LEAD SCORING ASSIGNMENT

JOSHLY MARY JOHNSON

TARIQ MOHAMMED

PROBLEM STATEMENT

- An education company named X Education sells online courses to industry professionals.
- Although X Education gets a lot of leads, its lead conversion rate is very poor.
- To make this process more efficient, the company wishes to identify the most potential leads, also known as ‘Hot Leads’. If they successfully identify this set of leads, the lead conversion rate should go up as the sales team will now be focusing more on communicating with the potential leads rather than making calls to everyone
- The company requires to build a model wherein a lead score is assigned to each of the leads such that the customers with higher lead score have a higher conversion chance and the customers with lower lead score have a lower conversion chance.

ASSUMPTIONS

- Columns having ‘Select’ are considered as null and are imputed with either NA or mode value.

APPROACH

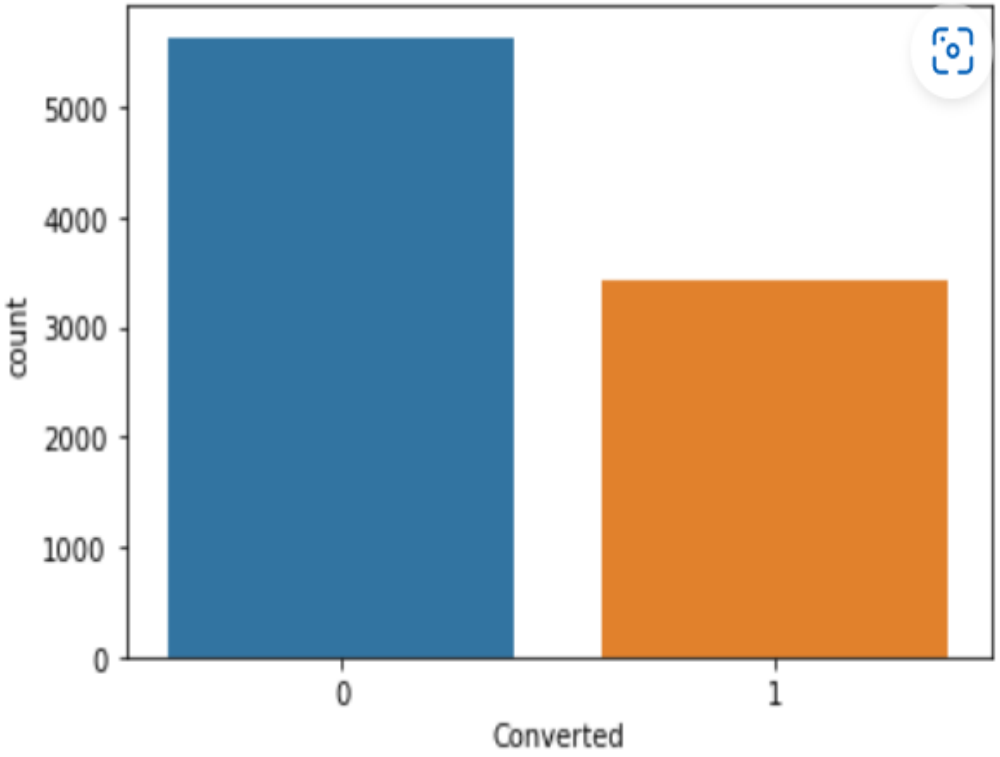
- Understanding the data
- Data cleaning –
  - Handling missing values,
  - highly skewed categorical columns,
  - imputation,
  - handling outliers
- EDA –
  - Univariate,
  - bivariate,
  - multivariate analysis

APPROACH

- Data Preparation –
  - Creating dummy variables,
  - performing train-test data split,
  - feature scaling
- Data modelling –
  - RFE for variable selection
  - Creating Logistic regression model
  - Checking p-value and VIF
  - Check model performance over test data
  - Generate score variable

