Lead Scoring Assignment

Shehin Hanief

DS47 UPGRAD

Problem Statement

- An education company named X Education which sells online courses to industry professionals. Although X Education gets a lot of leads its lead conversion rate is very poor. To make this process more efficient the company wishes to identify the most potential leads also known as Hot Leads.

- Build a logistic regression model to assign a lead score between 0 and 100 to each of the leads which can be used by the company to target potential leads. 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.

- Summarize conversion predication using evaluation metrices like accuracy sensitivity specificity and precision

Data Exploration

- Leads.csv contains all the information regarding the leads generated across various sources
  - It has 9240 entries and 37columns
  - Out of 37 columns 30 are categorical columns and remaining 7are numerical columns.
  - Current conversion rate is 39%
- Leads Date Dictionary.csv is the data dictionary which describes the meaning of the variable in Leads.csv dataset.

Data Cleaning and Preparation

- Leads.csv:
- Columns with more than 30% null Values:
  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
- Columns with Select as default Value:
  1. Specialization
  2. How did you hear about X Education
  3. Lead profile
  4. City
- All missing values of categorical columns have been imputed with NA

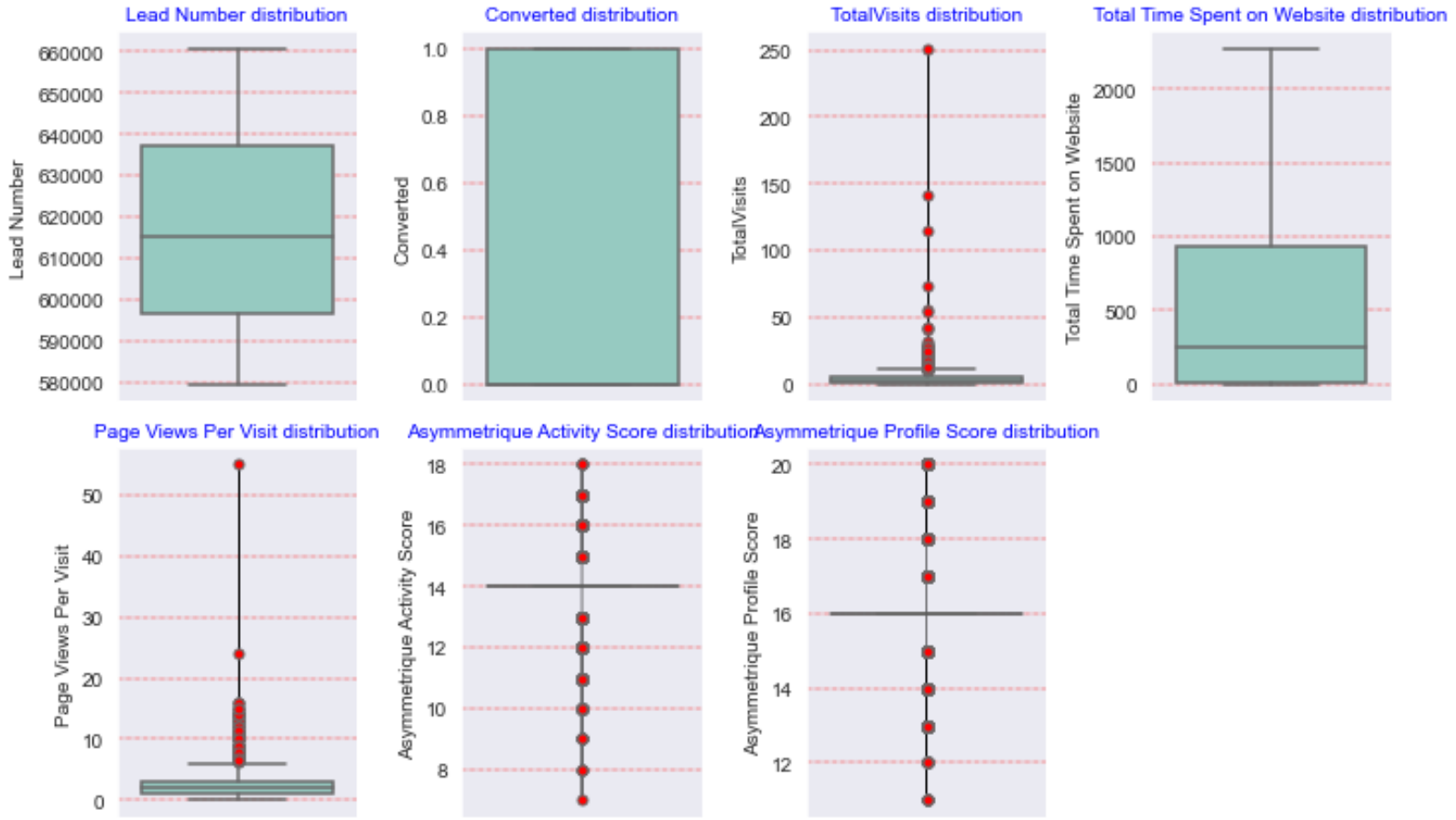
Data Cleaning and Preparation

- All missing values of quantitative columns have been imputed with median
- Following columns have been dropped which contain single values
  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 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 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 :
  1. TotalVisits
  2. Page Views Per Visit
  3. Asymmetrique Activity Score
  4. Asymmetrique Profile Score
- Inter Quantile Range (IQR method has been used to treat outliers in the data.
- Decision has been taken to not to remove any outliers as its percentage is high(9%)

