Social Media Marketing
Initial Pool of Leads
Lead Nurturing
Hot leads
£ € £ $
Converted leads

Lead Scoring Case Study using Logistic Regression

VISALAKSHI G

RAMANA KISHORE G

Problem Statement & Business Goal

Problem Statement

- X Education sells online courses to industry professionals. The company markets its courses on several websites and search engines like Google.
- X Education gets a lot of leads, but its lead conversion rate is very poor. The typical lead conversion rate at X education is around 30%. In order to increase lead conversion process more efficient, the company wishes to identify the potential leads, who can be categorized as ‘Hot Leads’.
- Once ‘Hot Leads’ are acquired, 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 Goal:

- X Education needs help in selecting the most promising leads, i.e. the leads that are most likely to convert into paying customers.
- The company wants to build a model to identify Hot Leads.
- Improve the lead conversion rate from 30% to 80%

<u>Approach</u>

- Exploratory Data Analysis (EDA)
  - Data Cleaning
  - Numerical Analysis
  - Categorical Analysis
- Data Preparation
- Model Building
- Metrics Validation
- Conclusions & Recommendations

EDA - Data Cleaning

Leads.csv (df)

- 37 columns X 9240 rows
- Data set has lot of categorical features.
- Data Imbalances for categorical features analyzed and dropped features with high imbalance
- Data set has conversion ratio of 38%
- 17 columns have missing values, Dropped columns Missing values greater than 40%.
- Imputing and standardization - Replaced **'Select'** with '**NaN'** based on data analysis as this value is result of not entering values as part of data collection
- 'Prospect ID' , 'Lead Number' are unique identifiers which are not useful in analysis, hence dropped from data set

EDA – Categorical Analysis

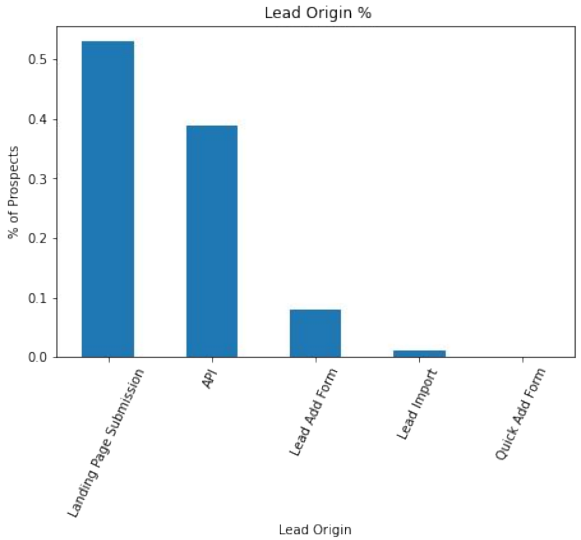
Approach used for Categorical columns:

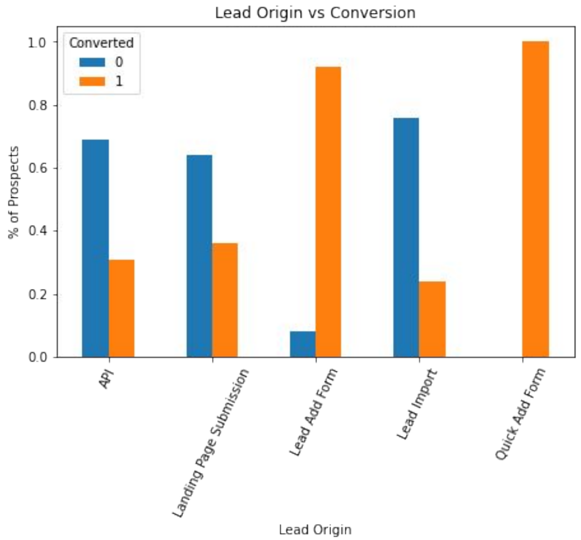
Merging of insignificant values of column(if any) into ‘Others’ category

