Lead Scoring Case Study

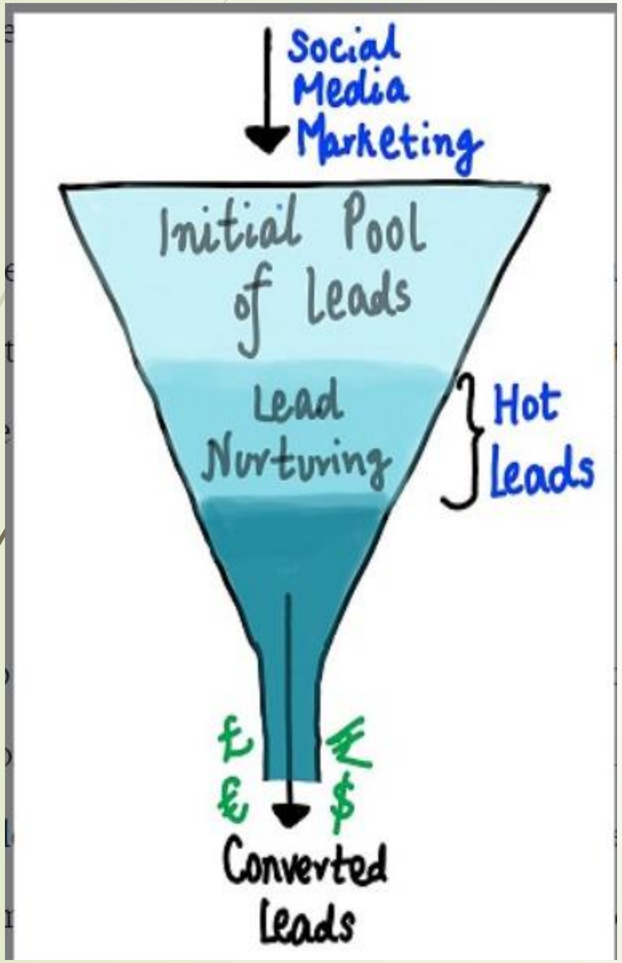
Group members

- Karishma Singh
- Ashish Pawar
- Shrikant Angre

Agenda

- ❑ X Education: Understanding The Customer’s Business & Business Problem
- ❑ Expectations From X Education
- ❑ Solution Methodology
- ❑ Data Cleaning & Manipulation
- ❑ Exploratory Data Analysis
- ❑ Creating Dummy Variables and Scaling Numerical variables
- ❑ Model Building & Evaluation
- ❑ Using ROC method to optimize
- ❑ Evaluating Model on Test Data Set
- ❑ Conclusion and Recommendations

Understanding The Customer’s Business & Business Problem



- ❑ X Education is a leading Ed-tech firm which sells online courses to industry professionals across multiple geographies
- ❑ They typically market their courses across various websites and search engines
- ❑ Potential customers fill a form for a course which funnels into potential leads for X Education, which is then taken up by the Sales team for conversion
- ❑ Although X Education gets a lot of leads, its lead conversion rate is very poor. The current conversion rate is 30%
- ❑ In order to make this more efficient, the company wishes to identify the most potential leads, also known as ‘Hot Leads’.
- ❑ By identifying ‘Hot leads’, the lead conversion rate can go up, as the sales team will now be focusing more on communicating with the potential leads rather than making calls to everyone.

Expectations From X Education

- ❑ Expectation is to help identify the most promising leads, i.e. the leads that are most likely to convert into paying customers.
- ❑ Build a model, to assign a lead score to each of the leads
  - ❑ Higher lead score → Higher conversion chance
  - ❑ Lower lead score → Lower conversion chance
- ❑ The CEO is expecting the target lead conversion rate to be around 80%.

