Summary

X Education generates a lot of leads. But the lead conversion rate is very poor(Only 30%). The company needs us to build a

model wherein every lead must be assigned a score such that the ones with a higher lead score will have a higher chance of conversion. The target for lead conversion rate is approximately 80%.

1) Data Cleaning:

* Columns with more than 35% nulls were dropped. Value counts within categorical columns were checked to decide the appropriate action to be done.
* Numerical categorical data were imputed with mode .
* Outliers’ treatment, fixing invalid data, mapping binary categorical values were carried out.

EDA:

* Data imbalance checked- only 38.5% 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.

* Time spend on website shows positive impact on lead conversion.

Data Preparation:

* Converting Yes/No to Binary variables and correcting datatype of Page Views Per Visit and TotalVisits.
* Creating dummy variables for the 8 categories and dropping the first level.
* Splitting the data in Train and Test sets.
* Feature Scaling of numerical data.
* Initial conversion rate was found to be 38.45%.

Model Building:

* Used RFE for feature elimination from 73 to 20. This will make the data frame more manageable.
* A Manual Feature Reduction process was used to build models by dropping variables with p–value> 0.05.
* Model 5 which was stable with (p-values < 0.05) and No sign of multicollinearity with VIF < 5 was selected
* X\_train\_5 & reg5 were selected as a final model with 18 variables, we used them for making predictions on the train and test set.

Model Evaluation:

* ROC curve was plotted and found the AUC score to be 0.89 which is good enough.
* Also found the probability cutoff point of 0.36, which was selected based on accuracy, sensitivity and specificity. This cut-off gave accuracy, specificity and precision all around 80%.
* Chose a sensitivity-specificity view for our optimal cut-off for final predictions
* Lead score was assigned to train data using 0.36 as cut-off.

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 80%.● Lead score was assigned.
* Top 3 features are:
  + Total Time Spent on Website
  + Lead Origin\_Lead Add Form
  + What is your current occupation\_Working Professional

Recommendations:

* More budget/spend must be done on Websites for advertising, promotion etc.
* Incentives/discounts should be provided for referencing new hot leads.