LEAD SCORE CASE STUDY

BY:

NIRBHAY TANDON AND NAVEEN SHARMA

PROBLEM STATEMENT & BUSINESS OBJECTIVE

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 these 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. Now, although X Education gets a lot of leads, its lead conversion rate is very poor. For example, if, say, they acquire 100 leads in a day, only about 30 of them are converted.

Business Objective:

- X Education wants it's Data Analyst team to help them select the most promising leads, i.e. the leads that are most likely to convert into paying customers.
- The company requires its Data Analyst team to build a model wherein they need 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.

METHODOLOGY

Data Analysis & EDA

- Checking the shape, columns, datatypes etc. of the dataset
- Check for duplicates
- Checking for null values
- Assessing outliers
- Dropping unnecessary columns
- Perform Univariate and Bivariate analysis

Data Preparation

- Creating Dummy variables for categorical columns
- Removing repeated columns
- Performing train-test split
- Performing scaling

Modeling

- Perform GLM analysis on the base model
- Use RFE to perform feature selection
- Calculate VIF & p values
- Build a model
- Assess parameters & repeat the previous two steps till we have a good model with no multicollinearity
- Generate prediction probabilities for our existing dataset
- Plot the ROC curve to assess our model
- Check the accuracy, specificity, sensitivity, precision & recall of our model

Generate Leads Table

- Join items with lead scores present in prediction score
- Present findings

Conclusion

- Summarizing the analysis
- Listing the top 5 analysis
- List top 3 features

EXPLORATORY DATA ANALYSIS

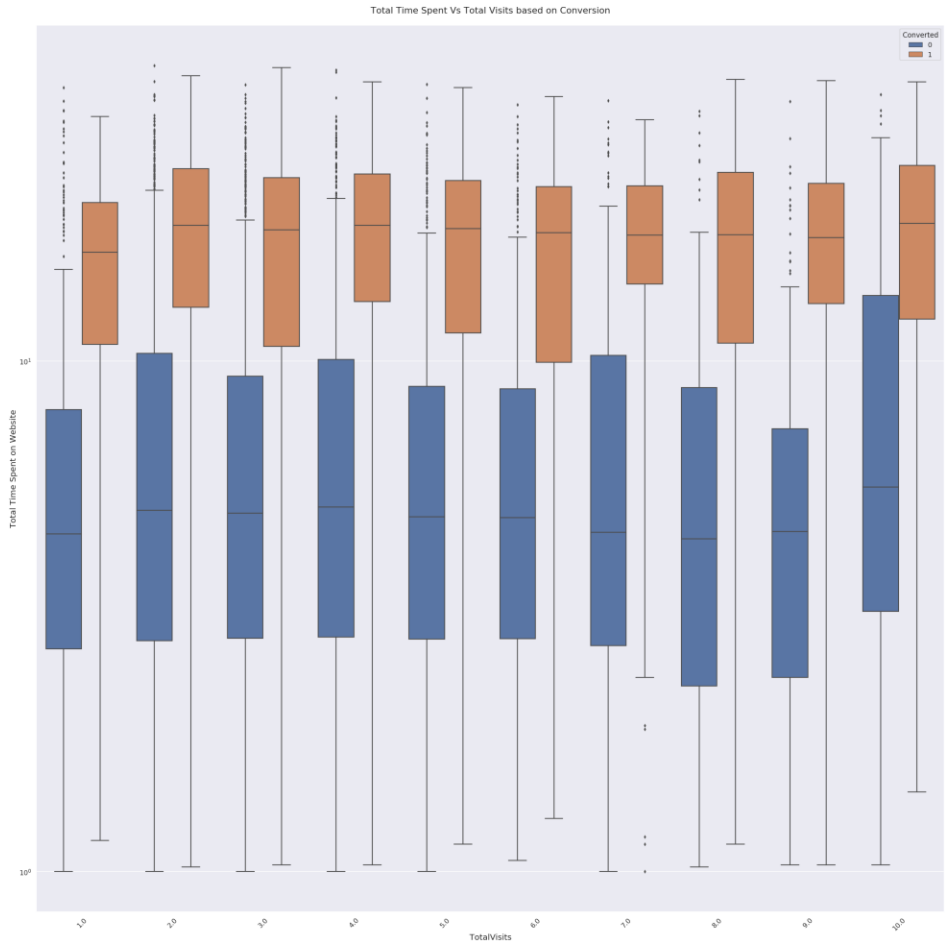
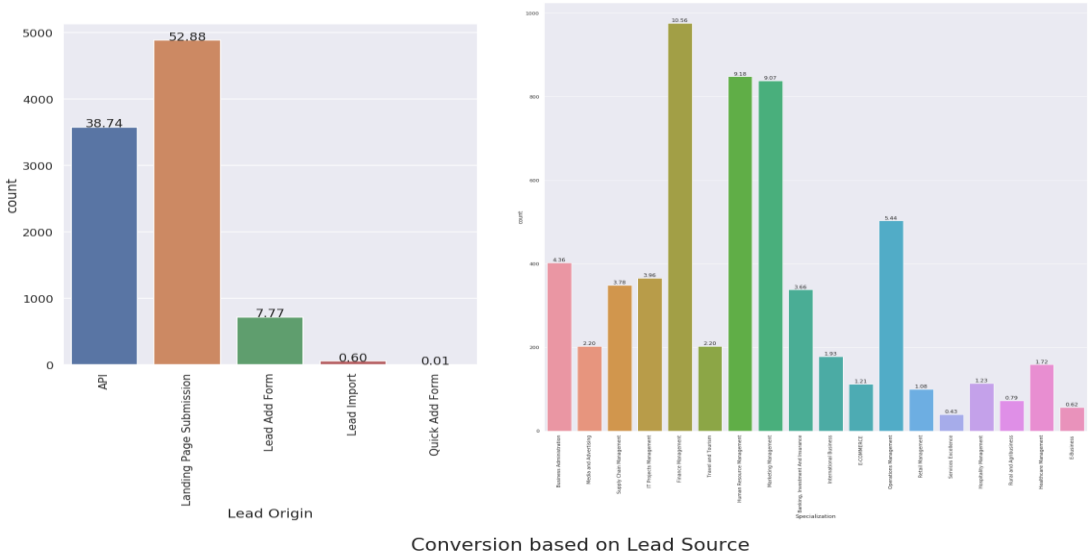


