Lead Scoring Case Study

BY :

1. S Sailesh
2. Karan Dhorage
3. Mirnalini Beniwal

Contents

- Problem statement
- Business Objective
- Problem approach
- Insights and Graphs from EDA
- Model Evaluation
- Conclusion
- Recommendation

Problem Statement

- X Education, an online course provider for industry professionals, faces a challenge with a low lead conversion rate despite generating a significant number of leads daily.
- With only around 30% of leads resulting in conversions, the company aims to enhance efficiency by identifying 'Hot Leads'—those most likely to become paying customers.
- The objective is to build a lead scoring model that accurately assigns scores to leads, distinguishing high-conversion probability leads from lower ones.
- The target is to achieve an 80% lead conversion rate by prioritizing and focusing the sales efforts on leads with higher scores, optimizing the middle-stage lead nurturing process for increased conversions.

Business Objective

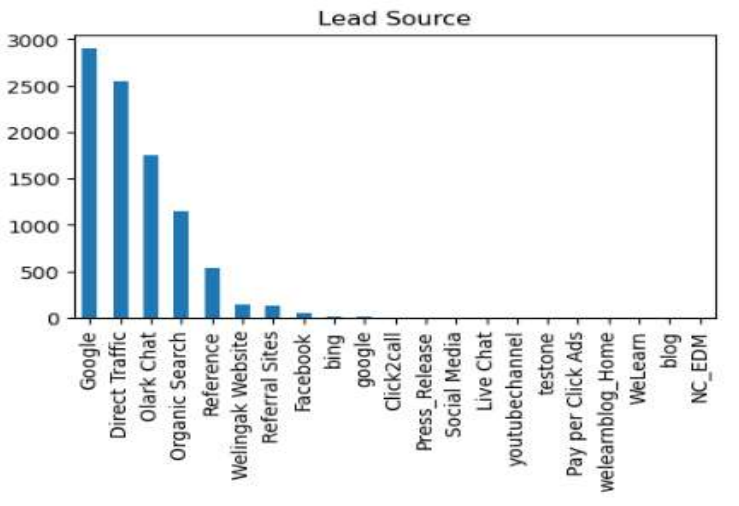
- X education wants us to help them select the most promising leads and require us to build a model that assigns a lead score between 0 to 100 for each lead.
- The CEO wants us to achieve a lead conversion rate of 80%.
- They want the model to be able to handle future constraints as well like Peak time actions required, how to utilize full man power and after achieving target what should be the approaches.

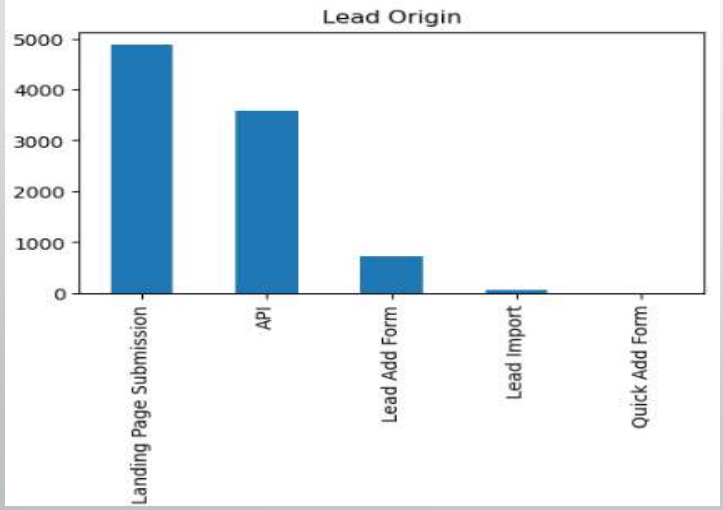
Problem Approach

1. Data Preparation and EDA
    1. Reading data, Data Cleaning and Data Manipulation
    2. EDA(Exploratory Data Analysis)
2. Model Building
3. Model Evaluation

