Lead Scoring Case Study

SUBMITTED BY :

- **Souradeep Bose**
- **Viraj Morgaonkar**
- **Shahrukh Naved**

Problem statement

- X Education sells online courses to industry professionals.
- Professionals visit the website through marketing efforts on various platforms.
- Visitors may browse courses, fill forms, or watch videos.
- Leads are generated from form submissions, referrals, and website interactions.
- Sales team initiates communication via calls and emails to convert leads.
- Average lead conversion rate is approximately 30%.

Business Goal:

- X Education seeks assistance to prioritize high-conversion probability leads.
- A lead scoring model is required to assign scores based on conversion likelihood.
- Higher lead scores correspond to increased conversion potential.
- CEO's target lead conversion rate is approximately 80%.

Approach

- Develop a logistic regression model which assigns score to leads based on different variables
- At first, we identify the relevant variables which are useful in modelling using EDA
- Prepare and clean the data like, removing nulls, outliers, variables with no variance, etc.
- Prepare the data for modelling: creating dummy variables, binary mapping, scaling etc.
- Build and train the model using train dataset and identifying top features using RFE and manual feature elimination
- Identifying the best model based on p value and VIF
- Evaluating the model based on multiple metrics like accuracy, sensitivity, specificity, etc.
- Predicting lead conversion using the test data
- Evaluate the model using test data based on similar metrics
- Present the model and findings to relevant stakeholders

Problem solving methodology

Data Sourcing , Cleaning and Preparation

- Read the Data from Source
- Convert data into clean format suitable for analysis
- Remove duplicate data
- Outlier Treatment
- Exploratory Data Analysis
- Feature Standardization

Feature Scaling and Splitting Train and Test Sets

- Feature Scaling of Numeric data
- Splitting data into train and test set

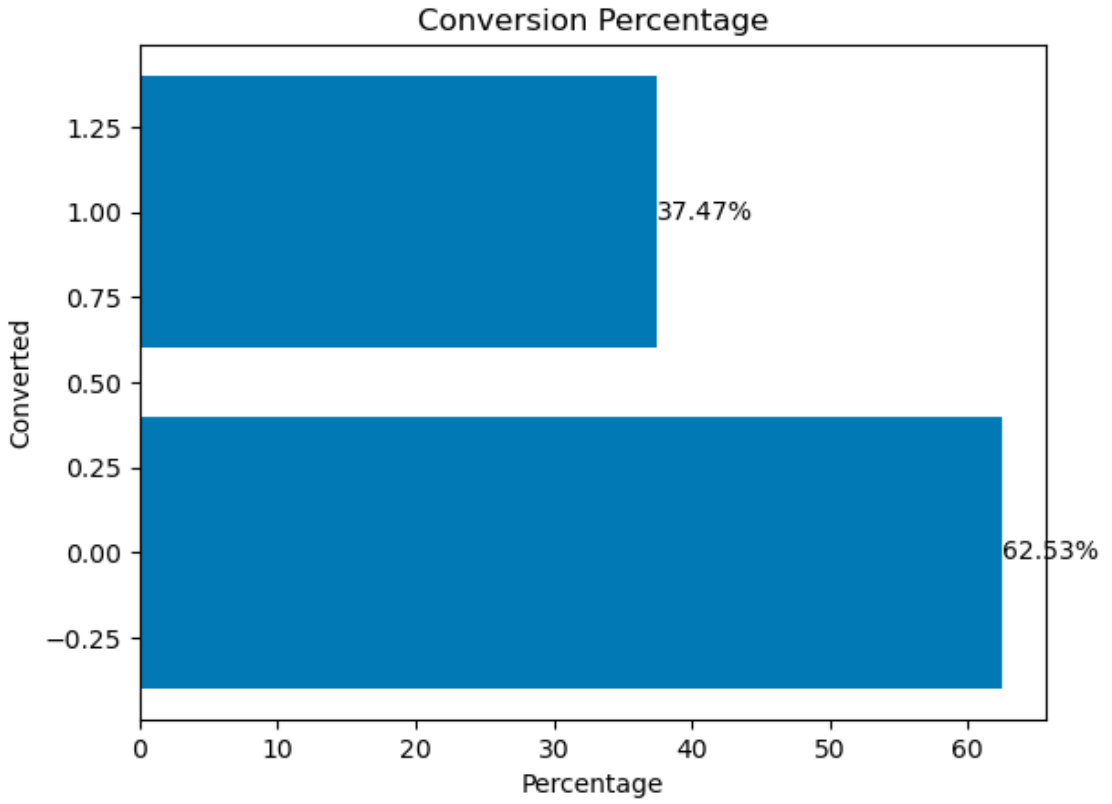
Model Building

- Feature Selection using RFE
- Determine the optimal model using Logistic Regression
- Calculate various metrics like accuracy, sensitivity, specificity, precision and recall and evaluate the model

