LEAD SCORING MODEL FOR X EDUCATION

ENHANCING LEAD CONVERSION EFFICIENCY USING LOGISTIC REGRESSION

PROBLEM STATEMENT

- X Education sells online courses to professionals.
- Many leads are generated, but only ~30% convert to customers.
- The objective is to improve lead conversion by identifying high-potential leads.
- The goal is to achieve an 80% conversion rate using a lead scoring model.

BUSINESS GOALS

- Develop a predictive model to assign lead scores.
- Prioritize high-potential leads for the sales team.
- Reduce time and effort spent on low-potential leads.
- Increase overall lead conversion efficiency.

OVERALL APPROACH

DATA OVERVIEW

- Dataset contains ~9000 leads.
- Key variables:
    - **Lead Source** (Google, Direct, Referral, etc.)
    - **User Behavior** (Total Visits, Time Spent, Page Views)
    - **Lead Quality & Tags**
    - **Activity Status** (Last Email, Last Call)
- Missing values & ‘Select’ values handled properly.

```
[285]  data.describe()
```

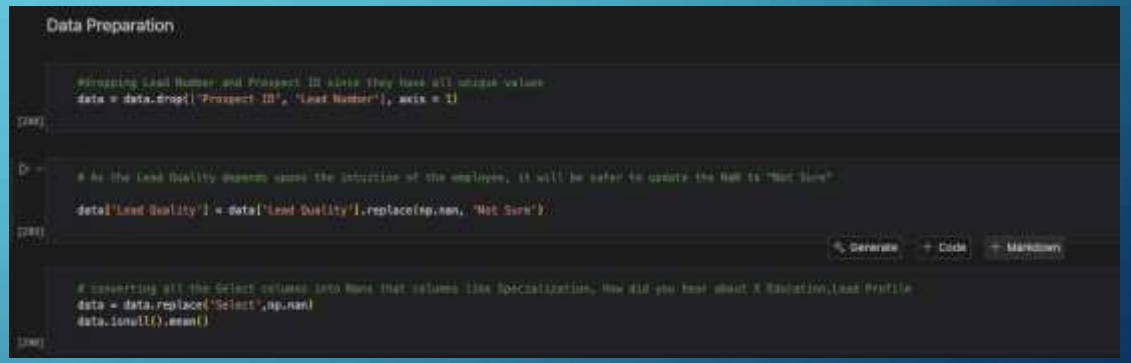
	Lead Number	Converted	TotalVisits	Total Time Spent on Website	Page Views Per Visit	Asymmetrique Activity Score	Asymmetrique Profile Score
count	9240.000000	9240.000000	9103.000000	9240.000000	9103.000000	5022.000000	5022.000000
mean	617188.435606	0.385390	3.445238	487.698268	2.362820	14.306252	16.344883
std	23405.995698	0.486714	4.854853	548.021466	2.161418	1.386694	1.811395
min	579533.000000	0.000000	0.000000	0.000000	0.000000	7.000000	11.000000
25%	596484.500000	0.000000	1.000000	12.000000	1.000000	14.000000	15.000000
50%	615479.000000	0.000000	3.000000	248.000000	2.000000	14.000000	16.000000
75%	637387.250000	1.000000	5.000000	936.000000	3.000000	15.000000	18.000000
max	660737.000000	1.000000	251.000000	2272.000000	55.000000	18.000000	20.000000

