Leads Score Case Study

Logistic Regression

By - Veena Nagamani M, Veena Bakhtar & Bhavana Yelagandula

PROBLEM STATEMENT

X Education is an organization which provides online courses for industry professional. The company marks its courses on several popular websites like google.

X Education wants to select most promising leads that can be converted to paying customers.

Although the company generates a lot of leads only a few are converted into paying customers, wherein the company wants a higher lead conversion. Leads come through numerous modes like email, advertisements on websites, google searches etc.

The company has had 30% conversion rate through the whole process of turning leads into customers by approaching those leads which are to be found having interest in taking the course. The implementation process of lead generating attributes are not efficient in helping conversions.

BUSINESS GOAL

The company requires a model to be built for selecting most promising leads.

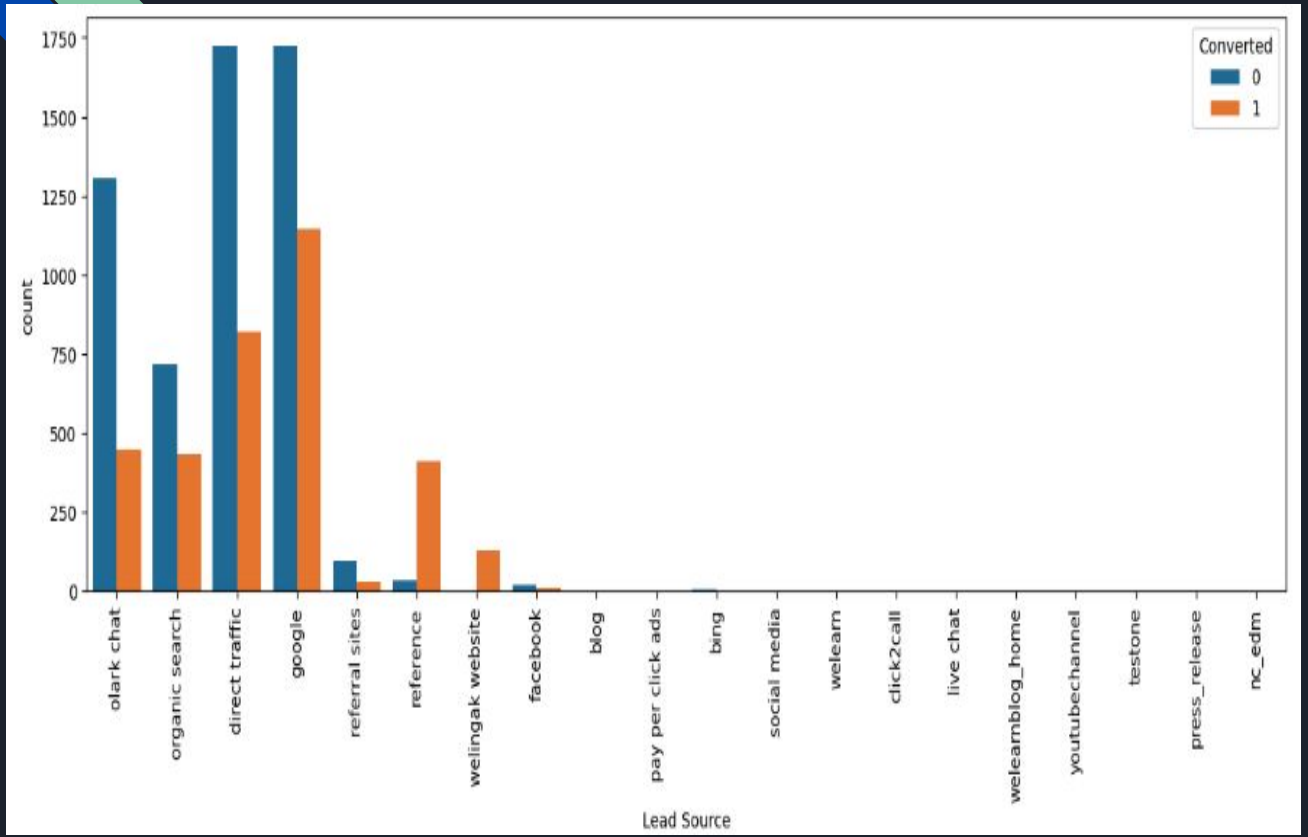
Lead score to be given to each leads such that it indicates how promising the lead could be. The higher the lead score the more promising the lead to get converted, the lower it is the lesser the chances of conversion.

The model to be built in lead conversion rate around 80% or more.

STRATEGY

- Import data
- Clean and prepare the acquired data for further analysis
- Exploratory data analysis for figuring out most helpful attributes for conversion
- Scaling features
- Prepare the data for model building
- Build a logistic regression model
- Assign a lead score for each leads
- Test the model on train set
- Evaluate model by different measures and metrics
- Test the model on test set
- Measure the accuracy of the model and other metrics for evaluation

