Lead Score

Case study - by Shruti Dandagi & Bharti Gokavi

Problem Statement

- X education sells online courses to Industry Professionals.
- X education gets a lots of leads, its lead conversion rate it very poor. For example, if they acquire 100 leads in a day , only about 30 of them are converted.
- To make this process more efficient ,the company wishes to identify the most potential leads, also known as “Hot Leads”.
- If they successfully identify this set of leads, the lead conversion rate should go up as the sales team will now be focusing more on communicating with the potential leads rather than making calls to everyone.

Business Objective

- Help X education to select the Most Promising Leads (Hot Leads)
- Build the Logistic regression model to assign a lead score value between 0 to 100 to each of the leads which can be used by the company to target Potential Leads.

The objective is thus classified into following subgoals

Create a Logistic Regression model to predict the Lead Conversion probabilities for each Lead	Decide on a Probability threshold value above which the Lead will be predicted as converted	Get the Lead score value for each Lead.

Problem solving methodology

- Reading and Understanding the Data
- Data Cleaning
- Data visualization
- Data Preparation
- Applying Recursive Feature Elimination to identify best set of features
- Building the model with features selected by RFE.
- Eliminate features with high p-value and VIF values and get the final model
- Model Evaluation with various metrics
- Decide on the Probability threshold value based on optimal cutoff point
- Use the model for prediction on the test dataset and perform model evaluation

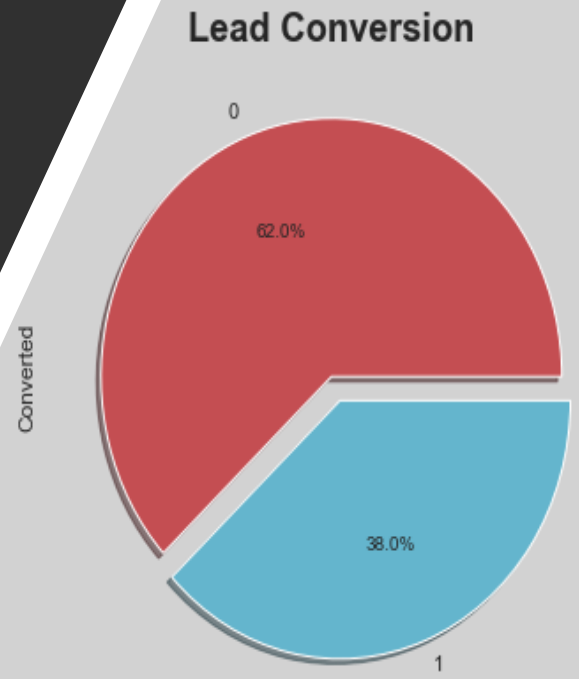
Data Cleaning

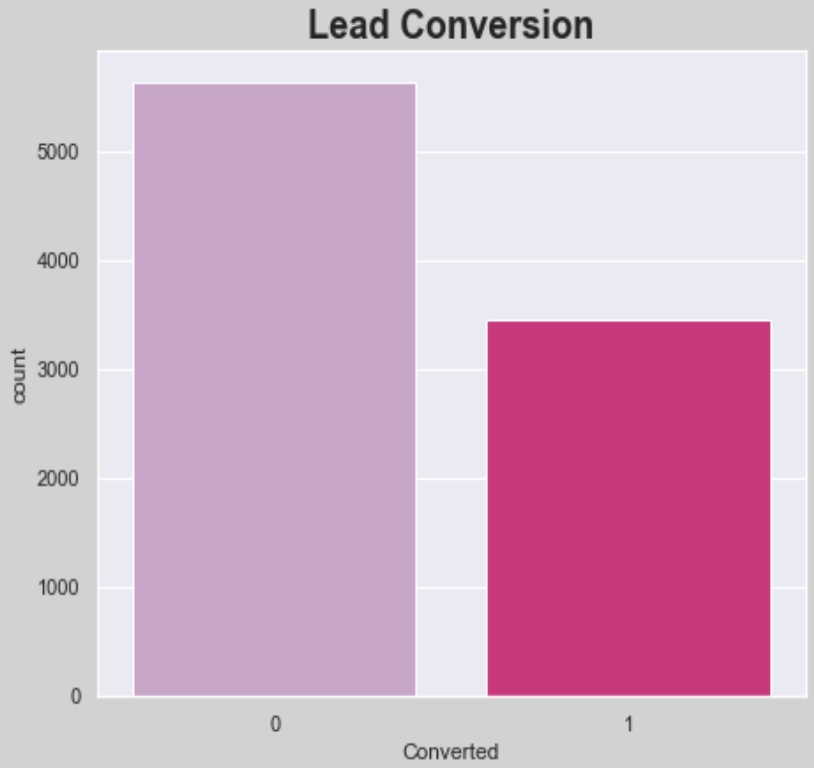
Data inspection	• The dataset contains 9240 rows and 37 columns
Handling “Select” Level	• Some Categorical columns have ‘Select’ level • Replace “Select” values with Null values
Check and Treat missing values	• There are missing values present in the data • Drop the columns having more than 45% missing values
Check unique categories	• Some columns have too much variation among the categories • Drop the columns that have highly skewed categories
Impute columns with less percent of missing	• For the columns with higher number of categories and with less percentage, impute with 'Others'
Check percentage of rows retained	• 98.5% of the rows retained after data cleaning

