Summary Report

In this case study our aim was to analyze and identify variables which leads to higher lead conversion rate for an online education company. It was progressed further by building a logistic regression model which can assign a lead score to each of the leads suchthat customers with high lead score have a higher chance of conversion.

Data cleaning

we dropped columns in which 45 or more percent of the data were missing. We imputed the missing values in continuous variables with median and those in categorical variables with mode. However, few categorical variables had missing value count more than that of mode, in such cases we replaced the missing values with 'Not mentioned'. In few of the categorical columns we had highly varying counts for levels, so we grouped less frequent levels together to avoid model picking on the highly underrepresented dummy levels. We handled outliers in continuous variables by capping them at 99th percentile.

EDA

Both numerical and categorical variables were plotted to see their trend. Many levels in categorical variable 'Tags' seemed to have higher conversion rate.

Customers who were tagged as 'will revert after reading the email' or 'Closed by Horizzon' were more likely to purchase the courses. Leads in 'Lost to EINS seemed to have a very high lead conversion rate. Leads who are tagged as 'Switched off','Ringing' and 'Already a student' were unlikely to purchase the course materials.Our target variable 'Converted' had a positive correlation with 'Total Time Spent On Website' with a correlation coefficient of 0.36.The more a lead spends time on X Education's website the more is the chances of lead conversion.Other numerical variables didn't seem to have any strong correlation with our target variable 'Converted'.

Model

We created dummy variables for categorical variables. Our data frame was split into train and test sets. Numerical variables were normalized so that they wouldn’t have any disproportionate effect on the model’s results. We had 54 features, since it was not feasible to use all the variables, we eliminated a few using RFE and selected 20 for Model building. We built a model using these features, In that model one feature was insignificant. we dropped that feature and rebuilt the model in which all the features were statistically significant and had VIF values below 2.Variables which helps most in predicting lead conversion in our model are:Tags_Lost to EINS (coef 6.3616), Tags_switched off (coef -5.7080), Tags_Closed by Horizzon (coef 5.5389), We used this model as our final model.Using logistic regression curve’s ‘predict’ function we computed the probabilities of lead conversion. We found that the optimum cut off value is 0.3, with that we classified the leads as converted or not. Then we evaluated our model’s predictions. Our model has sensitivity and specificity around 93%. We computed a lead score for every lead in the data set so that Leads with very high lead scores belong to hot leads, who are most likely to be converted.