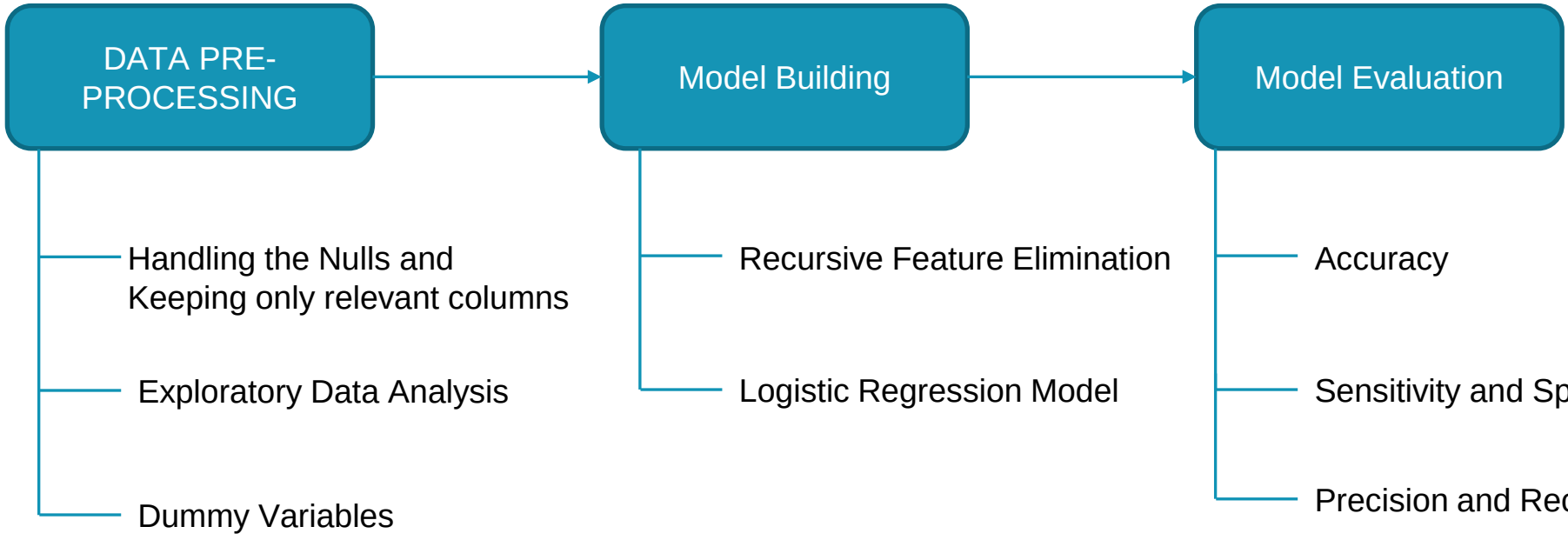
Lead Scoring Case Study

Submitted By:
- Sumit Kumar
- Akashnidhi Prasad

Building a logistic regression model on an education company (X Education) leads data and scoring each leads between 0 and 100 which can be used by the company for targeting. 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.

