SUMMARY

X Education – Case Study on Lead Scoring

Team members: Priyadarsini, Priyanka Raut and Sandhya K.

X Education gets a lot of leads, its lead conversion rate is very poor at around 37%. 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%.

Original dataframe and EDA gave us a lot of information about how the potential customers visit the site, the time they spend there and how it is related to the conversion rate. We adopted following steps broadly in this case study.

1. Data Cleaning:
    - Original df had no duplicates so no action
    - Columns with >40% nulls were dropped.
    - For <40% missing values, value counts within categorical columns were done for imputation. Where imputation causes skew, then column was dropped, created new category (others), impute high frequency value, drop columns that don't add any value.
    - Numerical categorical data were imputed with mode and columns with only one unique response from customer were dropped.
    - Outliers' treatment was done by capping few variables to 5% and 95% percentiles.
2. EDA:
    - Target variable "Converted" was checked for data imbalance- approx. 37% leads converted.
    - Univariate, bivariate analysis was done for categorical and numerical variables. Correlation Heatmap was created to check for highly corelated variables. 'Lead Origin', 'Current occupation', 'Lead Source', etc. provide valuable insight on effect on target variable.
3. Data Preparation:
    - Mapping binary categorical values were carried out
    - The dummy variables were created for categorical variables.
    - For numeric variables we used the MinMax Scaling.
4. Train Test Split: Dataset split was done at 70% and 30% for train and test data respectively.

THANK YOU

5. Model Building:
    - RFE was done to select the top 20 relevant variables.
    - Rest of the variables were removed one by one manually depending on the p-value (p-value < 0.05 were kept)
    - Finally the rest of the variables were removed one by one manually depending on the VIF (variables with VIF < 5 were kept) parallelly checking their p-values
    - logm9 was the 9^{th} and our final model which had 12 variables.
6. Model Evaluation:
    - A confusion matrix was made with arbitrary 0.5 cut-off value and got the accuracy and specificity which came to be around 80% each. Whereas sensitivity was less around 70%.
    - The optimum cut off value (using ROC curve) was arrived at which was 0.34 and got accuracy, specificity and precision all around 80%. Whereas sensitivity improved to 82%.
    - Lead score was assigned to train data using 0.34 as cut off.
7. Predictions on Test Data:
    - Prediction was done on the test data frame and with an optimum cut off as 0.34 with accuracy, sensitivity and specificity of 80%.
    - Lead score was assigned to test data using 0.34 as cut off.
    - Top 5 features are:
        i. Lead Source_Welingak Website
        ii. Total Time Spent on Website
        iii. Lead Source_Reference
        iv. Current_occupation_Working Professional
        v. Last Activity_SMS Sent

Recommendations:

- The company should make calls to the Hot leads:
    - coming from the lead sources "Welingak Websites" and "Reference" as these are more likely to get converted.
    - who are the "working professionals" as they are more likely to get converted.
    - who spent "more time on the websites" and make more visits to websites as these are more likely to get converted.
    - coming from the lead sources "Olark Chat" as these are more likely to get converted.
    - whose last activity was SMS Sent as they are more likely to get converted.
- Working professionals to be aggressively targeted as they have high conversion rate and will have better financial situation to pay higher fees too.
- Have Omni-channel presence across all touchpoints: the sales team should make sure to follow up with leads who have interacted with X Education through multiple channels.
- Have Referral bonus scheme for their Alumini as well as their employees whose generate potential Hot leads to encourage Referrals.

THANK YOU