Lead Scoring Case Study

Group Case study by:

1. Vaddi Pradeep Satya Chandra
2. Rishabh Salekar
3. Utkarsh Pandey

Problem Statement:

X Education receives a large number of leads daily, but their current lead conversion rate is only around 30%. The company needs to improve the efficiency of their sales process by identifying and targeting the most promising leads, also known as 'Hot Leads'. This requires developing a predictive model to score each lead based on their likelihood of conversion.

Business Goal:

The primary goal is to build a logistic regression model that assigns a lead score between 0 and 100 to each lead, allowing X Education to focus on leads with a higher probability of conversion. The target is to increase the lead conversion rate to approximately 80% by prioritizing high-potential leads and optimizing the sales team's efforts.

Analysis Approach in Brief:

- **1. Reading and Understanding the Data (Exploratory Data Analysis - EDA)**
- Get a comprehensive understanding of the dataset by loading the data, inspect its structure, understand its features, and check for initial patterns, distributions, and anomalies.
- **2. Data Cleaning**
- Prepare the data for analysis by addressing issues like handle missing values, correct data types, and resolve inconsistencies in values like “Select” by making them to NaN values.
- **3. Visualizing the Data (Exploratory Data Analysis - EDA)**
- Identify trends, relationships, and outliers through Create charts, histograms, scatter plots, and other visualizations to better understand the data's distribution and correlations.
- **4. Data Preparation**
- Transform and prepare data for modeling by standardize data, encode categorical variables into dummies.
- **5. Splitting the Data into Training and Testing Sets**
- Ensure a fair evaluation of the model by dividing the dataset into training and testing subsets, typically using a split ratio like 70/30.

- **6. Building the Logistic Regression Model**

- Develop and refine a logistic regression model through:
  - **Automated Feature Selection:** Use Recursive Feature Elimination (RFE) to automatically select significant features.
  - **Manual Method:** Combine automated selection with manual adjustments based on p-value < 0.05 & VIF <5.
  - **Detailed Statistics:** Build the model using a statistical package (e.g., statsmodels) to obtain detailed insights into model performance and feature significance.

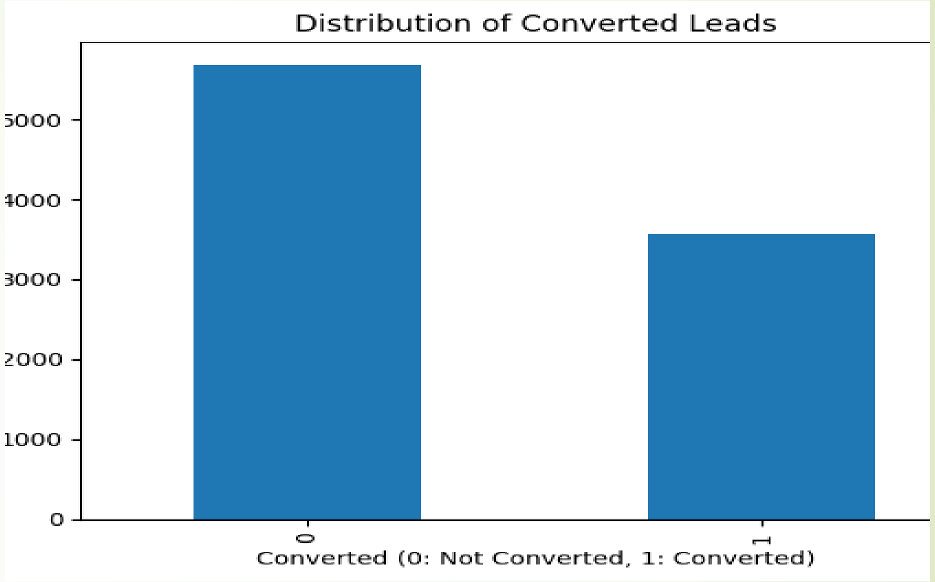
- **7. Model Evaluation**

- Assess model performance and optimize its predictive capability though:
  - **Finding the Optimal Cutoff:** Determine the best threshold for classification based on performance metrics.
  - **Precision-Recall View:** Analyze precision, recall, and other relevant metrics to evaluate the model's effectiveness.

- **8. Making Predictions on the Test Set**

- Apply the trained model to unseen data and make predictions. Use the model to generate predictions on the test dataset, evaluate these predictions against actual values, and interpret the results.

