International Institute of Information Technology Bangalore

upGrad

LEAD SCORING CASE STUDY

BY:

SUNITA MANE & PRAKASH R

PROBLEM STATEMENT

- An education company named X Education sells online courses to industry professionals.
- The company markets its courses on several websites and search engines like Google when people might browse or fill a form or watch some videos.
- Leads will be considered when these people either fill a form or watch videos or from the past referrals.
- The Lead conversion is 30% as of now.

Business Objectives ...

Business Objectives

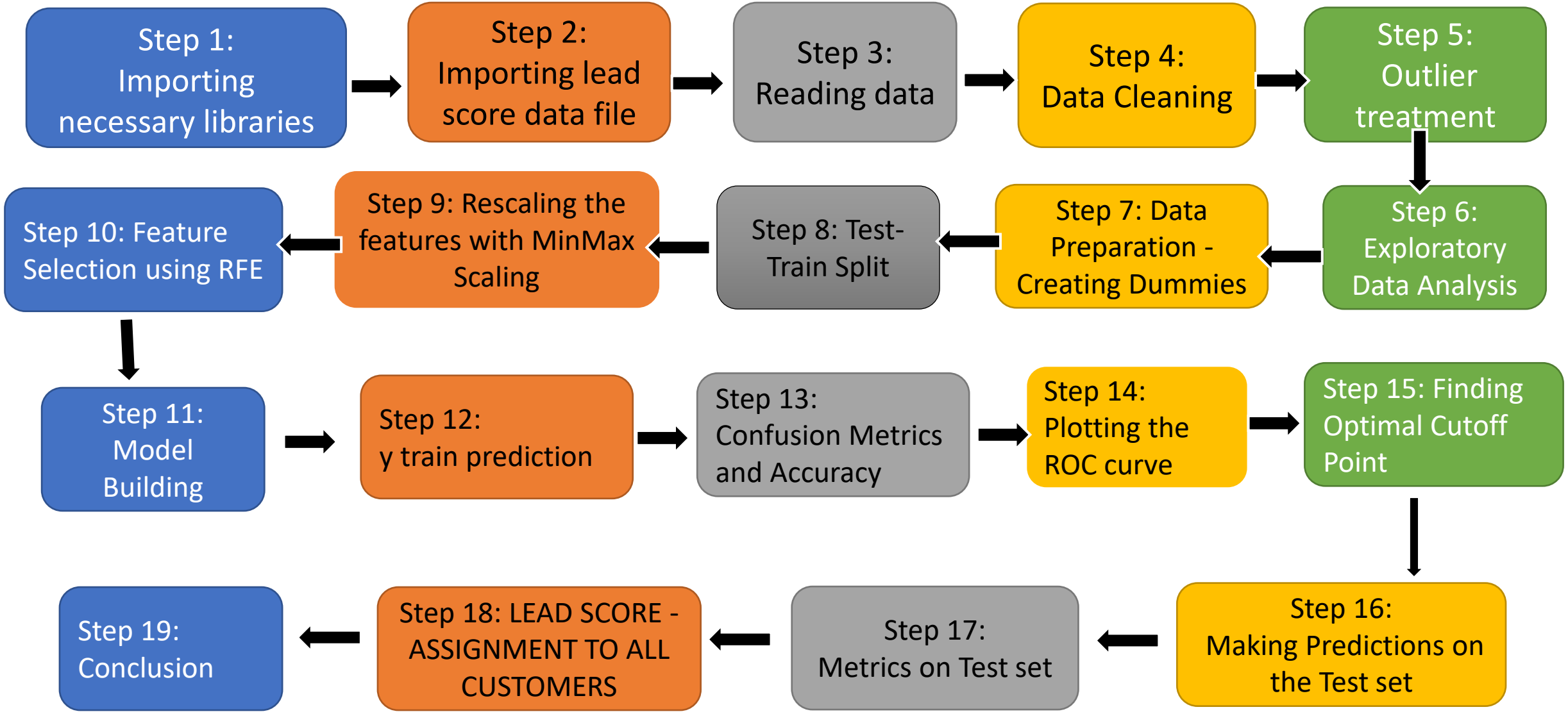
- X Education needs help in selecting the most promising leads

- The company needs a model wherein you a lead score is assigned to each of the leads such that the customers with higher lead score have a higher conversion chance and the customers with lower lead score have a lower conversion chance.
- The CEO has a target lead conversion rate to be around 80%.

