LEAD SCORING CASE STUDY USING LOGISTIC REGRESSION

By

Achama James

Santhossh J U

Revanth

Agenda

- **Problem Statement**
- **Problem Approach**
- **EDA**
- **Correlations**
- **Model Evaluation**
- **Observation**
- **Inference**

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. When people fill up a form providing their email address or phone number, they are classified to be a lead. 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%.

- The company requires you to build a model wherein you need to assign a lead score to each of the leads such that the customers with a higher lead score have a higher conversion chance and the customers with a lower lead score have a lower conversion chance. The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%.

Problem Approach

- Data cleaning
- Data Preparation
- Data analysis
- Dummy variable creation
- Test-Train split
- Feature Scaling
- Correlation
- Model Building
- Model Evaluation
- Prediction

Exploratory Data Analysis

After inspecting the data remove the null values and impute the data for better analysis. Check for outliers and treat them. Now that the data is clean and void of outliers we can do Exploratory Data Analysis in the prepared data to understand the general trends.

EDA – Data Cleaning

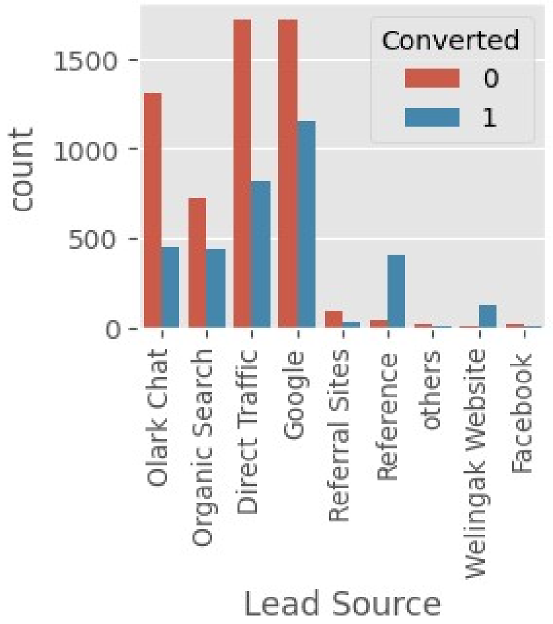
- There are two columns namely Last Activity and Last Notable Activity which is kind of repetitive so we dropped Last Activity column.

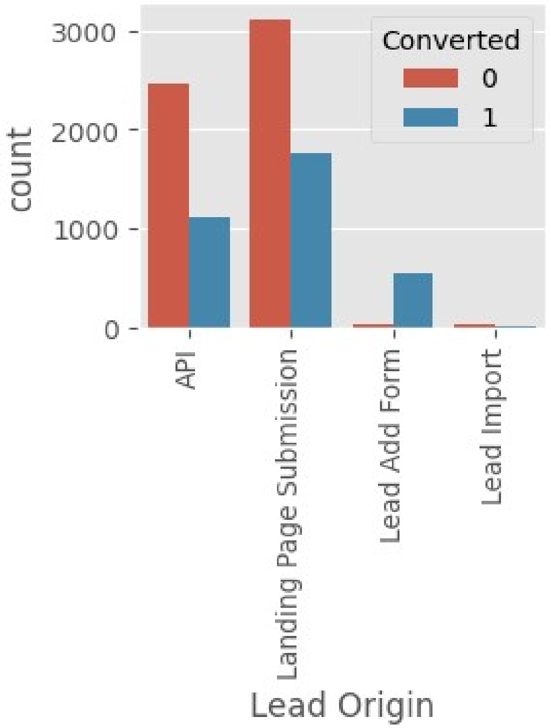
- Majority of the values in country columns has it as India and there are values that has very less contribution. As this might increase the number of dummy variables we are grouping values with less than 30 occurrences as a single new group.

- Columns like Do Not Call, Search are single category or have very low other category. As these columns won’t add much value to analysis we are dropping it.

- There are some columns having either Yes or No as value. As a numeric indicator is much more efficient in analysis point of view we are mapping Yes to 1 and No to 0.

