



Lead Score Case Study

Presented by

Abhinandan Gupta

Monica

Sonal Khot



Problem Statement


1. X Education sells online courses to industry professionals.
2. X Education gets a lot of leads; its lead conversion rate is very poor. For example, they acquire 100 leads in a day, only about 30 of them are converted.
3. To make this process more efficient, the company wishes to identify the most potential leads, also known as 'Hot Leads'.
4. 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.

Business Objective

1. X education wants to know most promising leads.
2. For that they want to build a Model which identifies the hot leads.
3. Deployment of the model for the future use.



Solution Methodology

1. Data Cleaning
 2. Data manipulation
 3. EDA
 4. Feature Scaling & Dummy Variables and encoding of the data
 5. Logistic Regression used for the model making and prediction
 6. Validation of the model
 7. Model presentation
 8. Conclusion and recommendation
- 

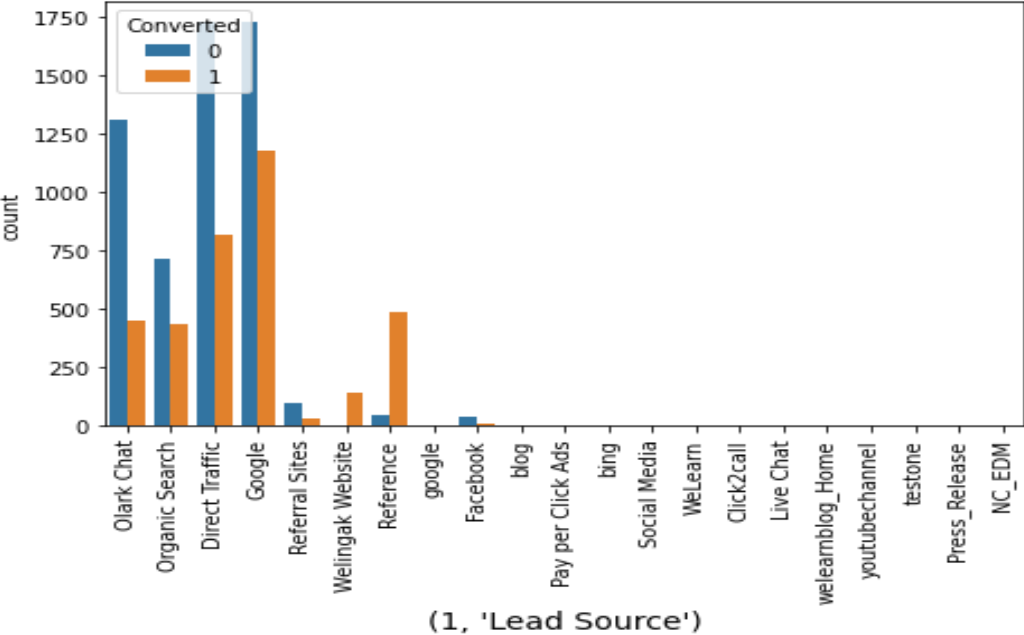
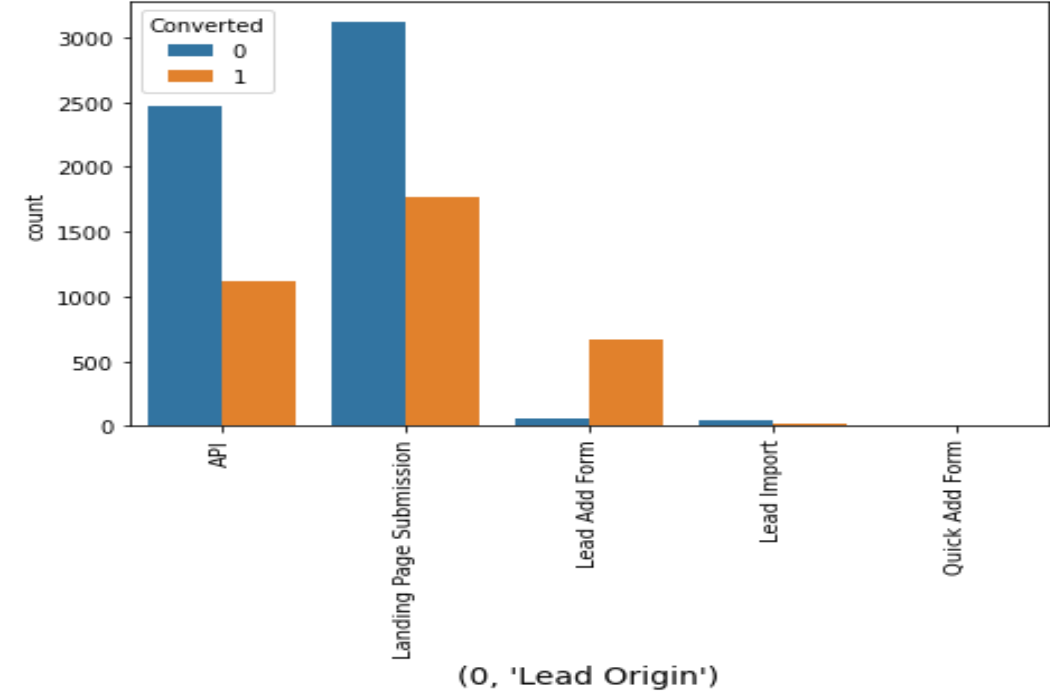


