LEAD SCORE PREDICTION MODEL

By,
Apurv Dhingra,
Anwin Aby George & Anupama Shrivastav

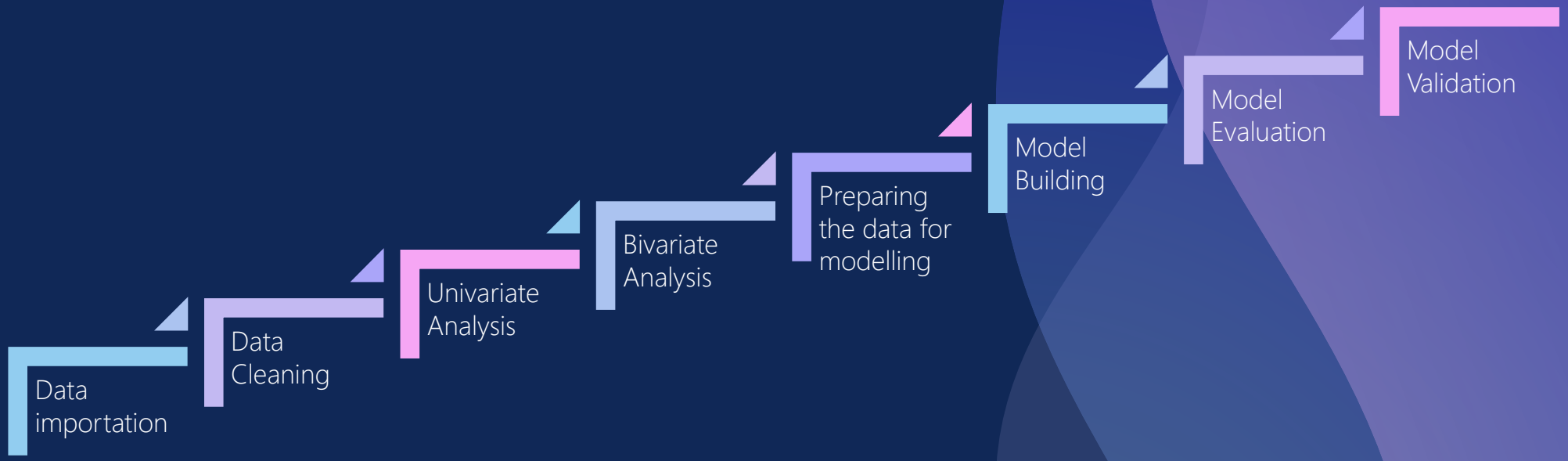
BUSINESS PROBLEM

X Education, an e-learning company, faces a critical challenge in optimizing its lead conversion process. Despite generating a substantial influx of leads from diverse channels, the current lead conversion rate stands at a mere 30%, significantly below the desired 80% conversion benchmark. X Education is determined to enhance the effectiveness of its lead conversion mechanism by identifying and prioritizing 'Hot Leads,' those prospects with the highest probability of conversion.

OBJECTIVE

The project's objective is to create a precise lead score prediction model for X Education. This model should accurately predict the conversion likelihood of each lead and help identify 'Hot Leads' with the highest potential for conversion. This initiative aims to significantly increase the lead conversion rate from its current 30% to the target of 80%. The project includes data analysis, feature engineering, and the development of a predictive model. Success will result in a streamlined lead conversion process, optimized resource allocation, and an improved conversion rate, leading to enhanced business performance and customer acquisition.

