LEAD SCORING ASSIGNEMENT

Batch – DS C52

Pratap Singh, Prateek Sisodia, Praveen Maranur

PROBLEM STATEMENT

- Dataset is for an education company X-Education, selling online cources to industry professionals.

- They market these courses on websites and search engines.

- There are several sources for the Lead Generation.

- Their Typical Lead Conversion Rate is 30%.

- The management is looking for efficiency in resource allocation.

- They want to concentrate resources more on leads which have a higher probability of conversion.
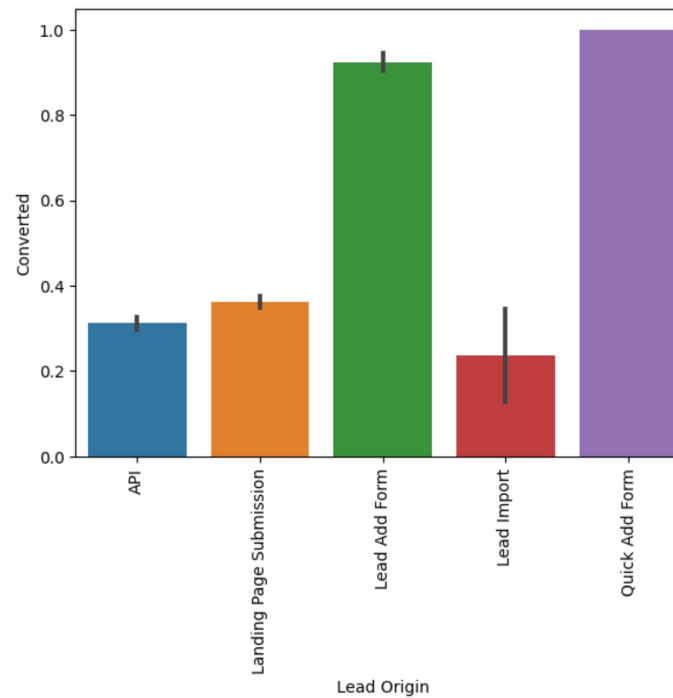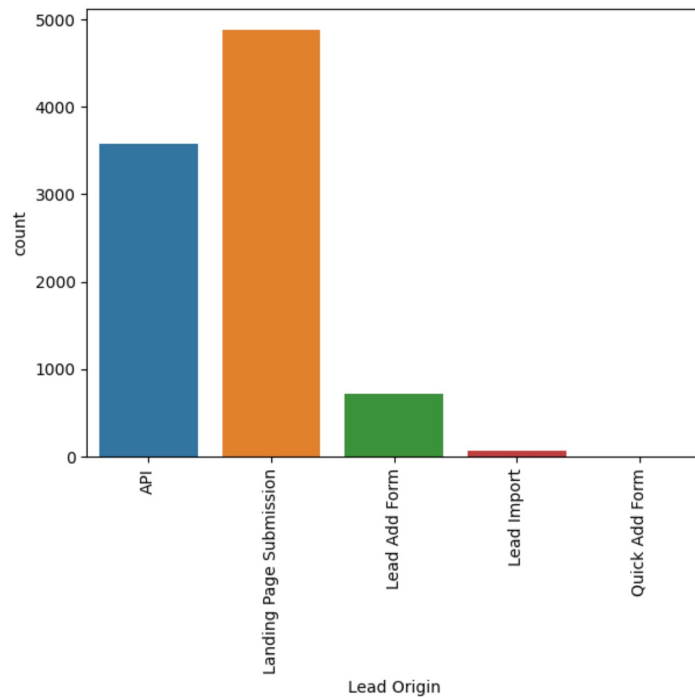
FINAL AIM OF ANALYSIS

- Improve the quality of the Hot Leads.

- Hot Leads are the ones which have a high probability of conversion.

- Conversion Rate the management is looking for is 80%.

- With better lead classification to Hot Leads lead nurturing to ensure minimum 80% Conversion

