

# LEADS SCORING CASE STUDY



M T W T F S S

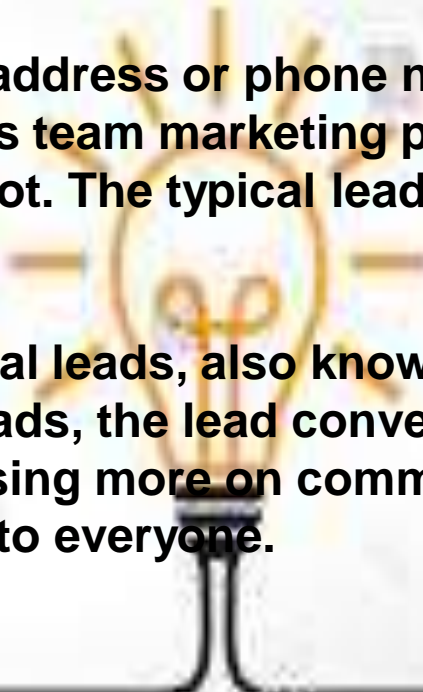
						1
2	3	4	5	6	7	8
9	10	11	12	13	14	15
16	17	18	19	20	21	22
23	24	25	26	27	28	29
30	31					

# Problem Statement

**An education company named X Education sells online courses to industry professionals.**

**When people fill up a form providing their email address or phone number, they are classified to be a lead. Through the sales team marketing process, some of the leads get converted while most do not. The typical lead conversion rate at X education is around 30%.**

**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 Objectives ...



- X Education needs help in selecting the leads that are most likely to convert into paying customers.
- The company requires you to build a model wherein you need to assign a lead score to each of the leads such as lead score have a higher conversion and vice versa.
- The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%.

## DATASETS CONTAIN

- Provided with a leads dataset from the past with around 9000 data points. This dataset consists of various attributes such as Lead Source, Total Time Spent on Website, Total Visits, Last Activity, etc.
- The target variable, is 'Converted' which tells whether a past lead was converted or not (wherein 1 → converted and 0 → wasn't converted).
- Details about the dataset are provided in the data dictionary.



# Importing dataset

## ▪ Data cleaning and pre processing

- Inspecting rows, columns, basic information ,type of the column
- Finding null values and its percentage,
- Deleting columns with null values > 40%
- Dealing with outliers and suggested possible treatment/correction required
- Finding statistical summary.
- Making standardization of values at required places.
- Changed columns containing “select” category as null value.
- Imputed missing values based on value counts ;such as mode value for categorical columns.
- Dropped unwanted columns after checking its value count and plots.
- Segregated variables according to its types – numerical and categorical.



# Data manipulation

Total Number of Rows=37, Total Number of Columns =9240.

- ▶ Single value features like “Magazine”, “ReceiveMoreUpdates About Our Courses”, “Update my supply”
- ▶ Chain Content”, “Get updates on DM Content”, “I agree to pay the amount through cheque” etc. have been dropped.
- ▶ Removing the “ProspectID” and “Lead Number” which are not necessary for the analysis.
- ▶ After checking for the value counts for some of the object type variables, we find some of the features which have enough variance, which have dropped, the features are: “Do Not Call”, “What matters most to you in choosing course”, “Search”, “Newspaper, Article”, “XEducation Forums”, “Newspaper”, “DigitalAdvertisement” etc.
- ▶ Dropping the column shaving more than 35% as missing values such as ‘How did you hear about X Education’ and ‘Lead Profile’.



## ▪ **Exploratory Data Analysis**

- Performed univariate, bivariate, multi variate analysis. Based on numerical/continuous and categorical variables.
- Pairplots and heat maps created to study the correlation.
- Bivariate analysis with target variable “converted” done.
- Plotting various graphs and visualizations to analyze , derive insights from the dataset.

## ▪ **Data transformation:**

- Encoding of variables to binary form “0” and “1”.
- Creation of dummy variables for categorical columns.

- **Data split into train and test in the ratio 70-30.**

- **Rescaling of numerical columns rescaled using minmax scaler.**

- **Model building:**

- Logistic regression model building done by using Recursive Feature Elimination (RFE) and selected top 18 features.

- Assessed the model using statsmodel summary and checked the p-values and dropped the insignificant p-values  $> 0.05$ .

- The Variance inflation Factor was also assessed to explain the predictor variable against all other predictors.

- Finally ended up with predictor variables with significant p value and good VIF.

