LEAD SCORING CASE STUDY PRESENTATION

Problem overview, solving strategy, statement

1. Company Overview:

- X Education: Sells online courses to industry professionals.
- Marketing Channels: Websites, Google, search engines.

2. Current Lead Acquisition Process:

- Visitors land on the website.
- Actions: Browse courses, fill forms, watch videos.
- Lead Classification: Email/phone submission or past referrals.
- Sales Team: Calls, emails, follow-ups.

3. Lead Conversion Statistics:

- Typical Lead Conversion Rate: ~30%
- Example: 100 leads/day -> ~30 converted.

Problem overview, solving strategy, statement

4. Problem Statement:

- High lead acquisition, poor conversion rate.
- Objective: Identify "Hot Leads" to improve efficiency.
- Goal: Focus on potential leads to increase conversion rates.

5. Funnel Representation: X Education Lead Conversion Improvement Plan

- **Top**: High lead generation.
- **Middle**: Nurture potential leads (education, communication).
- **Bottom**: Paying customers.

6. Solution Requirements:

- Develop a model to assign lead scores.
- Higher lead scores indicate higher conversion chances.
- Target Conversion Rate: ~80%

Problem solving process

1. Convert the categorical variable yes/no into numeric 1/0.

2. Remove the missing values.

3. Convert select into nan.

4. Variation in the column categories so will replace the categorical variable into nan.

5. Dropping columns having more than 70% null values.

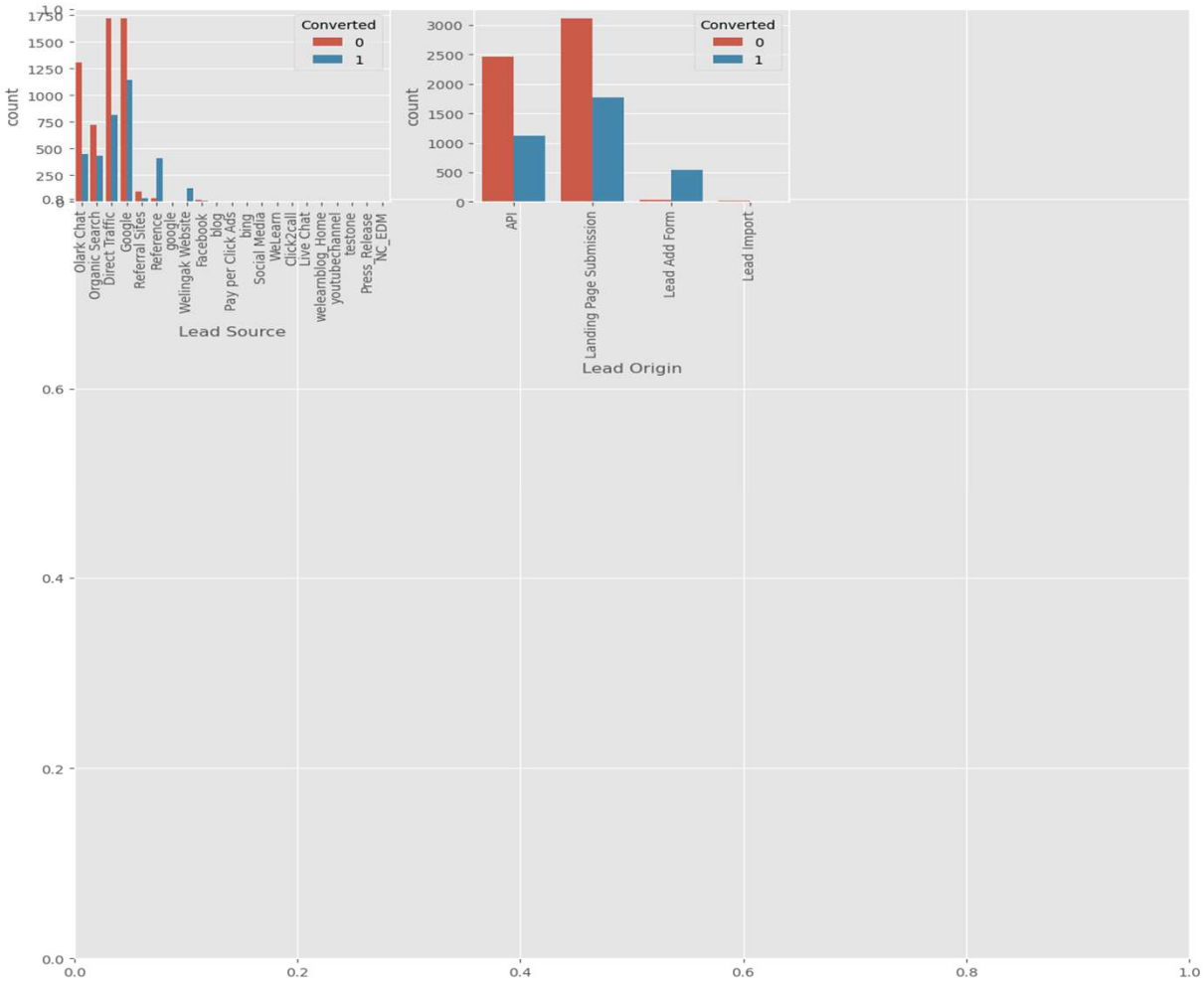
6. There are too many variations in the columns like Asymmetrique Activity Index, Asymmetrique Activity Score, Asymmetrique Profile Index, Asymmetrique Profile Score and it is not safer to impute any values in the columns and hence we will drop these columns with very high percentage of missing data.

