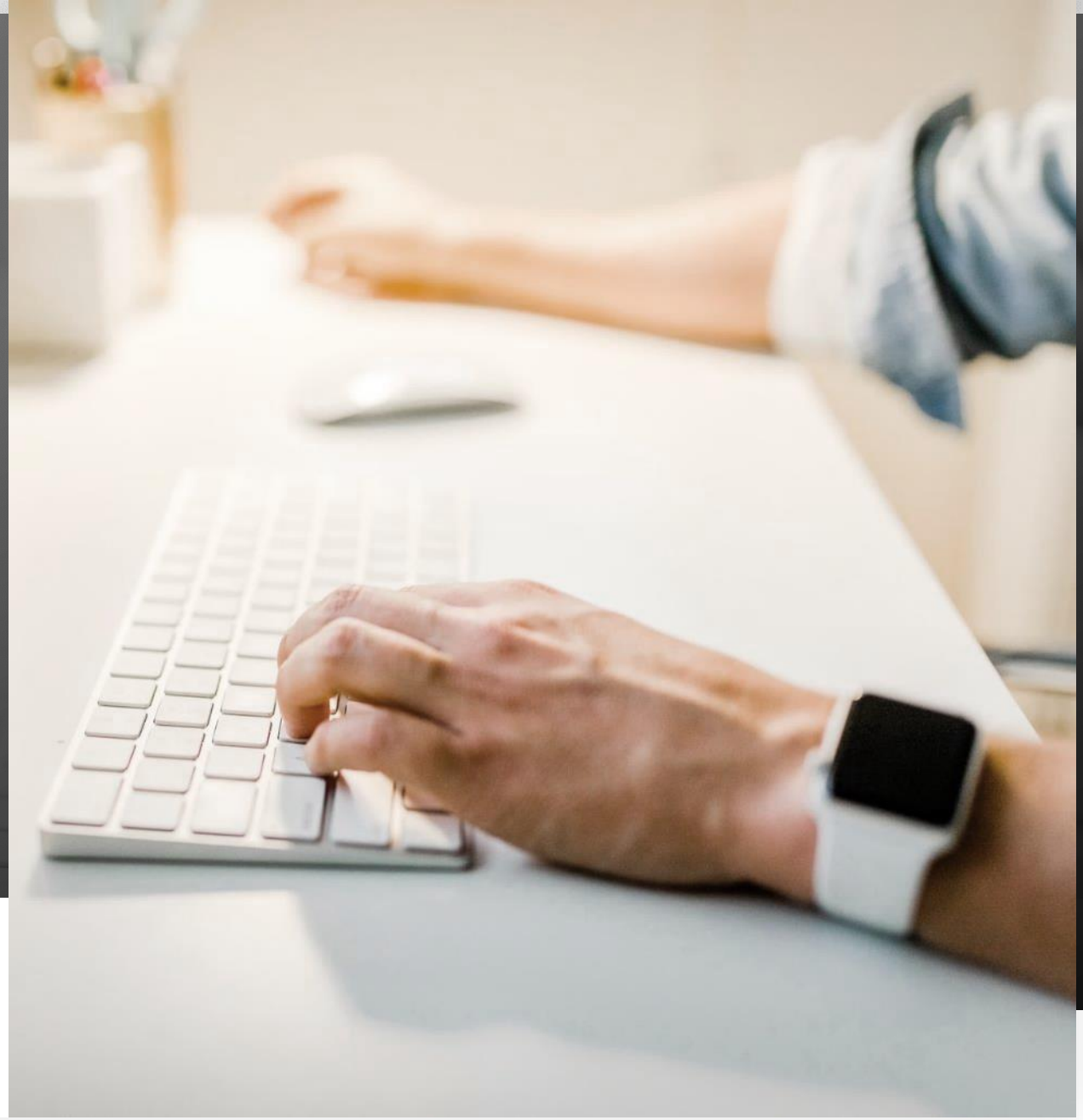
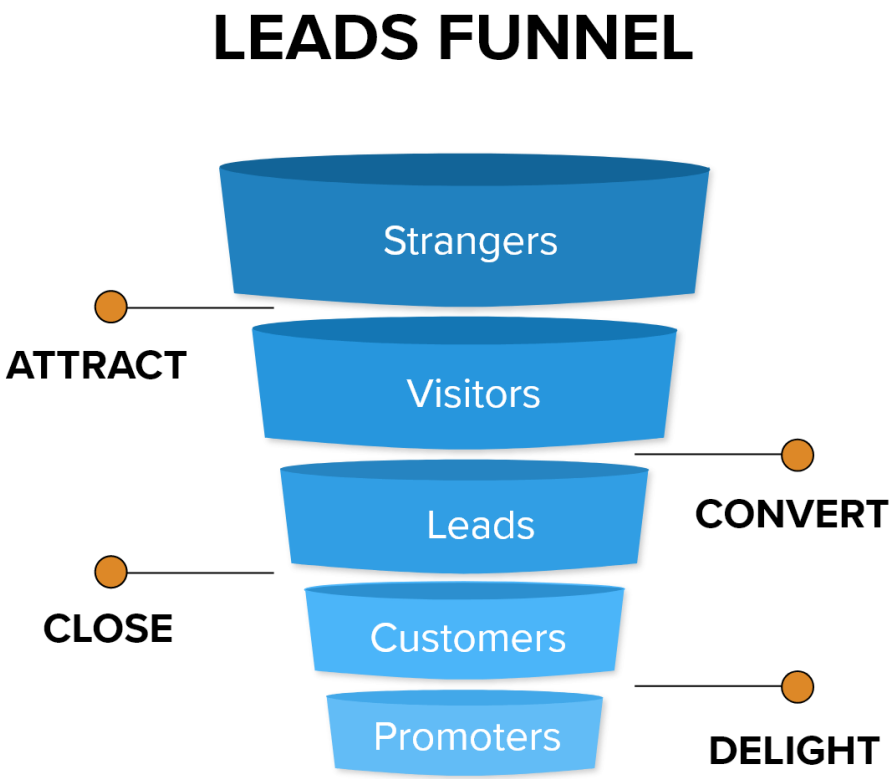
LEAD SCORING CASE STUDY