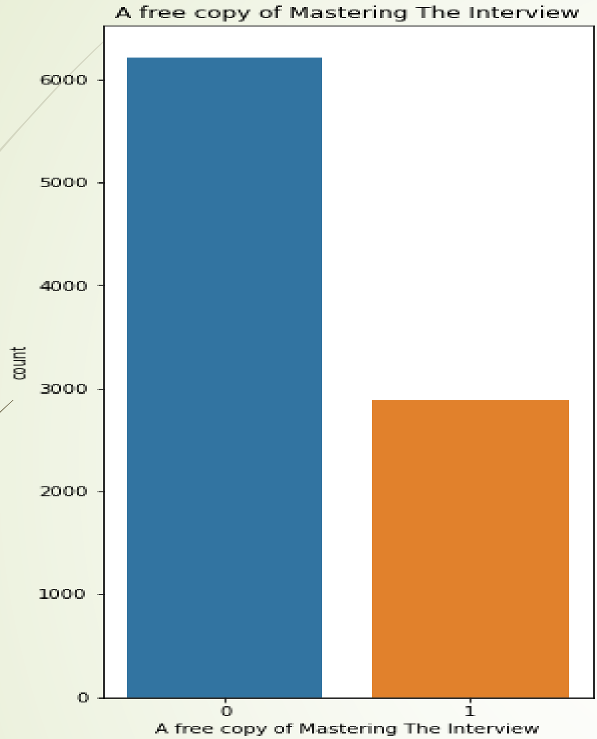
Solution Methodology

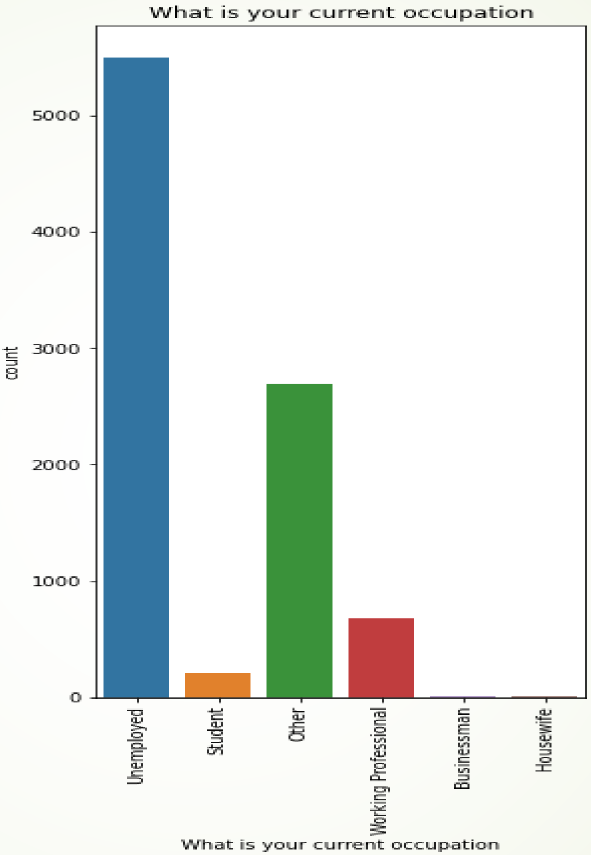
- Data cleaning and data manipulation
    - Check and handle duplicate data
    - Check and handle NA values and missing values
    - Drop features, if it contains large amount of missing values and not useful for the analysis or are duplicate in nature
    - Imputation of the values, if necessary with mode for categorical features, median/mean for numerical features or other, so as not to create any kind of imbalance
    - Converting long tail of certain categories to Other
    - Converting certain categorical variable into binary format to get better analysis
    - Check and handle outliers in data
- Exploratory Data Analysis
    - Univariate data analysis: value count, distribution of variable etc.
    - Bivariate data analysis: correlation coefficients and pattern between the variables etc.
    - Multivariate analysis using heatmap to see any kind of correlation between the variables etc.
- Creating Dummy variables for categorical features and encoding of the data.
- Train – Test Splitting of data
- Feature Scaling for numerical variables
- Classification technique: logistic regression used for the model making and prediction on Train data set
- Model Evaluation and Optimization
- Prediction on Test data set
- Conclusions and recommendations.

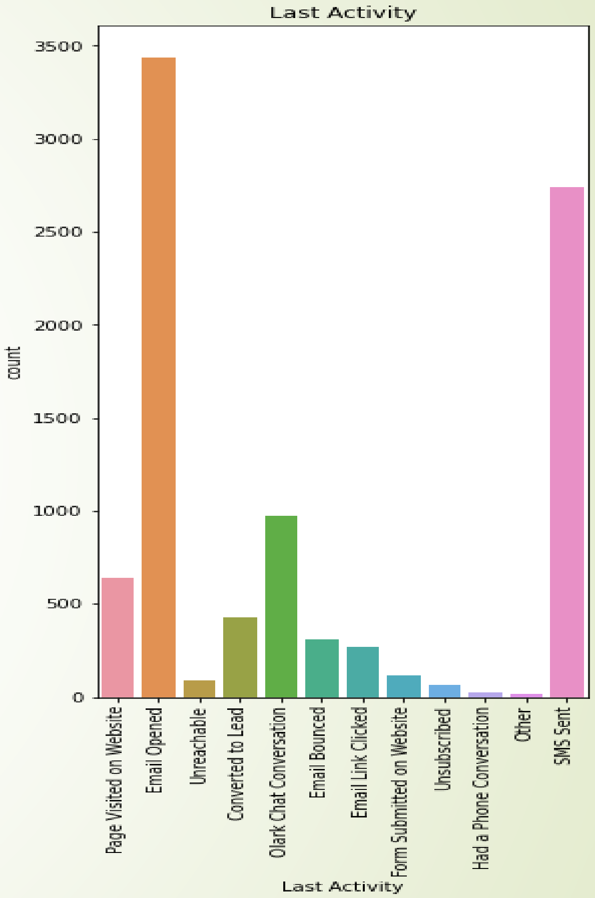
Data Cleaning & Manipulation

- ❑ Data has 37 columns and 9240 rows
- ❑ After analyzing the missing data and percentage for every features, we dropped any feature which has more than 40% or more missing value
- ❑ We dropped all the features which has more than 98% of single value like 'No'
- ❑ We also dropped all the features like "Tags", "City, "Country", "How did you hear about X Education", "Lead Profile" which has 70% or more of missing + non meaningful value like "Select"
- ❑ Imputed "Select" and NaN with "Not Selected" for "Specialization", so that no imbalance is created
- ❑ Drop feature "Last Notable Activity" as the data is almost similar to "Last Activity" feature
- ❑ Imputed missing value with mode for categorical variables like "Last Activity", "Lead Source"
- ❑ Converted long tail of features like "Last Activity", "Lead Source" into 'Other'
- ❑ Converted Categorical variables like "Do Not Email" & "A free copy of Mastering The Interview " into Binary format
- ❑ Dropped rows with missing values for "TotalVisits" & "Page Views Per Visit", as the percentage was just 1.48%
- ❑ Checked for Outliers for "TotalVisits" and "Total Time Spent on Website", however the data seemed to be legitimate, hence left the data as it is
- ❑ After this step there are 11 columns and 9103 rows remaining in the data set for further analysis