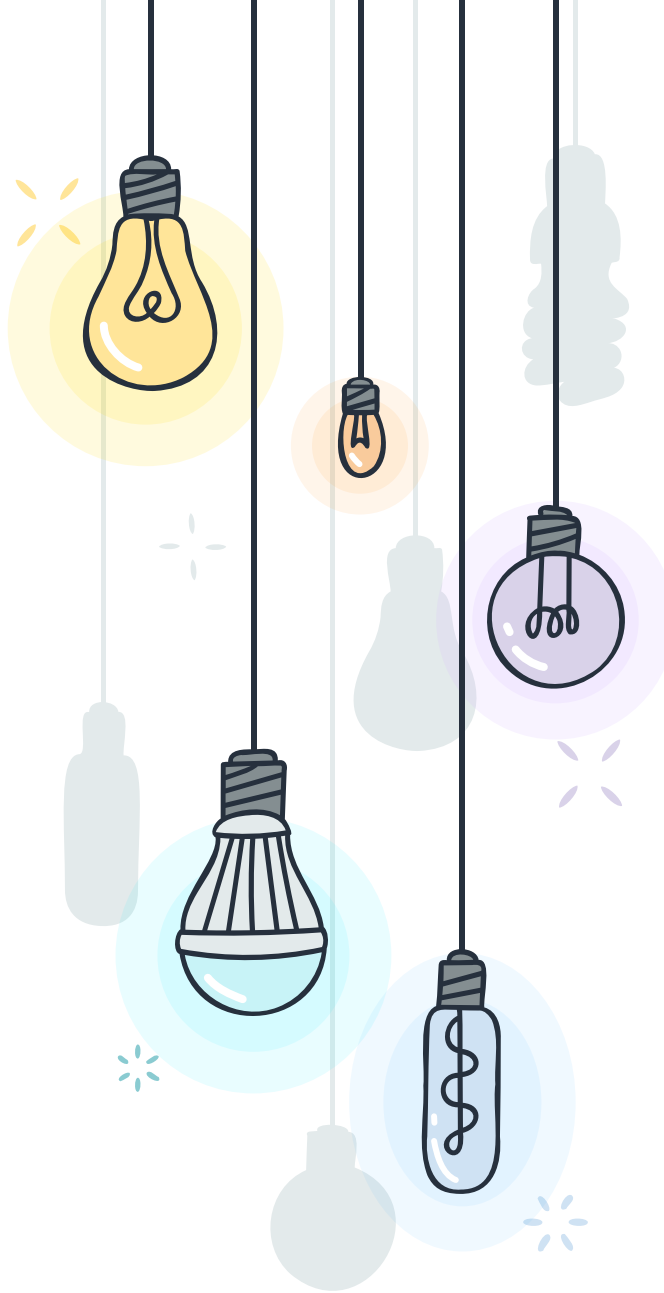
Flow Diagram



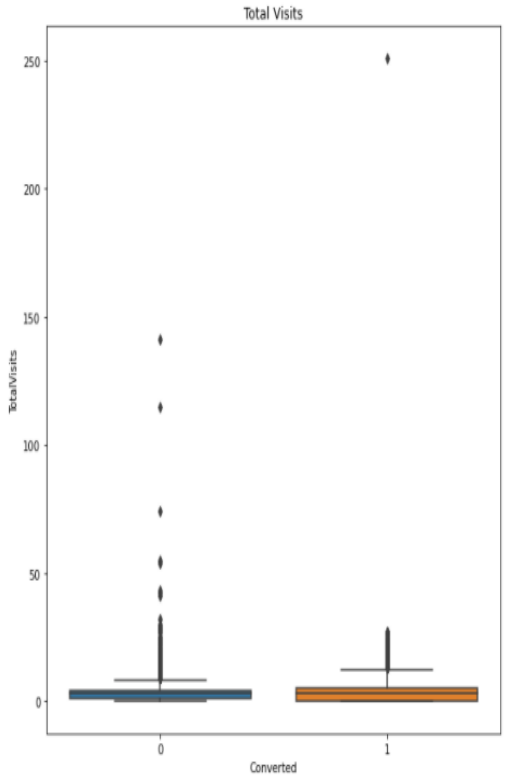
Data Cleaning & EDA

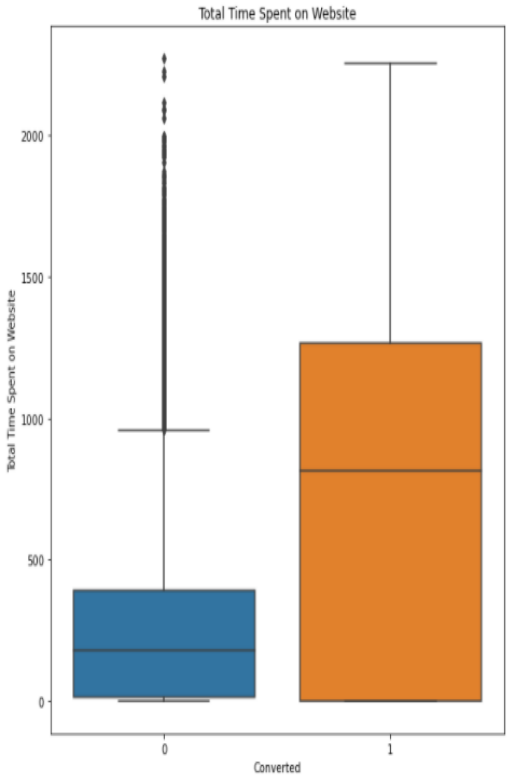
Data Cleaning

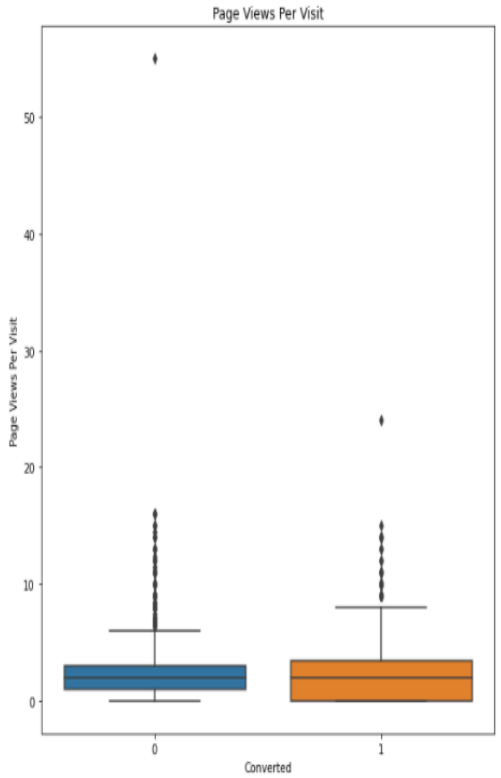
- Checked the sample data from this dataset
- Shape of the Dataset: **(9240, 37)**
- Removed all the columns that has more than 40% values as **null**.
- Removed ‘Prospect ID’ and ‘Lead Number’ as they were just identifiers.
- Removed all the columns where just one value had a much higher share like ‘Country’ etc.
- Imputed null values with ‘mode’ of the column wherever required.
- For some columns like ‘Occupation’ and ‘Specialization’ where we saw high share of nulls, we replaced null values with ‘Not Provided’ or ‘Not Specified’
- In order to avoid class imbalance, we grouped lower frequency values.