- On train set ,

- Optimal cutoff found on using ROC curve and the curve obtained was good with area under coverage (AUC) 95%. A trade-off between precision and recall was plotted to visually see the intersecting at optimal cutoff.



# MODEL EVALUATION

- To gain a comprehensive understanding of our model's performance, a confusion matrix was carefully plotted and analyzed on the training dataset. This matrix allowed us to assess various metrics, including precision, accuracy, recall, and specificity, providing valuable insights into its effectiveness.
- After the successful evaluation on the training set, we proceeded to implement the acquired knowledge and insights onto the test data. Running the model on this independent dataset allowed us to assess its generalization capabilities and validate its predictions in a real-world scenario.
- The same set of functions and evaluation metrics applied during the training phase were then meticulously applied to the test data. By comparing the model's performance on both the training and test datasets, we could ensure its consistency and reliability.
- One essential aspect of our model's evaluation was the calculation of the conversion probability. This critical metric provides valuable information about the likelihood of successful outcomes, which is particularly important for decision-making in various applications.





## ▪ Predictions on test data

### Comparison of metrics for train and test data:-

Parameters	Training data	Test data
Accuracy	0.8913	0.8821
Precision	0.8400	0.8180
Sensitivity (Recall)	0.8867	0.8685
Specificity	0.8942	0.8898
False Positive Rate	0.1057	0.1101
Negative predictive value	0.9265	0.9222

# Conclusion



- Important positive predictors of the case with decreasing coefficient values (top 5):-

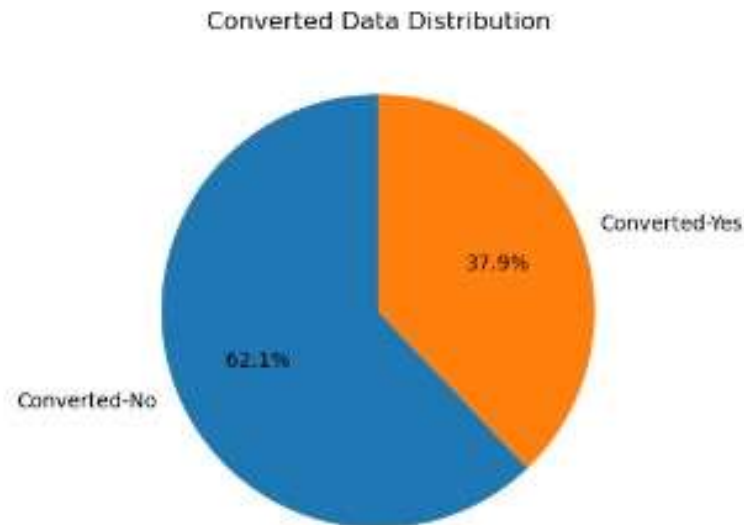
➤ Tags_Closed by Horizzon	8.2740
➤ Tags_Lost to EINS	8.2129
➤ Total Time Spent on Website	4.3879
➤ Lead Origin_Lead Add Form	3.8075
➤ Tags_Will revert after reading the email	3.6797

- Important negative predictors of the case with decreasing coefficient values:-

➤ Do Not Email	-1.8809
➤ Tags_Ringing	-1.6186
➤ What is your current occupation_Unemployed	-1.5692
➤ Tags_switched off	-1.0934



