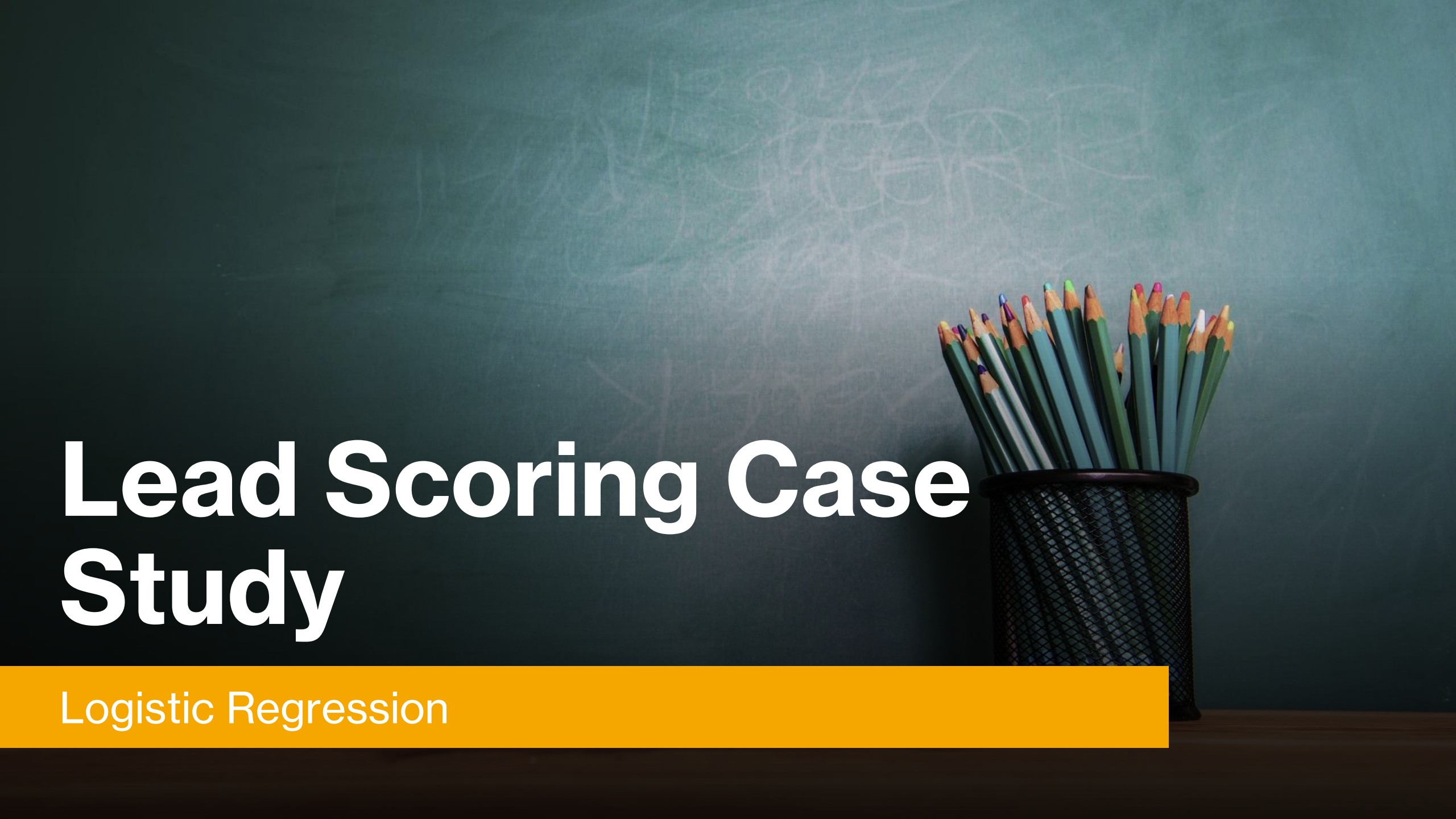


Lead Scoring Case Study

Logistic Regression

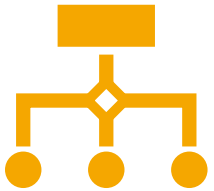
Problem Statement:

- X Education is an organization which provides online courses to industry professionals. The company marks it's courses on many 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 rate. 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 conversations.

Business Goal:

The company wants to build a model for selecting the most promising leads.

A lead score has to be given to each leads to indicate how promising the lead could be. The higher the lead score, the more promising the lead is to get converted. The lower the score, the lesser the chances of conversion.

