Summary Report: Lead Scoring Case Study

Objective

The primary goal of this study was to develop a logistic regression model to assign a lead score between 0 and 100 to each of the leads which can be used by the company to target potential leads.

-- A higher score would mean that the lead is hot, i.e. is most likely to convert whereas a lower score would mean that the lead is cold and will mostly not get converted.

Problem Statement

-An education company named X Education sells online courses to industry professionals.The company markets its courses on several websites and search engines like Google. Once these people land on the website, they might browse the courses or fill up a form for the course or watch some videos. When these people fill up a form providing their email address or phone number, they are classified to be a lead. Once these leads are acquired, employees from the sales team start making calls, writing emails, etc. Through this process, some of the leads get converted while most do not. The typical lead conversion rate at X education is around 30%. Its lead conversion rate is very poor. the company wishes to identify the most potential leads, also known as 'Hot Leads'

Data Processing & Cleaning

The dataset contained lead information, including source, website activity, email preferences, and past interactions. Several data cleaning and preprocessing steps were performed:

- Dropped irrelevant variables (e.g., "Prospect ID," "Lead Number").
- Replaced missing values with meaningful categories such as "Not Provided."
- Created dummy variables for categorical data.
- Removed outliers from numeric features like "Total Visits" and "Page Views Per Visit."

After cleaning, the dataset was split into **training (70%)** and **testing (30%)** sets.

Model Development

A logistic regression model was built using Recursive Feature Elimination (RFE) to identify the most important predictors of lead conversion.

Key Variables Influencing Lead Conversion:

1. Total Time Spent on Website:
    - Leads who spent more time browsing the website were more likely to convert.
2. Total Visits:
    - More visits indicate strong interest, but extremely high visits may reflect indecision.
3. Lead Source:

    - Leads from Google, Direct Traffic, and Organic Search had higher conversion rates compared to those from advertisements or referrals.

4. Lead Origin:
    - Leads generated through the API, Lead Add Form, and Landing Pages showed higher conversion potential.
5. Do Not Email (Yes/No):
    - Leads who opted out of emails were less likely to convert.
6. Last Activity:
    - Leads who engaged through Olark Chat Conversations or responded to emails had a higher probability of conversion.

After feature selection, the model was trained, and its predictive power was evaluated.

Model Performance & Evaluation

The model's performance was assessed using ROC-AUC, accuracy, precision, and recall metrics.

- Optimized Cutoff Score: 0.44 (balancing precision and recall).
- Accuracy: 80.5% – The model correctly classifies leads 80.5% of the time.
- Precision: 75% – 75% of predicted hot leads actually converted.
- Recall: 73% – 73% of actual hot leads were correctly identified.
- ROC-AUC Score: 0.88 – High ability to differentiate between hot and cold leads.

These metrics indicate that the model is highly effective in identifying valuable leads.

Business Implications

The study revealed key insights into **lead behaviour** and **conversion patterns**:

1. **Better Lead Prioritization:**
    - Sales teams can focus on high-scoring leads, improving efficiency.
    - Reducing wasted effort on low-quality leads saves time and resources.
2. **Key Factors Influencing Conversions:**
    - Website Engagement: Higher time spent = Higher conversion probability.
    - Lead Source Matters: Google and Direct Traffic yield better leads.
    - Email & Chat Activity: Interactions through Olark Chat or emails increase conversion rates.
3. **Strategic Recommendations:**
    - Optimize Digital Marketing – Invest in high-performing lead sources.
    - Enhance Website Experience – Encourage longer browsing times.

- Personalized Follow-ups – Engage leads who interact via chat and email.

Conclusion

The logistic regression model successfully predicts lead conversion with high accuracy and precision. X Education can now increase sales efficiency by focusing on high-scoring leads, leading to higher revenue and reduced operational costs. This data-driven approach improves lead conversion rates significantly, making sales and marketing efforts more effective.