Lead Scoring Case Study Assignment

-- Arnav Sinha

About the X Education Company

- X Education is an online education company that offers courses for sale.
- Potential customers visit the company's website or find their courses through online advertising and search engines like Google.
- Visitors may browse the course offerings, fill out a form, or watch informational videos about the courses.
- When individuals provide their contact details (email or phone number), they are considered "leads.“
- After acquiring leads, the sales team reaches out through calls, emails, and other communication methods.
- This outreach results in some leads converting into customers, but many do not.
- The average lead conversion rate at X Education is approximately 30%.

Problem Statement & Objective

Problem Statement

X Education’s lead conversion rate is very poor, approx 30%.

Objective

- The CEO has set a target lead conversion rate of approximately 80%.
- Improve the lead conversion process by identifying and prioritizing the most promising leads, or "Hot Leads.“
- Enable the sales team to focus their efforts on engaging with Hot Leads, rather than contacting all leads indiscriminately.

Approach of Analysis

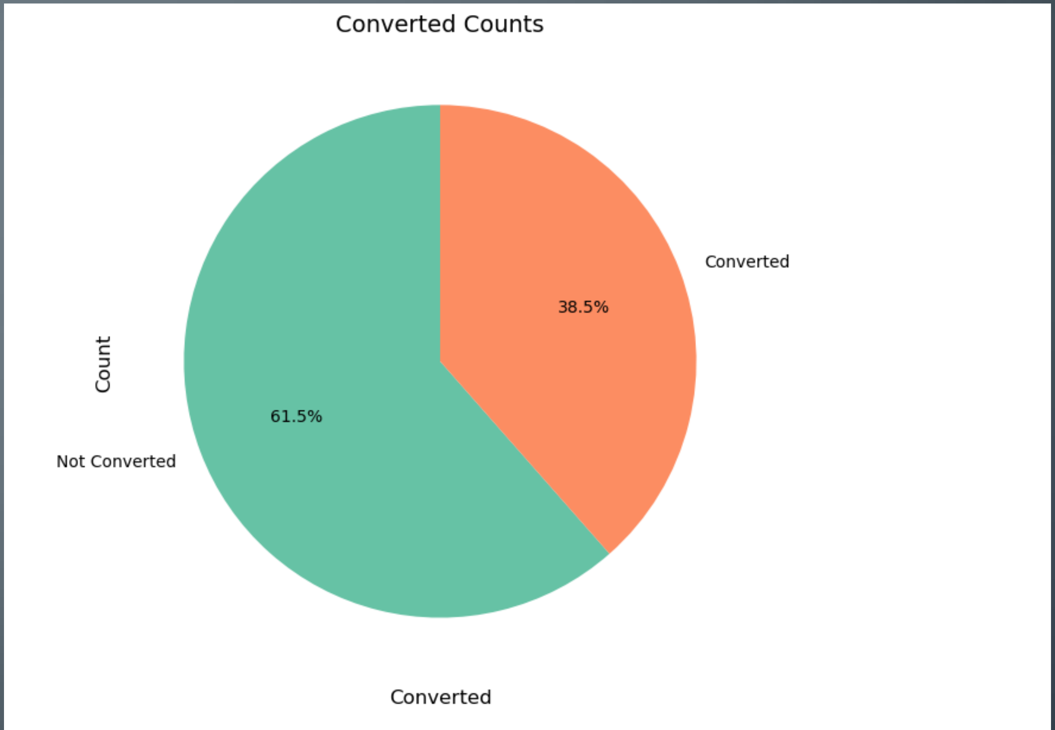
- Data Cleaning: Here we will load the data set and then will understand and clean the data.
- EDA: This process will include the data imbalance checking followed by univariate & bivariate analysis.
- Data Preparation: Here we will analyze the dummy variables, then data will be splitted into training and test followed by feature scaling.
- Model Building: Here we will to the RFE for Top 20 feature, manual feature reduction & model finalization.
- Model Evaluation: This step will have the creation of the confusion matrix, cutoff selection and assigning of the Lead Score.
- Predictions on Test Data: Here we will compare the train vs test metrices, will assign the Lead Score and get the top features.
- Recommendation: Finally, we will find and suggest the top 3 features or variables to focus for higher lead conversion.