Total Visits, Time Spent And Page Views

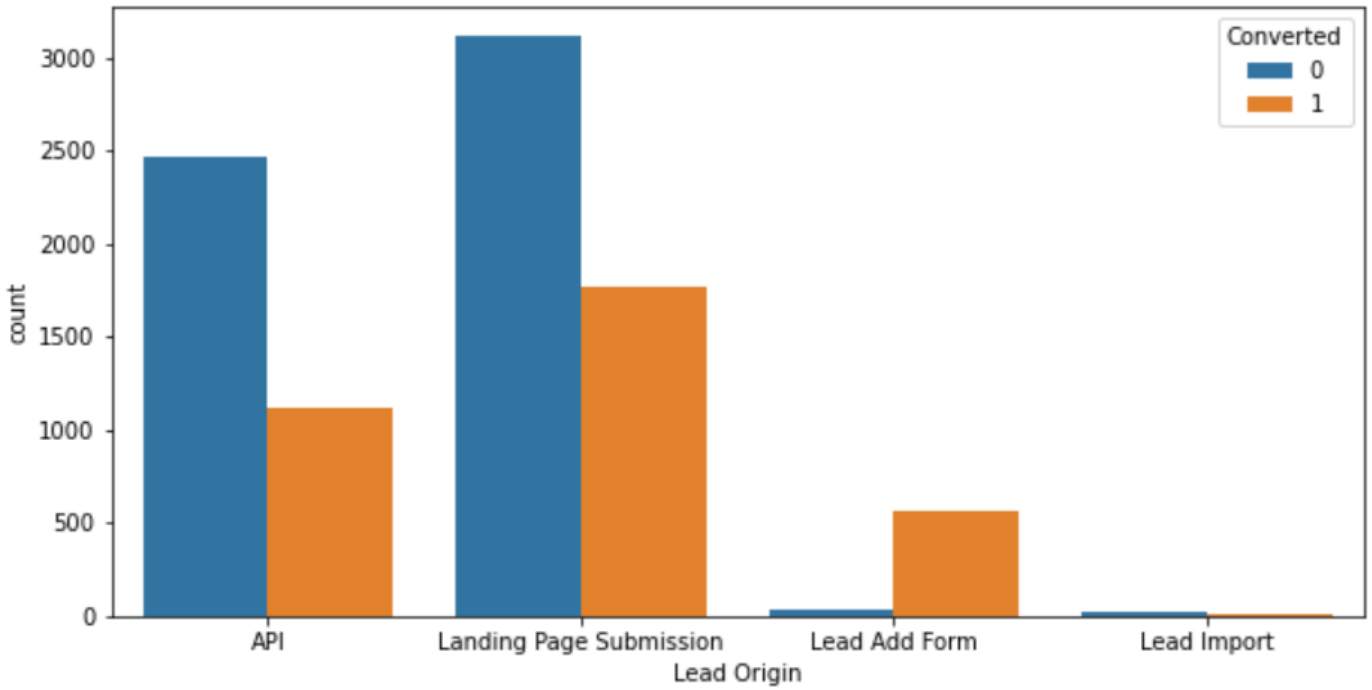






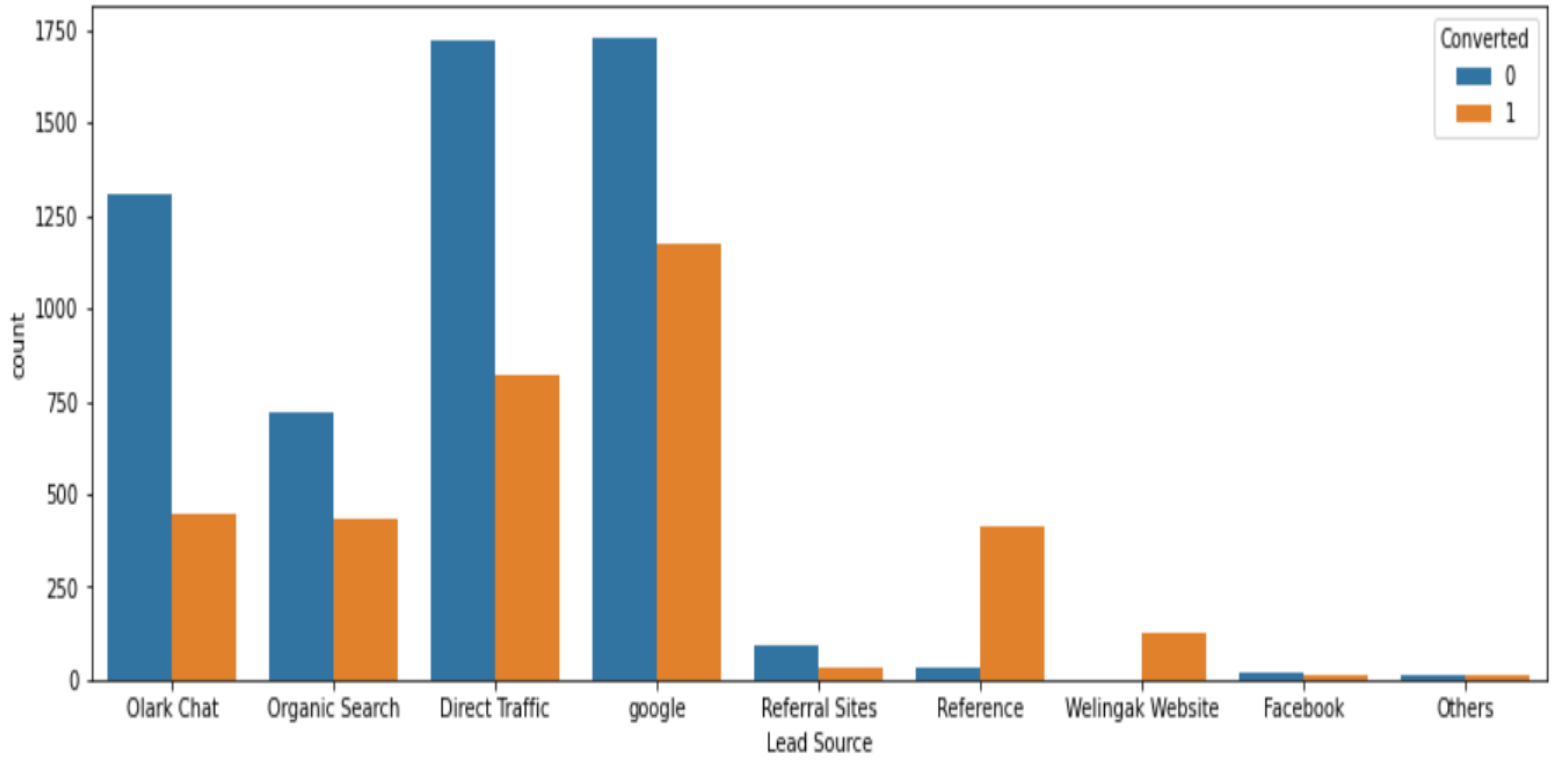
- People who have converted have higher number of visits, spent more time on website and also have higher page views.