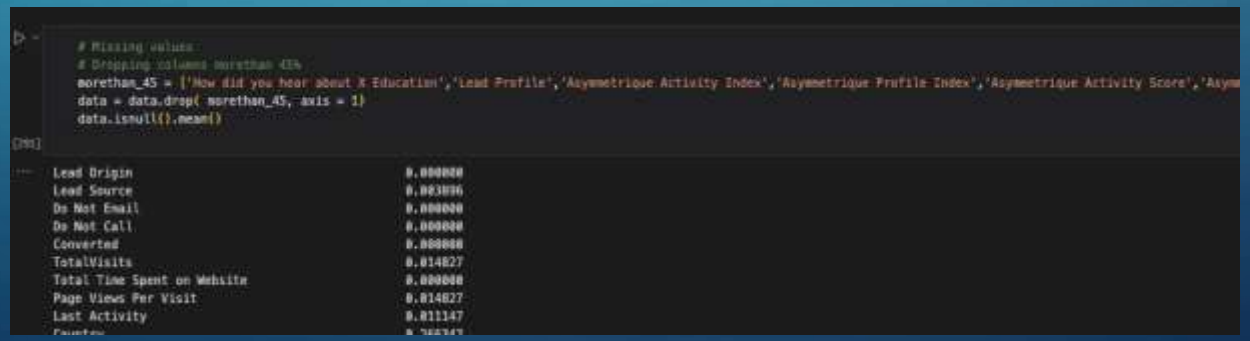
```
[284]  data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 9240 entries, 0 to 9239
Data columns (total 37 columns):
 #   Column                                         Non-Null Count  Dtype
---  ------                                         --------------  -----
 0   Prospect ID                                    9240 non-null   object
 1   Lead Number                                    9240 non-null   int64
 2   Lead Origin                                    9240 non-null   object
 3   Lead Source                                    9204 non-null   object
 4   Do Not Email                                   9240 non-null   object
 5   Do Not Call                                    9240 non-null   object
 6   Converted                                      9240 non-null   int64
 7   TotalVisits                                    9103 non-null   float64
```

DATA CLEANING & PREPROCESSING

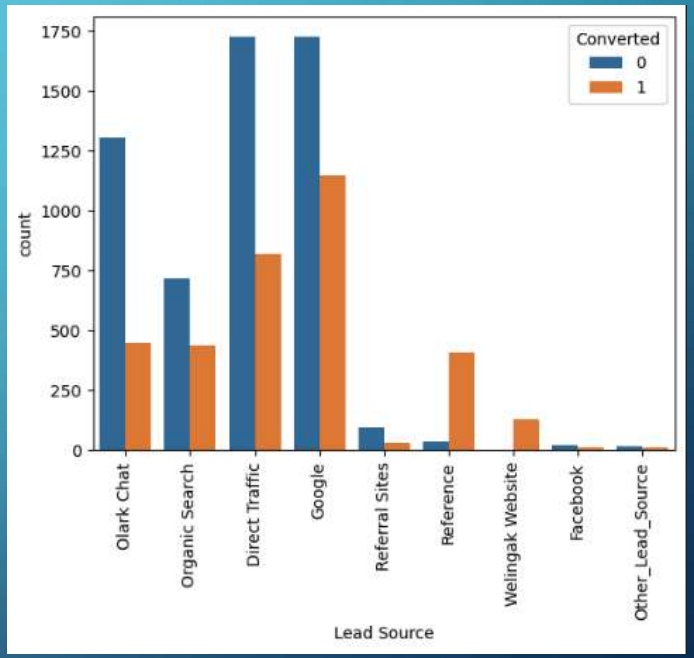
- Dropped unnecessary variables (e.g., Lead Number, Prospect ID).
- Handled missing values & encoded categorical variables.
- Standardized numerical features for model consistency.
- Removed multicollinearity (VIF Analysis).





EDA – EXPLORATORY DATA ANALYSIS

1. **Google and Direct Traffic Are Key Lead Sources:**
   - Google and Direct Traffic generate the highest number of leads.
   - A significant portion of these leads are converted (orange bars), making them high-value lead sources.
2. **Organic Search Shows Moderate Performance:**
   - Organic search brings in a good number of leads.
   - However, the conversion rate is lower compared to Google and Direct Traffic.
3. **Olark Chat Has a High Drop-off Rate:**
   - Many leads come from Olark Chat, but most do not convert (high blue bar, low orange bar).
   - This indicates low lead quality from this source.
4. **Referral Sites and Welingak Website Perform Well:**
   - While they have fewer leads, they show higher conversion rates, suggesting these sources bring in more serious prospects.
5. **Other Sources Have Negligible Impact:**
   - Facebook, Other Lead Sources, and Influence generate very few leads and conversions.
   - This suggests less focus should be placed on these sources.



