LEAD SCORING CASE STUDY

Upgrad Assignment by

Raghu vamsi, Nivya & jawaharlal

Problem Statement

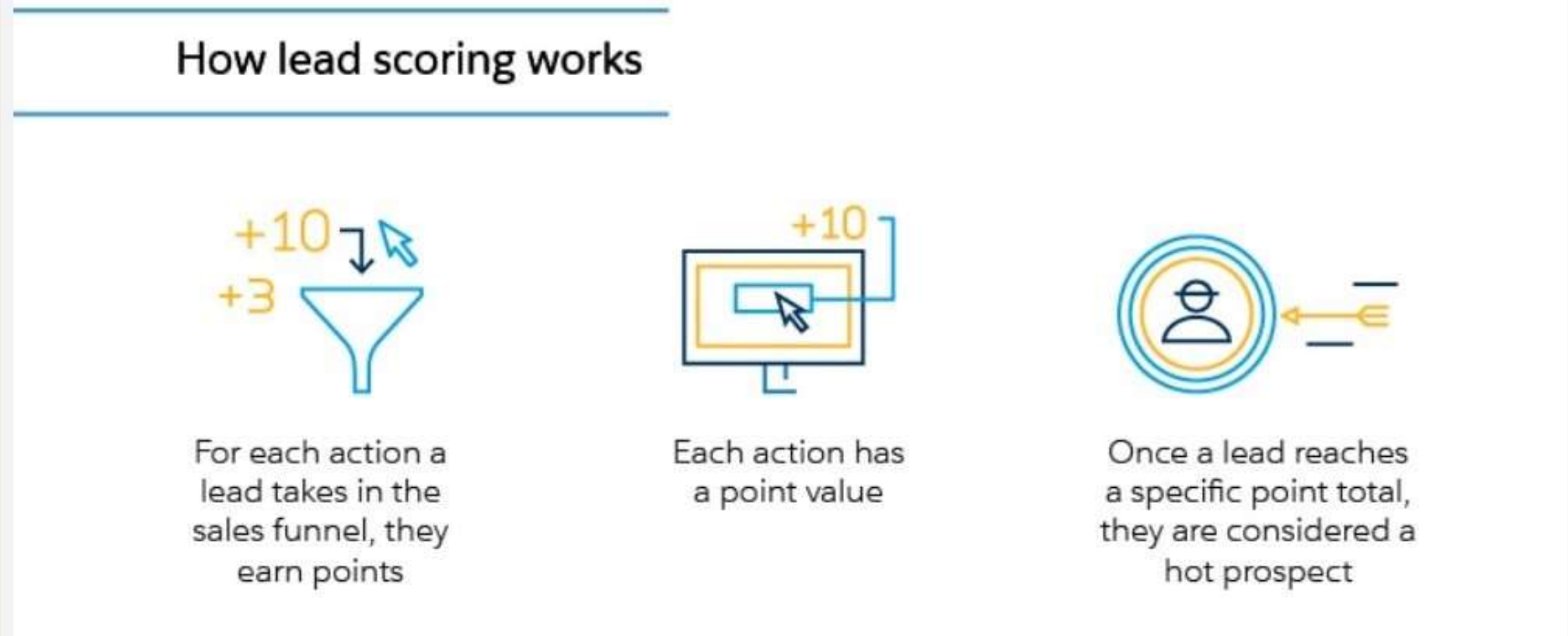
X Education an education company who sells **online courses** to industry professionals, need to select the **promising leads** to increase their target **lead conversion rate to 80%** in effective manner

Dataset Details

Dataset Name	:	Leads.csv
Number of Records	:	9240
Number of attributes	:	37
Data type Spread	:	30 Object type, 4 Float64 type, 2 int 64 type
Current Conversion Rate	:	38.54%

What is Lead Scoring?

- Lead scoring is the process of assigning values, often in the form of numerical "points," to each lead you generate for the business. When a lead reaches a specific point total, they are considered a hot prospect.
- Knowing when to reach out to a lead can help salespeople stay on track, focus their attention on the right group of potential customers, and be more productive.



