Lead Scoring Case Study

- Abhishek Mitra

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.
- Leads have features like Lead Source, Lead Origin, Employment etc.
- X Education wants to identify the best leads
- For this task, a machine learning model needs to be identified that can identify the best leads

Solution Methodology

Solution Methodology

- ▼ Data cleaning and data manipulation.
- Detect and remove rows with large number of missing values
- Impute missing values where possible
- Replace NaNs with suitable values

- ▼ Feature Scaling & Dummy Variables
- Encode categorical variables as dummy variables and perform feature scaling
- ▼ Modelling technique
- Logistic regression is the model used for modelling.
- ▼ Validation of the model.
- The model is tested on the test data

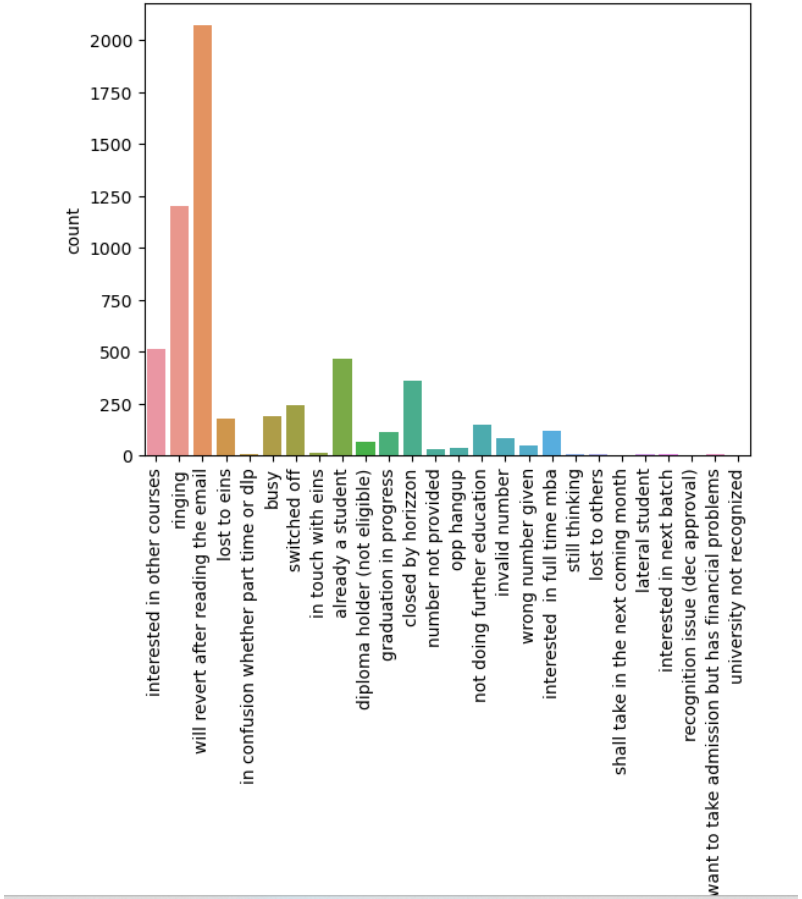
Data Manipulation

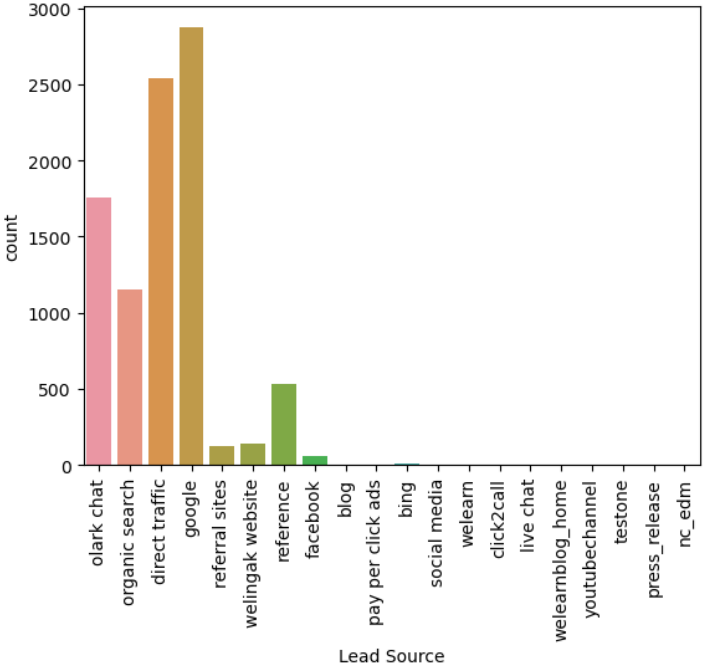
- “select” as a word is replaced with nan
- Columns with more than 35% of missing values are
- Nans are replaced with suitable values or imputed wherever feasible

