



upGrad

LEAD SCORING ASSIGNMENT

Presented By –

Shivani Jain

Josna Cardoza,

Jyothula Sunil Kumar

Batch –

DS C67 Apr 2024



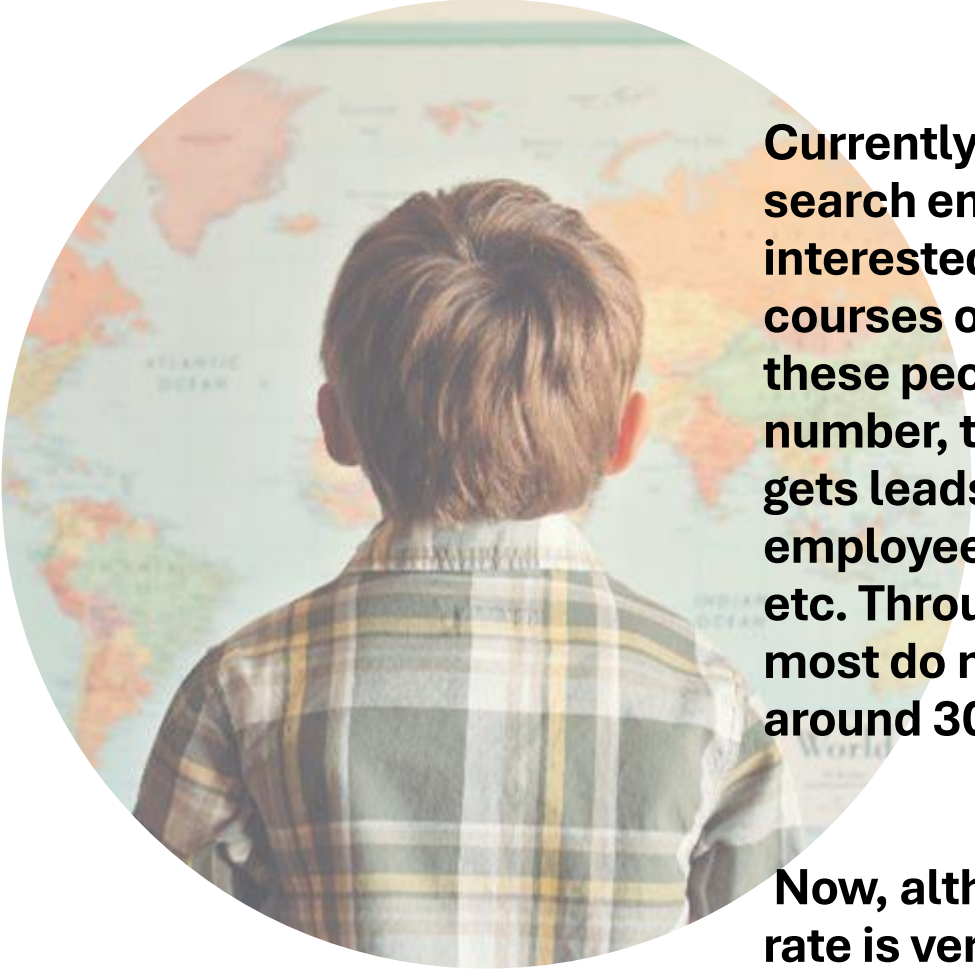
Agenda

- Introduction
- Business Understanding
- Business Objective
- Let's take a closer look at DATA
- Final tips & takeaways from the DATA
- Supervised ML model Logistic Regression
- Evaluation Metrics
- Suggestions



X Education

Business Problem



Currently, the company markets its courses on several websites and search engines like Google. Once these professionals who are interested in the courses 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. 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%.