Univariate Analysis & Bivariate Analysis using Bar Plots

Make Inferences based on analysis

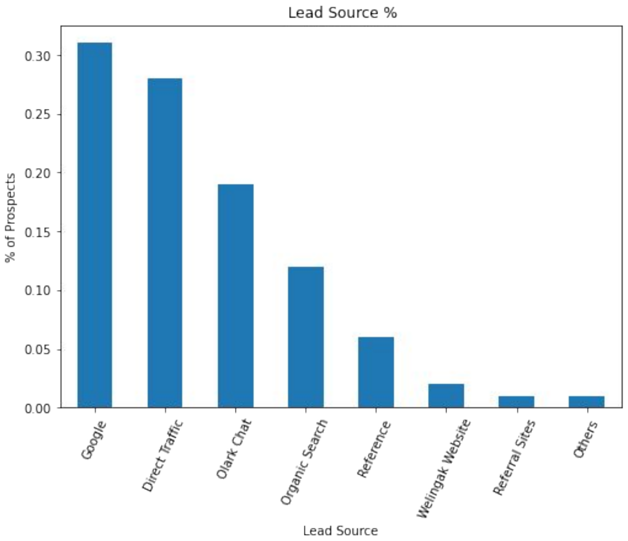
- Lead Origin (Details shown in right hand side plot)
- Lead Source(Details shown in the plot below)
- Last Activity
- Specialization
- Tags
- City
- A free copy of Mastering The Interview
- Last Notable Activity

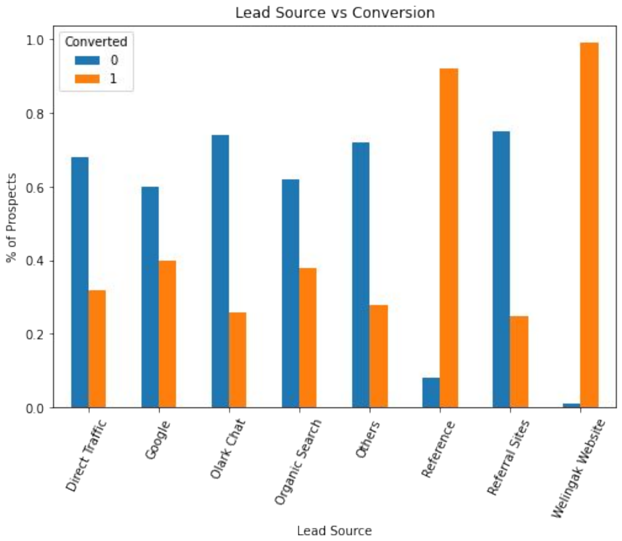




Lead Origin – Inferences

- **Univariate Analysis: ~53% of the Lead Origin is from 'Landing Page Submission' followed by ~39% from API.**
- **Bivariate Analysis: 'Landing Page Submission' has 36% of Conversion and 'API' has 31% of Conversion**





Lead Source – Inferences

- **Univariate Analysis: ~31% of the Lead Source is from ‘Google' followed by ~28% from 'Direct Traffic’**
- **Bivariate Analysis : 'Google' as a Lead Source has 40% of Conversion and 'Direct Traffic' has 32% of Conversion**

EDA - Numerical Analysis

- Numerical variable analysis using correlation matrix(Fig 1)
  Inference - Variables are not significantly correlated.
- Outlier analysis (Box plotting shown in Fig 2 below
  Inferences:
  - ‘Total Time spent on Website’ does not have outliers and not treated
  - ‘Total Visits’ and ‘Page Views Per Visit’ have outliers.
- Approach used - Capping with IQR*1.5
  Capping is done to retain data instead of dropping.
  99% of records are retained for modeling.
- Post Outlier treatment plot is shown for ‘Pages Views Per Visit’ is shown in Fig 3.

