Lead Score Case Study

BY: KUSHAL GAMBHIR

BATCH: EPGP DSC66

The Problem

- X Education is an education company which sells online courses to industry professionals. Many people come on the website through multiple sources and when they fill up a form providing their email address or phone number, they are classified to be a lead.
- The typical lead conversion rate at X education is around 30%, i.e., only 30 people buy the product/ course from X Education out of 100 people who filled and submitted the form.
- Since most of the leads are not buying although they show their interest so a lot of resources efforts in converting the leads are not fruitful and this needs to be optimized.
- To make this process more efficient, the company wishes to identify the most potential leads, also known as 'Hot 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.
- The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%.

Challenges Faced with the data

1. Inconsistent lead quality
2. Difficulty prioritizing leads
3. Wasted resources on low-potential leads

Understanding Lead Scoring and Solution

- For each lead, the system will be assigning a score based on certain parameters. This score will let us know the probability of a lead converting. If it is close to 1, we say the lead is most probable to convert and if it is close to zero we say the lead is low priority as it is less likely to convert.

- So, we build a logistic regression model to assign a lead score to each of the leads which can be used by the company to target potential leads.

Solution Approach

1. **Data Understanding:** After the data is imported functions like info(), describe(), unique(), head(), shape helps to understand the structure of the dataset, the different variables and their datatypes, missing values, etc.

2. **Data Cleaning:** The initial step involved cleaning the dataset, which was generally well-organized but contained some imperfections. Missing values were addressed by imputing the mode for categorical variables and median for numerical columns. In some cases, the null values were dropped. Some categorical datapoints like Last Activity, Tags were categorized into one bucket to simplify analysis and improve clarity.

3. **Exploratory Data Analysis (EDA):** Exploratory Data Analysis was performed to assess the quality and relevance of the data. It was noted that some categorical variables included irrelevant or redundant information, which could distort the results. However, numeric values were accurate although contained outliers that need to be handled, which facilitated further analysis. Exploring the patterns and relationship between two or more variables.

4. **Creation of Dummy Variables:** Categorical variables were converted into dummy variables to prepare the data for machine learning algorithms. Additionally, MinMaxScaler was applied to normalize the numeric data, ensuring all variables were on a comparable scale.

5. **Train-Test Data Split:** The data was divided into training and testing subsets, with 70% allocated for training the model and 30% for testing its performance. This approach allowed the model to be trained on one subset and evaluated on another, ensuring its effectiveness in predicting outcomes.

Solution Approach

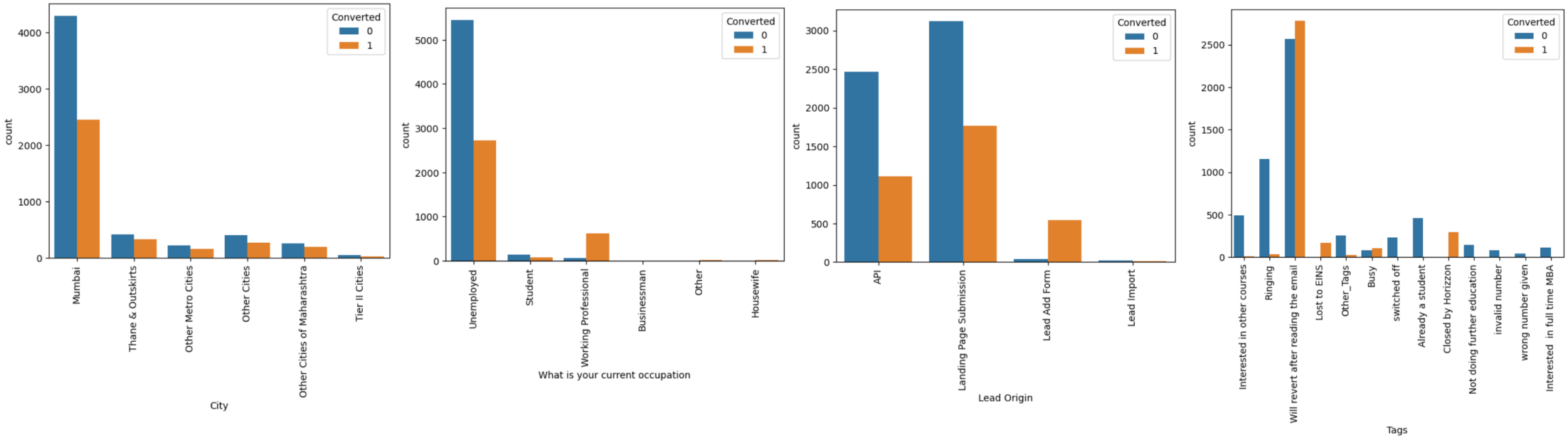
6. **Model Development:** For model development, Recursive Feature Elimination (RFE) was used to identify the top 15 relevant variables from the dataset. RFE systematically removed less significant variables to enhance model accuracy. Subsequently, additional filtering was performed based on Variance Inflation Factor (VIF) and p-values, retaining variables with VIF values below 5 and p-values below 0.05.

