LEAD SCORIGN CASE STUDY

<u>Team Members</u>:
Lavan Verma
Varun Seth

upGrad
RAHO AMBITIOUS

Executive
PG Program
in Data Science

PROBLEM STATEMENT

X Education struggles with a low lead conversion rate despite generating numerous leads daily. To boost efficiency, they seek to identify high-potential leads, or 'Hot Leads,' who are more likely to convert into paying customers.

They aim to develop a predictive model assigning lead scores to prioritize engagement efforts. With a target conversion rate of 80%, the company aims to bridge the gap between lead acquisition and conversion, optimizing resources and driving business growth.

OBJECTIVE TO ACHIEVE

- to assist X Education in choosing the most promising leads (also known as "Hot Leads"), or the prospects with the highest likelihood of becoming paying clients.
- To create a logistic regression model that will allow the business to target potential leads by giving each lead a score between 0 and 100.

The objective is thus classified into the following sub-goals :

01 Develop a logistic regression model for lead conversion probability prediction.

02 Determine a probability threshold to predict lead conversion. Leads surpassing this threshold are predicted as converted, while those below it are not.

03 Calculate the Lead Score which signifies the probability of conversion with a high accuracy

ANALYSIS APPROACH : A DATA-DRIVEN JOURNEY

Our analytical strategy meticulously uncovers layers in our lead generation and conversion process. Deep data dives reveal key leverage points for substantial conversion rate and ROI improvements. These insights drive informed decisions and strategic marketing adjustments.

Data Cleaning and Preparation:

- Handled outliers and invalid terms.
- Addressing missing values via imputation or exclusion, depending on their influence on the dataset.
- Standardizing categorical data for uniformity throughout the dataset.

Exploratory Data Analysis (EDA):

- Visualization of key metrics to identify trends and anomalies.
- Examining lead sources and origins to grasp their influence on conversion rates.
- Conducting correlation analyses to unveil connections between various variables and conversion results.

Feature Engineering:

- Developing composite metrics like engagement scores from website interactions.
- Categorizing leads by behavior and demographics for precise analysis.

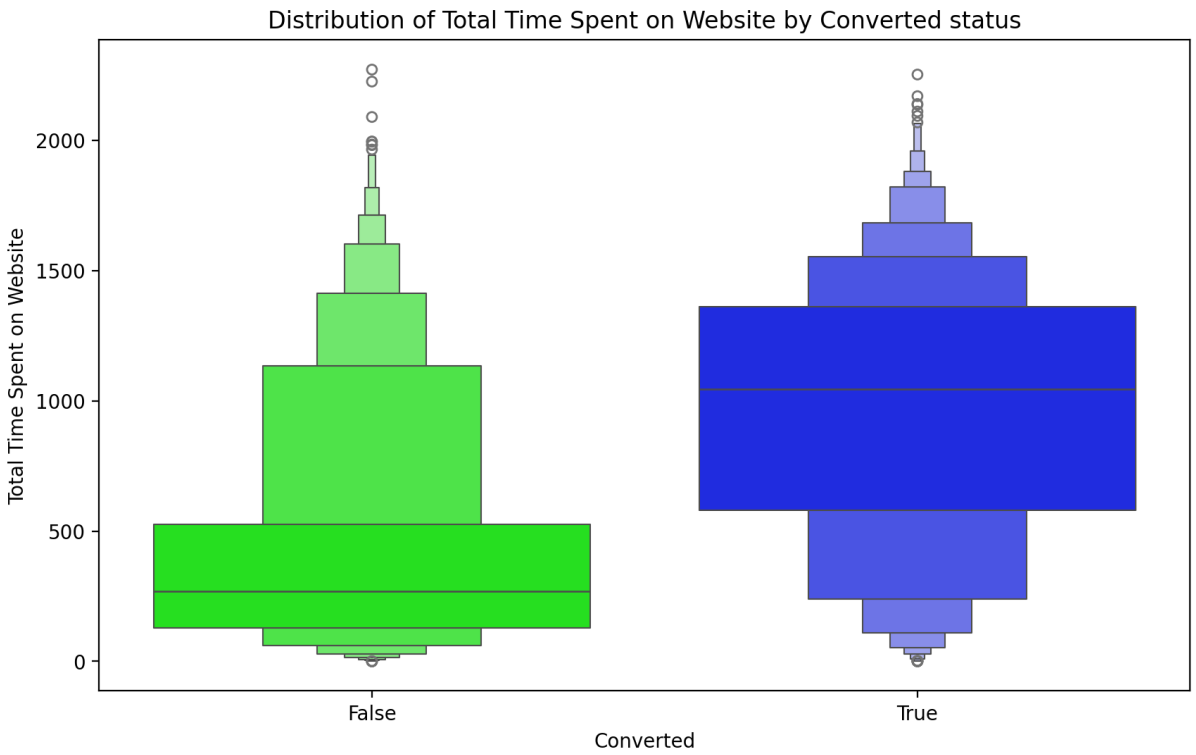
Predictive Modeling:

- Crafting classification models to predict the likelihood of lead conversion.
- Conducting feature importance analysis to identify the most predictive variables for conversion.

Impact Analysis:

- Assessing marketing channels and campaigns for lead quality and effectiveness
- Analyzing engagement metrics to gauge their impact on conversion rates.

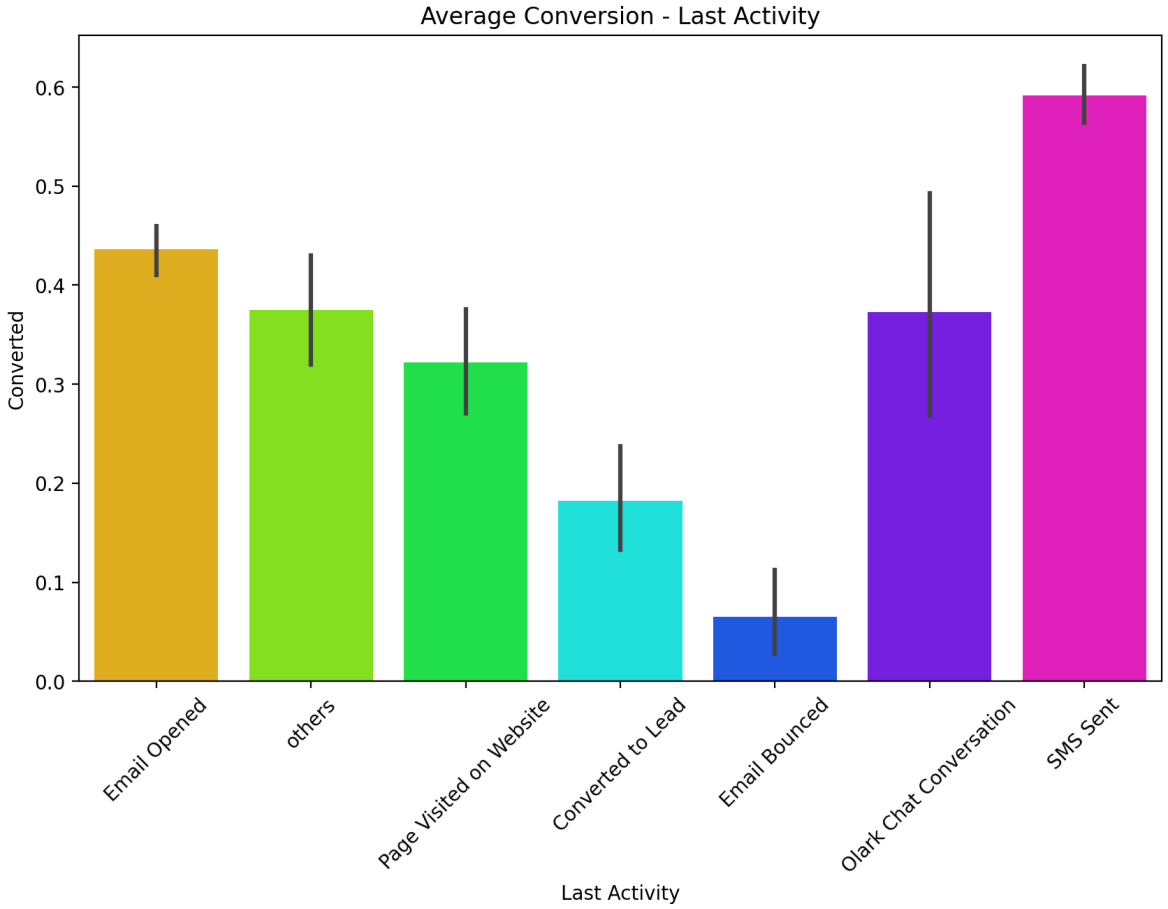
KEY “Data-Driven” INSIGHTS

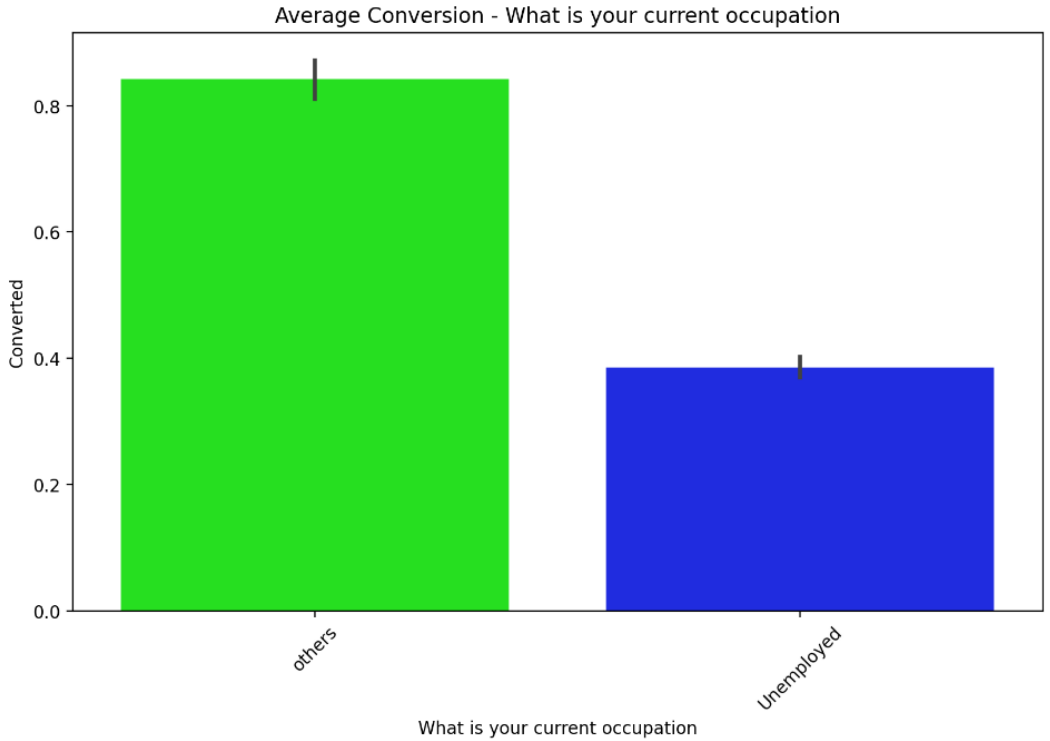


- Converted users (True) show a higher median total time spent compared to non-converted users (False). This is visually clear from the median line's position within each box. Additionally, the interquartile range (IQR) for converted users is wider, indicating greater variability in total time spent among converted users.
- Based on the visualization, increased site engagement (time spent) appears linked to higher conversion rates. However, it's crucial to understand that correlation doesn't imply causation. This chart doesn't encompass all factors influencing conversion probability.

- The key takeaway from this chart is the moderate positive link between total time spent on the website and conversion, emphasizing the significance of compelling content and user-friendly experience in lead conversion.
- Weak correlations with other variables imply that they may not individually predict conversion effectively and should be evaluated alongside other features in predictive modeling.