Image: Boxplot comparing Total time spent on website vs Total visits on the basis of conversion rate

- Our EDA over various columns consisted of the following steps:
  - Checking the shape, columns, datatypes etc. of the dataset
  - Check for duplicates
  - Checking for null values
  - Assessing outliers
  - Dropping unnecessary columns
  - Perform Univariate and Bivariate analysis
- The boxplot on the left shows a comparison between the Total Time Spent on a website vs Total Visits.
- We observe that the people who spent a lot of time on the website eventually converted to paying customers

EXPLORATORY DATA ANALYSIS



Conversion based on Lead Source

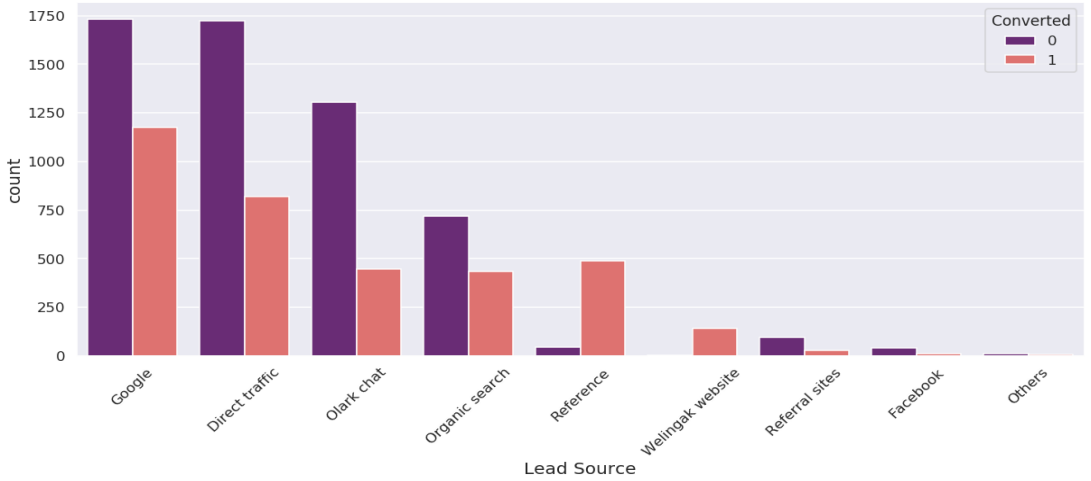


Image: Left: Conversions based on Lead Origin; Right: Specializations of current leads; Bottom: Conversion Rate of Leads based on Lead Source

- The landing page submissions lead to the highest source of our data. (ref: Top Left Graph)
- People who work in Finanace Management, Human Resources & Marketing constitue of almost 30% of our market. (ref: Top Right Graph)
- Referred candidates have the highest chance of converting to paying customers of the platform. (ref: Bottom Graph)
- People from Welingak website have 100% conversion rate as well
- 95% the applicants are looking for better career prospects
- 96% of our candidates are from India

EXPLORATORY DATA ANALYSIS

Heatmap After EDA

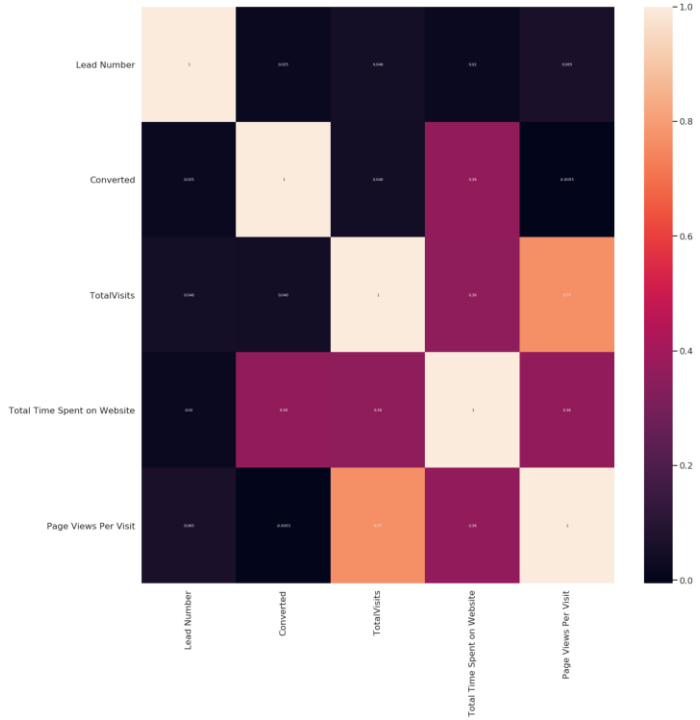


Image: Heatmap After EDA

Analysis

- After our EDA steps we finally selected 12 columns
- The heatmap on the left confirms that there is no multicollinearity in our data

MODELING

Inferences Drawn From Adjacent Plots

The two heatmaps on the right represent vaiables before and after RFE:

- We started with 56 variables
- Performed RFE on them and chose 20 variables
- We then assessed our model over various parameters such as p-values and VIF scores
- Finally we ended up with a model that had 15 variables