7. Checked the data imbalance.

8. Performed the Exploratory data analysis.

OBSERVATION:

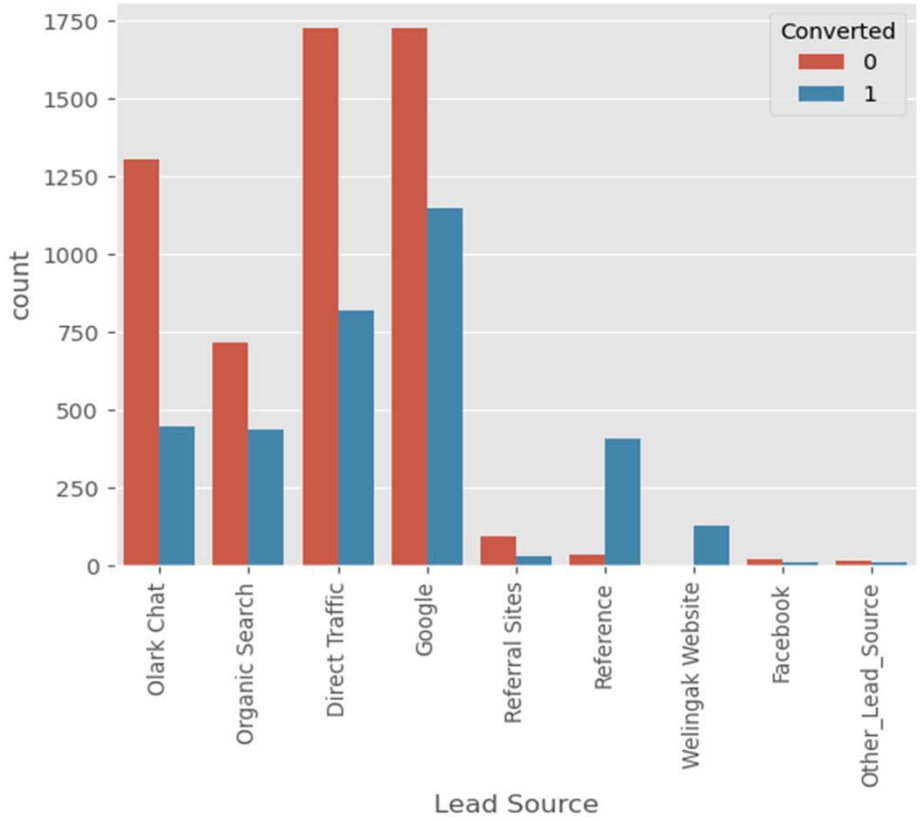
- API and Landing Page Submission has less conversion rate(~30%) but counts of the leads from them are considerable.
- The count of leads from the Lead Add Form is pretty low but the conversion rate is very high. Lead Import has very less count as well as conversion rate and hence can be ignored. To improve the overall lead conversion rate, we need to focus on increasing the conversion rate of 'API' and 'Landing Page Submission' and also increasing the number of leads from 'Lead Add Form'



OBSERVATION:

