# Lead Scoring Case Study

Sharanya Hegde

# Index

- Problem Statement
- Objective
- Analysis approach
- EDA findings
- Model Findings
- Summary

# Problem Statement

An education company named X Education sells online courses to industry professionals. The company markets its courses on several websites and search engines like Google. Once someone provides their email or phone number, they are considered to be a Lead. A lead is primarily obtained through their website or referrals. Through the marketing teams, these leads are converted into Sales. Currently, the conversion rate is only around 30%.

# Objective

To build a model wherein a lead score is assigned to each of the leads such that the customers with a higher chance of conversion (Hot Lead) is assigned a higher lead score, as compared to those that are less likely to be converted to a customer (Cold Lead).

# Analysis Approach

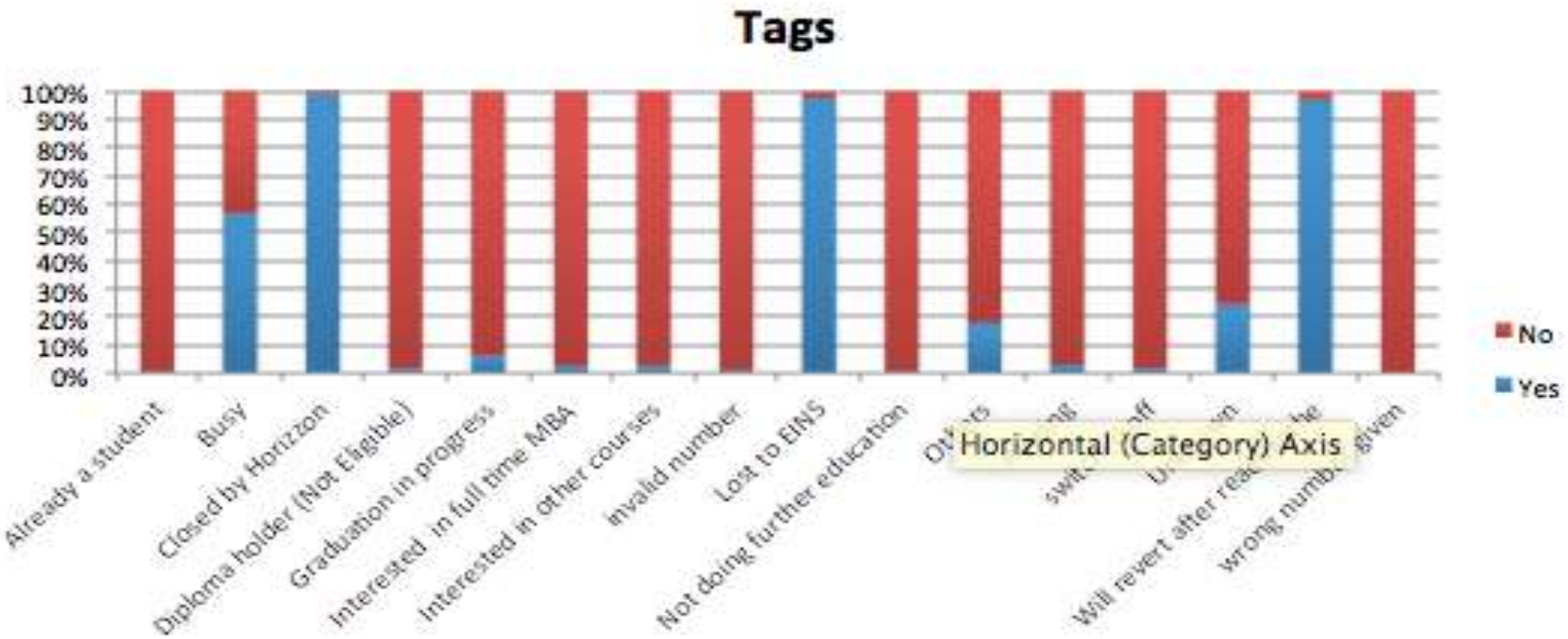
- Importing libraries
- Reading and understanding the data
- Cleaning the data
- EDA and dealing with outliers
- Getting dummy variables
- Model Building
  - Test-train split
  - Feature scaling
  - Dealing with highly correlated variables