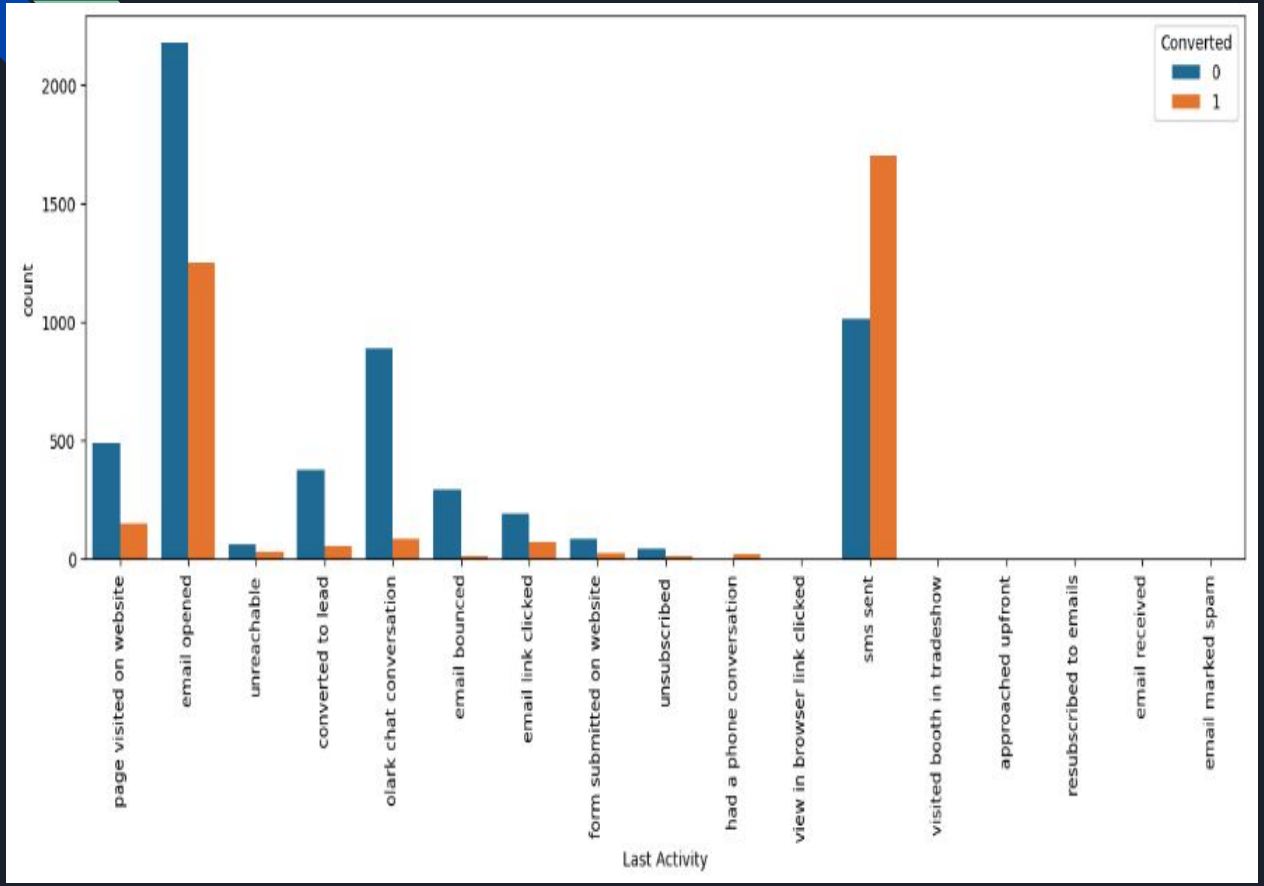
EXPLORATORY DATA ANALYSIS



Lead Source vs Converted

Google searches has had high conversions compared to other modes, whilst **references** has had high conversion rate.

