Lead Scoring Case Study

Introduction

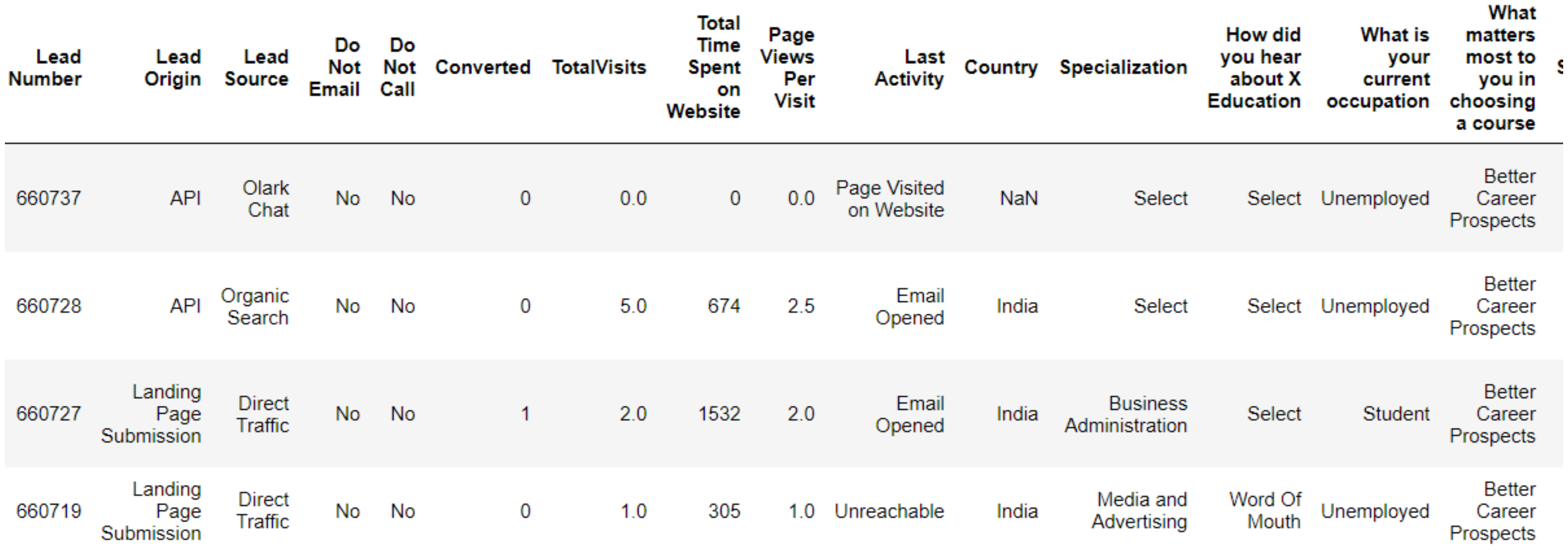
- **Problem Statement**

An education company named X Education sells online courses to industry professionals. Through marketing and past referrals, X education acquires the information of leads.
The aim of this case study is to identify 'hot leads' - those leads that are most likely to convert into paying customers, so that X education can target these promising leads to ensure a high conversion rate.

The end goal was to build a model assigns that a score between 0 and 100 for each of the leads, where a higher score is assigned to leads who are more likely to convert into paying customers, and vice versa

Analysis Approach

- **Understanding the data**



Lead Number	Lead Origin	Lead Source	Do Not Email	Do Not Call	Converted	TotalVisits	Total Time Spent on Website	Page Views Per Visit	Last Activity	Country	Specialization	How did you hear about X Education	What is your current occupation	What matters most to you in choosing a course
660737	API	Olark Chat	No	No	0	0.0	0	0.0	Page Visited on Website	NaN	Select	Select	Unemployed	Better Career Prospects
660728	API	Organic Search	No	No	0	5.0	674	2.5	Email Opened	India	Select	Select	Unemployed	Better Career Prospects
660727	Landing Page Submission	Direct Traffic	No	No	1	2.0	1532	2.0	Email Opened	India	Business Administration	Select	Student	Better Career Prospects
660719	Landing Page Submission	Direct Traffic	No	No	0	1.0	305	1.0	Unreachable	India	Media and Advertising	Word Of Mouth	Unemployed	Better Career Prospects