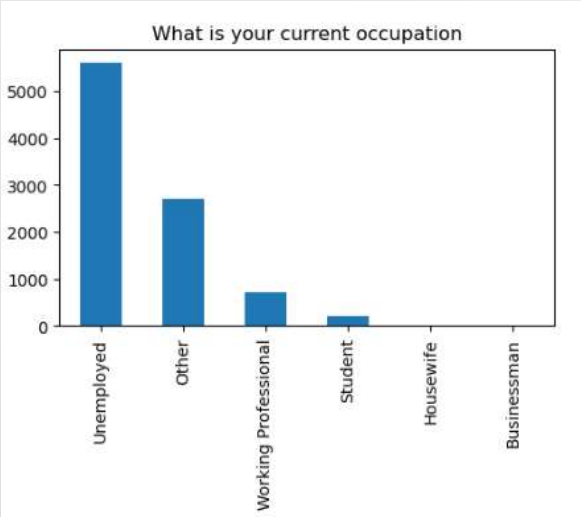
Insights And Charts From EDA

- Most of the Leads are from Google followed by Direct Traffic and Olark Chat



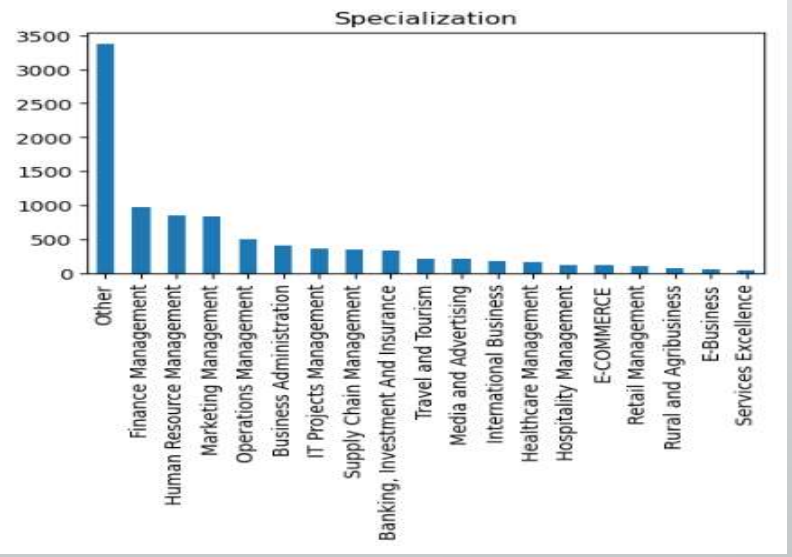


- Most of the Leads Origin from Landing page submission



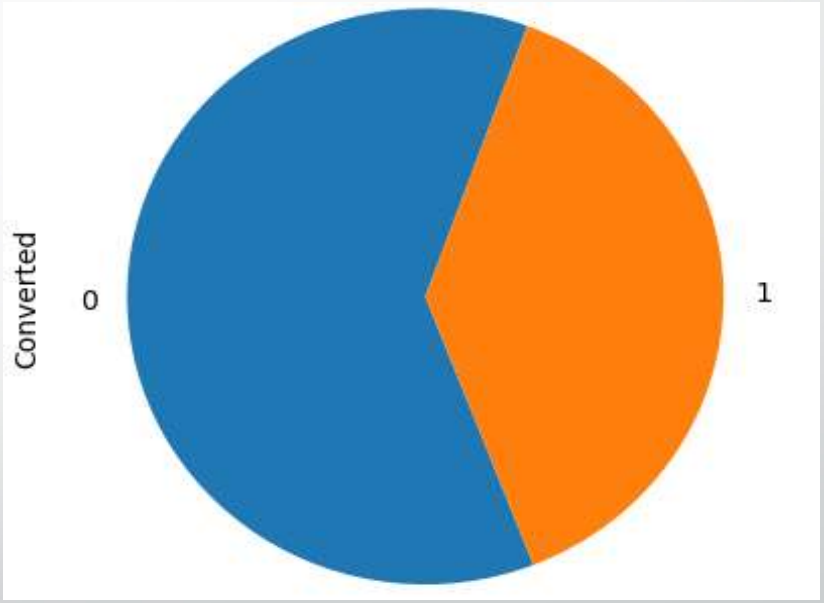
- Majority of Leads are Unemployed and there are very few students.