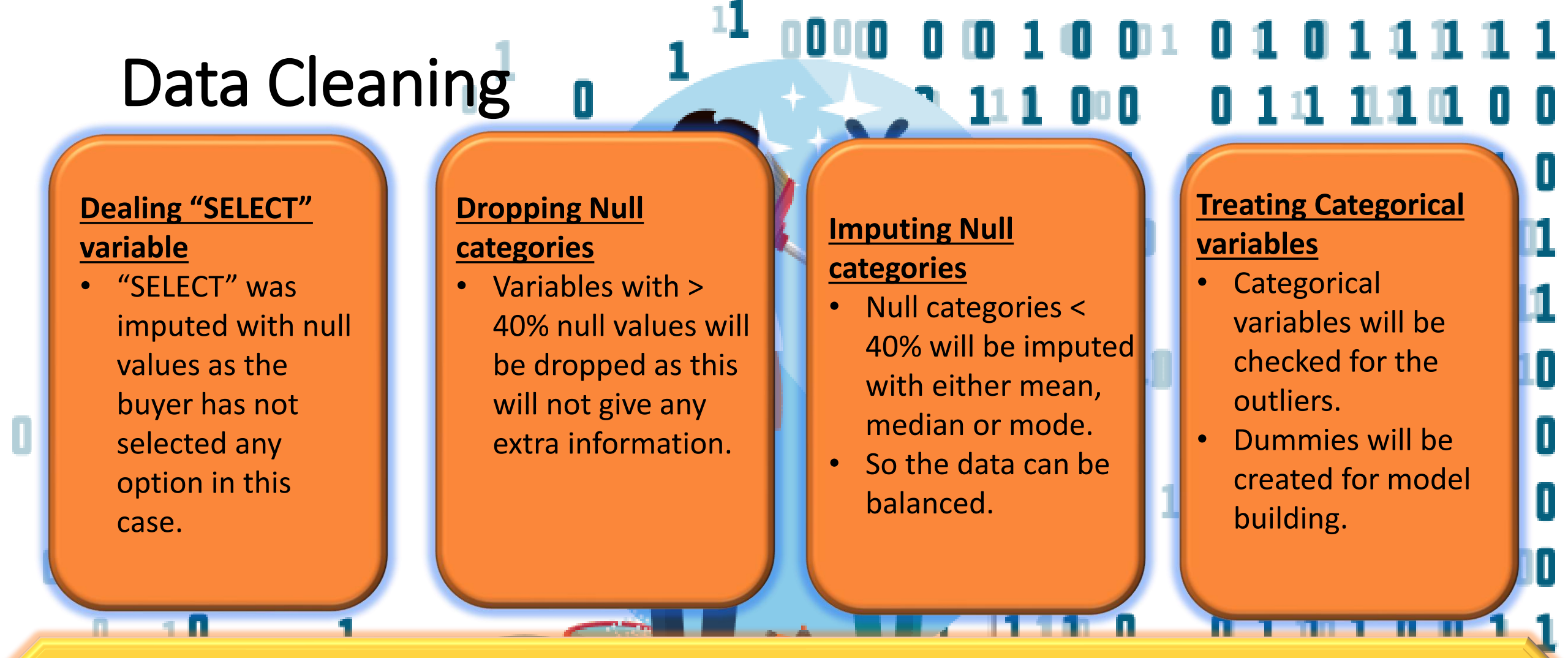
Understanding Dataset

- The file named Lead.csv was given for analysis.
- There were 9240 rows and 37 columns in data set.
- No duplicates were seen in the data set
- There were 4 variables of float type, 3 variables of int type and 30 variables of object type.
- There was huge difference between 75% and maximum values in "TotalVisits", "Total Time Spent on Website" and "Page Views Per Visit".
- There were 5 columns with only single unique values in the data set.
- There were 17 columns with null values in the data set.
- We have considered 35% - 38% as cutoff.
- The target variable was “Converted” where 0 represents not converted and 1 represents converted.

Approach for Business Problem Solving



Data Cleaning



Dealing "SELECT" variable

- "SELECT" was imputed with null values as the buyer has not selected any option in this case.

Dropping Null categories

- Variables with > 40% null values will be dropped as this will not give any extra information.

Imputing Null categories

- Null categories < 40% will be imputed with either mean, median or mode.
- So the data can be balanced.

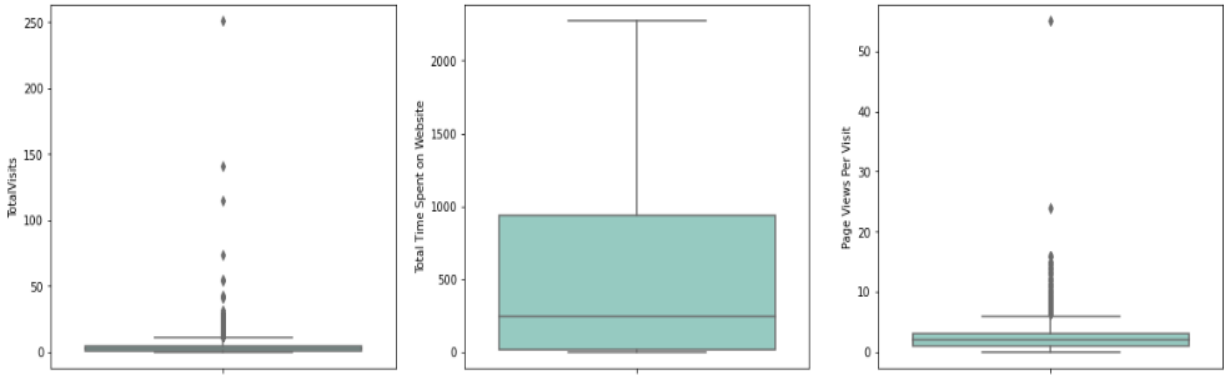
Treating Categorical variables

- Categorical variables will be checked for the outliers.
- Dummies will be created for model building.

