LEAD SCORING ASSIGNMENT

GROUP MEMBERS:

ADITYA GUPTA
AYUSH KUMAR
SATYA JEET VERMA

PROBLEM STATEMENT:

- X Education sells online courses to industry professionals.
- X Education gets a lot of leads, its lead conversion rate is very poor (~30%)
- 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.

Goal:

- 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%.

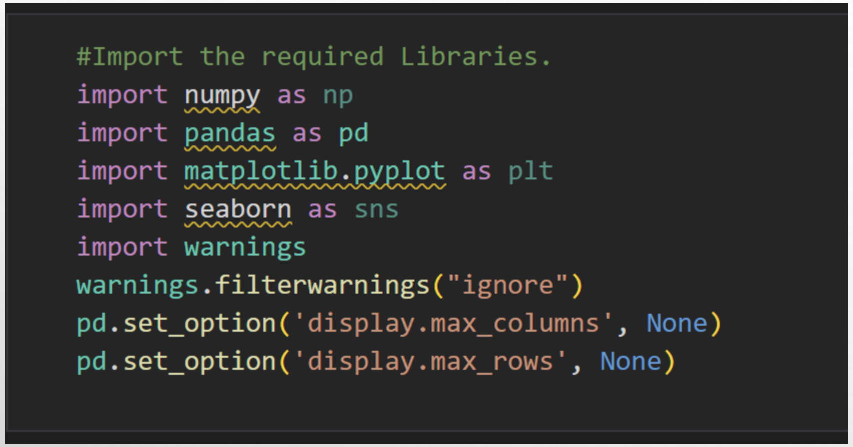
Contents:

1. Importing libraries and loading data
2. Data Cleaning & Outlier Treatment
3. Exploratory Data Analysis
4. Data Preparation
5. Model Building
6. Conclusion

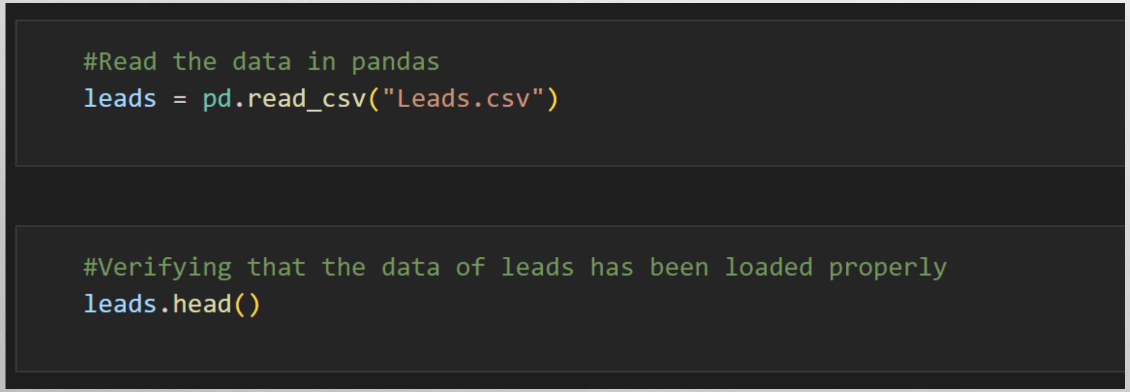
1. Importing libraries and loading data

- Required libraries are imported.
- Reading leads data and getting basic information of the data.

Code Samples below:



```
#Import the required Libraries.
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import warnings
warnings.filterwarnings("ignore")
pd.set_option('display.max_columns', None)
pd.set_option('display.max_rows', None)
```



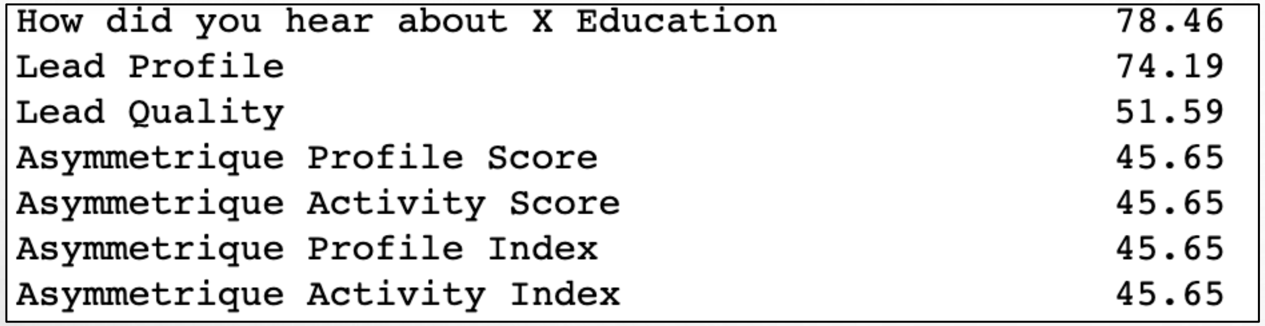
```
#Read the data in pandas
leads = pd.read_csv("Leads.csv")
```

```
#Verifying that the data of leads has been loaded properly
leads.head()
```

