Summary

- X Education is an online education company that focuses on offering courses to industry professionals. The company draws in a considerable number of professionals to its website, where they can explore and browse the available courses. Upon filling out a form with their email address or phone number, these individuals are categorized as leads
- 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 preprocessing:

- Dropping unnecessary columns, those with one unique value and zero variance, which cannot contribute to predicting the lead case. Additionally, columns with high missing values were removed.
- Replacing "Select" values throughout the data frame with "unknown" This adjustment is made because customers might not have selected any option from the list, and the default 'Select' was retained in such cases, like Unknown values. It is replaced with unknown so that the important column should not be dropped while dropping column based on missing value threshold.
- Imputation for numerical columns involved replacing null values with the median due to the presence of outliers in the datasets. For numerical categorical columns, null values were replaced with the mode.
- Other activities such as treating outliers, rectifying invalid data, grouping low-frequency values, and mapping binary categorical values were undertaken as part of the data preprocessing.

Exploratory Data Analysis:

- Checked for data imbalance, and found that only 38.5% of leads converted
- Conducted univariate and bivariate analysis for both categorical and numerical variables. Variables such as 'Lead Origin,' 'Current Occupation,' and 'Lead Source' offer valuable insights into their impact on the target variable
- Observed a positive impact on lead conversion associated with the time spent on the website

Data Preparation:

- For categorical variables with multiple levels, creating dummy features (one-hot encoded)
- The data is segregated into the target variable ("Converted") and the predicted variables, and it is stored in different variables (all columns except "Converted")
- The dataset was divided into training and test sets using a 70:30 ratio
- Feature scaling was applied through standardization.
-

Model Building:

- Initiated the construction of the linear regression model by incorporating the remaining independent variables post data preprocessing. Examined the summary statistics for each variable to assess their significance in predicting hot leads.
- Used RFE to reduce variables from 48 to 15. This will make data frame more manageable
- The models were constructed using a Manual Feature Reduction process, which involved eliminating variables with a p-value greater than 0.05. Additionally, multicollinearity among independent variables was assessed using the Variance Inflation Factor (VIF) score and correlation matrix.

- The Logm6 model, comprising 16 variables, was chosen as the final model. It was utilized for predictions on both the training and test sets. The evaluation of its performance, indicated by an area under the curve in the ROC curve, yielded a value of 0.96
- The optimal cutoff point of probability, set at 0.33, was determined by plotting the trade-off between sensitivity and specificity, as well as the trade-off between precision and recall

Model Evaluation:

- A confusion matrix was constructed, and a cut-off point of 0.33 was chosen, guided by the accuracy, sensitivity, and specificity plot. This cut-off yielded values of 0.90 for accuracy, specificity, and precision collectively, surpassing the 80% conversion rate, thereby addressing the business problem as per the CEO's objectives.

- The dependent variable value was predicted as per the above threshold values of conversion probability and lead score.

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 0.9
- Lead score was assigned based on the probability value predicted on test set.
- Based on our model, some features are identified which contribute most to a Lead getting converted successfully.
- The conversion probability of a lead increases with increase in values of the following features in descending order:

- Features with Positive Coefficient Values

  a. Tags_Lost to EINS
  b. Tags_Closed by Horizzon
  c. Tags_Will revert after reading the email
  d. Lead Source_Welingak Website
  e. Last Activity_SMS Sent
  f. What is your current occupation_Working Professional
  g. What is your current occupation_Unemployed

- The conversion probability of a lead increases with decrease in values of the following features in descending order:
  a. Features with Negative Coefficient Values
  b. Tags_switched off
  c. Tags_Ringing
  d. Tags_Already a student
  e. Tags_Not doing further education
  f. Lead Quality_Worst
  g. Tags_opp hangup
  h. Tags_Interested in full time MBA
  i. Tags_Interested in other courses
  g. Asymmetrique Activity Index_03

Recommendations:

To increase our Lead Conversion Rates:

- Focus on features with positive coefficients for targeted marketing strategies.
- Develop strategies to attract high-quality leads from top-performing lead source
- Engage working professionals with tailored messaging
- More budget/spend can be done on Welingak Website in terms of advertising, etc.
  Incentives/discounts for providing reference that convert to lead, encourage providing more references.
- Working professionals to be aggressively targeted as they have high conversion rate and will have better financial situation to pay higher fees too.