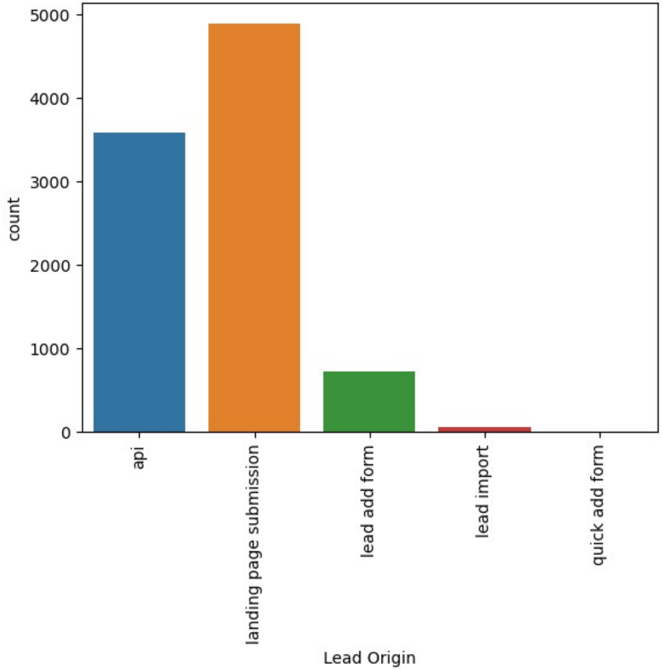
Data Conversion

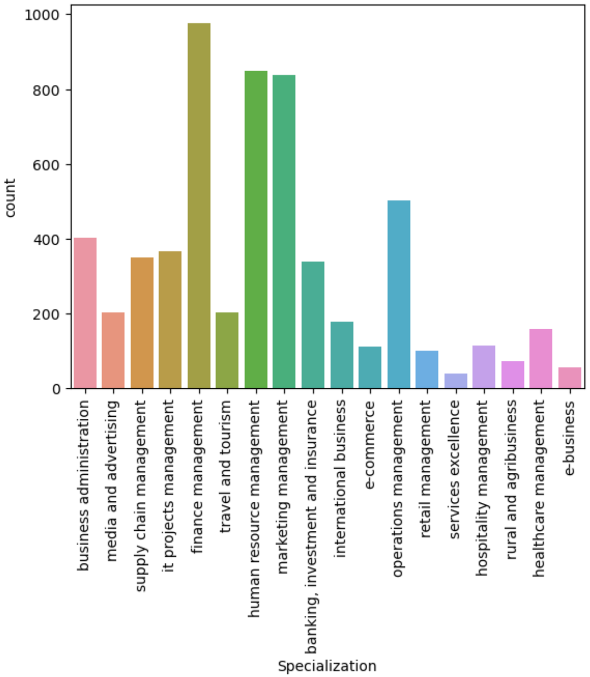
- Numerical Variables are normalised and scaled using sklearn’s scaling transformers.
- Categorical variables are transformed into dummy variables by means of one-hot encoding.

