LEAD SCORING CASE STUDY ASSIGNMENT

BY –
KESHAV ATCHUTUNI
PRANAV KESHAV
KHUSHI PATEL

AGENDA

Problem Statement

Data Exploration

Data Cleaning and Preparation

Data Preparation for Modelling

Model Building: Using Logistic Regression

Final Model and Interpretation

Conclusion & Discussion

PROBLEM STATEMENT

- Develop a logistic regression model to score leads from 0 to 100, with higher scores indicating a greater likelihood of conversion to paying customers.
- Analyze key predictive variables for lead conversion, manage any outliers in the data, and consider both technical and business factors in the model's construction.
- Evaluate the model's predictions using metrics such as accuracy, sensitivity, specificity, and precision to summarize lead conversion rates.

DATA EXPLORATION

1. ‘Leads.csv’ contains all the information about the leads generated through various sources and their activities.
2. This file contains 9240 rows and 37 columns.
3. Out of 37 columns, 7 are numeric columns and 30 are non-numeric or categorical columns.
4. Current conversion rate of the leads is 39%.
5. ‘Leads Data Dictionary.csv’ is data dictionary which describes the meaning of the variables present in the “Leads” dataset.

DATA CLEANING & PREPARATION

- In the Leads.csv file, the following columns contain more than 30% null values initially:
- What is your current occupation
- What matters most to you in choosing a course
- Tags
- Lead Quality
- Lead Profile
- Asymmetric Activity Index
- Asymmetric Profile Index
- Asymmetric Activity Score
- Asymmetric Profile Score

Certain columns have 'select' as a dominant value, equivalent to a null value. Therefore, 'select' has been converted to 'NA'. These columns include:

- Specialization
- How did you hear about X Education
- Lead Profile
- City

All missing values in categorical columns have been imputed with ‘NA’.

