Lead Scoring Case Study

Presented By:
Gargi Patel
Mukesh Bhatt
Foram Shah

Problem Statement

- X Education is an education company that sells online courses to industry professionals.
- Many professionals visit the X Education website daily to browse courses.
- When visitors fill out a form with their email address or phone number, they are classified as leads.
- The company also acquires leads through past referrals.
- Despite getting many leads, X Education's lead conversion rate is around 30%, which is considered poor.
- The company aims to identify the most potential leads, also known as "Hot Leads."
- By focusing on Hot Leads, the sales team can improve the lead conversion rate.

Business Objective

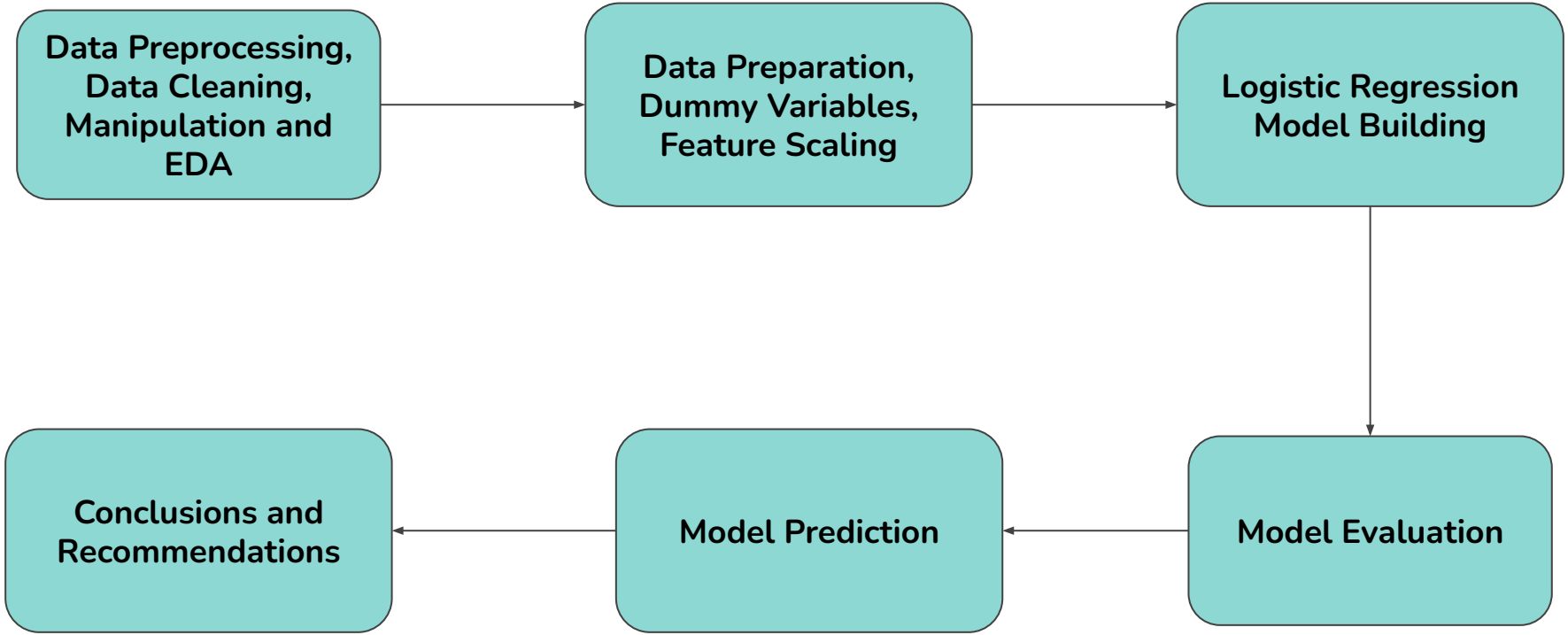
- Build a logistic regression model to score leads between 0 and 100. Higher scores indicate higher likelihood of conversion, allowing the sales team to prioritize these leads.