Analysis of Lead Source vs. Conversion Rate Graph

Business Recommendations

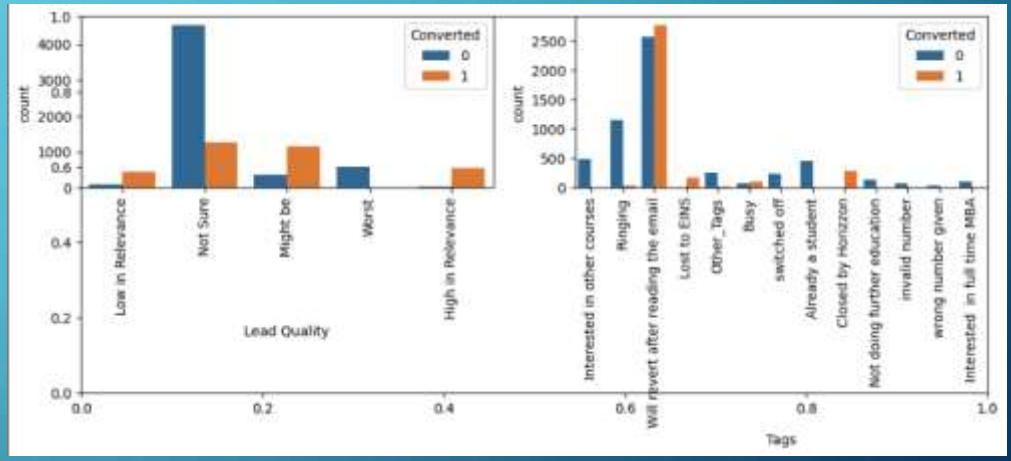
- Prioritize marketing spend on Google and Direct Traffic as they yield the most conversions.
- Optimize Organic Search strategy to improve conversion rates from search leads.
- Improve engagement on Olark Chat to better qualify and nurture chat-based leads.
- Leverage Referral and Welingak Website leads since they have a higher conversion rate despite low volume.
- Reduce efforts on Facebook and other low-performing lead sources.

Left Graph: Lead Quality vs. Conversion

1. **Most Leads Fall Under "Not Sure" and "Low in Relevance"**
    - The majority of leads are classified as "Not Sure" or "Low in Relevance."
    - These categories have a **low conversion rate** as indicated by the smaller orange bars.
2. **"High in Relevance" and "Might Be" Categories Convert Better**
    - Although these categories have fewer leads, their conversion rates are **higher**.
    - The orange bars are more balanced compared to the blue bars, indicating better lead quality.
3. **"Worst" Lead Quality Category Has Very Few Conversions**
    - Almost all leads in this category do not convert, suggesting that these leads should not be prioritized.

Right Graph: Tags vs. Conversion

1. **"Will Revert After Reading the Email" Tag Has the Highest Volume**
    - This tag has the largest number of leads but a **low conversion rate** (high blue bar, low orange bar).
    - Many of these leads are likely not engaging further despite showing initial interest.
2. **"Ringing" and "Switched Off" Tags Show Poor Conversion**
    - These leads are difficult to reach, as indicated by the **high number of non-converted leads**.
    - Efforts on these leads may not be fruitful.
3. **"Lost to EINS" and "Closed by Horizzon" Have High Conversion**
    - These tags show a **significant proportion of conversions** (orange bars are high).
    - These leads should be given priority in follow-ups.
4. **"Interested in Full-Time MBA" Tag Shows Strong Potential**
    - Leads with this tag have a **higher conversion rate**, indicating strong intent.
    - These leads should be nurtured more effectively.



