Group Case Study 2
on
Lead Scoring
using
Logistic Regression Model

By :

1. **Yogalakshmi Pasupathy (Reg. Email ID : laku1511@gmail.com)**
2. **Malathi Ashok (Reg. Email ID : malathiashok99@gmail.com)**

From :

DS C24 Sept 2020 Batch of upGrad

Table of Contents

Contents	Page Number
Problem Statement	3
Problem Solving Methodology	4
Model Building Approach	5
Data Visualization	6-9
Model Building on Train dataset	10
Metrics Evaluation on Train dataset	11-12
Prediction & Metrics Evaluation on Test dataset	13
Important Features and Metrics Summary	14
Recommendations	15
Word Problem Solution	16-17

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. 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.

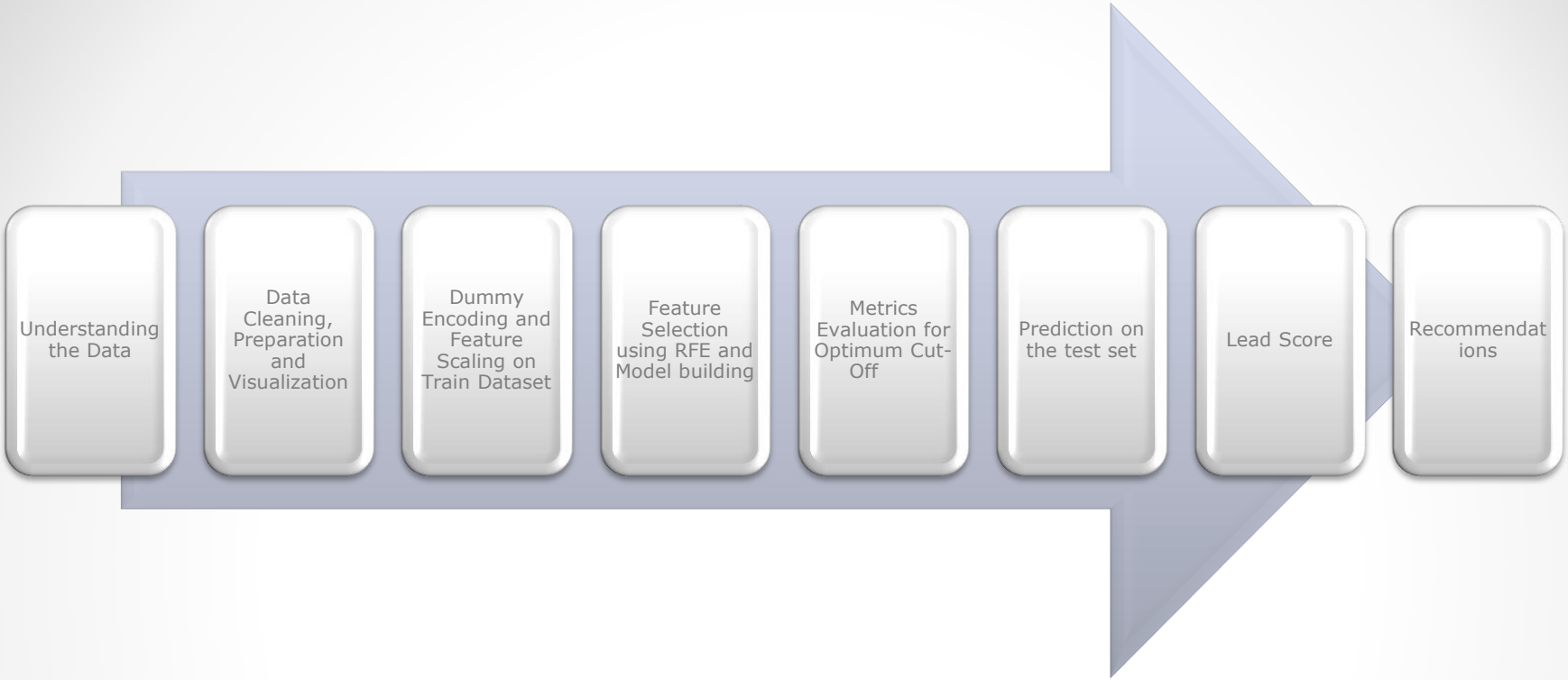
- Once these leads are acquired, employees from the sales team start making calls, writing emails, etc. Through this process, some of the leads get converted while most do not. The typical lead conversion rate at X education is around 30%.

