Lead Scoring – Case Study

Apoorva Joshi

Sathish Kumar

Introduction

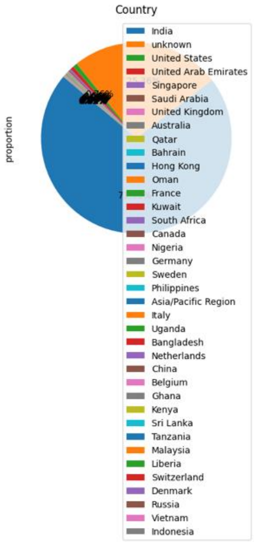
- Objective
  - X Education sells online courses to industry professionals
  - The people trying to subscribe to these courses are called **Leads**
  - This case study is to build models to improve the lead conversion rate (currently 30%) so that the sales team of the company will now be focusing more on communicating with the potential leads rather than making calls to everyone
- Overview of the dataset and problem statement
  - Historical leads dataset with 9000 data points
  - Has a column ‘Converted’ tells whether the past leads are converted or not
- Methodology employed for the analysis
  - Data preparation
  - Exploratory Data Analysis (Univariate, Multivariate analysis)
  - Multivariate Logistic Regression (Model building and evaluation)

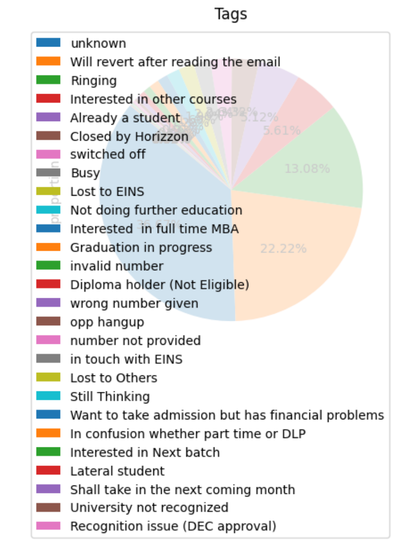
Assumptions

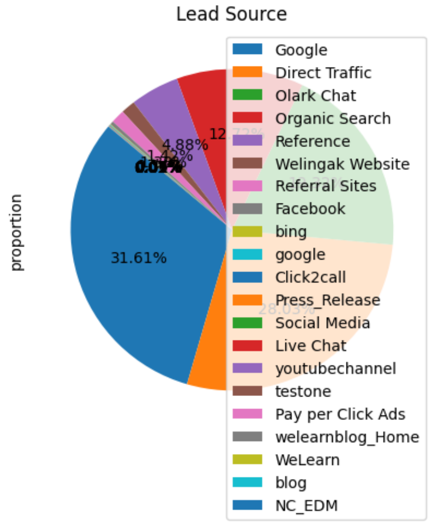
- List of assumptions made during the analysis
    - Null values are either **Imputed** to “unknown” or dropped
    - Feature selection is based on the e-learning domain relevance
        - **Lead Origin** – Initial point of knowing about X Education
        - **Lead Source** – Other sources of search
        - **Country** – Potential geography of high demand
        - **What is your current occupation** – Lead’s profession who are interested in enrolling
        - **Tags** – Categories of the leads to reach out for more conversion
        - **How did you hear about X Education** – Potential marketing strategy
- How assumptions influence results
    - **Lead Origin & Lead Source** – Help the marketing strategy to increase the popularity
    - **Country, Occupation, Tags and How did you hear about X Education** – Help Identify target leads