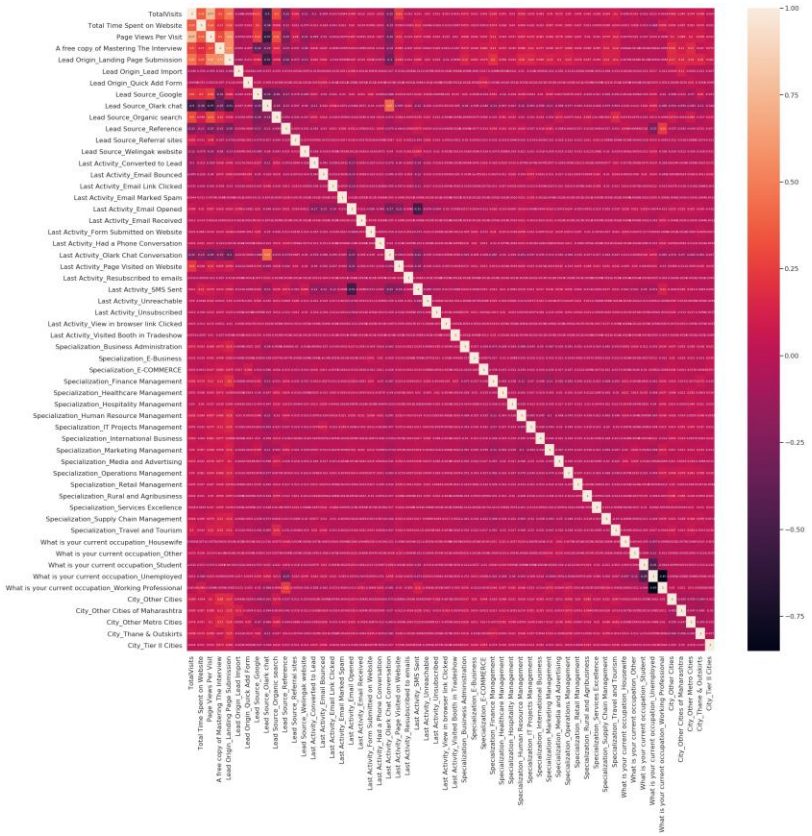


Image: Heatmap showing all 56 variables taken after dummy creation

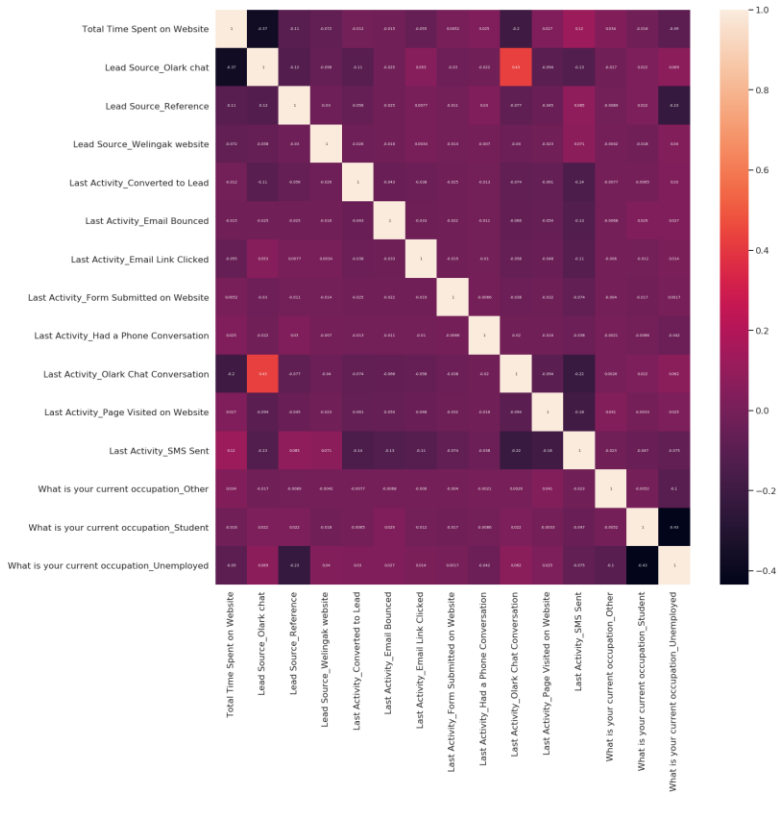


Image: Heatmap showing final 15 variables of the model

OPTIMAL CUTOFF PARAMETERS

Inferences Drawn From Adjacent Plots

- From the probability cutoffs we see that our optimal cutoff is 0.33
- The ROC curve has 87% of the area underneath it. This confirms our model can predict True Positive better.
- At the cut-off = 0.33, the various Model Performance parameters on test set are as per below
  - Sensitivity = 79%
  - Specificity = 80%
  - Accuracy = 80%
  - Recall = 80%
  - F1 – Score = 75%

