Summary Report: Lead Scoring Project

The primary objective of this project was to develop a lead scoring model and formulate a strategy for identifying the most likely conversion leads. The project consisted of several key steps, including data cleaning, feature engineering, and model creation and evaluation.

I. Data Cleaning

The project commenced with a thorough evaluation of the dataset's quality. To address missing data issues, columns with more than 40% missing values were removed, as imputing those values could introduce bias. An exception was made for the "Tags" feature, which contained valuable information with a noticeable impact on the results.

II. Exploratory Data Analysis (EDA)

Initial exploratory data analysis (EDA) revealed that certain features, such as "Tags," "Lead Origin," and "Lead Source," had specific values contributing significantly to lead conversion. Box plots of these variables displayed a skew towards conversion.

III. Feature Engineering

1. OneHot Encoding for Categorical Variables: Categorical variables like "Lead Source" and "Lead Origin" were converted into a numerical format using onehot encoding.
2. Standard Scaling for Numerical Variables: Standard scaling was applied to numerical variables to mitigate the impact of outliers on the results.
3. Recursive Feature Elimination: This technique was used to select the most relevant features for improving the model's performance while minimizing the number of features. The top 15 selected features were:
    - 'Lead Source_Welingak Website'
    - 'Last Activity_Email Opened'
    - 'Last Activity_SMS Sent'
    - 'Tags_Closed by Horizzon'
    - 'Tags_Diploma holder (Not Eligible)'
    - 'Tags_Interested in fulltime MBA'
    - 'Tags_Interested in other courses'
    - 'Tags_Lost to EINS'
    - 'Tags_Not doing further education'
    - 'Tags_Ringing'
    - 'Tags_Will revert after reading the email'

- 'Tags_invalid number'
- 'Tags_number not provided'
- 'Tags_switched off'
- 'Tags_wrong number given'

IV. Data Splitting

Before proceeding with model creation and evaluation, the dataset was divided into training and testing sets using a 7030 split. The training set was used for model training, and the testing set facilitated the evaluation of the model's performance on unseen data.

V. Model Creation

Two primary machine learning models were employed for lead scoring: Logistic Regression and Decision Trees.

1. Logistic Regression:
   After training the Logistic Regression model on the training data, four different models with various parameters were explored. The best model achieved an accuracy of 91%, with a minimal 1% gap between the training and test data, indicating excellent performance. The most important features identified in this model were:
   - 'Tags_Closed by Horizzon'
   - 'Tags_Lost to EINS'
   - 'Tags_Will revert after reading the email'

   However, as certain "Tags" appeared to be in progress, Decision Trees were used to identify other significant features.

VI. Model Evaluation for Logistic Regression

- Accuracy: 90.12%
- Sensitivity: 89.41%
- Specificity: 90.58%

VII. Decision Trees

For Decision Trees, a tree depth of 8 was used. This model helped identify important features beyond "Tags," "Lead Source," "Lead Origin," and the "Free Guide." Lead quality was also quantified based on all lead variables.

Important Features from Decision Trees:

- Lead Source_Organic Search
- Lead Source_Reference
- Lead Origin_Lead Add Form

- TotalVisits
- Total Time Spent on Website

VIII. Model Evaluation for Decision Trees:

- True Negative: 1598
- True Positive: 944
- False Negative: 151
- False Positive: 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

Conclusion

By comparing the results of both Logistic Regression and Decision Trees, common significant features were identified, and a consensus view on their importance in lead scoring was obtained.

Feature Importance

- Lead Source: Referral, Facebook, Organic Search
- Lead Origin: Add Form
- Tags: Will revert after reading the email
- Time Spent on Website, Number of Visits
- A free copy of Mastering The Interview

This comprehensive approach allowed the development of a robust lead scoring model and provided valuable insights into the most influential factors driving lead conversion.