LEAD SCORE CASE STUDY

POOJA MATHUR , VIRAJ PATEL, PURNIMA YADAV

PROBLEM STATEMENT

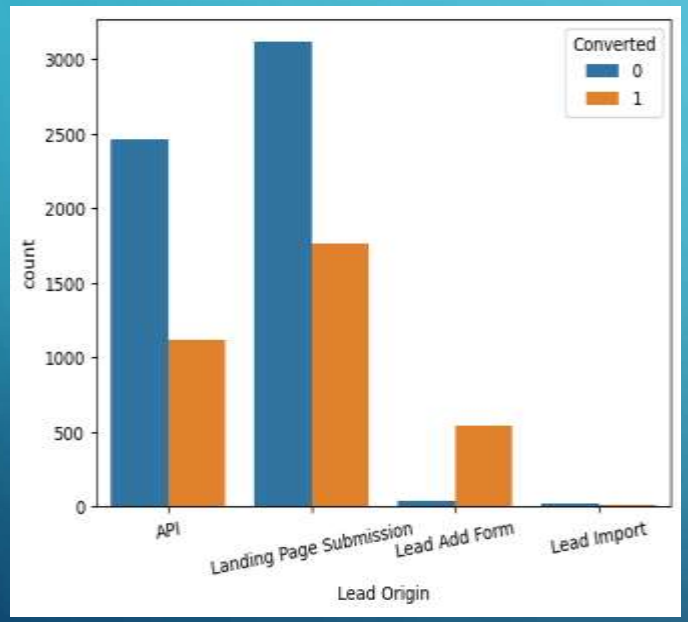
- An education company named X Education sells online courses to industry professionals. Now, although X Education gets a lot of leads, its lead conversion rate is very poor.
- You need to nurture the potential leads well (i.e. educating the leads about the product, constantly communicating etc.) in order to get a higher lead conversion.

METHODOLOGY FOLLOWED:

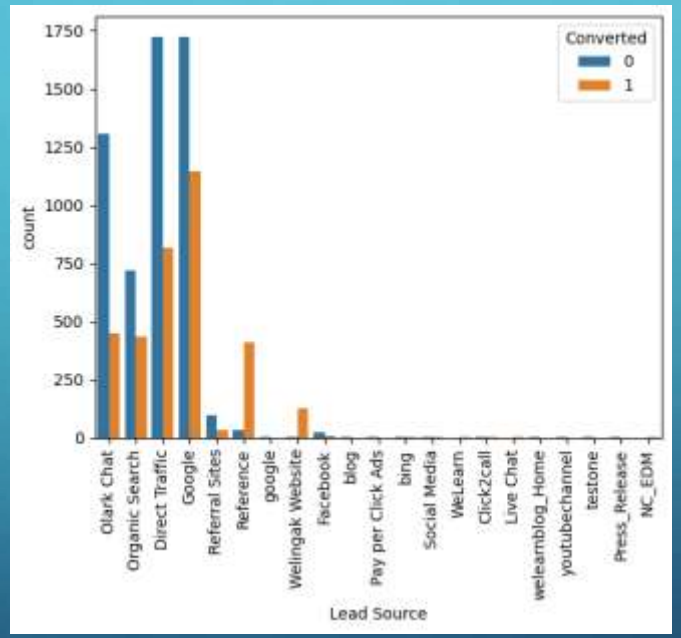
- Data Cleaning
  - Import the data
  - Check the data size, information, data types
  - Handle missing values by either
    - Dropping columns with huge missing data
    - Imputing the missing values based on current values
  - Handle Outliers
- Exploratory Data Analysis
  - Performed analysis on variables
    - Univariate
    - Bivariate
  - Group the less contributing/similar columns to 'Others'
  - Impact of variables with the conversion rate to see which variables maybe contributing to the conversion rate.