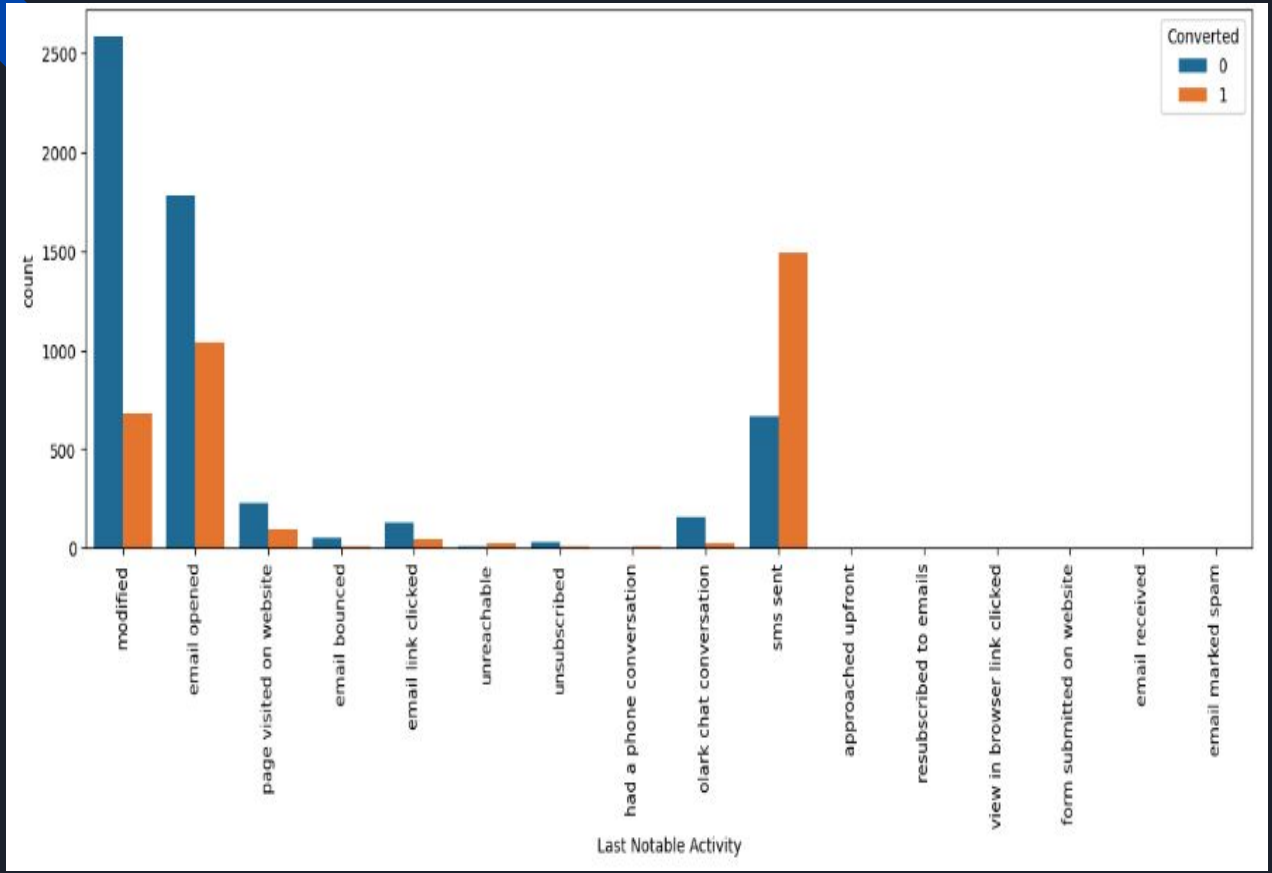
EXPLORATORY DATA ANALYSIS



Last Activity vs Converted

SMS has shown to be a promising method for getting higher confirmed leads, emails also has high conversions.

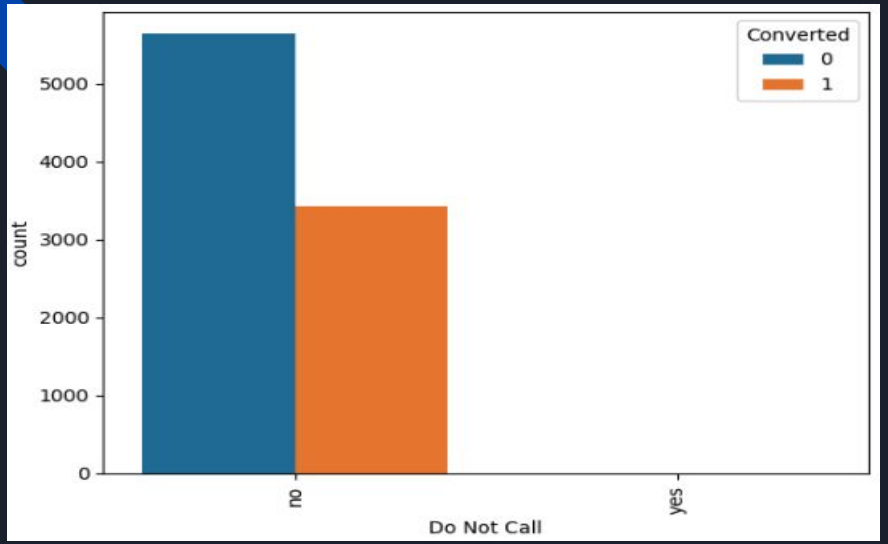
EXPLORATORY DATA ANALYSIS



Last Notable Activity vs Converted

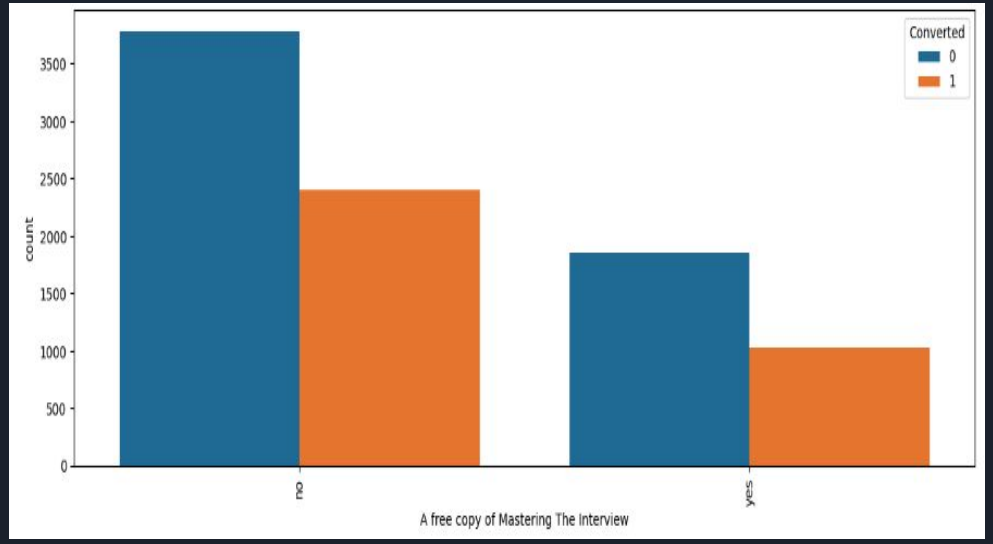
Most leads are converted with messages. Emails also induce leads.