Distribution of Lead Origin



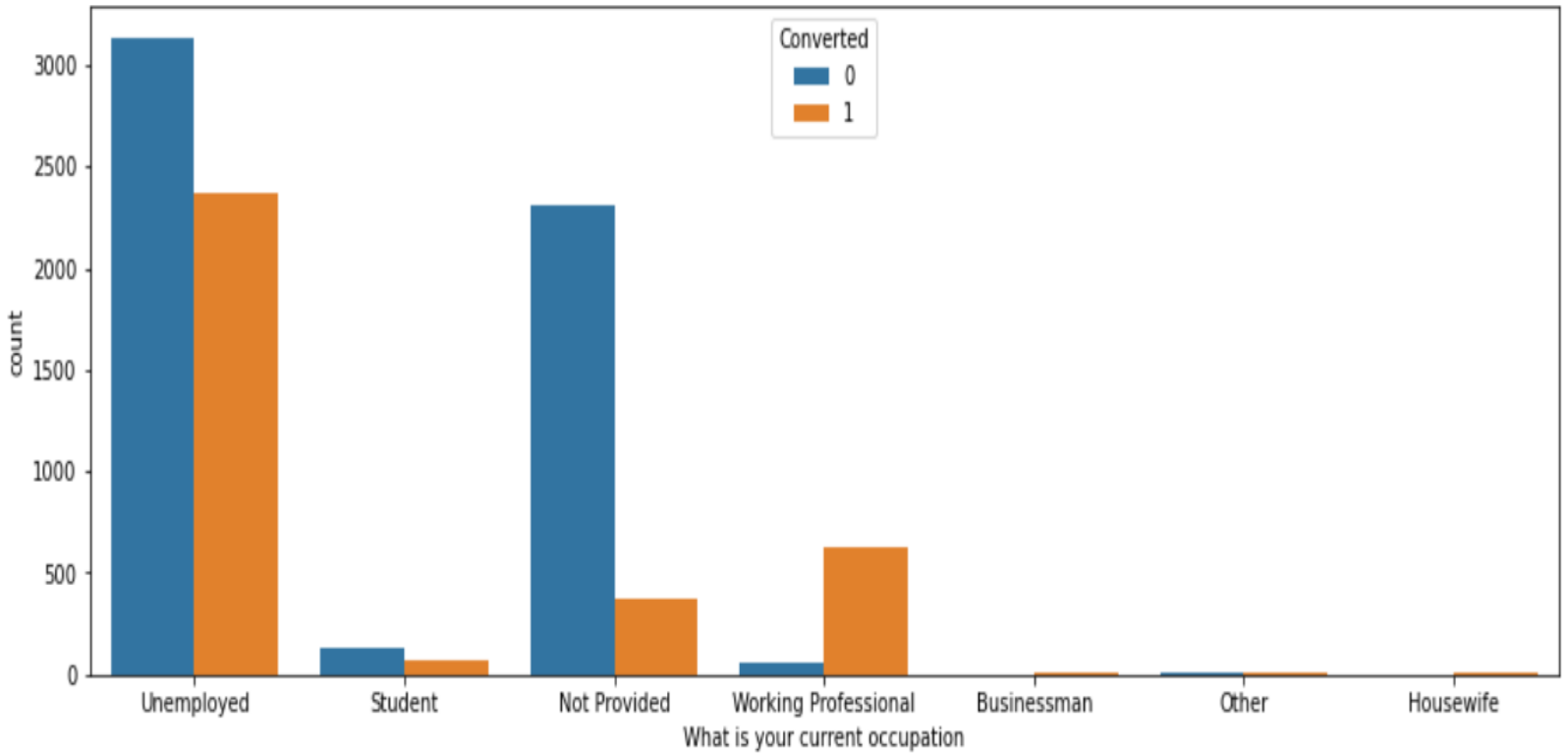
- A person who has done 'Lead Add Form' are very much likely to convert.

