Lead Scoring Case Study

GAYATHRI MNV

RAMASUBRAMANIAM P

Problem Statement

- Build a logistic regression model for X Education Company to assign a lead score between 0 and 100 to each customer.
- Customer with higher lead score have a higher conversion chance and the customers with lower lead score have a lower conversion chance
- 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.
- Model to be built with a ballpark of target lead conversion rate around 80% .
- Model should be able to adjust to if the company's requirement changes in the future.

Approach

- Data Import & Study
- Data Cleaning
- Data Analysis (EDA)
- Data Preparation
- Model Development
- Model Prediction
- Final Conclusion
- Recommendation

Data Import & Study

Dataset Name	Total Rows	Total Columns	Categorical Columns	Continuous Columns
Lead dataset	9240	37	30 (81 %)	7(19 %)

- Lead Dataset consists of 9240 rows & 37 Columns, 30 are Categorical variables which contribute 81% of the data & Continuous are 19% of the data
- Identified Target variable – 'Converted' , important columns and unnecessary columns.
- Observed the 'Select' label in the categorical variables & it will be handled as per the variable's importance.
- Statistical summary for continuous variables shows there are many missing values & no negatives values.

Data Cleaning

Lead Dataset	Total Rows	Total Columns	Categorical Columns	Continuous Columns
After Cleaning	9074	11	9	4

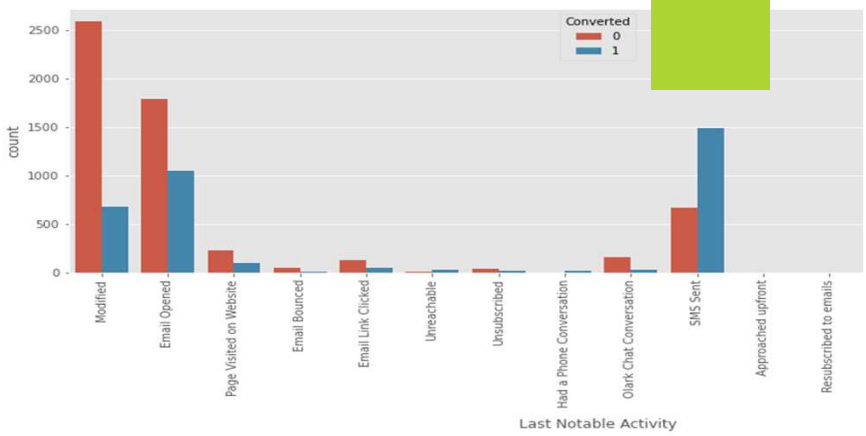
- Categorical & Numerical Columns with missing values above 35% and unnecessary columns like Prospect ID & Lead Number are dropped.
- Categorical columns with missing values between 5% - 35 % are imputed with the most frequent values.
- 'Select' label in **Specialization** is a label as this variable is important for the Education Company to make a lead to a Hot Lead.
- 'Select' in other 'Lead Profile' & **How did you hear about X Education'** columns is considered as equivalent to missing value Nan.
- Rows corresponding to Columns with missing values less than 5% are dropped.
- Binary Columns above 95% of 1 or 0 values are considered as Skewed columns and dropped from dataset.

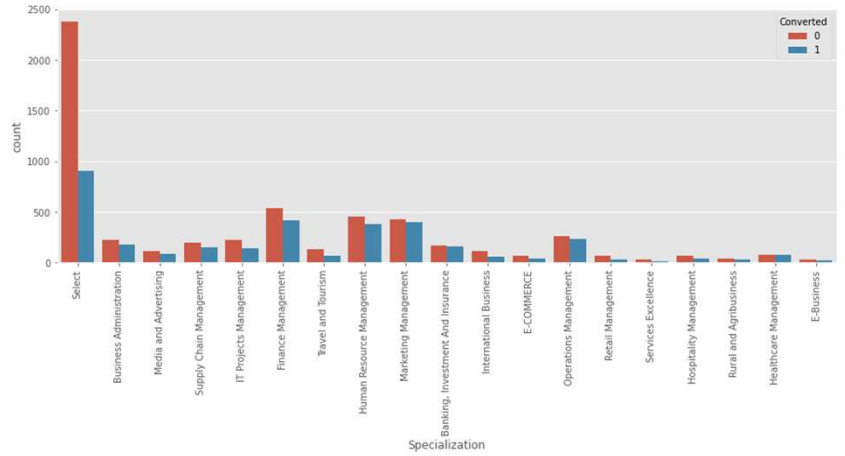
Data Analysis (EDA)