TARGET VARIABLE

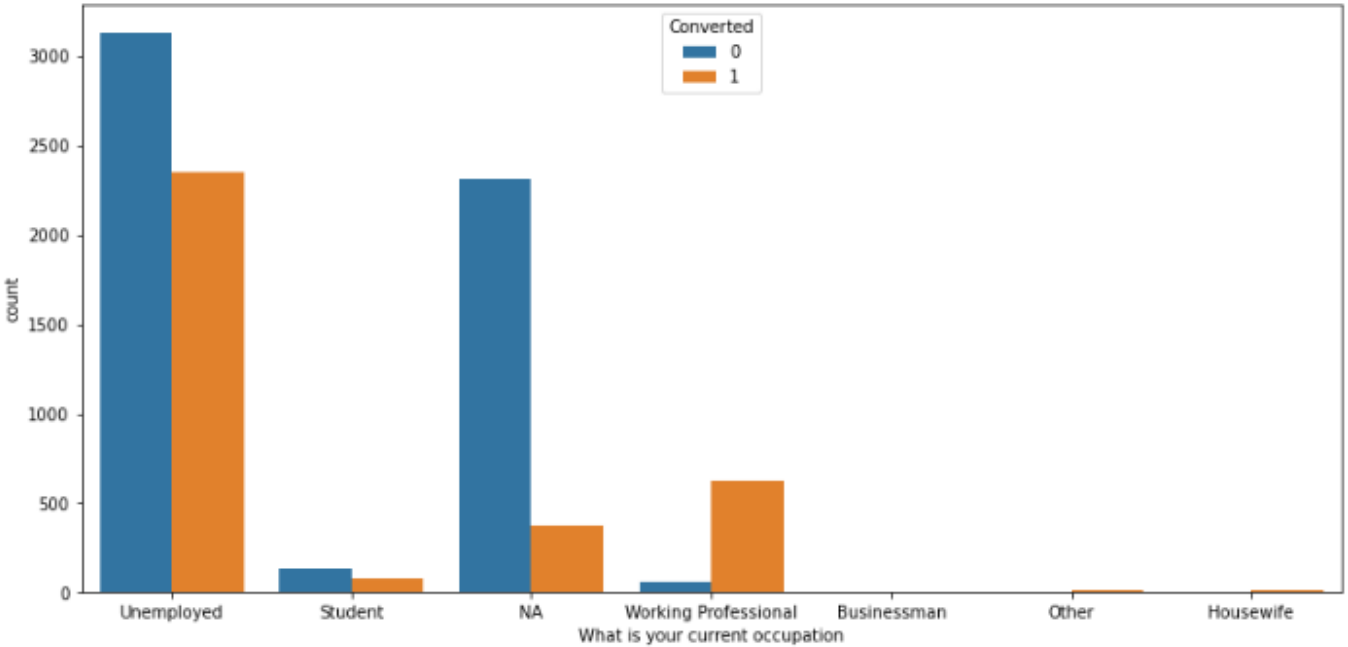
We can observe that the number of converted is less when compared to the ones not converted. The imbalance is not very drastic but is present