## The target variable 'Converted' data distribution



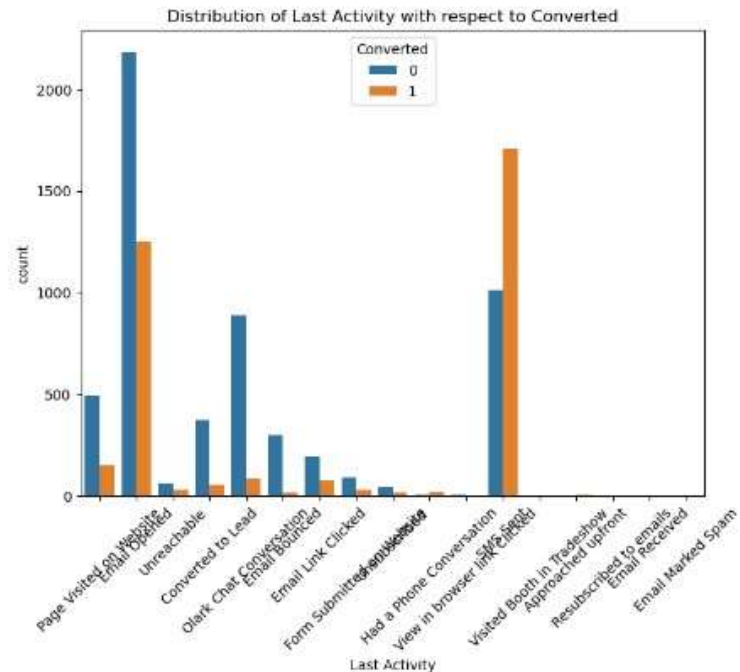
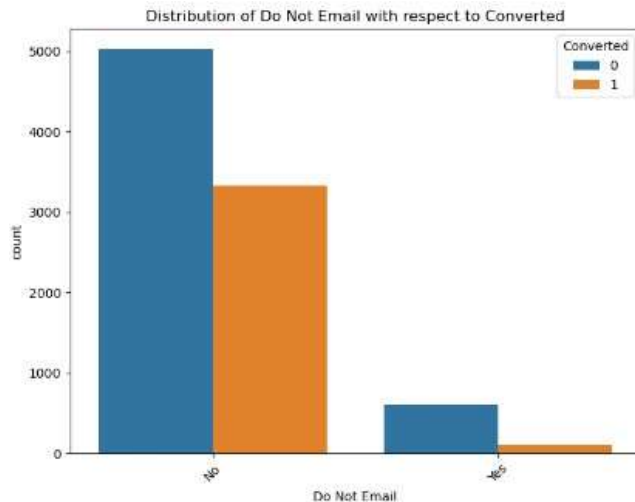
1 means – Yes converted  
0 means- Not converted

Orange- Yes converted – 37.9%  
Blue - Not converted – 62.1%

# Do not email, Last activity Vs Target Converted

1 means – Yes converted

0 means- Not converted



## Insights:

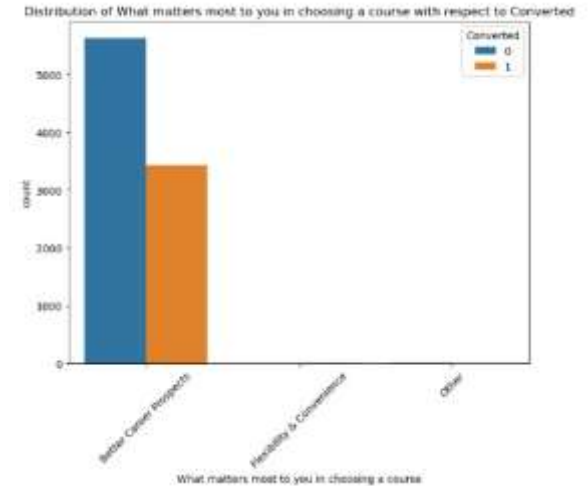
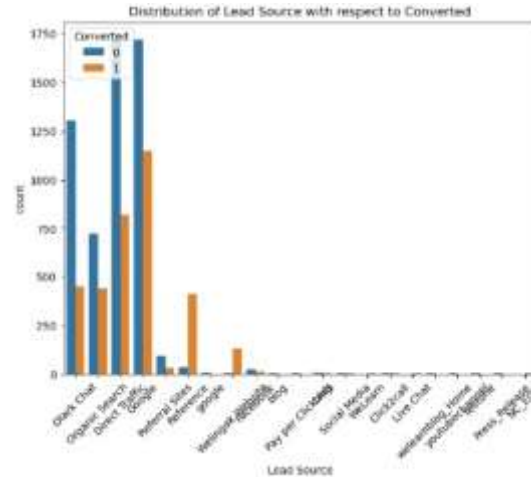
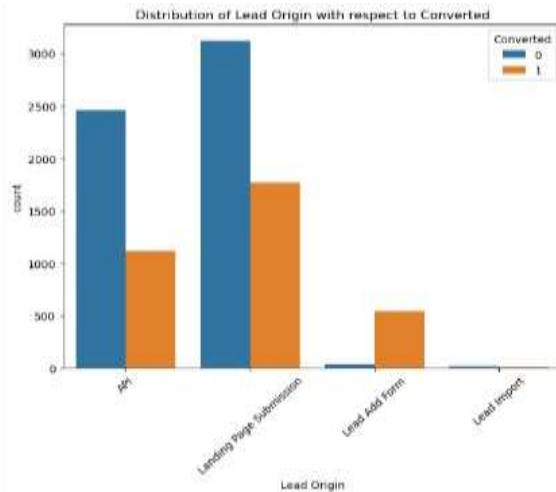
- Customers who have preferred do not email option are very less likely to be converted and company can expect very less revenue out of these customers.
- Focusing on the customers with whom Last Activity - Had a Phone Conversation - are likely to be converted with +ve coefficient value 2.9909(Among top 10 +ve predictors).