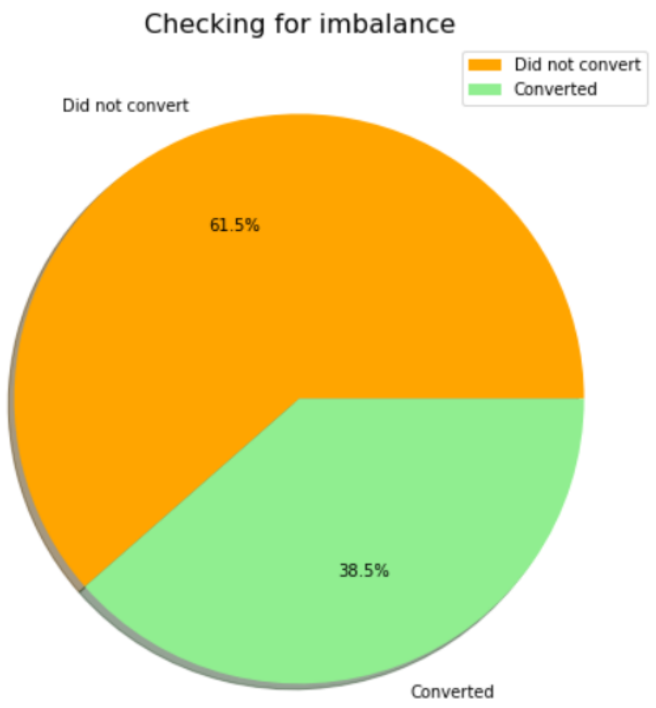
- The dataset provided contains the information of past leads : their last activity, occupation, etc. and also whether or not that lead has converted.

Analysis Approach – Data Cleaning

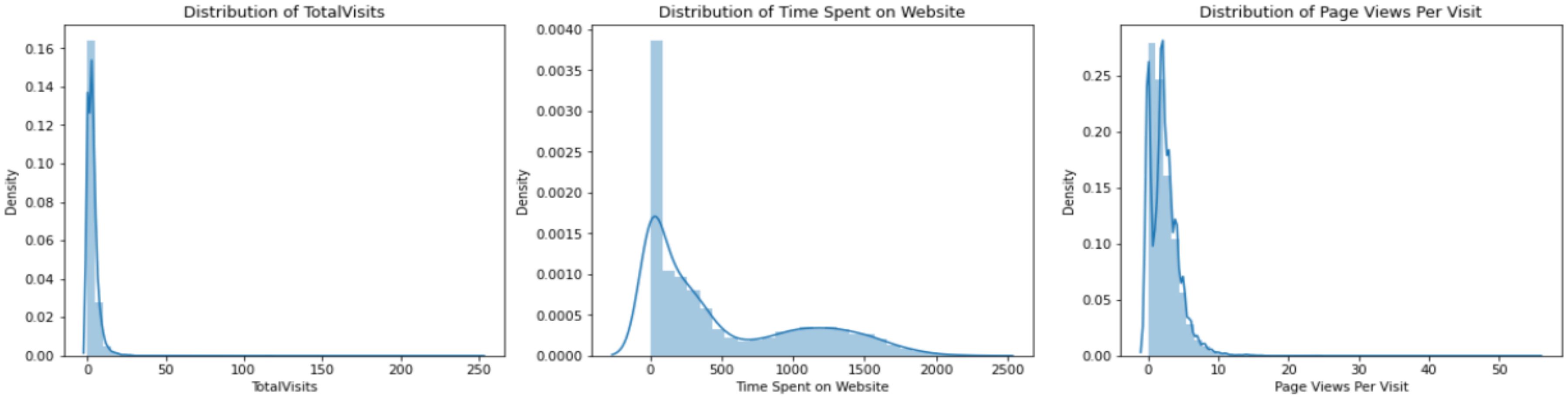
- **Data Cleaning – Steps involved**
  - Removed the columns that -
    - Had a very high null percentage (>40%)
    - That do not have enough variance (Have a single value for almost all the rows)
  - Performed imputation for-
    - Null values in Categorical columns with ‘Median’
    - Null values in Numerical columns with ‘Mode’
    - Null values in skewed columns with values like ‘Unknown’

Analysis Approach - EDA

- **Checking for imbalance in Target column**
  - We can see that the ratio of Converted and Non-Converted leads is almost 2:3.
  - Therefore the data is not skewed and there is no imbalance in the dataset.
  - As there is no imbalance we can go ahead with the next steps of EDA and Model building.