BY :

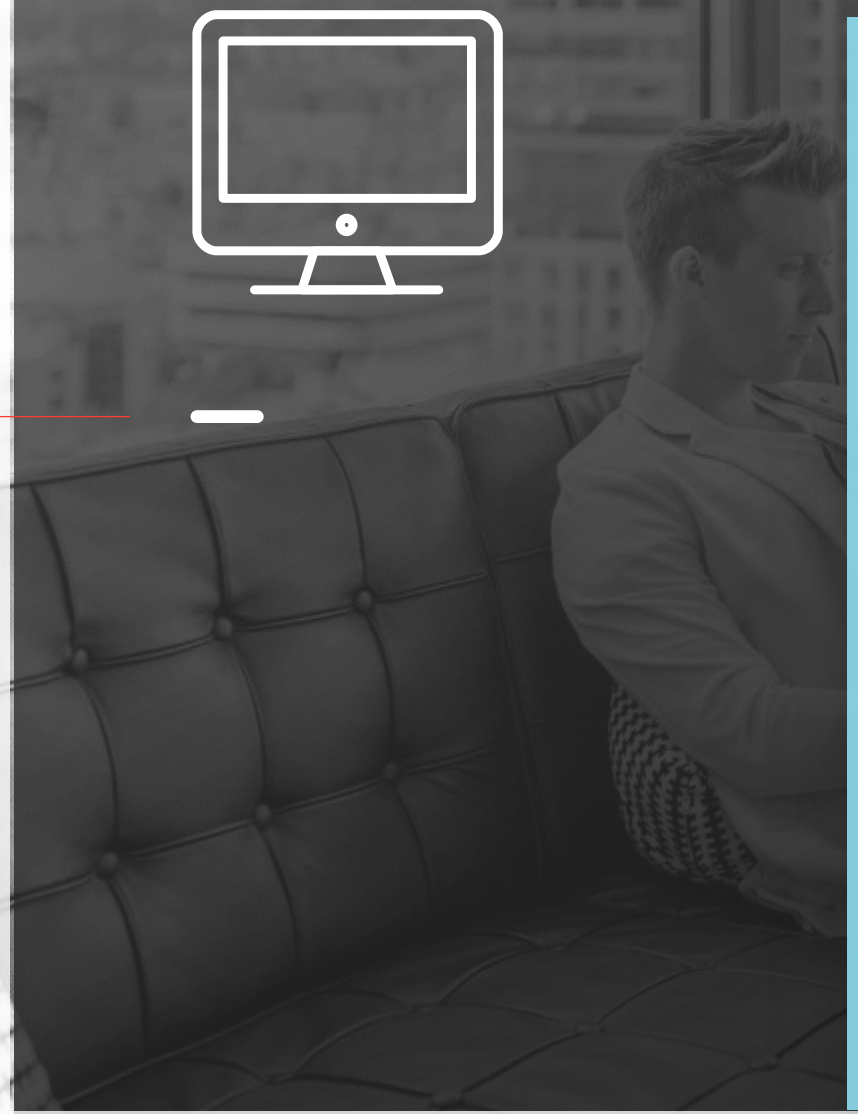
- ARUNA D
- PRERANA V

PROBLEM STATEMENT



- An education company named X Education sells online courses to industry professionals.
- Now, although X Education 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.
- The objective is to build a model to identify the hot leads and achieve lead conversion rate to 80%.

DATA INFORMATION



Information regarding the data:

- **Dataset used : Leads.csv**
- **Total number of customers present : 9240**
- **Total number of features : 37**
- **Model used : Logistic Regression**

- After initial analysis, we see that there are multiple factors that influence conversion rate
- The target column in our dataset : “Converted”
- We need to reduce the features to maximize the conversion rate.
- Current Conversion Rate = 38.53%

DATA CLEANING

Final list of features

- Lead number
- Lead Origin
- Lead Source
- Total Visits
- Total Time Spent
- Page Views per visit