EXPLORATORY DATA ANALYSIS



Do not call vs Converted

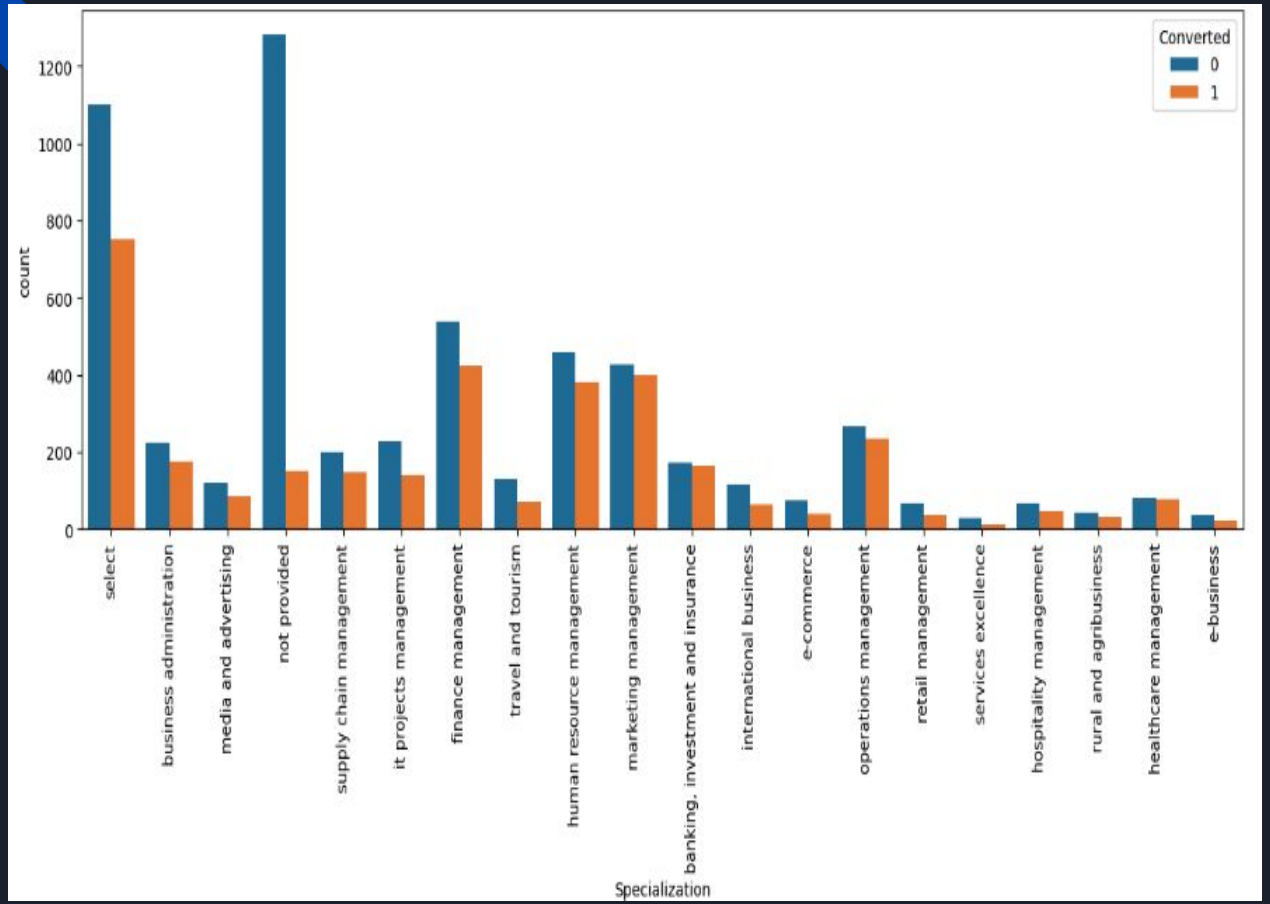
Most leads prefer not to be called through phone



A Free copy of Mastering the Interview vs Converted

We observe that Leads prefer less copies of interviews.