- Increase the current lead conversion rate from 30% to around 80%. Focus the sales team's efforts on the most promising leads to improve efficiency.

- Ensure the model is flexible enough to adjust to future changes in company requirements.

- The goal is to increase efficiency by having the sales team concentrate on communicating with potential leads rather than contacting everyone.

Approach

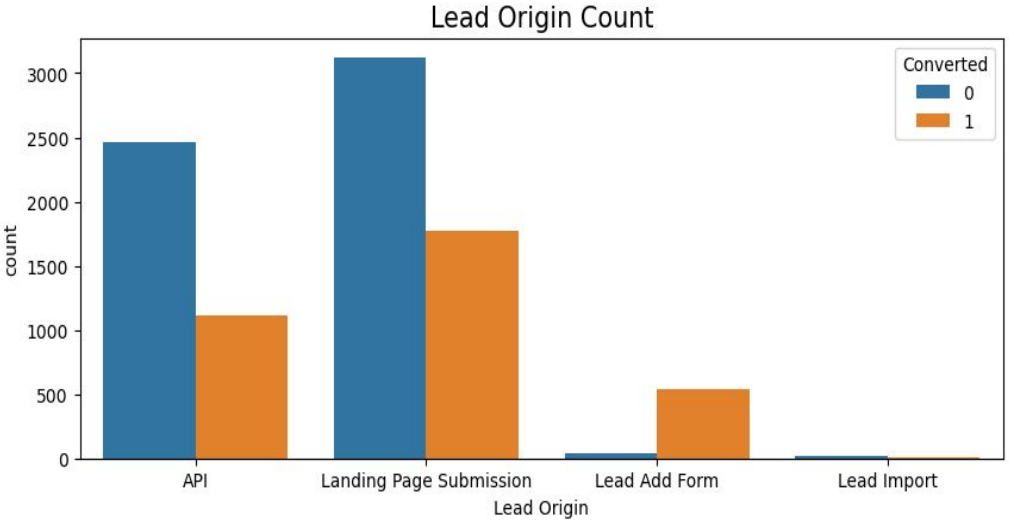

Data Preprocessing, Data Cleaning, Manipulation and EDA
Data Preparation, Dummy Variables, Feature Scaling
Logistic Regression Model Building
Model Evaluation
Model Prediction
Conclusions and Recommendations

Data Preprocessing

- Imported a few libraries which are required in the initial part.
- Imported the data and examined its shape, data types, and descriptive statistics to get an overview.
- There were select values in a few columns. These values were in the data because the customer didn't select any option. It is as good as Null. Hence, we converted the values to Null.
- Checked the values in the data and dropped all the columns that had null values of more than 40%.
- After dropping a few columns, we imputed the null values in a few columns with the mode and also replaced some null values with "Others"
- All the columns which have null values less than 2% we dropped the rows.
- Checked the data imbalance and also checked the duplicate values.

