LEAD SCORING CASE STUDY

LOGISTIC REGRESSION MODEL

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.
- There are a lot of leads generated in the initial stage (top) but only a few of them come out as paying customers from the bottom. In the middle stage, you need to nurture the potential leads well (i.e. educating the leads about the product, constantly communicating etc.) in order to get a higher lead conversion.
- X Education has appointed you to help them select the most promising leads, i.e. the leads that are most likely to convert into paying customers. 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 higher lead score have a higher conversion chance and the customers with lower lead score have a lower conversion chance.

APPROACH

- IMPORTING LIBRARIES
- READING AND UNDERSTANDING DATASET
- DATA CLEANING
- EXPLORATORY DATA ANALYSIS
- DUMMY VARIABLE CREATION
- DATA PREPARATION FOR MODEL BUILDING
- MODEL BUILDING
- MAKING PREDICTIONS ON THE TEST SET

IMPORTING LIBRARIES

- PANDAS
- NUMPY
- MATPLOTLIB
- SEABORN
- SKLEARN
- STATSMODEL

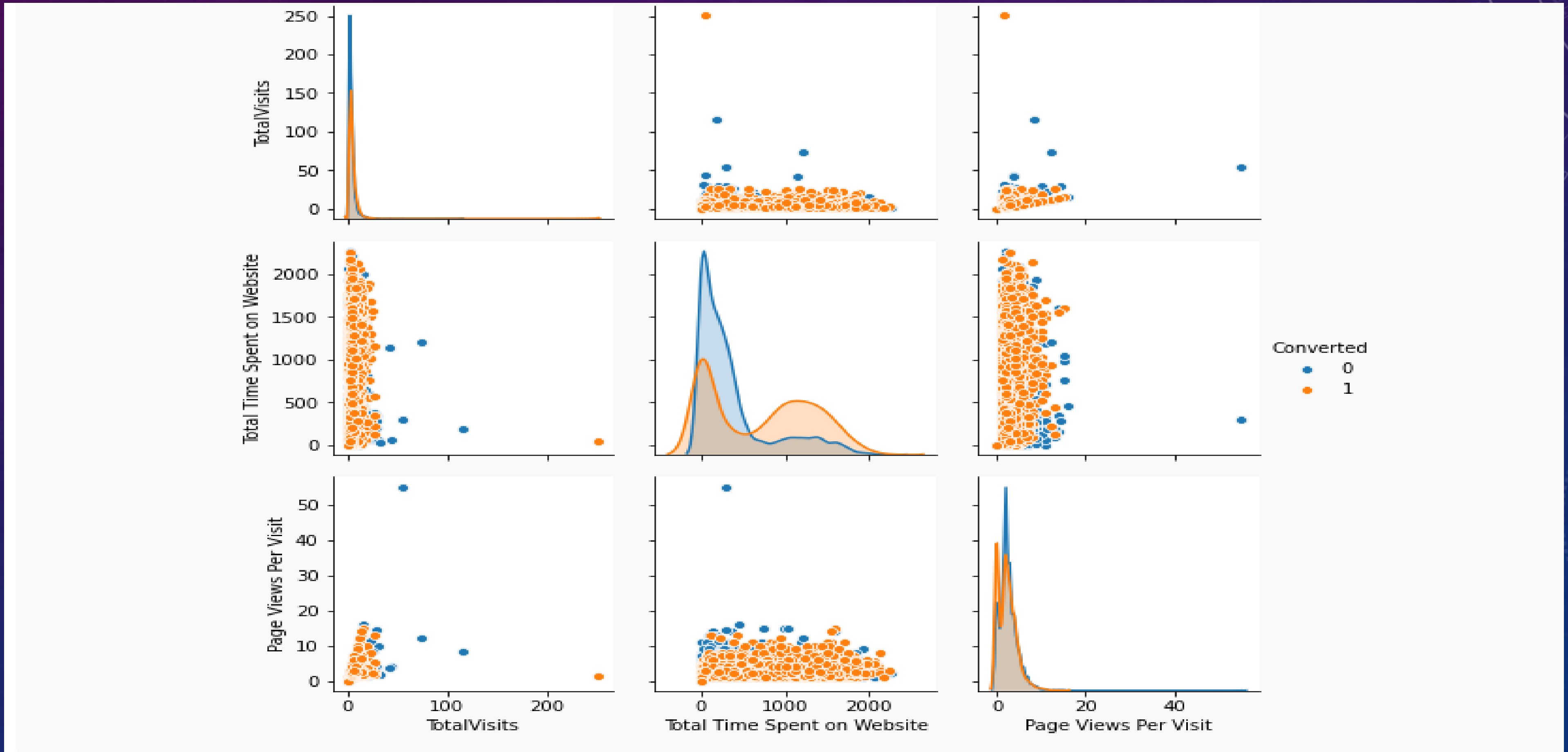
READING AND UNDERSTANDING DATASET

- First we have read the csv file in Pandas.
- Then we use certain functions to know more about the data like .shape, .dtype(), .describe(), .head(), .columns(), these function help us to understand data.
- Then we check for null values using .info() and we found that there is no null values in csv file.
- After analyzing data we found that there are 37 rows and 9240 columns in the dataset.
- Looks like there are quite a few categorical variables present in this dataset for which we will need to create dummy variables. Also, there are a lot of null values present as well, so we will need to treat them accordingly.

DATA CLEANING

- Check the null values using .isnull() command.
- We eliminate the columns having greater than 3000 missing values as they are of no use to us. Both city and country are not useful to us.
- Do Not Call, Search, Magazine, Newspaper Article, X Education Forums, Newspaper, Digital Advertisement, Through Recommendations, 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. Since practically all of the values for these variables are No, it's best that we drop these columns as they won't help with our analysis.
- ‘What Is current value’ has maximum number of missing values so we drop it as well.
- Now, clearly the variables Prospect ID and Lead Number won't be of any use in the analysis, so it's best that we drop these two variables.

PAIRPLOT



TotalVisits
Total Time Spent on Website
