Case Study - Lead Scoring

**Submitted by:** Janarthanan Balasubramanian & Siva Prakash

## Business Understanding

**Client:** X Education (selling online courses to industry professionals)

X Education markets its courses on several websites and search engines like Google. The prospects who see these advertisements and land on the website might browse the courses or watch some videos or fill up a form for the course (at this point they are classified as a lead). 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.

|  |  |
| --- | --- |
| **Problem** | Lead conversion rate is very poor (30%) i.e. out of 100 leads only 30 are converted. |
| **Objective** | To make the process efficient so that the lead conversion rate goes up. |
| **Strategy** | Communicate with potential leads rather than making calls to everyone. |
| **Assignment** | Help X Education select the most promising leads so that the lead conversion rate can be around 80%. |
| **Approach** | Build a logistic regression model and assign a lead score between 0 and 100 to each of the leads which can be used by the employees of X Education to target potential leads. A higher score would mean that the lead is hot, i.e. is most likely to convert whereas a lower score would mean that the lead is cold and will mostly not get converted. |
| **Data Set** | Provided with Leads data set from the past with more than 9000 data points. The target variable, in this case, is the column ‘Converted’ which tells whether a past lead was converted (= 1) or not (= 0). |
| **Deliverables** | (1) A well-commented Jupyter note with at least the logistic regression model, the conversion predictions and evaluation metrics. (2) Solution to the business problems (3) Summary Report in 500 words explaining how you proceeded with the assignment and the learnings (4) Overall approach of the analysis in a presentation to present the analysis and results to the chief data scientist (include both technical and business aspects). |

# Our Approach for Lead Scoring

**Step 1:** Load the data into Python Data Frame and understand (shape, data types, % of null values, levels in categorical variables etc.)

**Step 2:** Use visualization to get more insights.

**Step 3:** Identify and address the data quality issues (null or missing values, outliers etc.)

**Step 4:** Data Preparation (label encoding, creating dummies, train-test split, rescaling of numeric variables, address multicollinearity).

**Step 5:** Building the Model (recursive feature elimination, manual feature selection and / or elimination, building the model).

**Step 6:** Evaluating the Model (validating the predictions with train data set and computing the metrics to find optimal cut-off)

**Step 7:** Predicting with the Model for Test Dataset. Validate the predictions by creating confusion matrix and computing the metrics)

**Step 8:** Assignment of lead score.

**Step 9:** Computing odds & log odds for better interpretability of the model.

**This presentation has 3 sections:**

**Section 1:** **Exploratory Data Analysis** (covers the first 3 steps - Reading and Understanding the Data, Data Visualization and Data Clean up)

**Section 2:** **Logistic Regression** (covers the next 3 steps - Data Preparation, Model Building and Model Evaluation)

**Section 3:** **Predictions and Inferences** (covers the last 3 steps - Prediction, Lead Score Assignment and computation of odds and log odds)

1. Exploratory Data Analysis

**For each field in** **the dataset**

(which contains the features and target variable from Leads.csv),

we have presented the following:

* A brief description
* Unique values and value counts
* Identification of **data quality issues**, if any
  1. Whether there are any null values or missing values
  2. How to handle the null or missing values
* **Data Visualization:**
  1. Bar chart plotting the frequency of the values (in case of categorical variables)
  2. Distribution of the quantitative variables with distplot and boxplot.
  3. Correlation Heat Maps for quantitative variables
* **Identification of the Outliers** and insights regarding any skewness in the data.

This section of the presentation covers the first 3 steps in our solution approach.

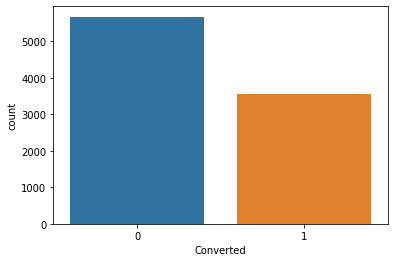
## Key Variables

|  |  |  |  |
| --- | --- | --- | --- |
| **Column Name** | **Explanation** | **Null** | **Remarks** |
| Prospect ID | Unique ID to identify the Customer | None | Primary Key for the data set. |
| Lead Number | A number assigned to each lead | None | This is also an unique identifier. |

**Note:** The key variables have no significance with respect to model building. However, once the model is built and is used to predict the target variable, we should be able to map the predictions to the leads. We have dropped the Prospect ID and converted the Lead Number into the index for the data frame. With the index value we can map the results to the leads.

## Target Variable

|  |  |  |  |
| --- | --- | --- | --- |
| **Column Name** | **Explanation** | **Null** | **Remarks** |
| Converted | Whether the lead is successfully converted or not. | None | 0: not converted 1: converted |



5679

3561

The data set has 9240 observations.

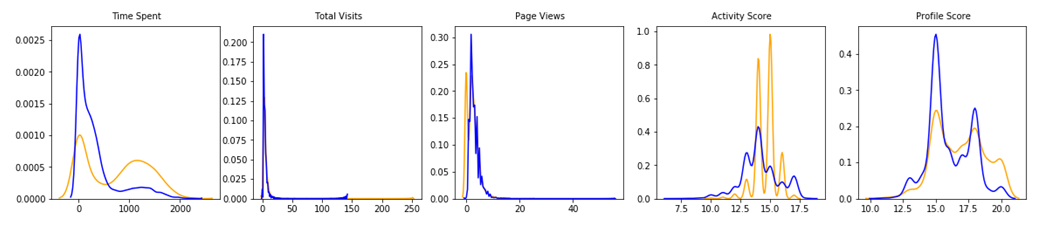
5679 records are with Converted = 0 and 3561 records are with Converted = 1.

The conversion rate is 3561 / 9240 = 38.5% only.

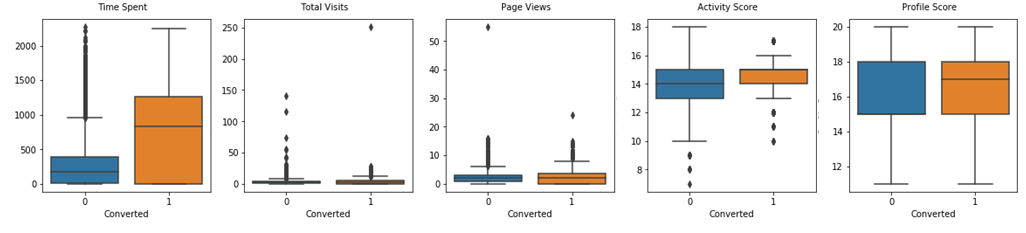
Our objective in this assignment is to **improve the conversion rate to at least 80%**.

