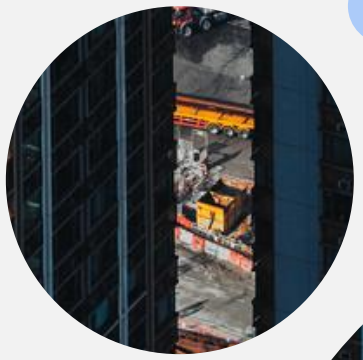
Lead Scoring Case Study

February,2024

Contents



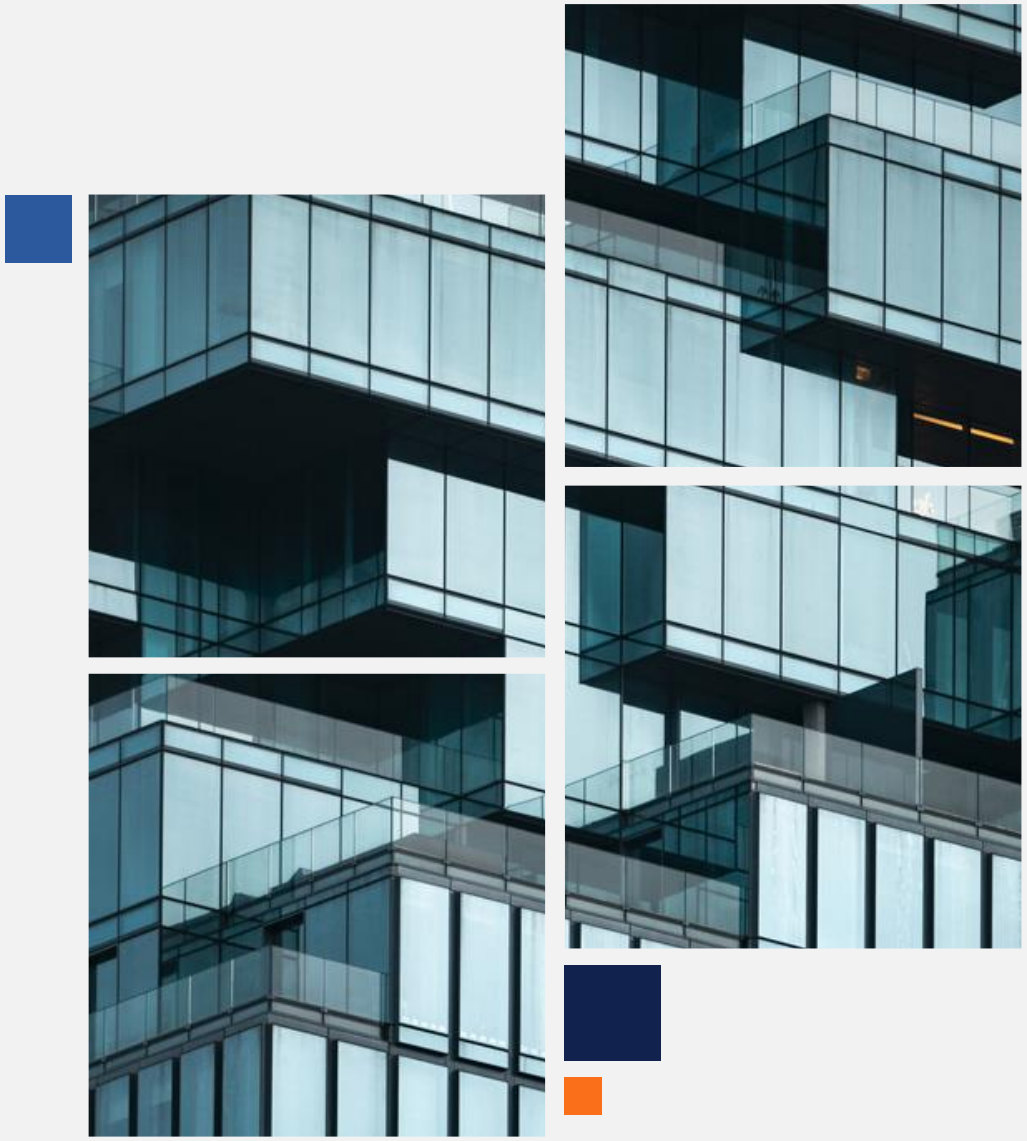
- Problem statement
- Problem approach
- EDA
- Correlation
- Model Eval.
- Observation
- Conclusion

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. They have process of form filling on their website after which the company that individual as a leadToday we'll review our wins and losses from last year and give you an overview of what you can expect for next year.

