Lead Scoring Case Study

Sridhar Chakravarthy

Birender Singh

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%.

- 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. To make this process more efficient, the company wishes to identify the most potential leads, also known as ‘Hot Leads’. If they successfully identify this set of leads, the lead conversion rate should go up as the sales team will now be focusing more on communicating with the potential leads rather than making calls to everyone.

Objective

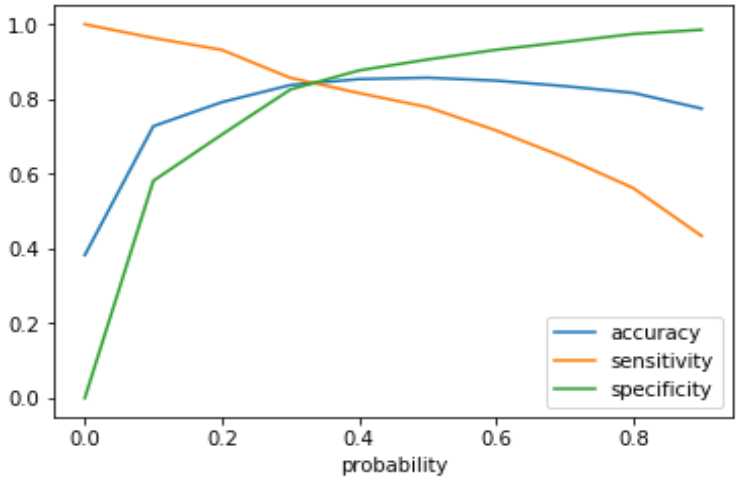
- X Education needs to data analysis to help them select the most promising leads, i.e. the leads that are most likely to convert into paying customers. The company requires a model wherein we need to assign a lead score to each of the leads such that the customers with higher lead score have a higher conversion chance and the customers with lower lead score have a lower conversion chance. The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%.

- Leads dataset from the past with around 9000 data points is provided. This dataset consists of various attributes such as Lead Source, Total Time Spent on Website, Total Visits, Last Activity, etc. which may or may not be useful in ultimately deciding whether a lead will be converted or not. The target variable, in this case, is the column ‘Converted’ which tells whether a past lead was converted or not wherein 1 means it was converted and 0 means it wasn’t converted.

Detailed Approach Of the Analysis

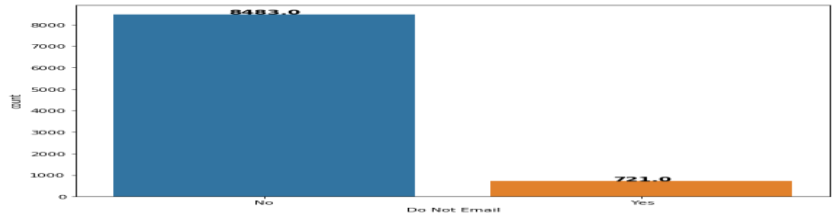
- Dropped all the unnecessary columns like country, city, prospect ID etc. as they were found to be not much useful and also columns with higher percentage of missing values were dropped.
- Imputing missing values
- Univariate analysis for categorical and numerical variables have been done
- Outlier check and treatment: Inter-Quartile Range (IQR) method is used to remove the outlier present in data for the range (0.05, 0.95).
- Binary Categorical variables were mapped with 0 and 1 while dummy variables were created for other categorical featured with more than 2 categories.
- Scaling of the data: Standard Scalar was used to scale the training data for numerical variables.
- Model building tool: Statsmodels api with RFE (Recursive Feature Elimination) method was used for model building and out of all the features given for training, Top 25 features were selected using RFE.
- Final 17 significant features were extracted with p-value almost equal to 0 and with no multicollinearity among the features (using VIF method)
- Prediction on training data was done with error terms normally distributed.
- Keeping the threshold probability as 0.5, we calculated the "Converted" flag value from the predicted probabilities. Also, accuracy, sensitivity and specificity using confusion matrix was compared to check the model performance

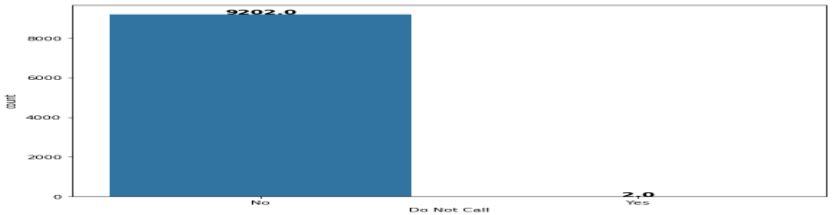
- Built a quiet stable model here. As the CEO expects the conversion rate to be around 80% - Sensitivity around 80% - we found the optimal cut-off point for probabilities using accuracy, sensitivity and specificity line plot as shown below.