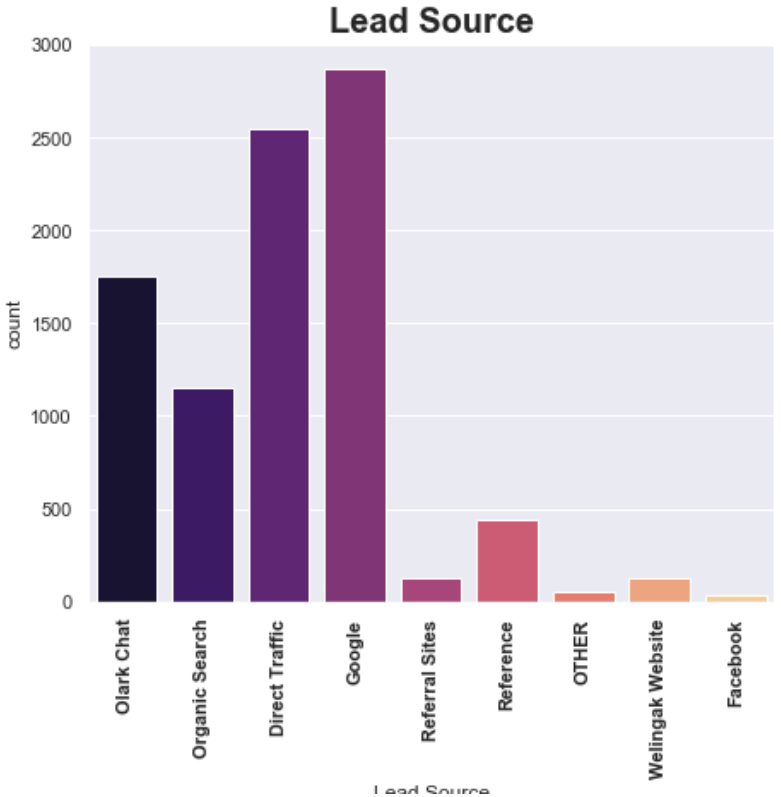
Data Visualization

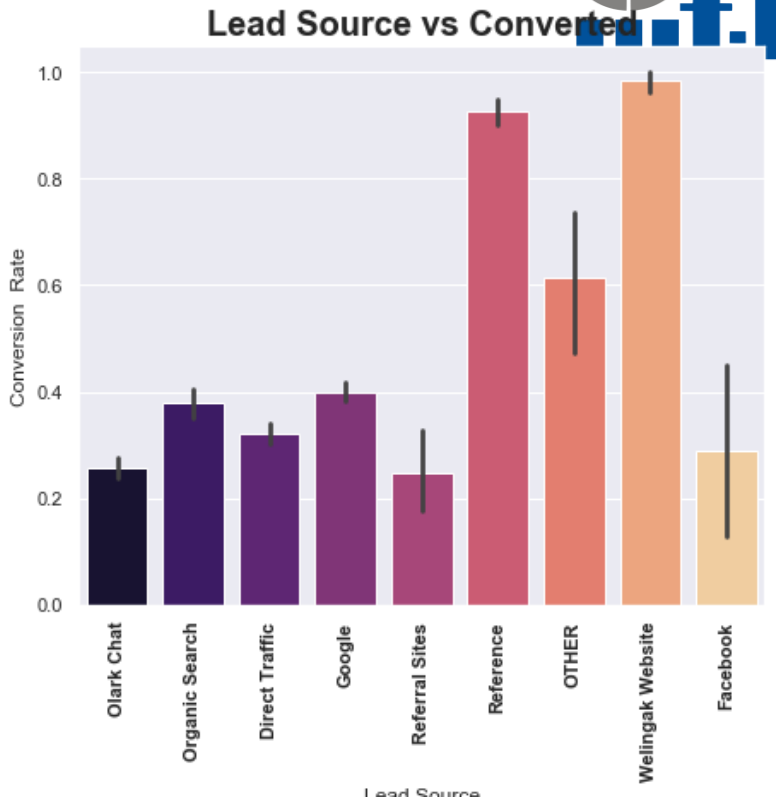
- **After Data cleaning we were left with 9103 Leads.**

- **We can see from the plot that not many Leads were 'Converted'. Out of 9103 Leads, only around 3500 (38%) of the Leads were converted**





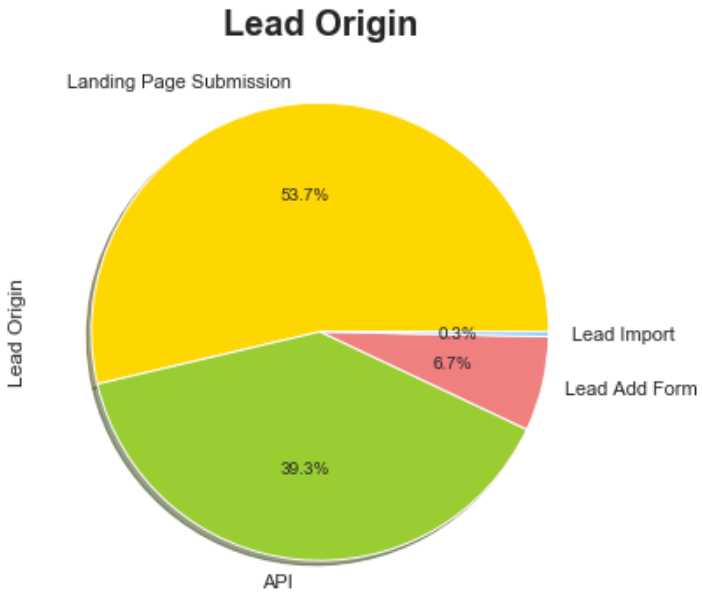


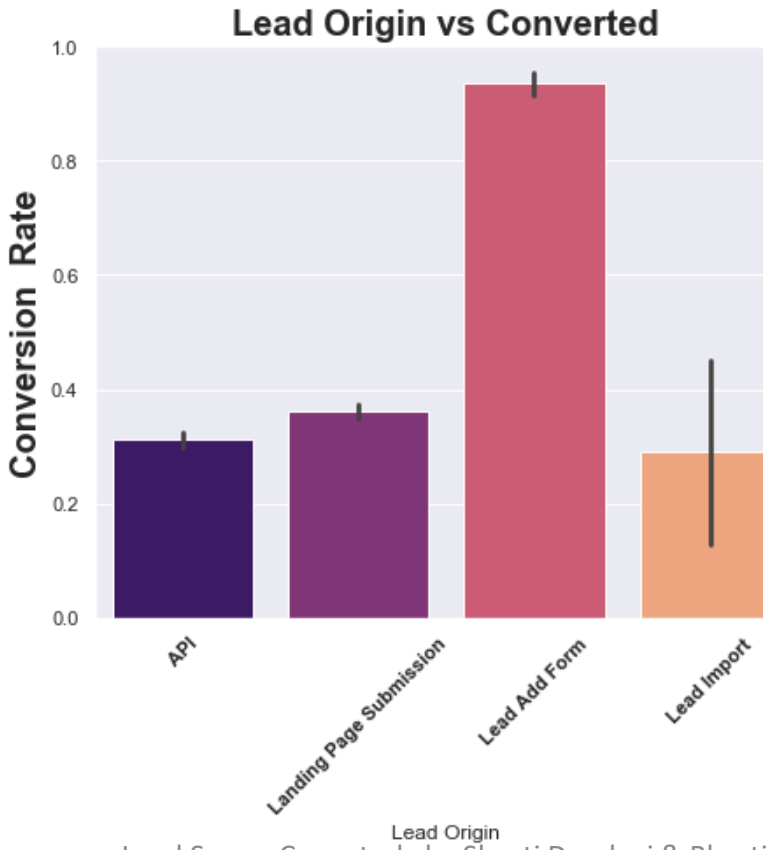


Lead Source

- **The source of the Leads were more from Google, Direct traffic and Olark Chat**
- **Most of the Leads were converted from Welingak Website, Reference and Google too**