EXPLORATORY DATA ANALYSIS
&
RECOMMENDATIONS

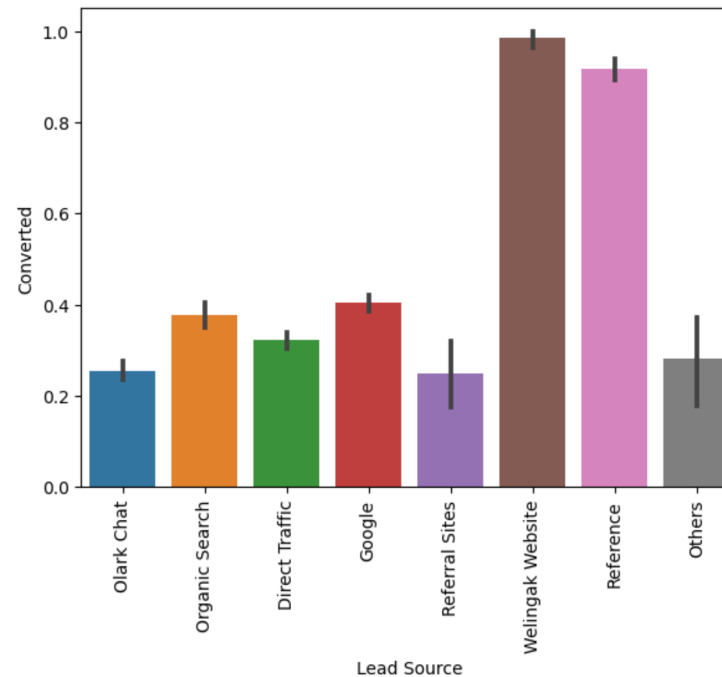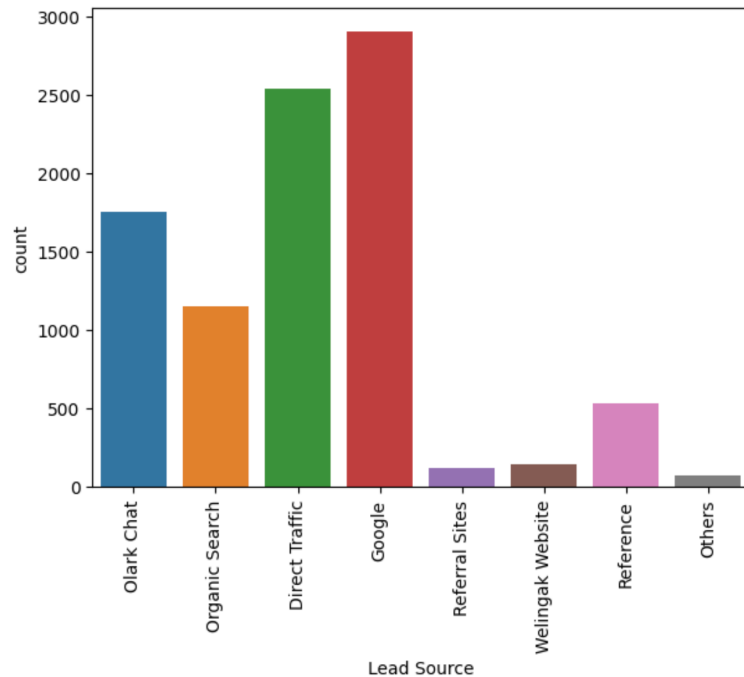# LEAD ORIGIN AND CONVERSION

Lead Counts are Highest
1. Landing Page Submission
2. API

Leads conversions:
1. Lowest for API
2. Highest for Lead Add Form

Recommendation:
Nurturing Leads generated through Lead Forms have a Drastically higher Conversion

LEAD SOURCE AND CONVERSION

Lead Source are Highest
1. Google
2. Direct Traffic
3. Olark Chat
4. Organic Search

Leads conversions:
1. Lowest @ 25-40% for Highest Lead Sources
2. Above 90% for References and Welingak Website

Recommendation:
Nurturing Leads generated through References and Welingak Website. Try generating more leads through these sources.

SPECIALISATION & CONVERSION

Lowest Conversions:
1. Unknown / Undisclosed
2. Services Excellence

Less Probable:
Travel/Tourism, International Business,
E-Commerce, Retail Management