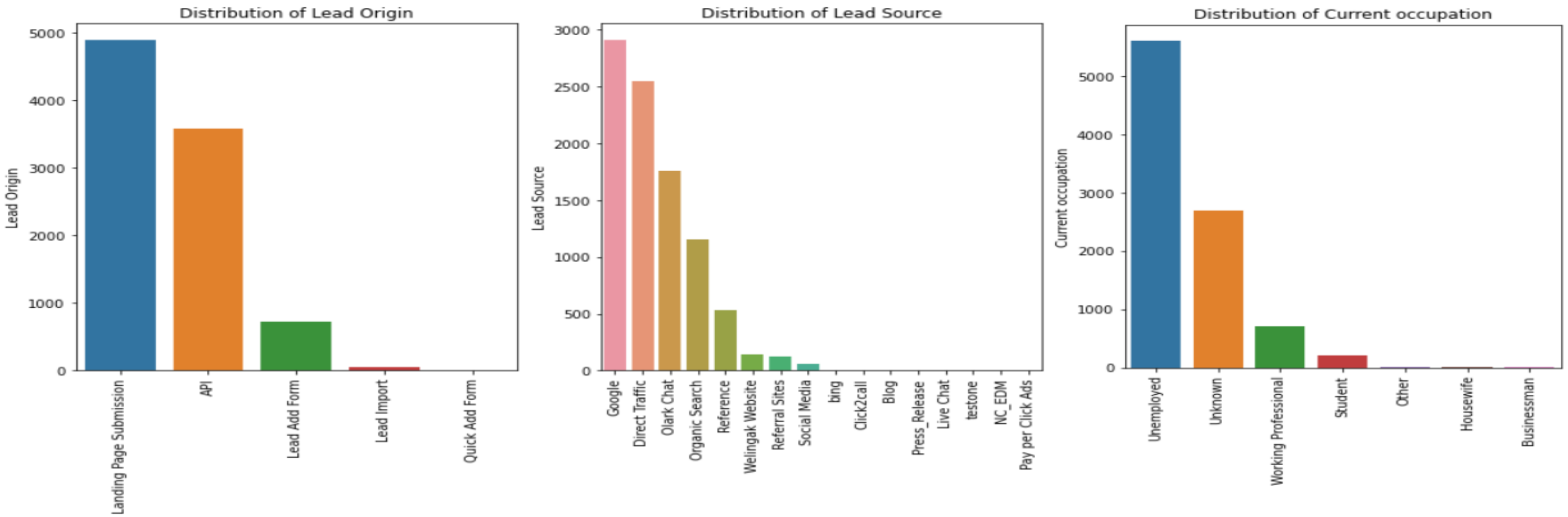


Analysis Approach - EDA



- **Univariate Analysis – Numerical Columns**
  - We can see that a majority of the users visit the website 0 to 50 times. However, there are some leads who visited the page almost 250 times.
  - The Time Spent on the Website column does not seem to be normally distributed. However, most of the users spend less time, and a small portion of the users spend around 1000 to 1500 minutes on the website.
  - Almost all the users viewed less than 10 pages per visit, with the exception of outliers who viewed more than 50 pages per visit