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 in order to let the conversation rate go up



Problem Approach

- Importing the data and inspecting the data frame
- Data preparation
- EDA
- Dummy variable creation
- Test-Train split
- Feature scaling
- Correlations
- Model Building (RFE R squared VIF and p-values)
- Model Evaluation
- Making predictions on test set

EDA

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 practically all of the values for these variables are No, it's best that we drop these columns as they won't help with our analysis

```
In [76]: leads.drop(['Lead Profile', 'How did you hear about X Education'], axis = 1, inplace = True)
```

```
In [81]: leads.drop(['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'], axis = 1, inplace = True)
```

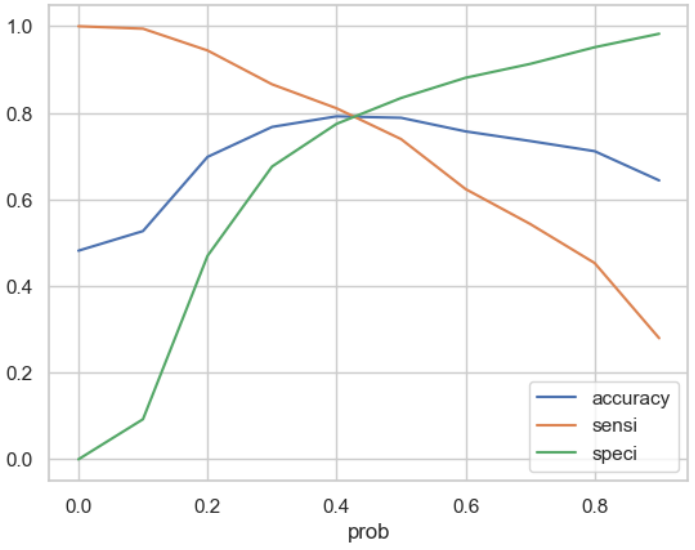
Correlation
There is no correlation between the variables

```
In [94]: # Observing Correlation
         # figure size
         plt.figure(figsize=(10,8))

         numeric_columns = leads.select_dtypes(include=[np.number])

         # Create the correlation matrix heatmap
         sns.heatmap(numeric_columns.corr(), annot=True, cmap="BrBG", linewidths=0.1, vmin=-1, vmax=1)
         plt.show()
```

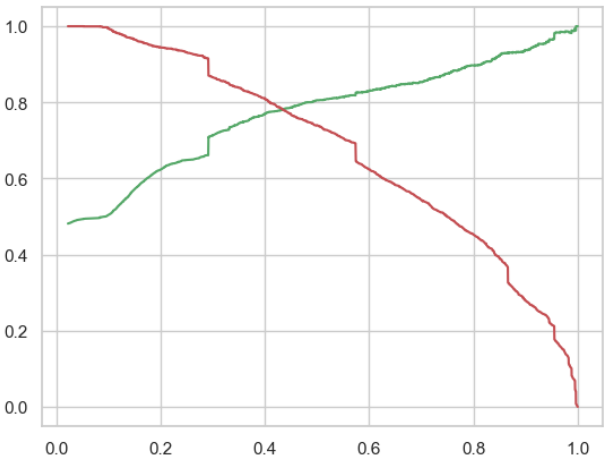
Model Evaluation

ROC curve 0.42 is the tradeoff between Precision and Recall -Thus we can safely choose to consider any Prospect Lead with Conversion Probability higher than 42% to be a HOT LEAD

```
In [161]: cutoff_df.plot.line(x='prob', y=['accuracy','sensi','speci'])
          plt.show()
```




```
In [202]: plt.plot(thresholds, p[:-1], "g-")
          plt.plot(thresholds, r[:-1], "r-")
          plt.show()
```



Observation

- ***Train Data:***
- Accuracy : 80%
- Sensitivity : 77%
- Specificity : 80%

Test Data:

Accuracy : 80%

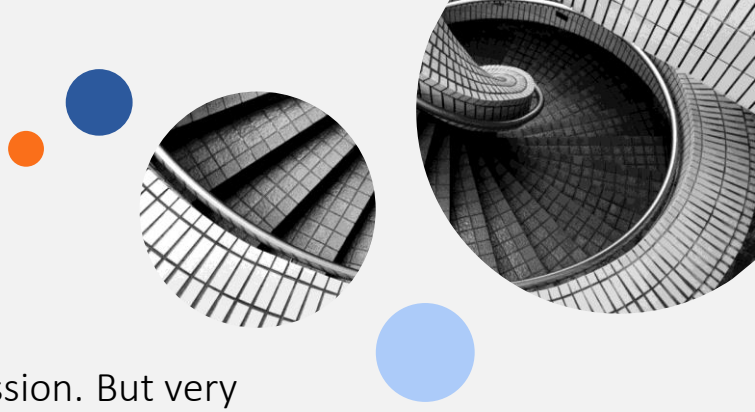
Sensitivity : 77%

Specificity : 80%

Final Features list

- Lead Source_Olark Chart
- Specialization_Others
- Lead Origin_LeadAdd Form
- Lead Source_WelingakWebsite
- Total Time Spent on Website
- Lead Origin_Landing Page Submission
- What is your current occupation_Working Professionals
- Do Not Email

Conclusion



- We see that the conversion rate is 30-35% (close to average) for API and Landing page submission. But very low for Lead Add form and Lead import. Therefore we can intervene that we need to focus more on the leads originated from API and Landing page submission
- We see max number of leads are generated by google/ direct traffic.
- Leads who spent more time on website, more likely to convert
- Most common last activity is email opened. Highest rate is SMS sent. Max conversion with working professional.