Categorical Variables Analysis

Based on the analysis of Categorical variables, most leads converted to Hot Leads are observed to be highest in,

- Lead Origin - Landing Page Submission , API and Lead Add Form
- Lead Search - Google ,Direct Traffic, Organic Search & Olark Chat
- Last Activity - SMS Sent and Email Opened
- Specialization - Finance Management , Human resources Management & Banking.
- What is your current Occupation - Unemployed
- A free copy of Mastering The Interview - No
- Last Notable Activity - SMS Sent , Email Opened & Modified

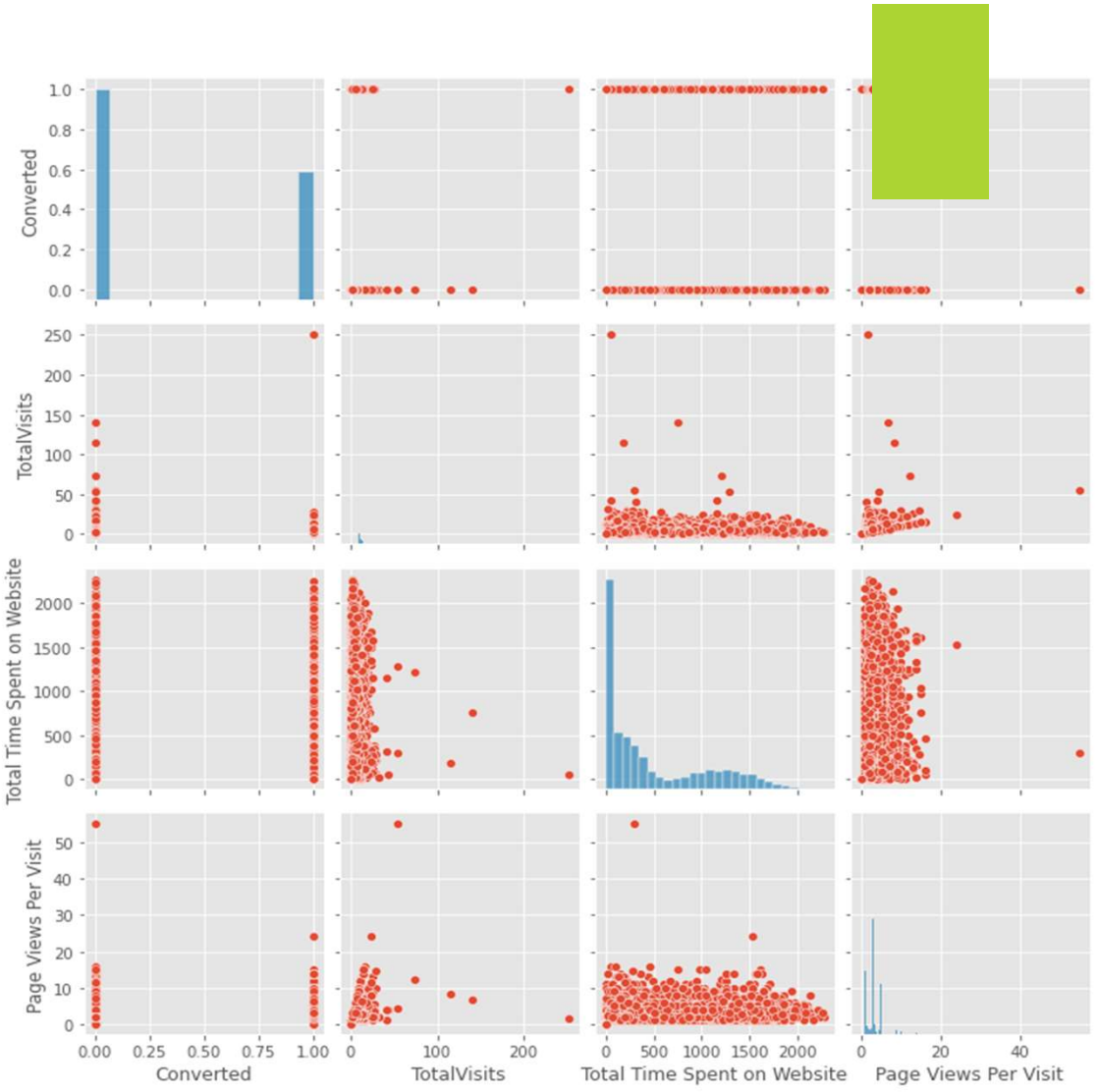




Continuous Variables Analysis

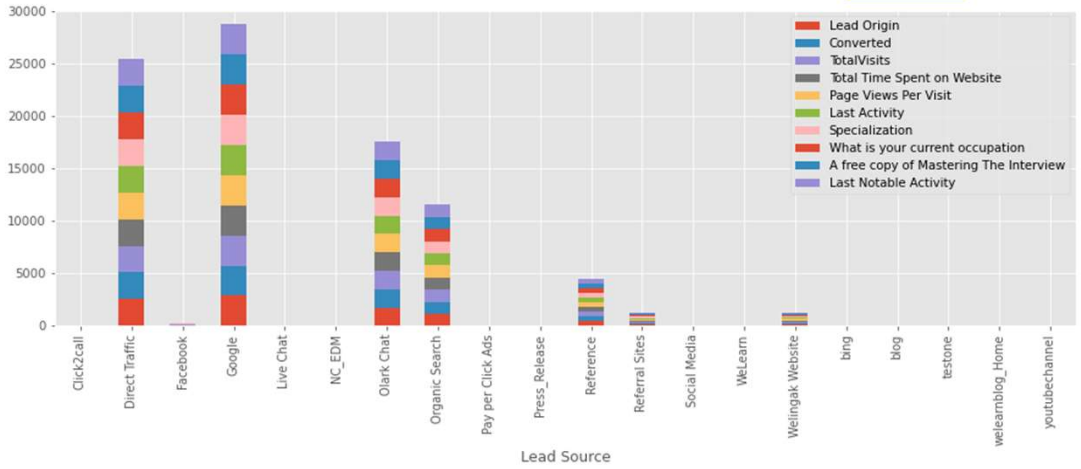
Based on the Analysis of Continuous variables,

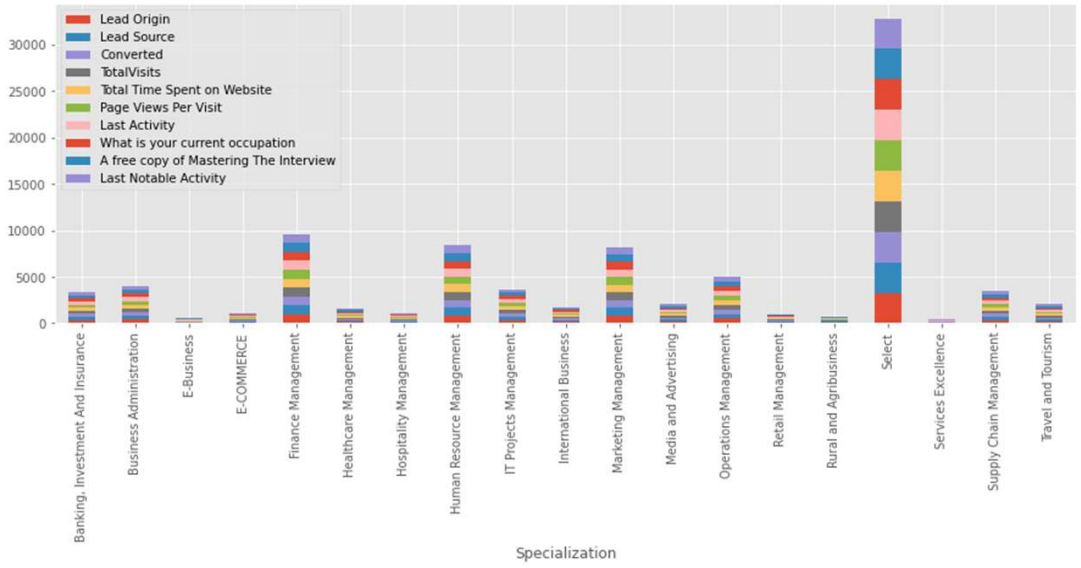
- Target Variable – 'Converted' is observed to have 38% of Hot Leads and 62% of Cold Leads.
- Continuous variables are right skewed & Clustered.
- Leads with high 'TotalVisits' are observed to be converted to Hot Leads.
- Most of the leads with "Total Time Spent on Website are those who converted to Hot Leads.
- Page Views per visit do not have any impact on Lead conversion



