Leads Scoring Assignment

By Akanksha Ojha

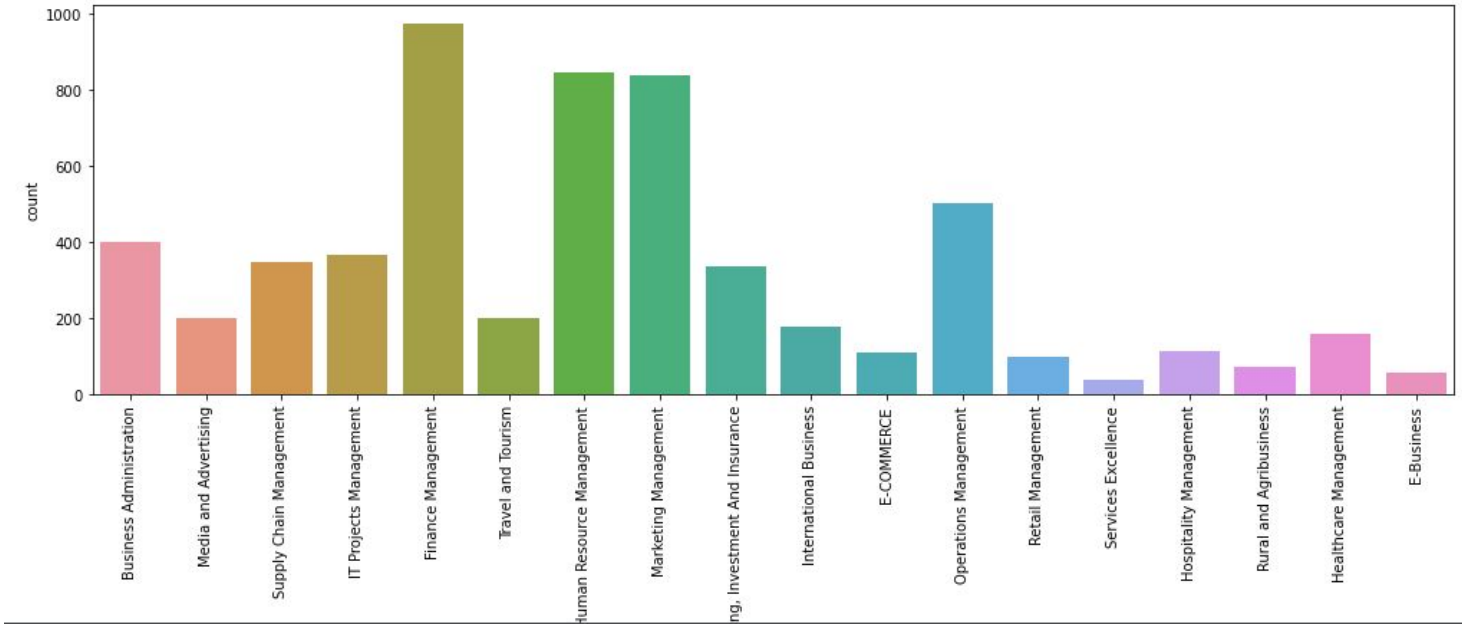
1) Column: 'Specialization'