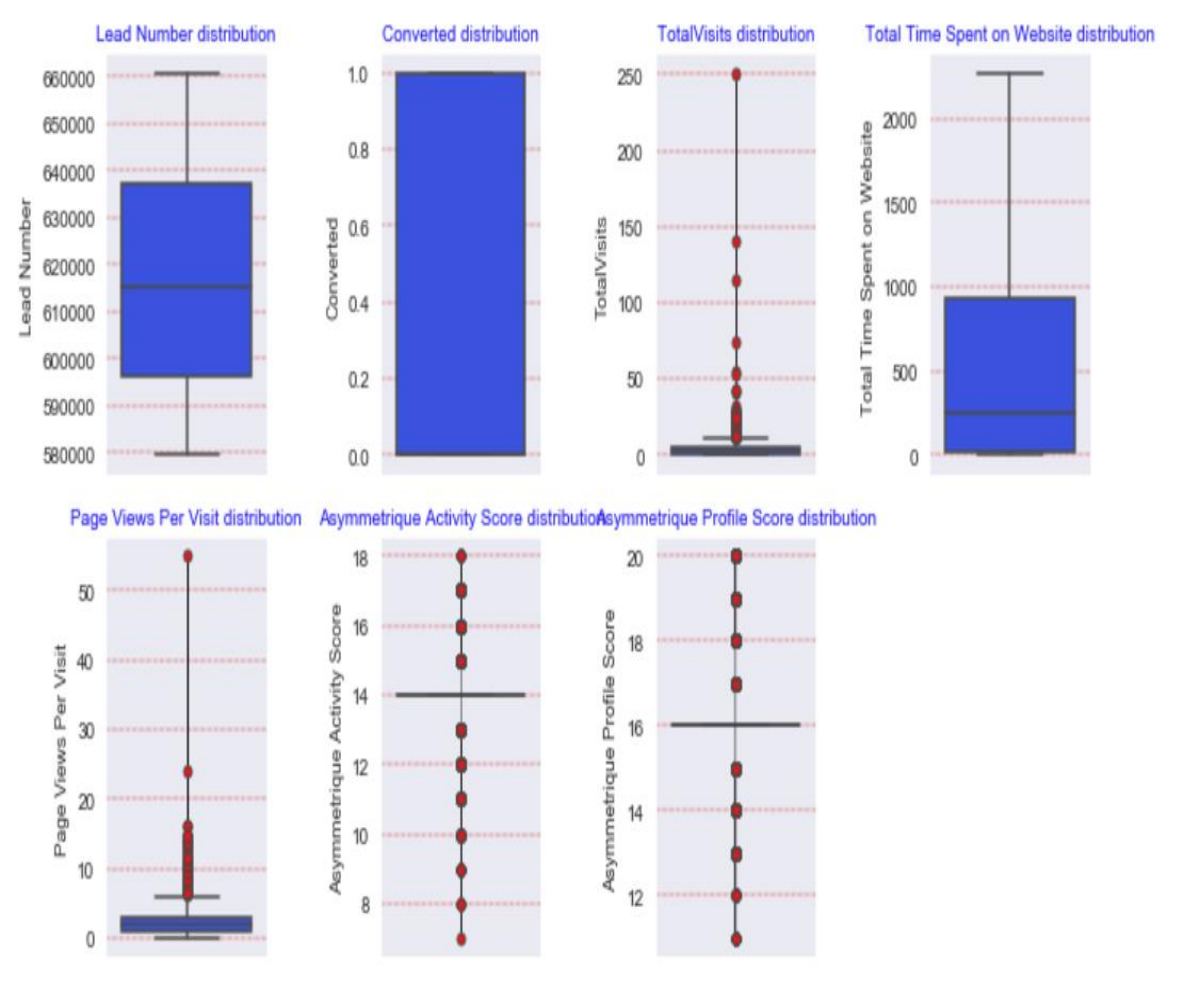
DATA CLEANING & PREPARATION

All missing values in quantitative columns have been filled with the median, as the difference between the mean and median is negligible.

- The following columns, containing a single value and contributing insignificantly, have been dropped:
- Magazine
- Receive More Updates About Our Courses
- Update me on Supply Chain Content
- Get updates on DM Content
- I agree to pay the amount through cheque
- The following columns have been dropped due to having more than 70% missing values:
- How did you hear about X Education
- Lead Profile
- The following columns, with a low percentage of missing values, have been imputed with the mode:
- Lead Source
- Lead Activity