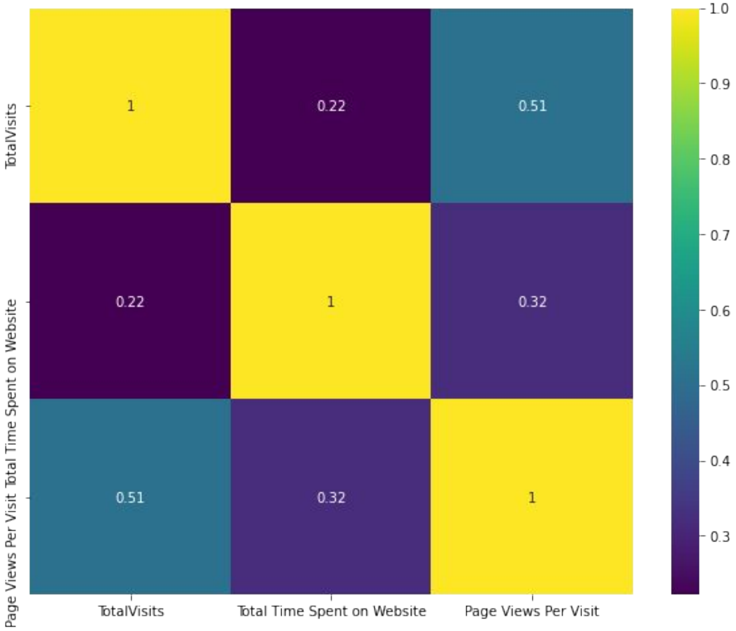


Fig 1

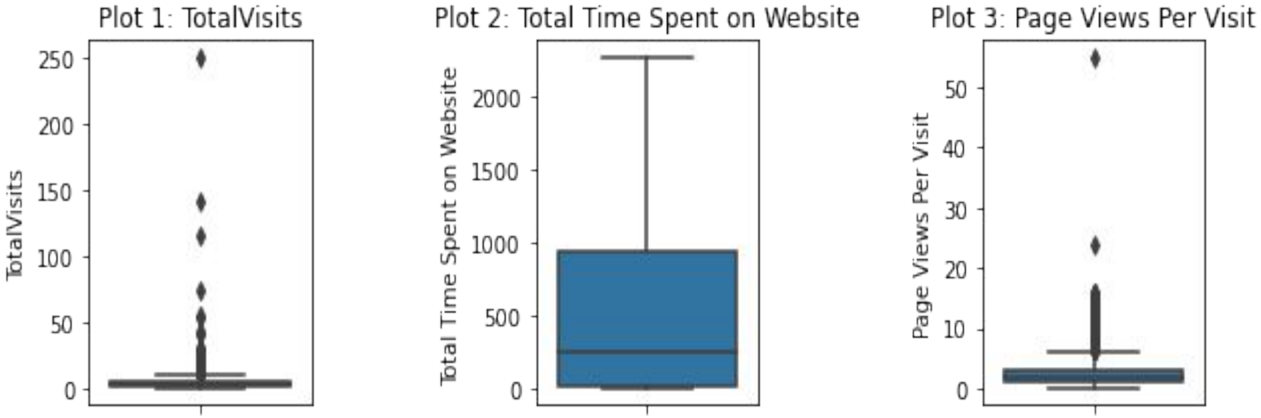


Fig 2

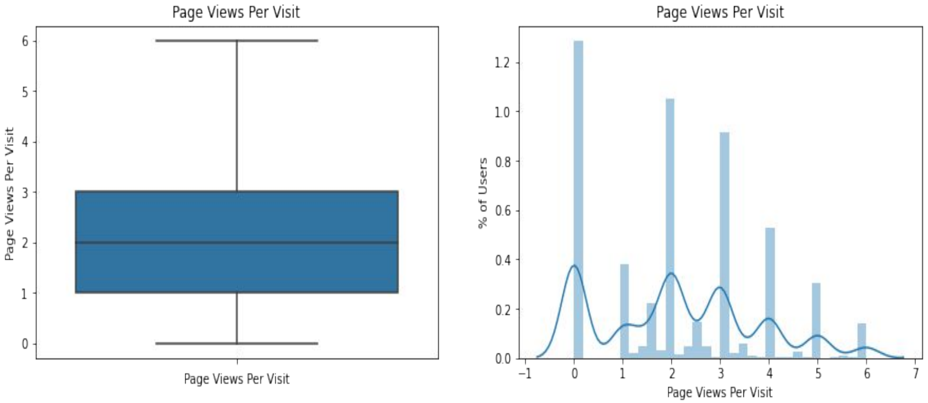


Fig 3

Data Preparation

Dummy Variable Creation

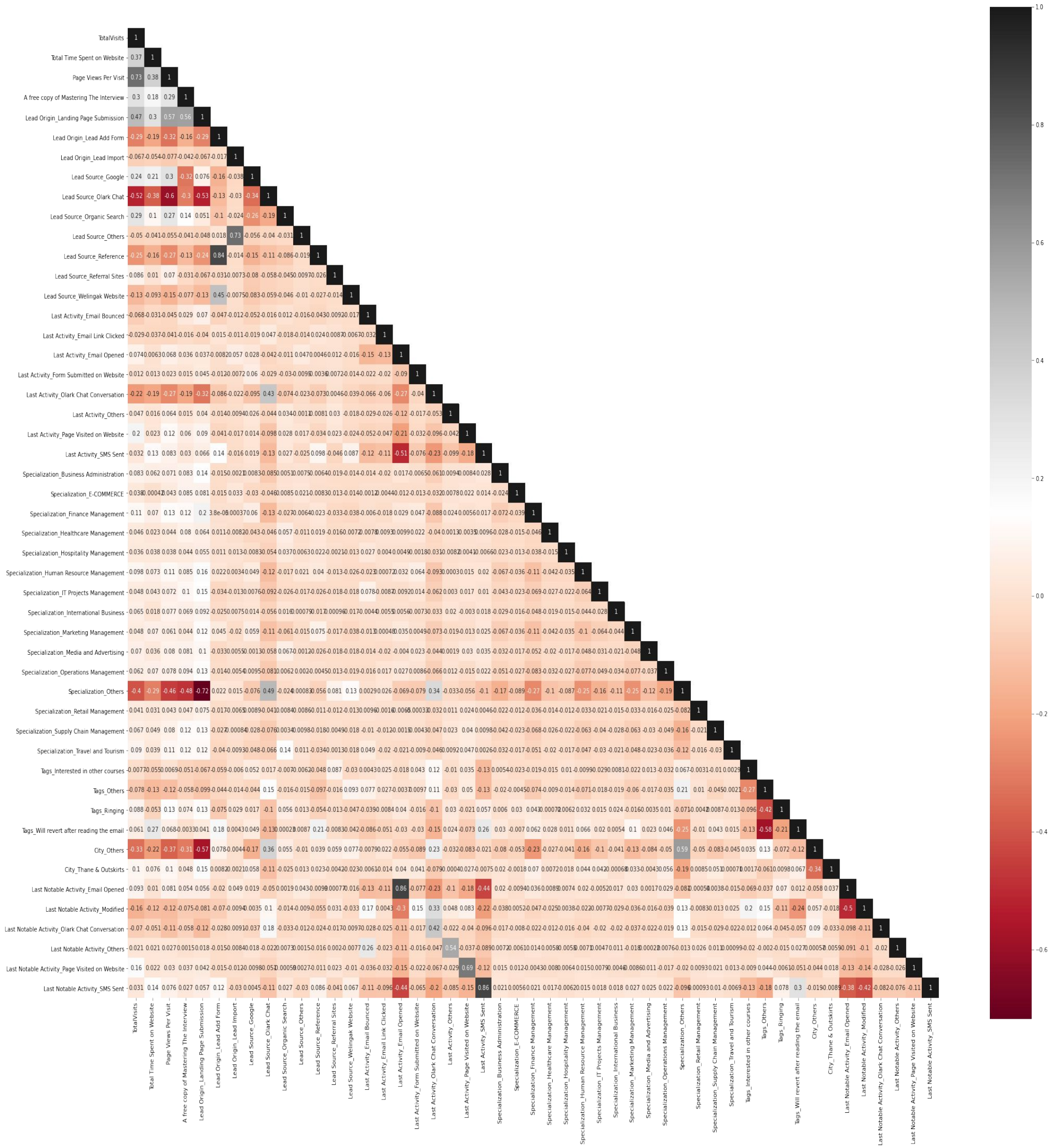
- Post EDA,Categorical features in data set:
  ['Lead Origin', 'Lead Source', 'Last Activity', 'Specialization’, ‘Tags', 'City', 'A free copy of Mastering The Interview’, 'Last Notable Activity’]
- 'A free copy of Mastering The Interview’ feature has binary {'Yes': 1, "No": 0}
- Categorical features processed for dummy variables conversion.
- Correlation matrix of dummy variables(shown in the plot)
  Inference - There are features with significant correlation.
  Treatment - Dropped features with correlation above 0.4