Lead origin



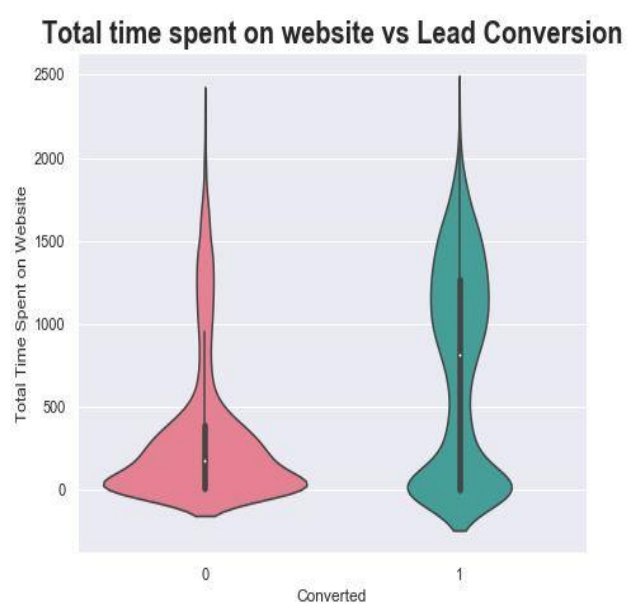


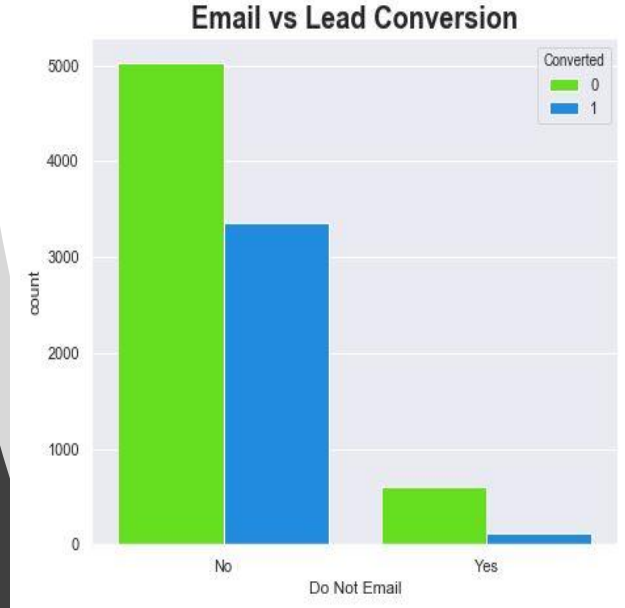
- 53.7% of the Leads have their origin from Landing page submission and 39% of Leads from API
- Lead Add Form has more than 90% conversion rate but count of lead are not very high.
- Lead Import are very less in count.

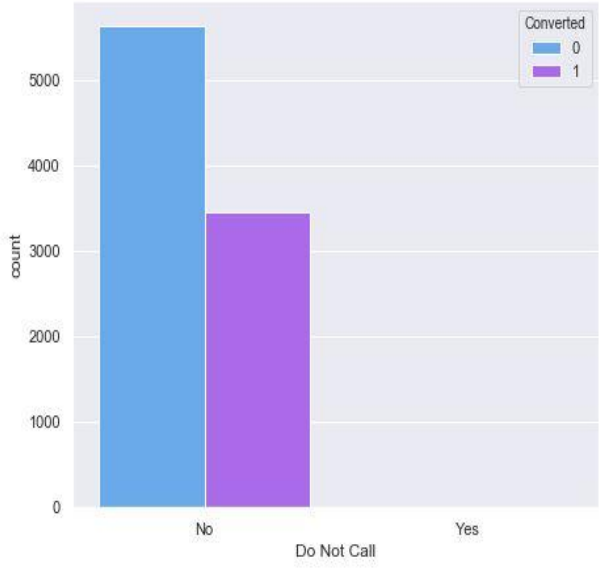
Total visits to website

- Median for converted and not converted leads are the same.
- Leads spending more time on the website are more likely to be converted.
- Website should be made more engaging to make leads spend more time.
- Nothing conclusive can be said on the basis of Total Visits.

- Whenever sales teams approach potential lead via email or call . The potential lead has high percentage of becoming hot lead