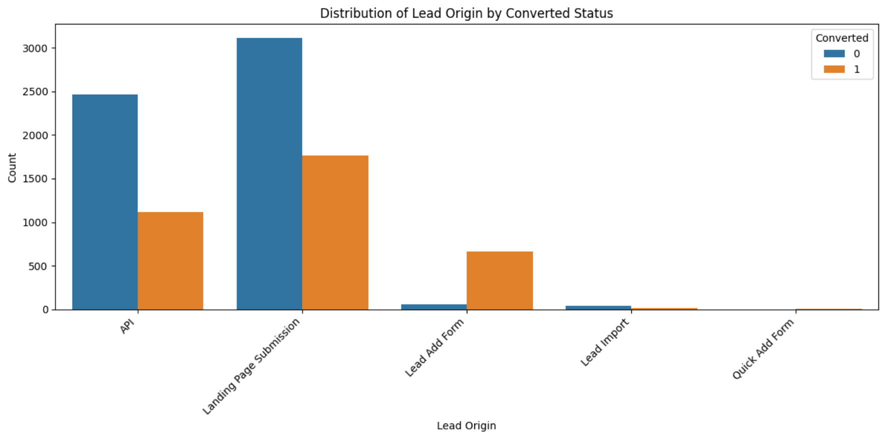
Results in business terms

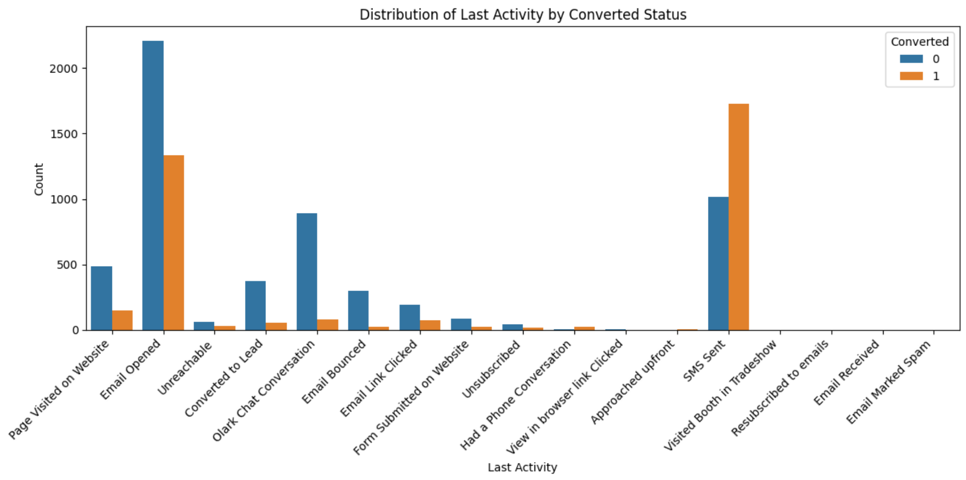


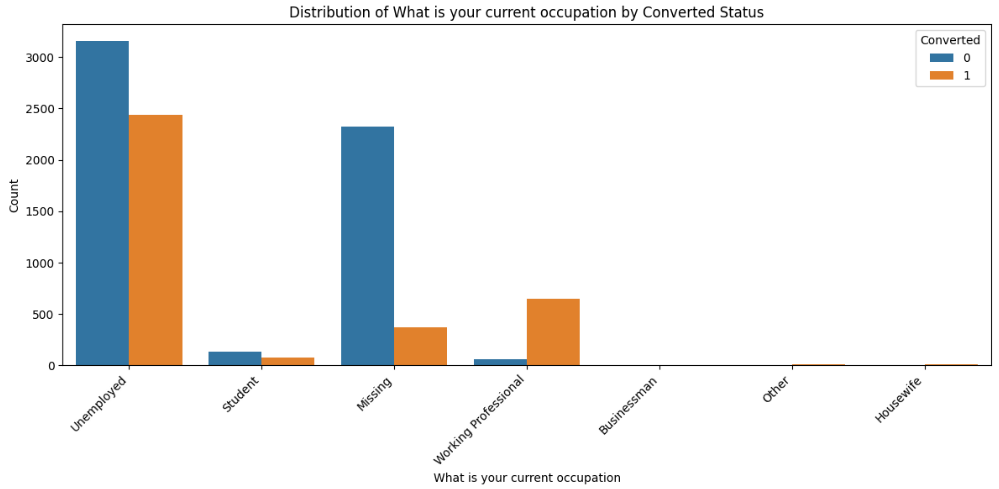
It has been observed that the conversion % for over all leads is 38.5% out of 100%