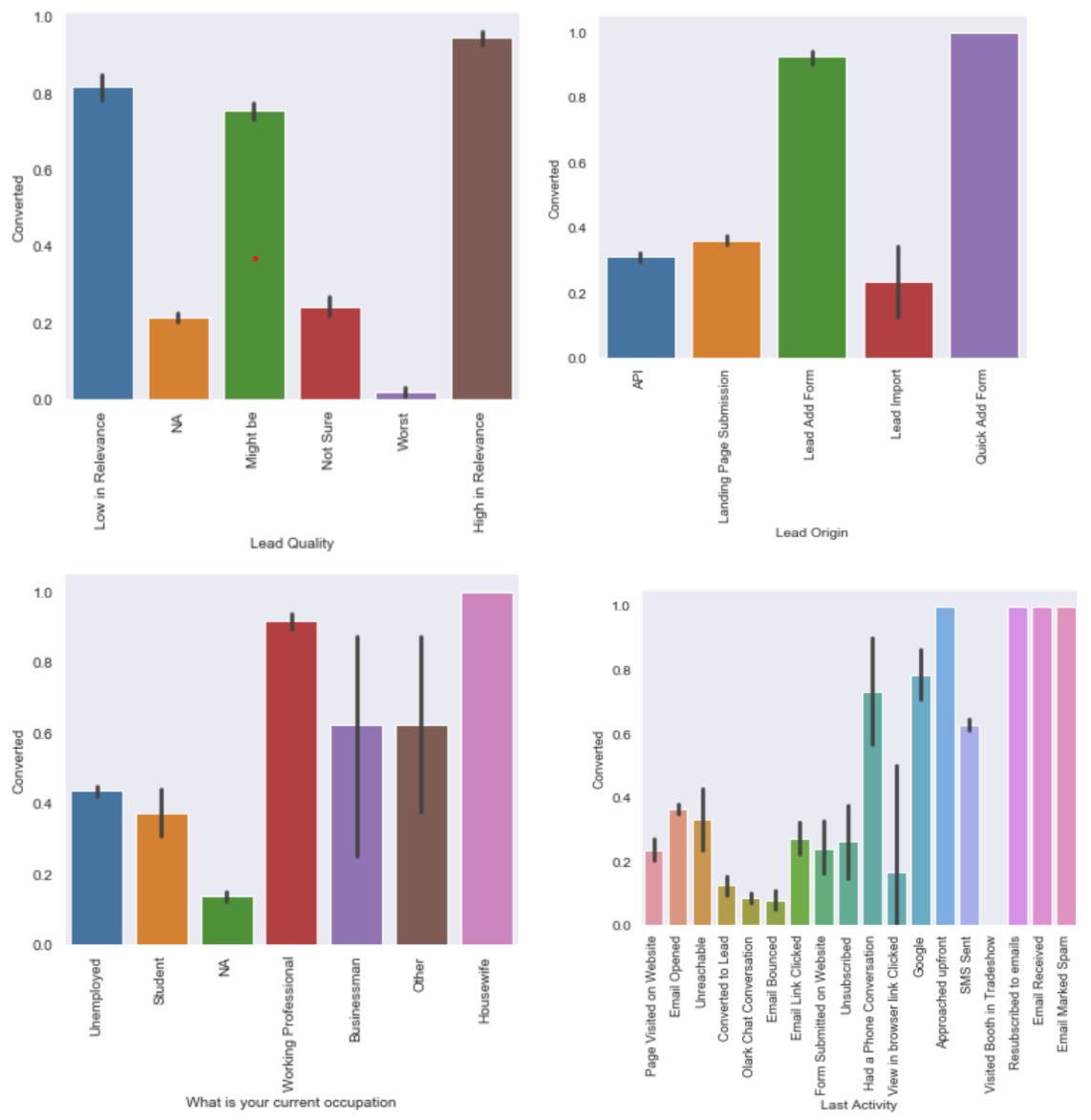
UNIVARIATE ANALYSIS - OUTLIERS

- Univariate analysis revealed data distribution and outliers in the 'Leads' dataset. Key columns with identified outliers are:
- Total Visits
- Page Views Per Visit
- Asymmetric Activity Score
- Asymmetric Profile Score
- The Interquartile Range (IQR) method was used to address outliers. The decision was made not to remove any outliers, as they constitute 9% of the data. We will review the final model to ensure this does not impact the score.



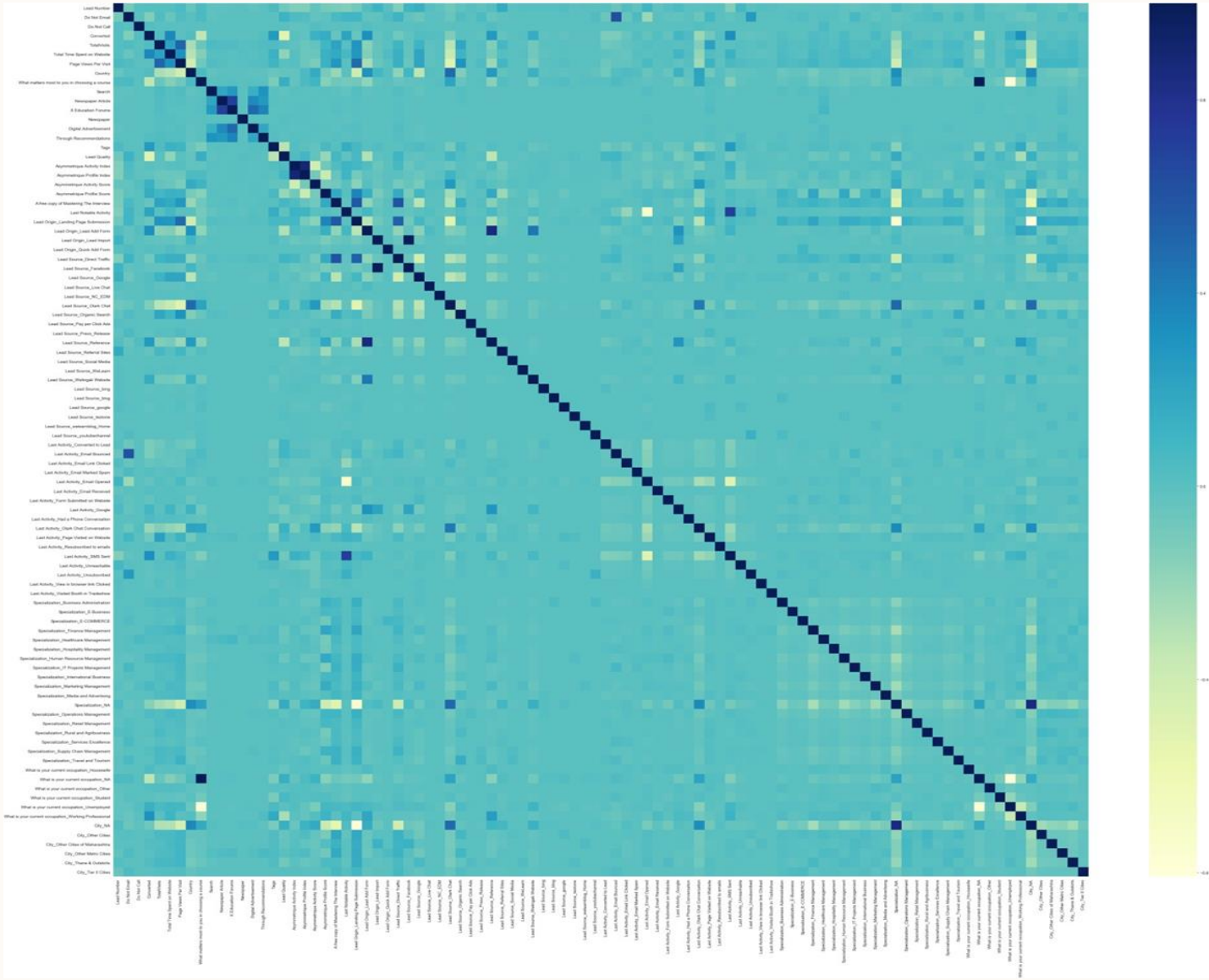
BIVARIATE ANALYSIS – CATEGORICAL VARIABLE

