LEAD SCORE CASE STUDY

DO TUAN ANH

PROBLEM STATEMENT

X EDUCATION

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. Moreover, the company also gets leads through past referrals. 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, 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’. 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.

METHODOLOGY FOR APPROACH

DATA CLEANSING

EDA

FEATURE & DUMMY

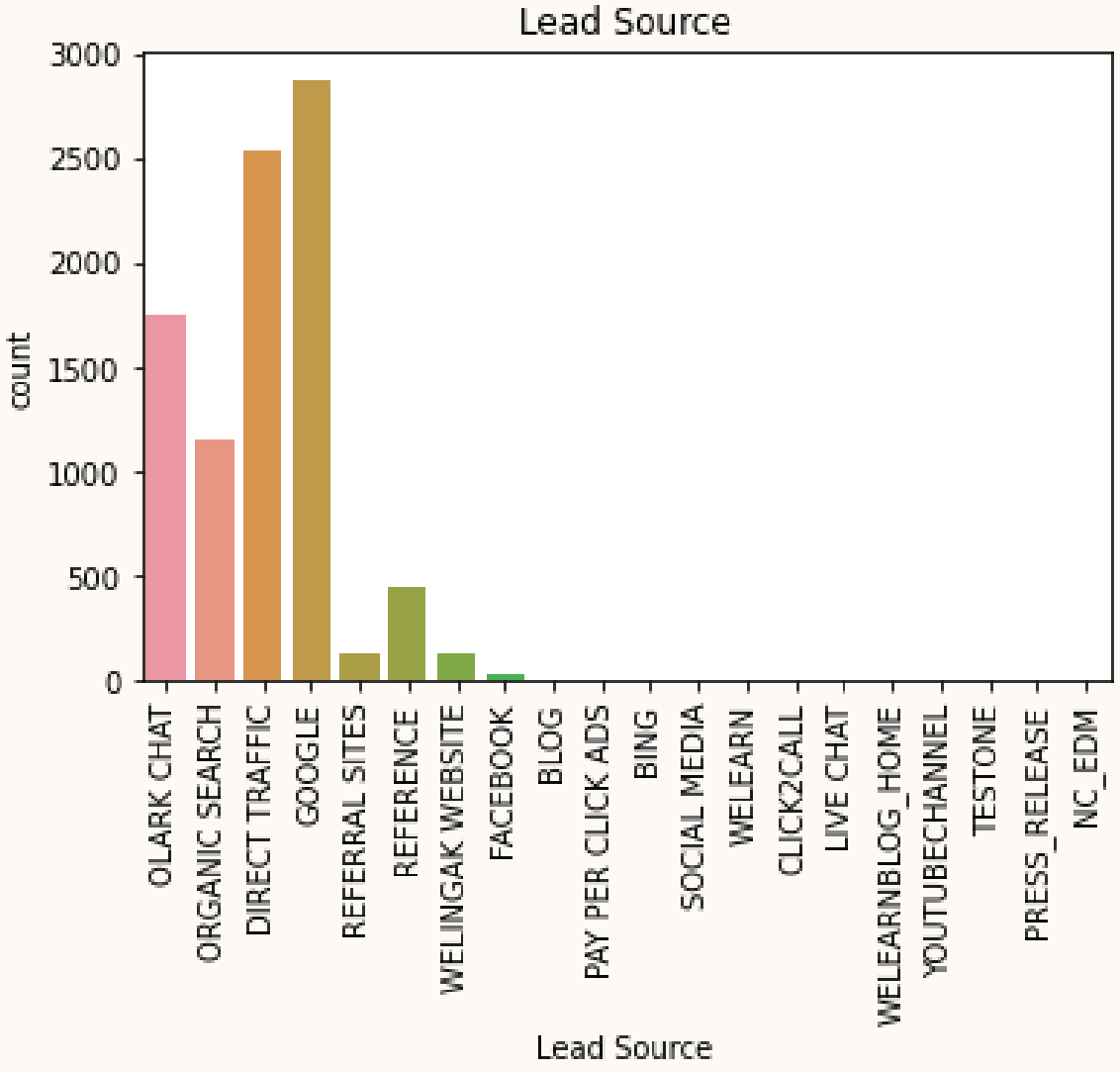
CLASSIFICATION

VALIDATION

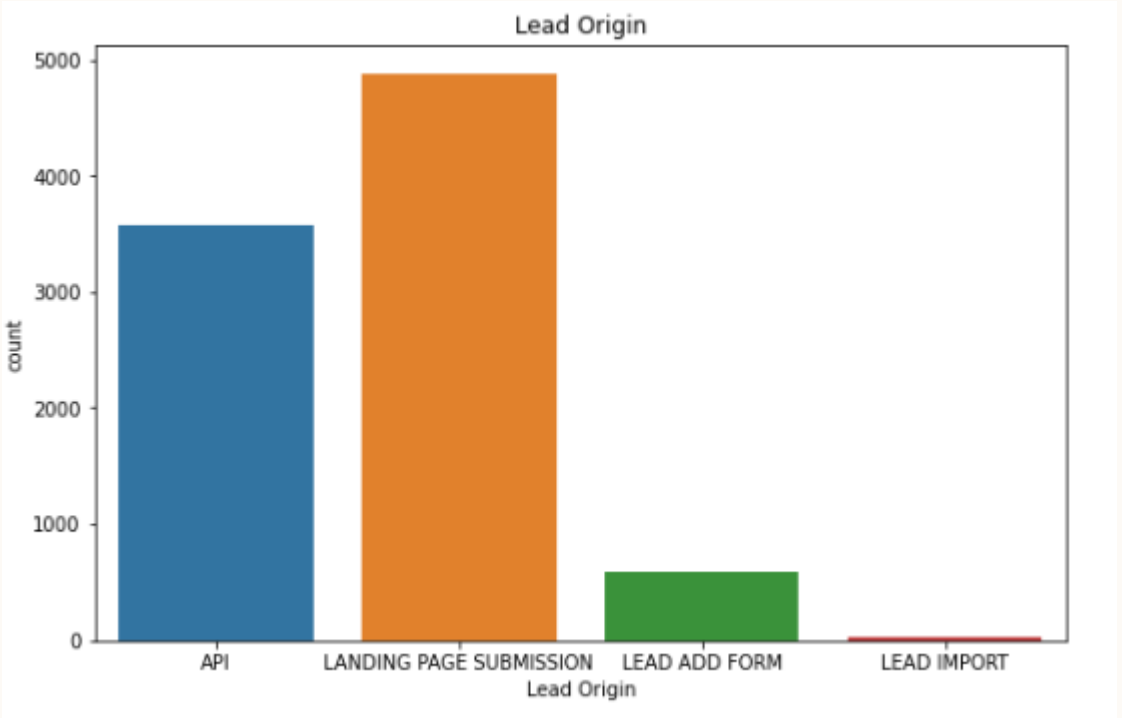
DATA CLEANSING

1. Normalization of data: Transform every string to upper case.
2. Transform every value == ‘Select’ to NA.
3. Drop all column that have <= 1 unique value.
4. Fill all missing value == ‘NA’
5. Drop ‘Prospect ID’ column as it is the PK of the table.