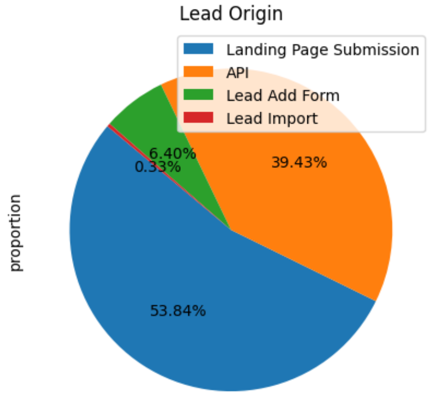
Analysis and Results

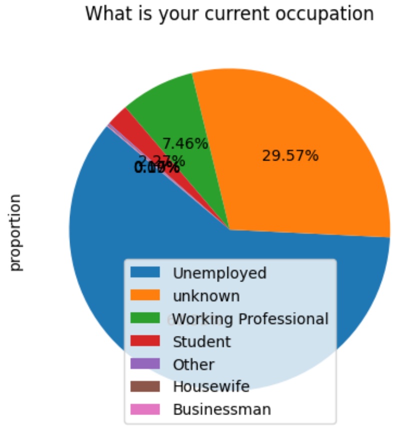
- Key insights from the Exploratory Data Analysis
  - *Above 50% of the lead origin is from, Landing page submission.*
  - *Google contribute to the 31% of the lead source*
  - *70% of the leads are from India.*
  - *Mumbai is the city from where most of the leads are generated.*
  - *unemployed leads has the highest conversion rate*
  - *Most of the leads tags with "will revert after reading the email" has good conversion rate*
  - *Nearly 79% of the leads, how did they hear about X education is unknown*