- The 'Converted' column has been selected as the target variable. Therefore, a bivariate analysis of key variables was conducted in relation to this target variable.
- Lateral students and visitors interested in the next batch have a higher likelihood of conversion. Leads tagged as "High in Relevance" demonstrate a strong history of conversion rates.
- Leads originating from the "Lead Add Form" and "Quick Add Form" are more likely to convert. Additionally, leads from the Welingak Website, WeLearn, Live Chat, and NC_EDM show higher conversion rates than those from other sources.



BIVARIATE ANALYSIS – CHECKING CORRELATION

- The following group of columns are highly positively correlated with each other:
- Search
- Newspaper Article
- X Education
- Digital Advertisement
- Through Recommendations
- Another set of columns also exhibit a strong positive correlation with each other:
- Total Visits
- Total Time Spent on Website
- Page Views Per Visit
- Additionally, there is a strong correlation between the Asymmetric Activity Index and the Asymmetric Profile Index.



DATA PREPARATION FOR MODELING

Creating Dummy Variable:

Converting independent variables into dummy variables enables easier interpretation and calculation of odds ratios, enhancing the stability and significance of the coefficients. Dummy variables have been created for the following columns:

- Lead Origin
- Lead Source
- Last Activity
- Specialization
- What is your current occupation
- City

Label Encoding:

Label encoding involves converting each value in a column to a numerical format. For variables with many levels, we will use label encoding to prevent a significant increase in the dataframe size. All relevant categorical variables have been encoded using 'LabelEncoder'.

DATA PREPARATION FOR MODELING

Binary Variables Encoding:
Variables with binary (Yes/No) values have been encoded as 1 and 0, where 1 represents Yes and 0 represents No.

Train – Test Split:
The modified 'Leads' dataset has been divided into training and testing sets in a 70:30 ratio. The training dataset is used to train the model, while the testing dataset is used to evaluate it.

Feature Scaling:
To ensure that variables with higher magnitudes do not dominate the model, it is crucial to scale all variables to the same level. The "StandardScaler" function has been applied to the data, standardizing it to a normal distribution with a mean of 0 and a standard deviation of 1.

MODEL BUILDING: USING LOGISTIC REGRESSION

Initial Model: Developed logistic regression using GLM with all 93 features from the dataset.

Feature Selection: Implemented Recursive Feature Elimination (RFE) to identify the top 20 features that contributed most to the model's predictive power. RFE iteratively ranks features based on their coefficients and eliminates the least significant ones until the desired number is reached.

Refinement Process: Evaluated remaining features based on statistical significance and multicollinearity.

- **P-values:** Features with P-values greater than 0.05 were considered insignificant and removed from consideration.
- **Variance Inflation Factor (VIF):** Features with VIF greater than 5 were dropped to address multicollinearity issues, ensuring that selected features were independent predictors.