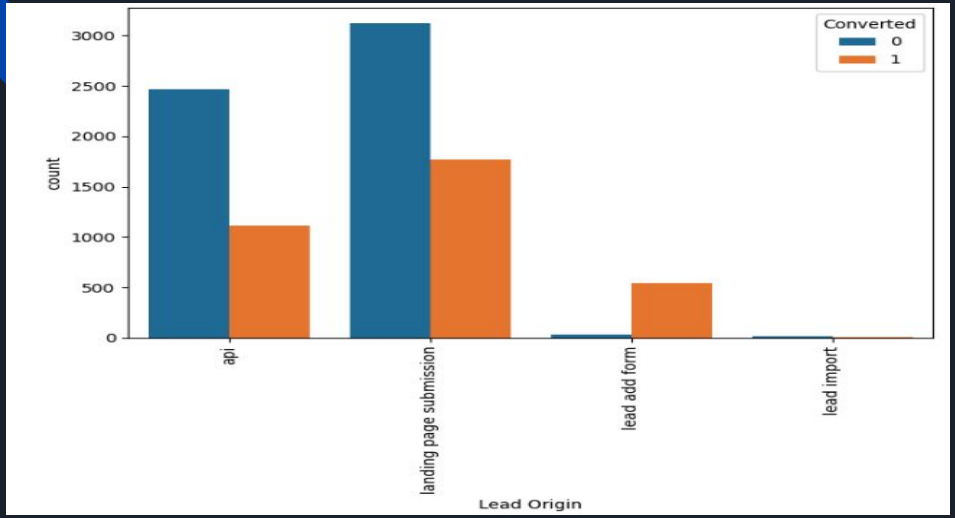
EXPLORATORY DATA ANALYSIS



SPECIALIZATION VS CONVERTED

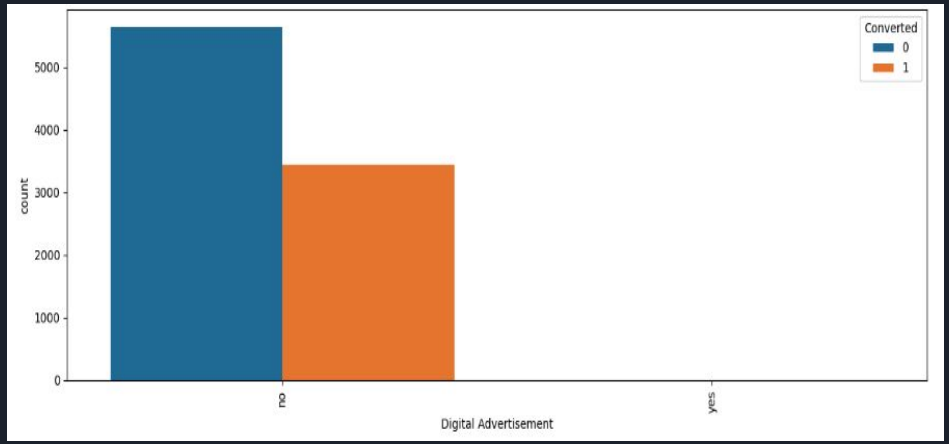
Most of the leads have no information about specialization. On the other hand, marketing management, human resources management has high conversion rates. People from these specializations can be promising leads

EXPLORATORY DATA ANALYSIS



LEAD ORIGIN VS CONVERTED

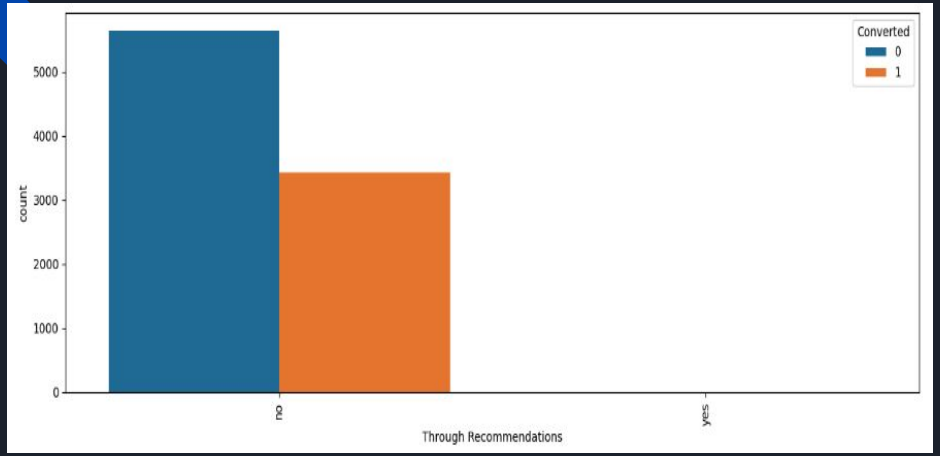
Landing page submissions has had high lead conversions



DIGITAL ADVERTISEMENTS VS CONVERTED

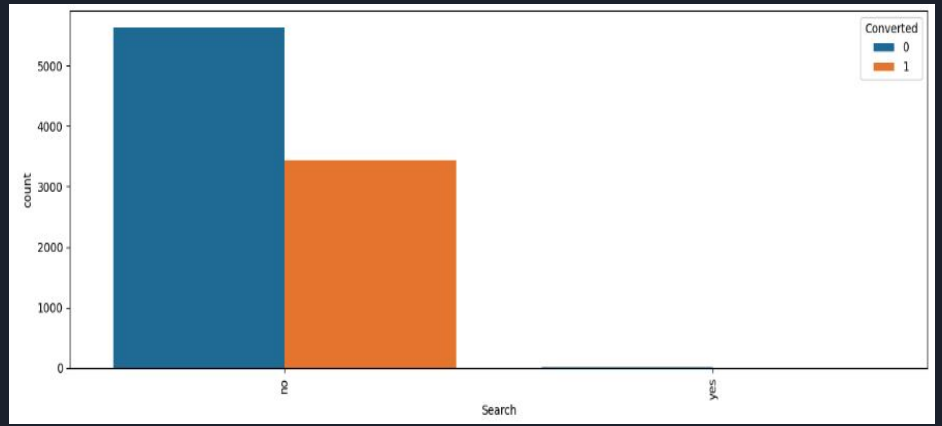
Digital advertisements do not have promising leads

