LEAD SCORING CASE STUDY USING LOGISTICS REGRESSION

Submitted By:
1. Souradeep Mitra
2. Avinash Kumar
3. Tarun Kumar

Problem Statement

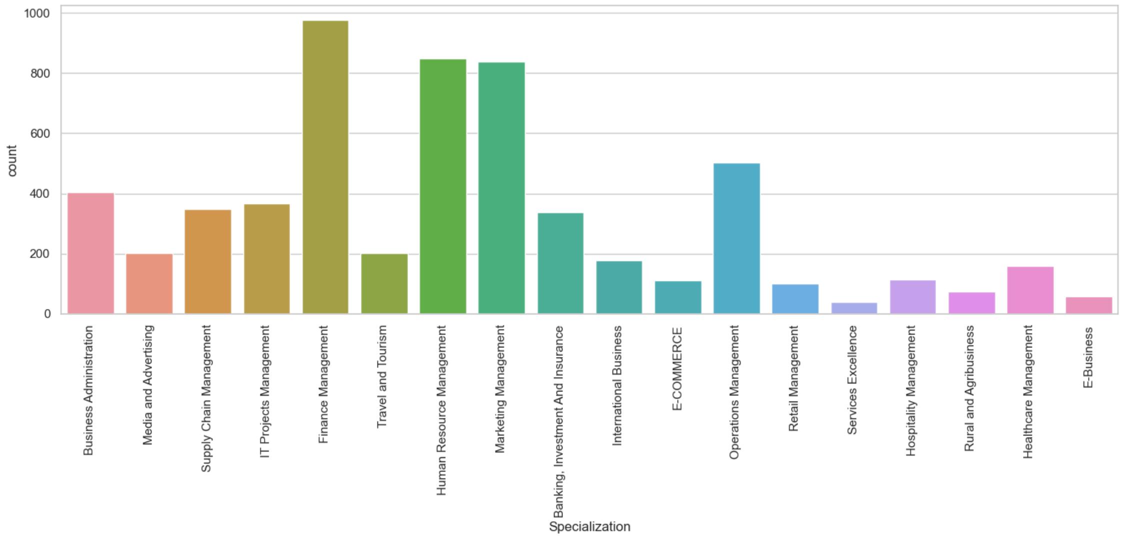
- An education company named X Education sells online courses to industry professionals. On any given day, many professionals who are interested in the courses land on their website and browse for courses. They have process of form filling on their website after which the company that individual as a lead. 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, this means 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

Business Objective

- Lead X wants us to build a model to give every lead a lead score between 0 -100 . So that they can identify the Hot leads and increase their conversion rate as well. The CEO want 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 manpower and after achieving target what should be the approaches.

Problem Approach

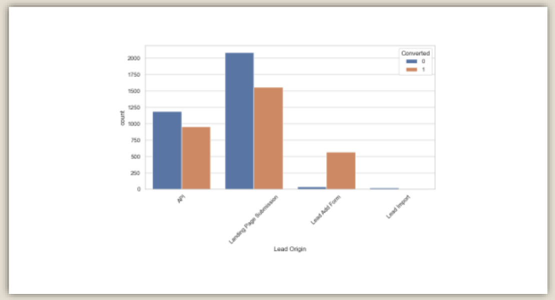
- Importing the data and inspecting the data frame
- Data preparation
- EDA
- Dummy variable creation
- Test-Train split
- Feature scaling
- Correlations
- Model Building (using RFE)
- Model Evaluation
- Making predictions on test set

