Lead Scoring Case Study

SUBMITTED BY: SNEHA CHATTERJEE, RAVI VERMA, MAROOF SAYED

Table of Content

Problem Statement

- Even though X education get lots of lead but poor conversion rate of 30%
- X education wants to make the process of lead conversion efficient, by focusing only on potential lead (also called Hot Leads)
- Sales team of X education wants to focus on hot lead, instead of making useless phone calls.

Analysis Approach

- Data Preparation: encoding categorical variables and handling null values
- EDA
- Dummy Variable Creation
- Test Train Split
- Feature Selection
- Correlation Analysis
- Model Building: Feature selection manually, Model Improvisation
- Model Finalization
- Model Evaluation
- Recommendation

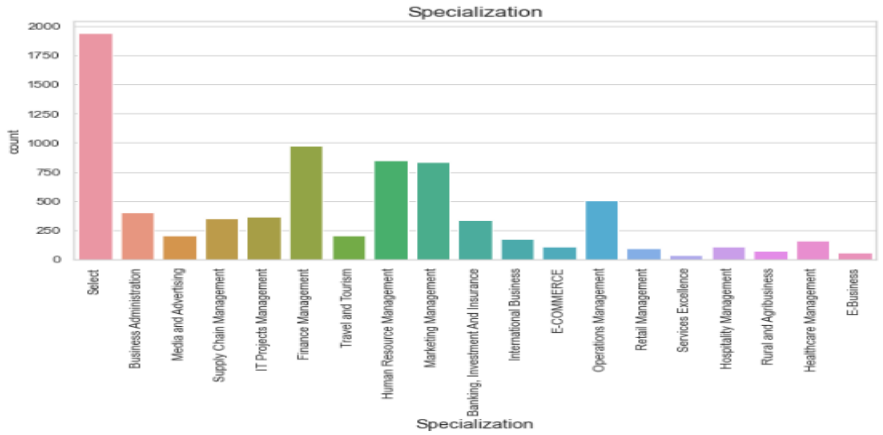
Since, we have target of 80% conversion rate or better, we would want obtain a high sensitivity of obtaining hot leads. Its important for our model to predict the potential lead that can be converted.

Data Preparation

- Checking missing values in each column. we eliminate the columns having greater than 3000 missing values as they are of no use to us.
- The variable City, Country is of no use in our analysis. Hence we drop it
- Checking the percentage of missing values in each column
- We notice that, when we got the value counts of all the columns, there were a few columns in which only one value was majorly present for all the data points. These include Do Not Call, Search, Magazine, Newspaper Article, X Education Forums, Newspaper, Digital Advertisement,. Since all of the values for these variables are No, it is best that we drop these columns as they won't help in our analysis.
- Dropping the null values rows in the column 'What is your current occupation',’Lead sorce’, ‘total Visit’
- Dropping the "Prospect ID" and "Lead Number"

EDA: Exploratory Data Analysis

- Analysing Categorical variables, Visualizing the features
- It can be seen that the levels of "Lead Profile" and "How did you hear about X Education" have a lot of rows which have the value 'Select' which is of no use to the analysis



EDA:

- We notice that, when we got the value counts of all the columns, there were a few columns in which only one value was majorly present for all the data points. These include Do Not Call, Search, Magazine, Newspaper Article, X Education Forums, Newspaper, Digital Advertisement, Through Recommendations, Receive More Updates About Our Courses, Update me on Supply Chain Content, Get updates on DM Content, I agree to pay the amount through cheque. Since all of the values for these variables are No, it is best that we drop these columns as they won't help in our analysis.
- The variable 'What matters most to you in choosing a course' has the level 'Better Career Prospects' 6528 times while the other two levels appear once twice and once respectively.

```
1 leads['What matters most to you in choosing a course'].value_counts()
```

```
Better Career Prospects      6528
Flexibility & Convenience       2
Other                           1
Name: What matters most to you in choosing a course, dtype: int64
```

EDA:

- Observing Correlation using heatmap
- We create dummy variable separately for the variable 'Specialization' since it has the level 'Select' which is useless so we drop that level by specifying it explicitly.

```
dummy_spl = pd.get_dummies(leads['Specialization'], prefix = 'Specialization')
dummy_spl = dummy_spl.drop(['Specialization_Select'], 1)
leads = pd.concat([leads, dummy_spl], axis = 1)
```

- Splitting the dataset into 70% train and 30% test
- Scaling the 3 numeric features

Model Building

- **Feature Selection Using RFA**
  - This will reduce the model performance and might take high computation time
  - Its important to perform Recurive Feature Elimination (RFE) and to select only the important columns
  - We will create Logistic Regression using statsmodels by checking VIF and p values
  - Dropping Lead Source_Reference since it has a high p-value as well as a high VIF.(Model 1)
  - VIFs are less than 5. Dropping 'Last Notable Activity_Had a Phone Conversation' as it has high p value (Model 2)
  - Dropping 'What is your current occupation_Housewife' as it has high p value(Model 3)
  - Dropping 'What is your current occupation_Working Professional' as it has high p value (Model 4)
  - Since VIFs are less than 5 and all p values are less than 0.05, we accept this as final model.(Model 5)