Result

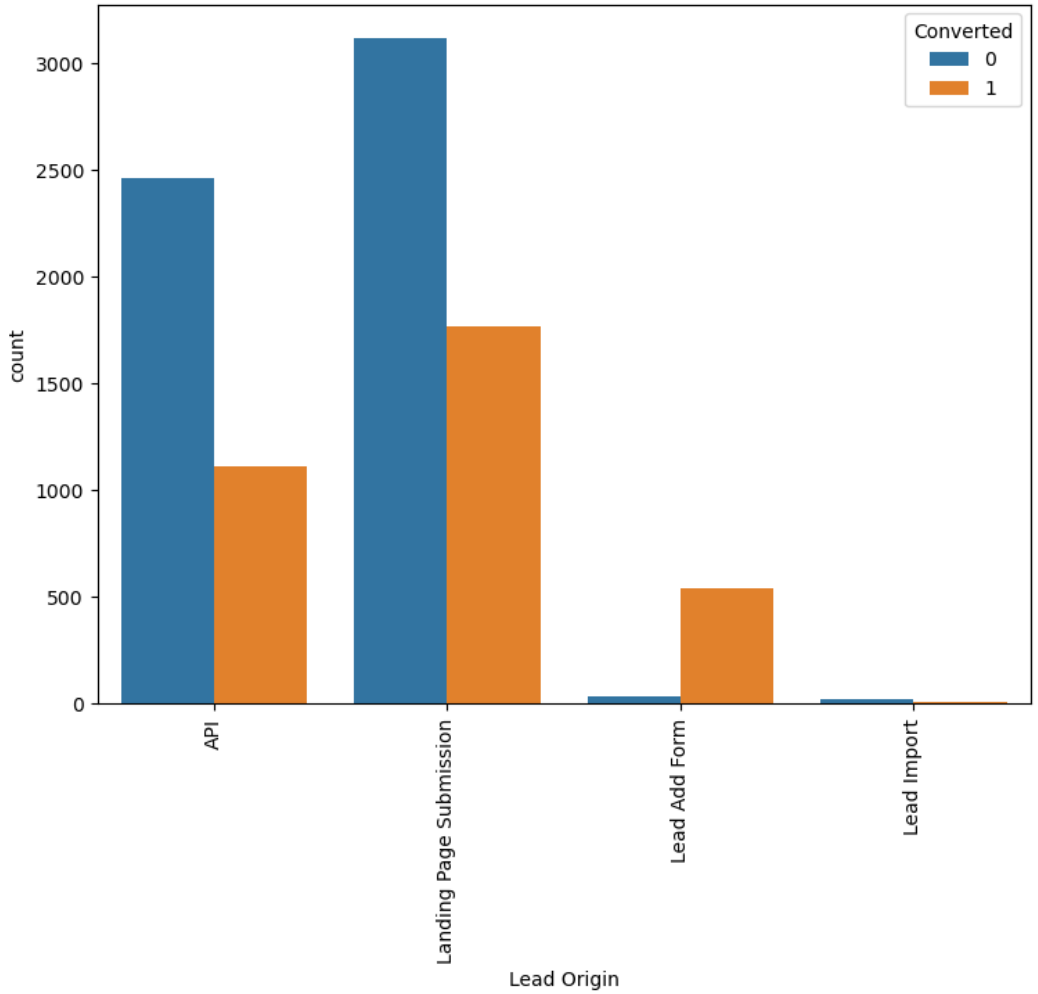
- Determine the lead score and check if target final predictions amounts to 80% conversion rate
- Evaluate the final prediction on the test set using cut off threshold from sensitivity and specificity metrics