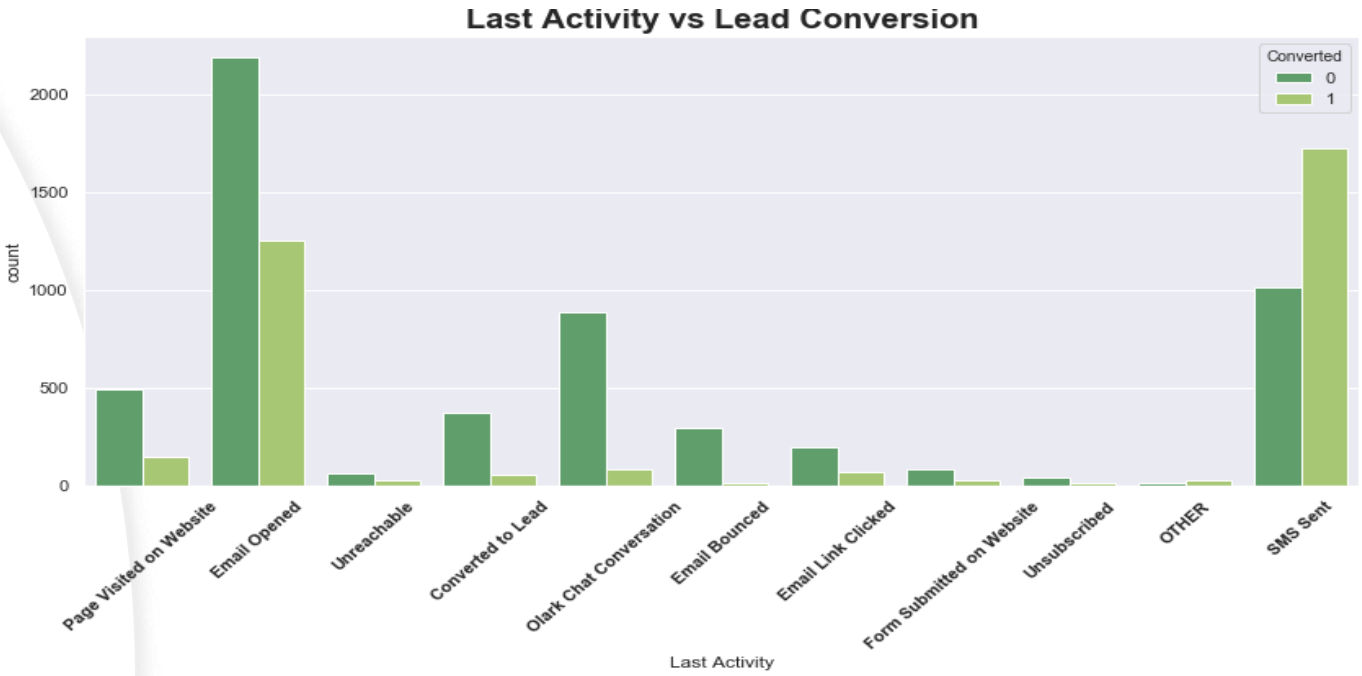


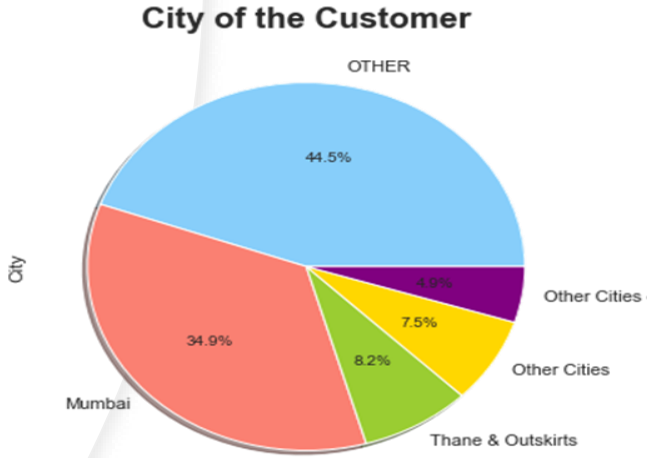


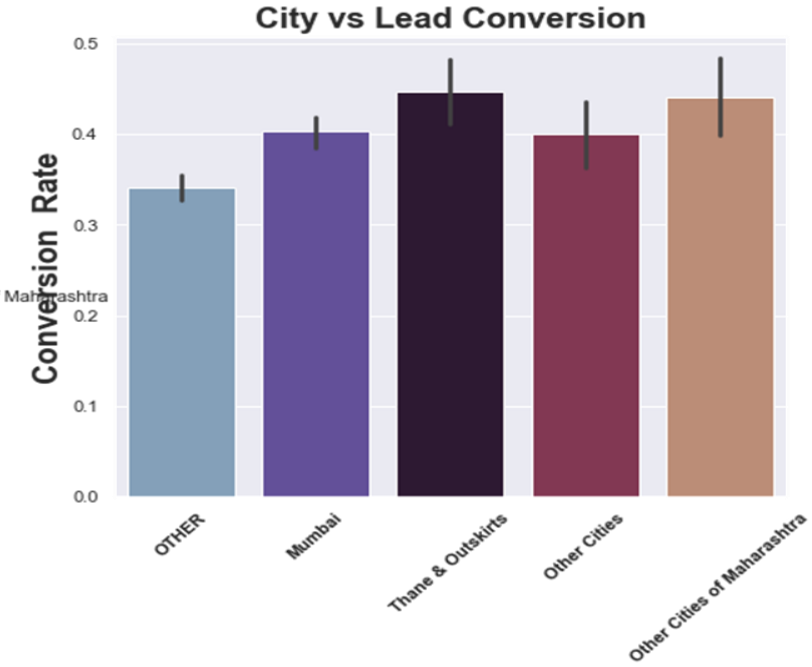


Last Activity and City

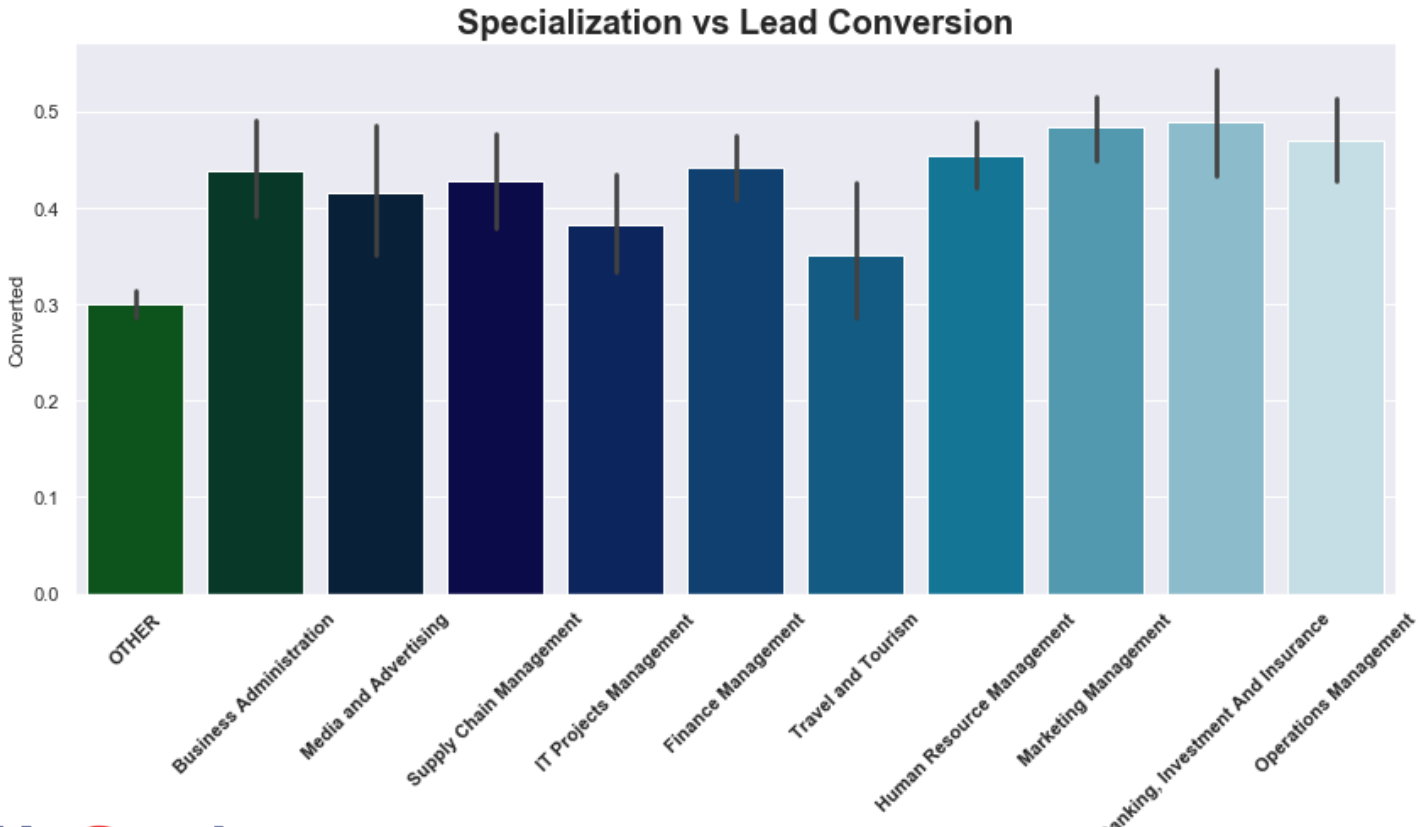
- **Most of the lead have their Email opened as their last activity.**
- **Conversion rate for leads with last activity as SMS Sent is almost 60%.**
- **35% of the total customers belong to Mumbai**
- **Most leads are from Mumbai, Thane & Outskirts, and other cities of Maharashtra with high conversion rate.**







