Lead Scoring Case Study

Submitted by
Dev Vyas
Shayamon Bastakoti
Aryma Rawat
Abhishek Srivastav

Contents

- Problem statement
- Problem approach
- EDA
- Correlations
- Model Evaluation
- Observations
- Conclusion

Problem Statement

- An education company called X Education offers online courses tailored for industry professionals. Each day, numerous professionals interested in these courses visit the company's website to browse through the offerings. Visitors can fill out a form on the website, after which the company classifies them as leads.
- Once a lead is acquired, the sales team engages in follow-up activities such as making calls and sending emails. Through this process, a portion of the leads is successfully converted, while the majority are not.
- Currently, X Education has an average lead conversion rate of about 30%. This means that out of 100 leads generated in a day, only around 30 are converted into customers. To improve efficiency and boost the conversion rate, the company aims to identify high-potential leads, referred to as *Hot Leads*, to focus their efforts more effectively.

Business Objective

- X Education aims to develop a model that assigns each lead a score between 0 and 100, enabling the identification of *Hot Leads* to improve their conversion rate.
- The CEO's goal is to achieve a lead conversion rate of 80%. Additionally, the model should be designed to accommodate future challenges, such as optimizing actions during peak times, effectively utilizing the entire workforce, and defining strategies to sustain success after reaching the target.

Problem Approach

- Importing the data and inspecting
- Data preparation
- Exploratory Data Analysis (EDA)
- Feature Engineering
- Splitting the Data
- Data Scaling
- Logistic Regression with Hyper-parameter Tuning
- Random Forest with Cross-Validation
- Gradient Boosting with Cross-Validation
- Model Comparison and ROC Curve

EDA- Data cleaning

Replace 'Select' with NaN and drop columns with > 40% missing values

```
data.replace('Select', np.nan, inplace=True)
columns_to_drop = [
    'How did you hear about X Education', 'Lead Profile', 'Lead Quality',
    'Asymmetrique Profile Score', 'Asymmetrique Activity Score',
    'Asymmetrique Activity Index', 'Asymmetrique Profile Index'
]
data.drop(columns=columns_to_drop, inplace=True)
```

Combine rare categories in 'Country' into 'Other'

+ Code + Markdown

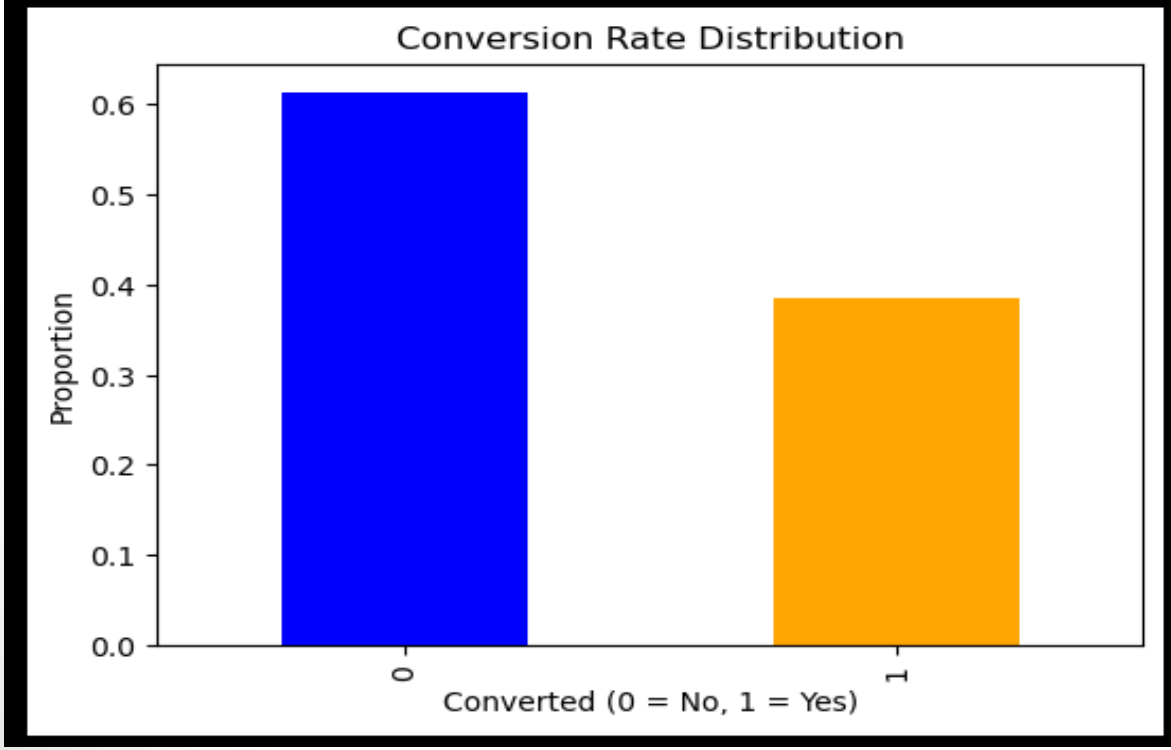
```
rare_countries = data['Country'].value_counts()[data['Country'].value_counts() < 10].index
data['Country'] = data['Country'].replace(rare_countries, 'Other')
```