Data Manipulation

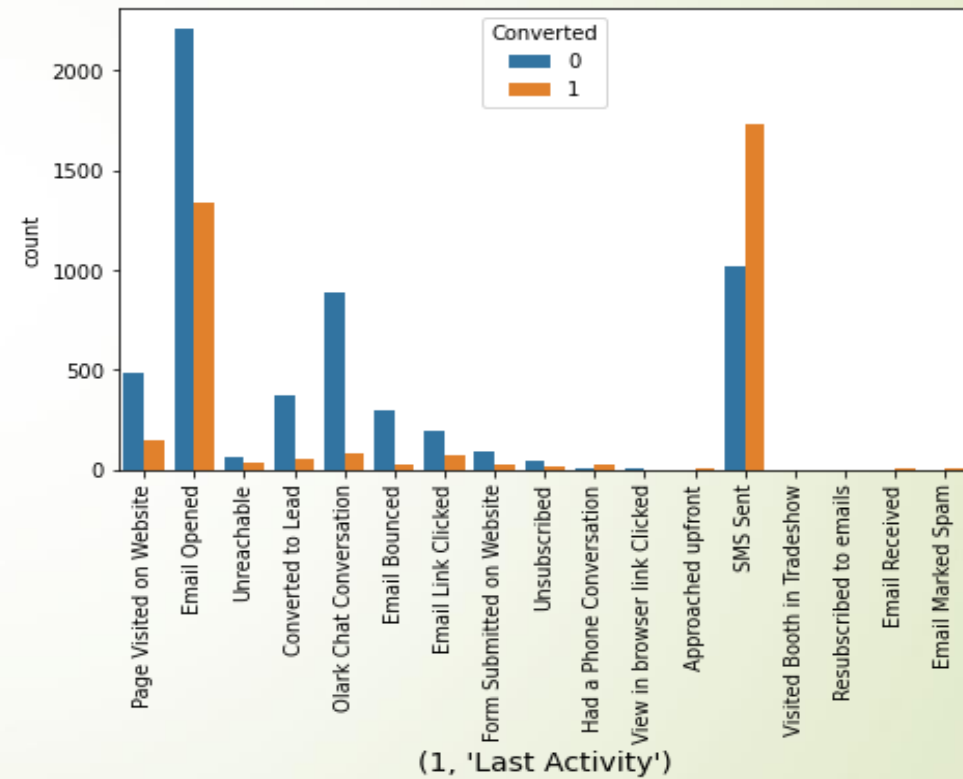
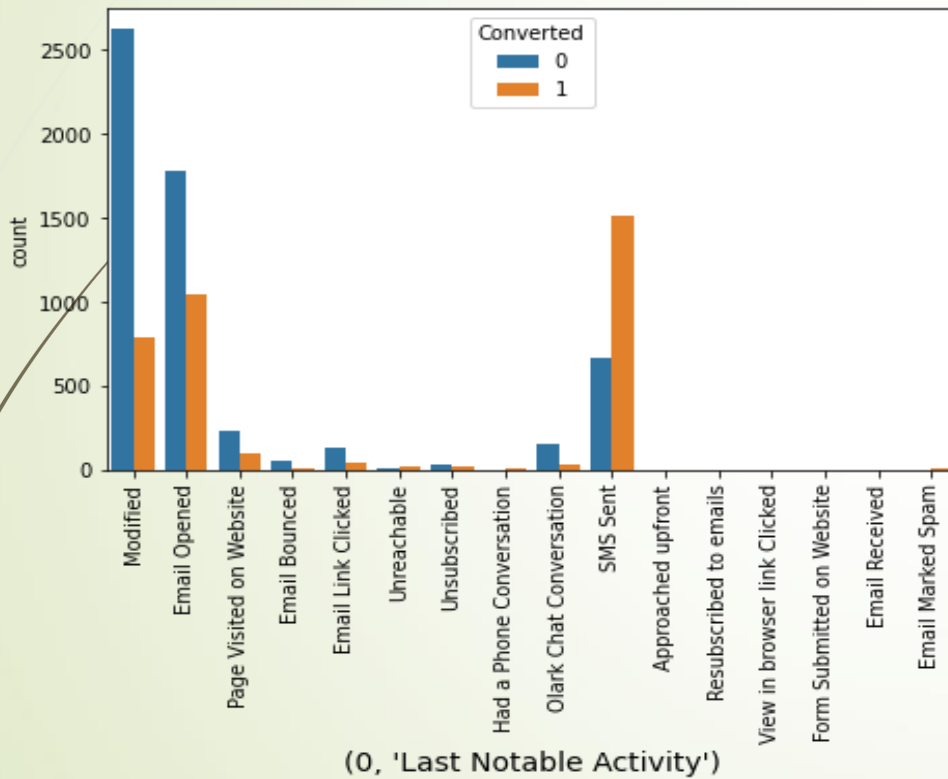
1. Loading CSV File namely - 'Leads.csv' .
2. Data Cleaning
3. Checking the features using Python commands(Describe, Info, Shape).
4. The variables has “select” values are replaced with Null values
5. Removed Columns Which have More than 40% of null values in the respective columns variables which are not significant for analysis as well as the variables has only one category
6. Checking Columns Which have less than 40% of null values and imputed based on the type of Columns Choosing Mean, Median and Mode.