Specialization

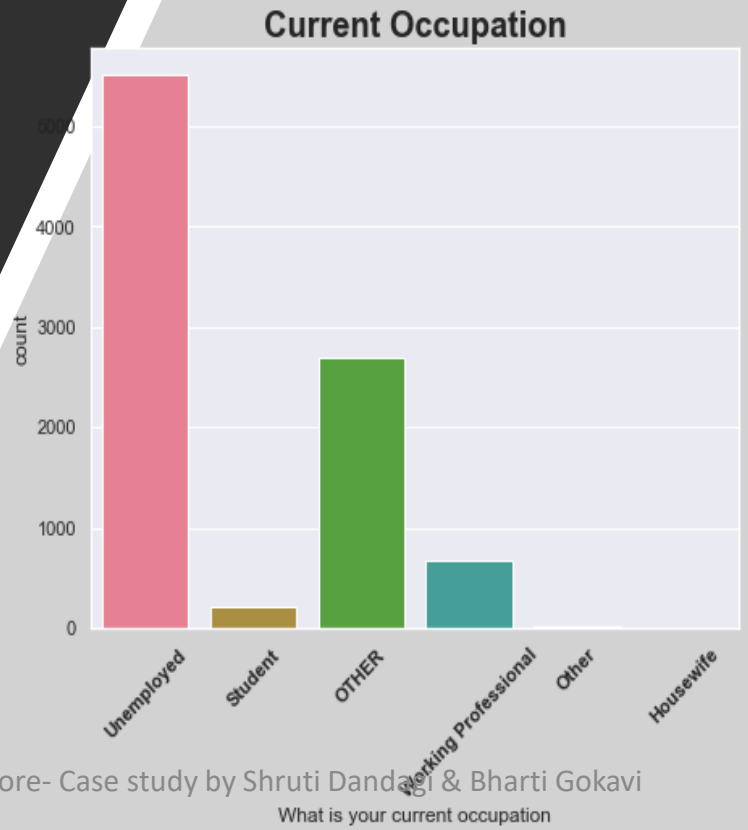


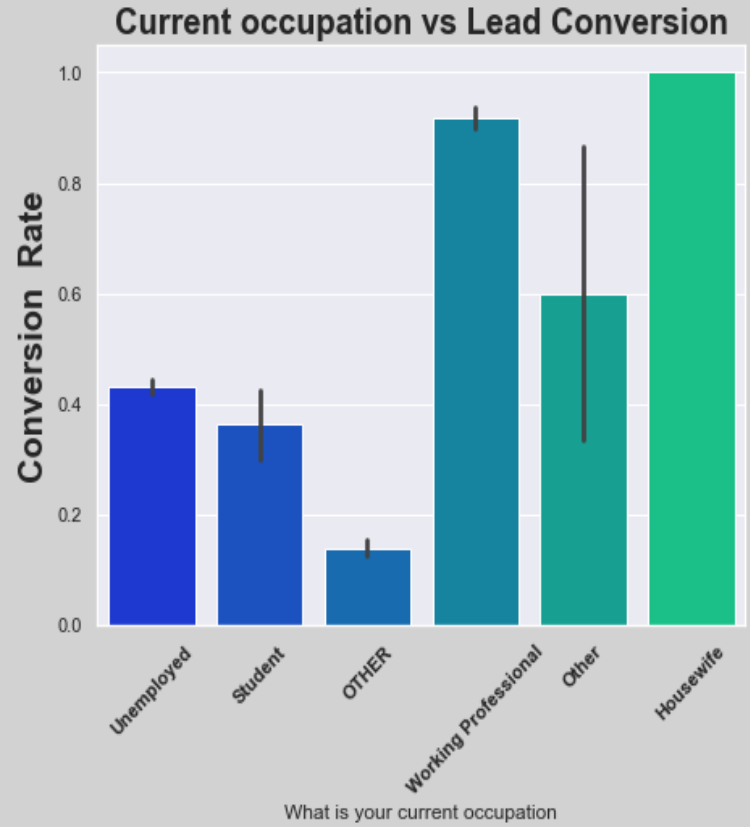
- **Customers who were in Travel and Tourism domain has less Lead Conversion rate compared to other domains**
- **Focus should be more on the Specialization with high conversion rate.**

Current occupation of the Leads

- There are many customers who are Unemployed

- Housewife and Working Professionals going for the course have high chances of joining it.





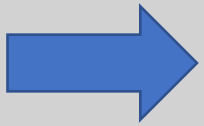
Data Preparation

Treating outliers	• The outliers present in the column ‘Total visits were treated by capping
Dummy encoding	• For some categorical variables with multiple levels dummy features (One hot encoded) were created.
Test- Train split	• The Original Dataframe was split into test and train Dataset. The Train Dataset was used to train the model and test dataset was used to evaluate the model
Feature Scaling	• Scaling helps in interpretation • Standardization was used to scale the data for modelling

Feature Selection using RFE

Recursive feature elimination is an optimization technique for finding the best performing subset of features. It is based on the idea of repeatedly constructing a model and choosing either the best (based on coefficients),setting the feature aside and then repeating the process with the rest of the features. This process is applied until all the features in the dataset are exhausted. Features are then ranked according to when they were eliminated.

Running RFE with the output number of variables 23



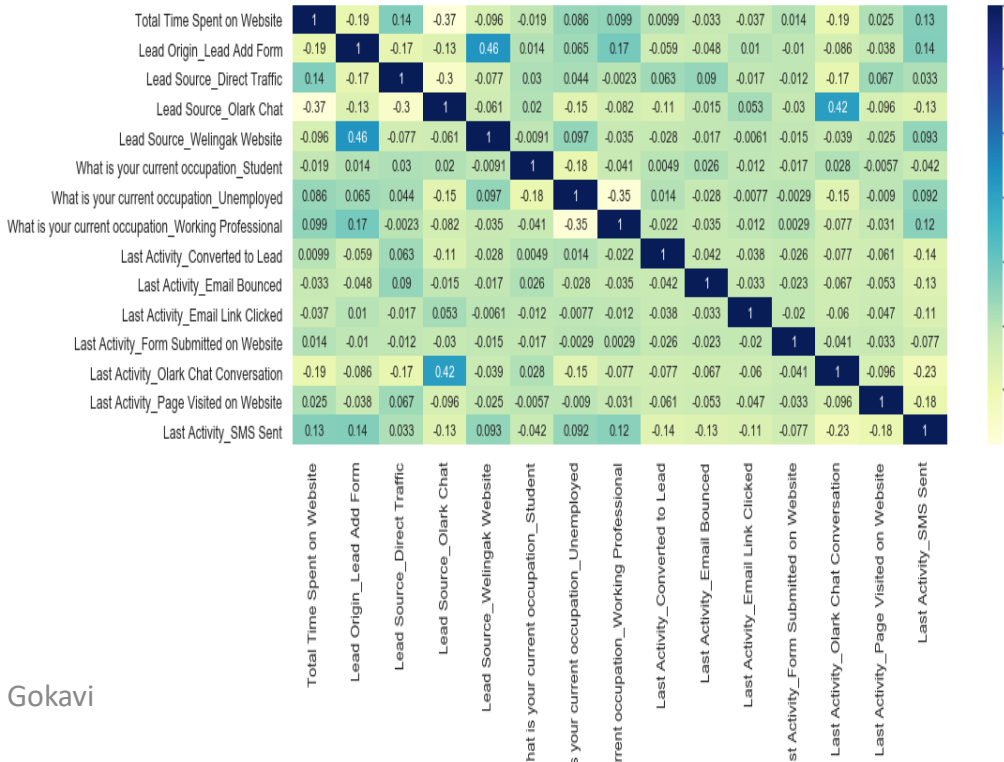
```
Index(['Total Time Spent on Website', 'Lead Origin_API',
       'Lead Origin_Landing Page Submission', 'Lead Origin_Lead Add Form',
       'Lead Source_Direct Traffic', 'Lead Source_Facebook',
       'Lead Source_Olark Chat', 'Lead Source_Organic Search',
       'Lead Source_Referral Sites', 'Lead Source_Welingak Website',
       'What is your current occupation_Housewife',
       'What is your current occupation_Student',
       'What is your current occupation_Unemployed',
       'What is your current occupation_Working Professional',
       'What matters most to you in choosing a course_Better Career Prospects',
       'What matters most to you in choosing a course_Flexibility & Convenience',
       'Last Activity_Converted to Lead', 'Last Activity_Email Bounced',
       'Last Activity_Email Link Clicked',
       'Last Activity_Form Submitted on Website',
       'Last Activity_Olark Chat Conversation',
       'Last Activity_Page Visited on Website', 'Last Activity_SMS Sent'],
      dtype='object')
```