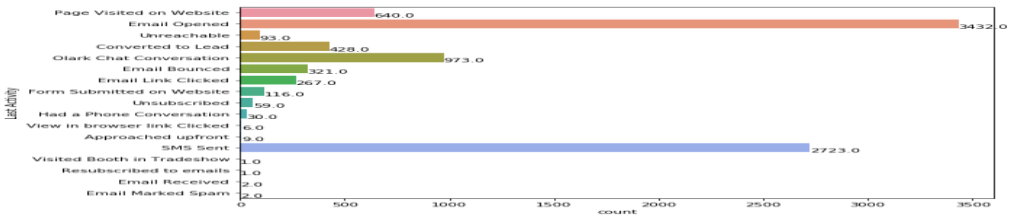


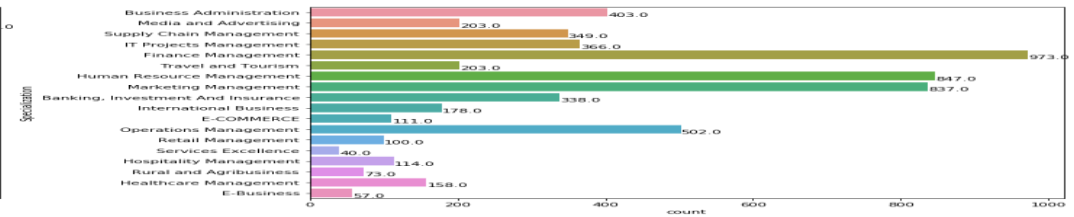
- we calculated final predicted probabilities again for training data, calculated “Converted” flag value and cross checked the confusion matrix
- The final model was applied on test data set with sensitivity received as 83% and accuracy score as 84%.
- Finally, the lead score was calculated for all the leads with range from 1 to 100. (1 being cold lead and 90-100 being hot leads)

