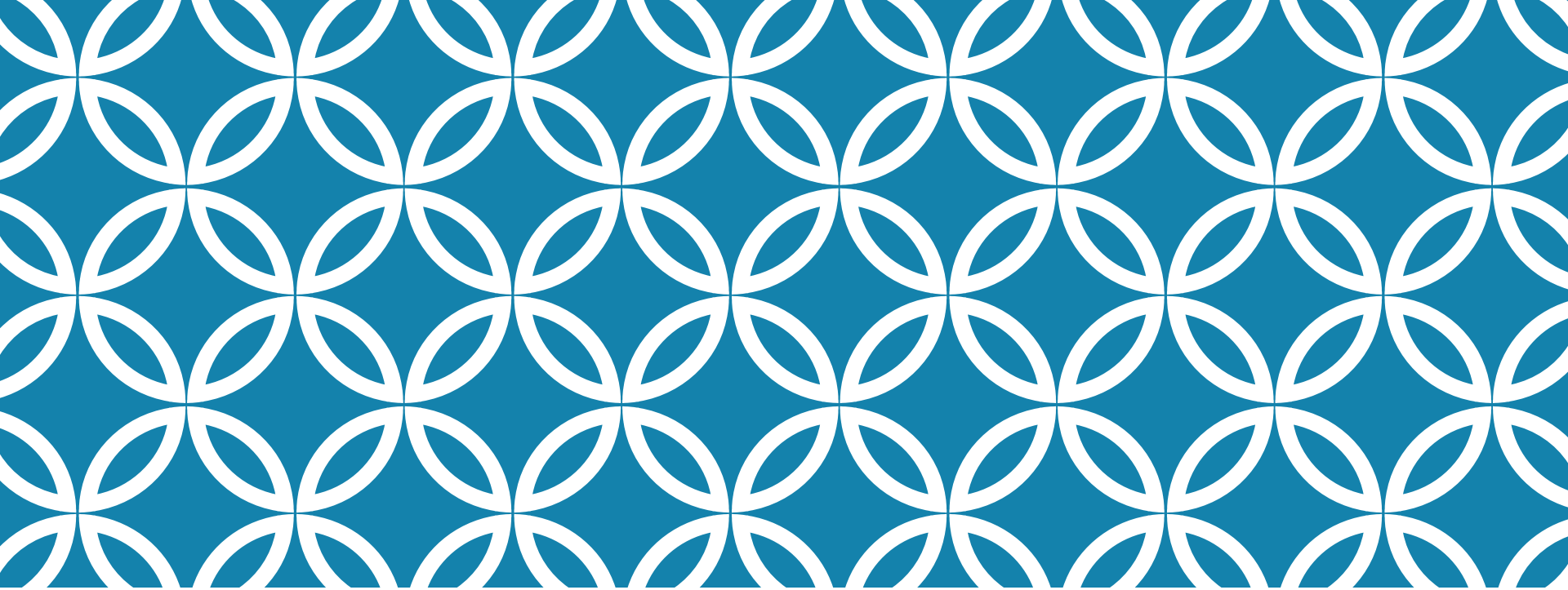


LEAD SCORING ASSIGNMENT

Stage 1 – Removing Missing values & Data Cleaning

- Checked for NaN values in the DataFrame using df.isna().sum() and null_percentages = (df.isnull().sum() / df.shape[0]) * 100
- In the first round, we removed all columns which had more than 40% null values ie 3600

```
[ ] #Drop all the columns in which greater than 3600 missing values,
    #which is 40% and above of the dataset
    for col in df.columns:
      if df[col].isnull().sum() > 3600:
        df.drop(col, 1, inplace=True)
```

- We removed redundant columns, such as country, city, id, prospect number as well as those with 70% missing values

Stage 1 – Removing Missing values & Data Cleaning

- **Columns with imbalanced data: I created a separate category for object columns called ‘temp’ to check the values.** There are few columns where only value has a majority for all data points. following columns were removed: Do Not Call', 'Search', 'Magazine', 'Newspaper Article', 'X Education Forums', 'Newspaper', 'Digital Advertisement', 'Through Recommendations', '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’

- Removed values with 2% missing values via ‘removing by row method’

