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%.

EDA:

- Data imbalance checked- only 38.5% leads converted.
- Performed univariate and bivariate analysis for categorical and numerical variables.

Data Cleaning:

- Features having more than 40% nulls were dropped. Except for 'lead quality' variable, this feature had 51% null value, which was made into a new category, since that provided good information the lead.
- Many flag columns had 100% no values, which will be of no use to the modelling. These columns where dropped after analysis.
- Categorical column like City and Country had around 30% null values, more than 95% of the values belonged to one category, imputing nulls with mode will skew the data and is not useful for making classification. So these columns where dropped.
- Tags column was dropped, since it has sentences as data and lot of null, which is hard to impute and the rows cannot be dropped.
- Numerical data had outliers which was capped using Inter quantile range.

Data Preparation:

- Created dummy features for categorical variables
- Splitting Train & Test Sets: 70:30 ratio
- Feature Scaling using MinMax Scaler

Model Building:

- Build an initial Logistic regression model with all features, there were many features which was highly correlated and had high p-values.
- Used RFE to reduce variables from 75 to 15. To only have most important features and reduce the collinearity between features.
- 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, at each stage single feature was dropped to reduce VIF and p-value.
- The final model has 13 features and all features have VIF< 5 and low p-values <0.05.
- Train metrics accuracy, Sensitivity and Specificity was calculated for different cutoff and the optimal cutoff of 0.39 with tradeoff between these metrics is selected.

Model Evaluation:

- This cutoff was used to make prediction using the probabilities obtained from logistic regression model.
- Both train and test accuracy were 85%
- Sensitivity and Specificity for both test and train are more than 80%
- Choose sensitivity-specificity view for our optimal cut-off for final predictions, this approach aligns with the problem statement. We can adjust the cutoff based on what is important to the company, on correctly predicting 1's or 0's or have a tradeoff between them.

Feature importance:

- The most important features for making predictions were 'LeadQuality_Worst', 'Total Time Spent on Website' and 'LeadSource_welingak website'
- If the lead is a working professional, they are likely to get converted.

Recommendations:

- More budget can be done on Welingak Website in terms of advertising, and focus more on olark chat too.
- Employee Intuitions are important if the lead will get converted or not, the lead quality should be entered by employees.
- Working professionals should be targeted as they have high conversion rate and will have better financial situation to pay higher fees.
