LOGISTICS REGRESSION : LEAD SCORING CASE STUDY

TABLE OF CONTENTS

IMPORTING AND READING DATA IN DATA FRAME

- Leads.csv has the required data needed for this assignment.
- We are reading it in data frame using pandas

INSPECTING DATA

- As part of inspecting data we are using pandas to understand the provided data and inspect into general characteristics like
- - Shape
- - Data type
- - Is any data type conversion needed
- - Find the count of missing data
- There 9240 rows and 37 column in the data
- int , float and object are the data types of the available data
- there are some columns with null values
- there are some identifier columns like "Prospect ID" , "Lead Number"
- there are some categorical columns like "Lead Origin" , "Lead Source" , Converted etc for which the data type is not categorical but object

- numerical columns are TotalVisits , Total Time Spent , Page Views Per Visit on Website
- Column country has 5 records with "unknown" value
- Column Specialization , has 1942 records where customers have not selected any value and has the value as Select. Its different from having null value, but functionally it can be same.
- Column Specialization , has 1942 records where customers have not selected any value and has the value as Select. Its different from having null value, but functionally it can be same.
- "How did you hear about X Education" has 5043 records with select value
- "Update me on Supply Chain" has all records with value No
- "Get updates on DM content" has all records with value No
- "Lead Profile" has value as Select for 4146 records
- City has value as Select for 2249 records.
- Lead Source has value like "Google" and "google". We would need to convert google to Google

IMPUTING MISSING VALUES

- Below table show the columns which has missing count in the data

```
Lead Source                                     0.389610
TotalVisits                                     1.482684
Page Views Per Visit                            1.482684
Last Activity                                   1.114719
Country                                        26.634199
Specialization                                 15.562771
How did you hear about X Education             23.885281
What is your current occupation                29.112554
What matters most to you in choosing a course  29.318182
Tags                                           36.287879
Lead Quality                                   51.590909
Lead Profile                                   29.318182
City                                           15.367965
Asymmetrique Activity Index                    45.649351
Asymmetrique Profile Index                     45.649351
Asymmetrique Activity Score                    45.649351
Asymmetrique Profile Score                     45.649351
dtype: float64
```

- We have dropped any column with more than 30% missing values

- Dropped Country as City and Country depicts the similar geo location information and City has more granular data compared to Country.
- Dropped Prospect ID which is a Unique Identifier
- Imputed missing values for "TotalVisits" , "Page Views Per Visit" with their respective median.
- Imputing Last Activity: So based on Tags we are going to impute the 'Last Activity' by using the most frequent item.Steps are stipulated below:

  - Calculate the value counts of Tags, where Last Activity is null.

  - Then calculate the most common value of Last Activity where Tags is as per the value calculated in last step

  - will use this calculated value as found in last step to impute Last Activity.

- Imputing Lead Source : Mode
- Imputing Specialization : Mode
- Imputing City : Mode

- Dropping 'How did you hear about X Education' : There are 7250/9240 values which are either null or user didnot provide any value for the column. This comprises of 78% of data. Hence dropping the column
- Imputing 'What is your current occupation': 29% of data is not provided appropriately, They are either missing or User did not select any value. Instead of dropping we are imputing with. New value i.e. "Unknown". Imputing such large amount of data will introduce lot of skewness.

- Imputing "What matters most to you in choosing a course": 2709/9240 is either missing or user did not select any value. We are imputing this with value "Unknown"
- Dropping "Lead Profile" :#74% of data is missing or users have not selected any value . Dropping this column
- Dropping Tags

TREATING OUTLIERS

- There are 3 numerical columns i.e. "TotalVisits" , "Page Views Per Visit" and "Total Time Spent on Website". Lead Number can be discarded as it is just a seqential number and has not functional relation with the data.

- Analysing the scatter plot we can see that there are outliers.
- We are considering data which lie between **inter quantile** range only.

DATA PREPARATION

- For all columns where data is Yes/No in nature we are mapping Yes to 1 and No to 0
- Lead Source has values Google and google. Converting google to Google
- We are converting categorical variables into dummy indicator variables. We are using panda **get dummie**s for the same

TEST-TRAIN SPLIT

- We are spliting data into train and test in 70:30 ration

FEATURE SCALING

- We are using StandardScaler to scale the numerical data.

MODEL BUILDING

- In the 1st iteration we are using stats library for building the model for all the columns
- We are randomly selecting features using RFE for feature selection.
- In 2nd iteration we are building model with RFE selected columns.
- Dropping the column "What is your current occupation_Housewife" as the p_value is quite high i.e.0.999.
- Building the model again.In 3rd Iteration , we are dropping "Last Activity_Had a Phone Conversation" as the p_value is quite high than ..05 i.e.0.207
- Building the model 4th time.At this level all the variables has appropraite p-value
- We calculate the VIF now. Its well within the limit, we don't need to treat anything for VIF .

MODEL EVALUATION

- As part of model we are getting the probabilities of Lead getting converted or not.
- In the first pass we are choosing the cut-off arbitrarily as 0.5 and prediciting Conveted as 1 if probability is greater than 1 , and 0 if. Probability is l
- Calculating the Confusion Matrix and bsed on which we are calculating Accuracy
- 0.82 is the Accuracy with cut-off as 0.5
- Sensitivity = 0.70
- Specificity = 0.89
- Then we are considering following cut-offs and calculating the probability , accuracy , sensitivity and specificity
- After drawing the line curve between probability and accuracy , sensitivity and specificity we are getting optimal probability as 0.3

- For cut-off 0.3 ,accuracy is 0.81
- Sensitivity = 0.85
- Specificity = 0.78

MAKING PREDICTIONS ON TEST SET

- We are using the same model we build and making predictions on test data and calculating Accuracy, Specificity and Sensitivity
- Accuracy = 0.81
- Sensitivity = 0.77
- Specificity = 0.86