EDA

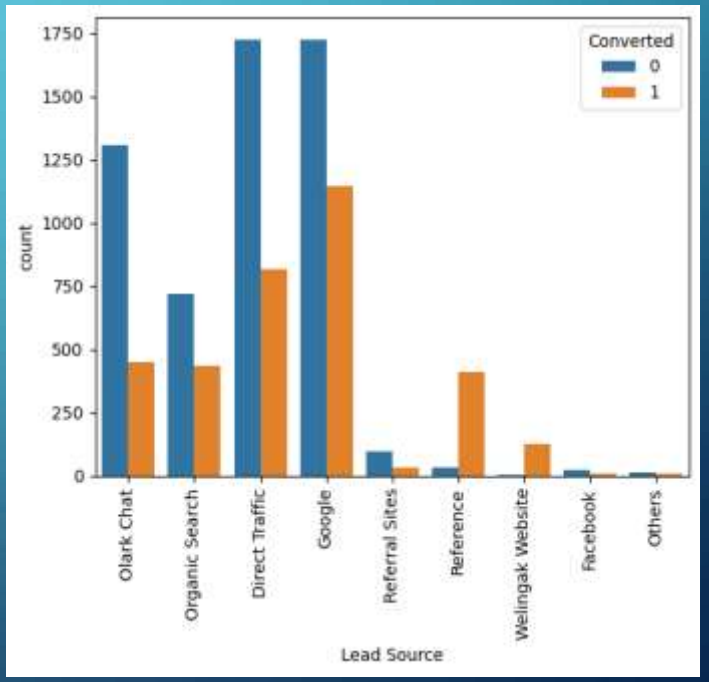
- API and Landing Page Submission have conversion rate but count lead 'not converted' is high



Conversion Rate of reference leads and leads through welingak website is high even though the overall count is less

