LEAD SCORING CASE STUDY

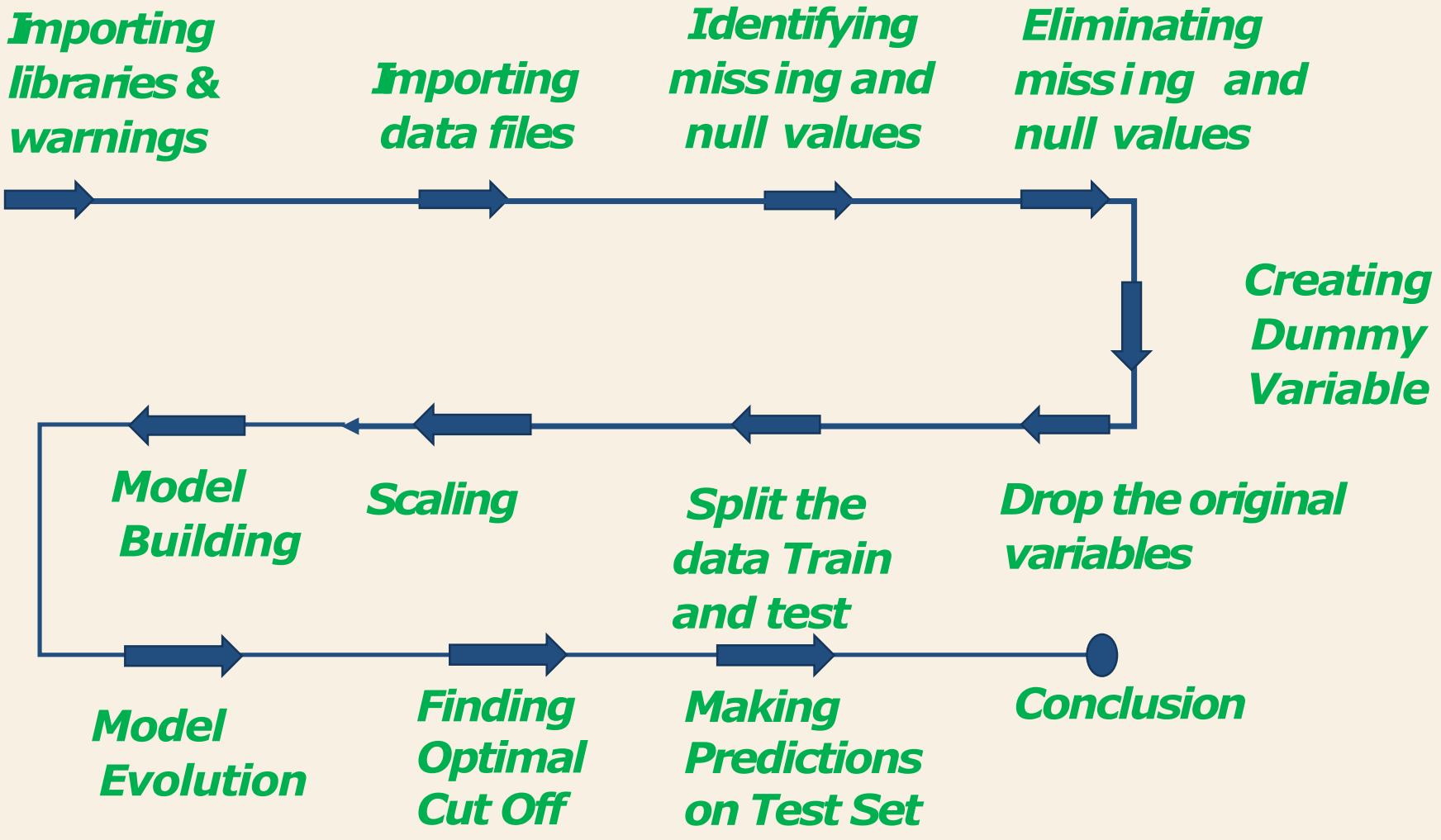
Presented by:
ABHISHEK AGRAWAL

TABLE OF CONTENTS

PROBLEM STATEMENT

- To help X education to select the most promising leads known as ‘hot leads’ who are most likely to convert into paid customers.
- Build a logistic regression model to assign a lead score between 0 and 100 to each of the leads where the leads with higher lead score have a higher conversion chance and the leads with lower lead score have a lower conversion chance.
- Identify the driver variables and understand their significance which are strong indicators of lead conversion. • Identify the outliers, if any, in the dataset and justify the same.
- Consider both technical and business aspects while building the model.
- Summarize the conversion predictions by using evaluation metrics like accuracy, sensitivity, specificity and precision.

WORK FLOW


Importing libraries & warnings
Importing data files
Identifying missing and null values
Eliminating missing and null values
Creating Dummy Variable
Drop the original variables
Split the data Train and test
Scaling
Model Building
Model Evolution
Finding Optimal Cut Off
Making Predictions on Test Set
Conclusion

IMPORTING LIBRARIES AND WARNINGS

Importing libraries
Imported pandas, numpy, matplotlib & seaborn for data loading & visualization
Imported sklearn for model building and statsmodels for model evolution.