2. Data Cleaning & Outlier Treatment

Identifying missing values & removing/imputing them

- Columns having > 45% null values are removed (examples mentioned below)



```
How did you hear about X Education          78.46
Lead Profile                                74.19
Lead Quality                                51.59
Asymmetrique Profile Score                  45.65
Asymmetrique Activity Score                 45.65
Asymmetrique Profile Index                  45.65
Asymmetrique Activity Index                 45.65
```

- Columns having < 2% nulls like ‘Lead Source’, ‘TotalVisits’, the null rows are dropped for them.
- For ‘What is your current occupation’ column, nulls are imputed with the mode of the same column.

Outlier Treatment

- Outlier treatment is done on ‘Total Time Spent on Website’ column by capping bottom & top 1 percentile values to 1 & 99 percentile respectively.
- For ‘Page Views Per Visit’ column, the capping was done to the max value of 10.

Data Manipulation

- For columns like ‘Lead Source’, ‘Last Activity’, ‘Specialization’, ‘What is your current occupation', 'Tags', 'City' and 'Last Notable Activity’ has many number of categories so lesser frequent categories are imputed as ‘Others’ category.
- 'Magazine', 'X Education Forums', 'Receive More Updates About Our Courses', 'Update me on Supply Chain Content', 'Get updates on DM Content', 'I agree to pay the amount through cheque’, 'Newspaper Article', 'Newspaper', 'Digital Advertisement', 'Through Recommendations' have mostly one value so they are dropped.

3. Exploratory Data Analysis

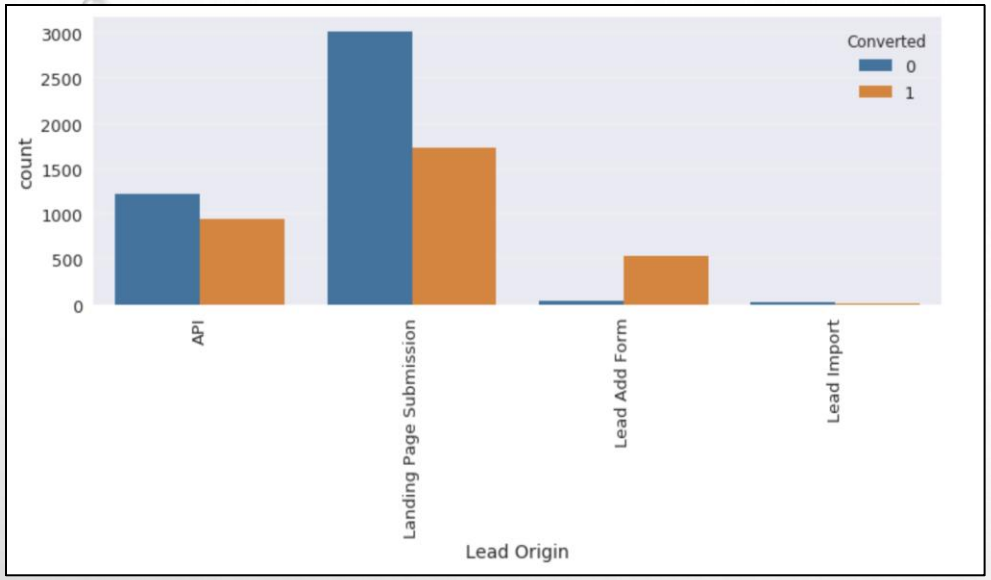


Fig1: Leads originating through API & Landing Page Submission has the highest number of conversion. Lead who have added forms has the highest conversion % but it doesn't have a greater number of leads.

