# LEAD CONVERSION ANALYSIS

FOR X EDUCATION

# INTRODUCTION

- Purpose: Improve lead conversion rate for X Education
- Dataset: 9240 leads with various attributes
- Tools used: Python, Pandas, NumPy, Matplotlib, Seaborn, Scikit-learn

# DATA PREPROCESSING

- Handled missing values:
  - Dropped columns with >45% missing data
  - Filled missing values in 'Tags', 'City', 'Lead Source', 'Specialization'
- Removed irrelevant features: 'Prospect ID', 'Lead Number', 'What matters most to you in choosing a course', 'Lead Profile', 'What is your current occupation', 'Country', 'How did you hear about X Education'
- Created dummy variables for categorical features
- Handled outliers in 'TotalVisits' and 'Page Views Per Visit'

# EXPLORATORY DATA

- Most leads from India (before dropping 'Country')
- 'Google' is the primary lead source
- 'Landing Page Submission' and 'API' are major lead origins
- 'Email Opened' and 'SMS Sent' are common last activities

# MODEL DEVELOPMENT

- Logistic Regression model
- Used Recursive Feature Elimination (RFE) to select top 15 features
- Applied StandardScaler to numeric columns
- Used statsmodels for detailed model statistics

# KEY PREDICTORS OF LEAD CONVERSION

- Total Time Spent on Website
- Total Visits
- Last Activity: SMS Sent
- Lead Origin: Landing Page Submission
- Lead Source: Direct Traffic
- Specialization: Others

# MODEL PERFORMANCE

1. Accuracy: 90.2%
2. Sensitivity (True Positive Rate): 81.2%
3. Specificity (True Negative Rate): 95.6%
4. ROC AUC Score: 0.95
5. Optimal probability cutoff: 0.3

# BUSINESS RECOMMENDATIONS

- Focus on increasing website engagement (time spent and visits)
- Optimize SMS marketing campaigns
- Improve landing page for lead submissions
- Enhance direct traffic channels
- Tailor approach for leads from various specializations

# NEXT STEPS

- Implement A/B testing for website improvements
- Develop personalized content strategy
- Set up real-time lead scoring system using the developed model
- Continuously monitor and update the model