

LEAD SCORING CASE STUDY

Presenting by:

Abhishek Goel

Abhishek Dhabal

Bhavani Jagu

Problem Statement

An education company named X Education sells online courses to industry professionals. On any given day, many professionals who are interested in the courses land on their website and browse for courses.

The company markets its courses on several websites and search engines like Google. Once these people land on the website, they might browse the courses or fill up a form for the course or watch some videos. When these people fill up a form providing their email address or phone number, they are classified to be a lead. Moreover, the company also gets leads through past referrals. Once these leads are acquired, employees from the sales team start making calls, writing emails, etc. Through this process, some of the leads get converted while most do not. The typical lead conversion rate at X education is around 30%.

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. To make this process more efficient, the company wishes to identify the most potential leads, also known as 'Hot Leads'. If they successfully identify this set of leads, the lead conversion rate should go up as the sales team will now be focusing more on communicating with the potential leads rather than making calls to everyone.

Analysis approach

- To achieve lead score around 80%, we build a logistic regression model after exploratory data analysis (EDA)
- Initially there is a need of data import, inspection and cleaning of data, handling of categorical variables
- Next step is exploratory data analysis (EDA) which includes univariate analysis of both categorical and numerical variables using count plots and hist plots respectively.
- Identification of correlation of numerical variables using heat maps projected the top most three variables helps in conversion of leads to hot leads.
- From the heap map, we can conclude that there is a positive correlation for total visits, total time spent on website and page views per visits
- Dummy variable creation is necessary in regression models when dealing with categorical variables because these models can only process numerical inputs. So all the necessary categorical variables got converted to dummy variables
- Dropping of original categorical columns after dummification of categorical variables is required.
- Building regression models using Scikit-learn and statsmodel libraries is major step for the predicative models results

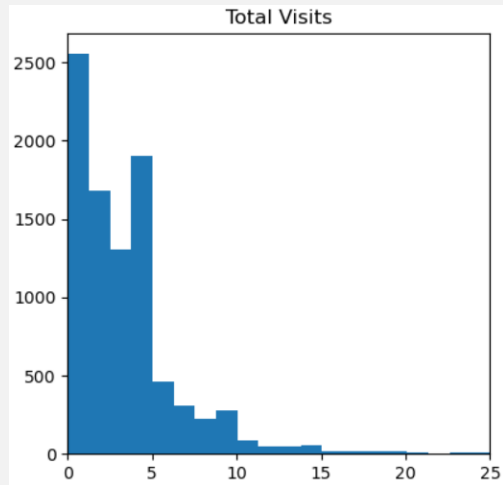
Analysis approach

- Many variables have p-value greater than 0.05 and high VIF for certain variables. So these variables are dropped for further analysis
- Dropping all variables with high P value and low RFE ranking
- Model evaluation by detecting conversion probabilities can be done on targeted variable using train set and test sets

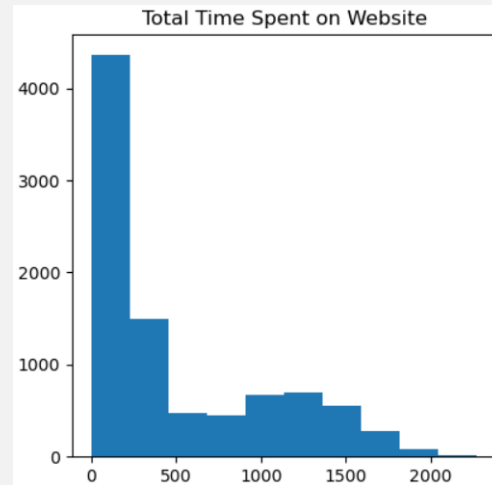
Interpretation of results

- After analyzing all the variables using count plots, few variables have significant right –skewed distribution.
- Those variables are Total Visits, Total Time Spent on Website, Page views per visit.
- Positive predictors to be focused are Total Visits, Total Time Spent on Website, Current Occupation - Working Professional, Lead Origin - Lead Add Form, Current Occupation – Student. All of these variables have positive correlation coefficient.
- The negative predictors to be addressed are Lead Source - Direct Traffic, Do Not Email – Yes, Last Notable Activity - Page Visited on Website