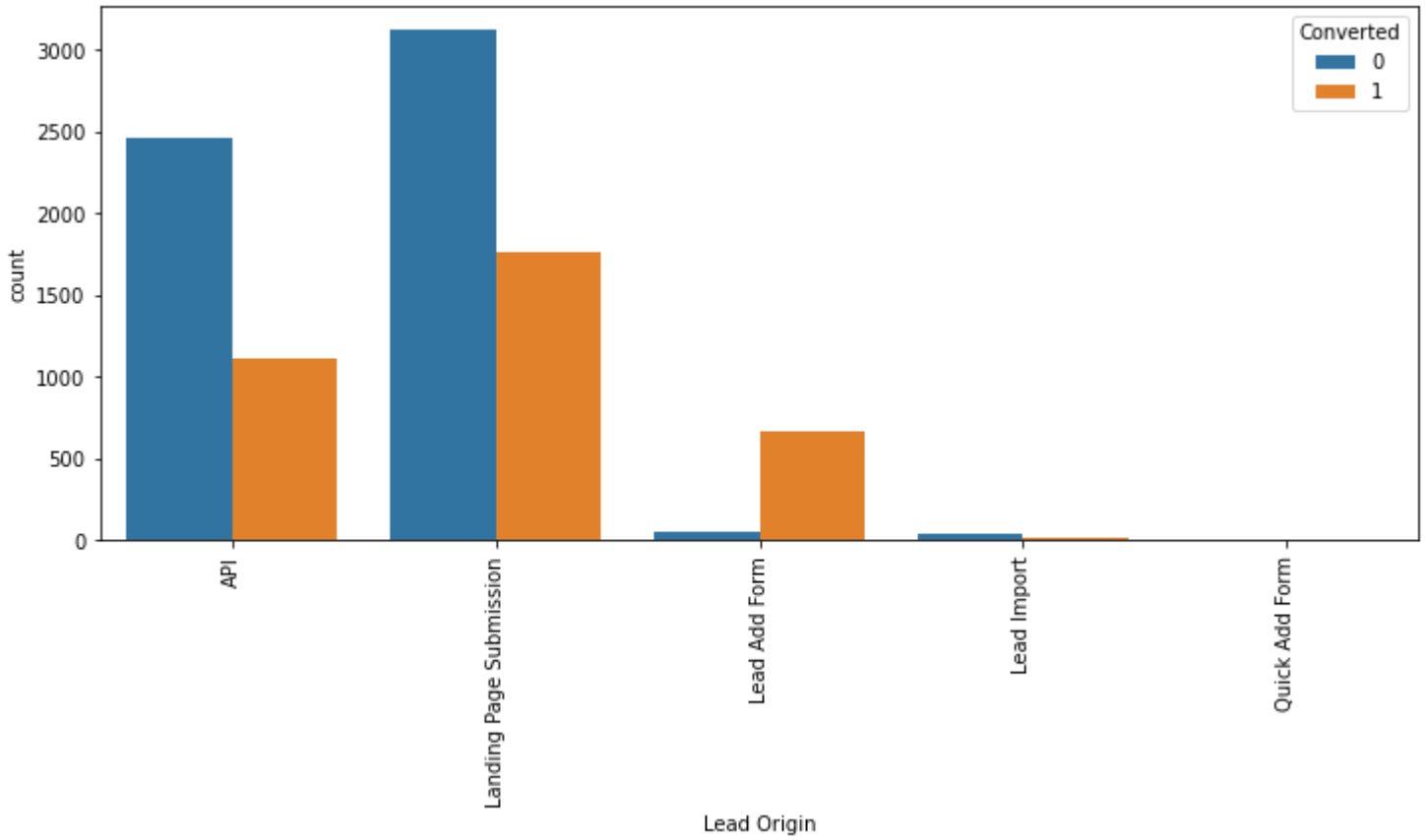


The built model should have a conversion rate of around 80% or more.

Strategy:

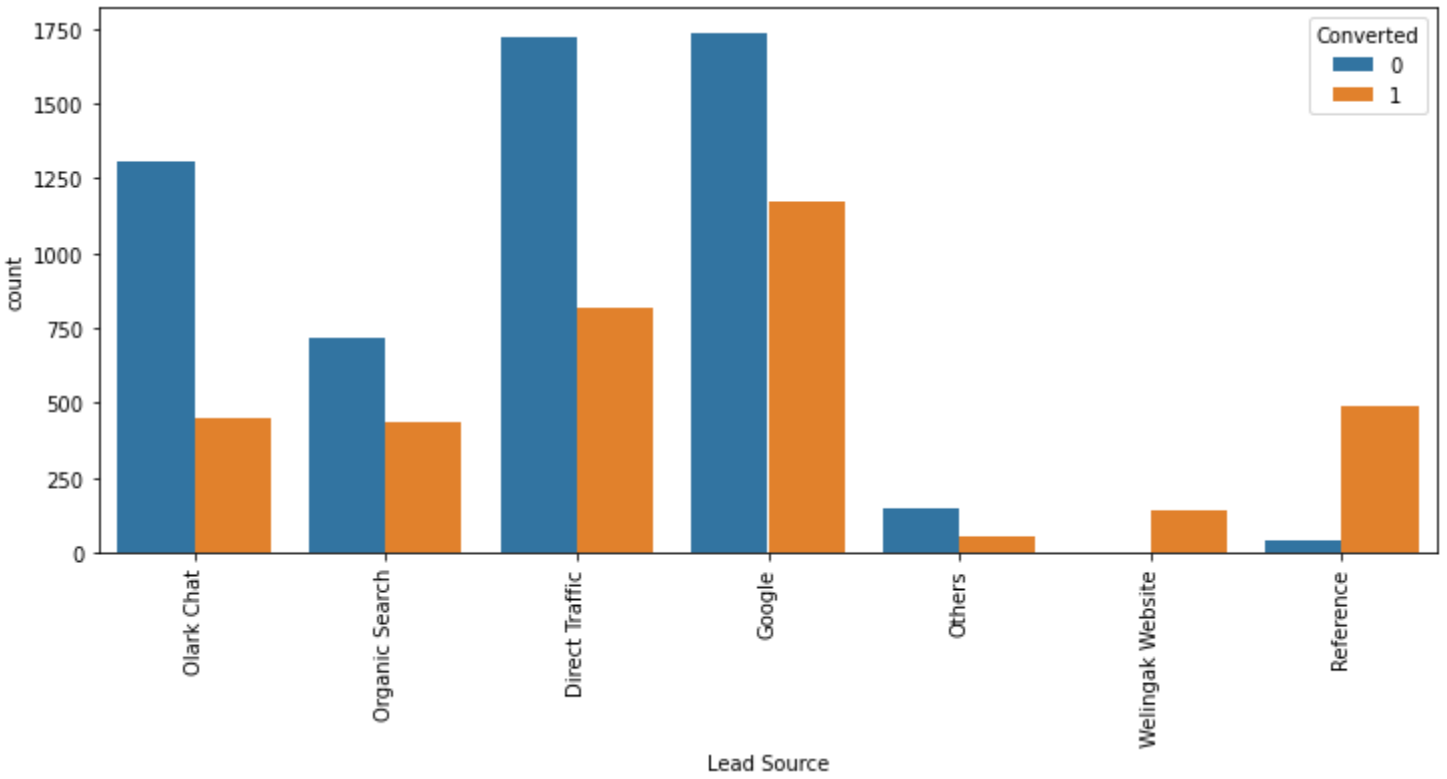
- Importing Data and Libraries
- Checking the Dataframe
- Preparation the Data
- Exploratory Data Analysis
- Outlier Detection and Treatment
- Creating Dummy Variables
- Train - Test Split
- Feature Scaling
- Model Building and Feature Selection using RFE
- Creating Confusion Matrix
- Plotting the ROC Curve and finding optimal cutoff point
- Precision and Recall and F1 Score
- Making predictions on the test set
- Assigning Lead Score with respect to Lead_Num_ID

Exploratory Data Analysis:



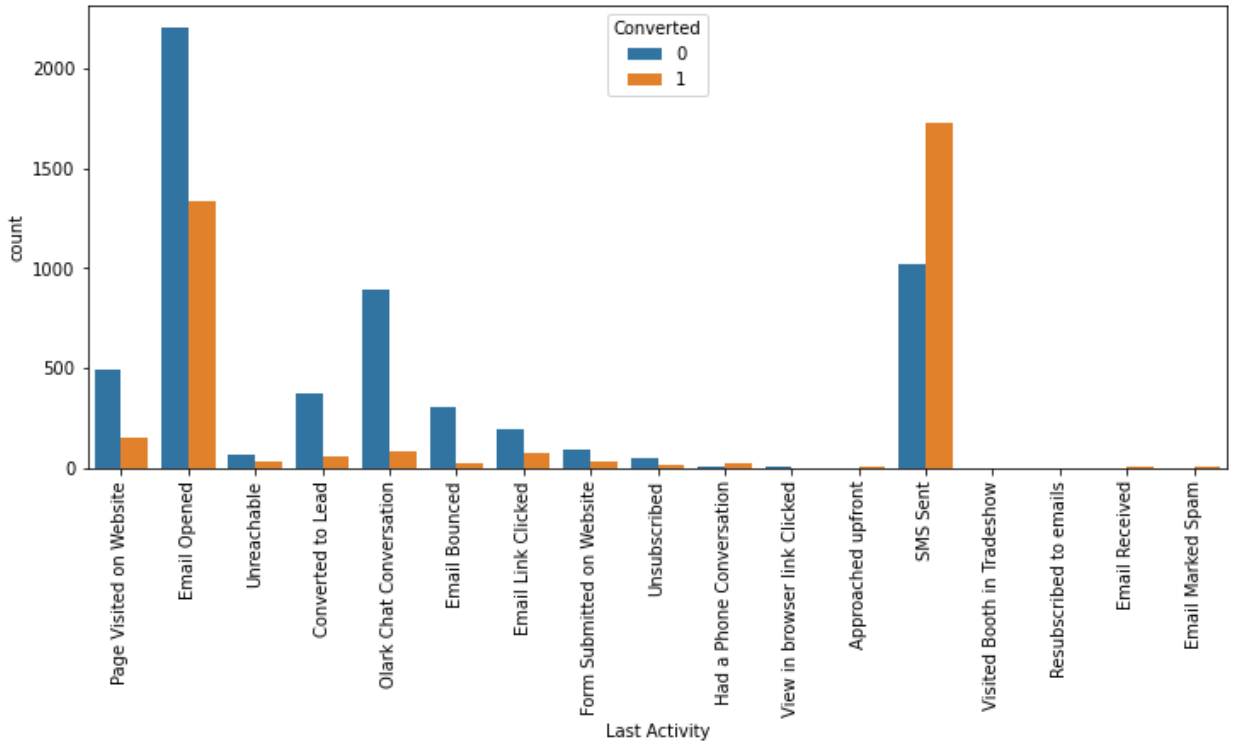
- **Lead Origin v/s Converted:**

- Leads who originate from Lead Add Form have high percentage of conversion compared to API and Landing Page Submissions.