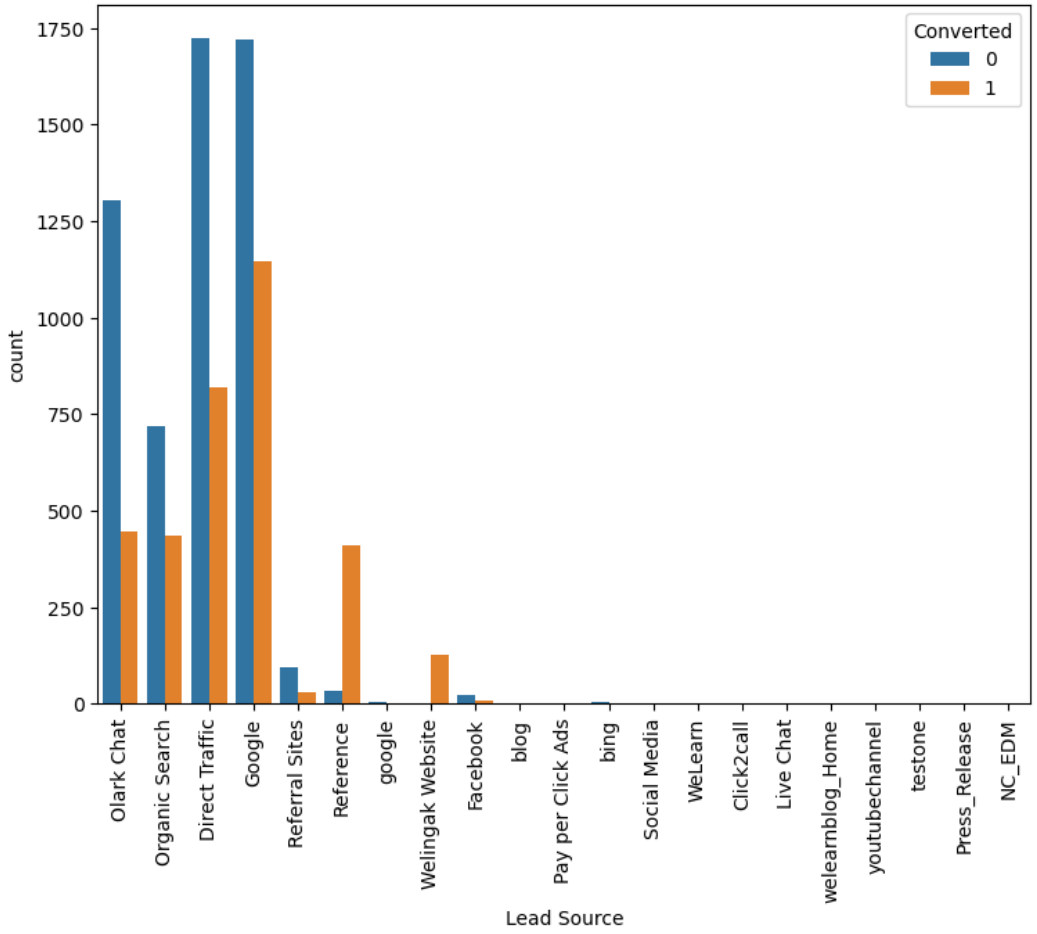
EDA



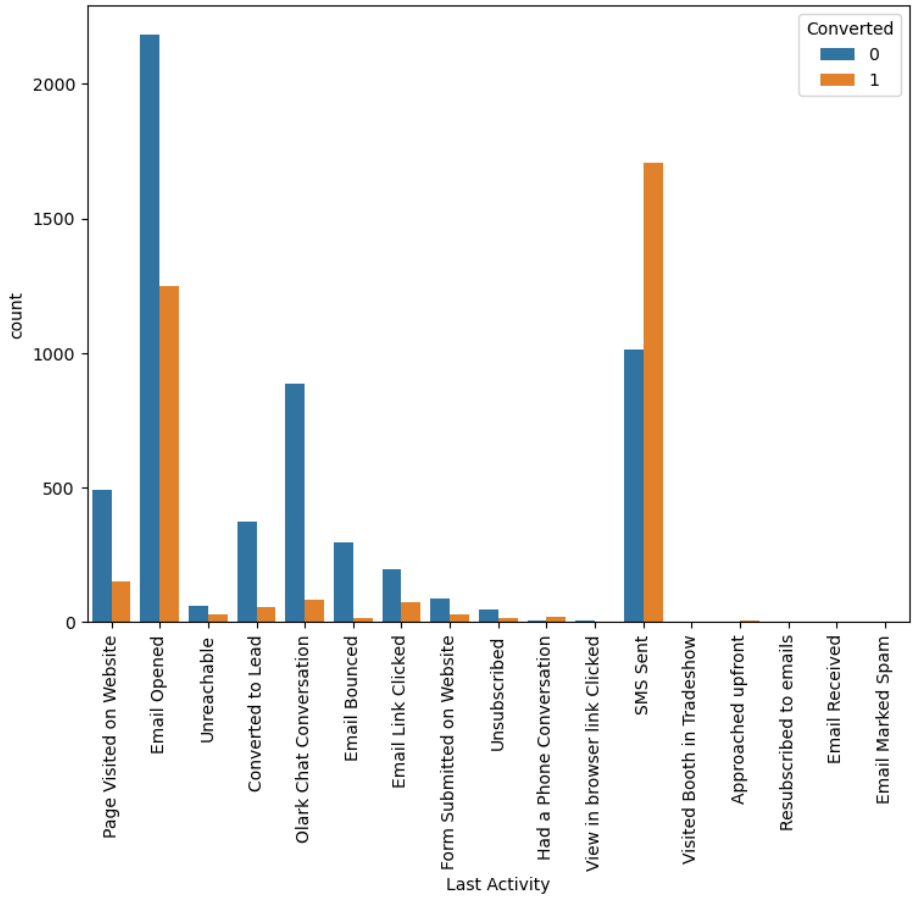
- The data has an imbalance with ~37% of converted vs not converted
- The % of converted is good enough for modelling