KEY “Data-Driven” INSIGHTS... *contd.*

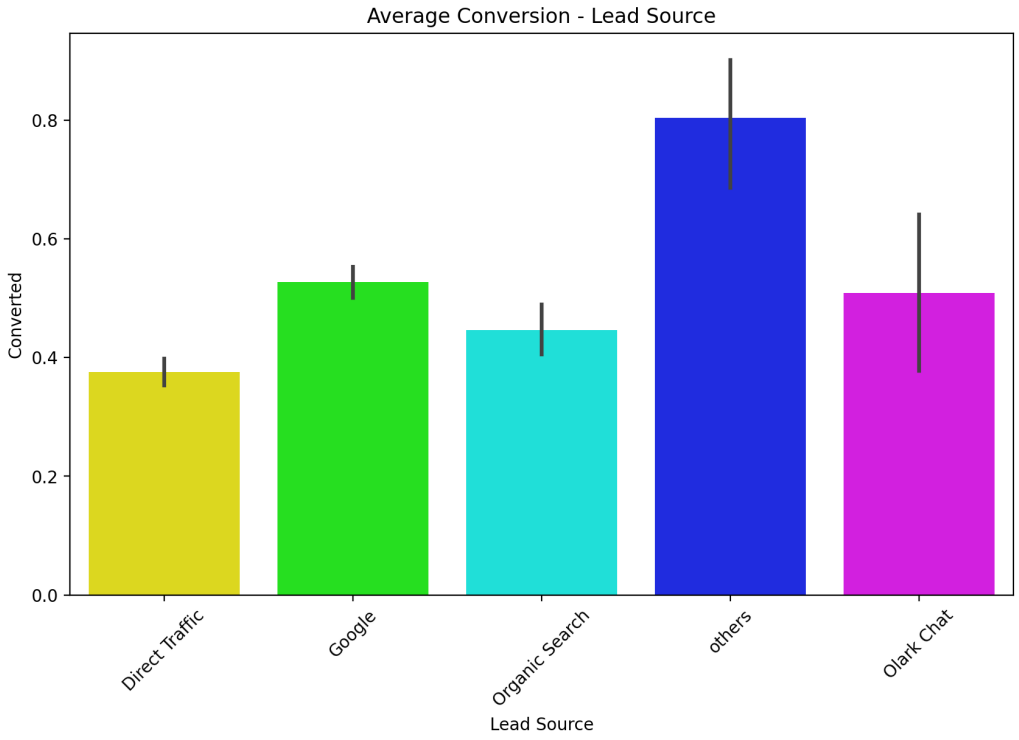


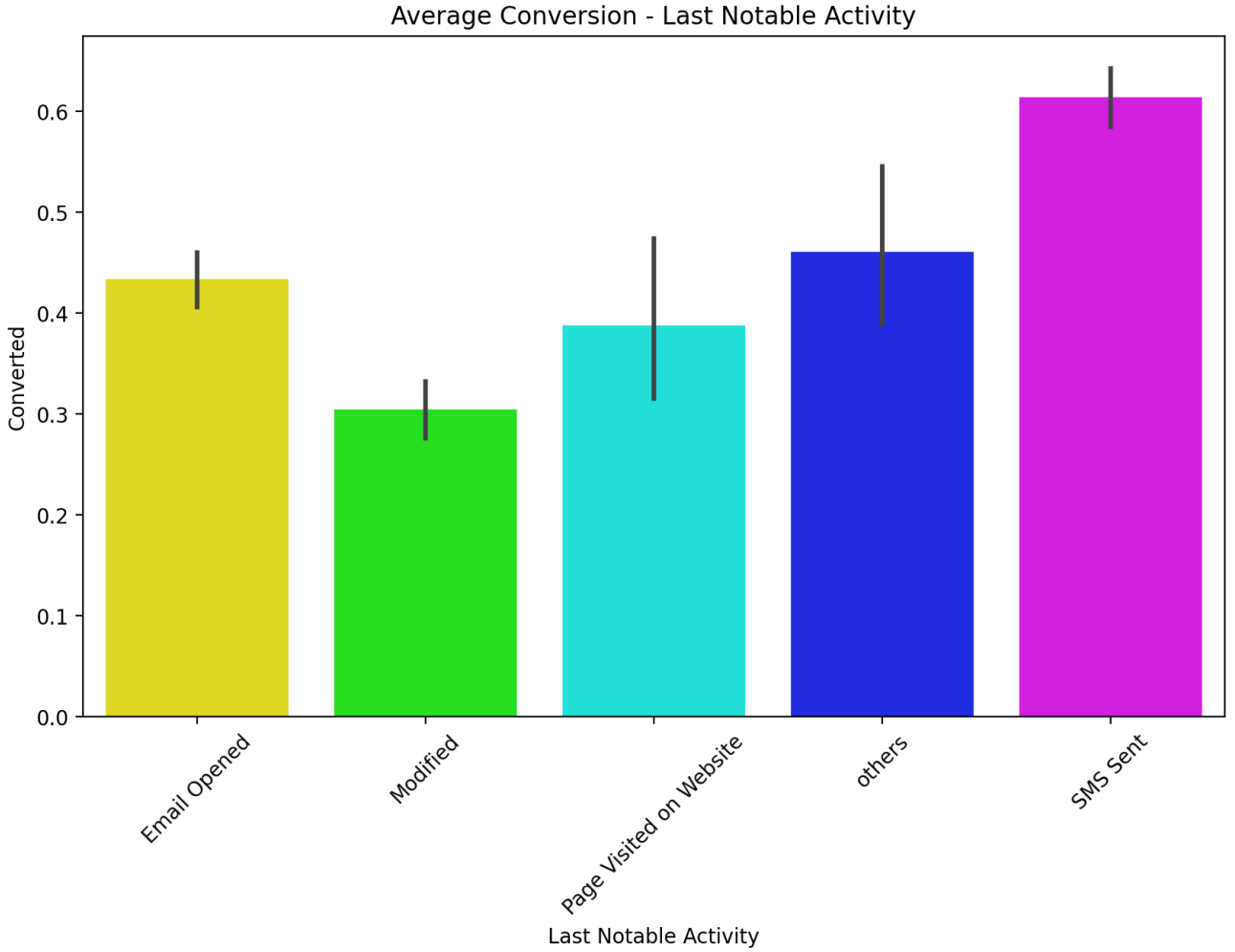


- Here we observe that the conversion rate is significantly different among different categories of “last Activity”. For example, “SMS Sent” has a very high conversion rate. While “Email bounced” has very low conversion rate.

- Employment status strongly influences conversion rates, particularly for individuals categorized as "others," who exhibit higher conversion likelihood. This insight underscores the need for tailored marketing tactics and sales resource distribution to capitalize on segments with superior conversion rates.

