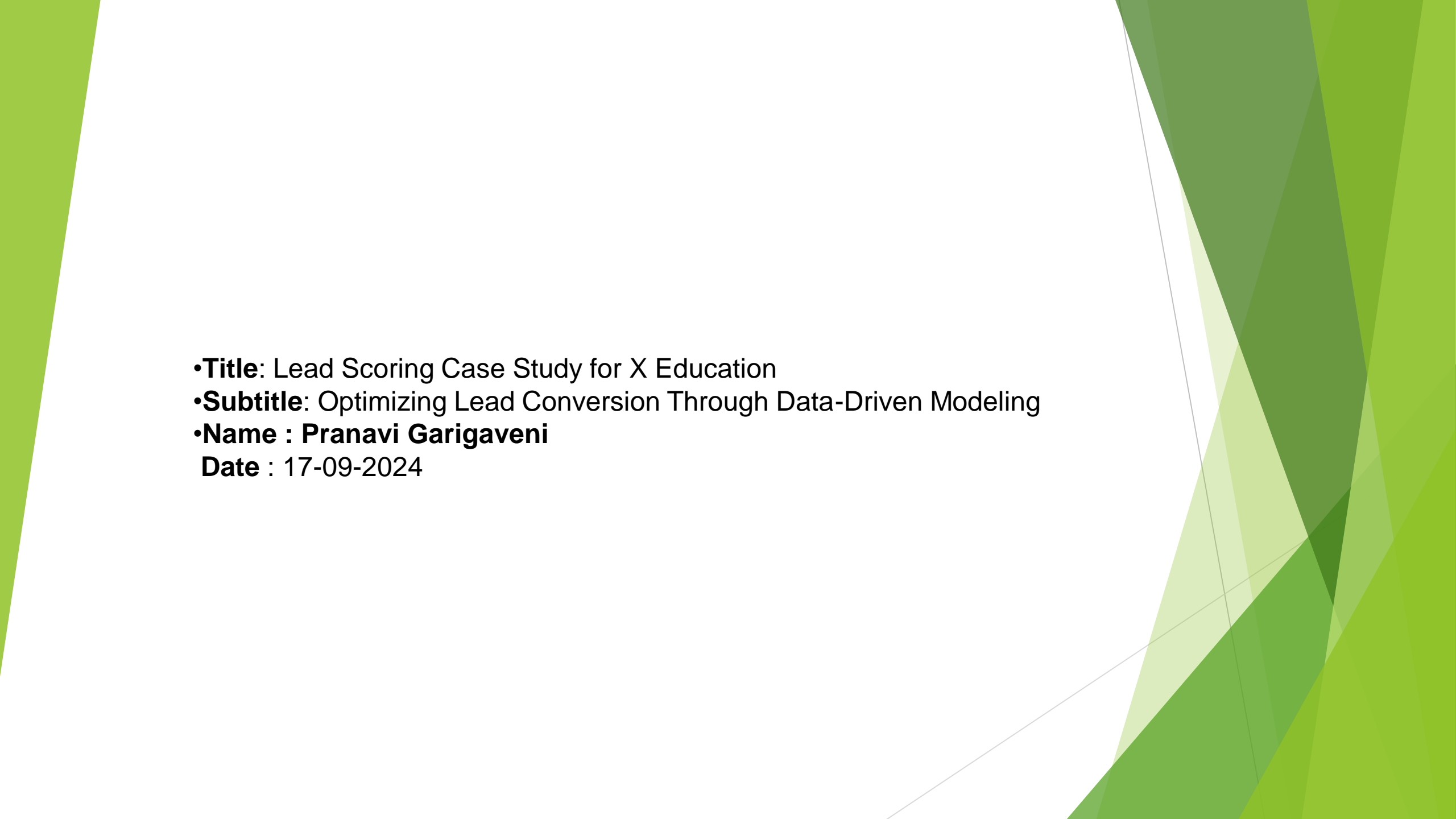


- 
- Title:** Lead Scoring Case Study for X Education
 - Subtitle:** Optimizing Lead Conversion Through Data-Driven Modeling
 - Name : Pranavi Garigaveni**
 - Date :** 17-09-2024

Introduction

- **What is Lead Scoring?**
 - Lead scoring is a method to rank leads (potential customers) based on their likelihood to convert.
 - X Education, an online course provider, faces a low lead conversion rate (~30%) and seeks to optimize this using a predictive model.

PROBLEM STATEMENT

•Challenges at X Education:

- High volume of leads but low conversion rate.
- Lack of an efficient mechanism to prioritize leads based on conversion probability.
- Objective: Build a model that assigns lead scores to predict the likelihood of conversion.