Importing Warnings
Highlights warnings however the program runs.

READING DATASET

Datafile is extracted from the given dataset. namely 'Leads.csv’

Highlighted datafile description, shape etc., in the notebook for elaborated experience in reading the data.

DATA CLEANING AND PREPARATION

Leads.csv :

- Following columns contain more than 3000 null values initially, hence dropped those columns:
  - Tags
  - Lead Quality
  - Asymmetrique Activity Index
  - Asymmetrique Profile Index
  - Asymmetrique Activity Score
  - Asymmetrique Profile Score

DUMMY VARIABLE CREATION

- Check the columns which are of type 'object’
- Create dummy variables using the 'get_dummies' command for following columns 'Lead Origin', 'Lead Source', 'Do Not Email', 'Last Activity’, 'What is your current occupation’, 'A free copy of Mastering The Interview', 'Last Notable Activity'
- Add the results to the master dataframe
- Creating dummy variable separately for the variable 'Specialization’
- Drop the variables for which the dummy variables have been created

TEST - TRAIN SPLIT & SCALING

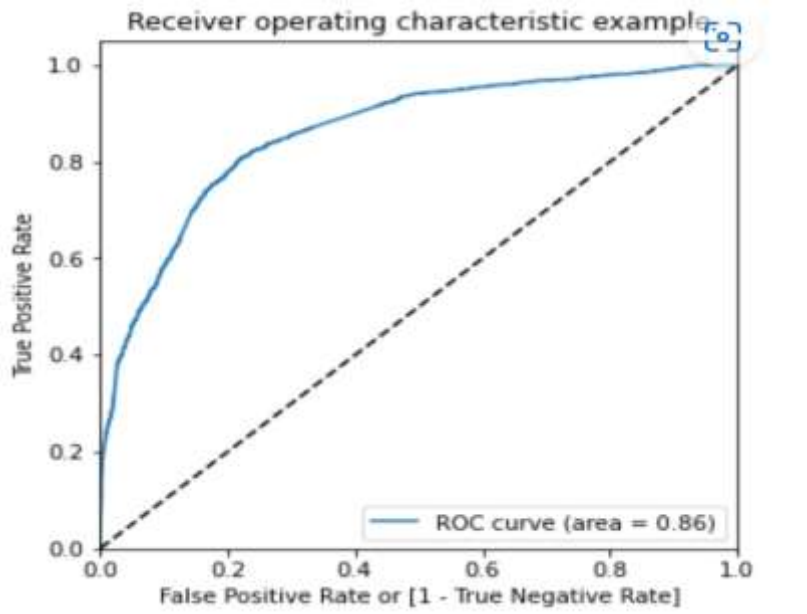
- Put all the feature variables in X
- Put the target variable in y
- Split the dataset into 70% train and 30% test
- Scale the three numeric features i.e. 'TotalVisits', 'Page Views Per Visit', 'Total Time Spent on Website' present in the dataset

