Lead Scoring Case Study

By-
Arun PS and Nitin Pratap Singh
DSC 53

Problem Statement

- A company named X education sells courses online.
- They get leads through several websites, search engines and past referrals.
- The sales team contacts the leads to get them to buy their courses, i.e. convert to customers.
- Their lead conversion rate is very poor.
- Our goal is to build a model to identify the ‘hot leads’ i.e. leads most likely to convert. This model will be deployed on the leads before the sales team makes contact.
- This is required so that the company can minimize resources spent on leads which aren’t likely to convert and maximize them on resources which are likely to convert.

Analysis Outline

- Goal - Build a logistic regression model to predict if a lead will be converted or not converted.
- Model must have high precision (i.e. out of the leads we predict will convert, roughly 80% of them must convert)
- This requires us to have a high cutoff on the probability that a lead is likely to convert.

	Probability	Accuracy	Precision	Recall
0.0	0.0	0.386767	0.386767	1.000000
0.1	0.1	0.432815	0.404904	0.993092
0.2	0.2	0.589187	0.483622	0.917920
0.3	0.3	0.758290	0.663943	0.759447
0.4	0.4	0.767720	0.696364	0.708249
0.5	0.5	0.753890	0.728666	0.579439
0.6	0.6	0.739903	0.758997	0.479886
0.7	0.7	0.718529	0.807339	0.357578
0.8	0.8	0.690240	0.854046	0.240146
0.9	0.9	0.644664	0.896825	0.091833

Fig i

Analysis Approach

Apart from the general template that we follow (Data cleaning, Exploratory Data Analysis, Data Preparation, Model Building, Prediction, Evaluation, Optimization, Interpretation and Conclusion) there are certain specific points we kept in mind while building a model for this problem that X-education faces-

- Data generated by the sales team after contacting the lead should be removed from our dataset. This is because our model is to be deployed on leads to identify the “hot leads” before the sales team contacts them (Fig ii). This allows them to focus on those customers which are likely to convert

Fig ii

- A lead score had to be created for each lead. It is to range from 0-100, where higher score means higher probability of the lead converting). Our model (Fig iii) gave us the probability that a lead would convert ‘Converted_Prob’. We multiplied that by 100 and rounded it off to 0 places to get an integer score for each lead ‘lead_score’.

	Converted	Converted_Prob	Predicted	final_predicted	lead_score
0	0	0.159000	0	0	16.0
1	0	0.222578	0	0	22.0
2	1	0.630950	1	0	63.0
3	1	0.335920	0	0	34.0
4	1	0.991971	1	1	99.0
5	0	0.214882	0	0	21.0
6	1	0.877058	1	1	88.0
7	0	0.871010	1	1	87.0
8	1	0.431427	0	0	43.0
9	0	0.194447	0	0	19.0

Fig iii

- We require flexibility from our model as well. Hence we want to be able to set the cut-off ourselves, i.e. cut-off optimization should be in our hands. In some cases we want higher precision (out of the predicted converts, most convert) and in some cases we require greater recall (out of all who could've converted, how many could we identify). Where we lean in the Precision-Recall tradeoff depends on the time crunch the company is in and the resources they're willing to spend.