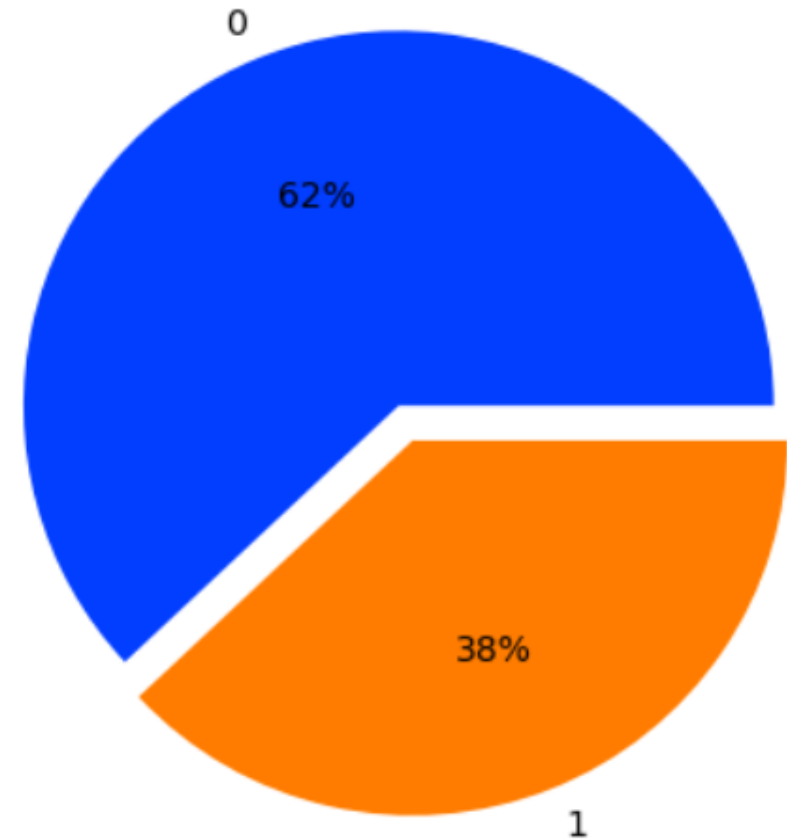
Now, although the company gets a lot of leads, its lead conversion rate is very poor. For example, if, say, they acquire 100 leads in a day, only about 30 of them are converted.

Business Objective

- The company wants to identify the most potential leads or Hot Leads
- To select the most promising leads, i.e. the leads that are most likely to convert into paying customers – they require a model.
- This ML model should assign a lead score to each of the leads such that the customers with a higher lead score have a higher conversion chance and the customers with a lower lead score have a lower conversion chance.

