Lead Scoring Case Study

Submitted by:

Supriya Kulkarni

Nimi C K

Rachit Pandey

Problem Statement

X education is an education company in the business of selling online courses to Industry professionals.

When these professionals land on the website of X education, some of them fill up a form providing their email address or phone number. These people are then classified as Leads. Some leads are also generated through past referrals.

Sales team contacts these leads and some of them get converted. Typical conversion rate is around 30%.

The CEO has given a target for lead conversion to be around 80%.

In order to achieve it, we have to make an ML model and assign a lead score between 0 and 100 to each lead. Based on the logistics regression model and the requirement of the company, a cut-off lead score can be selected to classify if the lead is hot or not.

Analysis Approach

This problem is a binary classification problem and can be solved using logistics regression model.

Our objective is to identify hot leads from the pool of initial leads. For doing this we will assign a probability of conversion to each of the initial lead.

For the given business case the objective is to achieve a Recall value of approx 0.8

Based on the data provided, following steps were followed:

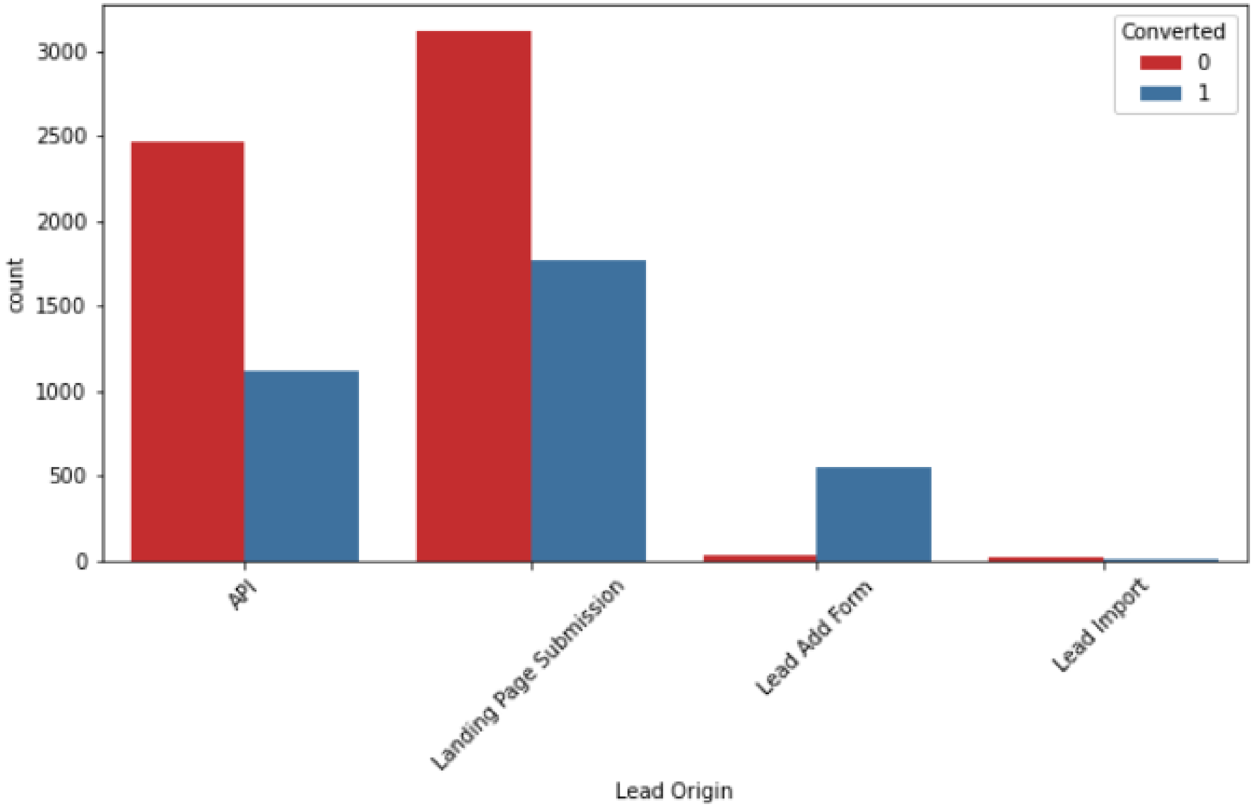
1. Data Cleaning
2. Exploratory Data Analysis (EDA)
3. Data Preparation
4. Modelling & Evaluation