- Last Activity
- Specialization
- Current Occupation
- Free Copy of Book

Handling Select variable

- "Select" variable indicates that the user has not selected any option.
- We impute the same with null values.

Dropping Score and Activity variables

- Score and Activity variables : This is the data that is obtained after contact with the lead. So we need to remove them.
- Score variables: Tags, Lead Quality, Lead Profile, Activity Index, Activity Score and Profile Score.
- Activity variables: Last Notable Activity

Treating Categorical data

- High Data Imbalance – Columns having high data imbalance must be removed.

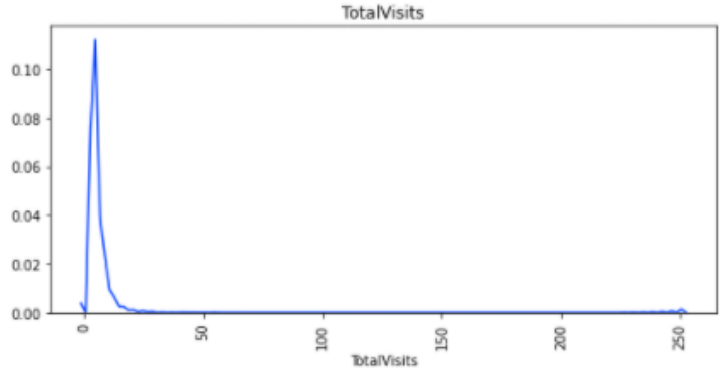
For e.g. : Category A has 98% , and Category B has 2% - This data is irrelevant to our analysis as one category is overpowering the other.

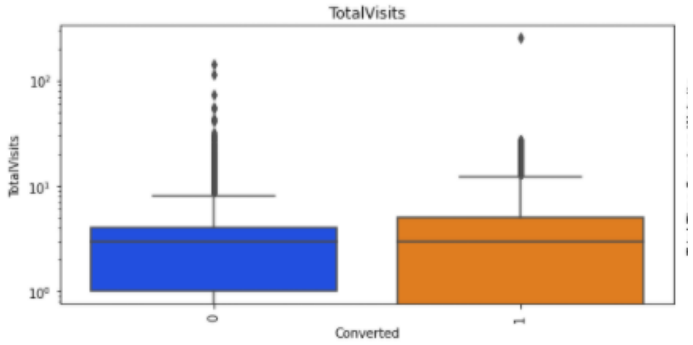
- In other categorical columns where there are columns with small percentages should be removed.