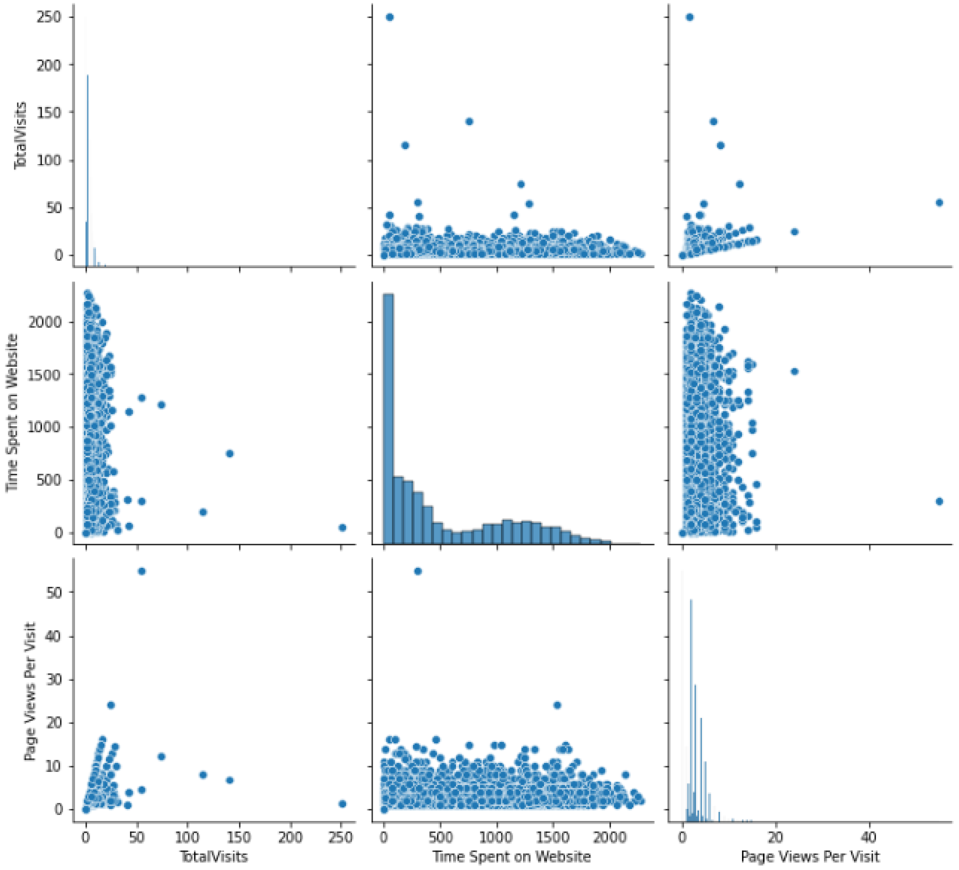
Analysis Approach - EDA



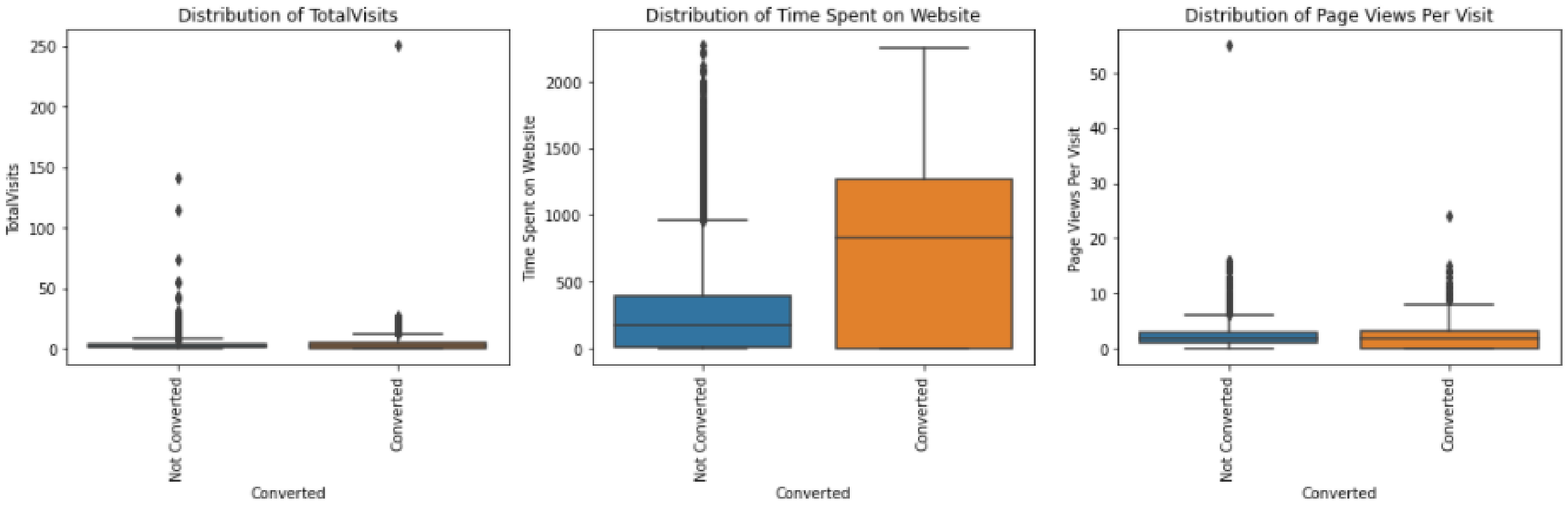
- **Univariate Analysis – Categorical columns**
  - Lead Origin: Over half of the leads that have originated through Landing page submission. Very few leads originated via Quick add forms.
  - Leads Source: Most of the leads have been sourced through Google, Direct traffic and Olark Chat.
  - Current Occupation: Over half of the users are Unemployed. Very few users are housewife and Businessman.

Analysis Approach - EDA

- **Bivariate Analysis – Numerical Vs Numerical Columns**
  - On an average, we can see that Total visits of user to the website is only 0-50 times, irrespective of the total time spent on the website.
  - We can also observe a similar, but weaker pattern between 'Time Spent on Website' and 'Page views per visit'. Most of the users viewed only 0-10 pages per visit, irrespective of the total time spent on the website.