Results in business terms - 1

1. Lead Origin:
   1. Observation: API and landing page leads convert better than Lead Add Form leads.
   2. Action: Improve API and landing page experiences to attract high-quality leads.
2. Lead Source:
   1. Observation: Google leads convert best; Direct traffic and Olark chat leads are also effective.
   2. Action: Boost marketing on Google, optimize for direct traffic, and enhance Olark chat engagement.
3. Last Activity:
   1. Observation: Email Opened and SMS Sent have higher conversion rates.
   2. Action: Focus on email and SMS campaigns for better engagement and conversions.
4. What is the current occupation:
   1. Observation: Unemployed leads have a higher conversion rate, indicating they are more receptive to educational opportunities.
   2. Action: Target marketing campaigns towards unemployed individuals to capitalize on their interest in pursuing new education.

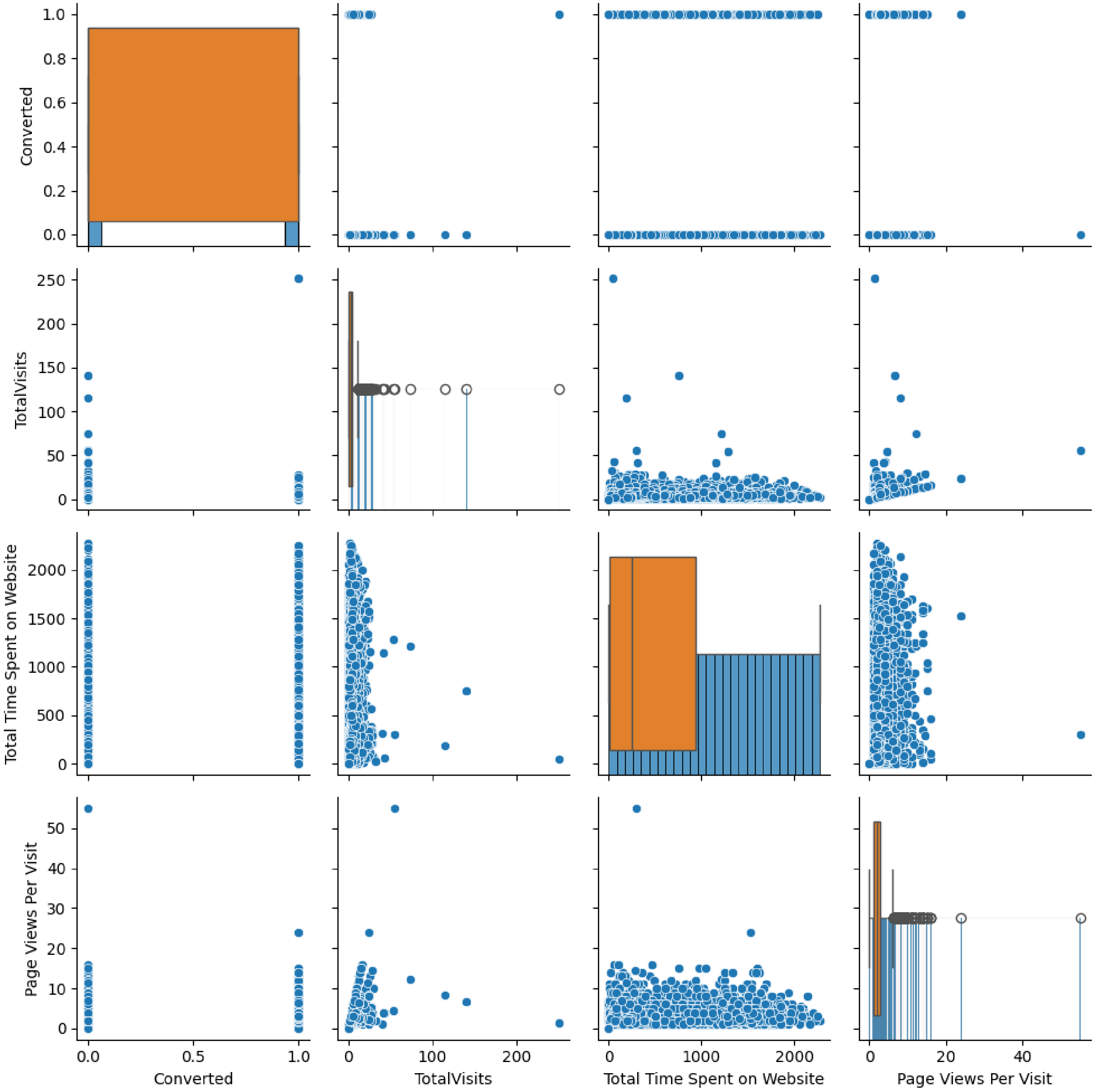






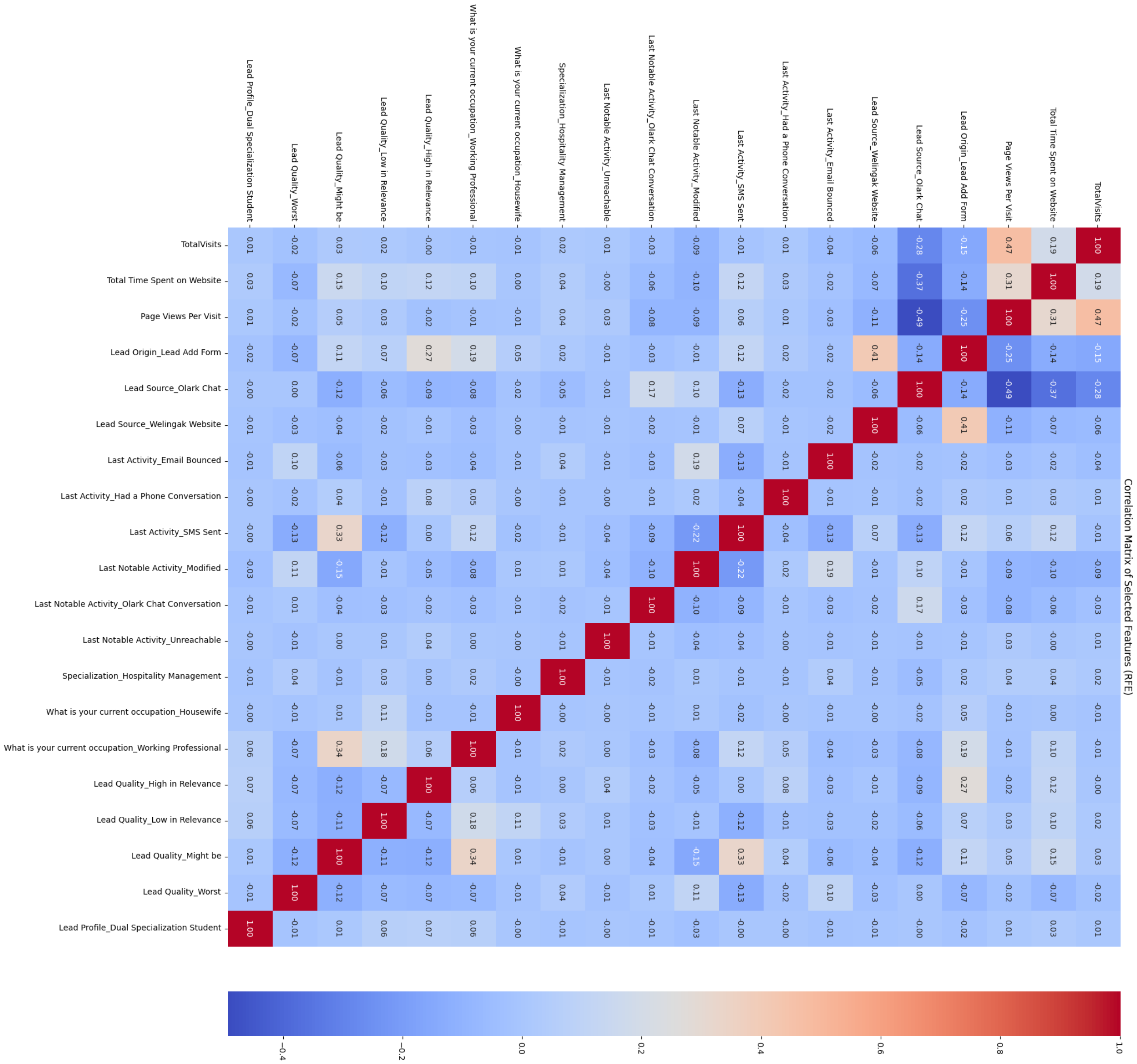
Results in business terms - 2

1. TotalVisits: Most leads have a low number of total visits, with a few outliers having significantly higher visits.
2. Total Time Spent on Website: There's a clear positive correlation between total time spent on the website and conversion. Leads spending more time on the website are more likely to convert.
3. Page Views Per Visit: Similar to total visits, most leads have a low number of page views per visit, with a few outliers having significantly higher page views.