CASE STUDY GOALS

• Primary Goals:

- Develop a logistic regression model to assign lead scores.
- Achieve a lead conversion rate of at least 80%.
- Provide recommendations for future adjustments.

DATA OVERVIEW

• Dataset Summary:

- 9240 leads, 37 columns.
- Types of data: numerical and categorical.
- Important features: Lead Source, Total Time Spent on Website, Page Views, Lead Activity.

	Prospect ID	Lead Number	Lead Origin	Lead Source	Do Not Email	Do Not Call	Converted	TotalVisits	Total Time Spent on Website	Page Views Per Visit	Last Activity	Country	Specialization	How did you hear about X Education	What is your current occupation	ma mc y cho a c
0	7927b2df-8bba-4d29-b9a2-b6e0beafe620	660737	API	Olark Chat	No	No	0	0.0	0	0.0	Page Visited on Website	NaN	Select	Select	Unemployed	E C Pros
1	2a272436-5132-4136-86fa-dcc88c88f482	660728	API	Organic Search	No	No	0	5.0	674	2.5	Email Opened	India	Select	Select	Unemployed	E C Pros
2	8cc8c611-a219-4f35-ad23-fdfd2656bd8a	660727	Landing Page Submission	Direct Traffic	No	No	1	2.0	1532	2.0	Email Opened	India	Business Administration	Select	Student	E C Pros
3	0cc2df48-7cf4-4e39-9de9-19797f9b38cc	660719	Landing Page Submission	Direct Traffic	No	No	0	1.0	305	1.0	Unreachable	India	Media and Advertising	Word Of Mouth	Unemployed	E C Pros
4	3256f628-e534-4826-9d63-4a8b88782852	660681	Landing Page Submission	Google	No	No	1	2.0	1428	1.0	Converted to Lead	India	Select	Other	Unemployed	E C Pros

```

xleads.shape
(9240, 37)

```

DATA CLEANING

Data Cleaning Steps:

- Handled missing values (e.g., dropped columns with more than 3000 missing values).
- Removed irrelevant or redundant columns (e.g., "City," "Country").
- Created dummy variables for categorical data (e.g., Lead Origin, Lead Source).