Data Cleaning

- The "Select" value in four categorical variables was replaced with a Null value, as it had no business significance.
- Columns having more than 30% missing data were removed from the dataset.
- Missing values in categorical columns were handled using value counts and other relevant considerations.
- Columns that provided no useful insights only a single value like 'No' were irrelevant to the study, so they were discarded.
- Imputation was carried out for missing numerical data using appropriate statistical measures, such as the median.
- New categories were created for certain variables with Null values, like "Last Activity" and "Occupation.“
- Outliers in the "TotalVisits" and "Page Views Per Visit" variables were capped at the 99th percentile to prevent skewing.
- After data cleaning, 98% of the original data was retained for analysis.

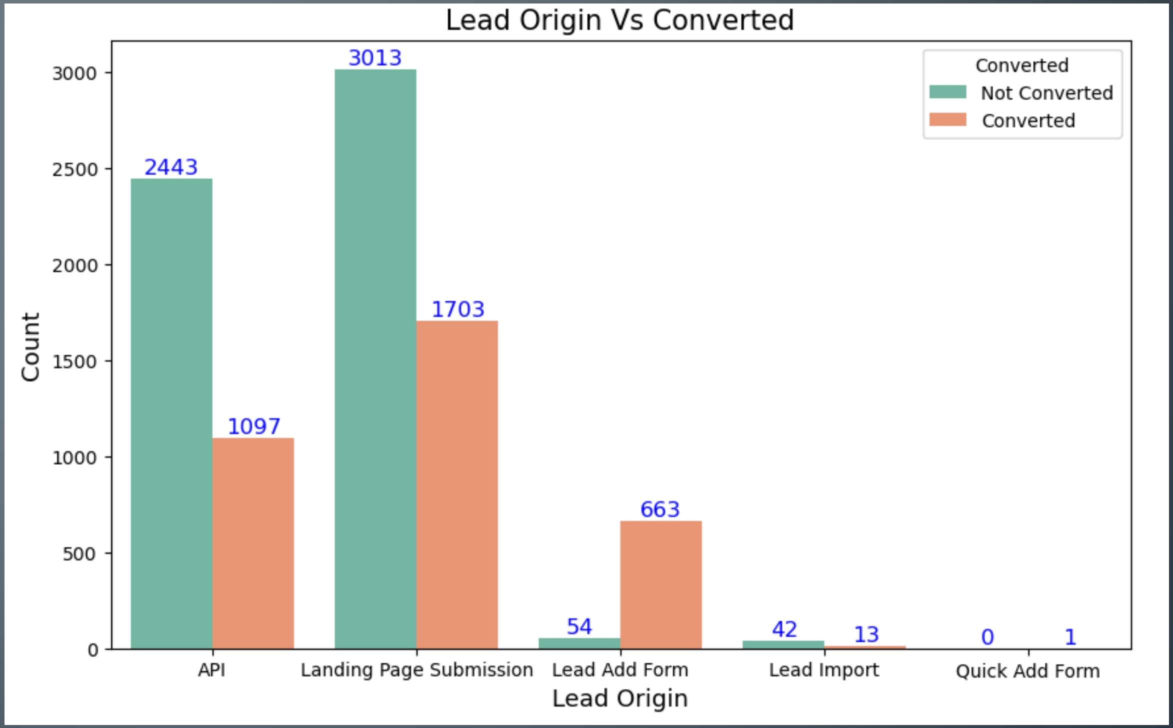
Data Imbalance

Data is imbalanced since the conversion rate is ~39%



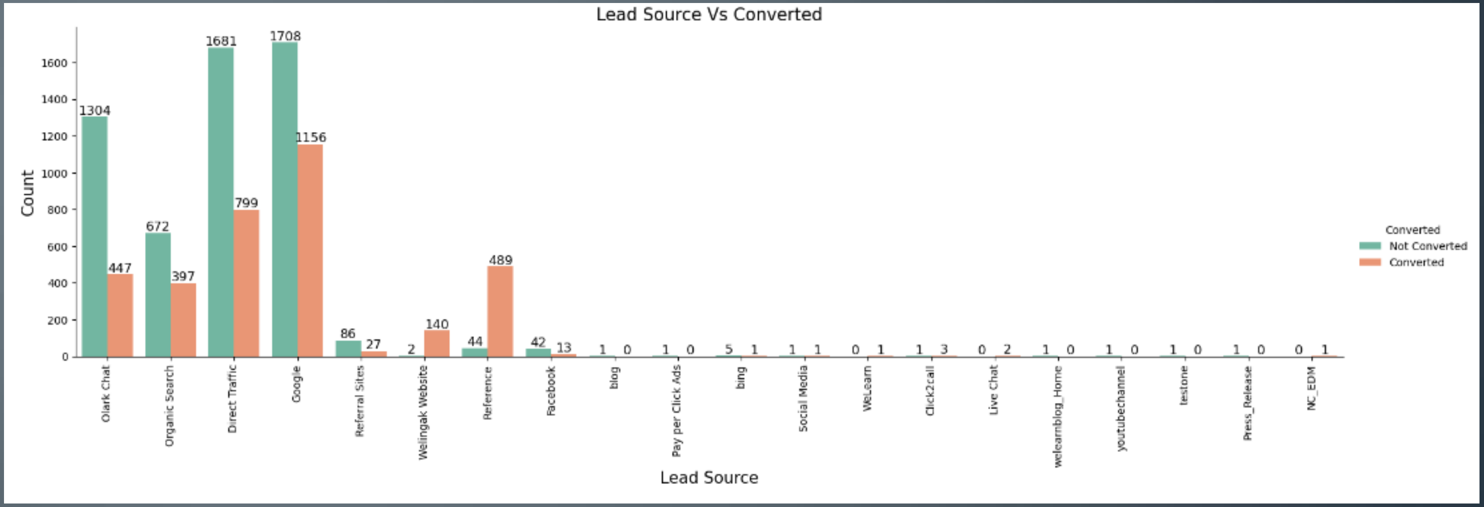
Univariate Analysis of Categorical Variables

- Highest number of conversion happened from Landing Page Submissions followed by API.



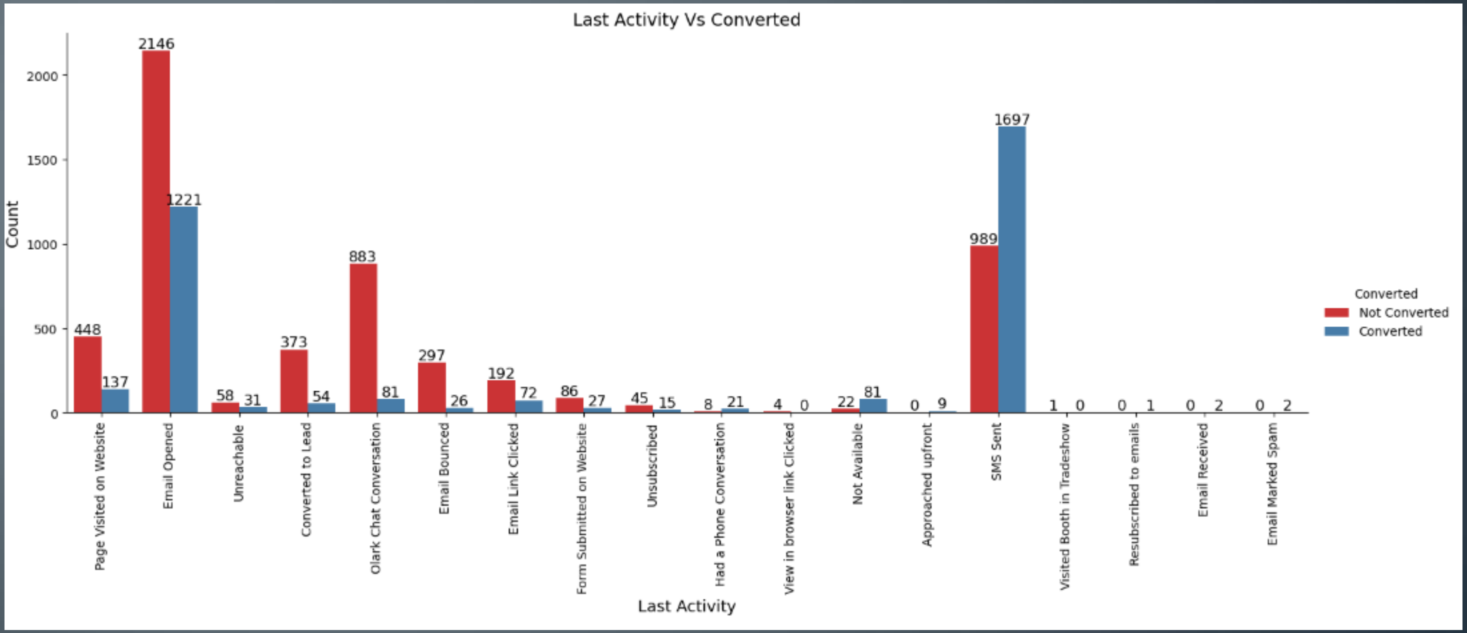
Univariate Analysis of Categorical Variables