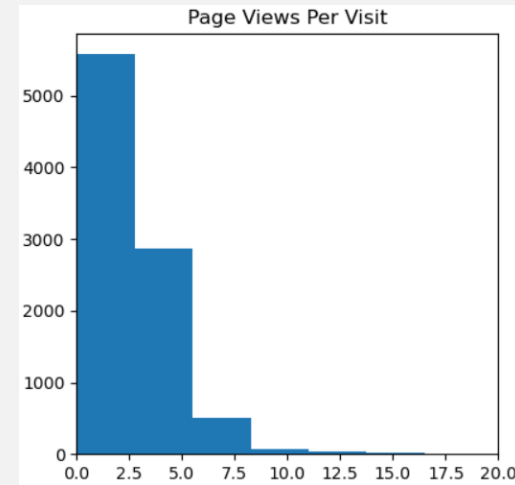
Visualization of variables before model building



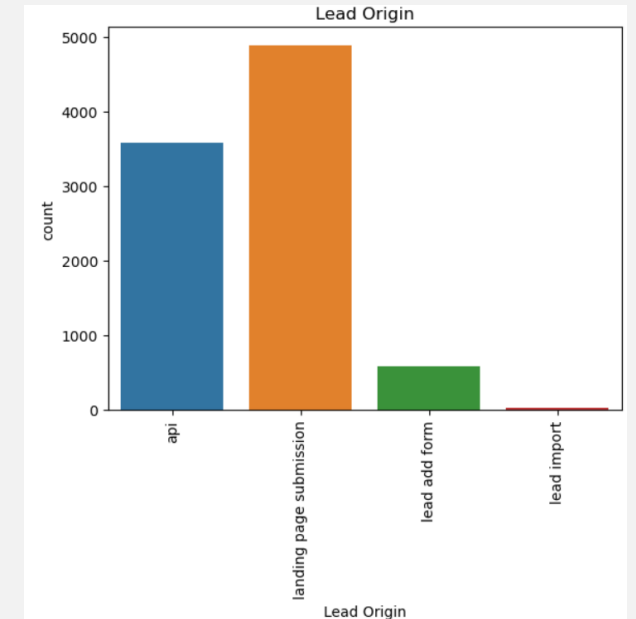
- The distribution is **right-skewed**, with most users having less than 5 visits.
- Very few users have more than 10 visits.



- The distribution is **right-skewed**, with most users spending less than 500 seconds.
- A small number of users spend significant time (close to or over 2000 seconds).