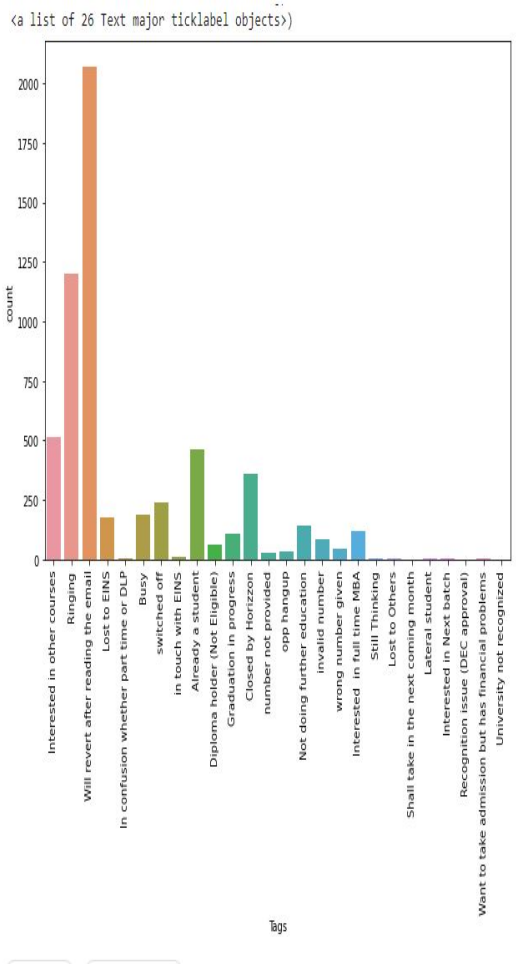
This column has 37% missing values

+ Code + Markdown

```
plt.figure(figsize=(17,5))
sns.countplot(lead_data['Specialization'])
plt.xticks(rotation=90)
```

```
(array([ 0,  1,  2,  3,  4,  5,  6,  7,  8,  9, 10, 11, 12, 13, 14, 15, 16,
        17]),
 <a list of 18 Text major ticklabel objects>)
```





Since most values are 'Will revert after reading the email' , we can impute missing values in this column with this value.

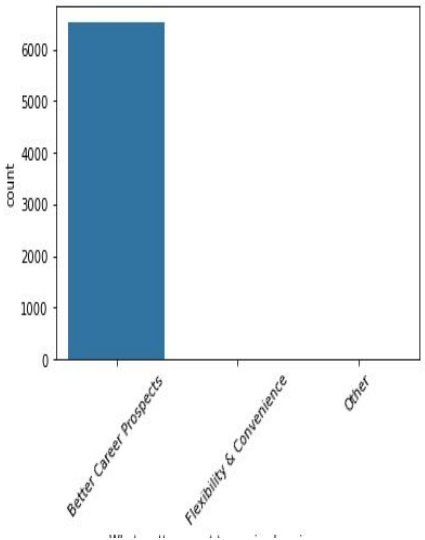
Missing data

3) Column: 'What matters most to you in choosing a course'

this column has 29% missing values

```
# Visualizing this column
sns.countplot(lead_data['What matters most to you in choosing a course'])
plt.xticks(rotation=45)
```

```
(array([0, 1, 2]), <a list of 3 Text major ticklabel objects>)
```



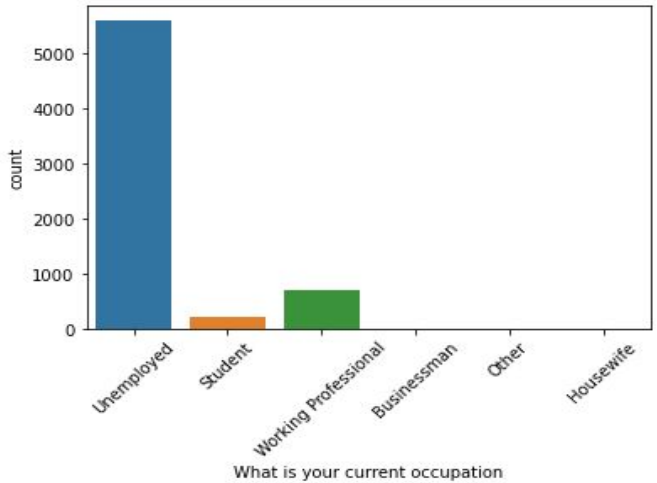
Current occupation

4) Column: 'What is your current occupation'

this column has 29% missing values

116]:
```
sns.countplot(lead_data['What is your current occupation'])
plt.xticks(rotation=45)
```

