Summary

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.

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.

Exploratory Data Analysis (EDA)

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

A. **Data Cleaning:**
   - There were no duplicates in the lead data.
   - Columns with a high proportion (>40%) of missing value were dropped.
   - Rows with a low proportion (<10%) of missing values in specific columns were dropped.
   - Missing values with proportion 10-30% have been imputed with most frequent value.
   - There were several columns in the dataset where the most frequent value was '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'.
   - Other activities like outliers’ treatment, fixing invalid data, grouping low frequency values, mapping binary categorical values were carried out.

B. **Data Visualization:**
   - Data imbalance checked- only 37.85% leads converted.
   - Visualized scatter plots between pairs of numerical variables using pair plots.
   - Plotted count plots for each categorical column w.r.t. the 'Converted' variable
   - Many irrelevant columns and columns having majority values as ‘No’ have been dropped as these columns have little variability.

Data Preparation

- Created dummy features (one-hot encoded) for categorical variables
- Splitting Train & Test Sets: 70:30 ratio
- Feature Scaled using Standardization

Model Building

- Used RFE to reduce variables from 67 to 15. This will make Data Frame more manageable.
- Manual Feature Reduction process was used to build models by dropping variables with p-value > 0.05.
- Total 2 models were built, second (final) model was stable with (p-values < 0.05). No sign of multicollinearity with VIF < 5.

Model Evaluation

- ROC curve was plotted, giving score of 0.98, suggesting strong predictive performance.
- 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) 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.

Making Predictions on Test Data

- To ensure consistency in the preprocessing steps between the training and test sets, numerical variables in the test set are scaling using the same scaling parameters obtained from the training 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."