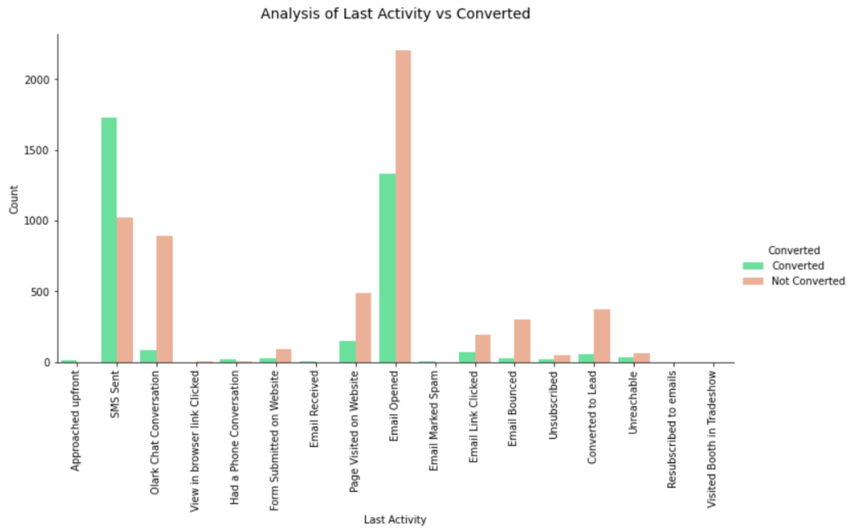


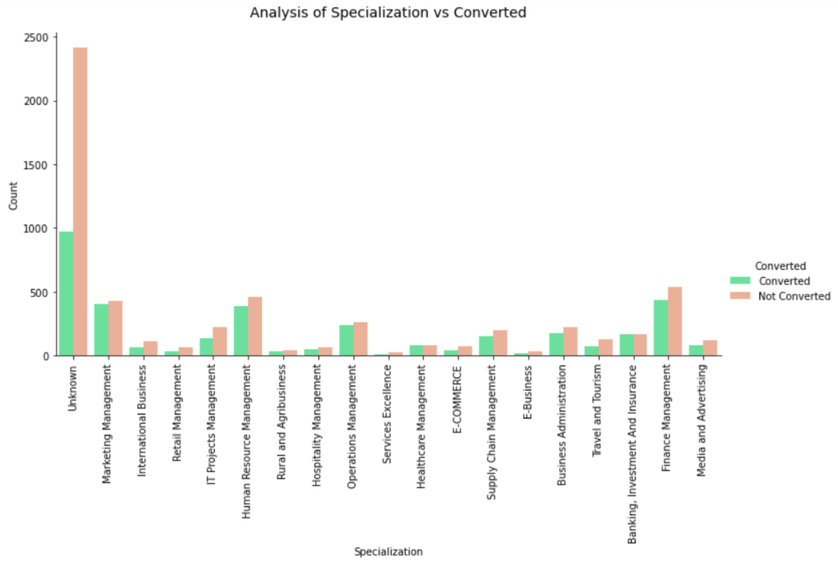
Analysis Approach - EDA



- **Bivariate Analysis – Numerical Vs Categorical Columns (Target)**
  - We can see that the a majority of the people who converted spend more time on website than those who do not.
  - Also, the distribution of time spent is more well-spread among converted users than non-converted users.

Analysis Approach - EDA





- **Bivariate Analysis – Categorical Vs. Categorical Columns (Target)**
  - The Leads whose 'Last Activity' is 'SMS sent' have an high conversion rate. On the other hand, The leads whose last activity is Olark Chat have an extremely low conversion rate.
  - Also, looking at the 'Specialization' column, we observe that the lead conversion rate is very bad for the leads whose specialization is Unknown. This is expected because people who are not interested to join would not have provided specialization details.

Data Preparation

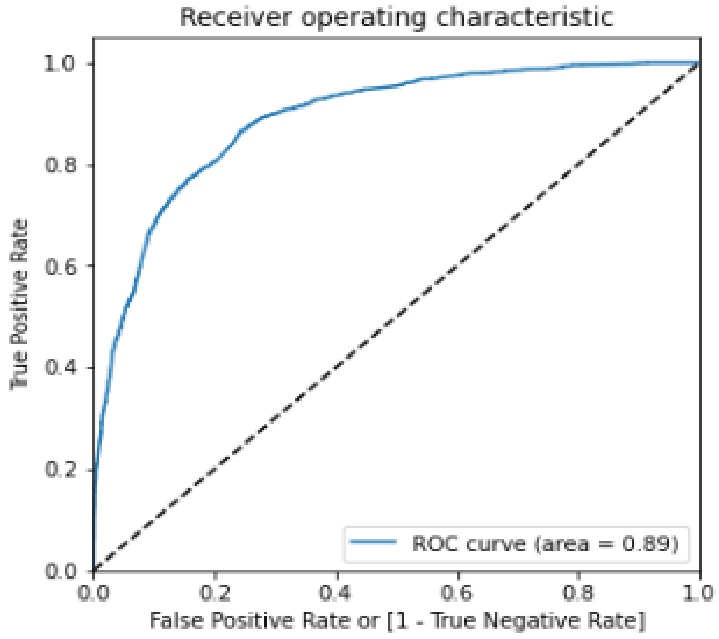
- **Data Preparation**
  - To build a Logistic Regression model, all the columns must be in numerical format.
  - We first mapped all the Yes/No variables to 1/0. Then we converted all Categorical columns into numeric format by using one hot encoding.
  - After conversion, we identified the highly correlated columns (> 0.75) and dropped them, as correlated features bring the same information to the dataset.
  - The dataset had 62 columns left after dropping.
  - We then split the data into train and test datasets with a 70:30 ratio. To first Train the Model on train dataset and test the model performance and metrics using test dataset.
  - After splitting, we performed feature scaling on Numerical columns using MinMaxScaler to standardize the independent features present in the data to a fixed range.