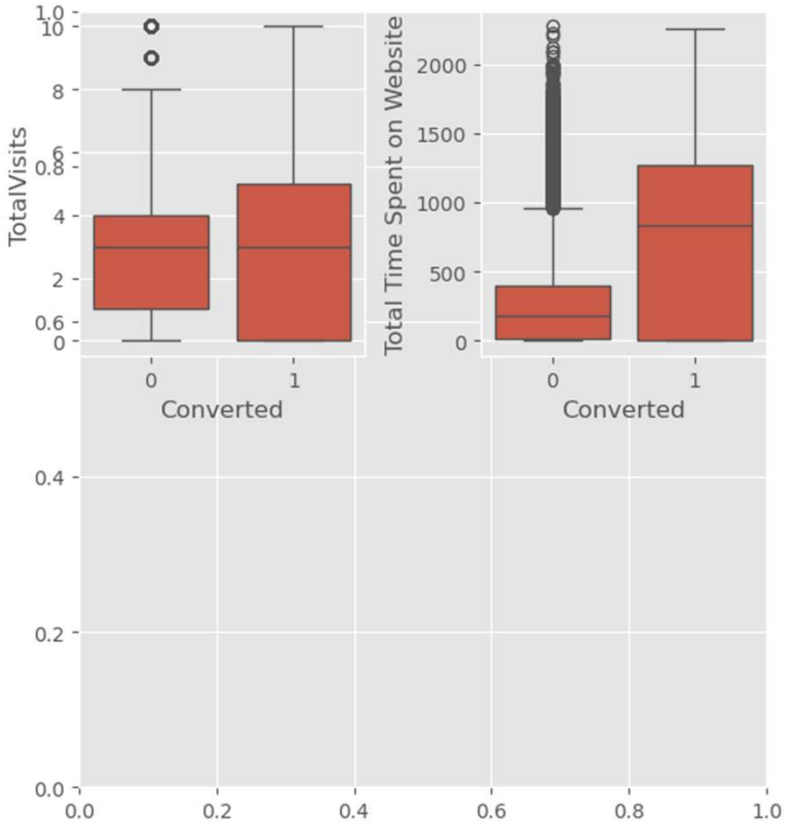
- The count of leads from the Google and Direct Traffic is maximum The conversion rate of the leads from Reference and Welingak Website is maximum.

- To improve the overall lead conversion rate, we need to focus on increasing the conversion rate of 'Google', 'Olark Chat', 'Organic Search', 'Direct Traffic' and also increasing the number of leads from 'Reference' and 'Welingak Website'.



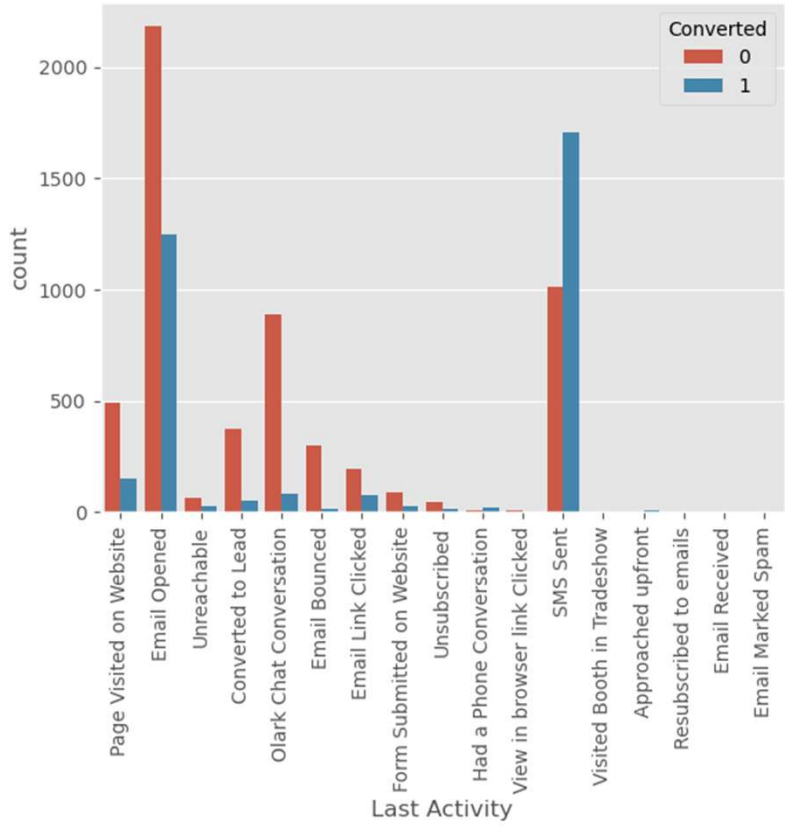
OBSERVATION:

- 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.