Concentrate More on Others
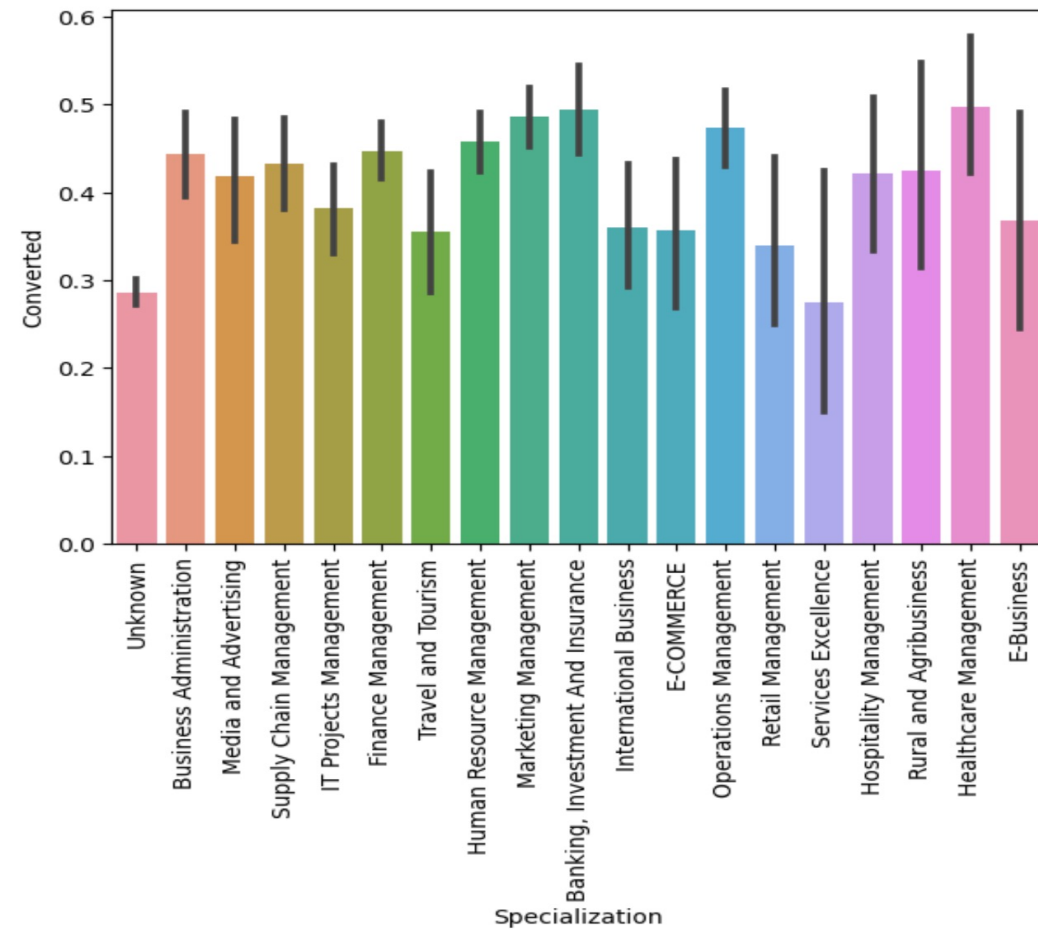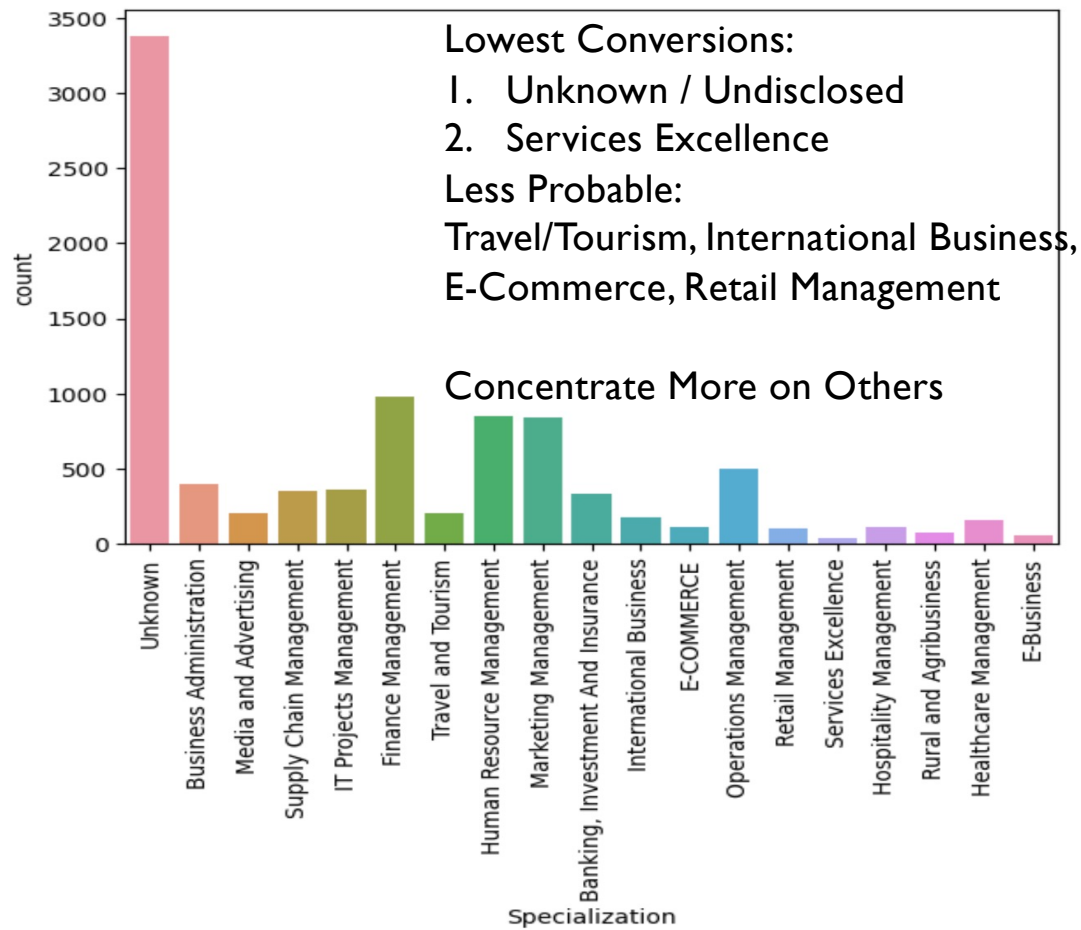
OCCUPATION AND CONVERSION

Lowest Conversion:
1. Unemployed
2. Students

Best Conversions:
1. Housewives Nearly 100%
2. Working Professionals 90%

ACTIVITY & CONVERSION

The Highest count in Activity is of Receiving and Reading Email but it has a very poor Conversion.

Conversion Response is Much More Positive for SMS than Email.

