Lead Score Case Study

Presented by:

Sonam Tripathi

Soumya Shambhavi

Sowmya R

Highlights

- Problem Statement
- Data Preparation
- Data Imbalance ratio
- EDA (Univariate Analysis, Bivariate and Multivariate Analysis, Handling Outliers)
- Correlation between Target and other numeric variables
- Creating Dummy Variables , Performing Train-Test Split and Feature Scaling
- Model Building (Feature Selection Using RFE)
    - ROC Curve
    - Precision Recall Trade off
- Model Evaluation
- Conclusion
- Recommendations

Problem Statement

X Education, an online education platform, faces a low lead conversion rate of 30%. To address this issue, X Education seeks to identify the most promising leads, referred to as "Hot Leads." The goal is to increase this to 80% by identifying and scoring "Hot Leads" based on various factors (demographics, behavior, preferences).

Business Objectives:

- Build a logistic regression model to assign a lead score between 0 and 100.
- Higher scores should indicate a higher likelihood of conversion (hot leads).
- Improve sales efficiency and conversion rates, driving revenue.

Data Preparation

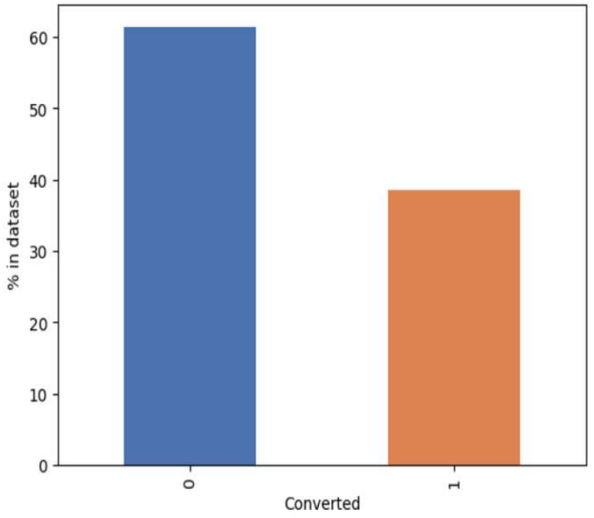
- Imported the necessary libraries for data manipulation, analysis, visualization, and model building. Loaded the lead dataset for further exploration and analysis.
- Identified and treated columns with 'Select' labels as null values.
- Dropped columns with >35% missing data to minimize analysis impact.
- Imputed missing values in categorical columns with 'Unknown' or mode and numeric columns with mode or median.
- Ensured no missing values post-imputation.
- Applied outlier capping to reduce extreme value impact.
- Dropped irrelevant columns and visualized remaining features.
- Removed columns with unique values, as they offer no modeling value.

Data Imbalance Ratio

- Lead Conversion Rate: 38.53%
- The dataset exhibits a moderate imbalance with a higher proportion of non-converted leads compared to converted ones.

```
(sum(lead_df['Converted'])/len(lead_df['Converted'].index))*100
```