- Most of the Leads did not have the option to select specific specializations and hence majority of them did not select any option

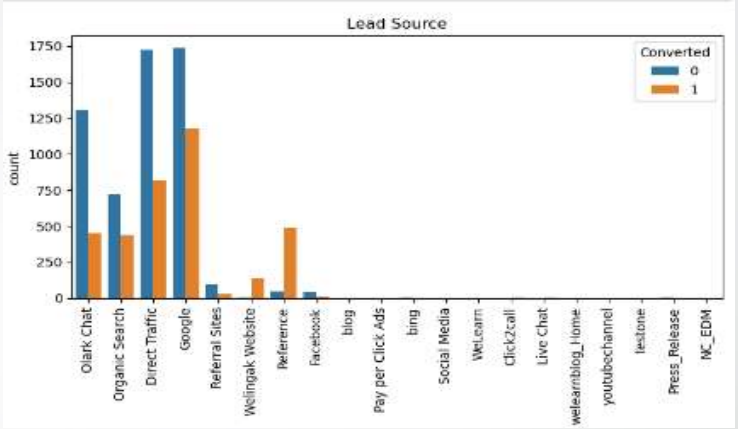


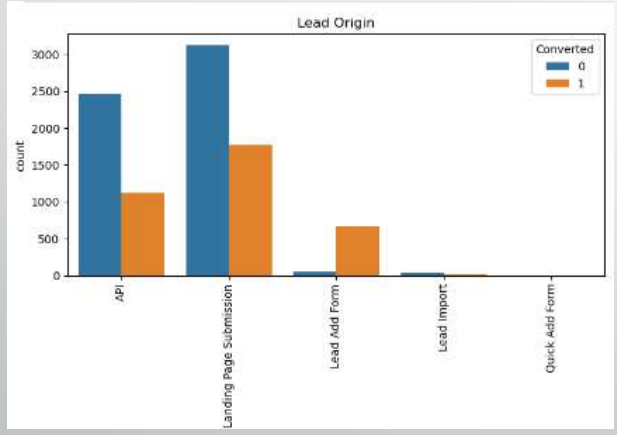
Analyzing the Target Column

- From the Pie chart we see that about 40% of the Leads have been converted
- This implies that there is not much imbalance in the data.



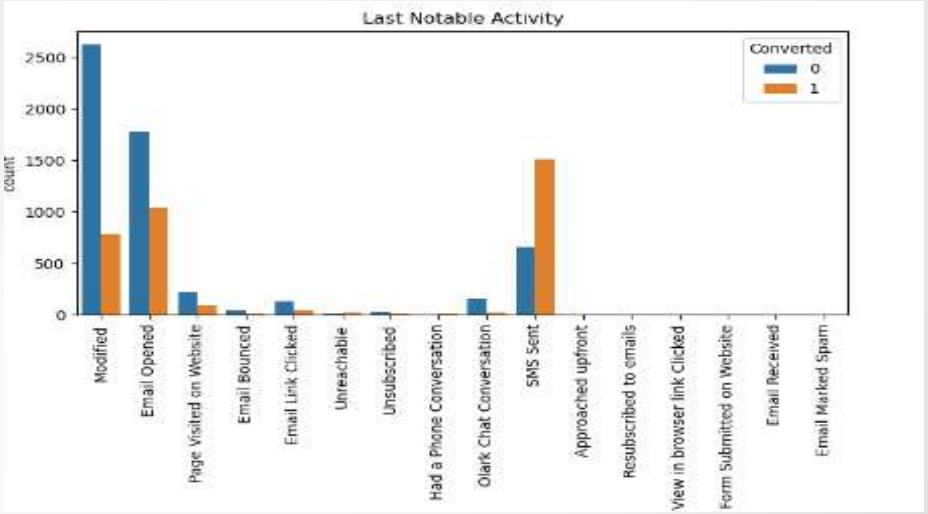
- Leads from ‘Reference’ and ‘Welingak Website’ Have the highest conversion rate.
- Whereas Leads from ‘Olark Chart’ and ‘Direct Traffic’ have the lowest probability of conversion