Final Model Construction: Constructed the final logistic regression model using the refined subset of features identified through RFE, P-value assessment, and VIF analysis.

Outcome: This approach enhanced the model's predictive performance by focusing on the most relevant features, thereby improving interpretability and efficiency in the logistic regression model. The refined model is expected to provide more reliable predictions by eliminating noise from the initial feature set and reducing the risk of multicollinearity, contributing to better decision-making capabilities in practical applications.

MODEL BUILDING: USING LOGISTIC REGRESSION (ON PCA DATA)

1. Principal Component Analysis (PCA) is a statistical method that transforms a dataset of potentially correlated variables into a set of linearly uncorrelated variables known as principal components through orthogonal transformations. Initially, PCA was applied to the X_train dataset (excluding Prospect ID and Lead Number columns). Subsequently, Incremental PCA was employed on this PCA-transformed dataset, retaining the first 10 principal components that collectively explain over 95% of the variance. The same process was repeated on the X_test dataset for consistency.
2. Following PCA, Logistic Regression models were built using the PCA-transformed datasets. While these models produced satisfactory results, it was observed that the performance of models constructed without PCA was superior. Consequently, for prediction and conclusive analysis, the logistic regression model without PCA was selected and further utilized.
3. This approach highlights the iterative nature of model refinement and the importance of evaluating different techniques to optimize predictive accuracy.

FINAL MODEL AND INTERPRETATION

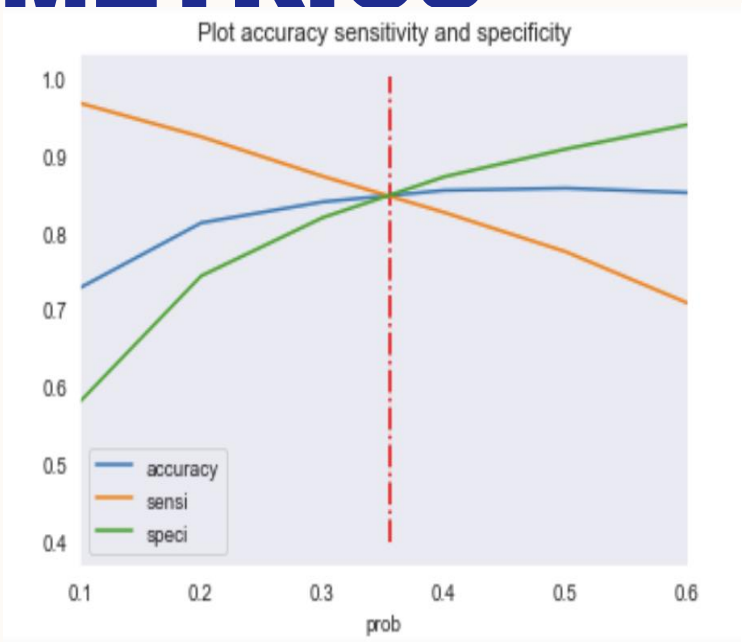
1. The final model includes 14 key features that meet all selection criteria. Leads with a lead score indicating a conversion probability greater than 0.43 are classified as "Converted". Using this threshold, predictions were made on leads from the test dataset to determine conversion likelihood. A confusion matrix was constructed using this cutoff value of 0.43 for evaluating the model's performance:
2. Confusion Matrix: [[3630 372] [556 1910]]
3. The following evaluation metrics were calculated:

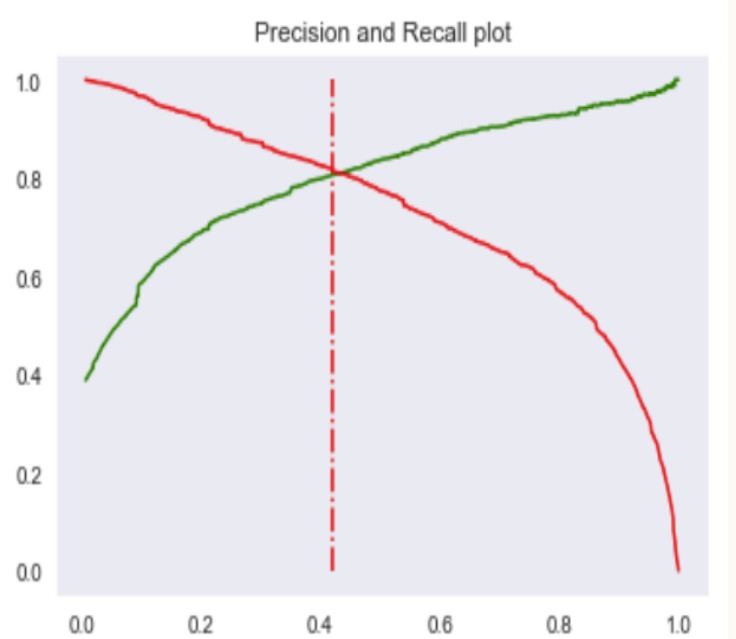
- Accuracy: 0.8565
- Sensitivity (True Positive Rate): 0.7745
- Specificity (True Negative Rate): 0.907
- Precision: 0.837

