SUMMARY

In this project we focussed on identifying the model for identifying the model and based on that a strategy for identifying the most probably conversion leads. This involved data cleaning, feature engineering, and model creation and evaluation.

Data Cleaning:

We verified the quality of our dataset. We addressed missing data issues by removing Columns with more than 40% missing values from the dataset as imputing them might lead to skewed results. This step streamlined our dataset and eliminated variables with significant data gaps.

We added mode for categorical features and mean for numerical features for the remaining features.

But Tags feature was an exception since it had some information which had impact on the results so decided to keep it.

EDA:

1. Base don EDA we found Tags, Lead Origin, Lead Source had specific values that contributed to the Values. Box plot of these variables showed skew towards the Conversion.
2.

Feature Engineering:

1. One-Hot Encoding for Categorical Variables: Categorical variables, such as lead source, Origin etc were converted into numerical format using one-hot encoding.
2. We used Standard Scaler for Numerical variable to reduce the impact of Outliers affecting the results.
3. Using Recursive Feature Elimination selected the features that helped in improving the model with least number of features selected the top 15 features.
    a. The selected features were
    b. 'Lead Source_Welingak Website'
    c. 'Last Activity_Email Opened',
    d. 'Last Activity_SMS Sent',
    e. 'Tags_Closed by Horizzon',
    f. 'Tags_Diploma holder (Not Eligible)',
    g. 'Tags_Interested in full time MBA',
    h. 'Tags_Interested in other courses',
    i. 'Tags_Lost to EINS',
    j. 'Tags_Not doing further education',
    k. 'Tags_Ringing',
    l. 'Tags_Will revert after reading the email',
    m. 'Tags_invalid number',
    n. 'Tags_number not provided',
    o. 'Tags_switched off',
    p. 'Tags_wrong number given'

Data Splitting:

Before moving on to model creation and evaluation, we divided our dataset into training and testing sets70-30 split. The training set was used to train our models, while the testing set allowed us to evaluate their performance on unseen data.

Model Creation:

We employed two primary machine learning models for lead scoring: Logistic Regression and Decision Trees.

Logistic Regression:

We trained a Logistic Regression model on our training data. We tried 4 different models with different parameters. The best model got accuracy of 91% and the gap between train and test was 1% which is really good.

Most important features

1. Tags_Closed by Horizzon

2. Tags_Lost to EINS

3. Tags_Will revert after reading the email

But since some Tags seems like in the progress so we use the Decision trees to find the features other than Tags

Model Evaluation:

Accuracy : 90.12%

Sensitivity : 89.41%

Specificity : 90.58%

Decision Trees:

Decision Tree with depth of 8 was used for the model. And this gives features importance for other than Tags, Lead Source, Lead Origin and Free Guide while Lead quality is just quantifiable value which combine all the Lead variables.

Important Features

- Lead Source_Organic Search

- Lead Source_Reference

- Lead Origin_Lead Add Form

- TotalVisits

- Total Time Spent on Website

Model Evaluation:

True Negative : 1598

True Positive : 944

False Negative : 151

False Positve : 79

Model Accuracy : 0.917

Model Sensitivity : 0.8621

Model Specificity : 0.9529

Model Precision : 0.9228

Model Recall : 0.8621

Model True Positive Rate (TPR) : 0.8621

Model False Positive Rate (FPR) : 0.0471

By comparing the results of both Logistic Regression and Decision Trees, we could identify common significant features and obtain a consensus view on their importance in lead scoring.

Feature Importance –

1. Lead Source – Referral, Facebook, Organic Search
2. Lead Origin - Add Form
3. Tags - Will revert after reading the email
4. Time Spent on Website, Number of Visits
5. A free copy of Mastering The Interview
6. Do Not Email