Distribution of Lead Sources



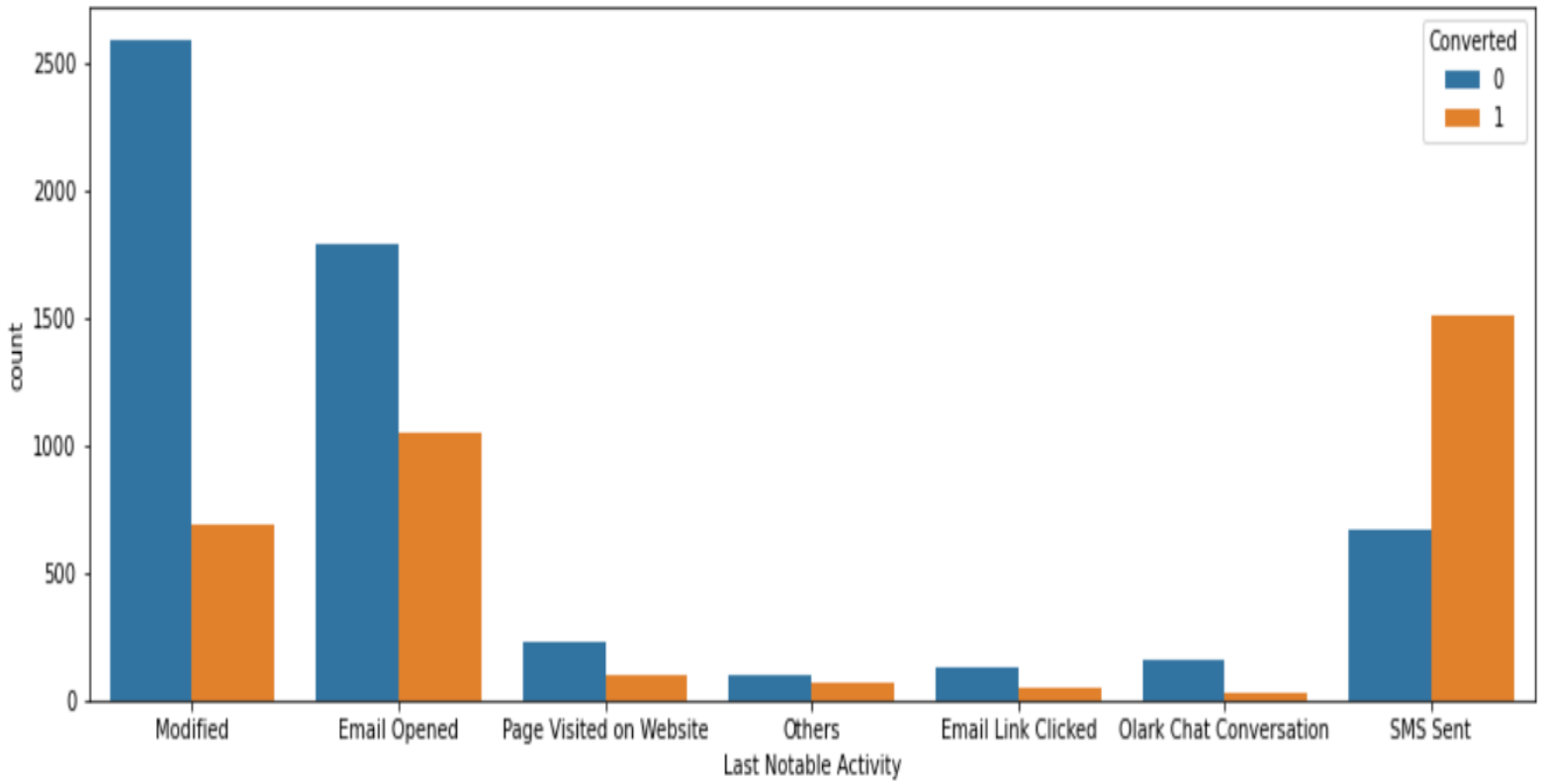
- Google has the highest number of leads with a good conversion rate

Distribution of Occupation



- Majority of the people are 'Unemployed'. It is surprising to see that 'Working Professional' has a really good conversion rate.

