Lead Score Case Study

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
- 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 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. To make this process more efficient, the company wishes to identify the most potential leads, also known as ‘Hot Leads’.
- The company requires you to build a model wherein you need to assign a lead score to each of the leads such that the customers with a higher lead score have a higher conversion chance and the customers with a lower lead score have a lower conversion chance.

Business Objective

- The company wants to increase their conversion rate from 30% to 80%
- They want a model which can give hot leads i.e. with high conversion rate
- They want an efficient process to reduce the time spent on lower conversion leads

Data

- The given data contains different information about each lead in 37 columns.
- ‘Converted’ is the target variable which gives information if the lead gets converted or not.
- The target variable is balanced with 60% of 0’s and 40% of 1’s

Lead Origin
Lead Source
Do Not Email
Do Not Call
Converted
TotalVisits
Total Time Spent on Website
Page Views Per Visit

EDA steps taken

- Missing value treatment

  - There were some columns with around 30% of missing values. Decided the drop these columns.

  - Select value in some columns can also be considered under missing values.

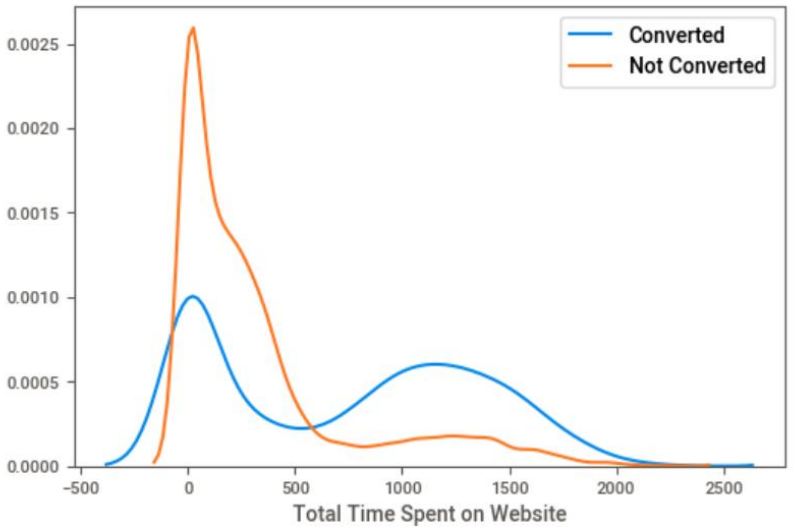
  - Imputed the missing values of numeric columns with mean value and categorical columns with mode value

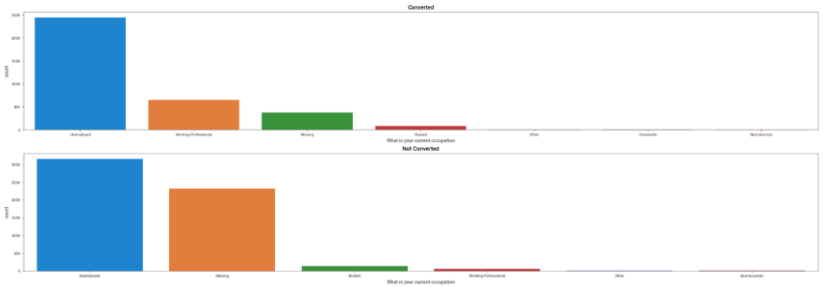
- Zero variance columns

  - Some columns like have only one value. This is not of any use while prediction. Hence dropping these columns.

Data Visualizations

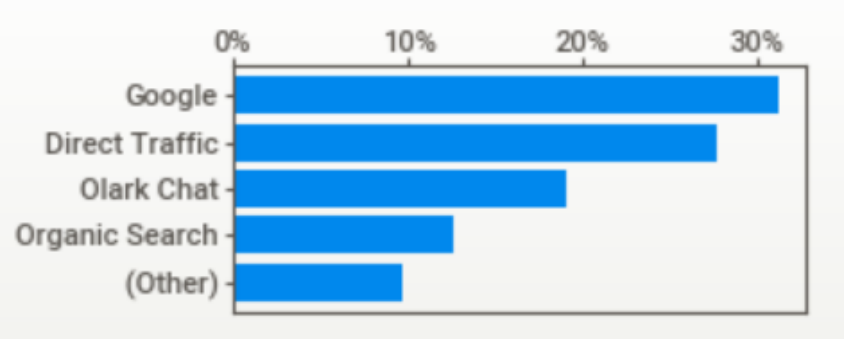
- Surprisingly, people who spent more time on the website are turning out to be not converted. People who spend moderate time on the website are turning out to be converted.
- We can clearly see Working professionals have the high probability of getting converted after unemployed

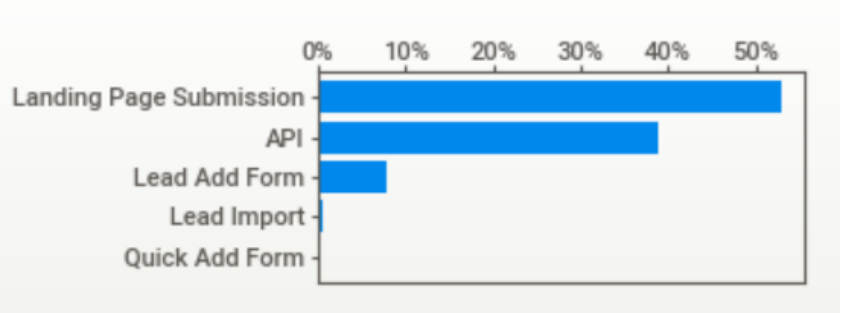




Data Visualizations

- Lead Origin has more than 50% values as ‘Landing page submission’ and ‘API’ around 40%
- Lead source has values such as Google, Direct Traffic, Olark chat





Data Preparation

- Dummy Value Creation

  - Created dummy values for all the categorical columns using get_dummies

- Data scaling

  - Using MinMaxScaler scaled all the numerical features so that all the values are on the same scale for comparison and prediction.

