Lead Scoring Case study to identify Potential leads for X-educaton company to sell its Online courses

09-Feb-2022

by Lakshmikar and Manohar

Problem Statement

An education company named X Education sells online courses to industry professionals. The company markets its courses on several websites and search engines like Google. Once these people land on the website, they might browse the courses or fill up a form for the course or watch some videos. When these people fill up a form providing their email address or phone number, they are classified to be a Assignment lead. Now, although X Education gets a lot of leads, its lead conversion rate is very poor. For example, if, say, they acquire 100 leads in a day, only about 30 of them are converted.

What we need to do?

● To help X-education select the most promising leads, i.e. the leads that are most likely to convert into paying customers.

● We need to build a model wherein we need to assign a lead score to each of the leads such that the customers with higher lead score have a higher conversion chance and the customers with lower lead score have a lower conversion chance.

Approach for this Case study

Steps:

Data Cleaning

- Understand the dataset by inspecting it and referring to provided data dictionary
- Look for missing values. If missing value count is more than 40%, then drop those columns(variables). Otherwise, missing values will be imputed accordingly with Mean/Media/Mode.
- `Select‘ category should also be considered as missing value.
- Check for Outliers and deal with them accordingly.
- To drop the Sales team generated variables as these are ot required for Model building.

Data Preparation

- Perform EDA.
- Create dummies for all categorical columns.
- Perform train-test split.
- Perform scaling.

Approach for this Case study

Steps:

Modeling

- Use techniques like RFE to perform variable selection.
- Build a Logistic Regression model with good sensitivity.

Assignment

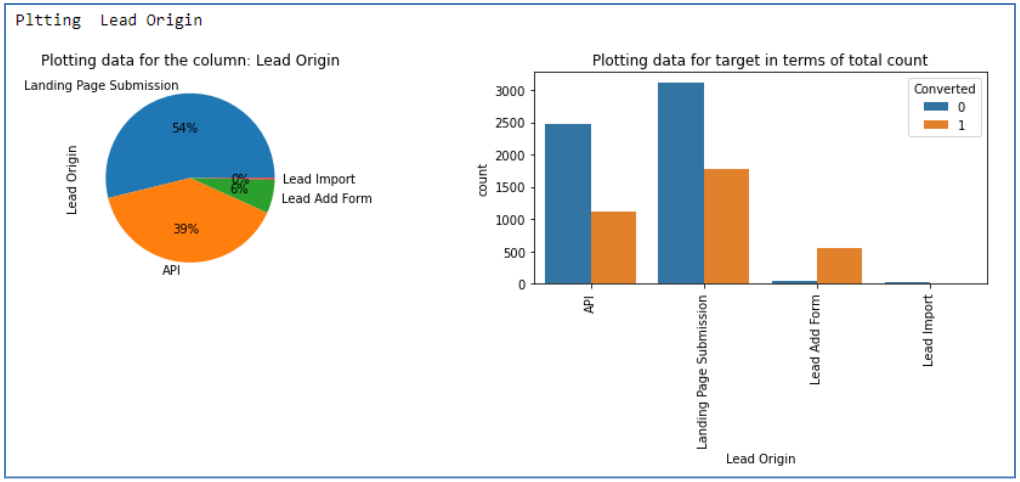
- Check p-value and VIF.
- Find the optimal probability cutoff.
- Check the model performance over the test data.
- Generate the score variable.

Results of Case study analysis

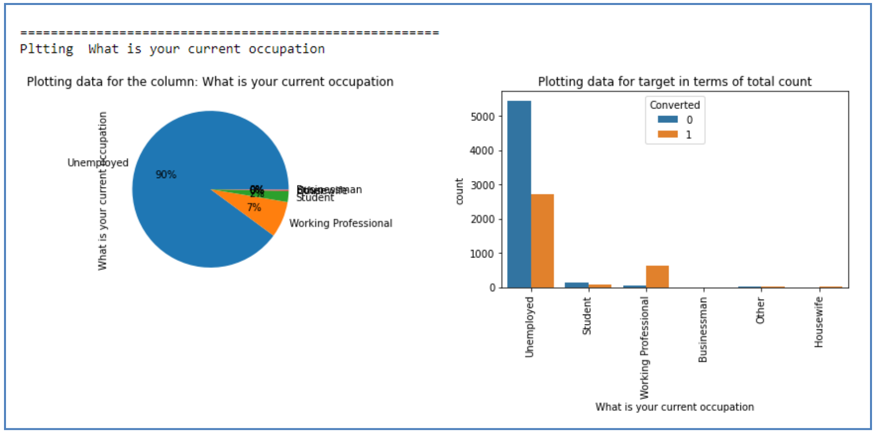
Data cleaning:

Given dataset size is 9240 rows x 37 columns. After performing Data cleaning steps, we have arrived at 9074 rows x 10 columns

Univariate analysis

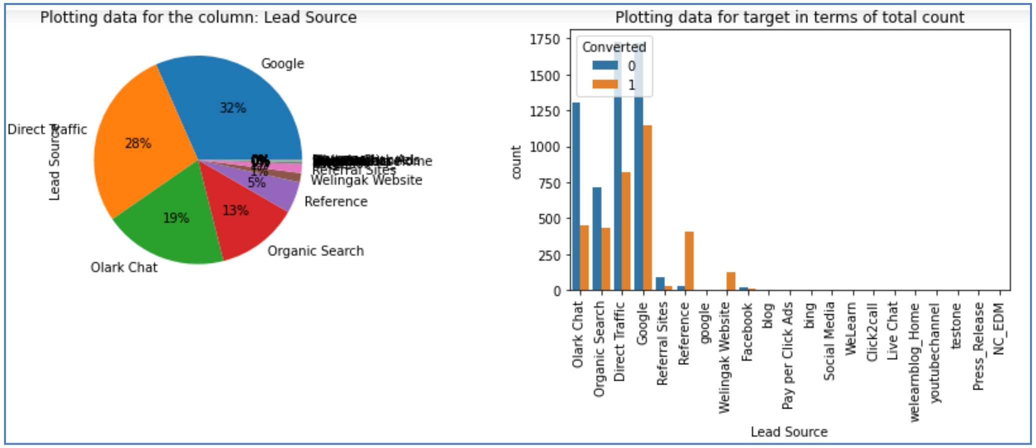


```
Pltting  Lead Origin
```



```
=======================================================
Pltting  What is your current occupation
```

Results of Case study analysis



After performing univariate analysis on these variables, below are the observations

•Lead Origin variable: Most leads have identified from Landing Page submission but Lead add forum has more percentage of converting Leads to Successful leads.
•Lead Source variable: Most leads have sourced from Google and Direct Traffic. Google has more potential to convert into successful leads
•Current Occupation variable: Mostly unemployed have visited but Working professionals are mostly converting to successful leads
•Total visits and Page views per visit have positive correlation

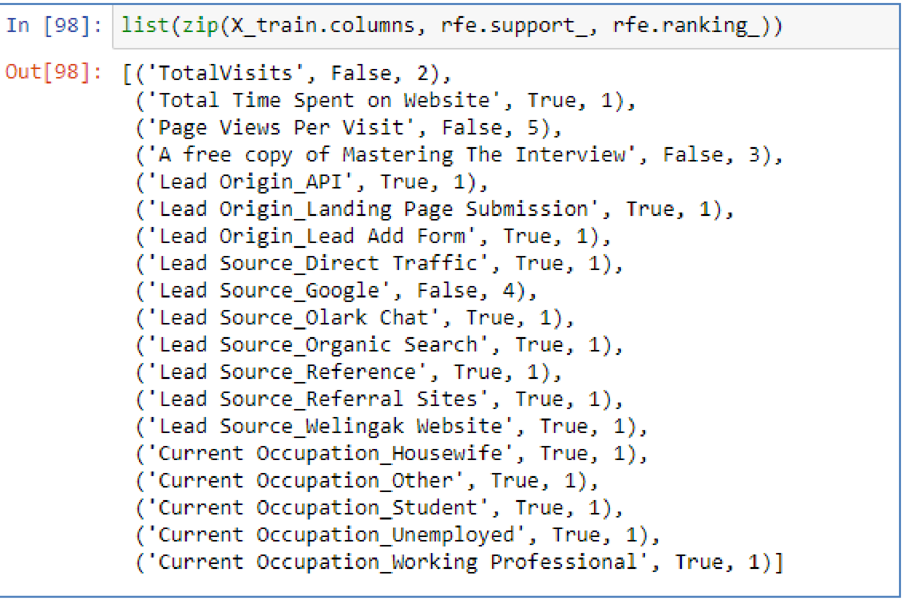
Creating Dummies

After creating dummy variables for all the categorical columns, we have arrived at 9074 rows x 22 columns