- Maximum conversion happened from Google, Direct Traffic and Reference.



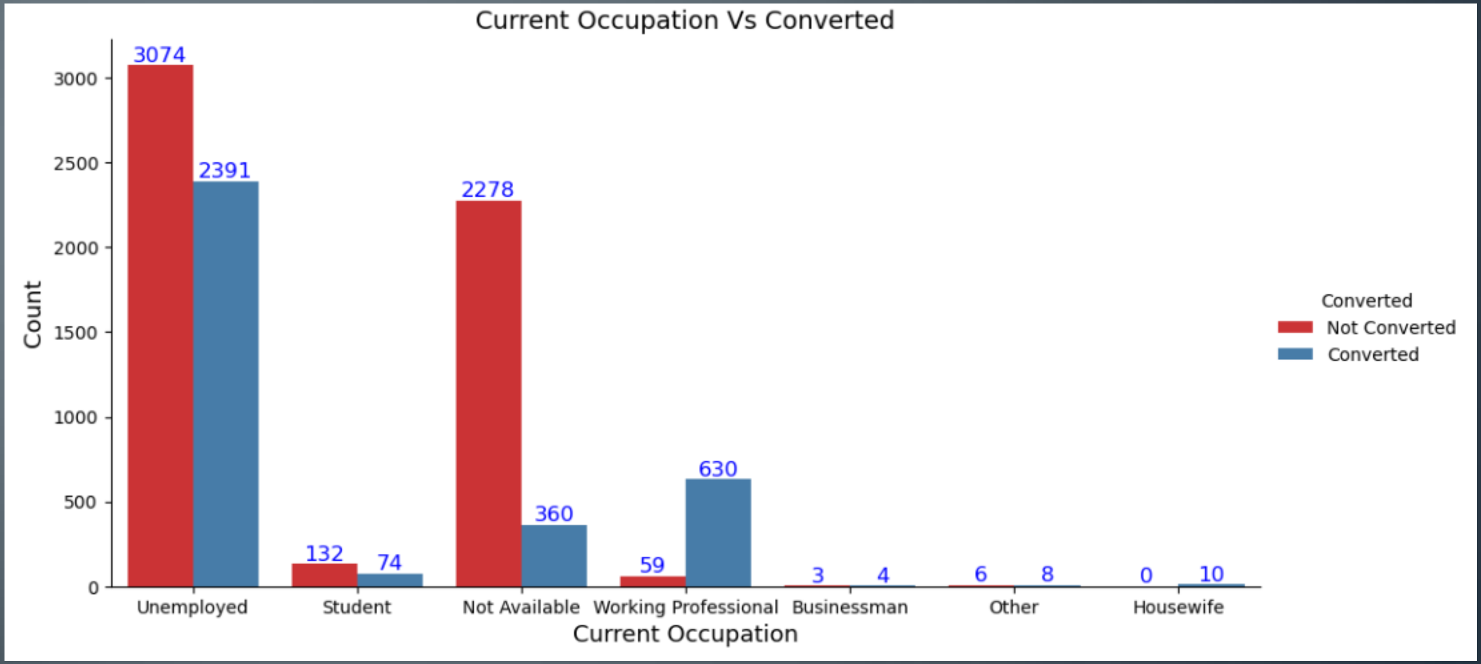
Univariate Analysis of Categorical Variables

- SMS Sent and Email Opened have the most number of conversions.



