Summary Report: Lead Scoring Case Study

Objective

The goal/objective is to build a predictive model to assign lead scores that reflect the probability of lead conversion. A logistic regression model was used to determine the probability of lead conversion and assign scores to prioritize leads.

Steps Followed

1. **Data Import and Exploration**
   - The dataset consisted of 37 columns and 9240 rows.
   - Columns with more than **40% missing values** were dropped.
   - Missing values were handled by replacing with "Unknown" for categorical variables.
2. **Outlier Handling**
   - Outliers in numerical columns (TotalVisits, Total Time Spent on Website, Page Views Per Visit) were identified and capped/removed.
   - The final dataset contained **8991 rows and 29 columns**.
3. **Data Transformation**
   - Binary categorical variables (Yes/No) were converted to 0/1.
   - Rare categories in categorical variables were grouped as "Other" if they contributed less than 5% of data.
4. **Feature Engineering**
   - Dummy variables were created for multi-level categorical variables using **one-hot encoding**, dropping "Other" categories to avoid multicollinearity.
   - Resulting dataset had **54 features**.
5. **Train-Test Split**
   - 70% training and 30% test data split.
   - Standard scaling applied to numerical variables (TotalVisits, Total Time Spent on Website, Page Views Per Visit).
6. **Feature Selection Using RFE**
   - Recursive Feature Elimination (RFE) selected **15 features**.
   - Variables were further pruned based on multicollinearity using **VIF**.

Final Selected Variables:

- Do Not Email

- Total Time Spent on Website
- LeadOrigin_Lead Add Form
- LastActivity_Email Opened
- LastActivity_SMS Sent
- Country_Unknown
- CurrentOccupation_Working Professional
- CourseGoal_Unknown
- Tags_Already a student
- Tags_Interested in other courses
- Tags_Ringing
- Tags_Will revert after reading the email
- LastNotableActivity_Modified

7. **Model Building**
   - Logistic Regression was used, achieving the following results:
     Train Metrics:
     - Accuracy: **89.62%**
     - Sensitivity: **89.78%**
     - Specificity: **89.53%**
     - False Positive Rate: **10.47%**
     - Positive Predictive Value: **83.79%**
     - Negative Predictive Value: **93.56%**
8. **Threshold Optimization**
   - Optimal cutoff threshold identified at **0.35** using ROC curve and sensitivity-specificity trade-off.
9. **Model Evaluation on Test Data**
   - Predictions were made on the test data, and results were evaluated.

Test Metrics:

- Accuracy: **88.88%**
- Sensitivity: **84.69%**
- Specificity: **91.48%**
- False Positive Rate: **8.52%**

10. **Lead Score Assignment**
    - Lead scores were assigned by multiplying the predicted probabilities by 100.

Key Observations

1. **Top Variables Contributing to Lead Conversion** (based on coefficients and RFE selection):
    - **Total Time Spent on Website**
    - **LastActivity_SMS Sent**
    - **LeadOrigin_Lead Add Form**
2. **Model Performance**
    - The model achieved consistent accuracy, sensitivity, and specificity on both train and test data.
    - Sensitivity on test data (84.69%) meets the business requirement of identifying most potential leads for conversion.
3. **Lead Scores**
    - Higher probabilities (>35% threshold) are mapped to higher lead scores, enabling targeted follow-ups.

Conclusion

The logistic regression model is robust and generalizable. The final model performs well, achieving an accuracy close to **89%**. The lead scores can help prioritize leads, ensuring better resource allocation and improved conversion rates.

Train Data Metrics:

- Accuracy: **89.62%**
- Sensitivity: **89.78%**
- Specificity: **89.53%**
- False Positive Rate: **10.47%**
- Positive Predictive Value: **83.79%**
- Negative Predictive Value: **93.56%**

Test Data Metrics:

- Accuracy: **88.88%**
- Sensitivity: **84.69%**

- Specificity: **91.48%**
- False Positive Rate: **8.52%**

Important Insights

1. **Key Variables Impacting Lead Conversion**
   The following variables were identified as the **top contributors** to lead conversion probability based on model coefficients and feature importance:
   - **Total Time Spent on Website**:
     - Positively correlated with lead conversion. Higher time spent on the website increases the chances of conversion.
   - **LastActivity_SMS Sent**:
     - Leads who received an SMS were more likely to convert, highlighting the importance of SMS communication.
   - **LeadOrigin_Lead Add Form**:
     - Leads coming from the "Lead Add Form" origin were significantly more likely to convert, indicating its effectiveness as a lead generation channel.

Some Other Important Variables:

- **Do Not Email**: Negatively impacts lead conversion. Leads opting out of emails are less likely to convert.
- **Tags (Ringing, Unknown, Will revert after reading the email)**:
  - Specific lead tags act as strong indicators of conversion potential. Tags like "Will revert after reading the email" are highly predictive.
- **CurrentOccupation_Working Professional**:
  - Working professionals were found to have higher conversion probabilities compared to other occupations.

2. **Optimal Cutoff for Lead Conversion**
   - The **optimal threshold** for predicting lead conversion was identified as **0.35**.
   - This cutoff provides a good balance between **sensitivity (recall)** and **specificity**:
     - Sensitivity: **89.78%** (Train), **84.69%** (Test)
     - Specificity: **89.53%** (Train), **91.48%** (Test)
3. **Lead Conversion Trends**

- **Website Engagement**: Leads with higher Total Time Spent on Website and frequent website activity show a higher conversion likelihood.
- **Communication Channels**:
    - SMS notifications and specific follow-up activities (e.g., "Email Opened") positively influence conversions.
    - Leads opting out of emails are less responsive, indicating that **email opt-in rates** should be improved.
- **Tags & Follow-up**:
    - Tags indicating **interest or responsiveness** (e.g., "Will revert after reading the email") are critical for prioritizing leads.
- **Demographics**:
    - **Country_Unknown** and **CurrentOccupation_Working Professional** are key demographic indicators influencing conversion.

4. **Model Performance**
    - The model generalizes well across training and test datasets with **minimal performance drop**.
    - Metrics demonstrate:
      Train Data:
        - Accuracy: **89.62%**
        - Sensitivity: **89.78%**
        - Specificity: **89.53%**
        - False Positive Rate: **10.47%**

      Test Data:
        - Accuracy: **88.88%**
        - Sensitivity: **84.69%**
        - Specificity: **91.48%**
        - False Positive Rate: **8.52%**
5. **Resource Allocation Insights**
    - Leads scoring above **35% probability** should be aggressively pursued.
    - Segmentation based on lead scores allows better targeting of high-conversion leads during peak campaigns (e.g., with interns or new hires).

Business Recommendations

1. **Enhance Website Engagement**:
   - Improve website content to encourage more time spent on the platform, as it is the strongest predictor of conversion.
2. **Focus on SMS Communication**:
   - SMS campaigns and follow-ups significantly improve conversions. Ensure SMS touchpoints are utilized effectively.
3. **Prioritize Leads from "Lead Add Form"**:
   - Leads generated via this origin have the highest conversion likelihood. Focus marketing efforts on optimizing and scaling this lead source.
4. **Improve Email Opt-In Rates**:
   - Since opting out of emails negatively impacts conversion, develop strategies to retain email subscriptions.
5. **Leverage Tags for Lead Prioritization**:
   - Focus on leads with tags like **"Will revert after reading the email"** and **"Interested in other courses"** as they have high conversion probabilities.
6. **Target Working Professionals**:
   - Tailor campaigns specifically for working professionals, as they show higher conversion rates.