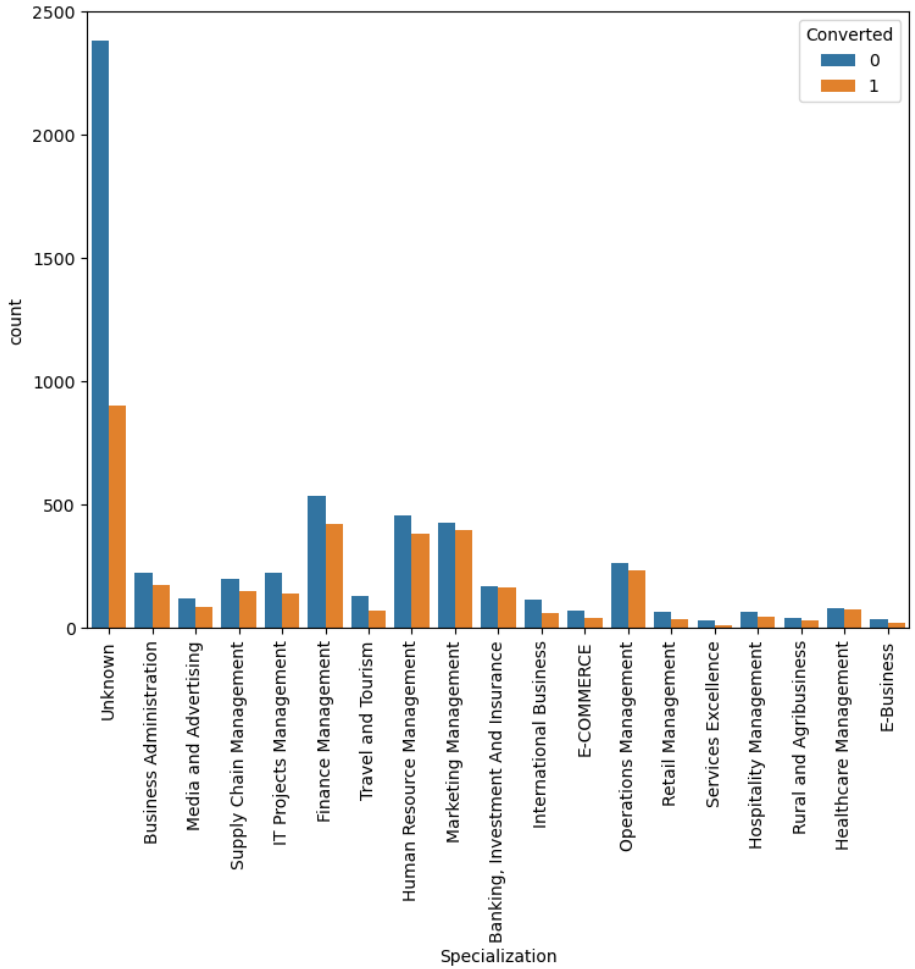
Analysing The Categorical Columns



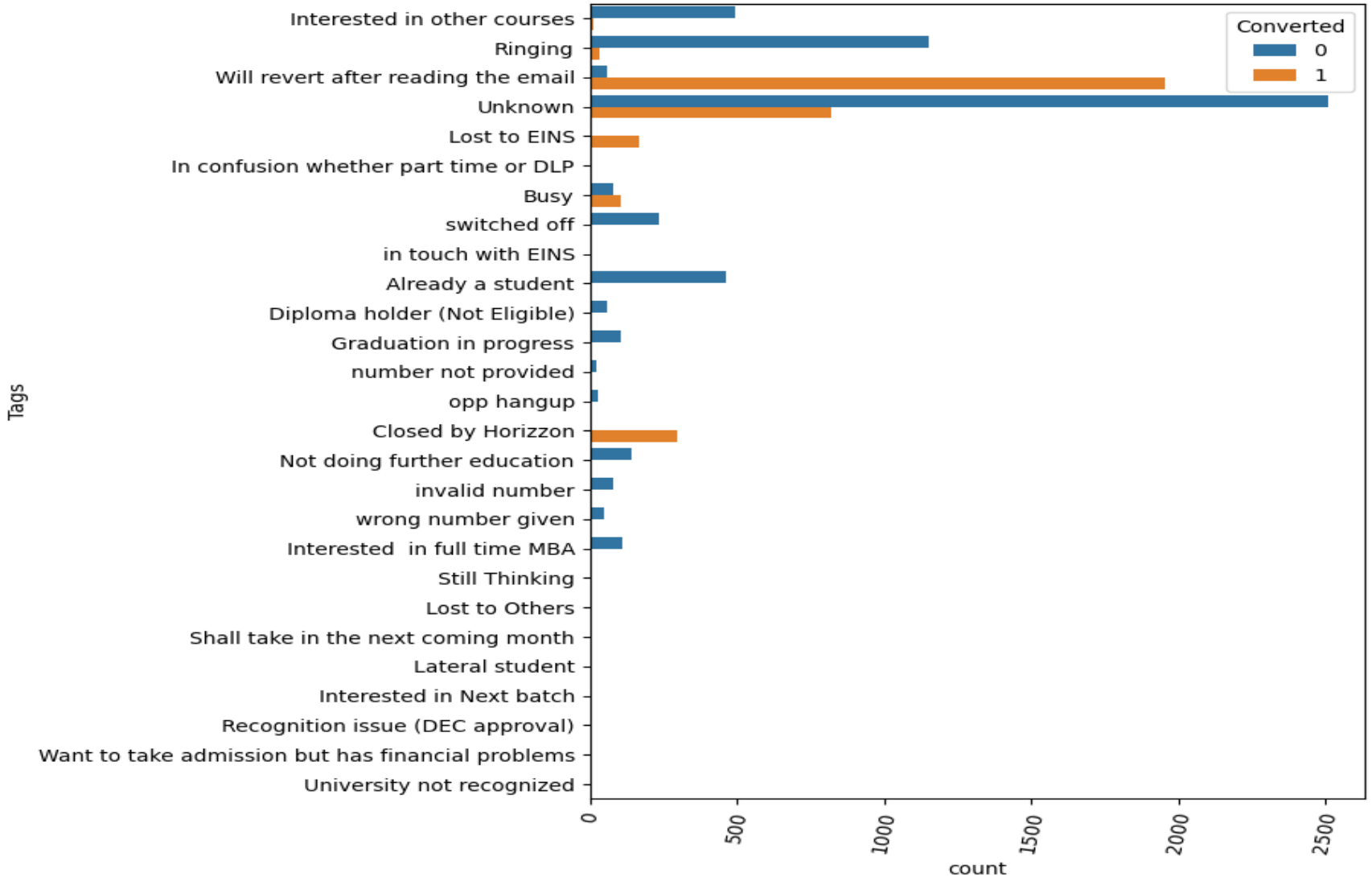


Graph 1 : Lead Origin vs Converted. The Landing page Submissions has highest Conversion followed by API.
Graph 2 : Lead Source vs Converted. The Collage, Direct traffic and Olark Chat, Organic Search has highest Positive Conversions.