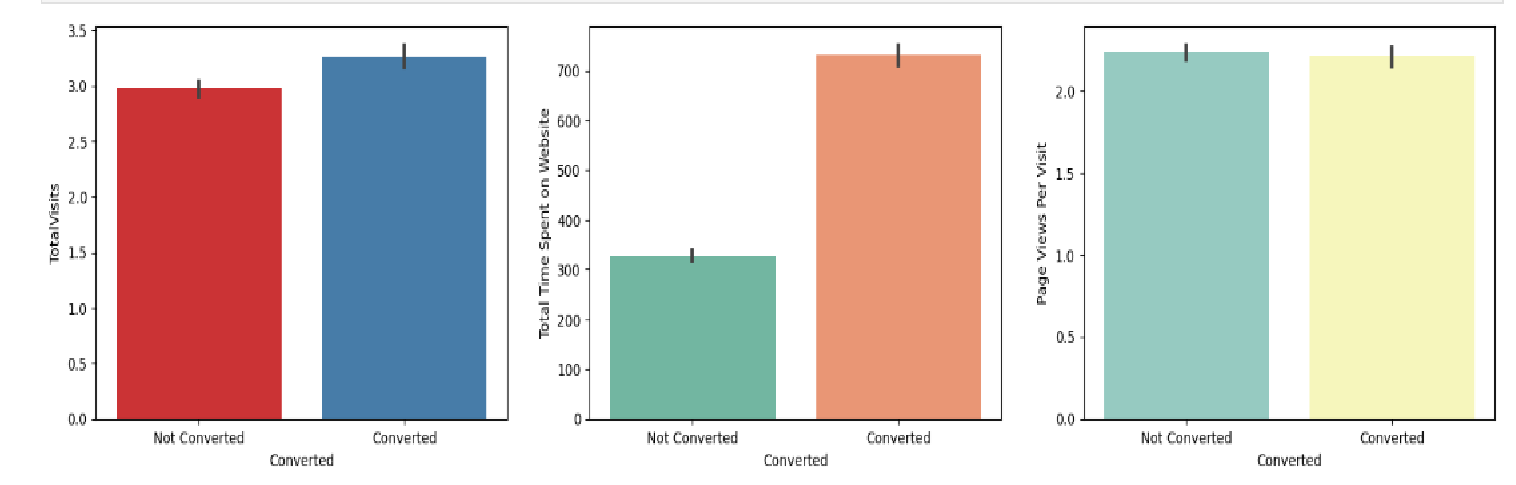
Univariate Analysis of Categorical Variables

- Unemployed and Working Professionals are the highest converted customers.



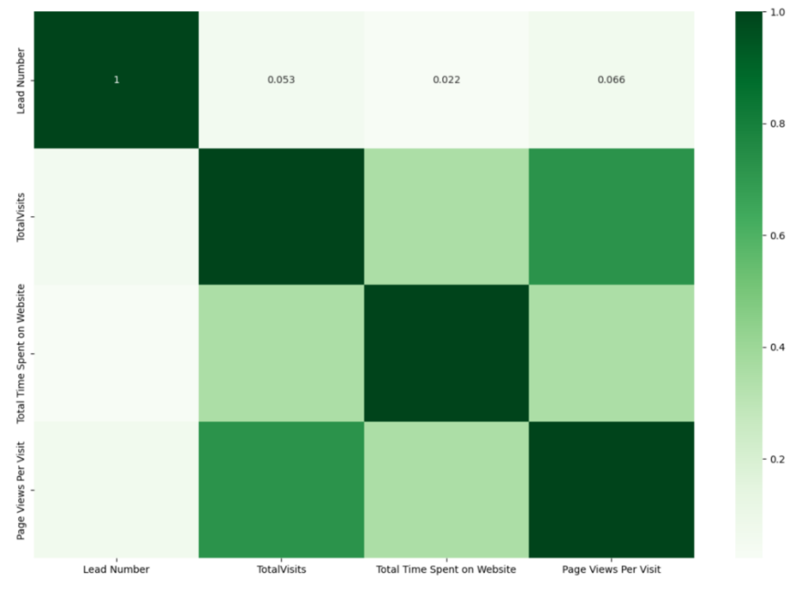
Bivariate Analysis of Numerical Variables

