LEADS SCORING CASE STUDY

-Kumar Himanshu
-Harmeet Singh
-Md. Aseer Shaik

Business Problem Statement

Business Problem Statement:

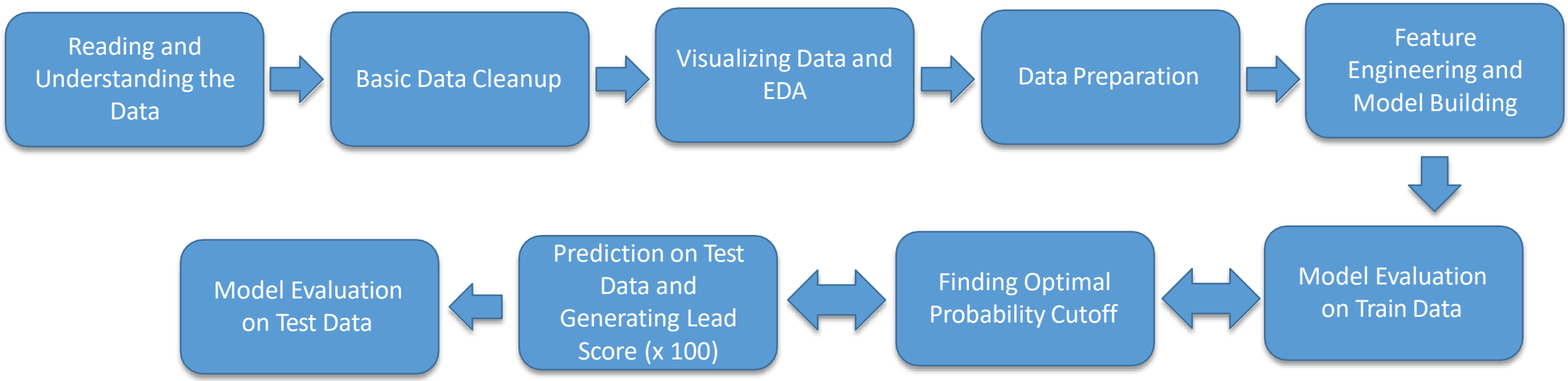
An education company named X Education sells online courses to industry professionals. On any given day, many professionals who are interested in the courses land on their website and browse for courses.

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. 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. Through this process, some of the leads get converted while most do not. The typical lead conversion rate at X education is around 30%.

Goal:

1. To identify the features that contributes to predict Lead Conversion.
2. Identifying Hot Leads by generating Lead Score for all leads, so that leads having higher Lead Scores can be contacted with priority for achieving Higher Lead Conversion Rate.