Categorical Segmented Univariate Analysis

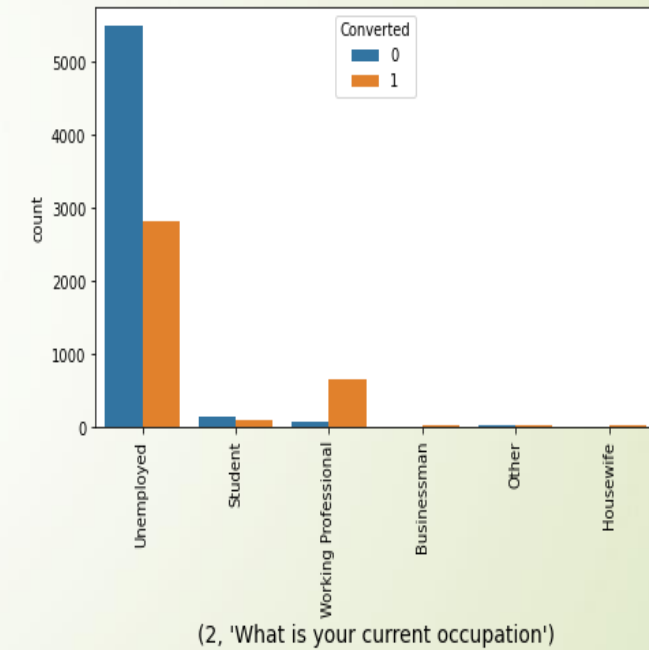
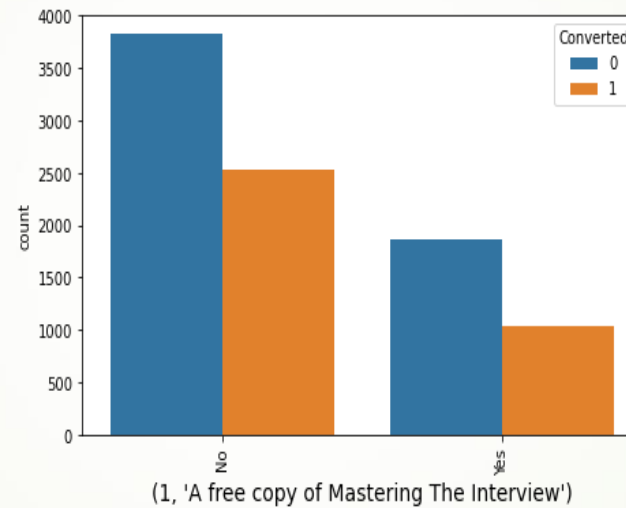
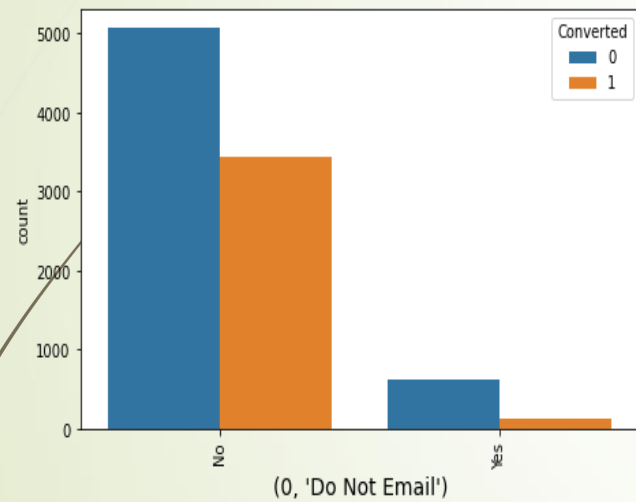
Most of leads search on Google, 'Landed on page submission are converted to hot lead