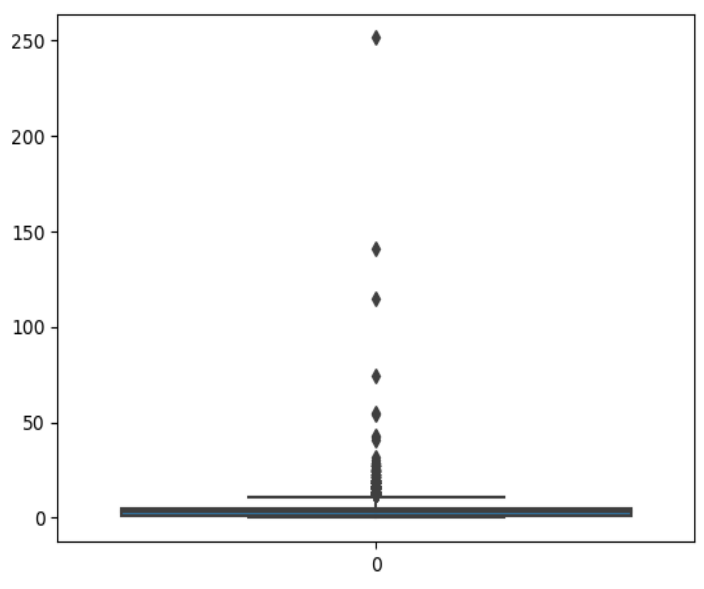


Google and Direct traffic generates maximum number of leads

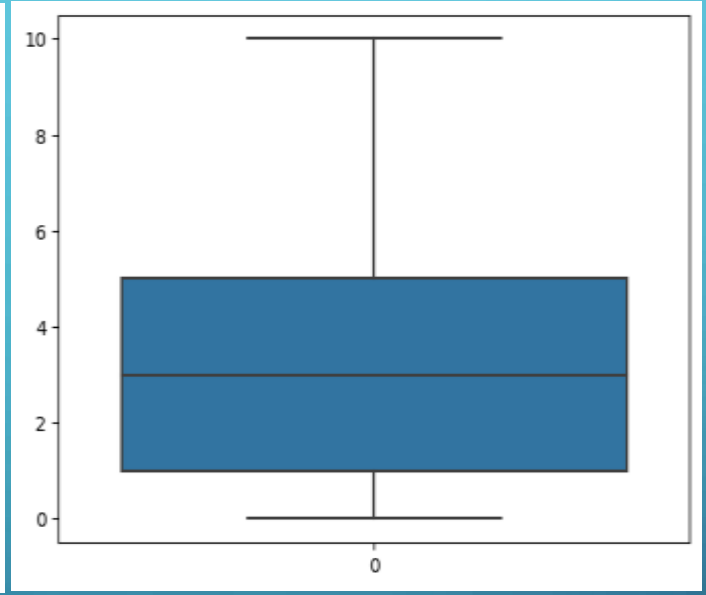


EDA

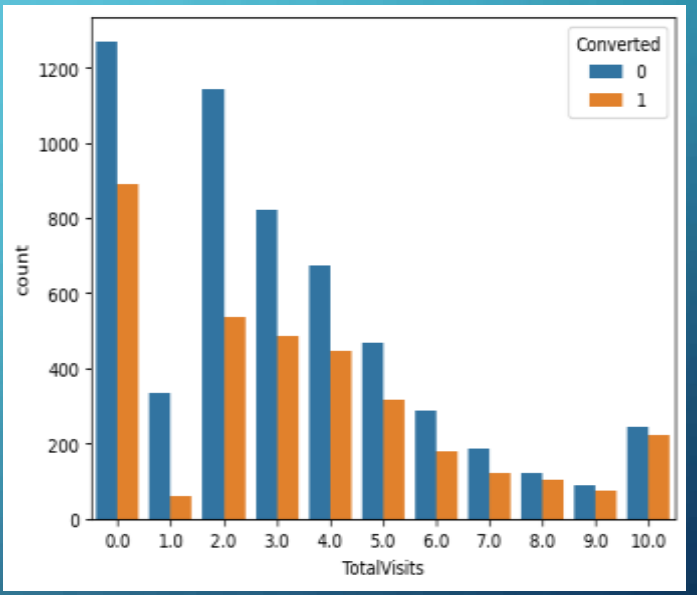
Boxplot to check outliers



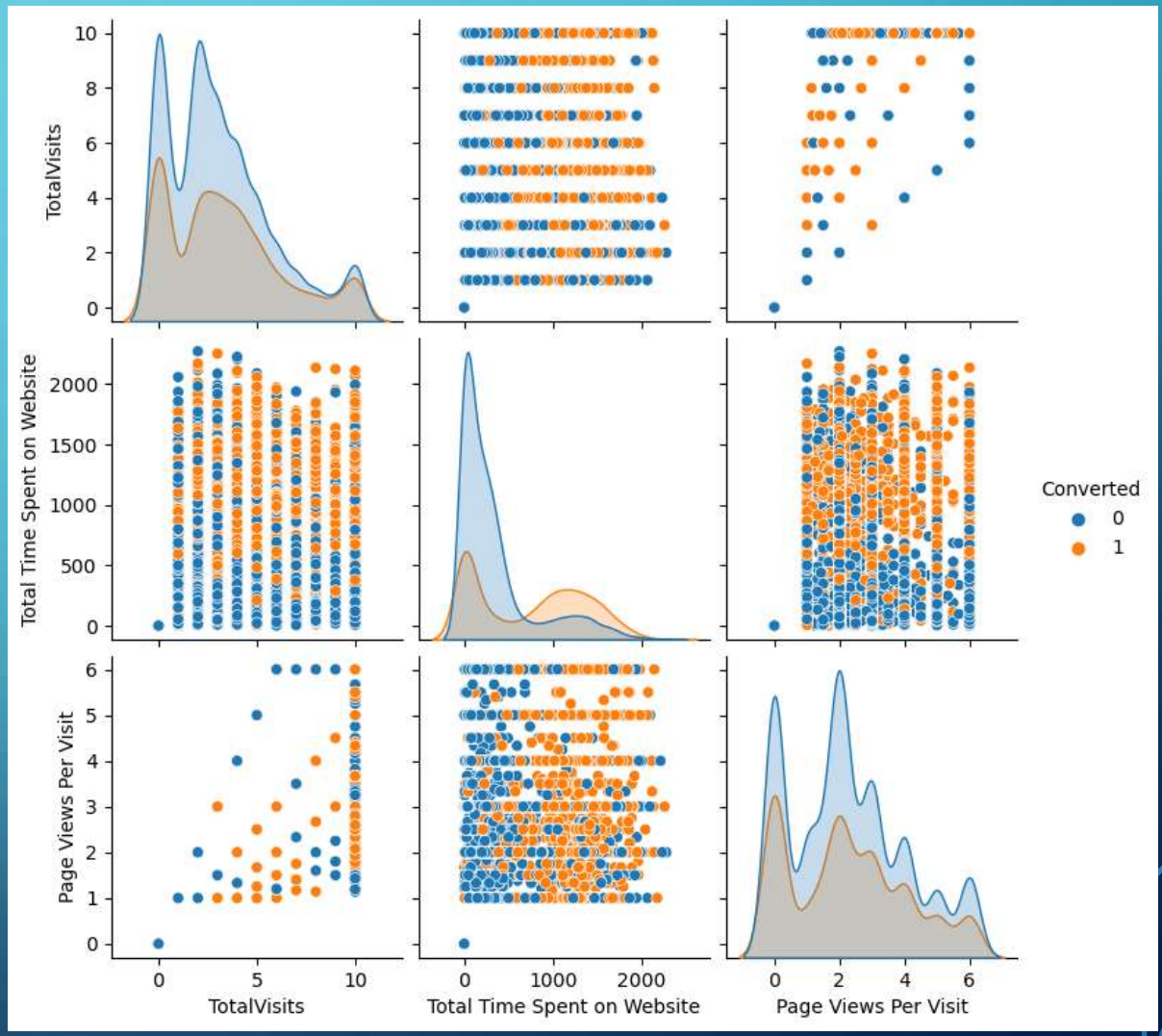
Outliers are present, so we can cap it



People who have visited the website more is more likely to get converted



EDA-FINDING THE CORRELATION THROUGH PYPLOT AND HEATMAP