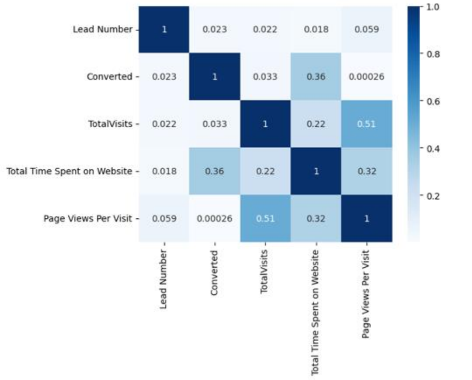








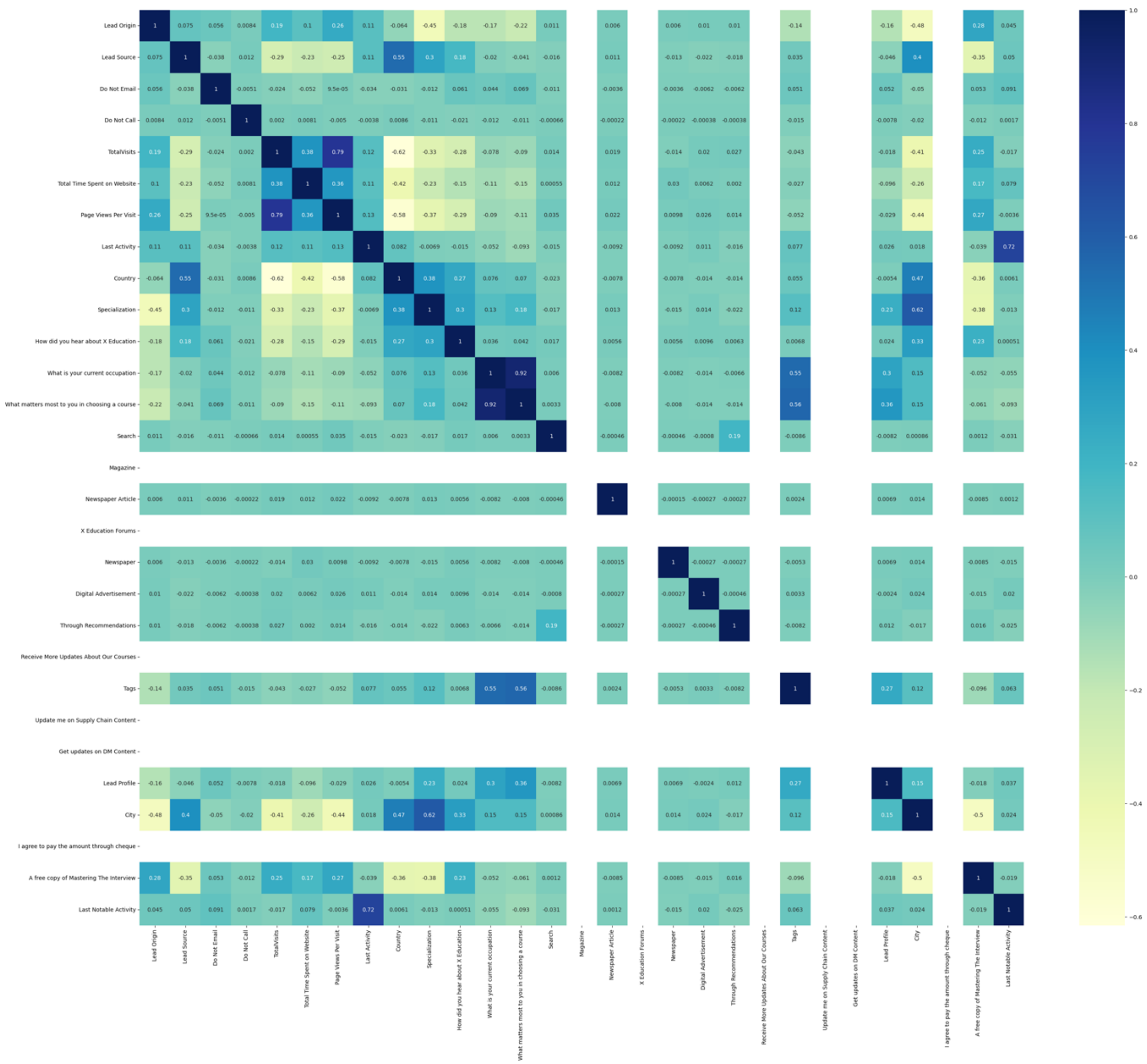




	Lead Number	Converted	TotalVisits	Total Time Spent on Website	Page Views Per Visit
Lead Number	1	0.023	0.022	0.018	0.059
Converted	0.023	1	0.033	0.36	0.00026
TotalVisits	0.022	0.033	1	0.22	0.51
Total Time Spent on Website	0.018	0.36	0.22	1	0.32
Page Views Per Visit	0.059	0.00026	0.51	0.32	1