Advantages and Limitations of Lead Scoring Model

Advantages

- **Focus on Opportunities that Matter**: Lead scoring ranks the quality of all your leads, making it easy to quickly identify which are most worth pursuing. Lead scoring means fewer lost opportunities, increased conversions, and higher ROI.
- **Increase Productivity**: Scoring leads, specifically with predictive methods, saves time by automatically sorting through the leads that are most likely to convert to clients.
- **Align marketing and sales**: In terms of business automation, lead scoring elevates the process for both sales and marketing teams. Efficient scoring methods utilize the expertise of marketing and insights from salespeople. Each department offers their perspectives to collaboratively create a mutually beneficial method to be integrated into workflows
- **Better understand your leads**: Businesses have a better understanding of their clients if they collect, analyze, and make sense of the demographic and behavioral data available to them. Lead scoring does just that, which means you can more comprehensively interpret current and future clients
- **Improve the sales experience for clients**: Lead scoring aligns sales and marketing, both teams get accurate information to work from. Sales and marketing teams can more personally cater to potential clients' needs with lead scoring.

Limitations

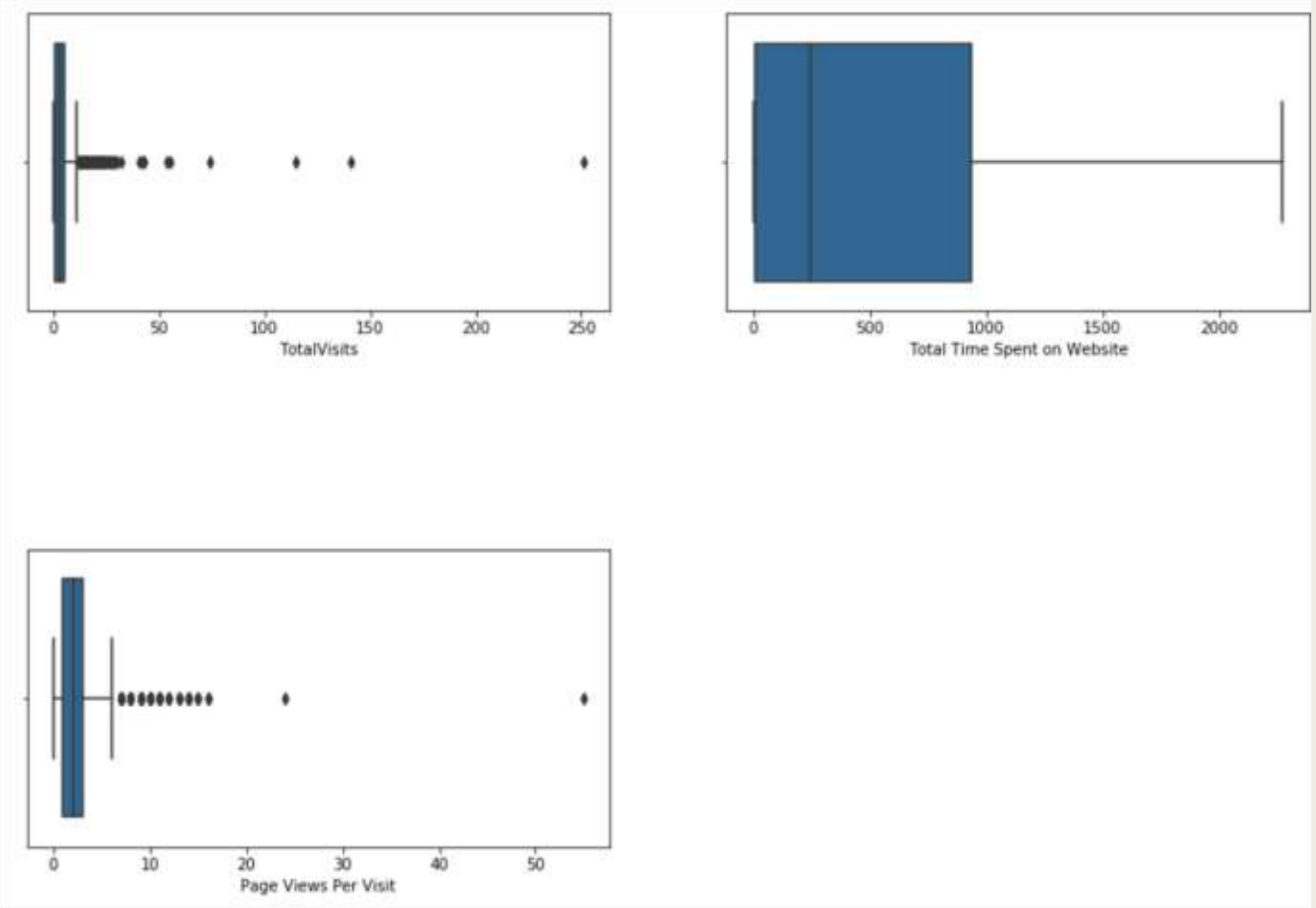
- Building a lead scoring model requires a **lot of data**. The reason why A.I. and machine learning algorithms haven't been a factor in lead scoring before recent times, is that there simple wasn't enough data and high-quality data too. If you try it without the requisite data, predictive lead scoring can go very awry
- It requires a **little expertise**; predictive lead scoring also requires **some expense**. Though there are platforms and experts that will help you, they can be costly