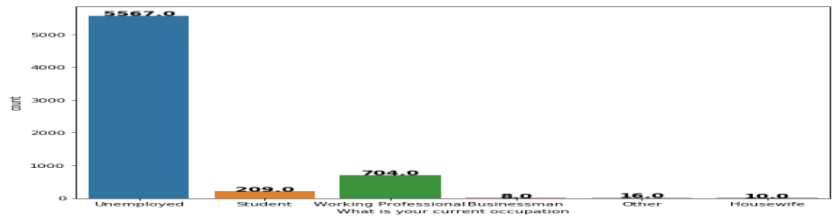
Visualization Of Plots

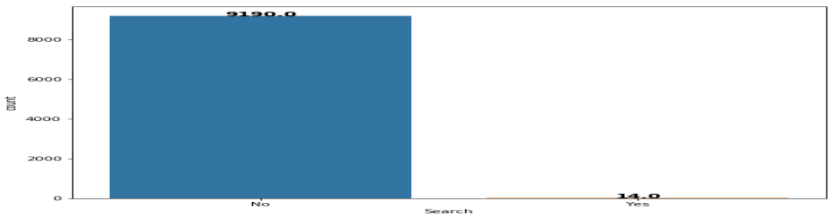


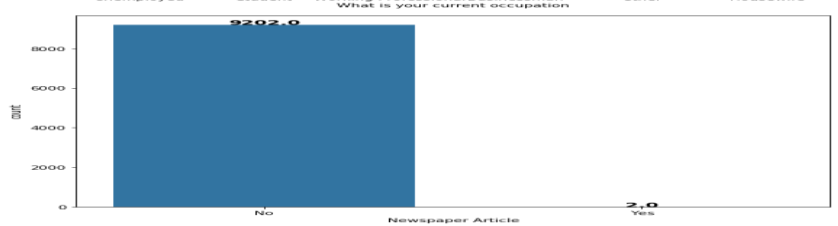


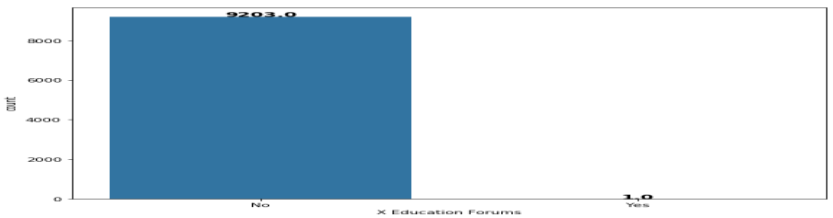


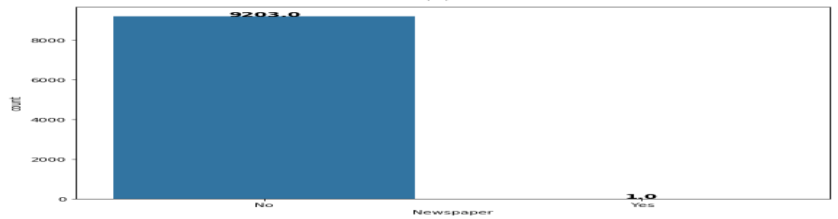


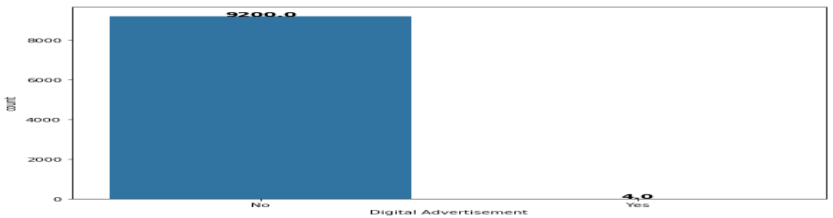


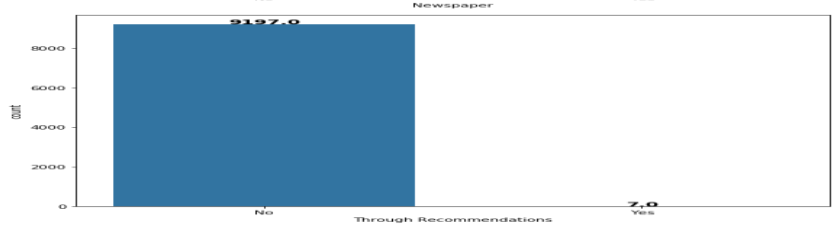


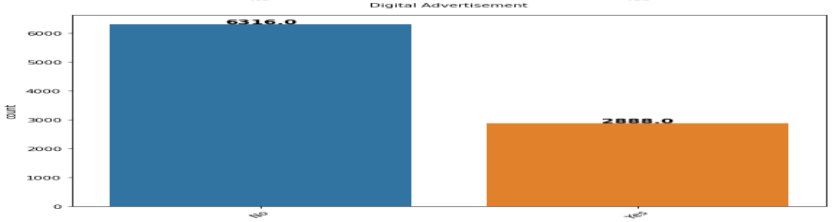






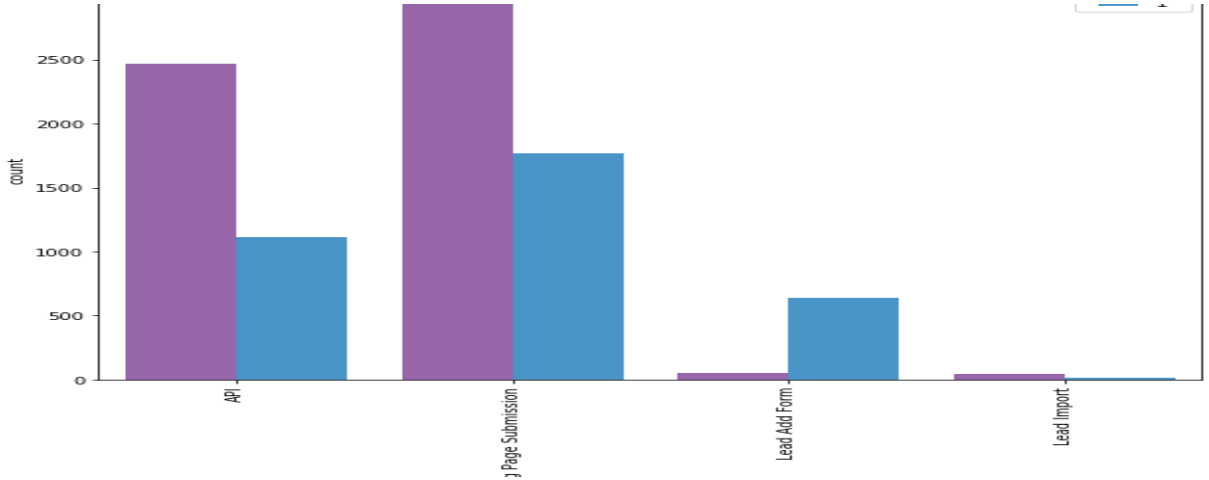




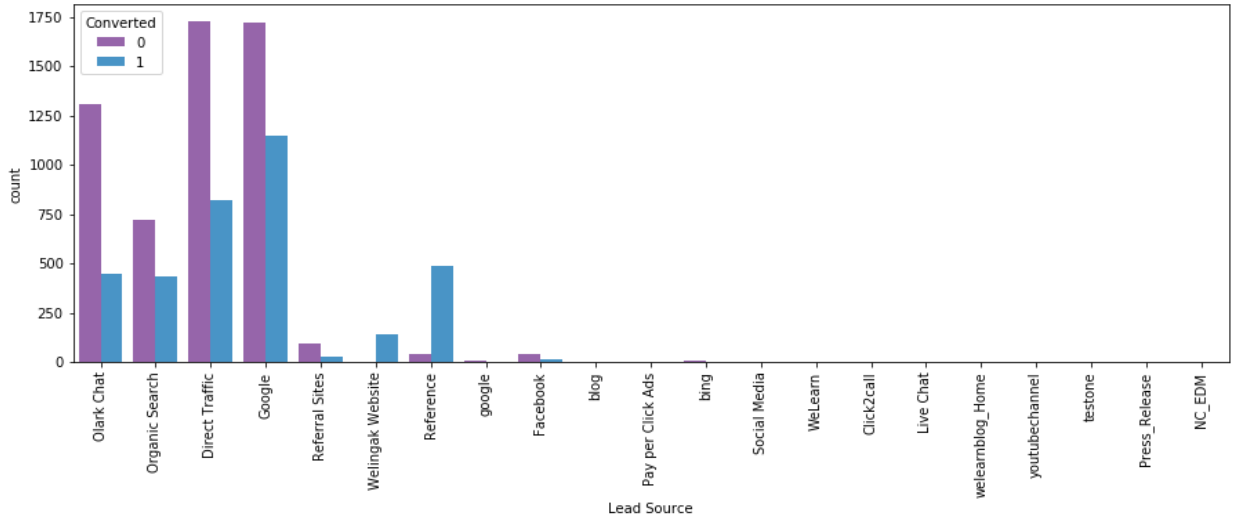


Segmented Univariate Analysis

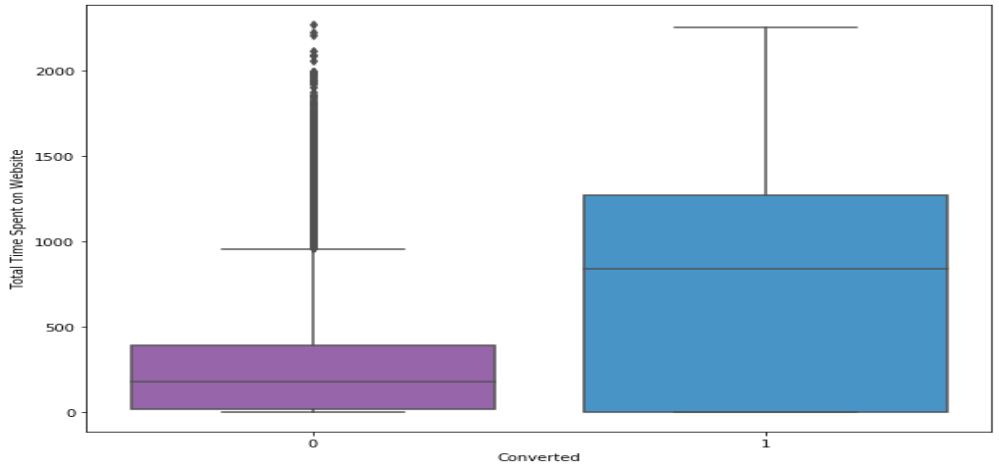
- API and Landing Page Submission have higher conversion rate but count of lead originated from them are considerable.