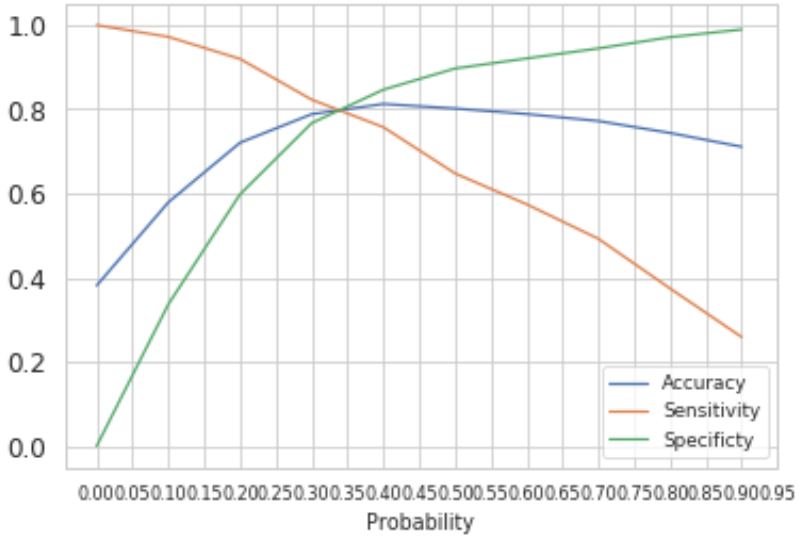
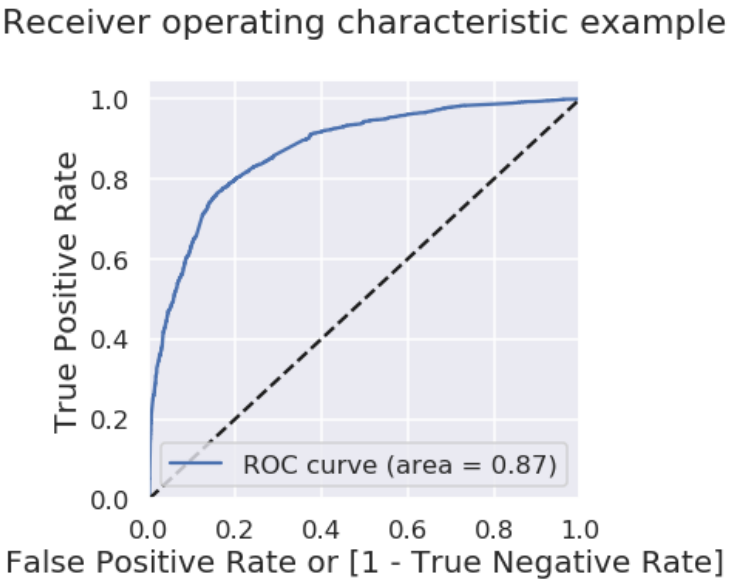


Image: Probability Cutoff Plot



GENERATING LEADS TABLE

From the adjacent table we can see:

- Lead numbers against Lead Scores
- The higher the lead score the better the chances to convert them into paying customers

Resulting Table Showing Lead numbers and Scores

Index	Lead Number	Lead Score
0	619003	69
1	636884	92
2	590281	67
3	579892	7
4	617929	79

CONCLUSION

Based on the analysis we have carried out, we can conclusively say that :

- Most applicants would like to join a course to have better career prospects
- X Education has the highest conversion rate of individuals who are referred to them
- Overall it is safe to say that the more time the user spends on the website, the better their chances of becoming a student.
- Hot Leads are identified as 'Customers having lead score of 33 or above'
- Sales Team of the company should first focus on the 'Hot Leads'
- Higher the Lead Score, higher the chances of conversion of 'Hot Leads' into 'Paying Customers'
- The 'Cold Leads'(Customer having lead score < 33) should be focused after the Sales Team is done with the 'Hot Leads' or during an off season.
- For converting customers the top 3 variables are:
  - Lead Source_Welingak website
  - Lead Source_Reference
  - Last Activity_Had A Phone Conversation