## Numeric Variables - Visualization

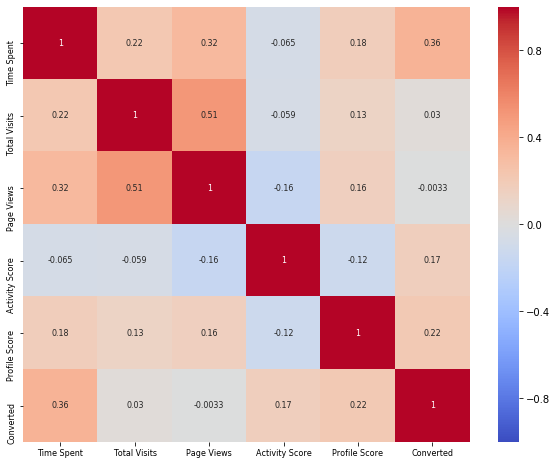
**Distribution Plot:** Shows how the data is distributed and whether there are any skewness.



**Box Plot:** Shows data distribution, skewness and outliers in the data.



## Numeric Variables - Correlation Heat Map



## Numeric Variables - Insights

There are very few numeric variables in the data set.

|  |  |  |  |
| --- | --- | --- | --- |
| **Column Name** | **Explanation** | **Null** | **Remarks** |
| Total Visits | Number of visits made by customer on the website. | 137 | Impute the null values with median. |
| Time Spent on Website | The total time spent by the customer on the website. | 0 | No null values. A significant predictor. |
| Page Views Per Visit | Avg. number of pages viewed during visit to website. | 137 | Impute the null values with median. |
| Activity Score | Assigned based on customer activity. | 4218 | More than 45% null values. Drop the column. |
| Profile Score | Assigned based on customer activity. | 4218 | More than 45% null values. Drop the column. |

## Observations from the Visualization

There are quite a few outliers in the Time Spent on Website variable.

The distributions are skewed to right for Total Visits, Time Spent and Page Views.

For imputing null values, it is better to use the median (since the data is skewed, mean is not a reliable measure).

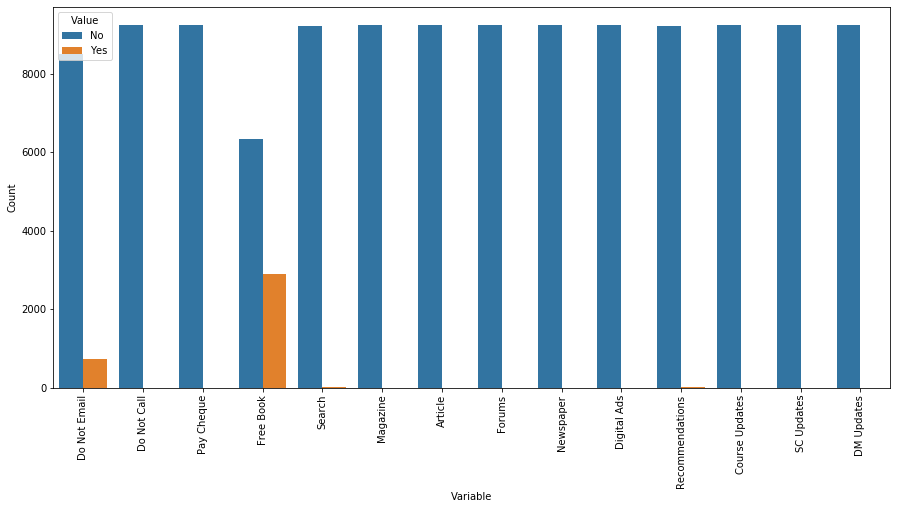
Time Spent on the website is a clear predictor of conversion. There is more conversion from leads who spend more time in website.

It is difficult to predict whether a lead will convert based on the Total Visits and Page Views.

We see a 51% correlation between the Total Visits and Page Views.

Time Spent has the highest positive correlation (0.36) with Converted (target variable) compared to other numeric variables.

## Binary Variables (Categorical Variables with 2 Levels) - Visualization



## Binary Variables (Categorical Variables with 2 Levels) - Visualization

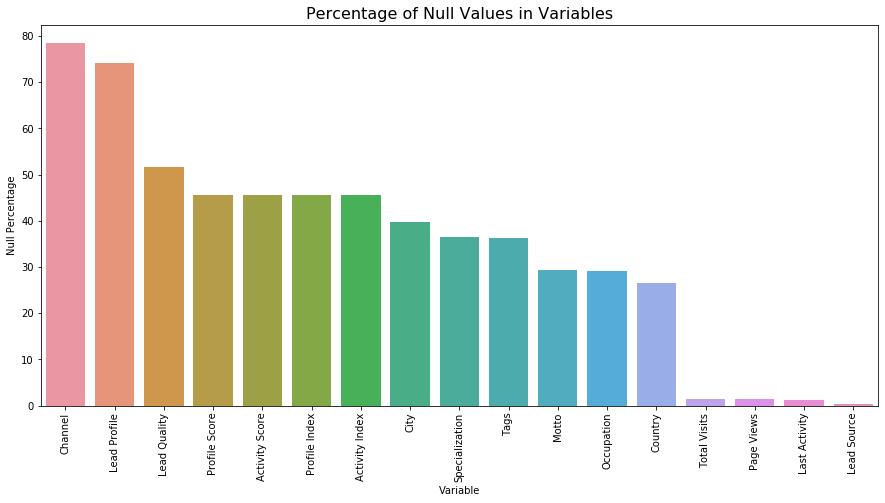
|  |  |  |  |
| --- | --- | --- | --- |
| **Column Name** | **Explanation** | **Null** | **Remarks** |
| Do Not Email | Whether customer want to be emailed or not. | None | 92% have opted for receiving the emails. |
| Do Not Call | Whether the customer want to be called or not. | None | Number of records with Yes is almost Nil |
| Search | Whether the customer had seen Search Ads. | None | Number of records with Yes is almost Nil |
| Magazine | Whether the customer had seen Magazine Ads. | None | Number of records with Yes is almost Nil |
| Article | Whether the customer had seen Newspaper Articles. | None | Number of records with Yes is almost Nil |
| Forums | Whether the customer had visited Forums. | None | Number of records with Yes is almost Nil |
| Newspaper | Whether the customer had seen Newspaper Ads. | None | Number of records with Yes is almost Nil |
| Digital Ads | Whether the customer had seen Digital Ads. | None | Number of records with Yes is almost Nil |
| Recommendations | Whether the customer had been recommended. | None | Number of records with Yes is almost Nil |
| Course Updates | Customer choice to receive course updates. | None | Number of records with Yes is almost Nil |
| SC Updates | Customer choice to receive supply chain content. | None | Number of records with Yes is almost Nil |
| DM Updates | Customer choice to receive DM content updates. | None | Number of records with Yes is almost Nil |
| Pay Cheque | Customer choice to pay through cheque. | None | Number of records with Yes is almost Nil |
| Free Book | Whether the customer wants a free copy of the book | None | 6352 (68%) opted out and 2888 opted in. |