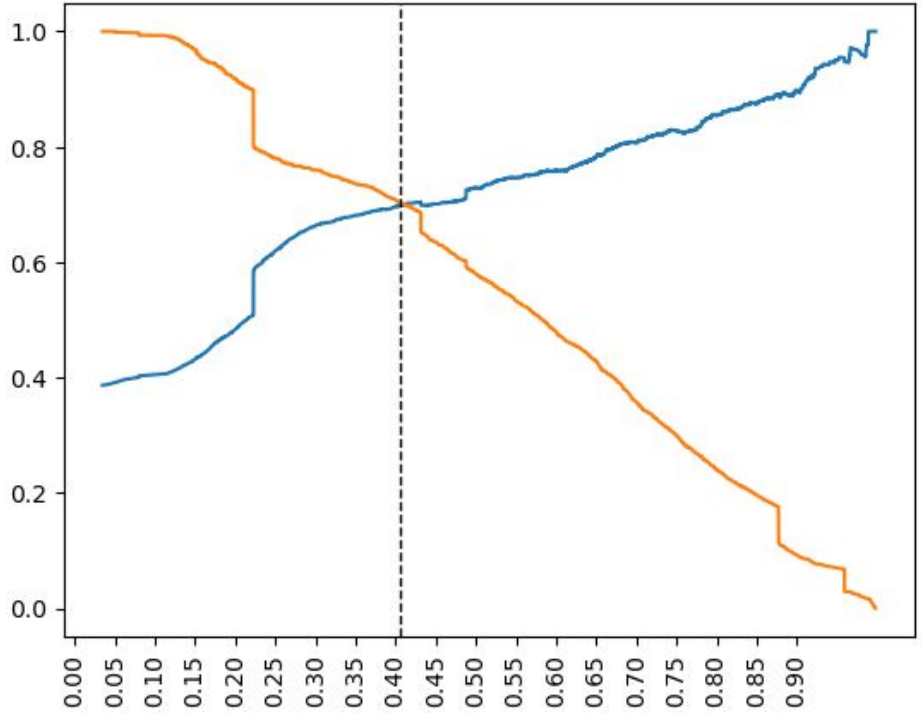

Fig iv

Blue - increasing precision with increasing probability cut-off (X-axis), Orange - Decreasing recall with increasing X

Results

- Leads brought in through references show the highest conversion rate.
- Wellingak website shows the second highest rate of conversion.
- Olark chat and google are follow after the above 2 channels. (Fig v for reference)