Drop irrelevant or unique identifier columns

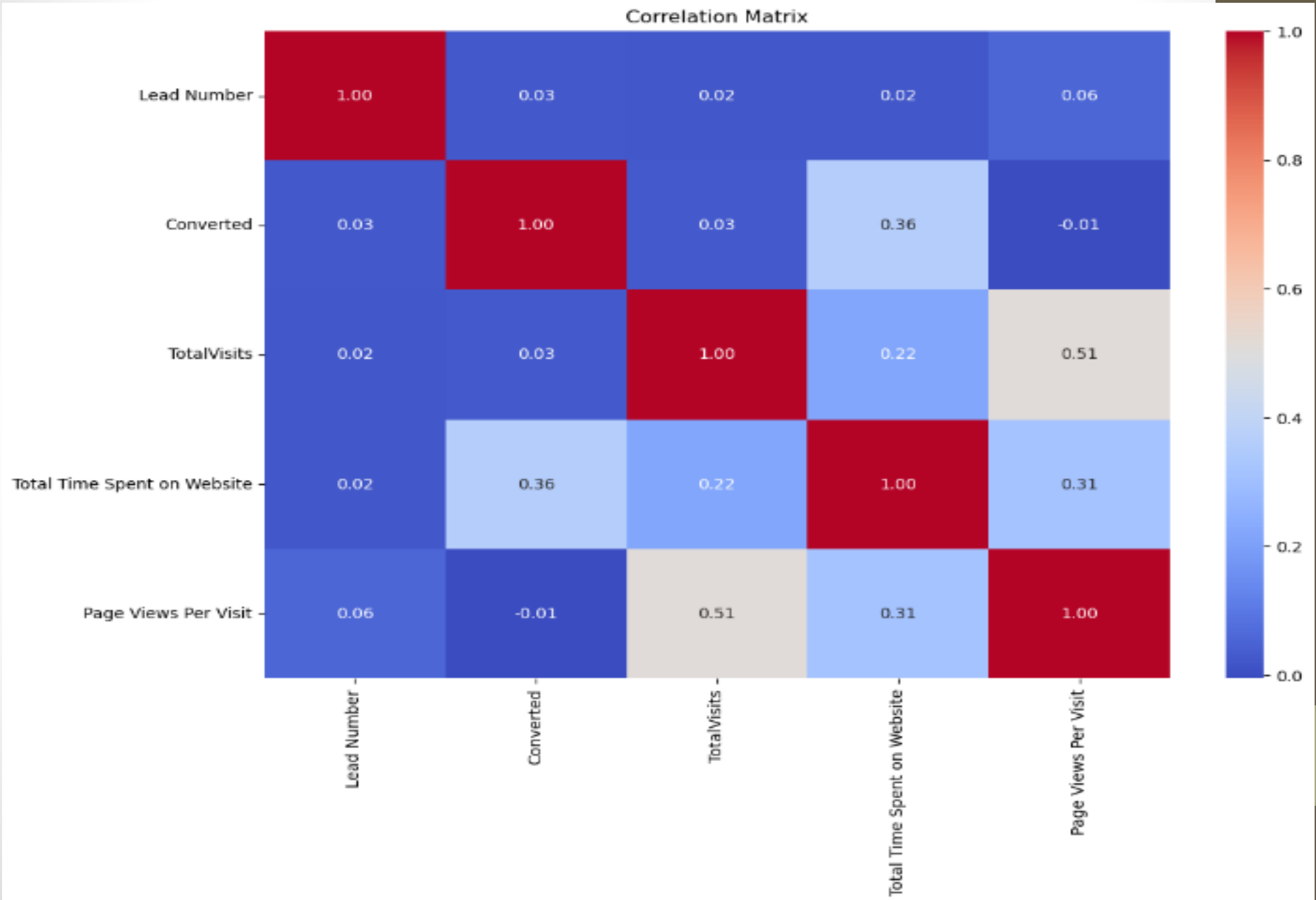
```
data.drop(columns=['Prospect ID'], inplace=True)
```

Conversion Rate Distribution

```
plt.figure(figsize=(6, 4))
data['Converted'].value_counts(normalize=True).plot(kind='bar', color=['blue', 'orange'])
plt.title('Conversion Rate Distribution')
plt.xlabel('Converted (0 = No, 1 = Yes)')
plt.ylabel('Proportion')
plt.show()
```