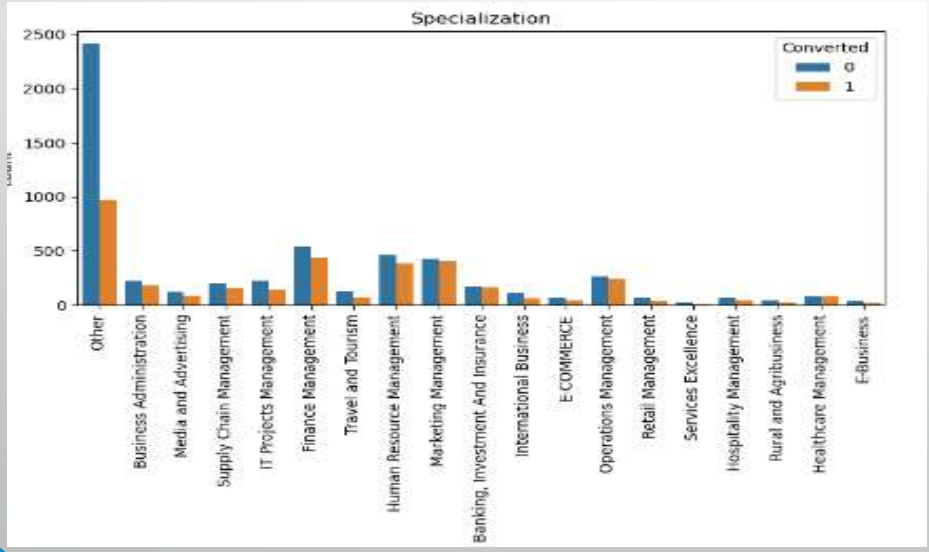




- Leads Originating from Lead Add Form are more likely to get converted when compared to other Lead Origins.

- Leads with Last Notable Activity labeled as 'Modified' have the lowest probability of conversion.
- On Contrary to the label 'SMS Sent', which has the highest probability of conversion

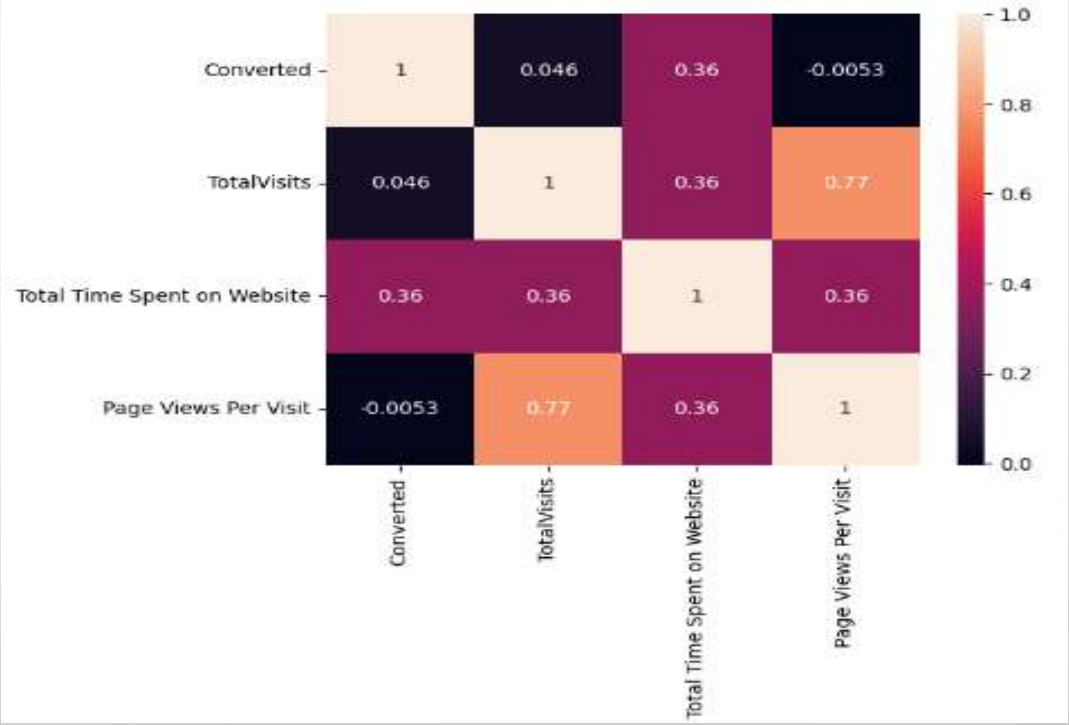




- Leads who fail to select a specialization are more likely to not get converted.

