Lead Scoring Case Study Assignment

By:

Bharti

Problem Statement

- To help X education to select the most promising leads known as ‘hot leads’ who are most likely to convert into paid customers.
- Build a logistic regression model to assign a lead score between 0 and 100 to each of the leads where the leads with higher lead score have a higher conversion chance and the leads with lower lead score have a lower conversion chance.
- Identify the driver variables and understand their significance which are strong indicators of lead conversion.
- Identify the outliers, if any, in the dataset and justify the same.
- Consider both technical and business aspects while building the model.
- Summarize the conversion predictions by using evaluation metrics like accuracy, sensitivity, specificity and precision.

Data Exploration

- '**Leads.csv**' contains all the information about the leads generated through various sources and their activities.
  - This file contains 9240 rows and 37 columns.
  - Out of 37 columns, 7 are numeric columns and 30 are non-numeric or categorical columns.
  - Current conversion rate of the leads is 39%.

- **'Leads Data Dictionary.csv'** is data dictionary which describes the meaning of the variables present in the "Leads" dataset.

Data Cleaning and Preparation

Leads.csv :

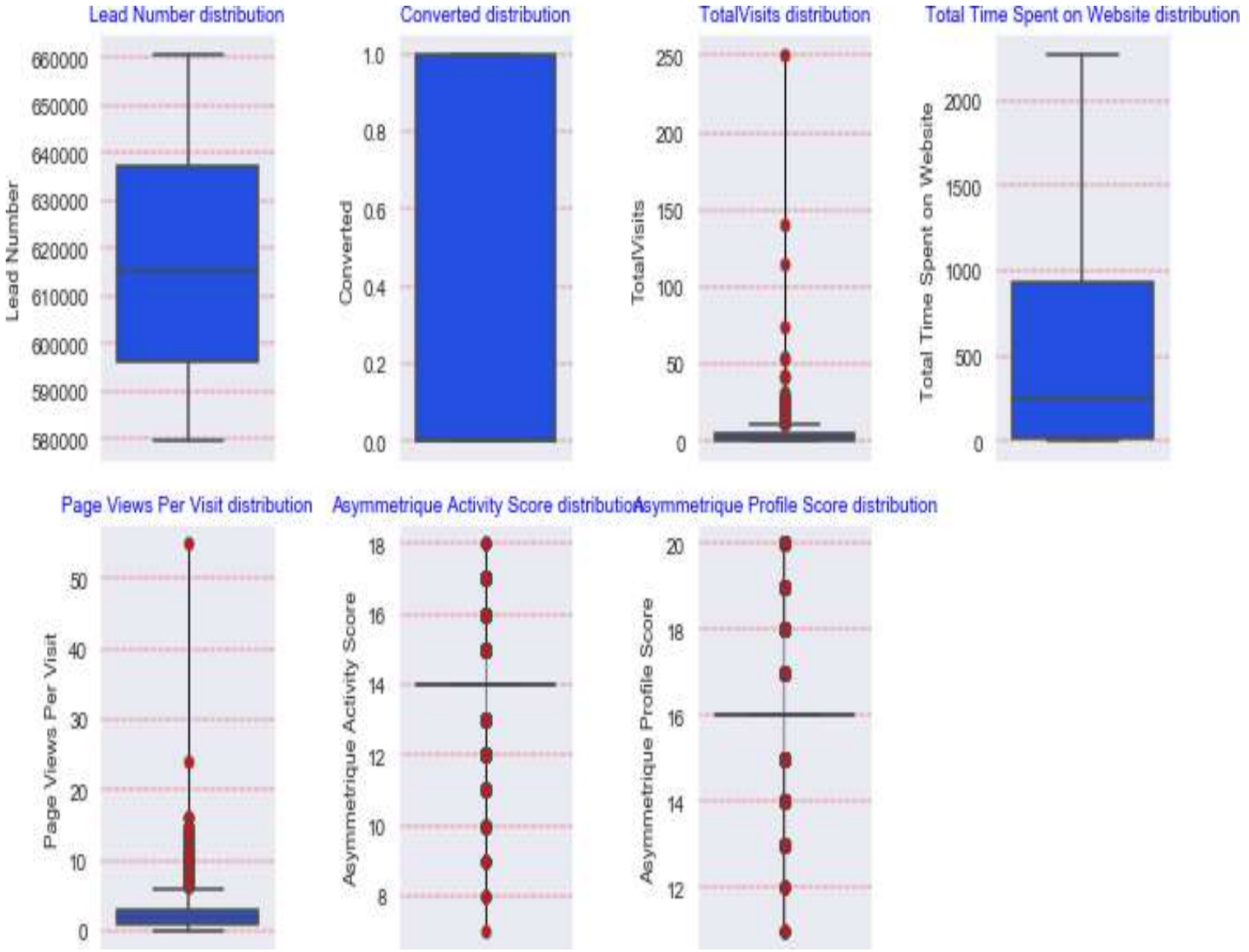
- Following columns contain more than 30% null values initially:
    1. What is your current occupation
    2. What matters most to you in choosing a course
    3. Tags
    4. Lead Quality
    5. Lead Profile
    6. Asymmetrique Activity Index
    7. Asymmetrique Profile Index
    8. Asymmetrique Activity Score
    9. Asymmetrique Profile Score
- Following columns have default value of ‘select’ as a dominating value which is same as null value. So, we have converted ‘select’ to ‘NA’.
    1. Specialization
    2. How did you hear about X Education
    3. Lead Profile
    4. City
- All the missing values of categorical columns have been imputed with ‘NA’.