- X Education needs help in selecting the most promising leads, i.e. the leads that are most likely to convert into paying customers.

- The company needs a model wherein a lead score is assigned 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%.

Problem Solving Methodology



Model Building Approach

Understanding the Data

- There are in total 9240 records with 37 columns in the Lead Dataset. There are lots of missing and null values in the dataset. Also it contains the records with the mix of integer, float and object datatype. There are no duplicate values in the dataset

Data Cleaning, Preparation and Visualization

- High % of Null value columns, The columns which had single value, Binary valued column which were highly skewed were dropped
- Replaced the Null/Select with relevant values and Binary conversion were also done
- Outliers are identified and treatment was done
- Numerical/Categorical column were analysed visually. The columns having single digit value were grouped together or dropped based on its significance/existence of correlation

Dummy Encoding and Feature Scaling on Train Dataset

- Dummy features for Categorical columns were created using one-hot encoded. Features with least significance were dropped
- Test-Train split were done for 30-70% percentage
- Scaling was done to Numerical features using Min-Max scaler

Feature Selection using RFE and Model building

- 15 Feature were selected using RFE. Model Building was made through Logistic Regression and several Iterations using Statsmodel were performed by eliminating high p-value features as VIFs were under control in train dataset

Metrics Evaluation for Optimum Cut-Off

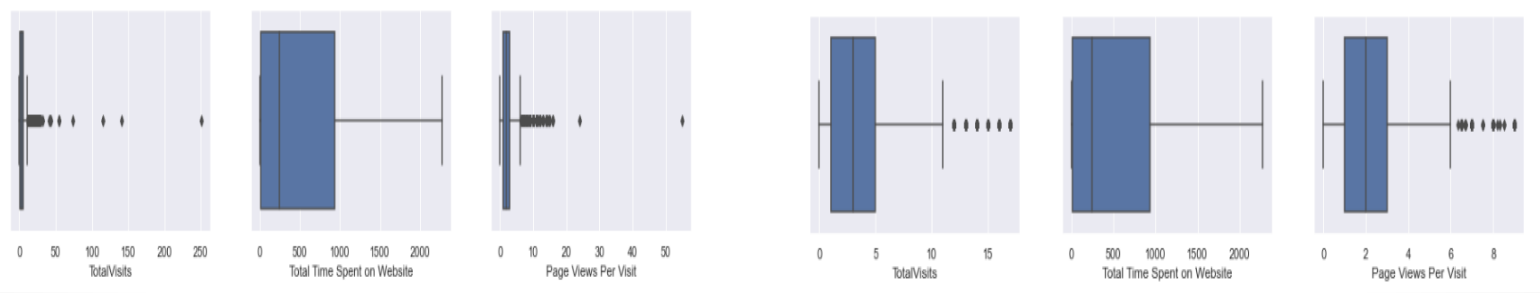
- Using ROC curve and various probabilities iterations, Optimal cut-off was identified as "0.3"
- Once the model was finalized, the metrics were calculated Accuracy, Sensitivity and Specificity

Prediction on the test set, Lead Score, Recommendations

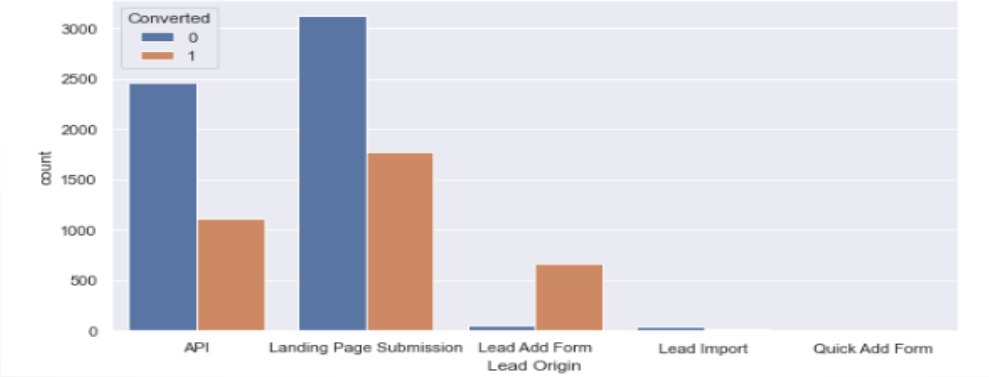
- Prediction was made after scaling on test dataset. The results were matching with the train dataset
- Lead Score was calculated for Test and Train dataset
- Final recommendations were provide for selecting the feature which will convert into a potential/hot lead