0 : Not converted

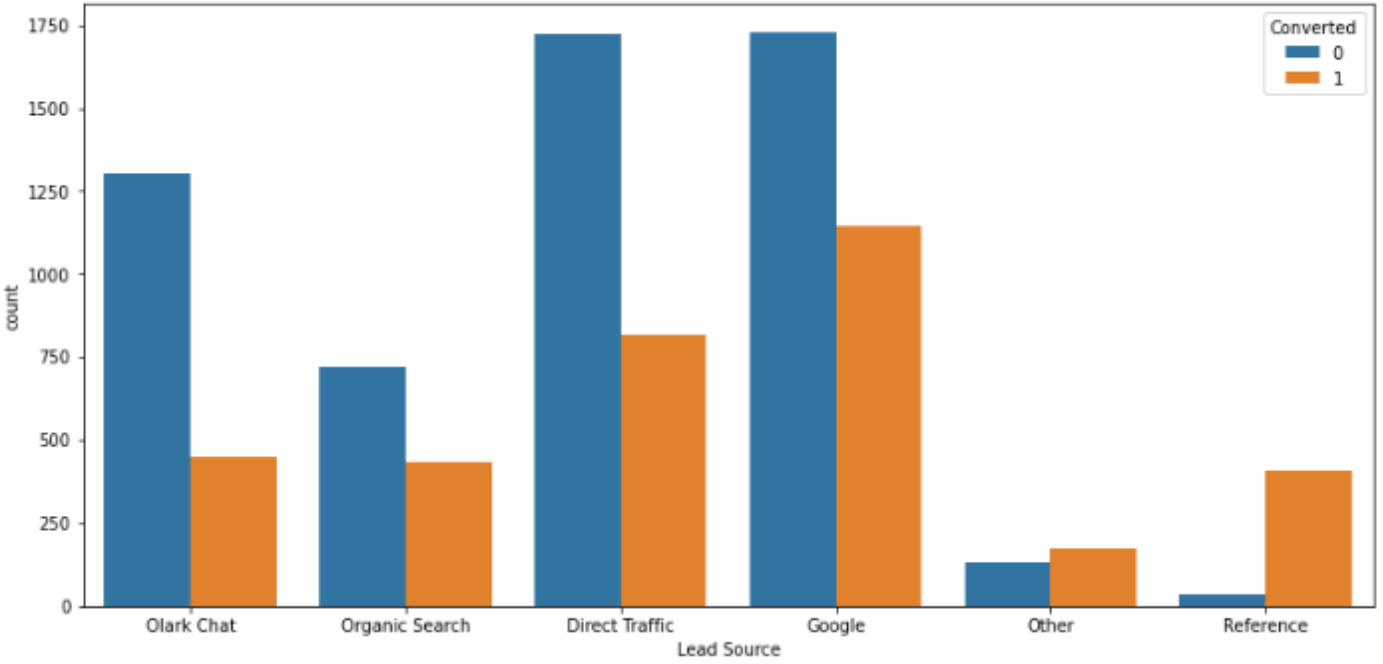
1 : Converted

Occupation



People who are unemployed are more likely to be a lead. We can also observe that working professionals who are contacted are most likely to be converted to a lead.

Lead source



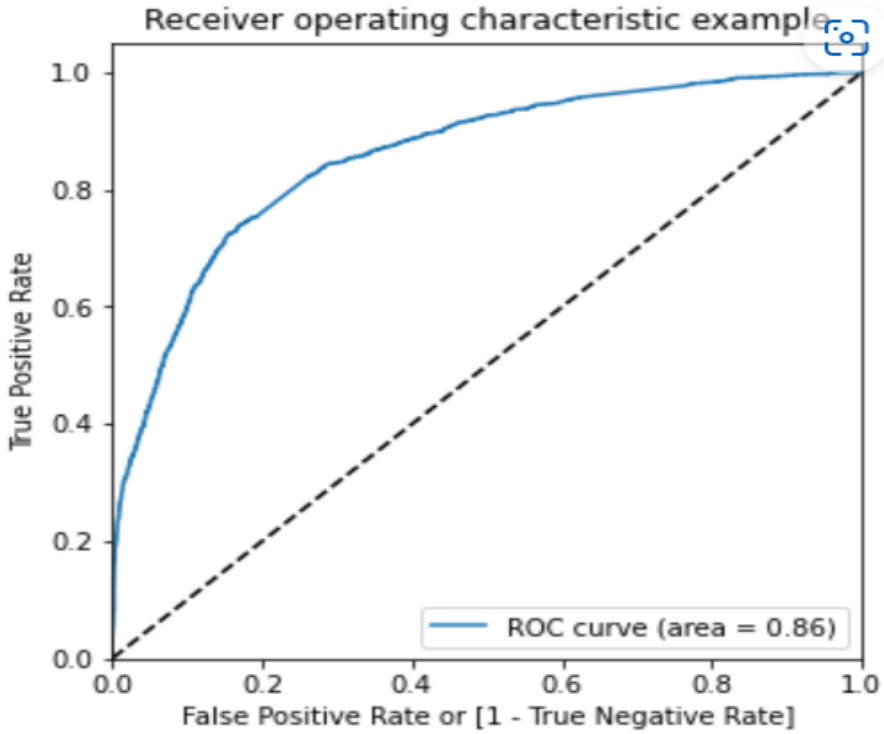
Most of the leads have their source from Google or from direct traffic