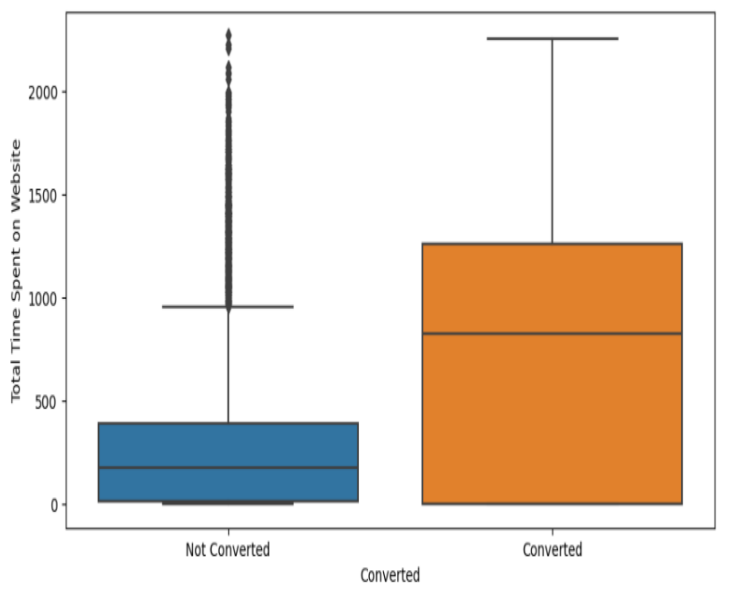
- More time spent on the website contributes to higher conversions



Bivariate Analysis of Numerical Variables

- Leads who **spends more time on the Website** have a higher chance of getting successfully converted than those who spends less time.





Data Preparation

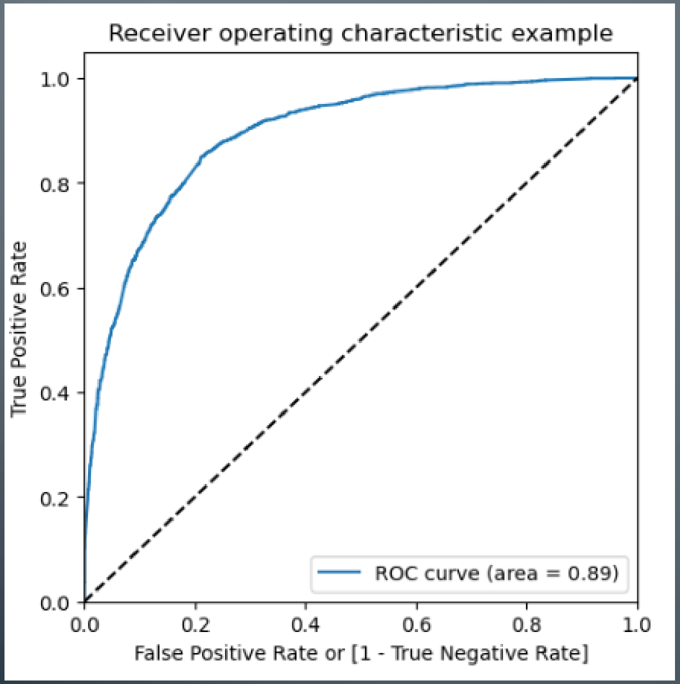
- ‘Yes/No’ categorical columns have been mapped to '1’ and ‘0‘
- Dummy Variables has been created for all the categorical columns where the no. of categories > 2
- 70:30 % ratio was chosen for the train-test split
- MinMax scaler was used to scale the features for continuous variables