Exploratory Data Analysis – Univariate Analysis



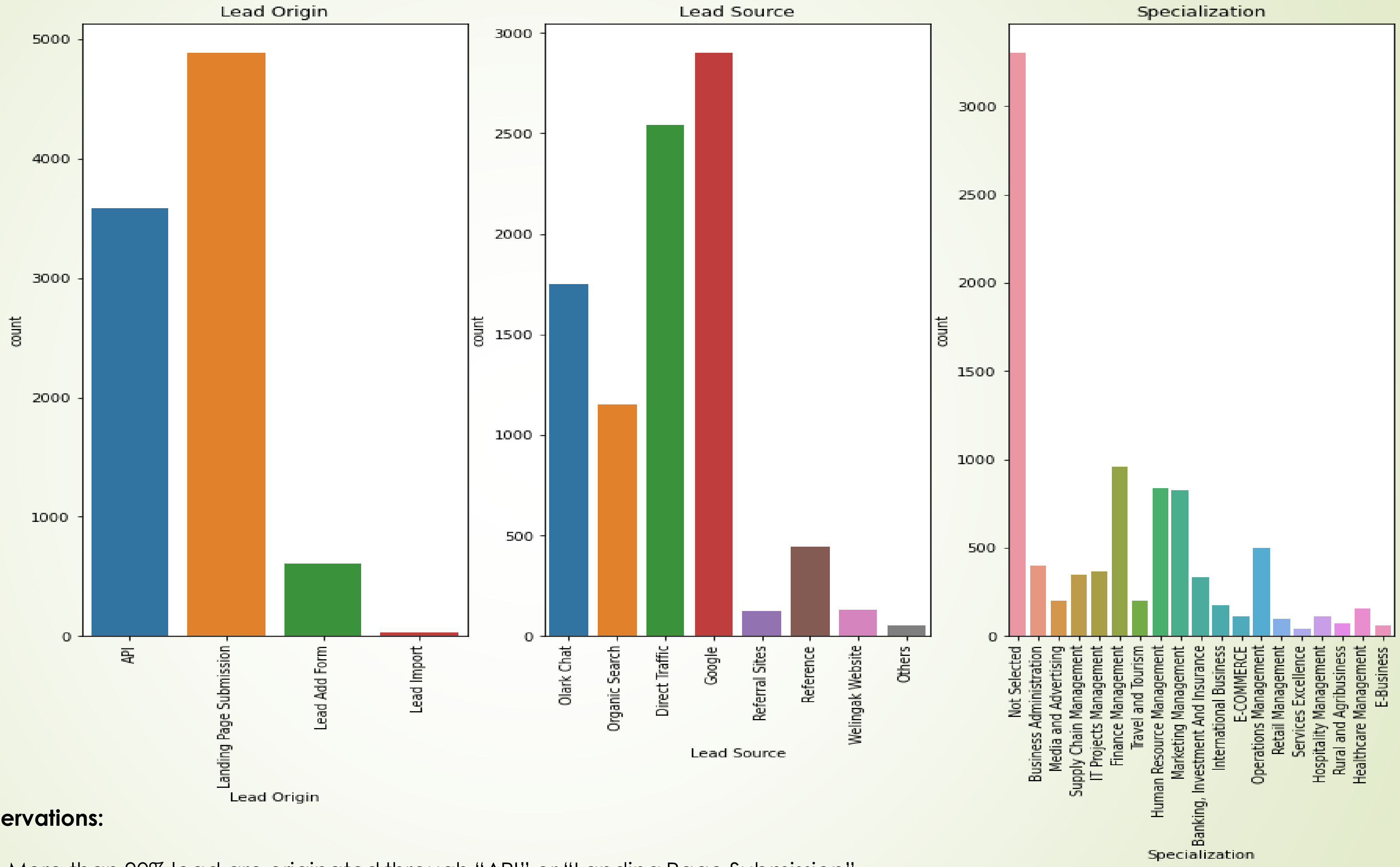




Observations:

- Less than 33% customers are opting for “A free copy of Mastering The Interview”
- Unemployed seems to be around 60% of the population and along with Other they constitute with more than 85% of the population
- Email Opened and SMS Sent are constituting to 66% of the population

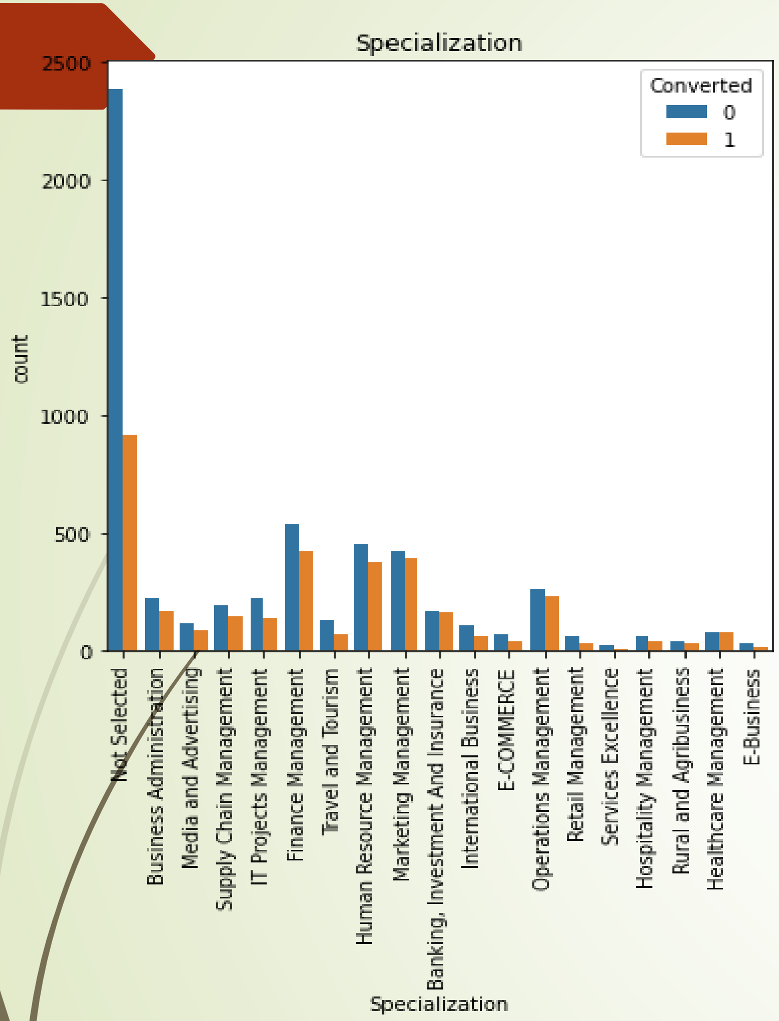
Exploratory Data Analysis – Univariate Analysis contd..