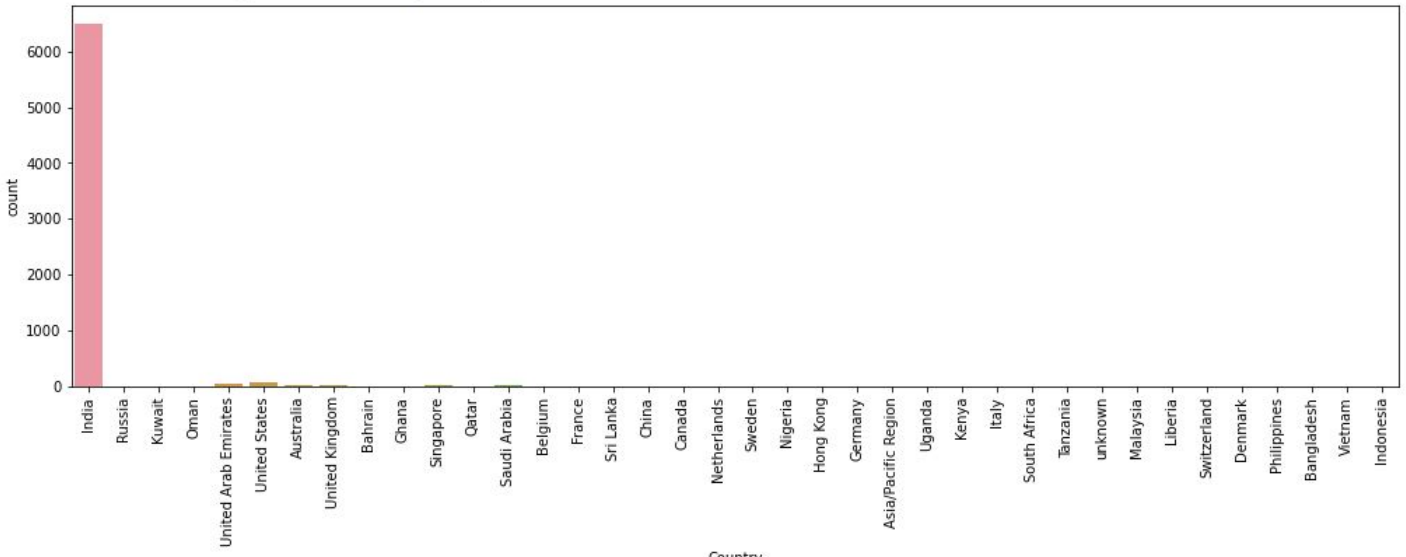
[116… (array([0, 1, 2, 3, 4, 5]), <a list of 6 Text major ticklabel objects>)



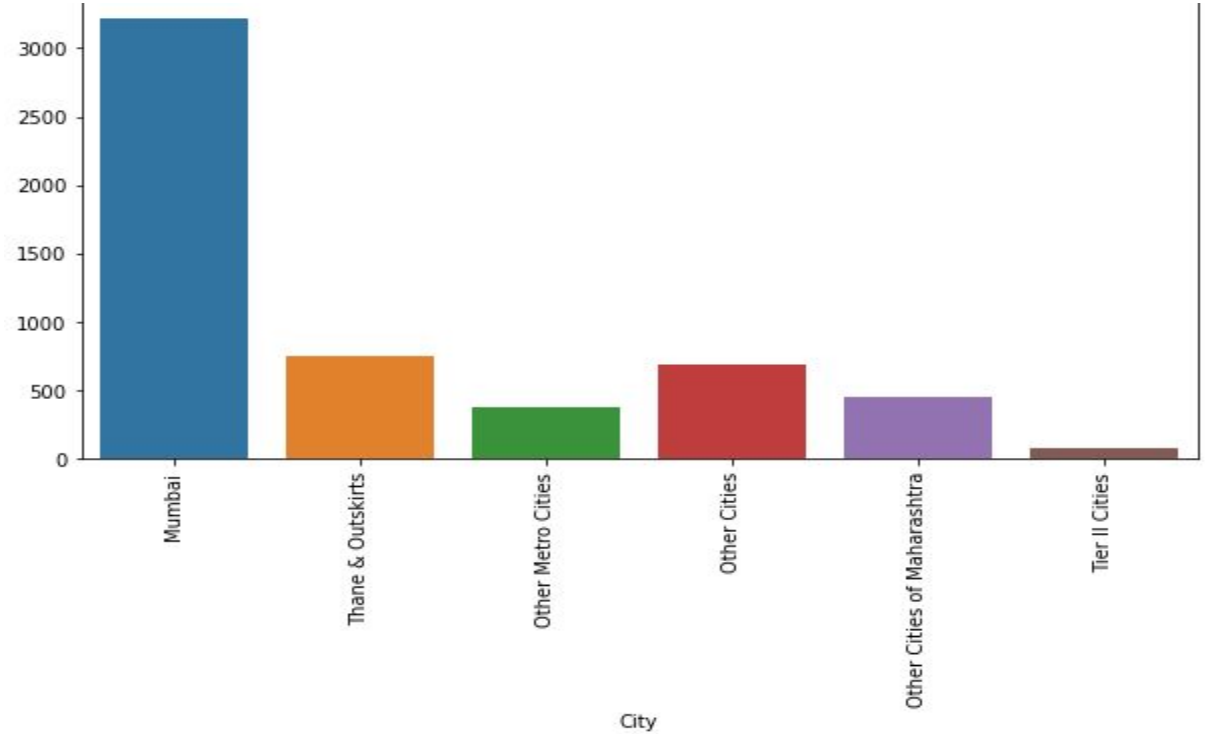
Although this column is clearly slanted, it contains crucial information on the lead. Since 'India' is the most common response, we may use it to impute missing values in this column.