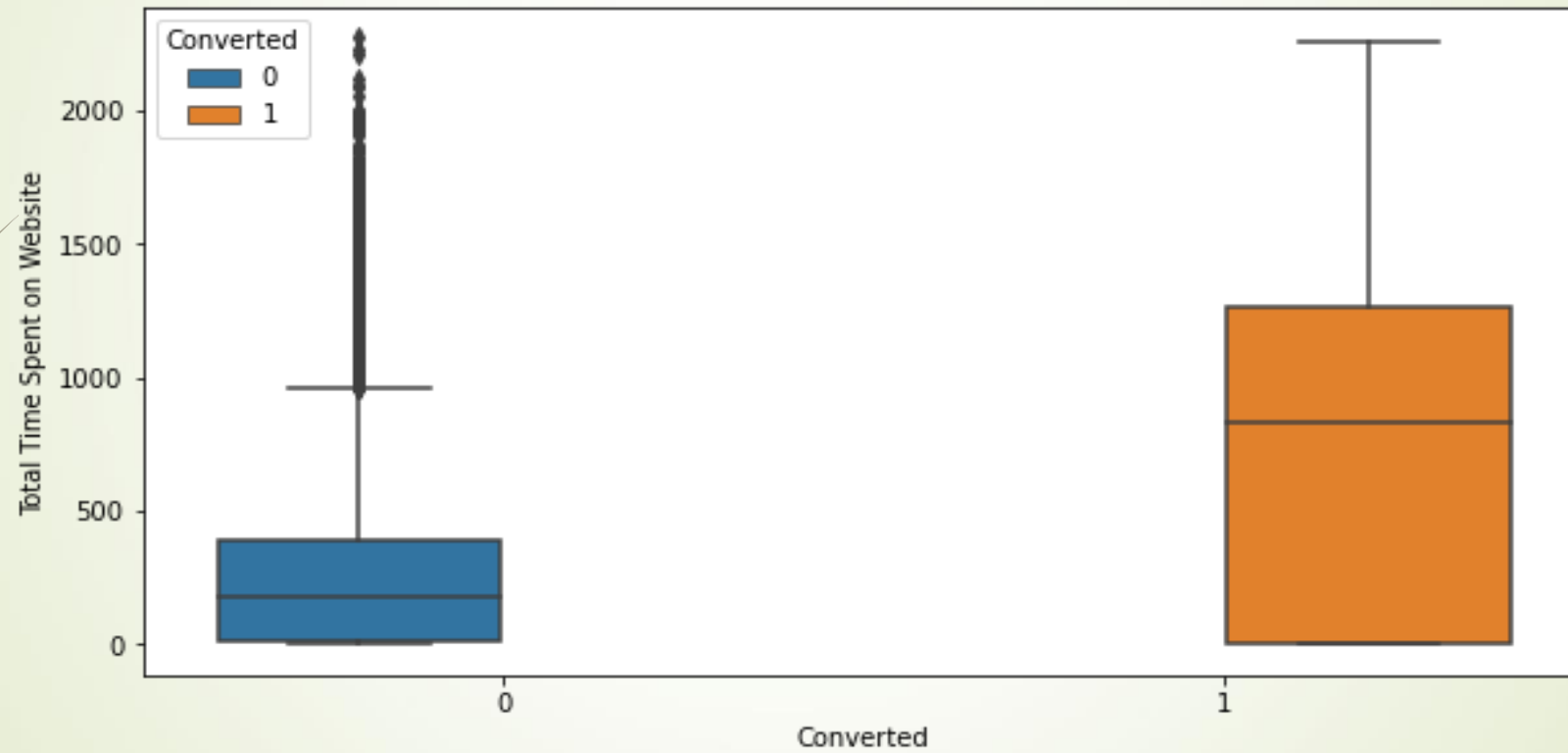
Most of leads to “sent sms” and those “email opened” are converted to hot lead



The leads who allow to sent mail and unemployed are converted to hot lead




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





Data Preparation

1. Converted binary variables (Yes/No) to 0/1
 2. Total Columns for analysis are 13
 3. Dummy Variables are created for Object Type variables
 4. After creating dummy variables total Column for analysis are 50
 5. Numerical Variables are Scaled by using Standard Scaler method
- 