Data Cleaning and Preparation

- All the missing values of quantitative columns have been imputed with median as the difference between mean and median is insignificant.
- Following columns have been dropped which contain single value as their contribution is insignificant:
    1. Magazine
    2. Receive More Updates About Our Courses
    3. Update me on Supply Chain Content
    4. Get updates on DM Content
    5. I agree to pay the amount through cheque
- Following columns have been dropped since percentage of missing value is more than 70%:
    1. How did you hear about X Education
    2. Lead Profile
- Following columns have been imputed with mode since the percentage of missing value is low.
    1. Lead Source
    2. Lead activity

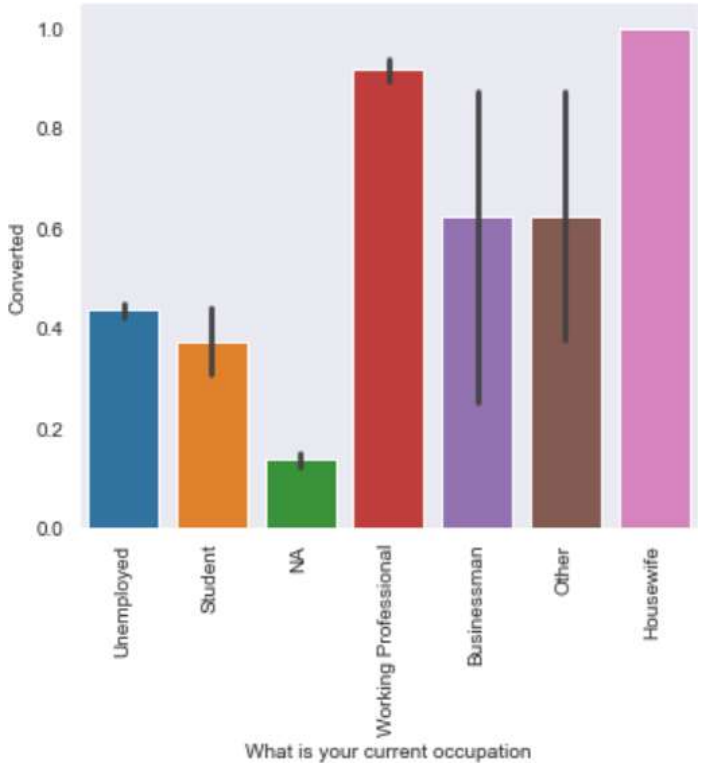
Univariate Analysis – Outliers

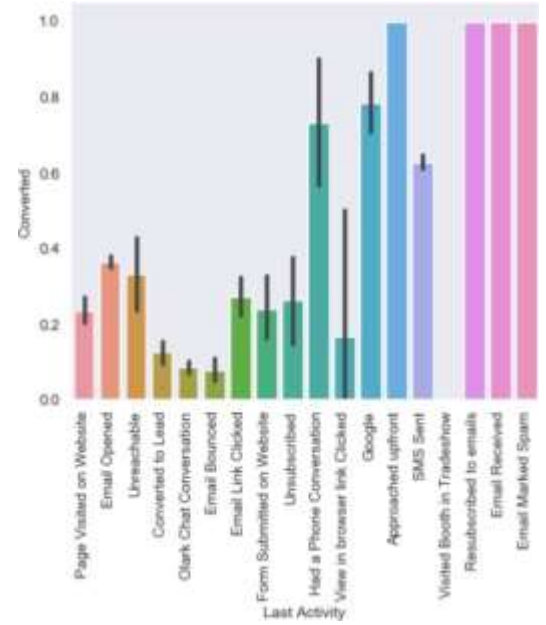
- Univariate analysis revealed data distribution and outliers in 'Leads' data. Key columns where outliers were identified are:-
  a. TotalVisits
