LEAD SCORING CASE STUDY

Dheeraj Sapate

Shiva Krishna

Arpith R K

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 company markets its courses on several websites and search engines like google. Once these people land on the website, they might browse the courses or fill up a form for the course or watch some videos. When these people fill up a form providing their email address or phone number, they are classified to be a lead. Moreover, the company also gets leads through past referrals. Once these leads are acquired, employees from the sales team start making calls, writing emails, etc. Through this process, some of the leads get converted while most do not. The typical lead conversion rate at X education is around 30%.

Now, although x education gets a lot of leads, its lead conversion rate is very poor. For example, if, say, they acquire 100 leads in a day, only about 30 of them are converted. 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 c

OBJECTIVES

- To help the company in selecting the most potential leads, also known as ‘Hot Leads’ whose lead conversion rate is around 80%.
- 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.
- Help the sales team to divert their focus on potential leads & avoid them from making useless phone calls.

APPROACH

- Analysing Patterns: Using Exploratory Data Analysis, we have analyzed the patterns present in the Dataset which will provide us intuition that the which features will help in driving the lead conversion.
- Correlations: Identifying correlations amongst variables to identify the variability in data and identify most important features that can help in driving the conversion of leads.

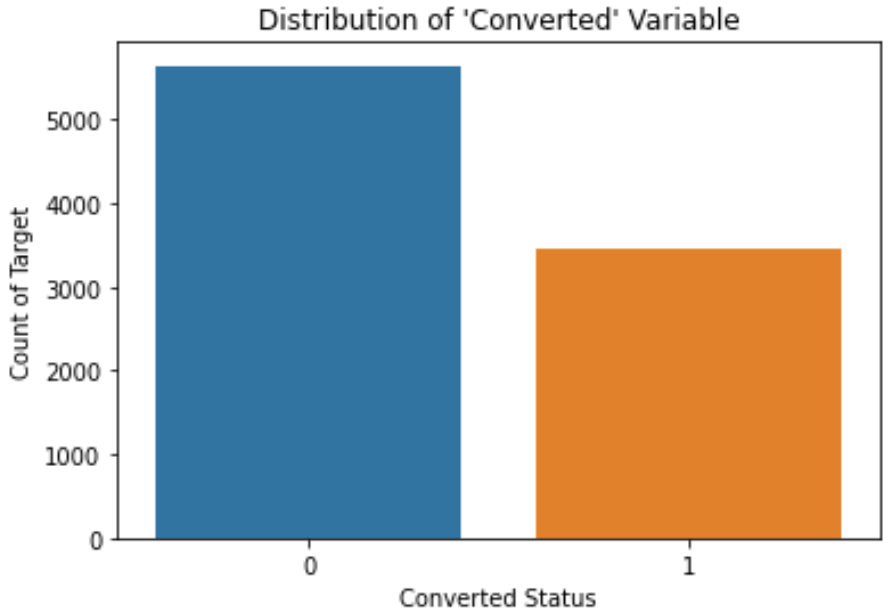
DATA INSIGHTS

We have total 9240 entries of unique customers and we needs to identify out of these which have the highest probability of getting converted.

- **Decision Criteria:** Potential Leads can be bifurcated on the basis of Leads Score (which is probability of getting converted).

Out of 9240 entries we see that around 37% of leads are converted and 73% of leads are not converted.