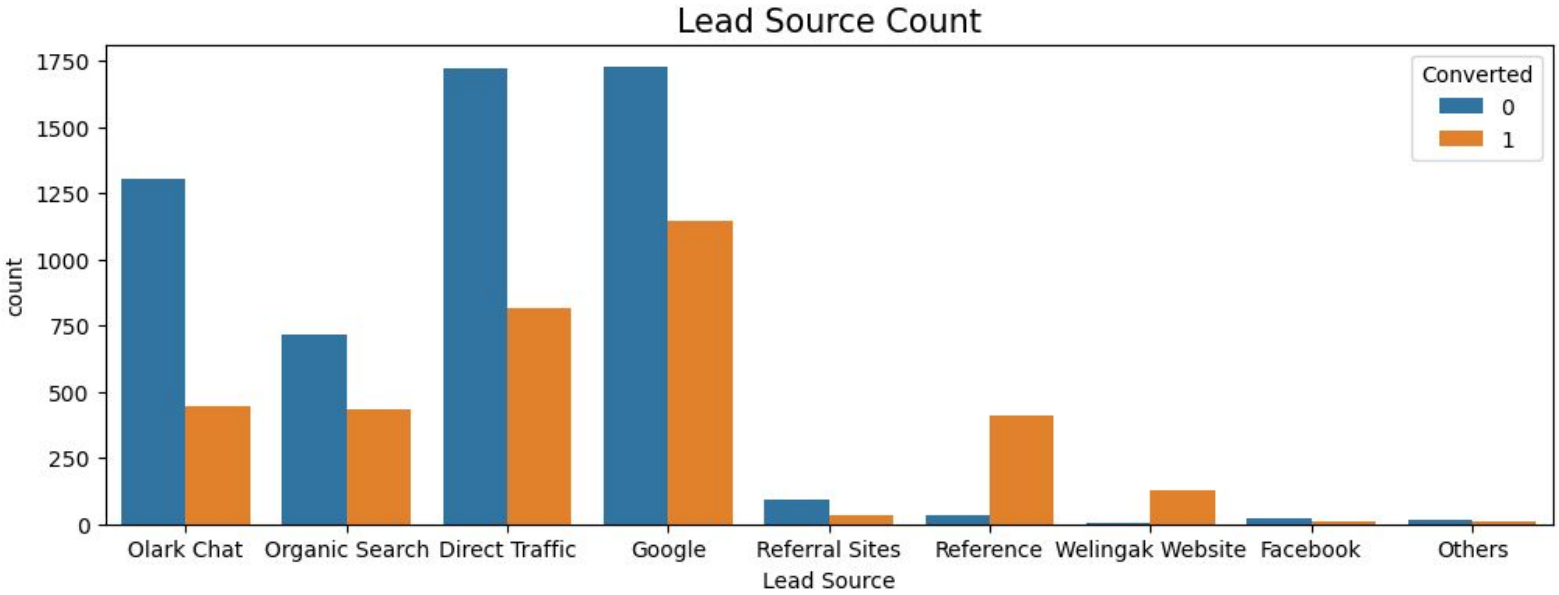
Exploratory Data Analysis

Countplot of Lead Origin



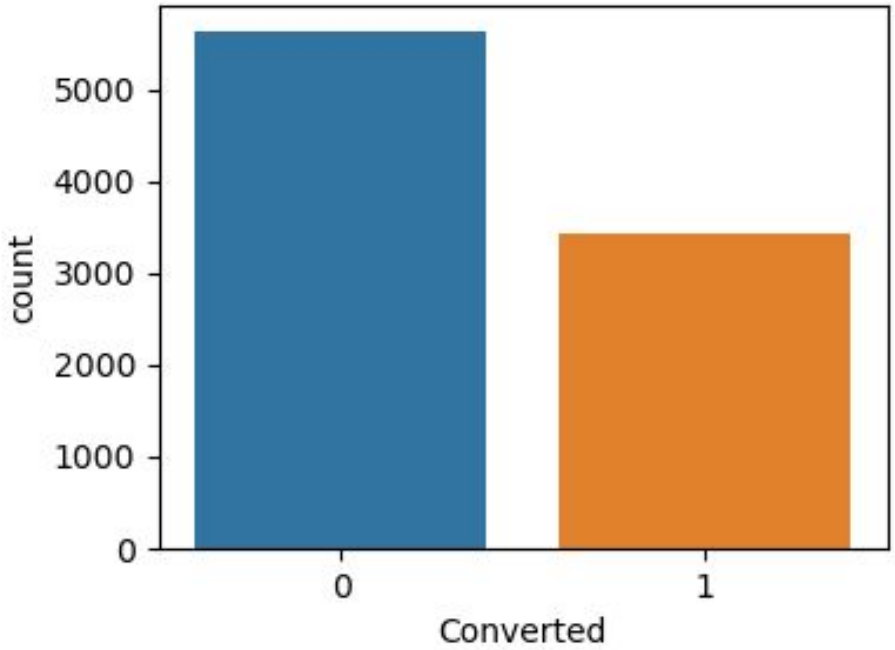
- API and Landing Page Submissions have a conversion rate of 28-34%, and the number of leads originating from them is significant.
- Lead Add form have approx 90% of conversion rate but the leads are less.