  b. Page Views Per Visit
  c. Asymmetrique Activity Score
  d. Asymmetrique Profile Score
- Inter Quantile Range (IQR) method has been used to treat outliers in the data.
- Decision has been taken to not remove any outliers since the % is high (9%).
- We will review the final model to ensure this does not impact the score.

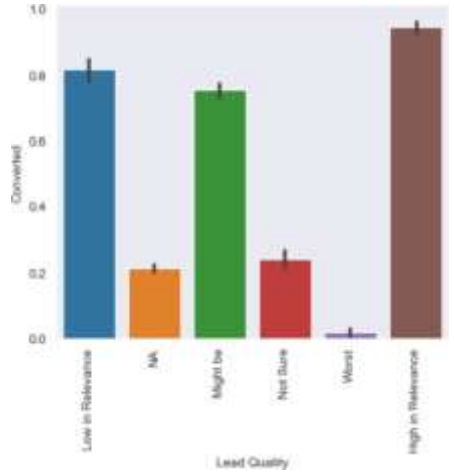


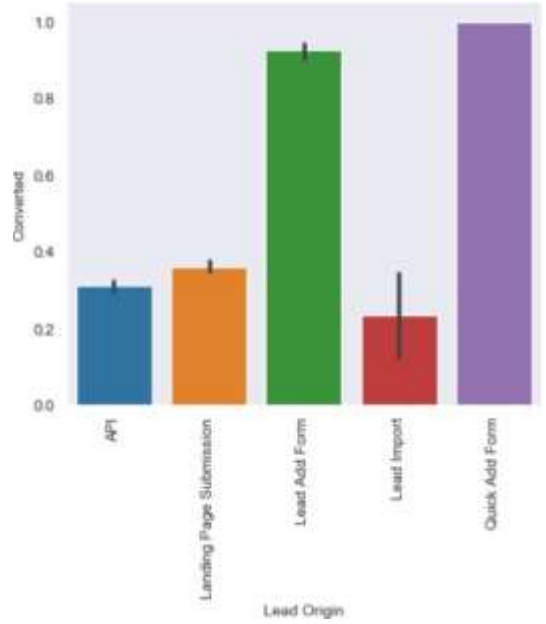
Bivariate Analysis: Categorical variables

- ‘Converted’ column has been chosen as target variable. So, bivariate analysis of important variables has been performed with respect to the target variable.
- Lateral students and the visitors showing interest on next batch have higher chances of getting converted.
- Lead quality tagged with “High in Relevance” has high conversion rate history.
- Lead originated through “Lead Add Form” and “Quick Add Form” has high possibility of getting converted.
- Lead belongs to Welingak Website, WeLearn, Live Chat and NC_EDM converts more than any other sources.