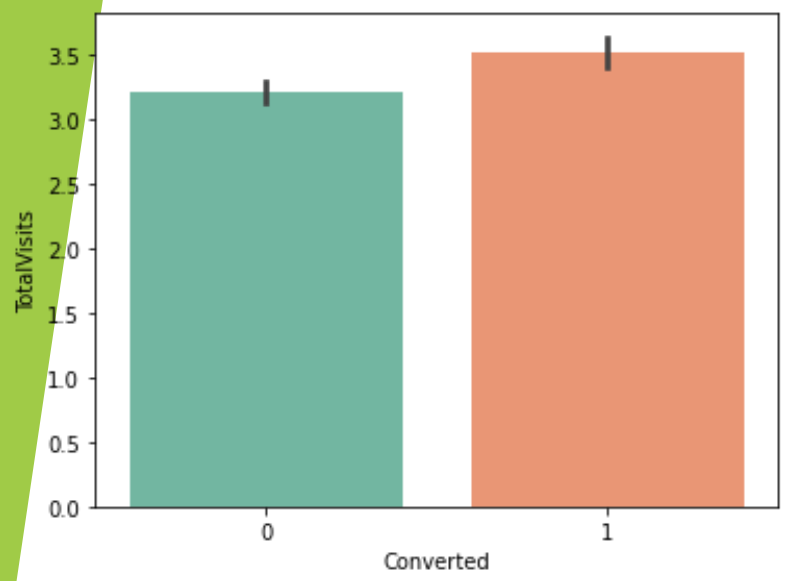
- **Task:** Identify solution so that the lead conversion rate could be increased.