Multivariate Analysis

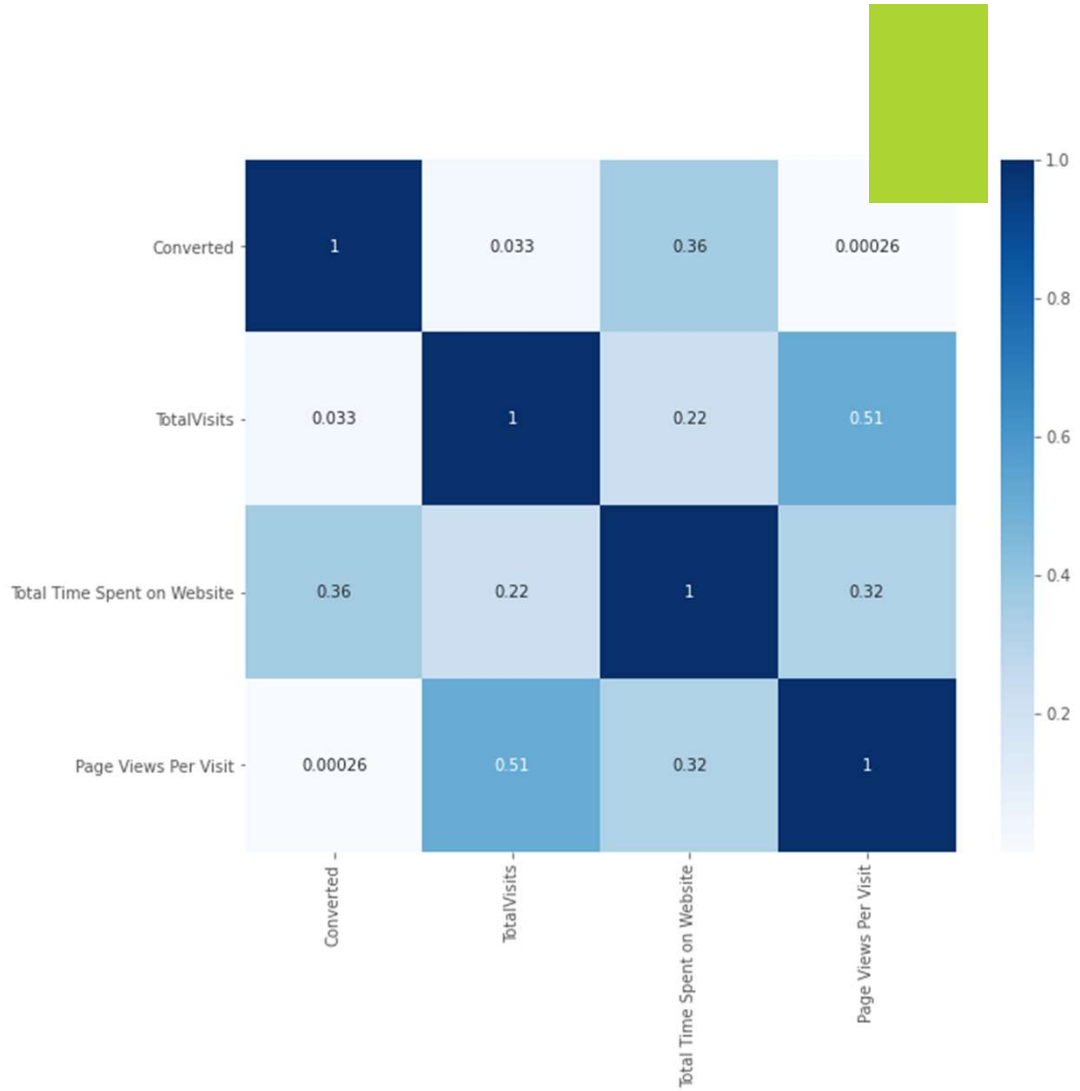
- From the stacked Bar charts, we can find that most of the leads converted to Hot Leads activities like,
- Specialization opted are Finance, HR, Marketing and Operations Management.
- Lead Source is Google, Direct Traffic, Olark Chat & Organic Search.
- Last Activity was tracked to be ‘Email Opened’, ‘SMS Sent’ Olark Chat Conversation and Page viewed on website.
- Last Notable Activity was Modified, Email Opened, SMS Sent & Page visited on Website.
- These actions will be helpful in predicting the leads conversion to Hot Leads