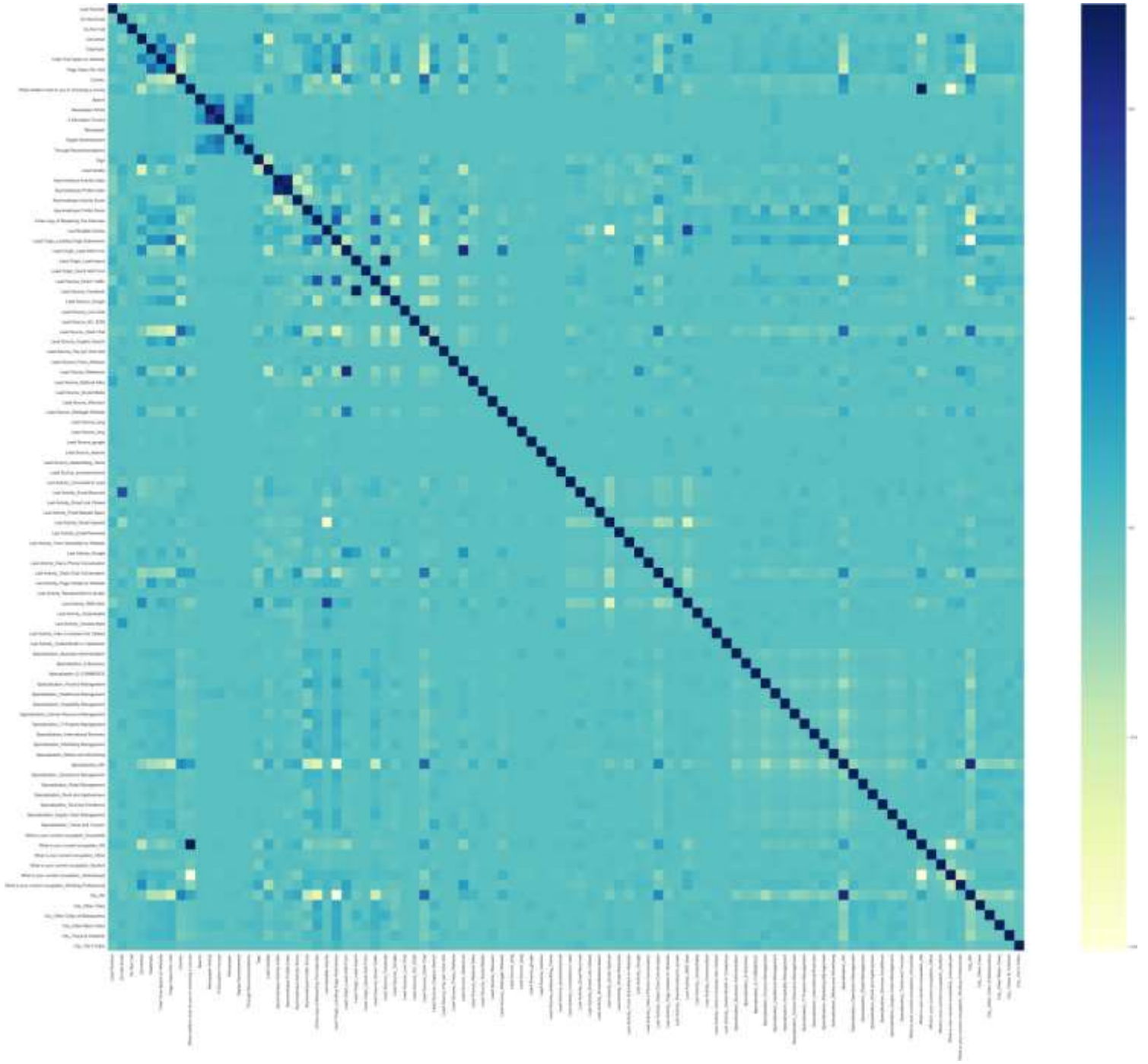






Bivariate Analysis: Checking correlation

- Following group of columns are positively highly correlated with each other:
  1. Search
  2. Newspaper Article
  3. X Education
  4. Digital Advertisement
  5. Through Recomendations
- Another set of columns are also positively highly correlated with each other:
  1. TotalVisits
  2. Total Time Spent on Website
  3. Page Views Per Visit
- There is a strong positive correlation between Asymmetrique Activity Index and Asymmetrique Profile Index.



Data Preparation for Modeling

Create Dummy Variables:

- Independent variables as dummy variables allows easy interpretation and calculation of the odds ratios, which increases the stability and significance of the coefficients.
- Dummy variables have been created for following columns:
    1. Lead Origin
    2. Lead Source
    3. Last Activity
    4. Specialization
    5. What is your current occupation
    6. City

Label Encoding:

- Label encoding is simply converting each value in a column to a number.
- We will use label encoding for variables with higher level. This is to avoid drastic increase in dataframe size.
- All the relevant categorial variables have been encoded using ‘LabelEncoder’.