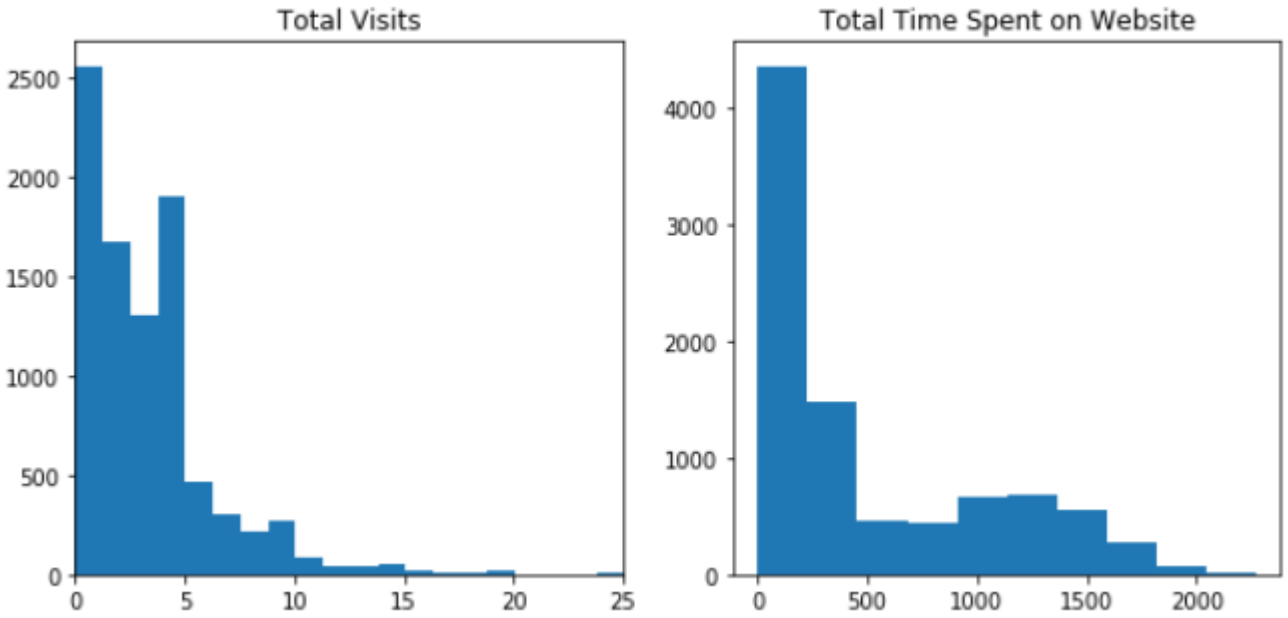
Overall Approach

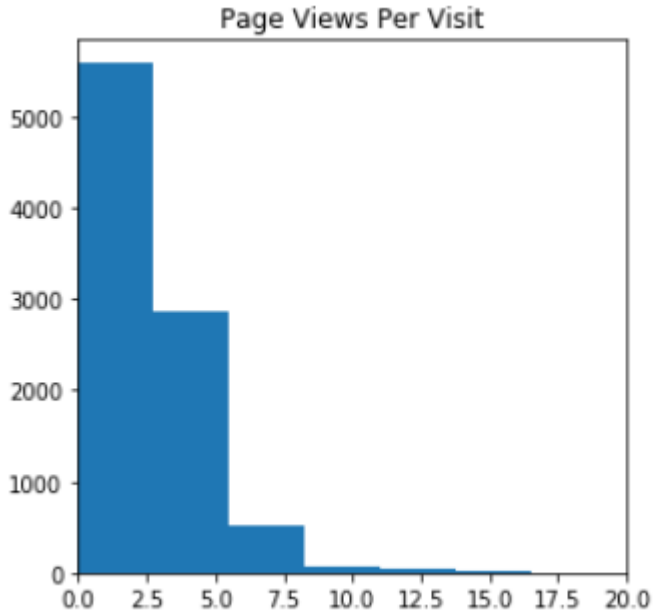


Understanding the Data & Basic Data Cleanup

- There are 37 columns (30 categorical and 7 Numeric) and 9240 observations in the dataset.
- Select is present as a class in different columns like:
  - Specialization
  - How did you hear about X Education
  - Lead Profile
  - City
- Since Select is an invalid class, we can deduce that it may be the form dropdown's default value; if the user hasn't chosen an option, Select will remain the default value. We used NaN in place of Select.
- Periodical, Get Additional Information About Our Courses, Give me a supply chain content update. Receive updates about DM Content, and I consent to paying by check. These columns contain only one unique value and no missing data. We removed these columns because they were inconsistent and didn't add any value to our EDA or model building.
- The Asymmetrique Activity Score, Asymmetrique Profile Score, Lead Profile, Lead Quality, X Education, and Asymmetrique Activity Index are all related terms. Please let us know how you learned about them. There are over 40% missing values in these columns. As a result, we have removed these columns from the model building and EDA.
- There is no datapoint/ observation (rows) in our dataset having more than 70% missing values.
- For the categorical variables Lead Origin, Lead Source, Last Activity, Last Notable Activity, Country, Specialization, and What is your present occupation—which contain a large number of classes and few datapoints—we have made additional buckets/bins.
- Applied business understanding to the missing value treatment. A new category is used in place of the NaN values for Specialization and Occupation. Not revealed.
- For our convenience during EDA and Model development, we renamed the columns that asked about your current occupation to Occupation and What matters most to you when picking a course to Reason_choosing.