KEY “Data-Driven” INSIGHTS... *contd.*

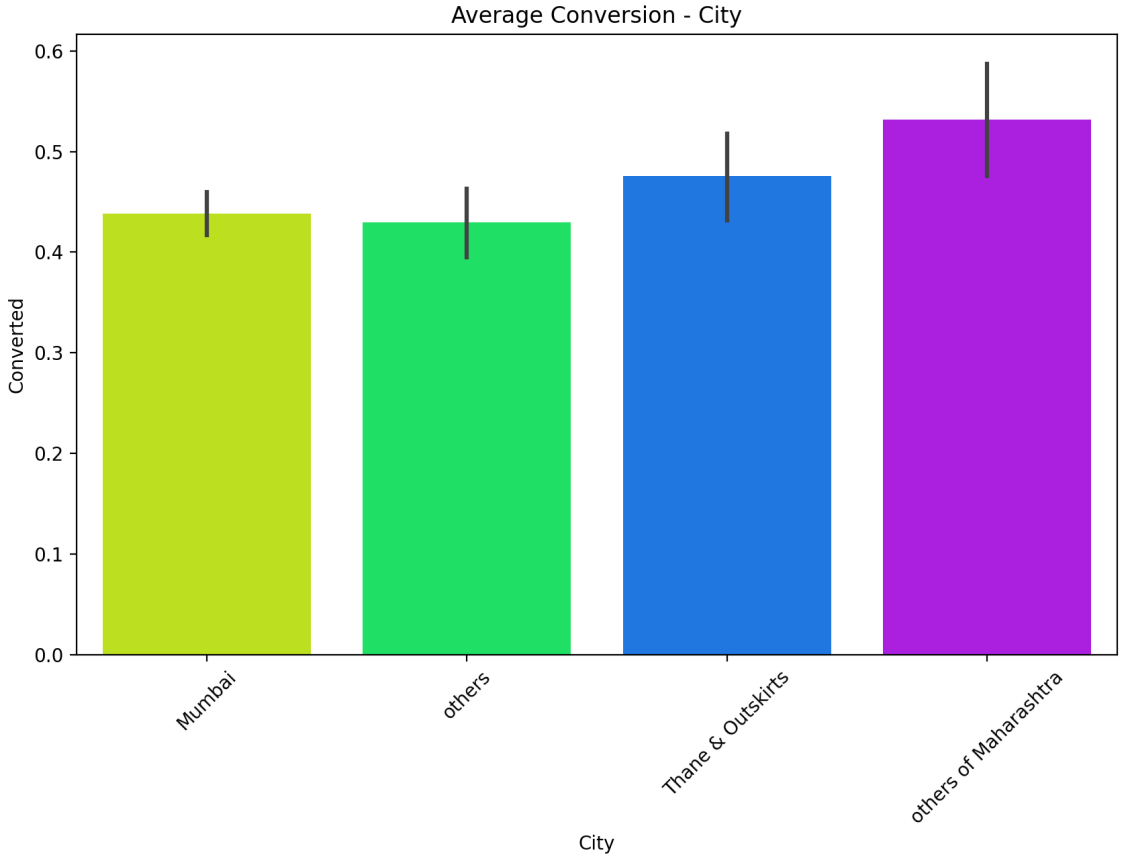




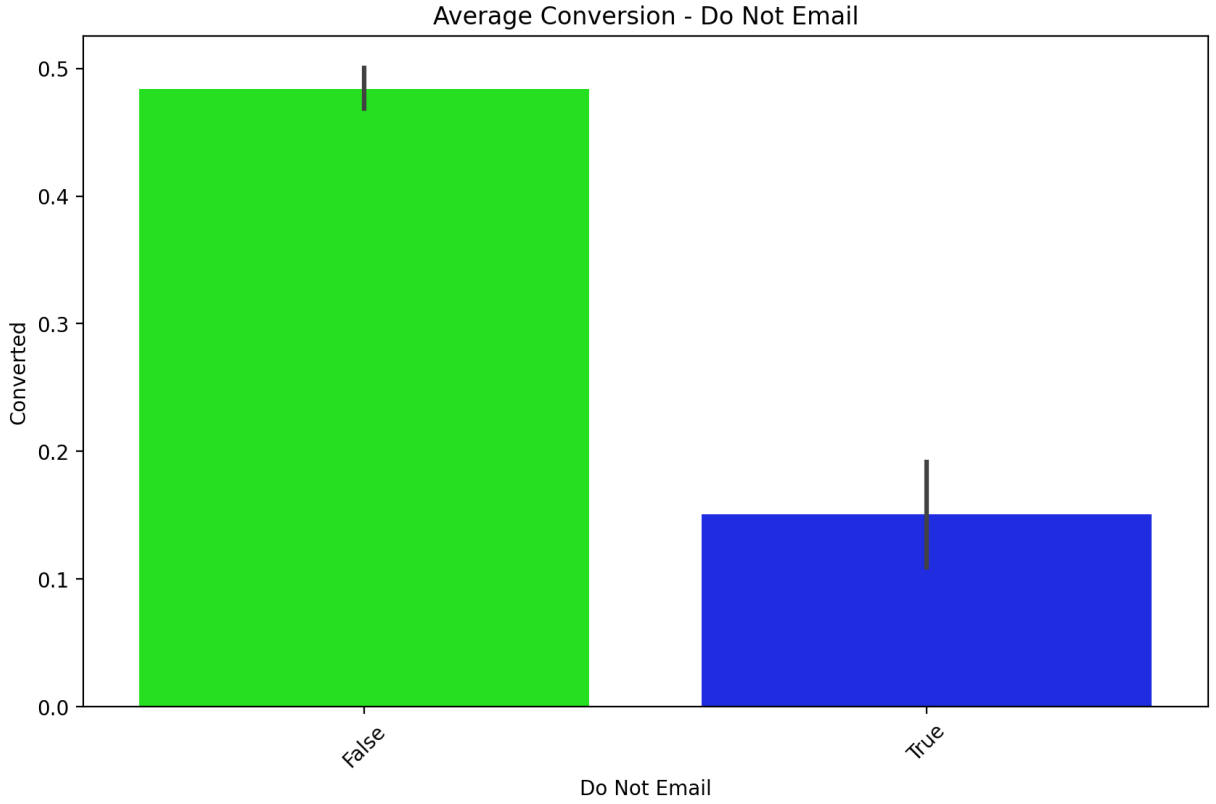
- The bar chart above depicts lead conversion rates per source, highlighting each source's impact on overall conversion success. Analyzing these rates reveals the most effective channels for generating convertible leads, aiding strategic resource allocation in marketing for optimal conversion impact.

- Based on our analysis, it seems that SMS Sent is the most potent activity driving conversions, followed by Email Opened, Page Visited on Website, and Modified. The effectiveness of SMS likely stems from its direct and personal nature, triggering prompt responses from leads.