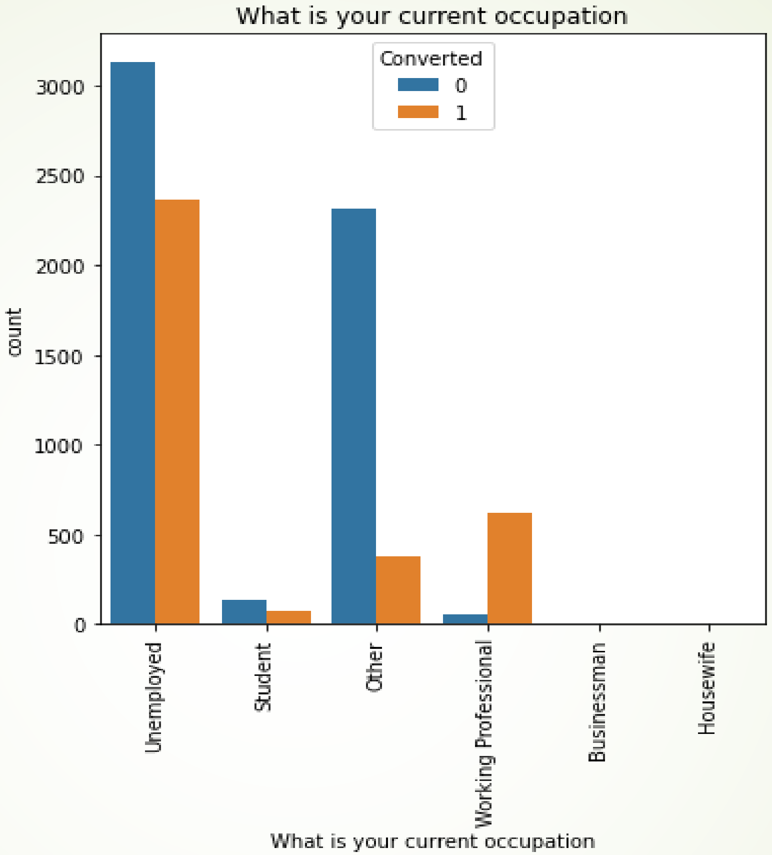


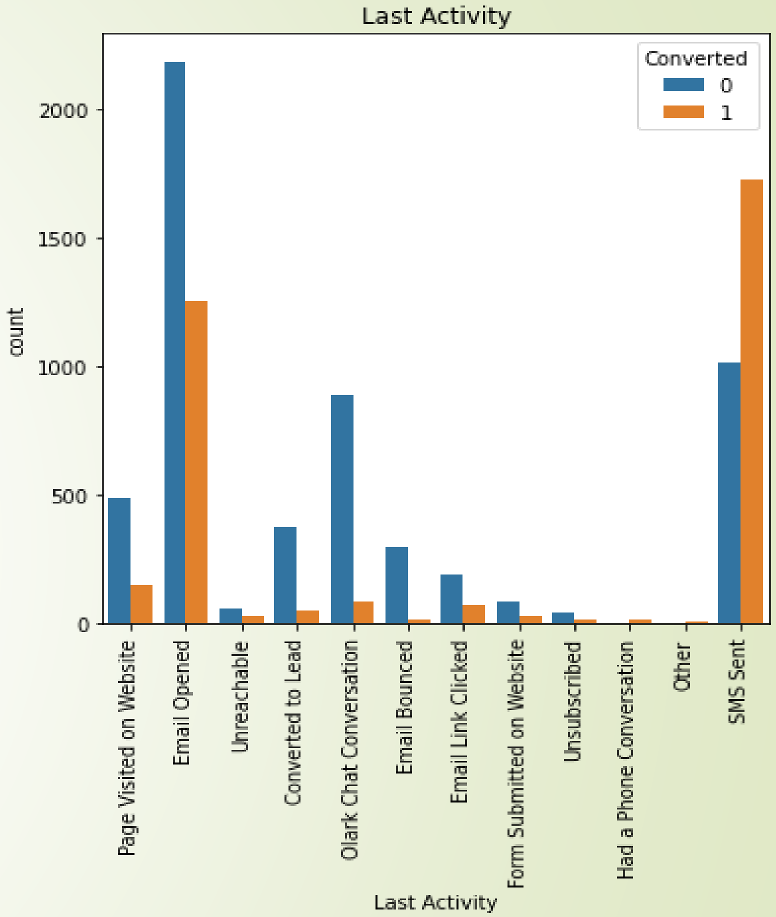
Observations:

- More than 90% lead are originated through “API” or “Landing Page Submission”
- Google and Direct Traffic constitutes to around 57% of the population of the Lead Source
- 37% of the leads have not selected any specialization and a lot of people are opting for different types of management courses