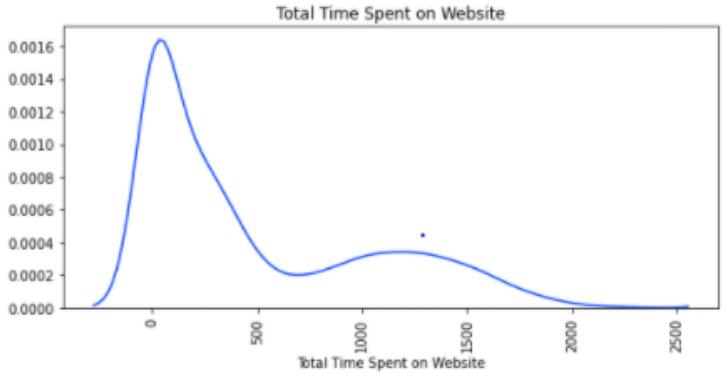
Dropping column with high null values.

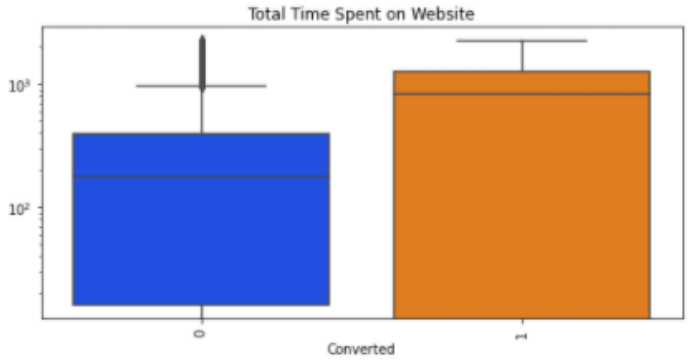
- Columns having null values greater than 40% does not have meaning to the data, hence we drop these columns
- For Specialization, we consider the column where people have not selected any value into one more column known as Not Specified and we use this for model building.

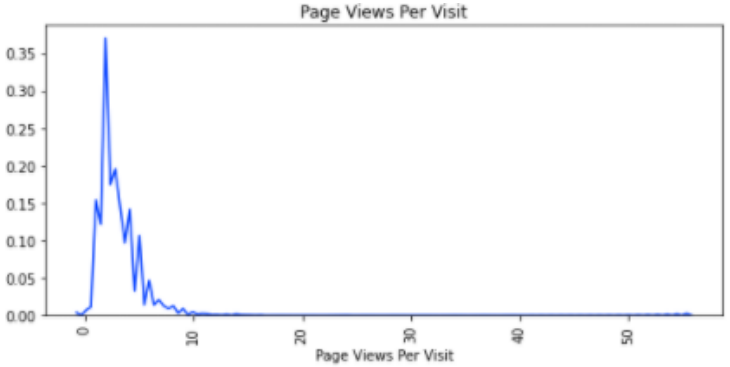
EDA – NUMERICAL ANALYSIS

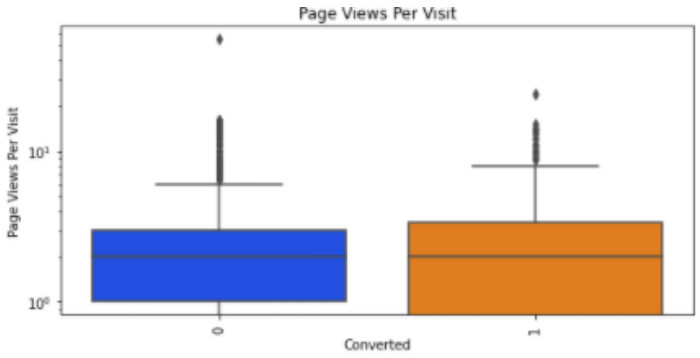












TOTAL VISITS

- The Total Visits is going to increase initially but decreases further. The average is found to be 2-3
- The mean for Total Visits is almost the same for both converted and non converted customers.