Visualizing Data and EDA : Numerical variables





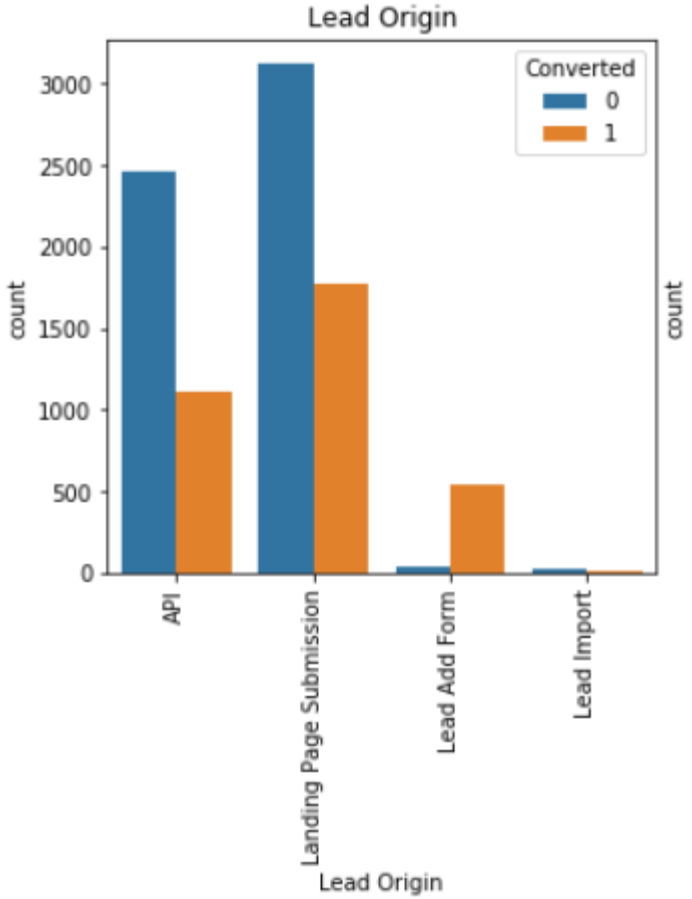
Inferences:

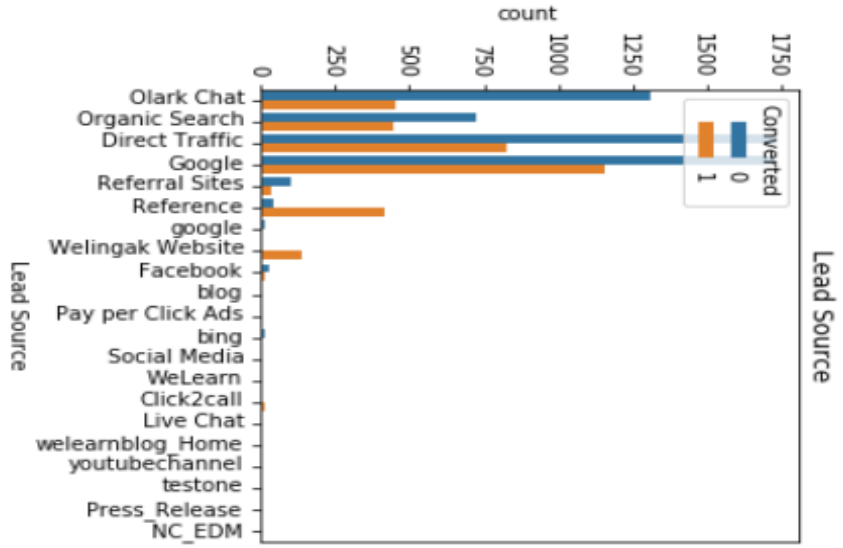
1.The average amount of time spent on the website for conversions Leads are significantly greater than those in the other group. The team ought to focus on the users who are spending more time on the website. The likelihood of those leads converting is higher.

2.A large number of anomalies are visible in the Total Visit for Converted = 0. Although a sizable portion of visitors are making more frequent visits to the website, they are not choosing to enroll in the course. The team ought to look at the cause of this. It can be due to financial difficulties, the fact that they are looking for courses that X Education does not now provide, the fact that they are finding other excellent possibilities from competitors, etc.