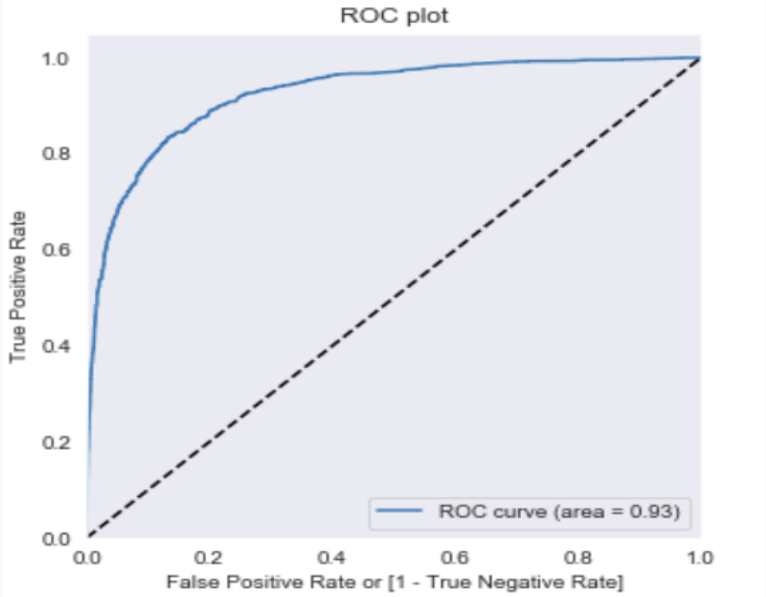
These metrics provide insights into how well the model distinguishes between converted and non-converted leads, demonstrating its effectiveness in predicting conversions based on the specified threshold.

EVALUATION METRICS

- Receiver Operating Characteristic (ROC) Curve: The Area Under the Curve (AUC) of the ROC curve indicates the model's performance. A higher AUC signifies a better model. In our case, the ROC curve is positioned towards the upper left, indicating strong model performance. The AUC value for our model is 0.93.
- Accuracy, Sensitivity, and Specificity Plot: The plot demonstrates the trade-off between sensitivity and accuracy, with a cutoff of 0.42.
- Precision-Recall Plot: From the precision-recall plot, an optimal cutoff of 0.43 is identified.
- Both cutoff values will be used to evaluate results for future predictions, ensuring comprehensive assessment of model performance.







EVALUATION USING PCA

Utilizing PCA addresses dimensionality reduction and mitigates multicollinearity issues. Although predictions made using the PCA-built model yield satisfactory results, they score lower than the model constructed without PCA. This discrepancy prompts the identification of original variables or factors contributing to the higher score in the non-PCA model.