Countplot of Lead Source



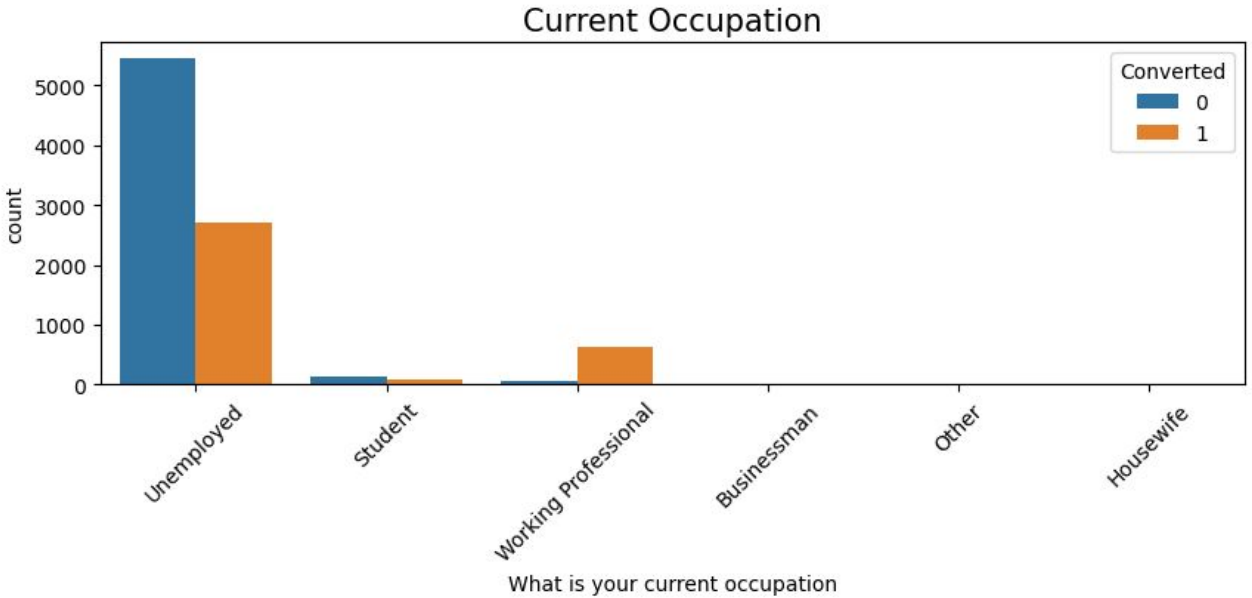
- Google and Direct traffic generate the most leads.
- Reference leads and those from the Welingak website have high conversion rates.
- To boost overall conversion rates, focus on enhancing lead conversion from Olark chat, organic search, direct traffic, and Google leads, while also generating more leads from references and the Welingak website.