Analysis of Lead Quality and Tags vs. Conversion Rate

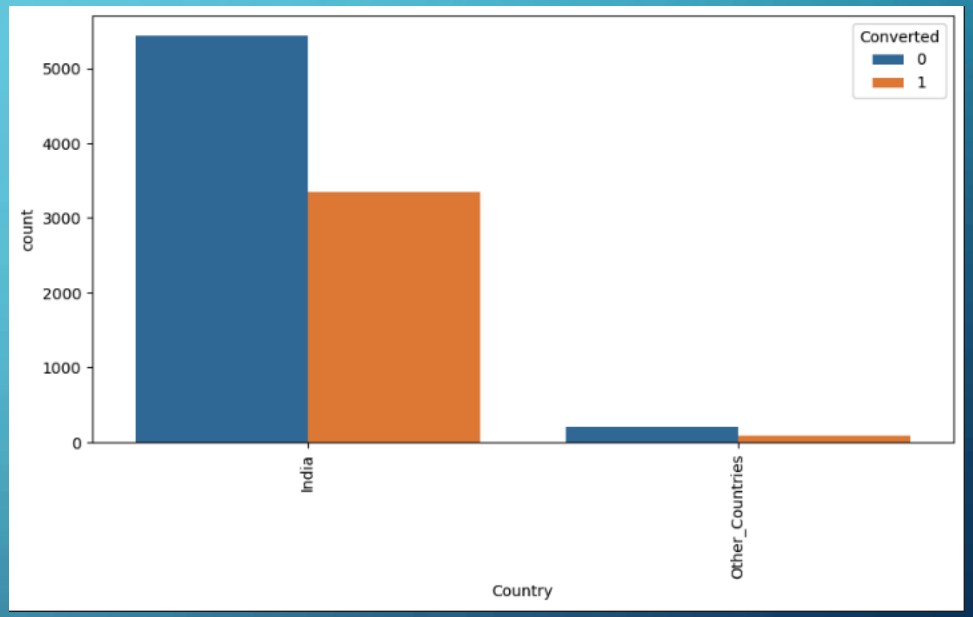
Business Recommendations

- **Prioritize leads marked as "High in Relevance" and "Might Be"** since they convert at a higher rate.
- **Reduce focus on "Not Sure" and "Low in Relevance" leads** unless further engagement shows strong interest.
- **Minimize efforts on "Ringing" and "Switched Off" leads** as they have poor response rates.
- **Focus on leads tagged as "Lost to EINS" and "Closed by Horizzon"** since they have a higher likelihood of conversion.
- **Create personalized follow-up strategies for "Will Revert After Reading the Email"** leads to push them towards engagement.

1. **India is the Primary Market**
    - The majority of leads are from **India,** with a significantly higher count compared to other countries.
    - India has a **higher conversion rate,** as seen from the substantial number of converted leads (orange bar).
2. **Other Countries Have Minimal Lead Generation and Conversions**
    - Very few leads come from **Other Countries,** and their conversion rate is also low.
    - The low lead volume from international markets suggests limited global reach.

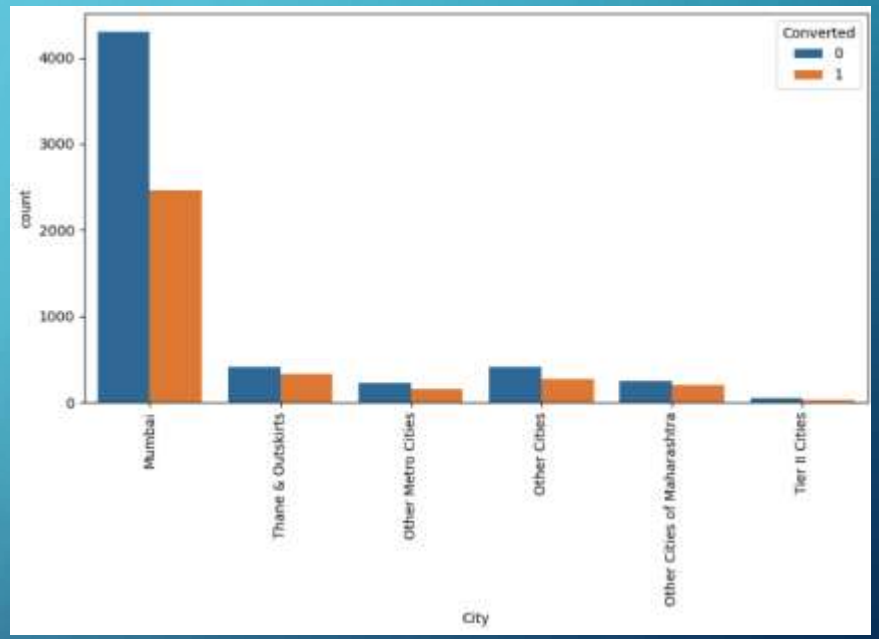
Business Recommendations

- **Focus marketing efforts on India** since it contributes the highest number of leads and conversions.
- **Analyze the potential for international expansion** by understanding why international leads are not converting.
- **Improve lead generation strategies for other countries** if the business aims to expand globally.
- **Optimize marketing spend** by allocating more resources to regions with high conversion potential.