EDA

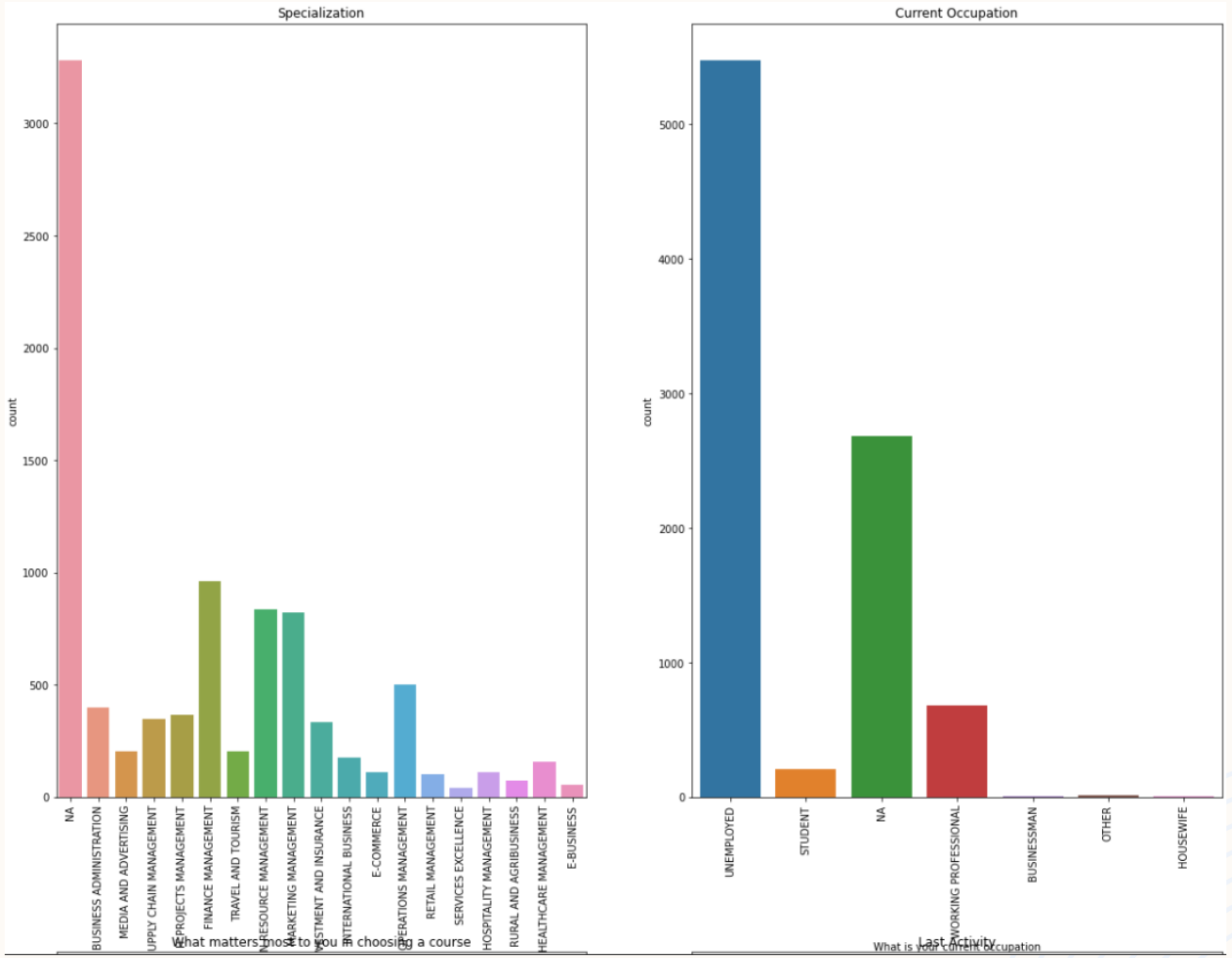


- Most of the lead source comes from the internet.
- It is also true for most lead to be originated from landing page.



EDA

It is also very clear that most lead are unemployed or have no specialization in work.



DUMMY VARIABLES

1. Dummy variables are generated for every object columns.
2. After dummy variables were generated we have total 9074 rows and 215 columns.

MODEL TRAINING

1. Split the data set into training and test set with 70:30 ratio.
2. Using logistic regression to get RFE.
3. Use RFE to get list of features.
4. Run RFE with total 15 variables.
5. Get VIF and start dropping meaningless features using VIF and p-value.
6. We have remaining 11 variables.
7. Find optimal cut-off point using auc score.
8. Optimal cut-off point is found at ~ 0.3.
9. Run prediction on test data set.
10. Final model have ~0.85 accuracy.

SUMMARY

After running the model on test data these are the summary of the model:

Accuracy : 85.86%

Sensitivity :88.35%

Specificity : 84.46%

Based on the above result, this model seems to be good.

Important features responsible for good conversion rate or the ones' which contributes more towards the probability of a lead getting converted are:

Total Time Spent on Website

Lead Origin_LEAD ADD FORM

Tags_RINGING

Tags_WILL REVERT AFTER READING THE EMAIL