lead_Score
0	7417	0	0.926517	1	1	92.651679
1	1032	1	0.888379	1	1	88.837938
2	6537	0	0.472254	1	1	47.225438
3	7284	0	0.160708	0	0	16.070783
4	3194	1	0.832044	1	1	83.204449

Business Problem

1. Which are the top three variables in your model which contribute most towards the probability of a lead getting converted?



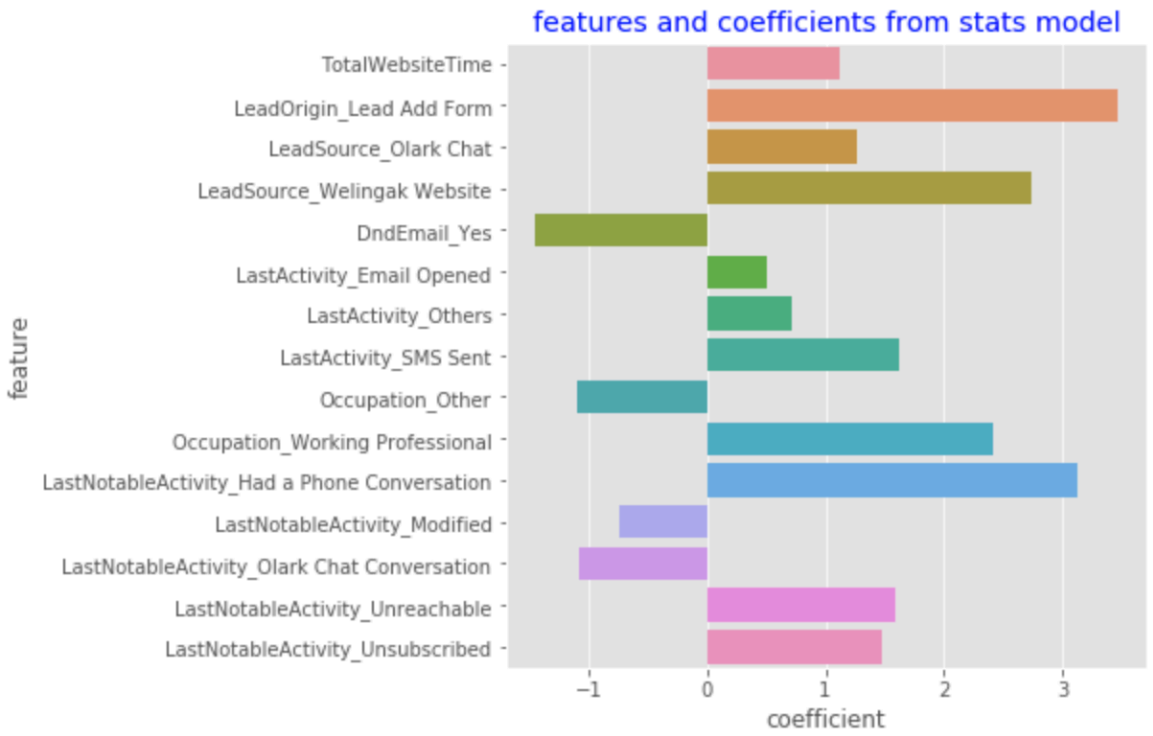
	feature	coefficient
2	LeadOrigin_Lead Add Form	3.466767
11	LastNotableActivity_Had a Phone Conversation	3.123466
4	LeadSource_Welingak Website	2.733684
10	Occupation_Working Professional	2.406889
8	LastActivity_SMS Sent	1.623893
14	LastNotableActivity_Unreachable	1.593580
15	LastNotableActivity_Unsubscribed	1.470591
3	LeadSource_Olark Chat	1.273367
1	TotalWebsiteTime	1.126039
7	LastActivity_Others	0.721333
6	LastActivity_Email Opened	0.511292
12	LastNotableActivity_Modified	-0.742999
13	LastNotableActivity_Olark Chat Conversation	-1.078162
9	Occupation_Other	-1.092805
5	DndEmail_Yes	-1.452211

Observations

- Top three predictors that contribute in lead conversion are Lead Origin, Last Notable Activity and Lead Source

Business Problem

2. What are 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?



	feature	coefficient
2	LeadOrigin_Lead Add Form	3.466767
11	LastNotableActivity_Had a Phone Conversation	3.123466
4	LeadSource_Welingak Website	2.733684
10	Occupation_Working Professional	2.406889
8	LastActivity_SMS Sent	1.623893
14	LastNotableActivity_Unreachable	1.593580
15	LastNotableActivity_Unsubscribed	1.470591
3	LeadSource_Olark Chat	1.273367
1	TotalWebsiteTime	1.126039
7	LastActivity_Others	0.721333
6	LastActivity_Email Opened	0.511292
12	LastNotableActivity_Modified	-0.742999
13	LastNotableActivity_Olark Chat Conversation	-1.078162
9	Occupation_Other	-1.092805
5	DndEmail_Yes	-1.452211

