upGrad

#LifeKoKaroLift

Lead Scoring Case Study

***AGENDA**: Develop a ML model to identify the most promising leads, to increase the conversion rate*

By
Nishit Chaudhry
Raj Praveen Pradhan

Overview

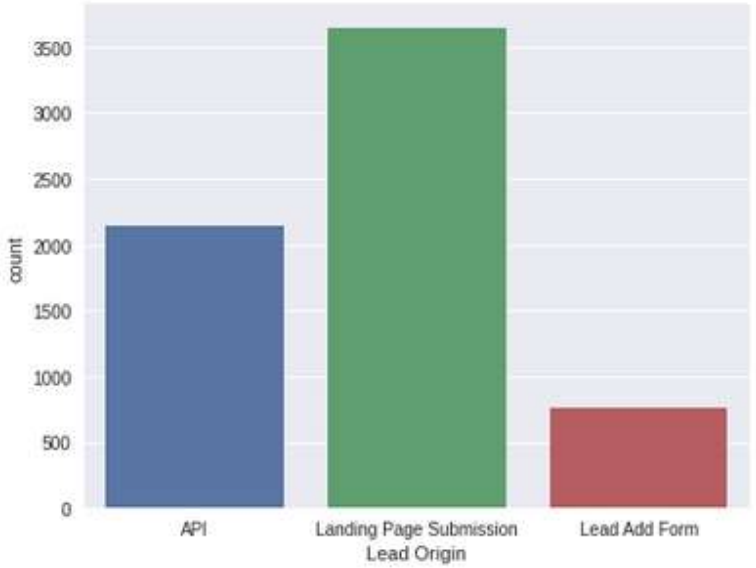
upGrad

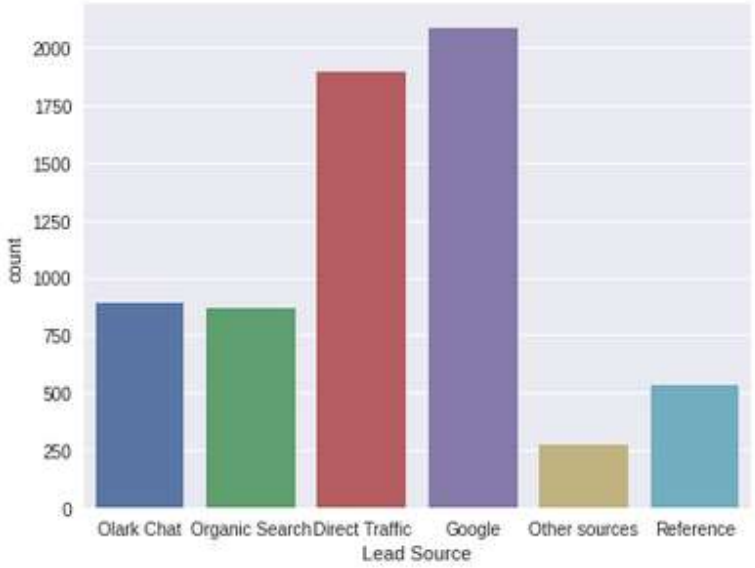
Problem Statement: *Our task is to build a model which needs to assign a lead score to each of the leads such that the customers with higher lead score have a higher conversion chance and the customers with lower lead score have a lower conversion chance.*

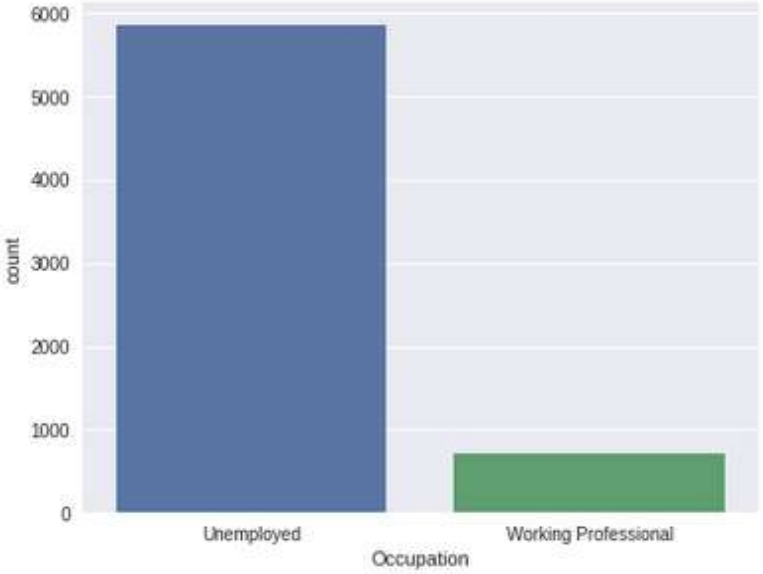
Approach*: Since this is a classification problem Logistic regression was used to calculate probabilities denoting lead conversion rates, which were later used to assign scores to leads. Furthermore to gain insights, EDA and various feature selections techniques were used based on which business recommendations were provided.*

