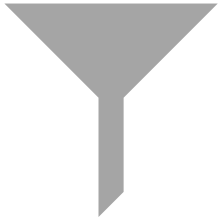
Lead Scoring Case Study

- Neha Lohia
- Pankaj Yavdav
- Samith S

Problem Statement:

An education company named X Education sells online courses to industry professionals. On any given day, many professionals who are interested in the courses land on their website and browse for courses.



The acquired leads are contacted by the sales team through a standard process, some of the leads are converted while others do not..

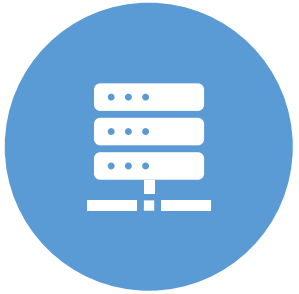
The conversion rate is very poor almost 30%.

Approach:

Once the data was imported, Data Cleaning was performed where the features having missing values more than 45% was dropped. Because imputing / dropping those records would result in misinformation/ loss of information, respectively.

Dropped the features which did not contribute much information and had very minimal variance.

Performed Univariate, Bivariate and Multi-Variate analysis on the provided data



Merged/ Transformed features to prepare the data for modeling.

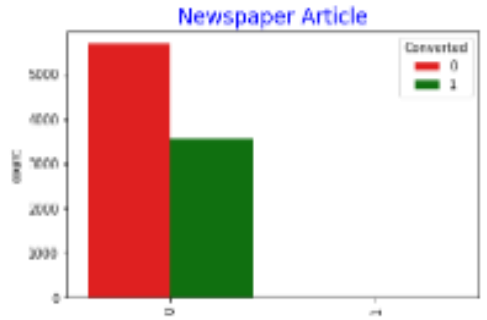
```
In [14]: for col in lead_df.columns:
             print("Duplicate data for column {0}: {1} out of {2}".format(col, lead_df.duplicated(subset=col).sum(), lead_df.shape[0]))

Duplicate data for column Prospect ID: 0 out of 9240
Duplicate data for column Lead Number: 0 out of 9240
Duplicate data for column Lead Origin: 9235 out of 9240
Duplicate data for column Lead Source: 9230 out of 9240
```

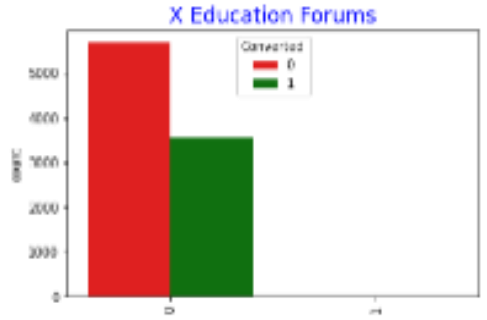
Results:

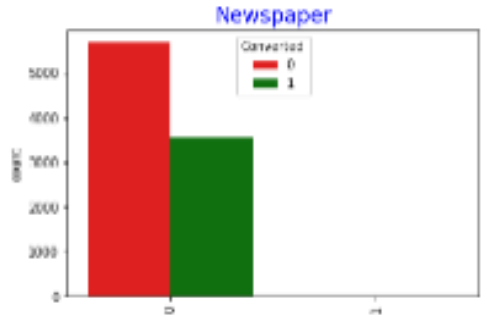
- The features like Prospect ID, Lead Number which do not have any duplicate values have been dropped since there is not much information that can be extracted from this features.

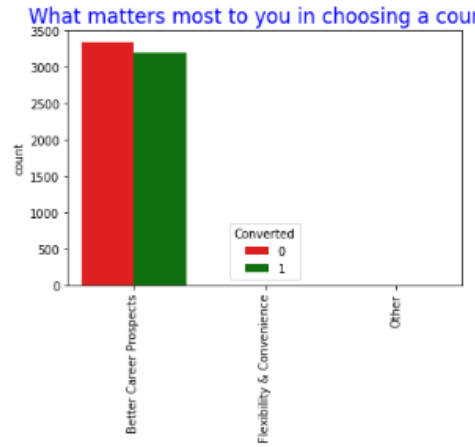
Results (Contd.)