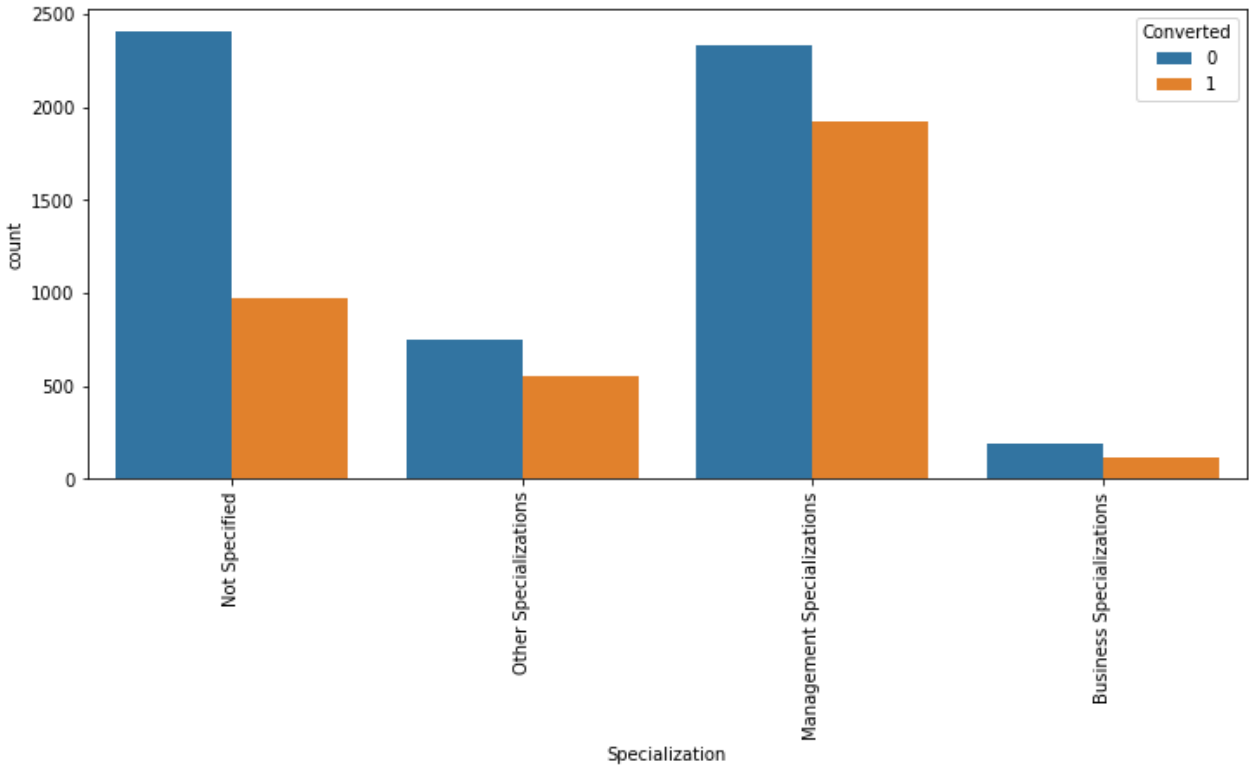
- **Lead Source v/s Converted:**

  - Conversion rate of Reference source is maximum alongside the ones which originates directly from Google.
  - Other sources such as Direct traffic and Olark chat etc. have comparatively low conversion rates.