# Lead origin, Lead source, What matters to choose the course Vs Target Converted

1 means – Yes converted

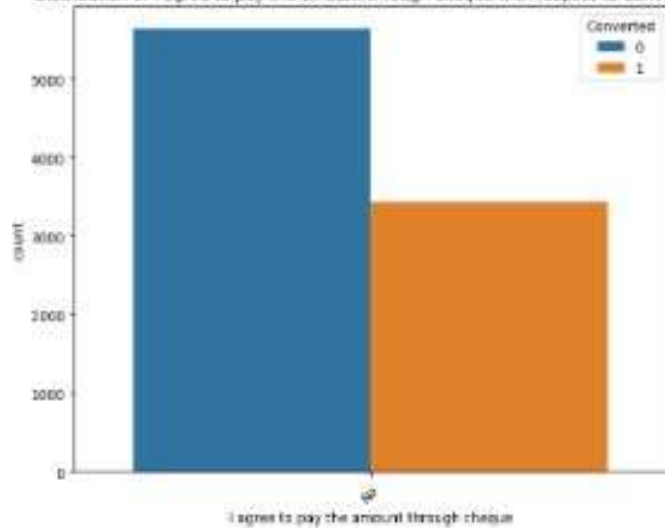
0 means- Not converted



## Insights:

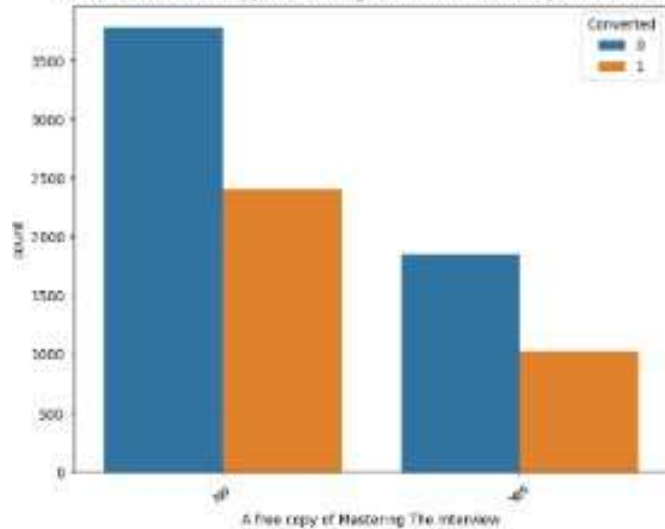
- Company can focus on marketing more on Lead Origin especially on the categories Lead Add Form , Lead Import since they are the top 5 potential positive predictors; and likely to be converted.
- On lead source – more conversion are likely seen in google search method.
- People prefer better career aspect for the

Distribution of I agree to pay the amount through cheque with respect to Converted