3. There are many outliers in **Total Time Spent on Website** column for Converted= 0.

Visualizing Data and EDA : Categorical Variables I





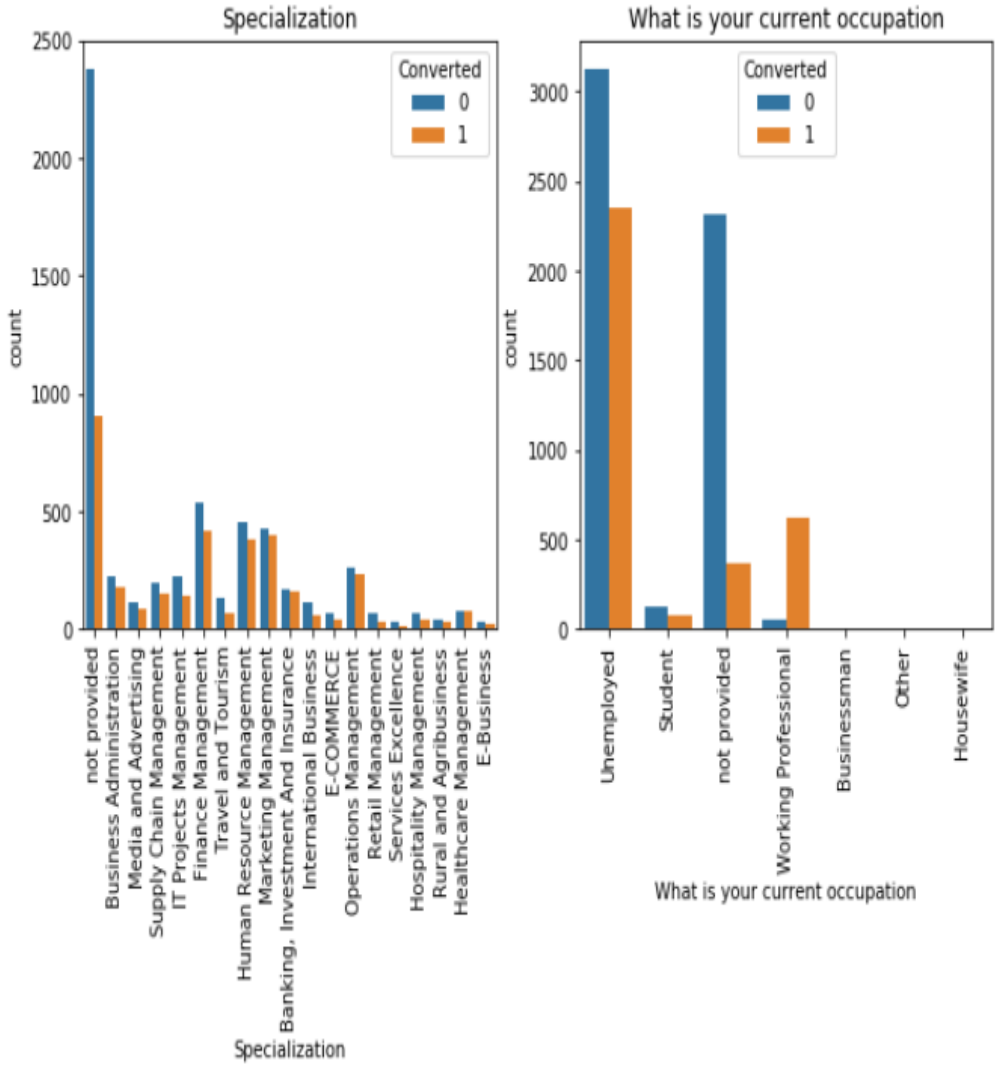
Inferences:

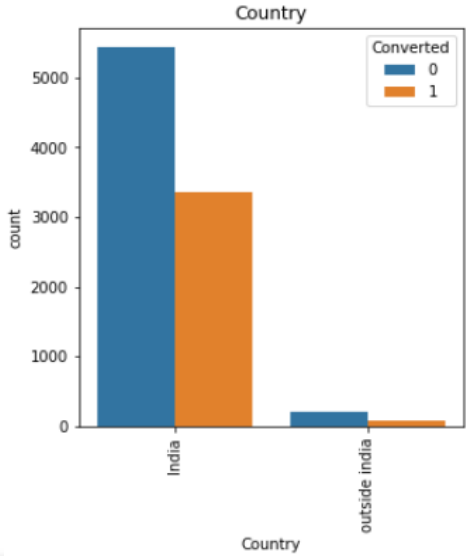
4. There is a very good possibility that the lead for Other type Lead Origin will be successfully converted.