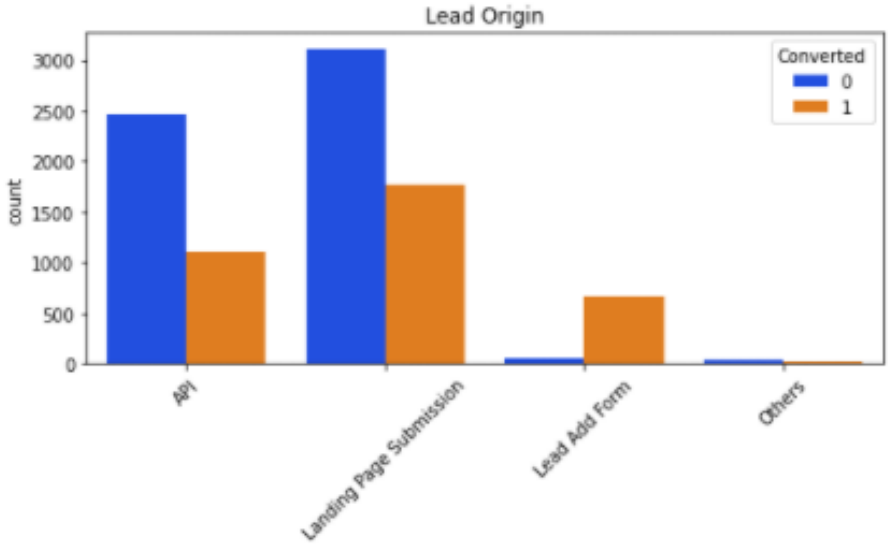
TOTAL TIME SPENT ON THE WEBSITE

- The total time spent is found to be increasing initially and decreases to a large extend and increases again.
- The average is found to be 280 seconds.

PAGE VIEWS PER VISIT

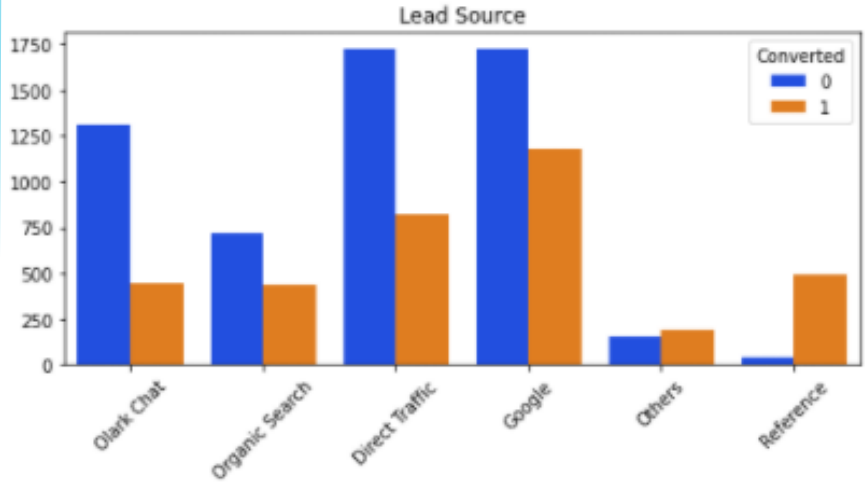
- The average Page views per visit is found to be around 3-5
- The page views for both converted and non converted customers is found to be the same.

EDA – CATEGORICAL DATA



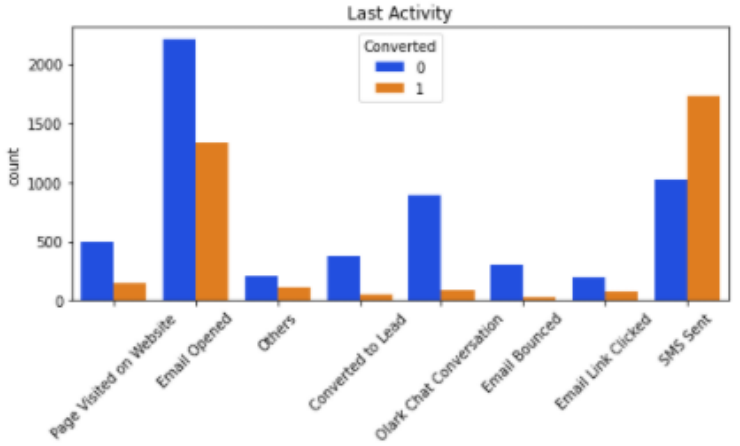
LEAD ORIGIN

- The percentage of converted people is found to be greater for Landing Page Submission.
- We can also see that if Lead Origin – Add Form, the ratio of lead conversion is very high. The number of people not converted is very less.