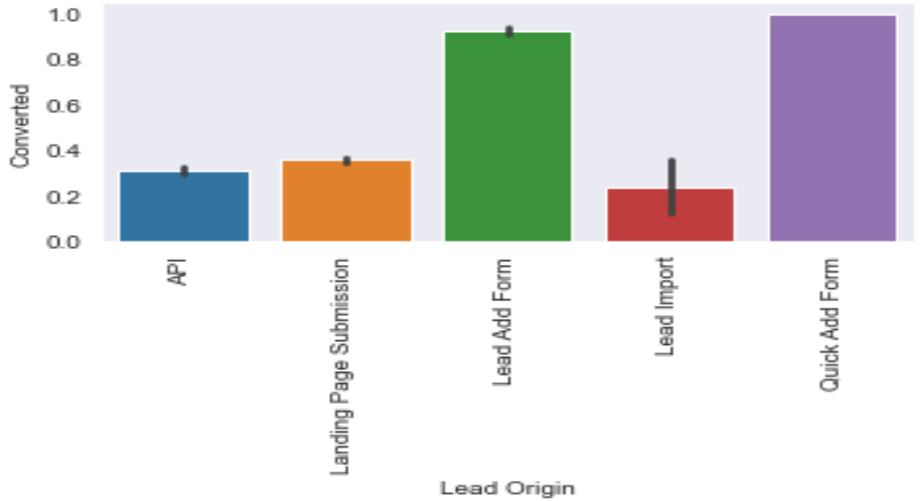
Univariate Analysis - Outliers

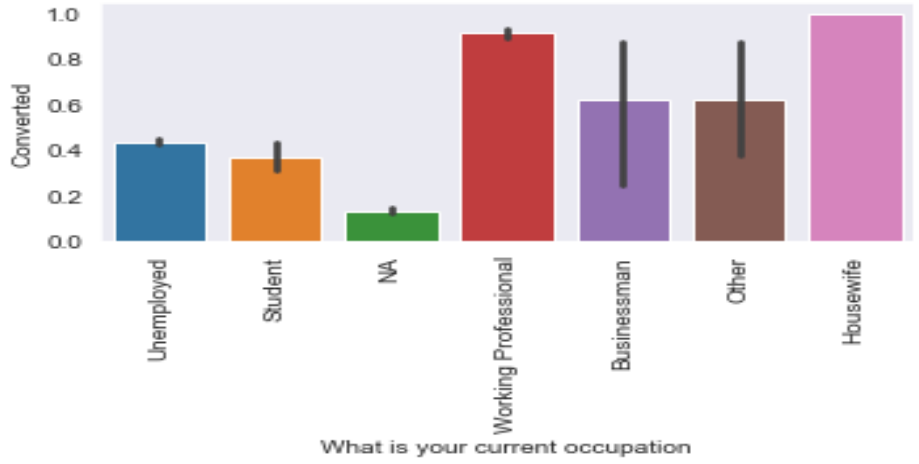


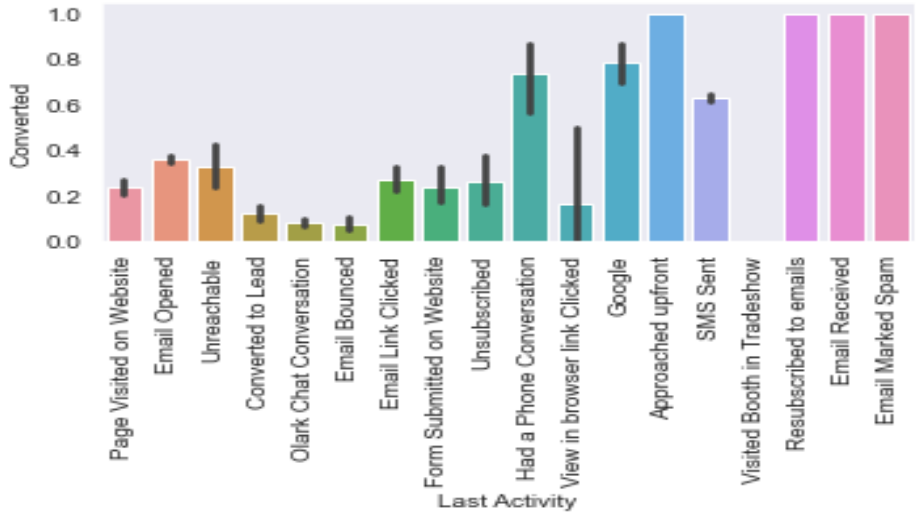
Bivariate Analysis – Categorical variables

- Bivariate analysis of important variable has been performed with that of Target variable ‘Converted’.
- Lateral students and visitors showing interest on next batch has high conversion chances
- Lead quality tagged with ‘High in Relevance’ has high conversion rate history
- Leads thorough ‘Lead Add Form’ and ‘Quick Add Form ’ has high possibility of conversion
- Lead Source ('Welingak Website' and 'Reference') are good indicators to determine if lead will convert.