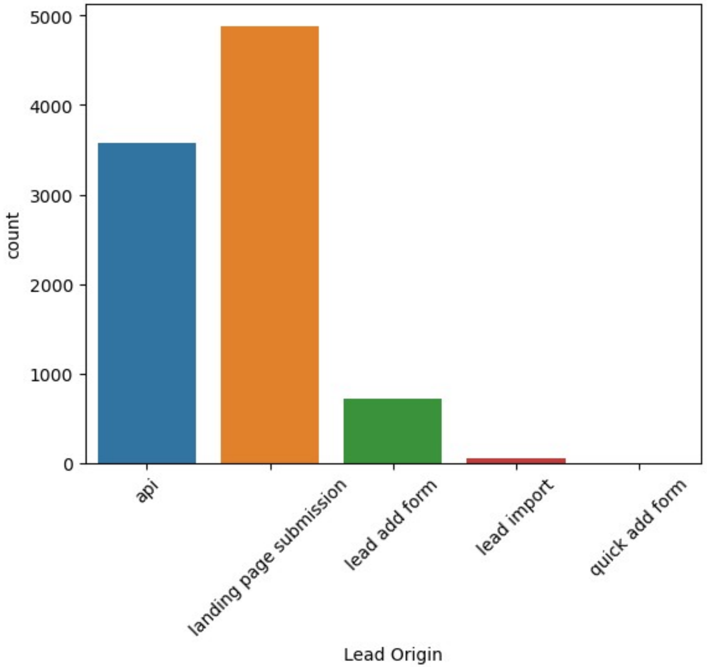
Exploratory Data Analysis




3000
2500
2000
1500
1000
500
0
count
olark chat
organic search
direct traffic
google
referral sites
welingak website
reference
facebook
blog
pay per click ads
bing
social media
welearn
click2call
live chat
welearnblog_home
youtubechannel
testone
press_release
nc_edm
Lead Source


5000
4000
3000
2000
1000
0
count
api
landing page submission
lead add form
lead import
quick add form
Lead Origin

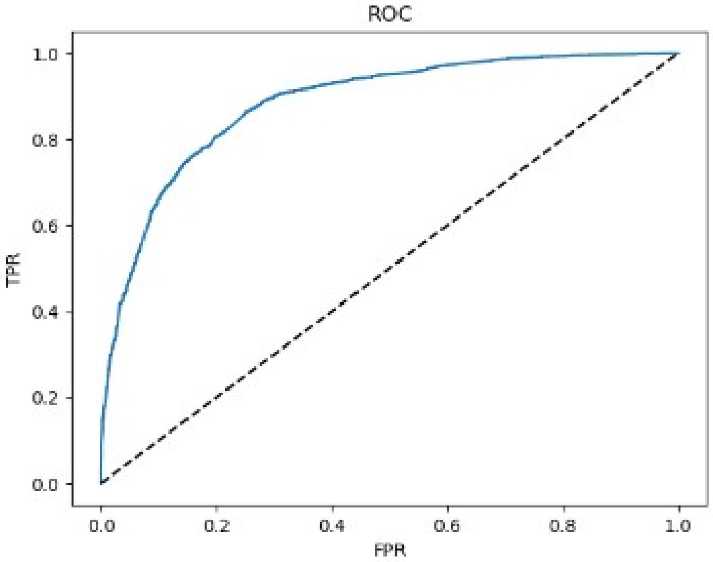

count
1000
800
600
400
200
0
business administration
media and advertising
supply chain management
it projects management
finance management
travel and tourism
human resource management
marketing management
banking, investment and insurance
international business
e-commerce
operations management
retail management
services excellence
hospitality management
rural and agribusiness
healthcare management
e-business
Specialization


5000
4000
3000
2000
1000
0
count
api
landing page submission
lead add form
lead import
quick add form
Lead Origin

Model Building

- The data is split into a train and test split with 70% data in train and 30% in test set.
- RFE is used to select 15 columns.
- VIF values are computed and columns are removed on the basis of this value.
- The test precision achieved is 75.45% and recall is 75.85%