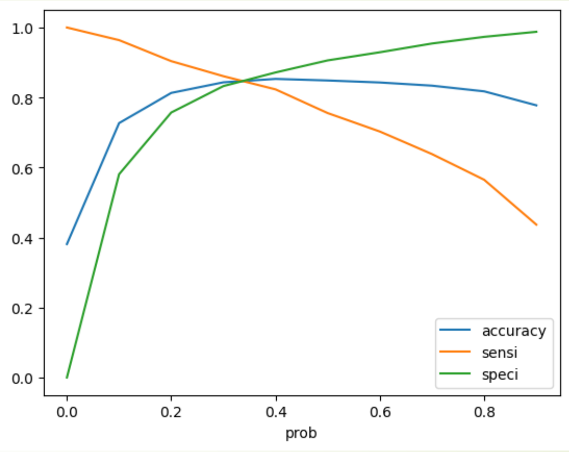
Results in business terms - 3

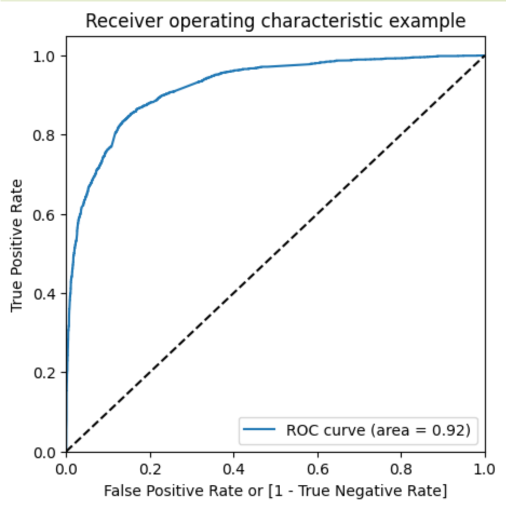
1. Strong Positive Correlation:
   1. Insight: Features move together (e.g., “TotalVisits” and “Page Views Per Visit”).
   2. Action: Enhance website engagement and content quality to boost lead conversion.
2. Strong Negative Correlation:
   1. Insight: Features move inversely (e.g., “Lead Origin_Lead Add Form” vs. “Lead Source_Olark Chat”).
   2. Action: Optimize effective lead generation channels and allocate resources accordingly.
3. No Correlation:
   1. Insight: Features are not significantly related (e.g., “Lead Source_Direct Traffic” vs. “TotalVisits”).
   2. Action: Focus on improving website user experience and content rather than traffic sources.
4. Multicollinearity:
   1. Insight: High correlation between features (e.g., “TotalVisits” and “Page Views Per Visit”).
   2. Action: Remove or combine redundant features to improve model stability and performance.