Multivariate Analysis

- Page Views Per Visit and TotalVisits have a higher correlation coefficient



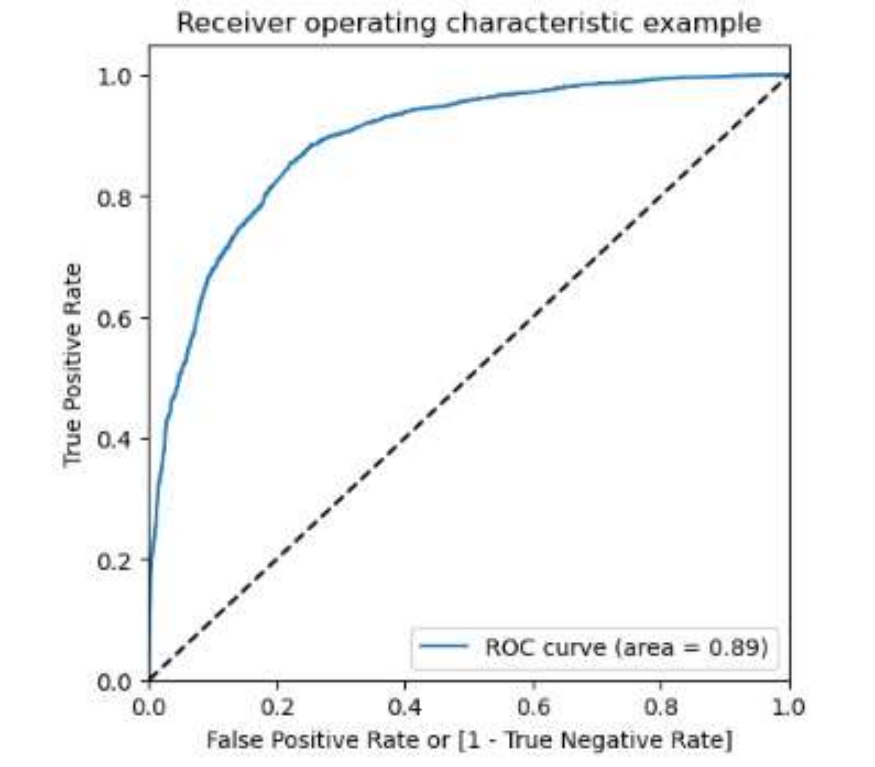
Final Model And its Features

Our Final Logistic Regression model had a total of 14 features, they are:

1. Last Notable Activity_Had a Phone Conversation
2. Lead Origin_Lead Add Form
3. What is your current occupation_Working Professional
4. Lead Source_Welingak Website
5. Last Notable Activity_Unreachable
6. Do Not Email
7. Last Activity_SMS Sent
8. Total Time Spent on Website
9. What is your current occupation_Other
10. Lead Source_Olark Chat
11. Last Activity_Olark Chat Conversation
12. Specialization_Hospitality Management
13. Last Notable Activity_Modified
14. Lead Origin_Landing Page Submission

Model Evaluation

- The area under the ROC curve is **0.89**, Indicating excellent model fit.



