Summary

Problem Statement

1. X Education is a platform which sells online courses to industry professionals.
2. People on this platform fill up a form providing their email address or phone number, they are classified to be a lead.
3. Lead Conversion Rate(LCR) = Total Students acquired/Total Leads Produced
4. LCR for X education is very poor(~30%)
5. In order to increase the LCR. X education needs to identify the most potential leads(Hot Leads) i.e. which have a very high chance of getting converted.

Approach

1. **Getting and Inspecting the data**.
2. **Missing value treatment** - Removed the columns missing values more than 45% of the values missing.Imputed the numerical missing values with the median value and for categorical variables after performing EDA imputed them with the highest occurring category or creating a separate category as ‘na_categoryname’.
3. **Performed EDA on the features, Insights Gathered**:
    a) **Lead origin**: Max. Conversion for Landing Page Submission Lead Add Form showing ~8% per share and ~19% Con rate.
    b) **Last Activity** : Max. Conversion for SMS sent.Trend reverses for SMS sent and Email opened going from percentage share to conversion rate.

c) **Specialisation :** Conversion Rate Highest for Finance Management.Finance and Marketing Management have very close conversion rates.
d) **Tags :** Highest Con. Rate for Will revert after reading the mail.Ringing has very high per share but the con rate is very low.
e)Even though X education is for working professionals. Max. users who are converting are Unemployed.
d) India is the country with max. Conversion rates.
f) **Lead Source**:Google has the highest Conversion rate. Reference has very low per share but good con rate.

4. **Modelling**
    a) **Bottom-Top Approach** : Built 3 three models,dropped after that as the accuracy wasn't coming good and the features were in high numbers.
    I. Numerical features: 'TotalVisits', 'Total Time Spent
    II. Categorical features 'Lead Origin', 'CurrentOccupation', 'Specialization', 'Lead Source'.
    Final Model Specifications:
    Feature Scaling
    One Hot Encoding
    Oversampling using SMOTE
    Accuracy 81%
    Sensitivity 74%

    Bottom up approach was showing improvement in results by every iteration but the overall approach was time consuming and results were not significant.

b) **Top-Bottom Approach** : Using Recursive feature elimination to select the final 14 features from all the features.
Final Model Specifications:
Feature Scaling
One Hot Encoding
Oversampling using SMOTE
Accuracy - 88.37%
Sensitivity - 89.07%
Specificity - 87.91%
Probability Threshold - 0.4

Gave scores to all the leads on the basis of probability. A score greater than 40 is a hot lead.A score less than 40 is not a good lead.