Distribution of Last Notable Activity



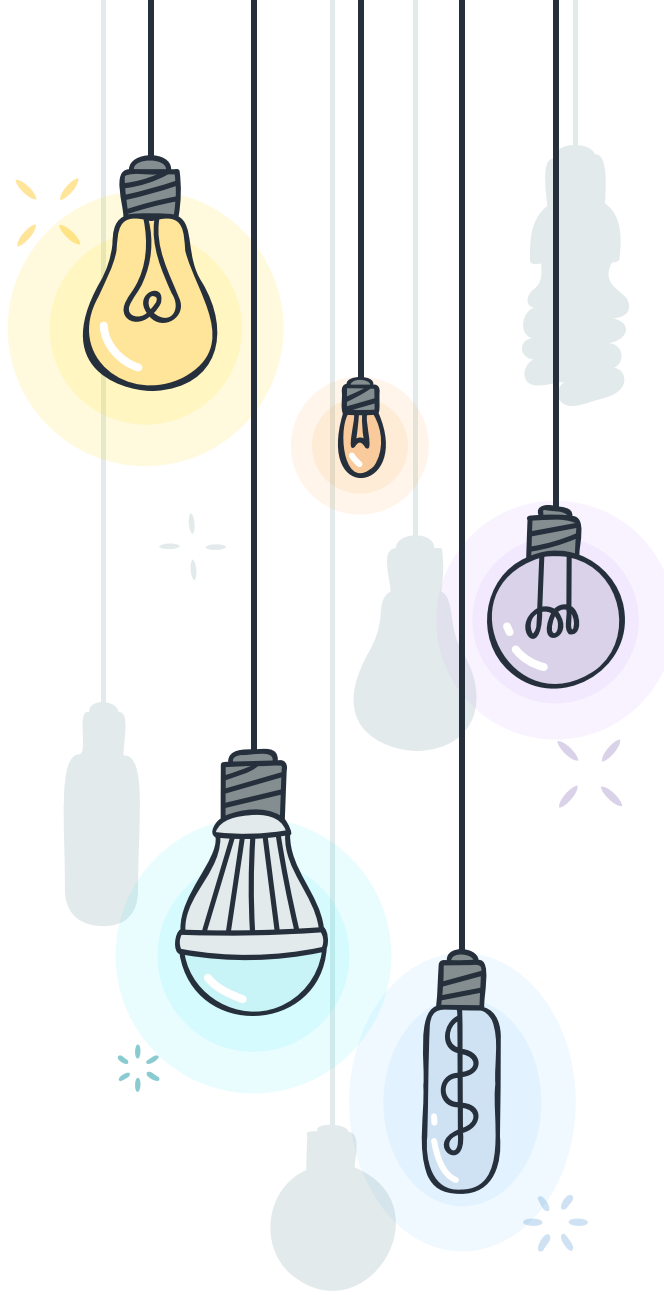
- 'SMS Sent' has the best conversion rate.

Creating Dummy Variables

- We created dummy variables using ‘get_dummies’ command for categorical variables having more than two unique values.
- We then dropped all the original columns from the dataset.

```
# Creating dummy variables using the 'get_dummies' command
dummy = pd.get_dummies(lead_data[['Lead Origin', 'What is your current occupation']], drop_first=True)
lead_data = pd.concat([lead_data, dummy], axis = 1)

dummy_lead_source = pd.get_dummies(lead_data['Lead Source'], prefix = 'Lead Source')
dummy_lead_source = dummy_lead_source.drop(['Lead Source_Others'], 1)
lead_data = pd.concat([lead_data, dummy_lead_source], axis = 1)

dummy_activity = pd.get_dummies(lead_data['Last Activity'], prefix = 'Last Activity')
dummy_activity = dummy_activity.drop(['Last Activity_Others'], 1)
lead_data = pd.concat([lead_data, dummy_activity], axis = 1)

dummy_notable_activity = pd.get_dummies(lead_data['Last Notable Activity'], prefix = 'Last Notable Activity')
dummy_notable_activity = dummy_notable_activity.drop(['Last Notable Activity_Others'], 1)
lead_data = pd.concat([lead_data, dummy_notable_activity], axis = 1)


dummy_specialization = pd.get_dummies(lead_data['Specialization'], prefix = 'Specialization')
dummy_specialization = dummy_specialization.drop(['Specialization_Not Specified'], 1)
lead_data = pd.concat([lead_data, dummy_specialization], axis = 1)

dummy_tags = pd.get_dummies(lead_data['Tags'], prefix = 'Tags')
dummy_tags = dummy_tags.drop(['Tags_Not Specified'], 1)
lead_data = pd.concat([lead_data, dummy_tags], axis = 1)
```

```
# Dropping the variables for which the dummy variables have been created

lead_data = lead_data.drop(['Lead Origin', 'Lead Source', 'Last Activity',
                   'Specialization', 'What is your current occupation', 'Tags', 'Last Notable Activity'], 1)
```