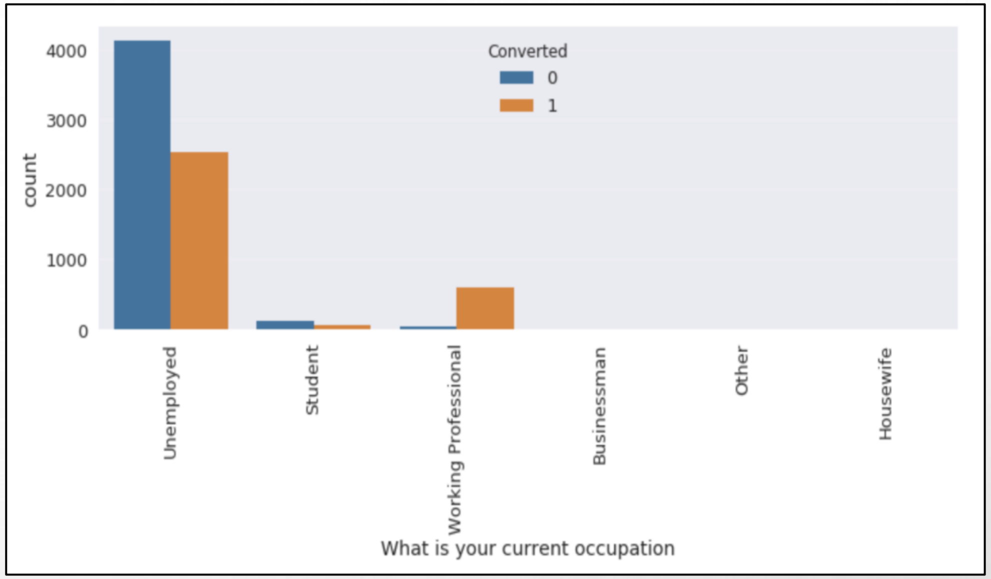


Fig2: Although unemployed leads have higher number of conversions but conversion rate is around 40-50%. Working Professionals have highest conversion rates.

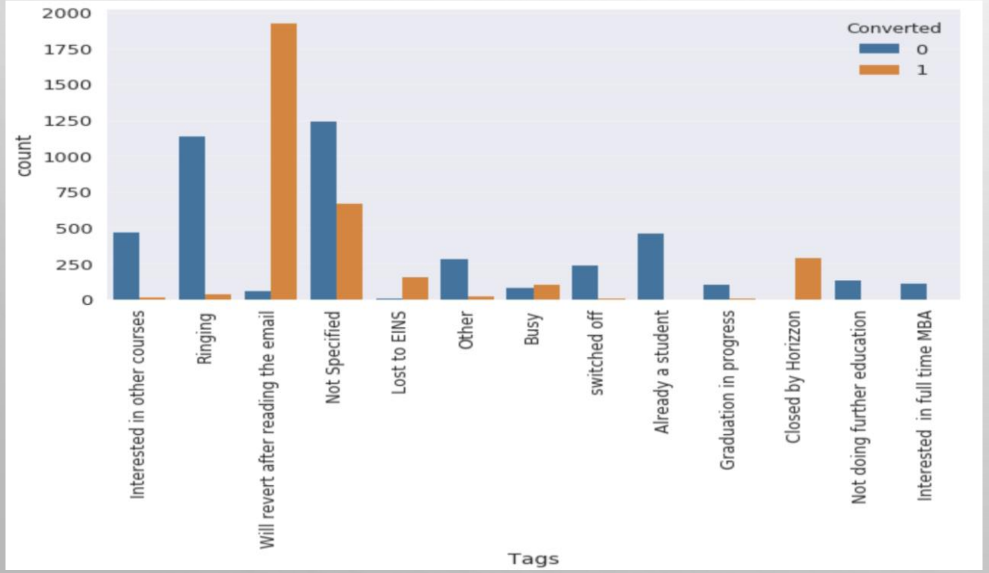


Fig3: Leads with tags as "Will revert after reading the email', 'Lost to EINS', 'Closed by Horizzon' have the highest conversion rates.