APPROACH FOR ANALYSIS

S.NO.	STEP	REASON
1	Data Exploration	The explore and understand the Dataset provided
2	Data Cleaning	Clean up the Data by analyzing the Null Value Column, Categorical Column Spread, Single Value Columns
3	Exploratory Data Analysis	To Identify relation between the attributes in Data with conversion rate
4	Data Preparation	Preparing the Dataset for building the model by creating dummy variables, splitting data into Train and Test set, Feature Scaling
5	Build Logistic Regression Model	To identify which attributes, contribute most of the lead conversion rate
6	Finding Optimal cutoff	Find the best probability for a better model
7	Drawing Inferences	Based on the model listing the inferences drawn
8	Conclusion	Concluding with the effective ways to increase the lead conversion rate

DATA CLEANING

Step	Columns	Action
"Select" Value	1] Specialization 2] How did you hear about X Education 3] Lead Profile 4] City	Replace with Null Value
Null Value > 40%	1 Lead Quality 2 Lead Profile 3 Asymmetrique Activity Index 4Asymmetrique Profile Index 5] Asymmetrique Activity Score 6] Asymmetrique Profile Score	Drop columns
Skewed Columns	1] Country 2] What matters most to you in choosing a course 3]Do Not Call 4] Search 5] Newspaper Article 6] X Education Forums 7] Newspaper 8] Digital Advertisement 9] Through Recommendations	Drop Columns
Columns with Single Value	1 Magazine 2Receive More Updates About Our Courses 3] Update me on Supply Chain Content 4 Get updates on DM Content 5 I agree to pay the amount through cheque	Drop Columns



Exploratory Data Analysis

Numerical Columns

1 TotalVisits
2Total Time spent on Website
3] Page Views Per Visit

The columns TotalVisits and Page Views Per Visit contain outliers can be handled by Capping them with 99 Percentile