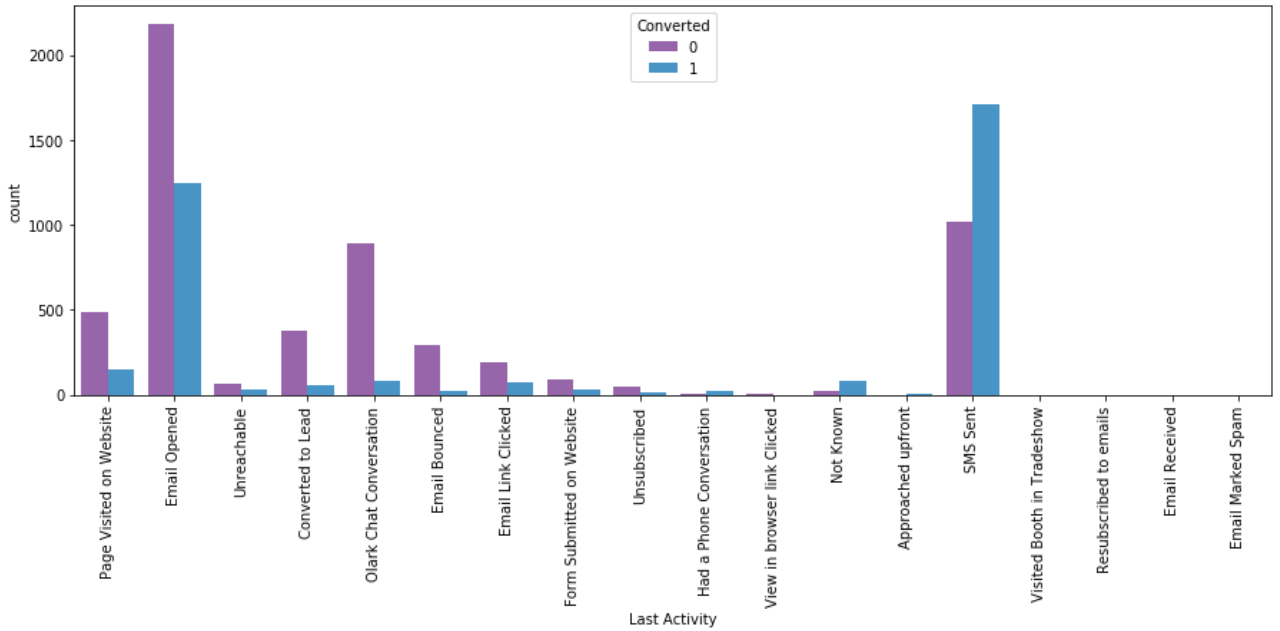
- Google, Direct traffic, olark chat generates maximum number of leads. Conversion Rate of reference leads and leads through welingak website is high.



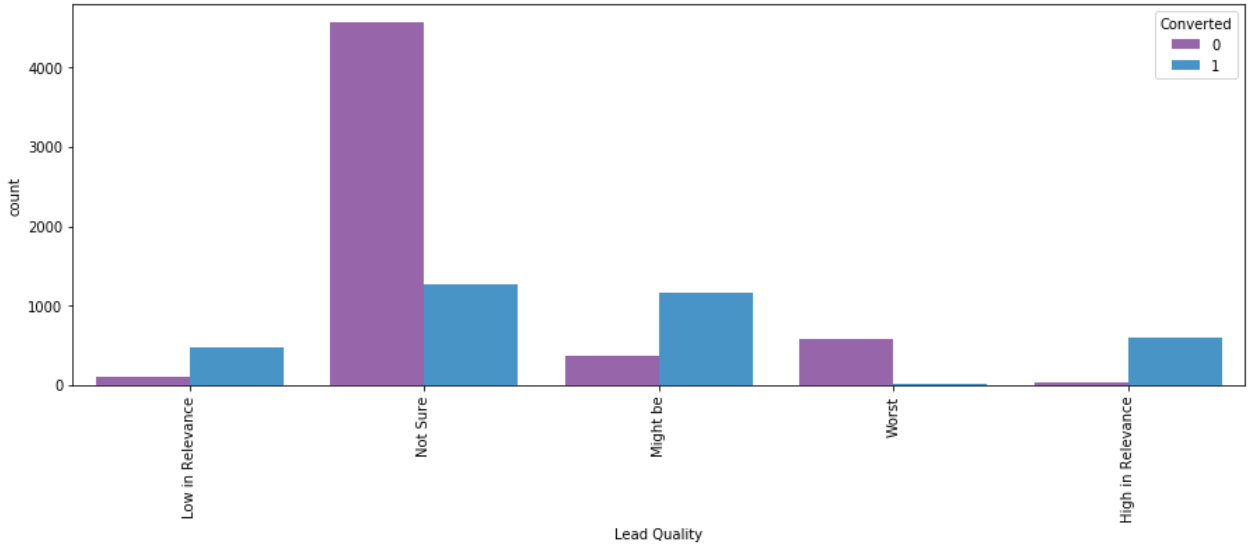
- Leads spending more time on the website are more likely to be converted.