Correlation Matrix

- Heat map shows that there is positive relationship between the continuous variables and the target variable 'Converted'.
- Total Time spent on Website has a low correlation with the target variable.
- TotalVisits & Page Views per visit variable has almost no relationship with the target variable.
- No Multi-Collinearity observed between the variables.
- We can conclude that there is no effect of these variables in lead to Hot Lead conversion

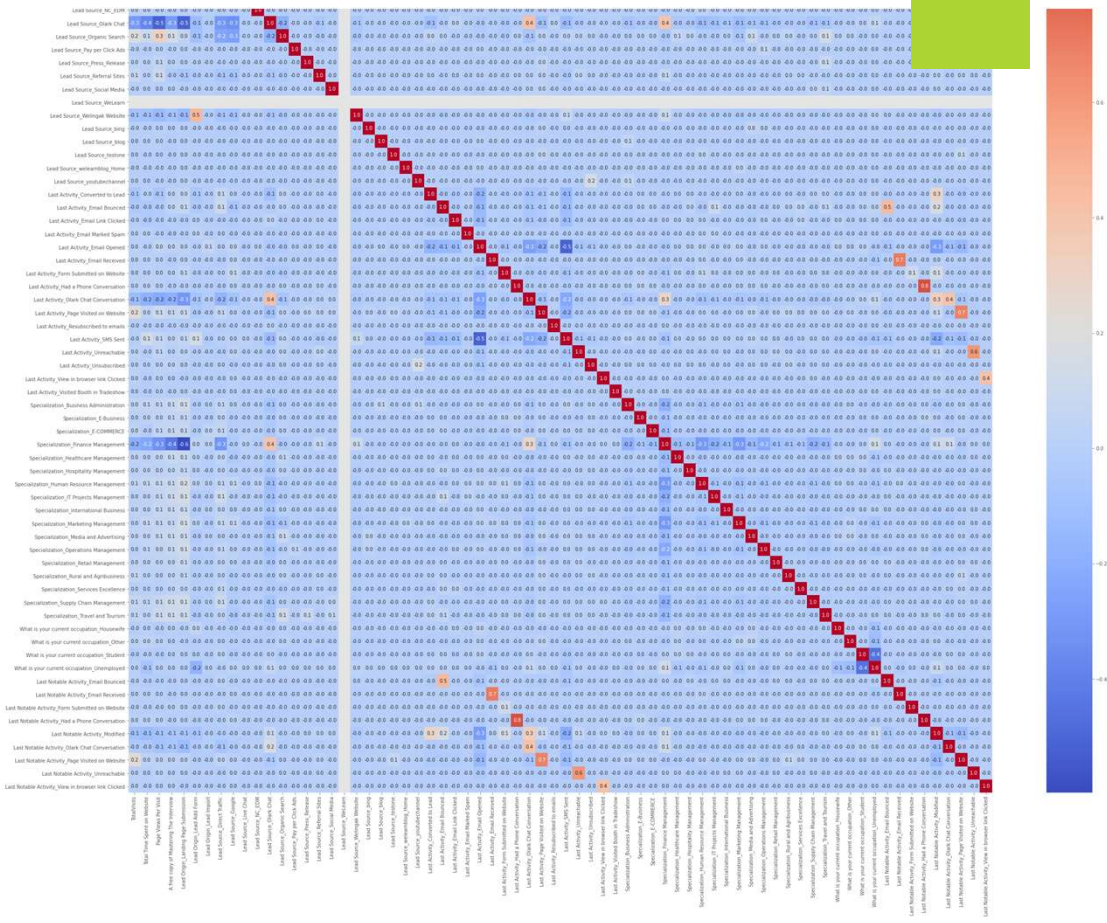


Data Preparation

- As a part of Data Preparation, did Binary variables encoding & Dummy variables creation.
- Binary Variables Encoding – Converted “A free copy of Mastering The Interview' variable with ‘Yes’/ ‘No’ to 1/0 .
- Dummy Variables – Created Dummies data frame for all the categorical columns (Specialization Separately).
- Created a new_lead_dataframe by concatenating the lead & the dummies and then dropped the categorical columns for which the dummies are created in the new data frame.
- The new_lead_dataframe has 9074 rows & 81 columns.
- Outliers are detected & those will be handled by StandardScaler.