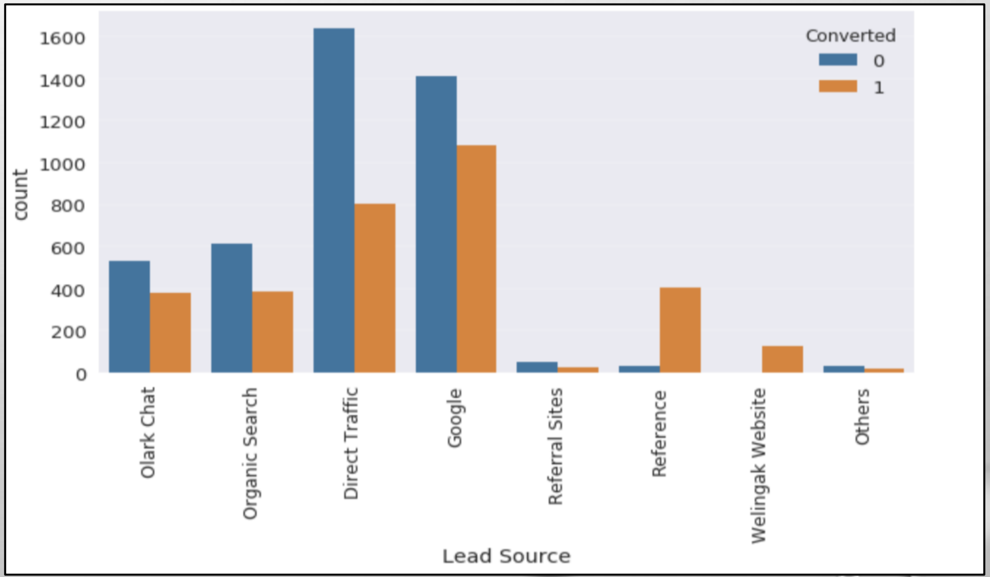


Fig4: Google & Direct Traffic has the highest number of leads

3. Exploratory Data Analysis

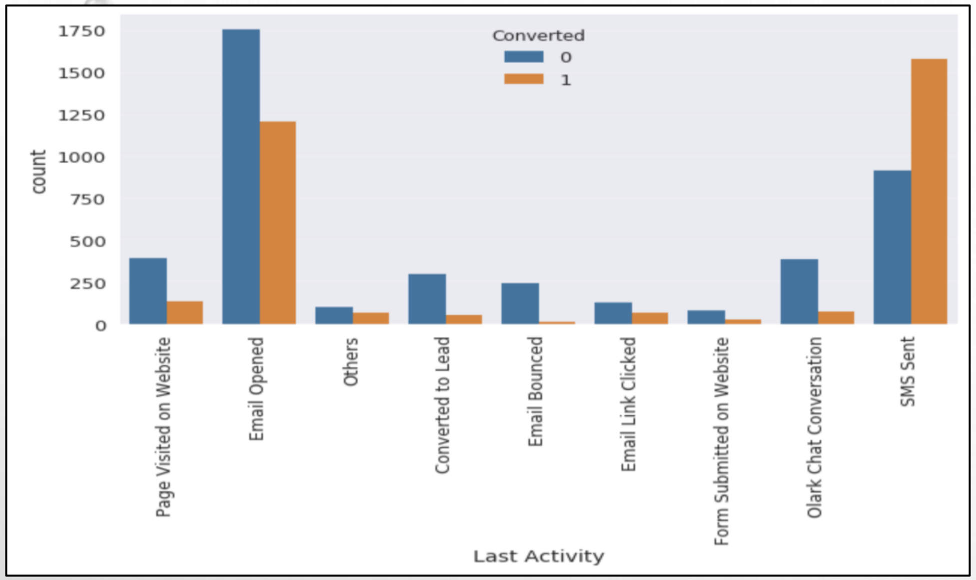


Fig5: Leads with last activity as Email Opened & SMS Sent has the highest number of conversions.