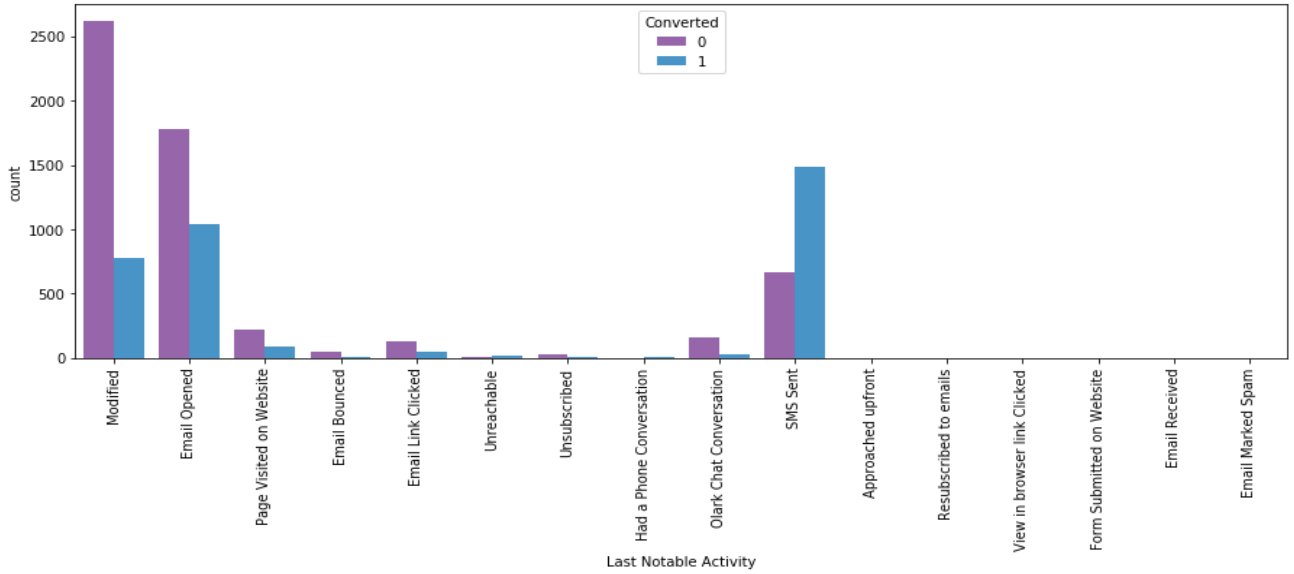
- Most of the leads have their last activity as "Email opened". Conversion rate for leads with last activity as "SMS Sent" is really high.



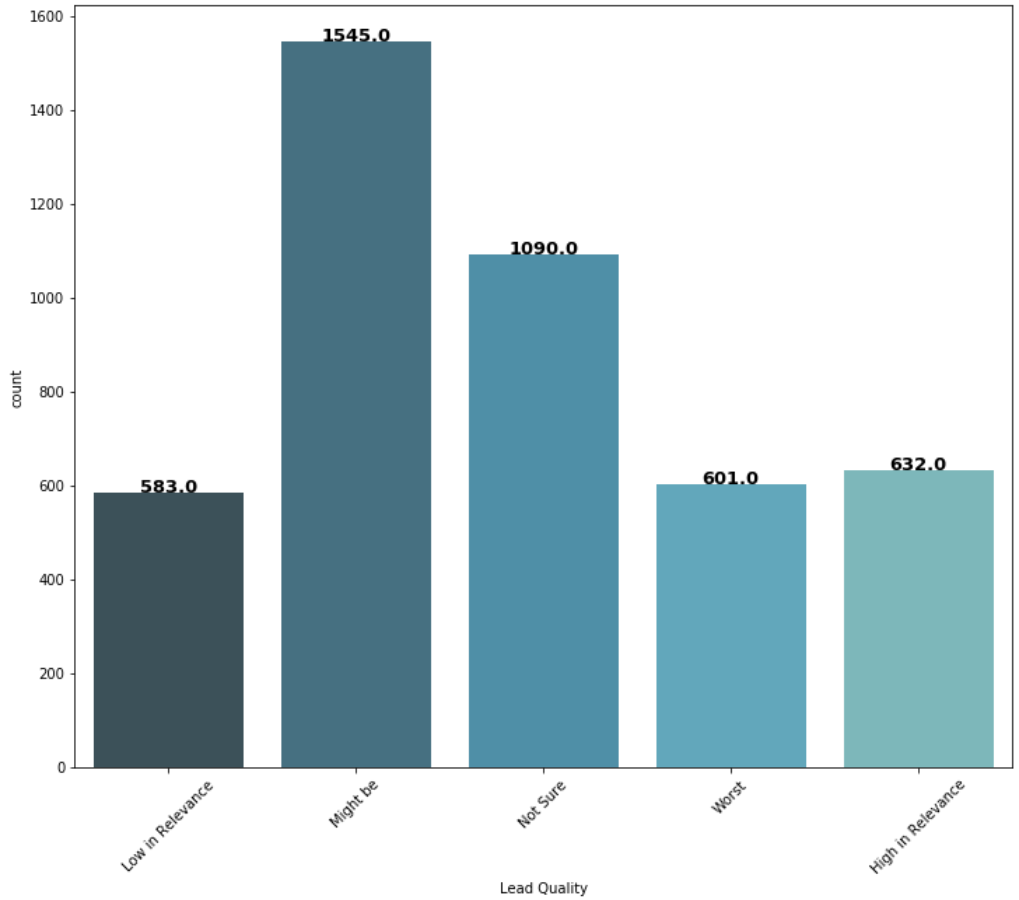
- We can infer that "Lead Quality" feature with values as high, low and might be as higher conversion rate among all the others though the count of leads is highest for "Not sure" category.