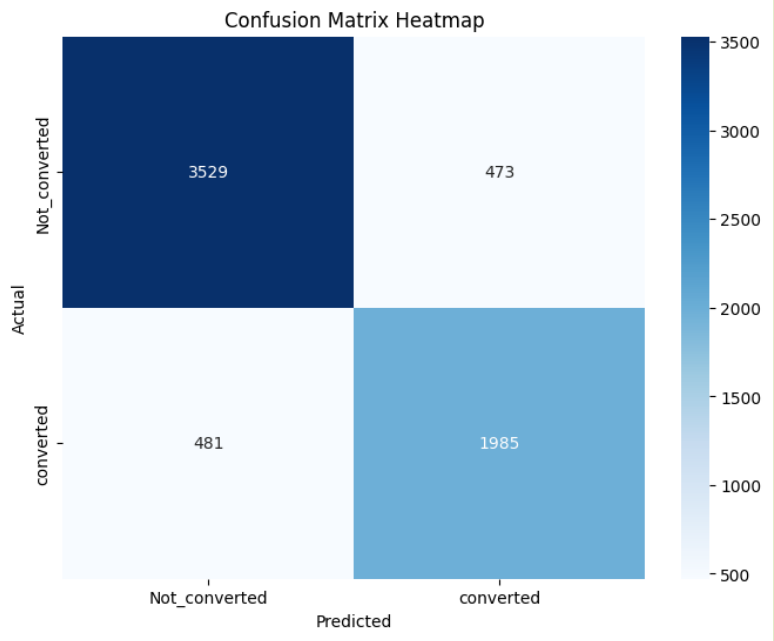


Model Evaluation

1. Accuracy:
    1. Our model accurately predicts lead conversion with approximately 85% accuracy.
2. Sensitivity (Recall):
    1. Our model correctly identifies around 80% of leads who actually converted.
    2. This means we're effectively capturing most of the leads who are likely to convert.
3. Specificity:
    1. Our model correctly identifies around 88% of leads who did not convert.
    2. This means we're effectively minimizing false positives, preventing wasted resources on leads who are unlikely to convert.
4. Precision:
    4. Our model has a precision of around 81%, which means that out of all the leads predicted to convert, around 81% actually did.
    5. This indicates that our model is quite good at identifying leads who are likely to convert.
5. AUC (Area Under the ROC Curve):
    5. The AUC of our model is approximately 0.92, which is considered very good.
    6. This indicates that our model has a strong ability to distinguish between leads who will convert and those who will not.

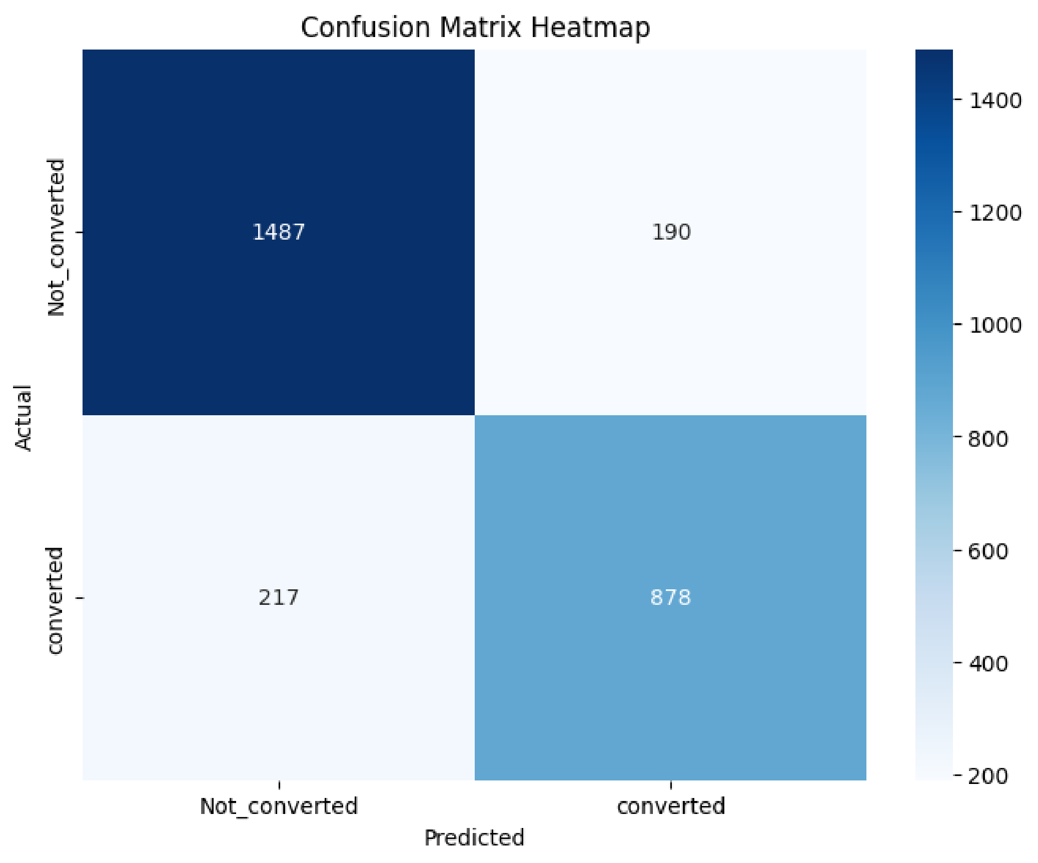






Results on test data

1. Accuracy:
   1. The model accurately predicts lead conversion with approximately 85% accuracy on the test data.
   2. This means that the model can correctly identify whether a lead will convert or not in 85% of cases.
2. Sensitivity (Recall):
   1. The model correctly identifies around 80% of leads who actually converted on the test data.
   2. This is a good indicator that the model is effectively capturing most of the leads who are likely to convert.
3. Specificity:
   1. The model correctly identifies around 88% of leads who did not convert on the test data.
   2. This means that the model is effectively minimizing false positives, preventing wasted resources on leads who are unlikely to convert.
4. Precision:
   1. The model has a precision of around 81% on the test data, which means that out of all the leads predicted to convert, around 81% actually did.
   2. This indicates that the model is quite good at identifying leads who are likely to convert.