Data Imbalance



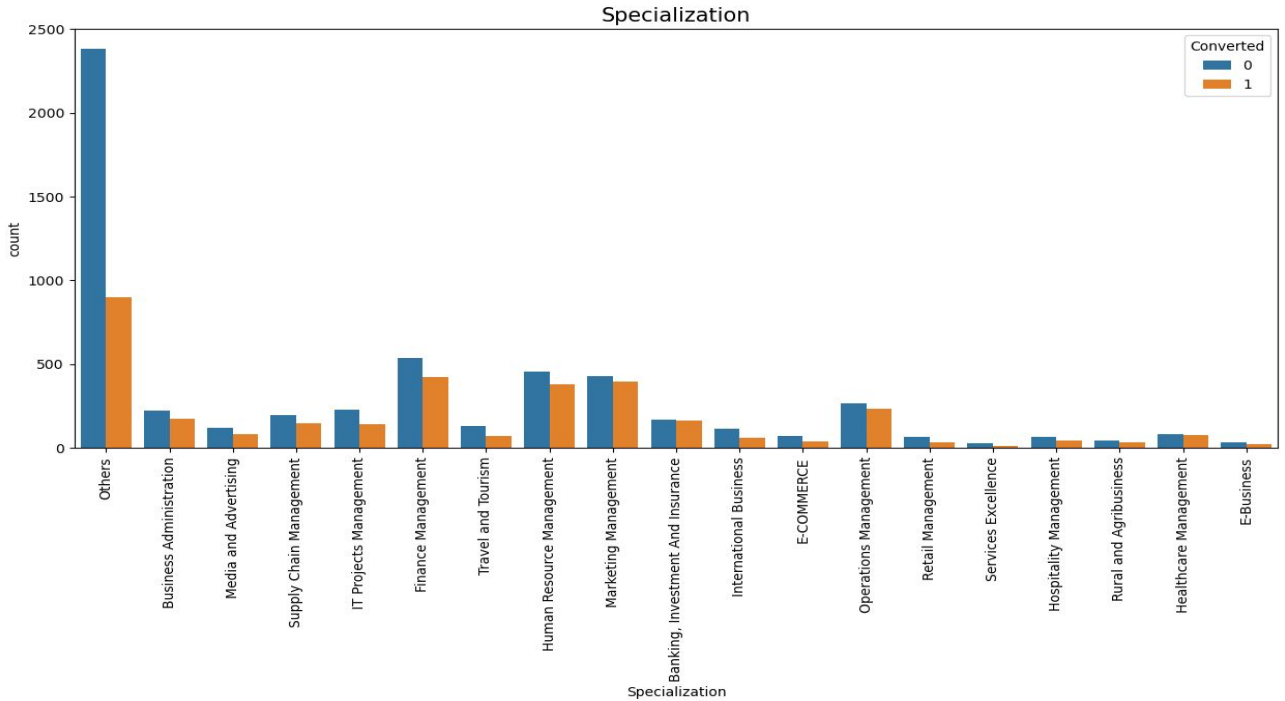
The Lead conversion rate of X Company is 38%

Countplot for Current Occupation



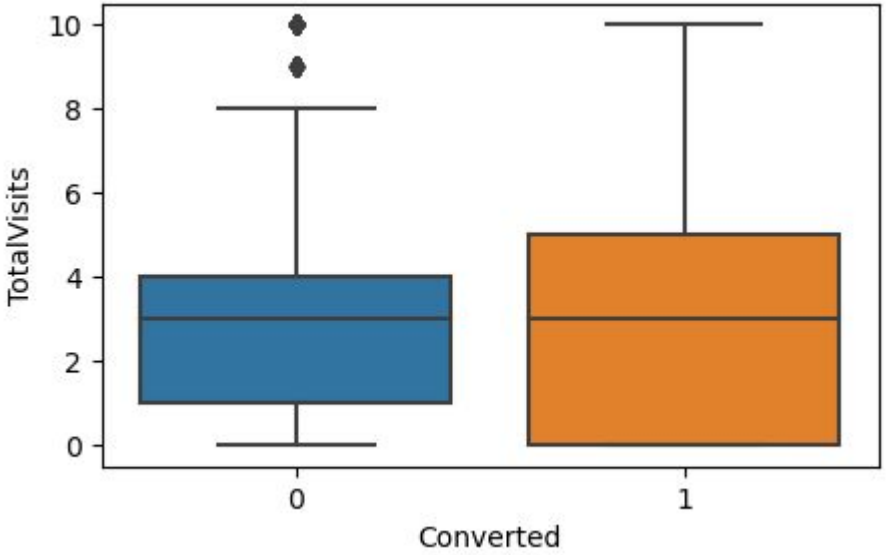
- **It can be observed from the above shown graph that 'Unemployed' people are generating maximum leads.**
- **Conversion rate of 'Working Professionals' is higher.**