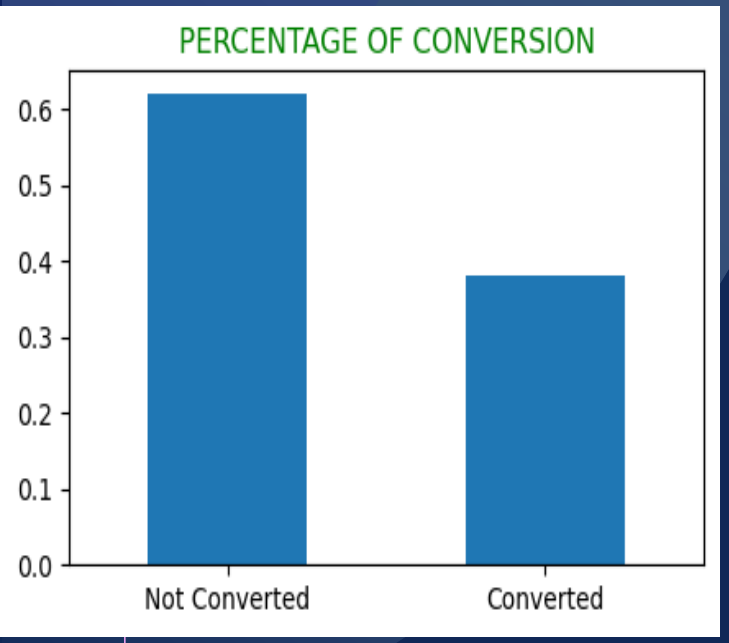
STEPS UNDERTAKEN DURING MODEL BUILDING

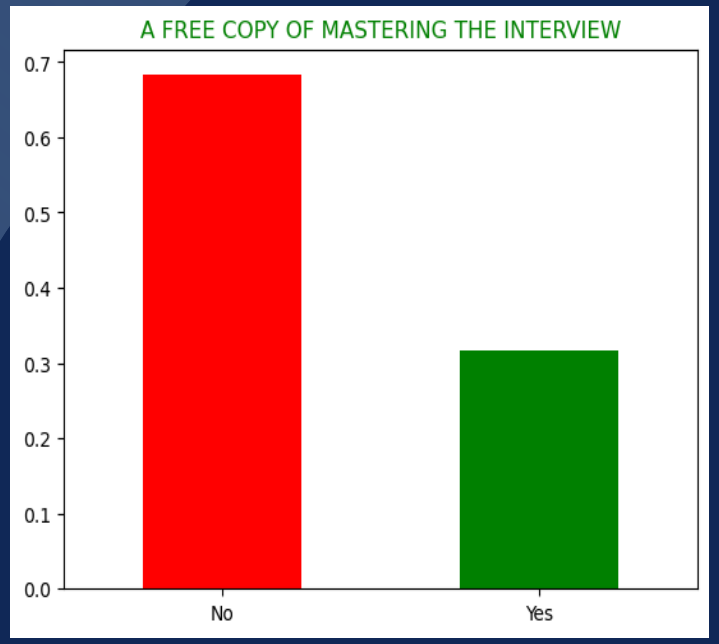


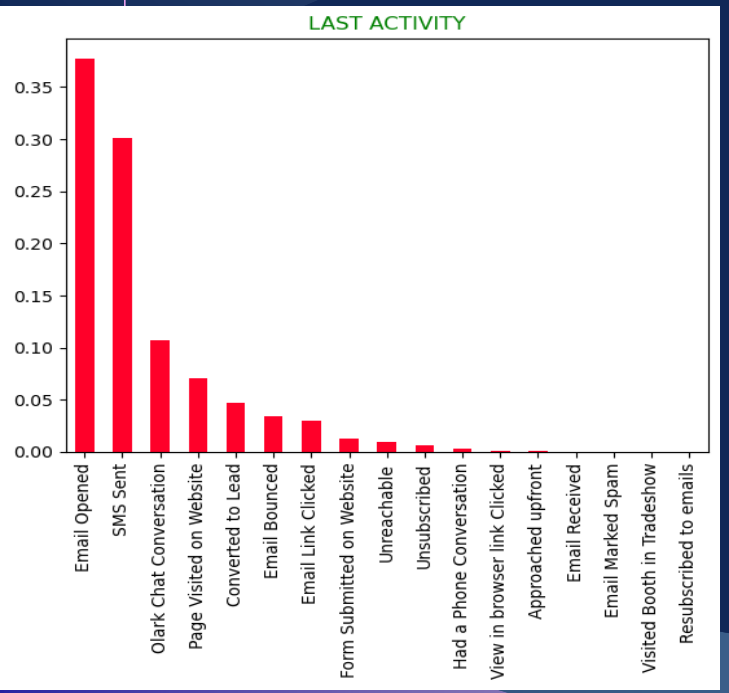
DATA CLEANING

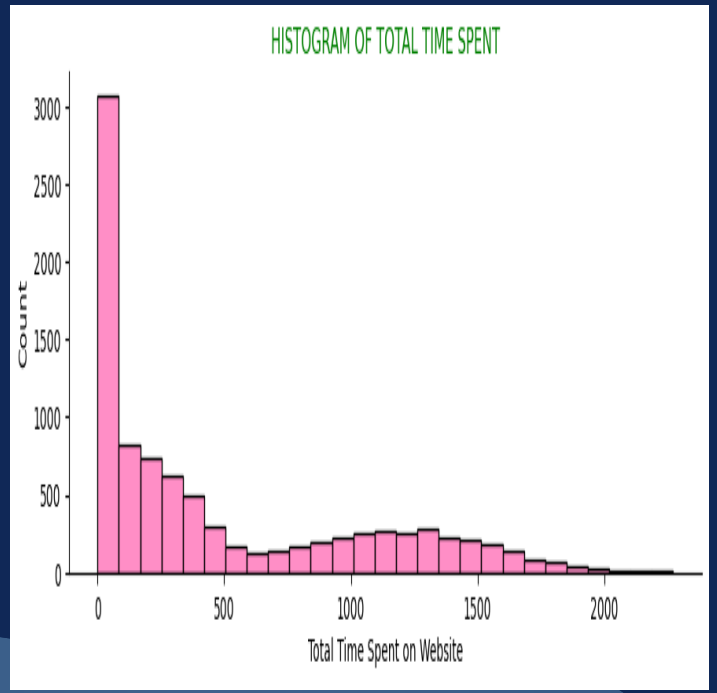
- Initially, we tackled the 'Specialization' column, which contained approximately 24% null values, treating them as ‘Select’ values to ensure data consistency.
- Recognized columns with more than 25% missing values and prioritized data quality by deciding to eliminate them.
- Dealt with the 'How did you hear about X Education' column, which had over 71% 'Select' entries, indicating non-informative data, and consequently, removed the column.
- Identified the lowest average conversion rates in Tier II cities, which constituted only 1% of leads, and decided to remove the 'City' column for data consistency.
- Ensured data completeness and consistency by filling missing values in the 'Lead Source' column with 'Google.'
- For data quality and consistency, null rows in the 'TotalVisits' column were removed.