LEAD SOURCE

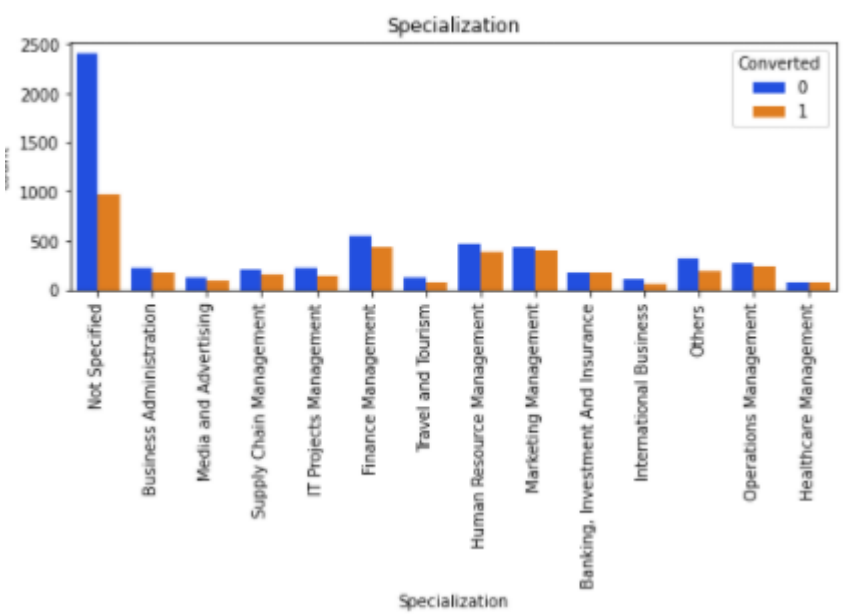
- Google is found to be one of the most important sources for lead conversion.
- Direct traffic also proves to be important to secure leads.



LAST ACTIVITY

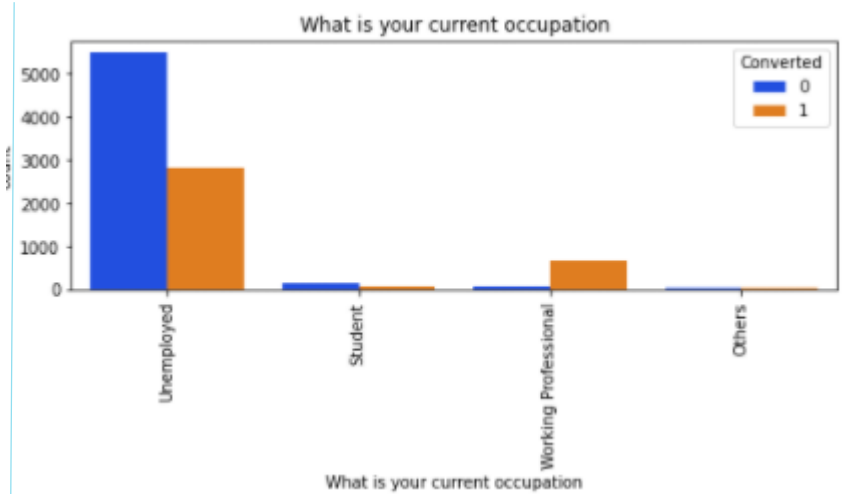
- It is clearly visible that the response activity for SMS is found to be high.
- People also reply to emails, so targeting people via emails is equally important.

EDA – CATEGORICAL DATA



SPECIALIZATION

- We cannot infer much from this specialization feature as people who do not specify their area of preference, can also be converted to a lead.
- The ratio of non converted leads is found to be greater that converted leads when customers do not fill the specialization.



CURRENT OCCUPATION

- We need to target unemployed and working professional as the conversion rate is found to be higher in these cases.

FREE COPY OF MASTERING THE INTERVIEW

- People usually do not subscribe for the free copy for mastering the interview.
- Since the interest is less, it also indicates that, investing on this is not of great use.

DATA PREPROCESSING BEFORE MODEL BUILDING

There are a few basic steps that needs to be followed for preparing the data before model building.
Model building:

- *Number of features after scaling and dummy variable creation : 35*
- *Target Variable : Converted*
- *Libraries used: StandardScaler()*
- *Columns that are not considered : Lead Number and Prospect ID (these variables do not help in model building)*

The steps are as follows:

1

Outlier Treatment:

- Total Visits and Page Views Per Visit had some outliers.
- We perform capping using Soft Capping (Checking for 99th percentile) and complete the outlier treatment process before we continue to the next step.