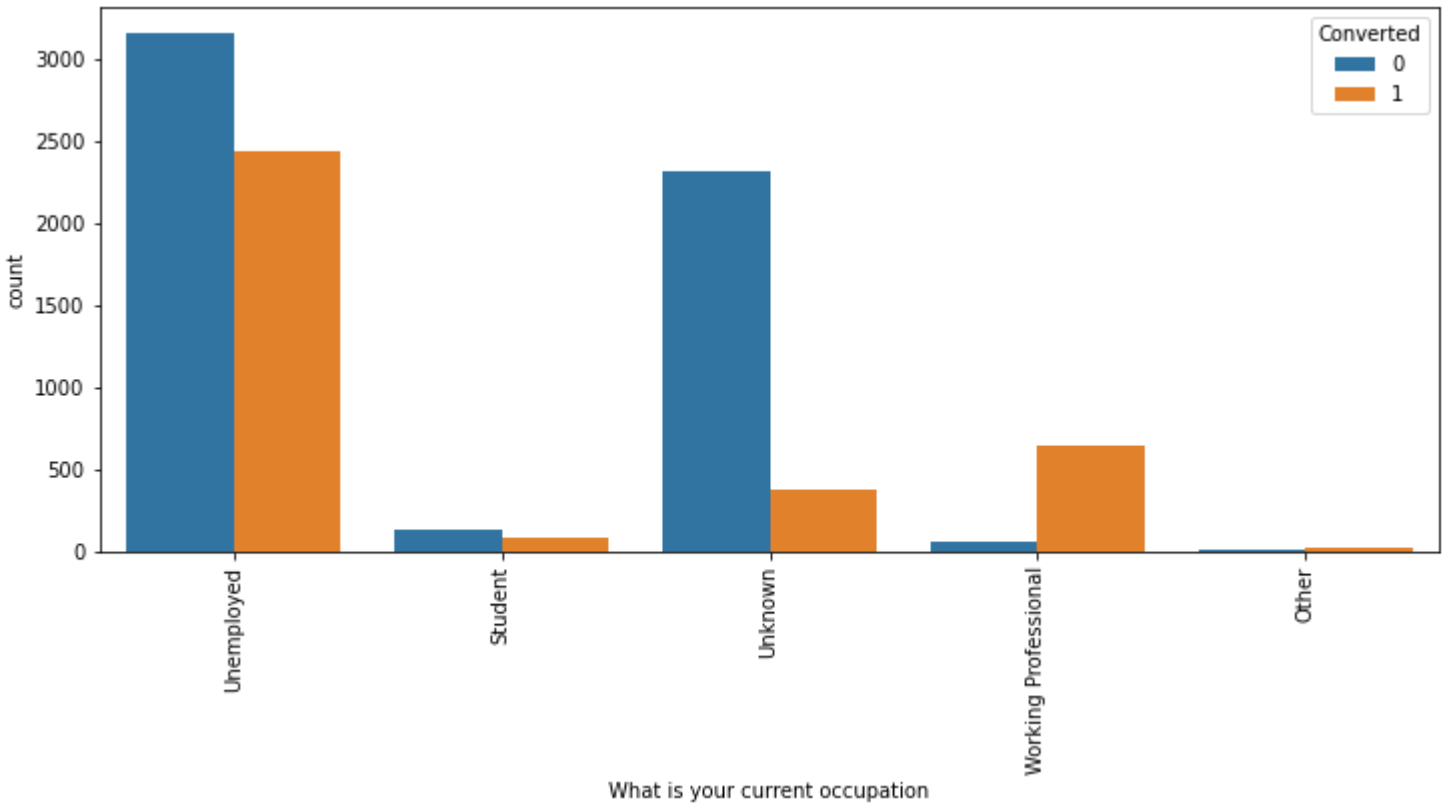
- **Last Activity v/s Converted:**

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



- **Specialization v/s Converted:**

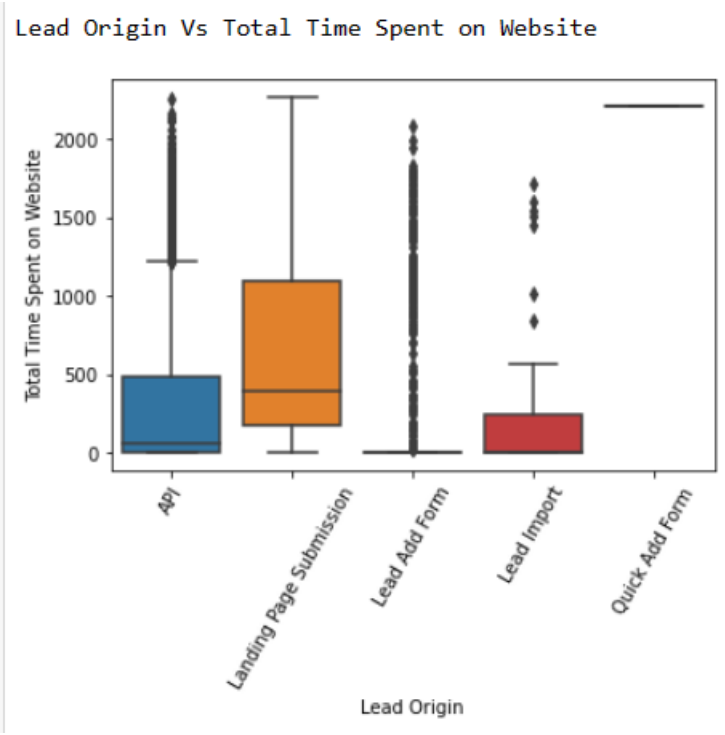
- The customers who have worked a particular specialization have low conversion rate. It could have happened that most of the customers who have not defined specialization are students or have no prior work experience.

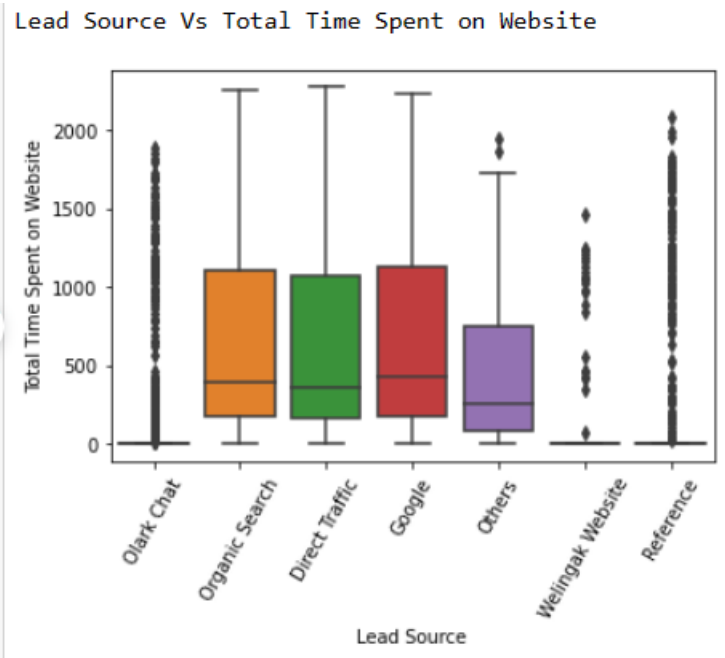


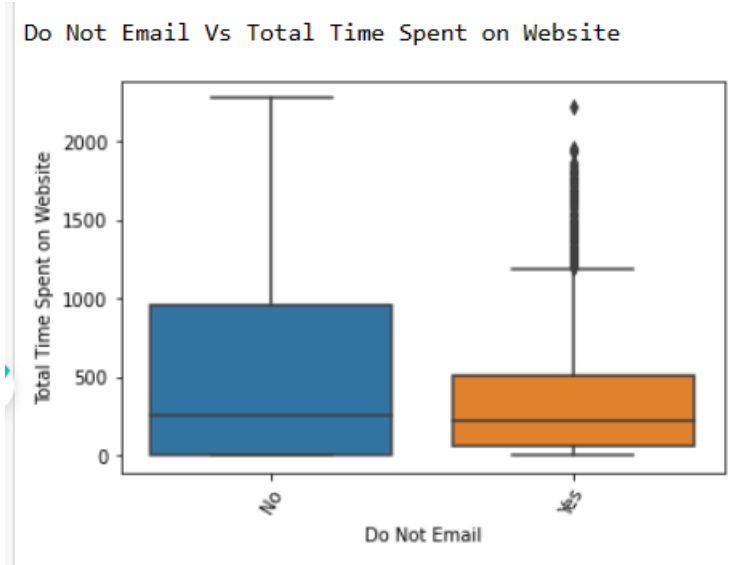
- **What is your current occupation v/s Converted:**

- The conversion rate of working professionals is high compared to unemployed , the reason might be the working professional are aware of current market requirements and accordingly they upskill themselves.