*Note: Certain column names from the original data set are renamed with short names for ease of coding.*

The columns with only one level has no variance and do not contribute to the model. So we can drop those columns.

**Conclusion:** The columns other than Do Not Email and Free Book can be dropped.

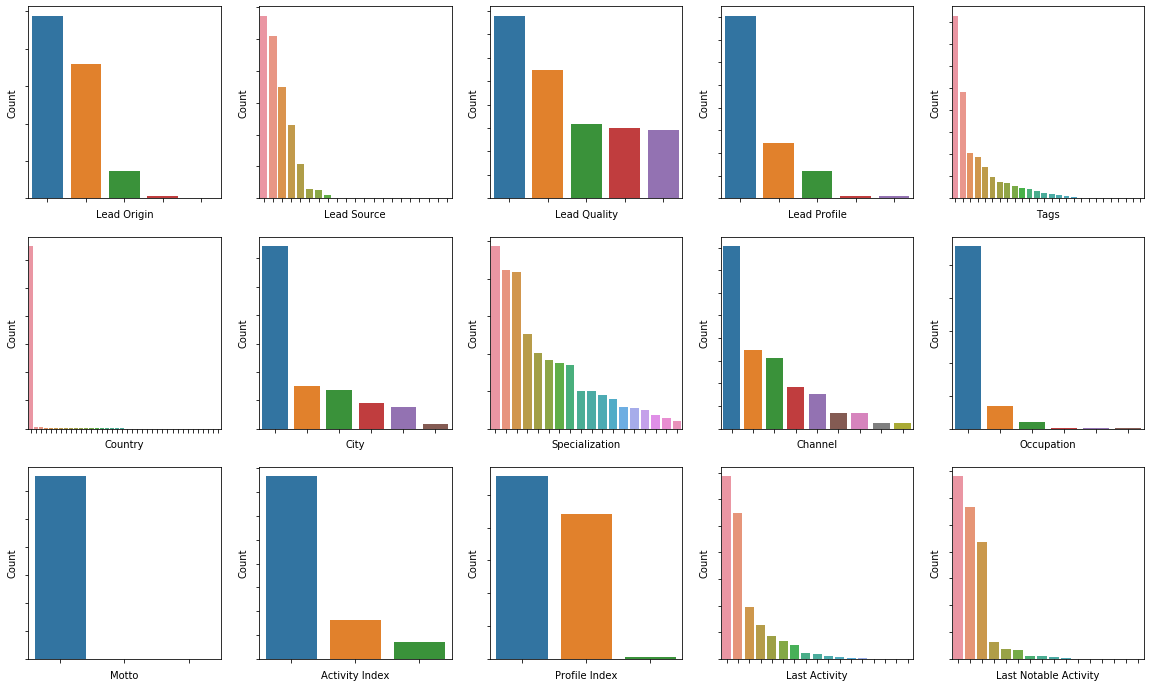
## Categorical Variables - Percentage of Null Values



**Drop the columns** **with more than 45% null values**

Channel, Lead Profile, Lead Quality, Profile Score, Activity Score, Profile Index and Activity Index

## Categorical Variables - Univariate Plots



## Categorical Variable - Lead Origin

|  |  |  |  |
| --- | --- | --- | --- |
| **Column Name** | **Description** | **Null** | **Visualization** |
| Lead Origin | Origin with which the customer was identified to be a lead. This has 5 different levels: Landing Page Submission (4886 records) API (3580 records) Lead Add Form (718 records) Lead Import (55 records) and Quick Add Form (1 record)  We have reduced the levels by adding the record in Quick Add Form also to Lead Add Form. | None | C:\Users\jnvd\AppData\Local\Microsoft\Windows\INetCache\Content.MSO\92A3EFCF.tmp |

## Categorical Variable - Lead Source

|  |  |  |  |
| --- | --- | --- | --- |
| **Column Name** | **Description** | **Null** | **Visualization** |
| Lead Source | The source of the lead. There are 21 unique values. We can merge Google and google, which are merged. Many of these sources have less than 10 observations in the given data set. After marking the records with less than 100 observations and the null values as ‘Others’ we have the following levels.  Google (2873) Direct Traffic (2543) Olark Chat (1755) Organic Search (1154) Reference (534) Welingak Website (142) Referral Sites (125) Others (114) | 36 | C:\Users\jnvd\AppData\Local\Microsoft\Windows\INetCache\Content.MSO\30189B59.tmp |

## Categorical Variable - Lead Quality

|  |  |  |  |
| --- | --- | --- | --- |
| **Column Name** | **Description** | **Null** | **Visualization** |
| Lead Quality | Based on the data and intuition of the employee assigned to the lead. There were 4767 null values. 4767 / 9240 > 50% null values. Further the Lead Quality is assigned by the employee based on data and intuition. We cannot use it in our model, because we are trying to build the model to provide this intuition. **We will drop this column.** | 4767 | C:\Users\jnvd\AppData\Local\Microsoft\Windows\INetCache\Content.MSO\5841AFD1.tmp |