LOGISTIC REGRESSION

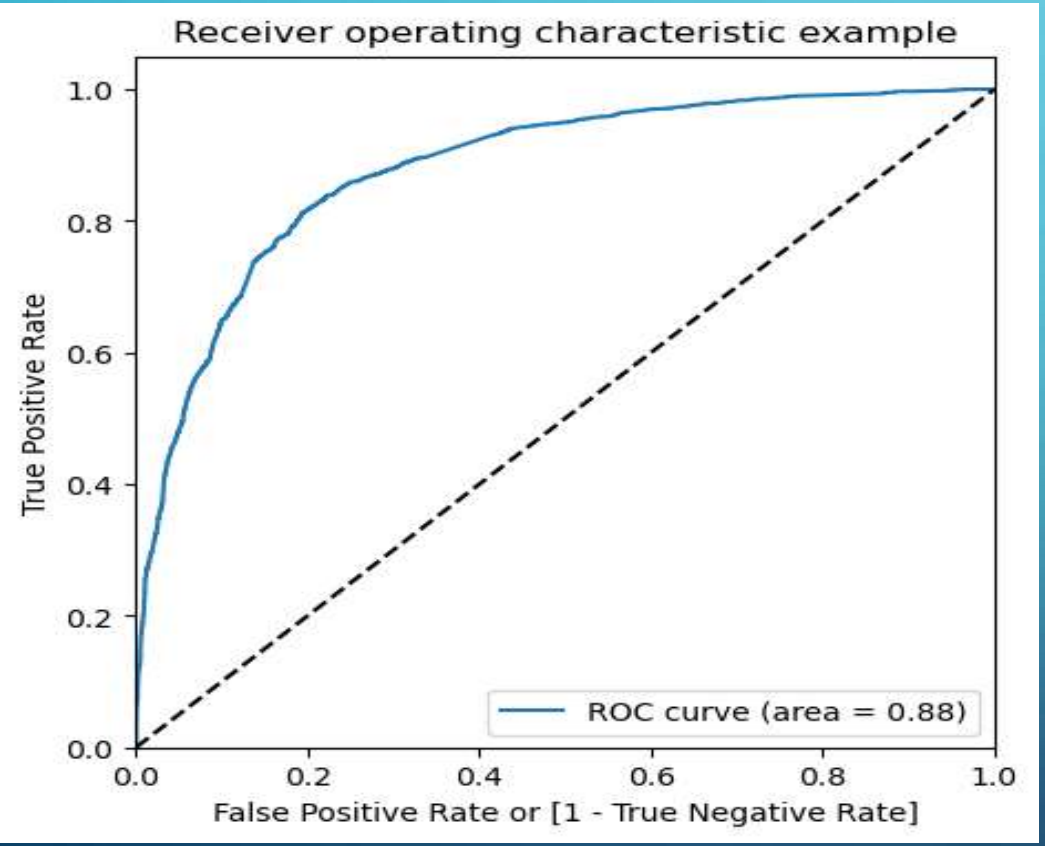
- Created dummy variables for the categorical variables
- Performed feature selection
- Put ‘Converted’ to the target variable
- Created training sets.
- Performed scaling
- Scaled 3 numeric columns : 'TotalVisits','Total Time Spent on Website','Page Views Per Visit’
- Performed RFE by feature selecting 15 variables
- Created test dataframe using the RFE selected variables