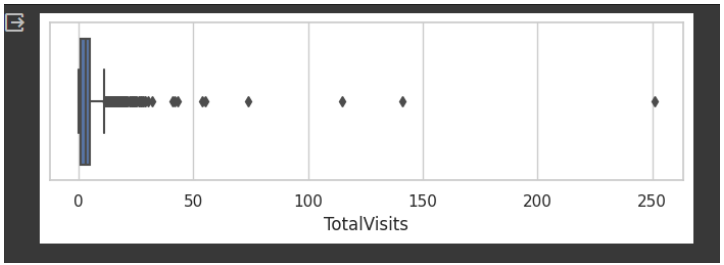
```
[231] #reducing the rows that have missing values if less than 2%
      df = df[~pd.isnull(df["TotalVisits"])]
      df = df[~pd.isnull(df["Lead Source"])]
```

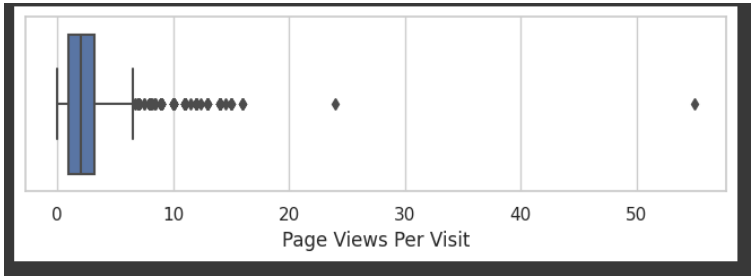
- Values that were in the middle or type 2, with say 30% missing values, a separate ‘Missing’ category’ was created to prevent data loss of 30%

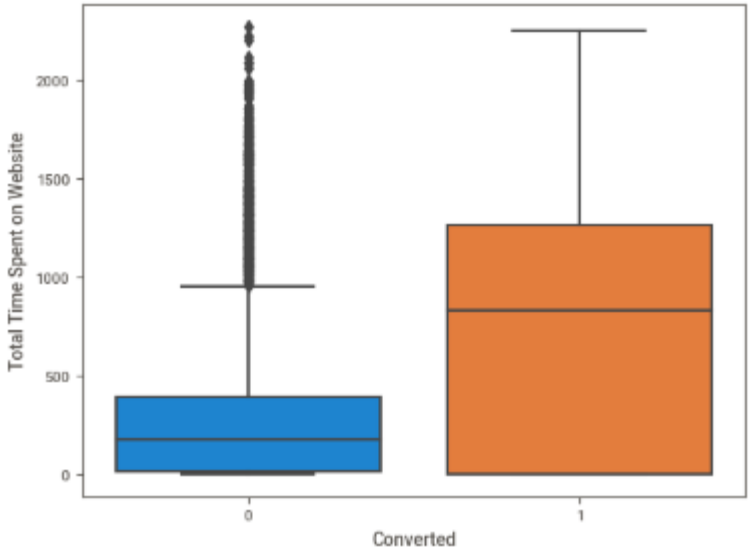
```
[235] df['Specialization'] = df['Specialization'].replace(np.nan, "Missing")
```

Stage 1 – Removing Missing values & Data Cleaning

- After removing all nulls from the entire data set, check numeric columns for outliers, which we will treat with min max scalar before modelling.
- Three columns had outliers:







Data Visualization

- Spending time on freebies, such as infographics or ePDFs should be productive. People are currently not getting converted with infographics or ebooks . Either one has to change the topic to a more relevant one or find another way to convert

- Most converted leads come from people who have opted out of email than the ones who have opted for it. The majority of the people opt out of email, hence it's not advisable to focus much on emails.
- Landing Page Submission has the highest conversion rate and the lead ad form is the lowest conversion.

- There are more conversions for Total time spent on the website than non-converted leads. Around 1300 conversions and 400 non-converted leads

- As we can see, the highest converted leads and total visits for Google, Direct Traffic and Reference. Reference has the most converted leads and negligible non converted leads out of all the categories. We can see the highest non-converted leads and total visits for Direct Traffic.
-
- One should focus on SMS as most converted leads are active from SMS and the most non-converted ones from email.