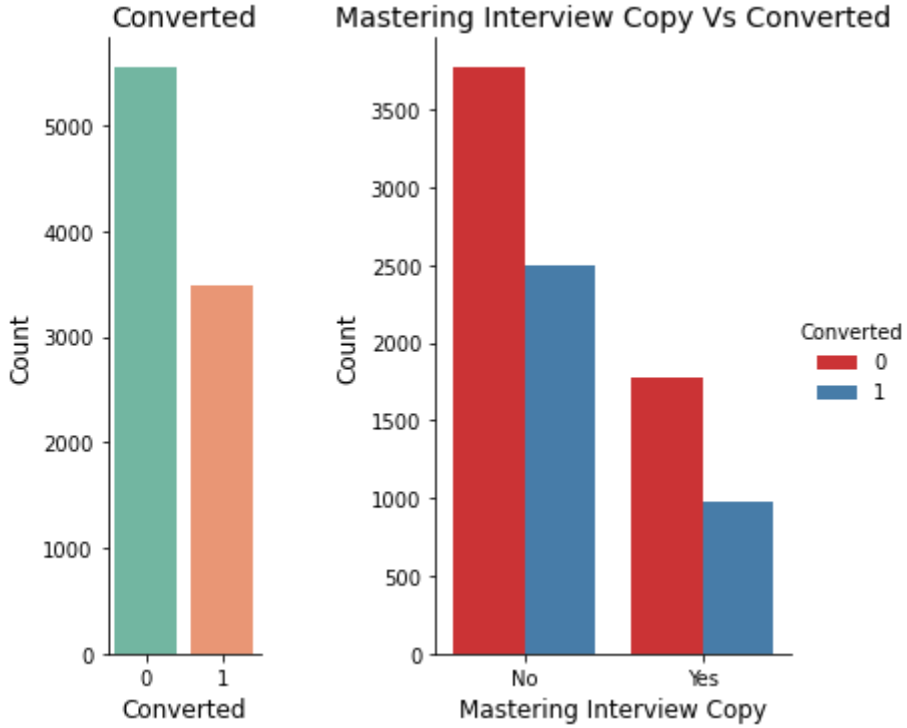
Final list

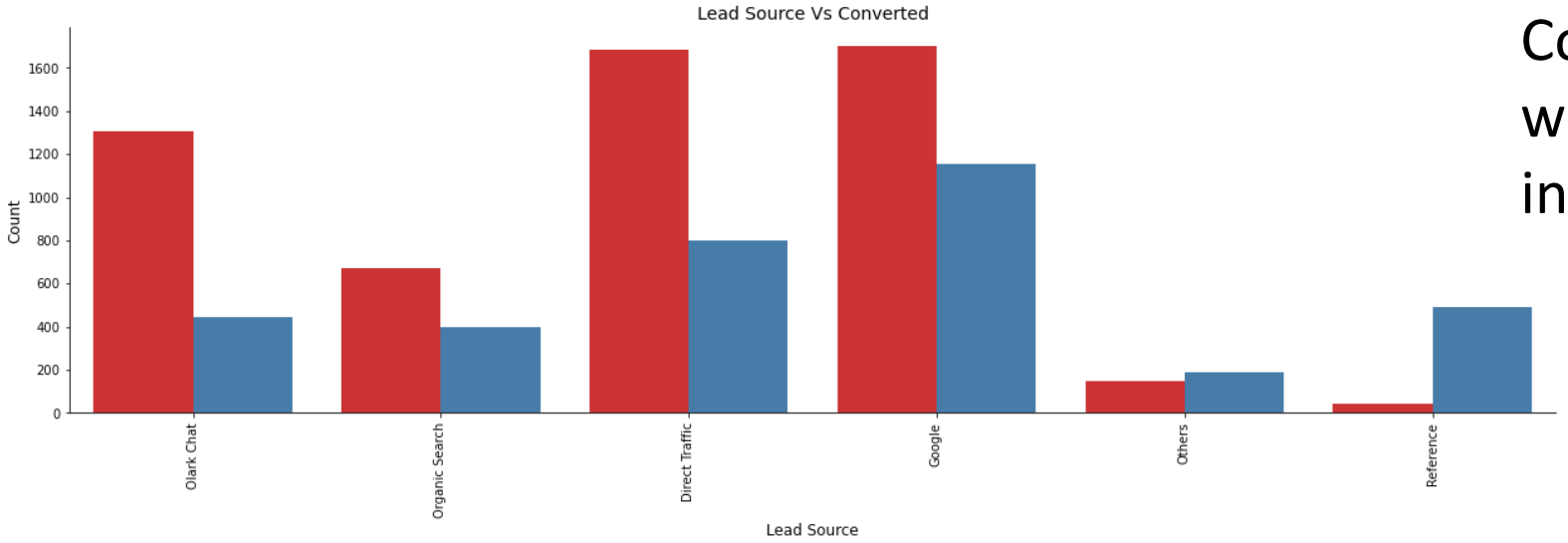
The final list of variables will be selected using RFE and VIF.