## Categorical Variable - Lead Profile

|  |  |  |  |
| --- | --- | --- | --- |
| **Column Name** | **Description** | **Null** | **Visualization** |
| Lead Profile | A lead level assigned to each customer based on their profile. There are 2709 null values and 4146 records with value 'Select'. 6855 / 9240 = 74% null values. The remaining records are distributed in 5 levels: Potential Lead, Other Leads, Student of Some School, Lateral Student, Dual Specialization Student. Since 74% are null values and there is no specific logic to impute, **we will drop this column**. | 6855 | C:\Users\jnvd\AppData\Local\Microsoft\Windows\INetCache\Content.MSO\62B06A8B.tmp |

## Categorical Variable - Country

|  |  |  |  |
| --- | --- | --- | --- |
| **Column Name** | **Description** | **Null** | **Visualization** |
| Country | The country of the customer. 6492 records have Country = 'India' and 2461 records have null values. 37 other countries put together are 287 records. We can just mark them as ‘Others’, however, this is not going to contribute to the model. The data is highly skewed. **We can drop this column.** | 2461 | C:\Users\jnvd\AppData\Local\Microsoft\Windows\INetCache\Content.MSO\73956543.tmp |

## Categorical Variable - City

|  |  |  |  |
| --- | --- | --- | --- |
| **Column Name** | **Description** | **Null** | **Visualization** |
| City | The city of the customer. Out of 9240 records, 1420 null values and 2249 missing values (more than 30% are null values) and 3222 records have City = 'Mumbai'. The other 5 values are distributed among the remaining 2349 records, but instead of giving a specific city name it says Other Cities, Other Cities of Maharashtra, Other Metro Cities, Tier II Cities, Thane & Outskirts. **We can drop this column.** | 3669 | C:\Users\jnvd\AppData\Local\Microsoft\Windows\INetCache\Content.MSO\44439D3F.tmp |

## Categorical Variable - Specialization

|  |  |  |  |
| --- | --- | --- | --- |
| **Column Name** | **Description** | **Null** | **Visualization** |
| Specialization | The industry domain in which the customer worked before. There are 1438 null values and 1942 missing values (marked as Select). It has 19 unique values like Finance Management, IT Projects Management, Human Resources Management and so on. The conversion rate ranges from 28% to 50% and hence this is a significant variable. We will mark the missing / null values as Unknown and later drop this when creating dummy variables. | 3380 | C:\Users\jnvd\AppData\Local\Microsoft\Windows\INetCache\Content.MSO\6C06D8A5.tmp |

## Categorical Variable - How did you hear about X Education

|  |  |  |  |
| --- | --- | --- | --- |
| **Column Name** | **Description** | **Null** | **Visualization** |
| Channel | The source from which the customer heard about X Education. Renamed as Channel.  There are 2207 null values and 5043 records with value 'Select' which means that the customer has not selected any input. Out of 9240 records, 7250 records have missing value for this feature, which is more than 75%. **We can drop this column.** | 7250 | C:\Users\jnvd\AppData\Local\Microsoft\Windows\INetCache\Content.MSO\10901395.tmp |

## Categorical Variable - What is your current occupation

|  |  |  |  |
| --- | --- | --- | --- |
| **Column Name** | **Description** | **Null** | **Visualization** |
| Occupation | Current occupation of the customer. Renamed as Occupation. The levels are Unemployed (5600) Working Professional (706), Student (210) Other (16), Housewife (10) and Businessman (8). We can mark 2690 null values as Other. Housewife and Businessman also can be marked as Others (as number of records is less than 10). | 2690 | C:\Users\jnvd\AppData\Local\Microsoft\Windows\INetCache\Content.MSO\57A3969D.tmp |

## Categorical Variable - What matters most to you in choosing a course

|  |  |  |  |
| --- | --- | --- | --- |
| **Column Name** | **Description** | **Null** | **Visualization** |
| Motto | What is the main motto behind doing the course. Renamed as Motto. Out of 9240 records, 6528 records have Motto = 'Better Career Prospects' and 2709 records have null values. Only 3 more records have a different values: Flexibility & Convenience (2 records), Other (1 record).  **We can drop this column.** | 2709 | C:\Users\jnvd\AppData\Local\Microsoft\Windows\INetCache\Content.MSO\FD91B6FB.tmp |

## Categorical Variable - Tags

|  |  |  |  |
| --- | --- | --- | --- |
| **Column Name** | **Description** | **Null** | **Visualization** |
| Tags | Tags assigned to customers indicating the current status of the lead. There are 3353 null values. Tags are free-form text that is updated based on what the lead said when the employee from the organization approached them. We are trying to build a model which will optimize the number of calls made so that the employees can contact the potential leads and the conversion rate can improve. This field will not be available in the future and hence **we cannot use this feature for our model.** | 3353 | C:\Users\jnvd\AppData\Local\Microsoft\Windows\INetCache\Content.MSO\56F565C1.tmp |

## Categorical Variable - Activity Index

|  |  |  |  |
| --- | --- | --- | --- |
| **Column Name** | **Description** | **Null** | **Visualization** |
| Activity Index | Activity Index assigned based on customer activity. The values are High, Medium and Low. However, there is a high percentage of null values: 4128 / 9240 - approximately 45.5% and we do not have any means to impute these values. **We will drop this column.** | 4218 | C:\Users\jnvd\AppData\Local\Microsoft\Windows\INetCache\Content.MSO\1B5D9307.tmp |

## Categorical Variable - Profile Index

|  |  |  |  |
| --- | --- | --- | --- |
| **Column Name** | **Description** | **Null** | **Visualization** |
| Profile Index | Profile Index assigned based on customer profile. The values are High, Medium and Low. However, there is a high percentage of null values: 4128 / 9240 - approximately 45.5% and we do not have any means to impute these values. **We will drop this column.** | 4218 | C:\Users\jnvd\AppData\Local\Microsoft\Windows\INetCache\Content.MSO\7C3DF9AD.tmp |

