Summary

The following analysis is performed for X Education company to find ways to get more industry professionals to join their courses. The company gets a lot of leads however its lead conversion rate is very poor at around 30%. The company requires us to build a model to achieve the lead conversion rate of around 80%. Apart from conversion rate the dataset provided also gave insights on how the potential customers visited the website, how they reached the site, how much time they spent on the website.

Data Cleaning:

- The data provided had no duplicates but had quite a few null values. There was a 'select' level which was replaced by null values because it was an option left blank by the customer.
- Columns with >40% nulls were dropped. Value counts within categorical columns with skew were dropped and for rest created new category (others), imputed them with high frequency value and dropped the columns that didn'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 outliers' treatment, fixing invalid data, grouping low frequency values, mapping binary categorical values were carried out.

EDA:

- Data imbalance checked- only 38.5% leads converted.
- Performed univariate, bivariate and multivariate analysis for categorical and numerical variables. Quite a few categorical variables were irrelevant.
- TotalVisits' and 'Page Views Per Visit' showed high correlation.

Data Preparation:

- Binary mapped a column and Created dummy features (one-hot encoded) for categorical variables
- Splitting Train & Test Sets: 70:30 ratio
- Feature Scaling using MinMax scaler
- Dropped high correlation columns.

Model Building:

- RFE and manual feature elimination both techniques were used to reduce variables.
- RFE reduced variables from 48 to 15.
- Model 5 (logmod5) was selected as it has all features with p – value < 0.05 and VIF > 5. It had 11 features.

Model Evaluation:

- Confusion matrix was made and the cut off point of 0.36 was selected based on accuracy, sensitivity and specificity plot. This cut off gave accuracy, specificity and precision all around 80%. The precision recall also gave 80% results.

Prediction on Test Data:

- Prediction was done on the test data frame with an optimum cut-off as 0.36 with accuracy, sensitivity and Specificity of 80%.

Precision-Recall:

- The method was also used to recheck and a cut-off of 0.42.

Conclusion:

Following are the top 5 variables for potential hot leads:

- Total Time Spent on Website
- Lead Origin_Lead add form
- Lead Source_Welingak website
- Last Notable Activity_Sms sent
- Lead Source_Olark chat

Recommendations:

- Phone calls to be made to those who spend more time on website
- More advertisements can be done on the Welingak Website and Olark chat as it attracts more customers.
- Customers with lead add from and whose last activity was sms sent can be tracked for calling.