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%.

For this, we have proceeded with the basic analysis of the given data set.

- Identifying the columns based on Data Dictionary.
- Removed records with > 40% missing data
- Identifying the relationship and distribution of column data using graphs
- outliers' treatment.
- mapping binary categorical values were carried out

After performing data analysis, we performed EDA (Exploratory Data Analysis) in which we analysed categorical and numerical variables. 'Lead Origin', 'Current occupation', 'Lead Source', etc. to provide valuable insight on effect on target variable.

Later, we proceeded with encoding the categorical data into Dummy variables so that we can easily convert them into features which can be fed into a Model used for predictions. For that we split the data into train and test set in the 70:30 ratio. Feature Scaling using MinMax Scalar was performed and dropped few columns, they were highly correlated with each other.

After creating dummy variable, we used RFE to reduce the variable to15 to make our data frame more manageable. Manual Feature Reduction process was used to build models by dropping variables with p – value > 0.05. Total 3 models were built with p-values < 0.05 and No sign of multicollinearity with VIF < 5. logm3 was selected as final model which, we used it for making prediction on train and test set.

At last, we made prediction the test data using final data and we have also observed more than 80% accuracy & precision there as well. And found top 3 features (Total Time Spent on Website, Lead Origin_Lead Add Form, Lead Source_Welingak Website)