LEAD SCORE CASE STUDY

GROUP MEMBERS:- LOVE GUPTA, SHRAVANTHI KONDA, MAMTA KAKADI

PROBLEM STATEMENT

1. Build a logistic regression model to assign a lead score between 0 and 100 to each of the leads which can be used by the company to target potential leads.
2. A higher score would mean that the lead is hot, i.e. is most likely to convert whereas a lower score would mean that the lead is cold and will mostly not get converted.

METHODOLOGY FOLLOWED

- Data Understanding
- Data cleaning such as handling missing values & outliers
- EDA
- Feature Scaling & Dummy Variables and encoding of the data.
- Test Train split
- Building Model using RFE and manual feature eliminations by looking at VIF & Pz
- Validation of the model.
- Model presentation.

DATA UNDERSTANDING

- Leads.csv' - contains all the information of the leads from the past with around 9200 data points.

- Leads_Data_dictionary.csv' - You can learn more details about the dataset from this data dictionary

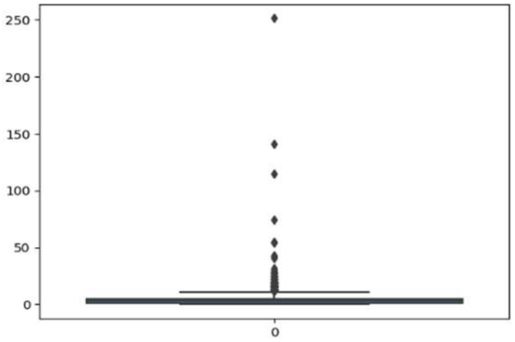
DATA CLEANING :- HANDLING MISSING VALUES

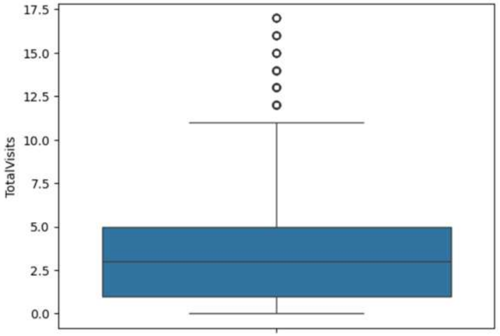
In this step, we will clean the data to do any form of analysis.

- We dropped columns having Null values/ “SELECT” value more than 25%
- Some of the column's values have Unique values as "NO“ which we dropped like Magazine, Receive More Updates About Our Courses, Update me on Supply Chain Content, Get updates on DM Content, I agree to pay the amount through cheque.
- We removed 2 more columns which are redundant: Prospect ID, Lead Number
- We replaced the missing values with mean, where missing values are between 1-2%
- We dropped records values where missing values are less than 1% in the variables

DATA CLEANING:- HANDLING OUTLIERS

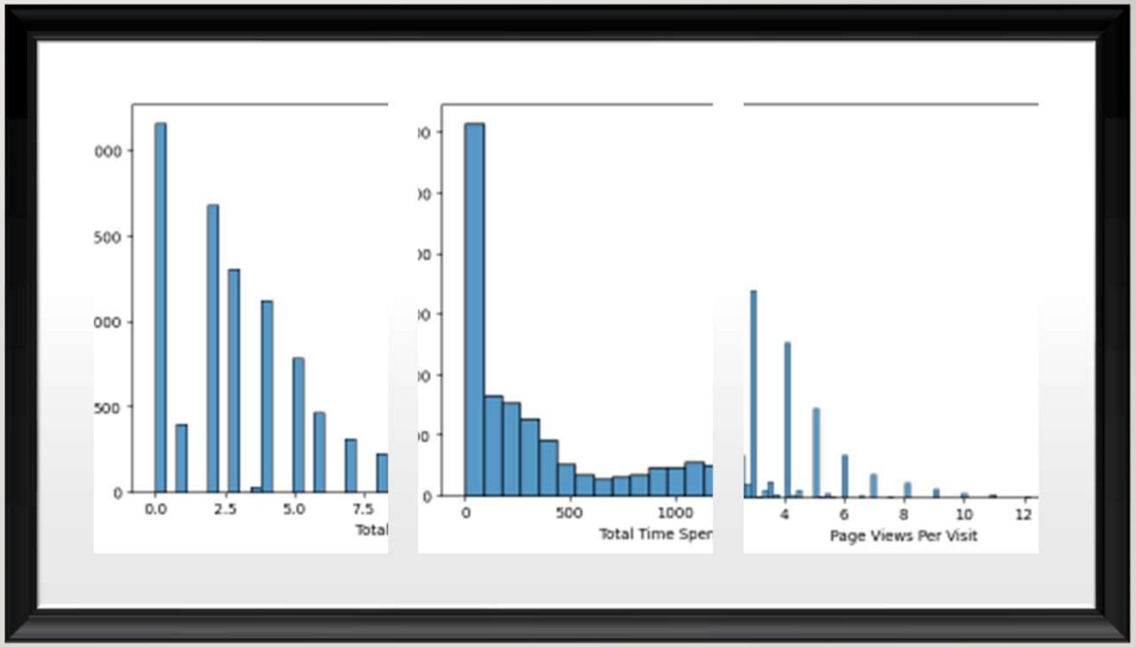
We have 4 numerical features in which we need to handle outlier for that we check the difference between q(0.95) & q(0.99) and Max & q(0.99), if difference between max and q(0.99) is far far greater than difference between q(0.99) & q(0.95) than we can drop records above q(0.99) else not required.