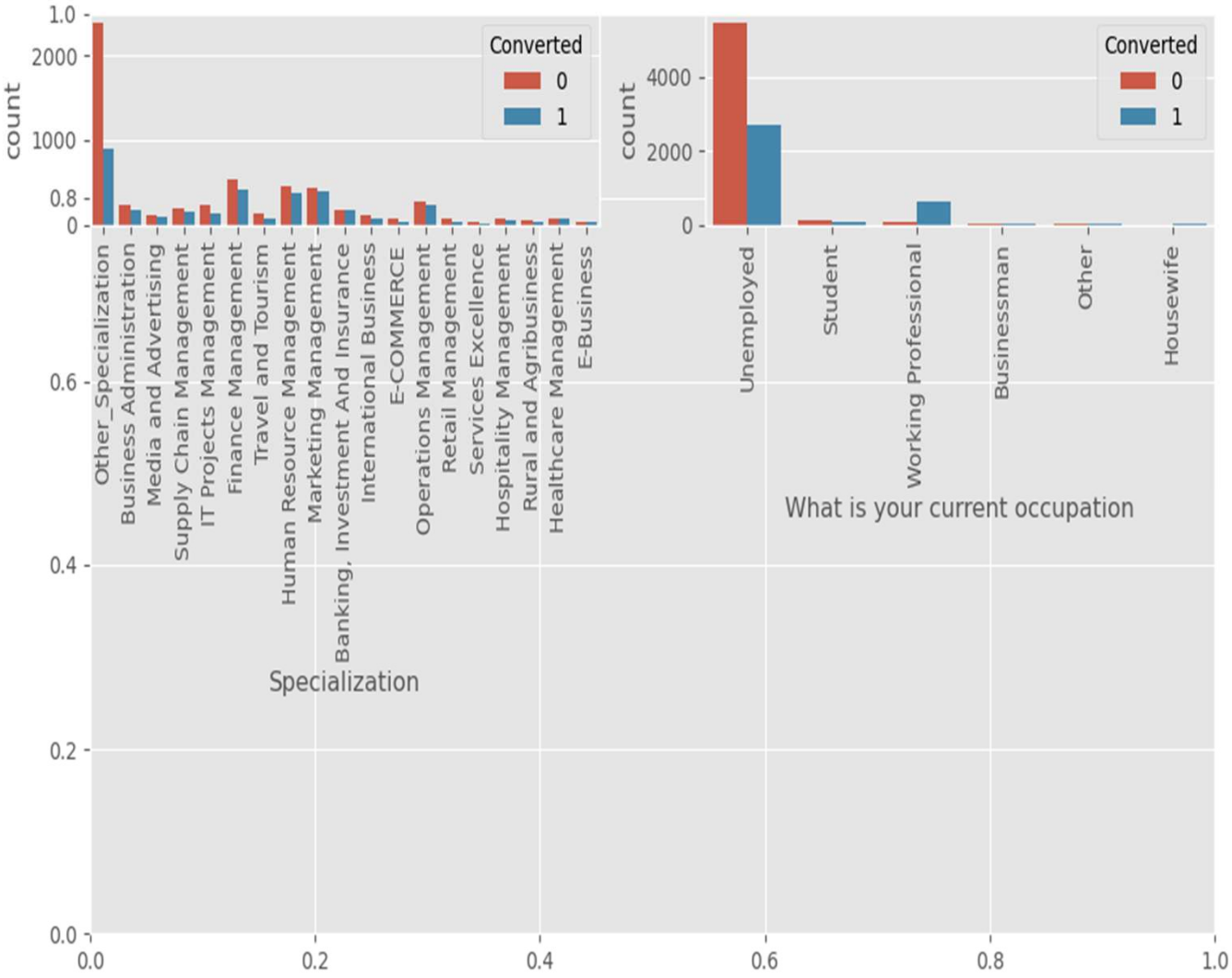
OBSERVATION:

- The count of last activity as "Email Opened" is max The conversion rate of SMS sent as last activity is maximum
- We should focus on increasing the conversion rate of those having last activity as Email Opened by making a call to those leads and also try to increase the count of the ones having last activity as SMS sent.