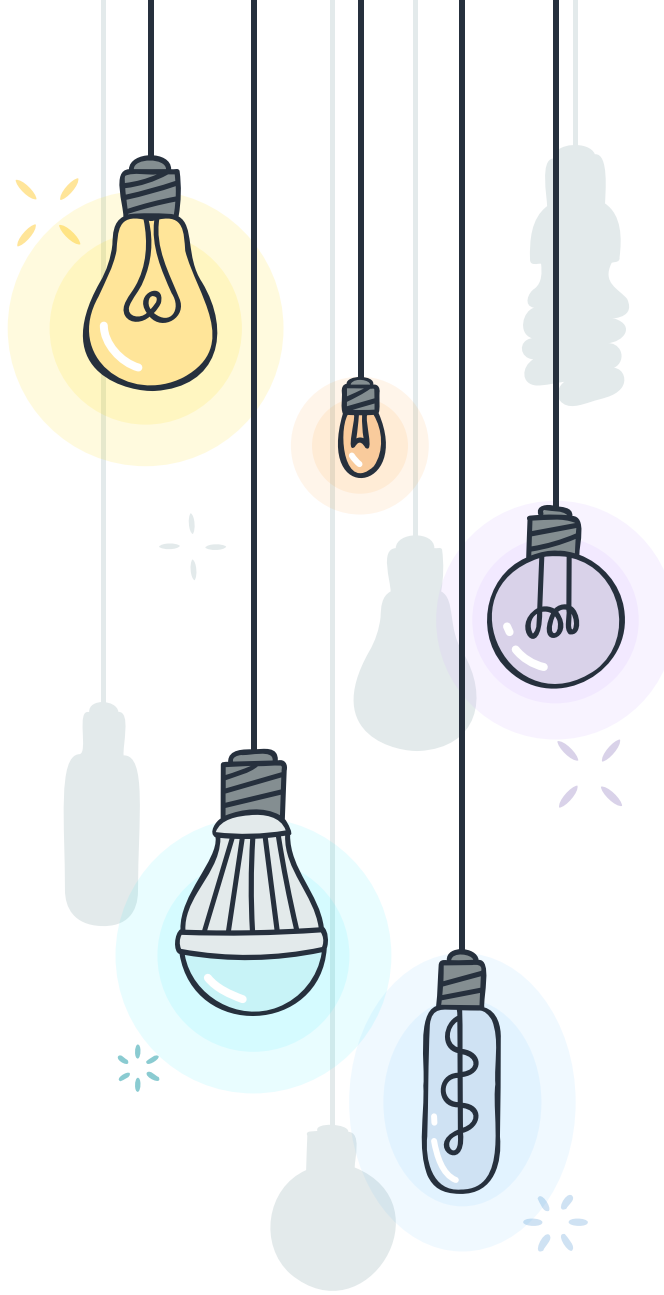
Model Building

Model Building

- We divided the data into train and test set and then scaled the training set using StandaradScaler.
- We then used RFE to get the top 15 features and then ran Logistic Regression using those columns.
- The p-value of all the variables were less than 0.05 and VIF were also less than 5.

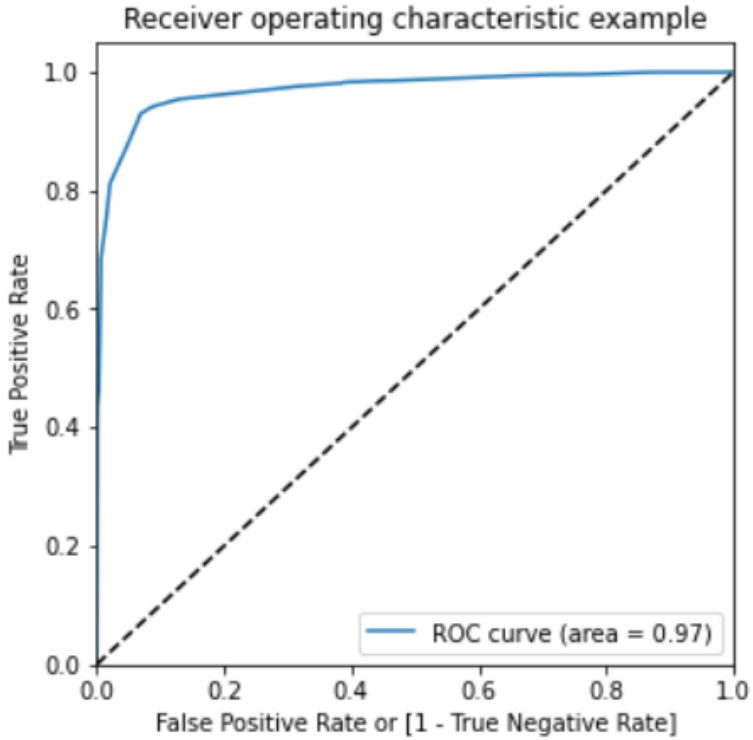
	coef	std err	z	P>|z|	[0.025	0.975]
const	0.1722	0.107	1.613	0.107	-0.037	0.381
What is your current occupation_Not Provided	-2.3789	0.131	-18.167	0.000	-2.636	-2.122
Lead Source_Welingak Website	2.7246	0.739	3.686	0.000	1.276	4.173
Last Activity_SMS Sent	2.1021	0.120	17.547	0.000	1.867	2.337
Last Notable Activity_Modified	-1.4333	0.127	-11.311	0.000	-1.682	-1.185
Tags_Already a student	-4.4228	0.590	-7.494	0.000	-5.580	-3.266
Tags_Closed by Horizzon	5.3749	0.724	7.427	0.000	3.956	6.793
Tags_Graduation in progress	-2.4171	0.498	-4.852	0.000	-3.394	-1.441
Tags_Interested in full time MBA	-3.0734	0.602	-5.102	0.000	-4.254	-1.893
Tags_Interested in other courses	-3.3362	0.361	-9.249	0.000	-4.043	-2.629
Tags_Lost to EINS	6.3048	0.734	8.594	0.000	4.867	7.743
Tags_Not doing further education	-4.4700	1.020	-4.381	0.000	-6.470	-2.470
Tags_Other_Tags	-3.6146	0.305	-11.836	0.000	-4.213	-3.016
Tags_Ringing	-4.9080	0.253	-19.404	0.000	-5.404	-4.412
Tags_Will revert after reading the email	3.1322	0.194	16.114	0.000	2.751	3.513
Tags_switched off	-5.3836	0.602	-8.938	0.000	-6.564	-4.203

