Lead Score Case Study

PROBLEM STATEMENT

Introduction:

- An education company, XEducation sells online courses to industry professionals. The company markets its courses on various websites and search engines such as Google.

- Once 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. The typical lead conversion rate at Xeducation is around 30%.

PROBLEM STATEMENT

BUSINESS GOALS:

- Company wishes to identify the most potential leads, also known as “Hot Leads”.
- The company needs a model wherein a lead score is assigned to each of the leads such that the customer with higher lead score have a higher conversion chance and customer with lower lead score have a lower conversion chance.
- The CEO, in particular, has given a ballpark number for the lead conversion rate i.e. 80%.

OVERALL APPROACH

1. Data cleaning and imputing missing values.
2. Exploratory data analysis :univariate,bivariate and multivariate analysis.
3. Feature scaling and dummy variables creation.
4. Logistic Regression model building.
5. Model Evaluation :specificity,sensitivity,precision and recall.
6. Conclusion and Recommendations.

PROBLEM SOLVING METHODOLOGY

Data Cleaning and Preparation

- Read data from source.
- Convert data into clean format suitable for analysis.
- Remove duplicate data.
- Outlier treatment.
- Exploratory data analysis.

Splitting the data and feature scaling

- Splitting the data into train and test dataset.
- Feature scaling of numerical variables.

Model Building

- Feature selection using RFE, VIF and p-value.
- Determine optimal model using Logistic Regression.
- Calculate various evaluation metrics.

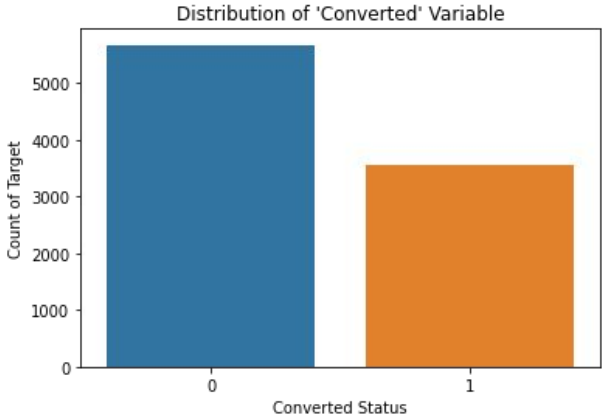
Result

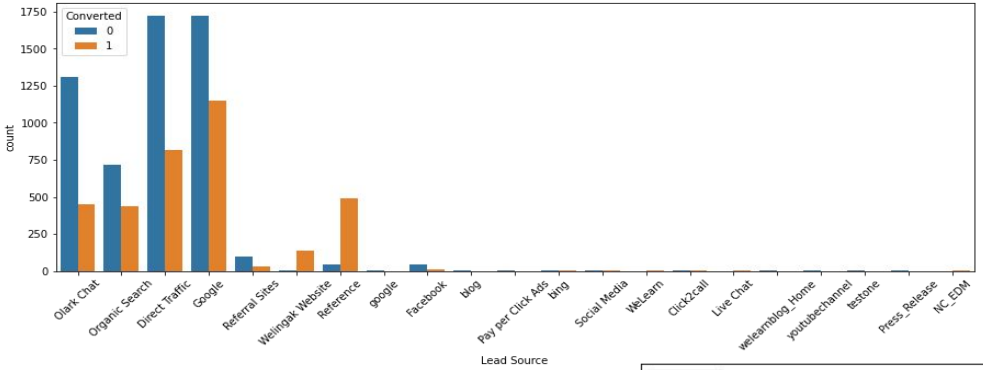
- Determine Lead score and check if target final prediction is greater than 80%conversion rate.
- Evaluate final prediction on test set.

