Lead Conversion Optimization: A Comprehensive Model Approach

INTRODUCTION

X Education needs help in selecting the most promising leads, i.e. the leads that are most likely to convert into paying customers. The company requires you to build a model wherein you need to assign a lead score to each of the leads such that the customers with a higher lead score have a higher conversion chance and the customers with a lower lead score have a lower conversion chance.

- **Data Preparation and Cleaning**

  Elimination of Columns with >30% Nulls to ensure data quality, columns with significant null values were removed to streamline the dataset. Even Select values are treated as null values.

- **Categorical Value Imputation Strategies**

  A meticulous approach to impute categorical values, considering actions such as dropping skewed columns, creating new categories, and imputing high-frequency values, was implemented.

- **Outliers' Treatment and Data Refinement**

  Activities such as outliers' treatment, invalid data fixing, low-frequency value grouping, and binary categorical value mapping were carried out to enhance data reliability.

Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) is a critical step in the data analysis process where analysts employ statistical and visual techniques to summarize and understand the main features of a dataset. By calculating descriptive statistics, exploring individual and paired variable distributions, and leveraging various visualization methods, EDA aims to reveal patterns, outliers, and potential relationships within the data. This comprehensive exploration forms the foundation for informed decision-making, subsequent modeling, and the extraction of meaningful insights from the dataset.

- **Assessment of Data Imbalance**

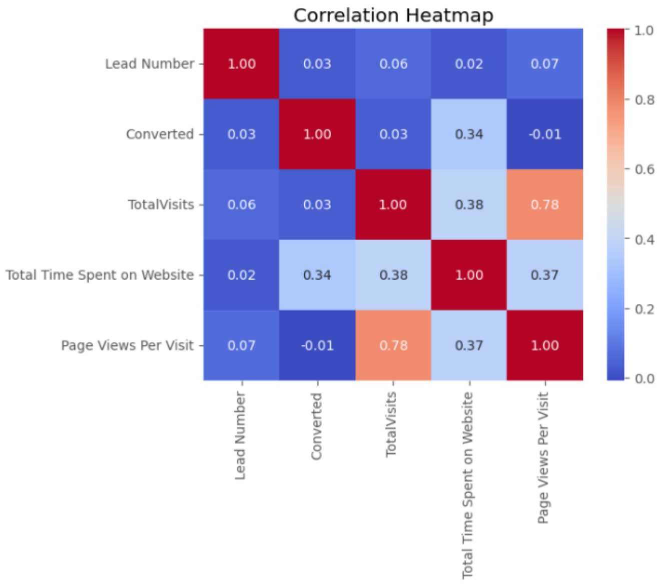
  EDA revealed an imbalance in lead conversions, with only 37.6% of leads resulting in successful conversions.

- **Univariate and Bivariate Analysis for Key Variables**

  Key variables such as 'Lead Origin,' 'Current Occupation,' and 'Lead Source' were subjected to thorough univariate and bivariate analysis, providing valuable insights into their impact on the target variable.

- **Positive Impact of Time Spent on Website on Lead Conversion**

  The analysis highlighted a positive correlation between the duration of time spent on the website and lead conversion.



Model Building

The selection of the Logistic Regression technique was driven by its suitability for the problem type, dataset characteristics, and overall objectives. By feeding the training dataset into this chosen algorithm, we aimed to enable the model to learn and make predictions or classifications based on the input features.

To enhance the efficiency of managing the dataset, Recursive Feature Elimination (RFE) was utilized. This technique systematically reduced the number of variables from an initial count of 33 to a more streamlined set of 15, thereby optimizing the model's performance and ensuring that it focuses on the most influential features for accurate predictions.

➢ Manual Feature Reduction

A manual feature reduction process was used to build models by dropping variables with p-value > 0.05. A total of 3 models were built before reaching the stable Model 3.

➢ Model Stability Check

The final Model 7, selected based on stability criteria (p-values < 0.05) and no signs of multicollinearity (VIF < 5), was labelled as logr7 with 9 variables and a constant.

Model Evaluation

Evaluation Metrics Alignment (Train & Test) at ~81%

The evaluation metrics for both the train and test sets closely aligned around the targeted 81%, indicating the model's robustness.

➢ For Train Set:

```
**************************************************

Confusion Matrix
[[3189  566]
 [ 591 1671]]

**************************************************

True Negative              :  3189
True Positive            :  1671
False Negative             :  591
False Positve             :  566
```

```
Model Accuracy                  :  0.8077
Model Sensitivity               :  0.7387
Model Specificity               :  0.8493
Model Precision                 :  0.747
Model Recall                    :  0.7387
Model True Positive Rate (TPR)  :  0.7387
Model False Positive Rate (FPR) :  0.1507


**************************************************
```

- **Lead Score Assignment to Test Data**

For Test Set:

```
**************************************************

Confusion Matrix
[[1382  225]
 [ 264  709]]

**************************************************

True Negative                   :  1382
True Positive                   :  709
False Negative                  :  264
False Positve                   :  225
Model Accuracy                  :  0.8105
Model Sensitivity               :  0.7287
Model Specificity               :  0.86
Model Precision                 :  0.7591
Model Recall                    :  0.7287
Model True Positive Rate (TPR)  :  0.7287
Model False Positive Rate (FPR) :  0.14


**************************************************
```

Consistent with the train data, lead scores were assigned to the test data, providing a holistic view of potential lead quality.

- **Top 5 Features Influencing Lead Conversion**

• Lead Origin_Lead Add Form	4.530120
• Total Time Spent on Website	4.422485
• What is your current occupation_Working Professional	2.753437
• Last Notable Activity_SMS Sent	1.348210
• Lead Source_Olark Chat	1.077277

Conclusion

It was found that the variables that mattered the most in identifying potential buyers are (in descending order):

1. **Lead Origin_Lead Add Form (4.530120):**

- The high positive score suggests that this particular lead origin has a strong positive impact on predicting lead conversion. Leads from this source are more likely to convert.

2. **Total Time Spent on Website (4.422485):**

- The high positive score indicates that the more time a user spends on the website, the more likely they are to convert. It suggests that user engagement on the website is a significant predictor of lead conversion.

3. **What is your current occupation_Working Professional (2.753437):**

- The positive score indicates that leads with a current occupation of "Working Professional" are more likely to convert compared to other occupations.

4. **Last Notable Activity_SMS Sent (1.348210):**

- The positive score suggests that sending an SMS as the last notable activity is positively correlated with lead conversion. Leads who received an SMS as the last notable activity are more likely to convert.

5. **Lead Source_Olark Chat (1.077277):**

- The positive score indicates that leads coming from Olark Chat are more likely to convert compared to other lead sources.

6. **Lead Source:** Certain lead sources significantly contribute to conversions, particularly:
   - Google
   - Direct traffic
   - Organic search
   - Visits through the Welingak website
7. **Last Activity:** The nature of the last activity performed by the potential buyer plays a crucial role, especially activities like:
   - SMS interactions
   - Olark chat conversations
8. **Current Occupation:** Individuals identified as working professionals are more likely to convert into buyers.
9. **What is your current occupation_Working Professional:** working professionals are more likely to convert into buyers.

Keeping these in mind, X Education can flourish as they have a very high chance to persuade potential buyers to change their mind and buy their courses.