Exploratory Data Analysis



Outliers are present in Total Visits and Page Views Per Visit Columns

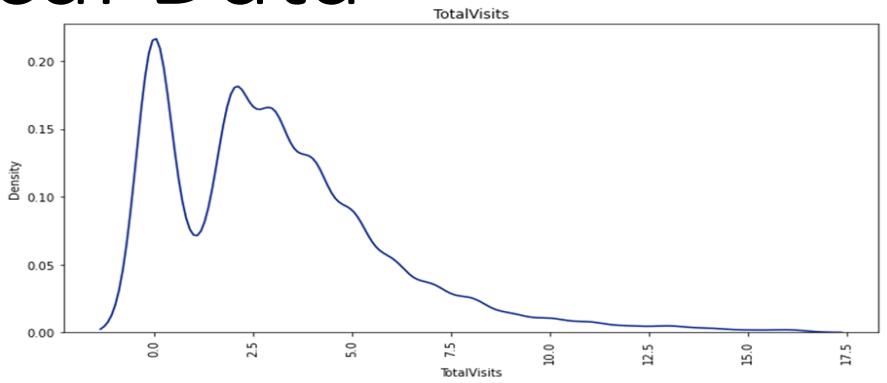


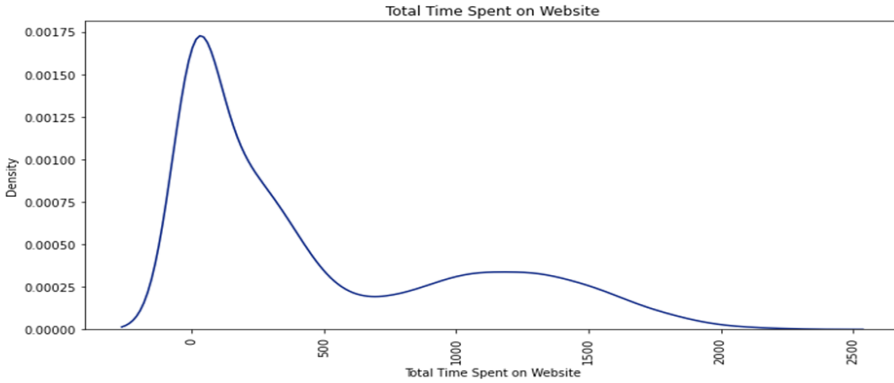


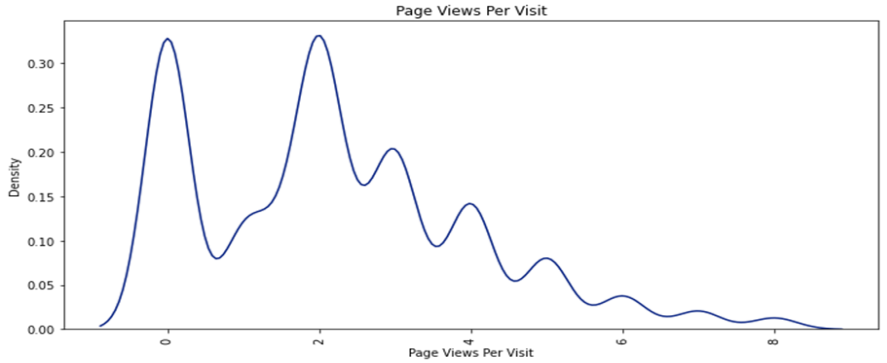
Conversion rates are 39% and those who did not opt for mastering interview have converted more