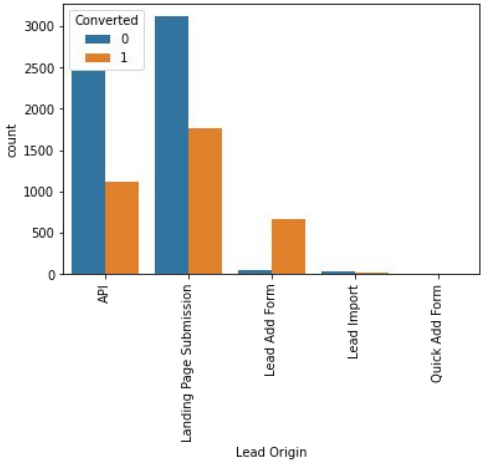
DATA CONVERSION

- Checking count of missing values.
- Finding the value_counts for all the missing values.
- Dropping the values which contain more null values.
- Dropping columns which are unnecessary.
- Dropping the rows in the dataset which has null values.
- Finding the index and total rows of the dataset.

EXPLORATORY DATA ANALYSIS



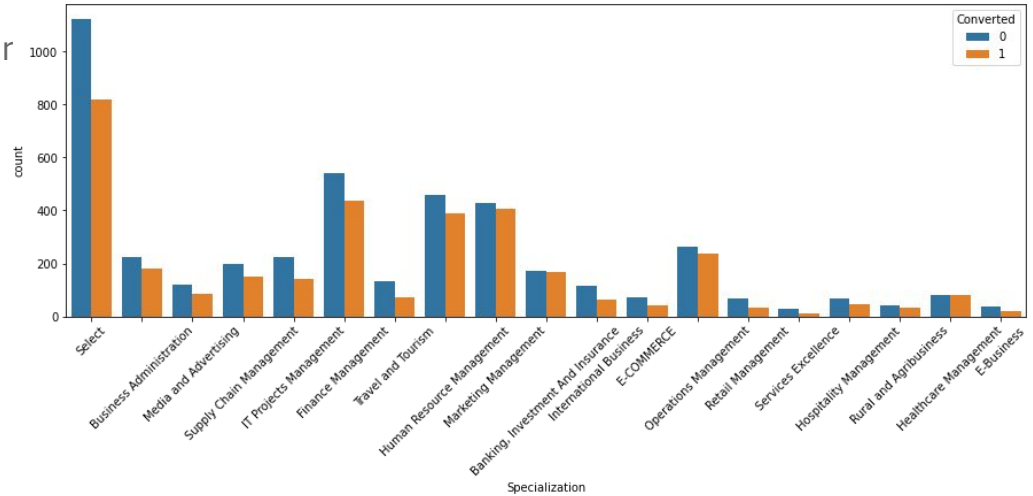


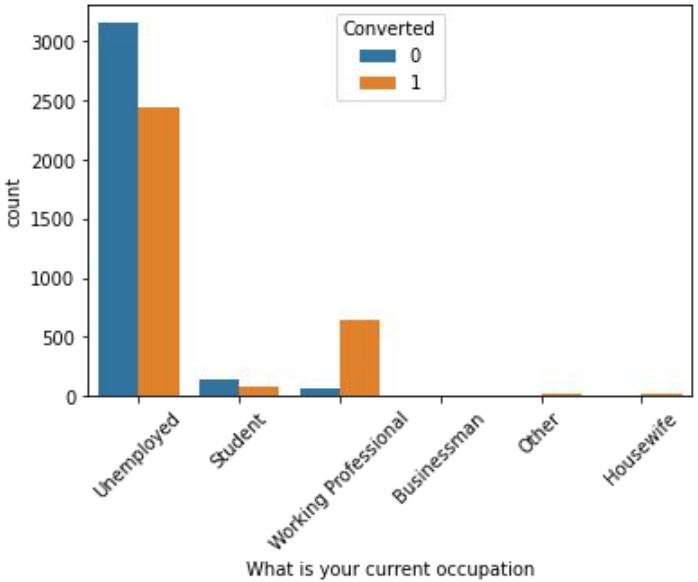


- We have a conversion rate of 30%.
- The count of leads from the Google and Direct Traffic is maximum.
- The conversion rate of the leads from Reference and Welingak Website is maximum.
- API and Landing Page Submission has less conversion rate(~30%) but counts of the leads from them are considerable.
- The count of leads from the Lead Add Form is pretty low but the conversion rate is very high.