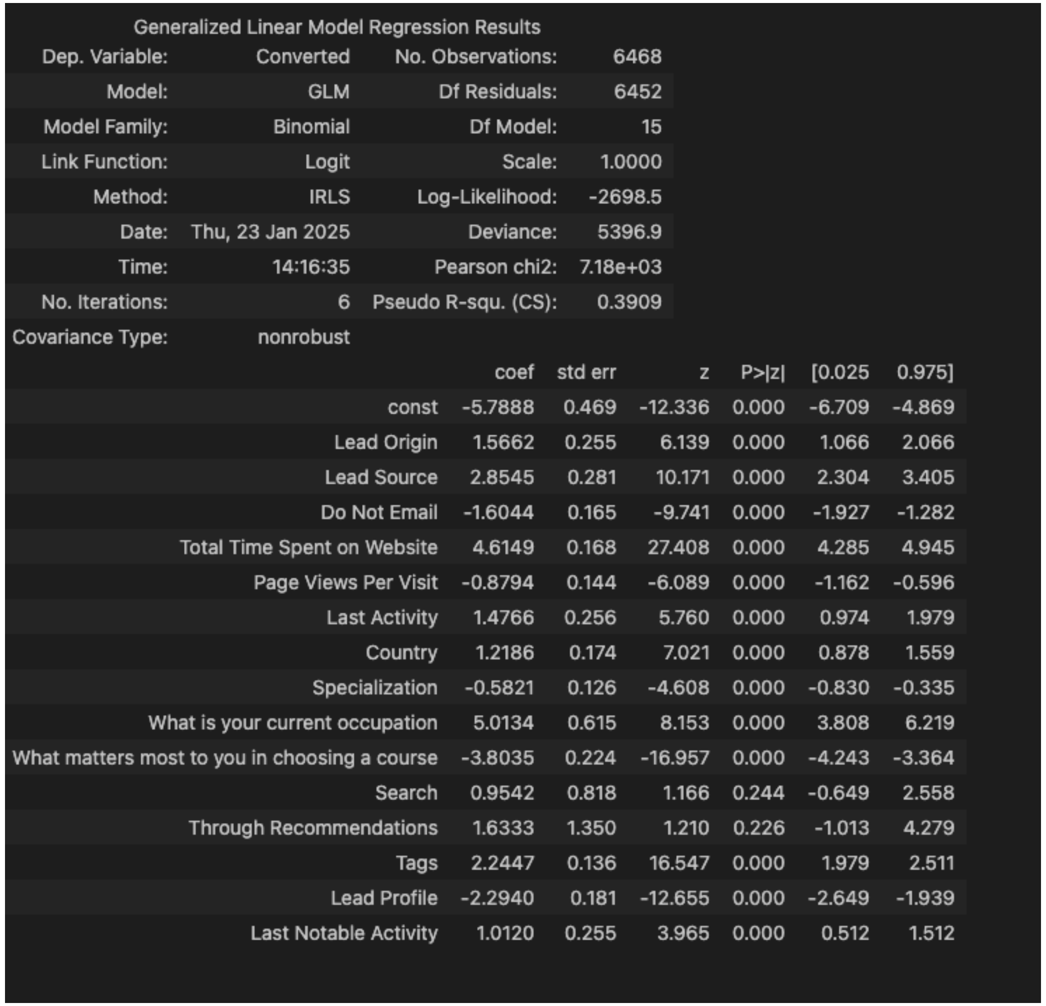
Analysis and Results

- Key insights from the logistic regression model
- Very high correlation between
  - **“last activity”** and **“last notable activity”**
  - **“city” and “specialization”**
  - **“tags” and “what matters most to you in choosing a course”**



Analysis and Results

- Key insights from the logistic regression model
  - Using RFE (Recursive Feature Elimination) to select 15 features to build the model
    1. Lead Origin
    2. Lead Source
    3. Do Not Email
    4. Total Time Spent on Website
    5. Page Views Per Visit
    6. Last Activity
    7. Country
    8. Specialisation
    9. What is your current occupation
    10. What matters most to you in choosing a course
    11. Search
    12. Through Recommendations
    13. Tags
    14. Lead Profile
    15. Last Notable Activity



```
                 Generalized Linear Model Regression Results
   Dep. Variable:        Converted    No. Observations:      6468
          Model:              GLM        Df Residuals:      6452
   Model Family:         Binomial            Df Model:        15
  Link Function:            Logit               Scale:    1.0000
         Method:             IRLS      Log-Likelihood:   -2698.5
           Date:  Thu, 23 Jan 2025           Deviance:    5396.9
           Time:         14:16:35        Pearson chi2:  7.18e+03
 No. Iterations:                6  Pseudo R-squ. (CS):    0.3909
Covariance Type:        nonrobust
                                                  coef  std err        z  P>|z|  [0.025  0.975]
                                         const  -5.7888   0.469  -12.336  0.000  -6.709  -4.869
                                   Lead Origin   1.5662   0.255    6.139  0.000   1.066   2.066
                                   Lead Source   2.8545   0.281   10.171  0.000   2.304   3.405
                                  Do Not Email  -1.6044   0.165   -9.741  0.000  -1.927  -1.282
                   Total Time Spent on Website   4.6149   0.168   27.408  0.000   4.285   4.945
                          Page Views Per Visit  -0.8794   0.144   -6.089  0.000  -1.162  -0.596
                                 Last Activity   1.4766   0.256    5.760  0.000   0.974   1.979
                                       Country   1.2186   0.174    7.021  0.000   0.878   1.559
                                Specialization  -0.5821   0.126   -4.608  0.000  -0.830  -0.335
               What is your current occupation   5.0134   0.615    8.153  0.000   3.808   6.219
What matters most to you in choosing a course  -3.8035   0.224  -16.957  0.000  -4.243  -3.364
                                        Search   0.9542   0.818    1.166  0.244  -0.649   2.558
                       Through Recommendations   1.6333   1.350    1.210  0.226  -1.013   4.279
                                          Tags   2.2447   0.136   16.547  0.000   1.979   2.511
                                  Lead Profile  -2.2940   0.181  -12.655  0.000  -2.649  -1.939
                         Last Notable Activity   1.0120   0.255    3.965  0.000   0.512   1.512
```