5.The success rate of reference-type lead sources is extremely high. Referenced customers should be given top consideration by the team. Although it brings in comparatively fewer customers, Other Sources has an excellent conversion rate. Additionally, customers that find you through organic search have a much better likelihood of converting effectively.

6.Customers who have shown **Positive Behaviour in Last Activity** have considerably higher chance of getting converted successfully.

Visualizing Data and EDA : Categorical Variables II





Inferences:

7. Most of the customers are from India and having **Management Specializations**.

8.Individuals who indicated areas of expertise on the registration form are more likely to choose the course.

9.The majority of consumers who have expressed interest in the campaigns are jobless. Extremely high chances of a successful conversion for working professionals. The sales team ought to start an initiative to connect with more working professionals.

10.Individuals who disclosed their work situation on the papers have a greater probability of being converted successfully.

Data Preparation

Outlier Treatment

Identified 2.8% of total data (< 5%) as outliers and removed those rows

Train-Test Split

The dataset was divided in a 70:30 ratio into Train and Test groups. The train dataset is used to develop the model, while the test dataset is used to assess it.

Missing Value Imputation (Statistical Imputation)

1. Calculated median, mode on Train dataset.
2. Used that value to impute missing values in Train and Test Dataset.

3.Statistical Imputation is done as below:

A.Nominal Categorical Columns: Mode Imputation

B. Numeric Columns: Median Imputation

Categorical Variables Encoding

1. Columns with Yes and No values are underneath. Yes has been swapped out for 1 and No for 0.
Never call, email, or search Newspaper Article, Digital Ad, X Education Forums, Newspaper, With Suggestions, Accessible version of "Mastering The Interview".
2. For the columns below, dummy variables have been constructed in place of the original variable and the first dummy variable for each column in the data frame.
3. Lead Origin, Lead Source, Country, Specialization, Reason_choosing, Occupation, City

MinMax Scaling on Train Data

Performed MinMax Scaling (fit and transform) on Train data on all numeric predictors: TotalVisits, Total Time Spent on Website, Page Views Per Visit