Majority of conversion in Lead Origin are from google followed by Direct traffic

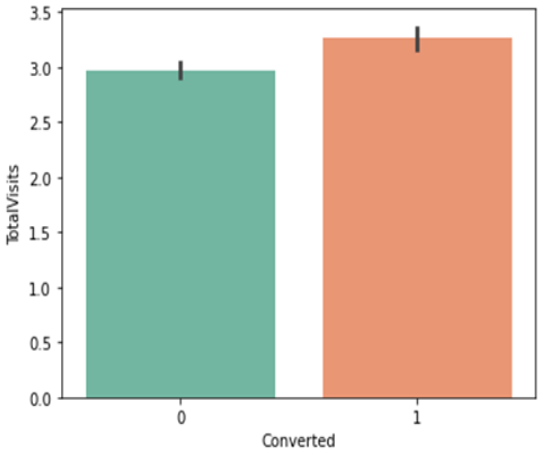
EDA Numerical Data

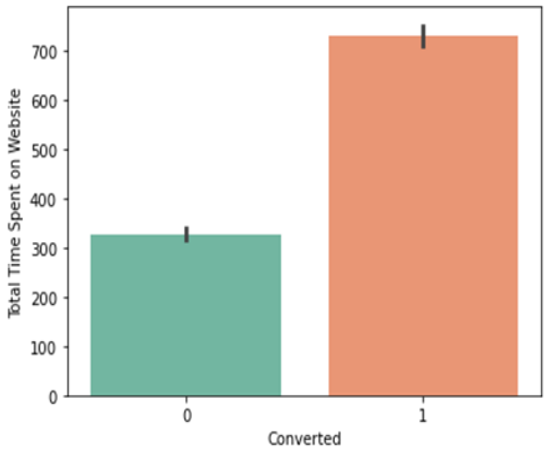


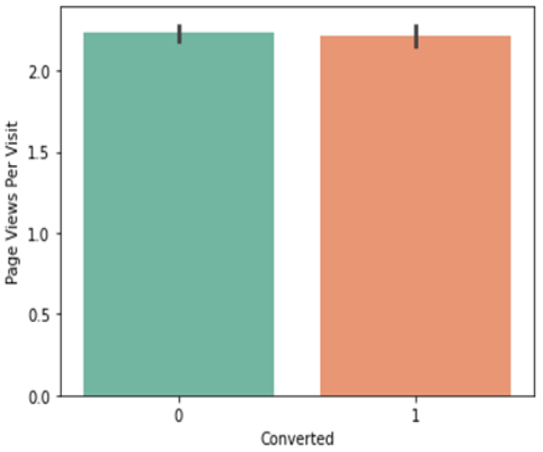




Majority of conversion are seen in all three numerical variables.







The max probability for TotalVisits is found to be around 15-20. It increases initially but decreases further.
The max probability for PageViewsPerVisit is found to be around to be 3-5
The probability of time spent is found to be high for time between 0-300 seconds and decreases further.

The percentage of Converted people is found to be greater for Landing Page Submission.

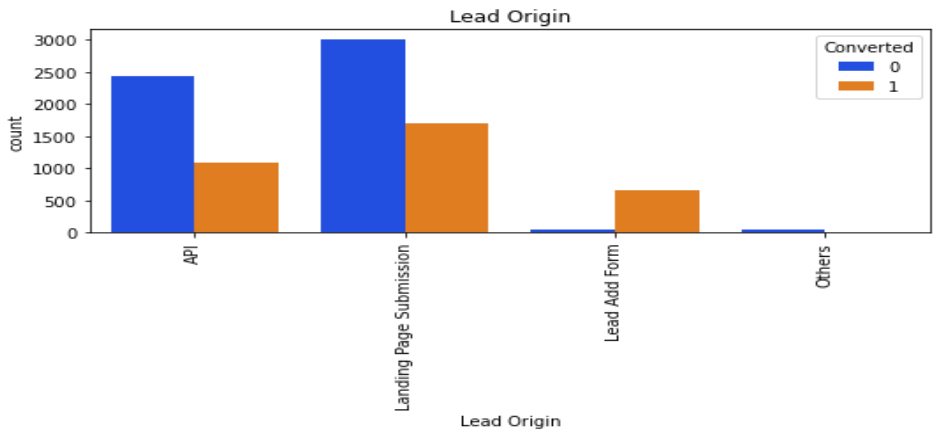
Google is found to be the important source for Lead Conversion

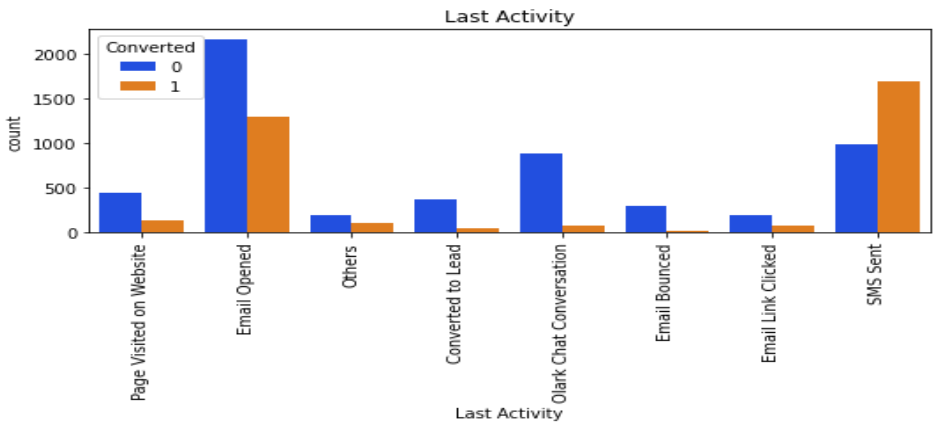
Target people via Emails and SMS as it is found that the probability of response in case Converted leads is found to be higher.