EXPLORATORY DATA ANALYSIS



THROUGH RECOMMENDATIONS VS CONVERTED

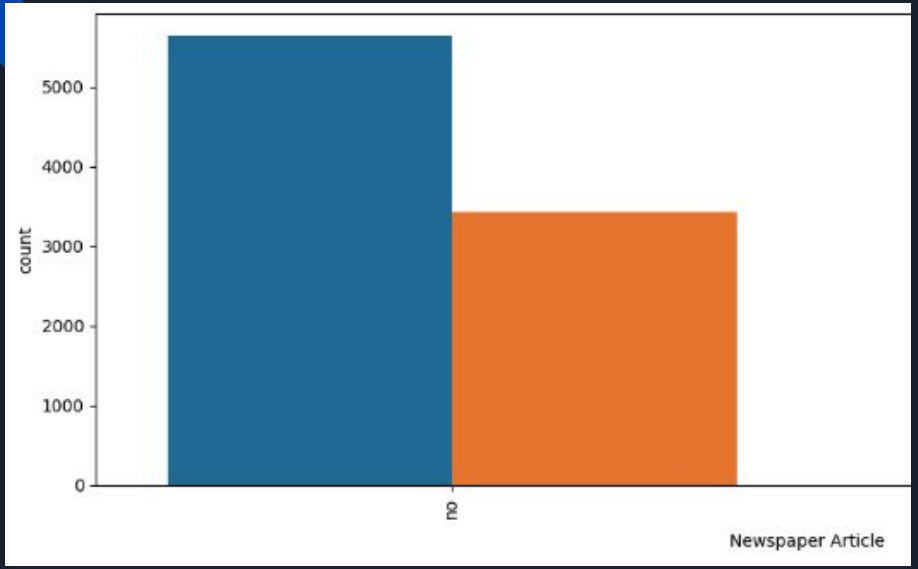
Recommendations are not a good source for promising leads



Search VS CONVERTED

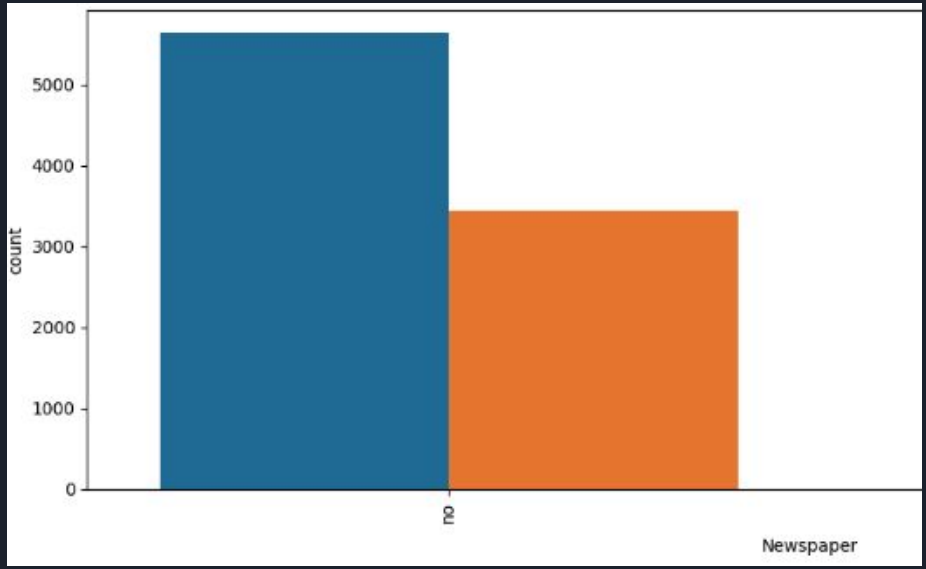
Searches are not a good source for leads