Data Visualization

- **Outlier Analysis** has been done on the Total Time Spent on Website, TotalVisits, Page Views Per Visit and treatment done by capping 99-1% quantile range on TotalVisits, Page Views Per Visit

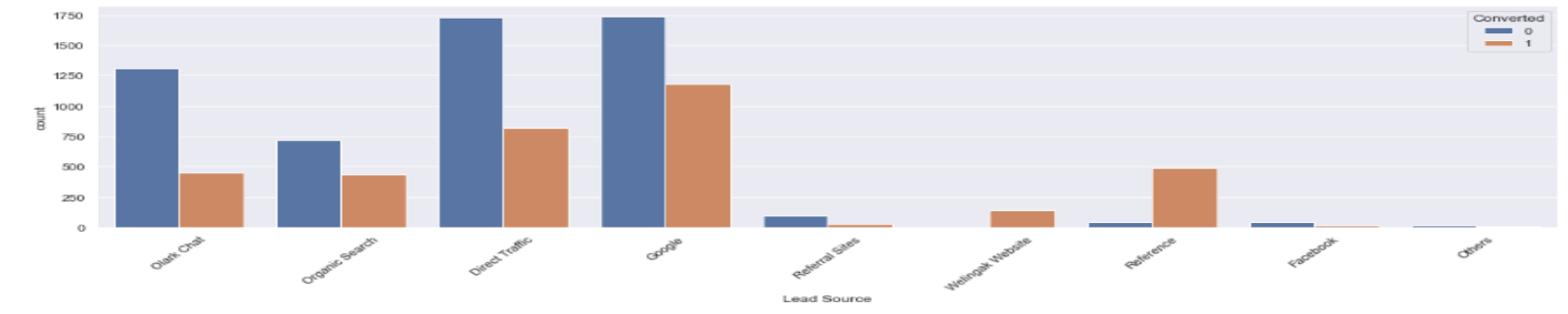


- **Lead Origin** - There were more number of leads from Landing Page Submission, API and Lead Add Form Lead Origin where in Lead Add Form Lead Origin the conversion rate has more when compared to other two main categories.

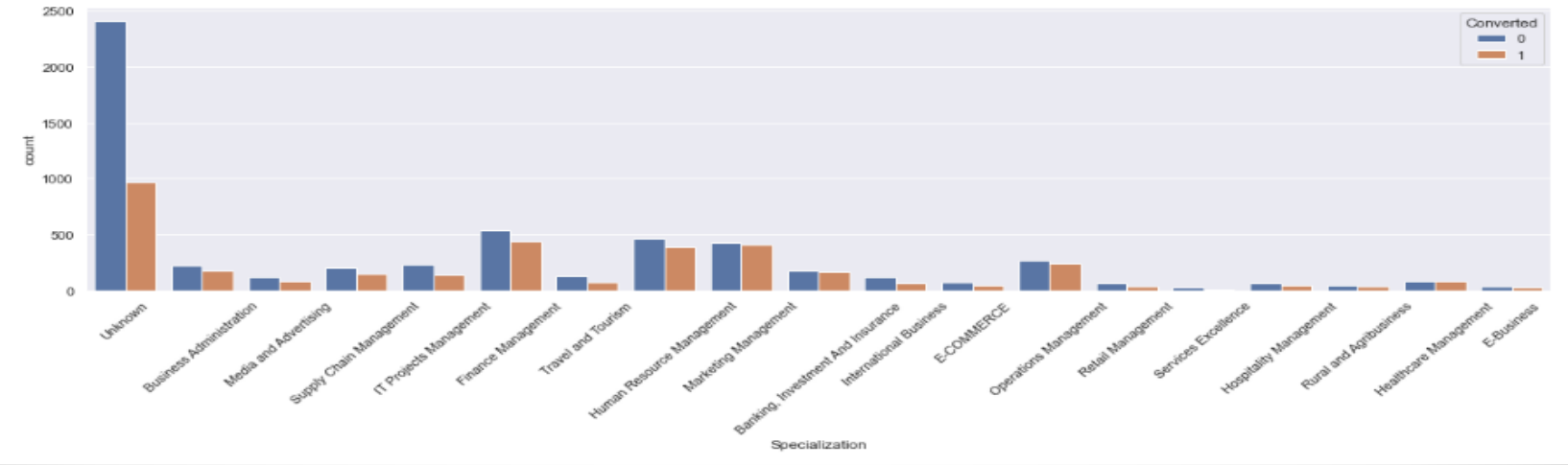


Data Visualization – contd.