Lead Score and Lead Origin



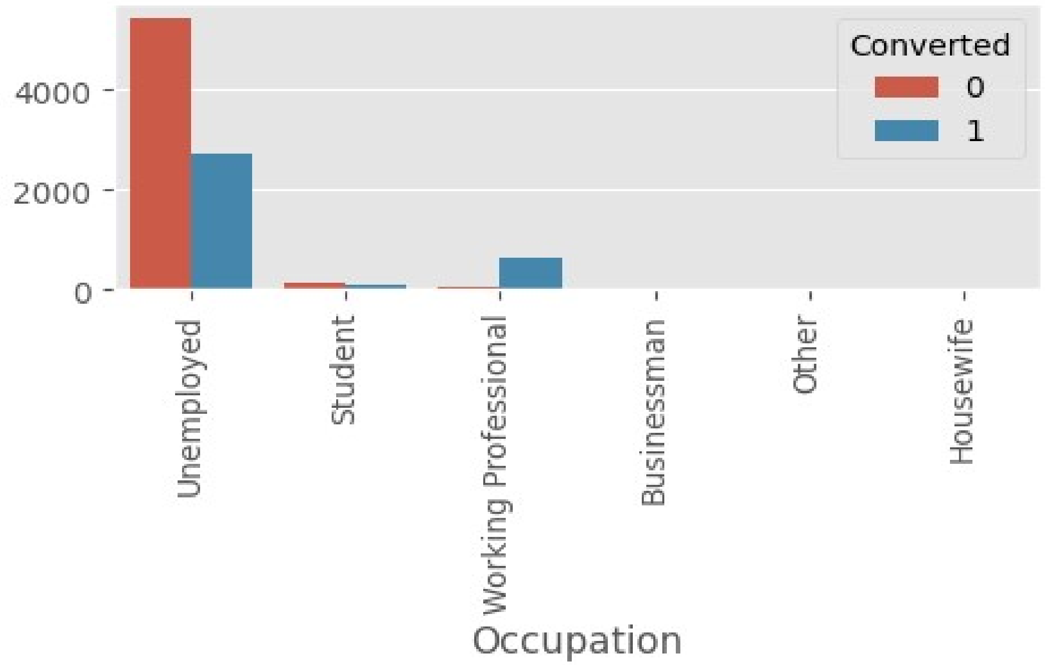


- Of all the available Lead Origins API and Landing page sees highest traffic but the conversion ratio is higher for Lead Add Form

- Google and Direct traffic are the top Lead Sources whereas References has the highest conversion rate.

- When either the lead source is reference or Lead origin is Lead Add Form then the lead is most likely to be converted