Data Transformation

- Created feature X & y variables.
- Split the data into X_train, X_test,y_train & y_test in the ratio of 70 : 30.
- X_train has 6351 rows & 80 columns
- y_train has 6351 rows
- X_test has 2723 rows and 80 columns
- y_test has 2723 rows.
- Features scaling on numeric columns is done using StandardScaler.
- X_train is fit transformed using the StandardScaler .
- Variables with high Multi Collinearity at a cut-off of 0.8 are dropped from train & test sets



Model Development

- Logistic Regression Model is developed using sklearn & stats models.
- Built first Model using RFE method with 15 variables.
- RFE picks the top & the most significant features ,hence reducing the time & cost.
- Reviewed the Model summary to understand the coefficients of the features and their respective p-values
- Checked for the VIF values for the features on the model built.
- Dropped the features with p-value above 0.05 or VIF above 5 and built a new model.
- Iteratively performed building the models till all the features p-value & VIF's are below 0.05 & VIF's below 5.
- Final Model is built on 12 significant features which will explain the top features that help in leads conversion to Hot or Cold Leads.

Final Model Summary & VIF Values

Generalized Linear Model Regression Results

Dep. Variable:	Converted	No. Observations:	6351
Model:	GLM	Df Residuals:	6338
Model Family:	Binomial	Df Model:	12
Link Function:	logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-2909.5
Date:	Wed, 08 Sep 2021	Deviance:	5819.1
Time:	01:41:05	Pearson chi2:	6.58e+03
No. Iterations:	7		
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-0.0345	0.087	-0.396	0.692	-0.205	0.136
Total Time Spent on Website	1.1500	0.038	29.945	0.000	1.075	1.225
Lead Origin_Lead Add Form	2.8862	0.219	13.202	0.000	2.458	3.315
Lead Source_Direct Traffic	-1.5038	0.111	-13.564	0.000	-1.721	-1.286
Lead Source_Google	-1.1155	0.105	-10.640	0.000	-1.321	-0.910
Lead Source_Organic Search	-1.3057	0.127	-10.303	0.000	-1.554	-1.057
Lead Source_Referral Sites	-1.4104	0.311	-4.537	0.000	-2.020	-0.801
Lead Source_Welingak Website	1.5748	0.756	2.084	0.037	0.094	3.056
Last Activity_Email Bounced	-2.1525	0.367	-5.872	0.000	-2.871	-1.434
Last Activity_Olark Chat Conversation	-1.5140	0.157	-9.671	0.000	-1.821	-1.207
Last Activity_SMS Sent	1.2446	0.070	17.772	0.000	1.107	1.382
Last Notable Activity_Had a Phone Conversation	3.5516	1.092	3.252	0.001	1.411	5.692
Last Notable Activity_Unreachable	1.8684	0.457	4.091	0.000	0.973	2.763

	Features	VIF
9	Last Activity_SMS Sent	1.48
1	Lead Origin_Lead Add Form	1.44
6	Lead Source_Welingak Website	1.30
3	Lead Source_Google	1.23
2	Lead Source_Direct Traffic	1.21
0	Total Time Spent on Website	1.17
4	Lead Source_Organic Search	1.11
8	Last Activity_Olark Chat Conversation	1.08
7	Last Activity_Email Bounced	1.07
11	Last Notable Activity_Unreachable	1.01
5	Lead Source_Referral Sites	1.00
10	Last Notable Activity_Had a Phone Conversation	1.00