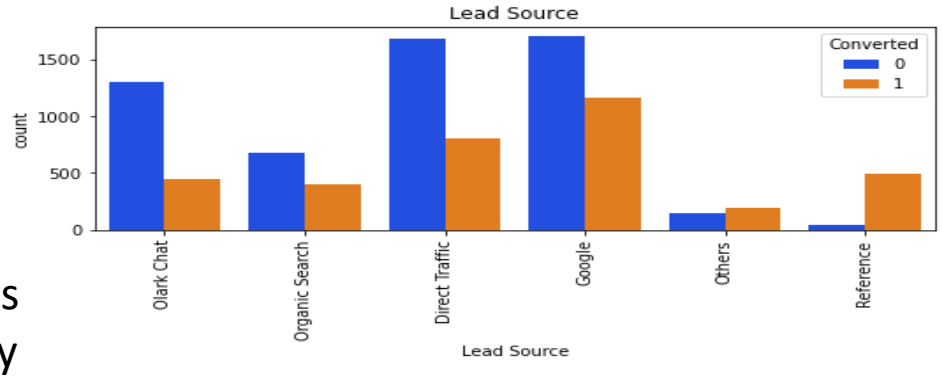
The ratio of non converted leads is higher than converted ones if they didn't choose specialization.

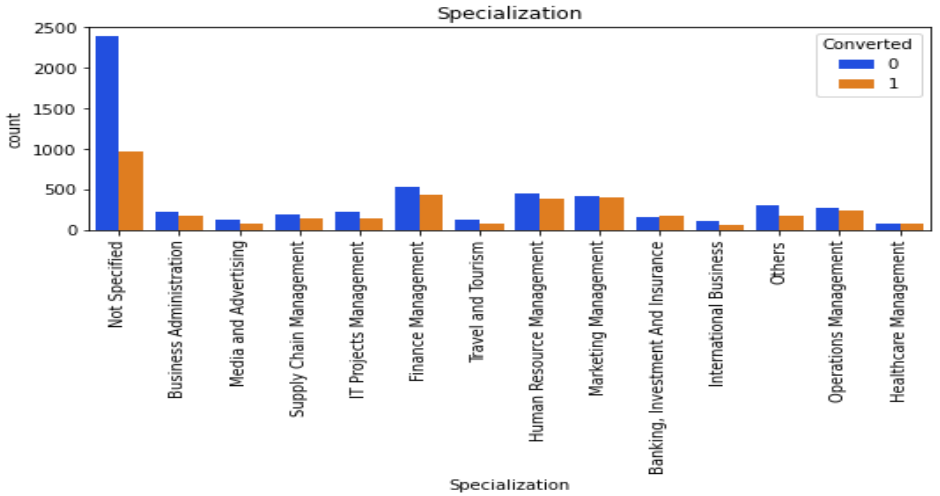
The ratio of conversion rate is higher than not converted people for working professionals.

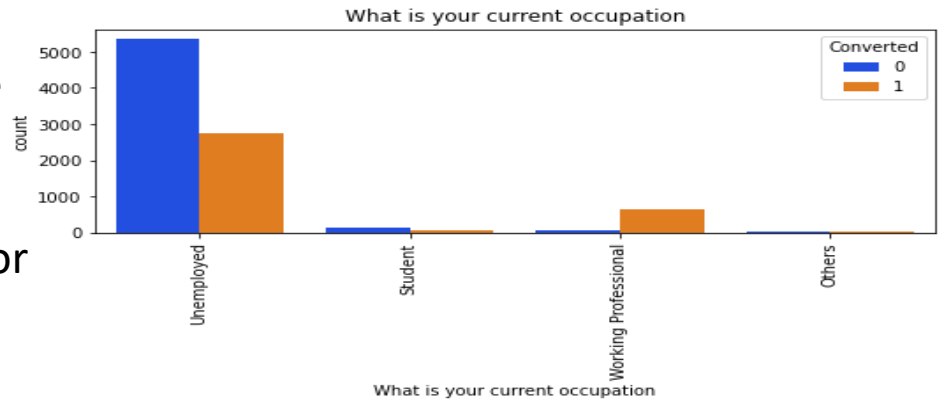
People usually do not subscribe for a free copy of mastering the interview.

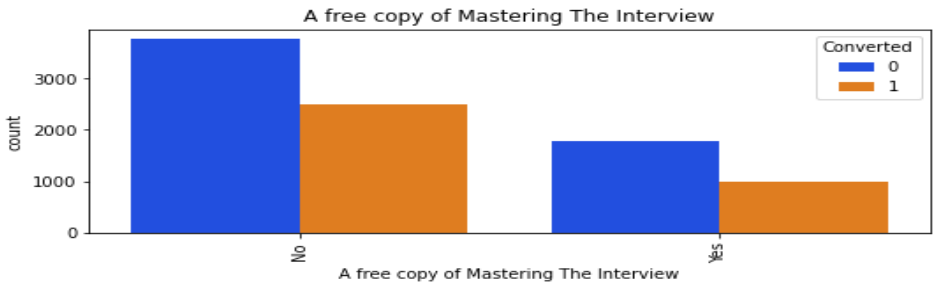












Model Building

Model 1 and 2 : Basic Model

We built a basic model with 35 variables. We used RFE to obtain the variables with top 20 variables.