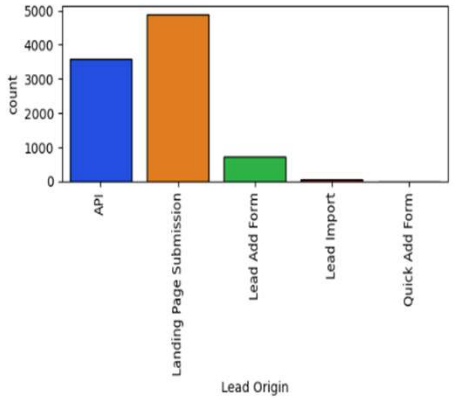
```
38.53896103896104
```

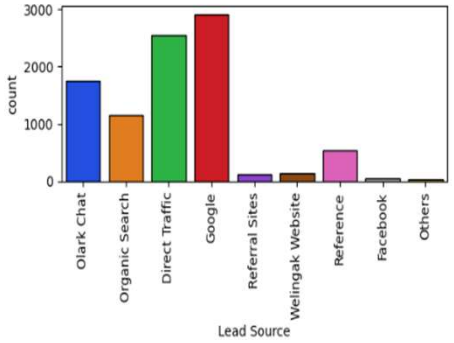


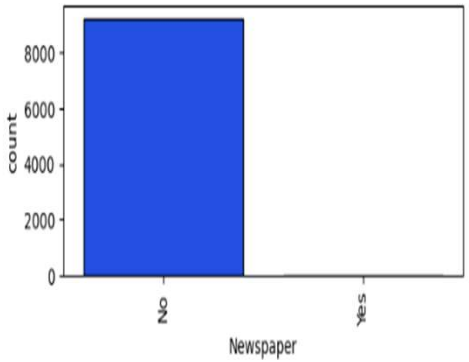
EDA :Univariate analysis

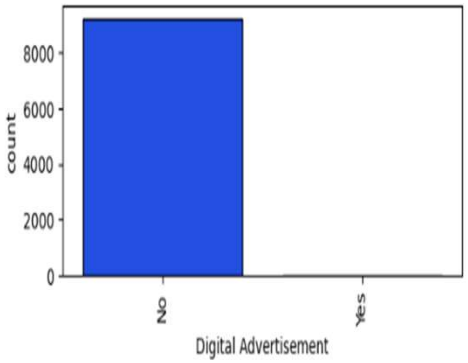
Insights:

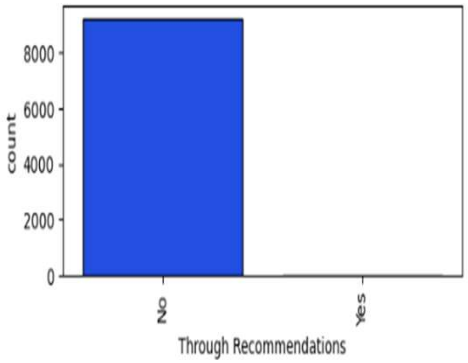
- The highest number of leads originated from Landing Page Submission, followed by API.
- Lead Add Form had a significant number of leads.
- We can observe that many of columns mostly have unique values('No').
- So, we can drop these columns as no inference can be drawn from those columns.







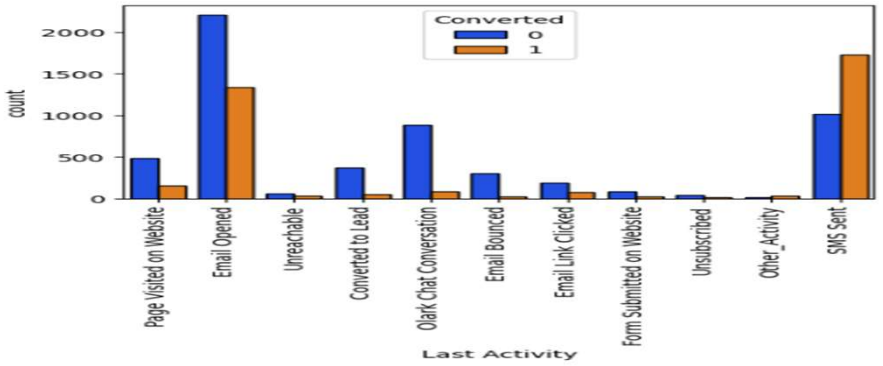


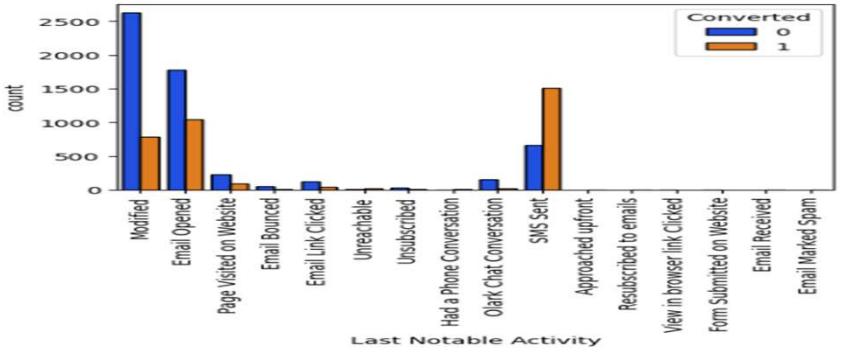


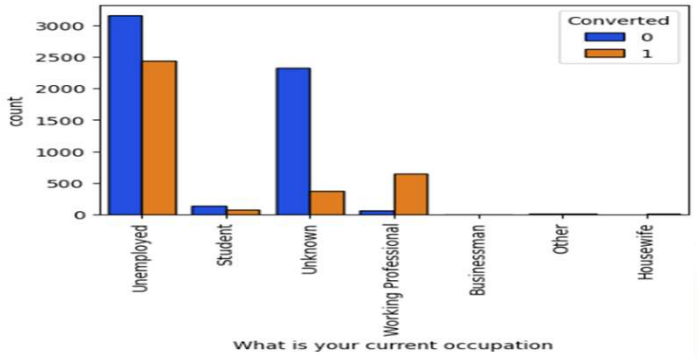
Bivariate analysis

Insights:

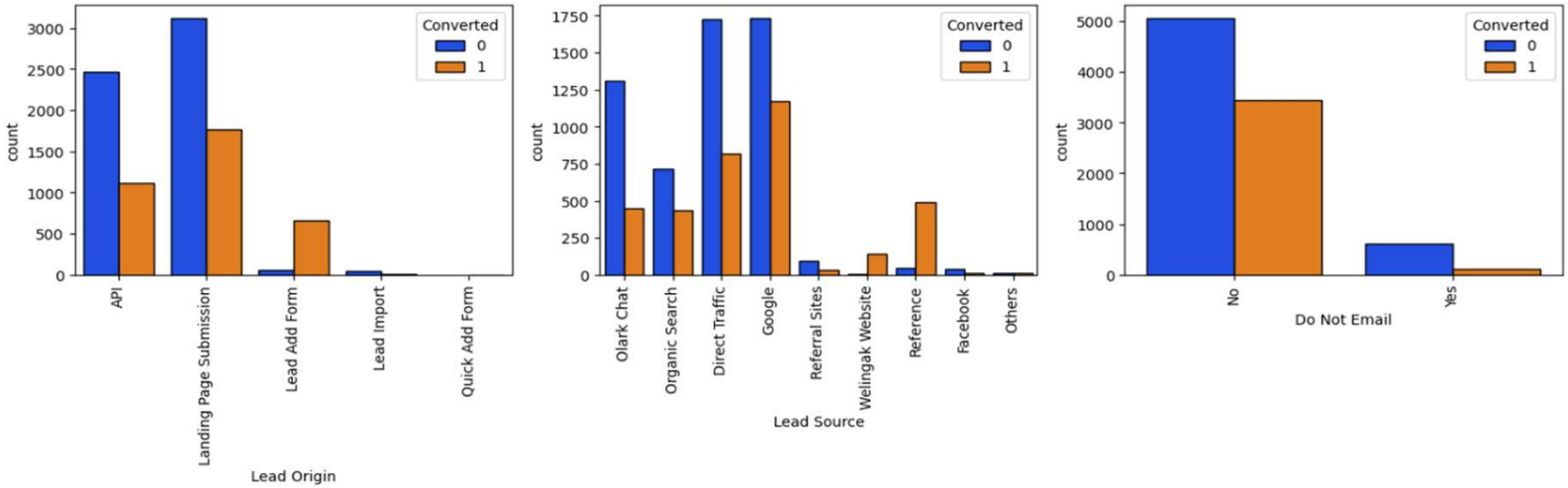
- for most of the lead ,Email Opened is the last activity.
- The conversion rate is high for leads with last activity as SMS Sent.
- The highest conversion rate is for 'Working Professional'.
- High number of leads are generated for 'Unemployed' but conversion rate is low.







Bivariate analysis



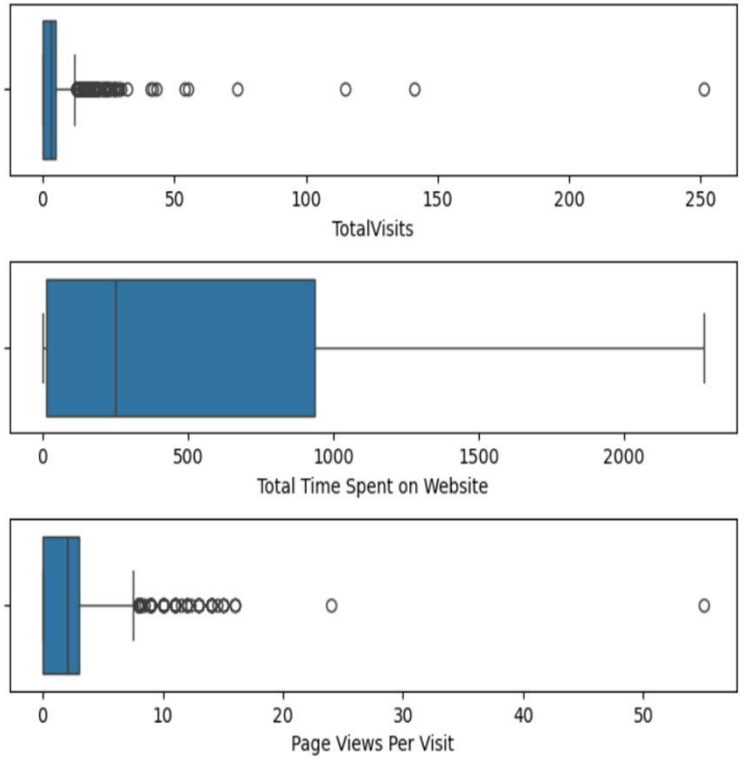
Insights:

- The conversion rates for API and Landing Page Submission are between 30-35%, but the number of leads generated from these sources is significant.
- The Lead Add Form, on the other hand, has a conversion rate of over 90%, though the number of leads generated from it is relatively low.
- The Lead Import category generates very few leads.
- Google and Direct traffic channels contribute the highest number of leads. However, the conversion rate is higher for reference leads and leads generated through the Welingak website.

Visualising Numerical Variables and Outlier Treatment

Insights:

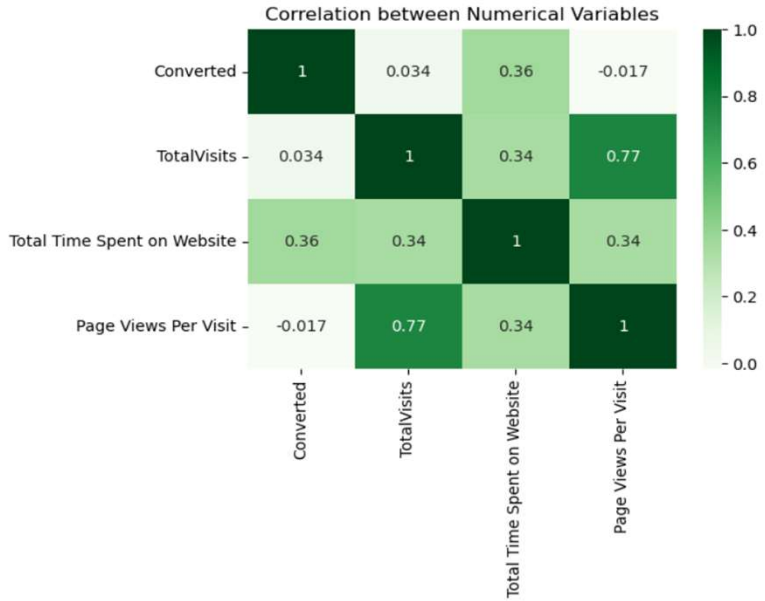
- From the boxplots, we can see that there are outliers present in the variables.
- For 'TotalVisits', the 95% quantile is 10 whereas the maximum value is 251. Hence, we should cap these outliers at 95% value.
- There are no significant outliers in 'Total Time Spent on Website' .
- For 'Page Views Per Visit', similar to 'TotalVisits', we should cap outliers at 95% value.



Correlation between Target and other numeric variables

Insight:

- The heatmap shows correlation between TotalVisits and Page Views Per Visit.
- We will further investigate multicollinearity using **VIF**.
- If the **VIF** of a feature is greater than 5, we will drop one of the correlated features to reduce multicollinearity.



Data Preprocessing: Dummy Variables, Train-Test Split and Feature Scaling

- Converting binary variable (Yes/No) to 0/1
- **Dummy Variables**: Convert categorical variables into numerical format using one-hot encoding .
- **Train-Test Split**: Split data into training and testing sets (70-30) to evaluate model performance.
- **Feature Scaling**: Normalize numerical features using Min-Max Scaler to scale values within a specific range (0 to 1) for consistent model training and improved performance.

Model Building

Feature Selection using RFE:

- Applied Recursive Feature Elimination (RFE) to select 15 key features.
- **Manual Feature Removal**: Dropped columns with p-values > 0.05 to improve model performance.
- **Multicollinearity Check**: Verified no multicollinearity using Variance Inflation Factor (VIF)
- final model retained 13 features.

