# Lead Scoring Case Study

# Problem Statement:

The education company, X Education, is facing challenges with its lead conversion rate. Although they generate a significant number of leads, the conversion rate is low. The company wants to improve efficiency by identifying the most potential leads, referred to as 'Hot Leads,' with a higher likelihood of conversion. The objective is to build a model that assigns a lead score to each lead, allowing the sales team to focus on communicating with the most promising leads.

# Business Objectives:

- **Improve Lead Conversion Rate:** The primary goal is to increase the lead conversion rate from the current 30% to around 80%. It involves identifying and targeting leads with higher conversion potential.
- **Build a Predictive Model:** To develop a predictive model that assigns lead scores to each lead based on various attributes. The lead score should reflect the likelihood of conversion.

# Solution Methodology

1

### Exploratory Data Analysis (EDA)

Understand the distribution of data, check for missing values, and explore relationships between variables.

2

### Data Preprocessing

Encoding categorical variables and scaling numerical features.

3

### Feature Selection

Identify the most relevant features that contribute to lead conversion.

4

### Model Building

Train a Logistic Regression model to predict the likelihood of lead conversion based on selected features.

5

### Model Evaluation and Prediction

Assess the model's performance using appropriate metrics.

# Exploratory Data Analysis (EDA)

## Data Cleaning

- Identified and resolved duplicate records using advanced techniques by ensured data integrity and consistency.
- Systematically addressed NA values through imputation or removal.
- Analyzed columns with significant missing values. Dropped non-essential columns to streamline the dataset.
- Investigated low-value count columns. Strategically transformed values for better analysis accuracy.