[119]:
```
plt.figure(figsize=(17,5))
sns.countplot(lead_data['Country'])
plt.xticks(rotation=90)
```

[119… (array([0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16,
17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33,
34, 35, 36, 37]),
<a list of 38 Text major ticklabel objects>)

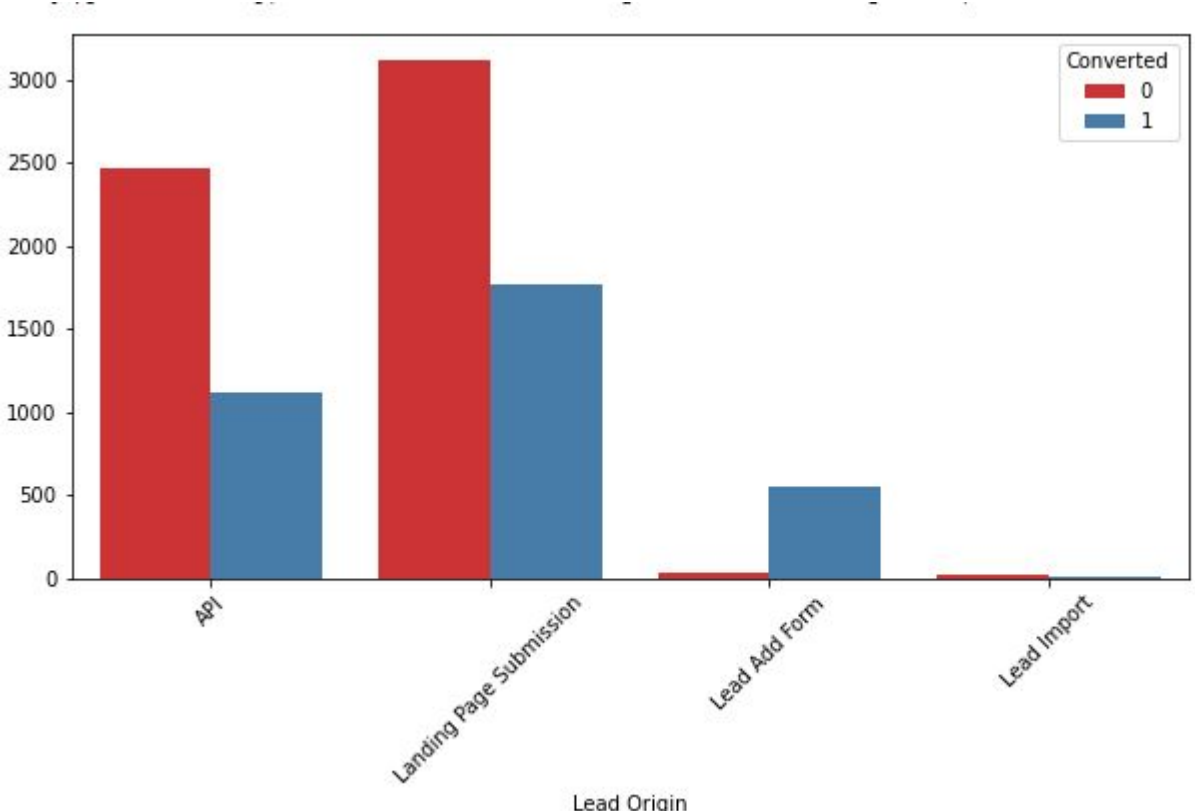


Check data for city



Lead Origin

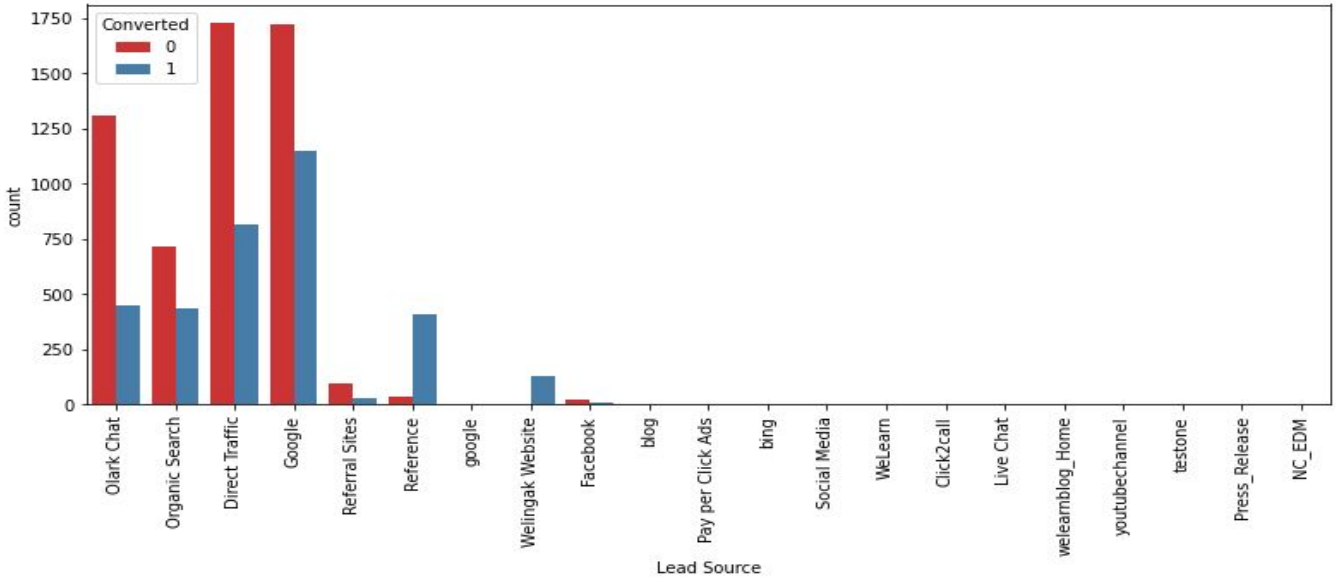
Although landing page submissions and APIs have a 30–35% conversion rate, they generate a sizable number of leads.



Lead Source

```
plt.figure(figsize=(13,5))
sns.countplot(x = "Lead Source", hue = "Converted", data = lead_data, palette='Set1')
plt.xticks(rotation = 90)
```

```
(array([ 0,  1,  2,  3,  4,  5,  6,  7,  8,  9, 10, 11, 12, 13, 14, 15, 16,
        17, 18, 19, 20]),
 <a list of 21 Text major ticklabel objects>)
```