Analysis and Results

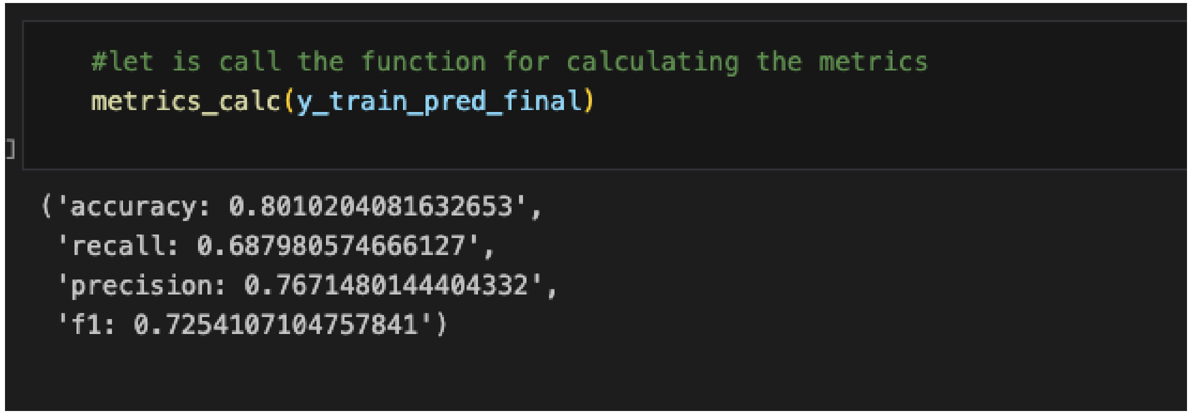
- Key insights from the logistic regression model
  - Based on the VIF (Variance Inflation Factor) of the features in the training dataset
    - Following features form a good model after 6 iterations
      - Lead Origin
      - Lead Source
      - Do Not Email
      - Total Time Spent on Website
      - Page Views Per Visit
      - Last Activity
      - Country
      - Specialization
      - What matters most to you in choosing a course
      - Tags

Analysis and Results

- Key insights from the logistic regression model
  - Following features are dropped because of
    - Insignificance
      - *Through Recommendations*
      - *Search*
    - High VIF
      - *What is your current occupation*
      - *Lead Profile*
      - *Last Notable Activity*

Analysis and Results

- Model evaluation metrics
  - accuracy: 80% (High correct predictions)
  - recall: 69% (High positive predictions)
  - precision: 77% (Predicted positives are positives)
  - f1: 0.73 (close to 1, ensuring a balanced model)