Train Test Split

- Data set is split into X_train, X_test, y_train and y_test model with test size of 20%

Scaling

- Scaling of features 'TotalVisits', 'Total Time Spent on Website' using StandardScaler() for training data (to bring all features to same scale)



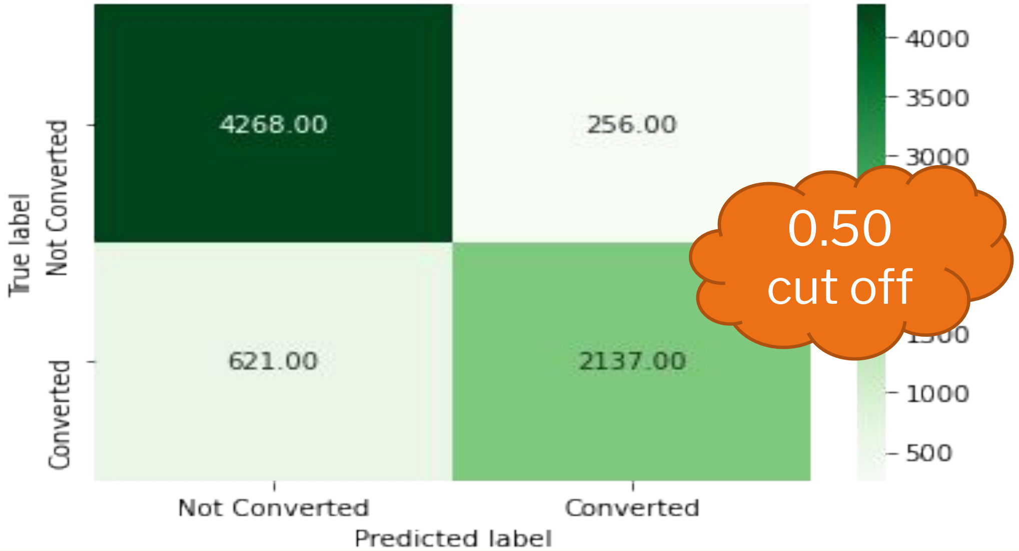
Model Building

Modeling Steps

- **Model 1** - Logistic regression model built using all the features.
  Observations - Multiple features having high p value which indicate that model is not statistically significant &
  strong evidence for null hypothesis.
- **Model 2** - Using RFE (Recursive Feature Elimination) - 15 features are selected for model building. ‘Lead Origin_Lead Import’ has insignificant p value.
- Iteratively ‘Lead Origin_Lead Import’,
  ‘Lead Source_Referral Sites’,
  ‘Specialization_International Business’
  and ‘Specialization_Healthcare Management' features are removed
  to create **Model 3**, **Model 4**, **Model 5** and **Model 6**
  while analyzing p-values and VIF (Variance Inflation Factor) scores
- **Model 6** is resulting with acceptable p-values (<0.05) and VIF scores less than 2. This model is considered for evaluation and metrics analysis.