Metrics

Since we built the final model, Now lets check the Metrics on the Train set

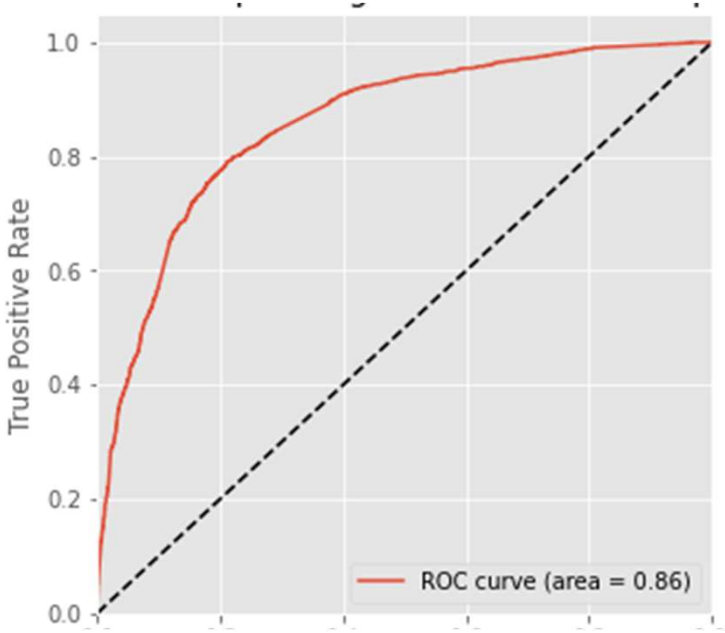
- Got the predicted values on the trainset, created a Data frame for the Hot Leads & the converted Probabilities.
- Created a new column 'Predicted' with 1 if Converted_prob > 0.5
- Calculated the metrics using Confusion Matrix
- Overall Accuracy - 79.48 (True Predictions/ Total Predictions)
- Sensitivity – 67.53 (True Positives/ True Positive + False negative)
- Specifiity – 86.96 (True Negative / True Negatives + False Positives

```
[[3396  509]
 [ 794 1652]]
```

	Converted	Converted_prob	Prospect ID	Predicted
0	0	0.151548	3009	0
1	0	0.013232	1012	0
2	0	0.546908	9226	1
3	1	0.831500	4750	1
4	1	0.883577	7987	1

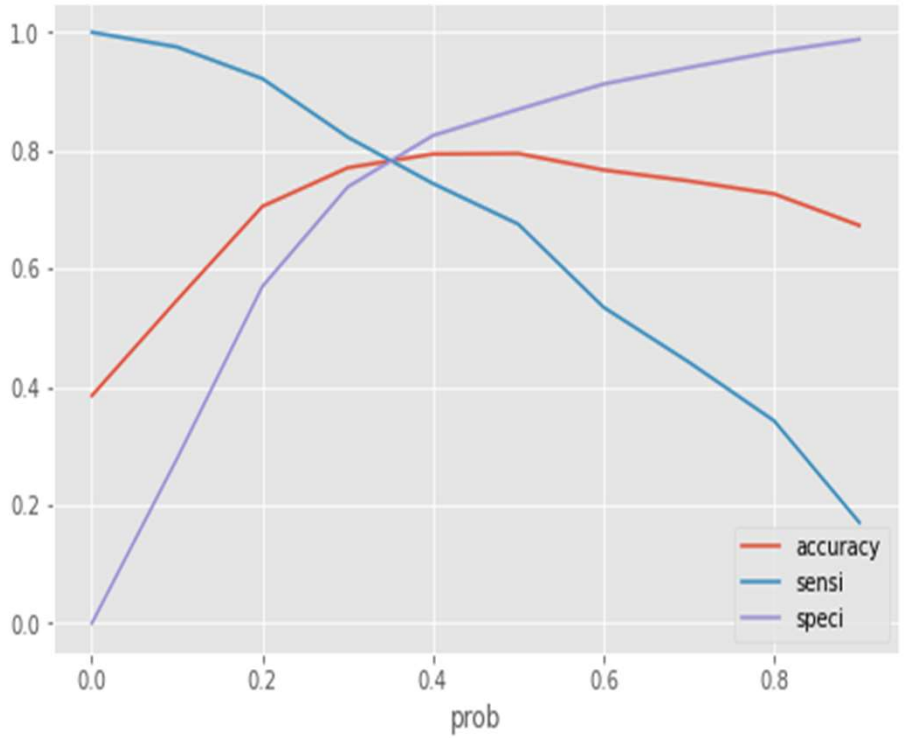
ROC

- Plotted ROC - Tradeoff between sensitivity and specificity (any increase in sensitivity will be accompanied by a decrease in specificity)
- Its best metric to determine if the new predicted value is Hot lead or a Cold Lead.
- ROC Curve should be closer to 1.
- ROC is from our model is 0.86 indicating a good predictive model
- It shows the tradeoff between sensitivity and specificity (any increase in sensitivity will be accompanied by a decrease in specificity).
- The curve follows the left-hand border and then the top border of the ROC space and gives the best & more accurate test results.