- **Lead Source -** There were more number of leads from Google, Direct Traffic, Olark Chat, Organic Search, Reference where in for the sources Welingak Website and Reference the conversion rate had more when compared to other sources.

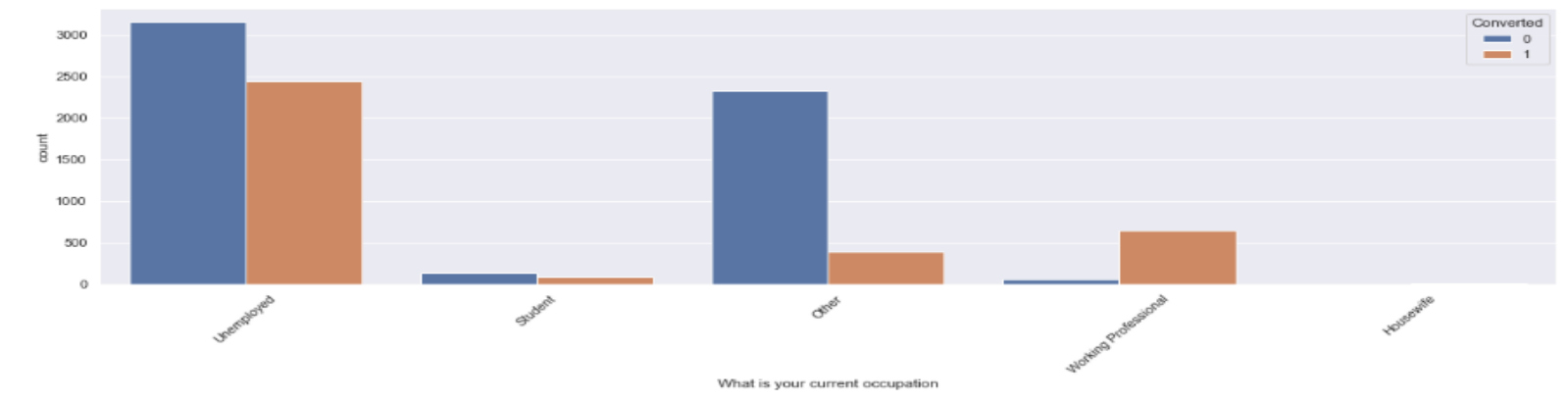


- **Specialization** - There were more number of leads from Finance Management, Human Resource Management, Marketing Management, Operations Management, Business Administration.