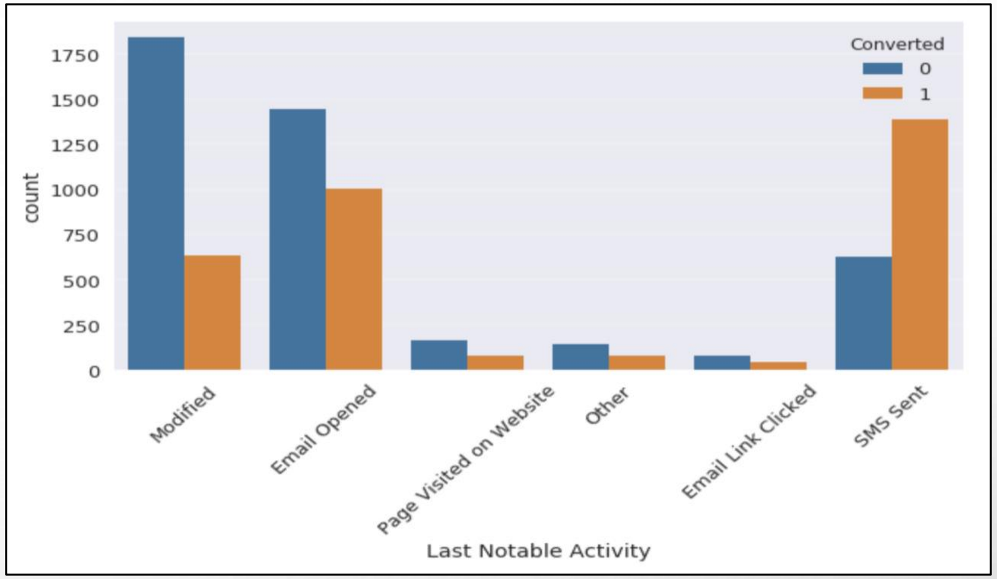
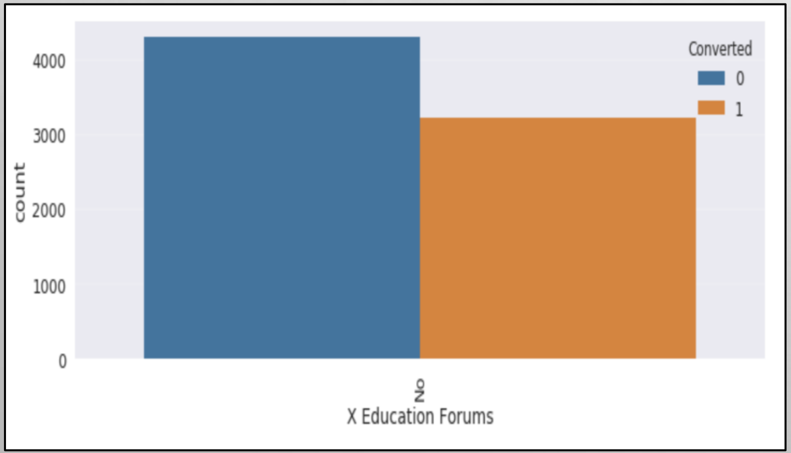
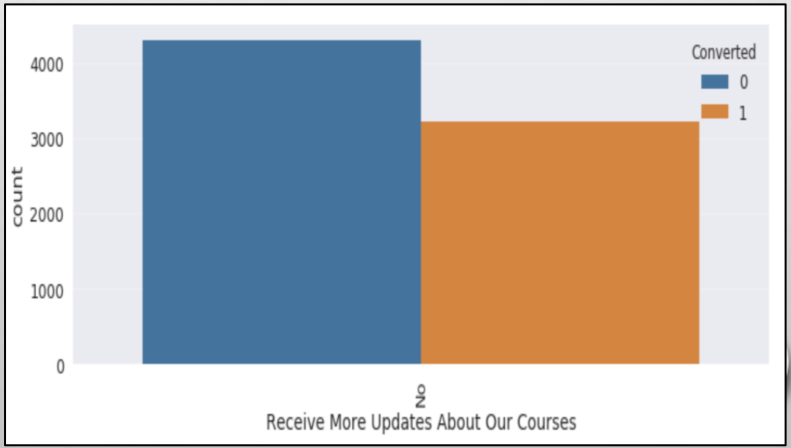
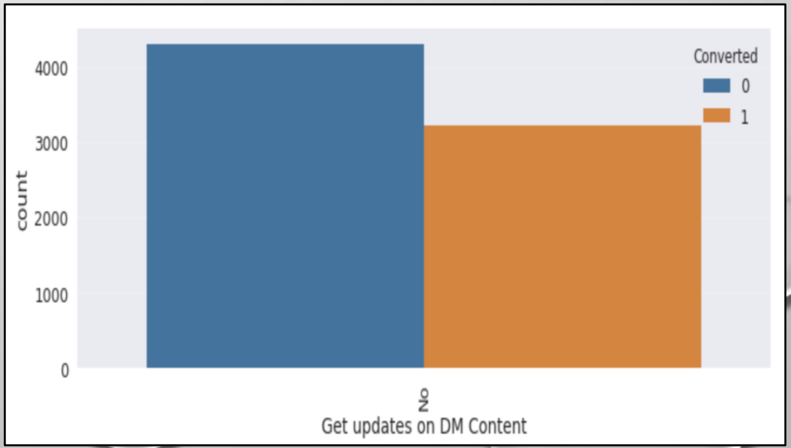


Fig6: Leads with last notable activity as SMS Sent has the highest conversion rate while last notable activity as Email opened & modified has the highest number of customers.