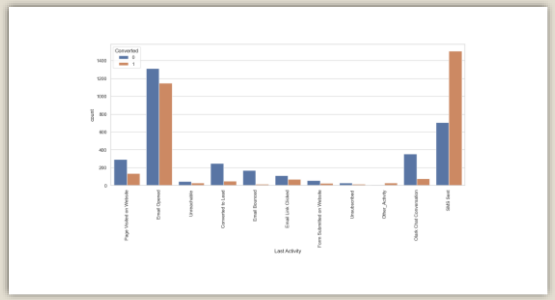


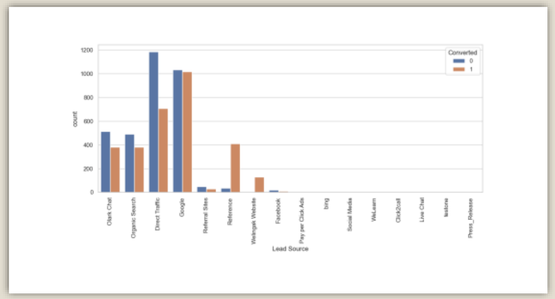
Data Cleaning

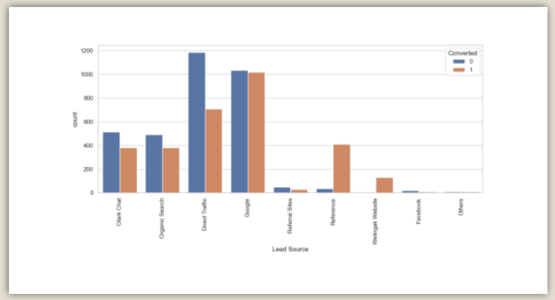
- It looks like that there are quite a few categorical variables present in this dataset for which we will need to create dummy variables. Also, there are a lot of missing values present as well, so we will need to process the dataset and clean the same.
- As it is clearly seen that there are a lot of columns which have a very high number of missing values, these columns are not thus seemingly useful. Since, there are 9000 datapoints in our data frame, we choose to drop the columns having greater than 3000 missing values.
- Highest number of leads from INDIA, so we can drop the country columns and we can also drop the city column as it is of no use to us for this analysis
- We observe that there are 'Select' values in many columns. It may be because the customer did not select any option from the list, hence it shows 'Select’. 'Select' values are as good as NULL. So, we choose to convert these values to null values.
- The 37% missing values in Specialization may be because the lead has not filled this specific column or is not working, so we can replace the missing values as "others" column instead.

