Summary

X Education gets a lot of leads, its lead conversion rate is very poor at around 30%. The company requires us to build a model wherein we need to assign a lead score to each of the leads such that the customers with a higher lead score have higher conversion chance. CEO's target for lead conversion rate is around 80%.

Data Cleaning:

- Some redundant columns with >40% nulls were dropped. A few columns like 'Country', 'Lead Quality', Lead Profile' etc were maintained even though they had missing values (imputed missing values with with 'unknown') since these are important. Value counts within categorical columns were checked to decide appropriate action: if imputation causes skew, then column was dropped, created new category (others), impute high frequency value, drop columns that don't add any value.
- Numerical categorical data were imputed with mode and columns with only one unique response from customer were dropped.
- Other activities like outliers' treatment, fixing invalid data, grouping low frequency values, mapping binary categorical values were carried out.

EDA:

- Data imbalance checked- only 38 % leads converted.
- Performed univariate and bivariate analysis for categorical and numerical variables. 'Lead Origin', 'Current occupation', 'Lead Source', etc. provide valuable insight on effect on target variable.
- Time spend on website shows positive impact on lead conversion.

Data Preparation:

- Created dummy features (one-hot encoded) for categorical variables
- Splitting Train & Test Sets: 70:30 ratio
- Feature Scaling using Standardization
- Dropped few columns, they were highly correlated with each other

Model Building:

- Used RFE to reduce variables from 80 to 15. This will make dataframe more manageable.
- Manual Feature Reduction process was used to build models by dropping variables with p – value > 0.05.
- Total 3 models were built before reaching final Model 4 which was stable with (p-values < 0.05). No sign of multicollinearity with VIF < 5.
- logm4 was selected as final model with 12 variables, we used it for making prediction on train and test set.

Model Evaluation:

- Confusion matrix was made and cut off point of 0.2 was selected based on accuracy, sensitivity and specificity plot. This cut off gave accuracy, specificity and precision all around 90%. Whereas precision recall view gave less performance metrics around 85-88 %.
- As to solve business problem CEO asked to boost conversion rate to 80%, but metrics dropped when we took precision-recall view. So, we will choose sensitivity-specificity view for our optimal cut-off for final predictions
- Lead score was assigned to train data using 0.2 as cut off.

Making Predictions on Test Data:

- Making Predictions on Test: Scaling and predicting using final model.
- Evaluation metrics for train & test are very close to around 80%.
- Lead score was assigned.
- Top 3 features are:
    - Tags_Closed by Horizzon (Coeff : 2.78)
    - Lead_Quality_High in Relevance (Coeff : 2.04)
    - Lead Origin_Lead Add Form (Coeff : 1.78)

Recommendations:

1. **Focus on Leads Closed by Horizzon (Tags_Closed by Horizzon)**

   Insight: Leads that were previously closed by Horizzon have the highest positive impact on the likelihood of conversion.

   Recommendation:
   - Prioritize similar leads in your current pipeline, as they have a higher probability of conversion.
   - Investigate what specific factors or strategies Horizzon used to successfully close these leads, and replicate those strategies across the organization.
   - Consider segmenting these leads for targeted campaigns or assigning them to the most experienced sales representatives.

2. **Emphasize High-Relevance Lead Quality (Lead_Quality_High in Relevance)**

   Insight: Lead quality is a critical factor, with those rated as "High in Relevance" being much more likely to convert.

   Recommendation:
   - Enhance the lead qualification process to ensure that leads marked as "High in Relevance" are given priority.
   - Use this insight to refine your lead scoring system, ensuring that high-relevance leads are scored appropriately and followed up promptly.
   - Consider allocating more resources to nurturing these leads, as they are more likely to result in successful conversions.

3. **Optimize the Lead Origin from 'Lead Add Form' (Lead Origin_Lead Add Form)**

   Insight: Leads originating from the 'Lead Add Form' are also strong predictors of conversion.

   Recommendation:
   - Increase focus on optimizing and promoting the 'Lead Add Form' across different channels, as it is a significant source of high-quality leads.
   - Analyze the form's performance and user experience to ensure it is easy to use and captures the necessary information effectively.
   - Consider implementing A/B testing or improving the user interface to further enhance conversion rates from this source.

4. **Change in approach towards features with Negative coefficient**

   The features like Tags_Ringing, Tags_Interested in other courses, Tags_Already a student, Last Activity_Email Bounced, and Current occupation_Unspecified have high negative coefficients, indicating that these factors significantly reduce the likelihood of a lead converting.
   - To improve conversion rates, management should consider deprioritizing or revising the approach for leads exhibiting these characteristics.
   - For instance, leads with email bounces might benefit from alternative communication channels, while those already engaged in other courses or identified as "already a student" might require different messaging or be excluded from certain campaigns.
   - Additionally, efforts should be made to gather more specific information for leads with "unspecified" occupations to better assess their potential.

     This approach allows for more targeted marketing and resource optimization, focusing on leads with higher conversion probabilities.