There are certain columns which have only one value present in the data. These columns are removed from the analysis.



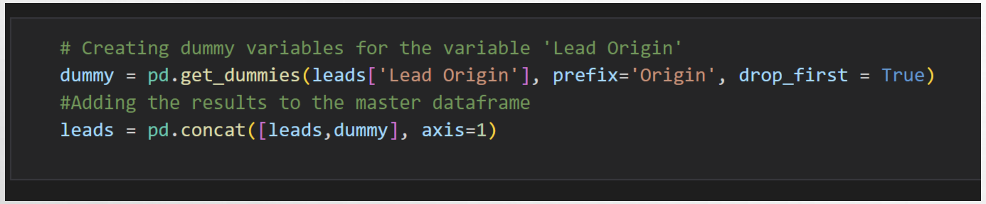




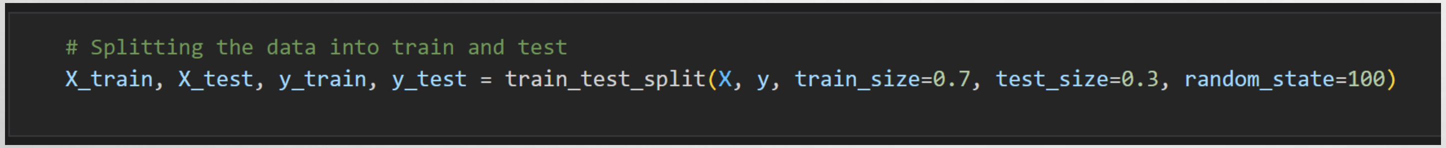
4. Data Preparation

- For 'Do Not Email’, 'Do Not Call’, 'Search', 'A free copy of Mastering The Interview’ columns, ‘Yes’ & ‘No’ values are converted into binary numbers 1 & 0 respectively.
- Dummy variable creation is done for categorical columns.
- To reduce correlated columns, we have used heatmap to drop certain unnecessary columns.
- Train–test split of 70%-30% is done.
- For feature scaling, we have used StandardScaler.

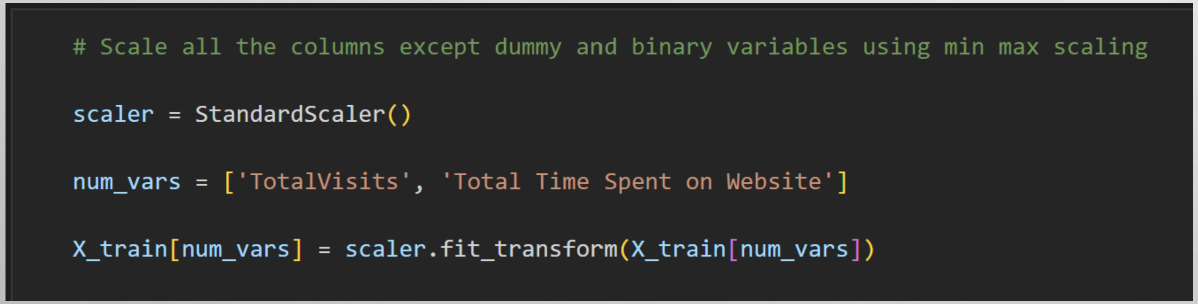
Code samples below:



```
# Creating dummy variables for the variable 'Lead Origin'
dummy = pd.get_dummies(leads['Lead Origin'], prefix='Origin', drop_first = True)
#Adding the results to the master dataframe
leads = pd.concat([leads,dummy], axis=1)
```



```
# Splitting the data into train and test
X_train, X_test, y_train, y_test = train_test_split(X, y, train_size=0.7, test_size=0.3, random_state=100)
```