## Categorical Variable - Last Notable Activity

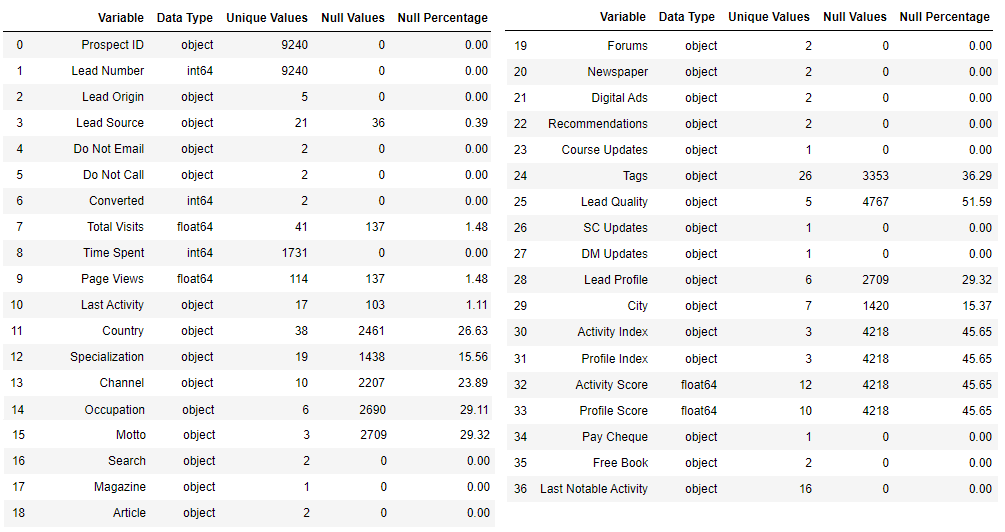
|  |  |  |  |
| --- | --- | --- | --- |
| **Column Name** | **Description** | **Null** | **Visualization** |
| Last Notable Activity | The last notable activity performed by the student. There are no null values. However, to keep it consistent with Last Activity, we will retain only the levels that are retained in Last Activity. Apart from that we have a value Modified, which we will retain.  The levels are: Modified, Email Opened, SMS Sent, Olark Chat Conversation, Page Visited on Website, Email Bounced, Email Link Clicked, Form Submitted on Website, Unreachable, Unsubscribed. | None | C:\Users\jnvd\AppData\Local\Microsoft\Windows\INetCache\Content.MSO\7A6466AF.tmp |

## Categorical Variable - Last Activity

|  |  |  |  |
| --- | --- | --- | --- |
| **Column Name** | **Description** | **Null** | **Visualization** |
| Last Activity | Last activity performed by the customer.  The null values are imputed with the value: Page Visited on Website (since for these records the Total Time Spent on Website is not null indicating that they visited the website).  Also the levels with less than 50 records can be marked as Others to avoid any skewness affecting the final model. We will retain the levels: Email Opened, SMS Sent, Olark Chat Conversation, Page Visited on Website, Converted to Lead, Email Bounced, Email Link Clicked, Form Submitted on Website, Unreachable, Unsubscribed, Others. | 103 | C:\Users\jnvd\AppData\Local\Microsoft\Windows\INetCache\Content.MSO\BC5FC739.tmp |

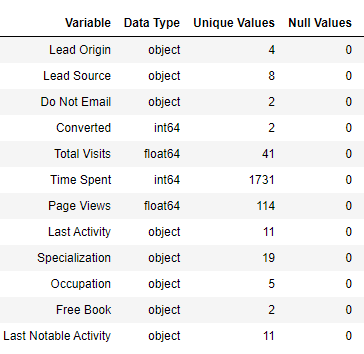
## Summary

We started with 37 different features from the Leads dataset. Note that we have renamed the columns for ease of coding & presentation.



## Summary - Action Taken

After a thorough analysis on the leads data set we have now addressed the data quality issues (null values, missing values) and also reduced the number of levels in some of the categorical variables.

****Categorical variables had a level ‘Select’ (customer did not select any value). The value is missing and hence updated with null values. (Lead Profile, City, Specialization, Channel)

**Drop the columns with more than 45% null values:** (1) Channel, (2) Lead Profile, (3) Lead Quality, (4) Profile Score, (5) Activity Score, (6) Profile Index and (7) Activity Index.

**Drop the binary variables with no variance** (will not contribute to the model): (8) Do Not Call, (9) Search, (10) Magazine, (11) Article, (12) Forums, (13) Newspaper, (14) Digital Ads, (15) Recommendations, (16) Course Updates, (17) SC Updates, (18) DM Updates, and (19) Pay Cheque are dropped.

**Drop the Categorical variables that are skewed** (will adversely affect the final model): (20) Motto, (21) Country, (22) City are dropped. (23) Tags is also dropped because it is the status update by the callers; the objective is to reduce the number of calls.

**Key Columns:** (24) Prospect ID and (25) Lead Number are both unique identifiers. Prospect ID is dropped and Lead Number is converted into index of the data set.

After dropping these 25 columns we are left with 11 features excluding the target variable (Converted). The null values are imputed with appropriate values. No rows were dropped. We have also **reduced the number of levels** in some of the variables by consolidating the levels with less than 10 observations into a separate level (Others). Lead Origin (5 to 4), Lead Source (21 to 8), Occupation (6 to 5), Last Activity (17 to 11), Last Notable Activity (16 to 11).

*This brings us to the end of the part 1 of this presentation title Exploratory Data Analysis.*

2. Logistic Regression

# Step 4: Data Preparation

## 1. Dealing with Categorical Variables

The categorical variables cannot be used directly in the model; hence need some treatment. For binary variables, we just need one indicator variable by **mapping the values to 0 and 1**. Do Not Email and Free Book (A free copy of Mastering The Interview) has Yes or No values. So we can map the values to 0 and 1.

For variables with more than 2 levels, we have two options: (1)

* perform **label encoding**: convert labels into numeric form.
* **create dummies**: for a categorical variable with n levels, we need to create n-1 dummy variables. The nth variable is redundant and may create multicollinearity issues. In Python, we can use the pd.get\_dummies() function to create dummy variables from the data frame.

We have created dummies for the features Lead Origin, Lead Source, Specialization, Occupation, Last Activity, Last Notable Activity.