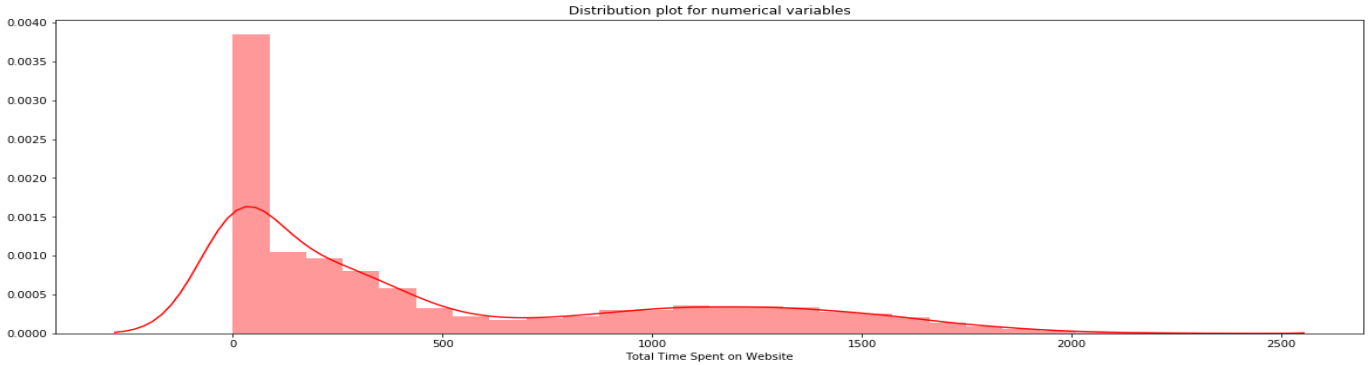


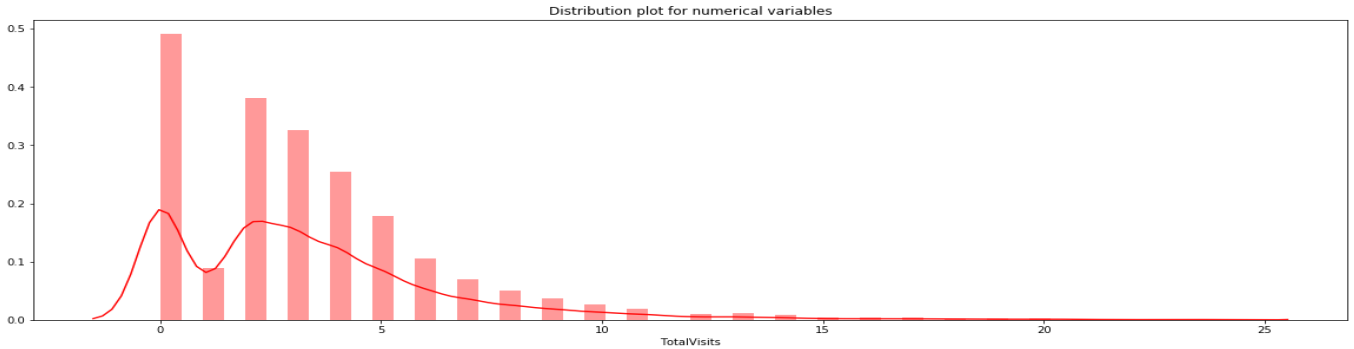
- "Last Notable Activity" feature categories are mainly related to SMS, chat conversation and Emails. Most of the leads have their last activity related to emails. But SMS category has highest conversion rate.

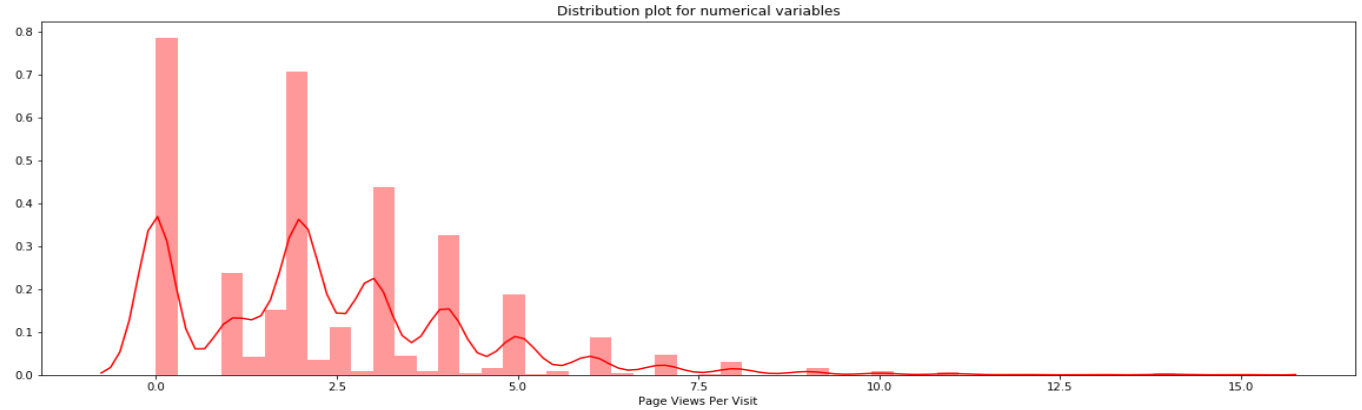


Univariate Analysis









We can infer from the plots that Website total visits are mainly around 0-10 visits. while total time spent on website is about 500 seconds and also it increases till 1500 seconds for few customers.