Analysis of Country vs. Conversion Rate

1. **Mumbai Dominates Lead Generation and Conversions**
    - Mumbai has the highest number of leads compared to other cities.
    - It also has the **highest conversion count**, indicating that leads from Mumbai have a strong potential to convert into paying customers.
2. **Other Cities Show Similar Patterns but Lower Volumes**
    - Cities like **Thane & Outskirts, Other Metro Cities, and Other Cities** generate fewer leads.
    - Their conversion rates appear to be relatively balanced, but the overall volume is much smaller compared to Mumbai.
3. **Tier II Cities Have Minimal Impact**
    - Leads from **Tier II Cities** are very few, and conversions are also negligible.
    - This suggests that these cities may not be a primary target for marketing and sales efforts.



Analysis of City vs. Conversion Rate

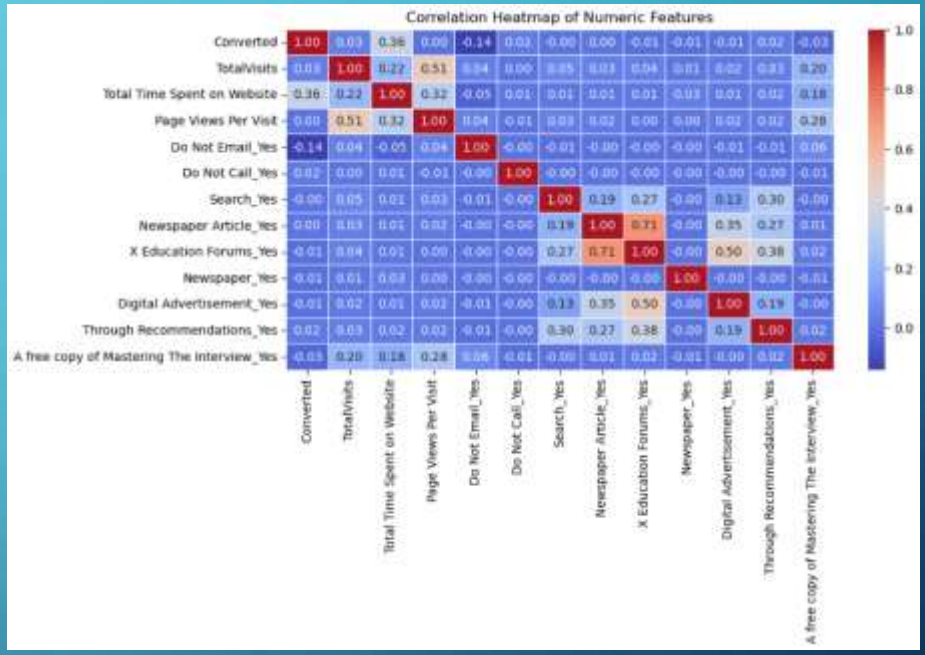
Business Recommendations

- **Focus marketing and sales efforts on Mumbai**, as it generates the most leads and conversions.
- **Explore strategies to increase lead generation in other metro cities**, as they have some conversion potential.
- **Limit efforts in Tier II Cities**, unless additional research suggests untapped potential.
- **Enhance digital marketing and local campaigns in high-converting regions** to maximize ROI.

1. **Strong Positive Correlation with Conversion:**
    - **Total Time Spent on Website (0.36):**
    - The highest positive correlation with conversion.
    - Leads that spend more time on the website are more likely to convert.
    - **Page Views Per Visit (0.32):**
    - More page views per visit indicate higher engagement, increasing conversion probability.
2. **Negative Correlation with Conversion:**
    - **Do Not Email (-0.31):**
    - Leads who opted out of emails are less likely to convert.
    - **Do Not Call (-0.08):**
    - A weak negative correlation, indicating that leads who opted out of calls are slightly less likely to convert.
3. **Other Observations:**
    - **Total Visits has low correlation (0.01) with conversion:**
    - Indicates that just visiting the website multiple times does not necessarily lead to conversion.
    - **Other variables such as Newspaper, X Education Forums, and Digital Advertisements have very low correlation with conversion, suggesting they have minimal impact on lead conversion.**