Model 3: Removing variable with highest p value more than 5%

What is your current occupation_unemployed was removed to build model 3.

Model 4: Removing the variable with high VIF value

Page views per visit was removed as VIF was 6.66. So we were left with 15 columns for model 4

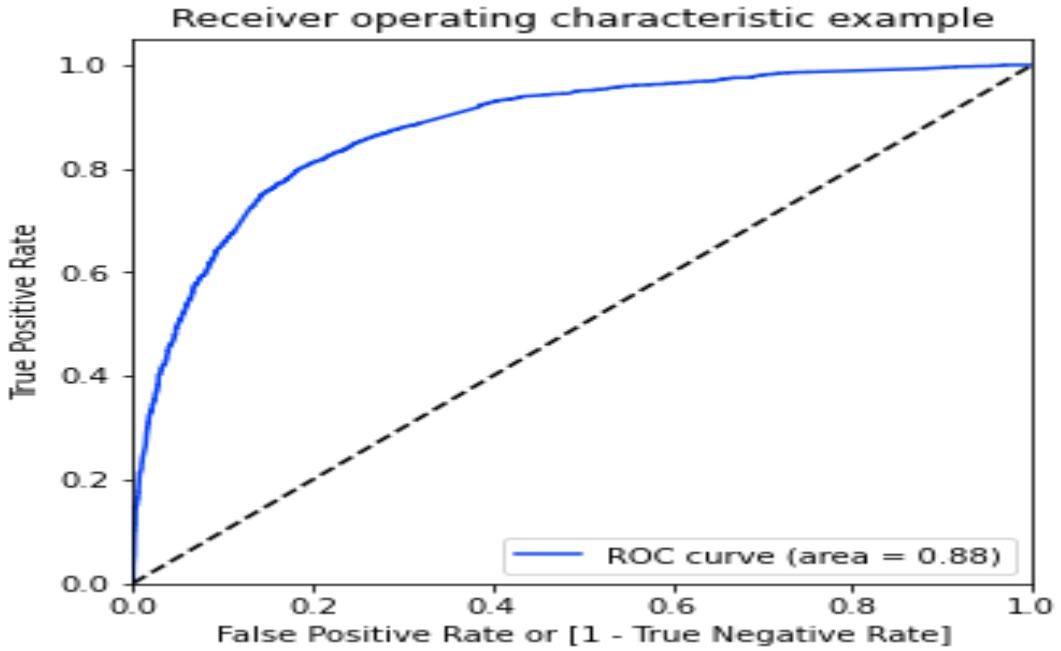
Model 5: Removing the variable with high p value

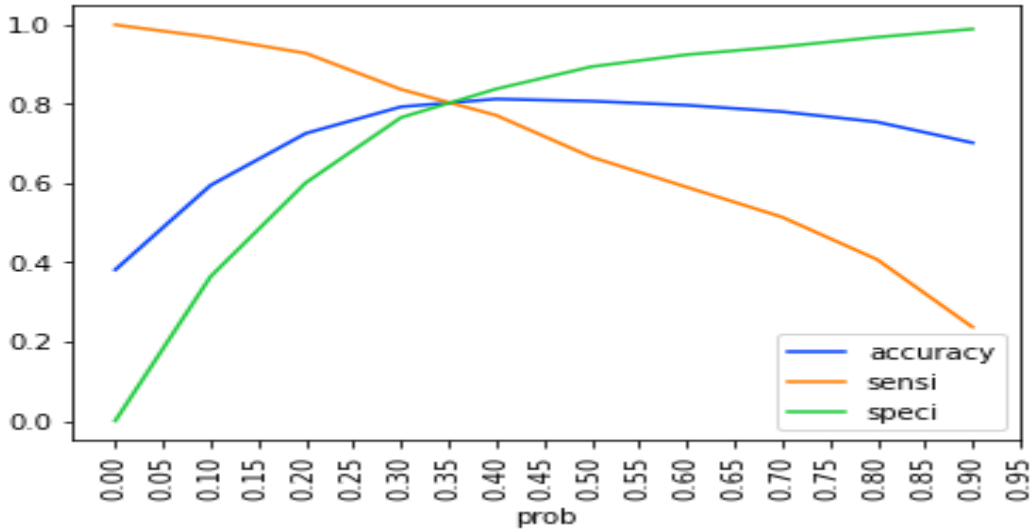
Lead Origin Others was removed has it had p value of 0.184.

Final Model

Model 5 was kept as the final model as the p value and VIF were less than 5% and 5.