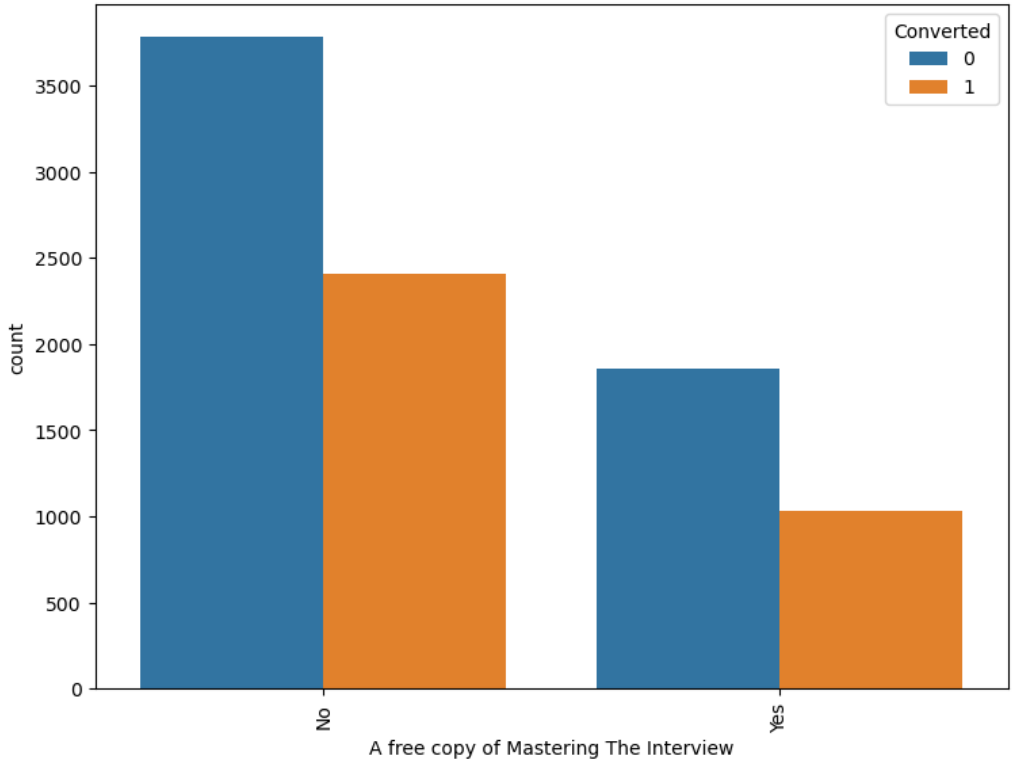


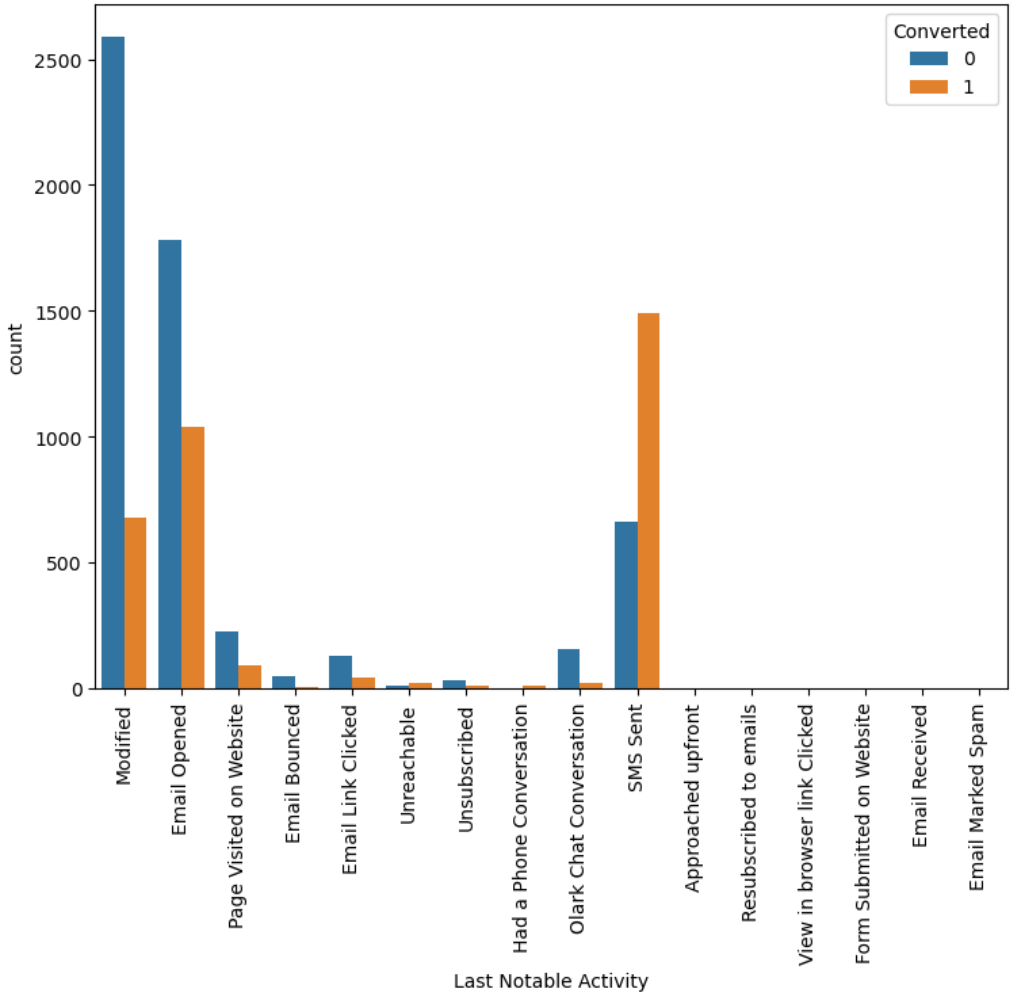


Graph 3 : Last Activity vs Converted. The SMS Sent shows highest positive conversion followed by Email Sent.
Graph 4 : Specialization vs Converted. The visitor who has not selected specialization is unknow and has highest Positive conversion.



Graph 5 : Tags vs Converted. Will revert after reading the email has highest positive conversion rate followed by Unknown tags.





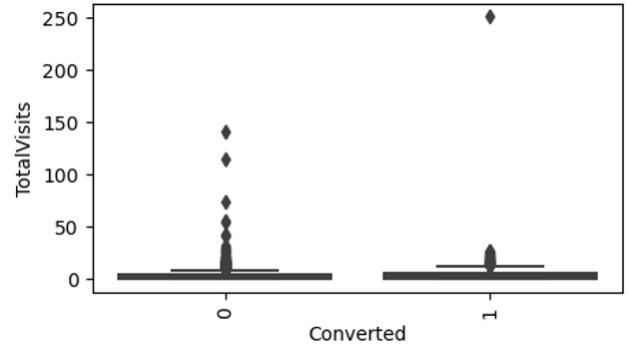