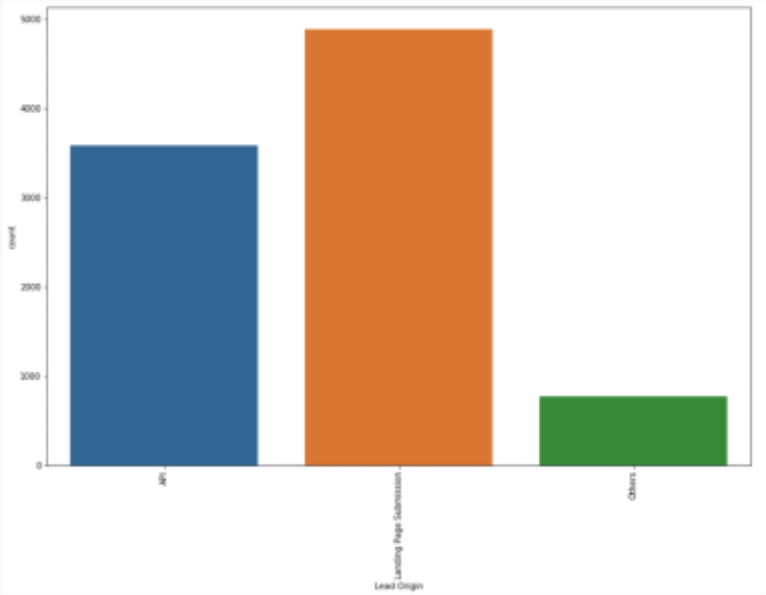
Exploratory Data Analysis

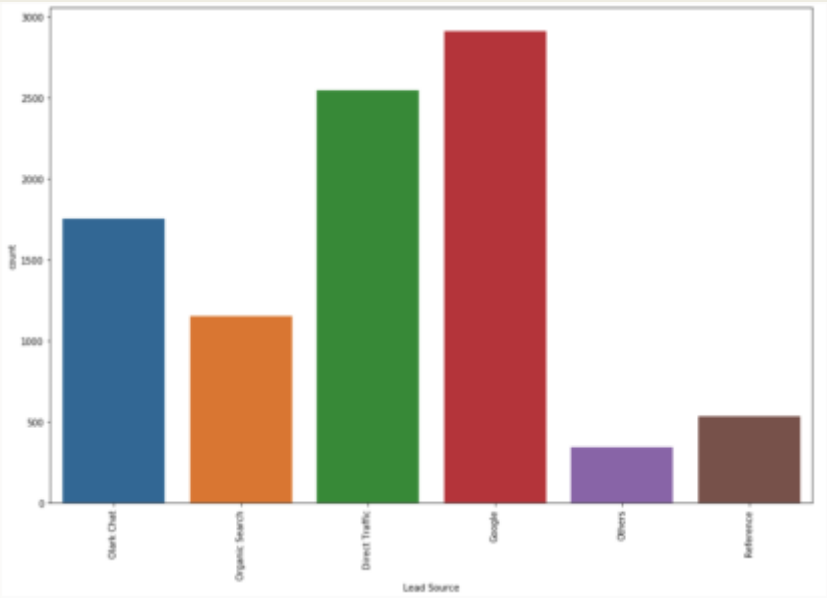
Categorical Columns

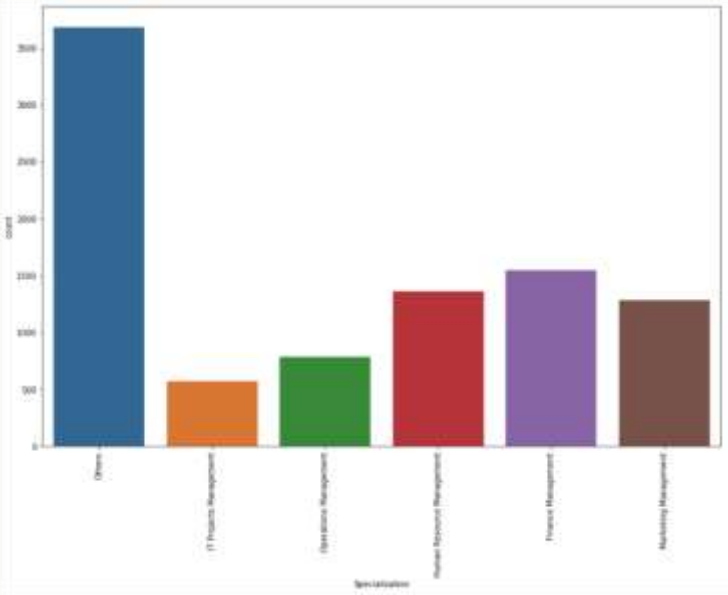
Columns	Action
1] Lead Source 2] Specialization 3] What is your current occupation 4] City 5] A free copy of Mastering The Interview 6] Last Notable Activity	Grouping the Categories with less percent into a common category Others and fill up Null values with the probability of the categories