ROC Curve And Optical Cut Off Probability





- 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.
- The optimal cut off point is found to be at 0.35 which 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 Test

Interpretation of Accuracy - Specificity - Sensitivity of train

Accuracy = 80.68%

Sensitivity = 80.34%

Specificity = 80.89%

- Interpretation of Accuracy - Specificity - Sensitivity on Test Set
- Accuracy = 81.80%
- Sensitivity = 81.52%
- Specificity = 81.98%

CONCLUSION

Equation:

ln(odds)= -2.3368 * const + 1.0708 * TotalVisits + 4.4986* Time Spent on Website -1.0903 * Lead Origin_Landing Page Submission + 3.6323 * Lead Origin_Lead Add Form - 0.8804 * Last Activity_Email Bounced + 0.4968 * Lead Source_Reference -1.1921 * Last Activity_Email Bounced + 0.8166 * Last Activity_Email Link Clicked + 0.8172 * Last Activity_Email Opened - 0.7460 * Last Activity_Olark Chat Conversation + 0.7252 * Last Activity_Others + 1.9733 * Last Activity_SMS Sent + 1.3358 * Lead Source_Olark Chat + 0.4272 * Lead Source_Others + 2.5195 * What is your current occupation_Working Professional - 1.0958 * Specialization_Not Specified

we can conclude following points :

- The customer/leads who fills the form are the potential leads.
- We must majorly focus on working professionals.
- We must majorly focus on leads whose last activity is SMS sent or Email opened.
- It’s always good to focus on customers, who have spent significant time on our website.
- It’s better to focus least on customers to whom the sent mail is bounced back.
- If the lead source is referral, he/she may not be the potential lead.
- If the lead didn’t fill specialization, he/she may not know what to study and are not right people to target. So, it’s better to focus less on such cases.

RECOMMENDATIONS

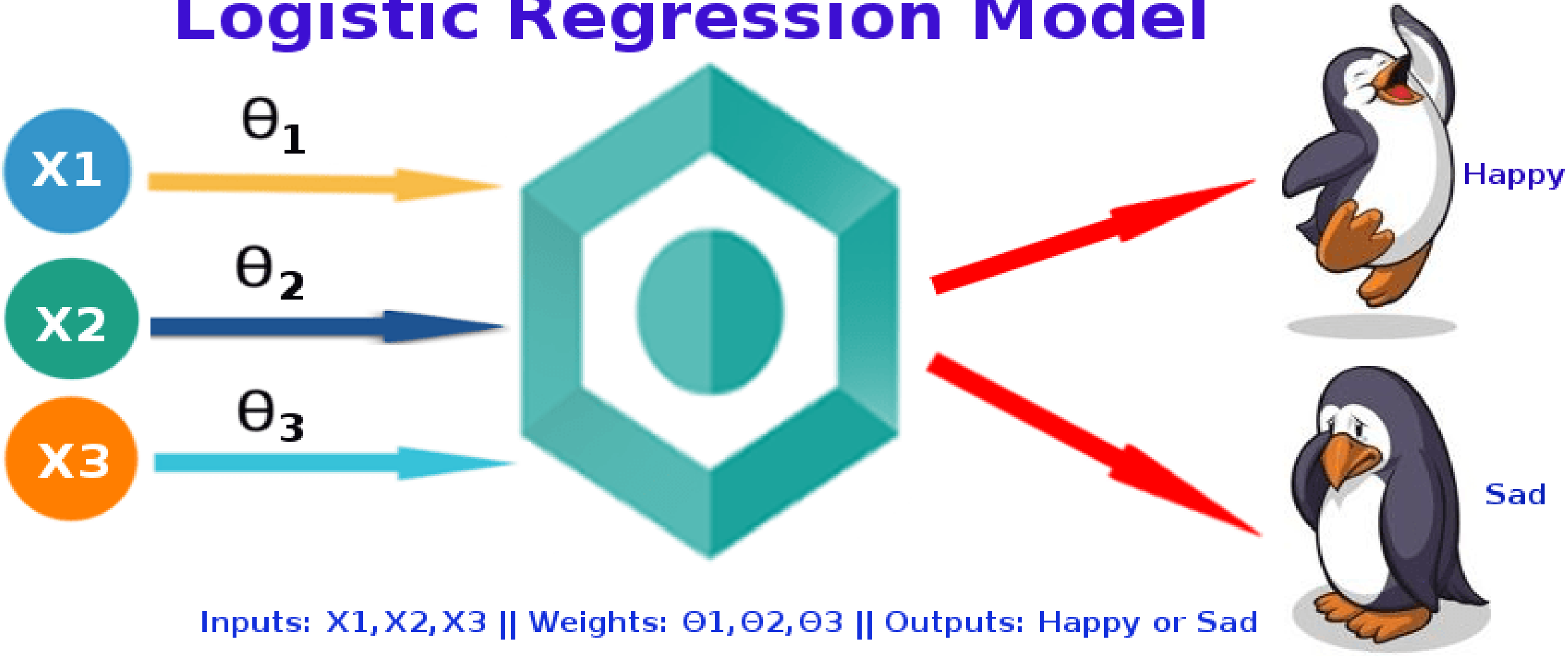
- It’s good to collect data often and run the model and get updated with the potential leads. There is a belief that the best time to call your potential leads is within few hours after the lead shows interest in the courses.**
- Along with phone calls, it’s good to mail the leads also to keep them reminding as email is as powerful as cold calling.

- 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.

- 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

Logistic Regression Model



@dataaspirant.com