EXPLORATORY DATA ANALYSIS

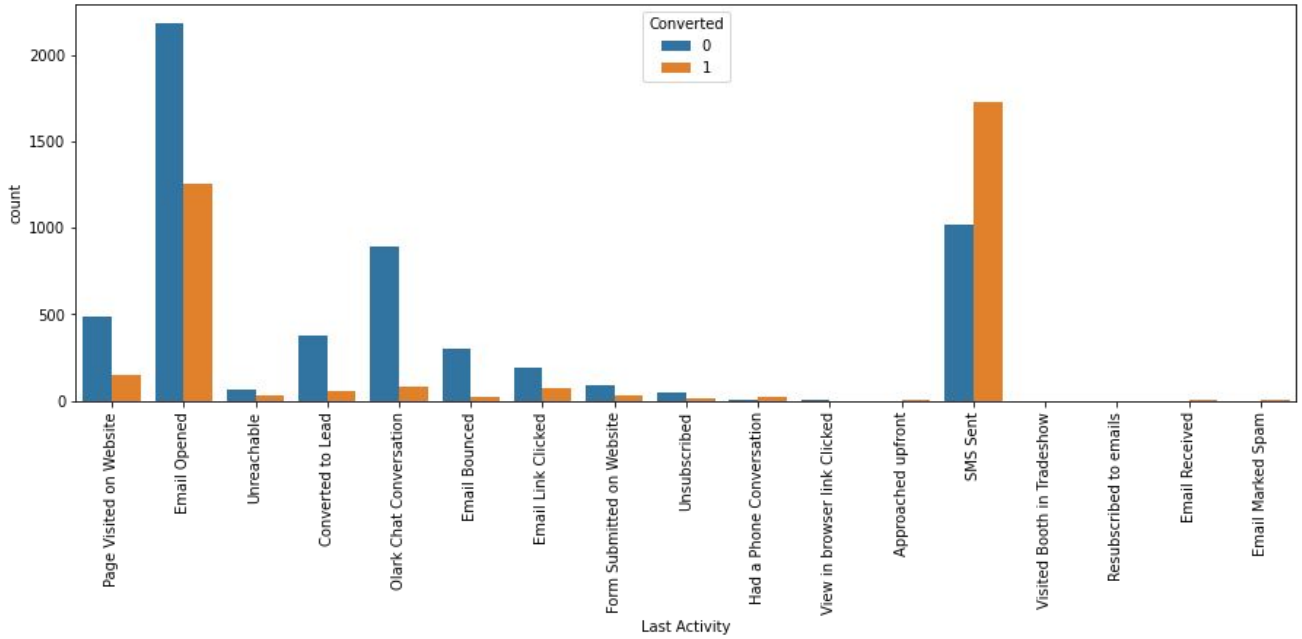
- Looking at the plot on the right, no particular inference can be made for Specialization.