# Analysis Approach

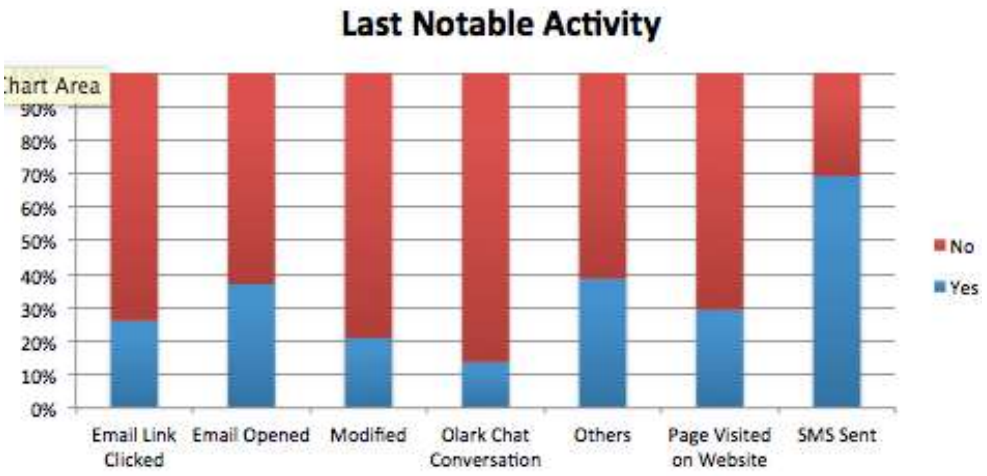
- Training the model:
  - Feature selection using RFE
  - Assessing the model using StatsModel and further eliminating variables
  - Using VIF to further eliminate variables
- Determining the performance of the model:
  - Confusion matrix
  - Plotting the ROC curve
  - Finding the optimal cut off point
  - Precision and recall trade off
- Making predictions using the test set
- Evaluating the model based on the test set

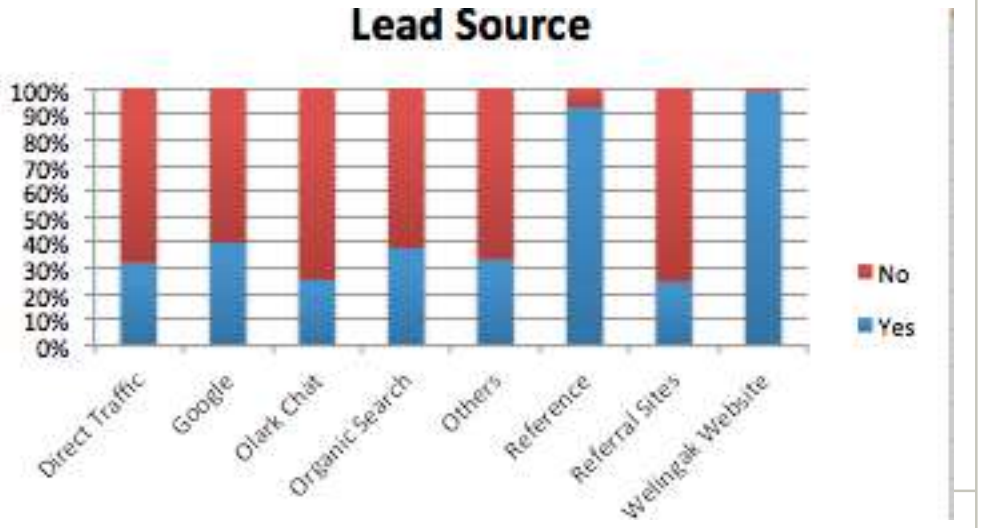
# EDA Findings

Through EDA we have observed that the following variables, have the most influence on increasing lead conversions, since many categories under these variables have conversion rates of more than the existing 38%.