For a categorical variable with n levels, n columns get created; but we only need only n - 1, because the nth column will create issues due to multicollinearity. We have to drop one columns per feature: Lead Origin\_Lead Import, Lead Source\_Others, Last Activity\_Others, Specialization\_Unknown, Occupation\_Unknown and Last Notable Activity\_Others are dropped. We have 58 features in the data set.

## 2. Train Test Split

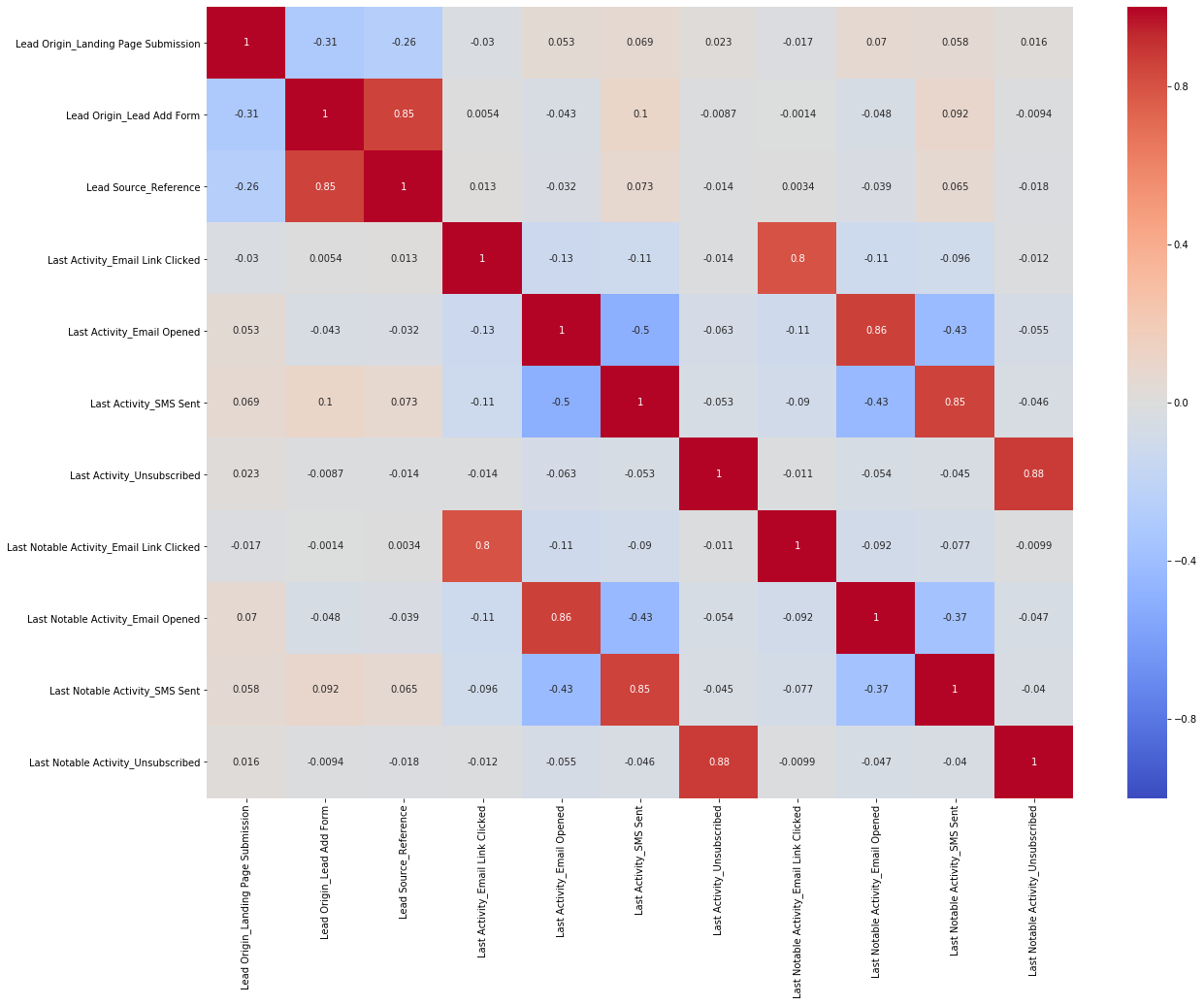
The next step is to split the data set into train data and test data. The algorithm learns from the train data to get deeper understanding on the target variable and uncovers patterns and relationships with other features in the dataset. With scikit learn libarary we split the data in the ratio 70:30 (70 for training the model and 30 for testing the model). After the split, there are 6468 records in the train data set and 2772 records in the test set.

## 3. Rescaling the Numeric Features

We need to rescale the numeric variables (Total Visits, Page Views, Time Spent in Website) so that they have a comparable scale. Otherwise, the coefficients obtained while fitting the model might be very large or very small as compared to the other coefficients. There are two methods of scaling that we can go for: (1) Standardization and (2) Normalization. In this assignment, we have applied **Standard Scaler** from the scikitlearn library. Scaling will not impact the model. Scaling should be done only on train data set after the train-test split.

## 4. Multicollinearity in Train Data Set

Created a heat map (see next slide) to check for multicollinearity.

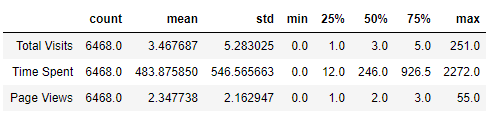


## Summary of Data Preparation

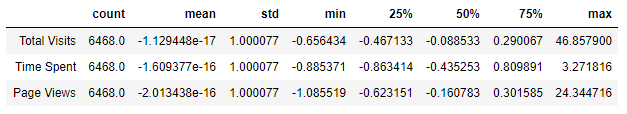
**Number of Features after Creating Dummies:** 58

**Test Train Split:** Ratio 70:30 (6468 records in the train data set and 2772 in test data set)

**Numeric Variables: Before Rescaling**



**Numeric Variables: After Rescaling**

****

**Multicollinearity in Train Data Set**

A quick check for multicollinearity between these 58 variables showed high positive correlation between:

* Lead Origin\_Lead Add Form and Lead Source\_Reference (85%)
* Last Notable Activity\_Email Link Clicked and Last Activity\_Email Link Clicked (80%)
* Last Notable Activity\_Email Opened and Last Activity\_Email Opened (86%)
* Last Notable Activity\_SMS Sent and Last Activity\_SMS Sent (85%)
* Last Notable Activity\_Unsubscribed and Last Activity\_Unsubscribed (88%)

# Step 5: Building the Model

## Recursive Feature Elimination

While evaluating the models, we can drop the variables with highest p-value. A p-value > 0.05 is considered insignificant. We can complement this information with the VIF (Virtual Inflation Factor). VIF, gives a basic quantitative idea about how much the feature variables are correlated with each other. It is an extremely important parameter to test our linear model. A value < 5 is considered good fit. Sometimes VIF is infinity, which means that there exists a higher degree of collinearity between the variable and others. However, first we will apply RFE (recursive feature elimination) an inbuilt feature to reduce the number of variables from 58 to 15.