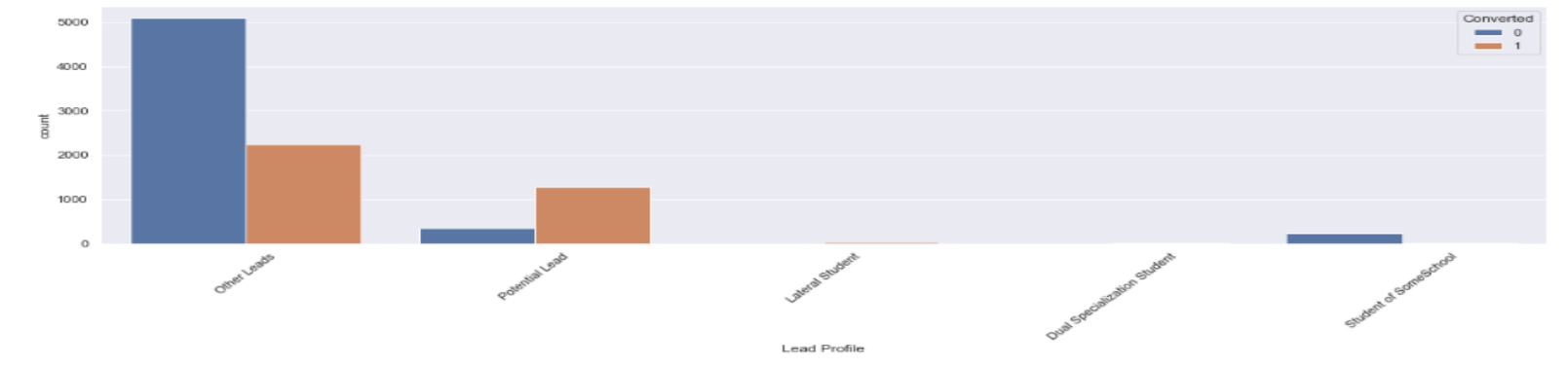


Data Visualization – contd.

- **What is your current occupation -** There are more leads from Unemployed, Working Professional occupation

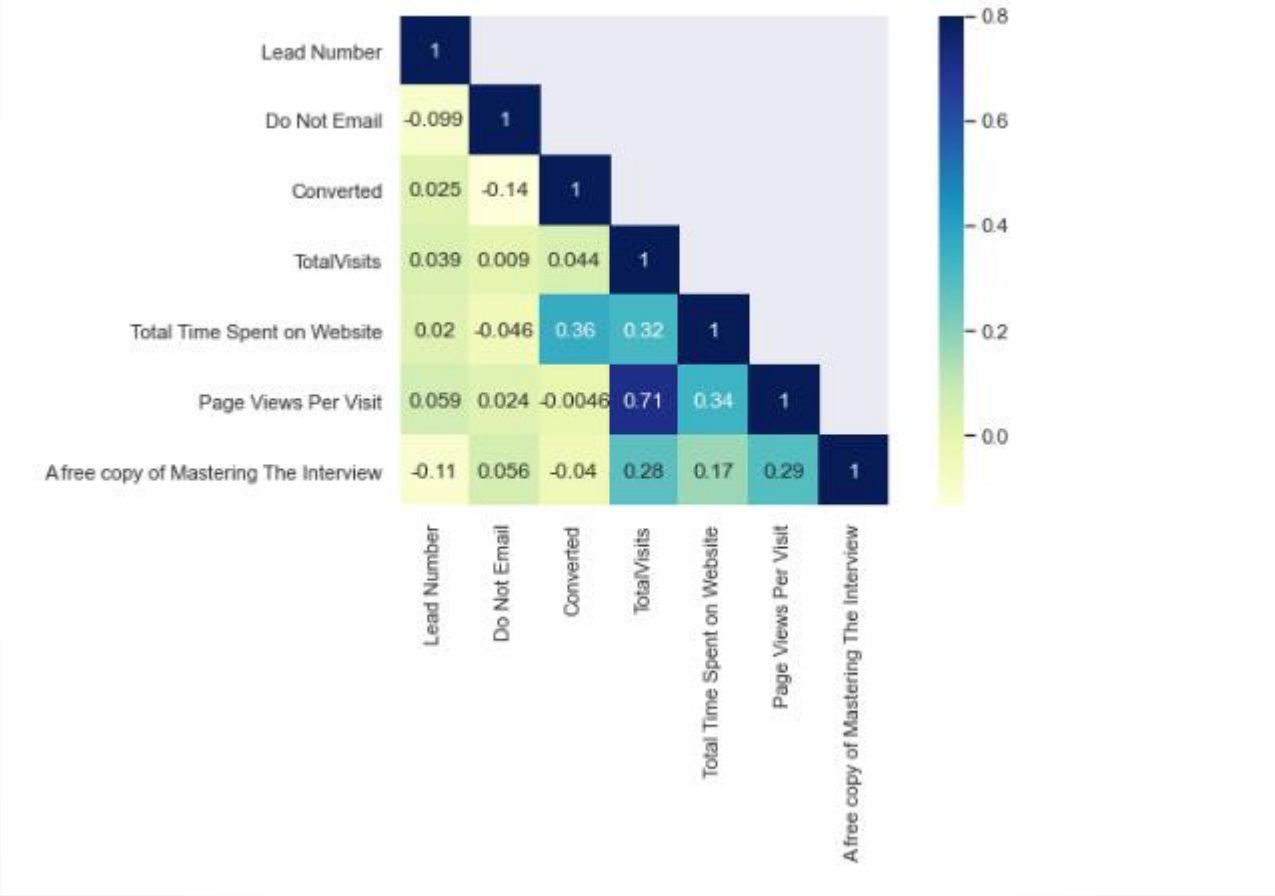


- **Lead Profile** - There were more number of leads from Potential Lead, Student of SomeSchool and in that Potential Lead got converted more than other profiles



Data Visualization – contd.