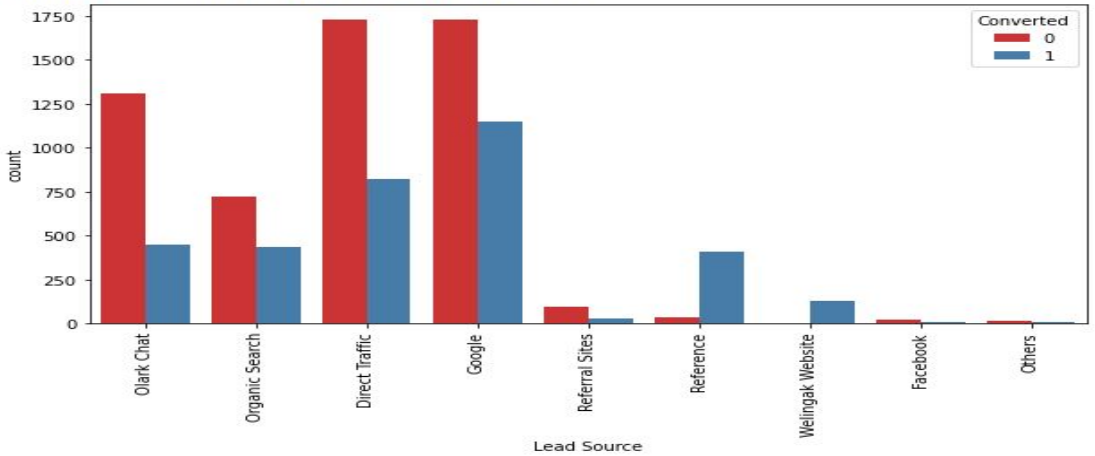


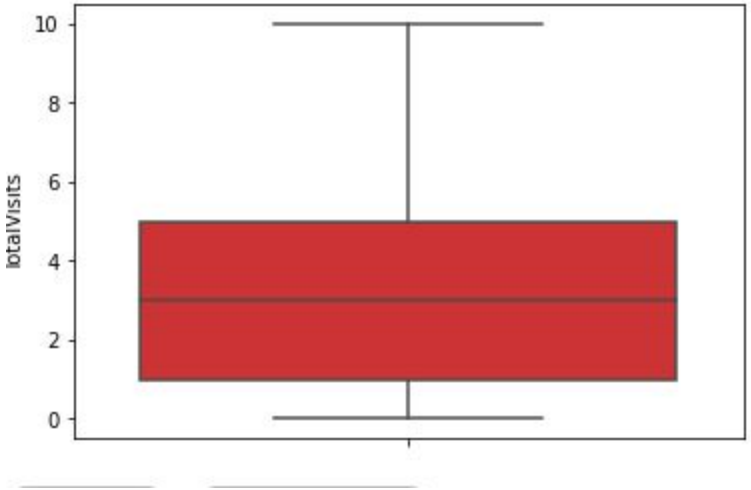
Lead Source

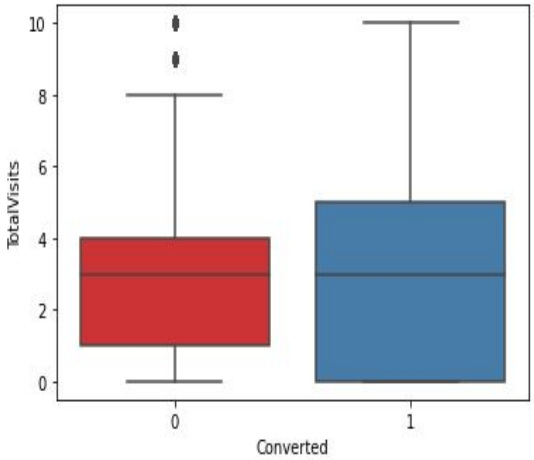
1. The most leads are produced by Google and direct traffic.
2. There is a high conversion rate for both reference and welingak website leads.

```
]:
# Visualizing again
plt.figure(figsize=(10,5))
sns.countplot(x = "Lead Source", hue = "Converted", data = lead_data,palette='Set1')
plt.xticks(rotation = 90)
```

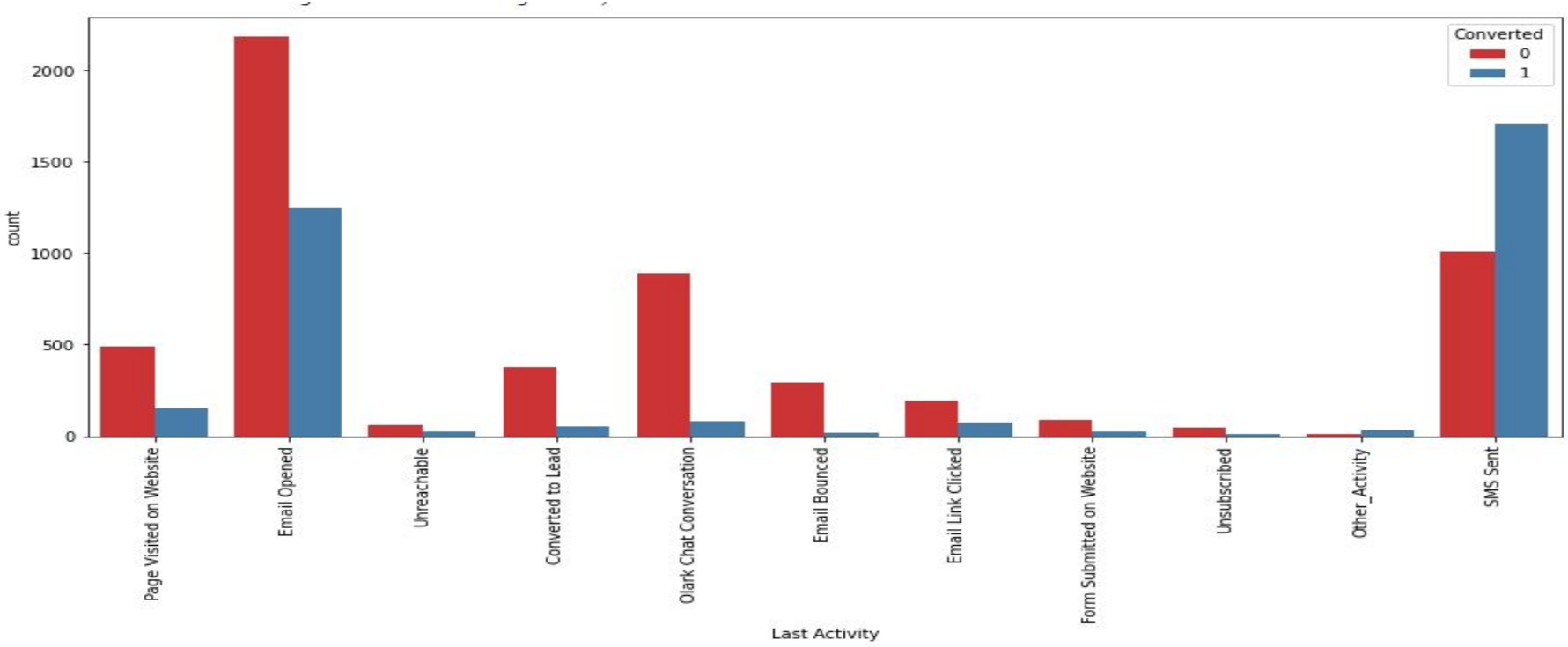
```
(array([0, 1, 2, 3, 4, 5, 6, 7, 8]),
 <a list of 9 Text major ticklabel objects>)
```