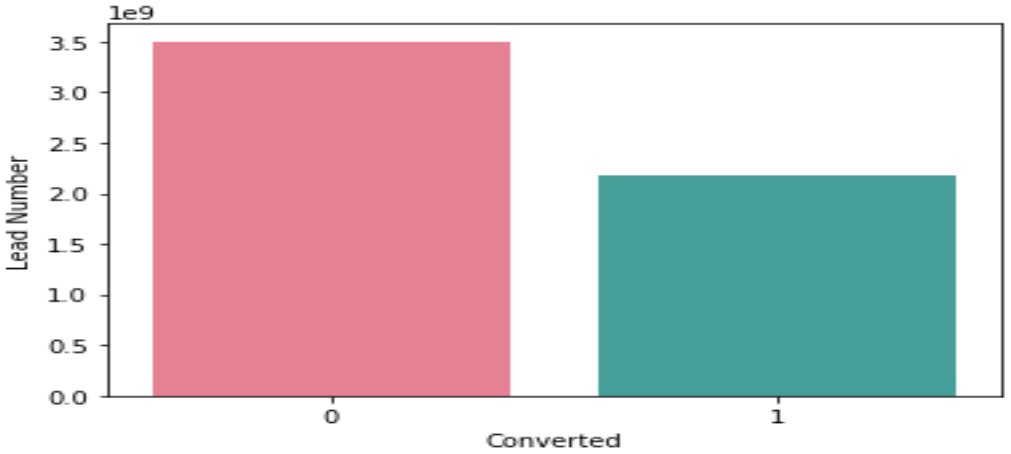
Almost all the users are viewing around 2-5 Pages minimum

Data Modelling

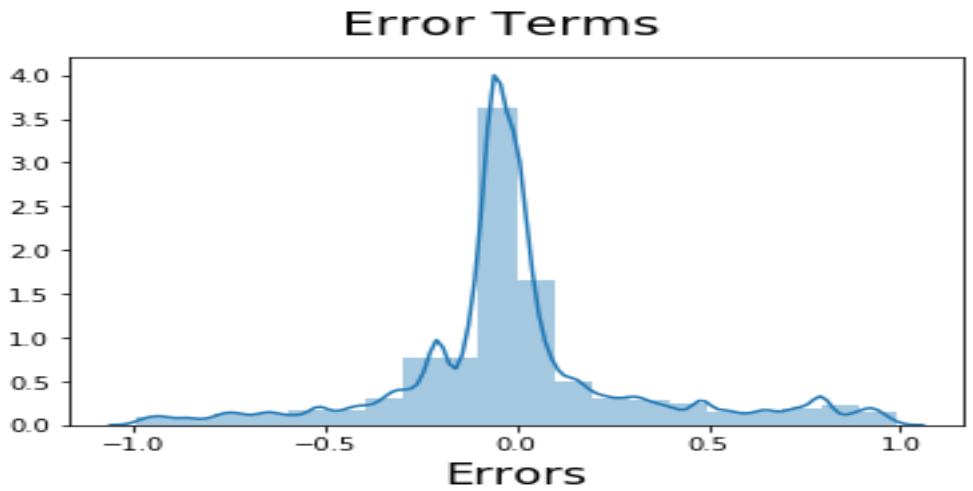
Data Modelling Steps are as followed

- Dummy variables creation
- Label encoding
- Binary value creation for columns with Yes/No
- Train Test split
- Scaling the data using Standard Scaler only for numeric variables
- Building the model
- Apply and test the model on training dataset
- Validate the model on test dataset
- Reporting the insights

- The below graph is the conversion rate of the data: (lead conversion rate is around 38.42%) as mentioned even in the problem statement stating it is a poor conversion ratio.



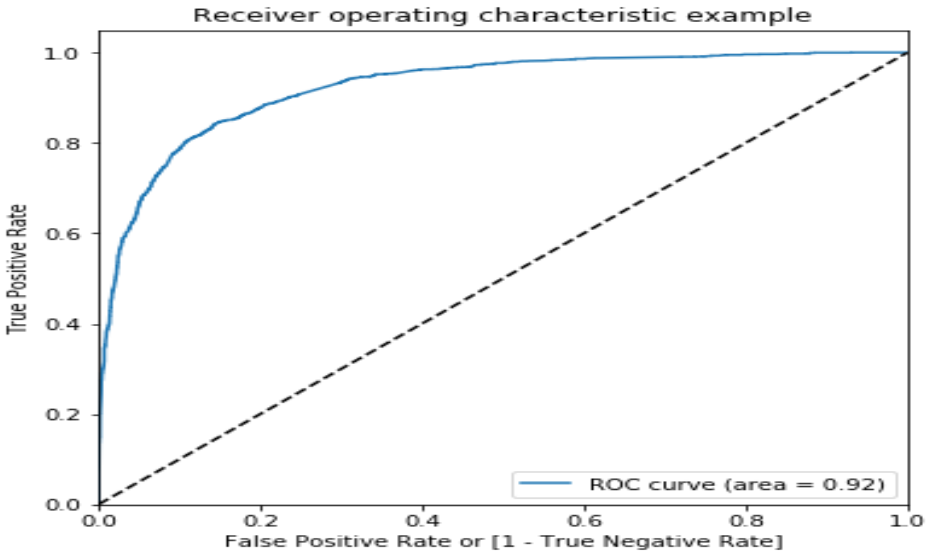
- The below graph on train data set shows that error terms are normally distributed, which concludes that our model doesn’t violate any regression rule.