Results of Case study analysis

Model building:

Using RFE, 15 top features are selected. Among these, below festures with rfe_raning as 1 are important for model building



```
In [98]:  list(zip(X_train.columns, rfe.support_, rfe.ranking_))

Out[98]:  [('TotalVisits', False, 2),
           ('Total Time Spent on Website', True, 1),
           ('Page Views Per Visit', False, 5),
           ('A free copy of Mastering The Interview', False, 3),
           ('Lead Origin_API', True, 1),
           ('Lead Origin_Landing Page Submission', True, 1),
           ('Lead Origin_Lead Add Form', True, 1),
           ('Lead Source_Direct Traffic', True, 1),
           ('Lead Source_Google', False, 4),
           ('Lead Source_Olark Chat', True, 1),
           ('Lead Source_Organic Search', True, 1),
           ('Lead Source_Reference', True, 1),
           ('Lead Source_Referral Sites', True, 1),
           ('Lead Source_Welingak Website', True, 1),
           ('Current Occupation_Housewife', True, 1),
           ('Current Occupation_Other', True, 1),
           ('Current Occupation_Student', True, 1),
           ('Current Occupation_Unemployed', True, 1),
           ('Current Occupation_Working Professional', True, 1)]
```

We have built the model with Logistc Regression and taken a random cut-off value of 0.5.
Then, we obtained:
Accuracy: 78%
Sensitivity: 62%
Specificity: 88%

Results of Case study analysis

Checking VIFs

We checked VIFs to determine the effect of multicolliniarity and to eliminate redundant variables to make the simpler model.

After performing this, we could drop below variables due to high VIF values

1. Current Occupation_Unemployed
2. Lead Origin_Lead Add Form

For this final model also, we obtained below values:

Accuracy: 78%

Sensitivity: 62%

Specificity: 88

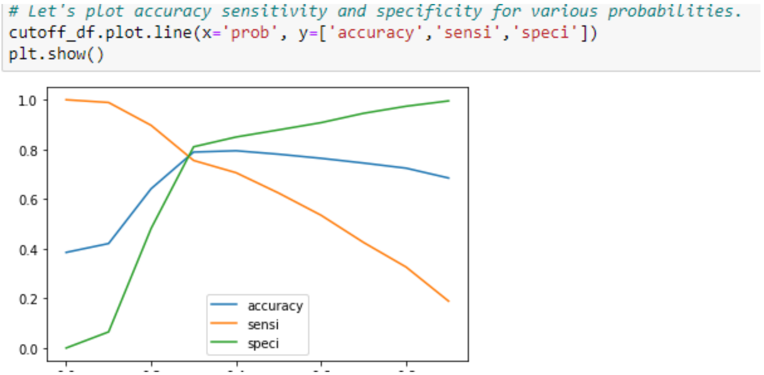
Finding Optimal cut-off point

We have plotted the above metrics at all cut offs from 0.1 to 0.9 and have drawn curves to identify optimal cut-off value as 0.3.

At appx. 0.3 cut off, Accuracy, sensitivity and Specifity curves are intersecting and Sensitivity and Specificity are also above 75%. We obtained below values:

Accuracy: 79%

Sensitivity: 76%



```
# Let's plot accuracy sensitivity and specificity for various probabilities.
cutoff_df.plot.line(x='prob', y=['accuracy','sensi','speci'])
plt.show()
```

Results of Case study analysis

Model evaluation on test set

After evaluating the model on the test set, we obtained below values:
Accuracy: 79%
Sensitivity: 75%

So, we got good results for this model and can be used for immediate action to act on Potential leads.

We have assigned Lead score to each Prospect ID based on which action can be taken to get successful leads**.**

Thank you