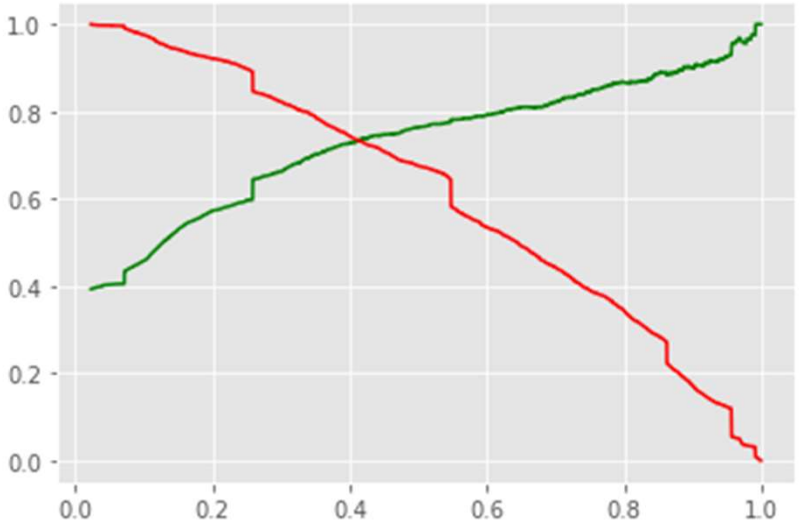
Optimal Cutoff

- Optimal cutoff probability is that prob where we get balanced sensitivity and specificity
- Created columns with different probability cutoffs, calculated accuracy sensitivity and specificity for various probability cutoffs.
- Plotted the the curve and found 0.35 is the optimum point to take it as a cutoff probability.
- Created new a new column "final_Predicted" & 'Lead Score' to understand the lead to Hot Lead conversion
- Calculated the final metrics on the Train data.
  - Accuracy : 78.87%
  - Sensitivity : 78.66%
  - Specificity : 79.00%



Precision - Recall Trade off

- Few more important Metrics like Precision & Recall are also calculated on Train set,
- False Positive Rate - 20.98%
- Positive Predictive value or Precision - 70.11
- Negative Predictive value – 85.52.
- Recall – 78.65%
- Precision Recall Trade off is plotted to get the accurate conversion for the Test set.
- Precision – Recall Trade off is – 0.41
- Hence, we make Predictions on the test set using this Trade off Value.



Predictions on the Test set

- On Test set, we make Predictions using the Precision – Recall trade off value 0.41
- First, transform the continuous columns on X_test data by Standard Scaler , add constant , run the model & make predictions on the y_test using stats model
- Convert the resultant data frame to have the Prospect ID, Converted, Converted _Prob ,Lead_Score & final_Predicted.
- Mapped the final_Predicted with 1 if probability is >0.41 else 0 .
- Calculated the final metrics on the test data,
  - Accuracy : 79.50%
  - Sensitivity : 72.19%
  - Specificity : 83.67%

	Prospect ID	Converted	Converted_prob	Lead_Score	final_Predicted
0	3271	0	0.136976	14	0
1	1490	1	0.653011	65	1
2	7936	0	0.117557	12	0
3	4216	1	0.861736	86	1
4	3830	0	0.117526	12	0

Our model is seems to be the best model and predicts the Conversion Rate very well and helps in Hot Leads Conversion

Final Conclusion

On observing & comparing the Train & test sets, all the metrics seem to be perfect & adhering the companies CEO's ball park conversion rate of 80%.

Train Data:

- Accuracy : 78.87%
- Sensitivity : 78.66%
- Specificity : 79.00%

Test Data:

- Accuracy : 79.50%
- Sensitivity : 72.19%
- Specificity : 83.67%

We can conclude that the Logistic Model is apt of the Education company to make accurate predictions & Conversions.

Recommendations

- ✓ Our Logistic Model is the best model for X Education Company adhering to the 80% ballpark Conversion Rate.
- ✓ This Model provides the X Education CEO confidence in making good calls based on this model.
- ✓ Our Model make accurate predictions & improving the business as per its timelines & future predictions

Top significant features are which helps in converting leads to Hot Leads are,

- Last Notable Activity_Had a Phone Conversation
- Lead Origin_Lead Add Form
- Last Notable Activity_Unreachable
- Lead Source_Welingak Website
- Last Activity_SMS Sent
- Total Time Spent on Website

The features which should be improved to convert leads to HOT Leads are,

- Lead Source_Google
- Lead Source_Organic Search
- Lead Source_Referral Sites
- Lead Source_Direct Traffic
- Last Notable Activity_Olark Chat Conversation
- Last Activity_Email Bounced

Thank You