Model Evaluation

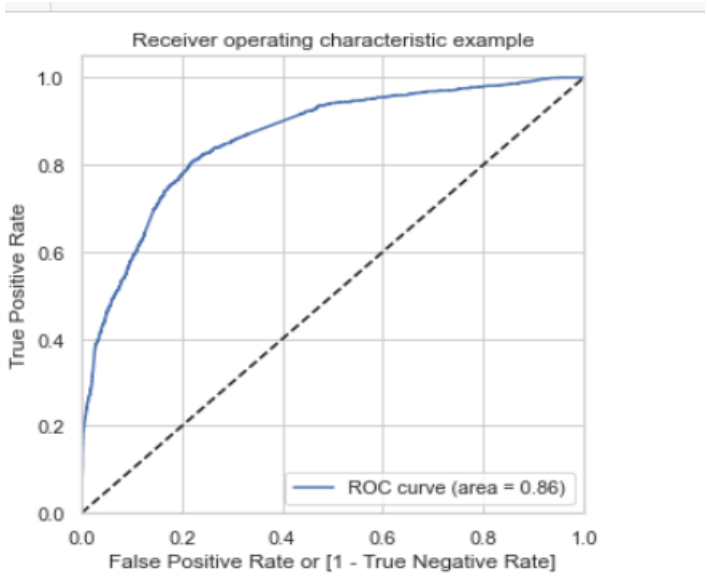
- Creating df with actual flag and conversion probabilities
- Creating new column with 1 if Conversion prob > 0.5 else 0

```
1 y_train_pred_final['Predicted'] = y_train_pred_final.Conversion_Prob.map(lambda x: 1 if x > 0.5 else 0)
2
3 # Let's see the head
4 y_train_pred_final.head()
```

	Converted	Conversion_Prob	Predicted
0	0	0.300117	0
1	0	0.142002	0
2	1	0.127629	0
3	1	0.291558	0
4	1	0.954795	1

Model Evaluation

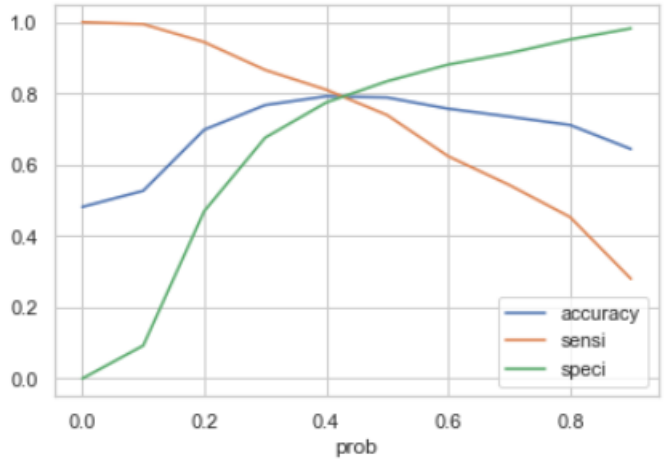
ROC function



Area under ROC is 0.86 which is quite good. Hence we conclude that it is a good model

Model Evaluation

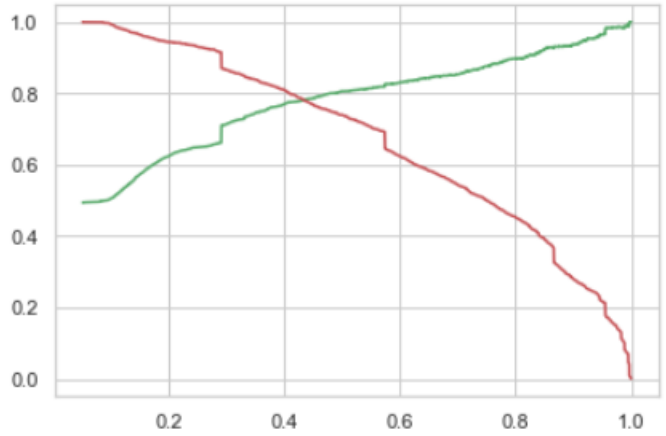
creating columns with different probability cutoffs .



We see that around 0.42 we get optimal values of the 3 metrics. Hence we choose 0.42 as our cutoff.

Model Evaluation

Precision and recall tradeoff:



From above plot, we can take 0.44 as the cutoff. Making predictions on the test set using 0.44 as the cutoff.

Recommendation Based on Final Model

By understanding the needs and wants of our target customer, we can better qualify them as a potential buyer. This will save time and money in the long run. Implementing a lead scoring system is an essential part of any sales process, so make sure to put one in place if we want to see results.

- Website could be made more attractive, with fun-quality content and interactive component might to increase the time spent by users on website.
- Offer a stronger value proposition. Make sure your message is clear and compelling.
- Focus on the lead having high potential based on the lead score.
- Leads who had last activity as “SMS sent” can be focused as they have higher conversion rate.
- Include prominent calls-to-action and contact forms on the website so people can reach out with any questions or concerns.
- Use A/B testing to compare different versions side-by-side and see which one performs better.

THANK YOU