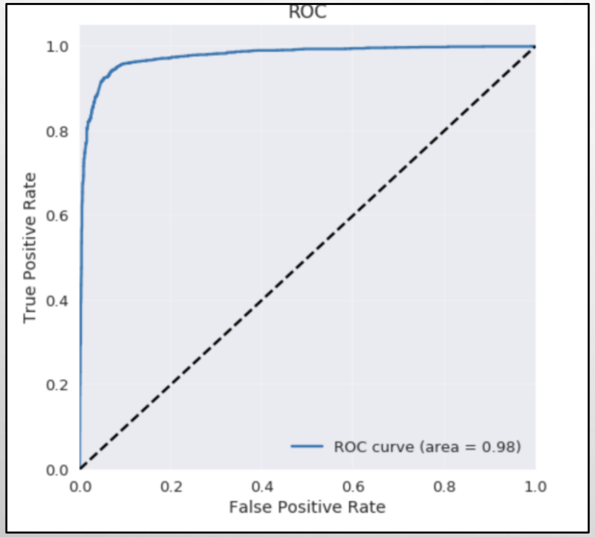


```
# Scale all the columns except dummy and binary variables using min max scaling

scaler = StandardScaler()

num_vars = ['TotalVisits', 'Total Time Spent on Website']

X_train[num_vars] = scaler.fit_transform(X_train[num_vars])
```

5. Model Building

- Firstly, we have created basic model with all the features.
- Feature selection using RFE is done where we have selected top 15 features.
- Features having high p-values > 0.05 are dropped since it doesn't carry high significance.
- In order to check multicollinearity, VIF is checked, and we found no multicollinearity. All the variables had VIF < 2

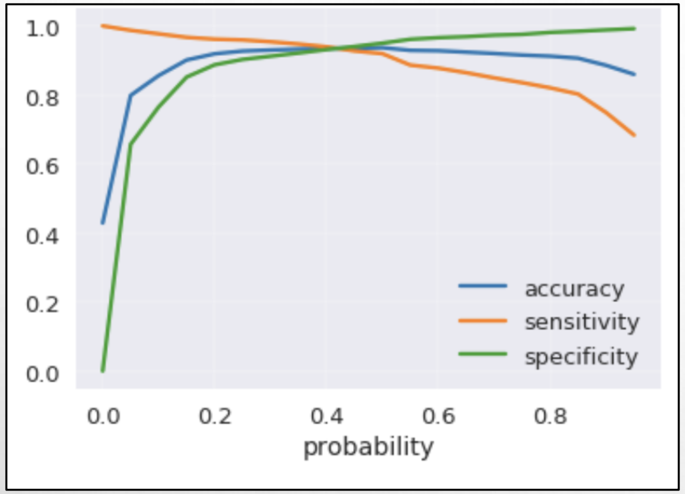
ROC Curve: We get ROC Curve Area as 0.98 suggesting the model is performing good.



- Cut-off point which is giving balanced, accuracy, sensitivity, specificity is chosen as 0.4.
- We get train & test accuracy of around 93% for both. Same is the case with sensitivity as well which is around 93%-94% for both train & test data.

5. Model Building

Finding Optimal Cut-Off Point:



- Optimal Cut-off point which is giving balanced, accuracy, sensitivity, specificity is 0.4.
- We get train & test accuracy of around 93% for both. Same is the case with sensitivity as well which is around 93%-94% for both train & test data.

6. Conclusion

- The probability for each lead, multiplied by 100, will give us the **lead score between 0 - 100**
- Columns with highest coefficient values suggests that they carry high weight on deciding the conversion into hot leads.
- Tags_Closed by Horizzon, Tags_Lost to EINS, Tags_Will revert after reading the email, Source_Welingak Website, Tags_Busy carry highest weightage as compared to other columns.
- Here we can find the columns that mattered the most in the potential buyers are:
    1. When the Tags given to the leads are
        a. Closed by Horizzon
        b. Lost to EINS
        c. Will revert after reading the email
        d. Busy
    2. When the lead source is:
        a. Welingak Website
        b. Reference
        c. Olark Chat
    3. Total time spent on the website by the customers

Recommendations:

Using the above-mentioned variables, we can determine whether the leads will be converted to potential buyers for X Education. Focus on potential leads having high converted scores. Give more effort and give them detailed information about the product, its functionalities and give information on future scenarios like provide job offerings that suits potential leads. Keep in constant touch with the promising leads, arrange meeting with ex-customer to give them more insight about the courses, the current requirement in the market and how can the course benefit them.

THANK YOU