MODEL CREATION, EVALUATION AND PREDICTIONS

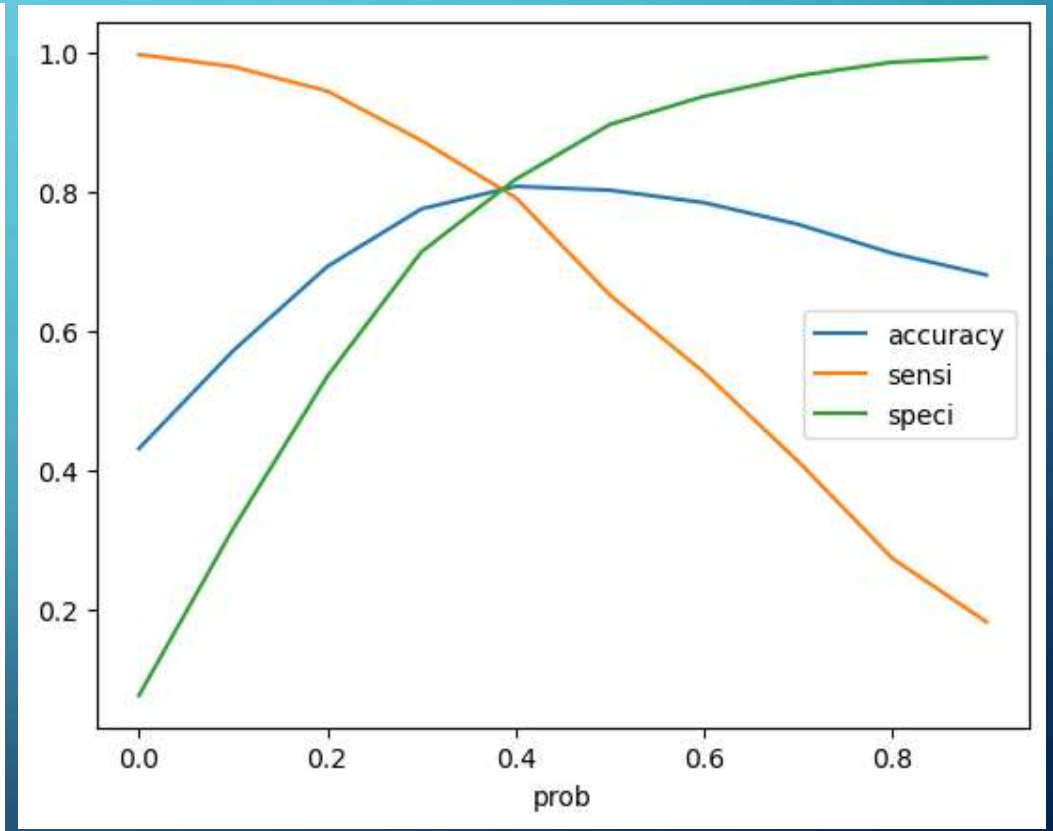
- Created a dataframe that will contain the names of all the feature variables and their respective VIFs
- Created model
- Removed columns based on high VIF value and high p value
- After series of similar steps, we got our final model.
- Predicted the train set probabilities
- Created confusion matrix
- Checked the accuracy, sensitivity and specitivity
- Found the optimal cut off by plotting the ROC Curve
- Made predictions using the cut off value
- Created confusion matrix
- Found the accuracy, sensitivity and specivity