Business Recommendations

- **Focus on engaging leads who spend more time on the website.**
- **Optimize email marketing, as opting out of emails correlates with lower conversion rates.**
- **Improve website experience to encourage more page views per visit.**
- **Reduce reliance on newspaper advertisements and forums, as they show minimal impact on conversion.**
- **Prioritize follow-ups for highly engaged leads rather than targeting those with high visit counts but low engagement.**



Analysis of Correlation Heatmap of Numeric Features

MODEL BUILDING

Feature selection was performed by employing recursive feature elimination to select the fifteen most important features.
The most important features are Lead Quality, Last Notable Activity, Tags (Lost to EINS, Closed by Horizzon, etc.).
Feature scaling was achieved by bringing the numerical features onto the same scale using StandardScaler, which would lead to increased performance of the models.

Features scaled include:

1. TotalVisits
2. Total Time Spent on Website
3. Page Views Per Visit

Model Used - Logistic Regression:
Why Logistic Regression?
Best suitable for binary classification (Lead Converted = Yes/No).
It produces easily interpretable coefficients, hence reveals the importance of features.
Used Generalized Linear Model (from Statsmodels) for deep insights.

In order to counter multicollinearity:
Take VIF(Variance Inflation Factor) and remove the variables that are highly correlated.
Each feature is guaranteed to contain only unique information.

Final Model Training:
Trained Logistic Regression with balanced class weights to address the imbalanced dataset.
Performed backward elimination in order to refine feature selection.

MODEL EVALUATION & PERFORMANCE

Confusion matrix is as follows
[[3086 247]
[537 2481]]

Interpretation
3086 True Negatives (non-conversions predicted correctly).
2481 True Positives (conversions predicted correctly).
247 False Positives (leads predicted to convert but that did not convert).
537 False Negatives (leads that converted but were not predicted).
Precision, Recall, and F1-Score
Precision=0.91, 91% of predicted conversions were actually correct.
Recall=0.82, 82% of actual conversions, were correctly identified.
F1-Score: 0.86, harmonic mean of precision & recall.
Accuracy at 87.8%-the model predicts lead conversions proficiently.

ROC Curve & AUC Score
The AUC Score=0.95: Excellent classification performance.
Model successfully distinguishes between converted & non-converted leads.

Feature Importance-Key Predictors (Logistic Regression Coefficients)
Lead Quality, Not Sure: -3.34, negative impact on conversion.
Tags: Closed by Horizzon, 7.95, highly predictive of conversion.
Tags: Lost to EINS, 9.17, strongest predictor of conversion.
Last Notable Activity, SMS Sent: 2.74, higher engagement, increases conversion probability.
Tags-Ringing: -1.69, negative impact- low probability of conversion.
Source of Lead: Welingak Website, 3.41, a good source of lead conversions.

Conclusion & Business Impact of the Lead Scoring Model

The Lead Scoring Model was devised in order to increase the lead conversion Efficiency for X Education. Under the Logistic Regression, the model indicated the key variables of Lead Quality, Last Notable Activity, and Tags-with-included features selected through Recursive Feature Elimination and standardized. An accuracy of 87.8% was achieved with the final model, with a precision of 0.91 while recall stood at 0.82, resulting in an AUC score of 0.95, which suggested that its predictive ability was pretty strong.

The analysis demonstrated that the conversion probability was highest for the leads tagged Lost to EINS, Closed by Horizzon; on the contrary, Not Sure and Ringing tags were low-potential leads. SMS and email engagement was a potent conversion driver, whilst leads labeled either as Switchoff or Ringing were less likely to convert. Of the sources, Google, Direct Traffic, and Welingak Website channels produced the most converting leads; of course, Facebook and many others had relatively minimal impact.
For increased efficiency, the sales team should follow up on high-converting leads while sending personalized follow-up SMS and email messages. Because so much marketing software has focused on lower-performing ones while squandering millions of dollars in the process, a shift in their focus to high performance will efficiently dispense the resource allocation decisions. Such data-driven insights provide the most enhanced sales prospecting for X Education, enabling an ever-expanding ROI and better conversion rate.

By,
Kankanala Saisudheer
Anish Khosla
Karthik R L