Current
Conversion
Rate

38%



1 - Converted
0 - Not Converted

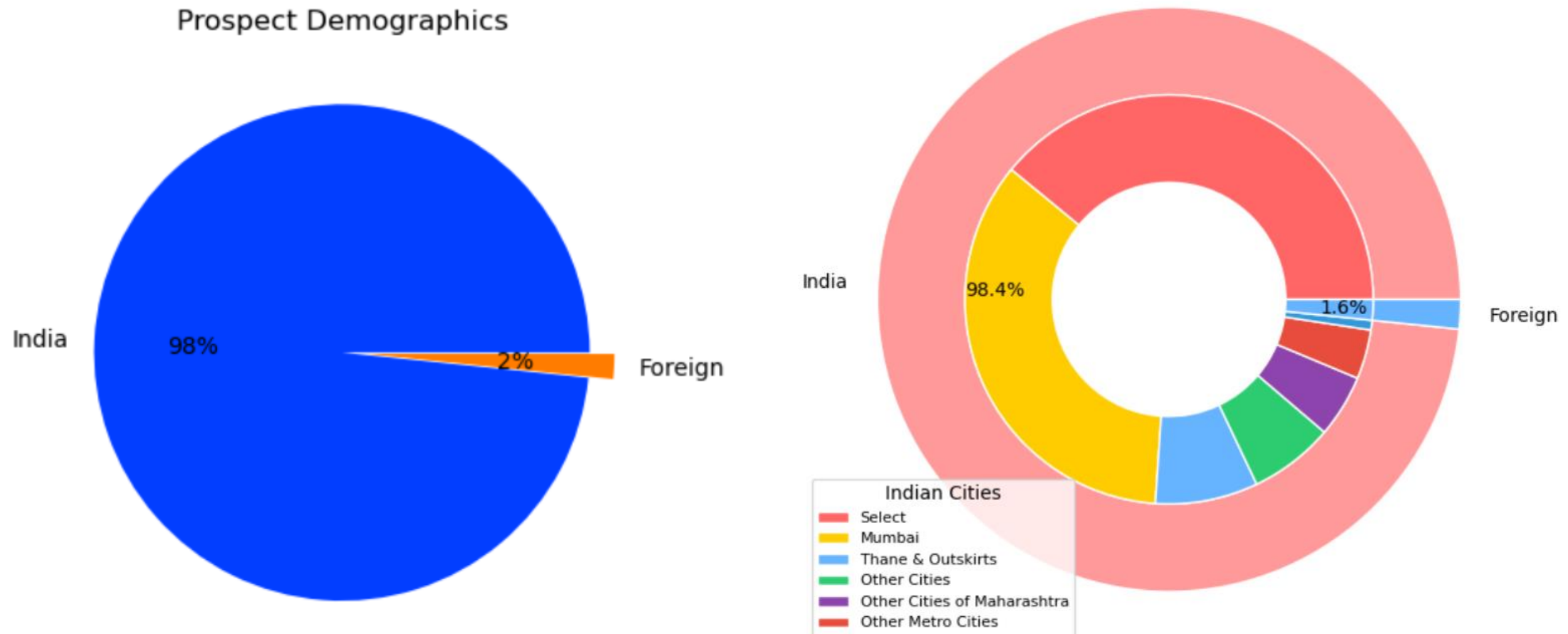
Let's take a closer look at DATA



Prospect Demographics

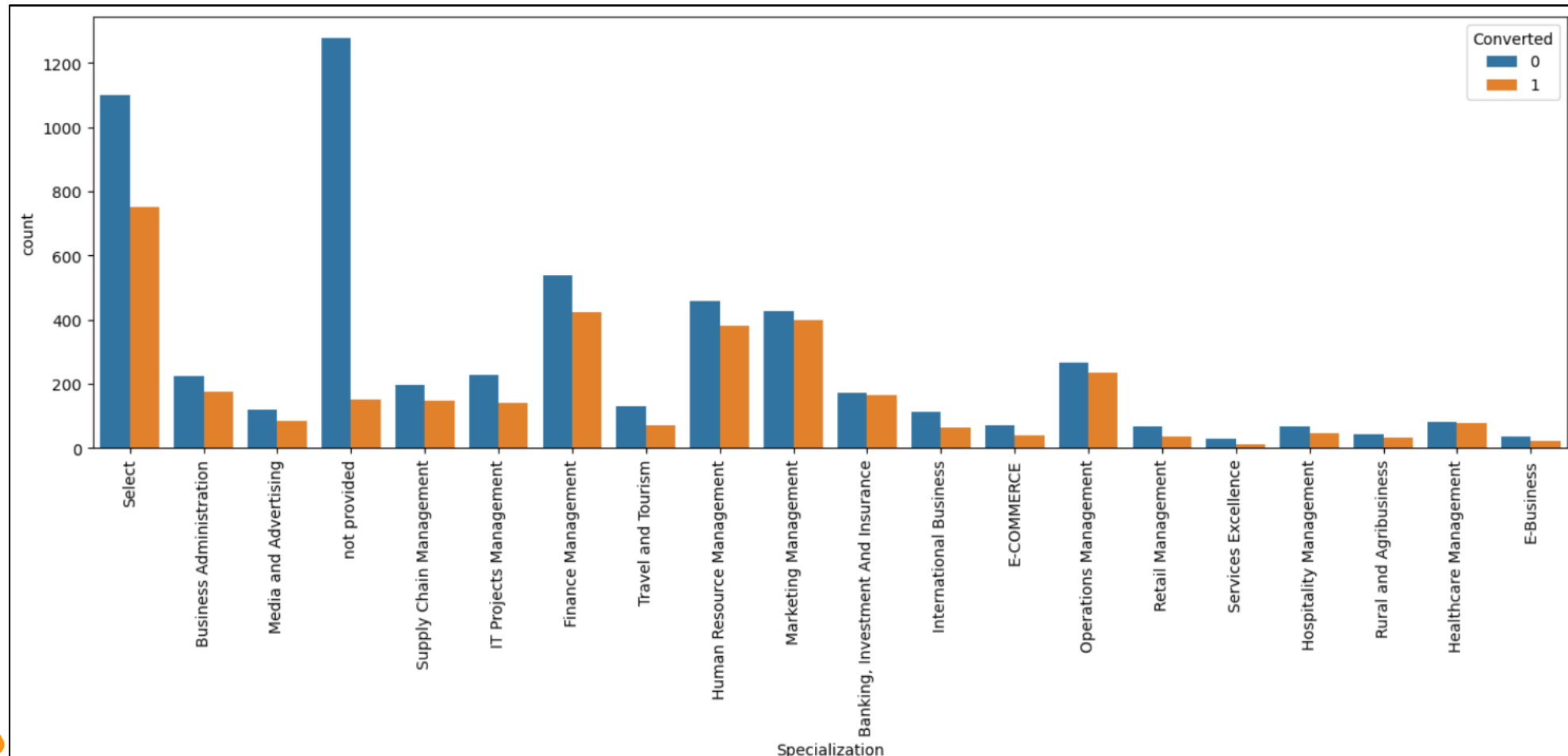
While 98% of the prospects are Indian, the distribution within Indian cities is also not uniform. State of Maharashtra in general has a high prospect count, majority of them coming from the IT hub of Mumbai. Tier II cities on the other hand show very small participation.