Olark Chat has the worst response. It may be discontinued.
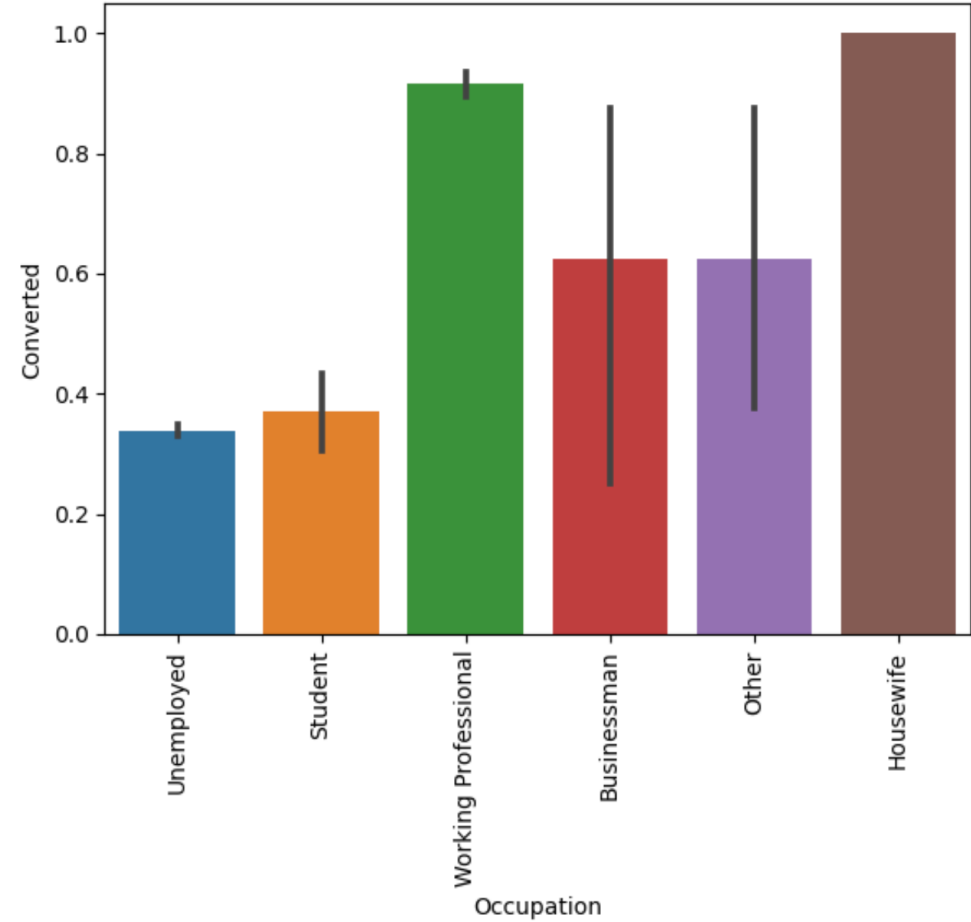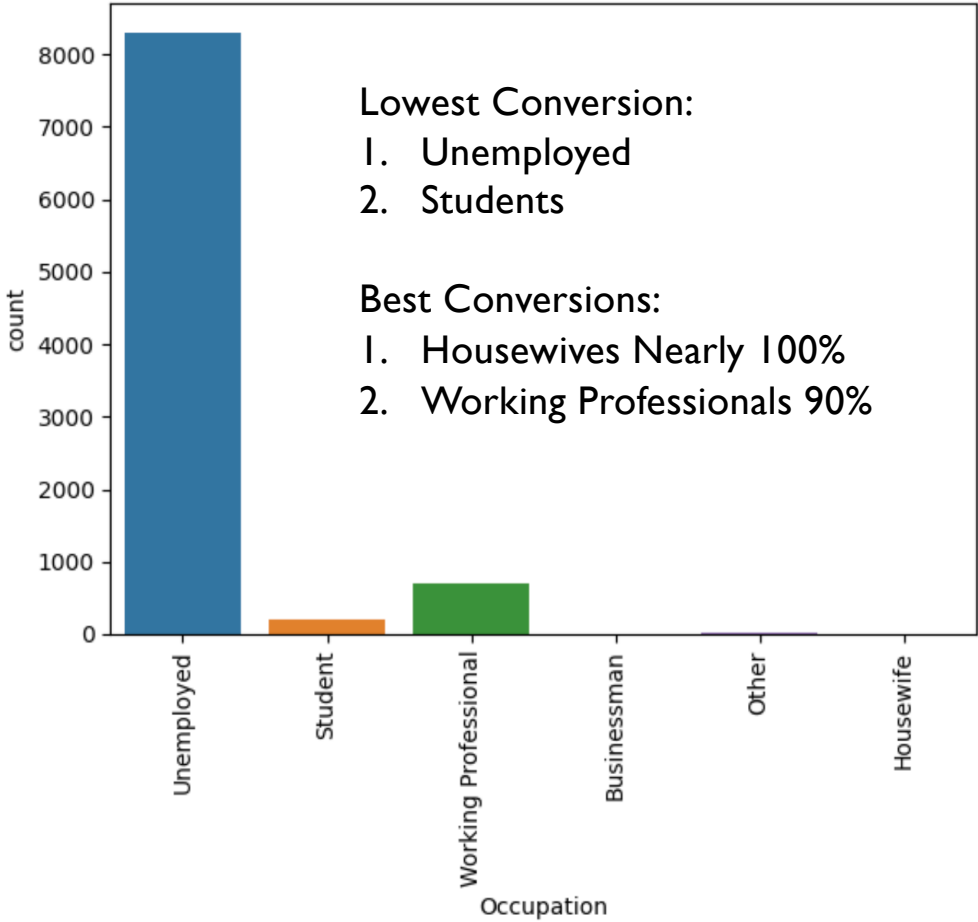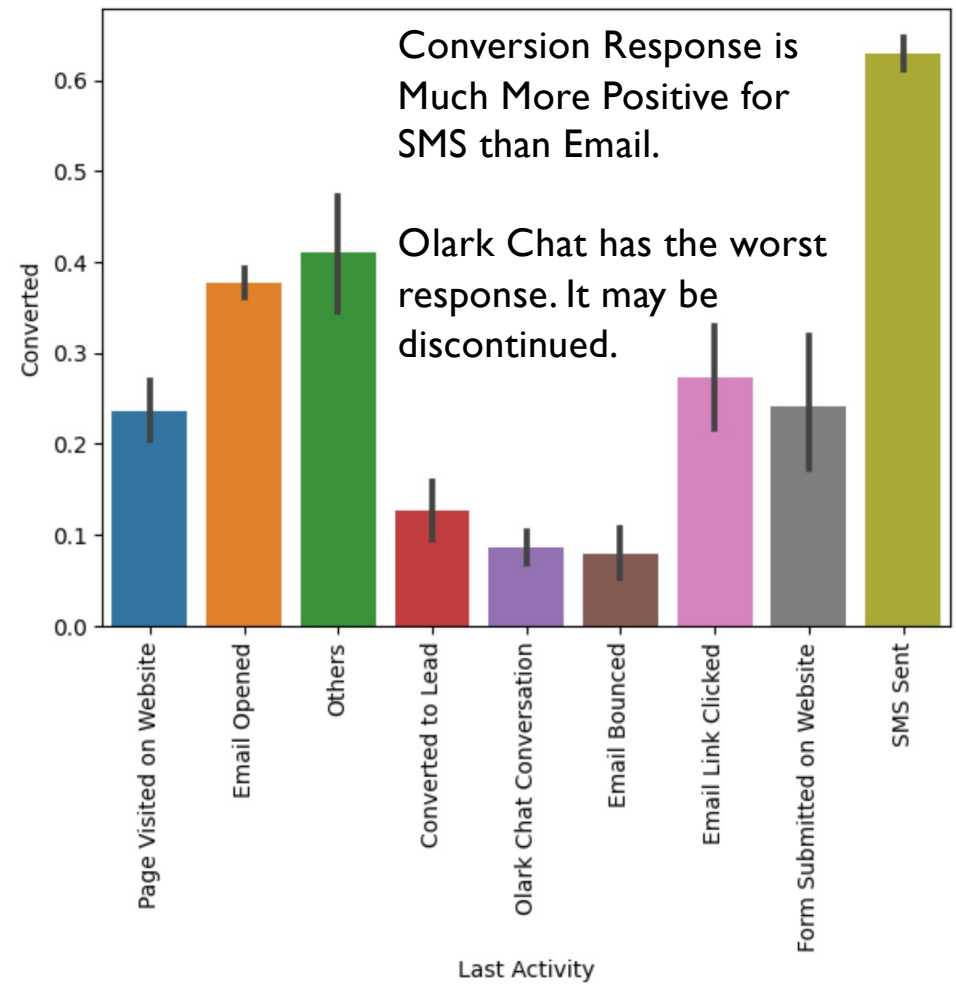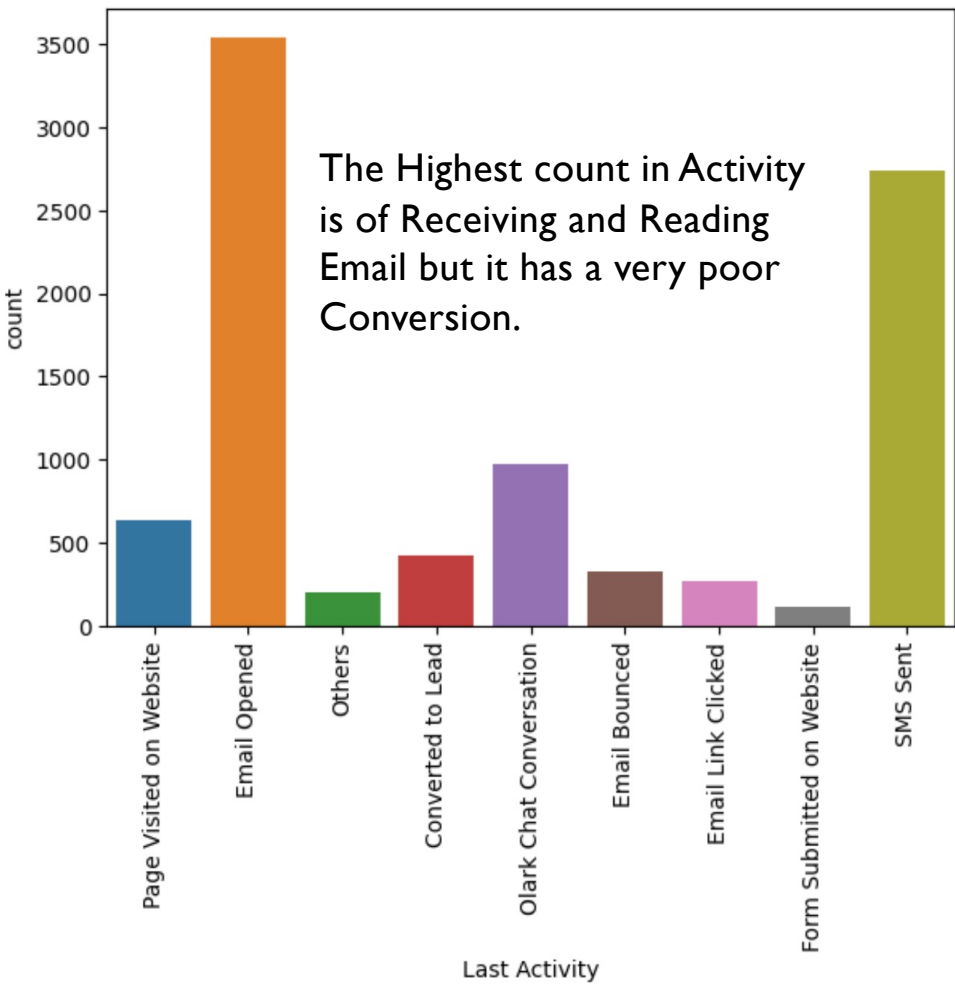