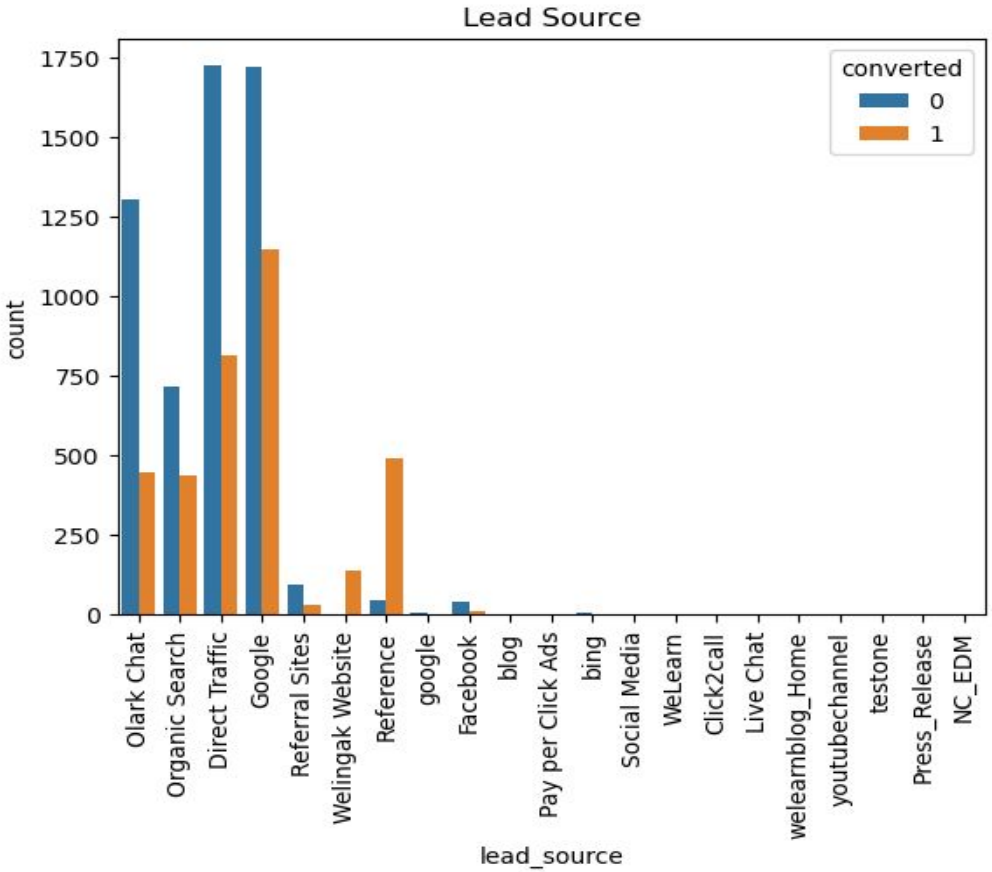

Fig v

- If a lead chooses to not receive emails, this perhaps shows that they are not as serious about the course and it is more likely that they will not convert (Fig vi)

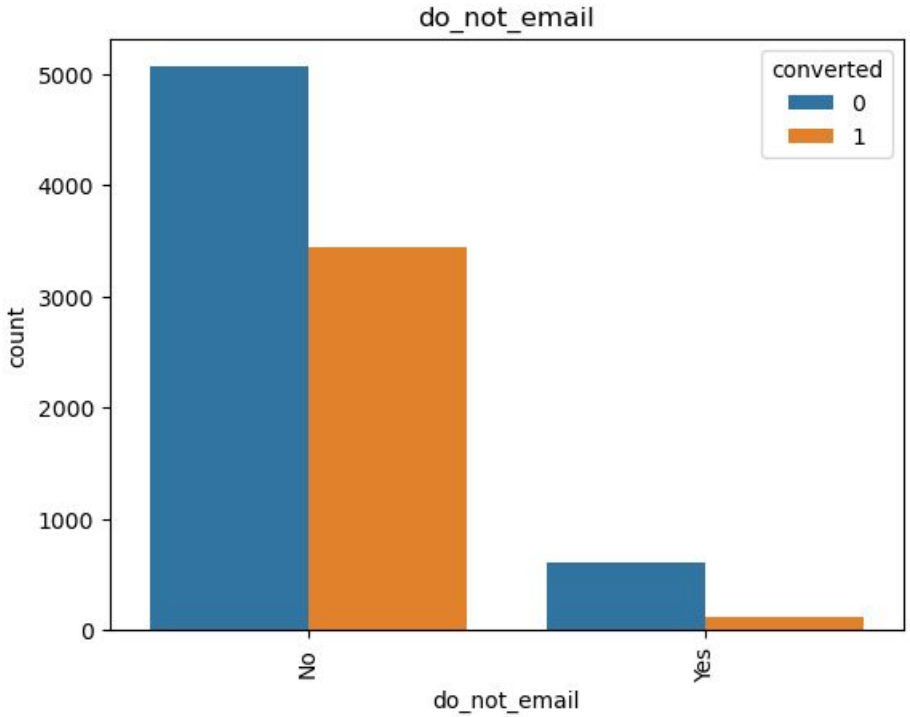

Fig vi

- If a lead is a working professional, it is more likely that they will convert as opposed to someone who isn't working (unemployed, students and others)(Fig vii)