Data Cleaning

- Handling "Select" level for categorical variables: "Select" is basically equivalent to null in this case and has been replaced accordingly.
- Columns having more than 40% missing values (null) were dropped.
- "Specialization" column has 37% null values. For this a separate category- "Others" was created as perhaps a suitable option was not available for the lead to choose.
- "Tags" column has 36% null values. These have been imputed with the mode value, as "Tags" is a categorical variable.
- Some columns have very high contribution from a unique category (high skew). So we dropped such columns.
- We know that missing values are imputed with mean/median for continuous variables and mode values for categorical variables. This principle was followed for remaining columns also where the missing values were still large (>25%)
- Missing values still left were less than 2%, hence these rows were dropped. Therefore 98% of the rows were retained after cleaning the data

Exploratory Data Analysis (EDA)

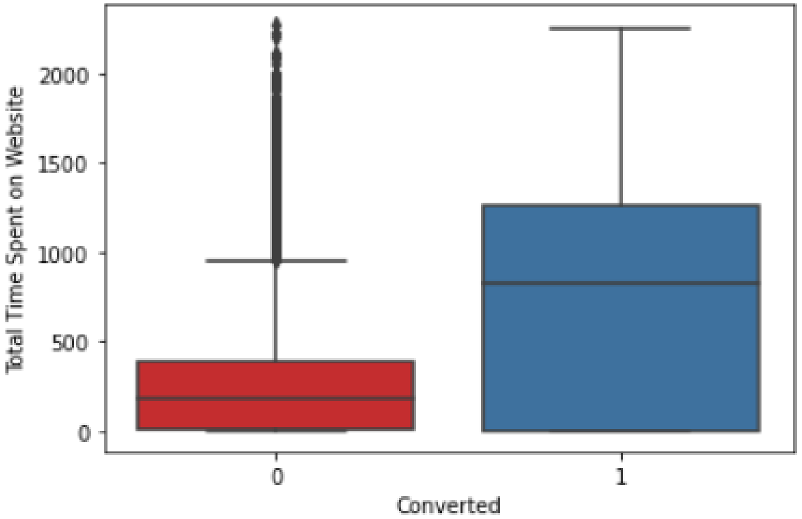
- The provided data shows that the conversion rate is approx 38%
- Lead origin:

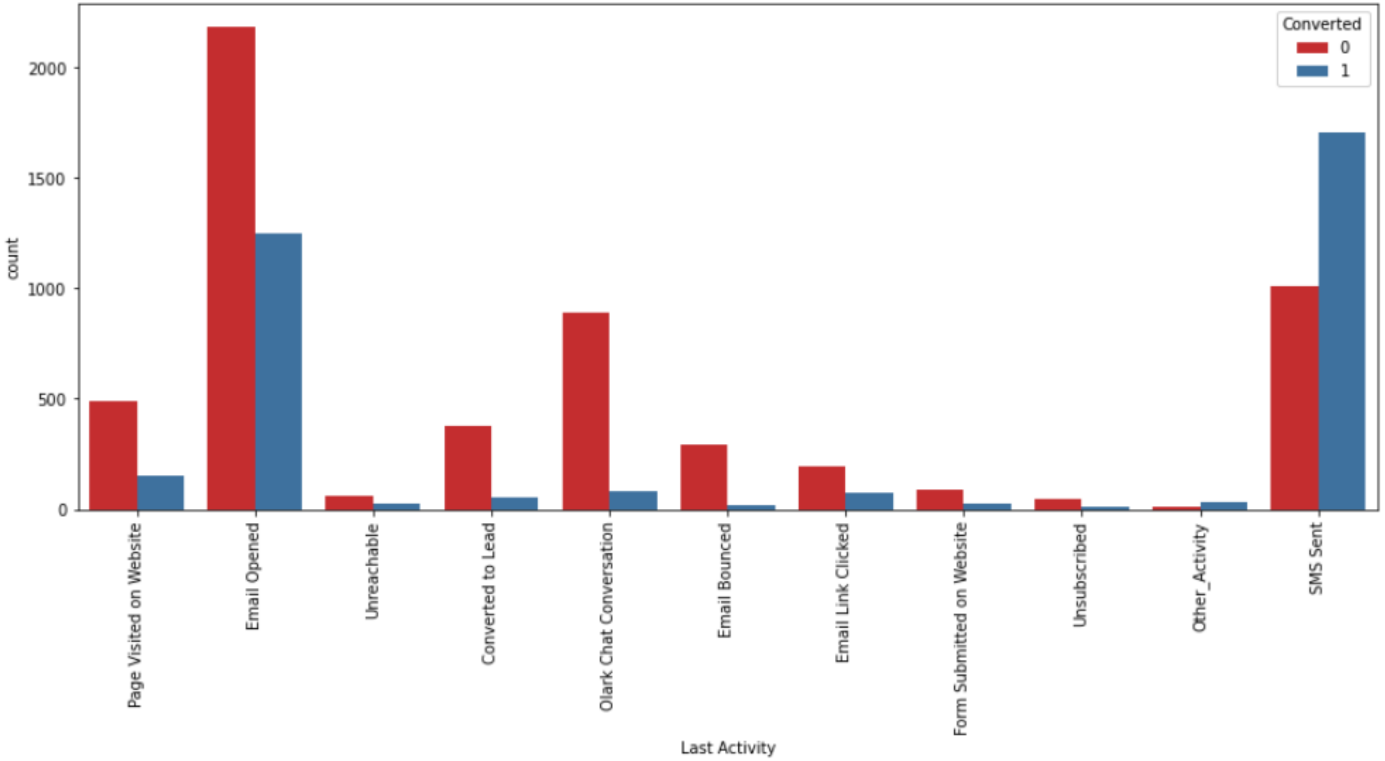


- Landing page submission and API are the most important origins and we need to focus more for improving conversions

Exploratory Data Analysis (EDA)

- Lead Source: Except for top 9 categories others have very less values. So all the remaining categories have been merged into a new category "Others".
- Outliers in some of the variables ("TotalVisits", "Total Time Spent on Website", "Page views per visit" etc) were dropped based on cut-off percentile.

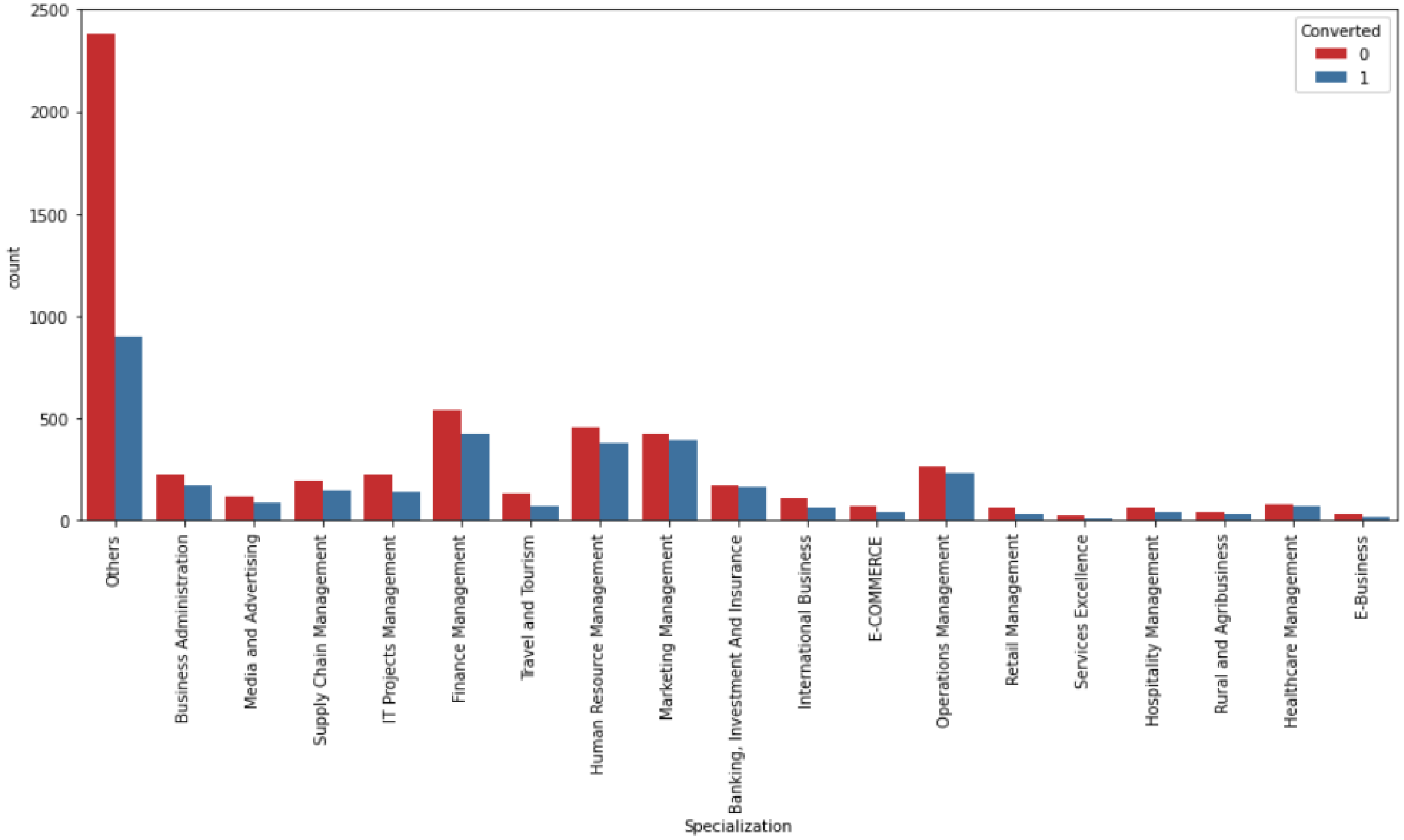




- As shown above, persons who spend more time on the website have more possibility for conversion
- Similarly Conversion rate for leads with last activity as "SMS sent" is very high.