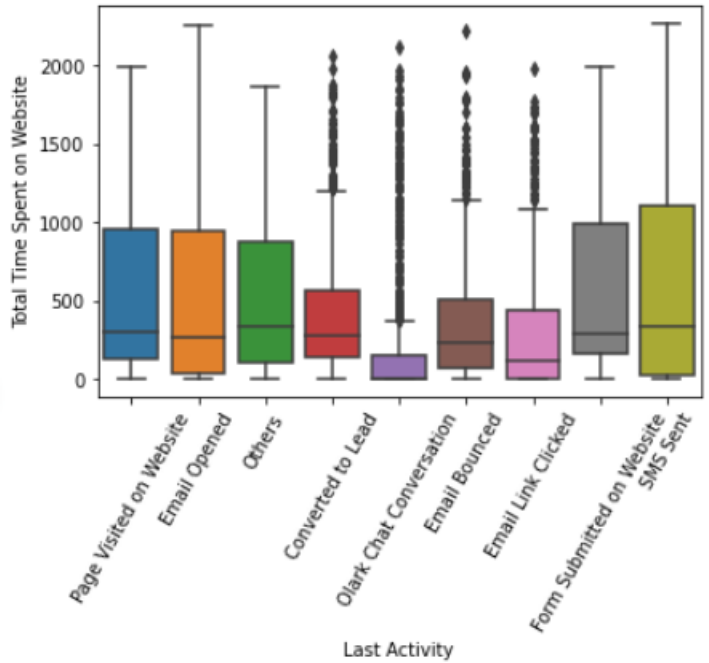
Bivariate Analysis (Categorical vs Continuous)



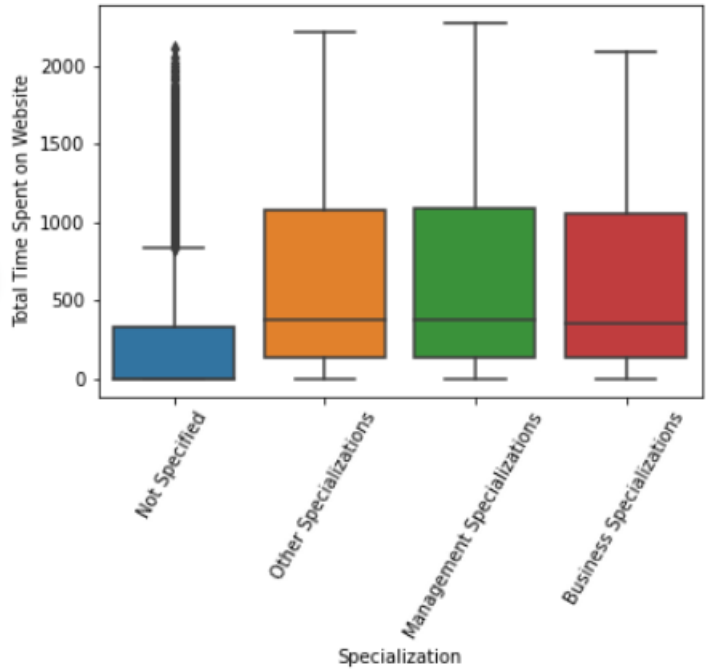




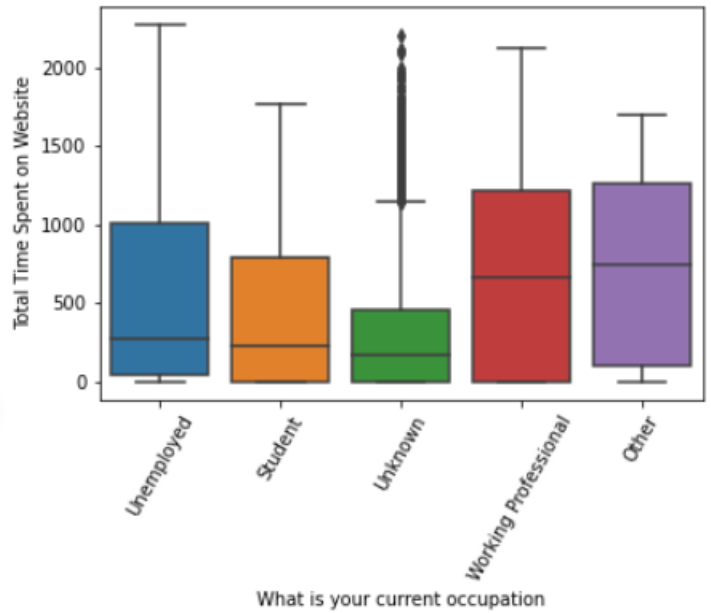
Last Activity Vs Total Time Spent on Website