EDA

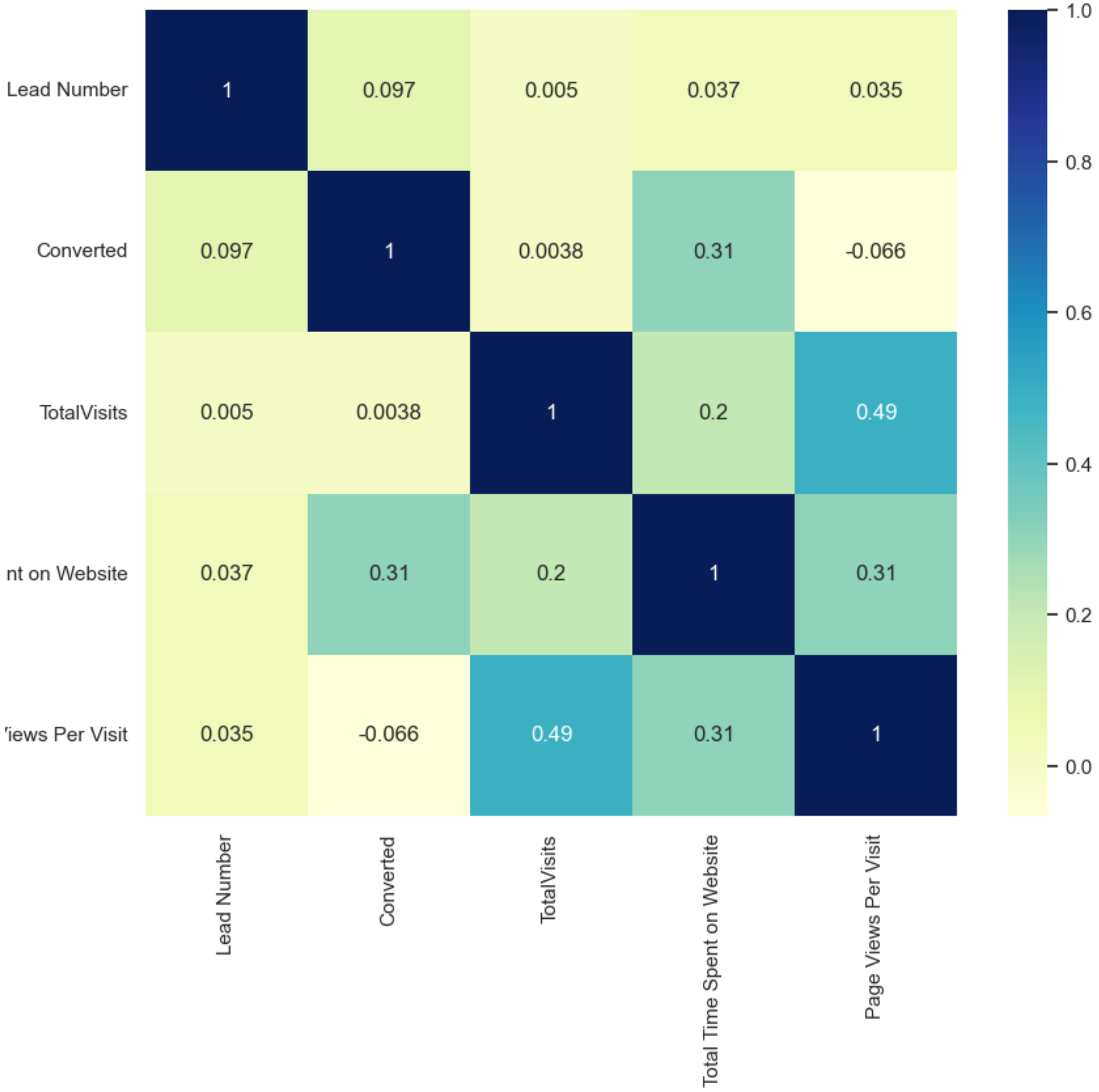




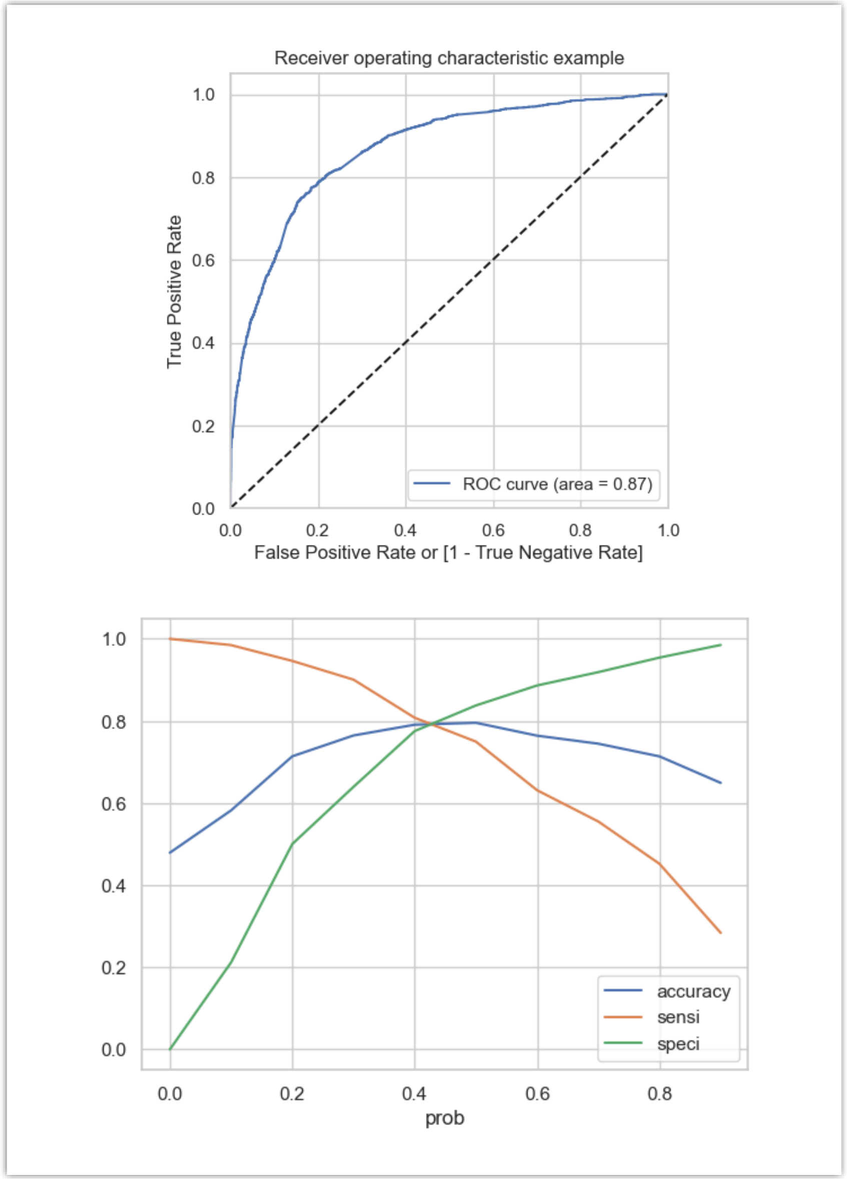




- API and Landing Page Submission have 30-35% conversion rate and count of the total number of lead originating from them are considerable.
- Lead Add Form has more than 90% conversion rate but the count of the total number of lead originating are not very high. Lead Import are very less in count.
- To improve overall lead conversion rate, we need to focus more on improving lead conversion of API and Landing Page Submission origin and generate more leads from Lead Add Form.
- Maximum converted leads are generated by Google, followed by direct traffic.
- Conversion Rate of reference and welingak website is high.
- 'To improve overall lead conversion rate, focus should be on improving lead conversion of olark chat, organic search, direct traffic, and google leads and generate more leads from reference and welingak website.'
- Conversion rate for leads with SMS Sent is the highest.
- Leads with last activity as email opened have major conversion rates.



- We see that no such high correlation is available for the variables in the dataset.
- Now we move on to model building and training the dataset to make the final predictions.