- **Heat Map for correlation check -** The Total Visit columns has positive correlation with Page Views Per Visit.
-



Model Building on Train dataset

- After creation of dummy variables using one-hot encoder, Test-Train split and feature scaling using MinMax Scaler, RFE feature elimination happened

- Initial feature selection using RFE

```
'Lead Origin_Lead Add Form', 'Lead Source_Olark Chat',
'Lead Source_Welingak Website', 'Country_Qatar', 'Country_Saudi Arabia',
'Specialization_Hospitality Management',
'What is your current occupation_Housewife',
'What is your current occupation_Working Professional',
'What matters most to you in choosing a course_Better Career Prospects',
'Lead Profile_Dual Specialization Student',
'Lead Profile_Lateral Student', 'Lead Profile_Potential Lead',
'Lead Profile Student of SomeSchool'],
```

- Final feature after model building using Logistic regression model

```
['Do Not Email', 'Total Time Spent on Website',
 'Lead Origin_Lead Add Form', 'Lead Source_Olark Chat',
 'Lead Source_Welingak Website', 'Specialization_Hospitality Management',
 'What is your current occupation_Working Professional',
 'What matters most to you in choosing a course_Better Career Prospects',
 'Lead Profile_Lateral Student', 'Lead Profile_Potential Lead',
 'Lead Profile Student of SomeSchool'],
```

Metrics Evaluation on Train dataset

- The metrics Accuracy, Sensitivity, Specificity, False Positive Rate, Positive Predictive Value and Negative Predictive Value were calculated from the Confusion Metrics

- **Confusion Metrics**

Predicted/ Actual	Not_Converted	Converted
Not_Converted	TN = 3543	FP = 459
Converted	FN = 810	TP = 1656

- **Where :** TP : True Positive, TN : True Negative, FP : False Positive and FN : False Negative

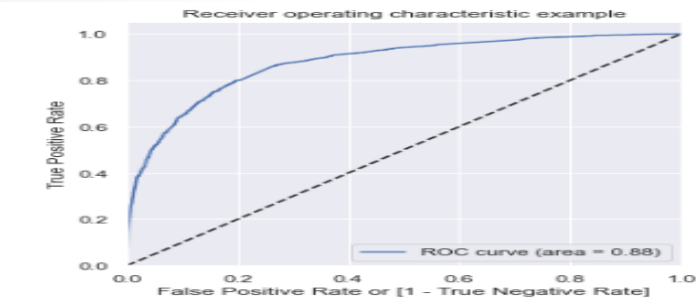
- **Metrics values**

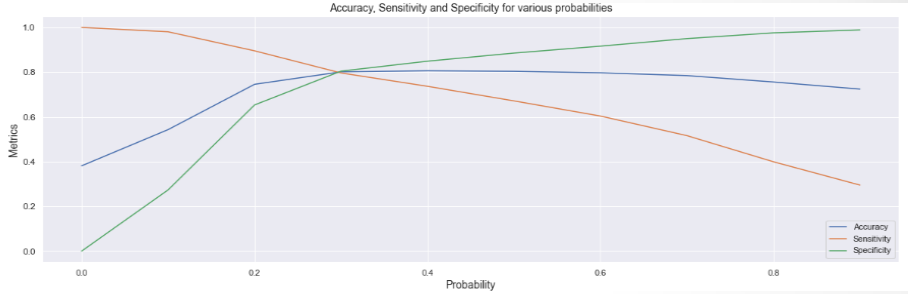
Metrics	Accuracy	Sensitivity	Specificity	False Positive Rate	Positive Predictive Value	Negative Predictive Value
Train before Cut-off	0.80	0.67	0.89	0.11	0.78	0.81

Metrics Evaluation on Train dataset

- ROC curve and different probability cut-off iterations were performed. Optimum probability cut-off was selected as "0.3"

- **ROC and Accuracy, Sensitivity, Specificity Curve** as follows :





- **Confusion Metrics**

Predicted/ Actual	Not_Converted	Converted
Not_Converted	TN = 3220	FP = 782
Converted	FN = 502	TP = 1964