Building the model

- Generalized Linear Models from StatsModels is used to build the Logistic Regression model.
- The model is built initially with the 23 variables selected by RFE.
- Unwanted features are dropped serially after checking p values (< 0.05) and VIF (< 5) and model is built multiple times.
- The final model with 15 features, passes both the significance test and the multi-collinearity test.

	Features	VIF
6	What is your current occupation_Unemployed	1.9700
3	Lead Source_Olark Chat	1.7000
1	Lead Origin_Lead Add Form	1.6300
14	Last Activity_SMS Sent	1.5900
2	Lead Source_Direct Traffic	1.4500
12	Last Activity_Olark Chat Conversation	1.3900
4	Lead Source_Welingak Website	1.3200
0	Total Time Spent on Website	1.2900
7	What is your current occupation_Working Professional	1.2400
13	Last Activity_Page Visited on Website	1.1000
8	Last Activity_Converted to Lead	1.0800
9	Last Activity_Email Bounced	1.0600
10	Last Activity_Email Link Clicked	1.0500
5	What is your current occupation_Student	1.0400

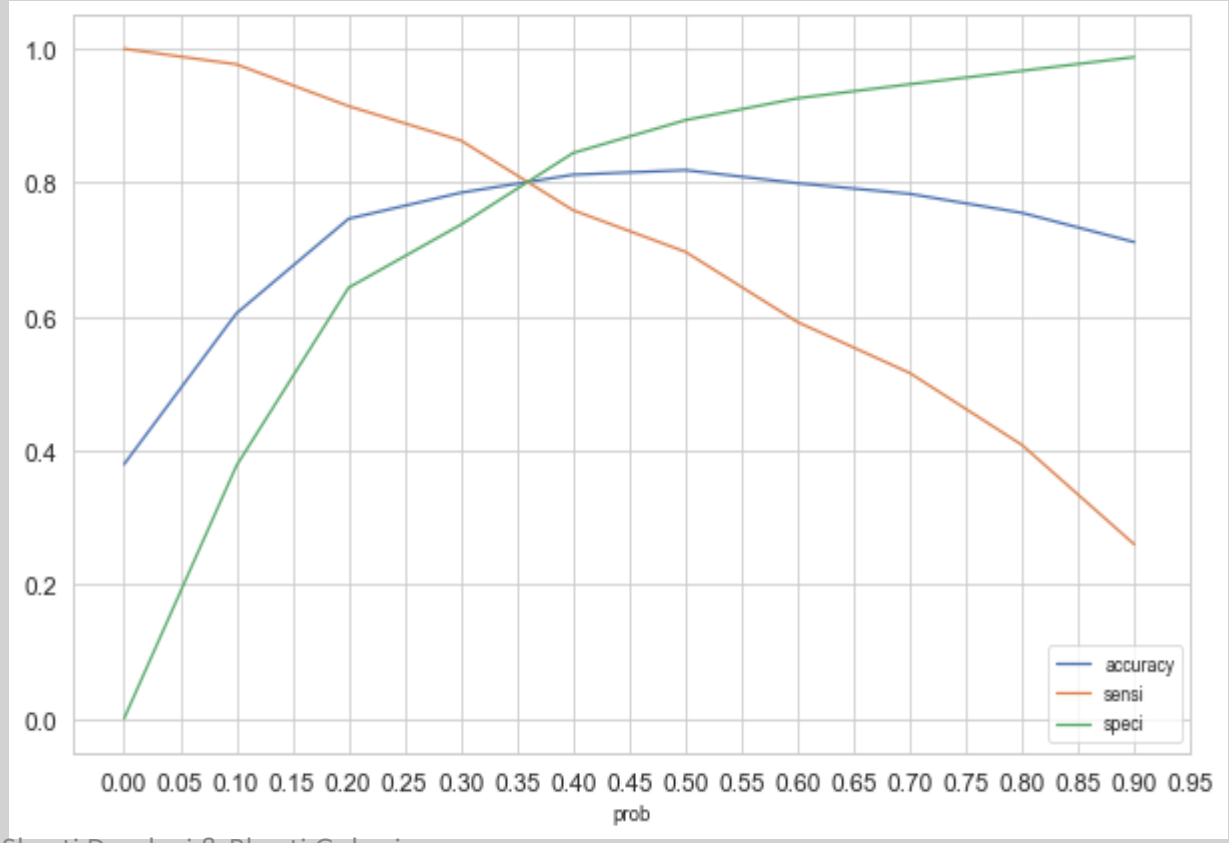
	coef	std err	z	P>|z|	[0
const	-2.0150	0.094	-21.472	0.000	-2
Total Time Spent on Website	1.1052	0.040	27.696	0.000	1
Lead Origin_Lead Add Form	3.5108	0.225	15.627	0.000	3
Lead Source_Direct Traffic	-0.3214	0.082	-3.936	0.000	-0
Lead Source_Olark Chat	1.1435	0.108	10.556	0.000	0
Lead Source_Welingak Website	1.7538	0.756	2.319	0.020	0
at is your current occupation_Student	1.2916	0.241	5.362	0.000	0
your current occupation_Unemployed	1.1711	0.086	13.629	0.000	1
ent occupation_Working Professional	3.7418	0.199	18.849	0.000	3
Last Activity_Converted to Lead	-1.4200	0.214	-6.643	0.000	-1
Last Activity_Email Bounced	-2.0748	0.318	-6.532	0.000	-2
Last Activity_Email Link Clicked	-0.5624	0.219	-2.564	0.010	-0
t Activity_Form Submitted on Website	-0.7167	0.322	-2.224	0.026	-1
ast Activity_Olark Chat Conversation	-1.5030	0.165	-9.119	0.000	-1
Last Activity_Page Visited on Website	-0.6006	0.148	-4.058	0.000	-0
Last Activity_SMS Sent	1.0014	0.078	12.859	0.000	0



Finding Optimal Probability Threshold

Optimal cut-off probability is that prob where we get balanced sensitivity and specificity.