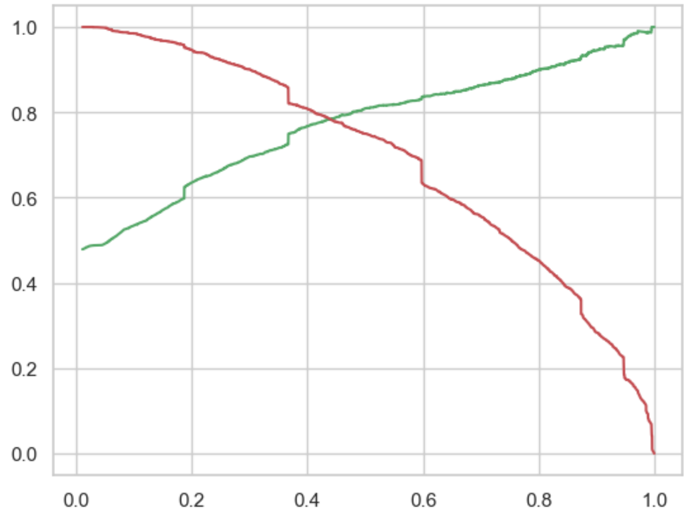


Model Evaluation

- We found out that our specificity was good (~83%) but our sensitivity was only 74%. Hence, this is needed to be taken care of. Now 0.5 was just arbitrary to loosely check the model performance. But to get good results, we need to optimize the threshold by plotting the ROC curve.
- The ROC is found to be 0.87 .
- Once the trade-off between accuracy, sensitivity, and specificity was plotted, it helped us find the best cutoff value to train the model to predict conversions.
- The Cutoff was chosen to be 0.42.

Model Evaluation on Test Set

- With the cutoff value of 0.42, we made predictions on the test set.
- We got a precision of 0.81 and a recall of 0.75
- Then on plotting the precision-recall trade-off we find an optimal cutoff value of 0.44 as show in the graph here.
- On generating the confusion matrix base on the cutoff value of 0.44 we find that the prediction made on the test set finally had a Precision of 78% and a Recall of 78%



Final Observation

Train Set Statistics

Precision : 78%

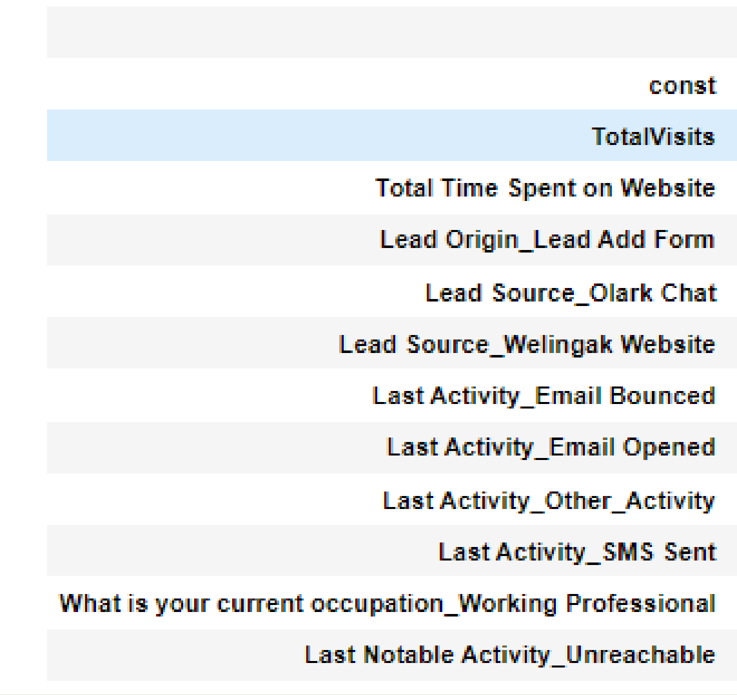
Recall : 80%

Test Set Statistics

Precision: 78%

Recall : 78%

- **Final Features List**



const
TotalVisits
Total Time Spent on Website
Lead Origin_Lead Add Form
Lead Source_Olark Chat
Lead Source_Welingak Website
Last Activity_Email Bounced
Last Activity_Email Opened
Last Activity_Other_Activity
Last Activity_SMS Sent
What is your current occupation_Working Professional
Last Notable Activity_Unreachable

Conclusion

Considering the outcome of the most potential variables provided by the model they should focus on implementing what the model has suggested like Time spent on site, total visits, working professionals etc.

Start sending SMS and making calls repetitively.

Try to get more familiar with the leads via the phone calls, discussing their requirements, suggestive background, and providing them with the required financial information.

Based on the model, it is suggested:
1. Not to focus on students and unemployed leads.
2. Avoid sending more updated via email to those who have their emails bounced.