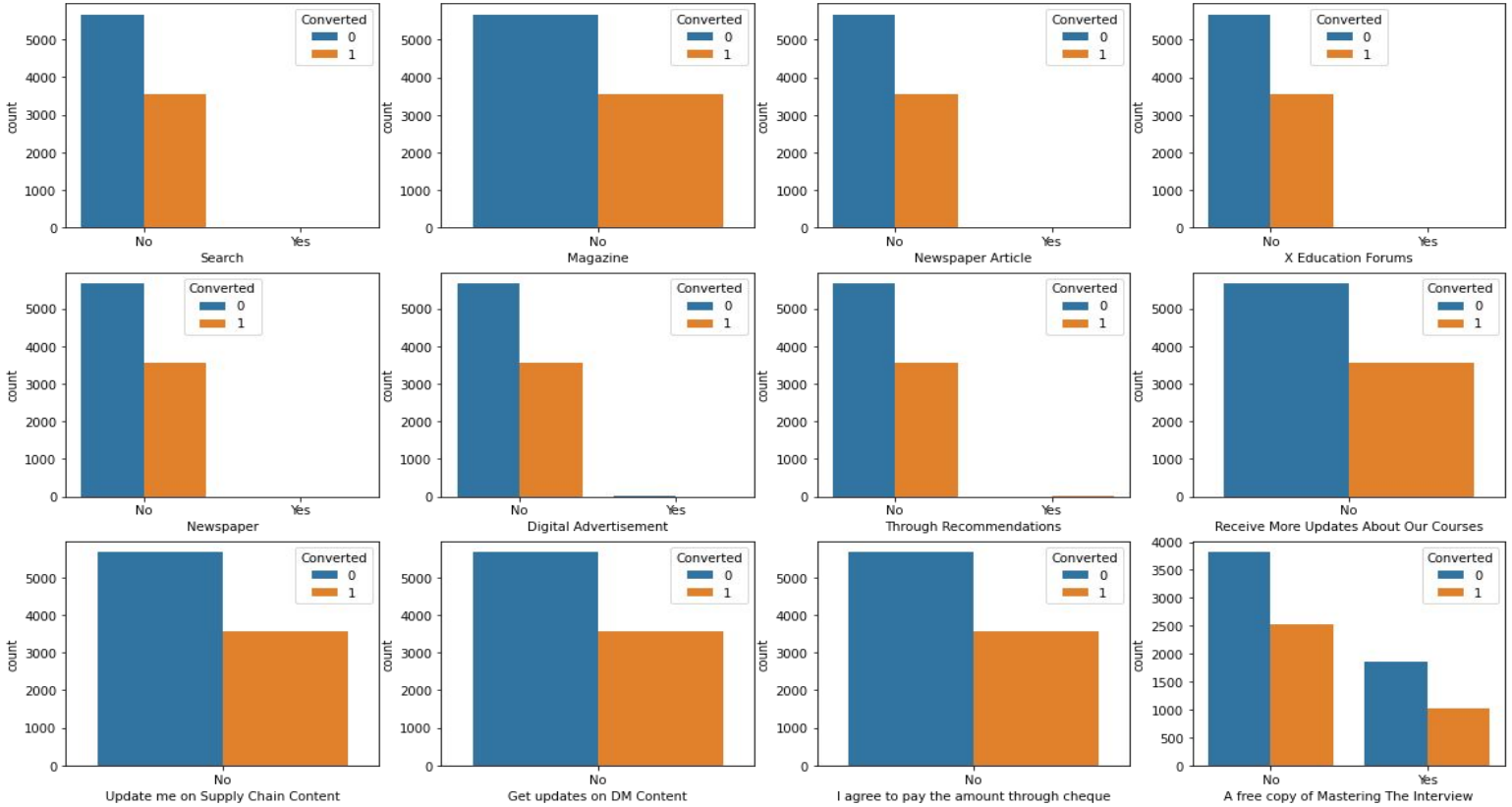
- Looking at the plot on the left, we can say that working professionals have high conversion rate.

EXPLORATORY DATA ANALYSIS



- 'Will revert after reading the email' and 'Closed by Horizzon' have a high conversion rate.

EXPLORATORY DATA ANALYSIS



EXPLORATORY DATA ANALYSIS

- From the graphs of the features on the previous slide most of the data is highly imbalanced hence they were dropped.
- “A free copy of Mastering The Interview” is the only one that doesn’t have a high data imbalance.

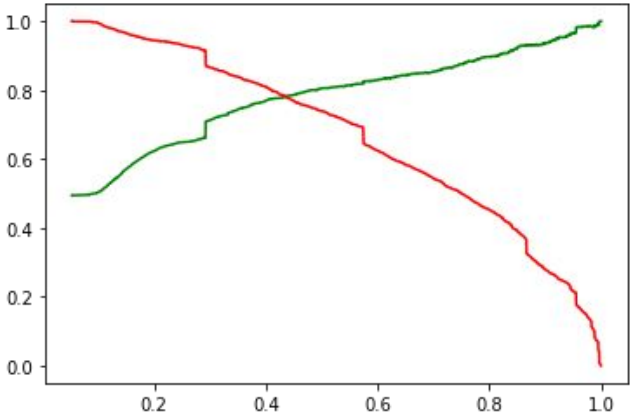
MODEL BUILDING

- Splitting the data into test and training sets.
- Split Ratio is 70:30 for train and test respectively.
- Use RFE to choose top 15 variables.
- Build the model by removing the variables whose p-value score is > 0.05 and

