SUMMARY

Problem Statement:

An education company named X Education sells online courses to industry professionals. On any given day, many professionals who are interested in the courses land on their website and browse for courses. The company markets its courses on several websites and search engines like Google. Once these people land on the website, they might browse the courses or fill up a form for the course or watch some videos. When these people fill up a form providing their email address or phone number, they are classified to be a lead. Moreover, the company also gets leads through past referrals. Once these leads are acquired, employees from the sales team start making calls, writing emails, etc. Through this process, some of the leads get converted while most do not. The typical lead conversion rate at X education is around 30%. X Education needs help to select the most promising leads, i.e. the leads that are most likely to convert into paying customers. A model is required to be built wherein a lead score is assigned to each of the leads such that the customers with higher lead score have a higher conversion chance and the customers with lower lead score have a lower conversion chance. The CEO has given a ballpark of the target lead conversion rate to be around 80%.

Solution Approach:

1. **Reading and Understanding Dataset:**
    - There are 9240 rows and 37 columns in the given dataset.
    - Object, integer and Float are the most common data types in the given dataset.
    - Check the description of the data.
2. **Data cleaning:**
    - Check the null values and duplicate values in the given dataset.
    - There are no duplicate values in the given dataset.
    - Columns with >40% nulls were dropped.
    - Value counts within categorical columns were checked to implement appropriate actions like dropping columns that don't have any values, imputing the columns, if imputation causes skew, then columns are dropped.

- Various tasks such as handling outliers, rectifying invalid data, consolidating infrequent values, and converting binary categorical values were performed.
- The data was checked for imbalance, revealing a conversion rate of only 38.5% for leads.
- Outlier analysis has been conducted on numerical data and 'Total visits' and 'Page Views Per Visit" Variables contain outliers.
- Low frequency values of Lead Source and Last Activity variables are grouped together to avoid unnecessary columns when dummy variables are created.
- 'Do Not Email' and 'Free_Copy' variables are binary categorical variables and both of the them mapped from yes/no to 1/0.

3. EDA:

- Univariate and bivariate analysis are performed on categorical and numerical variables. 'Lead origin', 'Lead Source', 'Current Occupation', 'Do Not Email', etc. provide valuable insight on effect on target variable.

- Univariate Analysis Insights:

  Lead Origin: "Landing Page Submission" identified 53% customers, "API" identified 39%.

  Lead Source: 58% Lead source is from Google & Direct Traffic combined

  Current_occupation: It has 90% of the customers as Unemployed

  Do Not Email: 92% of the people has opted that they dont want to be emailed about the course.

  Last Activity: 68% of customers contribution in SMS Sent & Email Opened activities

- Bivariate Analysis Insights:

  Lead Source: Google has Lead Conversion Rate of 40% out of 31% customers, Direct Traffic contributes 32% Lead Conversion Rate with 27% customers which is lower than Google,Organic Search also gives 37.8% of LCR but the contribution is by only 12.5% of customers ,Reference has LCR of 91% but there are only around 6% of customers through this Lead Source.

Lead Origin: Around 52% of all leads originated from "Landing Page Submission" with a lead conversion rate (LCR) of 36%.The "API" identified approximately 39% of customers with a lead conversion rate (LCR) of 31%.

Current_occupation: Around 90% of the customers are Unemployed with lead conversion rate (LCR) of 34%. While Working Professional contribute only 7.6% of total customers with almost 92% lead conversion rate (LCR).

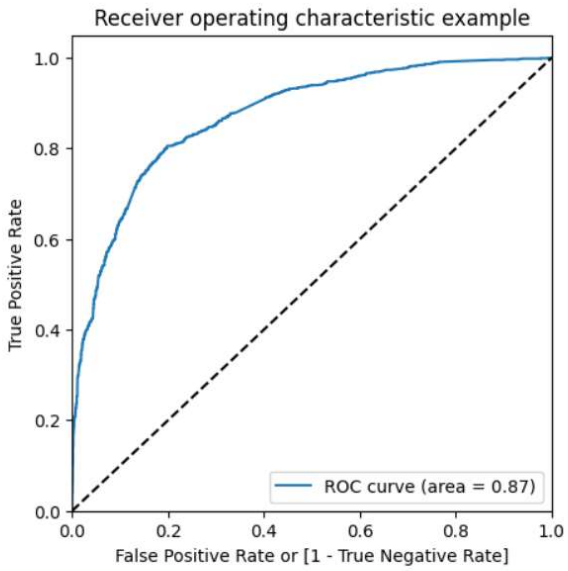
Do Not Email: 92% of the people has opted that they dont want to be emailed about the course.

Last Activity: 'SMS Sent' has high lead conversion rate of 63% with 30% contribution from last activities, 'Email Opened' activity contributed 38% of last activities performed by the customers with 37% lead conversion rate.

Specialization: Marketing Management, HR Management, Finance Management shows good contribution.

4. **Data Preparation:**
    - Created dummy variables for categorical variables like ‘Lead origin’, ‘Last Activity’, ‘Specialization’, ‘Current_occupation’.
    - The dataset was divided into training and testing sets using a 70:30 ratio.
    - Standardization was utilized for feature scaling.
    - Lead Conversion Rate (LCR) is 38.5 indicates a conversion rate of 38.5%.
    - ‘Lead Origin_Lead Import’, ‘Lead Origin_Lead Add Form’ variables are dropped as they were highly correlated.
5. **Model Building:**
    - Used REF for feature selection to reduce variables from 48 to 15, there by enhancing the manageability of the data frame.
    - Manual Feature Reduction process was used to build models by dropping variables with p-value >0.05.
    - Total 4 models were built. Model 4 was stable with (p-values <0.05). No sign of multicollinearity with VIF < 5.
    - Lrm4 was selected as final model with 12 variables, we used it for making prediction on train and test set.

6. **Model Evaluation:**
    - Confusion matrix was implemented and optimal cutoff point was selected as 0.345 based on accuracy, sensitivity, specificity plot. This cut off gave accuracy, specificity and precision all around 80% and precision recall view gave less performance metrics around 75%.
    - To address the business problem of achieving an 80% conversion rate, the CEO's directive, when focusing on precision-recall metrics, led to a drop in overall metrics. Consequently, we opted for a sensitivity-specificity perspective to determine the optimal cutoff for our final predictions.
    - Lead score was assigned to train data using 0.345 as cutoff.
7. **Predictions on Test Data:**
    - Final model(model-4) is used on test data for scaling and prediction.
    - ROC (Receiver Operating Characteristic) curve of test data is 0.87



NOTE: Area under ROC curve is 0.87 out of 1 which indicates a good predictive model

- For Test data Accuracy: 80.32%, Sensitivity: 80%, Specificity: 80.68%.

  These metrices are very close to training data metrices.
- Lead score was assigned.

- Top 3 feature are:

  Lead Source_Welingak Website

  Lead Source_Reference

  Currecnt_occupation_Working Professional

8. **Recommendation:**
   - By observing the results, most of the target leads are unemployed. So, concentrate more on unemployed target leads.
   - Working Profession are another potential target leads as they have high conversion rate.