WEBSITE RESPONSE
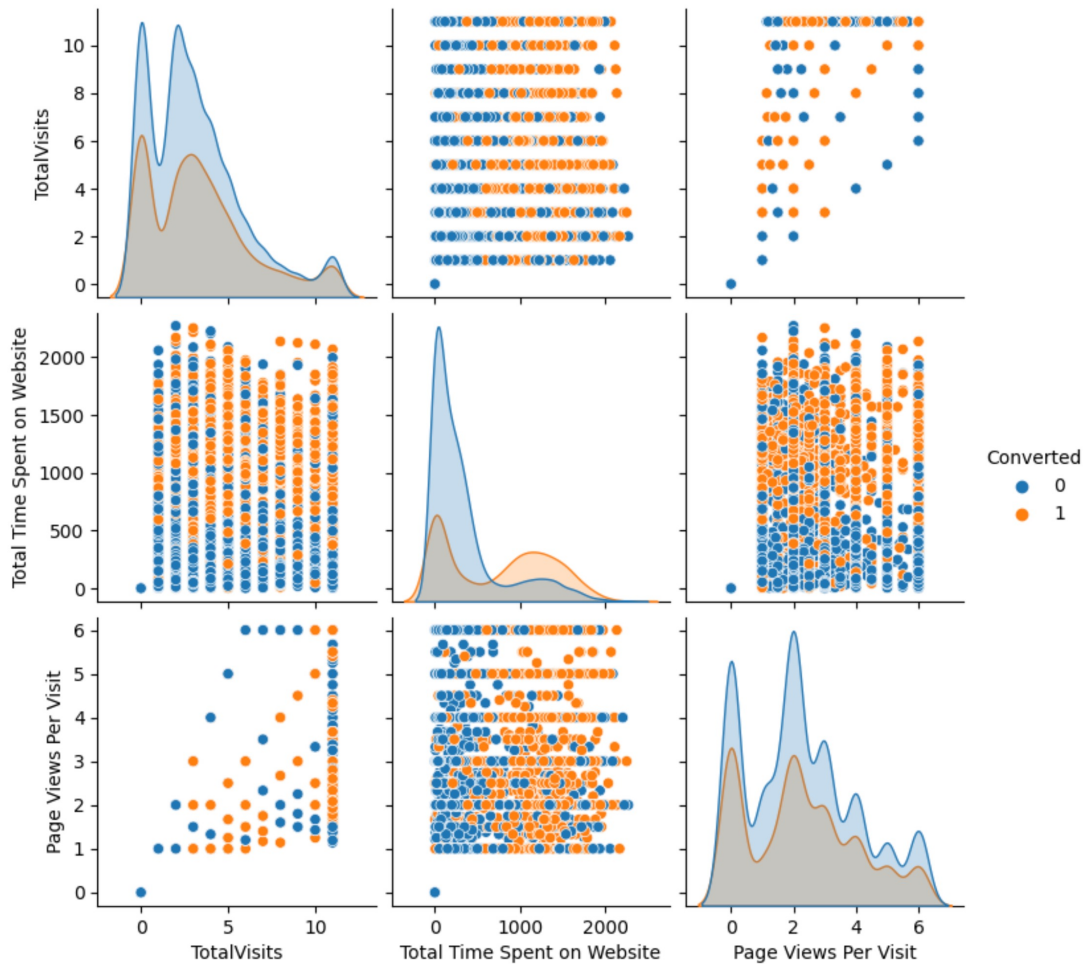
From the Distribution Plots
1. People Spending More than 500 Sec on the website have nearly 66% Chance of conversion.
2. Less than 500 Secs spent on website the chances reduce to 10%
3. More the Number of Visits and More the number of Page Views per visit the Chances of conversion increase to nearly 40-50%