Exploratory Data Analysis – Bivariate Analysis



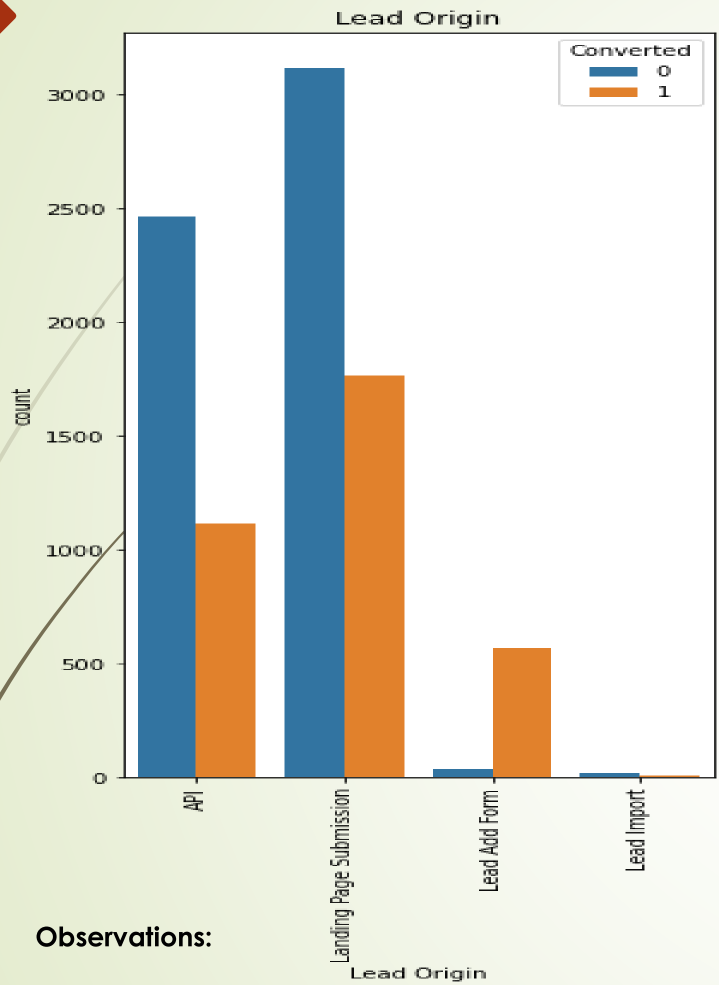


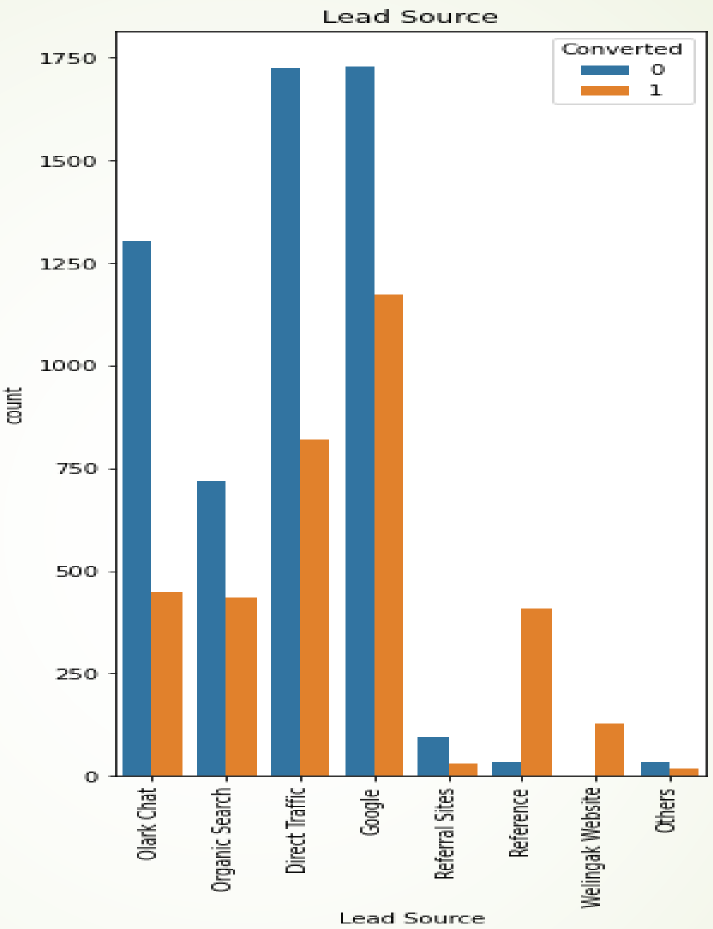


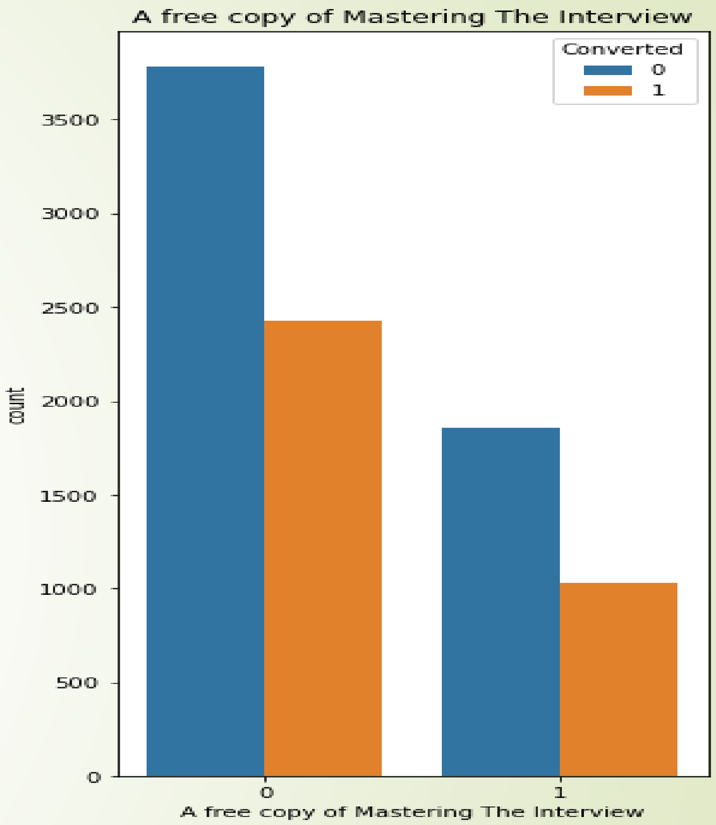
Observations:

- Lead conversion is far better for Working Professionals who are either looking to upskill or change their field
- Unemployed leads are the most in terms of Absolute numbers and can be tapped with better marketing offers to improve the conversion rate
- Leads which are looking for Human Resource, Marketing and Operations Management course have much better conversion rate.. which means that the X Education is well know for these streams
- Lead conversion for Finance Management stream is very less compared to other streams and X Education needs to work towards that
- Emailed Opened and SMS sent has higher leads and higher conversion rate. More focus should be on these 2 categories

Exploratory Data Analysis – Bivariate Analysis contd..