Assignment Subjective Questions & Answers

1. Which are the top three variables in your model which contribute most towards the probability of a lead getting converted?

A) Top 3 Variables are:

- Total Visits: An increase in the total number of visits to the website has a significant positive effect on the likelihood of conversion.
- Total Time Spent on Website: More time spent on the website also strongly correlates with higher chances of conversion. This suggests that engaged leads who spend more time exploring are more likely to convert.
- Lead Source: Leads from Olark Chat & Welingak website are positively associated with conversion, implying that interactions via chat can enhance conversion chances.

2. What are the top 3 categorical/dummy variables in the model which should be focused the most on in order to increase the probability of lead conversion?

A) The top three categorical/dummy variables to focus on for increasing the probability of lead conversion are:

- Lead Quality_High in Relevance: Leads with this quality are most likely to convert.
- Lead Quality_Low in Relevance: This variable also contributes positively with conversion as it has high coefficient
- Lead Source_Welingak Website: With a coefficient of 3.534517, this source is highly effective at driving conversions

3. X Education has a period of 2 months every year during which they hire some interns. The sales team, in particular, has around 10 interns allotted to them. So during this phase, they wish to make the lead conversion more aggressive. So they want almost all of the potential leads (i.e. the customers who have been predicted as 1 by the model) to be converted and hence, want to make phone calls to as much of such people as possible. Suggest a good strategy they should employ at this stage.

A) I would just to go with below strategy:

• Focus on Leads with High Lead Quality Scores: Prioritize leads with high "Lead Quality" scores (e.g., "High in Relevance" and "Low in Relevance").

o Reason: These categories have the highest positive impact on conversion probability, making them ideal targets for aggressive follow-up.

• Utilize Effective Lead Sources: Target leads from sources with high positive coefficients, such as "Welingak Website" and "Olark Chat".

o Reason: These sources have shown to significantly increase conversion likelihood, so focusing on them will be beneficial.

• Engage with Leads Showing Strong Indicators: Prioritize leads with high total visits and more time spent on the website.

o Reason: These behaviors strongly correlate with a higher chance of conversion, suggesting they are more engaged and likely to convert.

4. Similarly, at times, the company reaches its target for a quarter before the deadline. During this time, the company wants the sales team to focus on some new work as well. So, during this time, the

company's aim is to not make phone calls unless it's extremely necessary, i.e. they want to minimize

the rate of useless phone calls. Suggest a strategy they should employ at this stage.

A) I would just to go with below strategy:

• Focus on High Lead Scores: Call only leads with the highest predicted conversion probabilities (scores close to 1).

o Reason: These leads are most likely to convert, ensuring calls are more effective.

• Prioritize High-Quality Leads: Target leads with the best quality ratings (e.g., "High in Relevance").

o Reason: High-quality leads have a better chance of conversion.

• Minimize Calls Based on Recent Engagement: Contact leads showing recent high engagement metrics.

o Reason: Engaged leads are more likely to convert, making calls more worthwhile.

• Review and Adjust Strategy: Continuously monitor call outcomes and refine targeting criteria.

o Reason: Ensures calls are only made to leads with a high likelihood of conversion.

Summary & Recommendations

I would like to summarize that the model which was created is predicting the values with good significance level of 80% and this level is matching with CEO requirement.

Below are the highlights on the features on which the company needs to pay attention:

- Highly Predictive Variables: Variables such as total visits, total time spent on the website, lead origin, and lead quality are strong predictors of conversion. Focusing on these areas can help in targeting and nurturing leads more effectively.
- Negative Indicators: Negative coefficients associated with page views per visit, bounced emails, and certain specializations should be closely monitored. These factors may indicate areas where the company's approach needs adjustment or where additional support could be beneficial.

Thank you!