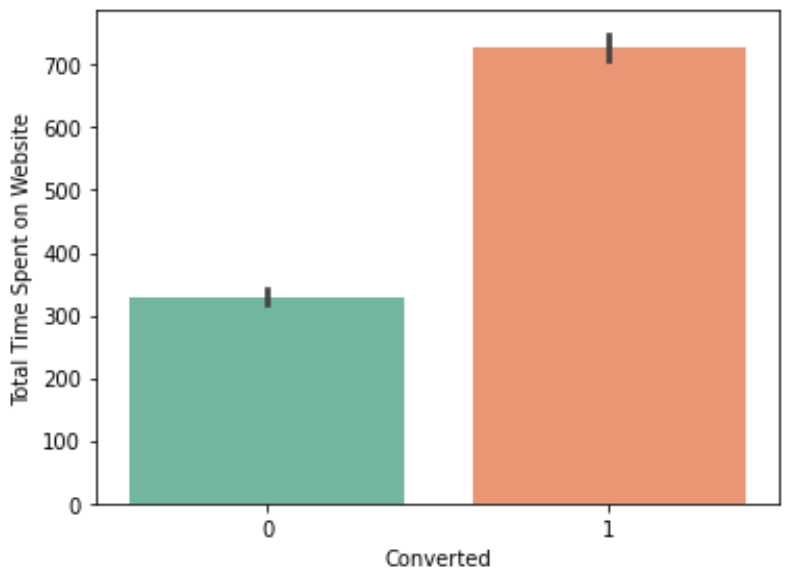


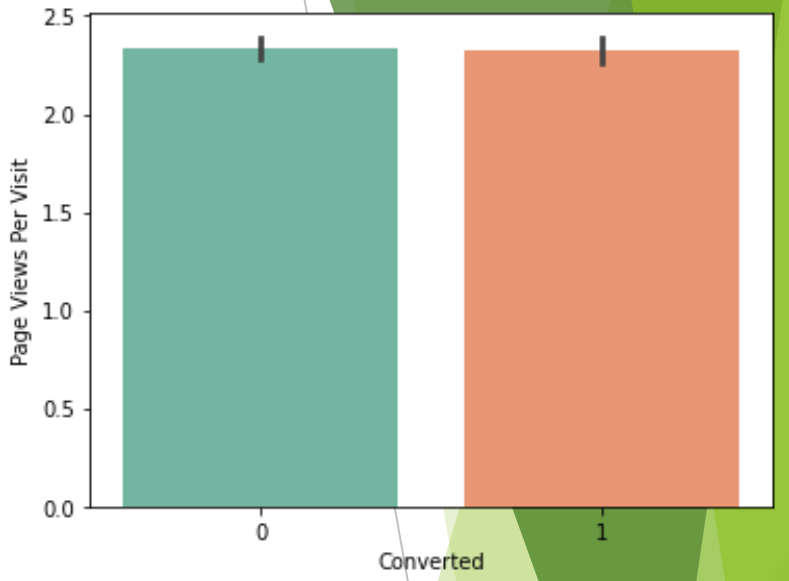
DATA CLEANING

- Handling ‘Select’ variable :“Select” variable indicates that the user has not selected any option, We impute the same with null values.
- Dropping column with high null values: Columns having null values greater than 40% does not have meaning to the data, hence we drop these
- Treating Categorical data: Columns having high data imbalance must be removed. For e.g. : Category A has 98% , and Category B has 2% -This data is irrelevant to our analysis as one category is overpowering the other.
- Columns which do not make any insights for our analysis are removed.

EDA : NUMERICAL DATA





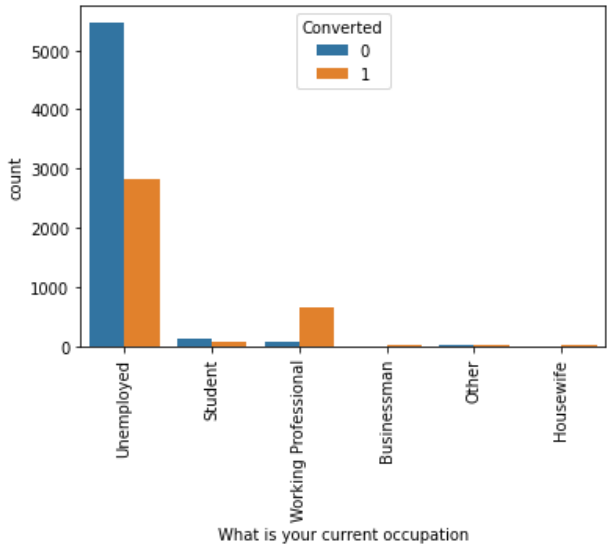


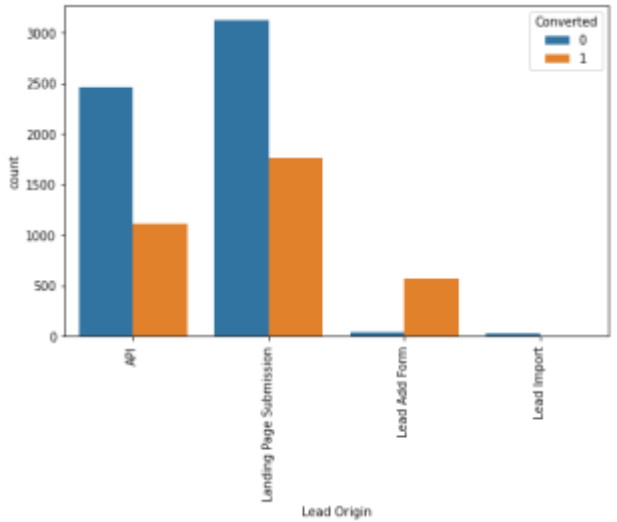
As the median for both converted and non-converted leads are same , nothing coclusive can be said on the basis of variable TotalVisits