Final Model

	coef	std err	z	P>|z|	[0.025	0.975]
const	-1.9781	0.094	-20.989	0.000	-2.163	-1.793
TotalVisits	0.1093	0.040	2.712	0.007	0.030	0.188
Total Time Spent on Website	1.1331	0.039	29.371	0.000	1.058	1.209
Lead Origin_Landing Page Submission	-0.4574	0.095	-4.794	0.000	-0.644	-0.270
Lead Origin_Lead Add Form	5.0077	0.519	9.650	0.000	3.991	6.025
Lead Source_Olark Chat	1.0657	0.116	9.207	0.000	0.839	1.293
Lead Source_Reference	-1.2828	0.555	-2.312	0.021	-2.370	-0.195
Specialization_Banking, Investment And Insurance	0.6639	0.172	3.851	0.000	0.326	1.002
Specialization_Finance Management	0.3612	0.116	3.108	0.002	0.133	0.589
Specialization_Human Resource Management	0.2916	0.119	2.457	0.014	0.059	0.524
Specialization_Marketing Management	0.4488	0.120	3.735	0.000	0.213	0.684
Specialization_Operations Management	0.3697	0.151	2.441	0.015	0.073	0.667
What is your current occupation_Unemployed	1.2103	0.078	15.499	0.000	1.057	1.363
What is your current occupation_Working Professional	3.6582	0.189	19.364	0.000	3.288	4.028
City_Mumbai	0.1923	0.081	2.387	0.017	0.034	0.350

The above features form the final model. We can observe that the p-value of all the features are low.

ROC curve


Receiver operating characteristic example
True Positive Rate
False Positive Rate or [1 - True Negative Rate]
0.0
0.2
0.4
0.6
0.8
1.0
ROC curve (area = 0.86)

Evaluation Metrics

99]:
```
# Sensitivity
print(TP_test / float(TP_test + FN_test))
# Specificity
print(TN_test / float(TN_test + FP_test))
# Recall
print(TP_test / float(TP_test + FN_test))
# Precision
print(TP_test / float(TP_test + FP_test))
```

```
0.8169868554095046
0.7352941176470589
0.8169868554095046
0.6377269139700079
```

The evaluation metrics on the test data are as above. We can see that we have 63% precision and about 81% recall an 81% sensitivity

Accuracy

```
# Checking accuracy
metrics.accuracy_score(y_pred_final.Converted, y_pred_final.final_predicted)
```

```
0.7649651120088138
```

We can also observe that we have about 76% accuracy on the test data.

Lead Conversion rate

```
# Checking if 80% of the customers have converted
checking_df = y_pred_final.loc[y_pred_final['Converted']==1,['Converted','final_predicted']]
checking_df['final_predicted'].value_counts()
```

```
1    808
0    181
Name: final_predicted, dtype: int64
```

```
808/float(808+181)
```

```
0.8169868554095046
```

We have got a Lead conversion rate of 81.7% which is as per the expectations of the client.

CONCLUSION

The following conclusion can be derived from the model that we have created.

- The sales team can concentrate on the Unemployed candidates and the working professionals as they have the highest conversion rates among all other occupations
- Those who have spent more time on the website are more likely to be converted as they would be intrigued to find about the courses. The marketing team could make their website more accessible and interesting and more informative.
- Most of the leads are from the city of Mumbai. Metropolitan cities have more probability to become a lead.