Occupation

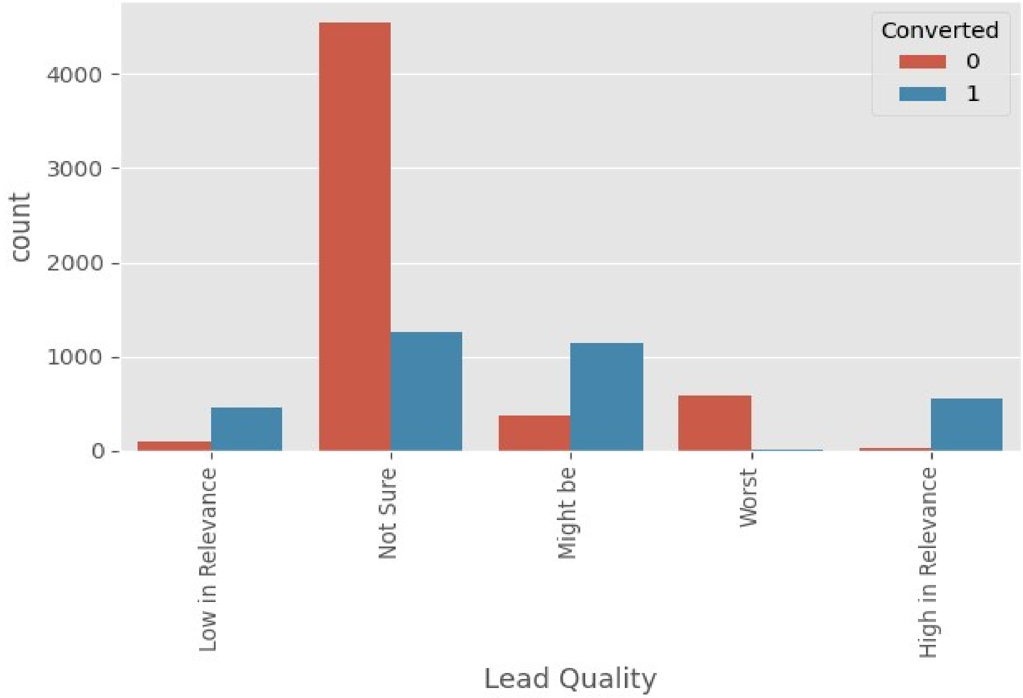


- Unemployed and Working professional are the most common people who search for courses.

- Even though the traffic of unemployed is high the conversion ratio is considerably high for working professional

- Working professional is most likely to buy a course also unemployed has a high chance to get converted

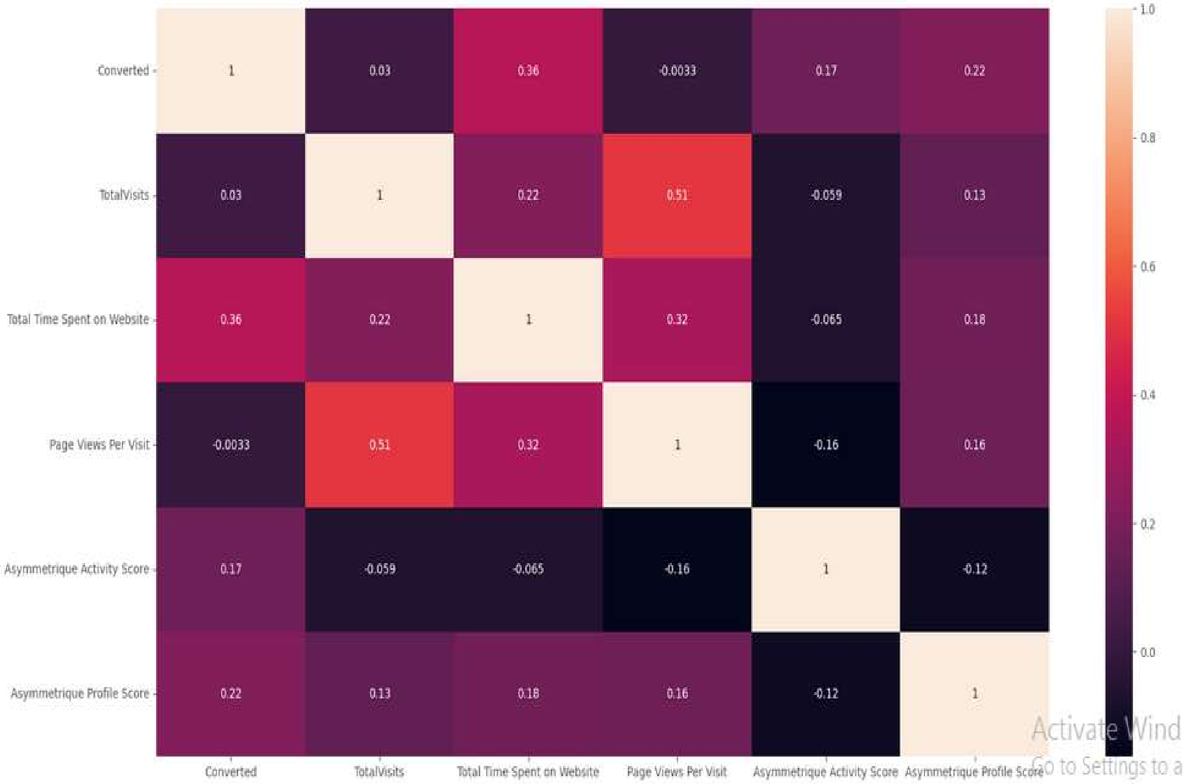
Lead Quality



- We can observe that leads with lead quality as Not sure shows high traffic but we need to consider we imputed the null values here with Not sure. So the actual traffic with Not Sure might be considerably low

