Lead Score Case Study

Problem Statement

- The company wants to classify which of the potential leads will eventually convert
- To achieve this, company wants to understand the key drivers that influence the conversion from the data set provided, with the history of past conversions
- We need to come up with recommendations so that the company focusses on the right set of drivers that lead to a higher conversion rate
- Recommendations should also include strategies for aggressive and moderate approaches to reaching out via phone to potential leads based on availability of interns in the company

Assumptions

- Value of “Select” has been treated as null. It has been assumed that the value “Select” indicates that the user has not selected any value for that column

Approach

- Columns with more than 35% Null values were dropped
- The rest of the remaining missing values have been imputed with the most frequently occurring value for categorical variables
- Outliers were detected. A subset of the outliers that were very extreme among the outliers, were removed so that they do not skew the analysis. These extreme values did not add any special value to the dataset
- There were many columns with spaces in their names. So column name transformations were done for ease of operation.
- EDA was done to gain some insights on the data

Approach

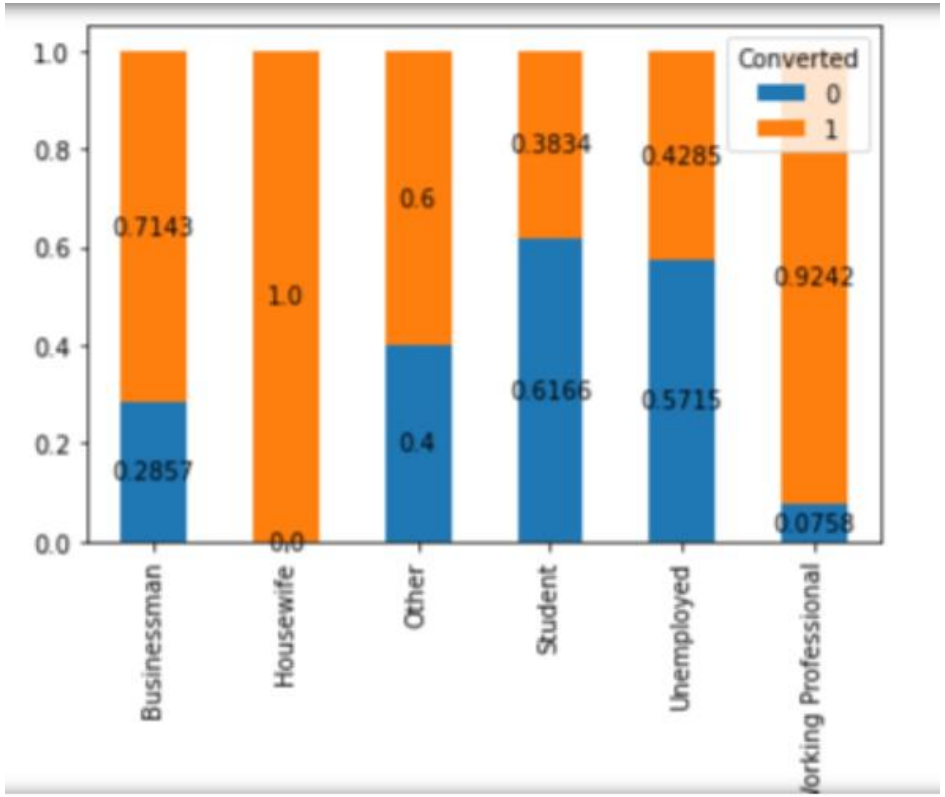
- Some obvious columns like 'Prospect ID’ and 'Lead Number’ were dropped from analysis
- Label encoding was used for categorical variables with large number of values, like 'Lead Source', 'Last Activity’ , 'Country’ and 'Last_NotableActivity’
- One-hot encoding was done for categorical variables like 'Lead Origin', 'Current_Occupation’ and 'Course_Intent’ that had less number of values
- Correlation matrix was generated and some highly correlated variables like 'Page Views Per Visit’, 'Lead Origin_Lead Add Form’, 'Last_NotableActivity’ and 'Current_Occupation_Unemployed’ were dropped from analysis
- Standard scaling was performed for continuous variables

Approach

- RFE was used to start the first model with 15 columns
- Subsequently models were built and columns were removed from the model, one at a time, which had high p-values
- All the columns had good VIF values for all the models