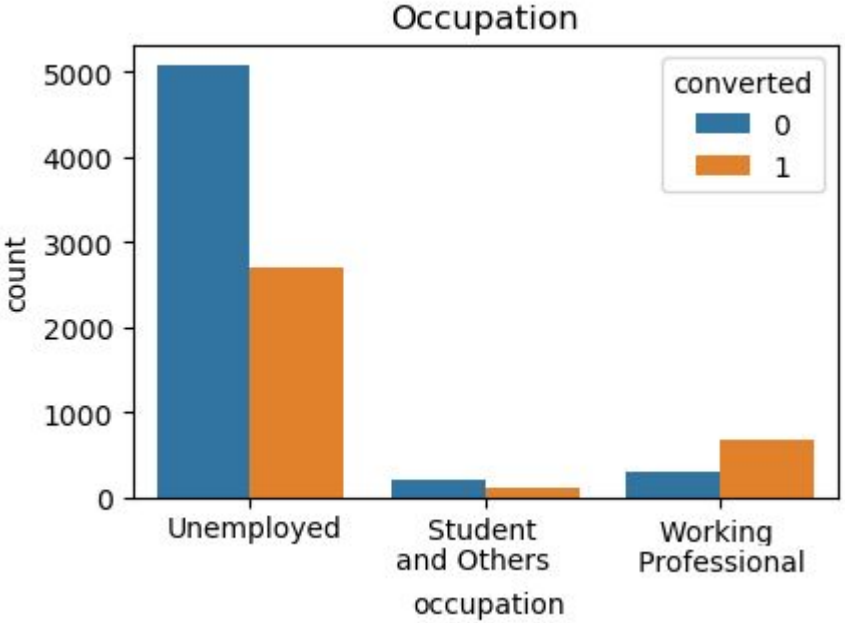

Fig vii

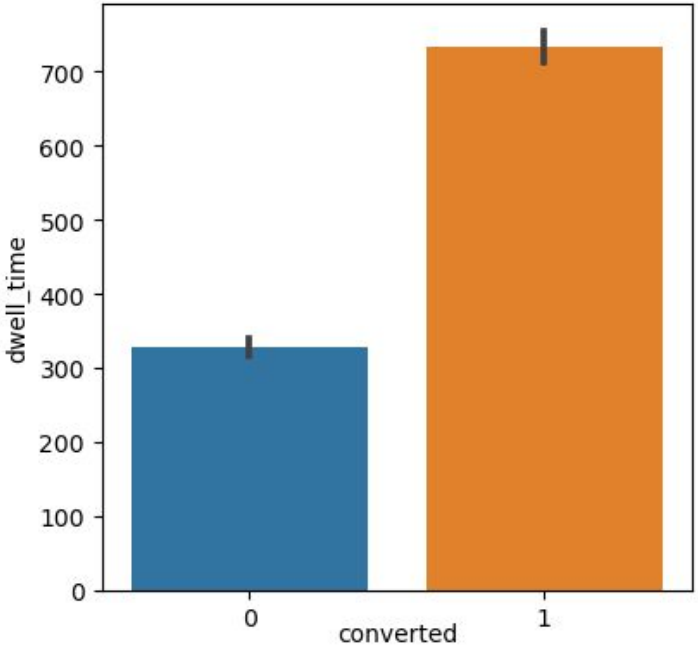

Fig viii

- Customers who spend a lot of time on our website show significantly greater chances of being 'hot leads' (Fig viii)

- Leads which visit our site more, our more likely to buy our course (perhaps signalling growing interest and conviction). (Fig ix)