- The accuracy sensitivity and specificity was calculated for various values of probability threshold and plotted in the graph to the right.
- From the curve above, 0.34 is found to be the optimum point for cutoff probability.
- At this threshold value, all the 3 metrics - accuracy sensitivity and specificity was found to be well above 80% which is a well acceptable value.



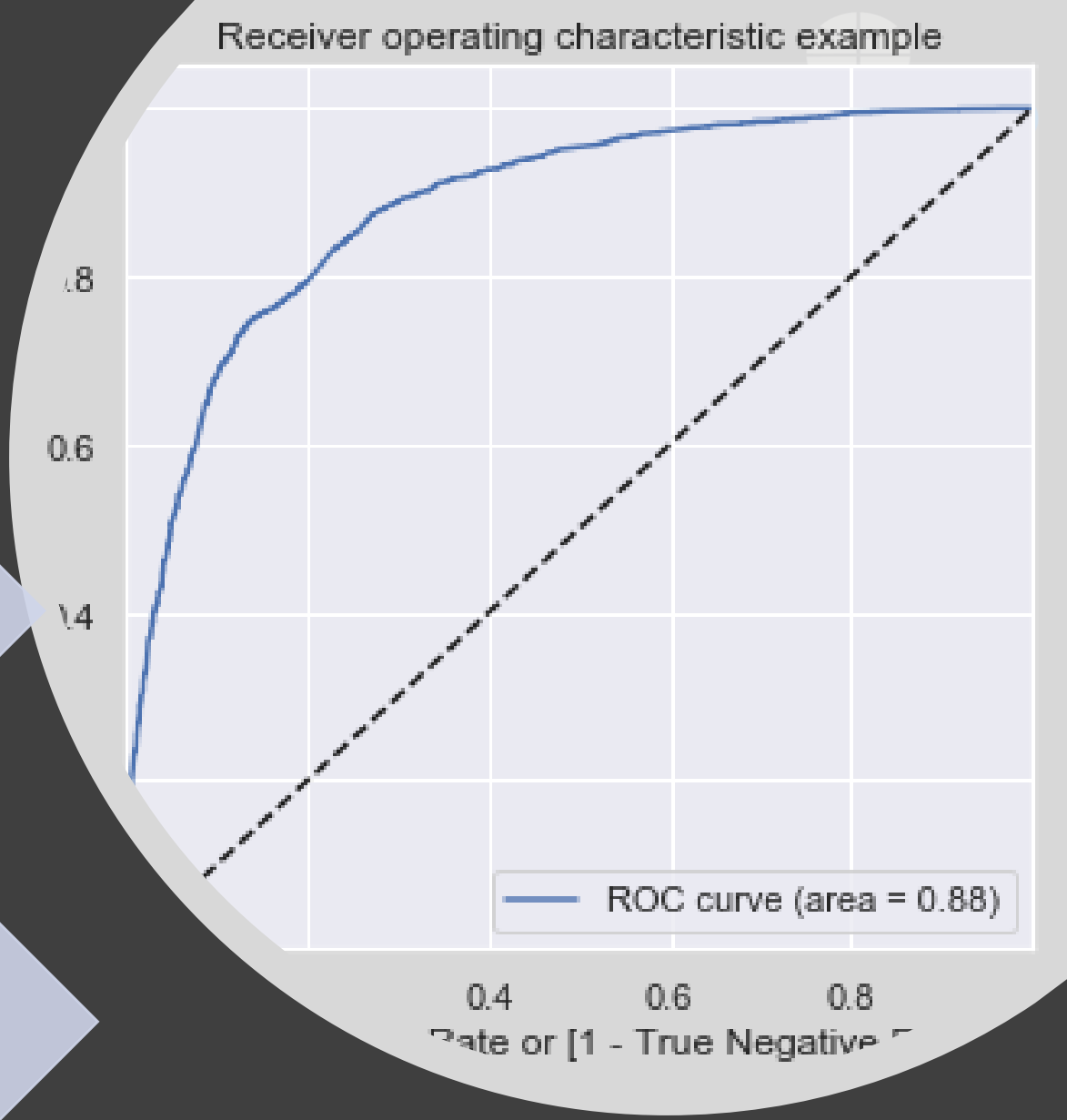
Plotting ROC curve and calculation AUC

Receiver Operating Characteristics(ROC) Curve

- It shows the tradeoff between sensitivity and specificity (any increase in sensitivity will be accompanied by a decrease inspecificity

Area under the Curve (GINI)

- By determining the Area under the curve (AUC) of the ROC curve, the goodness of the model is determined. Since the ROC curve is more towards the upper-left corner of the graph, it means that the model is very good. The larger the AUC, the better will is the model..
- The value of AUC for our model is 0.88.



Evaluating Model On Train Dataset

Confusion Matrix

Predicted	Not converted	Converted
Actual		
Not converted	3159	794
Converted	495	1924

Probability Threshold 0.34

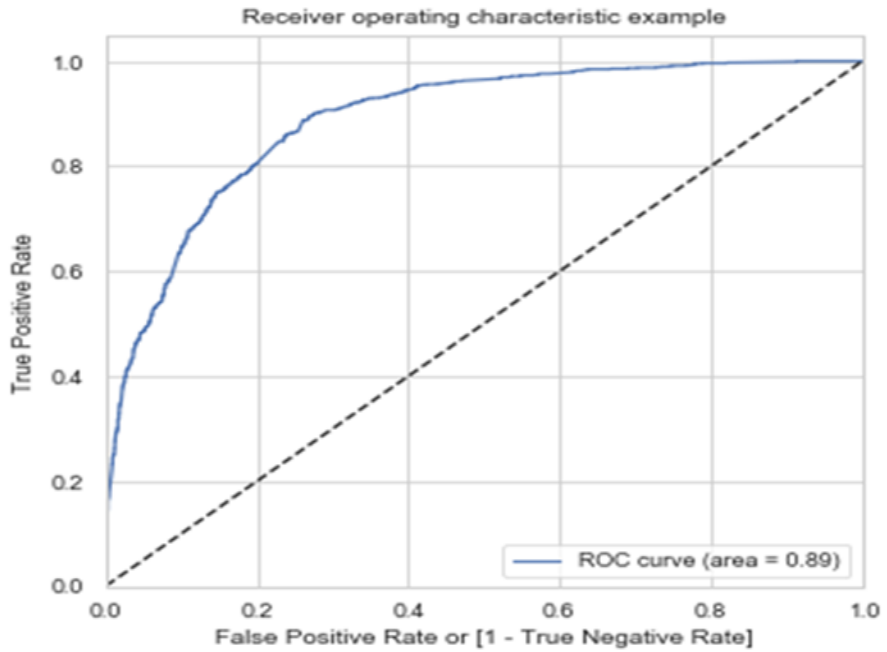
Accuracy	Sensitivity	Specificity	Area under curve	F1 Score
79.7	79.5	79.9	0.88	0.74

Positive Predicted Value	Negative Predicted Value	Precision	Recall
0.70	• 0.86	• 0.80	• 0.69

Making Predictions on Test set

- The final model on the train dataset is used to make predictions for the test dataset
- The train data set was scaled using the scaler.transform function that was used to scale the train dataset.
- The Predicted probabilities were added to the leads in the test dataframe.
- Using the probability threshold value of 0.34, the leads from the test dataset were predicted if they will convert or not.