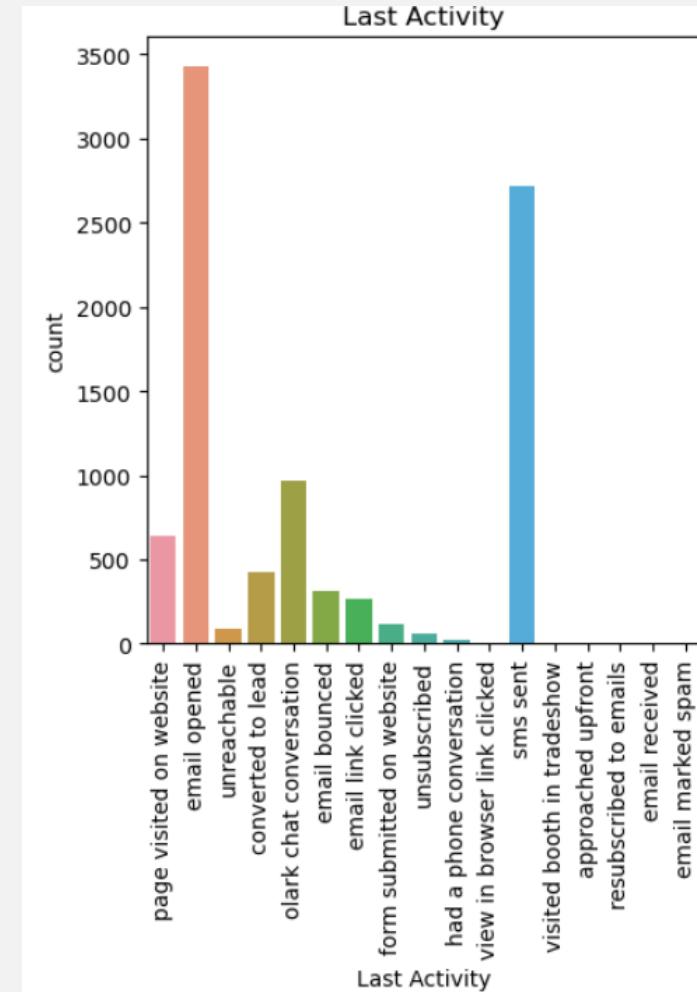
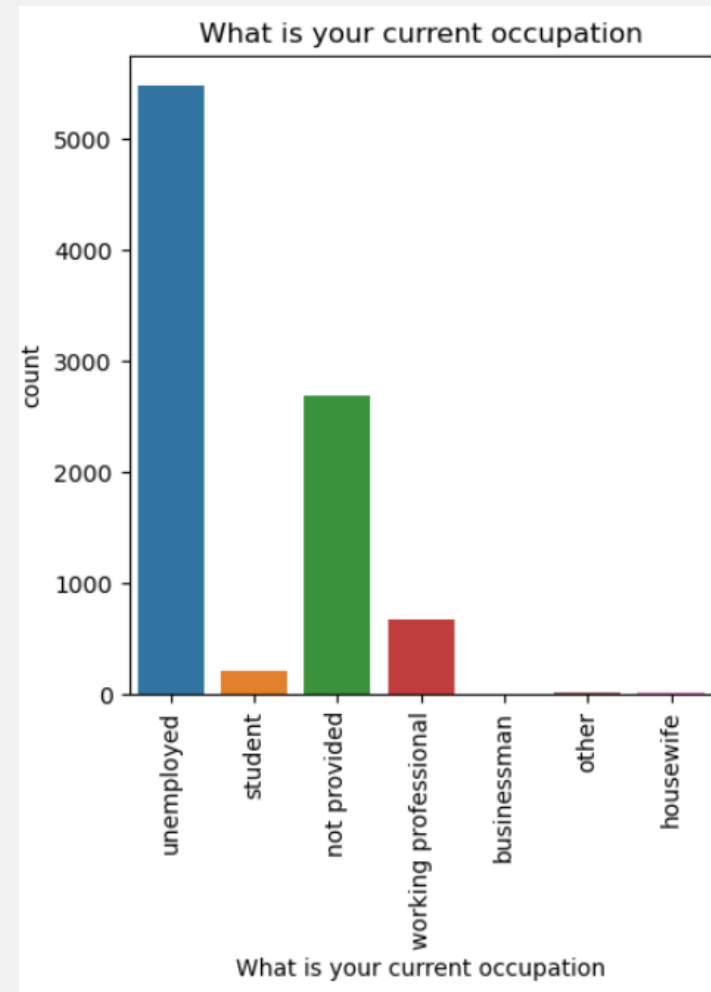
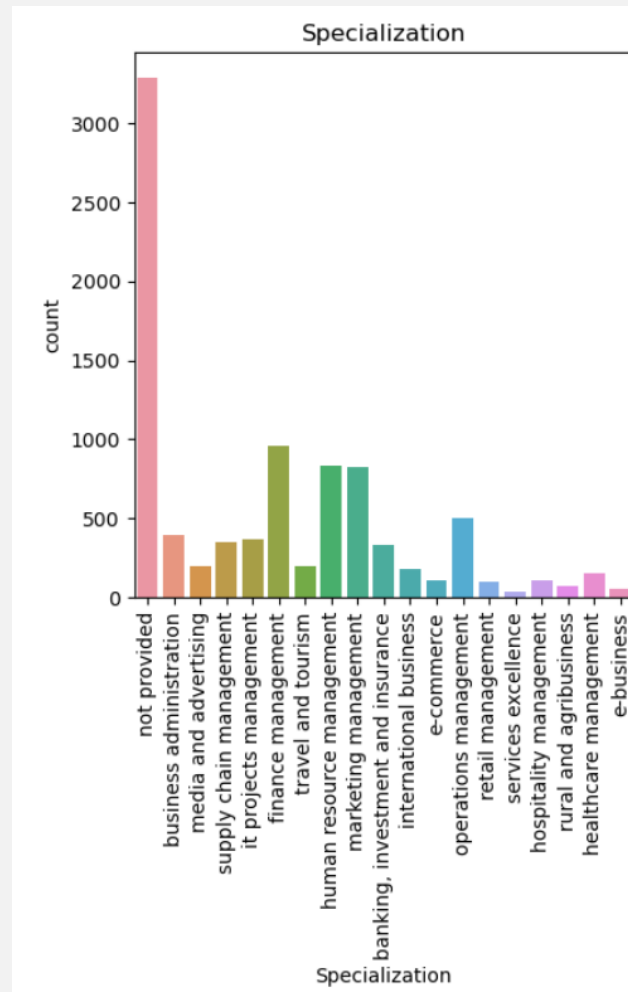


- The distribution is heavily **right-skewed**, with the majority of values clustered between **0 to 3 page views per visit**.
- Very few visits exceed 5 page views.



- The majority of leads are sourced from **Landing Page Submissions** (largest bar) and **API**.
- Lead Add Form** has a significantly smaller contribution to the total leads.

