Lead Scoring
Group Case Study

Huynh Huu Hien

Harsh Rawat

Hemlata Sah

Agenda

- Problems & Goals
- Data sources
- Data cleaning
- Data analyzing
- Previous application dataset
- Summary

Problems

- 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

Objectives

- 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 man-power and after achieving target what should be the approaches.

For this case study, 3 files were provided:

- Leads Data
- Lead Data Dictionary
- Assignment Subjective Questions

Solution Approach

- Import Data and Inspect The Data Frame
- Data Cleaning
- EDA
- Dummy Variable Creation
- Test-Train Split
- Feature Scaling
- Check Correlations
- Model Building (RFE & Manual approach)
- Make Prediction & Model Evaluation
- Conclusion

Data Cleaning

- Remove unnecessary columns with missing rows higher than 40%

```
col_missing_percent=pd.DataFrame(round(100*(df1.isnull().sum()/len(df1.index)), 2))
high_miss_col=[]
col_missing_percent=col_missing_percent.reset_index()
for i in range(len(col_missing_percent)):
    if col_missing_percent.iloc[i,1] > 40 :
     # print(col_missing_percent.iloc[i,0])
       high_miss_col.append(col_missing_percent.iloc[i,0])
print(high_miss_col)
```

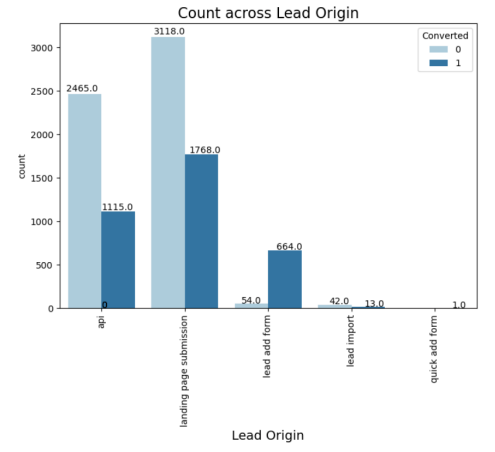
```
['How did you hear about X Education', 'Lead Quality', 'Lead Profile', 'Asymmetrique Activity Index', 'Asymmetrique Pr
ofile Index', 'Asymmetrique Activity Score', 'Asymmetrique Profile Score']
```

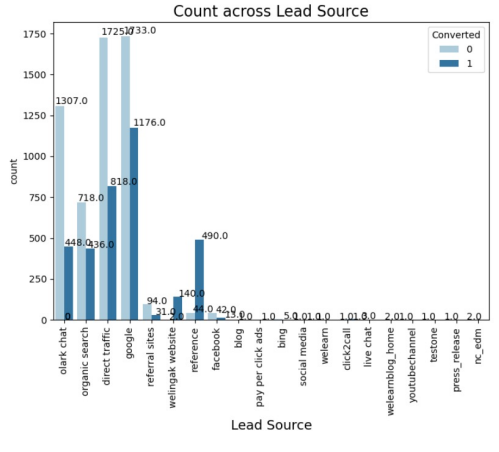
- Convert columns which have 'Select' to NaN , then check case by case to replace them with 'not_provided' (Specialization) and 'Other' ('what is your current occupation')
- Drop columns which have high missing values , imbalance data and no useful for analysis: 'City', 'Country', 'What matters most to you in choosing a course', 'Prospect ID','Lead Number', ' Better Career Prospects'
- Drop columns which have unique data :'Magazine', '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'
- Replace low missing values such as: 'Lead Source', 'Last Activity' (by mode) and 'TotalVisits', 'Page Views Per Visit' (by median)