Analysis - Results

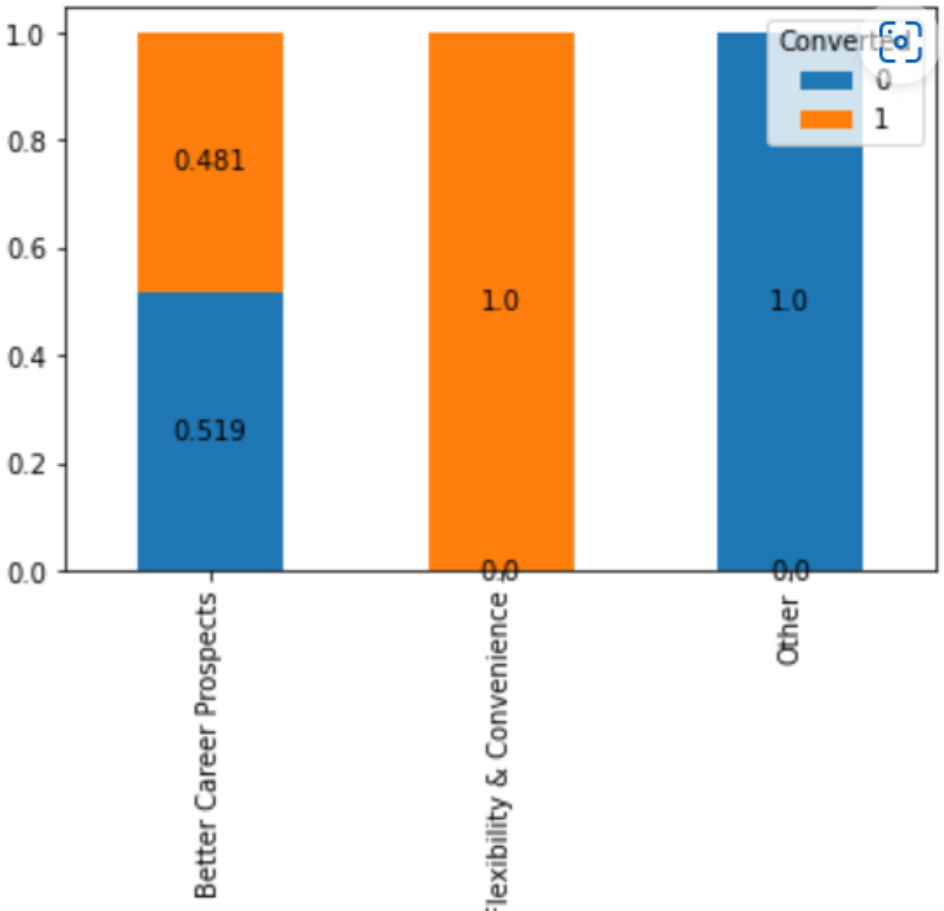
- The data set was fairly balanced with almost an even split between the data for converted (48%) and Non-converted(52%)
- Based on EDA some observations could be made:



- All the leads that were housewife ended up getting converted
- Working professionals had a high conversion rate as well

Analysis - Results

- Based on EDA some additional observations could be made:



- It turned out that leads who had mentioned that they selected course for Flexibility & Convenience ended up with 100% conversion
- Leads that did not give specific reason (chose Other) had 0% conversion

Analysis - Results

- Based on EDA and the information on Countries available in the data set, certain countries stood out
  - For lead conversions: 100% of leads from Bangladesh and Denmark get converted. More work could be done on Bahrain (80%), France (60%) and Hong Kong(66.67%) to get more conversions
  - For non conversions: None of the leads from Canada, China, Ghana, Italy, Kenya, Kuwait, Malaysia, Nigeria, Philippines, Russia, South Africa, Sri Lanka, Tanzania, Uganda and Vietnam get converted

- From the model, we could see that
  - The optimal cut-off probability was 0.4
  - The overall accuracy of predictions on training data set and test data set are 73.81 and 72.27. Similarly, the sensitivity values for predictions on training data set and test data set are 75.56 and 75.87. So the model is highly stable as well.

Analysis - Results

- From the model we could see that
  - The features Current_Occupation_Working Professional, Dont_email, Total Time Spent on Website, Lead Origin_Landing Page Submission and Lead Source are the top drivers of this model.

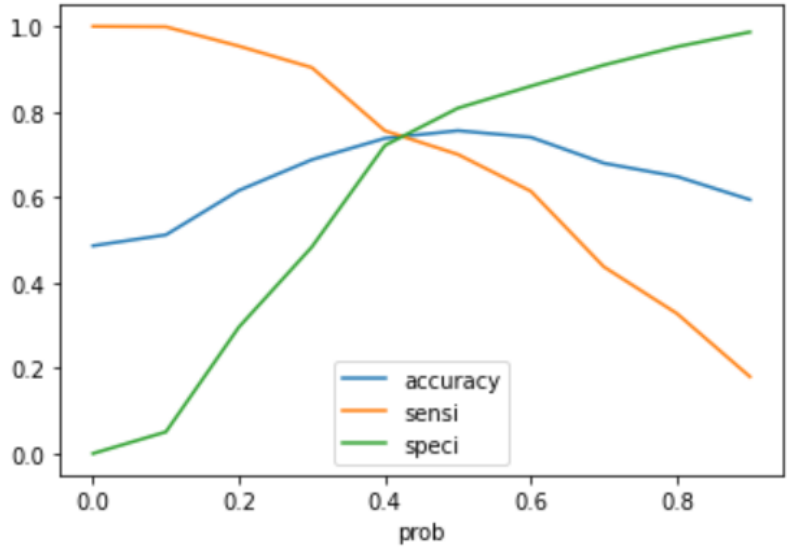
	coef	std err	z	P>|z|	[0.025	0.975]
const	-0.8957	0.101	-8.904	0	-1.093	-0.699
Current_Occupation_Working Professional	2.6562	0.186	14.251	0	2.291	3.022
Dont_email	-1.2383	0.174	-7.112	0	-1.58	-0.897
Total Time Spent on Website	1.0005	0.042	23.861	0	0.918	1.083
Lead Origin_Landing Page Submission	-0.4448	0.086	-5.163	0	-0.614	-0.276
Lead Source	0.2547	0.017	15.329	0	0.222	0.287

Key recommendations

- Focus on
  - working professionals followed by Housewife
  - leads who have spent good time with the web site. The web site could probably be improved so that people who come in engage more with the website and gain more insights about the various courses offered
  - Leads from Bangladesh and Denmark since as per EDA these countries had 100% lead conversion
  - leads who had mentioned that they selected course for Flexibility & Convenience
- Do not spend time and resources on leads
  - that have not chosen to be emailed
  - for which the lead originated from Landing Page Submission
  - that did not give specific reason, chose others as reason for choosing the course

Key recommendations

- Aggressive and moderate strategies could be drawn out based on the availability of interns, as needed by the company



- Aggressive strategy: Get a larger pool of potential leads by choosing a cut-off probability which is lower than the optimal cut-off probability

- Moderate strategy: Get a smaller pool of potential leads by choosing a cut-off probability which is lower than the optimal cut-off probability

Thank you!