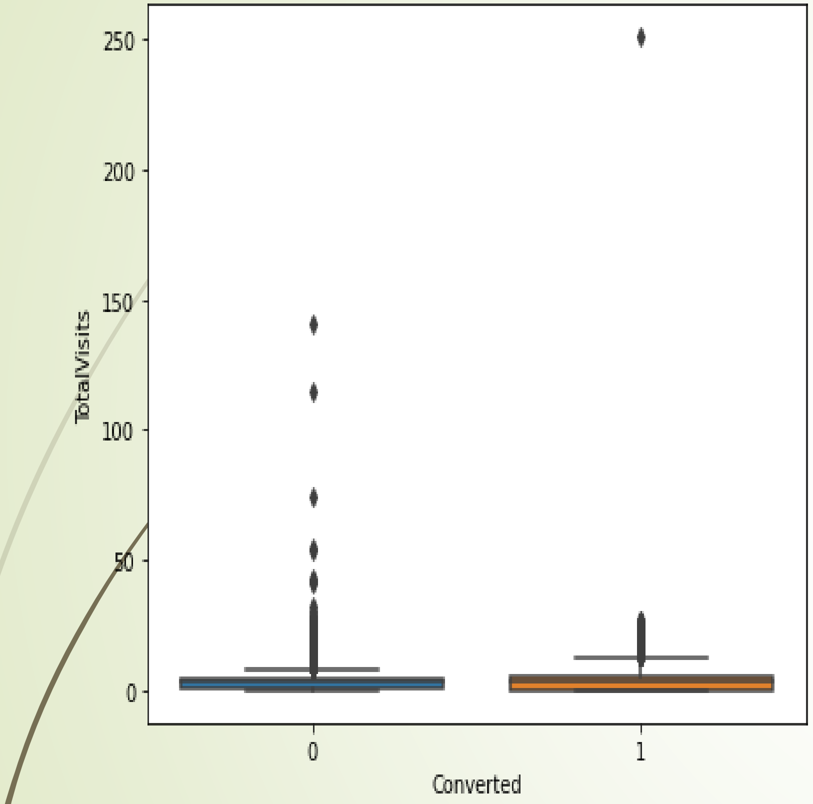


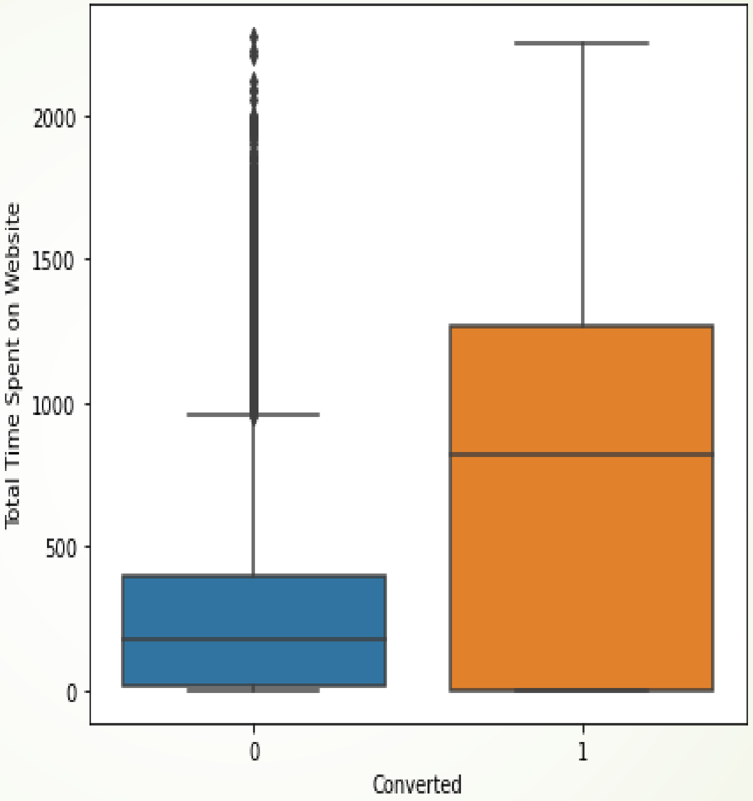


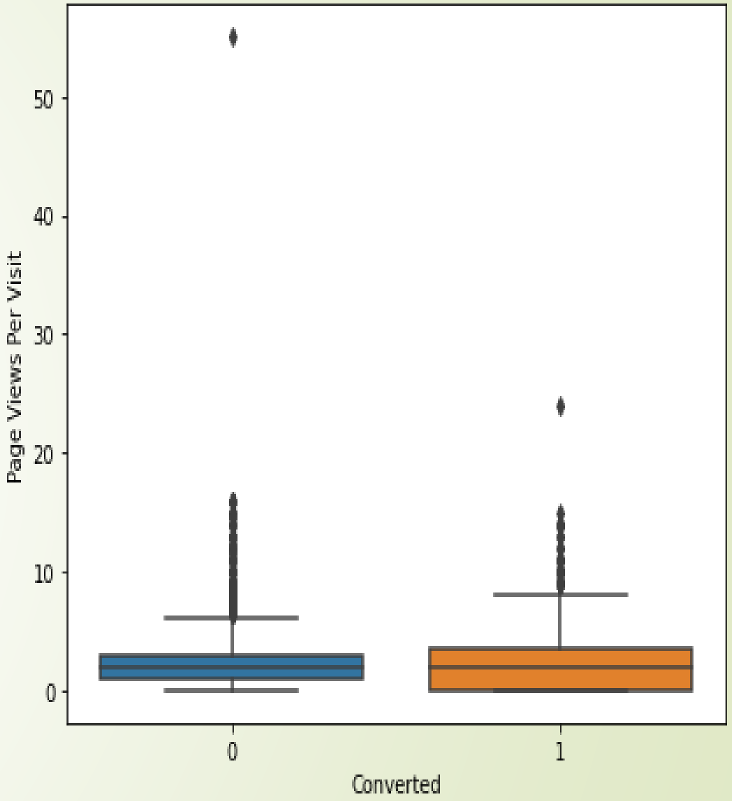
Observations:

- API and Landing Page Submission bring higher number of leads as well as conversion.
- In order to improve overall lead conversion rate, we have to improve lead conversion of API and Landing Page Submission origin and generate more leads from Lead Add Form.
- There is no impact of giving "A free copy of Mastering The interview" on lead generation..
- Maximum # of leads are generated by Google and Direct Traffic channels.
- Reference seems to be having very high conversion rate. Same is the case with Welinkar Website
- To improve the lead conversion rate we can focus on Google and Direct Traffic channels, we can also look at investing in Olark_Chat and Organic Search to improve the rate further

Exploratory Data Analysis – Bivariate Analysis contd..