Countplot for Specialization



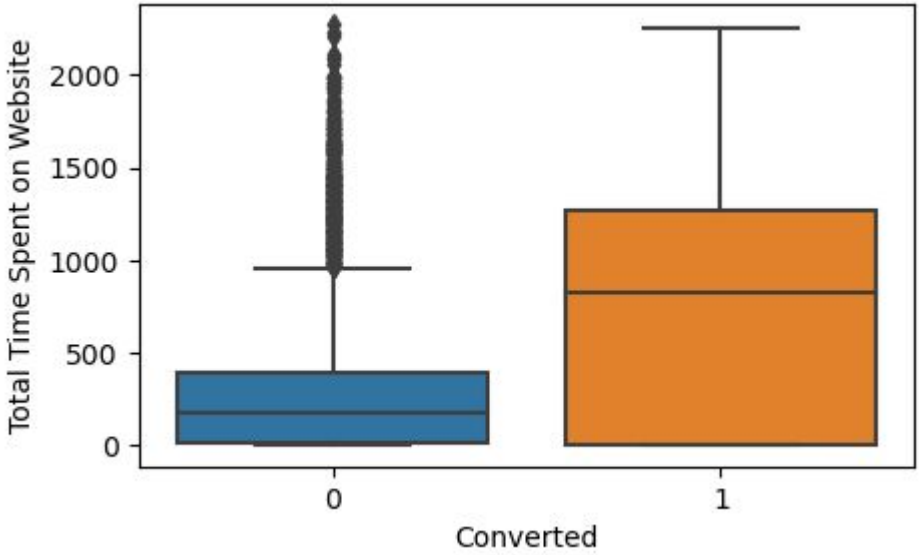
- The specializations for 'Management' altogether are having more number of leads with significant rate of conversion.
- The 'Other' specialization is also generating more number of leads.

Countplot for Total Visits



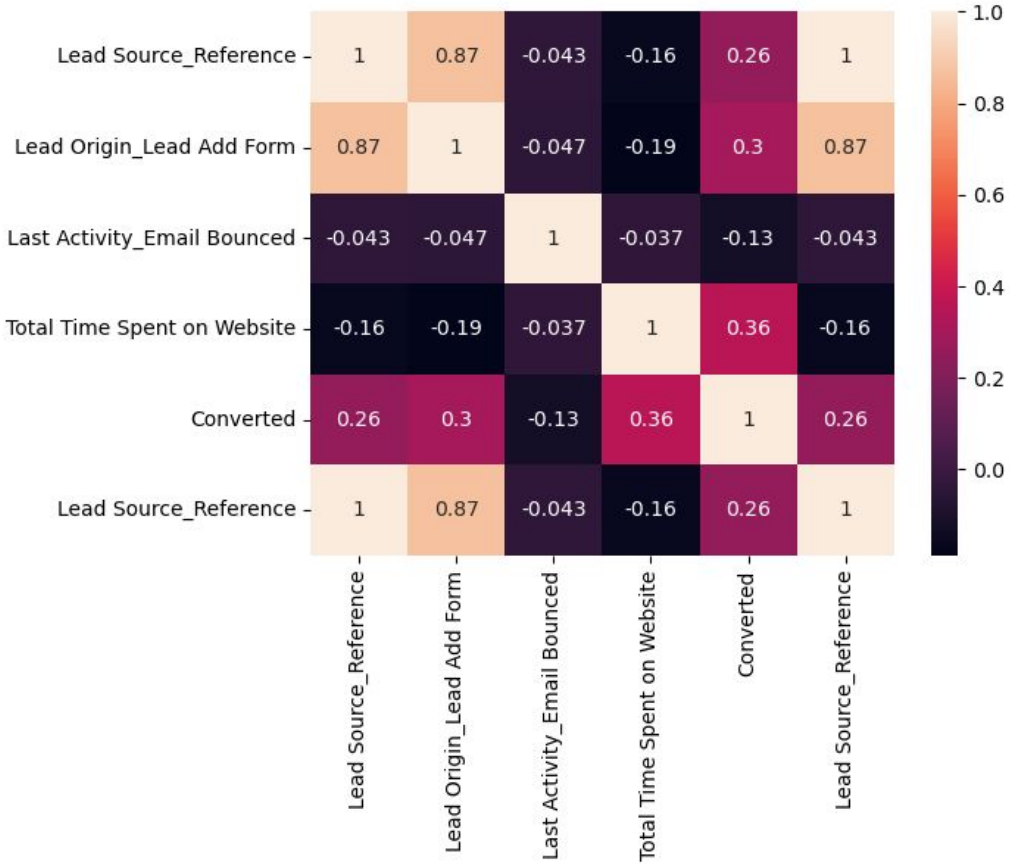
- We are unable to infer anything on totalvisits as the median for converted and not converted leads are same.