EDA : *Univariate Analysis*

upGrad





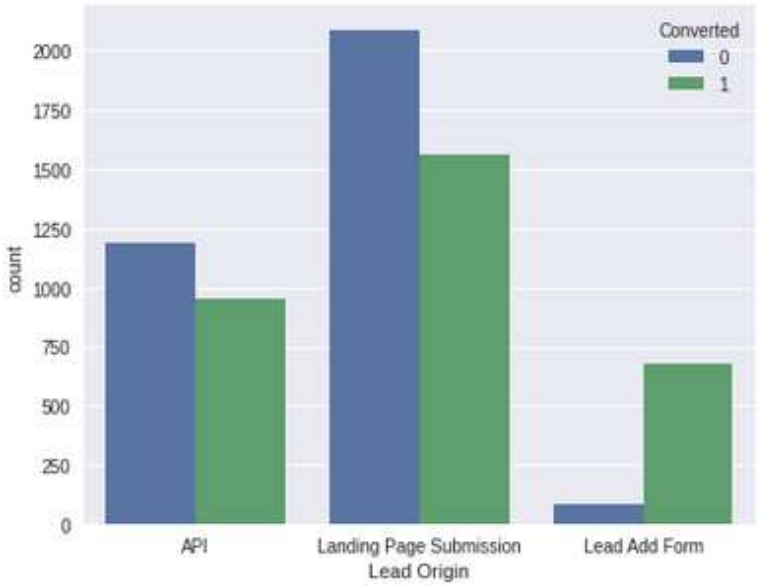


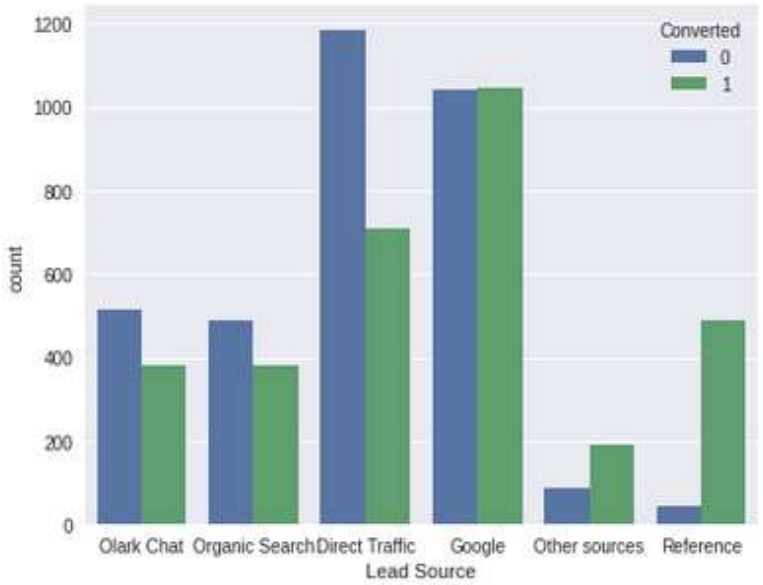
Inferences:

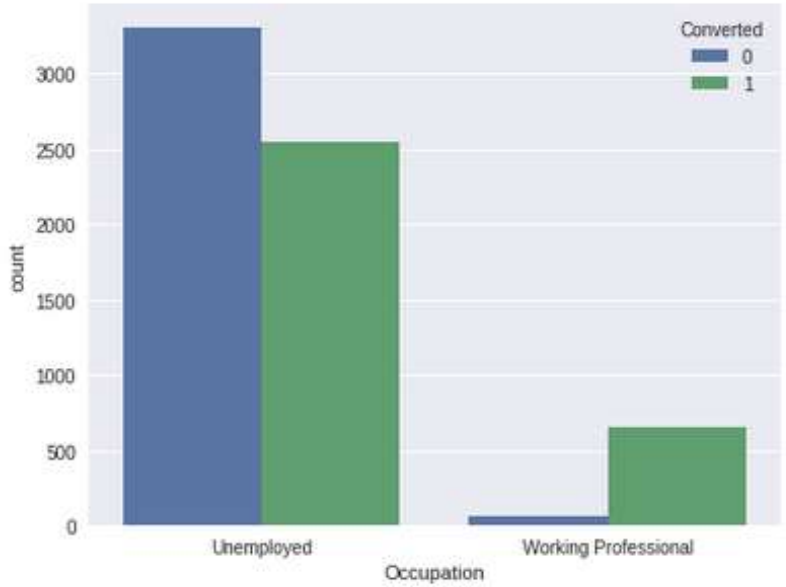
- Leads usually originates from landing page submission, followed by API.
- Google and Direct Traffic are two major sources of leads.
- As per plot, high distribution of lead falls under unemployed segment. Working professionals share only a small segment of customer base.