# EDA Findings (continued)



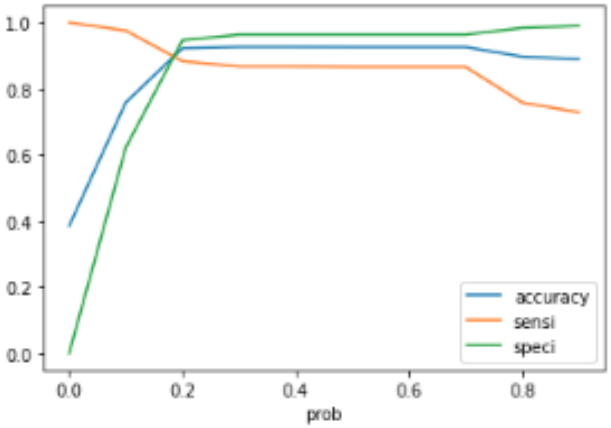
# Model findings

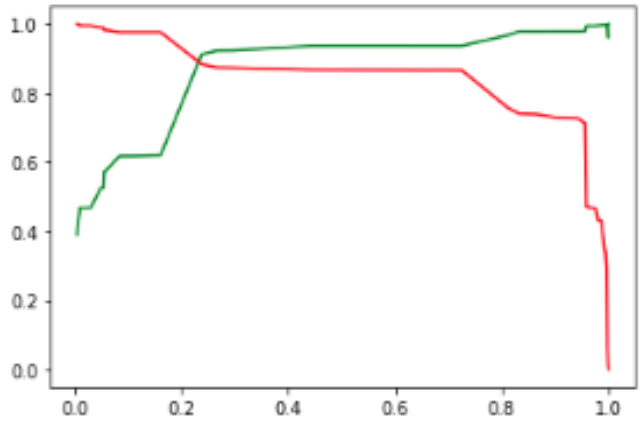
- The logistic regression model is used to predict the probability of converting a lead into a customer, given certain variables.
- Based on our understanding, the existing conversion rate was about 38%. Whereas the target was around 80%.
- We first used RFE to find the top 15 variables. And then used Stats Model (variables with high p-values) as well as VIF (variables with factors above 5) to further eliminate variables.

# Model findings (continued)

As per the Sensitivity-Specificity-Accuracy plot (fig 1), 0.2 was the optimal probability. Whereas as per the PR plot (fig 2), around 0.22 seems to be optimal.

We have concluded that the optimal cut off was found to be 0.2, i.e, any lead with a probability of higher than 0.2, is considered to be a hot lead (high chances of turning into a customer), and therefore should be targeted. Whereas any lead with a probability of 0.2 or lesser, is considered to be a cold lead.



# Model findings (continued)

Our final model was built using the following 11 features:

- Tags_Closed by Horizzon
- Tags_Lost to EINS
- Tags_Will revert after reading the email
- Lead Source_Welingak Website
- Last Notable Activity_SMS Sent
- Tags_Busy Tags_Unknown
- Tags_Unknown
- Last Activity_Olark Chat Conversation
- Tags_Ringing
- Tags_switched off
- Lead Quality_Worst

# Model findings (continued)

Out of these "Tags_Closed by Horizzon", "Tags_Lost to EINS", "Tags_Will revert after reading the email" are most significant with the highest absolute values of their factors. This indicates that these factors have the highest influence on the model.

```
Tags_Closed by Horizzon                          8.233323
Tags_Lost to EINS                                7.879748
Tags_Will revert after reading the email         6.616031
Lead Source_Welingak Website                     4.471552
Last Notable Activity_SMS Sent                   2.629012
Tags_Busy                                        2.307268
Tags_Unknown                                     1.824416
Last Activity_Olark Chat Conversation           -1.240321
Tags_Ringing                                    -1.833864
Tags_switched off                               -1.968857
Lead Quality_Worst                              -2.125494
const                                           -3.477180
dtype: float64
```

# Model findings (continued)

The final model has a sensitivity rate of around 88%, i.e. 88% have been correctly identified as hot leads that have been converted. And a specificity rate of 94%, i.e. 94% has been correctly identified as cold leads, that can't be converted.

| Train Set Confusion Matrix | | |
|---|---|---|
| | **Predicted No** | **Predicted Yes** |
| **Actual No** | TN = 3695 | FP = 210 |
| **Actual Yes** | FN = 286 | TP = 2160 |

| Test Set Confusion Matrix | | |
|---|---|---|
| | **Predicted No** | **Predicted Yes** |
| **Actual No** | TN = 1634 | FP = 100 |
| **Actual Yes** | FN = 122 | TP = 867 |

# Conclusion

The marketing team can now focus on those leads with higher Lead Scores to increase the conversion rate of X Education. Below is a sample of 10 of the leads obtained. Out of these, lead numbers 4223, 4216 and 1490 should be top priority.

| | Converted | Lead Number | Converted_Prob | final_predicted | Lead Score |
|---|---|---|---|---|---|
| 0 | 0 | 3271 | 0.160529 | 0 | 16 |
| 1 | 1 | 1490 | 0.956111 | 1 | 96 |
| 2 | 0 | 7936 | 0.160529 | 0 | 16 |
| 3 | 1 | 4216 | 0.994411 | 1 | 99 |
| 4 | 0 | 3830 | 0.160529 | 0 | 16 |
| 5 | 1 | 1800 | 0.956111 | 1 | 96 |
| 6 | 0 | 6507 | 0.062072 | 0 | 6 |
| 7 | 0 | 4821 | 0.004316 | 0 | 0 |
| 8 | 1 | 4223 | 0.998240 | 1 | 100 |
| 9 | 0 | 4714 | 0.160529 | 0 | 16 |

When the marketing team has more resources, the probability can be reduced, thereby increasing the pool of Hot Leads. Whereas, when there are limited resources in the marketing team, the probability can be increased, so that the pool of Hot Leads will be smaller. The marketing team must always approach those leads with the highest lead scores, first, as the chances of them being converted to customers, is the highest.