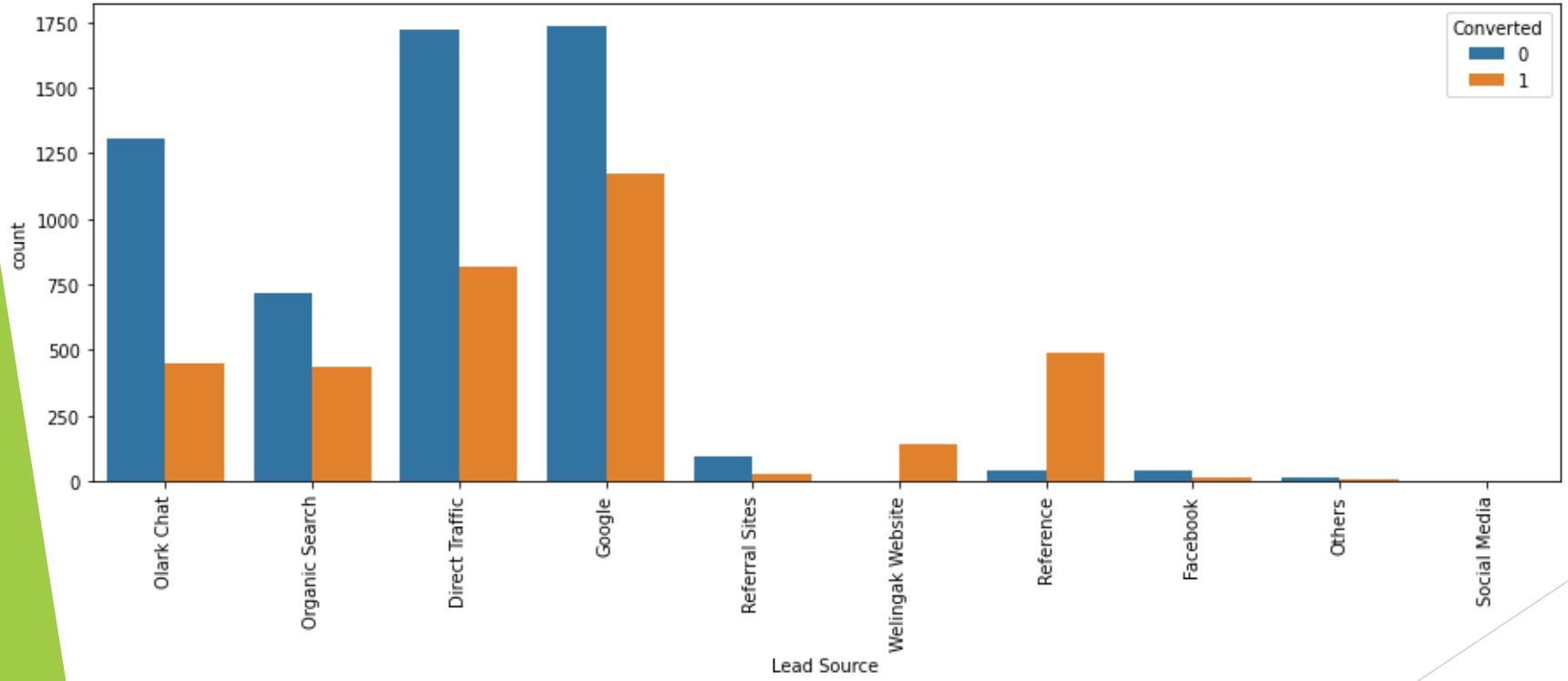
Leads spend more time on the website therefore more engagement should be shown on website.

Median for converted and unconverted leads is the same, Nothing can be said specifically for lead conversion from Page Views Per Visit

EDA : CATEGORICAL DATA







- Working Professional have high conversion rate.
- Unemployed personals are most in terms of numbers.
- Maximum Leads are generated by Google and Direct Traffic.
- Conversion rate of Reference leads and Welinkgak Website leads is very high.
- API and Landing Page Submission bring higher number of leads as well as conversion.
- Lead Add Form has a very high conversion rate but count of leads are not very high.
- Lead Import and Quick Add Form get very few leads.