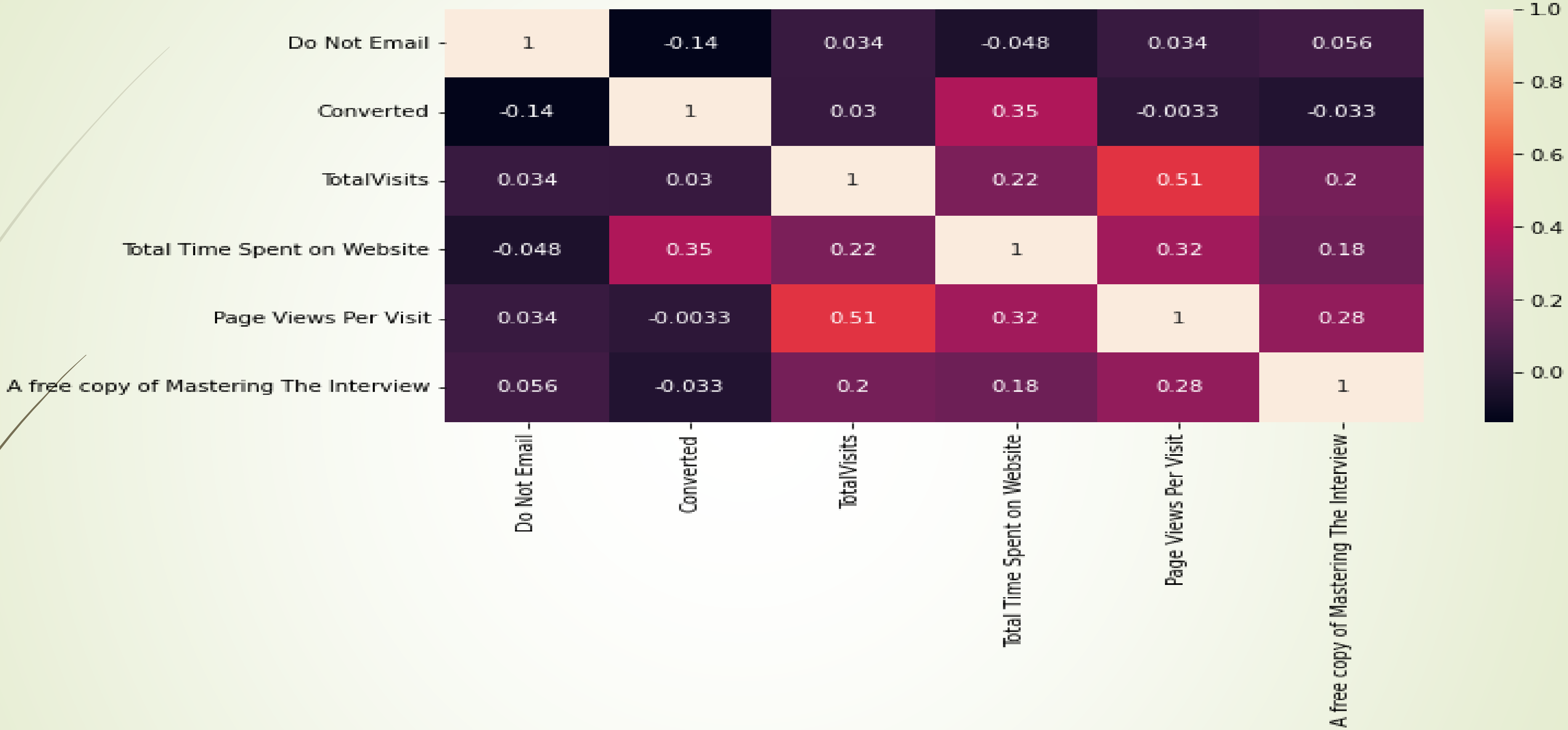




Observations:

- The median of both the conversion and non-conversion are same and hence nothing conclusive can be said using this information.
- Users spending more time on the website are more likely to get converted.
- Websites can be made more appealing so as to increase the time of the Users on websites

Exploratory Data Analysis – Multivariate Analysis



Observations:

- There is some correlation between Conversion and Total Time spent on the website.. albiet it is not so strong
- There is decent correlation between TotalVisits and Page Views per visit and hence we need to keep this in mind while building the model

Creating Dummy Variables, Scaling numerical variables

- Created Dummy variables for categorical features:
  - 'Lead Origin', 'Lead Source', 'Last Activity', 'Specialization’ &'What is your current occupation’
  - Used Drop First feature in order to help create the optimum model

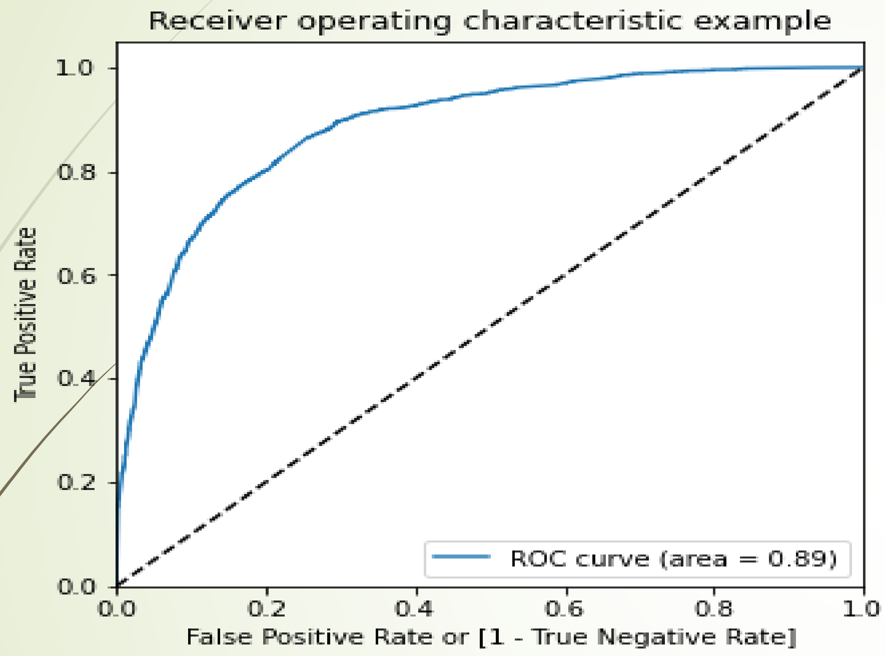
- Did scaling for numerical variables using MinMaxScaler
  - 'TotalVisits','Total Time Spent on Website’ & 'Page Views Per Visit’