2

Binary Mapping:

- "A free copy of mastering the interview" contains values in terms of Yes/No , we convert these to 1/0 so it converts into numerical values and helps in model building.

3

Dummy Variable Creation:

- We need to create dummy variables for all the categorical columns as they enable us to use a regression equation on multiple groups.

4

Test Train Split:

- Division of data into test data and train data to check the stability of the model.
- We have randomly sampled 70% of the data as the test data and 30% of the data as test data.
- Random State = 100

5

Scaling:

- Division of Train Data into X and Y where X has all the features and Y has the target variable – Converted.
- We perform scaling to normalize the data within a particular range
- Technique : Standard Scaler

MODEL BUILDING STRATEGY

Model – VIII is our final model :

- ***All p-values < 5% - Hence they are highly significant***
- ***All VIF values are < 5% - Hence the dependency of variable with another is tolerable.***
- ***Final model has 14 features in total.***

Model VIII : Removing variables having high VIF

- *After model –VII , all p-values < 5%, hence we need to check VIF*
- *VIF for Current Occupation_Unemployed = 12.20 which is > 5%*
- *Hence we drop this variable from our analysis.*

Model V, VI and VII: Removing variables having p-value > 10%

- *Since we have a cut off for significance value > 10 % does not improve our model*
- *Hence, we remove these variables which are:*
- *Current Occupation Student , Specialization International Business and LastActivityEmail*

Model – III and IV – Removing variable with p-values > 50%

- *2 columns having p –values > 50% : Lead Source_Others and Lead Source Origin*
- *Since p –value > 50% , it does not seem to significant at all.*

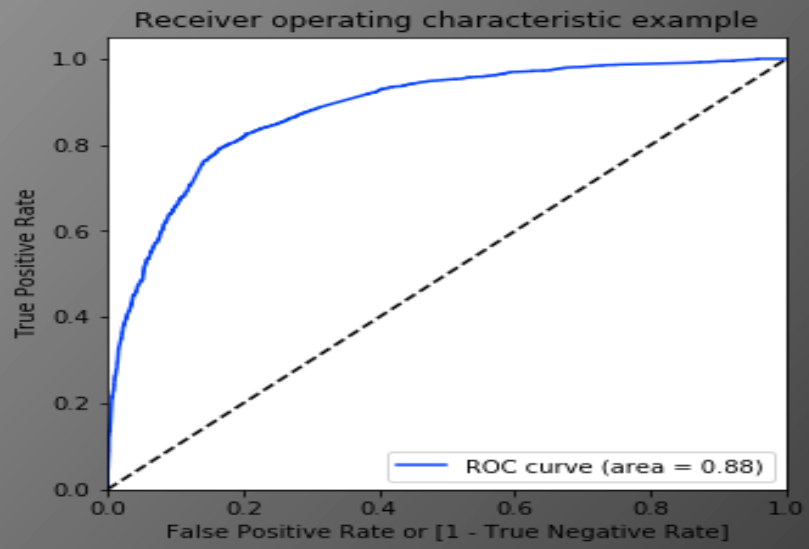
Model – I and II: We build a basic model using 35 features. Since it is not efficient we perform RFE to obtain a model with Top – 20 features. There are so many variables with high p-values and VIF value, we need to remove them.