7. **Model Evaluation:** The model's performance was assessed using a confusion matrix to determine its classification based on Recall value. The optimal cutoff value was determined using the ROC (Receiver Operating Characteristic) curve, resulting in accuracy, recall, sensitivity, and specificity rates of over 84%. These metrics indicated that the model was effective in predicting potential leads.

8. **Prediction Analysis:** Predictions were conducted using the test data with an optimal cutoff value of 0.42. This cutoff was chosen to balance accuracy and recall, which were maintained around 80%, demonstrating the model's reliability.

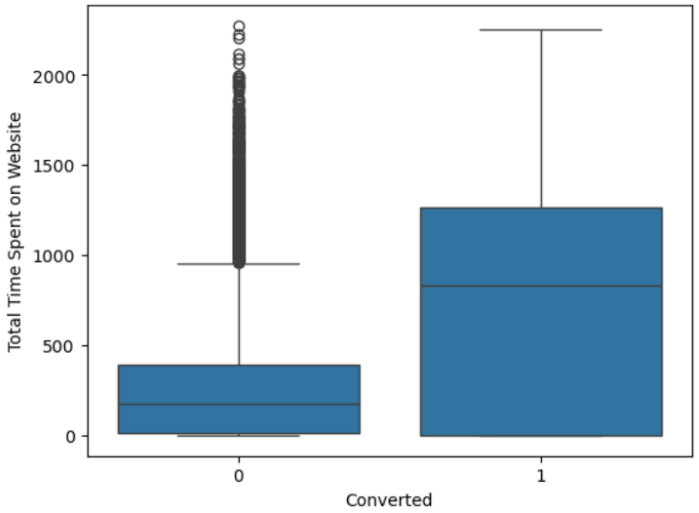
9. **Precision-Recall Assessment:** To further validate the model, a precision-recall analysis was performed. Adjusting the cutoff to 0.42 yielded a recall rate of 84%. Precision measures the proportion of true positive predictions among all positive predictions, while recall assesses the proportion of actual positives identified by the model

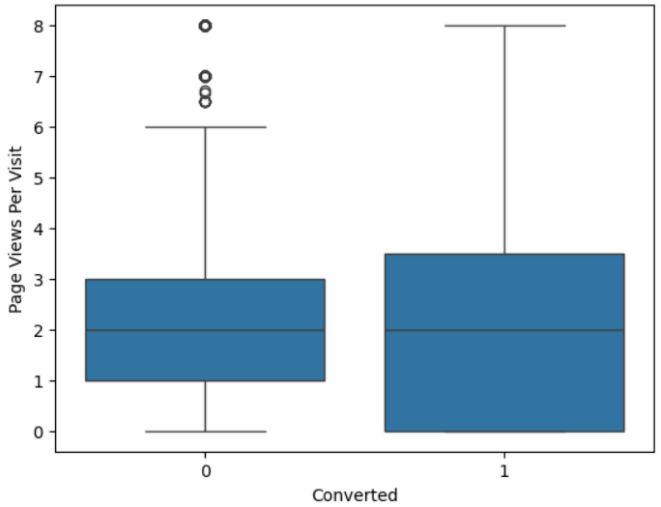
Insights from EDA

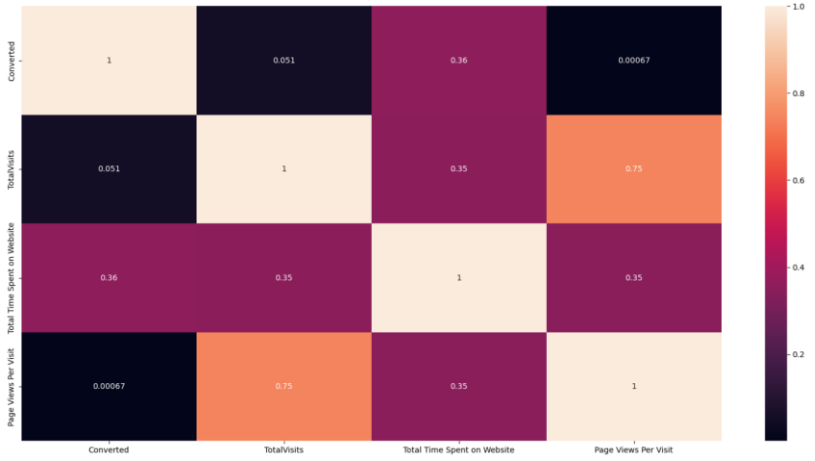


- **Tags**: Will revert after reading the email, closed by horizzon, lost to EINS have a good conversion but we need to focus on 'Interested in other courses' and 'ringing' which are good in numbers but low in conversion.
- **What is your current Occupation**: conversion rate is high for working professionals. Now the focus should be on unemployed leads.
- **Lead Origin**: API and Landing Page Submission have 30-35% conversion rate but count of lead originated from them are considerable.
- **City**: Most of the leads are from Mumbai city.