RECOMMENDATIONS

- Concentrate More on Leads which Fill Up the Form.
- Generate More Leads through References and Welingak Website.
- Is there is a Marketing budget for the Welingak Website then increase it.
- Those who do not disclose their Specialisation or are from Services Excellence, Travel/Tourism, International Business, E-Commerce, Retail Management have less conversion rates.
- Working Professionals and Housewives have excellent conversion.
- Connect More with SMS Services
- People Spending More time on the website have good conversion

LOGISTIC REGRESSION MODEL AND RESULTS

```
                     Generalized Linear Model Regression Results
==============================================================================
Dep. Variable:              Converted   No. Observations:                 6468
Model:                            GLM   Df Residuals:                     6454
Model Family:                Binomial   Df Model:                           13
Link Function:                  Logit   Scale:                          1.0000
Method:                          IRLS   Log-Likelihood:                -2771.3
Date:                Tue, 18 Jul 2023   Deviance:                       5542.6
Time:                        09:24:19   Pearson chi2:                 7.58e+03
No. Iterations:                     7   Pseudo R-squ. (CS):             0.3765
Covariance Type:            nonrobust
===============================================================================================
                                          coef    std err          z      P>|z|      [0.025      0.975]
-----------------------------------------------------------------------------------------------
const                                  -2.2515      0.103    -21.959      0.000      -2.452      -2.051
Total Time Spent on Website             1.1004      0.039     28.307      0.000       1.024       1.177
Lead Origin_Landing Page Submission    -0.3721      0.090     -4.146      0.000      -0.548      -0.196
Lead Origin_Lead Add Form               3.3334      0.196     16.970      0.000       2.948       3.718
Lead Source_Olark Chat                  0.9067      0.115      7.903      0.000       0.682       1.132
Lead Source_Welingak Website            2.0522      0.745      2.755      0.006       0.592       3.512
Last Activity_Email Opened              1.1312      0.093     12.185      0.000       0.949       1.313
Last Activity_Others                    1.3671      0.236      5.804      0.000       0.905       1.829
Last Activity_SMS Sent                  2.1773      0.097     22.441      0.000       1.987       2.367
Occupation_Working Professional         2.8725      0.187     15.393      0.000       2.507       3.238
Specialization_Banking, Investment And Insurance  0.4767  0.182  2.620    0.009       0.120       0.833
Specialization_Finance Management       0.4249      0.112      3.798      0.000       0.206       0.644
Specialization_IT Projects Management   0.4238      0.181      2.336      0.019       0.068       0.779
Specialization_Rural and Agribusiness   0.8585      0.374      2.293      0.022       0.125       1.592
===============================================================================================
```
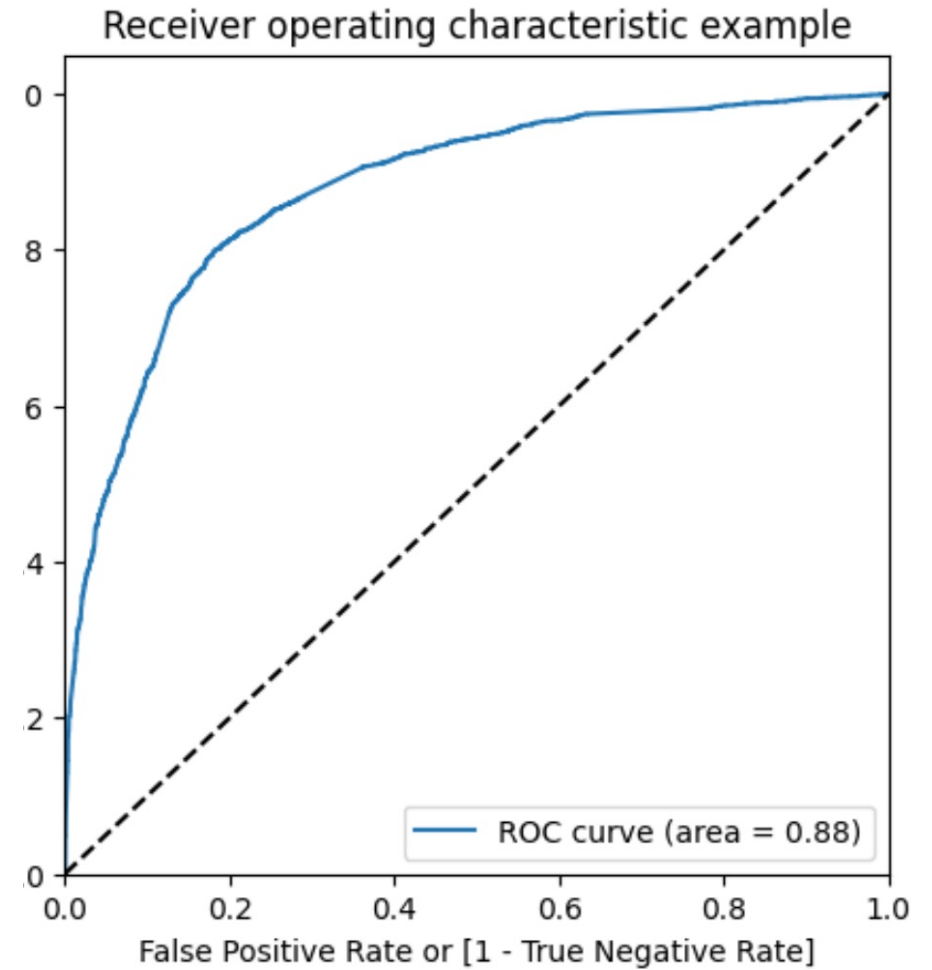
FINAL MODEL

- Highest Coefficients:

  - Origin - Lead Add Form

  - Occupation – Working Professionals

  - Activity - SMS Sent

  - Source – Welingak Website

  - Activity – Email Opened

  - Total Time spent on Website

ROC CURVE AND AREA UNDER THE CURVE
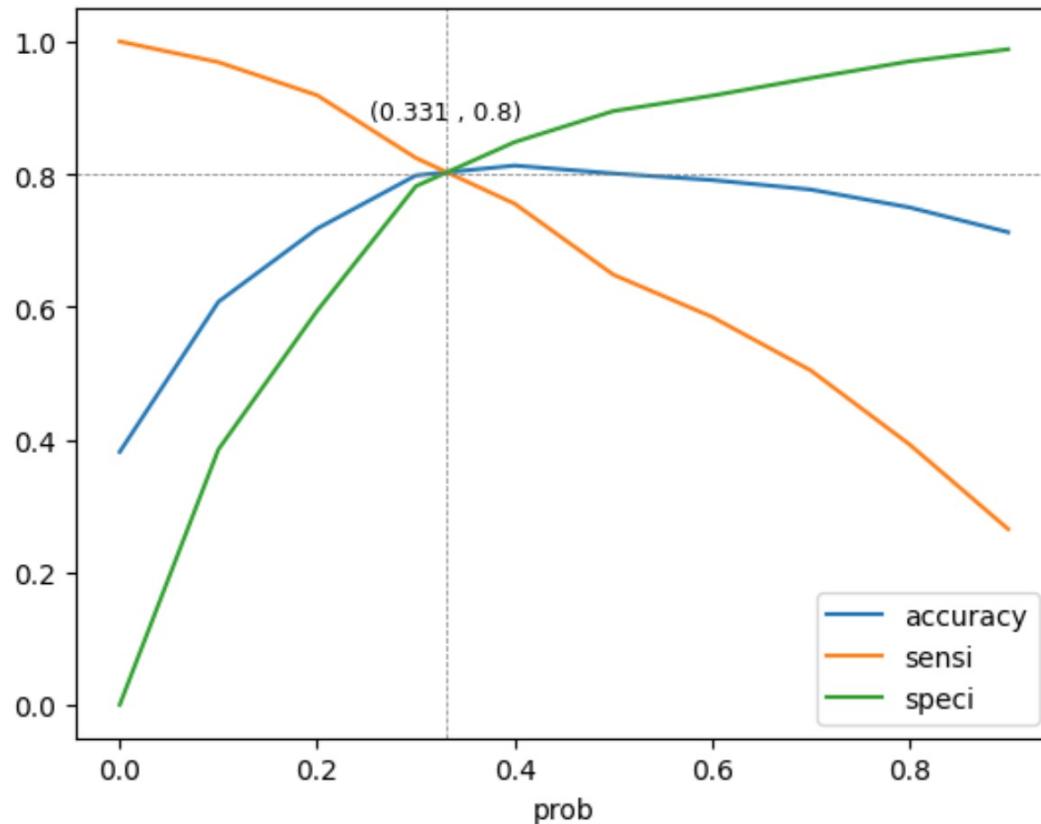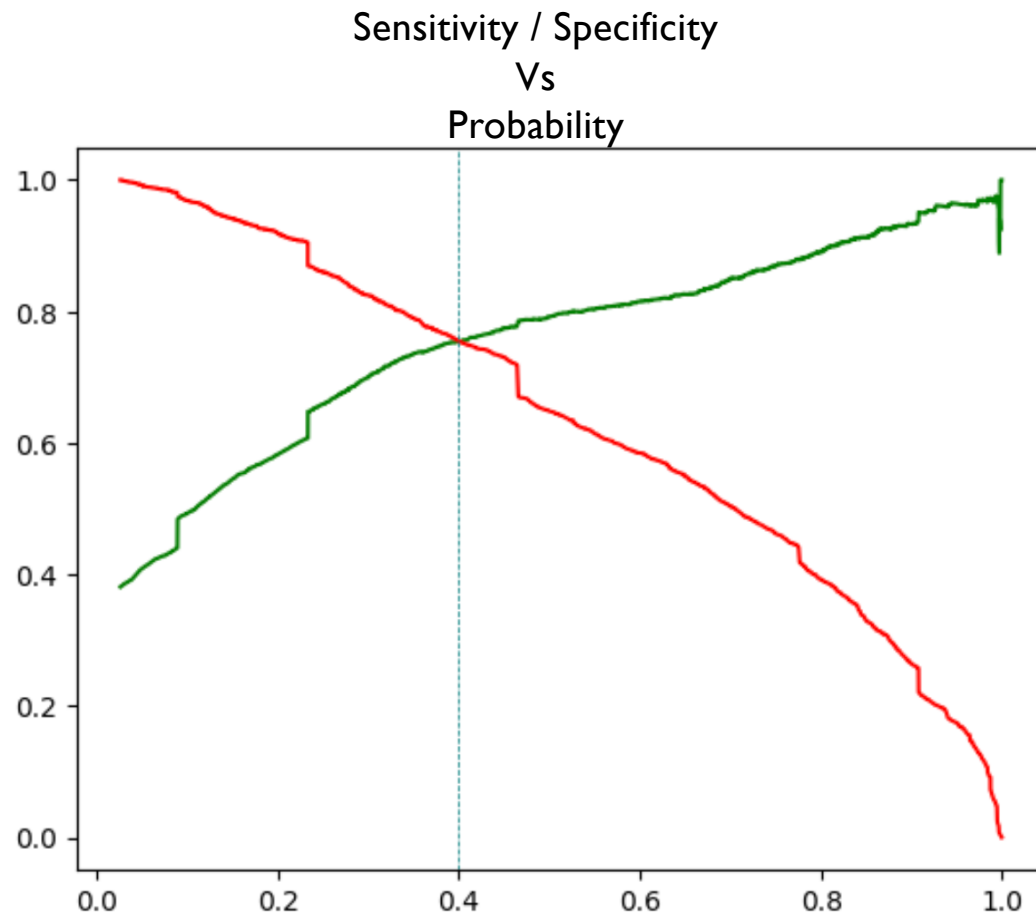
The Area Under the curve is 0.88

Receiver operating characteristic example

ROC curve (area = 0.88)

False Positive Rate or [1 - True Negative Rate]

OPTIMIZING FOR RIGHT PROBABILITY CUTOFF

Probability Cut–Off

0.331

- Accuracy – 80%
- Sensitivity – 80%
- Specificity – 80%

It complies with the Management expectation of min 80% Conversion Rate

Sensitivity / Specificity
Vs
Probability

METRICS ON TRIANING SET

- Accuracy: 0.81
- Sensitivity: 0.8
- Specificity: 0.81

- False Positive Rate: 0.19
- Positive Predictive Value: 0.72
- Negative Predictive Value: 0.87

METRICS ON TEST SET

- Accuracy: 0.8
- Sensitivity: 0.8
- Specificity: 0.8

- False Positive Rate: 0.2
- Positive Predictive Value: 0.72
- Negative Predictive Value: 0.86

ALL METRICS AS PER THE EXPECTATION OF THE MANAGEMENT:

20% FALSE POSITIVES