- We can observe High in Relevance has the highest conversion ratio among all the options

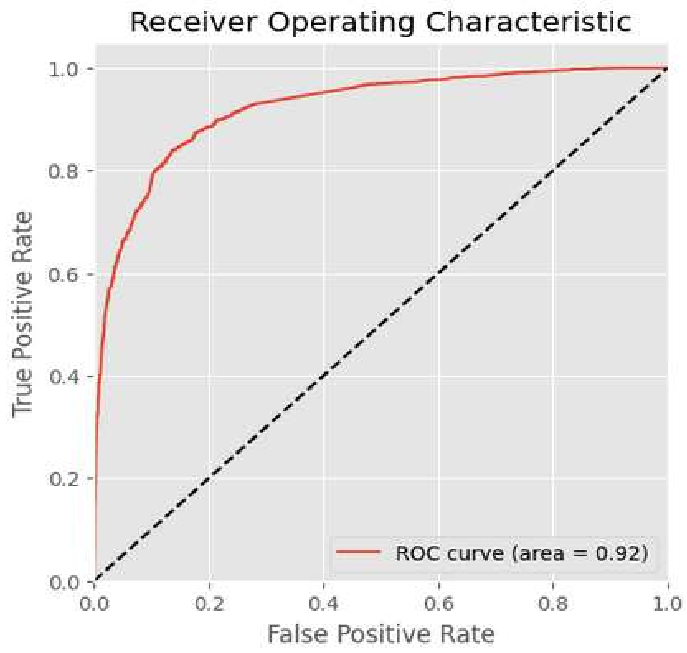
Correlation



- From heat map we can observe that the correlation between TotalVisits and Converted is pretty high which confirms the general assumption that more the customer visits the site more likely the customer will take up a course.