- Metrics from the PCA model are as follows:
- PCA Scores:-

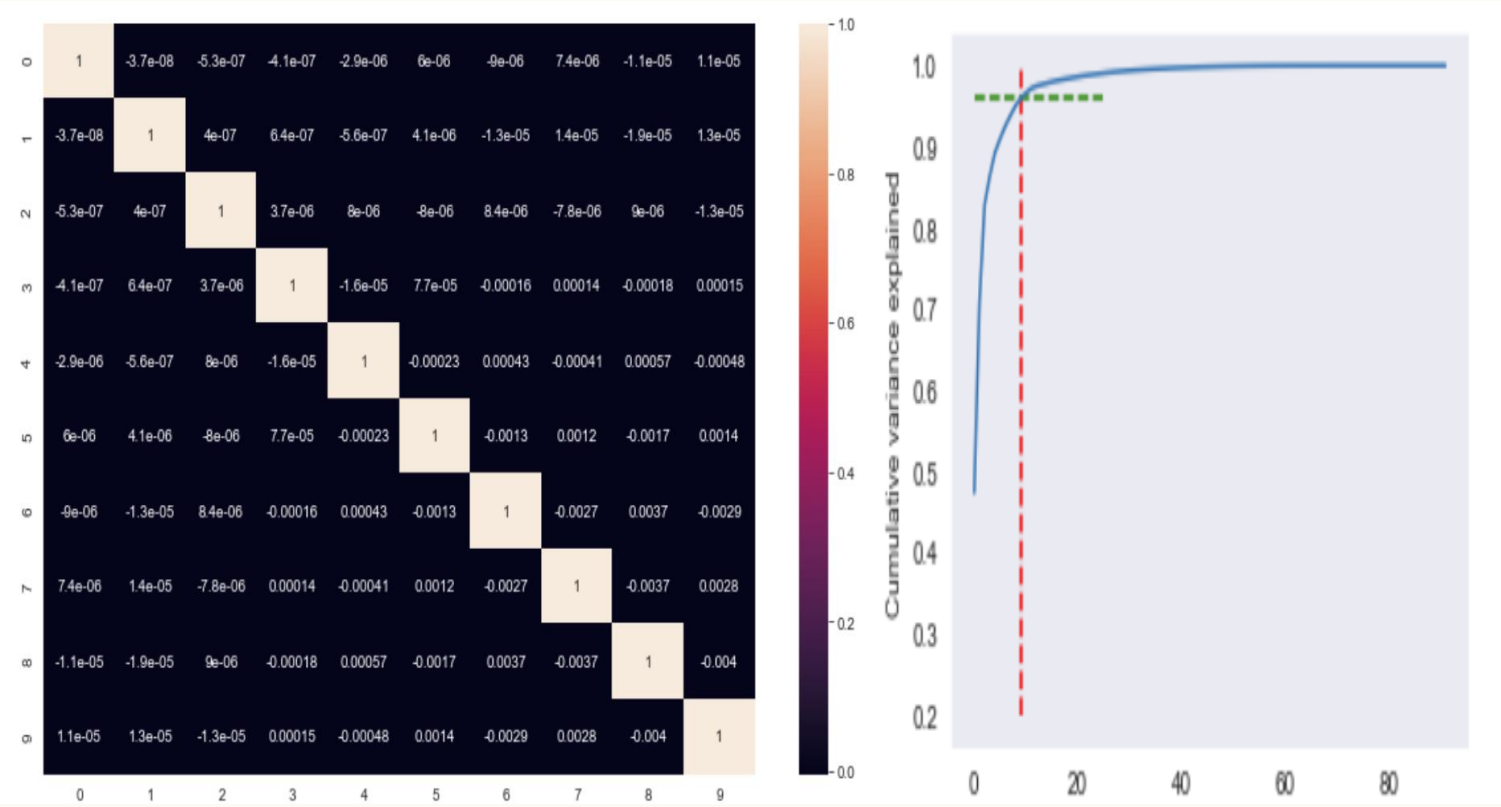
Accuracy: 0.829

Sensitivity: 0.7653

Specificity: 0.8706

Precision: 0.7943

Comparatively, the model without PCA demonstrates superior performance.




```
PCA - Confusion metric
[[1460  217]
 [ 257  838]]
```

CONCLUSION AND DISCUSSION

The following are the top three features that significantly influence the decision, leading to an increase in the conversion probability of a lead as their values increase:

1. Lead Origin: Specifically, leads from the "Lead Add Form" origin.
2. What is your current occupation: Particularly, leads who are "Working Professionals".
3. Last Activity: Especially, leads with the last activity being "SMS Sent".

These categories play a crucial role in determining the likelihood of lead conversion based on their respective values.

- This model aids in pinpointing hot leads, thereby improving speed-to-lead and response rates. Focusing solely on hot leads offers several advantages:

1. Accelerated sales cycle through effective prioritization.
2. Improved opportunity-to-deal ratio.
3. Enhanced control over unpredictable buying cycles.
4. Increased effectiveness of marketing efforts.
5. Enhanced accuracy in sales forecasting.
6. Reduction in missed opportunities.
7. Growth in revenue streams.