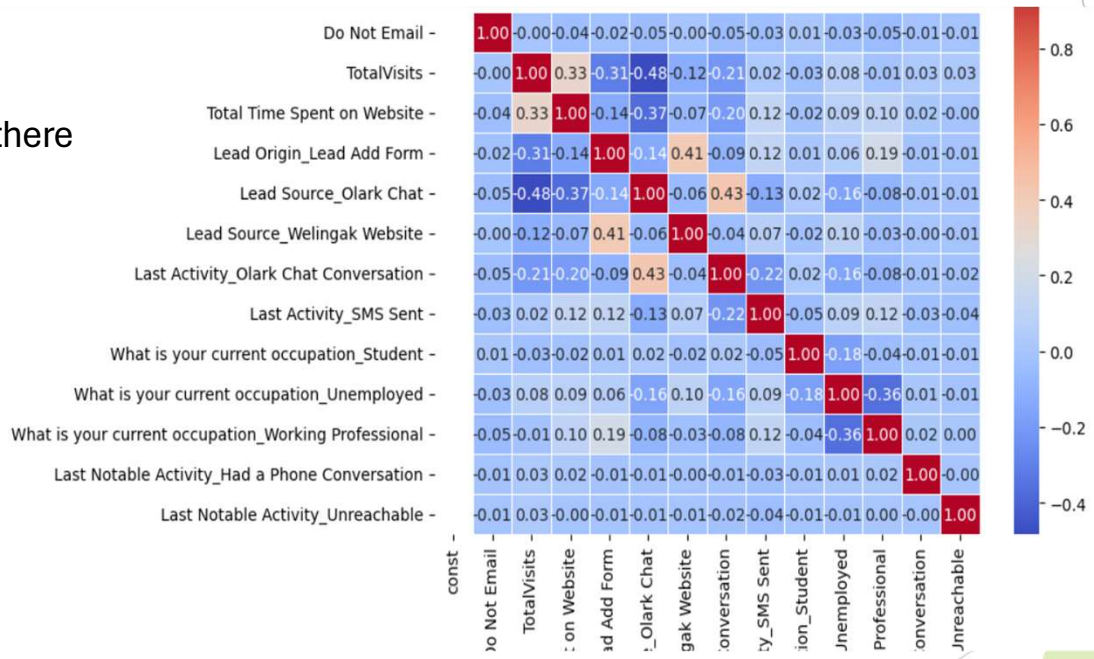
Threshold Selection:

- Initially set cutoff at 0.5 but observed low sensitivity.
- Adjusted threshold to **0.34** based on ROC curve analysis, improving model performance.

Correlation between final predictor variables

Insight:

- We can see from the heatmap, there is no such multicollinearity exist
- So we can proceed with these features.

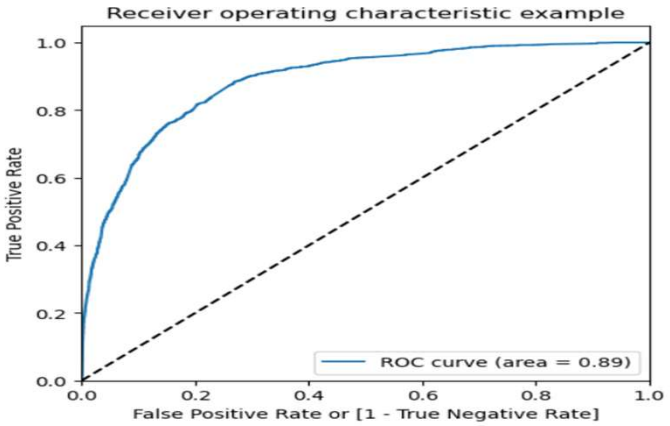


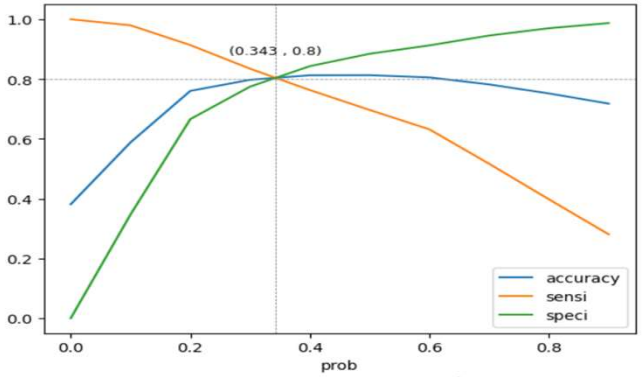
Model Evaluation: ROC Curve

ROC Curve - AUC Insight

AUC (Area Under Curve): 0.89

- AUC of **0.89** indicates **strong model performance**, with a high ability to distinguish between positive and negative cases.
- **89% probability** of correctly ranking a positive case higher than a negative case.
- At a threshold of 0.34, the model achieves **80.36% accuracy, 79.76% sensitivity**, and **80.73% specificity**.