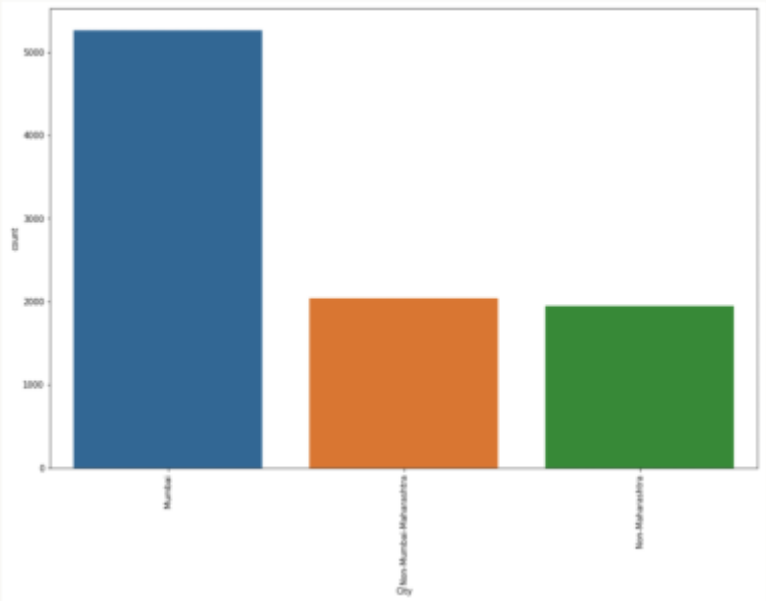
Exploratory Data Analysis

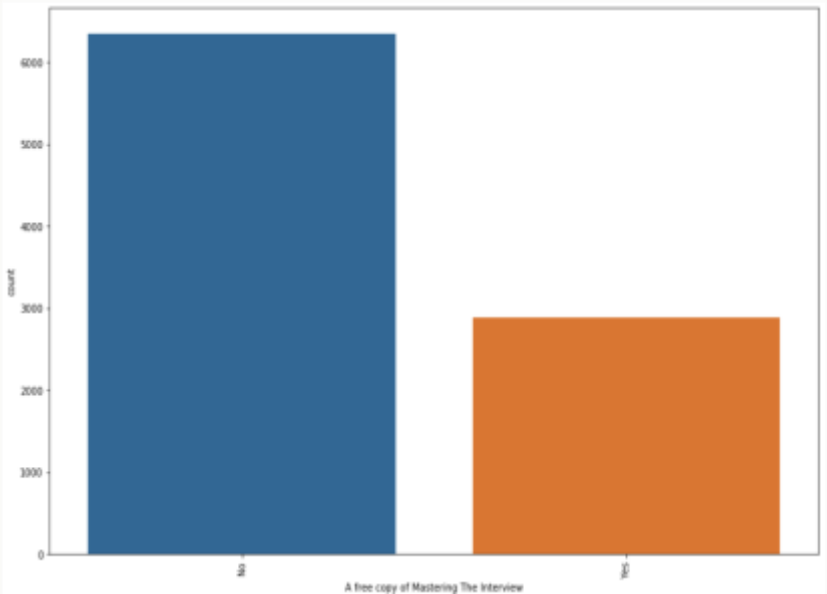
Categorical Columns

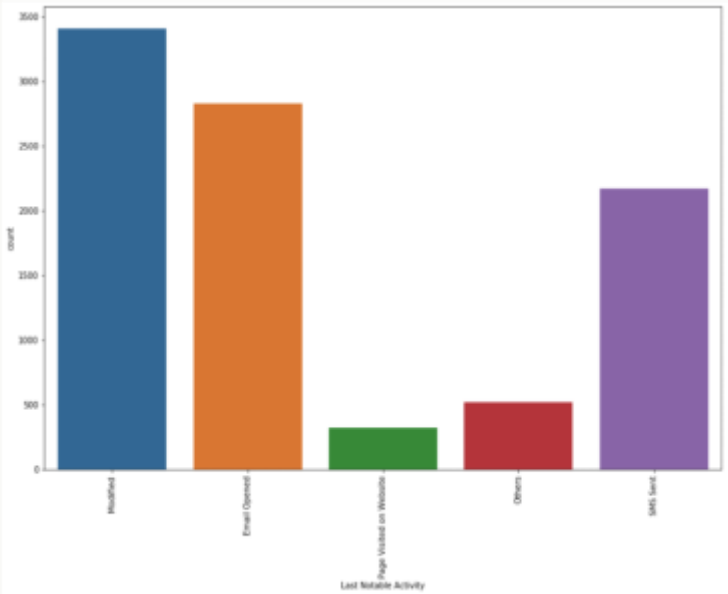






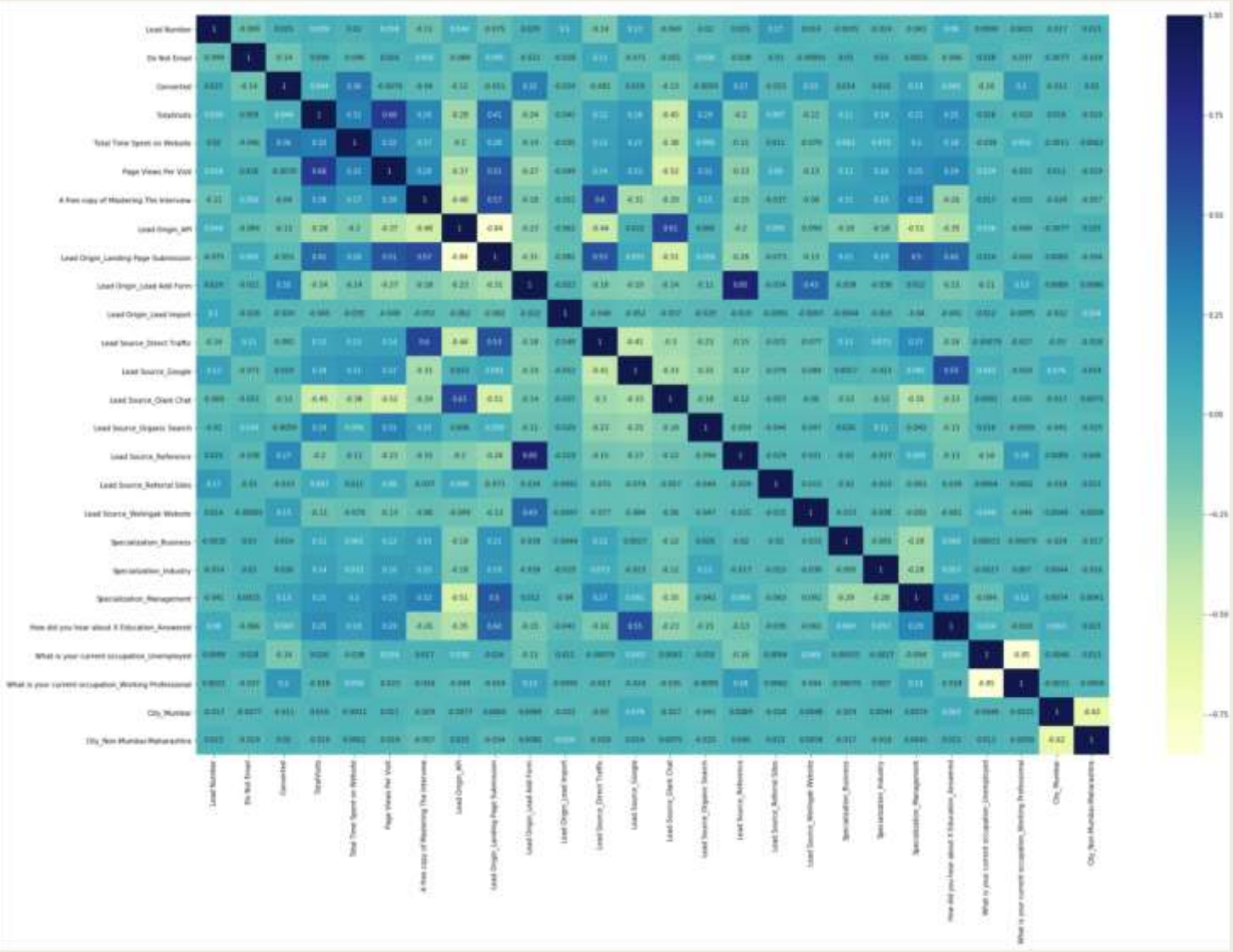






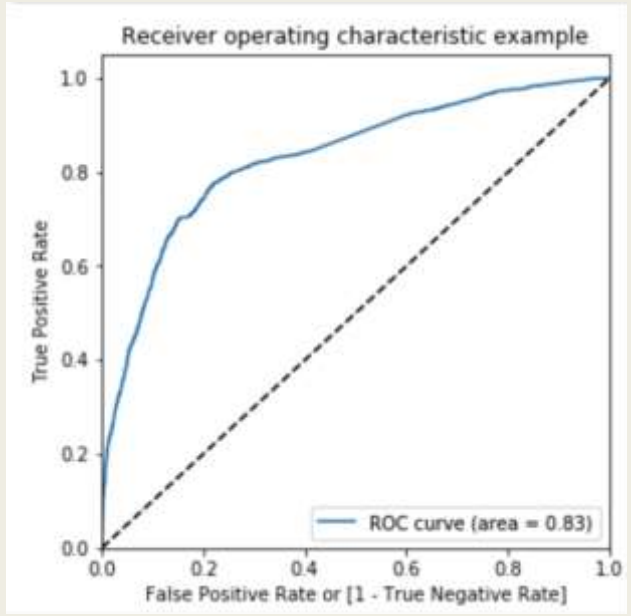
DATA PREPARATION

Step	Action
Convert Binary Columns to 0 or 1	A free copy of Mastering The Interview Column values mapped to 0 or 1
Dummy Variables creation	For the bellow Categorical Columns 1] Lead Origin 2] Lead Source 3] Specialization 4] What is your current occupation 5]City
Split Data into Train and Test Sets	Use train_test_split method with train size=0.7 and test size=0.3
Feature Scaling	Using StandardScaler for numerical Columns 1] TotalVisits 2] Total Time Spent on Website 3] Page Views Per Visit