Data Visualization

- we can see more converted leads as the total time spent on the website increases (above 600 minutes for converted leads). They typically have 1 to 6 page views per visit on the website
- Most converted leads come from people who have opted out of email than the ones who have opted for it. The majority of the people opt out of email, hence it's not advisable to focus much on emails.
- Landing Page Submission has the highest conversion rate and the lead ad form is the lowest conversion. Interns can focus on making landing page forms easy and API-based conversions and work on improving lead ad form conversions
- There are more conversions for Total time spent on the website than non-converted leads. Around 1300 conversions and 400 non converted leads
- we can see more converted leads as the total time spent on the website increases (above 600 minutes for converted leads). They typically have 1 to 6 page views per visit on the website

Data Visualization

- Checked the target column for imbalanced data. Ir eems to be fairly blanced
- The conversion rate is around 38%

```
[70] ### Checking the conversion_rate
     conversion_rate = (sum(df['Converted'])/len(df['Converted'].index))*100
     conversion_rate

     37.85541106458012
```

Arranged the high correlations in ascending order to find relations. Following were the relations:

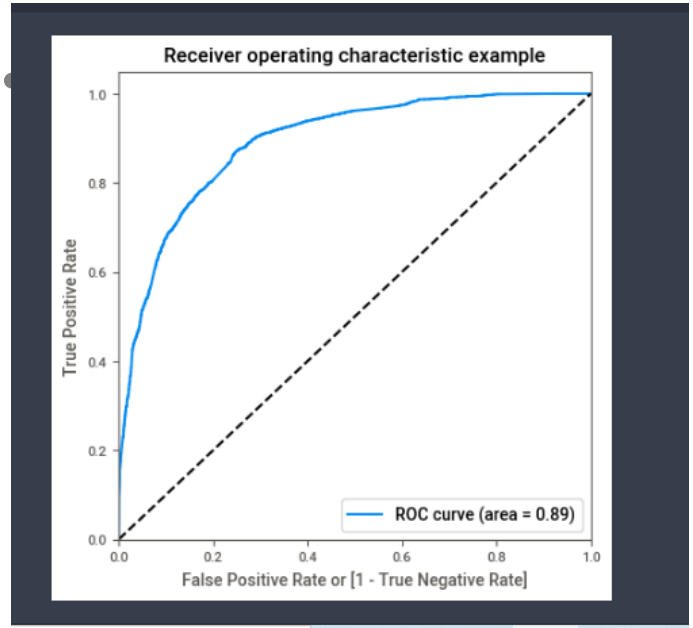
- There is a high correlation between the Lead origin as lead import and the lead source as Facebook.
- There is a high correlation between the Last Activity of the users which was **Unsubscribed** and the **lead source** which was **Facebook**. It means most people who are unsubscribing are from Facebook.
- High correlation between the lead origin in **Ad form category and lead source**, reference category. It simply means the converted leads from **ad form** are **references.**

Data Visualization

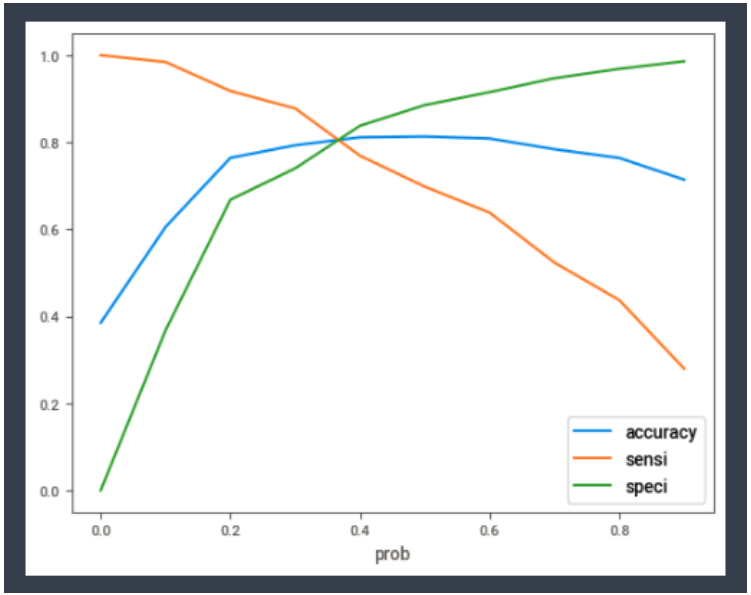
- High correlation between **Last Activity** column in the **Email Opened** category and **Last Notable Activity** column in the **Email Opened** category. It simply means the customers whose last activity was checking email were most likely to check emails or be active on email as compared to other platforms.
- High correlation between the **Last Activity** column in the **SMS Sent** category and **Last Notable Activity** column in the **SMS Sent category**. It simply means the customers whose last activity was SMS Sent were most likely to check SMS or be active on SMS as compared to other platforms.
- High correlation between **Last Activity** column in the **Email opened** category and Last Activity in the **Email Link Clicked** category. It simply means the customers whose last activity was **Email Link Clicked** were most likely to check emails or be active on emails as compared to other platforms.
- High correlation between the **Last Activity** column in the **Had a Phone Conversation** category and the **Last Notable Activity** column in the **Had a Phone Conversation** category. It simply means the customers whose last activity was **having a phone conversation** were most likely to interact on calls as compared to other platforms.
- High correlation between the **Last Activity** column in the **Email Received** and the **Last Notable Activity** column in the **Email Received** category. It simply means the customers whose last activity was **receiving an email** were most likely to interact on emails as compared to other platforms.