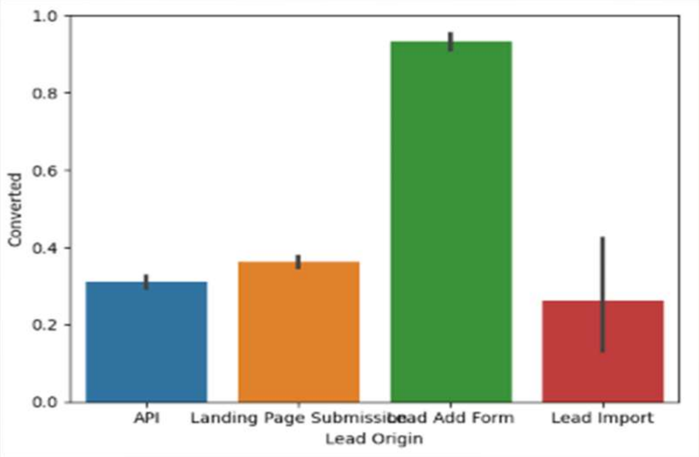
EDA:- UNIVARIATE ANALYSIS

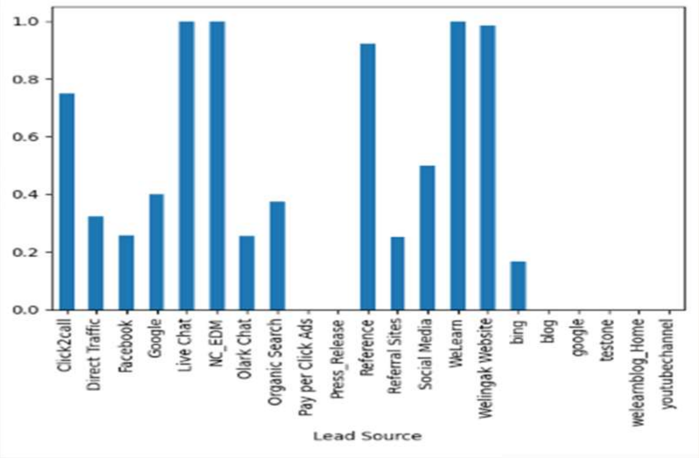
- Overall Conversion Ration is 37%
- As below all three charts showing same trends which is leads are on the left side of graph in terms of TotalVisits, Page views per visit & Total time spend on website.



EDA:- BIVARIATE ANALYSIS

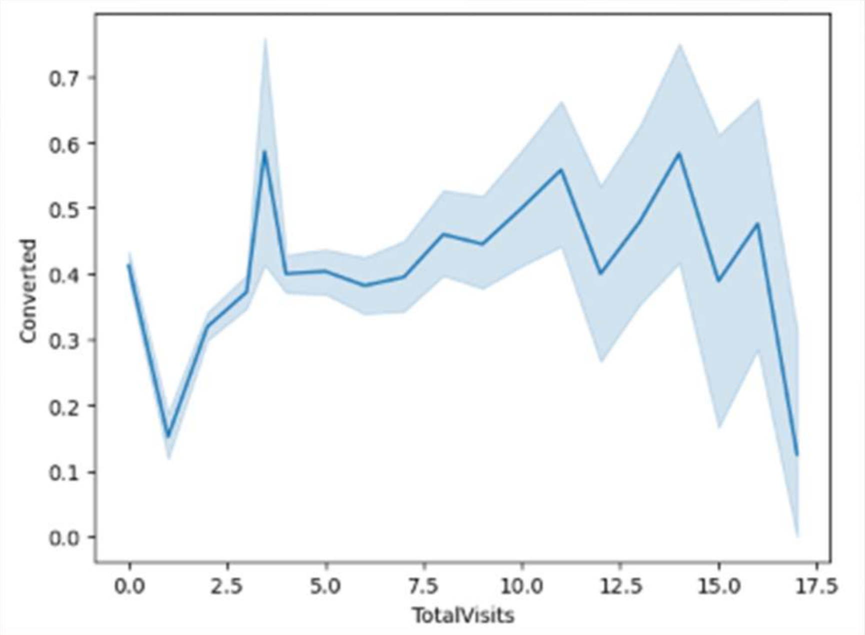
- Lead add form are converting the most though they are 6% in numbers in the whole data set but they are producing more than 93% conversions. so we should immediately target them
- All the leads from Live_chat, NC_EDM & we learn have 100% conversion but main chunk of leads from Google, Direct Traffic & Olark Chat which has very average conversion ratio which is around 30%