KEY “Data-Driven” INSIGHTS... *contd.*

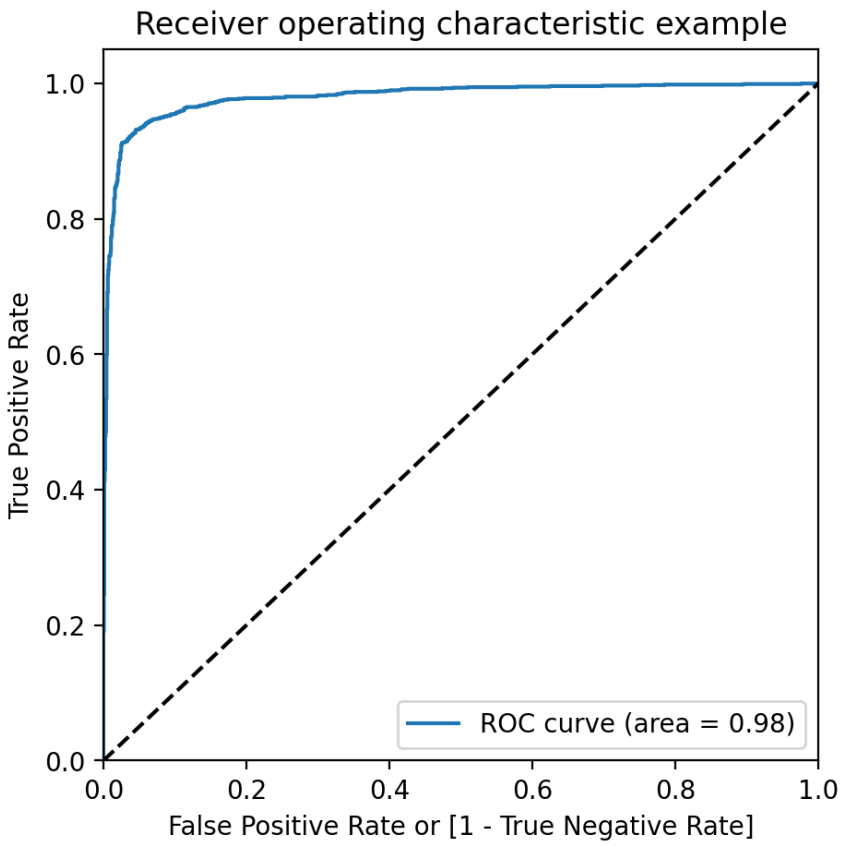


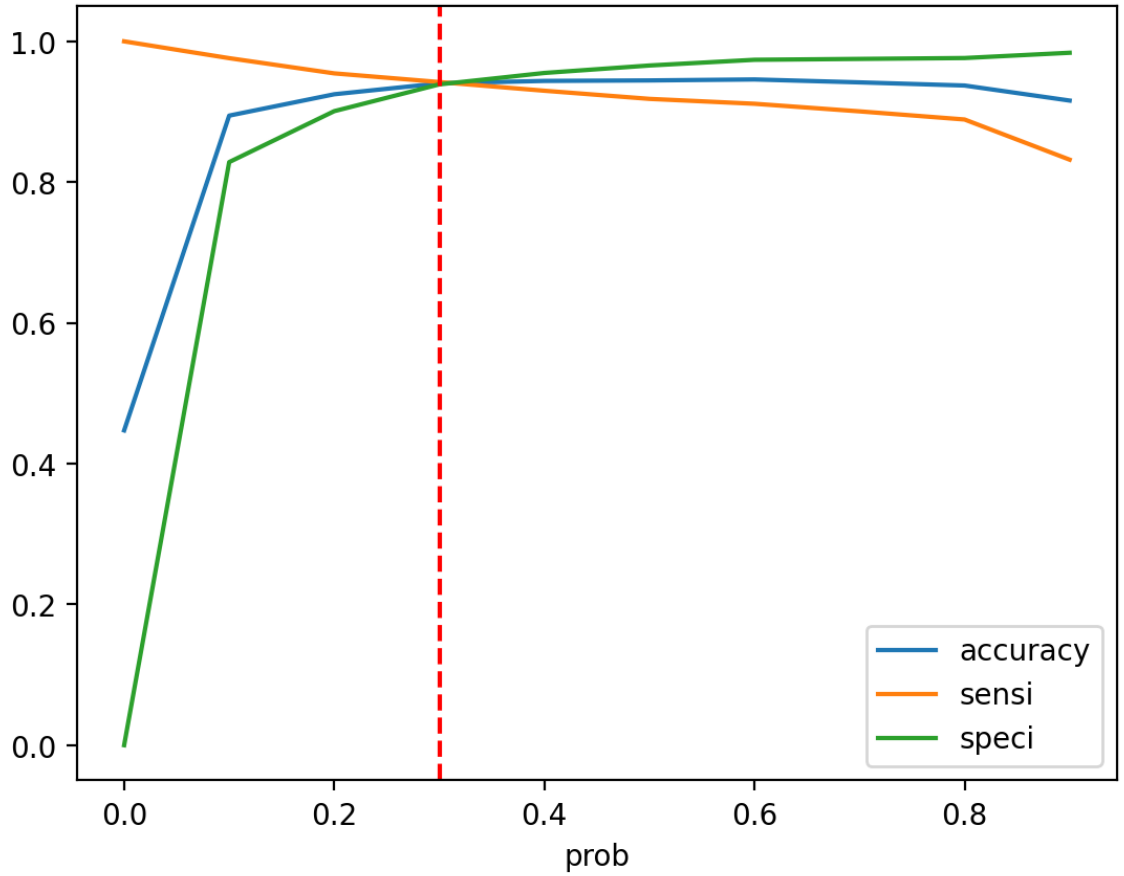
- Leads from Non-tier-1 cities seem to be more interested in distance learning programs as their conversion rate is higher.