ROC for test data



The top 5 records from the final test data set

	Converted	Prospect ID	Conversion_Prob	final_predicted
0	0	3504	0.2771	0
1	1	4050	0.9367	1
2	0	7201	0.5756	1
3	0	1196	0.2767	0
4	1	8219	0.0963	0

The top 5 records from the final test data set

Evaluating The Model On Test Dataset

The following evaluation metrices were recorded for the test dataset

Accuracy	Sensitivity	Specificity	Positive Predicted value	Negative Predicted value	Area under the curve
0.80	0.80	0.79	0.71	0.86	0.89

Lead Score Calculation

- Lead Score is calculated for all the leads in the original dataframe.
- Formula for Lead Score calculation is:

Lead Score = 100 * Conversion Probability

•The train and test dataset is concatenated to get the entire list of leads available.
•The Conversion Probability is multiplied by 100 to obtain the Lead Score for each lead.
•Higher the lead score, higher is the probability of a lead getting converted and vice versa,
•Since, we had used 0.34 as our final Probability threshold for deciding if a lead will convert or not, any lead with a lead score of 34 or above will have a value of '1' in the Final predicted column.

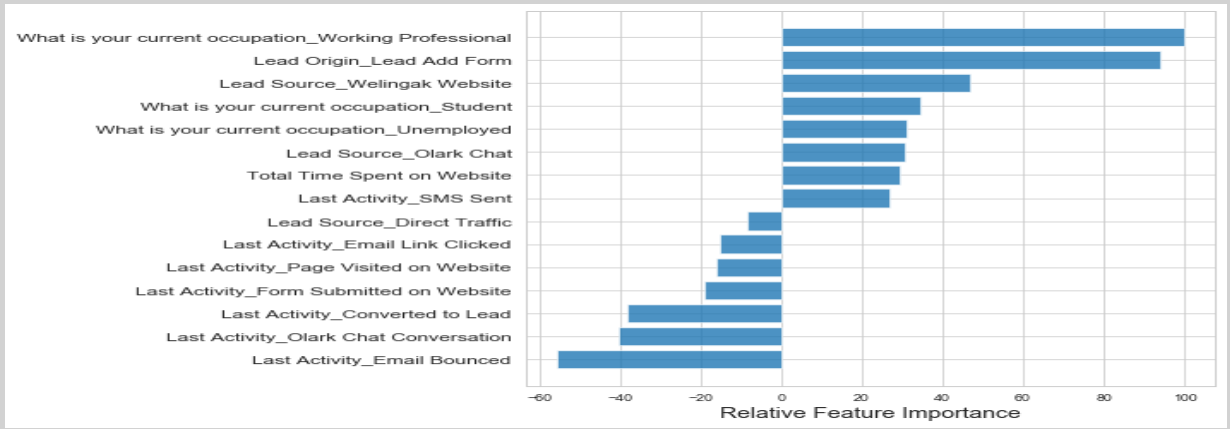
The figure showing Lead Score for top 10 records from the data set.

	Lead Number	Converted	Conversion_Prob	final_predicted	Lead_Score
0	660737	0.0000	0.2184	0.0000	22.0000
1	660728	0.0000	0.3888	1.0000	39.0000
2	660727	1.0000	0.7475	1.0000	75.0000
3	660719	0.0000	0.1793	0.0000	18.0000
4	660681	1.0000	0.4147	1.0000	41.0000
5	660680	0.0000	0.0340	0.0000	3.0000
6	660673	1.0000	0.8183	1.0000	82.0000
7	660664	0.0000	0.0340	0.0000	3.0000
8	660624	0.0000	0.0405	0.0000	4.0000
9	660616	0.0000	0.0536	0.0000	5.0000

Feature Importance

•15 features have been used by our model to successfully predict if a lead will get converted or not.
•The Coefficient (beta) values for each of these features from the model parameters are used to determine the order of importance of these features.
•Features with high positive beta values are the ones that contribute most towards the probability of a lead getting converted.
•Similarly, features with high negative beta values contribute the least.

Total Time Spent on Website	1.1052
Lead Origin_Lead Add Form	3.5108
Lead Source_Direct Traffic	-0.3214
Lead Source_Olark Chat	1.1435
Lead Source_Welingak Website	1.7538
What is your current occupation_Student	1.2916
What is your current occupation_Unemployed	1.1711
What is your current occupation_Working Professional	3.7418
Last Activity_Converted to Lead	-1.4200
Last Activity_Email Bounced	-2.0748
Last Activity_Email Link Clicked	-0.5624
Last Activity_Form Submitted on Website	-0.7167
Last Activity_Olark Chat Conversation	-1.5030
Last Activity_Page Visited on Website	-0.6006
Last Activity_SMS Sent	1.0014



The Relative Importance of each feature is determined on a scale of 100 with the feature with highest importance having a score of 100.

feature_importance = 100.0 * (feature_importance / feature_importance.max())

The features are then sorted using Quick Sort algorithm. Finally the sorted features are plotted in a bar graph in descending order of their relative importance.

Top 3 variables

After trying several models, we finally chose a model with the following characteristics:

- **All variables have p-value < 0.05.**
- **All the features have very low VIF values, meaning, there is hardly any mulitiicollinearity among the features. This is also evident from the heat map.**
- **The overall accuracy of 0.80 at a probability threshold of 0.34 on the test dataset is also very acceptable.**

Top three variables in our model that contribute most towards the probability of a lead getting converted

index	0
What is your current occupation_Working Professional	100.00
Lead Origin_Lead Add Form	93.82
Lead Source_Welingak Website	46.87

Inferences

- To improve overall lead conversion rate, focus should be on improving lead conversion of olark chat, organic search, direct traffic, and google leads and generate more leads from reference and welingak website.
- Lead conversion rate, can be improved by focusing more on improving lead conversion of API and Landing Page Submission origin and generate more leads from Lead Add Form.
- Though Google is the highest source to get leads, the lead conversion through google is low comparatively. Highlight the X education website everywhere.
- Can focus on unemployed customers too, can give them free course and can make the payment later.
- Website should be made more engaging to make leads spend more time
- Can focus more on housewives and Students by giving some offers like EMI.

Conclusions

Through our model we got the top variables which needs more focus,

1. Occupation – Working Professionals
2. Where the origin of the Leads is from Lead Add form
3. Lead Source- Wellingak Website
4. Students
5. Unemployed
6. Olark Chat as Lead Source
7. Time spent by the Lead on website
8. When the last activity was: a. SMS b. Email_bounced

Keeping these in mind the X Education can flourish as they have a very high chance to get almost all the potential buyers to change their mind and buy their courses.