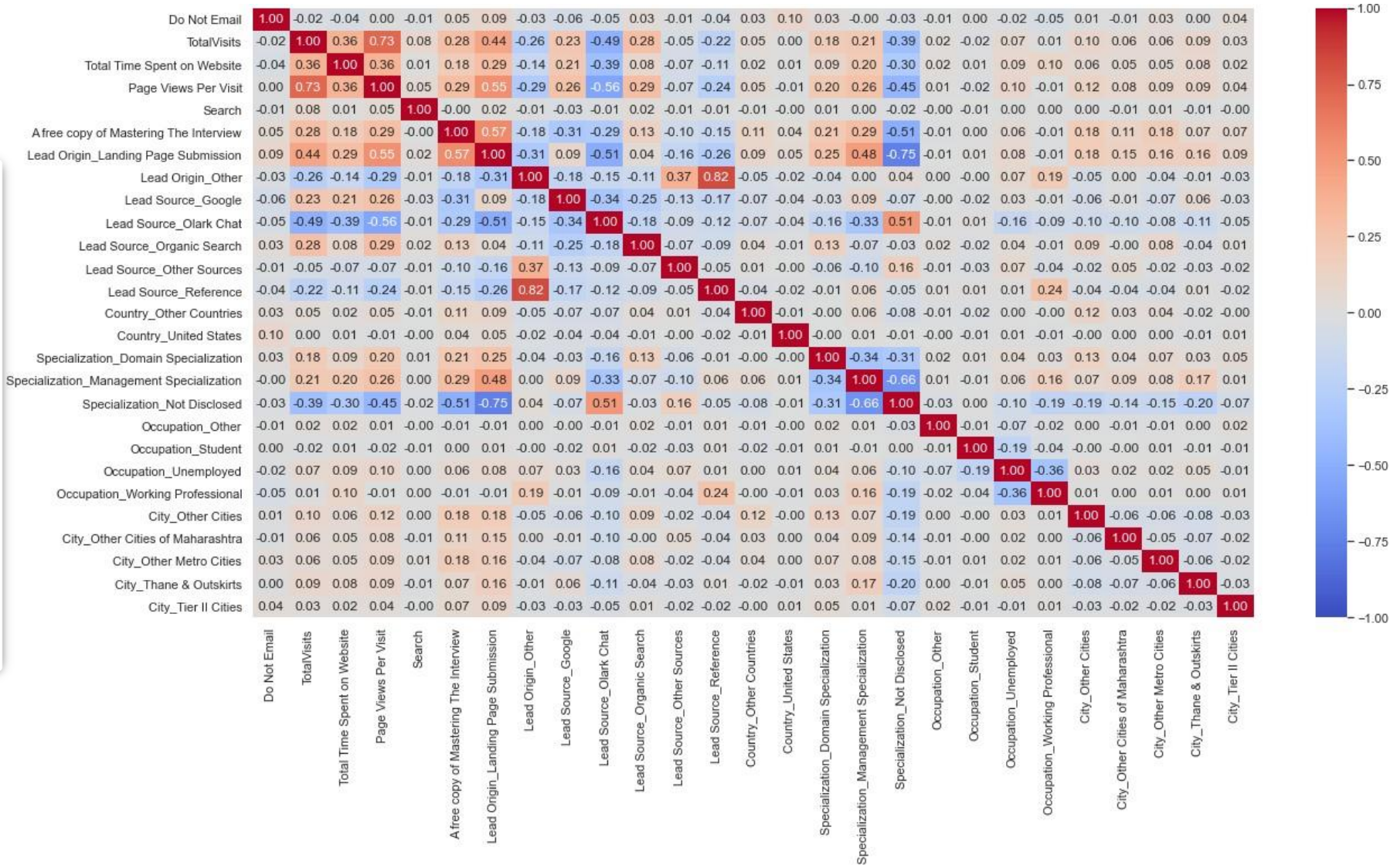
Variance Thresholding

Performed Variance Thresholding and removing columns having lower variance than threshold= .001
Removed Columns: Do Not Call, Newspaper Article, X Education Forums, Newspaper, Digital Advertisement, Through Recommendations, Reason_choosing_Flexibility & Convenience

Data Preparation: Correlation Map

Lead Origin_Other has very high correlation (.82) with **Lead Source_Reference**. Column **Lead Source_Reference** has been dropped.

Lead_Origin_Landing Page Submission has very high correlation (.75) with **Specialization_Not Disclosed**. **Specialization_Not Disclosed** column has been dropped.



Model Building : Approach

1. Recursive Feature Elimination (RFE) has been used to get top 16 features.

   - **Do Not Email :** An indicator variable selected by the customer wherein they select whether of not they want to be emailed about the course or not.
   - **TotalVisits:** The total number of visits made by the customer on the website.
   - **Total Time Spent on Website:** The total time spent by the customer on the website.
   - **Page Views Per Visit:** Average number of pages on the website viewed during the visits.
   - **Lead Origin_Landing Page Submission:** Dummy variable for Landing Page category of the origin identifier with which the customer was identified to be a lead.
   - **Lead Origin_Other:** Dummy variable for the Other category of the origin identifier with which the customer was identified to be a lead.
   - **Lead Source_Olark Chat:** Dummy variable for the Olark Chat category of the source of the lead.
   - **Lead Source_Other Sources:** Dummy variable for the Other category (other than Google, Direct Traffic, Olark Chat, Organic Search, Reference) of the source of the lead.
   - **Country_Other Countries:** Dummy variable for the Other category (other than India and United States) of the country of the customer.
   - **Specialization_Domain Specialization**: Dummy variable for Domain Specialization bin of Specialization variable.
   - **Specialization_Management Specialization:** Dummy variable for Management Specialization bin of Specialization variable.
   - **Occupation_Other:** Dummy variable for 'Other' category of customer's occupation.
   - **Occupation_Student:** Dummy variable for 'Student' category of customer's occupation.
   - **Occupation_Unemployed:** Dummy variable for 'Unemployed' category of customer's occupation.
   - **Occupation_Working Professional:** Dummy variable for 'Working Professional' category of customer's occupation.
   - **City_Tier II Cities:** Dummy variable for 'Tier II Cities' category of customer's city.