UNIVARIATE AND BIVARIATE ANALYSIS

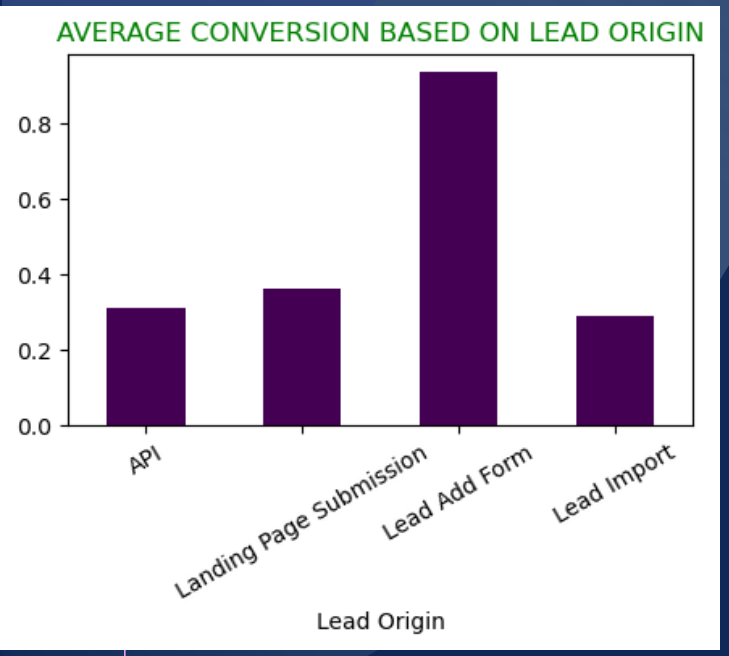


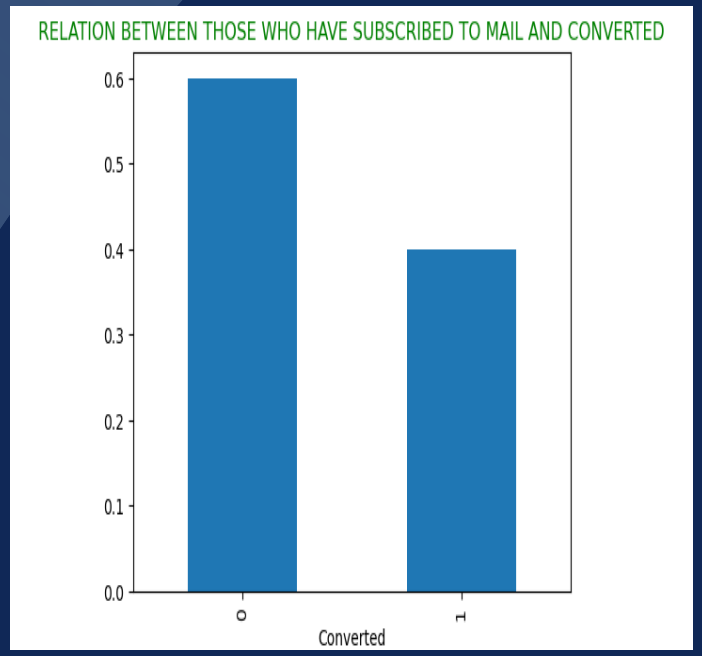


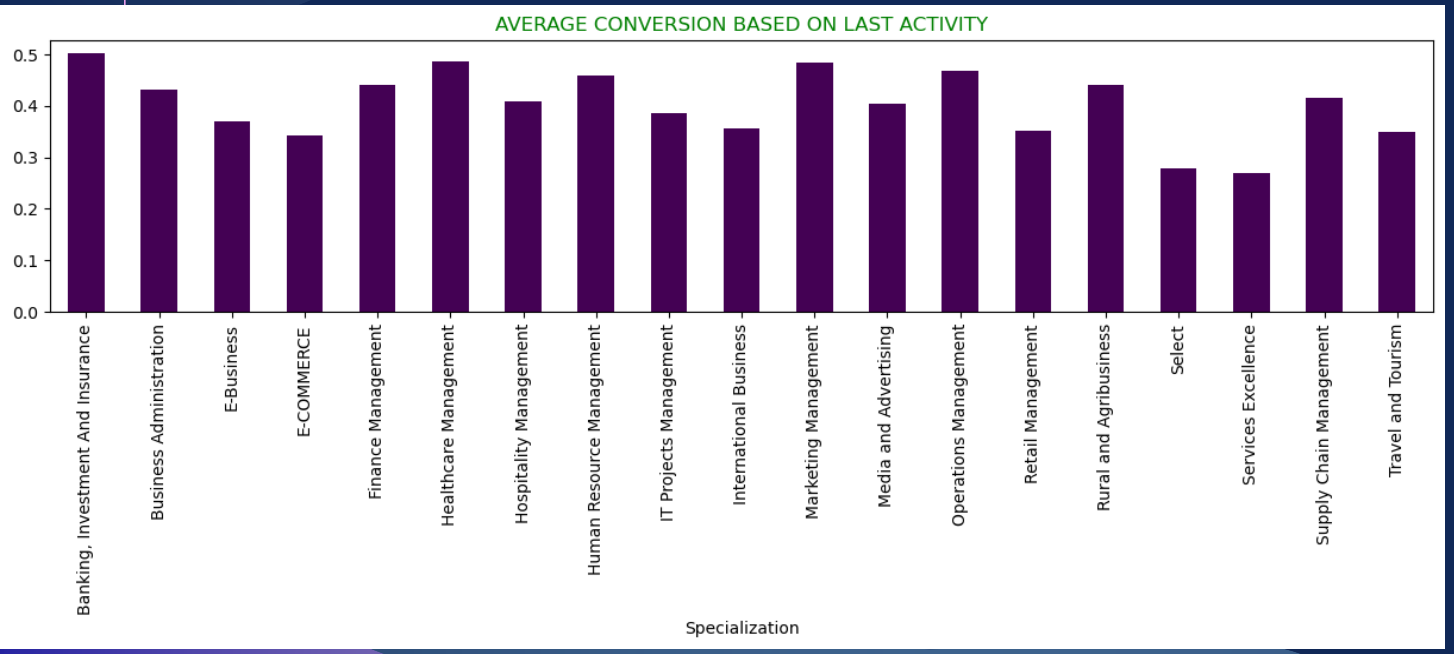




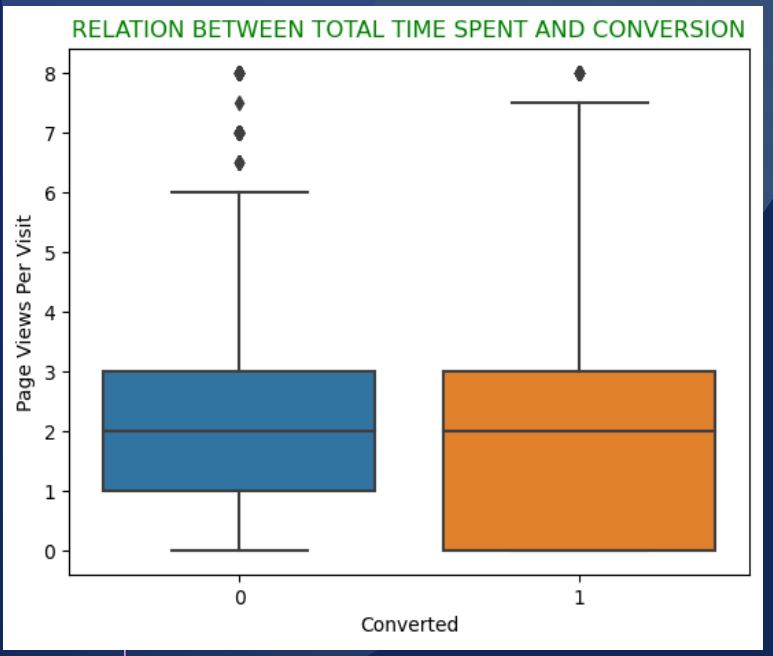
- A substantial 60% of leads originate from landing page submissions, emphasizing the importance of optimizing landing page strategies.
- Google serves as a dominant lead traffic source, contributing to more than 30% of the leads received.
- A significant data imbalance is observed, with 60% of leads not converting and the remaining 40% successfully converting.
- About 35% of leads have left the specialization column blank, possibly indicating diverse or unspecified fields of interest.
- Several columns, including 'Search,' 'Magazine,' 'Newspaper article,' 'X Education Forum,' 'Newspaper,' 'Digital Advertisement,' and 'Through Recommendations,' were removed due to extreme data imbalance, streamlining the dataset for analysis.