Data Preparation for Modeling

Binary Variables Encoding:

- Variables which have binary (Yes/No) values have been encoding with 1 and 0.
- 1 denotes Yes whereas 0 denotes No.

Train – Test Split:

- The modified ‘Leads’ dataset has been split into Train and test dataset in the ratio 70:30.
- Train dataset has been used to train the model whereas Test dataset has been used to evaluate the model

Feature Scaling:

- It is important to have all variables on the same scale in order to avoid the dominance of variables with high magnitude in the model.
- “StandardScaler” function has been used to scale the data for modeling which brings all the data points into a standard normal distribution with mean at ‘0’ and standard deviation at ‘1’.

Model Building: Using logistic Regression

- Generalised Linear Model (GLM) from StatsModels library has been used to build the Logistic Regression Model.
- Initially, the model was built using 93 features present in X_train dataset.
- Most of the features were found to be insignificant. So, we needed to perform feature selection technique.

Feature Selection using Recursive Feature Elimination (RFE):

- **RFE** is an optimization technique for finding the best performing subset of features. It is based on the idea of repeatedly constructing a model and choosing either the best (based on coefficients), setting the feature aside and then repeating the process with the rest of the features. This process is applied until all the features in the dataset are exhausted. Features are then ranked according to when they were eliminated.
- We ran RFE to identify top 20 features for further model building process.
- Insignificant features were dropped one by one after checking the P-value and Variance Inflation Factor (VIF).
- Accepted P-value should be kept below 0.05 and VIF should be less than 5.

Model Building: Using Logistic regression (On PCA data)

- Principal component analysis (PCA) is a statistical procedure that uses an orthogonal transformation to convert a set of observations of possibly correlated variables (entities each of which takes on various numerical values) into a set of values of linearly uncorrelated variables called principal components.
- Initially, PCA was performed on X_train data (excluding Prospect ID and Lead Number columns).
- Then, Incremental PCA was performed on the PCA dataset by taking first 10 principal components which
  are explaining more than 95% of the variance.
- Previous step was repeated on X_test data (excluding Prospect ID and Lead Number columns).
- After that Logistic Regression has been performed on PCA datasets.
- Although the results obtained from this model was good but, the results without using PCA were even better.
- Therefore, we have proceeded further for prediction and conclusion with the last logistic regression
  model which was built without using PCA.