Model Building - RFE

- **Model Building on Train dataset**
  - After completion of data preparation steps, we proceeded to build the model on Train dataset by first reducing the number of columns from 62 to 25.
  - This was done using recursive feature elimination(RFE).
  - RFE is an approach used for eliminating features from a training dataset and retaining those features which contribute the most to the model building.
  - We then built a model using features selected by RFE, identified the columns having p-value >0.05 and VIF >5 and dropped them.
  - In this manner, we arrived at the final model with 17 columns having 80% accuracy when the cut-off probability is 0.5.
  - The model has a sensitivity of 0.698, and a specificity of 0.892. This implies that the model is currently predicting some people as 'not hot leads' when they are actually hot leads. In order to change this, we need to lower the cutoff value from its current value of 0.5.

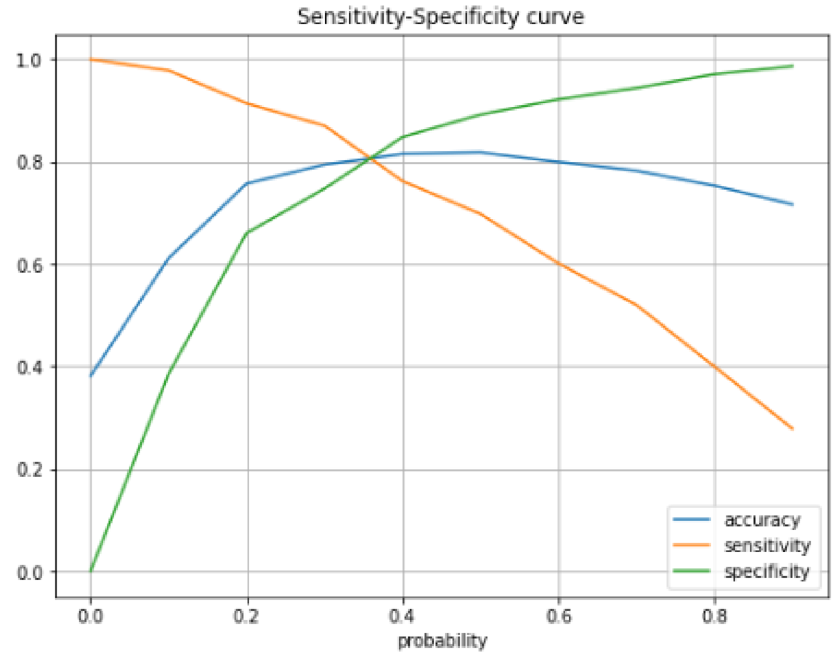
Model Building – ROC Curve

- **ROC Curve Interpretation:**
  - The closer the curve follows the left-hand border and then the top border of the ROC space, the more accurate the model becomes.
  - The Receiver Operating Characteristic (ROC) curve for the model is good, and has a high value of area under the curve - 0.89
  - This means the model's True Positive Rate is high, and the False Positive Rate is low, which is an indicator of a good model.
  - Therefore, we will use this as our model and we will now find the correct value of cut-off and calculate other metrics.