Visualization of variables before model building



Top three variables interpretation

1.Total Visits

1. Coefficient: 9.7183
2. P-value: 0.000 (highly significant)
3. Impact: A higher number of visits is the most impactful predictor.
4. Action: Encourage repeat visits through remarketing, follow-ups, and engaging website content.

2.Total Time Spent on Website

1. Coefficient: 4.6489
2. P-value: 0.000 (highly significant)
3. Impact: More time spent on the website increases conversion likelihood significantly.
4. Action: Optimize the website for engaging and informative content to retain user attention.

3.Lead Origin: Lead Add Form

1. Coefficient: 2.2127
2. P-value: 0.000 (highly significant)
3. Impact: Leads originating from the "Lead Add Form" source are significantly more likely to convert.
4. Action: Focus on optimizing and increasing the usage of lead forms, ensuring they are easily accessible and engaging.

Interpretation of results

- 81.9 % accuracy achieved on train data set
- 71.1% sensitivity achieved
- 88.7% specificity achieved
- 81.4% accuracy achieved on test data set
- 73% precision was observed
- 77.4% recall was observed

Key Takeaways:

1.Top Positive Predictors:

- Total visits, Total Time Spent on Website, lead from lead add form, and users who are working professionals are significantly more likely to convert.

2.Negative Predictors:

- Page Views Per Visit and certain Lead Sources (e.g., direct traffic) are associated with lower conversion likelihood.

3.Actionable Insights:

- Optimize for **lead forms** and target **working professionals** as they show high conversion rates.
- Consider revisiting marketing strategies for low-performing lead sources like direct traffic and organic search.
- Investigate why users with high page views or last activities like page visited on website are less likely to convert.

Interpretation of results

- **TotalVisits (9.7183, $p < 0.001$):**

- For each additional website visit, the log-odds of conversion increase significantly. The strong positive coefficient suggests this is a critical predictor.

- **Total Time Spent on Website (4.6489, $p < 0.001$):**

- Higher time spent significantly increases the likelihood of conversion. This aligns with expectations for engaged users.

- **Page Views Per Visit (-1.3782, $p = 0.021$):**

- Surprisingly, more pages viewed per visit are associated with lower likelihood of conversion. This may suggest users browsing extensively without converting.

- **Lead Origin:**

- landing page submission (-0.2790, $p = 0.006$): Less likely to convert compared to the reference category.

- lead add form (2.2127, $p < 0.001$): Strongly positive, suggesting leads from forms are more likely to convert.

- **Lead Source:**

- Negative coefficients for direct traffic, google, and organic search indicate lower conversion likelihood compared to the baseline.

- Positive coefficient for welingak website (2.1114) indicates this source has a strong positive impact on conversions.

- **Occupation:**

- working professional (3.6694, $p < 0.001$): This category has the highest positive association with conversion likelihood.

- **Last Notable Activity:**

- Most activities have negative coefficients, e.g., page visited on website (-1.8479, $p < 0.001$), suggesting these interactions are less likely to lead to conversions.

Business implications

- Increase total visits by using retargeting ads and remarketing emails to bring back previous visitors
- Offer incentives, such as discounts or exclusive content, for repeat visits.
- Optimize user experience with intuitive navigation and fast loading speeds.
- Include engaging content like videos, blogs, or FAQs relevant to the audience.
- Use recommendation engines to personalize the user journey and keep visitors on the site longer.
- Since a large proportion of leads come from landing pages but have a lower conversion probability, it's crucial to improve the landing page design and content
- Use A/B testing to evaluate and optimize calls-to-action (CTAs), page layout, and form fields
- Invest in channels that direct users to Lead Add Forms, as this category has the highest positive impact on conversions
- Minimize the number of pages needed for a user to perform key actions, such as filling out forms or viewing critical information
- Focus on improving the content and design of high-traffic pages to drive conversions earlier in the visit
- Ensure CTAs (Calls-to-Action) are clear and accessible on landing pages
- Analyze heatmaps and session recordings to understand why users are browsing more pages without converting. Address usability issues or gaps in information

A top-down view of a desk with a light-colored, textured wooden surface. In the center is a white spiral-bound notebook with the words "THANK YOU!" written in large, bold, black sans-serif font. The exclamation mark is red. To the top left of the notebook are gold-rimmed glasses with black temples. To the right is a silver and black ballpoint pen. To the bottom left is a small white pot containing a green succulent plant and brown pebbles. A black cable is visible in the bottom right corner.

**THANK
YOU!**