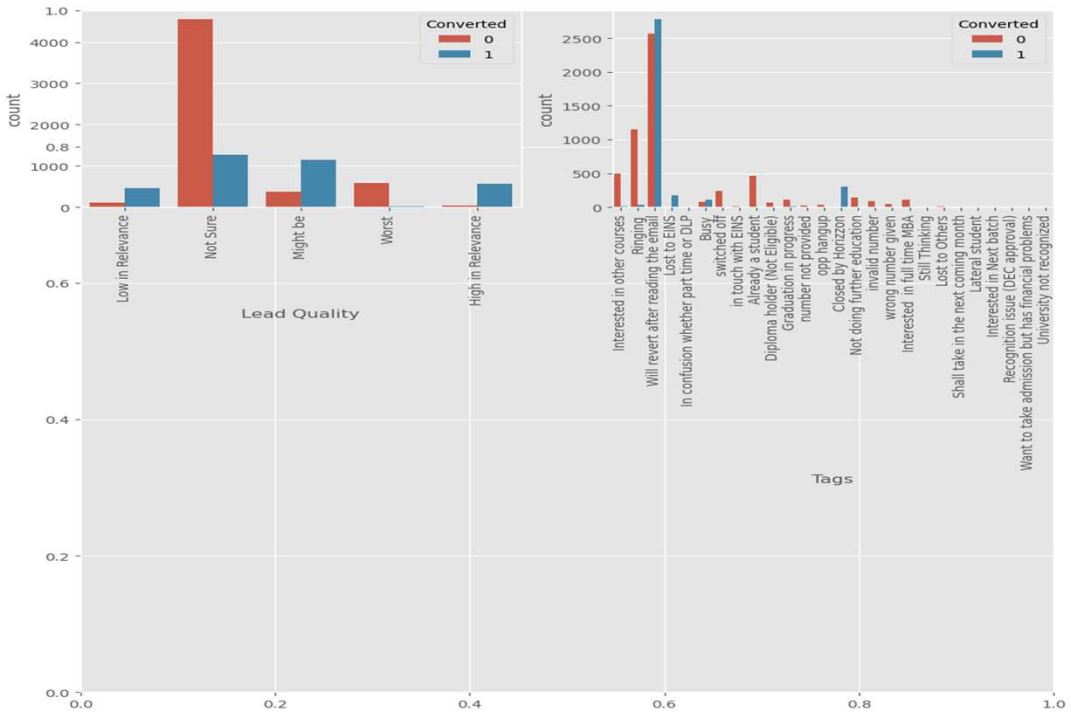


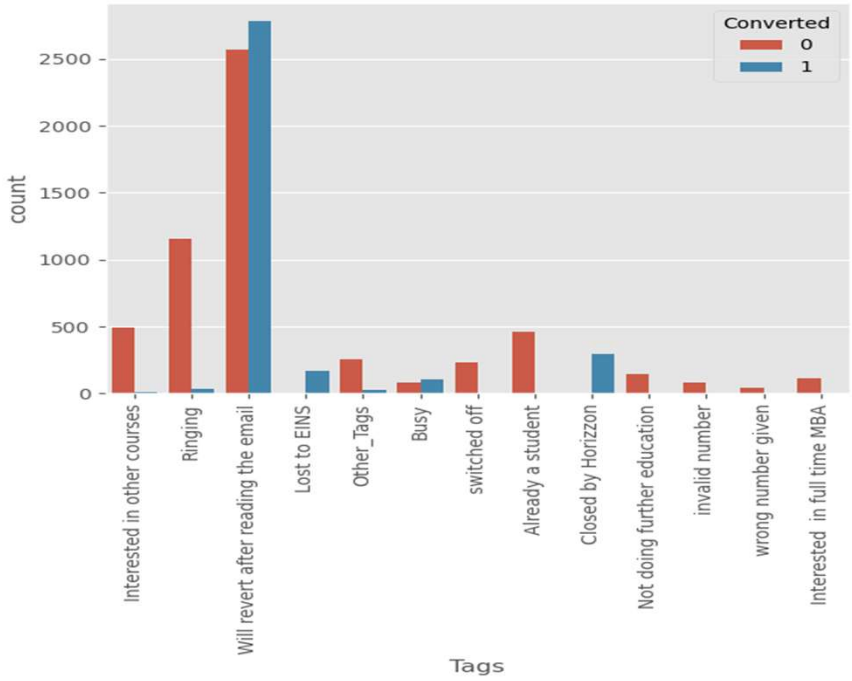
OBSERVATION:

