SUMMARY REPORT

Lead Scoring Case Study

Submitted by: Pavithra Sri S

1. Problem Statement:

"X Education" is an online education provider struggling with a low 30% lead conversion rate, increasing marketing costs and lost revenue. To improve efficiency, a predictive model will be developed to assign a lead score (0-100), indicating the likelihood of conversion. Higher scores will help prioritize follow-ups. The model will also be adaptable to evolving business needs.

2. Methodology:

Since, we are dealing with a classification problem, we shall be implementing a **Logistic Regression Model** using the Leads Dataset.

2.1 Data understanding and Pre-Processing:

- **Total Records:** 9,240
- **Total Features:** 37 (Categorical + Numerical)
- **Target Variable:** Converted (1 = Lead Converted, 0 = Not Converted)
- **Data Cleaning:** Removed redundant columns, handled missing values, and created dummy variables for categorical features.
- **Outlier Treatment:** Capped extreme values for key numerical features.
- **Final Feature Count:** 24 (After transformations and dummy encoding).

2.2 Feature Selection using VIF and RFE.

- **Variance Inflation Factor (VIF):** Removed highly correlated features (VIF > 10) to minimize multicollinearity. • **Recursive Feature Elimination (RFE):** Selected the top 20 features with the highest impact on lead conversion.

2.3 Model Building and Evaluation.

a) Train-Test Split:

- Train Data: 70% (6,351 records)
- Test Data: 30% (2,723 records)
- Feature Scaling: Standardized numerical variables.

b) Model Development:

- Built an initial model with **15 features selected via RFE**.
- Removed non-significant features (p-value > 0.05), retaining **14 key predictors**.

3. Results and Analysis.

3.1 Performance Metrics:

Metric	Train Data	Test Data
Accuracy	80.8%	81.6%
Sensitivity	78.2%	80%
Specificity	82.4%	82.7%
Recall	78.2%	80%

3.2 Lead Score and Conversion Probability as calculated by our model.

	Prospect ID	Converted	Converted_prob	final_predicted	Lead_Score
1	4050	1	0.977387	1	98
9	8187	0	0.954259	1	95
20	2052	1	0.899541	1	90
23	7005	1	0.993433	1	99
46	5353	1	0.909696	1	91
...	...	...	...	...	...
2717	6163	1	0.915467	1	92
2718	1467	1	0.965996	1	97
2719	4781	1	0.997850	1	100
2729	8043	1	0.958385	1	96
2730	5826	1	0.899336	1	90

3.3 Key Features of the Model:

1. **Total Time Spent on Website** - 1.078972
2. **Lead Origin - Lead Add Form** - 0.921377
3. **Current Occupation - Working Professional** - 0.629677
4. **Lead Source - Olark Chat** - 0.378323
5. **Last Activity - Page Visited on Website** - (-0.270248)
6. **Last Activity - Converted to Lead** - (-0.292159)
7. **Specialization - Others** - (-0.350695)
8. **Lead Origin - Landing Page Submission** - (-0.406182)
9. **Do Not Email** - (-0.443754)
10. **Current Occupation - Not Provided** - (-0.517307)
11. **Last Notable Activity - Email Opened** - (-0.521870)
12. **Last Activity - Olark Chat Conversation** - (-0.534478)
13. **Last Notable Activity - Modified** - (-0.560417)

4. Conclusions

4.1 Business Recommendations:

- **Prioritize High-Scoring Leads:** Leads with a score above 70 should be given higher priority for sales follow-ups to improve conversion rates.
- **Enhance Website Engagement:** Optimize user experience and interaction to increase conversions, especially focusing on increasing time spent on the website.
- **Improve Lead Tracking:** Utilize insights from last activity trends to refine follow-up strategies and ensure timely engagement.
- **Focus on High-Performing Acquisition Channels:** Channels such as the Welingak Website and Olark Chat have shown strong correlations with conversions. Invest in these channels to maximize lead quality.
- **Ensure Model Scalability:** Adapt the model to evolving business requirements and lead acquisition strategies, ensuring long-term effectiveness.

4.2 Final Conclusions:

The logistic regression model achieves over **80% accuracy** in predicting lead conversion. The model-generated lead scores provide actionable insights, enabling better decision-making, optimized sales efforts, and more efficient resource allocation.