**Features retained after Recursive Feature Elimination:**

Index(['Do Not Email', 'Time Spent', 'Lead Origin\_Lead Add Form', 'Lead Source\_Olark Chat', 'Lead Source\_Welingak Website', 'Last Activity\_Converted to Lead', 'Last Activity\_Email Bounced', 'Last Activity\_Olark Chat Conversation', 'Last Activity\_Page Visited on Website', 'Occupation\_Others', 'Occupation\_Student', 'Occupation\_Working Professional', 'Last Notable Activity\_Email Link Clicked', 'Last Notable Activity\_Email Opened', 'Last Notable Activity\_Modified']

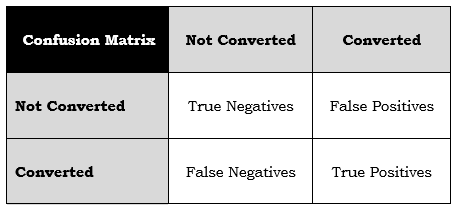
## Manual Feature Elimination

Build the Model and check the p-value and / or VIF to see whether all these variables are significant. Occupation\_Student had a p-value = 0.108 (> 0.05) and hence insignificant. Removing this variable, we built the model with the remaining 14 features.



The p-value < 0.05 indicates that all variables are significant. The VIF for these features are less than 2 and hence significant.

# Step 6: Evaluating the Model

**Using the above model we predict the probabilities. However the challenge is how to decide whether the lead will get converted or not, on the basis of probabilities? At what probability cutoff we can predict that the lead will get converted?**

**1. Create confusion matrix at different probability levels (0.1, 0.2, ... 0.9)**

**2. Compute the True Positive Rate and False Positive Rate.**

**True Positive Rate** = True Positives / (True Positives + False Negatives)

**False Positive Rate** = False Positives / (True Negatives + False Negatives)

**3. Plot the ROC curve. ROC curve**

ROC curve shows the tradeoff between the True Positive Rate (TPR - also called sensitivity) and the False Positive Rate (FPR = 1 - specificity). Higher values of TPR, will also result in higher values of FPR, which is not good. So we have to find a balance between these two metrics. A good ROC curve is the one which touches the upper-left corner of the graph.

**4. Compute the Other Metrics at different probability levels.**

**Accuracy** = Actual Instances Correctly Predicted / Total Instances

Total Instances = True Positive + True Negative + False Positive + False Negative

Correctly Predicted = True Positive + True Negative

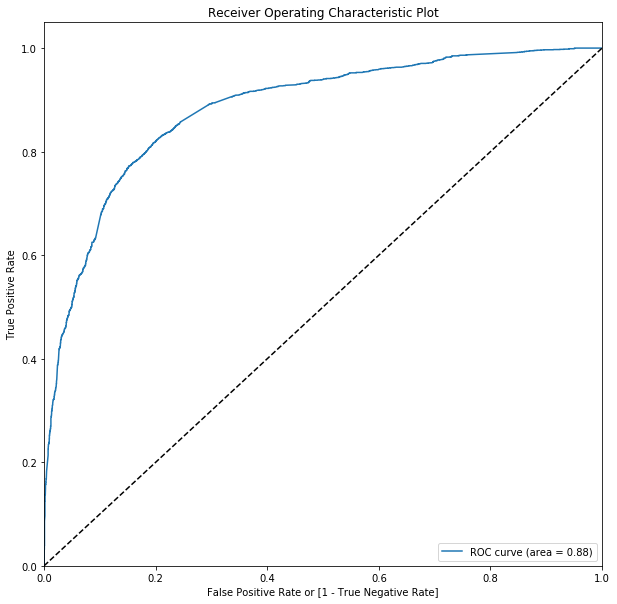
**Sensitivity** = Number of actual Yeses correctly predicted / Total Number of Yeses = True Positive / (True Positive + False Negative)

**Specificity** = Number of actual Nos correctly predicted / Total Number of Nos = True Negative / (True Negative + False Positive)

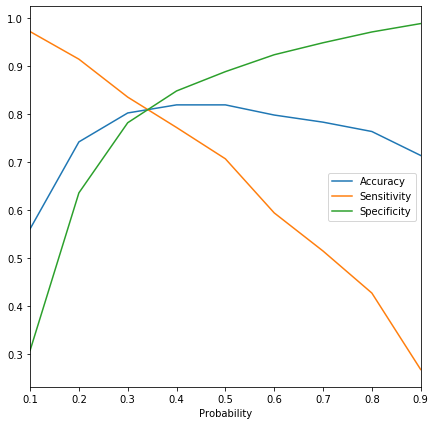
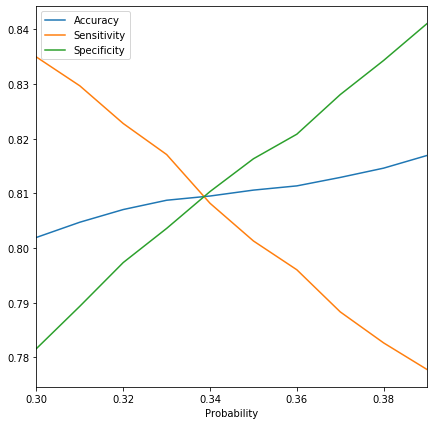
**5. Plot them to see where they intersect.** That is the optimal cutoff point with best values for accuracy, specificity and sensitivity.