We can observe that the variables are not highly correlated with each other. But still there is multicollinearity among some features

FACTORS RESPONSIBLE IN DRIVING LEADS

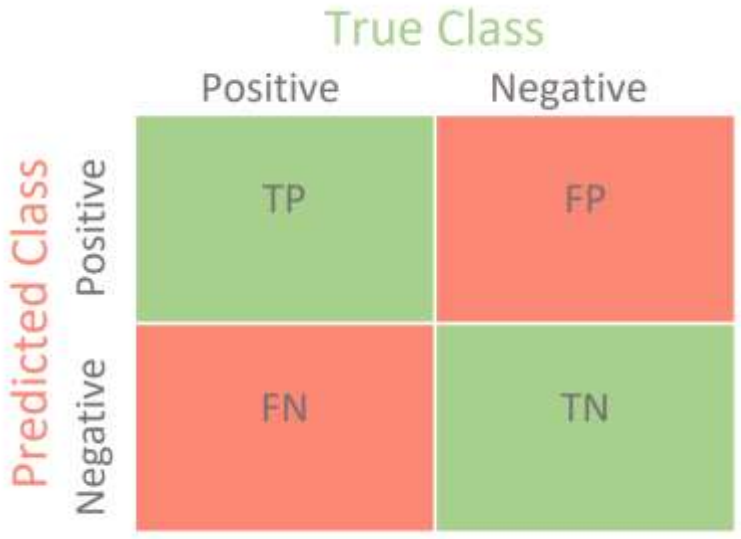
Below features are most important ones which are responsible for leads conversion

- Total Time Spent on Website
- Lead Origin_Lead Add Form
- Lead Origin_Lead Import
- What is your current occupation_Working Professional
- Lead Source_Olark Chat
- Last Activity_Converted to Lead
- Last Activity_Email Bounced
- Last Activity_Olark Chat Conversation
- Last Notable Activity_SMS Sent

TERMINOLOGIES REQUIRED

Before proceeding ahead, we need to understand few terminologies

- **Conversion of categorical columns to numerical.** This step is done as our algorithm runs only on numerical data.
- **Feature Scaling.** This is done to bring our data into same scale.
- **Data Splitting:** We have split the data into 80:20 and named it as train data and test data. We run model on train data and validate our model on test data.
- **Confusion Matrix:**



Where,

True positive (TP): correct positive prediction

False positive (FP): incorrect positive prediction

True negative (TN): correct negative prediction

False negative (FN): incorrect negative prediction

Above Metrics is Known as Confusion Metrics, using above metrics we derived following things:

1. **Accuracy** = (True Negative + True Positive)/Total

This metrics provides the accuracy of the model, where total is TP + FN + FP +FN

2. **Sensitivity** = True Positive / (True Positive + False Positive)

Sensitivity (SN) is calculated as the number of correct positive predictions divided by the total number of positives. It is also called recall (REC) or true positive rate (TPR). The best **sensitivity** is 1.0, whereas the worst is 0.0.

3. **Specificity** = True Negative/ (True Negative + False Negative)

Specificity (SP) is calculated as the number of correct negative predictions divided by the total number of negatives. It is also called true negative rate (TNR). The best **specificity** is 1.0, whereas the worst is 0.0.

4. **Precision** = True Positive/ (True Positives +False Positives)

Precision is defined as the number of true positives divided by the number of true positives plus the number of false positives.

5. **Recall** = True Positives/(True Positives +False Negatives)

The precise definition of **recall** is the number of true positives divided by the number of true positives plus the number of false negatives. True positives are data point classified as positive by the model that actually are positive (meaning they are correct), and false negatives are data points the model identifies as negative that actually are positive (incorrect).

1. Train Data:

- Accuracy : 80.5%
- Sensitivity : 79.4%
- Specificity : 81.1%
- Precision: 72.05%
- Recall: 79.41%

2. Test Data:

- Accuracy : 80.88%
- Sensitivity : 76.1%
- Specificity : 83.77%
- Precision: 74.34%
- Recall: 76.19%

The Model seems to predict the Conversion Rate very well and we should be able to give the CEO confidence in making good calls based on this model

Conclusion

- The model seems to predict the conversion rate very well and we should be able to give the CEO confidence in making good calls based on this model.
- We must majorly focus on working professionals.
- It’s always good to focus on customers, who have spent significant time on our website.
- We must majorly focus on leads whose last activity is SMS sent or Email opened.