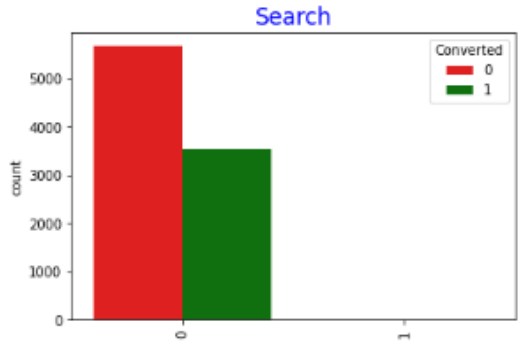
- The features that contain Imbalanced data are dropped.





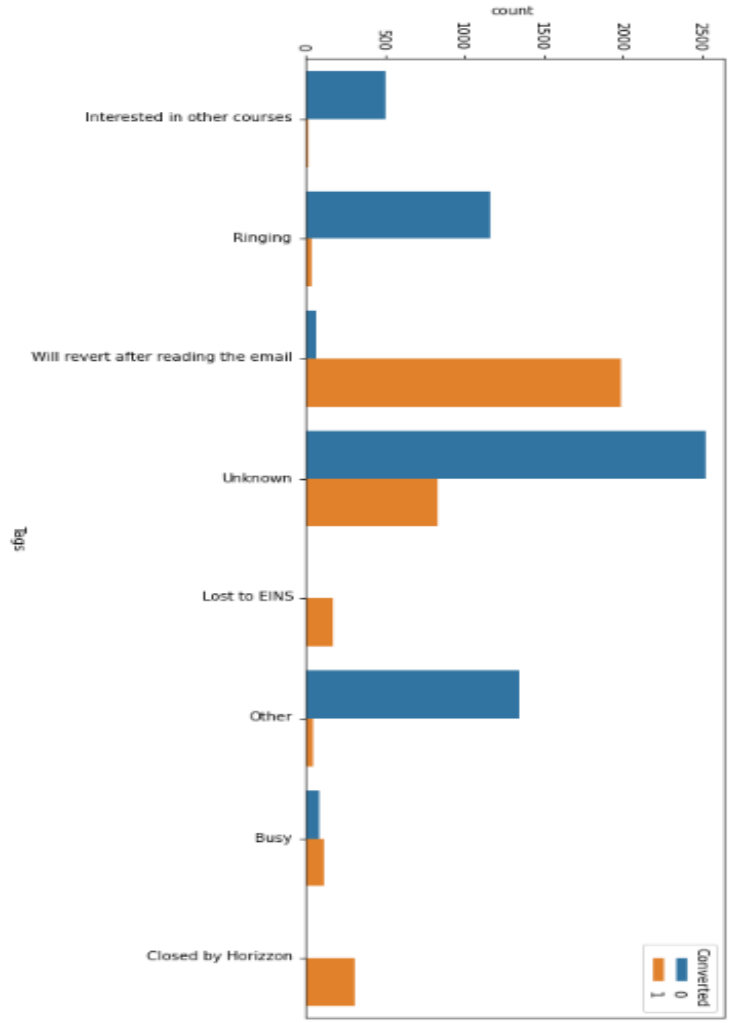






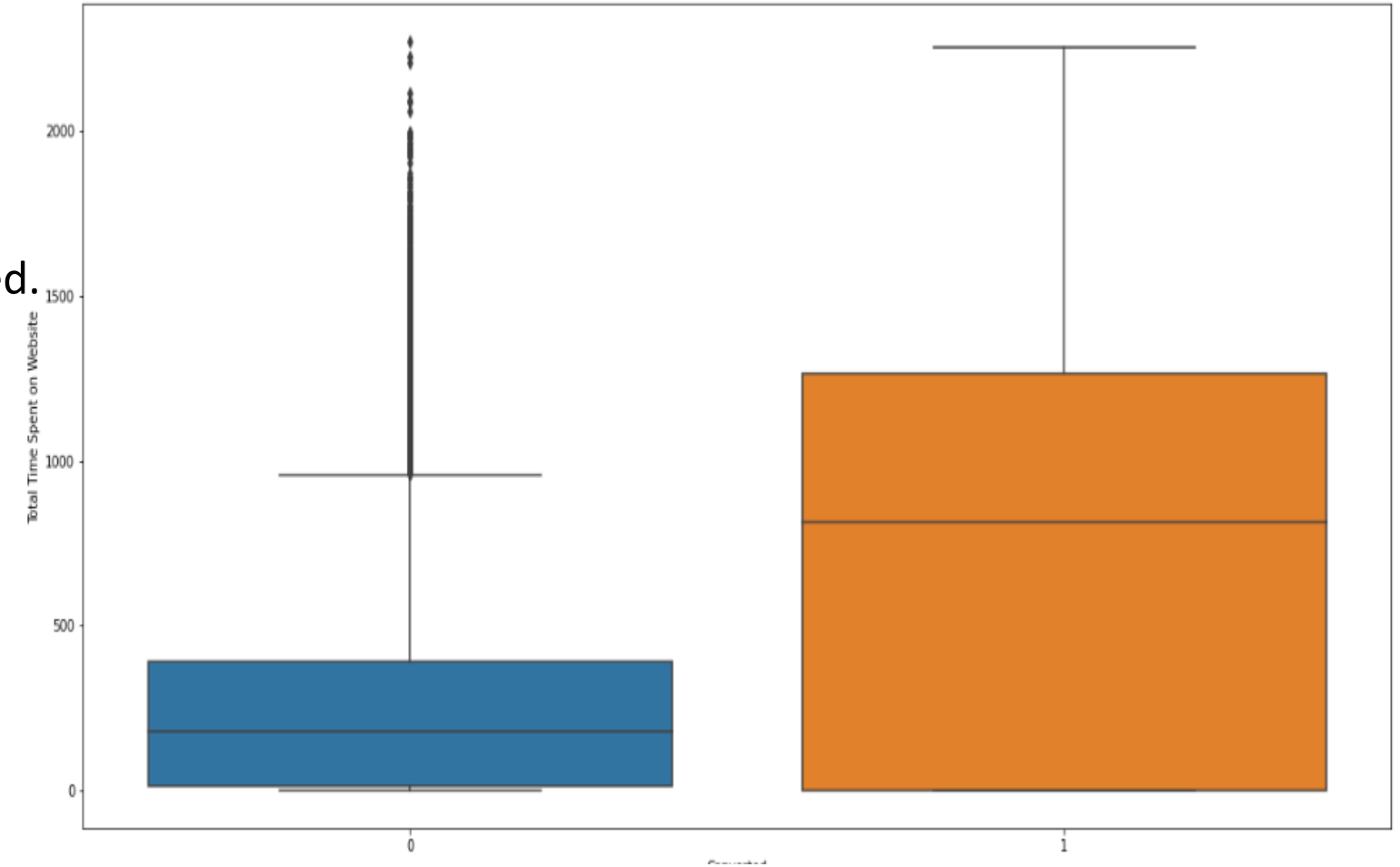
Results (Contd.)

- The class 'Will revert after reading the email’ has provided more conversion when compared to other Tags



Results (Contd.)

- The people who spend more time in website are more likely to be converted.



Final Model Summary:

- The goal is to convert the possible leads to actual leads, the target metric to be evaluated is Precision and Recall.
- Since recall represents the rate of predicted leads that will be converted to actual leads

```
In [104]: accuracy_score(y_test_df.Converted, y_test_df.Predicted)

Out[104]: 0.9245697546686196
```

```
In [105]: precision_score(y_test_df.Converted, y_test_df.Predicted)

Out[105]: 0.8694444444444445
```

```
In [106]: recall_score(y_test_df.Converted, y_test_df.Predicted)

Out[106]: 0.9352589641434262
```