

Summary

X education gets a lot of leads, its lead conversion rate is 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 >40% nulls were dropped. Values counts within categorical columns were checked to decide appropriate action: if imputation causes skew, then column was dropped, created new category, impute high frequency value, drop columns which don't add any value
- Numerical categorical data were imputed with mode and columns with only one unique response from customers 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 and bivariate analysis for categorical and numerical variables. 'leads origin' , 'lead source' , 'current occupation' etc. provide valuable insight on effect on target variable
- Time spend on website shows positive impact on lead conversion

Data Preparation:

- Created dummy features for categorical variables
- Splitting Train & Test sets: 70:30 ratio
- Feature scaling using standardization
- Dropped few columns, they were highly correlated with each other

Model Building:

- Used RFE to reduce variables from 48-15. This will make data frame more manageable
- Manual feature reduction process was used to build models by dropping variables with p-value >0.05
- Use RFE for feature selection
- Overall accuracy 81%

Data Conversion:

- Numerical Variables are normalised
- Dummy Variables are created for object type variables
- Total rows for analysis 8792
- Total columns for analysis 43

Conclusion:

It was found that the variables that mattered the most in the potential buyers are:

- The total time spend on the website
- The total number of visits
- When the lead source was: Google, Direct Traffic, Organic Search, Welingak Website
- When the last activity was: SMS, Olark Chat Conversation
- When the lead origin is Lead add format
- When their current occupation is as a working professional. Keeping these in mind the X education can flourish as they have a very high chance to get almost all the potential buyers to change their mind and buy their courses