Bivariate Analysis – Categorical variables



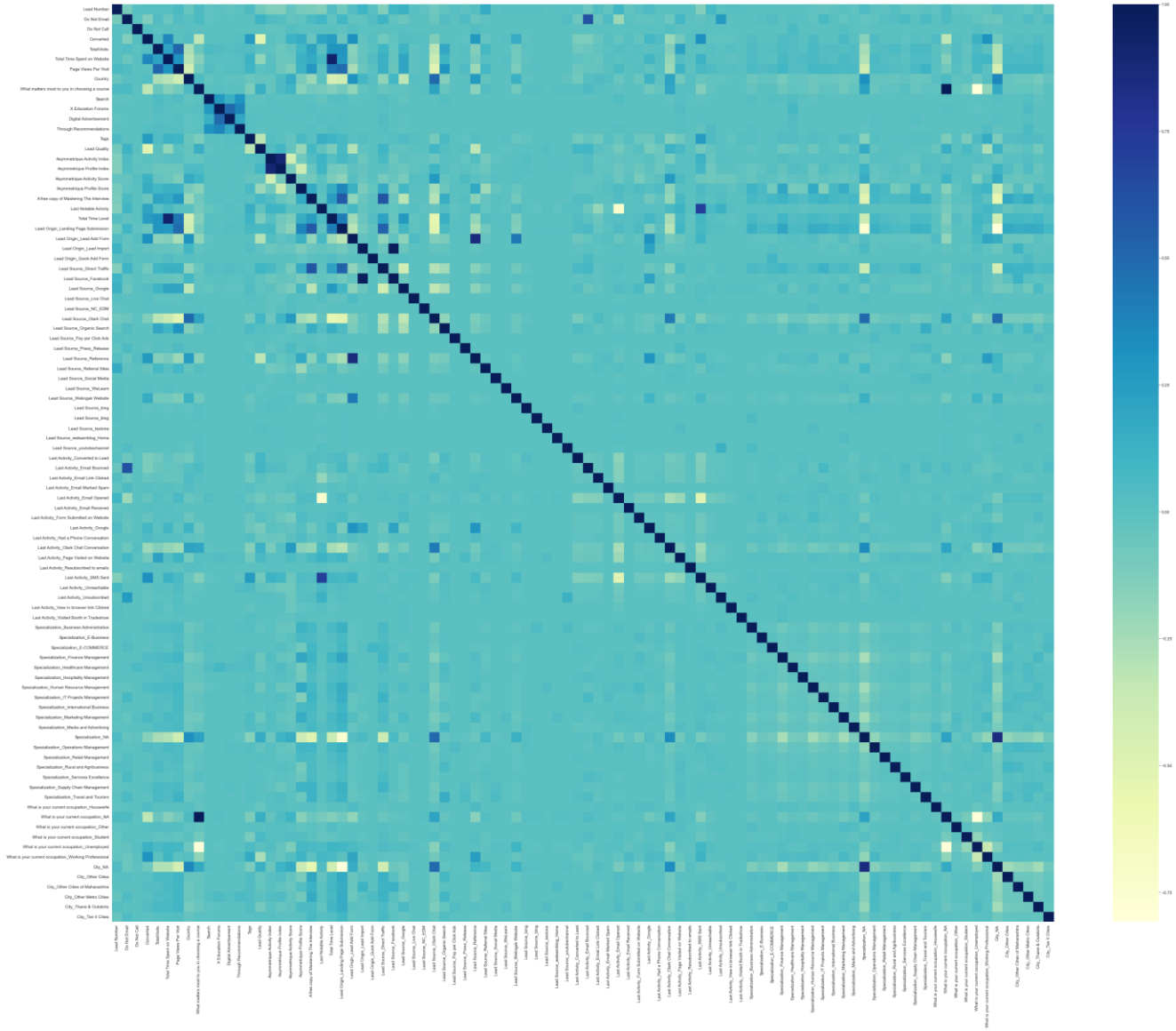




Bivariate Analysis – Checking correlation

- Following group of columns are positively highly correlated with each other:
  1. Search
  2. X Education
  3. Digital Advertisement
  4. Through Recommendations
- 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.

Bivariate Analysis – Checking correlation



Data preparation for Modeling

Create Dummy Variable:

- Dummy Variables has 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 user label encoding for variables with higher level to avoid drastic increase in dataframe size.
- All relevant categorical variable have been encoded using ‘LabelEncoder’

Data preparation for Modeling

Binary Variables Encoding:

- Variables with binary values have been encoded with 0/1

Train- Test Split:

- The modified Leads dataset has been split into Train and Test dataset with the ratio of 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 to avoid the dominance of variables with high magnitude in the model.
- 'StandardScaler' function has been used to scale the data for modeling. It brings all the data points into a standard normal distribution with mean at '0' and deviation at'1'

Create Dummy Variable:

- Dummy Variables has 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 user label encoding for variables with higher level to avoid drastic increase in dataframe size.
- All relevant categorical variable have been encoded using ‘LabelEncoder’

Model Building: Using Logistic Regression

- Generalized Linear Model (GLM) from StatsModel 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 , hence we need to perform feature selection technique.

Feature selection technique using Recursive Feature Elimination (RFE)

- RFE is an optimization technique for finding the best performing subset of features, based on the idea of repeatedly constructing a model, choosing either the best (based on the 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 variable into asset 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 OCA was performed on the PCA dataset by taking first 10 principal components which are explaining more than 95% of the variance
- The 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 obtaining from the model was good, the results without using PCA was even better.
- Therefor we have proceeded further for prediction and conclusion with the last logistic regression model which was built without using PCA.