Prediction and Metrics

- Converted probability calculated using Model 6 for training data set.
- **Predicted** values for training data set cutoff of 0.5
- Metrics are as below
  - Accuracy = 87.95 %
  - Sensitivity = 77.48 %
  - Specificity = 94.34 %
  - True Positive Rate = 89.30 %
  - False Positive Rate = 87.30 %

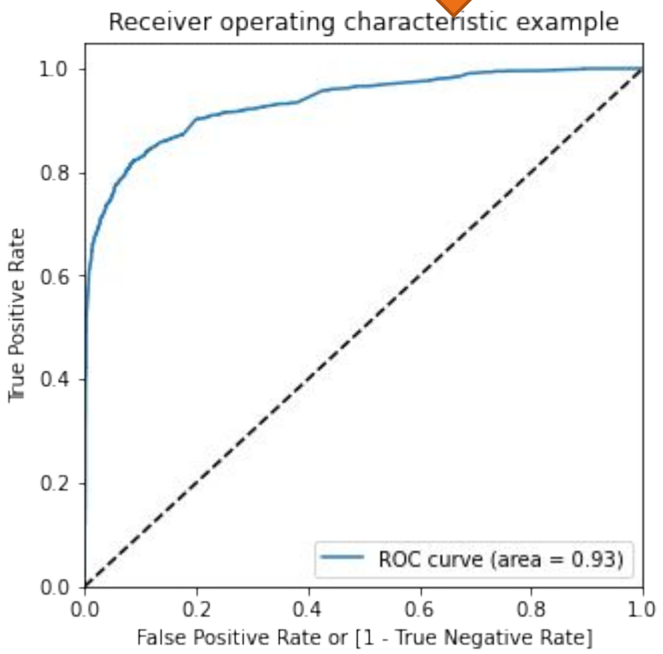


0.50
cut off

Metrics Validation - ROC Curve & Optimal Cut Off

ROC Curve

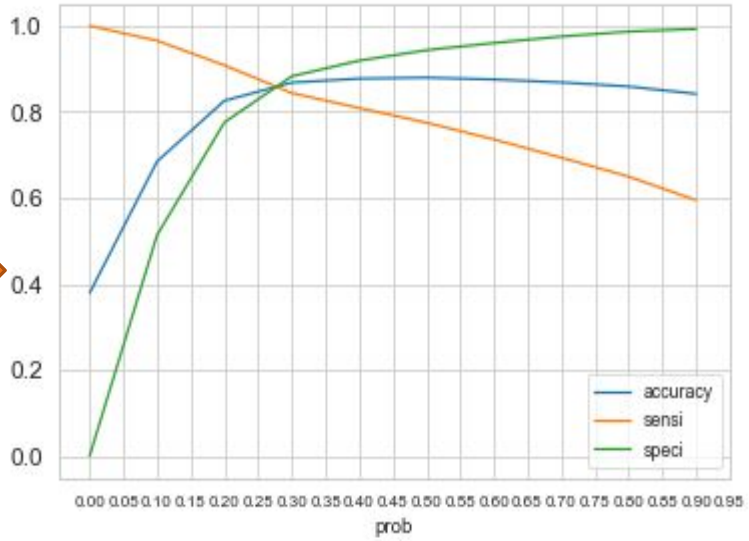
- Closer the curve follows the left and the top border of ROC space, the more accurate the test
- Closer the curve comes to 45° diagonal of the ROC space, the less accurate the test
- The ROC Curve should be a value close to 1. Good value of 0.93 indicates a good predictive model