- Looking at above plot, no particular inference can be made for Specialization Looking at above plot, w can say that working professionals hav high conversion rate Number of Unemployed leads are more than any other category.
- To increase overall conversion rate, we need to increase the number of Workin Professional leads by reaching out to them through different social sites such as LinkedIn etc. and also on increasing the conversion rate of Unemployed leads
- Country, What matters most to you in choosing a course, City columns have most values corresponding to one valu such as India for Country, Mumbai for city and hence there is no particular insights for these columns



Will revert after reading the email and Closed by Horizzon have high conversion rate.





Summary

- To improve the overall lead conversion rate, we need to focus on increasing the conversion rate of 'API' and 'Landing Page Submission' Lead Origins and also increasing the number of leads from 'Lead Add Form'.
- To improve the overall lead conversion rate, we need to focus on increasing the conversion rate of 'Google', 'Olark Chat', 'Organic Search', 'Direct Traffic' and also increasing the number of leads from 'Reference' and 'Welingak Website'.
- Websites can be made more appealing so as to increase the time of the Users on websites
- We should focus on increasing the conversion rate of those having last activity as Email Opened by making a call to those leads and also try to increase the count of the ones having last activity as SMS sent.
- To increase overall conversion rate, we need to increase the number of Working Professional leads by reaching out to them through different social sites such as LinkedIn etc. and also on increasing the conversion rate of Unemployed leads.
- We also observed that there are multiple columns which contains data of a single value only. As these columns do not contribute towards any inference, we can remove them from further analysis.

Performing Logistic Regression

- **Dummy Variable Creation.**
- **Test-Train Split**
- **Feature Scaling**
- **Looking at Correlation**
- **Model Building**
- **Feature Selection Using RFE**
- **Plotting the ROC Curve**
- **Finding optimal value of the cut off**
- **Final Model**

Logistic Regression Summary

- The logistic regression model predicts the probability of the target variable having a certain value, rather than predicting the value of the target variable directly. Then a cutoff of the probability is used to obtain the predicted value of the target variable.
- Here, the logistic regression model is used to predict the probabilty of conversion of a customer.
- Optimum cut off is chosen to be 0.27 i.e. any lead with greater than 0.27 probability of converting is predicted as Hot Lead (customer will convert) and any lead with 0.27 or less probability of converting is predicted as Cold Lead (customer will not convert)
- Our final Logistic Regression Model is built with 14 features.
- Features used in final model are ['Do Not Email', 'Lead Origin_Lead Add Form', 'Lead Source_Welingak Website', 'Last Activity_SMS Sent', 'Tags_Busy', 'Tags_Closed by Horizzon', 'Tags_Lost to EINS', 'Tags_Ringing', 'Tags_Will revert after reading the email', 'Tags_switched off', 'Lead Quality_Not Sure', 'Lead Quality_Worst', 'Last Notable Activity_Modified', 'Last Notable Activity_Olark Chat Conversation']
- The top three categorical/dummy variables in the final model are ‘Tags_Lost to EINS’, ‘Tags_Closed by Horizzon’, ‘Lead Quality_Worst’ with respect to the absolute value of their coefficient factors.

Logistic Regression Summary

- ‘Tags_Lost to EINS’, ‘Tags_Closed by Horizzon’ are obtained by encoding original categorical variable ‘Tags’. ‘Lead Quality_Worst’ is obtained by encoding the categorical variable ‘Lead Quality’.
- Tags_Lost to EINS (Coefficient factor = 7.085)
- Tags_Closed by Horizzon (Coefficient factor = 6.238)
- Lead Quality_Worst (Coefficient factor =-2.914)
- The final model has Sensitivity of 0.937, this means the model is able to predict 93% customers out of all the converted customers, (Positive conversion) correctly.
- The final model has Precision of 0.667, this means 67% of predicted hot leads are True Hot Leads.
- We have also built an reusable code block which will predict Convert value and Lead Score given training, test data and a cut-off. Different cutoffs can be used depending on the use-cases (for eg. when high sensitivity is required, when model have optimum precision score etc.)