- **Metrics values**

Metrics	Accuracy	Sensitivity	Specificity	False Positive Rate	Positive Predictive Value	Negative Predictive Value	Precision	Recall
Train after Cut-off	0.80	0.80	0.80	0.20	0.72	0.87	0.72	0.80

Prediction / Metrics Evaluation on Test dataset and Lead Score

- The results of test dataset were aligned with the train dataset

- **Confusion Metrics**

Predicted/ Actual	Not_Converted	Converted
Not_Converted	TN = 1363	FP = 314
Converted	FN = 228	TP = 867

- **Metrics values**

Metrics	Accuracy	Sensitivity	Specificity	False Positive Rate	Positive Predictive Value	Negative Predictive Value	Precision	Recall
Test	0.80	0.79	0.81	0.19	0.73	0.86	0.73	0.79

- **Lead Conversion percentage** on Train : 0.80 &Test dataset is : 0.79

- **Lead Score** was assigned on Train and Test dataset by multiplying conversion probabilities with 100. The customers with higher lead score have a higher conversion chance into hot/potential lead and the customers with lower lead score have a lower conversion chance into cold lead

Important Features and Metrics Summary

- **Coefficient** & **Relative Coefficient** of the **TOP** features as follows :

Total Time Spent on Website	4.510669	100.000000
Lead Origin_Lead Add Form	3.155602	69.958628
Lead Source_Welingak Website	2.548142	56.491443
What is your current occupation_Working Professional	2.347765	52.049145
Lead Profile_Lateral Student	2.321781	51.473087
Lead Profile_Potential Lead	1.362290	30.201515
What matters most to you in choosing a course_Better Career Prospects	1.041837	23.097181
Lead Source_Olark Chat	1.012415	22.444905
Specialization_Hospitality Management	-0.875781	-19.415769
Do Not Email	-1.201757	-26.642543
Lead Profile_Student of SomeSchool	-2.498673	-55.394727

- **Summary of Metrics values**

Metrics	Accuracy	Sensitivity	Specificity	False Positive Rate	Positive Predictive Value	Negative Predictive Value	Precision	Recall
Train before Cut-off	0.80	0.67	0.89	0.11	0.78	0.81		
Train after Cut-off	0.80	0.80	0.80	0.20	0.72	0.87	0.72	0.80
Test	0.80	0.79	0.81	0.19	0.73	0.86	0.73	0.79

Recommendations

- X-Education has a better chance of converting into a potential/hot lead when they focus on contacting people who fall into the below feature:

1. Total Time Spent on Website
2. Lead Origin_Lead Add Form
3. Lead Source_Welingak Website
4. What is your current occupation_Working Professional
5. Lead Profile_Lateral Student

- **Conclusion**
  - While we have checked both Sensitivity-Specificity as well as Precision and Recall Metrics, we have considered the optimal cut off based on Sensitivity and Specificity for calculating the final prediction.
  - Accuracy, Sensitivity and Specificity values of test set were around 80%, which were approximately closer to the respective values calculated using training dataset.
  - Also the lead score calculated in the trained set of data shows the conversion rate on the final predicted model is around 80%
  - Hence overall this model seems to be good.

Word Problem Solution

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

- The TOP three variables in the model which contribute most towards the probability of a lead getting converted were as follows : **a. Total Time Spent on Website, b. Lead Add Form (from Lead Origin) and c. Welingak Website (from Last Source)**

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?

- 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 were as follows : **a. Lead Add Form (from Lead Origin) b. Welingak Website (from Last Source) and c. Working Professional (from What is your current occupation)**

Word Problem Solution – contd.

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.

- In the below image, the final prediction is calculated based on an optimal cut off value (Our model's cut-off is 0.3). In order to make the sales aggressive, the Sales Team may contact all the leads which have a conversion probabilty **below the optimum cut off (0.3)** (highlighted in yellow). Since they have a low probability, the Sales Team have to go aggressive with them and pursue hard for converting them into Potentials Customers.

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.

- In order to minimize the rate of useless phone calls, the company may contact all the leads which have a **higher probabilty** (highlighted in yellow color) like in our case greater than 0.3. Since there is a scarcity of Sales Staff, it is recommended that they focus on higher probability leads which have a higher chance of conversion.