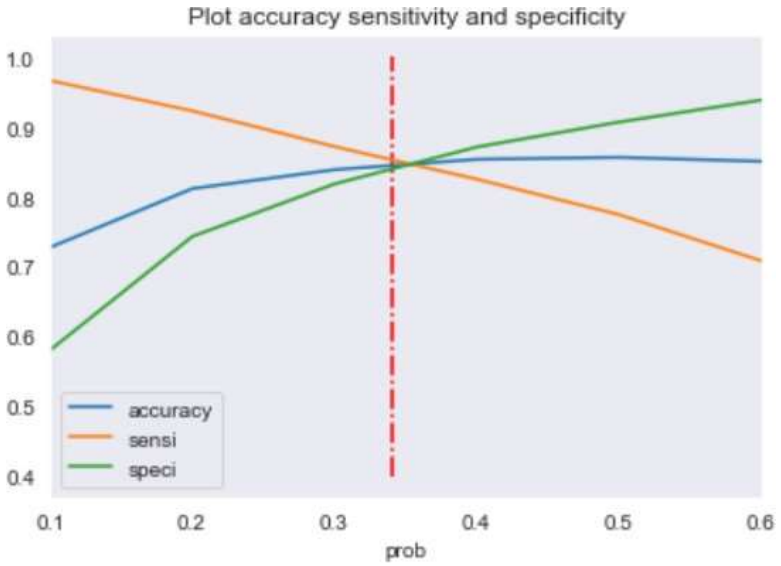
Final Model and Interpretation

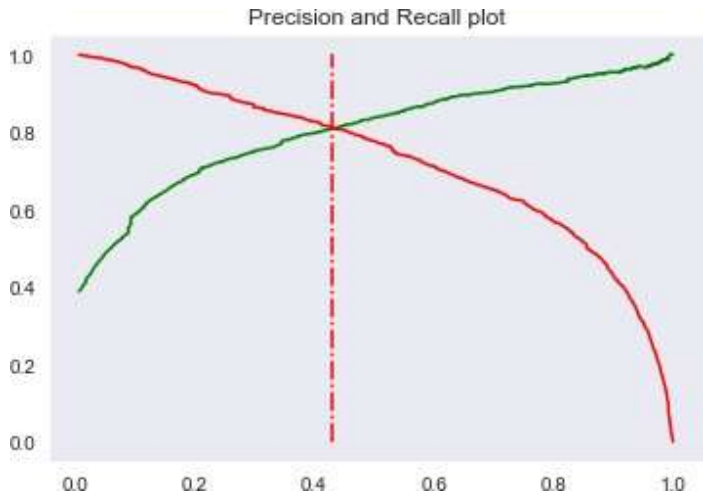
- Final model contains 14 most important features which satisfy all the selection criteria.
- Lead score having conversion probability greater than 0.43 are being predicted as “Converted”.
- Using this probability threshold value (0.43), the leads from the test dataset have been predicted whether they would get converted or not.
- Confusion matrix with cut-off 0.43 has been created to calculate evaluation metrics.
- Confusion matrix: `[[3528 474]`
  `[ 468 1998]]`

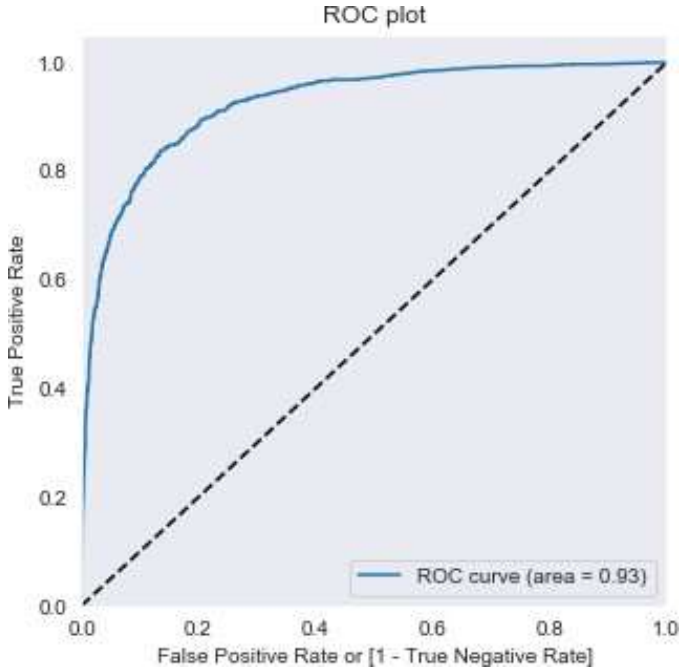
- Evaluation metrics:
  - Accuracy: 0.8544
  - Sensitivity: 0.8102
  - Specificity: 0.8816
  - Precision: 0.8083

Evaluation Metrics

- Receiver Operating Characteristics (ROC) Curve:
  - By determining the Area Under the Curve (AUC) of the ROC curve, the goodness of the model is determined.
  - Since the ROC curve is close to the upper left part of the graph, it means this model is a very good model.
  - The value of AUC for our model is 0.93.
- Plot accuracy sensitivity and specificity:
  - Tradeoff between sensitivity and accuracy can be observed (cutoff = 0.34).
- Precision and Recall plot:
  - Ideal cutoff of 0.43 is observed from recall and precision plot.
- We will use both the cutoff and evaluate results for further predictions.