Countplot for Total Time spent on website



- It can be observed that the Leads that are spending more time on the website are more likely to be converted

Correlation



- Total time spent on website is positively correlated with Converted.
- Lead Source_Reference has positive correlation with Converted.

Data Preparation

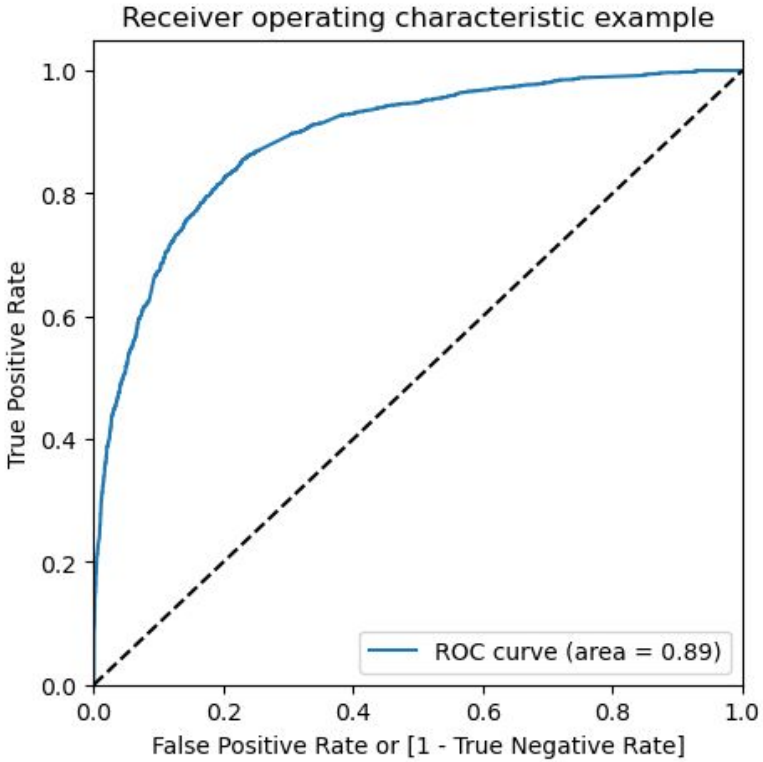
1. Converted all categorical columns to numeric values.
2. Created Dummy variables for categorical columns.
3. Splitting the data into Training and Test Sets with a 70:30 Ratio.
4. Scaling the variables with Standard Scaler.
5. Used RFE for Feature Selection and selected 20 features.