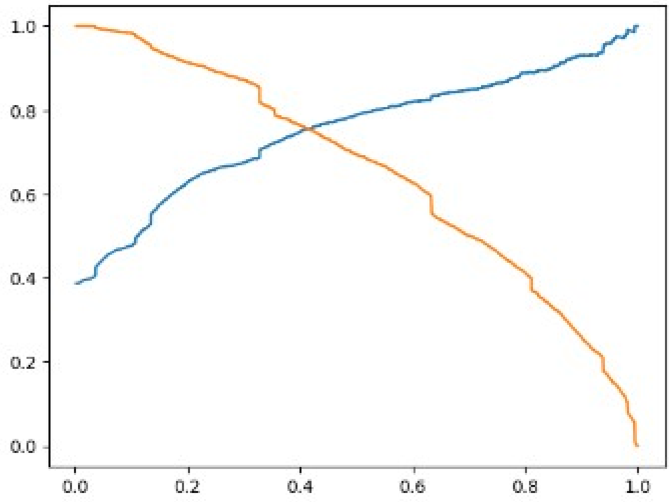
ROC Curve



- The optimal cutoff value is arrived at using the curve above of 0.35.

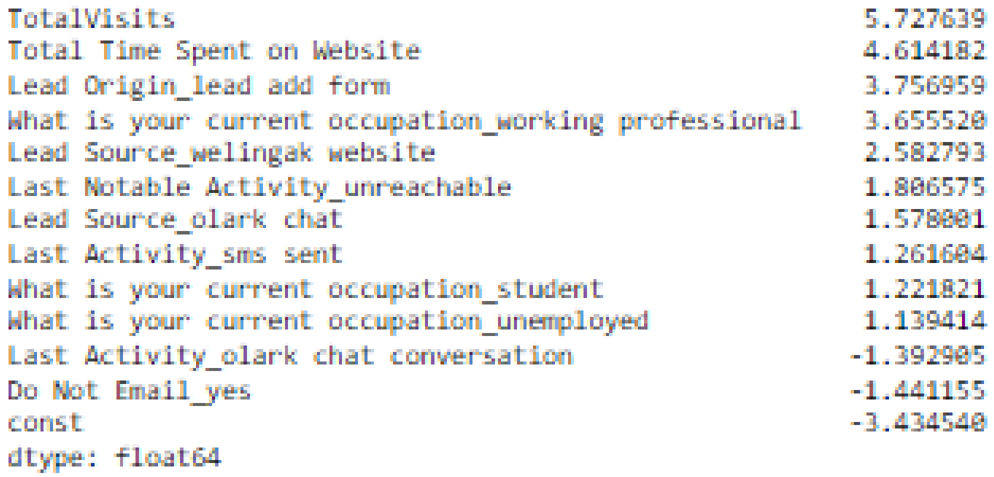
Precision-Recall Curve

The following precision-recall curve is plotted and the two curves intersect at 0.4



Most Important Parameters

The following variables are
the most important parameters



```
TotalVisits                                           5.727639
Total Time Spent on Website                           4.614182
Lead Origin_lead add form                             3.756959
What is your current occupation_working professional  3.655520
Lead Source_welingak website                          2.582793
Last Notable Activity_unreachable                     1.806575
Lead Source_olark chat                                1.578001
Last Activity_sms sent                                1.261604
What is your current occupation_student               1.221821
What is your current occupation_unemployed            1.139414
Last Activity_olark chat conversation                -1.392905
Do Not Email_yes                                     -1.441155
const                                                -3.434540
dtype: float64
```

Conclusion

- The 10 variables that matter the most are:

1. TotalVisits -	5.727639
2. Total Time Spent on Website -	4.614182
3. Lead Origin_lead add form -	3.756959
4. What is your current occupation_working professional -	3.655520
5. Lead Source_welingak website -	2.582793
6. Last Notable Activity_unreachable -	1.806575
7. Lead Source_olark chat -	1.578001
8. Last Activity_sms sent -	1.261604
9. What is your current occupation_student -	1.221821
10. What is your current occupation_unemployed -	1.139414