EXPLORATORY DATA ANALYSIS



Newspaper Article vs Converted

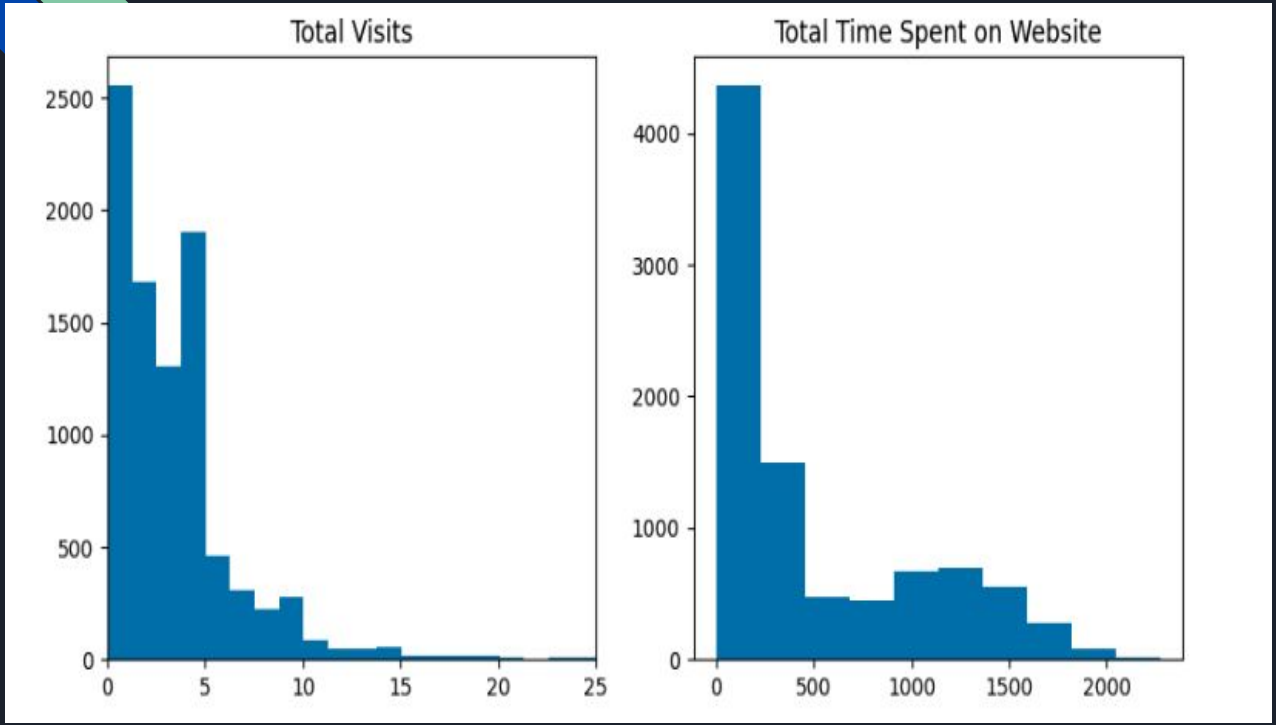
Newspaper Article does not have any conversion rate.



Newspaper vs Converted

Newspaper does not have any conversion rate.

EXPLORATORY DATA ANALYSIS

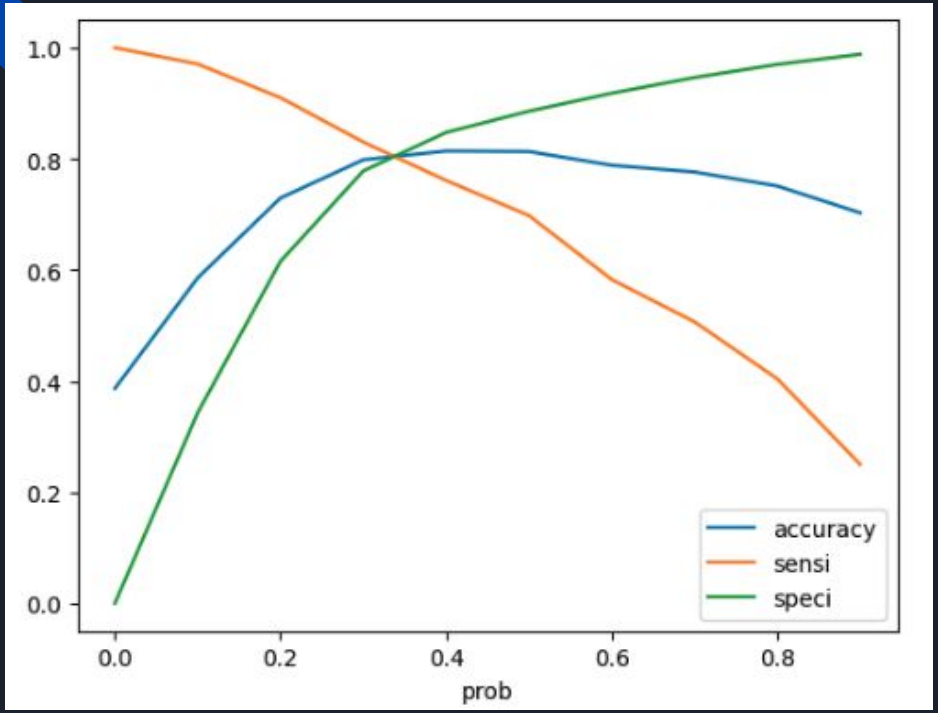


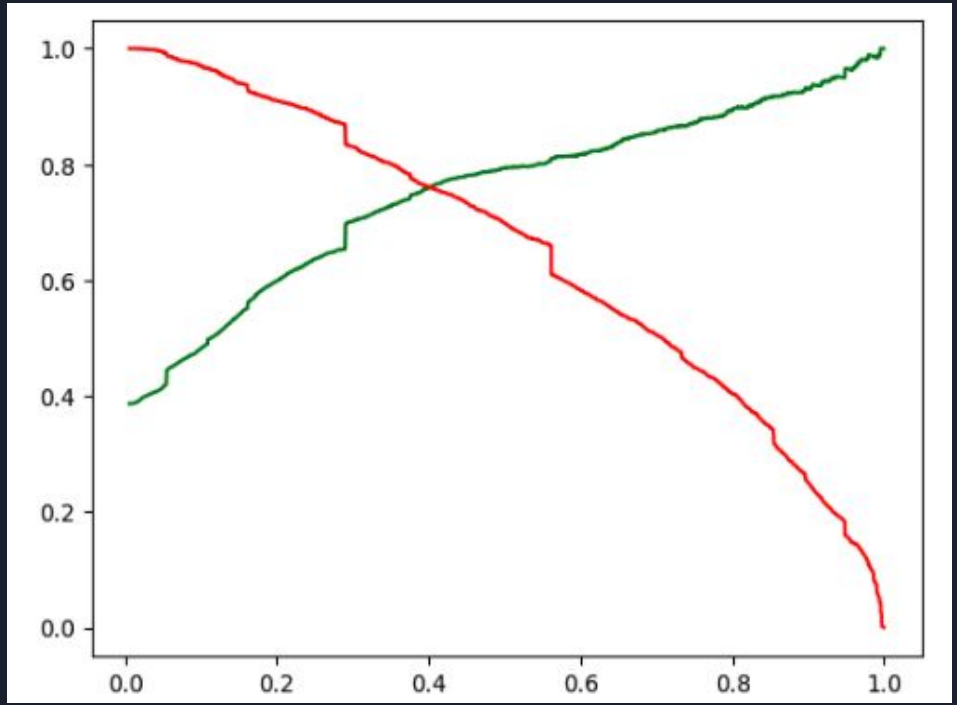
Total Visits and Total time spent on Website has high chances being promising leads.