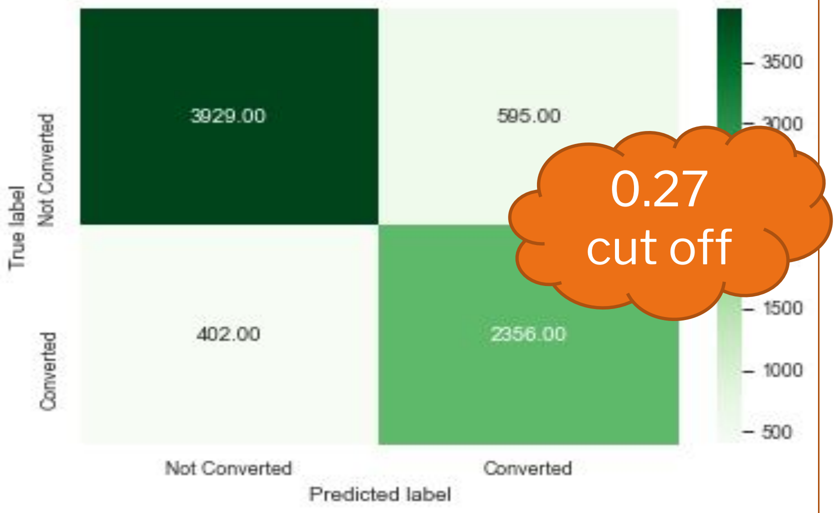


Optimal Cut Off Point

- Plotting accuracy, sensitivity and specificity
- Intersecting point is ~0.27, so cut off probability is 0.27 results in following metric values
  - Accuracy = 86.31 %
  - Sensitivity = 85.42 %
  - Specificity = 86.84 %
  - Precision = 79.84 %
  - Recall = 90.71 %

Confusion Matrix is depicted with 0.27 cutoff probability



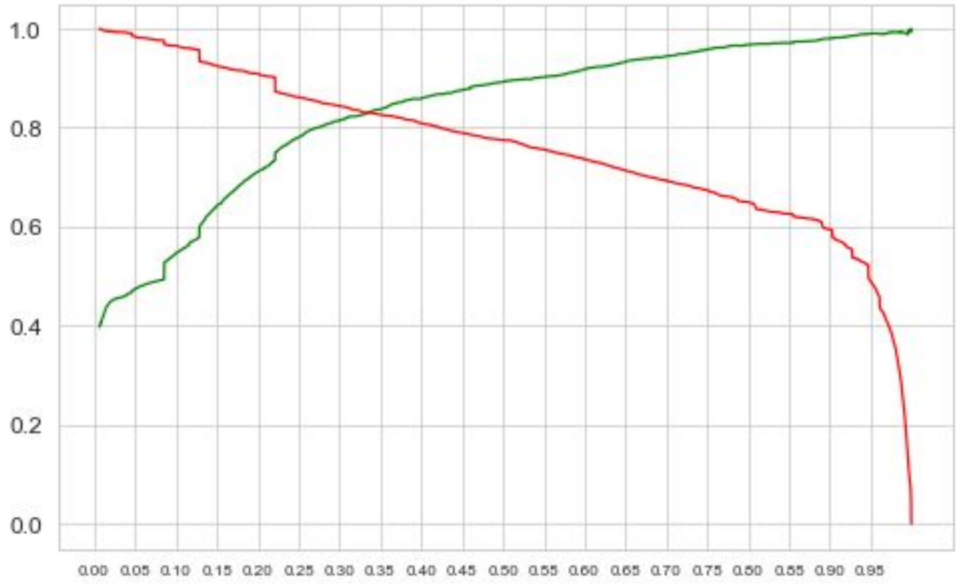


0.27 cut off

Precision Recall Curve & Prediction on Test Set

Precision Recall Curve

- Trade off between Precision and Recall is clearly visible
- Cut off is 0.34 based on precision recall curve
- 0.34 is used as threshold for test data

