By Priyadharshni S M: priyadharshnism1@gmail.com/9655730893
By Taher Chitalwala: mazahir3393@gmail.com / 8082437710

Problem Statement:

Evaluate the entire data sets of X Education to identify factors that lead to higher lead conversion. Multiple data sources are available with the Company; however, the lead conversion is very poor at 30%. The objective is to improve the lead conversion rate to ~80%

Overall Approach:

1. Exploratory data analysis was performed to understand the data
2. The data were reviewed for errors and inconsistencies. Missing values were treated in line with the methodology provided below
3. Three versions of the model were prepared.
4. The model was reviewed on the training and test set for accuracy

Data Preparation:

1. The shape, size, and stratification of the population were reviewed before proceeding
2. EDA was performed by distribution plots, Pair plots, and Other charting methods
3. Lead Sources, Channels were reviewed to see any uneven conversions
4. Data was reviewed for cleaning for null and missing values.
5. Numerical values were added through the Median function
6. Categorical variables were corrected by using the mode function
7. Blue represents failed and Orange represents successful conversions across charts

Model Building Approach:

1. A single model was first built to review all variables together
2. Recursive feature elimination was used to select the top 15 features
3. The variables selected by RFE were not skewed or concentrated
4. The categorical variables were additionally reviewed through EDA stated above
5. Post a review of the RFE results, variables were eliminated based on statistical significance and VIF

Model Testing and Evaluation:

1. The ROC curve demonstrated a strong model @ 93%
2. Various probabilities were considered as a cut-off for review
3. 20 probability scenarios were reviewed with different performance matrices
4. Based on the performance matrices, the optimum cut-off stood between 0.35 to 0.4
5. The results of the training matrix stood as follows:

Matrix for Performance	Training Set	Test Set
Overall Accuracy	86.5	85.8
Sensitivity/ Recall	84.47	85.11
Specificity	87.76	86.29
False positive rate	12.24	13.71
Positive Predictive rate/ Precision	80.96	80.21
Negative Predictive rate	90.17	89.88

Conclusion:

1. Data cleaning has been performed using fairly moderate assumptions.
2. Not all variables dropped due to redundancy.
3. RFE has factored to exclude redundant variables observed within the EDA section
4. The ROC curve and model parameters appear sound and logical
5. Accuracy, sensitivity, specificity, & positive predictive value are within acceptable levels.

Learnings and the Way forward:

1. Logistical and Linear Regression is key when it comes to predictive analytics
1. Pandas includes the relevant features to create the framework to complete a thorough modeled logistical regression
2. Priorities of the business user are always at the forefront. The results need to be revie wed after the inclusion of relevant business feedback.
3. The concept has strong applications in functions having binary decisions. For exampl e, the leads case, sales, and marketing for effectiveness, Banking and Credit profiling , etc.