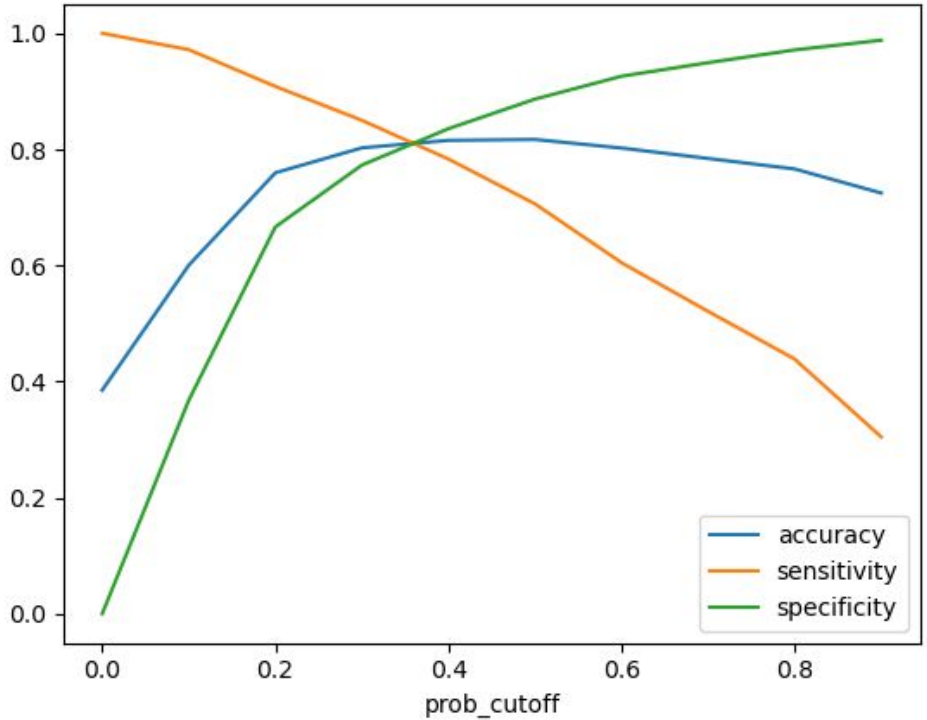
Model Building

1. Building model by systematically removing variables whose p-value is greater than 0.05 one by one.
2. Verified the VIF values and confirmed that all columns had VIF values less than 5.
3. Model 4 was selected as the final model with 18 columns. All the columns had p-values less than 0.05 and VIF less than 5.

Model Evaluation

- Predicted the values on the train set.
- Choose an arbitrary cut-off probability point of 0.5 to find the predicted labels.
- Made the confusion matrix and calculated Accuracy, Sensitivity, Specificity,positive predictive value Negative predictive value.
- Plotted the ROC curve and got a value of 0.89.
- Created columns with different probability cutoffs and calculated accuracy sensitivity and specificity for various probability cutoffs.
- Plotted accuracy sensitivity and specificity for various probabilities and chose 0.35 as the cutoff probability.
- Created a confusion matrix using the new cut-off and calculated accuracy, sensitivity, specificity, positive predictive value, and negative predictive value.
- Also showed the Precision and recall tradeoff.
- Got the Accuracy: 81% Sensitivity: 0.81% Specificity: 80.4%


Receiver operating characteristic example
True Positive Rate
False Positive Rate or [1 - True Negative Rate]
ROC curve (area = 0.89)
0.0
0.2
0.4
0.6
0.8
1.0


prob_cutoff
accuracy
sensitivity
specificity
0.0
0.2
0.4
0.6
0.8
1.0

Conclusion

Based on the analysis it was found that variables that mattered the most in the potential buyers are:

- Total time spent on the website
- When the lead source was:
  - Welingak Website
  - Reference
- When their current occupation is Working Professional
- When Last Activity is ‘Had a Phone Conversation