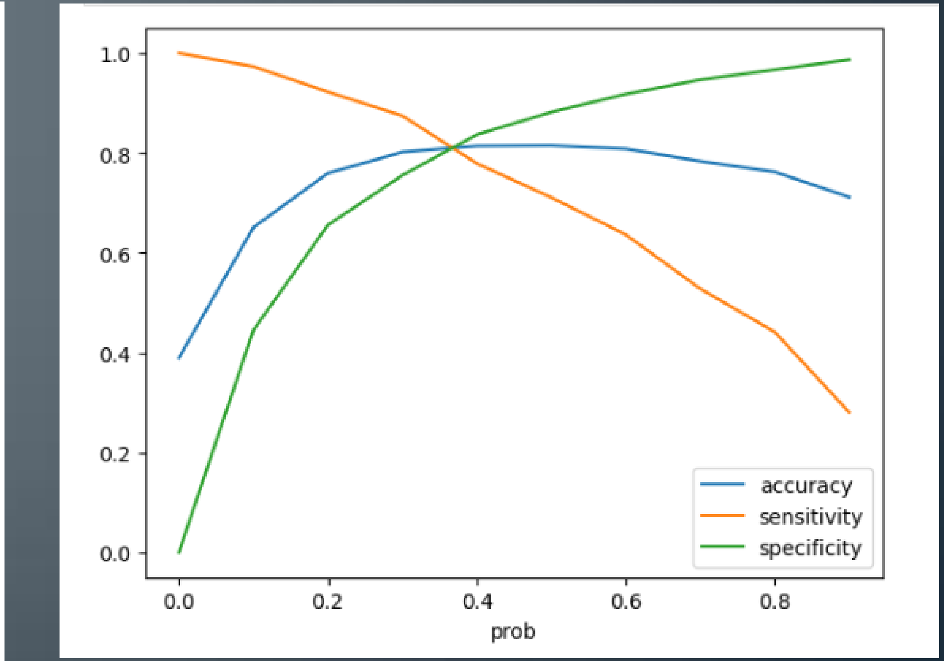
Model Building

- Recursive Feature Elimination (RFE) was used to conduct Feature selection.
- Manual Feature Reduction process was used to build models by dropping variables with insignificant p values.
- We saw that Pre RFE, there were 70 columns and Post RFE there became 11 columns.
- Model 11 was stable with significant p-values and acceptable multicollinearity with VIF values less than 5.

Model Evaluation – Train Data Set

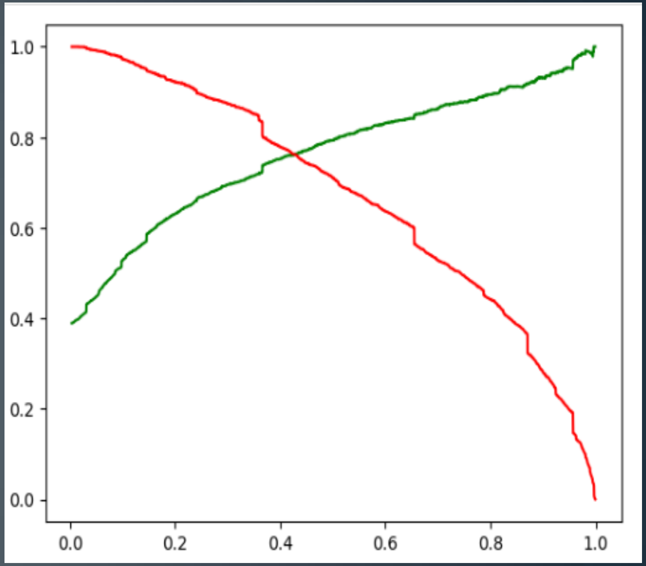
- 0.39 has been selected as the cutoff post checking evaluation metrics.
- Final prediction of conversions have a target of ~78%. Hence this is a good model.





Model Evaluation – Test Data Set

- Accuracy, Sensitivity and Specificity values of test set are around 81%, 77% and 83% which are closer to the respective values calculated using trained set.
- Lead score calculated in the trained set of data shows the conversion rate on the final predicted model is around ~78%.
- So, overall this model seems to be good.



Recommendations

- They should concentrate on leads with high scores which will maximize the chances of conversion. Targeting leads with higher scores optimizes outreach, focusing on those more likely to convert, leading to better conversion rates.
- A personalized script increases the quality of interactions, and fostering a positive impression.
- Increased portal engagement signifies stronger interest and commitment, improving the likelihood of successful conversions.
- Insights from current customers can provide valuable information to refine strategies, overcome objections, and tailor communication approaches.
- Diversifying the target audience taps into a potentially lucrative market, as working professionals may have higher conversion potential due to their specific needs and resources.

Thank You