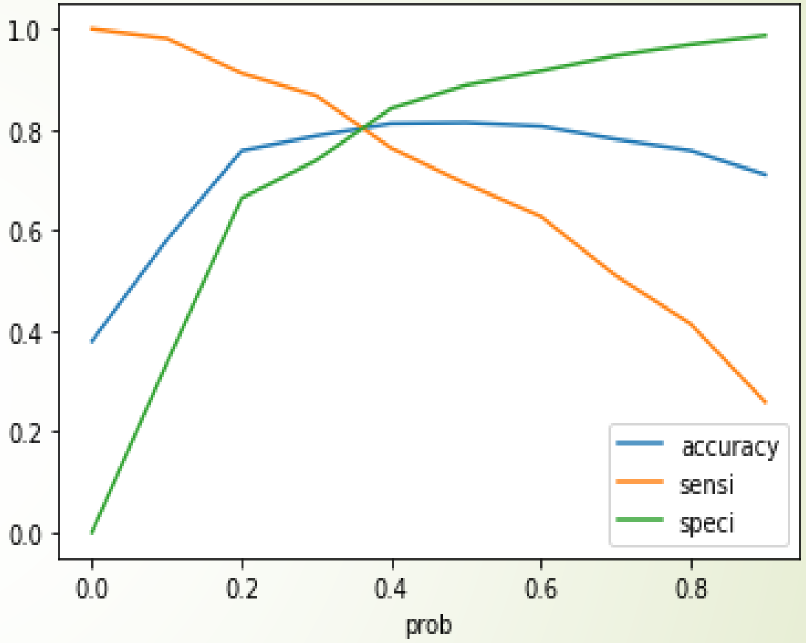
- After this step there are 50 columns and 9103 rows remaining in the data set for further analysis

Model Building & Evaluation

- We have used 70: 30 split for Train and Test Data Set
- Logistic Regression technique is used for building the machine learning model
- Recursive feature elimination (RFE) feature is used for feature selection and the cut off we have kept is 15
- We ended up running 3 iterations of the model and concluded when the p-value was less than 0.05 and VIF was less than 5
- The model was then ran first train data set to check Accuracy, Sensitivity and Specificity which came out as 81%, 69% and 89% respectively
- Next step was to use ROC to optimize the probability cut off

Using ROC to optimize





- ROC method is used to find the optimal cut off point for probability
- Area under ROC curve is 89%, which is a good coverage, also indicates good model
- Looking at the graph we used 0.36 as the probability cut off
- After using this probability cut off the Accuracy, Sensitivity and Specificity scores were 81%, 79% and 82%, which are very descent and indicative of a good model
- We also ran Precision and Recall tests and scores were 72% and 79% respectively

Evaluating Model on Test Data Set

Parameters	Train Data Set	Test Data Set
Accuracy	81%	80%
Sensitivity	79%	80%
Specificity	82%	81%
Precision	72%	72%
Recall	79%	80%

- Above table shows comparison of various parameters like Accuracy, Sensitivity, Specificity, Precision and Recall on Train and Test Data Set
- Looking at the above table we can conclude that the model is robust and stable for lead Score generation
- We also have generated lead_Scores to help X Education identify ‘Hot Leads’. This is a score between 0 to 100

Conclusion

The final formula for this Log Reg model is :

ln (p/(1-p)) = -2.0715 + -1.6069 * Do Not Email + 9.122 * TotalVisits + 4.5188 * Total Time Spent on Website

+ -0.3055 * Lead Origin_Landing Page Submission + 3.6615 * Lead Origin_Lead Add Form + 1.2565 * Lead Source_Olark Chat + 1.8397 * Lead Source_Welingak Website

+ 1.7057 * Last Activity_Had a Phone Conversation + -1.3292 * Last Activity_Olark Chat Conversation + 1.327 * Last Activity_SMS Sent + 1.2425 * Last Activity_Unsubscribed + -1.2159 * What is your current occupation_Other

+ 2.523 * What is your current occupation_Working Professional

- The top Five variables in our model which are contributing most towards the probability of lead getting converted are
  I. TotalVisits
  II. Total Time Spent on Website
  III. Lead Origin_Lead Add Form
  IV. What is your current occupation_Working Professional
  V. Lead Source_Welingak Website
- These are selected basis their coefficient's value in our Logistic Regression model, which are 9.12, 4.51, 3.66, 2.52 and 1.84 respectively

Recommendations

- If we make the website of X Education with better user experience (UX) and attract the prospective leads with contents like Learning sessions, Quiz, Industry updates, offers and Industry led webinars etc. We can drive more traffic to their website and also increase the time spent by the lead on the website. Thus, driving the lead to a conversion.
- We have seen the most of the leads are not filling the lead Add form, if we make the form user friendly and add some marketing offers scholarships etc. to it. that may lead them to provide their details
- We have also seen the potential leads are someone who are not working and have not enrolled for any school. Hence if we nurture these leads well and engage them well. They can easily be converted
- Similarly, if we make Welingak website with better UX, then the lead generation will be higher. We have also seen that the lead conversion from leads coming out of this source is far better. Thus, driving better conversion

When the sales team has around 10 interns allotted to them. They can be deployed to reach out to below type of customers to aggressively target potential leads

- They should call the potential leads who are visiting the website more than 2 times in a short span of time (say in a gap of 2-4 days)
- They can call the potential leads who are spending more than average time on their website or say a threshold of 30 mins or more.
- If they are visiting the enquiry page of one or other course and are spending more than a threshold time on that page
- They can call the folks who are not enrolled for any school and are unemployed
- They can also call working professionals who are looking for a career change or upskilling themselves
- They can reach out to existing students who can provide their referrals. From the data we know that referrals have better conversion rate

Recommendations contd.

When the company reaches its target for a quarter before the deadline. The Sales team can focus on following activities to boost lead generation and conversion

- ❑ Lead Add Form should be made more user friendly so that maximum number of leads can fill it and generate more leads into the sale funnel for further conversion
- ❑ One/two interns can work with the Marketing team to make the lead Add form user friendly, they can also brainstorm with the team, to see, how and when to position the lead add form for a prospect
- ❑ They can work with Marketing team to make the website of X Education with better user experience (UX) and attract the prospective leads with contents like Learning sessions, Quiz, Industry updates, offers and Industry led webinars etc. We can drive more traffic to their website and also increase the time spent by the lead on the website. Thus, driving the lead to a conversion.
- ❑ They can host sessions like Ex-student webinars who have made it big in the industry and guide the potential leads

THANK YOU