

A decorative graphic on the left side of the slide consisting of white lines and circles on a blue gradient background, resembling a circuit board or data flow diagram.

# CAPSTONE PROJECT – LEAD SCORES ANALYSIS PRESENTATION

BY GADE PAVAN KUMAR REDDY

# PROBLEM STATEMENT OVERVIEW

- **X Education** offers online courses for working professionals and receives thousands of leads daily from digital marketing, referrals, and direct traffic.
- Despite a high volume of leads, the **conversion rate is only ~30%**, which makes the sales process inefficient.
- The sales team currently **contacts all leads**, regardless of their likelihood to convert, resulting in wasted effort and time.
- The company wants to **focus only on high-potential leads** — the “Hot Leads” — to improve efficiency and boost conversions.

# PROJECT OBJECTIVE

- Build a **logistic regression model** that assigns a **lead score between 0 and 100**, helping the sales team identify and prioritize leads most likely to convert.

# UNDERSTANDING LEAD DATA

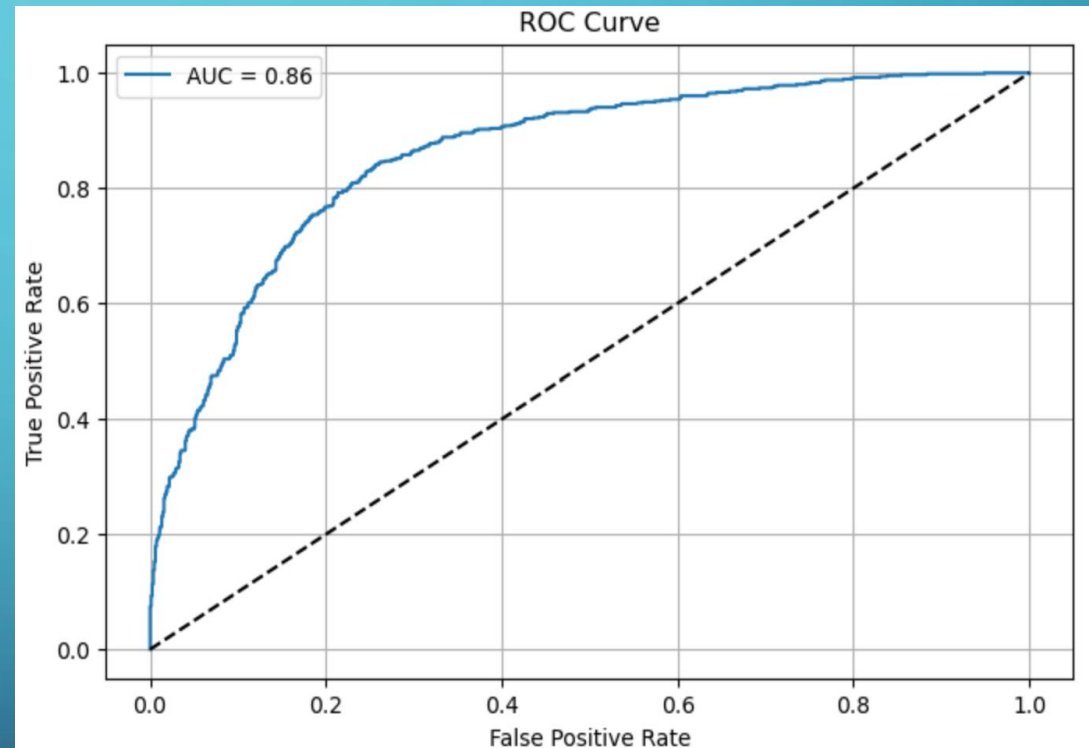
- Dataset consists of 9240 leads and 37 features (Categorical, Numerical, Binary)
- Pre – processing steps :
  - Dropped columns with over 35% null values and low variance.
  - Replaced “Selected” with nulls and imputed
  - One – hot encoded categorical variables
  - Normalized numeric features
- Final Dataset consists of 4925 leads and 55 features

# LOGISTIC REGRESSION MODEL

- Chosen Model : Logistic Regression (interpretable, good for probabilities)
- Used :
  - Train – Test Split (70/30)
  - StandardScaler for numeric features
  - Recursive Feature Elimination (RFE) for feature elimination
- Output : Lead Score (0-100), representing probability of conversion

# MODEL PERFORMANCE

- Accuracy : 78 %
- Precision : 78 %
- Recall : 70 %
- ROC AUC : 0.86



# FEATURE IMPORTANCE

- What is your current occupation
- Lead Origin
- Last Notable Activity

	Feature	Coefficient
10	What is your current occupation_Working Profes...	2.637134
2	Lead Origin_Lead Add Form	1.496496
12	Last Notable Activity_Had a Phone Conversation	1.406541
14	Last Notable Activity_Unreachable	1.159456
11	Last Notable Activity_Email Bounced	0.820740
13	Last Notable Activity_SMS Sent	0.801761
8	What is your current occupation_Housewife	0.516390
7	Last Activity_SMS Sent	0.362249
0	TotalVisits	0.045582
1	Total Time Spent on Website	0.001947

# BUSINESS IMPACT

- Use **lead score to prioritize follow-up**.
- Avoid low-score leads → Save time and resources.
- Personalize communication for hot leads based on top factors (like activity, source).

## Outcomes :

- Higher conversion with same team size.
- Better ROI on marketing channels



# RECOMMENDATIONS

- Monitor & retrain model every 3–6 months
- Improve input data quality (e.g., ensure proper tracking of Last Activity)
- Increase spend on channels like Google
- Automate emails based on engagement level

# LESSONS LEARNED

## Key Takeaways:

- Business domain understanding is vital in feature handling.
- Logistic regression offers both **accuracy and interpretability**.
- Threshold tuning helps align models with business strategy.

**Thank You**