Page Views Per Visit
250
200
150
100
50
0
2000
1500
1000
500
0
50
40
30
20
10
0
0
100
200
TotalVisits
0
1000
2000
Total Time Spent on Website
0
20
40
Page Views Per Visit
Converted
● 0
● 1

Split the dataset into 70% train and 30% test

DUMMY VARIABLES

- The next step is to deal with the categorical variables present in the dataset. So first take a look at which variables are actually categorical variables.
- Create dummy variables using the 'get_dummies' command.
- Creating dummy variable separately for the variable 'Specialization' since it has the level 'Select' which is useless so we drop that level by specifying it explicitly.
- Drop the variables for which the dummy variables have been created 'Lead Origin', 'Lead Source', 'Do Not Email', 'Last Activity’, 'Specialization', 'What is your current occupation’, 'A free copy of Mastering The Interview', 'Last Notable Activity’.
- Then we split the dataset into training and testing sets into 70% train and 30% test.

MODEL BUILDING

- There are a few numeric variables present in the dataset which have different scales. So we scale these variables.
- We have all the variables selected by RFE and since we care about the statistics part, i.e. the p-values and the VIFs, let's use these variables to create a logistic regression model using statsmodels.
- Then we fit a logistic Regression model on X_train after adding a constant.
- There are quite a few variable which have a p-value greater than 0.05
- VIFs seem to be in a decent range except for three variables.
- Let's first drop the variable ‘Lead Source Reference’ since it has a high p-value as well as a high VIF
- Now, both the p-values and VIFs seem decent enough for all the variables. So let's go ahead and make predictions using this final set of features.

MAKING PREDICTIONS ON THE TEST SET

- We take ‘converted’ as a target variable.
- Make predictions on the test set and store it in the variable 'y_test_pred’.
- Make predictions on the test set using 0.45 as the cutoff.
- Check y_pred_final
- We create a dataframe to see the values of accuracy, sensitivity, and specificity at different values of probability cutoffs.
- We get specificity = 80% and sensitivity = 75%.

RESULT

- Final Precision we got is 76%
- Final recall we got is 78%
- You must keep a list of leads handy so that you can inform them about new courses, services, job offers and future higher studies. Monitor each lead carefully so that you can tailor the information you send to them.
- Overall accuracy we got is 78% which is really good for our model.