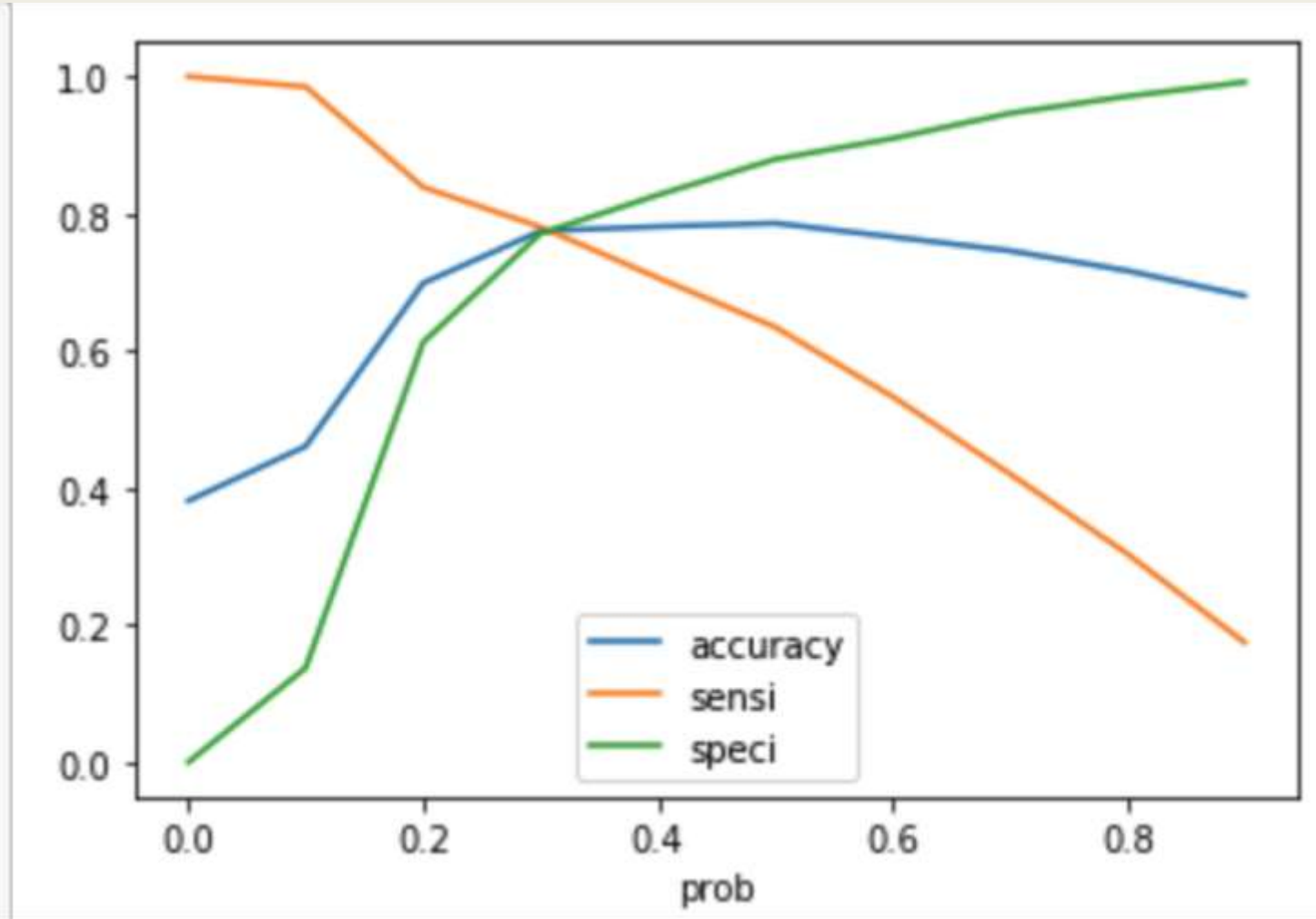
Correlation Between the Attributes

	Features	VIF
8	Specialization_Management	1.76
3	Lead Source_Direct Traffic	1.69
2	Lead Origin_Lead Add Form	1.41
5	Lead Source_Welingak Website	1.24
1	Total Time Spent on Website	1.23
9	What is your current occupation_Working Profes...	1.19
4	Lead Source_Olark Chat	1.18
6	Specialization_Business	1.14
7	Specialization_Industry	1.10
0	Do Not Email	1.09



Logistic Regression Model

1 Perform Feature Selection using RFE Method
2 Drop the columns with high p-Value or VIF value
3] Calculate Accuracy, Sensitivity, Specificity
4] Plot the ROC Curve



FINDING OPTIMAL CUTOFF

From the plot we can say that optimal cutoff is at probability

0.3

Measure	Train Dataset	Test Dataset
Accuracy	77.5%	77.1%
Sensitivity	78.1%	78%
Specificity	77.1%	77%
Precision	68%	69%
Recall	78%	78%

INFERENCES

Conclusion

- The source from which a lead is approaching the X Educations becomes an important criterion. **Prospects approaching X Education via Olark Chat or Welingak Website have high chances of conversion and will require rigorous follow-ups**. Leads from direct traffic have a very low probability of conversion, so the team can divert their energy on prospects from other sources
- It is highly recommended to follow up leads that have **originated via the Lead Add form**. They have a very chance of converting
- Time the prospect is spending on the X Education website is another criterion that the sales team can consider. The **more time lead is spending signifies the more interested** they are in taking up the course and hence will require good follow-ups
- Prospects that **choose a specialization while filling the form** have more clarity and have a higher chance of conversion rather than those that leave it blank
- **Working Professionals** generally tend to convert more than students, housewives, etc.
- Keep a check on prospects that are **opting out of Email**. There are highly **unlikely to convert**

Interpreting the Model

The model was built that assigns a lead score to each prospect. **The Sales team can run that model on new prospects and check if the score is greater than 30**. If yes, then the prospect will have a **78% chance of a conversion**. This will save a lot of time for the sales team as they can focus on leads that have a good chance of conversion