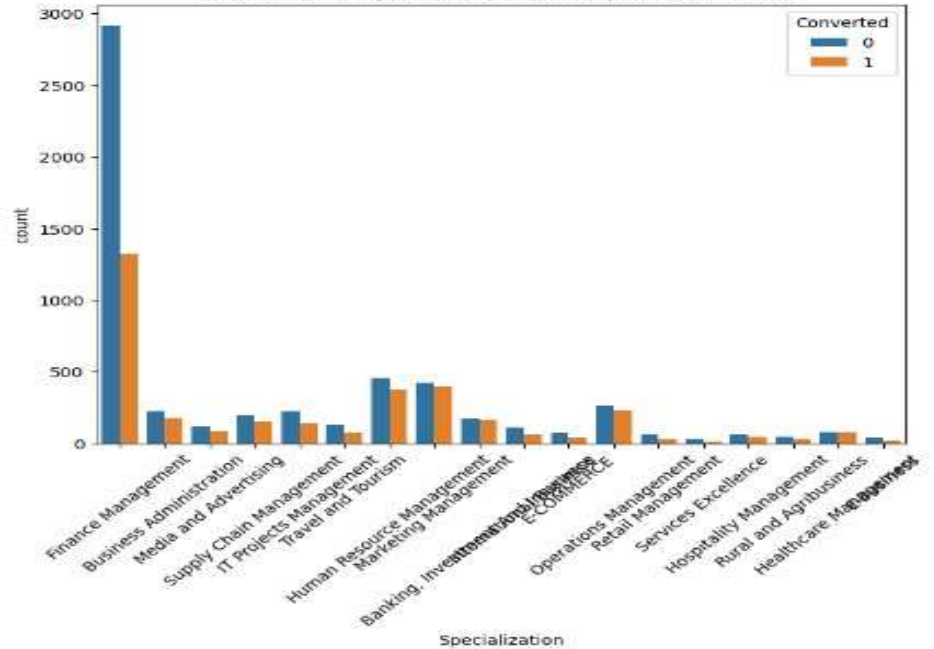


a) Converted 0 = 43.12%, Converted 1 = 38.88%  
 b) Converted 0 = 54.38%, Converted 1 = 35.62%

