Lead Scoring- Capstone Project

1. Business Problem

- The company wanted to **improve lead conversion rates** by identifying and prioritizing prospects most likely to convert into paying customers.
- The existing process treated all leads equally, leading to inefficient use of sales resources.
- Our objective: **Build a predictive model** that assigns conversion probabilities to each lead, enabling the sales team to focus on high-potential prospects.

---

2. Data and Initial Exploration

- We started with a lead dataset containing various attributes — such as **Total Time Spent on Website, Total Visits, Lead Origin, Lead Source, presence of interactions like Olark Chat, and demographic information**.
- Some features were optional dropdown fields where the default “Select” indicated no selection — these were handled appropriately during preprocessing.

---

3. Data Preprocessing

Steps included:

- **Handling missing values** and removing redundant variables.
- **Encoding categorical features** into numeric form.
- **Scaling numeric variables** (e.g., Total Time Spent on Website, Total Visits) to normalize ranges.
- Splitting into **training (≈70%)** and **test (≈30%)** sets to ensure unbiased evaluation.

---

4. Model Building

We selected **Logistic Regression** due to:

- Interpretability (understandable coefficients for business stakeholders).
- Suitability for binary classification (Converted vs. Not Converted).
- Ability to output probabilities for cutoff tuning.

The model was first fitted on **X_train** and validated using training predictions before testing on unseen data.

---

5. Cutoff Selection Using Training Data

By default, logistic regression uses a **0.5 probability threshold**, but this may not give the best trade-off between:

- **Accuracy** – Overall correctness.
- **Sensitivity (Recall)** – Ability to correctly identify converters.
- **Specificity** – Ability to correctly reject non-converters.

We plotted metrics vs. cutoff and found that **0.42** gave an optimal balance:

Metric	Value
Accuracy	**0.624**
Sensitivity	**0.608**
Specificity	**0.638**

Confusion Matrix @ 0.42 (Train)

```
[[1476  836]   TN=1476, FP=836
 [ 842 1307]]  FN=842,  TP=1307
```

---

6. Testing on Unseen Data

On the **test set**, the chosen cutoff was **0.45** after slight tuning for stability.
Results:

Metric	Value
Accuracy	**0.624**
Sensitivity	**0.618**
Specificity	**0.630**

Confusion Matrix @ 0.45 (Test)

```
[[627 369]     TN=627, FP=369
```

```
 [350 566]]    FN=350, TP=566
```

Performance was consistent between train and test data, showing **good generalization**.

---

7. Precision–Recall View

We also examined the **Precision–Recall trade-off** to consider marketing requirements (often higher recall is preferred so fewer actual converters are missed).

- Baseline (before PR tuning):
  Precision ≈ **0.729**, Recall ≈ **0.547**

- After optimizing cutoff for PR balance:
  Cutoff = ~0.40
  Accuracy: **0.656**
  Precision: **0.660**
  Recall: **0.591**

Confusion Matrix after PR Optimization (Train)

text
```
[[1657  655]
 [ 880 1269]]
```

While accuracy improved slightly, recall was boosted without large precision loss — favorable for lead prioritization.

---

8. Key Insights

- **Total Time Spent on Website** emerged as a **high-impact predictor** — engaged users tend to convert more.
- The logistic regression model **generalized well** between train and test, indicating stable predictive power.
- Through cutoff tuning, we achieved a balance tailored to the business goal: maximize potential conversions while controlling false positives.

---

9. Recommendations

1. **Integrate the Model into the CRM**: Assign conversion probabilities to incoming leads in real-time.

2. **Sales Prioritization**:

    - High-probability leads (above cutoff) = immediate follow-up.
    - Medium probability = nurture campaigns.
    - Low probability = minimal resource allocation.

3. **Monitor Over Time**: Revalidate cutoffs as lead behavior or marketing channels change.

4. **Feature Enrichment**: Future models could incorporate interaction sequences, campaign engagement, and response times for improved accuracy.

---

10. Conclusion

This logistic regression model, with a tuned cutoff of **~0.42–0.45**, provides a balanced classification of leads with consistent performance across datasets.
It can significantly **increase sales efficiency** by enabling **data-driven lead prioritization**, reducing wasted effort on low-potential prospects, and focusing resources where they matter most.