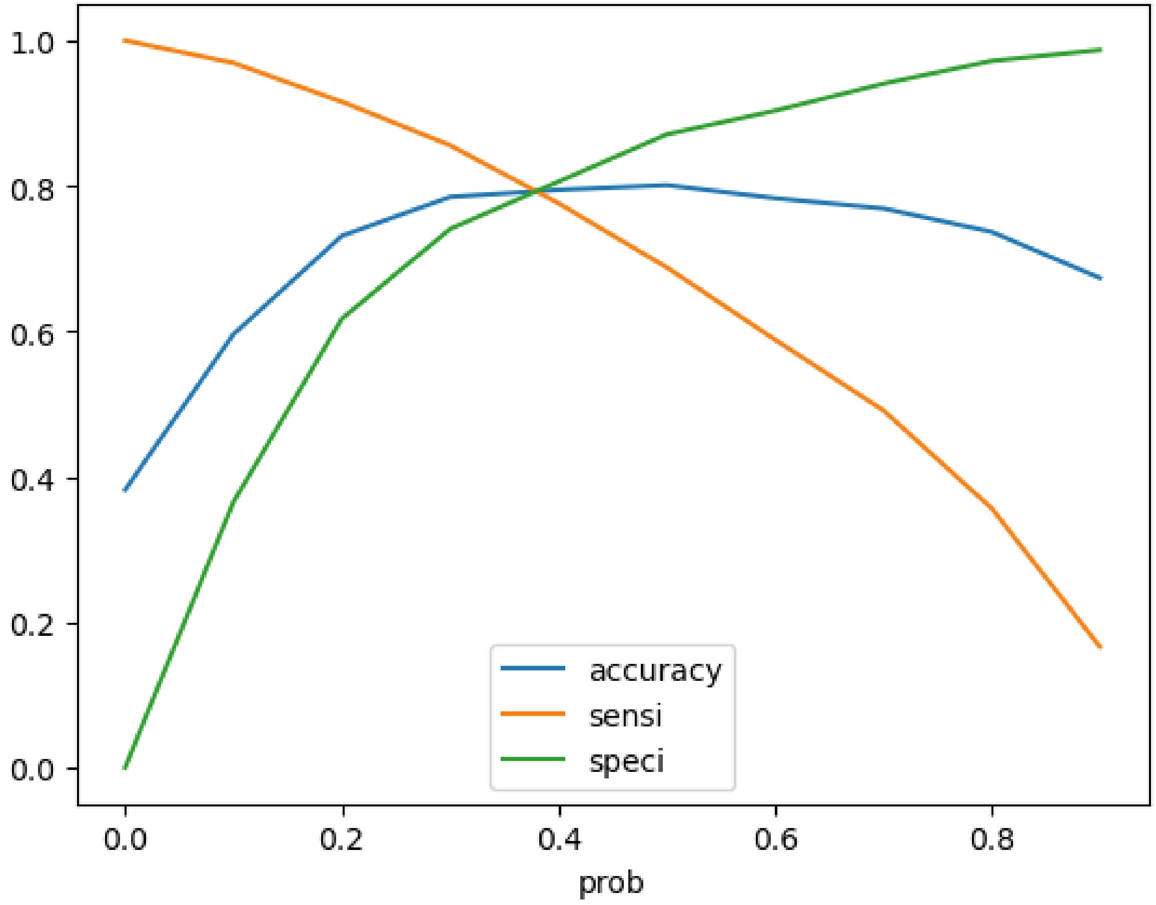


```
#let is call the function for calculating the metrics
metrics_calc(y_train_pred_final)
```

```
('accuracy: 0.8010204081632653',
 'recall: 0.687980574666127',
 'precision: 0.7671480144404332',
 'f1: 0.7254107104757841')
```

Analysis and Results

- Statistical significance of predictors
    - Accuracy of the train set is 80% while that of test dataset is 79%
    - Recall of the train set is 68% while that of test dataset is 76%
    - Precision of train set is 76% while that of test dataset is 72%
    - F1 score of train set is 0.72 while that of test dataset is 0.75



Recommendations

1. Which are the top three variables in your model which contribute most towards the probability of a lead getting converted?
    - The top three variables are determined by the absolute values of coefficients of the model.
    - In this case top three variables are “Total Time Spent on Website”, “Lead Source” and “What matters most to you in choosing a course”.
2. What are the top 3 categorical/dummy variables in the model which should be focused the most on in order to increase the probability of lead conversion?
    - To increase the probability of lead conversion we should focus on variables with positive coefficients.
    - They are “Total Time Spent on Website”, “Lead Source” and “Tags”.
3. X Education has a period of 2 months every year during which they hire some interns. The sales team, in particular, has around 10 interns allotted to them. So during this phase, they wish to make the lead conversion more aggressive. So they want almost all of the potential leads (i.e. the customers who have been predicted as 1 by the model) to be converted and hence, want to make phone calls to as much of such people as possible. Suggest a good strategy they should employ at this stage.
    - For rapid lead conversion, company can focus on the lead_score which is given in the model and try to convert those leads.
    - Also, focus on the variables with positive coefficients such as “Total Time Spent on Website”, “Lead Source” and “Tags” and try to invest time on these fields.
4. Similarly, at times, the company reaches its target for a quarter before the deadline. During this time, the company wants the sales team to focus on some new work as well. So during this time, the company’s aim is to not make phone calls unless it’s extremely necessary, i.e. they want to minimize the rate of useless phone calls. Suggest a strategy they should employ at this stage.
    - The company can try other communication methods such as emails, posts and advertisements.
    - Phone calls must be restricted only to the leads which have high probability of conversion based on lead_score.
    - Reduce the focus on unemployed leads and students as the conversion probability rate can be low