Distribution of A free copy of Mastering The Interview with respect to Converted

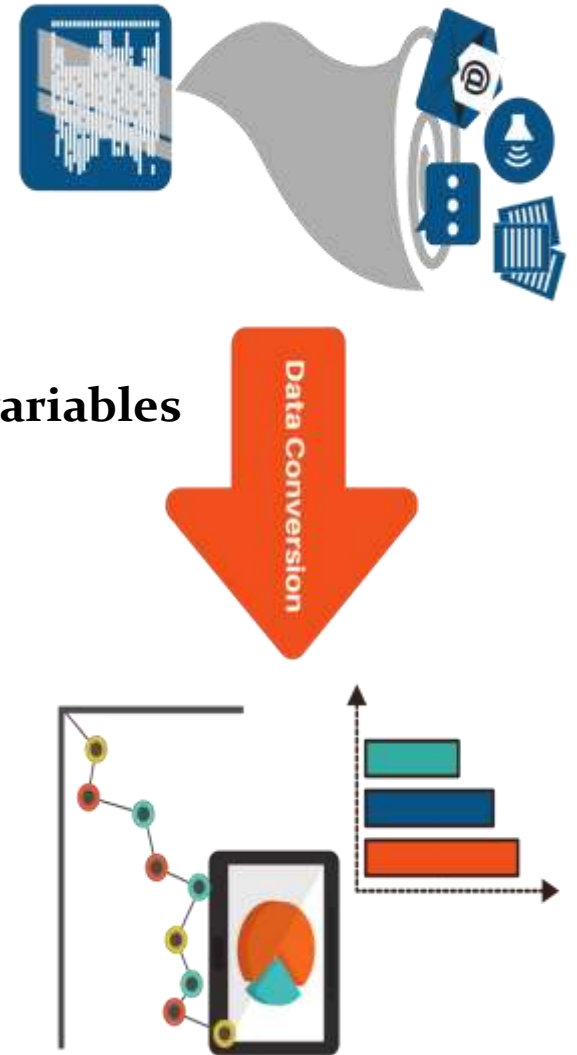


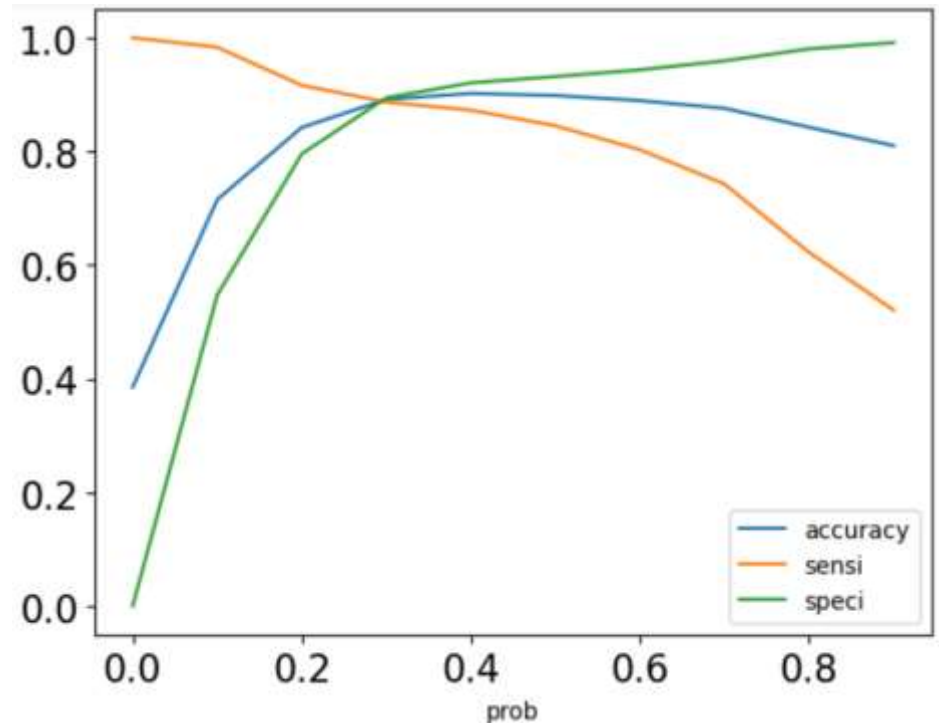
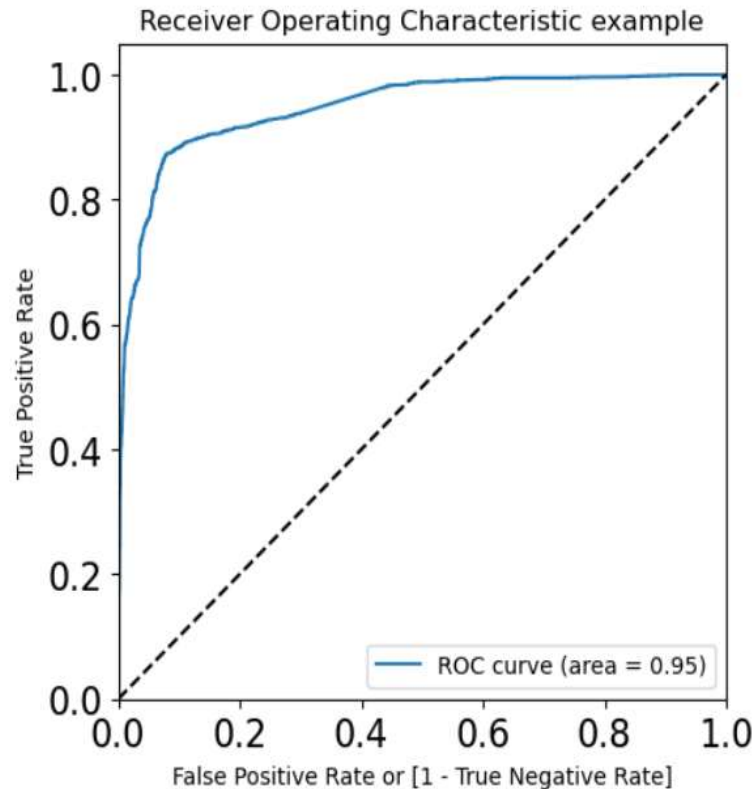
Distribution of Specialization with respect to Converted



## Data Conversion

- Numerical Variables are normalized
- Dummy Variables are created for object type variables
- Total Rows for Analysis: 9240
- Total Columns for Analysis: 37





## Finding Optimal Cut off Point

- Optimal cut-off probability is that
- Probability where we get balanced sensitivity and specificity.
- From the second graph it is visible that the optimal cut off is at 0.3



# CONCLUSION



- It was found that the variables that mattered the most in the potential buyers are (In descending order) :
- The total time spent on the Website.
- Total number of visits.
- When the lead source was: Google
- Direct traffic Organic search Welingak website
- When the last activity was: SMS
- Olark chat conversation
- When the lead origin is Lead add format.
- When their current occupation is as a working professional.
- Keeping these in mind 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

## **Analyzing the above model, the Company should focus on following features:**

### **➤ Increase Marketing For:**

- Focus on leads with specific tags like "Closed by Horizzon" and "Lost to EIN" for higher sales potential.
- Concentrate on lead sources such as "Lead Add Form" and "Lead Import" as they show promise for positive outcomes.
- Target customers who spend more time on the website, as they have a higher conversion potential.
- Prioritize leads with a history of "Last Activity - Had a Phone Conversation" as they are more likely to convert.
- Direct marketing efforts toward working professionals, especially those in current occupations, as they are likely to convert.

### **➤ Decrease Marketing For:**

- Reduce focus on customers who have opted out of email and phone communication, as well as those who are frequently unreachable.
- Expect lower conversion and revenue from these segments.
- Explore offering attractive promotions and referral credit cashbacks to potentially increase revenue from less likely customers.



**Thank You**

**By – Nandini, Nanda, Naren**