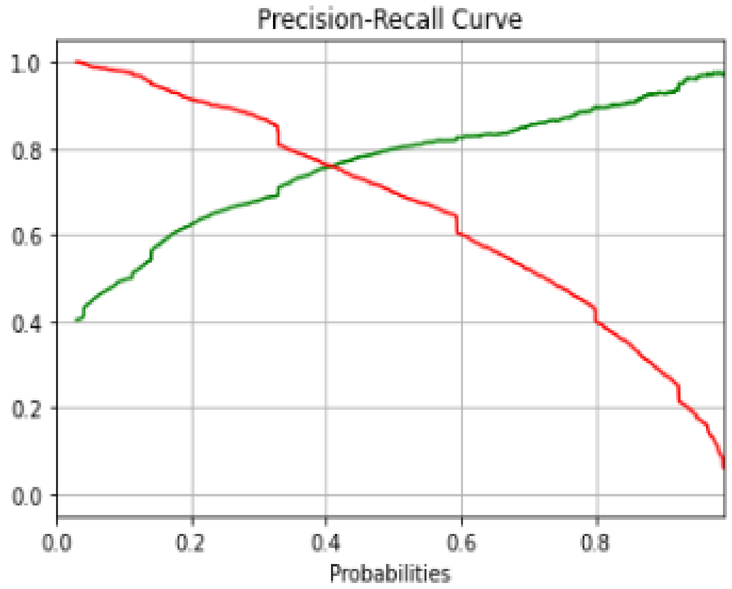
Model Building – Sensitivity-Specificity Curve

- **Sensitivity – Specificity – Accuracy Curve:**
  - To Plot this Curve, we need to create multiple columns with different probability cut-offs and calculate their sensitivity, specificity and accuracy.
  - We can then plot those details on a graph to obtain a cut-off point
  - The Specificity, Sensitivity and Accuracy curves intersect at probability = 0.35 which will be the cut-off value for our model.
  - After applying the cut-off, we see that the model is performing better.
  - The model has the specificity and sensitivity as 0.8, and has a low false-positive rate, which is an indicator that the model is performing well.
  - The model has 0.8 sensitivity, which means that the model is able to correctly predict 80% of the leads as 'Hot leads'



Model Building – Precision-Recall Curve

- **Precision – Recall Curve:**
  - After plotting the Precision – Recall curve, we see that the curves intersect at probability = 0.4.
  - However in alignment with the business requirements, we will choose the probability as 0.35, which has a slightly higher recall and a slightly lower precision.
  - All the metrics needed for a good model are already being met by using 0.35 as probability.
  - Therefore, we will be using 0.35 as the cut-off. This ensures that a majority of the 'Hot leads' are predicted correctly, and these can be targeted by X education.



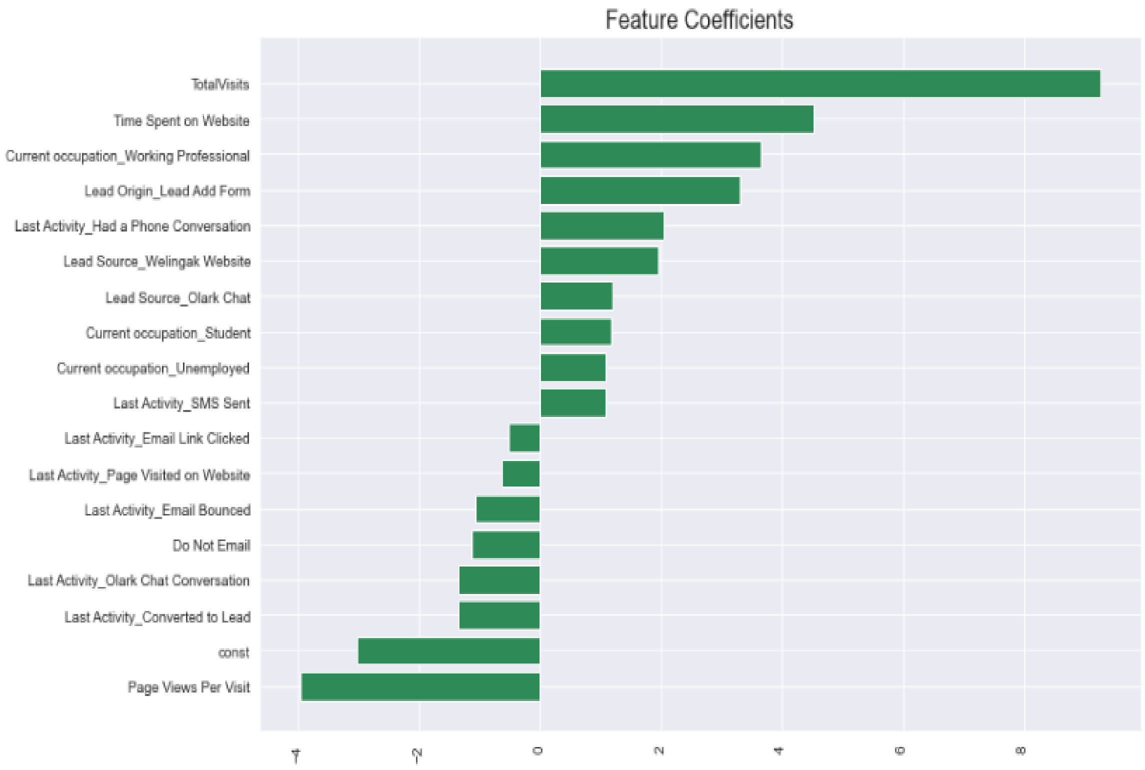
Model Evaluation – Predictions on Test set