- Leads who mark “do not email” are significantly less likely to be converted. We can include more such fields in the form, where user can themselves share their level of interest in the course.

ROC CURVE





- The AUC is 0.98, nearing 1, signaling exceptional model performance. Its high value implies strong capability in discerning between positive and negative classes accurately.

- It appears that as the threshold increases, specificity increases (fewer false positives), but sensitivity decreases (more false negatives). Conversely, as the threshold decreases, sensitivity increases, but specificity decreases. The optimal threshold should balance these metrics according to the cost of false positives and false negatives for the particular application.

PREDICTIONS

Predictions

In [322…]
```
X_test = sm.add_constant(df_test)

X_test = X_test[lm.model.exog_names]

y_test_pred = lm.predict(X_test)
y_train_pred = lm.predict(sm.add_constant(X_train_new))
```

In [323…]
```
y_test_pred_df = pd.DataFrame({
    'Converted': y_test.values,
    'Probability': y_test_pred}
)
y_test_pred_df.head()
```

Out[323…]

	Converted	Probability
8826	0	0.093393
1016	0	0.130365
3793	0	0.007424
9042	0	0.091903
4819	1	0.405301

In [324…]
```
# Substituting 0 or 1 with the cut off as 0.5
y_test_pred_df['Prediction'] = y_test_pred_df['Probability'].apply(lambda x: int(x > THRESHOLD) )
y_test_pred_df.head()
```

Out[324…]

	Converted	Probability	Prediction
8826	0	0.093393	0
1016	0	0.130365	0
3793	0	0.007424	0
9042	0	0.091903	0
4819	1	0.405301	1

In [325…]
```
print('Test model accuracy is', round(metrics.accuracy_score(y_test_pred_df['Converted'], y_test_pred_df['Prediction']), 4
```

```
Test model accuracy is 91.59 %
```

In [326…]
```
# Same as before, just kept here for completition of this section.
print(classification_report(y_test, (y_test_pred > THRESHOLD)))
```

```
              precision    recall  f1-score   support

           0       0.93      0.91      0.92       395
           1       0.90      0.92      0.91       330

    accuracy                           0.92       725
   macro avg       0.91      0.92      0.92       725
weighted avg       0.92      0.92      0.92       725
```

- The test model achieves a high accuracy rate of 91.59%, indicating effective lead conversion prediction based on given features.
- Precision and recall scores for both classes (0 and 1) exceed or approach 0.90, demonstrating high reliability in model predictions. The model exhibits both precision (accurate true positive predictions) and sensitivity (effective true positive identification).
- The f1-score, a blend of precision and recall, remains elevated for both classes, affirming the model's balanced performance in achieving precision and recall objectives.
- Based on the model's strong accuracy, precision, and recall scores, it appears robust and effective in lead conversion prediction. However, despite encouraging accuracy, it's vital to consider data context and the impact of false positives versus false negatives. Depending on business needs, adjusting the threshold (currently assumed at 0.5) may optimize for either precision or recall, prioritizing either lead quantity or quality.

Recommendations

- Maximize Engagement Focus: Utilize model-generated lead scores to prioritize engagement with high-potential leads. Concentrate resources and marketing efforts on leads scoring above the 30% probability mark for optimal conversion rates.
- Continual Monitoring and Model Refinement: Regularly assess the model's performance and adjust it as more data is gathered or market dynamics evolve. This guarantees sustained accuracy and relevance.
- Expand Data Exploration: Explore additional data points or external variables that could impact lead conversion, such as economic indicators or industry shifts, to enhance model precision.
- Deploy A/B Testing: Validate the efficacy of lead prioritization based on model scores through A/B testing on chosen lead segments. Analyze conversion rates and ROI to gauge the strategy's effectiveness.
- Improve Data Integrity: Enhance the data collection process to minimize gaps and ensure top-notch inputs for the model. This might involve refining lead capture forms, enhancing digital marketing platform integration, and providing comprehensive training to sales teams on data input.