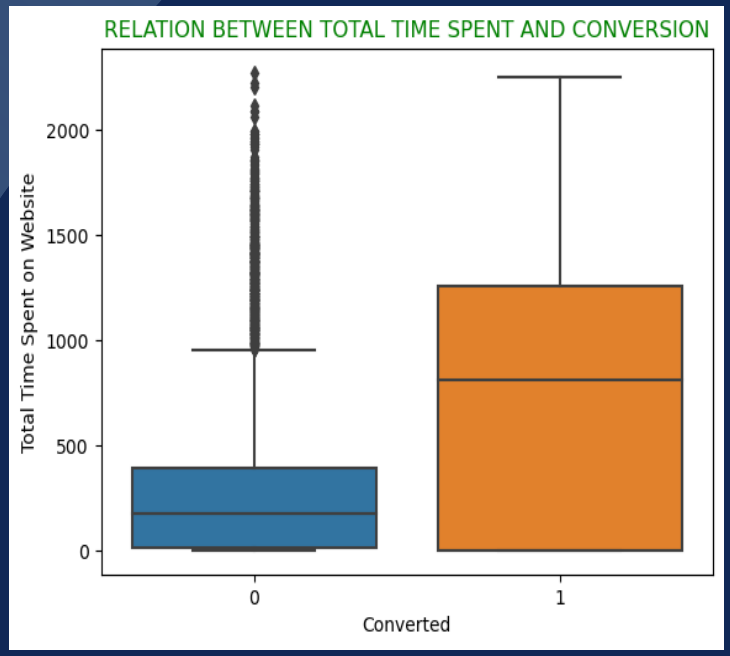


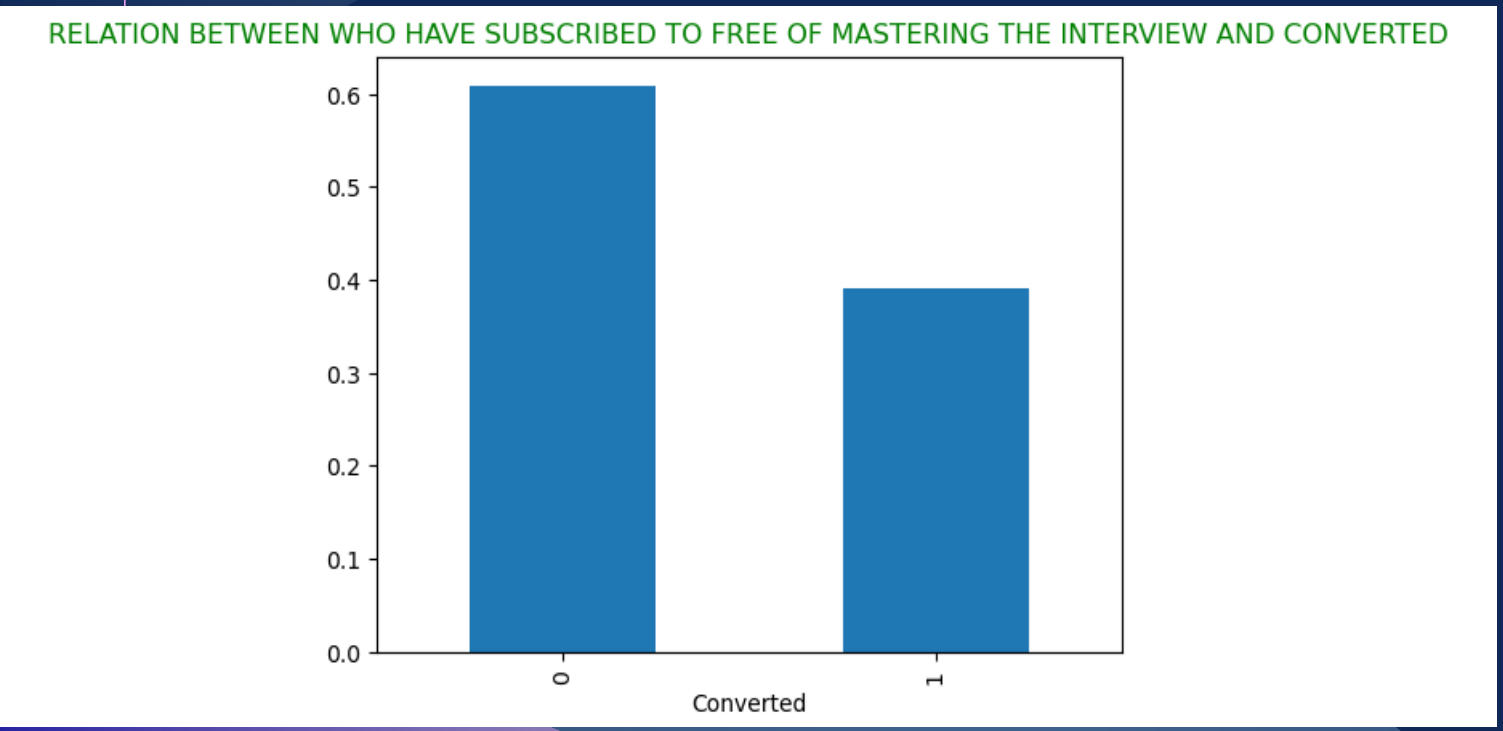




- An impressive conversion rate of over 80% is observed among leads originating from the Lead Add Form, highlighting its effectiveness in generating successful conversions.
- Lead sources, including Live Chat, NC_EDM, Reference, and We Learn, stand out by producing a maximum number of converted leads, indicating their potential as valuable lead generation channels.
- Notably, converted leads tend to exhibit higher website visitation rates and spend more time on the website, indicating a positive correlation between engagement and conversion.







- Last notable activities such as receiving emails, engaging in phone conversations, and resubscribing to emails are associated with significantly higher average conversion rates, emphasizing their influence in driving conversions.
- Leads who selected specializations like Banking, Investment, and Insurance, Healthcare Management, Marketing Management, among others, demonstrated notably high average conversion rates, suggesting a correlation between certain specializations and a greater likelihood of conversion.

PREPARING DATA FOR MODELING

- Created Dummy Variables for Categorical Variables: In the data preparation phase, dummy variables were generated for categorical features, enabling the model to work with these variables effectively.

- Split the Data into Train and Test Data: The dataset was divided into training and testing sets to assess the model's performance on unseen data and prevent overfitting.

- Used Standard Scaler for Scaling Numerical Variables: Standard scaling techniques were applied to numerical variables, ensuring that all numerical features were on the same scale and preventing biases in the model due to varying magnitudes of these features.

MODEL BUILDING

STEPS OF MODEL BUILDING

Build a model using all the features and used Recursive Feature Elimination to select top 15 features

Removed the column Lead Source_Welinga k Website

Removed the column Last_Notable_Activity_ Had_a_Phone_Conver sation

Removed the column Last Activity_Email Opened.