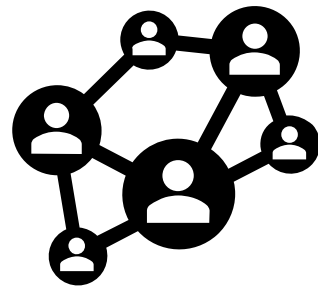
** Still a staggering number of visitors refrained from providing city details and hence 41% of the values are 'Select'



Specialization

Prospects have varied specializations ranging from business administrators to health care and agriculture. Here again we can see how the data is more biased towards cases with no specialization selected.





Lead Origin

- We are capturing leads from all possible sources.
- Google being one of our top sources.
- Others include channels like face book, other social media platforms, blogs etc.

Lead Source	
Google	2909
Direct Traffic	2543
Live Chat	1757
Organic Search	1154
Reference	534
Welingak Website	142
Referral Sites	125
Others	76

Final tips & takeaways from the DATA

A lot of missing data or with 'Selected' value

- Columns with more than 30% of missing data were dropped
- Missing values imputed using the Mode central measure.
- This includes replacing null values with 'Select' value.

Categorical columns with more than 10 categories

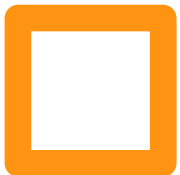
- In such cases, categories with small % were clubbed together in a broader category

Columns with zero to minimal variation in the values

- Columns like 'Search', 'Magazine' had only one value in them. Such columns do not provide any insight into the data and hence were dropped.

Columns with similar context

- Columns like 'Last Activity' & 'Last Notable Activity' offer very similar information in terms of context, and hence one of them were dropped.



Supervised ML model

Logistic Regression

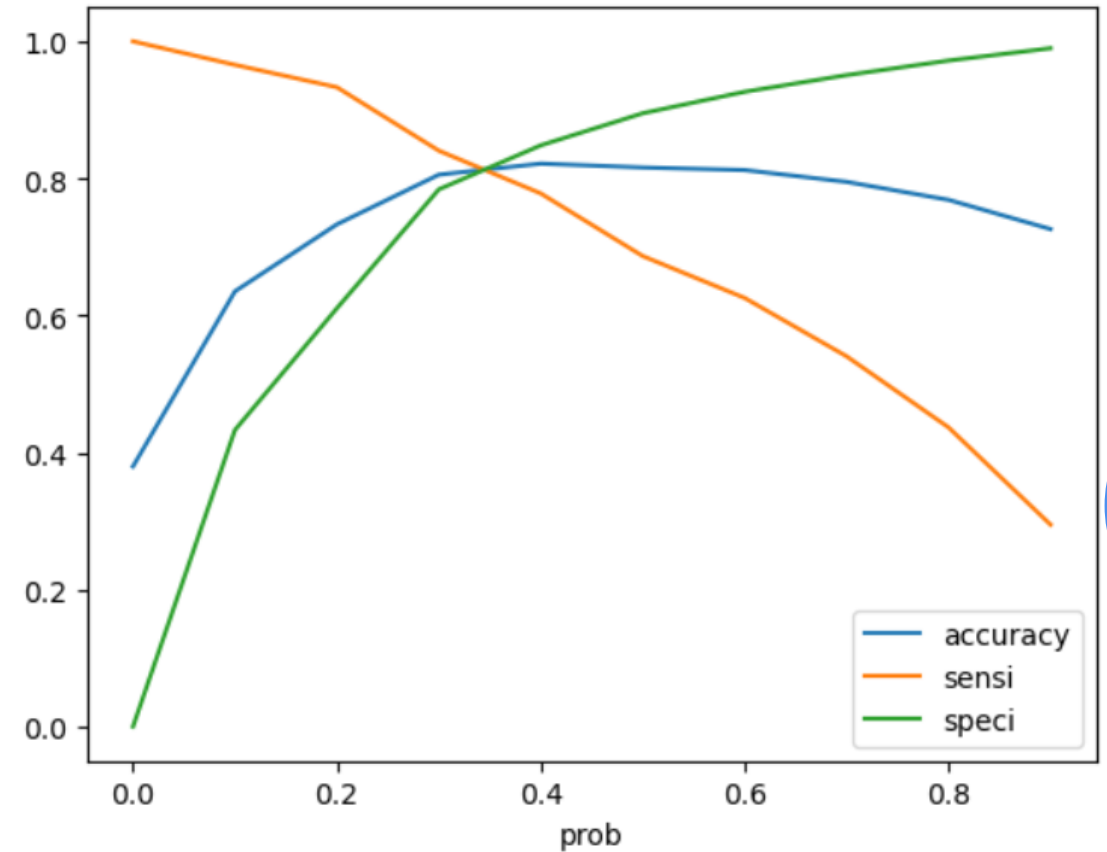
Since the target variable 'Converted' is a categorical column which classifies a lead as "converted" or "not converted", we will build a Logistic Regression classification model.