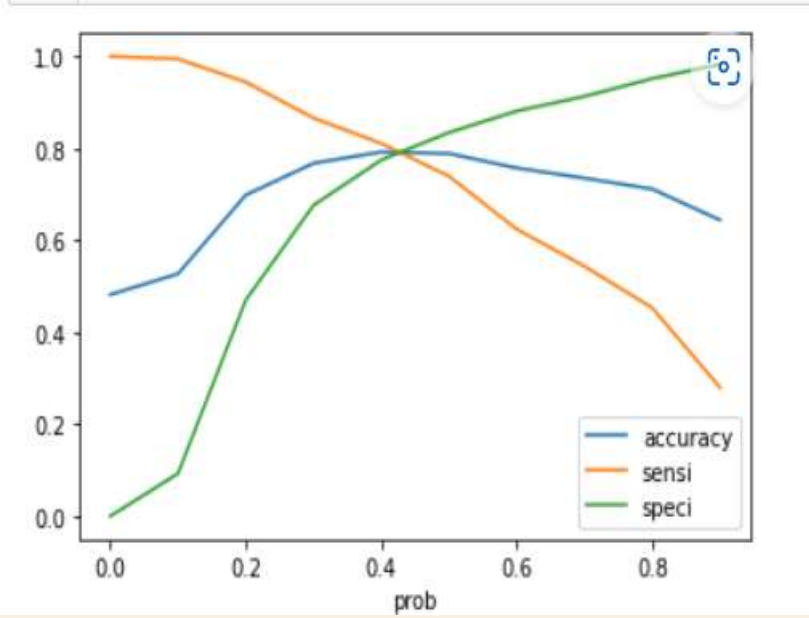
MODEL BUILDING

- Import RFE and select 15 variables
- Put all the columns selected by RFE in the variable 'col'
- Fit a logistic Regression model on X_train after adding a constant and output the summary
- Make a VIF dataframe for all the variables present
- VIFs seem to be in a decent range except for 'Lead Origin_Lead Add Form', 'Lead 'Source_Reference' and 'Lead Source_Welingak Website'. Let's first drop the variable 'Lead Source_Reference' since it has a high p-value as well as a high VIF.
- Refit the model with the new set of features
- Building Model by removing the variable whose p- value is greater than 0.05 and vif value is greater than 5%
- Predictions on test data set
- Overall accuracy 79%

	Features	VIF
2	Lead Origin_Lead Add Form	84.19
4	Lead Source_Reference	65.18
5	Lead Source_Welingak Website	20.03
11	What is your current occupation_Unemployed	3.65
7	Last Activity_Had a Phone Conversation	2.44
13	Last Notable Activity_Had a Phone Conversation	2.43
1	Total Time Spent on Website	2.38
0	TotalVisits	1.62
8	Last Activity_SMS Sent	1.59
12	What is your current occupation_Working Profes...	1.56
3	Lead Source_Olark Chat	1.44
6	Do Not Email_Yes	1.09
10	What is your current occupation_Student	1.09
9	What is your current occupation_Housewife	1.01
14	Last Notable Activity_Unreachable	1.01

ROC CURVE





- Finding Optimal Cut off Point
- Optimal cut off probability is that probability where we get balanced sensitivity and specificity.
- From the second graph it is visible that the optimal cut off is at 0.41.

CONCLUSION

It was found that the variables that mattered the most in the potential buyers are (In descending order) :

- The total time spend on the Website.
- Total number of visits.
- When the lead source was:
    i. Google
    ii. Direct traffic
    iii. Organic search
    iv. Welingak website
- When the last activity was:
    i. SMS
    ii. Olark chat conversation
- When the lead origin is Lead add format.
- When their current occupation is as a working professional. Keeping these in mind the X Education can flourish as they have a very high chance to get almost all the potential buyers to change their mind and buy their courses.

THANK YOU