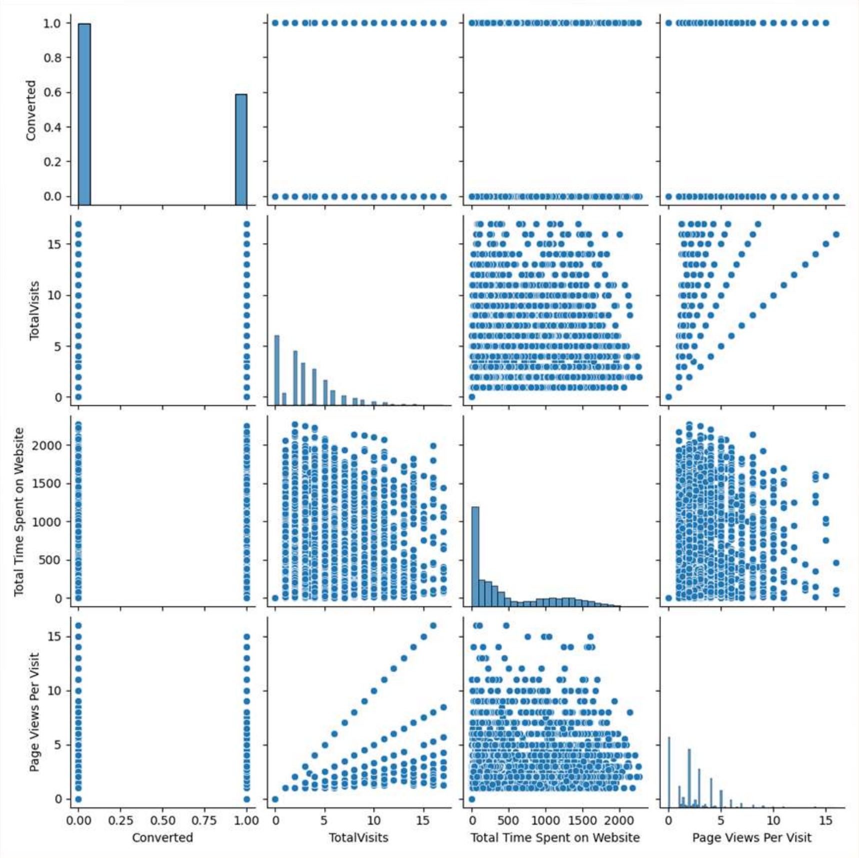
EDA:- BIVARIATE ANALYSIS

- we can see from Above chart conversion ration is more than 50% on total visits count between 10-15 & between 2.5-5 nos All the leads from Live_chat, NC_EDM & we learn have 100%



EDA:- MULTIVARIATE ANALYSIS

WE PLOTTED PAIRPLOT BETWEEN ALL THE VARIABLES ALL OBSERVED TRENDS.



FEATURE SCALING, DUMMY VARIABLE & TEST TRAIN SPLIT

- Converted Yes/no into 0 & 1
- For categorical variables with multiple levels with using get_dummies in pandas and dropped first column
- Then we dropped repeated columns after creative dummy variable
- We did feature scaling using standard scaler for the numerical variables.
- We split the data into train & test data for the ratio of 0.7 to 0.3 using sklearn library train_test_split.