After creating dummy variables in the dataset, the number of features reached to 76. We used RFE technique to obtain top 15 related features.

Thereafter we used manual feature elimination to reach at a final model with recall = 0.84.

Final model has 12 independent features with p-value below 0.05 and variance inflation factor below 5.

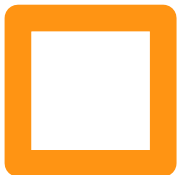
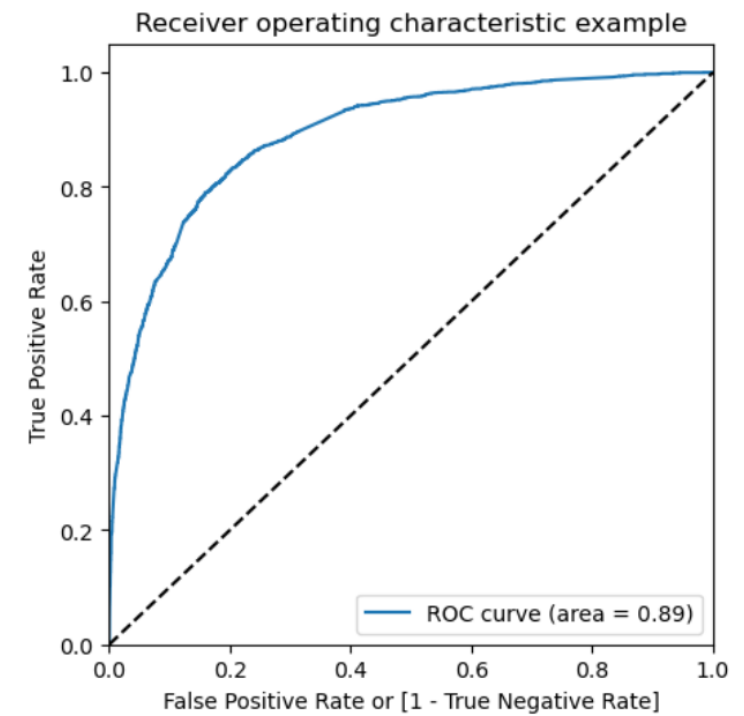
The optimal cut-off limit for the probabilities is defines at 30%.



Elements of Logistic Regression model

List of final 12 features

Lead Source_Live Chat
Lead Source_Welingak Website
Total Time Spent on Website
Last Activity_Olark Chat Conversation
Lead Profile_Potential Lead
Last Activity_SMS Sent
What is your current occupation_Working Professional
Last Activity_Converted to Lead
Do Not Email_Yes
Lead Profile_Student of SomeSchool
Last Activity_Had a Phone Conversation



Evaluation Metrics

Evaluation factor	Train Set	Test Set
Accuracy	80.5%	81.7%
Sensitivity	84.0%	86.7%
Specificity	78.4%	78.7%
Top three features	<ul style="list-style-type: none">• Lead Origin_Lead Add Form• What is your current occupation_Working Professional• Lead Source_Welingak Website	

Suggestions

- Company must try to perform strong strategies around the top 10 identified features.
- Make user experience more friendly and reliable for lead add form source.
- Target working professionals who are interested in upskilling.
- Spend more time with prospects from Welingak Website.