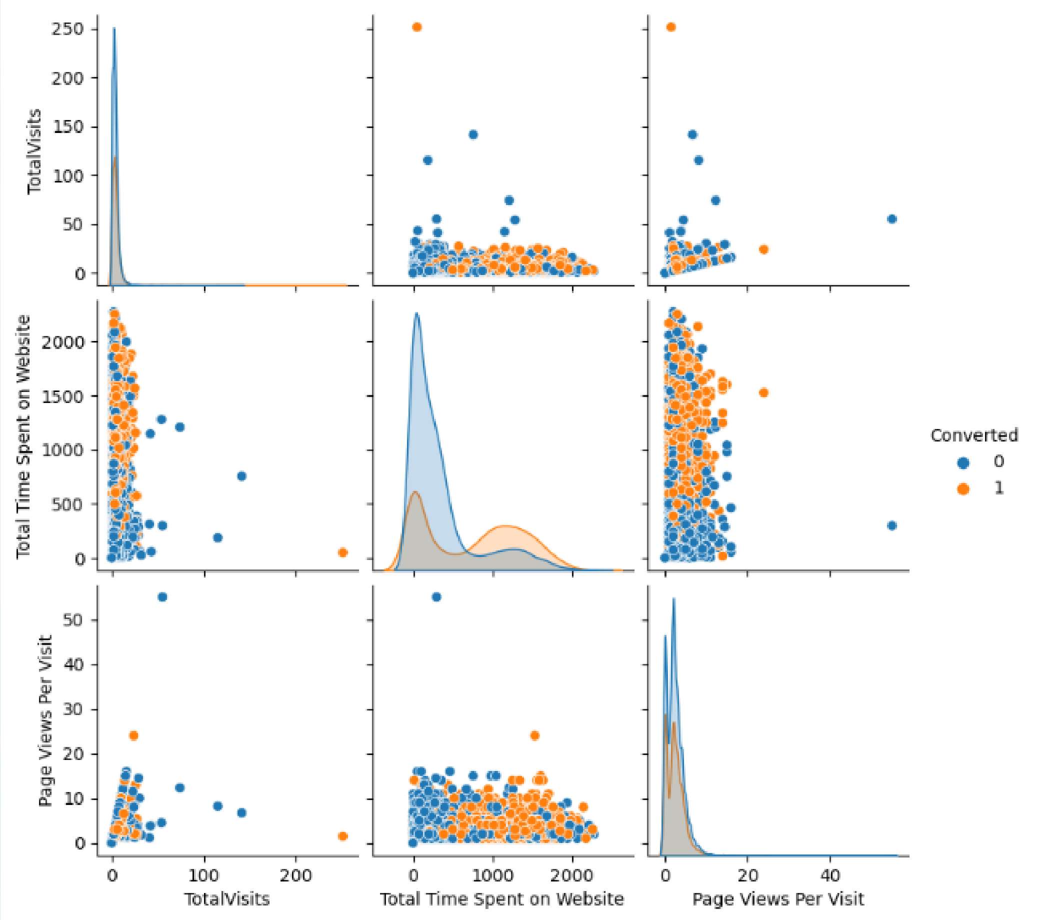
## Data Visualization

- Utilized pairplots to visualize relationships between numerical variables.
- Created count plots to represent the distribution of categorical variables with respect to target variable 'Converted'.
- Visualized the distribution of numerical values with the help of boxplots.
- Leveraged boxplots to identify potential outliers in the dataset.