2. We built first Logistic Regression model using GLM (Generalized Linear Model) in statsmodels with these 16 features.

3. After that, the model was manually adjusted to extract statistically significant features (by examining the p-values) and eliminate multicollinearity (by examining the Variance Inflation Factors) at the same time. Accepted VIF is less than 5, and accepted p-value is less than.05.

4. Seven models in all were constructed; following each model building, the p-values of all beta coefficients and VIFs were examined, and any features that were found were eliminated for the subsequent model building. After every new model, we have also examined the confusion matrix and overall model accuracy to see how the new model is doing in comparison to the old one.

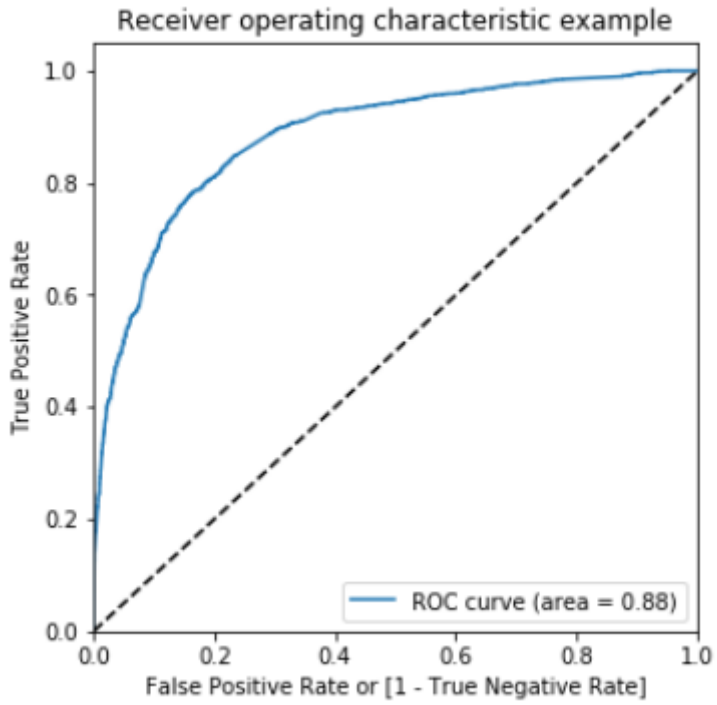
Prediction & Model Evaluation : (on Training data - cutoff .5)

1. In order to calculate our objective variable "Converted," we first calculated the probability for each observation in our training dataset using Model 7 and a probability cutoff of.5. Thus, "Converted" = 1 (Yes) if probability is >.5, and 0 (No) otherwise.
2. After predicting the target on our training data set, we calculated different evaluation metrics as below:

```
Overall model accuracy: 0.7989821882951654
Sensitivity / Recall:  0.6554025865665415
Specificity:  0.887432536622976
False Positive Rate:  0.1125674633770239
Positive Predictive Value:  0.7819810851169736
Positive Predictive Value:  0.8069642439822389
```