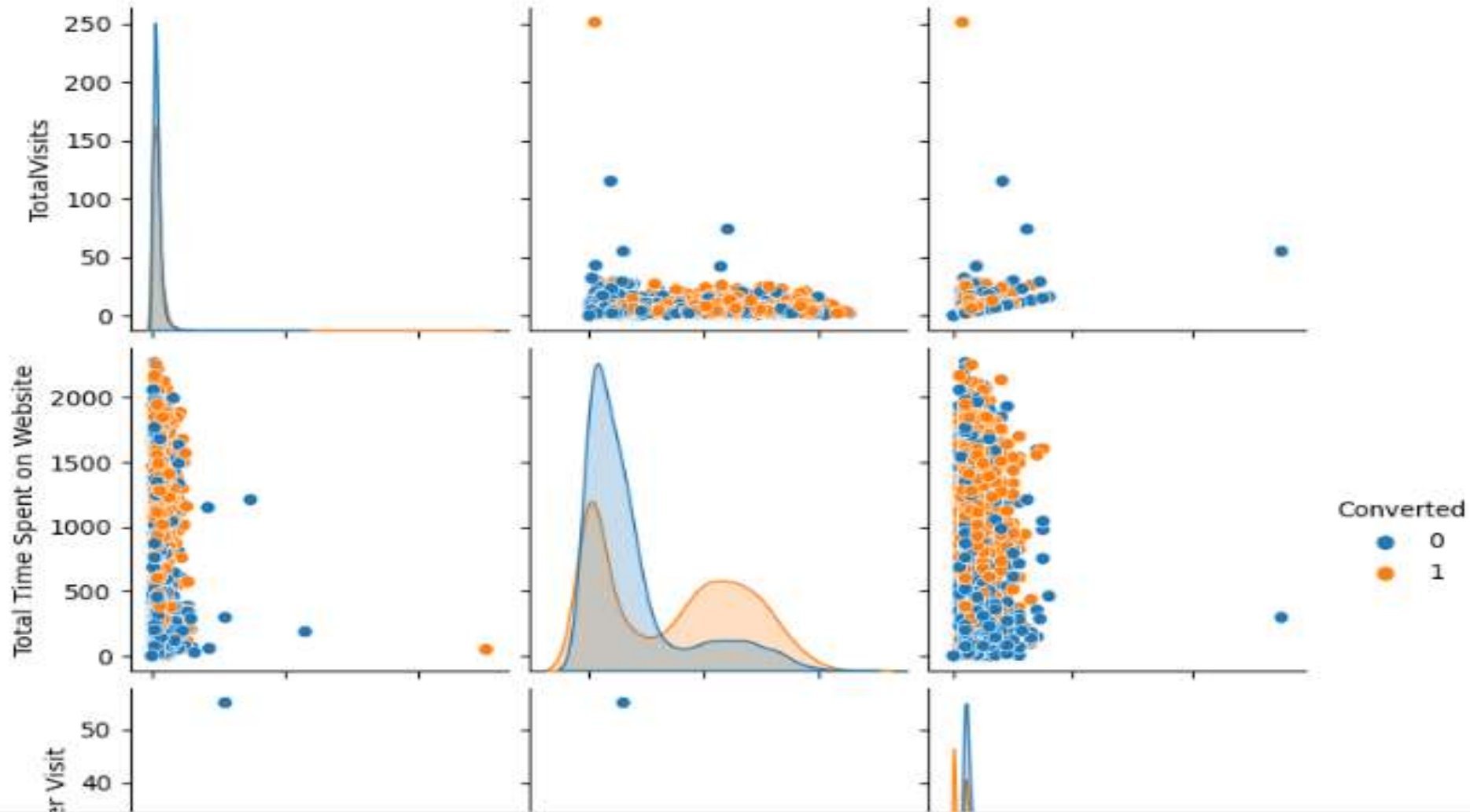
PREPARATION

•Steps Taken:

- Dropped irrelevant columns (e.g., columns with >3000 missing values like Lead Profile, How did you hear about X Education).
- Removed records with null values in critical columns (e.g., Specialization, TotalVisits).
- Created dummy variables for categorical columns.
- Scaled numeric variables like **TotalVisits**, **Page Views Per Visit**, and **Total Time Spent on Website**.

Prepare the data for modelling

```
from matplotlib import pyplot as plt
import seaborn as sns
sns.pairplot(xleads,diag_kind='kde',hue='Converted')
plt.show()
```



FEATURE SELECTION

Variables Selected Using RFE (Recursive Feature Elimination):

- Selected top 15 features after initial model evaluation.
- Dropped variables with high p-values and high multicollinearity ($VIF > 5$).
- **Key Variables:**
 - Total Time Spent on Website
 - Total Visits
 - Lead Source (e.g., Reference, Welingak Website)

MODEL BUILDING

- Model Type: Logistic Regression
- Train/Test Split:
 - Training: 70%, Testing: 30%
 - **Scaling:** Applied MinMax scaling to continuous variables (Total Time Spent, Page Views).
 - **Final Features:** Selected based on RFE and VIF scores.

MODEL PERFORMANCE (TRAINING SET)

•Evaluation Metrics:

- Accuracy: ~79%
- Sensitivity: 74%
- Specificity: 83%
- ROC-AUC Score: 0.86 (good model performance).

•Confusion Matrix:

- True Positives: 1589
- True Negatives: 1929

OPTIMAL CUTOFF SELECTION

- Optimal Probability Cutoff: 0.42 (based on ROC curve and tradeoff between sensitivity and specificity).
- Adjusted Accuracy: 79%
- Conclusion: Optimal balance between precision and recall for effective lead scoring.

MODEL PERFORMANCE (TEST SET)

• Test Set Results:

- Accuracy: 78%
- Sensitivity: 77%
- Specificity: 79%

• Confusion Matrix (Test Set):

- True Positives: 714
- True Negatives: 786

• Conclusion: The model performs consistently well on unseen data.

KEY INSIGHTS:

Top Variables Influencing Lead Conversion:

- Total Time Spent on Website
- Total Visits
- Page Views Per Visit

•Top Categorical Variables:

- Lead Source: Reference and Welingak Website.
- Lead Origin: Lead Add Form.

RECOMMENDATIONS

For Aggressive Lead Conversion:

- Focus on leads with high scores during high-intensity sales periods (e.g., when interns are active).
- Prioritize leads with high engagement metrics (Total Time Spent, Total Visits).

•For Efficient Call Management After Sales Targets:

- Increase the probability threshold for calls to avoid unnecessary outreach.
- Utilize automated emails/SMS for low-priority leads unless they show renewed interest.

FUTURE IMPROVEMENTS

• Potential Enhancements:

- Integrate new data sources (social media engagement, customer feedback).
- Use advanced models (e.g., Random Forest or XGBoost) to improve accuracy.
- Regularly retrain the model with fresh data to ensure optimal performance.

CONCLUSION

•Summary:

- Successfully built a lead scoring model that can enhance lead prioritization and conversion.
- The model's performance provides a foundation for improving the sales team's efficiency.
- X Education can now tailor its strategies based on lead conversion probabilities.

THANK YOU

BY
PRANAVI GARIGAVENI