SUMMARY REPORT ON LEAD SCORE CASE STUDY

Aim:

A company X Education sells online courses to industry professionals. The company wishes to focus on clients who have filled a form for the course on the company's website. Once the customer's leads are acquired it goes through different processes to get these leads converted.

Company wants us to build a ML model to convert atleast 80% of leads into paying customers.

Approach:

Stages of building a ML model:

1. Data Understanding
2. Data Visualization
3. Data Preparation
4. Dummy Variable Creation
5. Test-Train Split
6. Standardization
7. Model Building
8. Feature Selection using RFE
9. Manual Feature Selection
10. Performance metrics on the model

Data Visualization Insights

After this, we proceeded with data visualisation and gathered insightful observations from the plots. The observations were:

1. There is no discernible gap between total visits and converted. Customers who spend more time on website are more likely to be converted. The median number of visits by converted customers as well as non-converted customers are same.
2. Maximum lead conversion happened:
    - Lead Origin as Landing Page Submission
    - leads prefer calls, digital advertisement and emails
    - Leads come through recommendations and by searching courses online
    - Leads who were unemployed and working professional
    - Leads who are visiting the site for better career prospects from city Mumbai
3. The company must not put some resources in the Lead Add form as it has least Converted counts.
4. Top sites from which positive leads are generated are
    - Olark Chat
    - Organic Search
    - Direct traffic
    - Google
5. The conversion rate from reference,Welingak Website are the highest

6. There is no discrenable inferences can be made between specialization chosen and conversion rate

Model Inferences

Model was built using logistic regression technique.

	coef	std err	z	P>|z|	[0.025	0.975]
const	-0.7009	0.057	-12.228	0.000	-0.813	-0.589
Do Not Email	-1.2385	0.159	-7.802	0.000	-1.550	-0.927
Total Time Spent on Website	1.0464	0.036	28.935	0.000	0.976	1.117
Specialization_Hospitality Management	-0.9593	0.324	-2.959	0.003	-1.595	-0.324
What is your current occupation_Working Professional	2.5686	0.187	13.712	0.000	2.201	2.936
Lead Profile_Potential Lead	1.6081	0.093	17.305	0.000	1.426	1.790
Lead Profile_Student of SomeSchool	-1.9887	0.430	-4.625	0.000	-2.831	-1.146
Lead Origin_Lead Add Form	3.3090	0.179	18.456	0.000	2.958	3.660
TotalVisits_1-2_Visits	-0.8588	0.095	-9.022	0.000	-1.045	-0.672
TotalVisits_3-4_Visits	-0.6159	0.087	-7.116	0.000	-0.786	-0.446
Page Views Per Visit_5-6_Page_views_per_visit	-0.3946	0.121	-3.263	0.001	-0.632	-0.158

The coefficient values tell us how much the respective feature affects the probability of lead getting converted to customer. As you can see,

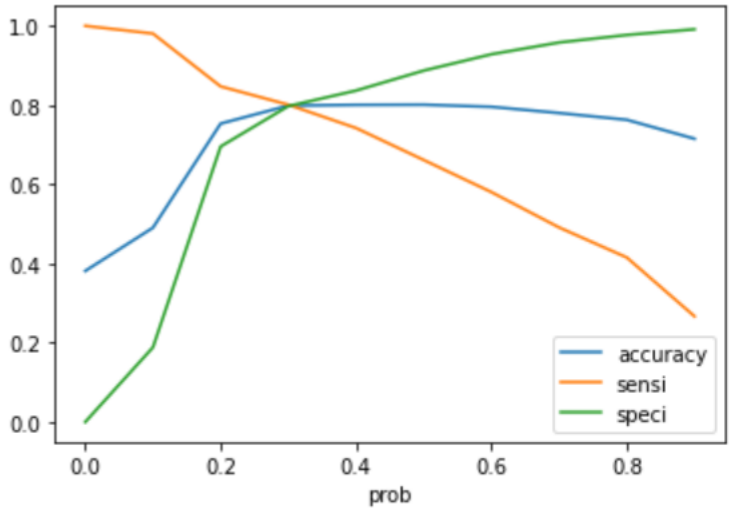
Lead Origin_Lead Add Form has highest coefficient value. This means that, if we target more leads with origin from Lead Add Form, probability of lead getting converted will highly increase.