Evaluating the model

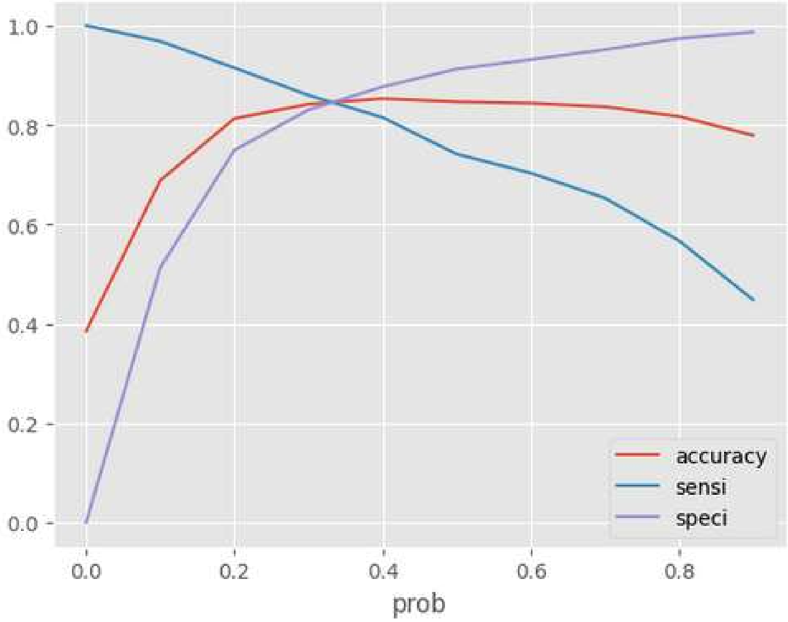
➢ ROC Curve

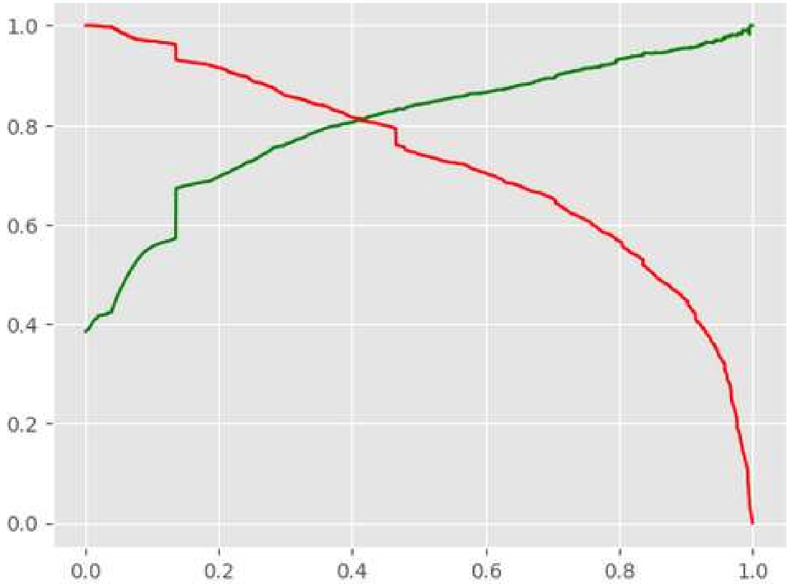


- We created ROC curve to find the model stability with Area under the curve. As we can see in the graph we are getting an area of 0.92 which is good.

- Also the cure is pretty far left from the diagonal line which also confirms us that the model has good accuracy

Evaluating the model

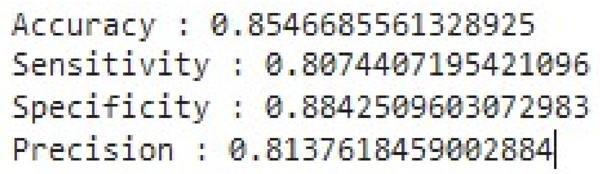




In Accuracy-Sensi-Speci plot 0.37 looks optimal. In precision-recall curve 0.42 looks optimal. We are taking 0.42 is the optimum point as cut-off probability.

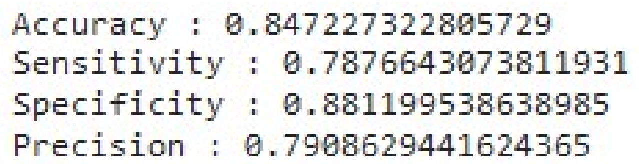
Evaluating the model

Train Data



```
Accuracy : 0.8546685561328925
Sensitivity : 0.8074407195421096
Specificity : 0.8842509603072983
Precision : 0.8137618459002884
```

Test Data



```
Accuracy : 0.847227322805729
Sensitivity : 0.7876643073811931
Specificity : 0.881199538638985
Precision : 0.7908629441624365
```

The difference in metrics between Train and Test data are pretty close. The model seems good.

Observation

- Lead originating from API and Landing page submission are high but the conversion rate for these channels is low compared to Lead add form.
- Google and Direct traffic are the major contributors in lead source sector but their conversion is rate is lower compare to that of referrals or Welingak website.
- Most people that search for the courses are either unemployed or professionals. Of the two professionals have higher conversion rate.
- Lead quality which is inputted by the employees shows great value to understand the customers. Lead quality with High in Relevance shows highest conversion ratio and Worst shows the lowest conversion ratio of all.

Inferences

- Welingak website and References have highest conversion ratio but they don’t have high traffic. We need to advertise more in website and increase references.
- Similarly Lead add form has the highest conversion ratio but they are not the cause for high traffic. We need to increase traffic in this channel.
- Working professionals have the highest conversion ratio but most unemployed people are visiting. We need to attract more professionals at the same time we should promote offers for unemployed people to increase the conversion.