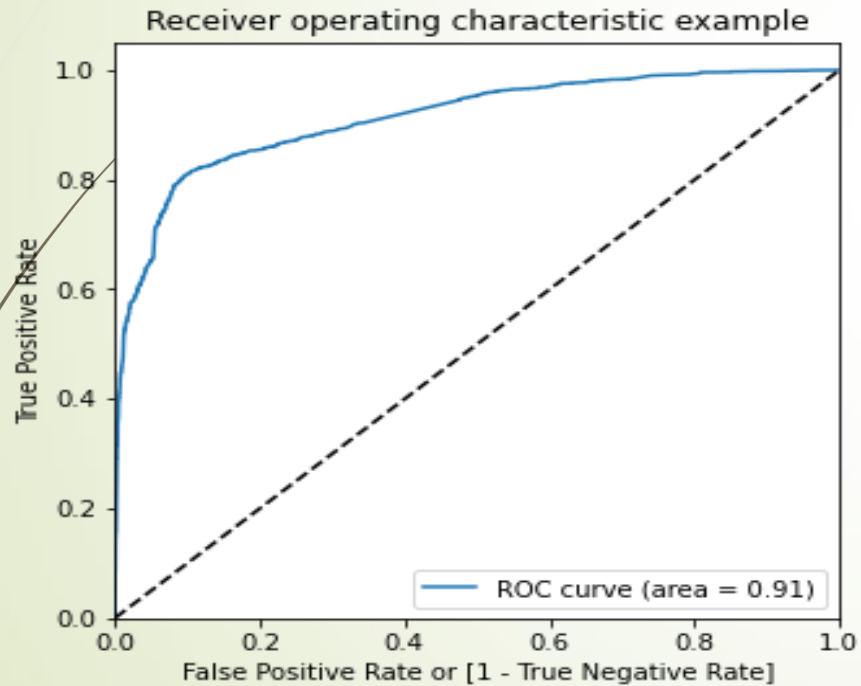


Model Building

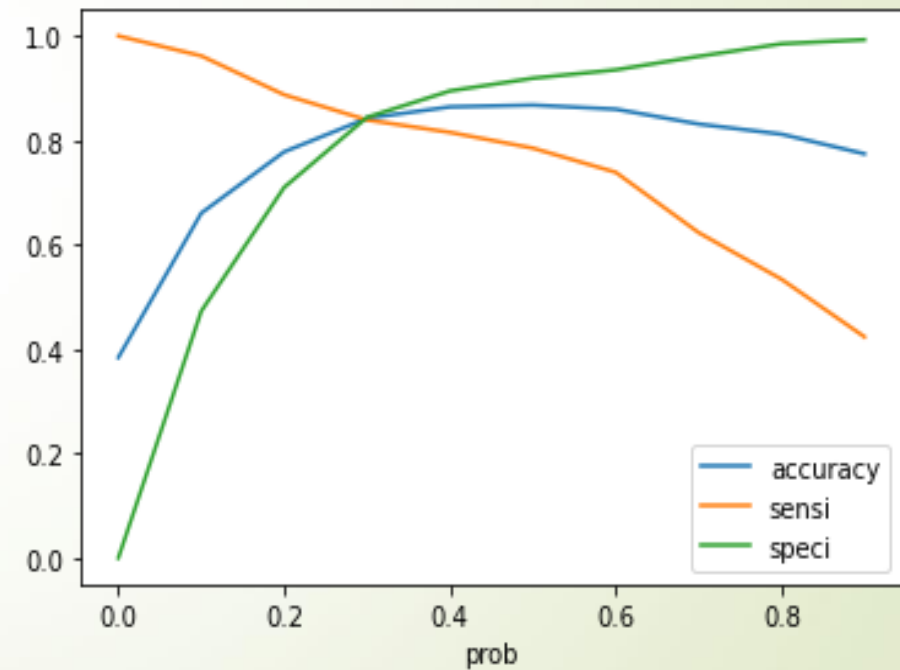
1. First Step in Logistic Regression was performing a train test split, we have chosen 70:30 ratio.
2. Splitting data into Train and test sets
3. Use RFE for feature Selection
4. Running RFE with 15 Variables as Output
5. Building Model by removing the variable whose p-value is greater than 0.05 and VIF value is greater than 5
6. Predictions on test data set

ROC curve

It shows the tradeoff between sensitivity and specificity



0.3 is the optimum point to take it as a cutoff probability





Result of model

Result on Train data


1. Accuracy : 84.21%
2. Sensitivity : 83.89%
3. Specificity: 84.41%
4. Precision : 85.78%
5. Recall: 78.50%

Result on Test data

1. Accuracy : 84.55%
2. Sensitivity : 85.88%
3. Specificity : 83.71%
4. Precision : 77%
5. Recall: 85.88%
6. ROC : 0.91



Conclusion

1. The logistic regression model predicts the probability of the target variable having a certain value, rather than predicting the value of the target variable directly. Then a cutoff of the probability is used to obtain the predicted value of the target variable.
 2. Here, the logistic regression model is used to predict the probability of conversion of a customer.
 3. Optimum cut off is chosen to be 0.3
 4. “Do Not Email” , “Total Time Spent on Website” , “Lead Origin”, “Last Notable Activity” and “When their current occupation is as a working professional” are important features for converting to hot leads
- 

Thank You!!!