ROC CURVE

The area under the curve of the ROC is 0.88 which is quite good. So we seem to have a good model



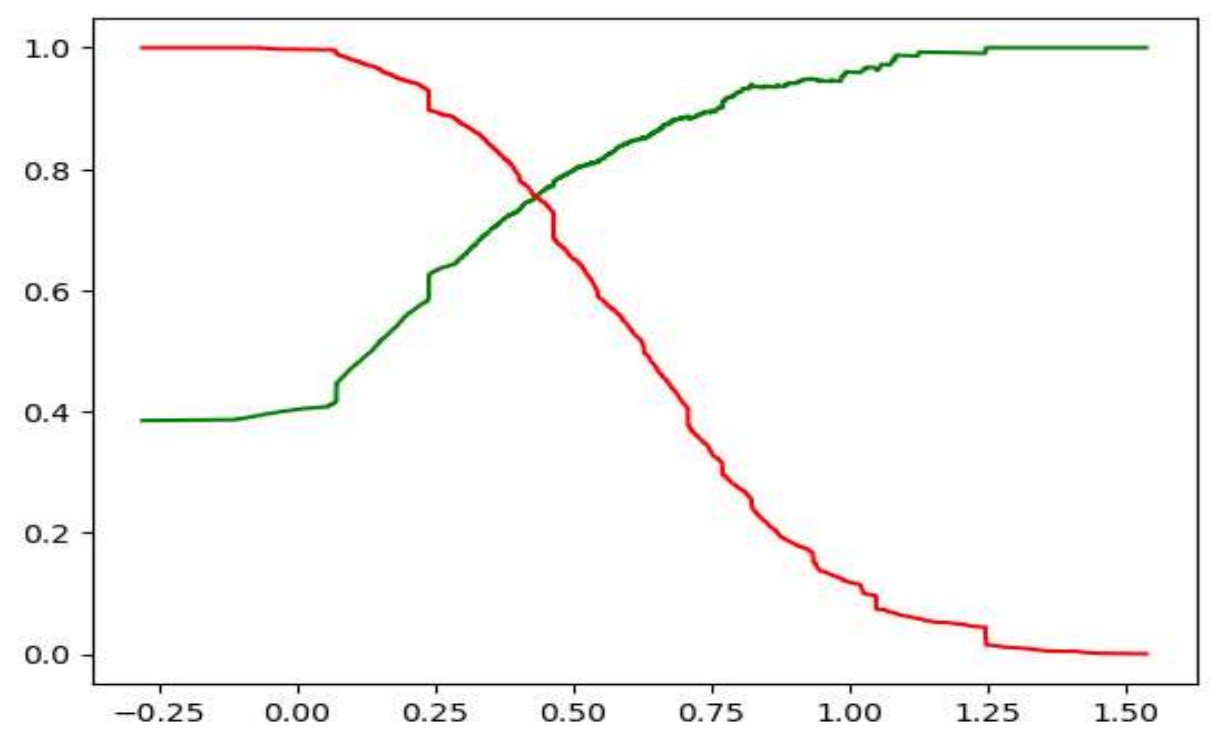
Accuracy, sensitivity and specificity tradeoff to find the optimal cutoff point which is around 0.385



PRECISION RECALL VIEW

- Built the training model using the precision-recall view
- Created confusion matrix again
- Calculated the accuracy, sensitivity, specificity
- Performed precision and recall tradeoff
- Found thresholds
- And again, calculated the accuracy, sensitivity, specificity
- Made predictions on the train set

PRECISION RECALL CURVE



CONCLUSION

- The following variables matter the most to get the leads are:

  Total Time Spent on Website
  Lead Origin - Lead Add Form
  Last Activity - Had a Phone Conversation
  Last Notable Activity - Unreachable
  What is your current occupation - Working Professional
  Lead Source - Welingak Website
  Last Activity - SMS Sent
  Do Not Email - Yes
  Lead Source - Olark Chat
  Last Activity - Olark Chat Conversation
  Specialization - Others
  Lead Origin - Landing Page Submission

- **Accuracy of the model is around 80% with Precision - 72% and Recall - 74% with a cutoff of 43%**