- **Making Predictions**
  - After using the model for making predictions on Test set we observe that the Accuracy is 81.39 %
  - And sensitivity and specificity are close to 0.8.
  - As Observed there is little to no difference between the metrics of train and test dataset.
  - Hence, it is safe to say that the logistic regression model performs well on both train and test datasets.
  - Now, we can calculate the Lead Score for each Lead and add it to the initial dataframe for the X Education to make calls for leads having a high score.
  - Lead Score= Round(Conversion probability*100)

Feature Importance

- **Determining Feature Importance:**
  - The Graph on the right side represents the plotting of coefficient of each feature.
  - Higher the bar, more the feature is representing the model.
  - The Logistic Regression Equation is:

$$
\begin{aligned}
Ln\left(\frac{P}{1-P}\right) = & -3.02 + (9.28 \times \text{TotalVisits}) + (4.52 \times \text{Time Spent On Website}) - (3.97 \times \text{Page Views Per Visit}) \\
& + (3.66 \times \text{Current occupation\_Working Professional}) + (3.31 \times \text{Lead Origin\_Lead Add Form}) \\
& + (2.04 \times \text{Last Activity\_Had a Phone Conversation}) + (1.96 \times \text{Lead Source\_Welingak Website}) \\
& - (1.35 \times \text{Last Activity\_Converted to Lead}) - (1.35 \times \text{Last Activity\_Olark Chat Conversation}) \\
& + (1.21 \times \text{Lead Source\_Olark Chat}) + (1.18 \times \text{Current Occupation\_Student}) \\
& - (1.12 \times \text{Do Not Email}) + (1.1 \times \text{Current occupation\_Unemployed}) \\
& + (1.09 \times \text{Last Activity\_SMS Sent}) - (1.06 \times \text{Last Activity\_Email Bounced}) \\
& - (0.62 \times \text{Last Activity\_Page Visited on Website}) - (0.51 \times \text{Last Activity\_Email Link Clicked})
\end{aligned}
$$



Conclusion – Top 5 Features

- **Using Feature Importance we can determine the Top 5 important features that indicate a 'Hot Lead'**
  - **Total Visits -** This column has a Positive coefficient of 9.28. This means if a person visits the web page more, he is more likely to convert.
  - **Time Spent on Website -** This column has a Positive coefficient of 4.52. This implies that people who spend more time on the website are likely to convert.
  - **Page Views Per Visit -** This column has a Negative coefficient of -3.97. This indicates that as the person views more pages, he is less likely to convert.
  - **Current occupation: Working Professional -** This column has a Positive coefficient of 3.66. This implies that Working Professionals are more likely to convert than non working Professionals
  - **Lead Origin: Lead Add Form -** This column has a Positive coefficient of 3.31. This means that the conversion rate is high for the leads that originated via Lead Add Form.

Conclusion – Final Metrics

- **Model performance on Train Dataset:**
  - Accuracy : 80.71 %
  - Sensitivity : 0.794
  - Specificity : 0.815
  - Precision : 0.726
  - F1 Score: 0.758
- **Model performance on Test Dataset:**
  - Accuracy : 81.39 %
  - Sensitivity : 0.804
  - Specificity : 0.821
  - Precision : 0.745
  - F1 Score: 0.773

Conclusion – Next Steps for X Education

- In order to maximize the lead conversion, X education needs to focus on the leads with following patterns:
  - Visiting the website more frequently,
  - Spending more time on the website, but view less pages in each visit,
  - Is a working professional,
  - Originated via Lead Add Form.
- Also, it is observed that leads that are sourced through "Welingak Website" are more promising, so this website can be promoted to attract more 'Hot leads'.
- In case X education want almost all of the potential leads (i.e. the customers who have been predicted as 1 by the model) to be converted, they should also consider making calls to leads that have a score slightly less than 35.
- Similarly, if X education wants to limit the number of calls they make, they should consider only making calls to leads that have a score slightly higher than 35.
- Further, it is observed that the leads that are sourced through Olark Chat are likely to convert, but the leads whose last activity is 'Olark Chat Conversation' are not likely to convert. Therefore Olark chat should be used more as a way to source leads, than to nurture them. The leads that are likely to convert prefer other methods of contact, such as phone conversation, which is also indicated by the model.