Insights from EDA



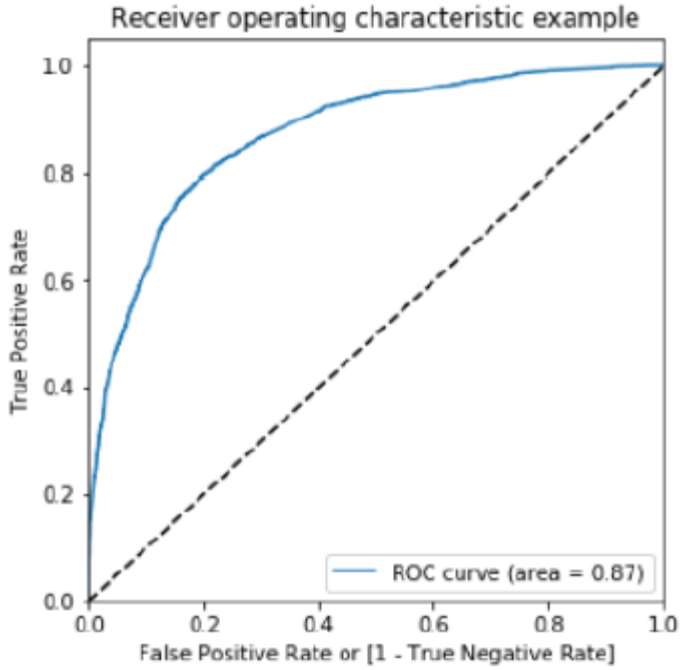


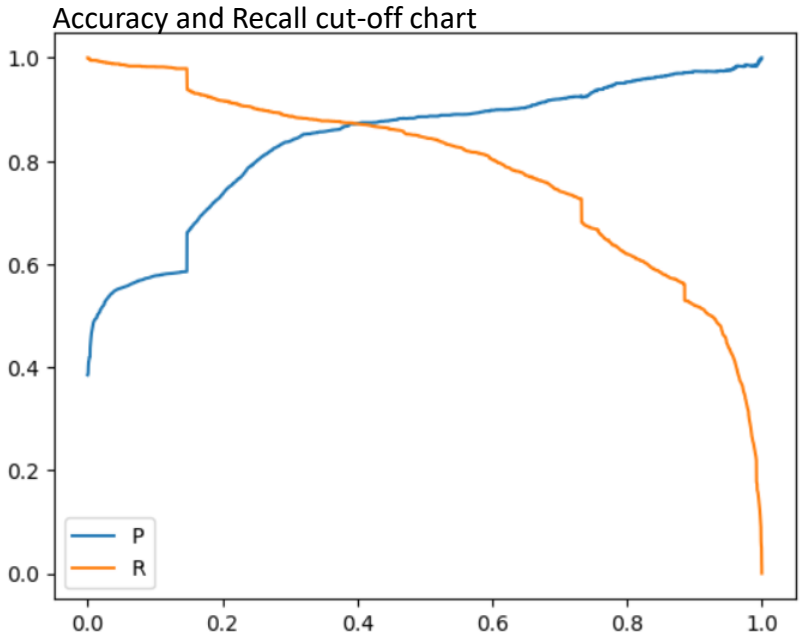


Heat Map

- Page Views Per Visits: Converted leads have a thicker spread although the median is same as non-converted leads.
- Time spent on Website: converted leads spend more time than non-converted leads. On an average the time spent is 15 mins which is 5 times than non-converted ones.
- High Positive correlation (0.75) is observed between “Page Views Per Visit” and “TotalVisits”.

ROC chart





- The cut-off is taken as 0.42 to get the optimal results.

Key Variables

- The variables considered in the lead scoring model:
  - What is your current occupation
  - Total time spent on the website
  - Tags (Closed by Horizzon, Ringing, etc)
  - Lead Activity (SMS Sent, Unsubscribed, etc)
  - Lead Origin (e.g., Lead Add Form, Lead Import, etc)
  - Email communication (Do Not Email)

Recommendations and Impacts

Since this model predicts the probability of a lead to converts so we can segment the leads to high-priority, mid-priority, low-priority and can focus on leads in the decreasing order of their priorities. This will:

- Increase conversion rates
- Optimize resource allocation
- Enhance sales efficiency