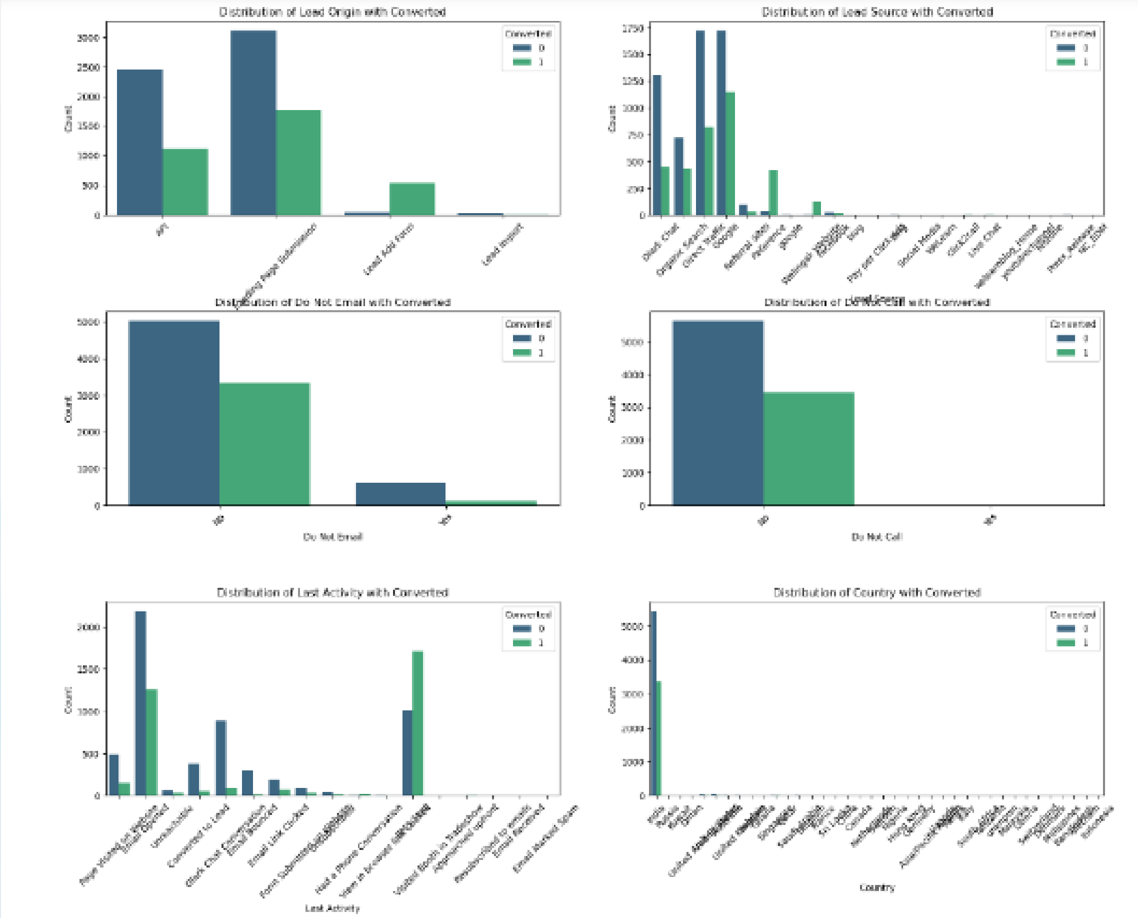
# Data Cleaning

- The dataset did not contain any duplicate rows.
- Dropped columns ("Lead Quality", "Asymmetrique Activity Index", "Asymmetrique Profile Index", "Asymmetrique Activity Score", "Asymmetrique Profile Score") with a high proportion (>40%) of missing values.
- Dropping rows with a low proportion (<10%) of missing values in specific columns ("Lead Source", "TotalVisits", "Page Views Per Visit", "Last Activity").
- Imputed missing values for column with a moderate proportion (10 - 35%) with the most frequent value.
- There were several columns in the dataset where the most frequent value is 'Select'. This likely indicates that customers either missed filling in these details or left them blank. To retain information and avoid losing data by removing these instances, we have converted the 'Select' values in those columns to 'Not Specified'.
- Dropped several columns as the majority of values are the same (e.g., 'No'), and there is little variability, it may not contribute much to the analysis. Such as 'Country', 'Do Not Email', 'Do Not Call', 'How did you hear about X Education', 'What matters most to you in choosing a course', 'Search', 'Magazine', etc.