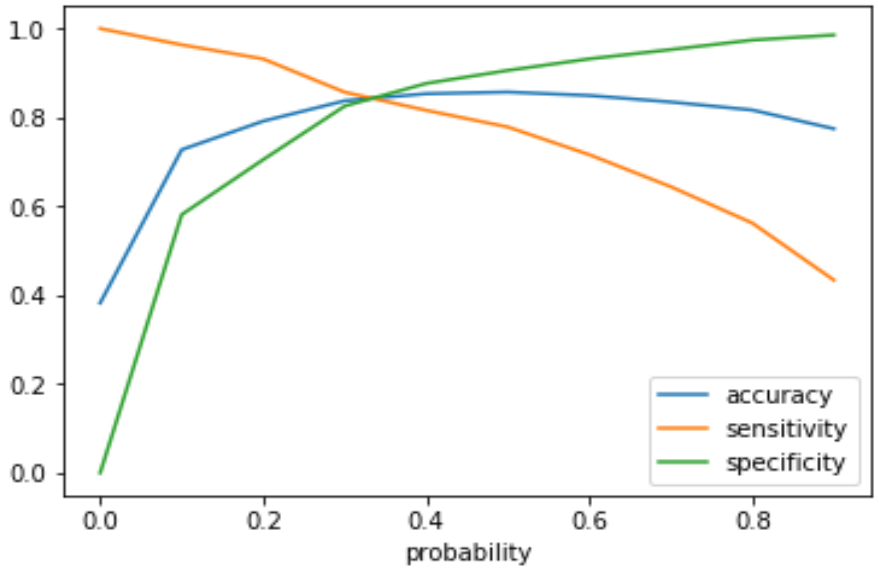


The model gives below variables list as the most significant factors :

- Do Not Email
- Total Time Spent on Website
- Lead Origin_Lead Add Form
- Lead Source_Olark Chat
- Lead Source_Welingak Website
- Last Notable Activity_Email Link Clicked
- Last Notable Activity_Email Opened
- Last Notable Activity_Modified
- Last Notable Activity_Olark Chat Conversation
- Last Notable Activity_Page Visited on Website
- Lead Quality_High in Relevance
- Lead Quality_Low in Relevance
- Lead Quality_Might be
- Lead Quality_Worst
- Last Activity_Olark Chat Conversation
- Last Activity_SMS Sent
- What is your current occupation_Working Professional

Some important statistical graphs like ROC curve that we have derived after the calculation of regression equation are as below in the form of graphs.





The Area under curve is around 0.92, which is quite high and a very well representative measure of how our final model is able to distinguish between the two diagnostic groups i.e. Conversion or No-Conversion.

Here, the curve is a convergence point for accuracy, sensitivity and specificity.

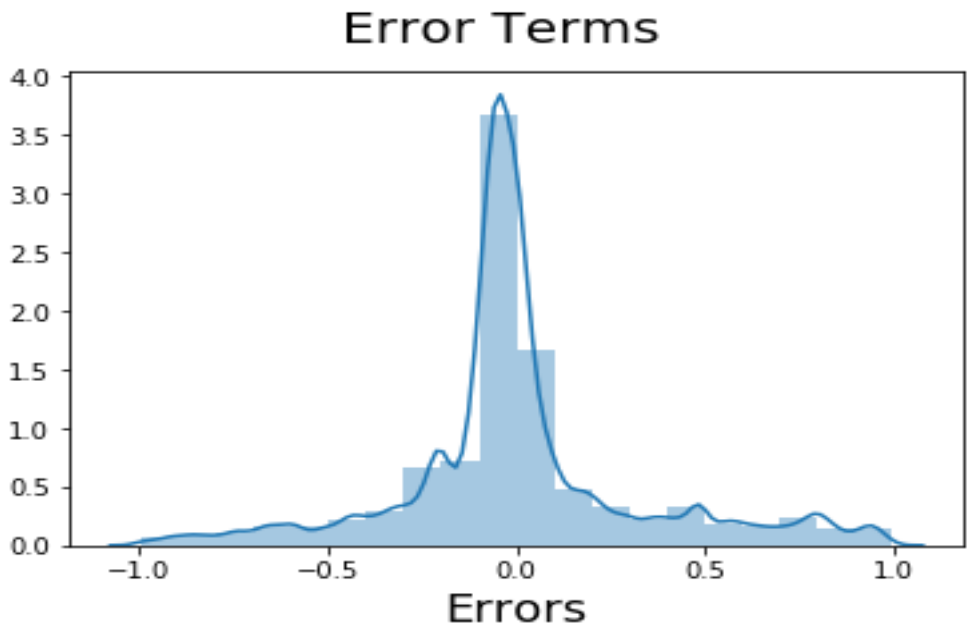
The intersection point represents where the probability cut-off should be, and as per our graph, the cut-off point should be around **0.35**.

Using the same cut-off of 0.35, we calculated final predicted probabilities again for training data, calculated “Converted” flag value and cross checked the confusion matrix again to get conversion rate around 80%. This time sensitivity was found to be around 83% as expected.

At last, the final model was applied on test data set with sensitivity received about 83% and accuracy score about 84%.

Finally, the lead score was calculated for all the leads with range from 1 to 100. (1 being cold lead and higher score around 90-100 being hot leads)

We could observe in the below graph that Error terms are normally distributed even for Test data.



Conclusion

We can conclude that below features are of utmost importance when it comes to increasing the lead conversion rate for the online company selling courses. Also while doing the univariate and segmented univariate analysis, Below variables were proved to be more influential for increasing conversion rate.

- Total Time spent on Website
- Lead origin with add form
- Lead Source with values (Olark chat, Welingak Website)
- Lead Quality (High, low)
- Last Activity performed by the user, can be anything like ranging from chat, SMS etc.
- Current occupation like working professional

Thank You