EDA : *Bivariate Analysis*

upGrad







Inferences:

•Leads which originates from "Lead add form" have significant chance of conversion.

•Chances of conversion are high when the lead sources from "references", whereas opposite is true for "Direct Search".

• Even though customer base of "Working Professionals“ is small they have significant conversion rate, which is probably due to better financial situation.

Feature Selection

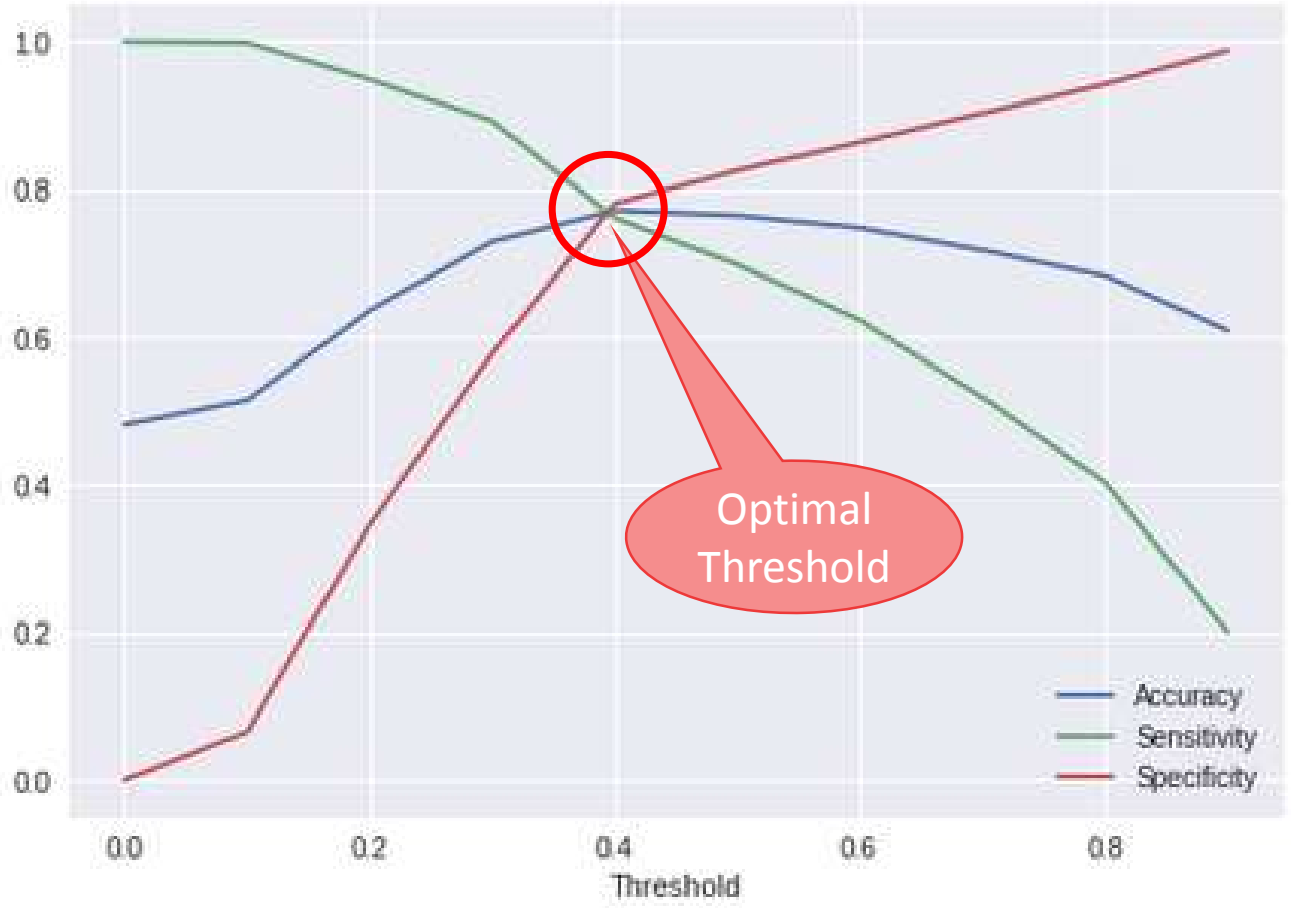
upGrad

Below is the list of features that are used for model building, These features holds statistical as well as business significance, same results were obtained in EDA also.