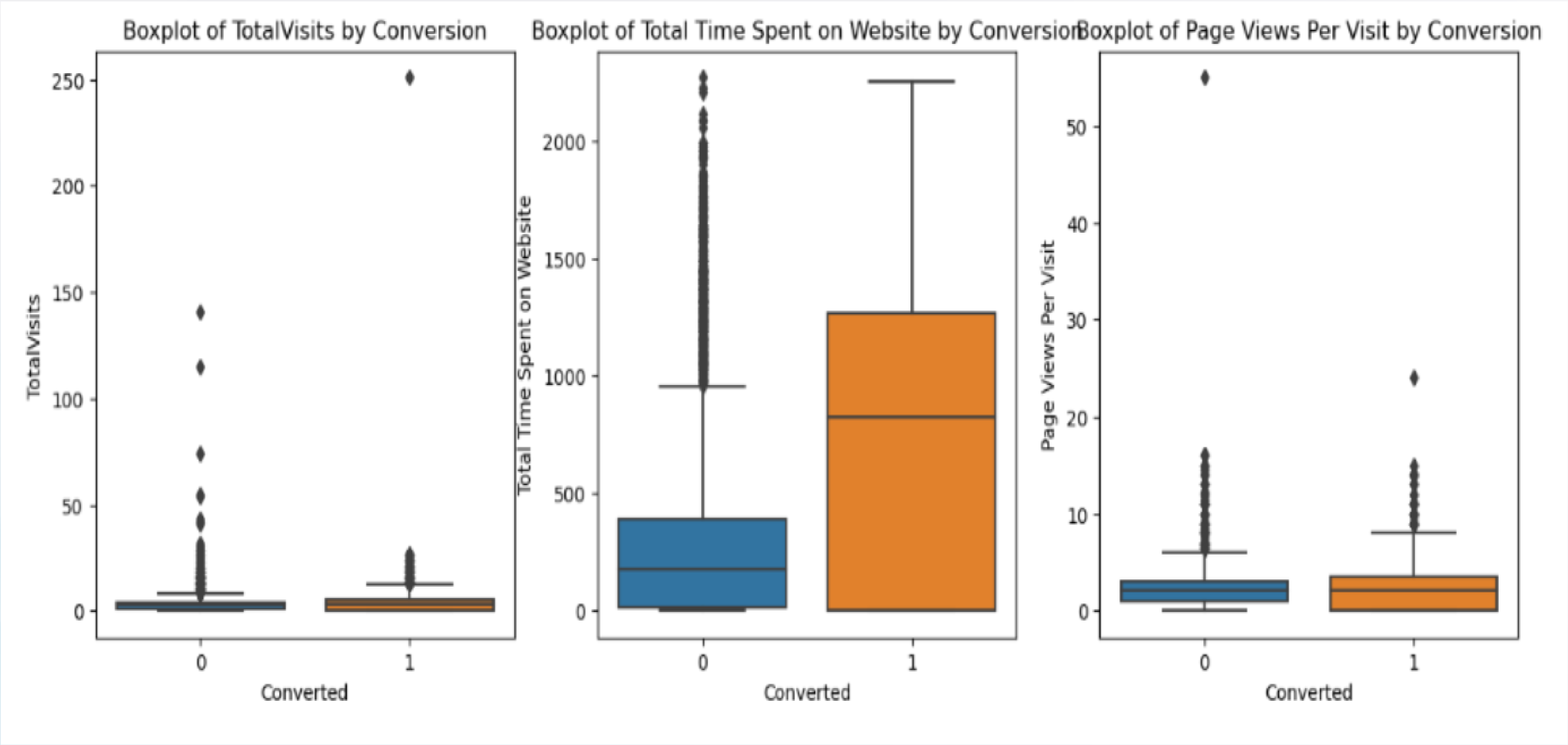
# Data visualization




Using pair plots to visualize scatter plots between pairs of numerical variables




Plotting count plots for each categorical columns w.r.t. the 'Converted' variable




Visualization the distribution of numerical variables across different categories, identifying outliers and understanding the spread of the data using box plots

# Data Preprocessing and Feature Selection

- Created dummy variables for categorical columns ('Lead Origin', 'Lead Source', 'Last Activity', 'Specialization', 'What is your current occupation', 'Tags', 'City', 'Last Notable Activity').
- Splits the data into training and testing sets (70% training, 30% testing).
- Scaling the variables is performed to standardize the range of features and to ensuring that all features contribute equally to the model. StandardScaler is used for scaling.
- Identified the most relevant features that contribute to lead conversion with the help of Recursive Feature Elimination (RFE).

# Model Building

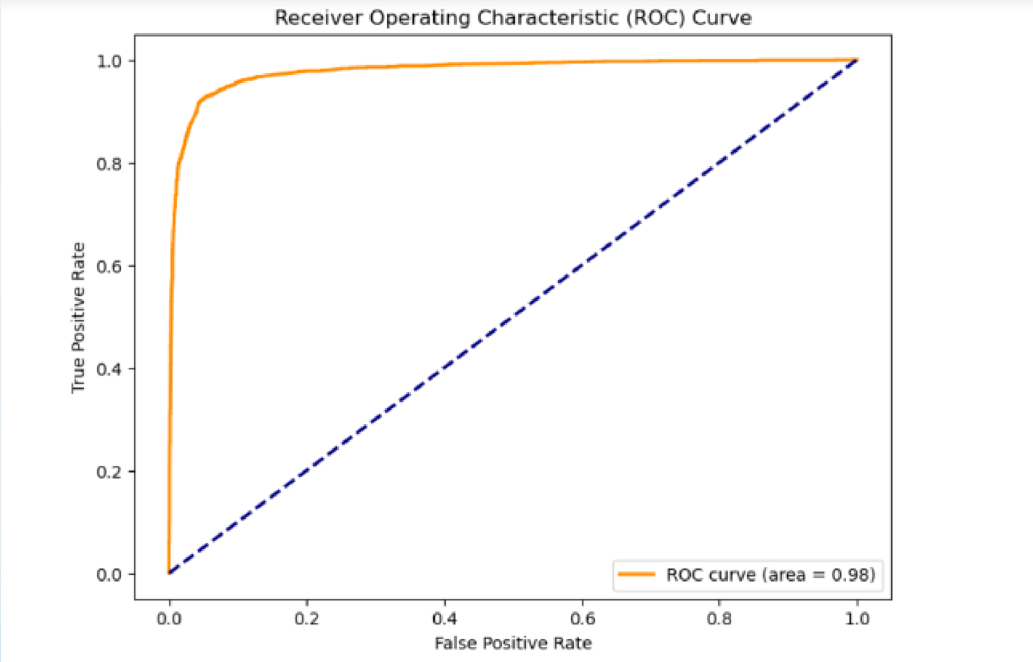
Trained a Logistic Regression model to predict the likelihood of lead conversion based on selected features. Two models were built, second model (final model) "lead_model2" was stable with p-values of coefficients less than 0.05 and VIF less than 5.