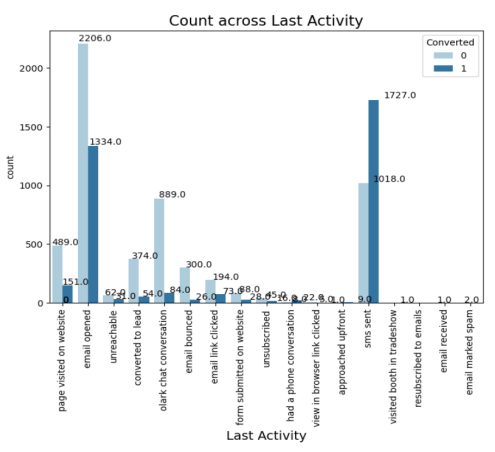
EDA (Categorical variables and Target variable)

Observation:

- Lead Origin: Focus on 'lead add form' as it has highest conversion rate.
- Lead Source: Conversion rate of 'Reference' and 'Welingak Website' leads is high although 'Google' and 'Direct traffic' generates maximum number of leads
- Last Activity: focus on 'sms sent ' as it’s generating a lot of converted leads . 'Email Opened ' also generates numbers of leads.



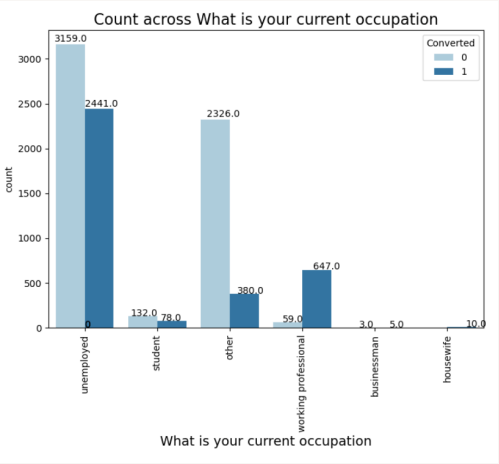


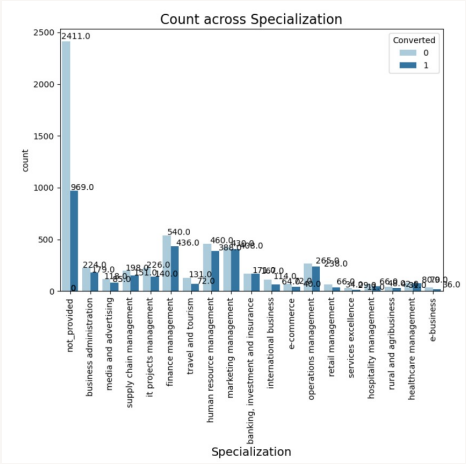


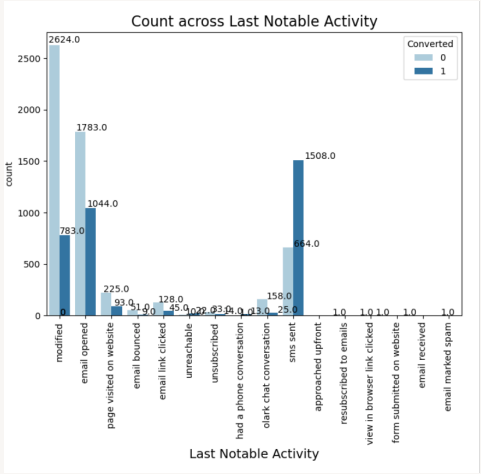
EDA (Categorical variables and Target variable)

Observation:

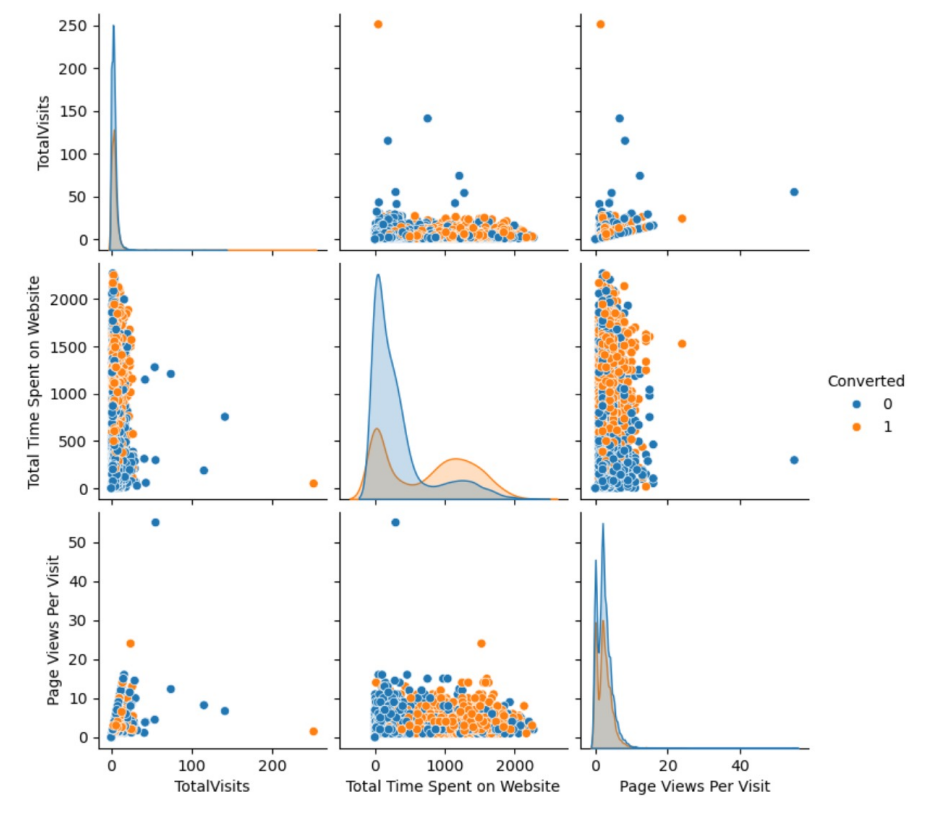
- Occupation: Although there are more converted leads with 'unemployed', the 'working professionals ' has higher conversion rate.
- Specialization: 'Management' specialization altogether having more number of leads generating and not_provided category is also generating more numbers of lead (not_provided = Select, customers who didn’t choose their specialization) .
- Last notable activity: focus on 'sms sent ' as it’s generating a lot of converted leads . 'Email Opened ' also generates numbers of leads.







EDA (Numerical variables and Target variable)



Observation:

- If customer more and more visit the websites, they're potential to be converted.
- If customer spend more time on website, they’re more potential to be converted.

Correlation

- No high correlation between the variables

Model Building

- Splitting the data into training & testing set, choose 70:30 ratio
- Use RFE with 15 variables output
- Manual approach by removing the variable whose VIF value is greater than 5 or p-value is greater than 0.5

Final model line equation :

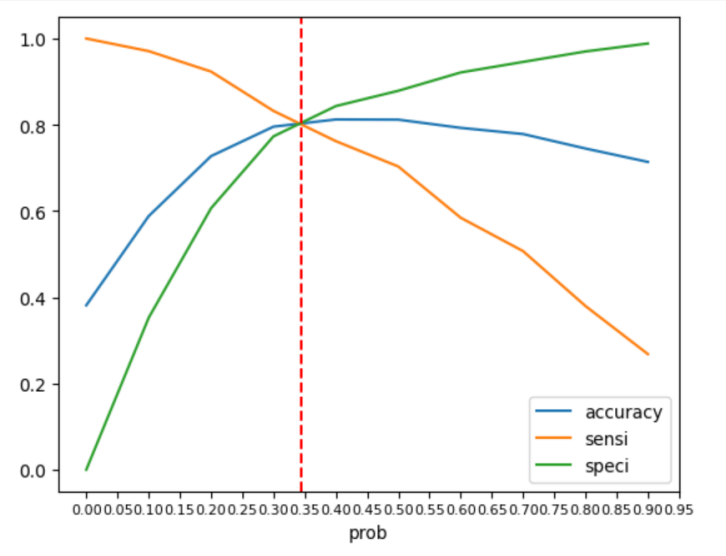
Converted *= -2.47 + 5.18*'TotalVisits' + 4.53*'Total Time Spent on Website' +3.75*'Lead Origin_lead add form' +1.25*'Lead Source_olark chat' + 1.87*'Lead Source_welingak website' -1.24*'Do Not Email_yes' -1.18*'Last Activity_converted to lead' - 1.41*'Last Activity_email bounce' -1.38*'Last Activity_olark chat conversation' +1.22*'Last Activity_sms sent' +1.86*'Last Notable Activity_email bounced' +3.56*'Last Notable Activity_had a phone conversation' +1.79*'Last Notable Activity_unreachable' + 2.84*'What is your current occupation_working professional'*

