SUMMARY OF LEAD SCORING CASE STUDY

X Education company gets a lot of leads, lead conversion rate is around 30%. The company requires us to build a model where 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. Target for lead conversion rate is around 80%.

1. Data Cleaning and EDA

Handling missing data, Variable transformation, Data visualization etc are performed

2. Data Preparation

Data prepared for further analysis

3. Model Building and Evaluation Data Cleaning and EDA:

• Handling the select level. Dropping the columns with high percentage of missing values. Value counts within categorical columns were checked to decide appropriate action: if imputation causes skewness, 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.

• EDA: Data imbalance checked. Performed univariate and bivariate analysis for categorical and numerical variables. 'Lead Origin', 'Current occupation', 'Lead Source', 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 Model Building and Evaluation: • Used RFE to perform variable selection.

• Built a Logistic Regression Model.

• Manual Feature Reduction process was used to build models by dropping variables with p – value > 0.05.

• Final Model 4 was stable with (p-values < 0.05). No sign of multi collinearity with VIF < 5. logm4 was selected as final model with 13 variables and used it for making prediction on train and test set.

• Confusion matrix was made and cut off point of 0.34 was selected based on accuracy, sensitivity and specificity plot. This cut off gave accuracy, specificity and precision all around 80%. Whereas precision recall view gave less performance metrics around 75%.

● Lead score was assigned to train data using 0.34 as cut off.

Top 3 features are:

1. Lead Source_Welingak Website

2. Lead Source_Reference

3. Current_occupation_Working Professional Observations/Recommendations:

• Strategies to be developed to attract high-quality leads from top-performing lead sources. • Optimize communication channels based on lead engagement impact.

• More budget can be spent on Welingak Website in terms of advertising, etc.

• Incentives/discounts for providing references that convert to leads.

• Working professionals to be targeted as they have high lead conversion rate