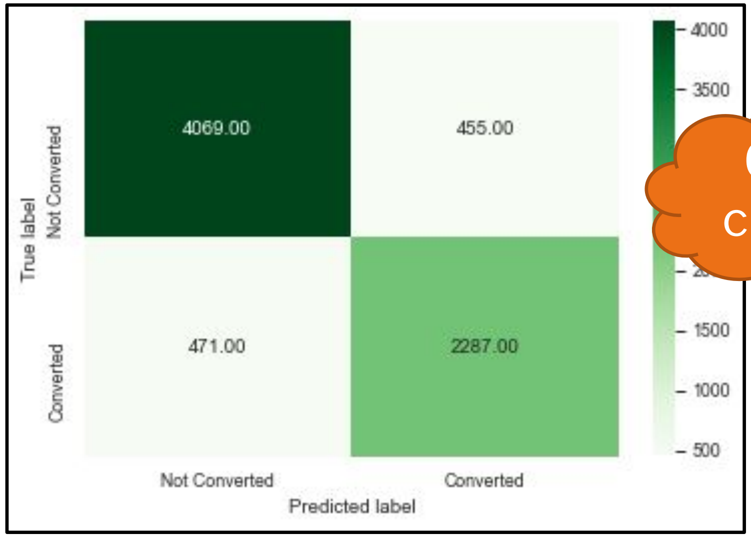


Predictions on Test Set

- After scaling Test set and running, here are metrics
  - Accuracy = 87.80 %
  - Sensitivity = 83.64 %
  - Specificity = 90.43 %
  - Precision = 84.60 %
  - Recall = 83.64 %

~84% of Recall value means that model is able to predict 84% of actual conversion cases correctly.

~85% of Precision value indicates that 85% of the conversions that our model predicted are actually converted.



0.34 cut off

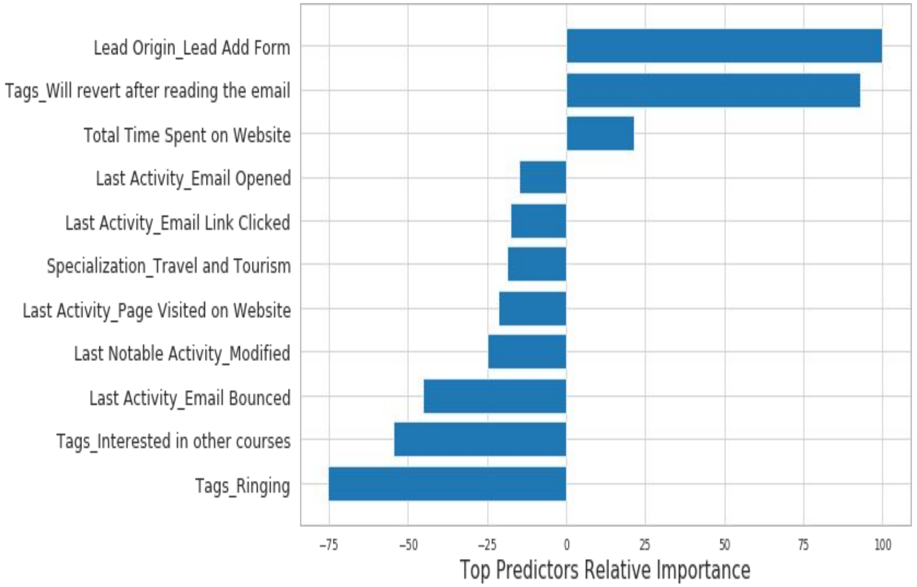
Conclusions & Recommendations

Model Conclusions:

- ~84% of Recall value means that model is able to predict 84% of actual conversion cases correctly.
- ~85% of Precision value indicates that 85% of the conversions that our model predicted are actually converted.

Top 3 Predictor variables:

- Lead Origin_Lead Add Form
- Tags_Will revert after reading the email
- Total time Spent on Website



Recommendations:

X Education Company needs to focus on following key aspects to improve the overall conversion rate

- ❖ Increase user engagement on their website since this helps in higher conversion
- ❖ Increase on sending email notifications since this helps in higher conversion
- ❖ Get Total Visits increased by advertising etc., as this helps in higher conversion
- ❖ Improve the effectiveness of phone service since this is affecting the conversion negatively.

Thank you

“It’s one small step for man, one giant leap for mankind.”
- Neil Armstrong