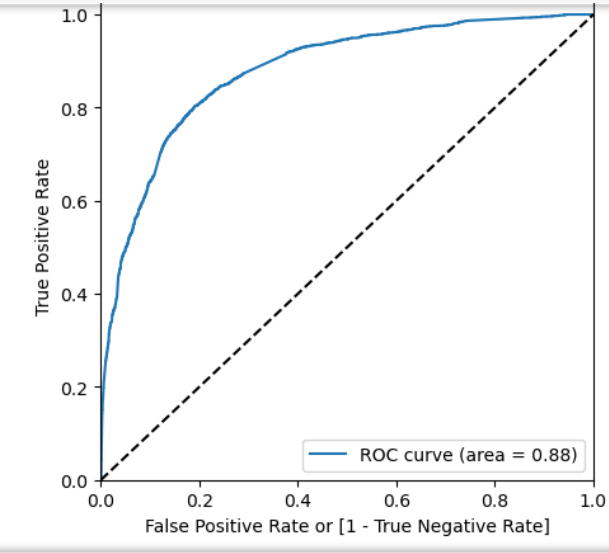
Make Prediction & Model Evaluation

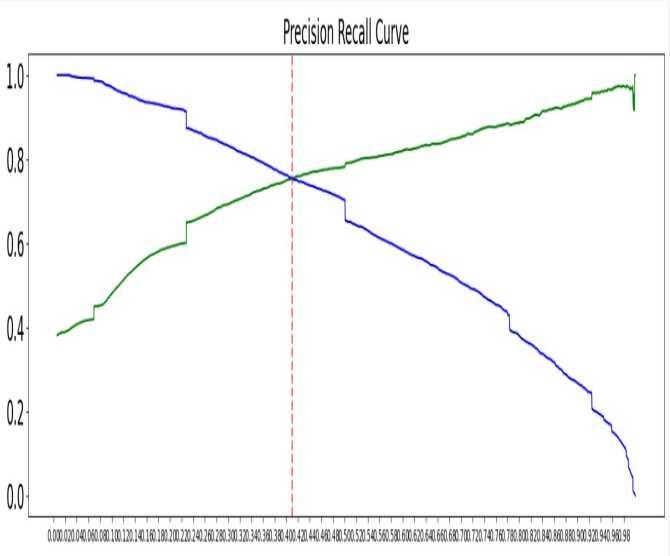
- Finding optimal cut-off point: 0.345.
- Use accuracy, sensitivity, specificity approach because it provide more positive predicts than precision recall

Training model:

- Accuracy : 80.7%
- Sensitivity: 80.3%
- Specificity: 81%
- Precision: 72.2%
- Recall: 80.3%.







Conclusion

It was found that the variables that mattered the most in the potential buyers are (In descending order) :

- Total number of visits
- Total time spend on the website.
- When the lead origin is Lead add form
- When last notable activity is had a phone converstation
- When customer occupation is working professional
- When lead source is Welinkak website
- When last notable activity is email bounced
- When last notable activity is unreachable
- When lead source is Olark chat
- When last activity is sms sent




Thank you