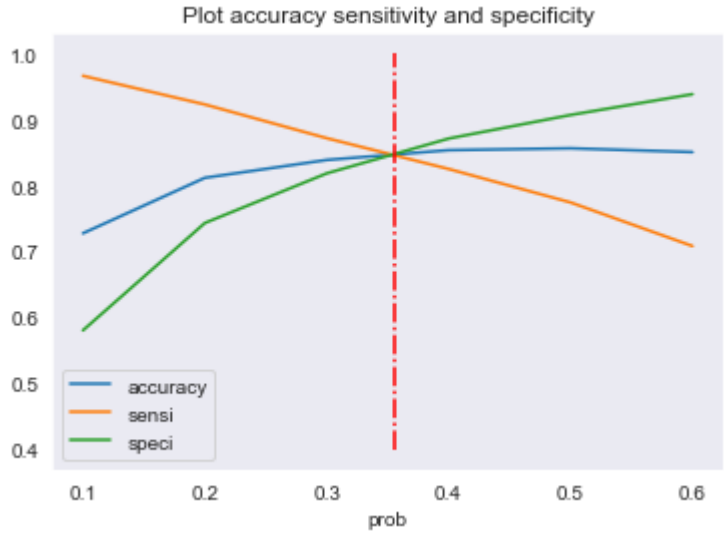
Final Model and Interpretation

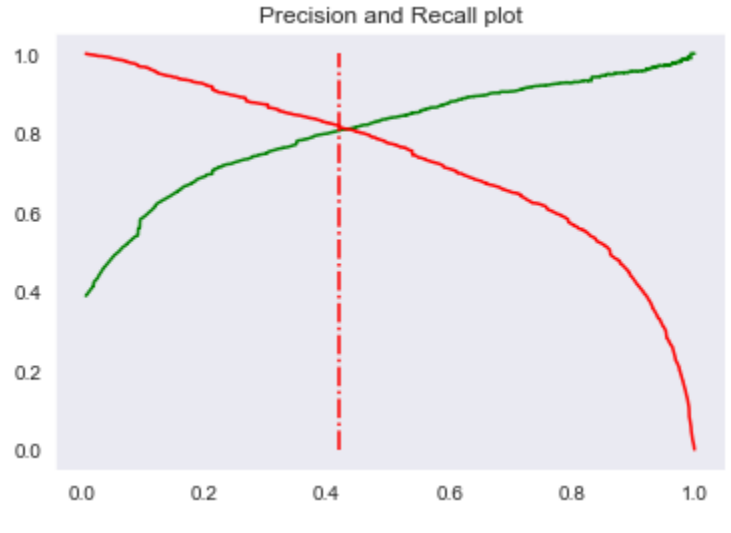
- Final Model contains 14 most important features which satisfy all the selection criteria.
- Lead score have conversion probability greater than 0.42 are being predicted as 'Converted'
- Using this probability threshold value(0.42), the leads from the test dataset have been predicted whether they would get converted or not.
- Confusion Matrix with cut-off 0.42 has been create to calculate evaluation metrics.
- Confusion Metrics: [[3515 487] [448 2018]]
- Evaluation Metrics:
  - Accuracy: 0.8554
  - Sensitivity: 0.8183
  - Specificity: 0.8783
  - Precision: 0.8056