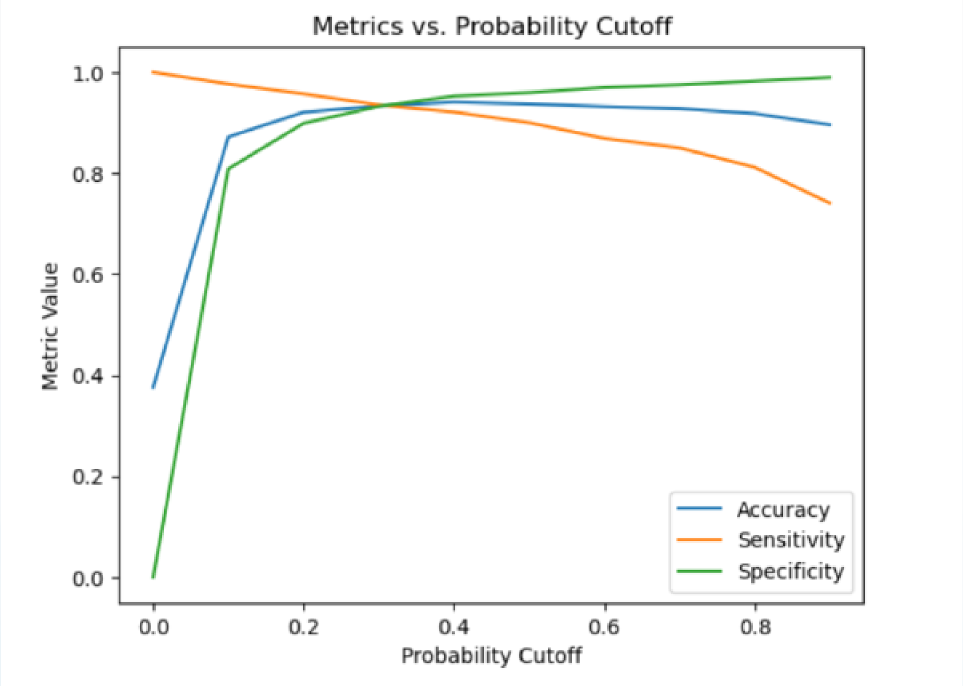
```
                 Generalized Linear Model Regression Results
==============================================================================
Dep. Variable:              Converted   No. Observations:                 5909
Model:                            GLM   Df Residuals:                     5889
Model Family:                Binomial   Df Model:                           19
Link Function:                  Logit   Scale:                          1.0000
Method:                          IRLS   Log-Likelihood:                -1017.9
Date:                Mon, 15 Jan 2024   Deviance:                       2035.8
Time:                        18:24:53   Pearson chi2:                 1.14e+04
No. Iterations:                     8   Pseudo R-squ. (CS):             0.6246
Covariance Type:            nonrobust
=================================================================================================================
                                                coef    std err          z      P>|z|      [0.025      0.975]
-----------------------------------------------------------------------------------------------------------------
const                                        -0.3274      0.134     -2.445      0.014      -0.590      -0.065
Total Time Spent on Website                   1.0457      0.068     15.467      0.000       0.913       1.178
Lead Origin_Lead Add Form                     1.5955      0.471      3.387      0.001       0.672       2.519
What is your current occupation_Not Specified -2.1449     0.142    -15.099      0.000      -2.423      -1.866
Lead Source_Olark Chat                        1.2835      0.167      7.700      0.000       0.957       1.610
Lead Source_Welingak Website                  3.2369      1.130      2.865      0.004       1.022       5.452
Last Activity_Converted to Lead              -1.5373      0.370     -4.156      0.000      -2.262      -0.812
Last Activity_Email Bounced                  -1.8454      0.434     -4.249      0.000      -2.697      -0.994
Last Activity_Olark Chat Conversation        -1.6584      0.246     -6.752      0.000      -2.140      -1.177
Last Activity_Page Visited on Website        -1.1009      0.279     -3.944      0.000      -1.648      -0.554
Tags_Already a student                       -4.6532      0.725     -6.422      0.000      -6.073      -3.233
Tags_Closed by Horizzon                       4.5340      0.746      6.079      0.000       3.072       5.996
Tags_Interested in other courses             -3.1964      0.360     -8.884      0.000      -3.902      -2.491
Tags_Lost to EINS                             5.0280      0.619      8.122      0.000       3.815       6.241
Tags_Others                                  -3.3903      0.303    -11.188      0.000      -3.984      -2.796
Tags_Ringing                                 -5.2471      0.304    -17.259      0.000      -5.843      -4.651
Tags_Will revert after reading the email      3.1762      0.220     14.459      0.000       2.746       3.607
Tags_switched off                            -5.7849      0.747     -7.749      0.000      -7.248      -4.322
Last Notable Activity_Email Link Clicked     -0.8834      0.447     -1.976      0.048      -1.760      -0.007
Last Notable Activity_SMS Sent                2.0735      0.147     14.147      0.000       1.786       2.361
=================================================================================================================
```

