

# 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:**

- Columns with more than 35% nulls were dropped. 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 outlier treatment, grouping low frequency values, mapping binary categorical values were carried out.

## **EDA:**

- Data imbalance checked- only 37.8% 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.
- Performed multivariate analysis for numerical variable. Time spend on website shows positive impact on lead conversion.
- Dropped few columns that were not adding value for building a good model.

## **Data Preparation:**

- Converted some binary variables (Yes/No) to 1/0.
- Created dummy features for categorical variables.
- Splitting Train & Test Sets: 70:30 ratio.
- Feature Scaling using Standardization.

## **Model Building:**

- Used RFE to reduce variables from 35 to 20. This made data frame more manageable.
- Manual Feature Reduction process was used to build models by dropping variables with  $p$ -value  $> 0.05$ .
- Total 7 models were built before reaching final Model 8 which was stable with ( $p$ -values  $< 0.05$ ). No sign of multicollinearity with  $VIF < 5$ .

- 'logm8' was selected as final model with 14 variables, we used it for making prediction on train and test set.

### **Model Evaluation:**

- Confusion matrix was made and cut off point of 0.35 was selected based on accuracy, sensitivity and specificity plot. This cut off gave accuracy, specificity and precision all around 80%. Whereas precision recall view gave less performance metrics around 75%.
- 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 0.35 as cut off.

### **Making Predictions on the Test set:**

- 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:
  - Lead Source\_Welingak Website
  - Lead Source\_Reference
  - Last Activity\_SMS Sent

### **Recommendations:**

- More budget/spend can be done on Welingak Website and Google in terms of advertising, etc.
- Incentives/discounts for providing reference that convert to lead, encourage to provide more references.
- Those we spend more time on the website should be aggressively targeted as they have high conversion rate.