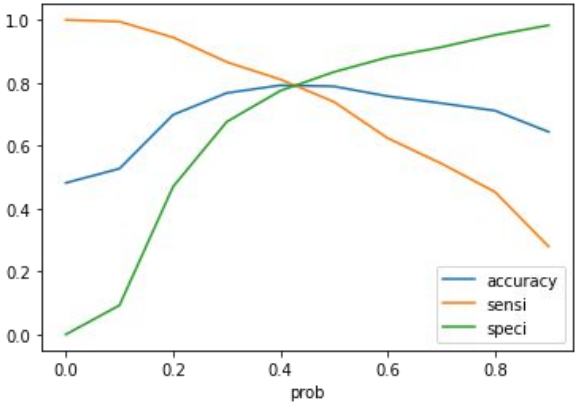
  VIF is > 5.

- Used the model on the test dataset to predict outcomes.

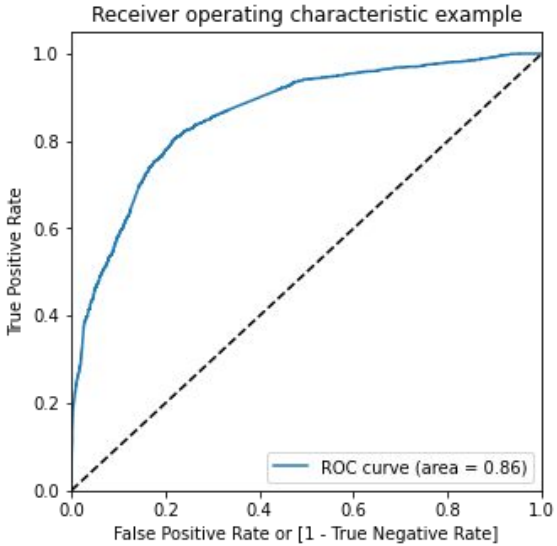
MODEL BUILDING



Precision Recall Tradeoff



Optimal Cutoff



ROC Curve

Threshold has been set to 0.44 for the final model

MODEL PREDICTION

Top Features:

```
Index(['TotalVisits', 'Total Time Spent on Website',
       'Lead Origin_Lead Add Form', 'Lead Source_Olark Chat',
       'Lead Source_Reference', 'Lead Source_Welingak Website',
       'Do Not Email_Yes', 'Last Activity_Had a Phone Conversation',
       'Last Activity_SMS Sent', 'What is your current occupation_Housewife',
       'What is your current occupation_Student',
       'What is your current occupation_Unemployed',
       'What is your current occupation_Working Professional',
       'Last Notable Activity_Had a Phone Conversation',
       'Last Notable Activity_Unreachable'],
      dtype='object')
```

Accuracy	78.66%
Precision	78.29%
Sensitivity	76.75%
Specificity	80.42%

Test_data confusion matrix

Predicted Actual	Not converted	converted
Not converted	1823	489
converted	444	1705

CONCLUSION AND RECOMMENDATIONS

1. The logistic regression model is used to predict the probabillty of conversion of a customer.
2. While we have calculated both sensitivity-specificity as well as Precision-Recall metrics, we have considered optimal cut off on the basis of sensitivity-specificity for final prediction.
3. Accuracy, Sensitivity and Specificity values of test set are around 79%, 77% and 80% respectively which are approximately closer to the respective values calculated using trained set.
4. Important features responsible for good conversion rate or the ones' which contributes more towards the probability of a lead getting converted are:
    - Lead Origin_Lead Add Form.
    - What is your current occupation_Working Professional.
    - Total Time Spent on Website.
5. Hence, the overall this model seems to be good.