MODEL BUILDING USING RFE

- We imported LogisticRegression library using sklearn
- We selected top 15 features out of 66 and build the Model using GLM with which we got 80% accuracy
- Later we checked VIF and came out to be 83 , 64 & 19 those were too high so started dropping features one by one manually from existing model.
- Like that we created final model in which VIF is less than 2 & pz is less than 1%

VALIDATION OF MODEL

We got following matrix on training data set:

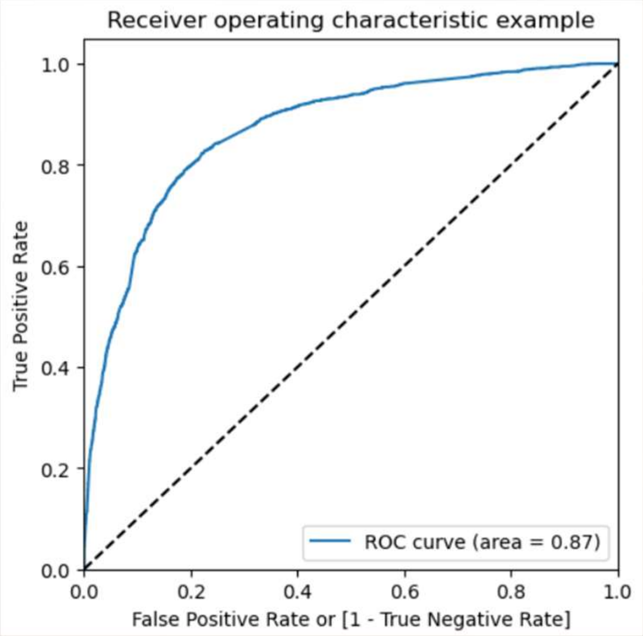
- Accuracy = 80.4%
- Sensitivity =67%
- Specificity = 88.31%
- False positive rate = 11.68%

ROC CURVE

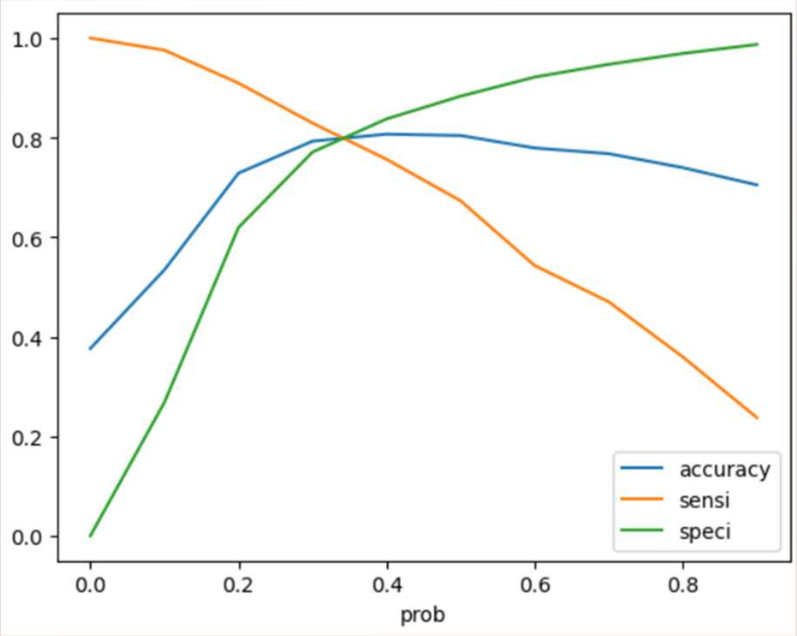
An ROC curve demonstrates several things:

- It shows the trade-off between sensitivity and specificity (any increase in sensitivity will be accompanied by a decrease in specificity).
- The closer the curve follows the left-hand border and then the top border of the ROC space, the more accurate the test.
- The closer the curve comes to the 45-degree diagonal of the ROC space, the less accurate the test.

For the perfect model ROC curve area should be 1 but we got 0.87 which is very good.



OPTIMAL CUTOFF POINT



As of now we took cut-off as 0.5 but we need to check which is the correct cut-off to use for the model prediction.

Which turned out to be 0.35 .

VALIDATION OF MODEL AFTER OPTIMAL CUTOFF POINT

We got following matrix on training data set:

- Accuracy = 80%
- Sensitivity =79.67%
- Specificity = 80.21%
- False positive rate = 19.76%

PREDICTION RESULT OF FINAL MODEL USING TEST DATASET

We got following matrix on test data set:

- Accuracy = 78.23%
- Sensitivity =77.3%
- Specificity = 78.81%
- False positive rate = 21.18%

As we can see our model is working fine with unknown test data set and predicting 78% of the correct result which is similar to the training results.

RECOMMENDATION

Important features responsible for good conversion rate or the ones' which contributes more towards the probability of a lead getting converted are:

- When the Lead Source is Reference.
- When the Lead Source is Welingak Website
- When Last Notable Activity is, Had a Phone Conversation.

THANK YOU