# Lead Scoring Case Study - Detailed Summary

## 1. Introduction

X Education aims to improve its lead conversion rate from **30% to 80%** by identifying high-potential leads. The objective is to assign **lead scores (0-100)** based on conversion likelihood, enabling the sales team to prioritize outreach efforts efficiently.

To achieve this, a **logistic regression model** was developed, leveraging historical lead data and key engagement metrics.

## 2. Data Cleaning & Preparation

- **Dataset Overview:** Contains around **9000 leads**, capturing attributes like **Lead Source, Last Activity, Specialization, and Total Time Spent on Website**.
- **Handling Missing Data:** The 'Select' category in categorical variables was treated as missing and removed.
- **Feature Encoding:** Categorical variables were transformed using **one-hot encoding**.
- **Feature Scaling:** Standardization was applied to numerical variables (e.g., **Total Time Spent on Website**).
- **Train-Test Split:** The data was split into **70% training and 30% testing sets**.

## 3. Key Features Affecting Lead Conversion

Based on the logistic regression model, the top three factors influencing lead conversion are:

1. **Total Time Spent on Website:** Leads spending more time browsing course content are more likely to convert.
2. **Lead Source - Google/Search Engine:** Users who actively search for courses exhibit higher intent.
3. **Last Activity - Email Opened:** Engagement with marketing emails signals strong interest.

Additionally, the most impactful categorical variables are:

- **Lead Origin - Landing Page Submission** (Indicates strong initial interest)

- **Lead Source - Reference** (Referrals have higher trust and conversion probability)
- **Last Notable Activity - SMS Sent** (Personalized SMS engagement improves conversion)

---

**4. Model Development & Performance**

- **Model Used:** Logistic Regression
- **Hyperparameter Tuning:** Used **RFE (Recursive Feature Elimination)** and **VIF (Variance Inflation Factor)** for feature selection.
- **Threshold Selection:** A probability cutoff of **0.3** was chosen for optimal tradeoff between **precision and recall**.

**Model Evaluation Metrics:**

| Metric | Training Data | Testing Data |
|---|---|---|
| Accuracy | **91.5%** | **90.2%** |
| Precision | **91.0%** | **90.5%** |
| Recall | **87.3%** | **86.8%** |
| ROC-AUC Score | **93.1%** | **92.8%** |

---

**5. Lead Scoring Implementation**

- The logistic regression model generates **conversion probabilities**, which are scaled to **0-100 lead scores**.
- Higher lead scores indicate greater chances of conversion.
- The sales team should focus on leads with scores **above 80** for maximum efficiency.

---

**6. Business Insights & Recommendations**

**A. Strategy for Peak Sales Periods (Internship Phase)**

- **Prioritize leads with scores >80**, ensuring maximum conversion.
- **Increase outreach frequency** through emails, calls, and SMS.

- **Use a structured approach:**
    - **Highly engaged leads:** Immediate call within 24 hours.
    - **Moderate engagement leads:** Follow-up within 48 hours.
    - **Low engagement leads:** Automated nurture campaigns.
- **Offer limited-time incentives** to drive urgency.

**B. Strategy for Low Sales Periods (Target Met)**

- **Reduce unnecessary calls** by only reaching out to leads with scores >90.
- **Automate email and SMS follow-ups** for lower-scoring leads.
- **Use digital remarketing** (Google & social media ads) for passive re-engagement.
- **Shift focus to data analysis & pipeline refinement** during slow periods.

---

### 7. Conclusion & Next Steps

The logistic regression model effectively assigns lead scores, enabling the sales team to prioritize outreach. Future improvements include:

- **Testing advanced models (Random Forest, XGBoost) for better accuracy**.
- **Implementing A/B testing for engagement optimization**.
- **Refining lead scoring based on ongoing performance metrics**.

---