Specialization Vs Total Time Spent on Website



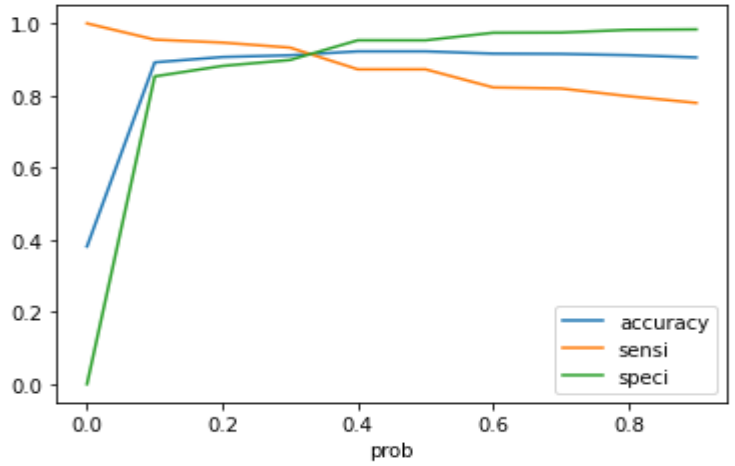
What is your current occupation Vs Total Time Spent on Website



Model Building:

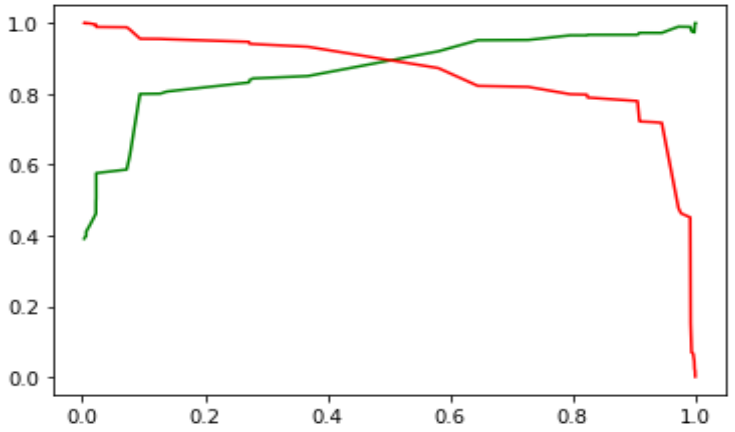
- Splitting into train and test set.
- Scale and fit transform variables in train set.
- Build the first model.
- Use RFE to eliminate less relevant variables.
- Build the next model.
- Eliminate variables based on high p-values.
- Check VIF value for all the existing columns.
- Predict using train set.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

Model Evaluation (Train):



Accuracy, Sensitivity and Specificity:

- 88% Accuracy
- 86% Sensitivity
- 88% Specificity



Precision, Recall and F1 score:

- 82.7% Precision
- 86% Recall
- 84.52% F1 score

According to the predictions of our model on the Test Data below are the performance metrics

Metrics	Percent %
Accuracy	88.77
Sensitivity	83.69
Specificity	92
False_postive_rate	7.9
Precision	87.12
Recall	83.69
F1_Score	84.52

Conclusion:

- People spending more than average time are promising leads, so targeting and approaching them could be helpful in conversions.
- SMS messages can have a high impact on lead conversions.
- Landing page submissions can help find out more leads.
- Marketing management, human resources management has high conversion rates. People from those specializations could be promising leads.
- References and offers for referring a lead can be a good source for high conversions.
- An alert messages or information has seen to have high lead conversion rate.
- The threshold has been selected from Accuracy, Sensitivity, Specificity measures and precision, recall curves.
- The train and test data has an approx accuracy of 88% which concludes that Logistic Regression model is a good fit.
- The model shows high accuracy which is 88% and a recall rate of 83% with precision around 87%.
- This implies the model is around 83% good in predicting the positive class (Hot Leads).
- Overall this model proves to be accurate.


THANK YOU!