# Model Evaluation and Prediction

Assess the model's performance using appropriate metrics.

- The model achieved an overall accuracy of approximately 93.72%, indicating that it correctly predicted the conversion status for about 93.72% of the observations.
- Sensitivity (True Positive Rate, Recall) is high at 93.52%, indicating that the model effectively identified a significant portion of the actual positive cases (converted leads).
- Specificity (True Negative Rate) is also high at 93.24%, suggesting that the model performed well in identifying non-converted leads.
- The False Positive Rate (FPR) is relatively low at 6.76%, indicating a good balance between correctly identifying positive cases and minimizing false alarms.
- Precision, which measures the accuracy of positive predictions, is 89.3%. This indicates that out of all predicted conversions, about 89.3% are true conversions.
- Recall, representing the proportion of actual positive cases correctly predicted by the model, is 93.52%, indicating a high ability to capture actual conversions.




**Receiver Operating Characteristic (ROC) Curves** which show the tradeoff between the True Positive Rate (**TPR**) and the False Positive Rate (**FPR**)

**Plots of accuracy, sensitivity, and specificity for various probabilities.**

The optimal threshold refers to the probability threshold used to classify the predicted probabilities into binary outcomes.

# Prediction and Evaluation on Test Set

- **Accuracy (ACC):** 93.57%. The overall accuracy of the model on the test set is quite high, indicating a good fit.
- **Sensitivity (True Positive Rate, Recall):** 93.41%. The model correctly identifies 93.41% of the actual conversions, suggesting a strong ability to capture positive instances.
- **Specificity (True Negative Rate):** 93.66%. The model exhibits a high specificity, correctly identifying 93.66% of non-conversions, indicating its ability to distinguish negative instances.
- **False Positive Rate (FPR):** 6.34%. The low FPR suggests that the model has a relatively low rate of incorrectly classifying non-conversions as conversions.
- **Precision:** 89.69%. Among the instances predicted as conversions, 89.69% are true conversions. This indicates the reliability of the model in its positive predictions.
- **Recall:** 93.41%. The recall value is consistent with sensitivity, emphasizing the model’s effectiveness in capturing actual conversions.

Overall, the model demonstrates robust performance on the test set, with a good balance between precision and recall, as well as high accuracy and specificity.

# Recommendations

- **Focus on Lead Add Form:** Since leads originating from the lead add form have a higher likelihood of conversion, consider strategies to increase engagement with this form.
- **Engage Leads from Welingak Website:** As leads from the Welingak website are more likely to convert, ensure that the website is optimized for lead generation and provides valuable information.
- **Address Tags and Last Activities:** Pay attention to specific tags and last activities that significantly impact conversion. For instance, leads tagged as "Already a student" or with the last activity as "Converted to Lead" may need different strategies.
- **Optimize Olark Chat Conversations:** Since leads from Olark Chat have higher conversion odds, focus on optimizing and engaging with leads through this channel.
- **Improve Last Notable Activity - Email Link Clicked:** "SMS Sent" and "Email Link Clicked" have significant impacts. Consider focusing on these activities for effective communication.
- **Mitigate Negative Influences:** Identify and address factors that negatively impact conversion, such as leads with no specified occupation or those with the last activity as "Email Bounced" or "Ringing."

# Thank You