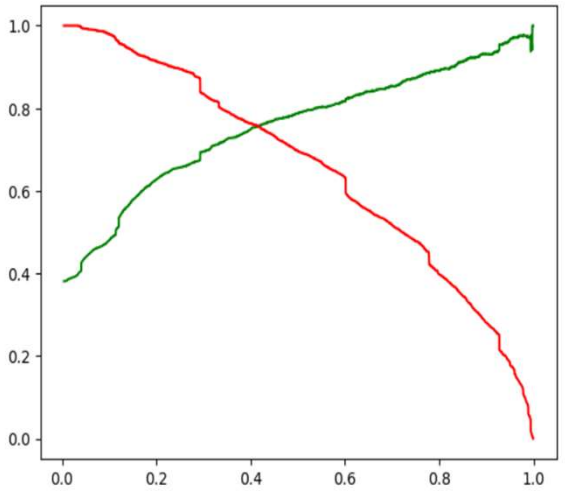




Precision Recall Trade-off

Observation:

- The precision-recall trade-off suggests 0.41 as the optimal threshold for balancing both precision and recall.
- With a threshold of 0.41, accuracy improves slightly to 81.43%, but sensitivity drops to 75.99%, while specificity increases to 84.78%.
- Since higher sensitivity is important for capturing more positive cases (which aligns with a higher conversion rate).
- **Threshold 0.34 is the better choice for prioritizing true positives**.



Model Evaluation on train data

Train data Evaluation (Threshold: 0.34):

- **Accuracy**: 80.36%
- **Sensitivity**: 79.76%
- **Specificity**: 80.73%

Insight:

- With the threshold value set at 0.34, the model achieves an accuracy of 80.36%, a sensitivity of 79.76%, and a specificity of 80.73%.
- This indicates a strong performance, with a good balance between correctly identifying positive cases (sensitivity) and correctly identifying negative cases (specificity).
- The overall accuracy suggests that the model is effectively distinguishing between the two classes.

Model Evaluation on test data

Test Data Evaluation (Threshold: 0.34)

- **Accuracy**: 80.91%

  Consistent model performance on test data.

- **Sensitivity**: 80.54%

  High true positive rate, correctly identifying leads likely to convert.

- **Specificity**: 81.15%

  Strong true negative rate, accurately identifying leads unlikely to convert.

Business Insight:

- Reliable model for **lead prioritization**, helping to target high-conversion leads efficiently.

Lead Score Calculation

Lead scores were calculated by applying a **0.34 threshold** to the **conversion probability**.

Lead Score Insight (Threshold: 0.34)

- Higher Sensitivity : The 0.34 threshold prioritizes identifying leads likely to convert, achieving **80.54% sensitivity** on test data.
- **Consistent Performance:** Both training and test data show consistent performance in identifying positive leads, ensuring that 80% of potential converters are correctly captured.
- The lead score calculation with a **0.34 threshold** helps prioritize the most promising leads for conversion while maintaining a reliable overall model performance.

Conclusion

Strong Model Performance:

- The model demonstrates high performance with an **accuracy** of **80.91%**, **sensitivity** of **80.54%**, and **specificity** of **81.15%** on test data, indicating good generalization and reliable predictions for lead conversion.

Key Features Driving Conversion:

- **Total Time Spent on Website (4.4841)**: Increased engagement is strongly correlated with higher conversion rates.
- **Lead Origin_Lead Add Form (3.7886)**: Form submissions are more effective for capturing high-converting leads.
- **Occupation_Working Professional (3.6383)**: Targeted campaigns for professionals are likely to yield higher conversion rates.

Additional Insights:

- **Phone Conversations** and **Olark Chat** interactions are strong predictors of conversions.
- **Leads from Welingak Website** have higher conversion potential.
- **SMS follow-ups** enhance lead conversion likelihood.

Recommendations

Strategy for X Education Company:

- **Enhance User Engagement**: Increase time spent on the website with interactive features and personalized content.
- **Optimize Lead Forms**: Improve and streamline lead capture forms for better quality leads.
- **Target Working Professionals**: Tailor campaigns for professionals with flexible schedules and certifications.
- **Leverage Communication Channels**: Prioritize follow-ups through phone calls and live chat for higher conversions.
- **Focus on High-Performing Lead Sources**: Invest more in leads from the **Welingak Website** for better conversion potential.
- **Utilize SMS Follow-ups**: Increase conversion rates with timely SMS follow-ups.

Implementing these strategies will help X Education increase lead conversion rates and achieve business goals.

Thank You