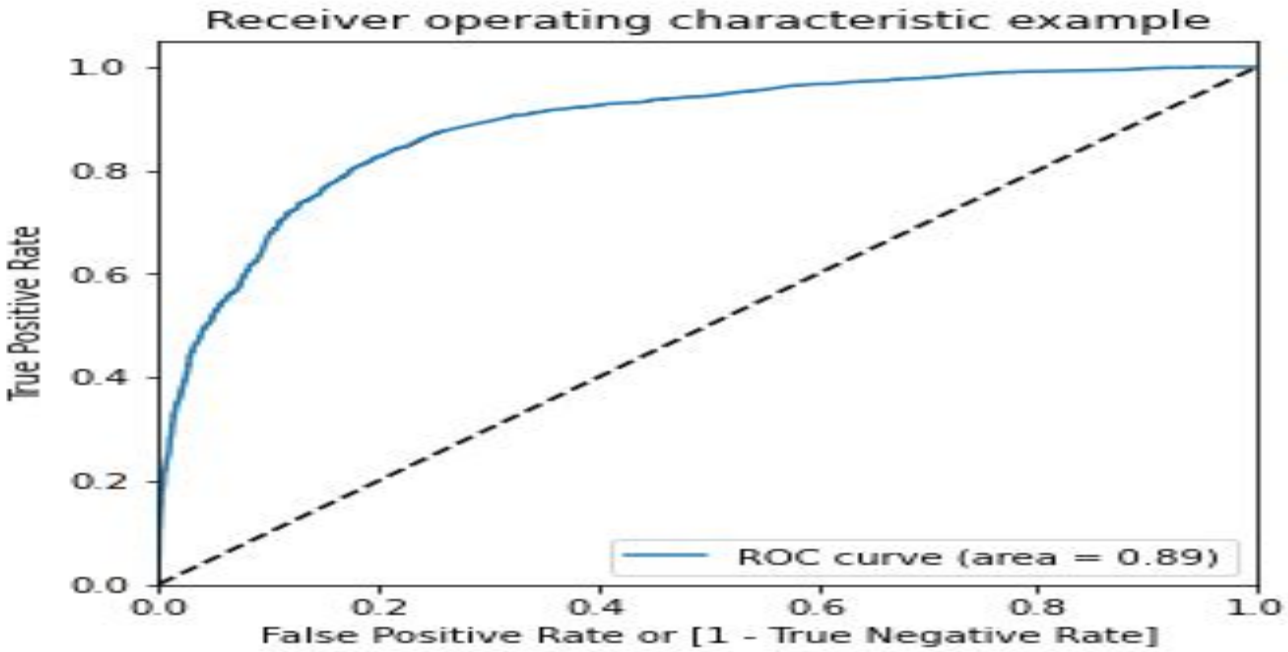




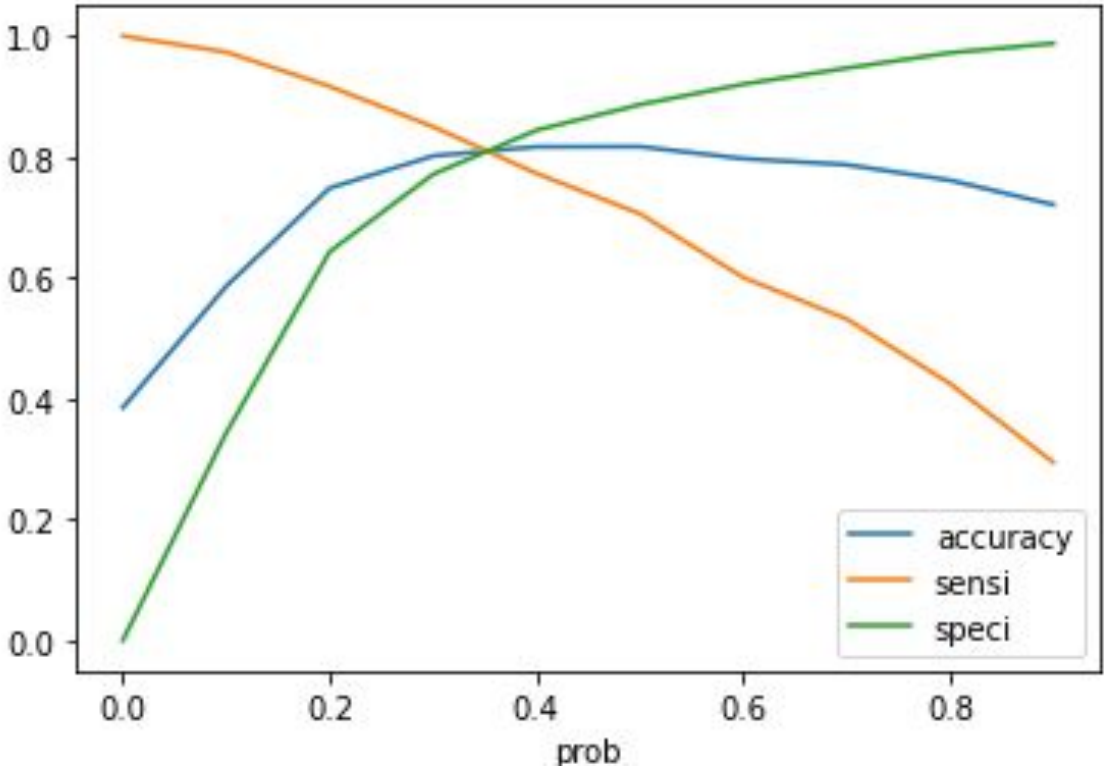
The median for leads that are converted and those that are not is the same.



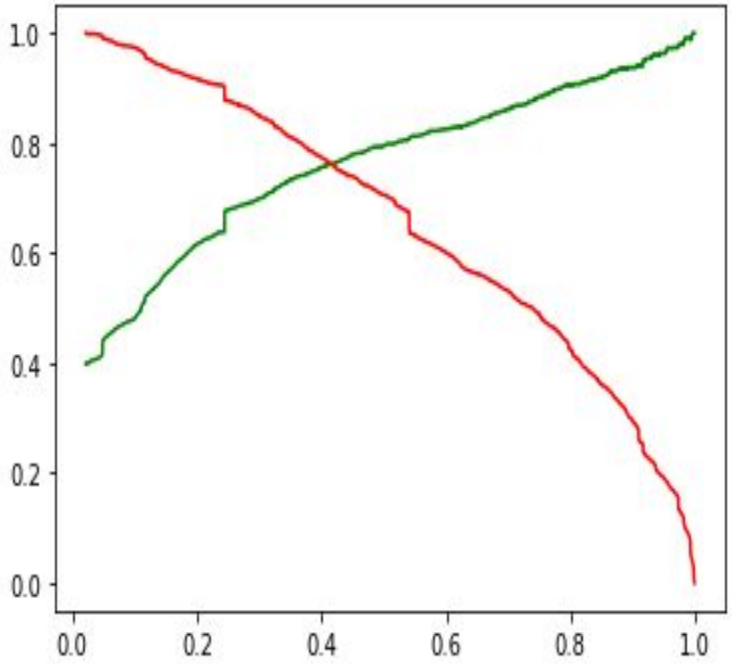
1. The majority of leads have opened their emails as their final action.
2. The conversion rate is around 60% for leads whose most recent activity was sending an SMS.



Model is a good because our area under the ROC curve is higher(0.89)



From the curve above, 0.34 is the optimum point to take it as a cutoff probability.



+ Code + Markdown

**The above graph shows the trade-off between the Precision and Recall .