Final FEATURE LIST

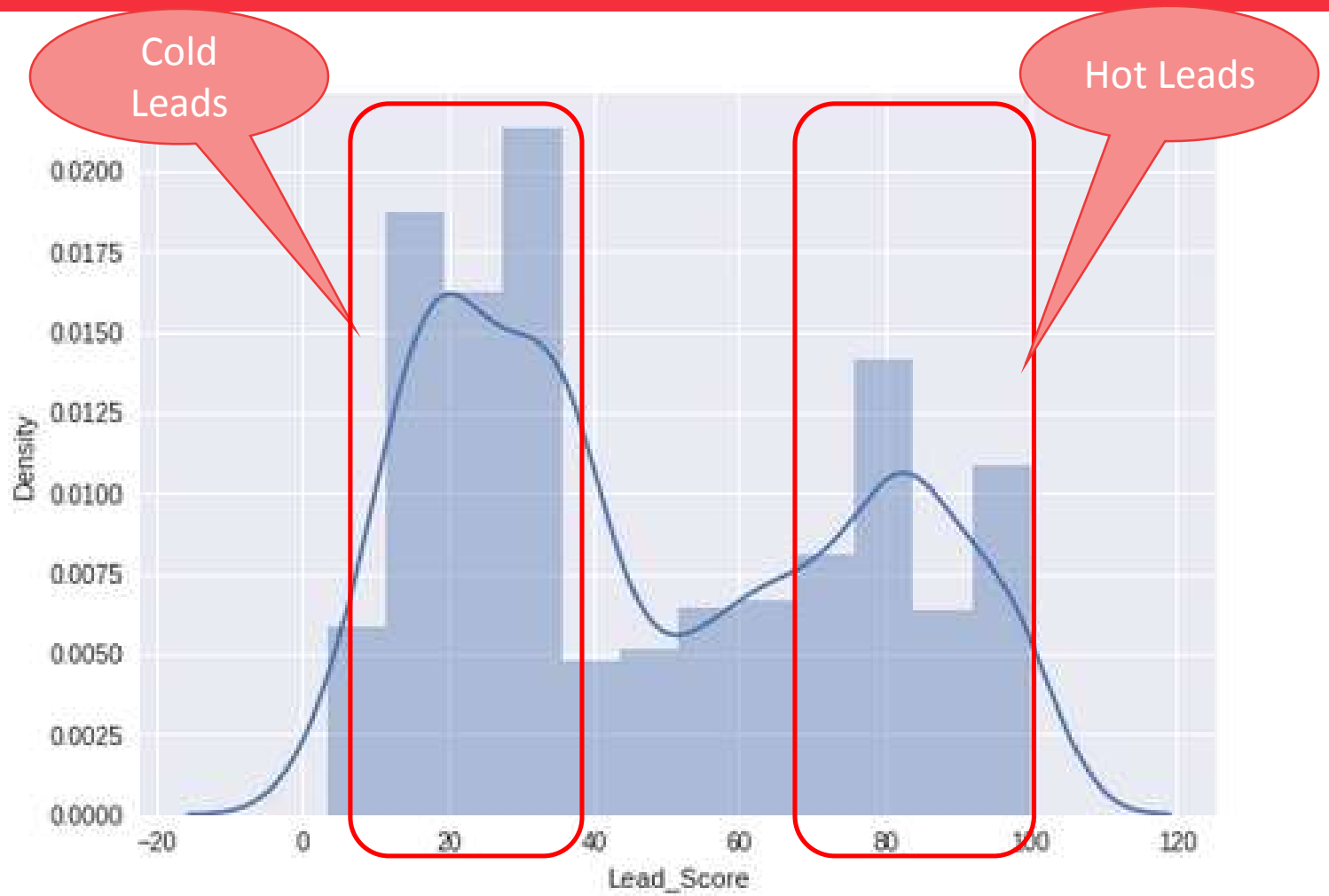
1. 'TotalVisits'
2. 'Total Time Spent on Website'
3. 'Page Views Per Visit'
4. 'Lead Origin_Lead Add Form'
5. 'Lead Source_Google'
6. 'Lead Source_Olark Chat'
7. 'Lead Source_Organic Search'
8. 'Do Not Email_Yes'
9. 'Occupation_Working Professional'

Determining optimal threshold for model



Based on mentioned plot 0.4 seems to be the optimal value of threshold. But for our business requirement i.e. to obtain a recall of round 80%, we proceeded with a suboptimal threshold i.e. 0.36 which provided generalized results for the model.

Generating Lead Score



After obtaining the lead scores on a scale of 0-100 with the probabilities that our model predicted, from the distribution plot of lead score we could observe that there were roughly two groups with lead score such as 10-36 and 70-100 that we can label as cold and hot leads. Thus to increase the conversion rate more resources need to “Hot leads”