On the other hand**, Lead Profile_Student** of SomeSchool has lowest coefficient value. This means that, if we don't target students of some school, probability of lead getting converted will increase as well. Instead, we can target working professionals as they contribute towards increasing probability of lead conversion.

Similarly, we can draw inferences on other features based on the coefficient values.

Assigning lead score

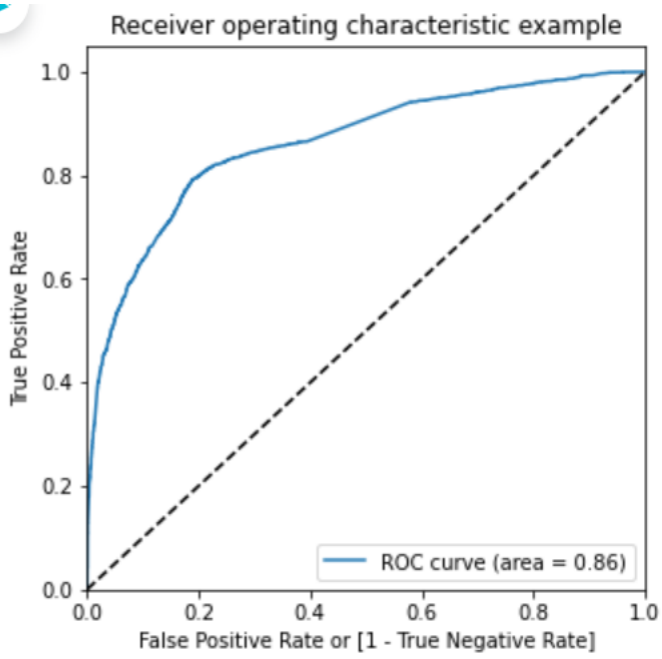
We moved on to finding the optimal threshold point, which gave us the boundary below above which a lead is taken to be converted or not. Threshold used is: **0.3**



As you can see sensitivity (Probability of being converted when lead is converted) and specificity (Probability of being not converted when lead is not converted) are almost equal at 0.3, hence we selected 0.3 as threshold.

Metrics

1. ROC Curve

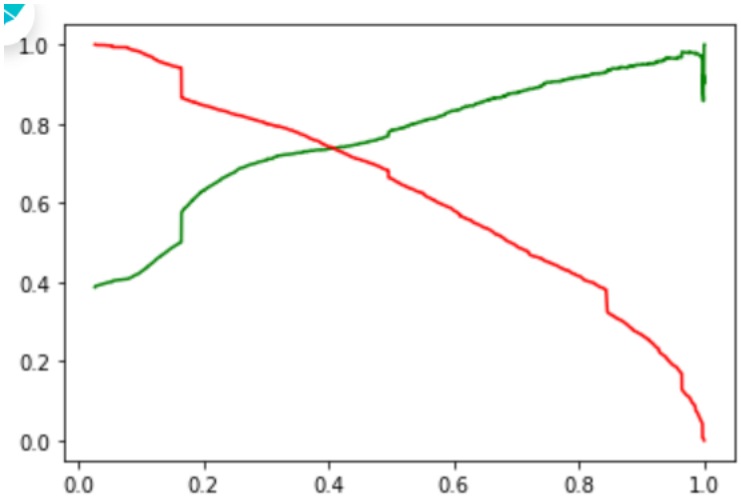


It shows the trade-off between sensitivity and specificity

As the curve of the ROC is more towards the upper left corner of the graph, it means the model is very good.

The value of area under the curve is 0.86 which is quite decent.

2. Precision and Recall Trade off



a) The recall is bit higher when the precision is high.

b) Recall is high when precision is low.

3. Model Evaluation on Test Data

```
Sensitivity is: 79.18%
Specificity is: 80.08%
Precision: 72.19%
Recall: 79.18%
Accuracy: 79.73%
```

4. Confusion matrix

Actual/Predicted	Not Converted	Converted
Not converted	1343	334
Converted	228	867

As we can see here model predicted True positive more than false positive. And recall it almost 80%, which means that lead conversion rate will be around 80% as well.

Important features for good conversion rate

1. Lead Origin_Lead Add Form

2. What is your current occupation-Working Professional

3. Lead Profile_Student of SomeSchool

4. Lead Profile_Potential Lead