## Plot the ROC curve

Higher the area under the curve of an ROC curve, the better is our model. **The area under curve is 0.88. The model is good.**

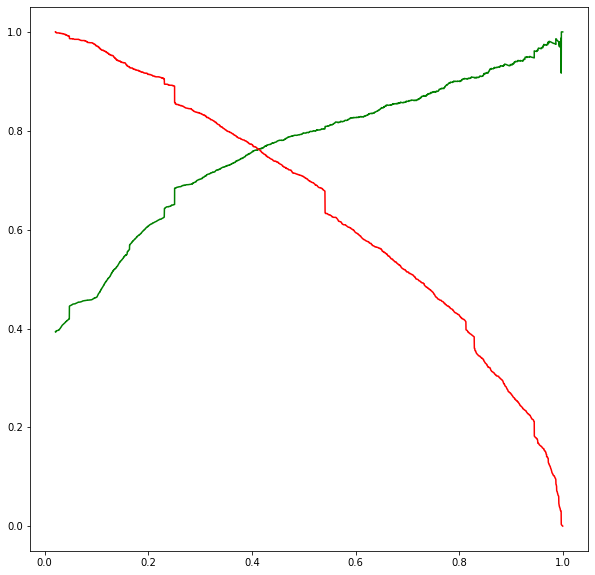
****

## Optimal Cutoff Point based on Metrics

We have plotted the accuracy, sensitivity and specificity with different probability levels - 0.1, 0.2, 0.3 and so on. The lines intersect between o.3 and 0.4. To take a closer look, we again plotted the metrics for 0.31. 0.32, 0.33 ... 0.39. The cutoff point is at 0.34. This is the optimal cutoff. If the probability is above 0.34 we can predict that the lead will be converted or else it will not be converted.

## Precision vs. Recall

This is another method to find the optimal cutoff.

**Precision** = True Positive / (True Positive + False Positive)

Precision is also called Positive Predictive Value

**Recall** = True Positive / (True Positive + False Negative)

Recall is same as Sensitivity.

Here, we get the optimal cutoff at 0.4. This is just another method to arrive at the optimal cutoff.

## Metrics for our model with optimal cutoff at 0.34 on Train Data vs. Test Data

|  |  |
| --- | --- |
| **Train Data** | **Test Data** |
| Confusion Matrix  |  |  |  | | --- | --- | --- | | **Confusion Matrix** | **Not Converted** | **Converted** | | **Not Converted** | **TN: 3243** | **FP: 759** | | **Converted** | **FN: 473** | **TP: 1993** |  Metrics Accuracy : 0.8095238095238095  Sensitivity : 0.8081914030819141  Specificity : 0.8103448275862069  False Positive Rate : 0.1896551724137931  Positive Predictive Value : 0.7242005813953488  Negative Predictive Value : 0.8727125941872982 Precision vs. Recall Precision Score : 0.7242005813953488  Recall Score : 0.8081914030819141 | Confusion Matrix  |  |  |  | | --- | --- | --- | | **Confusion Matrix** | **Not Converted** | **Converted** | | **Not Converted** | **TN: 1383** | **FP: 294** | | **Converted** | **FN: 212** | **TP: 883** |  Metrics Accuracy : 0.8174603174603174  Sensitivity : 0.806392694063927  Specificity : 0.8246869409660107  False Positive Rate : 0.17531305903398928  Positive Predictive Value : 0.7502124044180118  Negative Predictive Value : 0.8670846394984326 Precision vs. Recall Precision Score : 0.7502124044180118  Recall Score : 0.806392694063927 |

3. Predictions & Inferences

## Assigning Lead Score

The lead score is assigned as the percentage value of the probabilities computed. The lead score is between 0 and 100.

## Computing Log Odds

Since likelihood of conversion as a probability is difficult to interpret, the logistic regression equation can be converted into a more suitable form for interpretaion by *linearising* it. For this we can compute the log odds and odds. log odds = beta0 + beta1 x1 + beta2 x2 + ....

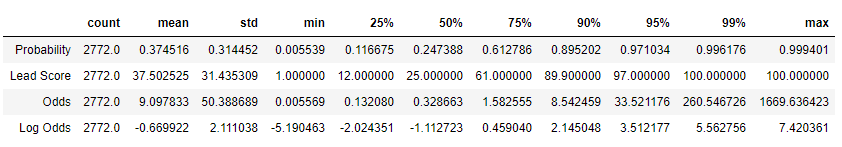
Our equation for computing log odds is:

**0.0338 + (Lead Origin\_Lead Add Form \* 3.764) + (Occupation\_Working Professional \* 2.8031) + (Lead Source\_Welingak Website \* 1.9529) + (Occupation\_Others \* 1.2718) + (Lead Source\_Olark Chat \* 1.092) + (Time Spent \* 1.0876) - (Do Not Email \* 1.2511) - (Last Notable Activity\_Email Opened \* 1.2588) - (Last Activity\_Page Visited on Website \* 1.2981) - (Last Activity\_Converted to Lead \* 1.3001) - (Last Activity\_Email Bounced \* 1.3556) - (Last Notable Activity\_Modified \* 1.3677) - (Last Activity\_Olark Chat Conversation \* 1.7954) - (Last Notable Activity\_Email Link Clicked \* 1.8287)**

## Computing the Odds

The odds are computed as the exponential value of the log odds.

A linear increase in log odds value will have a multiplicative effect on the odds of conversion.



# Inferences

The variables that contribute most towards the probability of lead getting converted are:

* Lead Origin\_Lead Add Form (3.764)
* Occupation\_Working Professional (2.8031) and
* Lead Source\_Welingak Website (1.9529)

To increase the conversion rate it is recommended to focus on:

* **Occupation:** Focus on working professionals, business men, house wife (than focusing on unemployed and students)
* **Lead Source:** Focus on leads sourced from Welingak Website and Olark Chats
* **Lead Origin:** Focus on leads received from the lead add form in the website.

Also more the time a customer spends on the website, the chances of conversion is high. So, try to provide free courses to specific focus segment of customers to have them spend time in the website.

The model is quite flexible and we can change the cutoff based on different scenarios.