Correlation

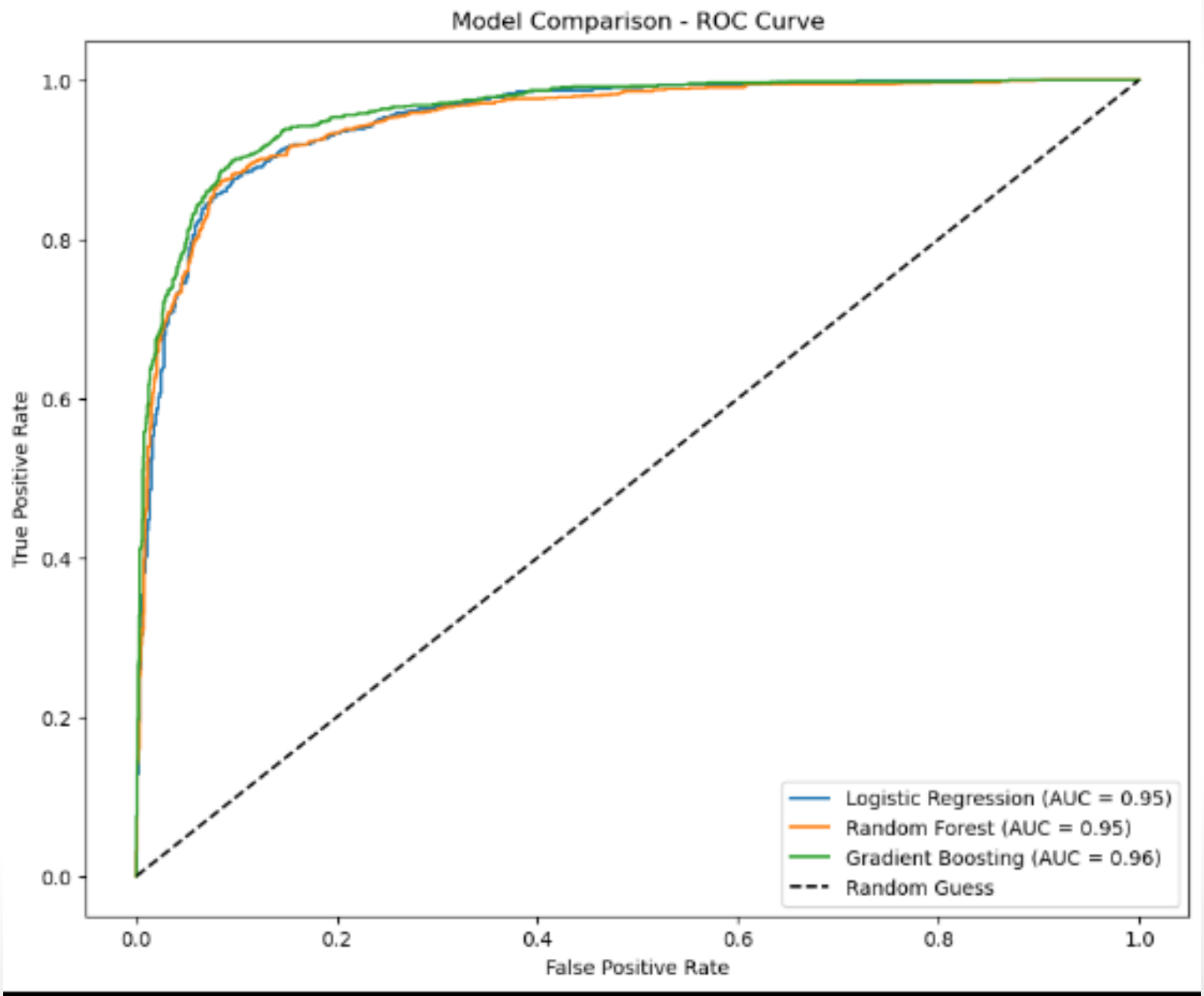


Insights from Correlation Matrix

- Insights:

Total Time Spent on Website and Conversion show strong positive correlation, indicating its importance in predicting lead conversion.

Model Comparison and ROC Curve



Feature Importance for Random Forest

Top Features from Random Forest:

	Feature	Importance
2	Total Time Spent on Website	0.137572
4	Interaction_TotalTime_Visits	0.094698
103	Tags_Will revert after reading the email	0.088330
127	Last Notable Activity_SMS Sent	0.078061
42	Last Activity_SMS Sent	0.063938
6	Lead Origin_Lead Add Form	0.063107
98	Tags_Ringing	0.061937
86	Tags_Closed by Horizzon	0.048360
76	What is your current occupation_Working Profes...	0.047238
75	What is your current occupation_Unemployed	0.032613

Logistic Regression with Hyper-parameter Tuning

Classification Report:

	precision	recall	f1-score	support
0	0.91	0.93	0.92	1704
1	0.88	0.85	0.86	1068
accuracy			0.90	2772
macro avg	0.89	0.89	0.89	2772
weighted avg	0.90	0.90	0.90	2772

Conclusion

- X Education can enhance lead conversion by focusing on key predictors like lead tags, website engagement, and communication activities. Prioritizing high-impact categorical variables ensures targeted efforts and efficient resource allocation.
- Strategies like prioritizing leads, automating outreach, and monitoring performance can maximize conversions. Once targets are met, a precision-based approach and reallocation of resources will sustain success while minimizing unnecessary calls. These measures will help X Education achieve its 80% conversion goal effectively.