Removed the column Last_Notable_Activity _Email Opened. Ended up with fine model consisting of 11 features

FINAL MODEL

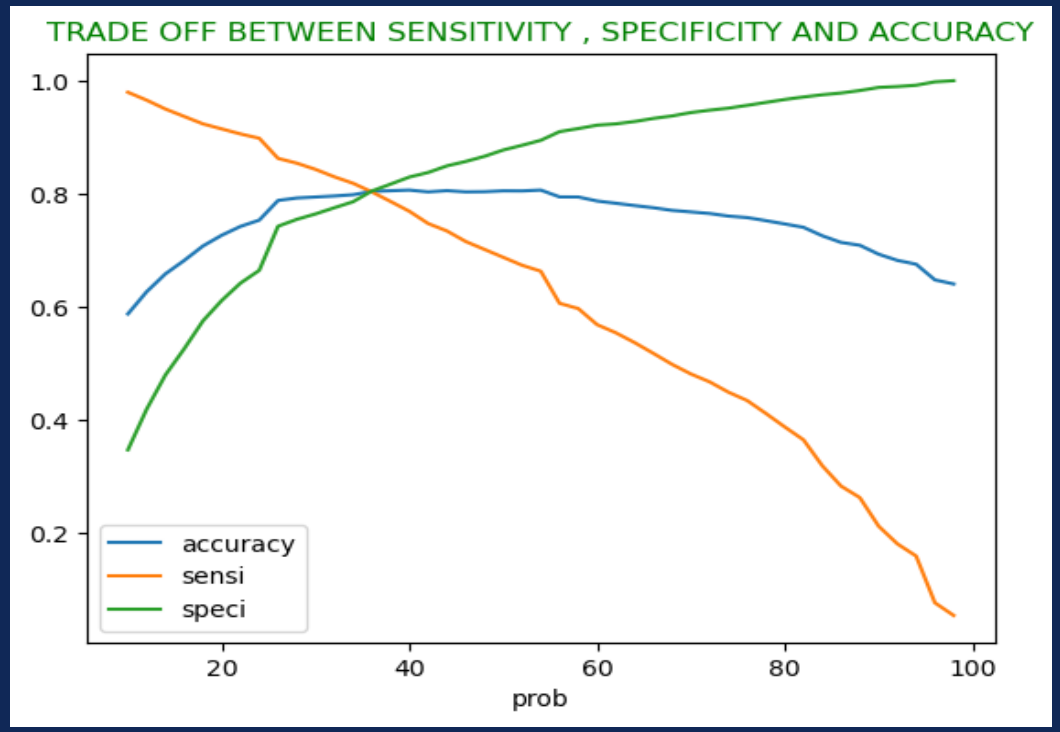
The final lead scoring model incorporates the following essential features for prioritizing high-conversion potential leads: 'Specialization_Select' indicates specific interest, 'Lead Source_Olark Chat' signifies active engagement, and 'Last Notable Activity_Modified' reflects ongoing interest due to recent modifications.