- Correlation Check

  - Check for correlation among the columns. We will eliminate the correlated columns after checking the VIF score

Data Modelling

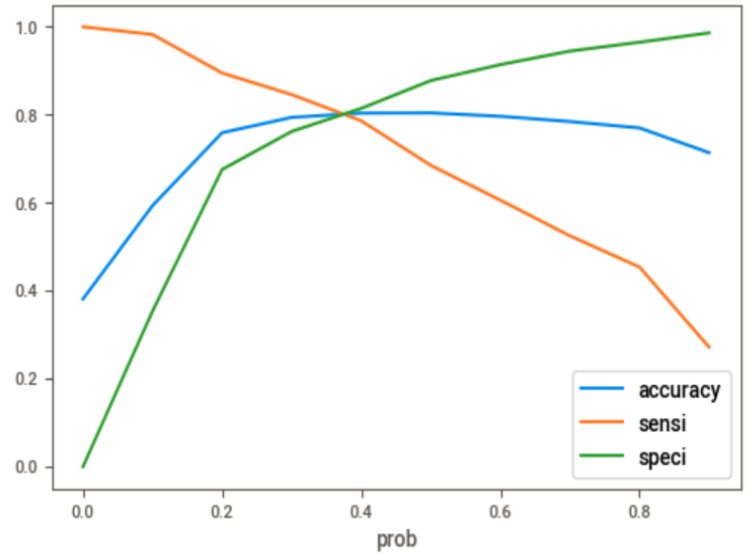
- Using Logistic regression to find the initial performance of the model. The performance looks good with 82% accuracy in train set and 81% in test set.
- Confusion matrix also looks good with fewer false Negatives (336) i.e., Leads that are predicted not converted when they are Converted.

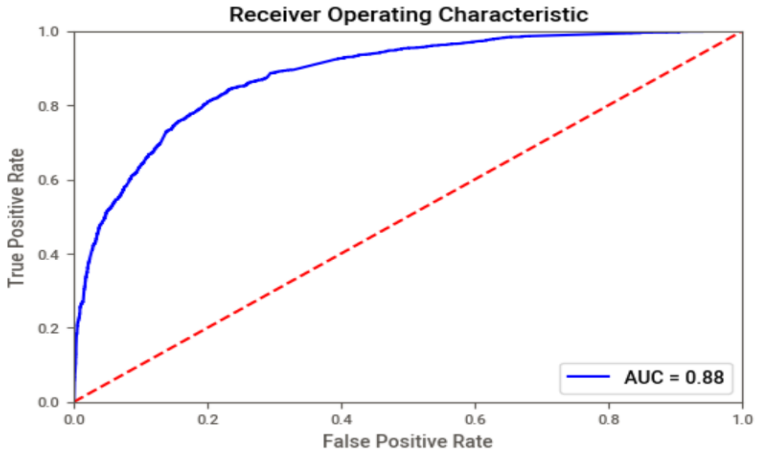
```
Train score :
0.8254483611626469
Test score :
0.8174603174603174
[[1507  170]
 [ 336  759]]
```

- Using RFE to select 15 important features.
- Using GLM model to calculate p value and eliminate the variables with p value >0.05
- VIF score looks good and below the threshold 5 for all the variables

Optimal cut-off value

- After the p value and VIF score looks good, Make a predictions again and get the model metrics
- Choose the initial cut off as 0.5, Obtained an AUC of 0.88 with this cut off
- Find the optimal cut off by checking the trade off graph between accuracy, sensitivity and specificity
- Obtained the cut off value of 0.38
- Made final predictions in train set with this cut off





Predictions on test data

- Make predictions on the test data using the same model
- Given are the top 15 features list

```
Sensitivity :0.8032432432432433
Specificity : 0.8094206821873308
Accuracy : 0.8073593073593074
```

Features
Lead Origin_Landing Page Submission
Specialization_Missing
Last Activity_Had a Phone Conversation
Last Notable Activity_Had a Phone Conversation
Total Time Spent on Website
Lead Source_Olark Chat
What is your current occupation_Missing
TotalVisits
Last Activity_SMS Sent
Lead Origin_Lead Add Form
Lead Source_Welingak Website
What is your current occupation_Working Profes...
Do Not Email_Yes
What is your current occupation_Housewife
Last Notable Activity_Unreachable

Conclusions

- Working professionals are more likely to convert after unemployed.
- Most of the traffic is coming from Google and direct traffic.
- The conversion rate for Landing page submission and API are less. Business needs to focus these areas.