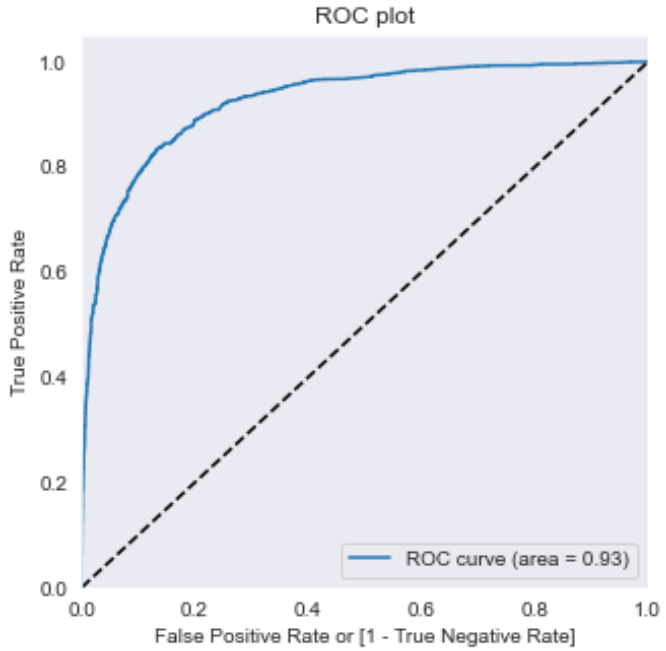
Evaluation Metrics

- 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 0.42 is observed from recall and precision plot.
- We will use both the cutoff and evaluate the results for further predictions

Final Model and Interpretation



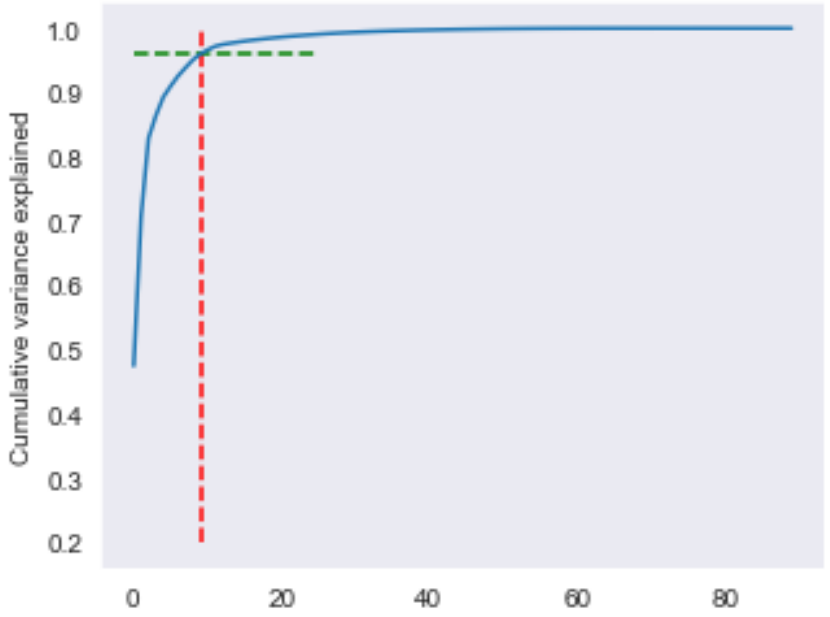




Evaluation Using PCA

- Using PCS helps in dimensionality reduction and solved for multicollinearity issue.
- Making prediction using model build using PCA gives decent results but presents below challenge.
  - Score less that model build without using PCA
  - Identify original variable / factors leading to high score
  - Metric derived from PCA
    - Accuracy: 0.829
    - Sensitivity: 0.7653
    - Specificity: 0.8706
    - Precision: 0.7943
  - PCA - Confusion metric : [[1460 217] [257 838]]
- Model without PCA yields better result.

Evaluation Using PCA



Conclusion and Recommendations:

- Following are the top three features that contribute most towards the probability of a lead getting converted based on the coefficient values
  - Lead Origin_Lead Add Form
  - What is your current occupation Working Professional
  - Last Activity_SMS Sent
- Following are the top three categorical/dummy variables that should be focused the most in order to increase the probability of lead conversion based on the coefficient values
  - Lead Add Form (from Lead Origin)
  - Had a Phone Conversation(from Last Activity)
  - Working Professional (from What is your current occupation)

Conclusion and Recommendations:

- 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 forcasting
  - Minimize oppurtunities loss
  - Increase in revenue.

THANK YOU