Graph 6 : A free Copy of Mastering Interview vs Converted. Not Providing A free Copy of Mastering Interview give higher positive Conversion rate.

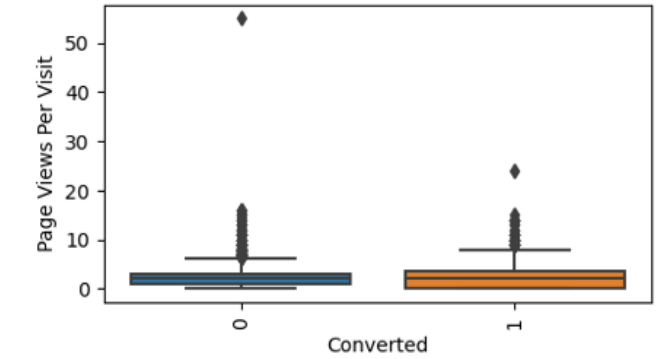
Graph 7 : Last Notable Activity vs Converted. A SMS Sent, Email Opened has positive Conversion. The Modified shows high Negative conversion.

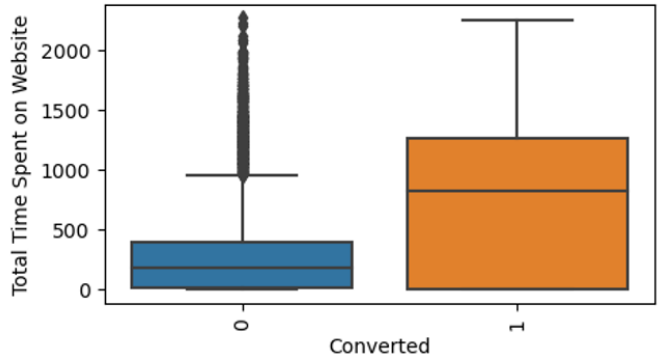
Analysing The Numerical Columns

- From the boxplots it's visible that the data has outliers and conversion rates are dependent on time spent on the website.