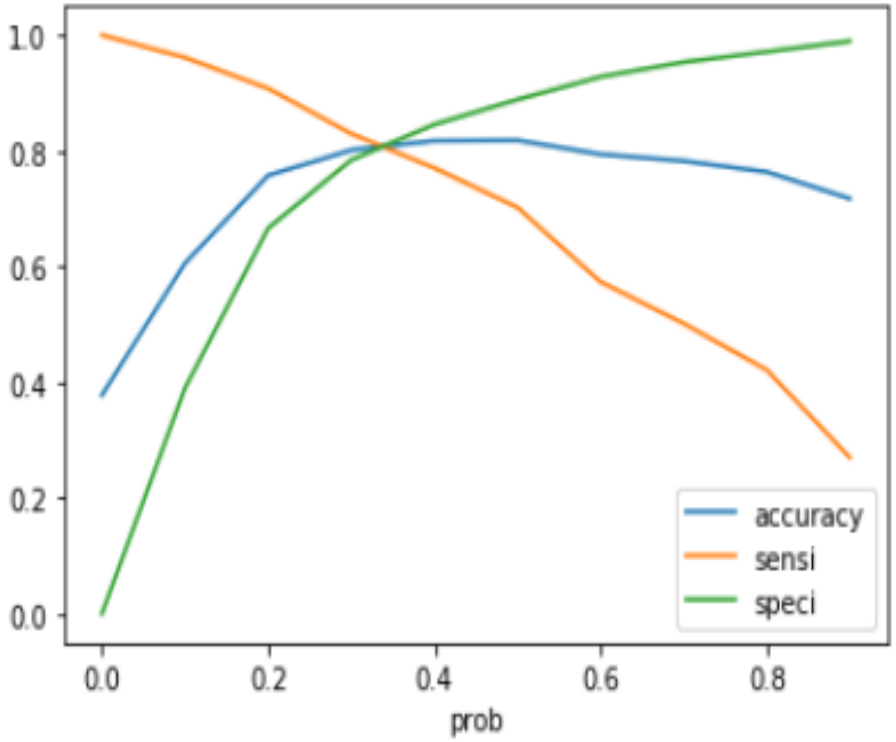
```
Confusion Matrix:
True Negative: 3453      False Positive: 438
False Negative: 826      True Positive: 1571

Overall model accuracy: 0.7989821882951654
```



3. The sensitivity of our model is poor (probability cut-off =.5). Through a trade-off analysis between Sensitivity and Specificity, the ideal probability cut-off value was determined.

Finding Optimal Probability cutoff & Evaluating on Train Data



In above plot, it's visible that 0.32 is the optimal point to set as cutoff probability for our model.

Model Evaluation Metrics on Train dataset with probability cutoff .32

```
Model Evaluation Metrics on Train dataset
#####################################################
Confusion Matrix:
True Negative: 2946      False Positive: 945
False Negative: 402      True Positive: 1995

Overall model accuracy: 0.7857824427480916
Sensitivity / Recall:  0.8322903629536921
Specificity:  0.7571318427139553
False Positive Rate:  0.24286815728604472
Positive Predictive Value:  0.6785714285714286
Positive Predictive Value:  0.8799283154121864
```

Observations:

Sensitivity of our model has been increased without any significant reduction in overall accuracy. New Specificity is also in well accepted range.

Model Evaluation : (on Test data) & Interpretation

Model Evaluation Metrics on Test dataset with probability cutoff .32

```
Model Evaluation Metrics on Test dataset
###################################################
Confusion Matrix:
True Negative: 1258      False Positive: 402
False Negative: 203      True Positive: 832

Overall model accuracy: 0.7755102040816326
Sensitivity / Recall:  0.8038647342995169
Specificity:  0.7578313253012048
False Positive Rate:  0.2421686746987952
Positive Predictive Value:  0.6742301458670988
Positive Predictive Value:  0.8610540725530459
```

Model is performing well on test data with Sensitivity= 80%, Specificity= 76% and overall accuracy: 78%,

	Converted	Conversion_Prob	final_predicted
0	0	0.123887	0
1	1	0.588440	1
2	1	0.370721	0
3	0	0.060348	0
4	0	0.442248	1
...	...	...	...
2693	1	0.111744	0
2694	1	0.829332	1
2695	0	0.039085	0
2696	1	0.965347	1
2697	0	0.007473	0

2698 rows × 3 columns

Top 3 variables which contribute most towards the probability of a lead getting converted:

- **Total Time Spent on Website**
- **What is your current occupation (Working Professional)**
- **Lead origin (Other)**

Conclusion and Recommendations

1. It was discovered that the following factors, in descending order, were most important to potential customers: #TotalVisits #The Total Amount of Time Spent on the Website. Lead Source_Google #Lead Source_Welingak Website #Lead Source_Organic Search #Lead Source_Referral Sites #Lead Origin_Lead Add Form Website of Welingak, the Lead Source #Email Bounced #Last Activity: Olark Chat Conversation #Do Not Email Yes

2. With this in mind, X Education can grow since they have a great possibility of persuading nearly every prospective customer to alter their mind and purchase their courses.