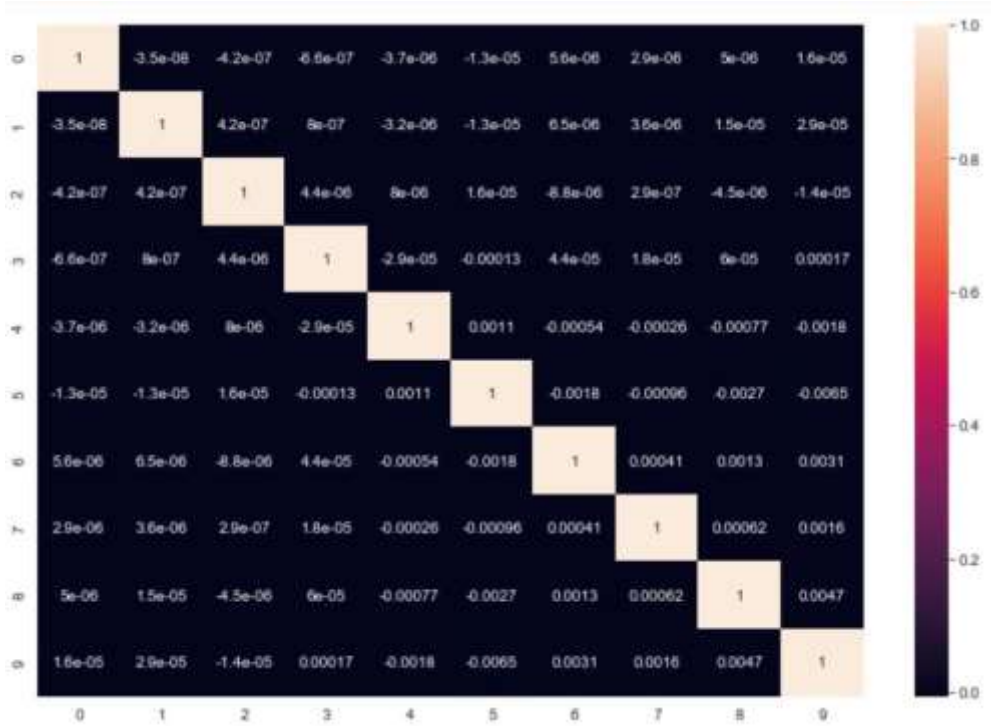


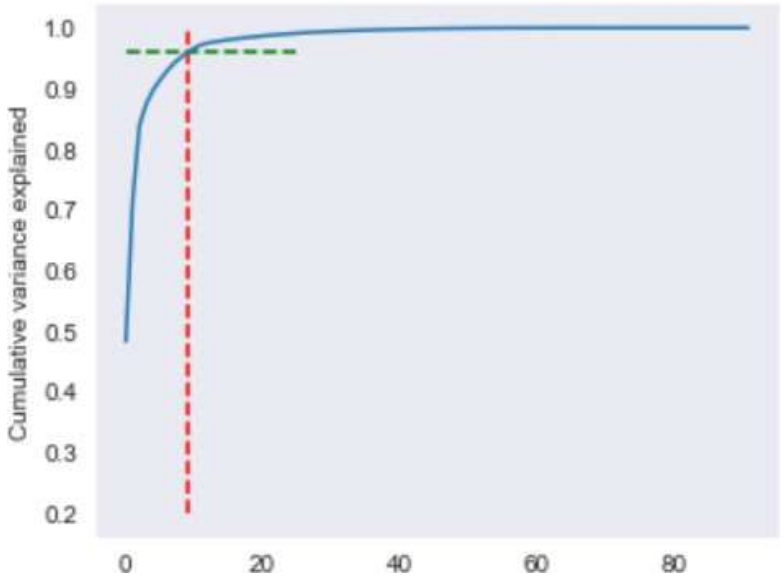




Evaluation using PCA

- Using PCA helps in dimensionality reduction and solves for multicollinearity issue.
- Making predictions using model build using PCA gives decent results but presents below challenge.
  - Score less that model build without using PCA.
  - Identify original variables/factors leading to high score.
- Metric derived from PCA
  - Accuracy: 0.8355
  - Sensitivity: 0.7635
  - Specificity: 0.8825
  - Precision: 0.8093
- Model without PCA yields better result.






```
PCA - Confusion metric
[[1480  197]
 [ 259  836]]
```

Conclusion and Recommendations:

- Followings are top three features that contribute to decision which mean the conversion probability of a lead increases with increase in values of these features:
  - Lead Origin
  - What is your current occupation
  - Last Activity
- Top three categories that contribute to decision
  - Lead Origin ==> Lead Add Form
  - What is your current occupation ==> Working Professional
  - Last Activity ==> SMS Sent

Conclusion and Recommendations: contd..

- This model will help to identify the hot leads which would enhance **speed-to-lead** and the **response rate**.
- Approaching only to hot lead would result in:
  - Shorter **sales cycle** through intuitive prioritization.
  - Better **opportunity-to-deal ratio**
  - Control over volatile **buying cycle**
  - Increase **marketing effectiveness**
  - Better **sales forecasting**
  - Minimize opportunities loss
  - Increase in revenue