Modelling – Train set

ROC is curve for train data



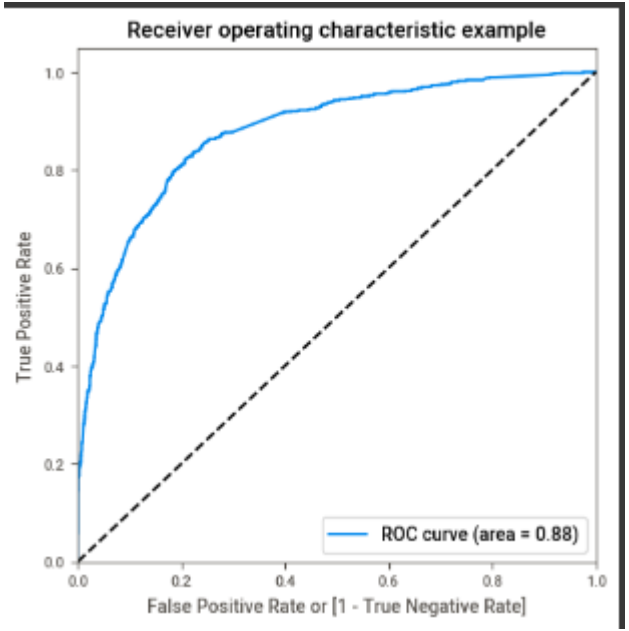
Cut off for train data is 3.8



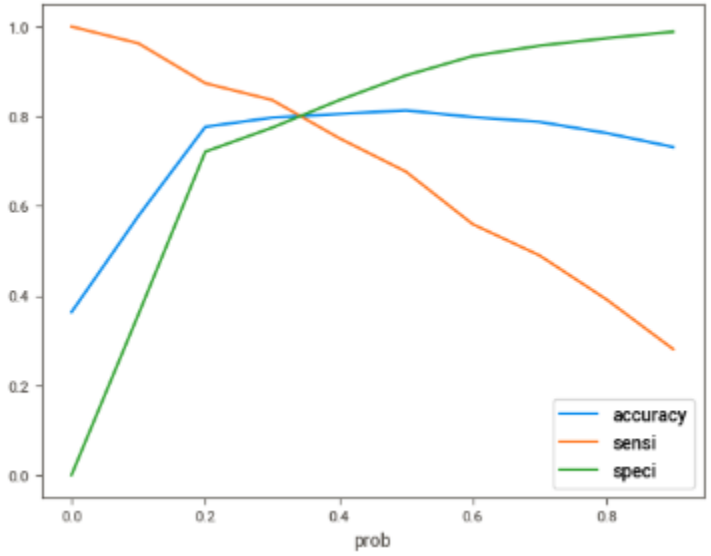
Accuracy for train is 81.04%

Modelling- Test set

ROC is curve for test data 88%



Cut off for train data is 3.8



Accuracy for test is 80.08%

Modelling- Comparing the confusion Matrix & other scores

Train Data

Sensitivity	**78%**
Specificity	83%
False Positive	16%
Positive predictive value	74%
Negative Predictive value	86%
Accuracy	81.04%

Test Data

Sensitivity	80%
Specificity	82%
False Positive	18%
Positive predictive value	71%
Negative Predictive value	87%
Accuracy	80.08%