Exploratory Data Analysis (EDA)

- Leads who specify clear specialisation have higher possibility of getting converted as compared to "others"



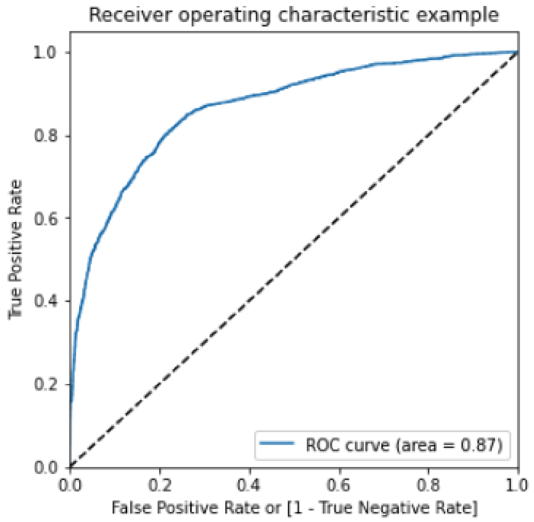
- Similarly, Working Professionals have very high conversion rate as compared to Unemployed
- Column "Tags" was removed as it was made by the sales team and hence should not be used.
- Some other columns like "Country","Search","Magazine" etc were dropped as they were not adding any relevant information to the model"

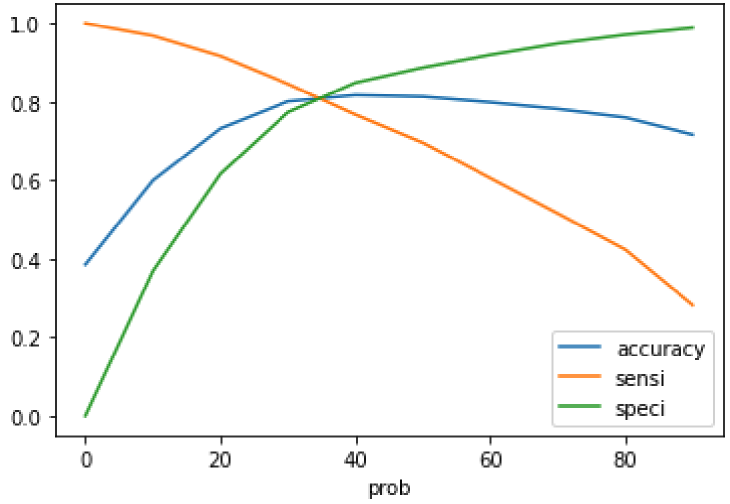
Data Preparation

- Binary variables(Yes/No) like "Do Not Email", "Do Not Call" were converted to 1/ 0
- Dummy variables were created for categorical features like "Lead Origin", "Lead Source","Last Activity" etc. And then these columns were dropped
- Data was split into Train and Test data in the ratio of 70% and 30% using train_test_split from sklearn
- Scaling of features was performed using StandardScaler from sklearn

Modelling & Evaluation

- LogisticRegression model from sklearn was used for modelling
- RFE (recursive feature elimination) was done for Feature Selection.
- Feature elimination was done one by one using criterion of p-value <0.05 and VIF <= 5.
- p-value was given priority over VIF for feature elimination.
- Initially we started with Lead probability cutoff at >50% for RFE.
- ROC curve and Accuracy, Sensitivity and Specificity were plotted:





- Based on above the optimal cutoff is between 30-40. Further checking with a cutoff of 35 was done.
- This resulted into Accuracy of 81.29% and Recall of 80.58% which is as per our target.

Modelling & Evaluation

- After this predictions were made on the test set
- The accuracy of 80.82% and Recall of 79.07% . This shows that the model prepared by us is stable . Also Recall value is closer to our target of 80%.
- In this particular case we have decided a cutoff of 0.35 to achieve a recall of approx 80%. Depending on business need this cutoff can be changed. If we have more resources in sales team we can lower it to include more leads so that absolute number of conversions can be increased.

Other Key Insights

- Persons who spend more time on the website have higher probability of conversion
- Conversion rate for leads with last activity as "SMS sent" is very high
- Working Professionals have very high conversion rate as compared to Unemployed