	TotalVisits	Total Time Spent on Website	Page Views Per Visit
count	9074.000000	9074.000000	9074.000000
mean	3.456028	482.887481	2.370151
std	4.858802	545.256560	2.160871
min	0.000000	0.000000	0.000000
50%	3.000000	246.000000	2.000000
95%	10.000000	1557.000000	6.000000
99%	17.000000	1839.000000	9.000000
max	251.000000	2272.000000	55.000000



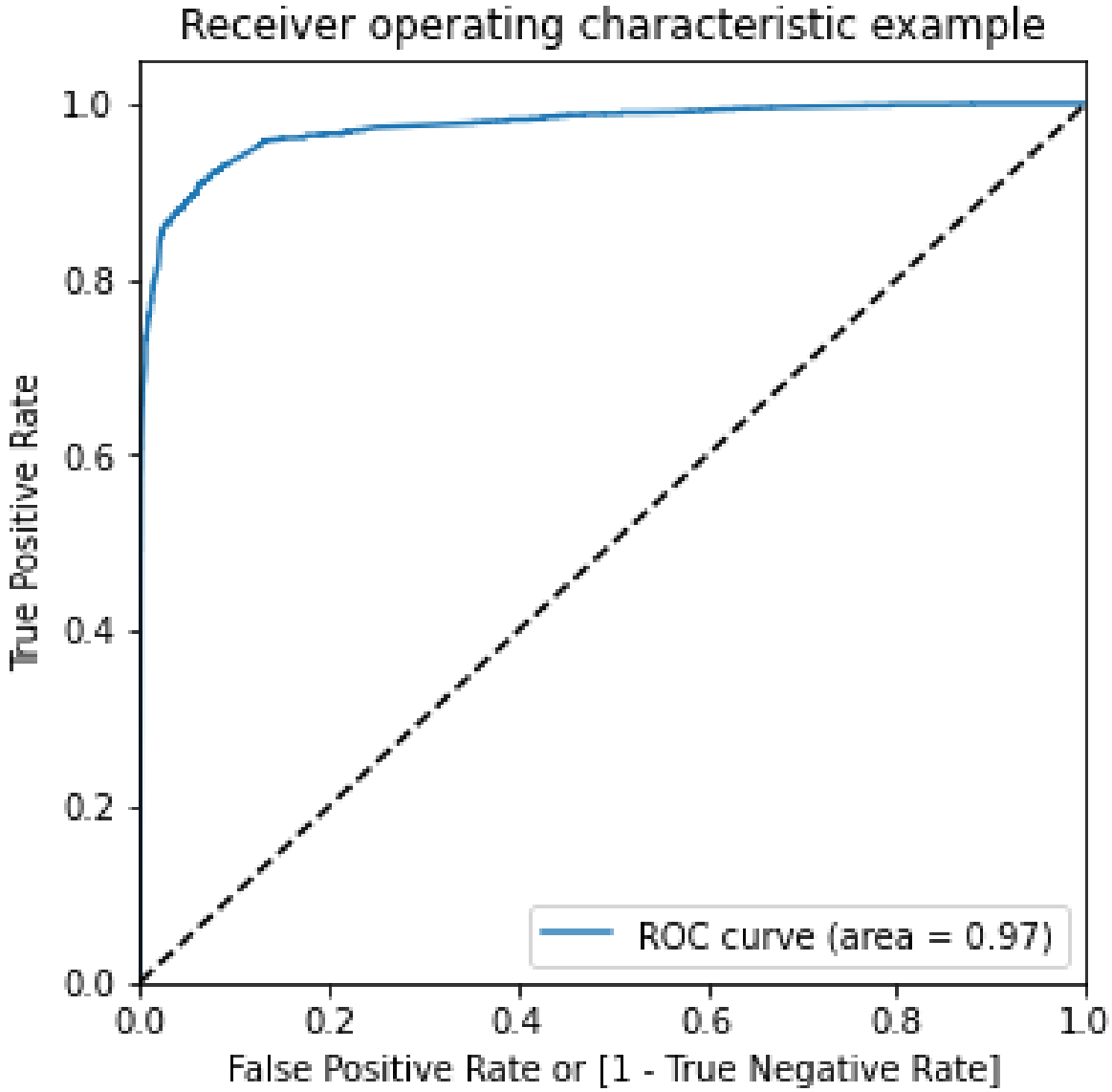




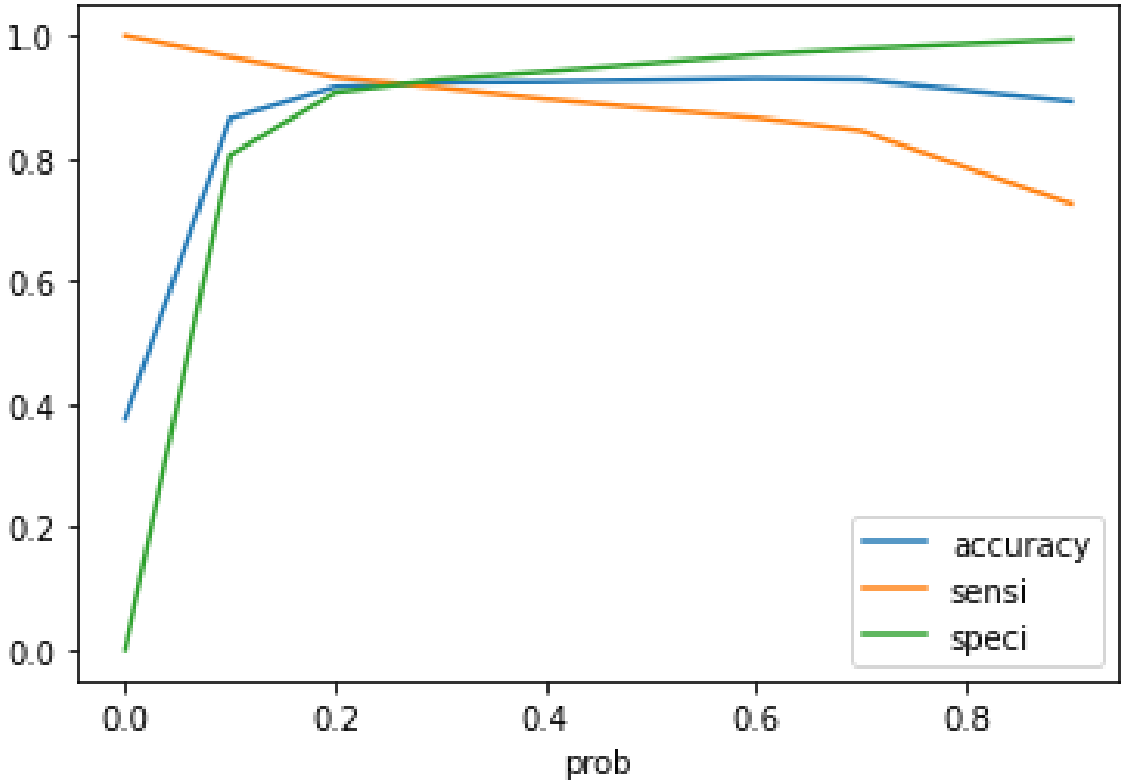
- The above data shows that there are clearly outleirs in the data above the 99th percentile and can be removed

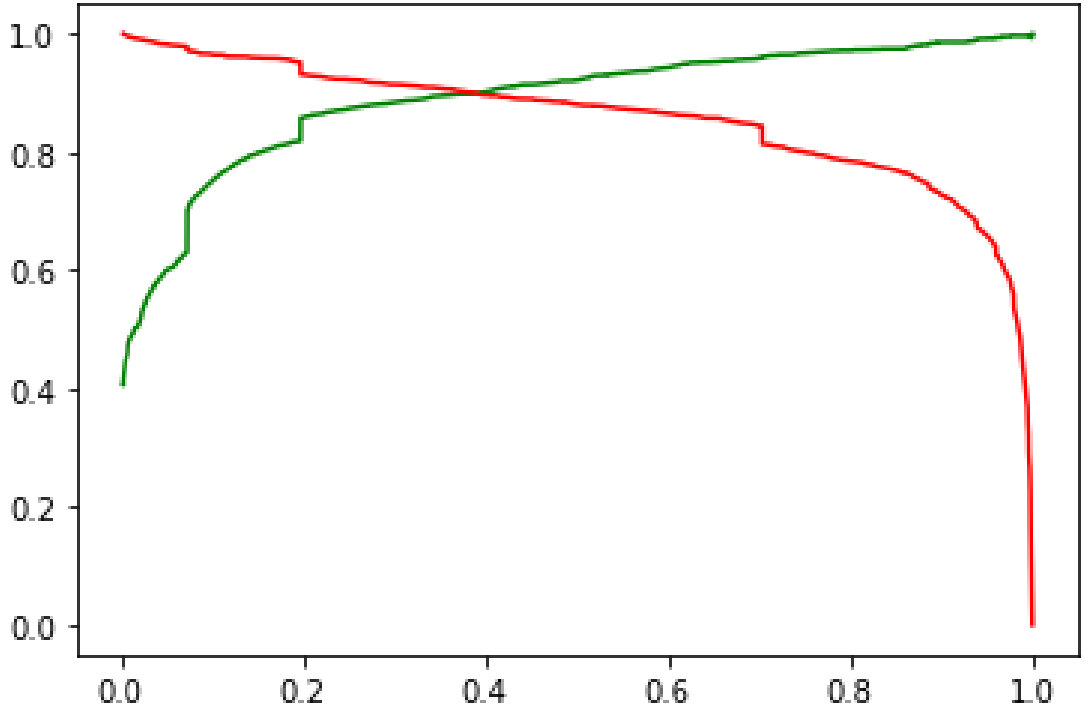
Model Evaluation

- The AUC for the ROC curve is 0.97 which shows that the model is significant



Accuracy, Sensitivity and Specificity





- The cut off point of 0.25 can be considered optimal

Evaluation Metrics

Model Evaluation Parameters

- Accuracy - Train data: 92%, Test Data: 92%
- Sensitivity - Train data: 92%, Test Data: 91%
- Specificity - Train data: 92%, Test Data: 93%
- Precision - Train data: 87%, Test Data: 88%
- Recall - Train data: 92%, Test Data: 91%

CONCLUSION:

- The model is able to accurately predict >90% of the outcomes
- The sales team should employ this model to determine the lead scores of customers
- A score of >25 should be considered a potential lead and efforts should be focused on them to ensure lead conversion
- The top variables which decides if a lead will be converted are time spent on the website, current status of the lead and the source from where the lead is generated
- The company should focus on these variables: encouraging customers to browse the website, increase advertising on the lead sources which leads to higher conversion and focus marketing efforts on engaging customers to ensure that the lead status remains active