Probability Vs Accuracy, Sensitivity, Specificity

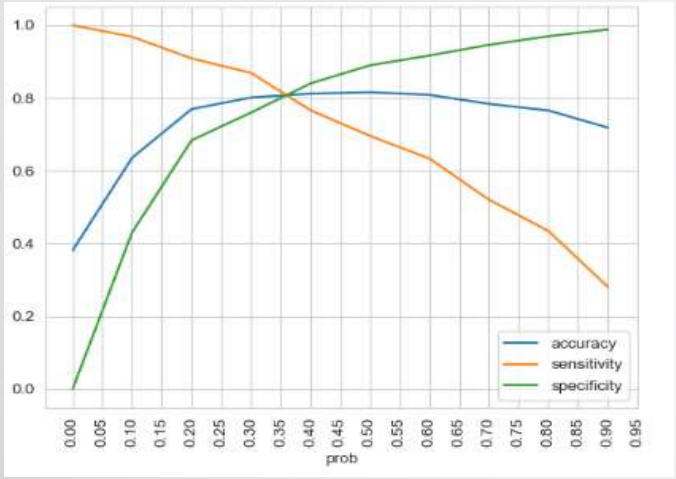


Fig 1

Precision-Recall curve

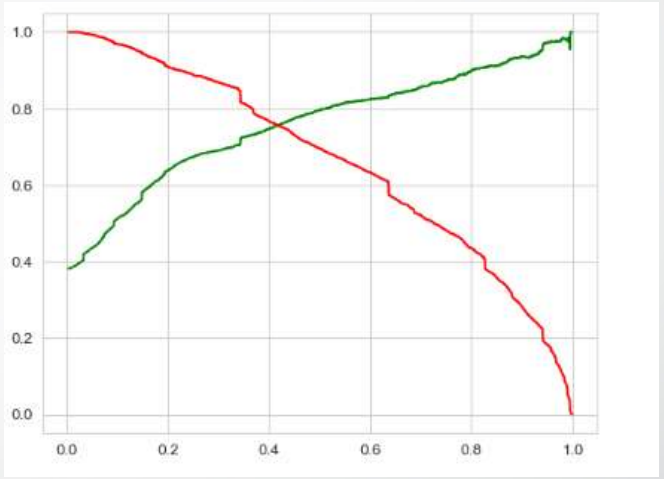


Fig 2

- From the above graphs we get 2 thresholds 0.35(fig 1) and 0.42(fig 2).
- We went ahead with 0.35 as the threshold, since the model performed the best at this point.

Confusion Matrix and Evaluation Metrics

Train Data

accuracy : 81.1%
recall : 81.3%
precision : 72.5%
f1_score : 76.6%
Sensitivity : 81.3%
Specificity : 81.0%

Test Data

accuracy : 81.6%
recall : 81.2%
precision : 74.6%
f1_score : 77.7%
Sensitivity : 81.2%
Specificity : 81.9%

Conclusion

- Successfully Built a Logistic Regression Model to predict the conversion of lead into paying customer with an accuracy of 81%.
- Some of the variables influencing the predicted outcomes are:
  - **Lead Sources:** Google, Direct traffic, and Olark Chat are primary sources; landing page submissions dominate, with a higher proportion of unemployed leads.
  - **Conversion Rates by Source:** 'Reference' and 'Welingak Website' lead to the highest conversion rates, while 'Olark Chat' and 'Direct traffic' result in the lowest conversions.
  - **Occupation Impact:** Working professionals' leads have higher conversion potential compared to those without specified occupations.
  - **Lead Origin Impact:** Leads from the Lead Add Form show the highest conversion likelihood among various lead sources.
  - **Last Notable Activity:** Leads marked 'Modified' show lower conversion potential, while 'SMS Sent' or 'Had a Phone Conversation' signify higher chances of conversion.

Recommendations

- **Refine Marketing Strategies:** Focus on optimizing strategies for Google, Direct traffic, and Welingak Website, which yield higher conversion rates, while reassessing approaches for Olark Chat and Direct traffic to improve their effectiveness.
- **Target Working Professionals:** Given their higher conversion potential, tailor marketing campaigns or engagement tactics to attract and engage working professionals specifically.
- **Last Activity Engagement:** Prioritize leads with 'SMS Sent' or 'Had a Phone Conversation' as these activities exhibit higher conversion chances, while strategizing on how to engage or re-engage 'Modified' leads effectively.
- **Utilize Predictive Lead Scoring:** Leverage the Logistic Regression model's insights and Lead Scores to prioritize leads more effectively, enabling a targeted approach towards high-potential prospects, thus improving overall conversion rates.