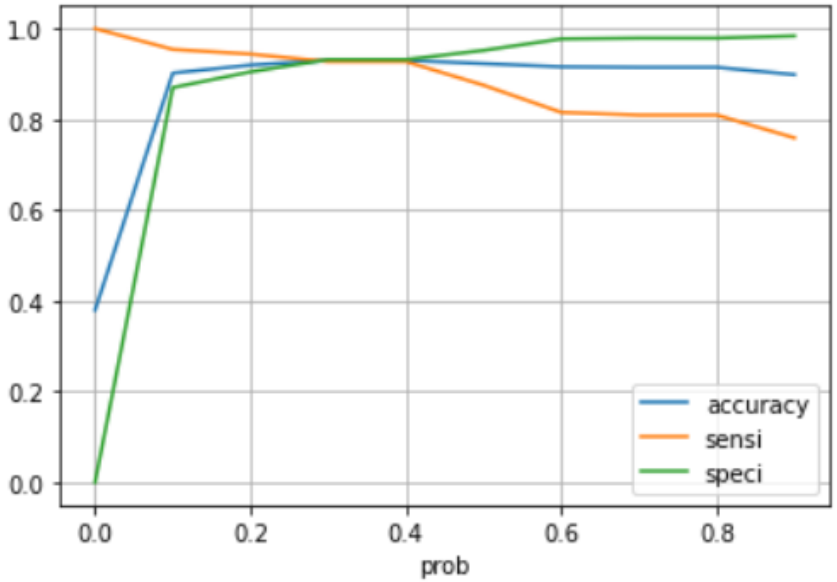
	Features	VIF
5	Tags_Closed by Horizzon	1.06
1	Lead Source_Welingak Website	1.04
9	Tags_Lost to EINS	1.04
10	Tags_Not doing further education	1.04
11	Tags_Other_Tags	1.04
6	Tags_Graduation in progress	1.03
14	Tags_switched off	1.03
7	Tags_Interested in full time MBA	1.02
8	Tags_Interested in other courses	0.38
4	Tags_Already a student	0.20
3	Last Notable Activity_Modified	0.18
13	Tags_Will revert after reading the email	0.11
2	Last Activity_SMS Sent	0.10
12	Tags_Ringing	0.07
0	What is your current occupation_Not Provided	0.05



Model Evaluation - 1

- We checked the ROC curve and area under the curve was 0.97.
- In order to find optimal cutoff, we calculated accuracy, sensitivity and specificity for various probability cutoffs. **0.26** came out to be optimal cutoff point.
- Accuracy: **92.27%**, Sensitivity: **92.72%**, and Specificity: **93.17%**





Model Evaluation - 2

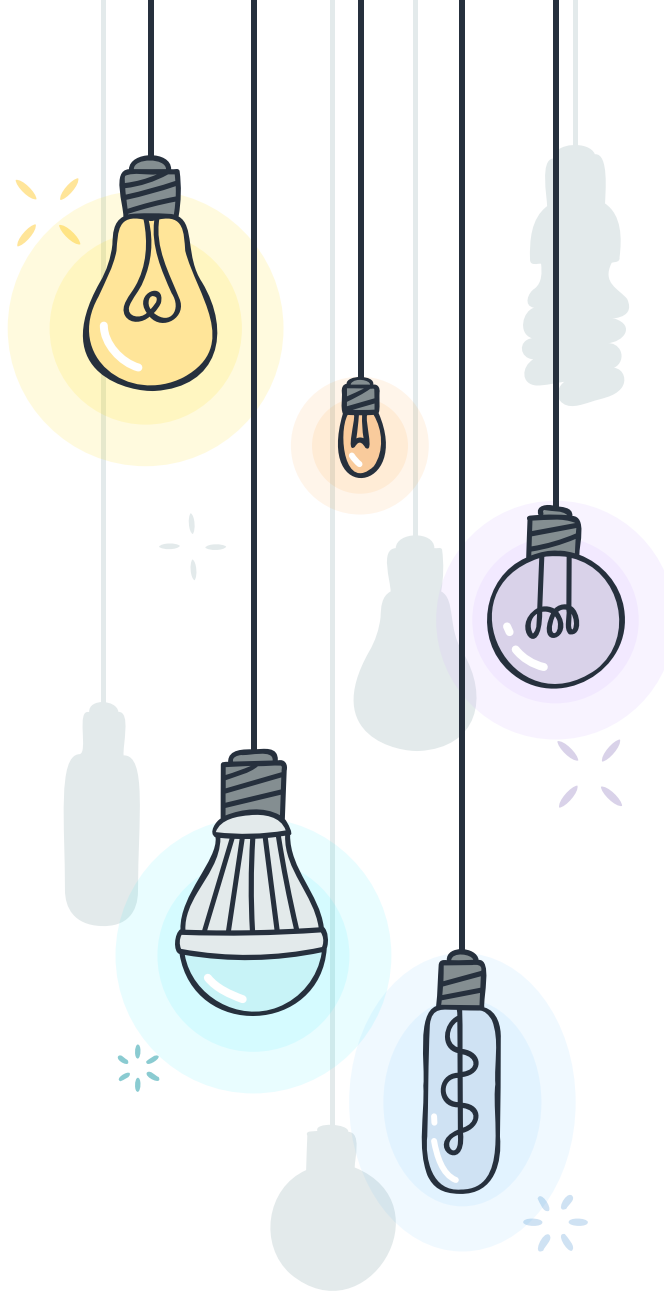
- We then made predictions on test set.
- Accuracy: **93.23%**, Sensitivity: **91.94%**, and Specificity: **94.02%**
- We saw that the accuracy along with sensitivity and specificity of the training set and testing set was very close.

Train Data:

*Accuracy: 93.0%
*Sensitivity: 92.72%
*Specificity: 93.17%

Test Data:

*Accuracy: 93.23%
*Sensitivity: 91.94%
*Specificity: 94.02%



The company should focus more on ‘Working Professionals’ as their conversion rate is very high.

Customers identified through ‘Add Lead Form’ are very much likely to convert.

The company should also focus more on people who have specialization in a Management course

People who are spending more time on the website also needs to be targeted aggressively

Tags, LeadSource and LastActivity are the most important factors causing conversions.


THANK YOU