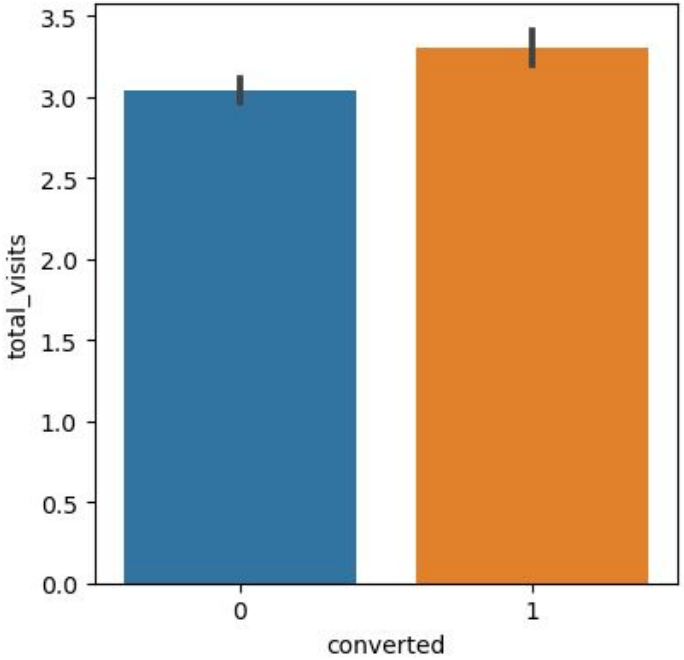

Fig ix

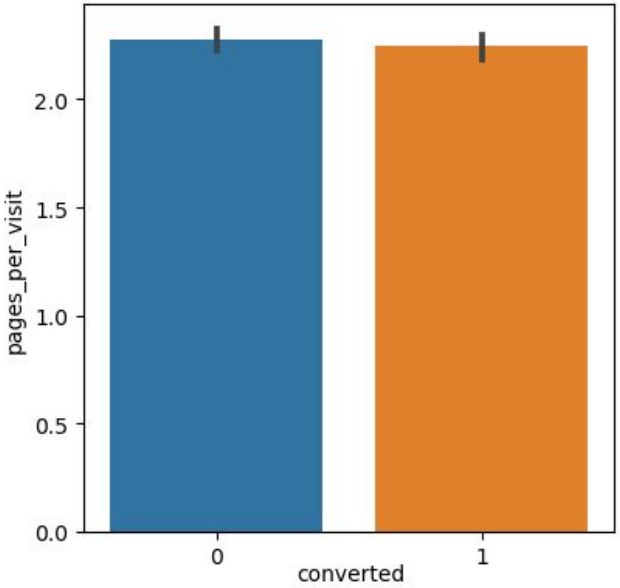

Fig x

- Leads which view more pages per visit on our site are less likely to convert. Perhaps due to paralysis by analysis (Fig x)

Recommendations

- References and Wellingak websites show the highest rate of conversion. They just don’t have the same traffic as google or Olark. Perhaps we could help drive more traffic there and as a result get more customers.

- Leads through Google do show good conversion rate, however perhaps it isn’t as high as compared to the previous 2, because the competition is outdoing us in terms of SEO (search engine optimization), other marketing or course structure and material. We’ll have to invest in digital marketing via SEO and make our presence known on other websites where our presence isn’t significant (like facebook etc.)

- One of the reasons people choose to not receive emails is probably because they consider reminders to be annoying/ not necessary. We can add a small question about practical applications of the course that our lead was checking out. Practical applications help make it more palpable and gets the person interested. We could offer insight from the course from different chapters with our mails which is not necessarily known in the public domain. However, the mechanism behind the insight would of course be behind the paywall.

- Since the highest conversion is amongst working professionals, we should structure our course to suit their needs more and target them in their internet spaces (LinkedIn, facebook groups, subreddits, blogs, stackexchanges about specific professions)

- The UI of our site should be great, as those who are interested visit it more often and spend more time there. If the UI is not good, then the prestige that our website, courses or company could hold in their minds would be lost. A comfortable experience also help others think better of the company and hence the courses.

Flexibility/Tunability of the Model

1. In situations where the company wants to conserve resources and time and reach out to only those leads which are most likely to convert i.e. 'hot leads' -
   a) We'd want to maximize precision (i.e. out of the leads we predict will convert, most of them should)
   b)Keep the cut-off lead score high at 70, i.e. reach out to customers with lead scores above 70

2. In situations where the company wants to reach out to as many leads as possible even though they don't have a very strong chance to convert -
   a) We'd want to maximize recall (i.e. out of all the leads which could convert, we'd want to get to most of them)
   b) Keep the cut-off lead score low, below 30, i.e. reach out to customers with lead scores below 30

Thank You