MODEL BUILDING

- Splitting into train and test set
- Scale variables in train set
- Build the first model
- Use RFE to eliminate less relevant variables
- Build the next model
- Eliminate variables based on high p-values
- Check VIF value for all the existing columns
- Predict using train set
- Evaluate accuracy and other metric
- Predict using test set
- Precision and recall analysis on test predictions

MODEL EVALUATION (TRAIN)





80.7% Accuracy

80% Sensitivity

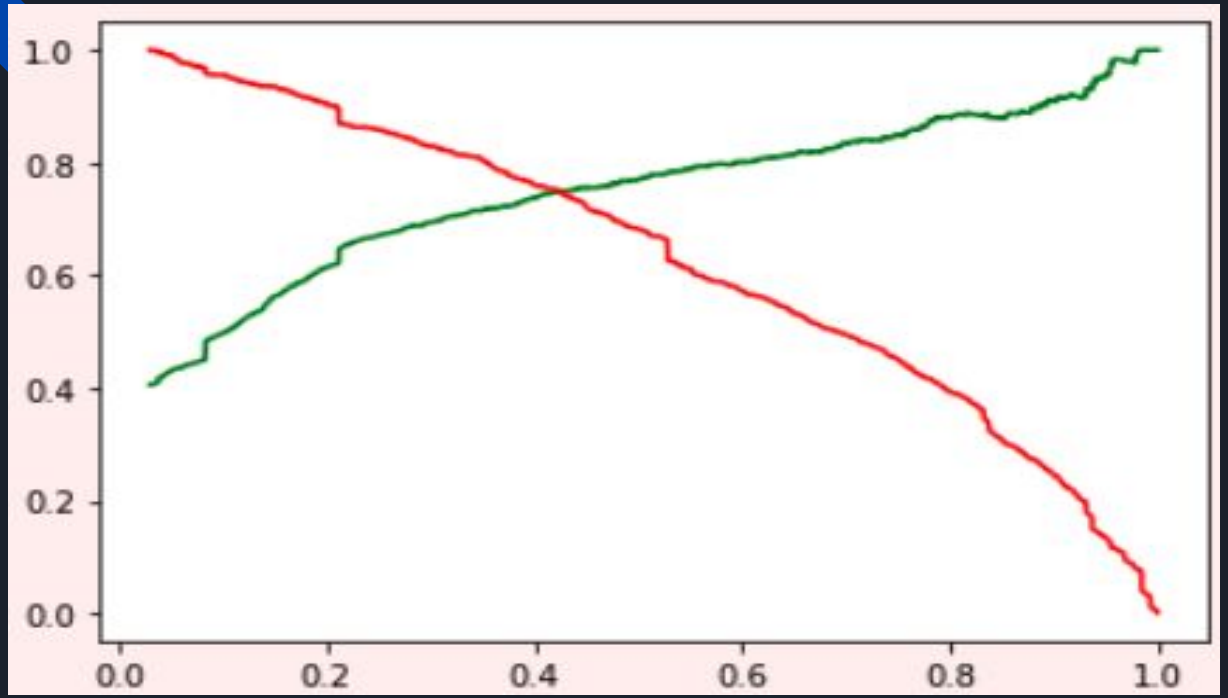
81.1% Specificity

With 0.35 cut off we have

79.3% Precision

69.7% Recall

MODEL EVALUATION (Test)



PRECISION AND RECALL

With 0.4 cut off we have

74% Precision

76% Recall

81% Accuracy

CONCLUSION

EDA Recommendations:

Company should contact the following leads as they are more likely to get converted:

- Leads coming from the lead sources "Welingak Websites" and "Reference".
- Leads who spent "more time on the websites".
- Leads who are the "working professionals".
- Leads whose last activity was "SMS Sent".
- Leads coming from the lead sources "Olark Chat".

Company should not contact the following leads as they are not likely to get converted:

- Leads whose last activity was "Olark Chat Conversation".
- Leads who chose the option of "Do not Email" as "yes".

LOGISTIC REGRESSION MODEL:

- The model shows high close to 81% accuracy
- The threshold has been selected from Accuracy, Sensitivity, specificity measures and precision, recall curves.
- The model finds correct promising leads and leads that have less chances of getting converted
- Overall this model proves to be accurate

THANK YOU!