Observations

- Top three categorical/dummy variables are LeadOrigin_Lead Add Form, LastNotableActivity_Had a Phone Conversation and LeadSource_Welingak Website

Business Problem

3. X Education has a period of 2 months every year during which they hire some interns. The sales team, in particular, has around 10 interns allotted to them. So during this phase, they wish to make the lead conversion more aggressive. So they want almost all of the potential leads (i.e. the customers who have been predicted as 1 by the model) to be converted and hence, want to make phone calls to as much of such people as possible. Suggest a good strategy they should employ at this stage

Solution: Since the company wants to focus on lead conversion aggressively, this means that company is focusing more on True Positive.
For this, we should choose the cutoff such that the sensitivity of the model should increase without compromising much on accuracy of model. Cutoff as 0.3 gives good sensitivity score without compromising much on accuracy of model

	prob	accuracy	sensi	speci
0.0	0.0	0.382794	1.000000	0.000000
0.1	0.1	0.637011	0.972514	0.428930
0.2	0.2	0.767291	0.921584	0.671597
0.3	0.3	0.803652	0.867017	0.764352
0.4	0.4	0.816958	0.774454	0.843319
0.5	0.5	0.820672	0.707357	0.890950
0.6	0.6	0.814173	0.644301	0.919529
0.7	0.7	0.788179	0.536378	0.944347
0.8	0.8	0.763423	0.436136	0.966408
0.9	0.9	0.715457	0.278092	0.986713

	cutoff	train_acc	train_sen	train_spec	train_prec	test_acc	test_sen	test_spec	test_prec
0	0.36	0.81	0.81	0.82	0.73	0.81	0.8	0.82	0.74
1	0.35	0.81	0.84	0.79	0.72	0.81	0.83	0.79	0.72
2	0.37	0.81	0.79	0.83	0.74	0.81	0.79	0.83	0.75
3	0.3	0.8	0.87	0.76	0.7	0.8	0.86	0.76	0.7
4	0.6	0.81	0.64	0.92	0.83	0.8	0.61	0.92	0.84

Business Problem

4. Similarly, at times, the company reaches its target for a quarter before the deadline. During this time, the company wants the sales team to focus on some new work as well. So 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. Suggest a strategy they should employ at this stage.

Solution:

- Since the company wants to to avoid unnecessary phone call, this means that company is focusing on less number of false positive rate.
- Since FPR = 1 - Specificity, this we must set a cutoff such that the Specificity is high from the model thereby resulting in less False Positive Rate

	prob	accuracy	sensi	speci
0.0	0.0	0.382794	1.000000	0.000000
0.1	0.1	0.637011	0.972514	0.428930
0.2	0.2	0.767291	0.921584	0.671597
0.3	0.3	0.803652	0.867017	0.764352
0.4	0.4	0.816958	0.774454	0.843319
0.5	0.5	0.820672	0.707357	0.890950
0.6	0.6	0.814173	0.644301	0.919529
0.7	0.7	0.788179	0.536378	0.944347
0.8	0.8	0.763423	0.436136	0.966408
0.9	0.9	0.715457	0.278092	0.986713

	cutoff	train_acc	train_sen	train_spec	train_prec	test_acc	test_sen	test_spec	test_prec
0	0.36	0.81	0.81	0.82	0.73	0.81	0.8	0.82	0.74
1	0.35	0.81	0.84	0.79	0.72	0.81	0.83	0.79	0.72
2	0.37	0.81	0.79	0.83	0.74	0.81	0.79	0.83	0.75
3	0.3	0.8	0.87	0.76	0.7	0.8	0.86	0.76	0.7
4	0.6	0.81	0.64	0.92	0.83	0.8	0.61	0.92	0.84

Conclusion and Recommendations

- The cutoff probability must be set as 0.35 for conversion rate i.e. Sensitivity of model to be 80%
- For aggressive lead conversion, the cut-off must be set as 0.3 increasing the Sensitivity of model without compromising much on accuracy of the model
- To avoid unnecessary phone calls, the cut-off must be set as 0.6 thereby increasing Specificity of the model without compromising much on accuracy of model
- Lead Origin, Last Notable Activity and Lead Source are top predictors in lead conversion model
- By marketing more on Welingak Website or approaching more Housewife, Working Professional will increase the chances of lead conversion
- Also marketing with Lead Source as Quick Add Form will increase the chances of conversion rate