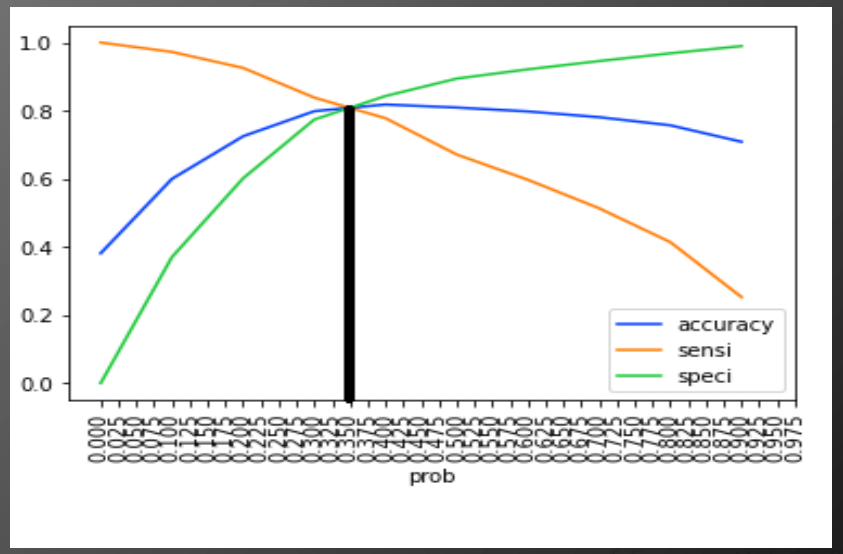
ROC CURVE AND PROBABILITY CUT OFF

AUC => 0.88 whereas Optimal Cutoff Probability => 0.35

ROC Curve and Optimal Cutoff Point

- ROC Curve represents how much the model is able to distinguish between the classes.
- AUC – Area under the curve represents that it is distinguishing the 1's and 0's correctly.
- On plotting the ROC curve for our data we see that, AUC is around 0.88 which means at around 88% of the times, the model is able to distinguish the 1's as 1's and 0's as 0's.
- AUC of 0.88 is found to be very stable model.
- When we plot the sensitivity, accuracy and specificity of the model together, the optimal cut off point is found to be at 0.35. This means that at 35% probability, the sensitivity and specificity are found to be balanced.
- With probability = 0.35 , we predict y-values with X-Train, in such a way that, any conversion prob > 35% is said to be converted to a lead.





Model Performance Parameters – Train Set v/s Test Set

Train Set

- **Accuracy : 81.19 %**
- **Sensitivity : 80.45 %**
- **Specificity : 81.7 %**

Test Set

- **Accuracy : 80.08 %**
- **Sensitivity: 80.0 %**
- **Specificity: 80.3 %**

- The sensitivity value after model building process is found to be greater than 80% as required.
- When the model is evaluated for Test Set, the model evaluation parameters remains to be the same. Hence the model is highly stable

LEAD SCORE AND CONVERSION RATE

Steps taken to assign a lead score variable for all customers.

1 Train the data with the model.

- Run the model on the entire Leads dataset.
- Do not divide into Test and Train and run the obtained LR model on the entire data frame

2 Predict the Conversion Probability using Cutoff

- Predict the Conversion probability for all the customers using the cutoff value = 0.35.
- Create a new data frame and store the Conversion_Probability and actual converted values in this.

3 Adding Lead Score for all variables.

- Create a new column called Lead Score.
- Convert the probability score into Lead Score by multiplying by 100 and store it in this column.

4 Calculate the conversion rate.

- Once we obtain the complete model result on the data, we filter only the leads as predicted by the model.
- Calculate the Conversion Rate using this filtered result.

Note on Conversion Rate

Conversion Rate is the number of customers who are converted to leads and interested in the course.

Before model building the Conversion Rate was found to be 38.53%

After model building, the conversion rate is increased to 72.87%

Hence we can conclude that our final model has served to the business purpose.

HOT LEADS

1. *Hot leads are people who have a high probability to be converted as a Lead and thus needs to be identified. They have a higher conversion rate.*

2. *The leads whose lead score is greater than 35% are considered as potential leads. The conversion rate is around 73%. When we increase this threshold from 35% to 95% we get Hot Leads.*

3. *Conversion Rate for hot leads is increases from 73% to 96%. This means they have a 96% probability of getting converted to a lead.*

4. *Focusing on Hot Leads will increase the chances of obtaining more value to the business as the number of people we contact are less but the conversion rate is high.*

OTHER BUSINESS RECOMMENDATIONS

1. *It's good to collect data often and run the model and get updated with the potential leads. There is a belief that the best time to call your potential leads is within few hours after the lead shows interest in the courses.*

2. *Along with phone calls, it's good to mail the leads also to keep them reminding as email is as powerful as cold calling.*

3. *Reducing the number of call attempts to 2-4 and increasing the frequency of usage of other media like advertisements in Google, or via emails to keep in touch with the lead will save a lot of time.*

4. *Focusing on Hot Leads will increase the chances of obtaining more value to the business as the number of people we contact are less but the conversion rate is high.*

THANK YOU!