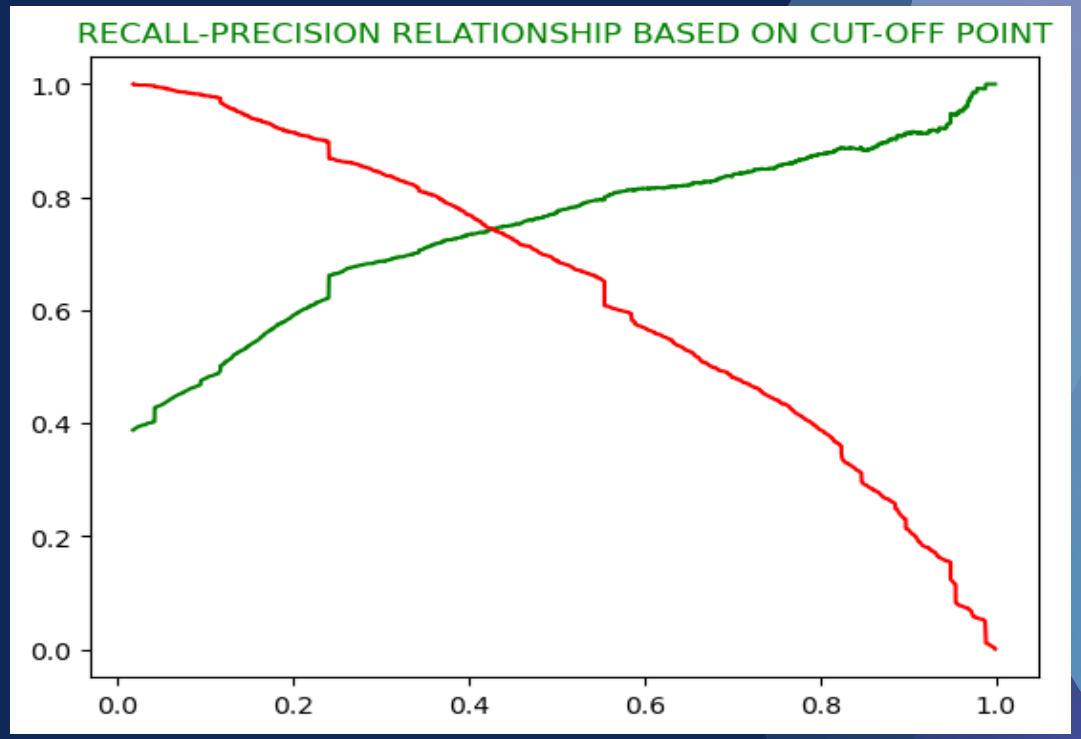
MODEL EVALUATION

METRICS FOR EVALUATION (ON TRAIN DATA)

CONFUSION METRICS FOR CUT-OFF 0.4

- **Accuracy Score of 80.63%:** The model correctly predicted outcomes with an accuracy rate of 80.63%, reflecting its overall performance in classifying leads.
- **Sensitivity Rate of 76.8%:** The model's sensitivity rate, also known as the true positive rate, indicates that it effectively identified 76.8% of actual positive cases, demonstrating its ability to recognize leads that are likely to convert.
- Specificity of 82.9%: The model's specificity, or true negative rate, stands at 82.9%, showcasing its proficiency in correctly identifying negative cases, or leads that are unlikely to convert.
- **Positive Predictive Value (PPV) of 73.4%:** The positive predictive value represents the model's precision in predicting positive outcomes.
- **Negative Predictive Value (NPV) of 85.39%:** The negative predictive value signifies the model's precision in predicting negative outcomes.





The analysis of sensitivity, specificity, precision, and recall metrics recommends a threshold of 0.4 for optimizing the balance between true positive and true negative rates in the classification model.

MODEL VALIDATION

METRICS FOR VALIDATION (ON TEST DATA)

CONFUSION METRICS FOR CUT-OFF 0.4

- **Accuracy Score of 80.51%:** The model correctly predicted outcomes with an accuracy rate of 80.51%, reflecting its overall performance in classifying leads.
- **Sensitivity Rate of 77.53%:** The model's sensitivity rate, also known as the true positive rate, indicates that it effectively identified 77.53% of actual positive cases, demonstrating its ability to recognize leads that are likely to convert.
- **Specificity of 82.3%:** The model's specificity, or true negative rate, stands at 82.3%, showcasing its proficiency in correctly identifying negative cases, or leads that are unlikely to convert.
- **Positive Predictive Value (PPV) of 72.6%:** The positive predictive value represents the model's precision in predicting positive outcomes.
- **Negative Predictive Value (NPV) of 85.8%:** The negative predictive value signifies the model's precision in predicting negative outcomes.

RECOMMENDATIONS

- **During Internship Period:** To enhance lead conversion during the internship phase, it is recommended to focus on specialized leads, particularly those in HR management, marketing, banking, insurance, and investment management, as they have demonstrated higher conversion rates. Prioritize leads from sources like references, WeLearn, Welingak's website, and live chat, which have a higher likelihood of converting. Optimize the lead add form to give prominence to leads generated through this channel, as it consistently exhibits the highest conversion rate. Engage in personalized communication by segmenting leads based on their specialization, offering tailored information. Experiment with adjustments to the conversion threshold in your logistic regression model to capture more potential converters. These strategies will help strike a balance between lead quantity and quality, improving lead conversion outcomes during the internship phase.

- **After Achieving Quarterly Targets:** To enhance lead conversion efficiency after surpassing quarterly targets, consider adjusting the cutoff point in the logistic regression model to classify only the most promising leads. Prioritize leads with specializations in HR Management, Marketing Management, Banking, Insurance, and Investment Management, as they exhibit higher conversion rates. Additionally, continue emphasizing leads from high-conversion sources such as references, WeLearn, Welingak's website, and live chat. These strategies will help maximize lead quality and resource allocation while minimizing unnecessary phone calls, ensuring a more productive sales team.

THANK YOU